A General Iterative Technique for Approximate Throughput Computation of Stochastic Marked Graphs

J. Campos*, J.M. Colom*, H. Jungnitz† and M. Silva*
Departamento de Ingeniería Eléctrica e Informática
Centro Politécnico Superior, Universidad de Zaragoza
María de Luna, 3, 50015 Zaragoza, SPAIN

Abstract

A general iterative technique for approximate throughput computation of stochastic strongly connected marked graphs is presented. It generalizes a previous technique based on net decomposition through a single input-single output cut, allowing the split of the model through any cut. The approach has two basic foundations. First, a deep understanding of the qualitative behaviour of marked graphs leads to a general decomposition technique. Second, after the decomposition phase, an iterative response time approximation method is applied for the computation of the throughput. Experimental results on several examples generally have an error of less than 3%. The state space is usually reduced by more than one order of magnitude; therefore the analysis of otherwise intractable systems is possible.

1 Introduction

Stochastic Marked Graphs (SMG's) are a well-known subclass of stochastic Petri net models. They allow concurrency and synchronization but not decisions. From a queueing network perspective, it can be seen [13] that, provided strong connectivity, they are isomorphic to *fork/join queueing networks with blocking* (FJQN/B).

In this paper we consider strongly connected SMG's with time and marking independent *exponentially distributed* service times associated with transitions. For this class of models, several computation techniques have been presented in the literature. Exact performance results can be obtained from the numerical solution of the underlying continuous time Markov chain (CTMC) [3], but the *state explosion problem* makes intractable the evaluation of large systems. The efficient computation of exact performance indices of SMG's cannot be done analytically because *local balance property* does not hold in general [15]. The alternative approach of *bounds computation* has been studied by several authors using different techniques (see, e.g., [5, 8]).

Concerning approximation techniques, several proposals have been done. In [4], a method is proposed for nets that admit a *time scale decomposition* based on *near-complete decomposability* of Markov chains. Near decomposability properties are also used in [10] for an iterative approximate solution of weakly connected nets. In [6], some particular queueing networks with subnetworks having *population constraints* are analyzed using *flow equivalent aggregation* (i.e., a non-iterative technique) and Marie's method [20] (the idea is to replace a subsystem by an equivalent exponential service station with load-dependent service rates obtained by analyzing the subsystem in isolation under a load-dependent Poisson arrival process). An alternative approach is presented in [18] to compute approximate throughput for SMG's. In that work, the original system is also *split in subsystems* and a *delay equivalence* criterion is used for throughput approximation. The service rates for the aggregated subsystems are *marking dependent*. In [16], *response time approximation* is applied for an iterative computation of the throughput of SMG's. The main differences with respect to the work in [18] are two: first, the splitting of the MG is more firmly based on *qualitative theory* of MG's and second, the service rates for aggregated subsystems are *constant* (similar accuracy of the throughput is obtained with simpler and more robust algorithms).

A discussion of the above recalled techniques is presented in [17] for the throughput approximation of SMG's. We summarize now some of the conclusions. Flow equivalent aggregation is clearly the most efficient method (it is not an iterative method). In this

*This work was partially supported by the European ESPRIT BRA Project 7269 QMIPS, the Spanish PRONTIC 354/91, and the Aragonese CONAI-DGA P-IT 6/91.

†The work by this author was performed while he was visiting the University of Zaragoza

method, the behaviour of the subsystem is assumed to be independent of the arrival process and depends only on the number of customers in the system. In many cases, this assumption is violated (see [16]), therefore the method cannot be applied.

Marie's method behaves correctly in many cases. As with many iterative methods, the uniqueness of the solution cannot be proven although numerical experience has shown that a unique point does indeed exist. The main drawback is that convergence sometimes presents a problem [7].

Concerning the delay equivalence technique presented in [18], its convergence may sometimes constitute a problem. The robustness of the method is improved in [19], where the service rates of the aggregated subsystems are made constant. Some problems of this approach have been reported in [16] where it is shown that the speed of convergence strongly depends on the initial values estimated for the service rates that represent the aggregated subsystem. Moreover, for several models the authors were not able to find initial values for which the method would converge.

Finally, the response time approximation method introduced in [16] shows similar accuracy as delay equivalence, but at a greatly reduced computational cost. The method seems to be insensitive with respect to the initial values of service rates and is the one which requires the least amount of iterations. The main drawback of this method is that the original MG must be decomposed into two subsystems each one with only one input place and only one output place (*single input-single output* or *SISO cut*), and such decomposition is not always possible. A generalization to *SIMO* (single input-multiple output), *MISO* (multiple input-single output), and *MIMO* (multiple input-multiple output) cuts has been proposed but it presents serious problems concerning the quality of the results [17]. These problems are due to the fact that the structure of the net must be modified, by adding a "dummy synchronization" transition, to get a SISO cut, and this transformation can lead to a system with a considerably different behaviour.

In this paper, we follow an *iterative response time approximation technique* that avoids the problems derived from the application of the method in [16] for the general cases (SIMO, MISO, or MIMO cuts). The approach is deeply based on *qualitative theory of MG's*. More precisely, given an arbitrary cut (subset of places producing a net partition), a *structural decomposition* technique is developed in this paper that allows us to split a strongly connected MG into two *aggregated subsystems* and a *basic skeleton system*. And what is more important, *the behaviours of the subsystems, including steps, language of firing sequences and reachable markings, are equivalent to the whole system behaviour* (projected on the corresponding subsets of nodes). The better the qualitative behaviour of the system is represented by the aggregated subsystems, the more accurate the quantitative approximation will be.

The paper is organized as follows. In Section 2, basic notation and fundamental properties on MG's and implicit places are presented. Section 3 includes the structural decomposition of MG's used in the rest of the paper. The iterative technique for approximate throughput computation is described in Section 4. Section 5 includes several application examples to illustrate the introduced technique. Finally, concluding remarks are presented in Section 6.

2 Basics on stochastic marked graphs

2.1 Basic notations

We assume that the reader is familiar with concepts of P/T nets. In this section we present notations used in later sections, for further extensions the reader is referred to [21, 22].

$\mathcal{N} = (P, T, F)$ is a net if P and T are disjoint sets of places and transitions, respectively and $F \subseteq (P \times T) \cup (T \times P)$. We shall only consider nets with finite and nonempty sets of places and transitions. A net is connected if and only if the least equivalence relation which includes F is $(P \cup T) \times (P \cup T)$.

Let $\mathcal{N} = (P, T, F)$ be a net. A path of $\mathcal{N}$ is a sequence $x_1 \ldots x_k$ of elements (places and transitions) of $\mathcal{N}$ satisfying $(x_1, x_2), \ldots, (x_{k-1}, x_k) \in F$. It is a circuit if $(x_k, x_1) \in F$. A path (circuit) is called simple if all elements in the sequence defining the path (circuit) are different. In this paper we only consider simple paths and circuits. We denote by $\mathcal{P}(x, y)$, $x, y \in P \cup T$, the set of simple paths from x to y. This notion is extended to sets of elements: $\mathcal{P}(X, Y)$ is the union of the $\mathcal{P}(x, y)$ for all $x \in X$ and for all $y \in Y$.

$\mathcal{N}$ is strongly connected if for every two elements x, y of $\mathcal{N}$ there exists a path $x \ldots y$. Pre- and Post-sets of elements are denoted by the dot-notation: ${}^\bullet x = \{y | (y, x) \in F\}$ and $x^\bullet = \{y | (x, y) \in F\}$. This notion is extended to sets of elements; ${}^\bullet X$ is the union of the pre-sets of elements of X, $X^\bullet$ is the union of the post-sets of elements of X.

A function $M : P \to \{0, 1, \ldots\}$ (usually represented in vector form) is called a marking. A net system is a couple $\langle \mathcal{N}, M_0 \rangle$ of a net $\mathcal{N}$ and an initial marking M_0. A transition t is enabled at marking M if for all $p \in {}^\bullet t, M[p] > 0$. An enabled transition can be fired. The fact that M' is reached from

M by firing t is represented by $M[t\rangle M'$. A sequence of transitions $\sigma = t_1 t_2 \dots t_k$ is a firing sequence of $\langle \mathcal{N}, M_0 \rangle$ if there exist a sequence of markings such that $M_0[t_1\rangle M_1[t_2\rangle \cdots M_{k-1}[t_k\rangle M_k$, it can be written as $M_0[\sigma\rangle M_k$, and M_k is said to be reachable from M_0 by firing σ. A step at a marking M is a maximal set of transitions concurrently firable from M.

The reachability set $R(\mathcal{N}, M_0)$ is the set of all markings reachable from M_0. $L(\mathcal{N}, M_0)$ is the language of firing sequences of $\langle \mathcal{N}, M_0 \rangle$ ($L(\mathcal{N}, M_0) = \{\sigma | M_0[\sigma\rangle\}$).

A marked graph (MG) is a Petri net such that each place has exactly one input transition and exactly one output transition. MG's allow synchronization but no choice. MG's are a subclass of ordinary Petri nets for which a simple, powerful, and elegant theory allows very efficient analysis and synthesis algorithms. A summary of structure theory of MG's can be found in [21].

2.2 Implicit places and MG's

An *implicit place* never is the unique restricting the firing of its output transitions. Let $\mathcal{N}$ be any net and $\mathcal{N}^p$ be the net resulting from adding a place p to $\mathcal{N}$. If M_0 is an initial marking of $\mathcal{N}$, M_0^p denotes the initial marking of $\mathcal{N}^p$ and $m_0(p) = M_0^p[p]$. The incidence matrix of $\mathcal{N}$ is C and l_p is the incidence vector of place p.

Definition 2.1 [22] *Let $\langle \mathcal{N}, M_0 \rangle$ be a net system and $p \notin P$ be a place to be added. Then p is an implicit place (IP) with respect to $\langle \mathcal{N}, M_0 \rangle$ (or equivalently, it is an implicit place in $\langle \mathcal{N}^p, M_0^p \rangle$) iff the languages of firing sequences of $\langle \mathcal{N}, M_0 \rangle$ and $\langle \mathcal{N}^p, M_0^p \rangle$ coincide. That is, $L(\mathcal{N}, M_0) = L(\mathcal{N}^p, M_0^p)$.*

A place is an IP depending on the initial marking, M_0. Places which can be implicit for any M_0 are said to be *structurally implicit* (SIP). Inside the class of SIP's we are interested in the so called *marking structurally implicit places* (MSIP) whose structural characterization is given in the following result.

Theorem 2.1 [12] *Let $\mathcal{N}$ be a net and p be a place with incidence vector l_p. The place p is an MSIP in $\mathcal{N}^p$ iff there exists $Y \geq 0$ such that $Y^T \cdot C = l_p$.*

From this characterization of an MSIP, p, a method to compute an initial marking of p making it implicit with respect to $\langle \mathcal{N}, M_0 \rangle$ is presented in [12].

In the following, we characterize a special class of MSIP's with respect to strongly connected MG's called *TT-MSIP's*. These places have only one input arc and one output arc and therefore, $\mathcal{N}^p$ will be also an MG. The row of the incidence matrix corresponding to a TT-MSIP can be obtained from the summation of rows corresponding to the places in any path from the input transition to the output transition of the place. Moreover, we characterize the minimum initial marking making these places implicit with respect to $\langle \mathcal{N}, M_0 \rangle$ and preserving its steps.

Theorem 2.2 *Let $\mathcal{N} = (P, T, F)$ be a strongly connected MG and $p \notin P$ be a place to be added with one input transition $t_\mathrm{i} \in T$ (${}^\bullet p = \{t_\mathrm{i}\}$) and one output transition $t_\mathrm{o} \in T$ ($p^\bullet = \{t_\mathrm{o}\}$). The place p is a TT-MSIP with respect to $\mathcal{N}$ and $\forall \pi \in \mathcal{P}(t_\mathrm{i}, t_\mathrm{o})$, $l_p = \sum_{p_j \in \pi} l_{p_j}$.*

Proof: If $\mathcal{N}$ is a strongly connected MG, for all path, $\pi \in \mathcal{P}(t_\mathrm{i}, t_\mathrm{o})$, of the form $t_\mathrm{i}(= t_1) p_1 t_2 \dots t_{k-1} p_{k-1} t_k (= t_\mathrm{o})$: ${}^\bullet p_j = \{t_j\}$, $p_j{}^\bullet = \{t_{j+1}\}$, $j = 1, \dots, k-1$. Therefore, the summation of the rows in the incidence matrix corresponding to the places in π, $V = \sum_{p_j \in \pi} l_{p_j}$, verifies:
(1) $V[t] = 0$, $\forall t \notin \pi$;
(2) $V[t_\mathrm{i}] = V[t_1] = \sum_{p_j \in \pi} l_{p_j}[t_1] = \texttt{if } t_\mathrm{i} \neq t_\mathrm{o} \texttt{ then } l_{p_1}[t_1] = 1$ $\texttt{else } l_{p_1}[t_1] + l_{p_{k-1}}[t_\mathrm{o}] = 0$;
(3) $V[t_r] = \sum_{p_j \in \pi} l_{p_j}[t_r] = l_{p_{r-1}}[t_r] + l_{p_r}[t_r] = 0$, $\forall t_r \in \pi$, $r = 2 \dots (k-1)$;
(4) $V[t_\mathrm{o}] = V[t_k] = \sum_{p_j \in \pi} l_{p_j}[t_k] = \texttt{if } t_\mathrm{i} \neq t_\mathrm{o} \texttt{ then}$ $l_{p_{k-1}}[t_k] = -1 \texttt{ else } l_{p_1}[t_\mathrm{i}] + l_{p_{k-1}}[t_k] = 0$.
That is, vector V coincides with the incidence vector, l_p, of p, and according to Theorem 2.1, p is an MSIP (with $Y[p_j] = \texttt{if}$ $p_j \in \pi \texttt{ then } 1 \texttt{ else } 0$, $\forall p_j \in P$) and also a TT-MSIP. Q.E.D.

The following result characterizes the minimum initial marking of a TT-MSIP to be implicit *preserving all steps of the net system* $\langle \mathcal{N}, M_0 \rangle$. This marking is computed from the contents of tokens of the existing paths from the input transition of p to its output transition.

Theorem 2.3 *Let $\langle \mathcal{N}, M_0 \rangle$ be a strongly connected and live MG, and $p \notin P$ be a TT-MSIP to be added with ${}^\bullet p = \{t_\mathrm{i}\}$ and $p^\bullet = \{t_\mathrm{o}\}$. The minimum initial marking of p to be an IP in $\langle \mathcal{N}^p, M_0^p \rangle$ preserving all steps of $\langle \mathcal{N}, M_0 \rangle$ is*
$m_0^{\min}(p) = \min\{\sum_{p_j \in \pi} M_0[p_j] | \pi \in \mathcal{P}(t_\mathrm{i}, t_\mathrm{o})\}$.

Proof: First we prove that p is an IP with an initial marking $m_0(p) = m_0^{\min}(p)$ (i.e., $L(\mathcal{N}, M_0) = L(\mathcal{N}^p, M_0^p)$).
$L(\mathcal{N}^p, M_0^p) \subseteq L(\mathcal{N}, M_0)$. Removing place p from $\mathcal{N}^p$, we remove constraints for firing transitions. Therefore, all sequence $\sigma \in L(\mathcal{N}^p, M_0^p)$ are also firable in $\langle \mathcal{N}, M_0 \rangle$.
$L(\mathcal{N}, M_0) \subseteq L(\mathcal{N}^p, M_0^p)$. We prove this part by contradiction. Let σ be a sequence firable in $\langle \mathcal{N}, M_0 \rangle$ but not firable in $\langle \mathcal{N}^p, M_0^p \rangle$. Let σ_1 be the maximal prefix of σ firable in $\langle \mathcal{N}, M_0 \rangle$ and $\langle \mathcal{N}^p, M_0^p \rangle$: $M_0[\sigma_1\rangle M$ and $M_0^p[\sigma_1\rangle M^p$. Obviously, $M^p[p_i] = M[p_i]$ for all $p_i \in P$. The only transition preventing to finish the firing of σ after the firing of σ_1 in $\langle \mathcal{N}^p, M_0^p \rangle$ is t_o. This means that $m(p) = m_0(p) + l_p \cdot \vec{\sigma_1} = 0$. Now, we select a path $\pi \in \mathcal{P}(t_\mathrm{i}, t_\mathrm{o})$ such that $m_0(p) = \sum_{p_j \in \pi} M_0[p_j]$. Moreover, according to Theorem 2.2, $l_p = \sum_{p_j \in \pi} l_{p_j}$. Therefore,

substituting these last expressions in the above expression of $m(p)$ we obtain, $0 = m(p) = \sum_{p_j \in \pi} M_0[p_j] + \sum_{p_j \in \pi} l_{p_j} \cdot \vec{\sigma_1} = \sum_{p_j \in \pi} M[p_j]$. But this contradicts the hypothesis from which σ is firable in $\langle \mathcal{N}, M_0 \rangle$ and therefore the place $p_j \in {}^\bullet t_o$ in the path π must contain at least one token.

In order to prove that $m_0(p) = m_0^{\min}(p)$ is the minimum initial marking making p an IP preserving the steps of $\langle \mathcal{N}, M_0 \rangle$ we distinguish two cases.

Case 1 ($t_i \neq t_o$, i.e., p is self-loop free). In this case, since p is an IP for $m_0(p)$, it is step preserving [11]. We prove that $m_0(p)$ is the minimum initial marking.

First we build a sequence, σ, of maximal length in $\langle \mathcal{N}, M_0 \rangle$ firing only transitions of $T \setminus \{t_i\}$. All reached markings throughout the sequence are different, on the contrary we have a reproducible sequence without transition t_i and this is not possible in MG's. This sequence is finite because the number of different markings in a bounded net is finite. Since the sequence is maximal, we reach a marking M from which t_i is the unique firable transition (the net system is live), $M_0[\sigma\rangle M$.

In $\langle \mathcal{N}, M \rangle$ there exists a path, π', from t_i to t_o where all places contain zero tokens. In effect, the only firable transition from M is t_i, then all transitions of $T \setminus \{t_i\}$ have at least one input place with zero tokens. Therefore, t_o has an empty input place with an input transition that has another empty input place, and so on. This sequence cannot be a circuit because the MG is live and then one of the places in the sequence is an output place of t_i.

Taking into account that p is an IP with respect to $\langle \mathcal{N}, M_0 \rangle$, σ is also firable in $\langle \mathcal{N}^p, M_0^p \rangle$ and the number of tokens in p is: $m(p) = m_0(p) + l_p \cdot \vec{\sigma}$. Let π be a path of $\mathcal{P}(t_i, t_o)$ such that $m_0(p) = \sum_{p_j \in \pi} M_0[p_j]$. Moreover, according to Theorem 2.2, $l_p = \sum_{p_j \in \pi} l_{p_j} = \sum_{p_k \in \pi'} l_{p_k}$, because $\pi' \in \mathcal{P}(t_i, t_o)$, but in general $m_0(p) \leq \sum_{p_k \in \pi'} M_0[p_k]$. Considering these expressions, we can rewrite the contents of tokens of p in the following way: $m(p) = \sum_{p_j \in \pi} M_0[p_j] + \left(\sum_{p_j \in \pi} l_{p_j}\right) \cdot \vec{\sigma} \leq \sum_{p_k \in \pi'} M_0[p_k] + \left(\sum_{p_k \in \pi'} l_{p_k}\right) \cdot \vec{\sigma} = \sum_{p_k \in \pi'} M[p_k] = 0$.

Therefore, $m_0(p)$ is a minimal initial marking because there exists a firable sequence in $\langle \mathcal{N}^p, M_0^p \rangle$ that empties the place.

Case 2 ($t_i = t_o$, i.e., p is a self-loop). In this case, the minimal initial marking to make p an IP is equal to one. We prove that in order to preserve the steps of $\langle \mathcal{N}, M_0 \rangle$ we need at least the initial marking stated.

From $\langle \mathcal{N}, M_0 \rangle$ we can obtain a new net $\langle \mathcal{N}', M_0' \rangle$ by splitting the transition t_i into two transitions t and t' such that: ${}^\bullet t = {}^\bullet t_i$ and $t'^\bullet = t_i{}^\bullet$; and a new ordinary place p_i such that: ${}^\bullet p_i = \{t\}$ and $p_i{}^\bullet = \{t'\}$ and $M_0'[p_i] = 0$. Let $\mathcal{M}$ be the set of reachable markings of $\langle \mathcal{N}', M_0' \rangle$ in which the marking of place p_i is equal to zero. It is trivial to verify that the set $\mathcal{M}$ projected with respect to the set of places P coincides with the set of reachable markings of $\langle \mathcal{N}, M_0 \rangle$. Therefore, the set of steps of $\langle \mathcal{N}, M_0 \rangle$ is enclosed in the set of steps of $\langle \mathcal{N}', M_0' \rangle$ renaming the appearings of t by t_i and removing the appearings of t'.

Let us consider a place p with ${}^\bullet p = \{t'\}$ and $p^\bullet = \{t\}$ with respect to $\langle \mathcal{N}', M_0' \rangle$. Applying the previous Case 1 to place p we conclude that the minimum initial marking to make implicit place p with respect to $\langle \mathcal{N}', M_0' \rangle$ and preserving the steps of the net is equal to the minimal contents of tokens in the paths from t' to t (i.e., the circuits of the net system $\langle \mathcal{N}, M_0 \rangle$ traversing the transition t_i). Therefore, according to the previous paragraph a self-loop, p, with this initial marking preserves the steps of $\langle \mathcal{N}, M_0 \rangle$. Moreover, it is minimal because the steps requiring the maximum amount of tokens in p are the steps of $\langle \mathcal{N}, M_0 \rangle$ (they contain the output transition of p). Q.E.D.

Corollary 2.1 *Let $\langle \mathcal{N}, M_0 \rangle$ be a strongly connected and live MG, and $p \notin P$ be a TT-MSIP to be added with ${}^\bullet p = \{t_i\}$ and $p^\bullet = \{t_o\}$. The place p is an IP in $\langle \mathcal{N}^p, M_0^p \rangle$ preserving all steps of $\langle \mathcal{N}, M_0 \rangle$ for all initial marking $m_0(p) \geq m_0^{\min}(p)$.*

Proof: If we remove $m_0(p) - m_0^{\min}(p)$ tokens from p, then p is an IP in $\langle \mathcal{N}^p, M_0^p \rangle$ preserving all steps of $\langle \mathcal{N}, M_0 \rangle$ (Theorem 2.3). Therefore, all sequences and steps of $\langle \mathcal{N}, M_0 \rangle$ are firable in $\langle \mathcal{N}^p, M_0^p \rangle$. On the other hand, $m_0(p) - m_0^{\min}(p)$ tokens in p are frozen, hence the sequences and steps of $\langle \mathcal{N}^p, M_0^p \rangle$ coincide with those of $\langle \mathcal{N}, M_0 \rangle$. In effect, add the place p to a net system $\langle \mathcal{N}^{p'}, M_0^{p'} \rangle$ where p' is a place such that ${}^\bullet p' = \{t_i\}$, $p'^\bullet = \{t_o\}$ and its initial marking is $m_0^{\min}(p')$. $\langle \mathcal{N}^{p'}, M_0^{p'} \rangle$ has the same sequences and steps that $\langle \mathcal{N}, M_0 \rangle$ (Theorem 2.3), and there exists a sequence in $\langle \mathcal{N}^{p'}, M_0^{p'} \rangle$ that empties the place p'. The place p is identical to the place p', hence the minimum marking of p is reached when p' is empty (i.e. it contains $m_0(p) - m_0^{\min}(p)$ tokens). Q.E.D.

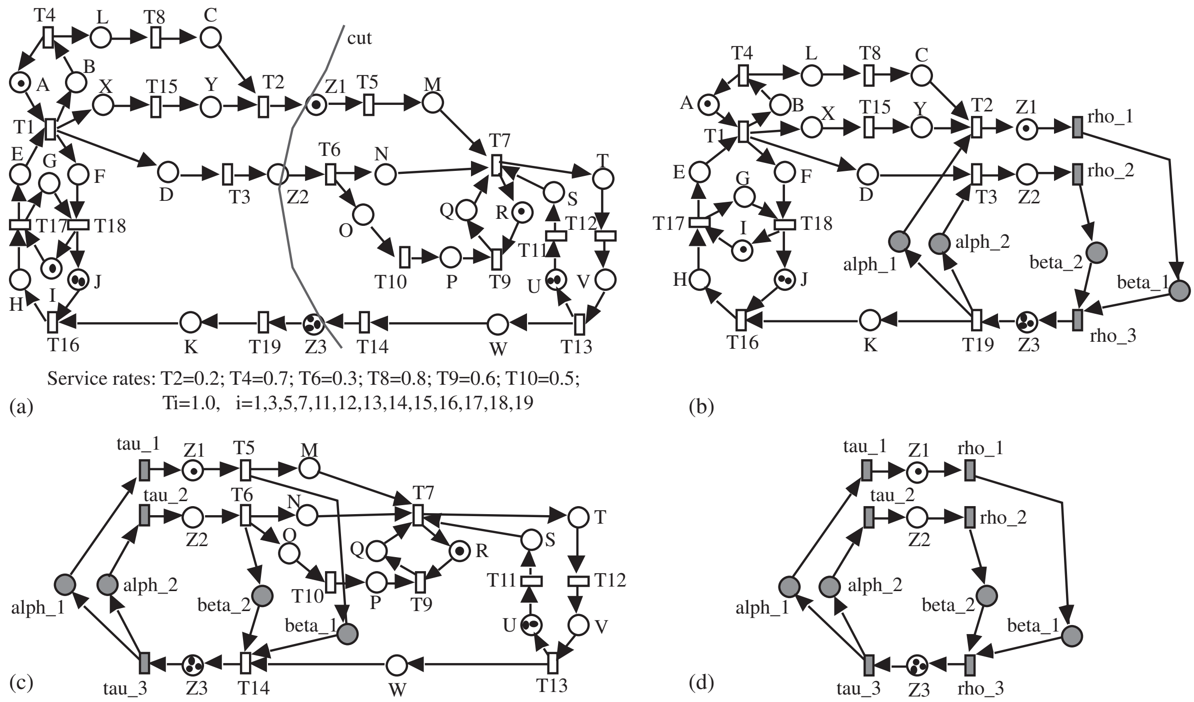
The Theorem 2.3 characterizes the minimum initial marking of a TT-MSIP to be an IP with respect to $\langle \mathcal{N}, M_0 \rangle$ in terms of the contents of tokens of the paths $\mathcal{P}(t_i, t_o)$. The computation of this minimum initial marking can be done applying an algorithm from the graph theory to determine the cost of the shortest path from a source vertex to a sink vertex of a directed graph, $G = (V, E)$, obtained from the original MG (see [2] for implementations of these algorithms). In this graph, each vertex corresponds to a transition of the net. There exists a directed arc between two vertices if and only if there exists a place in the net connecting the two transitions that represent the two vertices. The sense of the arc is the sense of the tokens' flow between the transitions through the place. Each arc has a non-negative cost equal to the initial marking of the place that represents. Moreover, we add an arc $t \to t$ for each vertex t with a cost equal to ∞.

Therefore, if we apply the algorithm to solve the shortest path problem in the directed graph G, we obtain the smallest length of any path from t_i to t_o, denoted $\text{length}(t_i, t_o)$. Observe, that $\text{length}(t_i, t_o) = \min\{\sum_{p_j \in \pi} M_0[p_j] | \pi \in \mathcal{P}(t_i, t_o)\} = m_0^{\min}(p)$.

3 Structural decomposition of MG's

The basic idea is the following: a strongly connected and live MG (see Fig. 1.a) is split into two subnets by a *cut* Q defined through some places ($Q = \{Z1, Z2, Z3\}$, in Fig. 1.a). From the cut we define three nets: two *aggregated subnets* ($\mathcal{AN}_1$ and $\mathcal{AN}_2$;

Figure 1: An SMG (a), its decomposition in aggregated systems $\mathcal{AS}_1$ (b), $\mathcal{AS}_2$ (c), and the basic skeleton (d).



see Figs. 1.b and 1.c) and a *basic skeleton* net ($\mathcal{BN}$; see Fig. 1.d). These nets will be obtained by substitution of the so called aggregable subnets, defined from the cut Q, by a set of places. We select an initial marking for each added place such that the behaviour of the aggregated subnet is the behaviour of the original MG hiding the behaviour of the aggregable subnet.

Definition 3.1 *Let $\mathcal{N} = (P, T, F)$ be a strongly connected MG. A subset of places, $Q \subseteq P$, is said to be a cut of $\mathcal{N}$ iff there exist two subnets, $\mathcal{N}_1 = (P_1, T_1, F_1)$ and $\mathcal{N}_2 = (P_2, T_2, F_2)$, of $\mathcal{N}$ verifying*
i) $T_1 \cup T_2 = T$, $T_1 \cap T_2 = \emptyset$
ii) $P_1 = T_1^\bullet \cup {}^\bullet T_1$, $P_2 = T_2^\bullet \cup {}^\bullet T_2$
iii) $P_1 \cup P_2 = P$, $P_1 \cap P_2 = Q$
iv) $F_i = F \cap ((P_i \times T_i) \cup (T_i \times P_i))$, $i \in \{1, 2\}$

Definition 3.2 *Let $\mathcal{N} = (P, T, F)$ be a strongly connected MG, $Q \subseteq P$ a cut of $\mathcal{N}$, and $\mathcal{N}_1 = (P_1, T_1, F_1)$, $\mathcal{N}_2 = (P_2, T_2, F_2)$ the two subnets associated with the cut (by Def. 3.1). The subnets $\mathcal{N}_{A_i} = (P_{A_i}, T_{A_i}, F_{A_i})$, $i \in \{1, 2\}$, are called the aggregable subnets of the cut Q, where*
i) $P_{A_i} = P_i \setminus Q$
ii) $T_{A_i} = T_i \setminus (T_Q \cap T_i)$, *where* $T_Q = {}^\bullet Q \cup Q^\bullet$
iii) $F_{A_i} = F_i \cap ((P_{A_i} \times T_{A_i}) \cup (T_{A_i} \times P_{A_i}))$
The places $p \in P_{A_i}$ such that ${}^\bullet p \cap T_{A_i} = \emptyset$ (resp., $p^\bullet \cap T_{A_i} = \emptyset$) are called source places (resp., sink places) of $\mathcal{N}_{A_i}$. The set of input transitions of the source places and output transitions of the sink places are called interface transitions of $\mathcal{N}_{A_i}$.

We denote $\mathcal{P}_{A_i}$ the set of paths in the net $\mathcal{N}_{A_i}$ from a source place to a sink place. $\mathcal{IP}_{A_i}$ denotes the set of TT-MSIP's with respect to $\mathcal{N}$ obtained from each path of $\mathcal{P}_{A_i}$ by the linear combination of the rows in the incidence matrix corresponding to the path's places. In the sequel, we define the so called *aggregated subnets* of an MG $\langle \mathcal{N}, M_0 \rangle$ with respect to a cut Q. These subnets will be obtained by substituting in $\mathcal{N}$ of an aggregable subnet $\mathcal{N}_{A_i}$ by the set of places $\mathcal{IP}_{A_i}$. This substitution is an abstraction of the subnet $\mathcal{N}_{A_i}$. We select an initial marking for each place $p \in \mathcal{IP}_{A_i}$ (called *aggregation's initial marking*, $m_0^a(p)$) equal to $m_0^a(p) = \min\{\sum_{p_j \in \pi} M_0[p_j] | l_p = \sum_{p_j \in \pi} l_{p_j}$ and $\pi \in \mathcal{P}_{A_i}\}$. With this initial marking we prove that the behaviour of the aggregated subnet is the behaviour of the original MG by hiding the behaviour of $\mathcal{N}_{A_i}$.

Definition 3.3 *Let $\langle \mathcal{N}, M_0 \rangle$ be a strongly connected and live MG, $Q \subseteq P$ a cut of $\mathcal{N}$, and $\mathcal{N}_{A_i}$, $i = 1, 2$, be the aggregable subnets defined by the cut Q. The aggregated subsystem $\mathcal{AS}_i = \langle \mathcal{AN}_i, M_0^{\mathcal{AN}_i} \rangle$ is the net system obtained from $\langle \mathcal{N}, M_0 \rangle$ by substituting the subnet $\mathcal{N}_{A_j}$ by the set of places $\mathcal{IP}_{A_j}$ with $m_0(p) = m_0^a(p)$, for all $p \in \mathcal{IP}_{A_j}$, $i = 1, 2; j = 1, 2$ and $j \neq i$. The basic skeleton system, $\mathcal{BS} = \langle \mathcal{BN}, M_0^{\mathcal{BN}} \rangle$, is the system obtained from $\langle \mathcal{N}, M_0 \rangle$ by substituting the sub-*

nets $\mathcal{N}_{A_1}$ and $\mathcal{N}_{A_2}$ by the set of places $\mathcal{IP}_{A_1}$ and $\mathcal{IP}_{A_2}$ with $m_0(p) = m_0^a(p) = \min\{\sum_{p_j\in\pi} M_0[p_j] | l_p = \sum_{p_j\in\pi} l_{p_j}$ and $\pi \in \mathcal{P}_{A_i}\}$, for all $p \in \mathcal{IP}_{A_1} \cup \mathcal{IP}_{A_2}$.

Theorem 3.1 *Let $\langle\mathcal{N}, M_0\rangle$ be a strongly connected and live MG, $Q \subseteq P$ a cut of $\mathcal{N}$ and $\mathcal{AS}_i$ be the aggregated subsystem obtained from $\langle\mathcal{N}, M_0\rangle$ by substituting the subnet $\mathcal{N}_{A_j}$ by the set of places $\mathcal{IP}_{A_j}$ with $M_0^{\mathcal{AN}_i}[p] =$ `if` $p \in \mathcal{IP}_{A_j}$ `then` $m_0^a(p)$ `else` $M_0[p]$, $i = 1, 2; j = 1, 2$, and $j \neq i$.*
i) $L(\mathcal{N}, M_0)|_{T\setminus T_{A_j}} = L(\mathcal{AN}_i, M_0^{\mathcal{AN}_i})$.
ii) $R(\mathcal{N}, M_0)|_{P\setminus P_{A_j}} = R(\mathcal{AN}_i, M_0^{\mathcal{AN}_i})|_{P_{\mathcal{AN}_i}\setminus\mathcal{IP}_{A_j}}$.

Proof: $L(\mathcal{N}, M_0)|_{T\setminus T_{A_j}} \subseteq L(\mathcal{AN}_i, M_0^{\mathcal{AN}_i})$. If we add the places of $\mathcal{IP}_{A_j}$ to $\langle\mathcal{N}, M_0\rangle$ then $L(\mathcal{N}, M_0)$ is preserved because all places of $\mathcal{IP}_{A_j}$ are IP with respect to $\langle\mathcal{N}, M_0\rangle$ preserving its steps (Corollary 2.1, taking into account that $m_0^a(p) \geq m_0^{\min}(p)$). All sequences firable in this net are also firable in the net $\mathcal{AS}_i$ after the removing of transitions in T_{A_j}. This is because in $\mathcal{AS}_i$ we have removed all firing constraints appearing in $\langle\mathcal{N}, M_0\rangle$ imposed by $\mathcal{N}_{A_j}$.

$L(\mathcal{AN}_i, M_0^{\mathcal{AN}_i}) \subseteq L(\mathcal{N}, M_0)|_{T\setminus T_{A_j}}$. We prove this part by contradiction. Let σ be a sequence of $L(\mathcal{AN}_i, M_0^{\mathcal{AN}_i})$ for which there is no $\sigma' \in L(\mathcal{N}, M_0)$ such that $\sigma = \sigma'|_{T\setminus T_{A_j}}$. Let σ_0 be the maximal prefix of σ for which there is a sequence $\sigma'_0 \in L(\mathcal{N}, M_0)$ verifying $\sigma_0 = \sigma'_0|_{T\setminus T_{A_j}}$. If $M_0[\sigma'_0\rangle M$ and $M_0^{\mathcal{AN}_i}[\sigma_0\rangle M^{\mathcal{AN}_i}$, it is trivial to verify that $M[p] = M^{\mathcal{AN}_i}[p]$ for all $p \in (P \setminus P_{A_j})$. The next transition to σ_0, t, in σ must be an output transition of a sink place of $\mathcal{N}_{A_j}$, because these transitions are the unique transitions of $\mathcal{AS}_i$ with additional constraints to fire in $\langle\mathcal{N}, M_0\rangle$. This constraints arise from $\mathcal{N}_{A_j}$ but not from the places $\mathcal{IP}_{A_j}$ because they are implicit with respect to $\langle\mathcal{N}, M_0\rangle$. All maximal firable sequences in $\langle\mathcal{N}, M\rangle$ containing only transitions of $\mathcal{N}_{A_j}$ never can enable the transition t because σ_0 is the maximal prefix of σ for which there is a sequence $\sigma'_0 \in L(\mathcal{N}, M_0)$ verifying $\sigma_0 = \sigma'_0|_{T\setminus T_{A_j}}$. Let M' be a marking reachable in $\langle\mathcal{N}, M\rangle$ firing a maximal sequence, σ_1, in $\langle\mathcal{N}, M\rangle$ containing only transitions of $\mathcal{N}_{A_j}$. At M' there exists an empty path in the $\mathcal{N}_{A_j}$ from a source place to a sink place that inputs to transition t. In effect, at M' all transitions of $\mathcal{N}_{A_j}$ are not enabled, hence have at least one empty input place. Moreover, t has at least one empty input place being a sink place of $\mathcal{N}_{A_j}$ because t is not enabled at M'. Therefore, t has an empty input place whose input transition has an empty input place, and so on, until we reach a source place of $\mathcal{N}_{A_j}$. This means that a place in $\mathcal{IP}_{A_j}$ corresponding to this path is an input place of t containing zero tokens, but this contradicts the hypothesis from which t is firable in $\mathcal{AS}_i$.

$R(\mathcal{N}, M_0)|_{P\setminus P_{A_j}} = R(\mathcal{AN}_i, M_0^{\mathcal{AN}_i})|_{P_{\mathcal{AN}_i}\setminus\mathcal{IP}_{A_j}}$. To prove this, observe that $P \setminus P_{A_j} = P_{\mathcal{AN}_i} \setminus \mathcal{IP}_{A_j}$ from the definition of $\mathcal{AN}_i$. Taking into account the part (i) of this theorem, the stated equality of markings' sets holds. Q.E.D.

Corollary 3.1 *Let $\langle\mathcal{N}, M_0\rangle$ be a strongly connected and live MG, $Q \subseteq P$ a cut of $\mathcal{N}$, and $\mathcal{BS}$ the basic skeleton system obtained from $\langle\mathcal{N}, M_0\rangle$ by substituting the subnets $\mathcal{N}_{A_1}$ and $\mathcal{N}_{A_2}$ by the set of places $\mathcal{IP}_{A_1}$ and $\mathcal{IP}_{A_2}$, respectively, and $M_0^{\mathcal{BN}}[p] =$ `if` $p \in \mathcal{IP}_{A_1} \cup \mathcal{IP}_{A_2}$ `then` $m_0^a(p)$ `else` $M_0[p]$.*
i) $L(\mathcal{N}, M_0)|_{T\setminus(T_{A_1}\cup T_{A_2})} = L(\mathcal{BN}, M_0^{\mathcal{BN}})$.
ii) $R(\mathcal{N}, M_0)|_{P\setminus(P_{A_1}\cup P_{A_2})} = R(\mathcal{BN}, M_0^{\mathcal{BS}})|_{P_{\mathcal{BN}}\setminus(\mathcal{IP}_{A_1}\cup\mathcal{IP}_{A_2})}$.

Proof: The proof of the corollary can be decomposed into two steps: (1) The proof of the behaviour equivalence between $\langle\mathcal{N}, M_0\rangle$ and $\mathcal{AS}_1$ (that is, the previous theorem); (2) The proof of the behaviour equivalence between $\mathcal{AS}_1$ and $\mathcal{BS}$. Taking into account that $\mathcal{AS}_i$ is a strongly connected and live MG, the proof of this second part is the same that the above theorem renaming, for example, $\mathcal{AS}_1$ as $\langle\mathcal{N}, M_0\rangle$ and $\mathcal{BS}$ as $\mathcal{AS}_2$. Q.E.D.

The main drawback of the above theorems concerns the great number (exponential in the worst case) of places in $\mathcal{IP}_{A_i}$. In the following we present a method to reduce the number of places to add, characterizing a subset of $\mathcal{IP}_{A_i}$, denoted $\mathcal{BIP}_{A_i}$, with the property that all places of $\mathcal{IP}_{A_i} \setminus \mathcal{BIP}_{A_i}$ are implicit with respect to the places $\mathcal{BIP}_{A_i}$. Therefore, in order to build the aggregated subnet we only add the set of places $\mathcal{BIP}_{A_i}$ instead of $\mathcal{IP}_{A_i}$.

Let us consider the aggregable subnet $\mathcal{N}_{A_i}$ together with its interface transitions. We derive from this net a directed graph $G_{A_i} = (V, E)$ in the same way to that presented at the end of previous section.

If we apply the algorithm of R.W. Floyd to solve the *all-pairs shortest paths* problem (see [2] for implementations of this algorithm) to the directed graph G_{A_i}, we obtain for each ordered pair of vertices (i.e., transitions) (t, t') the smallest length of any path from t to t', denoted $\text{length}(t, t')$ (if this value is equal to ∞, there is no path from t to t'). Observe, that $\text{length}(t, t') = \min\{\sum_{p_j\in\pi} M_0[p_j] | \pi$ is a path from t to $t'\}$. The computational complexity of this algorithm is $O(m^3)$, where m is the number of transitions of the considered net. From this values we define the set of places $\mathcal{BIP}_{A_i}$ as $\mathcal{BIP}_{A_i} = \{p | {}^\bullet p = \{t\}; p^\bullet = \{t'\}; t, t' \in T_Q; \text{length}(t, t') \neq \infty\}$.

For all $p \in \mathcal{BIP}_{A_i}$ we select an initial marking $m_0(p) = \text{length}(t, t')$. It is trivial to verify that this initial marking coincides with the previously defined *aggregation's initial marking*, $m_0^a(p)$. For instance, in the case of Fig. 1.b, $\mathcal{BIP}_{A_2} = \{$ *beta_1*, *beta_2* $\}$ and $m_0(\textit{beta\_1}) = m_0(\textit{beta\_2}) = 0$.

The following result states that all places of $\mathcal{IP}_{A_i} \setminus \mathcal{BIP}_{A_i}$ are implicit with respect to the places $\mathcal{BIP}_{A_i}$. Therefore, in order to build the aggregated subsystem we only add the set of places $\mathcal{BIP}_{A_i}$ instead of $\mathcal{IP}_{A_i}$.

Property 3.1 *Each place $p \in \mathcal{IP}_{A_i} \setminus \mathcal{BIP}_{A_i}$ with an initial marking equal to $m_0^a(p)$ is implicit with respect* *to the set of places $\mathcal{BIP}_{A_i}$ each one with an initial marking equal to the aggregation's marking.*

Proof: Let $p \in \mathcal{IP}_{A_i} \setminus \mathcal{BIP}_{A_i}$ be a place obtained from the summation of the rows in the incidence matrix corresponding to the places of a path. Let $t, t' \in T_Q$ be the interface transitions of this path. Because of the existence of this path, after the application of the Floyd's algorithm we have $\text{length}(t, t') \neq \infty$, therefore there exists an identical place in $\mathcal{BIP}_{A_i}$ with the same initial marking. Q.E.D.

In many cases the set $\mathcal{BIP}_{A_i}$ is bigger than necessary because some places can be implicit in $\mathcal{AS}_i$. In order to remove one of these unnecessary places, p, we can apply the method described at the end of the previous section to compute the shortest path from ${}^\bullet p$ to $p^\bullet$. The place p can be removed if the output of this algorithm is less than or equal to the aggregation's marking of p. Observe that in the case of Fig. 2.b, the set $\mathcal{BIP}_{A_2}$ contains 16 places but a further removing of places leads to a minimum set of 6 places, named β_i, $i = 1, \ldots, 6$ in the figure.

4 Iterative technique for approximate throughput computation

In previous section, an algorithm to decompose an MG into two aggregated subsystems and a basic skeleton system (being also MG's) has been presented. In aggregated subsystem $\mathcal{AS}_i$ $(i = 1, 2)$, the subnet $\mathcal{N}_j$ $(j \neq i)$ is represented by the places in the cut Q, by the interface transitions of $\mathcal{N}_j$, $T_{I_j} = T_Q \cap T_j$, and by the new places that substitute the subnet $\mathcal{N}_{A_j}$.

The technique for an approximate computation of the throughput that we present now is, basically, a *response time approximation* method [1, 16, 17]. The interface transitions of $\mathcal{N}_j$ in $\mathcal{AS}_i$ approximate the response time of all the subsystem $\mathcal{N}_j$ $(i = 1, 2; j \neq i)$. A direct (non-iterative) method to compute the constant service rates of such interface transitions in order to represent the aggregation of the subnet gives, in general, low accuracy. Therefore, we are forced to define a *fixed-point search iterative process*, with the possible drawback of the presence of convergence and efficiency problems.

4.1 First approach: Ping-Pong algorithm

The first algorithm that we explored, called "Ping-Pong", follows.

select a cut Q;
derive aggregated subsystems $\mathcal{AS}_i, i = 1, 2$;
give value μ_t^0 for each $t \in T_{I_1}$ in $\mathcal{AS}_2$;
compute value of throughput χ_2^0 of $\mathcal{AS}_2$;
$k := 0$; {counter for iteration steps}
repeat
  $k := k + 1$;
  compute μ_t^k for each $t \in T_{I_2}$ such that the throughput
    χ_1^k of $\mathcal{AS}_1$ is close enough to χ_2^{k-1};
  compute μ_t^k for each $t \in T_{I_1}$ such that the throughput
    χ_2^k of $\mathcal{AS}_2$ is close enough to χ_1^k;
until convergence of χ_1^k and χ_2^k;

In the above procedure, once a cut has been selected and given some initial values for the service rates of interface transitions of $\mathcal{N}_1$ (which approximate the response time of all the subsystem $\mathcal{N}_1$), the underlying CTMC of aggregated subsystem $\mathcal{AS}_2$ is solved. From the solution of that CTMC, the first estimation χ_2^0 of the throughput of $\mathcal{AS}_2$ can be computed. Then, the initial estimated values of service rates of interface transitions that approximate the response time of subsystem $\mathcal{N}_2$ must be derived. This must be done in such a way that the throughput χ_1^1 of $\mathcal{AS}_1$ is "close enough" to χ_2^0. Then, a better estimation of rates μ_t^k for each $t \in T_{I_1}$ must be computed such that the throughput χ_2^1 of $\mathcal{AS}_2$ is close enough to χ_1^1. The process is iterated until χ_i^{k-1} and χ_i^k are "close enough".

The first problem of the above sketch of approximation algorithm is that a *multidimensional search on the parameters* of a complex CTMC in order to get a given throughput cannot be done in an efficient way. A possible solution to this problem is the following. In the iterative process, each time that an aggregated subsystem $\mathcal{AS}_i, i = 1, 2$, is solved, *the ratios* among the service rates μ_t^k of all the transitions in T_{I_i} are estimated. After that, when the other subsystem $\mathcal{AS}_j, j \neq i$, is solved, only a *scale factor* for these service rates must be computed. The goal is to find a scale factor of μ_t^k for all $t \in T_{I_j}$ (and fixed k) such that the throughput of $\mathcal{AS}_j$ and the throughput of $\mathcal{AS}_i$, computed before, are the same. And this can be achieved with a linear search of the scale factor in $\mathcal{AS}_j$.

At this point, the main technical problem is the following: How to estimate from the solution of $\mathcal{AS}_i$ the ratios among the service rates of all transitions in T_{I_i} that in the next step (solution of $\mathcal{AS}_j$) will be scaled to obtain an approximation of the response time of the subsystem $\mathcal{N}_i$?

We explain our answer to this question by means of the example depicted in Fig. 1. Figure 1.b represents the aggregated subsystem $\mathcal{AS}_1$ derived from the original MG. It is necessary to compute the ratio between the service rate of $T2$ and $T3$ to be used as input data for the linear search of the scale factor in $\mathcal{AS}_2$ (Fig. 1.c). In order to do that, the aggregated subsystem $\mathcal{AS}_1$ is transformed (as depicted in Fig. 1.b) with the addition of places $\mathcal{BIP}_{A_1} = \{\textit{alph\_1}, \textit{alph\_2}\}$. The obtained system is behaviourally equivalent to $\mathcal{AS}_1$

because the added places (which are those that will substitute $\mathcal{N}_{A_1}$), are implicit. These new places allow to estimate the ratio between the "aggregated service times" of transitions $T2$ and $T3$ (representing the response time approximation of $\mathcal{N}_1$), as the quotient of the mean marking of *alph_1* by the mean marking of *alph_2*, because the throughput of all transitions is the same.

Now, two problems arise. First, the linear search of the scale factor must be done in the aggregated subsystems, that can have a considerably large state space, thus the efficiency of the method falls down. Additionally, we have found convergence problems in many cases. A solution for both problems is proposed in the next subsection.

4.2 A solution: Pelota[1] algorithm

The more practical solution of the problem we found makes use of the third system (another MG) derived from the original one, in previous section: *the basic skeleton*. The basic skeleton contains the interface subsystem and a simplified view (using the places $\mathcal{BIP}_{A_i}, i = 1, 2$, computed by the algorithm in previous section) of subsystems $\mathcal{N}_{A_i}, i = 1, 2$.

The idea is to use the basic skeleton as an intermediate point (*fronton*) between the two aggregated subnets (rackets), as explained in this algorithm:

select a cut Q;
derive $\mathcal{AS}_i, i = 1, 2$ and $\mathcal{BS}$;
give initial value μ_t^0 for each $t \in T_{I_2}$;
$k := 0$; {counter for iteration steps}
repeat
  $k := k + 1$;
  solve aggregated subsystem $\mathcal{AS}_1$ with
    input: μ_t^{k-1} for each $t \in T_{I_2}$,
    output: ratios among μ_t^k of $t \in T_{I_1}$ and χ_1^k;
  solve basic skeleton system $\mathcal{BS}$ with
    input: μ_t^{k-1} for each $t \in T_{I_2}$,
        ratios among μ_t^k of $t \in T_{I_1}$, and χ_1^k,
    output: scale factor of μ_t^k of $t \in T_{I_1}$;
  solve aggregated subsystem $\mathcal{AS}_2$ with
    input: μ_t^k for each $t \in T_{I_1}$,
    output: ratios among μ_t^k of $t \in T_{I_2}$ and χ_2^k;
  solve basic skeleton system $\mathcal{BS}$ with
    input: μ_t^k for each $t \in T_{I_1}$,
        ratios among μ_t^k of $t \in T_{I_2}$, and χ_2^k,
    output: scale factor of μ_t^k of $t \in T_{I_2}$;
until convergence of χ_1^k and χ_2^k;

In this iterative process, each time that an aggregated subsystem $\mathcal{AS}_i, i = 1, 2$, is solved, only the throughput χ_i^k and the ratios among the service rates μ_t^k of all the transitions in T_{I_i} are estimated (with the method explained in previous subsection). After that, a scale factor for these service rates must be computed. This is achieved by using the basic skeleton system $\mathcal{BS}$. The goal is to find a scale factor of μ_t^k for all $t \in T_{I_i}$ such that the throughput of the basic skeleton and the throughput of $\mathcal{AS}_i$, computed before, are the same. A linear search of the scale factor must be implemented, but now in a net system with considerably fewer states (the basic skeleton). In each iteration of this linear search, the basic skeleton is solved by deriving the underlying CTMC.

Now, the existence and uniqueness of the solution, and the convergence of the method should be addressed. Although no formal proof gives positive answers so far to the above questions, extensive testing allows the conjecture that there exists one and only one solution, computable in a finite number of steps, typically between 2 and 5 if the convergence criterion is that the difference between the two last estimations of the throughput is less than 0.1 %.

5 Examples

In this section we present several numerical results of the application of the iterative technique previously introduced. Among all the tested examples, we have selected two different Petri net structures because of their following characteristics: the first one (already introduced in Fig. 1) is structurally asymmetric while the second has symmetries; for the second one, the effect of timing asymmetries on the iterative algorithm can be studied by changing the service rates of transitions (preserving the strong structural symmetry). In all cases, the obtained approximations are compared with exact values obtained from the numerical solution of the underlying CTMC (*GreatSPN* package was used [9]).

Let us consider again the SMG depicted in Fig. 1.a. The exact value of the throughput is equal to 0.138341 (if single-server semantics is assumed). The underlying CTMC has 89358 states. The aggregated systems $\mathcal{AS}_1$ and $\mathcal{AS}_2$ are depicted in Figs. 1.b and 1.c, respectively. The corresponding basic skeleton system is that in Fig. 1.d.

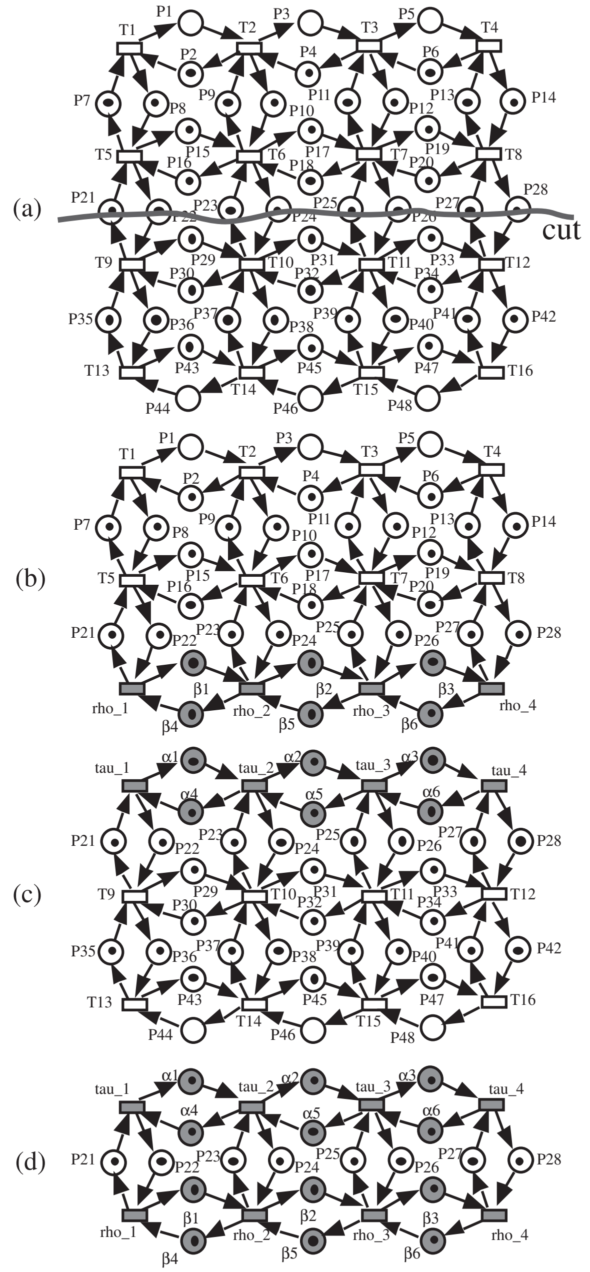
Table 1 shows the iterative results obtained for this example. The values in $\mathcal{AS}_1$ columns have been obtained from the solution of the aggregated system in Fig. 1.b: χ_1 is the throughput of $\mathcal{AS}_1$; columns *tau_1*, *tau_2*, and *tau_3* are the estimated values of the service rates of the aggregated transition tau_1, tau_2, and tau_3, computed in $\mathcal{AS}_1$; column *coeff* is the scale factor of previous estimated service rates, obtained by the

[1] Game played by two players who use a basket strapped to their wrists or a wooden racket to propel a ball against a specially marked wall, called *fronton*.

Table 1: Iteration results for the SMG in Fig. 1.

$\mathcal{AS}_1$					$\mathcal{AS}_2$				
χ_1	tau_1	tau_2	tau_3	coeff	χ_2	rho_1	rho_2	rho_3	coeff
0.17352	0.05170	0.16810	0.88873	1.01167	0.12714	0.89026	0.21861	0.14354	0.98468
0.14093	0.06265	0.19707	0.91895	1.01318	0.13795	0.88267	0.21363	0.13509	0.98582
0.13856	0.06325	0.19821	0.92054	1.01306	0.13841	0.88239	0.21343	0.13467	0.98592
0.13844	0.06328	0.19827	0.92062	1.01306	0.13843	0.88237	0.21342	0.13465	0.98592
0.13843	0.06328	0.19827	0.92064	1.01307	0.13843	0.88238	0.21342	0.13465	0.98593

Figure 2: A second example of SMG and its decomposition.



linear search in the basic skeleton of Fig. 1.d. Columns related with $\mathcal{AS}_2$ represent the analogous values for the aggregated system in Fig. 1.c. Convergence of the method can be observed from the third iteration step. The error is -0.064333 %, after the fifth step. The following additional fact must be remarked: the underlying CTMC's of $\mathcal{AS}_1$, $\mathcal{AS}_2$, and the basic skeleton have 8288, 3440, and 231 states, respectively, while the original SMG has 89358 states.

As a second example, let us consider the SMG depicted in Fig. 2.a. Any splitting of the net will generate two strongly coupled aggregated subnets. We select the following cut: $Q = \{P21, P22, P23, P24, P25, P26, P27, P28\}$. The corresponding aggregated systems are depicted in Figs. 2.b and 2.c. The basic skeleton is that in Fig. 2.d. The CTMC underlying the original SMG has 49398, while those underlying the aggregated systems have 6748. The basic skeleton has 771 reachable states.

We consider three different situations arising from different transition service rates (we assume infinite-server semantics in all cases). In the first case, we suppose that the service rates of all transitions are equal to 1.0. In this case, the exact throughput of the SMG is 0.295945.

Table 2 shows the iteration results for three different selections of initial values of aggregated service rates of transitions *rho_1*, *rho_2*, and *rho_3*. It can be seen that in all cases convergence occurs at the third iteration step, independently of the initial values given to the aggregated service rates. This fact illustrates the robustness of the method with respect to the seed. The error of the approximation in all cases is 0.4 %.

As a second case, consider again the SMG in Fig. 2.a but now with asymmetric service rates associated with transitions. Assume that the service rates of transitions $T1$ to $T8$ are all equal to 1.0, while service rates of transition $T9$ to $T16$ are equal to 2.0. In this case the exact throughput of the original system is 0.333356. The iteration results are shown in Table 3. Now, the initial values of aggregated service rates of transitions *rho_1*, *rho_2*, and *rho_3* are equal to 1.0. Convergence can be observed from the second itera-

Table 2: Iteration results for the SMG in Fig. 2 with all service rates of transition equal to 1.0.

Initial values of service rates of rho_1, rho_2, and rho_3 equal to 0.1											
$\mathcal{AS}_1$						$\mathcal{AS}_2$					
χ_1	tau_1	tau_2	tau_3	tau_4	coeff	χ_2	rho_1	rho_2	rho_3	rho_4	coeff
0.07930	1.02121	1.02452	1.01112	0.80930	1.06357	0.33294	0.29834	0.50973	0.61599	0.71668	1.03611
0.29244	0.84574	0.72462	0.55755	0.30802	1.05833	0.30079	0.29864	0.54035	0.70609	0.83610	1.06250
0.29710	0.84301	0.71383	0.54364	0.29813	1.06382	0.29733	0.29758	0.54270	0.71310	0.84299	1.06427
0.29711	0.84340	0.71354	0.54286	0.29751	1.06436	0.29711	0.29747	0.54281	0.71352	0.84343	1.06440
Initial values of service rates of rho_1, rho_2, and rho_3 equal to 1.0											
$\mathcal{AS}_1$						$\mathcal{AS}_2$					
χ_1	tau_1	tau_2	tau_3	tau_4	coeff	χ_2	rho_1	rho_2	rho_3	rho_4	coeff
0.33318	0.70982	0.61546	0.51044	0.29917	1.03518	0.29265	0.30871	0.55771	0.72423	0.84521	1.05804
0.30095	0.83571	0.70581	0.54034	0.29877	1.06233	0.29712	0.29817	0.54366	0.71378	0.84293	1.06378
0.29734	0.84296	0.71307	0.54270	0.29759	1.06425	0.29712	0.29751	0.54286	0.71354	0.84339	1.06436
0.29712	0.84343	0.71352	0.54281	0.29747	1.06440	0.29710	0.29746	0.54282	0.71354	0.84345	1.06439
Initial values of service rates of rho_1, rho_2, and rho_3 equal to 10.0											
$\mathcal{AS}_1$						$\mathcal{AS}_2$					
χ_1	tau_1	tau_2	tau_3	tau_4	coeff	χ_2	rho_1	rho_2	rho_3	rho_4	coeff
0.33419	0.68611	0.59756	0.49474	0.28053	1.03311	0.28561	0.30812	0.56325	0.73687	0.85741	1.06091
0.30136	0.83550	0.70455	0.53890	0.29791	1.06293	0.29679	0.29807	0.54392	0.71447	0.84356	1.06392
0.29735	0.84299	0.71304	0.54263	0.29753	1.06430	0.29710	0.29750	0.54287	0.71358	0.84343	1.06437
0.29711	0.84343	0.71352	0.54281	0.29747	1.06440	0.29710	0.29746	0.54282	0.71355	0.84346	1.06441
0.29710	0.84346	0.71355	0.54282	0.29746	1.06441	0.29710	0.29745	0.54281	0.71355	0.84346	1.06441
0.29710	0.84346	0.71356	0.54282	0.29746	1.06441	0.29710	0.29746	0.54282	0.71355	0.84346	1.06441

Table 3: Iteration results for the SMG in Fig. 2 with service rates of transition $T1$ to $T8$ equal to 1.0 and of transitions $T9$ to $T16$ equal to 2.0.

$\mathcal{AS}_1$						$\mathcal{AS}_2$					
χ_1	tau_1	tau_2	tau_3	tau_4	coeff	χ_2	rho_1	rho_2	rho_3	rho_4	coeff
0.33318	0.70983	0.61546	0.51045	0.29917	1.03519	0.34424	0.70118	1.49390	1.84123	1.92737	1.06187
0.33352	0.71500	0.60522	0.49835	0.28554	1.03637	0.33345	0.68342	1.50320	1.85362	1.93598	1.06255
0.33345	0.71616	0.60538	0.49834	0.28550	1.03656	0.33345	0.68281	1.50288	1.85352	1.93592	1.06251
0.33345	0.71621	0.60539	0.49834	0.28550	1.03656	0.33345	0.68278	1.50284	1.85348	1.93588	1.06249

tion step and the error of the obtained value is 0.02 %.

Finally, consider once more the SMG of Fig. 2.a, but now with the following service rates associated with transitions: the rates of $T1$, $T2$, $T5$, $T6$, $T9$, $T10$, $T13$, and $T14$ are equal to 2.0, while the rest are equal to 1.0. In this case, the exact throughput is 0.362586. The iteration results are shown in Table 4. Again, convergence can be observed from the second iteration step. The error is now 0.19 %.

6 Conclusions

In order to derive a general, efficient, and accurate technique for throughput approximation of stochastic marked graphs using the divide and conquer principle, qualitative theory ought to guide the decomposition phase (this principle underlies several previous works on the topic [6, 14, 16, 17, 18, 19]).

This was the first objective of the paper: the presentation of a general structural decomposition technique allowing to split a given marked graph through an arbitrary cut (subset of places) and to derive two aggregated subsystems whose qualitative behaviours are projections of the whole system qualitative behaviour. The technical tool to achieve this problem has been the use of implicit places: a subsystem of the original marked graph can be substituted by a minimal set of implicit places that represent an abstraction of the subsystem, leading to an aggregated subsystem. If the same process is applied to two complementary subsystems, two aggregated subsystems are derived, each one representing a portion of the behaviour of the whole system.

The second phase of the analysis problem is the selection of an approximate throughput computation algorithm. Iterative response time approximation technique was selected after a wide comparison with other approaches present in the literature. In order to assure the convergence of the method, a third subsystem was used for a correct tuning of parameters, the basic skeleton. It is obtained after the substitution of both subnets by the corresponding implicit places. Its behaviour is simple enough to allow a linear search of the correct value of a parameter in order to get a given throughput (the one obtained in previous iteration step).

Extensive numerical experiments using the method sketched in previous paragraphs showed very good results with respect to efficiency and accuracy. Convergence is generally observed after a couple of iteration

Table 4: Iteration results for the SMG in Fig. 2 with service rates of transition $T1$, $T2$, $T5$, $T6$, $T9$, $T10$, $T13$, and $T14$ equal to 2.0, and the rest equal to 1.0.

$\mathcal{AS}_1$						$\mathcal{AS}_2$					
χ_1	tau_1	tau_2	tau_3	tau_4	coeff	χ_2	rho_1	rho_2	rho_3	rho_4	coeff
0.40526	1.64486	1.58029	0.60759	0.36042	1.00568	0.35214	0.36948	0.69530	0.61363	0.80667	1.07872
0.36392	1.81297	1.72253	0.66348	0.38291	1.01927	0.36239	0.37446	0.68764	0.59809	0.79673	1.08484
0.36326	1.80988	1.72268	0.66584	0.38570	1.01711	0.36321	0.37514	0.68748	0.59702	0.79565	1.08508
0.36328	1.80942	1.72245	0.66596	0.38596	1.01688	0.36328	0.37520	0.68748	0.59694	0.79556	1.08510
0.36329	1.80938	1.72243	0.66598	0.38599	1.01686	0.36329	0.37521	0.68747	0.59693	0.79555	1.08510
0.36329	1.80938	1.72243	0.66598	0.38599	1.01686	0.36329	0.37521	0.68748	0.59694	0.79555	1.08510

steps and the approximate computation of throughput can be achieved with a considerable saving of time and memory (more than one order of magnitude) and with a very little error (less than 3 %).

Eventhough we have considered only strongly connected nets, the approach can be applied to non-strongly connected marked graphs: The iterative technique is used to compute the approximate throughput of each strongly connected component in isolation and, after that, [8, Theorem 5.1] applies.

An obvious generalization of the presented technique can be derived if the original system is partitioned into more than two subsystems, leading to the classical tradeoff between efficiency and accuracy. The extension to more general net subclasses, like macroplace-macrotransition nets proposed in [14], is being considered by the authors.

References

[1] S. C. Agrawal, J. P. Buzen, and A. W. Shum. Response time preservation: A general technique for developing approximate algorithms for queueing networks. In *Proc. of the 1984 ACM Sigmetrics Conf. on Measurement and Modeling of Computer Systems*, pp. 63–77, Cambridge, MA, Aug. 1984.

[2] A. V. Aho, J. E. Hopcroft, and J. D. Ullman. *Data Structures and Algorithms*. Addison-Wesley, 1983.

[3] M. Ajmone Marsan, G. Balbo, and G. Conte. *Performance Models of Multiprocessor Systems*. MIT Press, Cambridge, 1986.

[4] H. H. Ammar and S. M. R. Islam. Time scale decomposition of a class of generalized stochastic Petri net models. *IEEE Trans. Software Eng.*, 15(6):809–820, June 1989.

[5] F. Baccelli and Z. Liu. Comparison properties of stochastic decision free Petri nets. *IEEE Trans. Automat. Contr.*, 37(12):1905–1920, Dec. 1992.

[6] B. Baynat and Y. Dallery. Approximate techniques for general closed queueing networks with subnetworks having population constraints. Tech. report, MASI 90-49, Univ. Paris 6, Paris, France, Oct. 1990.

[7] B. Baynat and Y. Dallery. A unified view of product-form approximation techniques for general closed queueing networks. Tech. report, MASI 90-48, Univ. Paris 6, Paris, France, Oct. 1990.

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

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

[10] G. Ciardo and K. Trivedi. A decomposition approach for stochastic Petri nets models. In *Proc. of the 4th Int. Workshop on Petri Nets and Performance Models*, pp. 74–83, Melbourne, Australia, Dec. 1991. IEEE Comput. Soc. Press.

[11] J. M. Colom. *Análisis Estructural de Redes de Petri, Programación Lineal y Geometría Convexa.* PhD thesis, Dpto. de Ingeniería Eléctrica e Informática, Univ. Zaragoza, Spain, June 1989.

[12] J. M. Colom and M. Silva. Improving the linearly based characterization of P/T nets. In G. Rozenberg, editor, *Advances in Petri Nets 1990*, Vol. 483 of *LNCS*, pp. 113–145. Springer-Verlag, Berlin, 1991.

[13] Y. Dallery, Z. Liu, and D. Towsley. Equivalence, reversibility and symmetry properties in fork/join queueing networks with blocking. Tech. report, MASI 90-32, Univ. Paris 6, Paris, France, June 1990.

[14] A. Desrochers, H. Jungnitz, and M. Silva. An approximation method for the performance analysis of manufacturing systems based on GSPN's. In *Proc. of the Rensselaer's Third Int. Conf. on Computer Integrated Manufacturing*, pp. 46–55, Troy, NY, May 1992. IEEE Comput. Soc. Press.

[15] S. Donatelli and M. Sereno. On the product form solution for stochastic Petri nets. In *Proc. of the 13th Int. Conf. on Applications and Theory of Petri Nets*, pp. 154–172, Sheffield, UK, June 1992.

[16] H. Jungnitz, B. Sánchez, and M. Silva. Approximate throughput computation of stochastic marked graphs. *Journal of Parallel and Distributed Computing*, 15:282–295, 1992.

[17] H. J. Jungnitz. *Approximation Methods for Stochastic Petri Nets.* PhD thesis, Dept. of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, May 1992.

[18] Y. Li and C. M. Woodside. Iterative decomposition and aggregation of stochastic marked graphs Petri nets. In *Proc. of the 12th Int. Conf. on Applications and Theory of Petri Nets*, pp. 257–275, Gjern, Denmark, June 1991.

[19] Y. Li and C. M. Woodside. Performance Petri net analysis of communications protocol software by delay-equivalent aggregation. In *Proc. of the 4th Int. Workshop on Petri Nets and Performance Models*, pp. 64–73, Melbourne, Australia, Dec. 1991. IEEE Comput. Soc. Press.

[20] R. A. Marie. An approximate analytical method for general queueing networks. *IEEE Trans. Software Eng.*, 5(5):530–538, Sep. 1979.

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

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