On the use of Kronecker operators for the solution of generalized stochastic Petri nets

Gianfranco Ciardo [*]
Dept. of Computer Science
College of William and Mary
Williamsburg, VA 23187-8795, USA
ciardo@cs.wm.edu

Marco Tilgner [†]
Dept. of Mathematical and Computing Sciences
Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo 152, JAPAN
marco@is.titech.ac.jp

Abstract

We discuss how to describe the Markov chain underlying a generalized stochastic Petri net using Kronecker operators on smaller matrices. We extend previous approaches by allowing both an extensive type of marking-dependent behavior for the transitions and the presence of immediate synchronizations. The derivation of the results is thoroughly formalized, including the use of Kronecker operators in the treatment of the vanishing markings and the computation of impulse-based reward measures. We use our techniques to analyze a model whose solution using conventional methods would fail because of the state-space explosion. In the conclusion, we point out ideas to parallelize our approach.

[*] This research was partially supported by the National Aeronautics and Space Administration under NASA Contract No. NAS1-19480.

[†] This research was initiated during a visiting stay at the College of William and Mary. It was supported by the Japanese Ministry of Education, Science and Culture and the DAAD.

1 Introduction

Generalized stochastic Petri nets (GSPNs) [2, 3, 6] are ideally suited to model a large class of performance and reliability problems, but their numerical analysis requires the solution of a very large continuous-time Markov chain (CTMC). The size of the transition rate matrix $\mathbf{R}$ for the CTMC is the main obstacle, since its memory requirements can easily exceed the capacity of today's (and tomorrow's) machines even when sparse storage techniques are employed.

A possible approach to this problem is to store $\mathbf{R}$ implicitly. Plateau [14] proposed the use of Kronecker operators for the description of the transition rate matrix of a structured model composed of a set of "synchronized" stochastic automata, a subclass of GSPNs. Buchholz [4] used a similar idea for Markovian closed asynchronous queueing networks, and Takahashi [17] used it for open queueing networks with communication blocking. Donatelli [10, 11] adapted the approach to GSPNs, defining first the "superposed stochastic automata", then the "superposed GSPNs". Later, Buchholz [5] applied the concept to the special case of marked graphs and Kemper [12] addressed some of the problems in [11], extending the applicability of the results.

These approaches have in common a decomposition of the model into a set of submodels, so that the state space of the CTMC underlying the entire model is a subset of the cross-product of the state spaces of the CTMCs underlying each submodel. This implies that the transition rate for the entire model can be described using Kronecker operators on smaller matrices.

Focusing on GSPNs, the result of decomposing a GSPN is a set of largely independent sub-GSPNs, but some transitions will be shared by multiple sub-GSPNs, to model interactions among them. If a transition t_j is shared by two sub-GSPNs A_1 and A_2, and t_j has input and output places in both of them, this models a synchronization. A_1 and A_2 "wait for each other" and the event corresponding to t_j occurs only when they are both "ready". An alternative case arises when t_j has its input places in A_1 and its output places in A_2. This describes an asynchronous (and asymmetric) communication, since A_2 must "wait for permission" from A_1. However, if an output place of t_j in A_2 has a capacity defined for it, A_1 will wait for A_2 as well. One of the contributions of our work is to present a unified framework for all these types of interactions. Indeed, we do not assume that the net possesses any particular structure.

The solution of a decomposed GSPN is then based on the following idea [11]. Let n_i be the number of states in $\mathcal{S}_T^i$, the state space for the CTMC underlying the i-th sub-GSPN, $i = 1, \ldots, N$. $\mathcal{S}'_T = \mathcal{S}_T^1 \times \ldots \times \mathcal{S}_T^N \supseteq \mathcal{S}_T$ is the "potential state space" for the model, usually (much) larger than the actual state space $\mathcal{S}_T$, so that a transition rate matrix $\mathbf{R}'$ is defined using Kronecker algebra:

$$\mathbf{R}' = \bigoplus_{i=1}^{N} \mathbf{R}^i + \sum_{t_j \in \mathcal{T}^\bullet} \bigotimes_{i=1}^{N} \mathbf{R}^{i,j},$$

where $\mathbf{R}^i$ describes the local transitions for the i-th sub-GSPN (including the effect of immediate transitions), $\mathcal{T}^\bullet$ is the set of synchronizing transitions, which must be timed, and the "corrective matrix" $\mathbf{R}^{i,j}$ describes the effect of t_j on the i-th sub-GSPN.

The actual transition rate matrix $\mathbf{R}$ can be obtained from $\mathbf{R}'$ by eliminating the rows and columns corresponding to the unreachable states in $\mathcal{S}'_T \setminus \mathcal{S}_T$, but this requires an additional overhead, since the composition of a marking does not directly indicate whether it is reachable or not. Hence, an alternative approach was initially suggested. The power or Jacobi method is used to compute the steady-state solution in a vector $\boldsymbol{\pi}'$ of size $|\mathcal{S}'_T|$. By assigning a nonzero initial probability only to markings in $\mathcal{S}_T$, the solution $\boldsymbol{\pi}'$ is guaranteed to be zero for any marking $\mathbf{m} \in \mathcal{S}'_T \setminus \mathcal{S}_T$. This simplifies the algorithm, since $\mathbf{m}$ can now be interpreted as the mixed-base integer index of the corresponding entries in $\mathbf{R}'$ and $\boldsymbol{\pi}'$, but the memory requirement might be excessive when $|\mathcal{S}'_T| \gg |\mathcal{S}_T|$.

To reduce the impact of unreachable markings, Kemper [12] proposed a technique that only requires a probability vector $\boldsymbol{\pi}$ of size $|\mathcal{S}_T|$. In the numerical iterations, for each $\mathbf{m} \in \mathcal{S}_T$, each entry $\mathbf{R}'_{\mathbf{m},\mathbf{n}} > 0$ is generated (this implies $\mathbf{n} \in \mathcal{S}_T$) and, given $\mathbf{n}$, its index k in $\boldsymbol{\pi}$, or its lexicographic position in $\mathcal{S}_T$, is computed in $O(\log |\mathcal{S}_T|)$ operations, using a binary search.

We unify previous work by offering a thorough discussion of the structure of the underlying CTMC, including the management of immediate transitions and vanishing markings. Our formalism is more general than those assumed in [4, 5, 10, 11, 12], since we allow for marking-dependent arc cardinalities and rates, subject to certain restrictions, hence our results include those previously mentioned as special cases. Then, using an approach based on discrete-time Markov chains (DTMCs), we also remove the main restriction previously imposed on the decomposition of the GSPN: we allow immediate synchronizing transitions. Finally, we consider a reward structure defined on the GSPN, and we show how to compute the expected reward in steady-state in the Kronecker framework. This is of particular importance for impulse rewards associated with immediate transitions, whose firing are only implicitly represented in $\mathbf{R}$.

The paper is structured as follows. Section 2 describes the notation used and recalls the main concepts of Kronecker algebra, GSPNs, Markov chains, and rewards. Section 3 presents the expression for the transition rate matrix of the CTMC underlying a generic GSPN, provided its transitions satisfy certain restrictions on the type of marking-dependency. The result is quite general, but not directly applicable, since it requires one to compute the inverse of a matrix described as the sum of Kronecker products. However, Sections 4 and 5 use it to derive computationally effective expressions for GSPNs with timed and immediate synchronizing transitions, respectively. Implementation and application of these results are shown in Section 6, including detailed information about computation time and memory requirements. Section 7 contains a summary and discusses future extensions, including distributed implementations and approximate solutions.

2 Notation and definitions

Except for $I\!N$, the sets of natural numbers, $\{0, 1, 2 \ldots\}$, and $I\!R$, the set of real numbers, all sets are denoted by upper case calligraphic letters (e.g., $\mathcal{A}$); vectors and matrices are denoted by lower and upper case bold letters, respectively (e.g., $\mathbf{a}$, $\mathbf{A}$); their entries are denoted by subscripts (e.g., $\mathbf{a}_k$, $\mathbf{A}_{k,l}$); a set of indices can be used instead of a single index, for example, $\mathbf{A}_{\mathcal{X},\mathcal{Y}}$ denotes the submatrix of $\mathbf{A}$ corresponding to set of rows $\mathcal{X}$ and the set of columns $\mathcal{Y}$. Superscripts denote families of related quantities (e.g., $\mathbf{A}^1$, $\mathbf{A}^2$). $\mathbf{0}_{x\times y}$ and $\mathbf{1}_{x\times y}$ denote matrices with x rows and y columns, having all entries equal to 0 or 1, respectively, while $\mathbf{I}_x$ denotes the identity matrix of size $x \times x$; the dimensions of these matrices are omitted when they are clear from the context. Given a vector $\mathbf{a}$, $\text{diag}(\mathbf{a})$ is a square matrix having $\mathbf{a}$ on the diagonal and zeros elsewhere. Given an $n \times n$ matrix $\mathbf{A}$, $\text{rowsum}(\mathbf{A}) = \text{diag}(\mathbf{A} \cdot \mathbf{1}_{n\times 1})$ is a matrix having the diagonal equal to the sums of the entries on each row of $\mathbf{A}$, and zeros elsewhere, while $\delta(\mathbf{A})$ is a matrix having the same nonzero pattern as $\mathbf{A}$, but with entries equal to either 0 or 1.

2.1 Kronecker algebra

We recall the definition of the Kronecker product $\mathbf{B} = \bigotimes_{k=1}^{K} \mathbf{A}^k$ of K matrices $\mathbf{A}^k \in \mathbb{R}^{n_k \times m_k}$, using the convention that the rows and columns of both $\mathbf{B}$ and the matrices $\mathbf{A}^k$ are indexed starting at 0. The generic element of $\mathbf{B} \in \mathbb{R}^{\prod_{k=1}^{K} n_k \times \prod_{k=1}^{K} m_k}$ is

$$\mathbf{B}_{(\dots((i_1)n_2+i_2)n_3\cdots)n_k+i_k,(\dots((j_1)m_2+j_2)m_3\cdots)m_k+j_k} \quad = \quad \mathbf{A}^1_{i_1,j_1}\mathbf{A}^2_{i_2,j_2}\cdots\mathbf{A}^K_{i_K,j_K}$$

with $0 \le i_k < n_k$ and $0 \le j_k < m_k$, for $k = 1, \dots, K$. Assuming a mixed-base numbering scheme so that the tuple $(l_1, l_2, \dots l_K)$ corresponds to row $(\dots((l_1)n_2 + l_2)n_3\cdots)n_k + l_k$ or column $(\dots((l_1)m_2 + l_2)m_3\cdots)m_k + l_k$, respectively, we will also write the above quantity, more succinctly, as $\mathbf{B}_{(i_1,i_2,\dots i_k),(j_1,j_2,\dots j_k)}$.

The Kronecker sum $\bigoplus_{k=1}^{K} \mathbf{A}^k$ of K square matrices $\mathbf{A}^k \in \mathbb{R}^{n_k \times n_k}$ is defined as

$$\bigoplus_{k=1}^{K} \mathbf{A}^k = \sum_{k=1}^{K} \mathbf{I}_{n_1} \otimes \cdots \otimes \mathbf{I}_{n_{k-1}} \otimes \mathbf{A}^k \otimes \mathbf{I}_{n_{k+1}} \cdots \otimes \mathbf{I}_{n_K}.$$

2.2 Generalized stochastic Petri nets

A generalized stochastic Petri net (GSPN) is a tuple $(\mathcal{P}, \mathcal{T}, \mathcal{I}, \mathbf{C}^-, \mathbf{C}^+, \mathbf{m}^0, \mathbf{w})$, where:

- $\mathcal{P} = \{p_1, \dots, p_{|\mathcal{P}|}\}$ is a finite set of *places*, drawn as circles in the graphical representation of the GSPN. A non-negative integer vector $\mathbf{m} \in \mathbb{N}^{|\mathcal{P}|}$ called *marking* describes the number of *tokens* in each place. Given a place $p_i \in \mathcal{P}$, $\mathbf{m}_i$ is the number of tokens in p_i for marking $\mathbf{m}$.

- $\mathcal{T} = \{t_1, \dots, t_{|\mathcal{T}|}\}$ is a finite set of *transitions*, $\mathcal{P} \cap \mathcal{T} = \emptyset$.

- $\mathcal{I} \subseteq \mathcal{T}$ is the subset of *immediate* transitions, drawn as thin bars, while $\mathcal{X} = \mathcal{T} \setminus \mathcal{I}$ are the *timed transitions*, drawn as rectangles. The *firing time* of immediate transitions is zero, while that of timed transitions is exponentially distributed.

- $\mathbf{C}^-$ and $\mathbf{C}^+$ are incidence matrices of size $|\mathcal{P}| \times |\mathcal{T}|$. Their elements are functions from $\mathbb{N}^{|\mathcal{P}|}$ to $\mathbb{N}$. $\mathbf{C}^-_{i,j}(\mathbf{m})$ and $\mathbf{C}^+_{i,j}(\mathbf{m})$ denote the marking-dependent integer cardinality assigned to the input arc from p_i to t_j and the output arc from t_j to p_i respectively. In the graph, these arcs are drawn using an arrowhead pointing to the destination if their cardinality is not identically equal to zero. The cardinality function is indicated on the arc unless it is identically equal to one.

- $\mathbf{m}^0$ is the *initial marking*. In the graph, the value of $\mathbf{m}^0_i$ is written inside place p_i, if positive.

- For any $t_j \in \mathcal{T}$, $\mathbf{w}_j$ is a function from $\mathbb{N}^{|\mathcal{P}|}$ to $\mathbb{R}$. $\mathbf{w}_j(\mathbf{m})$ is the *weight* associated with transition t_j in marking $\mathbf{m}$. According to whether t_j is immediate or timed, this weight represents an (unnormalized) *firing probability*, or a *firing rate*.

A transition $t_j \in \mathcal{T}$ has *concession* in marking $\mathbf{m}$ iff

$$\forall p_i \in \mathcal{P},\ \mathbf{C}^-_{i,j}(\mathbf{m}) \le \mathbf{m}_i, \qquad \text{or} \ \ \mathbf{C}^-_{\mathcal{P},j}(\mathbf{m}) \le \mathbf{m}.$$

If any immediate transition has concession in $\mathbf{m}$, it is *enabled*, and $\mathbf{m}$ is said to be *vanishing*. Otherwise, $\mathbf{m}$ is said to be *tangible* and any timed transition t_j with concession is enabled in $\mathbf{m}$. In other words, a timed transition is not enabled in a vanishing marking even if it has concession.

Some definitions of GSPNs allow one to disable a transition t_j with concession in $\mathbf{m}$ by specifying a zero weight $\mathbf{w}_j(\mathbf{m})$ for it, or by introducing inhibitor arcs, drawn with a circle instead of an arrowhead. In a marking $\mathbf{m}$, an inhibitor arc from place p_i to transition t_j with cardinality $c(\mathbf{m})$ disables t_j if $\mathbf{m}_i \ge c(\mathbf{m})$. Since these behaviors can be represented by an appropriate definition of input arc cardinalities in our formalism, we assume, without loss of generality, that $\mathbf{w}_j(\mathbf{m}) > 0$ if t_j is enabled in $\mathbf{m}$, and we use inhibitor arcs in our models merely as a shorthand.

Let $\mathcal{E}(\mathbf{m})$ denote the set of enabled transitions in marking $\mathbf{m}$. An enabled transition t_j firing in marking $\mathbf{m}$ yields a new marking $\mathbf{n}$ such that

$$\forall p_i \in \mathcal{P},\ \mathbf{n}_i = \mathbf{m}_i - \mathbf{C}^-_{i,j}(\mathbf{m}) + \mathbf{C}^+_{i,j}(\mathbf{m}) = \mathbf{m}_i + \mathbf{C}_{i,j}(\mathbf{m}) \qquad (\text{or} \ \ \mathbf{n} = \mathbf{m} + \mathbf{C}_{\mathcal{P},j}(\mathbf{m})),$$

where $\mathbf{C} = \mathbf{C}^+ - \mathbf{C}^-$ is the *incidence matrix* of the GSPN. We can also write $\mathbf{m}\xrightarrow{t_j}\mathbf{n}$ to express that t_j has concession in $\mathbf{m}$ and that $\mathbf{n}$ is obtained from $\mathbf{m}$ by firing t_j, regardless of whether $t_j \in \mathcal{E}(\mathbf{m})$ or not (t_j is not enabled if it is timed and $\mathbf{m}$ is vanishing, or if $\mathbf{w}_j(\mathbf{m}) = 0$).

The firing probability of a transition t_j enabled in marking $\mathbf{m}$ is

$$\frac{\mathbf{w}_j(\mathbf{m})}{\sum_{t_l \in \mathcal{E}(\mathbf{m})} \mathbf{w}_l(\mathbf{m})}. \tag{1}$$

If $\mathbf{m}$ is tangible, this corresponds to a *race* between the exponentially distributed firing times of the enabled transitions, with rates given by $\mathbf{w}$. In a vanishing marking, instead, weights define a probabilistic choice, since the race model does not specify how to choose which transition to fire next when multiple enabled transitions have the same zero firing time.

2.3 Reachability set

The reachability set $\mathcal{S}$ is defined as the set of markings reachable from the initial marking $\mathbf{m}^0$ by firing any sequence of enabled transitions. Formally, $\mathcal{S}$ is the smallest subset of $I\!N^{|\mathcal{P}|}$ containing $\mathbf{m}^0$ and such that $\mathbf{m} \in \mathcal{S}$, $t_j \in \mathcal{E}(\mathbf{m})$, and $\mathbf{m}\xrightarrow{t_j}\mathbf{n}$ imply $\mathbf{n} \in \mathcal{S}$. Fig. 1 shows the skeleton of an algorithm to build the set of reachable markings $\mathcal{S}$ (which we assume finite from now on). Particular care must be placed on the implementation of statement 9, since the size of the set $\mathcal{S}$ to be searched is very large in practice. Efficient methods include hashing or balanced search trees (e.g., AVL trees [18]). While not explicitly stated in the algorithm, $\mathcal{S}$ should be stored as the union of two disjoint sets, $\mathcal{S}_T$ and $\mathcal{S}_V$, corresponding to the tangible and vanishing markings, respectively.

A function Ψ assigns an index to each reachable marking, according to a lexicographic order, indicated by "$\succ$":

$$\Psi : \mathcal{S} \to \{0, \ldots, |\mathcal{S}| - 1\} \quad \text{such that} \quad \Psi(\mathbf{m}) > \Psi(\mathbf{n}) \Longleftrightarrow \mathbf{m} \succ \mathbf{n}.$$

```
Algorithm BuildRS (input: $(\mathcal{P}, \mathcal{T}, \mathcal{I}, \mathbf{C}^-, \mathbf{C}^+, \mathbf{m}^0, \mathbf{w})$;    output: $\mathcal{S}$);

$\mathcal{S} \leftarrow \{\mathbf{m}^0\}$;          /* $\mathcal{S}$ contains the markings found so far */
$\mathcal{U} \leftarrow \{\mathbf{m}^0\}$;          /* $\mathcal{U} \subseteq \mathcal{S}$ contains the found but unexplored markings */
while $\mathcal{U} \neq \emptyset$ do
    "choose a marking $\mathbf{m}$ from $\mathcal{U}$";
    $\mathcal{U} \leftarrow \mathcal{U} \setminus \{\mathbf{m}\}$;
    "compute $\mathcal{E}(\mathbf{m})$";
    for each $j \in \mathcal{E}(\mathbf{m})$ do
        $\mathbf{n} \leftarrow \mathbf{m} + \mathbf{C}_{\mathcal{P},j}(\mathbf{m})$;
        if $\mathbf{n} \notin \mathcal{S}$ then
            $\mathcal{U} \leftarrow \mathcal{U} \cup \{\mathbf{n}\}$;
            $\mathcal{S} \leftarrow \mathcal{S} \cup \{\mathbf{n}\}$;
        end if
    end for
end while
```

Figure 1: Algorithm BuildRS

If an AVL tree is used, Ψ can be precomputed with a simple preorder visit of the tree, and its value can be stored in the nodes of the tree. Then, given $\mathbf{m} \in I\!N^{|P|}$, the value $k = \Psi(\mathbf{m})$ can be found in $O(\log |\mathcal{S}|)$ operations using the AVL tree augmented with this additional information ($\Psi(\mathbf{m})$ = "undefined" for any $\mathbf{m} \notin \mathcal{S}$). The restrictions of Ψ to the tangible and vanishing markings, can be defined accordingly: $\Psi_T : \mathcal{S}_T \rightarrow \{0, \ldots, |\mathcal{S}_T| - 1\}$ and $\Psi_V : \mathcal{S}_V \rightarrow \{0, \ldots, |\mathcal{S}_V| - 1\}$.

In the following, with a slight overloading in the notation, we use a marking $\mathbf{m}$ to index data structures (vectors and matrices) referring to $\mathcal{S}$, $\mathcal{S}_T$, or $\mathcal{S}_V$. Strictly speaking, we should use instead $\Psi(\mathbf{m})$, $\Psi_T(\mathbf{m})$, and $\Psi_V(\mathbf{m})$, respectively, but this would make the expressions excessively cumbersome. Nevertheless, it is important to stress this fundamental difference from a computational point of view; finding the index of a marking is a potential source of additional complexity in any structured approach.

2.4 Underlying continuous-time Markov chain and rewards

We focus on the steady-state analysis of the continuous-time Markov chain (CTMC) underlying a GSPN, described by the *infinitesimal generator* matrix $\mathbf{Q} \in I\!R^{|\mathcal{S}_T| \times |\mathcal{S}_T|}$, which we assume ergodic (and finite, since $\mathcal{S}$ is finite):

$$\mathbf{Q} = \mathbf{R} - \boldsymbol{\Lambda} = \mathbf{R}_{T,T} + \mathbf{R}_{T,V}(\mathbf{I}_{|\mathcal{S}_V|} - \mathbf{U}_{V,V})^{-1}\mathbf{U}_{V,T} - \boldsymbol{\Lambda}, \qquad (2)$$

where $\mathbf{R}$ is the *transition rate* matrix, $\boldsymbol{\Lambda} = \text{rowsum}(\mathbf{R})$, and $\mathbf{R}_{T,T}$ and $\mathbf{R}_{T,V}$ ($\mathbf{U}_{V,T}$ and $\mathbf{U}_{V,V}$) describe the rates (probabilities) of going from tangible (vanishing) markings to tangible or vanishing markings, respectively. The entry of $\mathbf{R}_{T,T}$ ($\mathbf{R}_{T,V}$) corresponding to the row for $\mathbf{m}$ and the column for $\mathbf{n}$ describes the rate from $\mathbf{m} \in \mathcal{S}_T$ to $\mathbf{n} \in \mathcal{S}_T$ ($\mathbf{n} \in \mathcal{S}_V$):

$$\sum_{t_j \in \mathcal{E}(\mathbf{m}), \mathbf{m} \xrightarrow{t_j} \mathbf{n}} \mathbf{w}_j(\mathbf{m}).$$

The entries of $\mathbf{U}_{V,V}$ and $\mathbf{U}_{V,T}$ are defined analogously, using the firing probability of immediate transition t_j, given by (1), instead of the weight $\mathbf{w}_j(\mathbf{m})$. For a discussion of how to generate $\mathbf{Q}$ in practice, see [2, 6].

We observe that $\boldsymbol{\Lambda}_{\mathbf{m},\mathbf{m}} = \sum_{t_j \in \mathcal{E}(\mathbf{m})} \mathbf{w}_j(\mathbf{m})$ is then the total rate leaving marking $\mathbf{m} \in \mathcal{S}_T$, and it equals the inverse of $\mathbf{h}_{\mathbf{m}}$, the expected *holding time* in $\mathbf{m}$.

Let $\boldsymbol{\pi}_{\mathbf{m}}$ be the steady-state probability of a tangible marking $\mathbf{m}$ (vanishing markings have zero probability). Then, the steady-state probability (row) vector $\boldsymbol{\pi} \in \mathbb{R}^{|\mathcal{S}_T|}$ satisfies the balance equation

$$\boldsymbol{\pi} \cdot \mathbf{Q} = \mathbf{0}_{1 \times |\mathcal{S}_T|} \qquad \text{subject to the normalization} \qquad \boldsymbol{\pi} \cdot \mathbf{1}_{|\mathcal{S}_T| \times 1} = 1. \tag{3}$$

We can specify a quantity of interest for the GSPN using a *reward structure* $(\rho, \mathbf{r})$, where $\rho(\mathbf{m})$ is the *reward rate* gained while the GSPN is in marking $\mathbf{m}$, and $\mathbf{r}_j(\mathbf{m})$ is the *reward impulse* gained when transition $t_j \in \mathcal{T}$ fires in marking $\mathbf{m}$. The expected reward rate in steady state is then

$$\sum_{\mathbf{m} \in \mathcal{S}_T} \boldsymbol{\pi}_{\mathbf{m}} \rho(\mathbf{m}) + \sum_{\mathbf{m} \in \mathcal{S}} \sum_{t_j \in \mathcal{E}(\mathbf{m})} \Phi_{j,\mathbf{m}} \mathbf{r}_j(\mathbf{m}), \tag{4}$$

where $\Phi_{j,\mathbf{m}}$ is the rate at which transition t_j fires in steady state in marking $\mathbf{m}$. If we let $\boldsymbol{\phi} \in \mathbb{R}^{|\mathcal{S}|}$ be the vector describing the rate at which each marking is entered in steady state, $\Phi_{j,\mathbf{m}}$ is obtained as

$$\Phi_{j,\mathbf{m}} = \boldsymbol{\phi}_{\mathbf{m}} \frac{\mathbf{w}_j(\mathbf{m})}{\sum_{t_l \in \mathcal{E}(\mathbf{m})} \mathbf{w}_l(\mathbf{m})}.$$

For $\mathbf{m} \in \mathcal{S}_T$, $\boldsymbol{\phi}_{\mathbf{m}} = \boldsymbol{\pi}_{\mathbf{m}} \sum_{t_l \in \mathcal{E}(\mathbf{m})} \mathbf{w}_l(\mathbf{m}) = \boldsymbol{\pi}_{\mathbf{m}} \boldsymbol{\Lambda}_{\mathbf{m},\mathbf{m}}$. For $\mathbf{m} \in \mathcal{S}_V$, instead,

$$\boldsymbol{\phi}_{\mathbf{m}} = \sum_{\mathbf{n} \in \mathcal{S}_T} \boldsymbol{\pi}_{\mathbf{n}} \cdot \mathbf{F}_{\mathbf{n},\mathbf{m}},$$

where, for $\mathbf{n} \in \mathcal{S}_T$ and $\mathbf{m} \in \mathcal{S}_V$, the corresponding entry of matrix

$$\mathbf{F} = \mathbf{R}_{T,V} (\mathbf{I}_{|\mathcal{S}_V|} - \mathbf{U}_{V,V})^{-1} \in \mathbb{R}^{|\mathcal{S}_T| \times |\mathcal{S}_V|} \tag{5}$$

describes the rate at which a vanishing marking $\mathbf{m}$ is entered after leaving a tangible marking $\mathbf{n}$ and before reaching the next tangible marking. If no reward impulse is defined for immediate transitions, then Eq. (4) reduces to

$$\sum_{\mathbf{m} \in \mathcal{S}_T} \boldsymbol{\pi}_{\mathbf{m}} \left(\rho(\mathbf{m}) + \sum_{t_j \in \mathcal{E}(\mathbf{m})} \mathbf{w}_j(\mathbf{m}) \mathbf{r}_j(\mathbf{m}) \right).$$

In this work, we consider the structure of both $\mathbf{Q}$ and $\mathbf{F}$, and the computation of $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$.

3 Kronecker expression for the CTMC underlying a GSPN

We now show how the ideas in [4, 5, 10, 11, 12] can be applied to individual places, not just to sub-GSPNs. Our goal is to clarify the relationship between Kronecker algebra and GSPNs, while relaxing several important restrictions on the type of interactions. Later we will merge individual places into "macroplaces", which corresponds to the notion of sub-GSPNs.

3.1 Using the state spaces of individual places

The first extension regards the type of marking-dependency allowed in the GSPN. We allow the weight of a transition to be expressed as the product of "local effects" due to the number of tokens in each place:

$$\forall \mathbf{m} \in \mathcal{S}, \forall t_j \in \mathcal{T},\ \mathbf{C}^-_{\mathcal{P},j} \le \mathbf{m} \;\;\Rightarrow\;\; \mathbf{w}_j(\mathbf{m}) = w^*_j \cdot \prod_{p_i \in \mathcal{P}} w_{i,j}(\mathbf{m}_i), \tag{6}$$

where w^*_j can be interpreted as a constant "reference" weight, while the values $w_{i,j}$ are dimensionless scaling functions [1]. This "independence of effects" in the marking dependence implies, for example, that if markings $\mathbf{m}$ and $\mathbf{n}$ differ only in the number of tokens in p_i, and if t_j is enabled in both, $\mathbf{w}_j(\mathbf{n}) = \mathbf{w}_j(\mathbf{m}) \cdot w_{i,j}(\mathbf{n}_i)/w_{i,j}(\mathbf{m}_i)$. If a weight $\mathbf{w}_j$ does not depend on the number of tokens in p_i, we assume, without loss of generality, that $w_{i,j}$ is identically equal to one. Note that we do not require the weight $\mathbf{w}_j(\mathbf{m})$ of a timed transition t_j in a vanishing marking $\mathbf{m}$ to be zero. Doing so would make the specification of $\mathbf{w}$ for a given GSPN more difficult in practice, and is not required by our approach.

Analogously, the dependence of the matrices $\mathbf{C}^-$ and $\mathbf{C}^+$, hence $\mathbf{C}$, is assumed to be of the form

$$\forall \mathbf{m} \in \mathcal{S}, \forall p_i \in \mathcal{P}, \forall t_j \in \mathcal{T},\ \mathbf{C}^\star_{i,j}(\mathbf{m}) = \beta^\star_{i,j}(\mathbf{m}_i), \tag{7}$$

where "$\star$" is one of "$-$", "$+$", or nothing, and $\beta^\star_{i,j}$ is a function from $I\!N$ to $I\!N$.

We can now state a theorem expressing the matrices $\mathbf{Q}$ and $\mathbf{F}$ of the CTMC underlying a GSPN in terms of smaller matrices related to each place-transition pair.

Theorem 3.1 Consider a GSPN with finite reachability set $\mathcal{S}$ satisfying Eq. (6) and (7), and let $n_i - 1$ be the bound of place $p_i \in \mathcal{P}$, that is, for any $\mathbf{m} \in \mathcal{S}, \mathbf{m}_i \in \{0, 1, \ldots n_i - 1\} = \mathcal{S}^i$. Define

$$\mathbf{R}' = \sum_{t_j \in \mathcal{X}} w^*_j \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{W}^{i,j} \qquad\qquad \mathbf{U}' = \sum_{t_j \in \mathcal{I}} w^*_j \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{W}^{i,j}, \tag{8}$$

where $\mathbf{W}^{i,j}$ is a square matrix of size $n_i \times n_i$ whose entry in position (r, c), for $r, c \in \mathcal{S}^i$, is given by

$$\mathbf{W}^{i,j}(r,c) \;\;=\;\; \begin{cases} w_{i,j}(r) & \text{if } r \ge \beta^-_{i,j}(r) \text{ and } c = r + \beta_{i,j}(r) \\ 0 & \text{otherwise} \end{cases} .$$

Also, define

$$\boldsymbol{\Lambda}' = \text{rowsum}(\mathbf{R}') = \sum_{t_j \in \mathcal{X}} w^*_j \cdot \bigotimes_{p_i \in \mathcal{P}} \text{rowsum}(\mathbf{W}^{i,j}) = \sum_{t_j \in \mathcal{X}} w^*_j \cdot \bigotimes_{p_i \in \mathcal{P}} \boldsymbol{\Lambda}^{i,j} \tag{9}$$

$$\boldsymbol{\Gamma}' = \mathrm{rowsum}(\mathbf{U}') = \sum_{t_j \in \mathcal{I}} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \mathrm{rowsum}(\mathbf{W}^{i,j}) = \sum_{t_j \in \mathcal{I}} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \boldsymbol{\Gamma}^{i,j} \tag{10}$$

and $\mathbf{T}' = \mathbf{I}_{|\mathcal{S}'|} - \delta(\boldsymbol{\Gamma}')$. Then, the matrix

$$\mathbf{Q}' = \mathbf{R}' \cdot \left(\mathbf{I} - (\mathbf{T}' + \boldsymbol{\Gamma}')^{-1} \cdot \mathbf{U}'\right)^{-1} - \boldsymbol{\Lambda}' \tag{11}$$

satisfies $\mathbf{Q} = \mathbf{Q}'_{\mathcal{S}_T,\mathcal{S}_T}$ and $\mathbf{F} = \mathbf{Q}'_{\mathcal{S}_T,\mathcal{S}_V}$, where $\mathbf{Q}$ and $\mathbf{F}$ have the meaning defined in Eq. (2) and (5).

Proof: Matrices with superscript "$'$" have row and column set $\mathcal{S}' = \left\{0, 1, \ldots \left(\prod_{p_i \in \mathcal{P}} n_i\right) - 1\right\}$, or $\mathcal{S}^1 \times \cdots \times \mathcal{S}^{|\mathcal{P}|}$, if we identify a tuple with its mixed-base value. In the following, however, we partition matrices and permute their rows and columns so that the markings appear in lexicographic order within the sets $\mathcal{S}_T$, $\mathcal{S}_V$, and $\mathcal{S}' \setminus \mathcal{S}$. This is for illustration purposes only.

First, we prove that

$$\mathbf{R}'_{\mathbf{m},\mathbf{n}} = \left(\sum_{t_j \in \mathcal{X}} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{W}^{i,j}\right)_{\mathbf{m},\mathbf{n}} = \sum_{t_j \in \mathcal{X}} w_j^* \cdot \prod_{p_i \in \mathcal{P}} \mathbf{W}^{i,j}_{\mathbf{m}_i,\mathbf{n}_i} = \sum_{t_j \in \mathcal{X}, \mathbf{m} \xrightarrow{t_j} \mathbf{n}} \mathbf{w}_j(\mathbf{m}). \tag{12}$$

Let's consider the contribution to this value for each timed $t_j \in \mathcal{X}$, by doing a case analysis:

1. If $\mathbf{m} \xrightarrow{t_j} \mathbf{n}$, the contribution should be $\mathbf{w}_j(\mathbf{m})$. Indeed, for all $p_i \in \mathcal{P}$, $\mathbf{W}^{i,j}_{\mathbf{m}_i,\mathbf{n}_i} = w_{i,j}(\mathbf{m}_i)$, hence the contribution of t_j is
$$w_j^* \cdot \prod_{p_i \in \mathcal{P}} w_{i,j}(\mathbf{m}_i) = \mathbf{w}_j(\mathbf{m}).$$
2. If t_j does not have concession in $\mathbf{m}$ the contribution should be zero. Indeed, there must exist a place p_i such that $\mathbf{m}_i < \mathbf{C}^-_{i,j}(\mathbf{m}) = \beta^-_{i,j}(\mathbf{m}_i)$. This implies $\mathbf{W}^{i,j}_{\mathbf{m}_i,\mathbf{n}_i} = 0$, and the contribution of t_j is $w_j^* \cdot \prod_{p_i \in \mathcal{P}} \mathbf{W}^{i,j}_{\mathbf{m}_i,\mathbf{n}_i} = 0$.
3. If $\mathbf{m} \xrightarrow{t_j} \mathbf{n}' \neq \mathbf{n}$, the contribution of t_j should be zero as well. Indeed, there must exist a place p_i such that $\mathbf{n}_i \neq \mathbf{m}_i + \mathbf{C}_{i,j}(\mathbf{m}) = \mathbf{m}_i + \beta_{i,j}(\mathbf{m}_i)$. Hence, $\mathbf{W}^{i,j}_{\mathbf{m}_i,\mathbf{n}_i} = 0$, and $w_j^* \cdot \prod_{p_i \in \mathcal{P}} \mathbf{W}^{i,j}_{\mathbf{m}_i,\mathbf{n}_i} = 0$.

Thus, the contribution of each transition in the summation is correct. An analogous argument allows us to show that

$$\mathbf{U}'_{\mathbf{m},\mathbf{n}} = \sum_{t_j \in \mathcal{I}, \mathbf{m} \xrightarrow{t_j} \mathbf{n}} \mathbf{w}_j(\mathbf{m}). \tag{13}$$

From Eq. (12) and (13), we can conclude that $\mathbf{U}'_{\mathcal{S}_T,\mathcal{S}'} = \mathbf{0}$, $\mathbf{U}'_{\mathcal{S}_V,\mathcal{S}'\setminus\mathcal{S}} = \mathbf{0}$, $\mathbf{R}'_{\mathcal{S}_T,\mathcal{S}'\setminus\mathcal{S}} = \mathbf{0}$, and that the matrices $\mathbf{R}_{T,T}$, $\mathbf{R}_{T,V}$, $\mathbf{U}_{V,T}$, and $\mathbf{U}_{V,V}$ for the underlying GSPN, with their rows and columns ordered according to Ψ, can be expressed as:

$$\mathbf{R}_{T,T} = \mathbf{R}'_{\mathcal{S}_T,\mathcal{S}_T} \qquad \mathbf{R}_{T,V} = \mathbf{R}'_{\mathcal{S}_T,\mathcal{S}_V} \qquad \mathbf{U}_{V,T} = \boldsymbol{\Gamma}'^{-1}_{\mathcal{S}_V,\mathcal{S}_V} \cdot \mathbf{U}'_{\mathcal{S}_V,\mathcal{S}_T} \qquad \mathbf{U}_{V,V} = \boldsymbol{\Gamma}'^{-1}_{\mathcal{S}_V,\mathcal{S}_V} \cdot \mathbf{U}'_{\mathcal{S}_V,\mathcal{S}_V} \tag{14}$$

(the normalization ${\boldsymbol{\Gamma}'}^{-1}_{\mathcal{S}_V,\mathcal{S}_V}$ is required because the weights of the immediate transitions enabled in a vanishing marking are not required to sum to one, while the entries in $\mathbf{U}_{V,T}$ and $\mathbf{U}_{V,V}$ are probabilities). We can conclude $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}'_{\mathcal{S}_T,\mathcal{S}_T}$ as well.

Hence, letting "$\bullet$" denote submatrices whose value is irrelevant,

$$
\begin{aligned}
\left(\mathbf{I} - (\mathbf{T}' + \boldsymbol{\Gamma}')^{-1} \cdot \mathbf{U}'\right)^{-1}
&= \left(\mathbf{I} - \left[\begin{array}{c|c|c} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & {\boldsymbol{\Gamma}'}^{-1}_{\mathcal{S}_V,\mathcal{S}_V} & \mathbf{0} \\ \hline \bullet & \bullet & \bullet \end{array}\right] \cdot \left[\begin{array}{c|c|c} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{U}'_{\mathcal{S}_V,\mathcal{S}_T} & \mathbf{U}'_{\mathcal{S}_V,\mathcal{S}_V} & \mathbf{0} \\ \hline \bullet & \bullet & \bullet \end{array}\right]\right)^{-1} \\
&= \left(\mathbf{I} - \left[\begin{array}{c|c|c} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline {\boldsymbol{\Gamma}'}^{-1}_{\mathcal{S}_V,\mathcal{S}_V}\mathbf{U}'_{\mathcal{S}_V,\mathcal{S}_T} & {\boldsymbol{\Gamma}'}^{-1}_{\mathcal{S}_V,\mathcal{S}_V}\mathbf{U}'_{\mathcal{S}_V,\mathcal{S}_V} & \mathbf{0} \\ \hline \bullet & \bullet & \bullet \end{array}\right]\right)^{-1} \\
&= \left[\begin{array}{c|c|c} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \hline -{\boldsymbol{\Gamma}'}^{-1}_{\mathcal{S}_V,\mathcal{S}_V}\mathbf{U}'_{\mathcal{S}_V,\mathcal{S}_T} & \mathbf{I} - {\boldsymbol{\Gamma}'}^{-1}_{\mathcal{S}_V,\mathcal{S}_V}\mathbf{U}'_{\mathcal{S}_V,\mathcal{S}_V} & \mathbf{0} \\ \hline \bullet & \bullet & \bullet \end{array}\right]^{-1} \\
&= \left[\begin{array}{c|c|c} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \hline -\mathbf{U}_{V,T} & \mathbf{I} - \mathbf{U}_{V,V} & \mathbf{0} \\ \hline \bullet & \bullet & \bullet \end{array}\right]^{-1} \\
&= \left[\begin{array}{c|c|c} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \hline (\mathbf{I} - \mathbf{U}_{V,V})^{-1}\mathbf{U}_{V,T} & (\mathbf{I} - \mathbf{U}_{V,V})^{-1} & \mathbf{0} \\ \hline \bullet & \bullet & \bullet \end{array}\right]
\end{aligned}
$$

Substituting this value in the definition of $\mathbf{Q}'$ given in Eq. (11) completes the proof:

$$
\begin{aligned}
\mathbf{Q}' &= \left[\begin{array}{c|c|c} \mathbf{R}'_{\mathcal{S}_T,\mathcal{S}_T} & \mathbf{R}'_{\mathcal{S}_T,\mathcal{S}_V} & \mathbf{0} \\ \hline \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet \end{array}\right] \cdot \left(\mathbf{I} - (\mathbf{T}' + \boldsymbol{\Gamma}')^{-1} \cdot \mathbf{U}'\right)^{-1} - \left[\begin{array}{c|c|c} \boldsymbol{\Lambda}'_{\mathcal{S}_T,\mathcal{S}_T} & \mathbf{0} & \mathbf{0} \\ \hline \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet \end{array}\right] \\
&= \left[\begin{array}{c|c|c} \mathbf{R}_{T,T} & \mathbf{R}_{T,V} & \mathbf{0} \\ \hline \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet \end{array}\right] \cdot \left[\begin{array}{c|c|c} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \hline (\mathbf{I} - \mathbf{U}_{V,V})^{-1}\mathbf{U}_{V,T} & (\mathbf{I} - \mathbf{U}_{V,V})^{-1} & \mathbf{0} \\ \hline \bullet & \bullet & \bullet \end{array}\right] - \left[\begin{array}{c|c|c} \boldsymbol{\Lambda} & \mathbf{0} & \mathbf{0} \\ \hline \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet \end{array}\right] \\
&= \left[\begin{array}{c|c|c} \mathbf{R}_{T,T} + \mathbf{R}_{T,V}(\mathbf{I} - \mathbf{U}_{V,V})^{-1}\mathbf{U}_{V,T} - \boldsymbol{\Lambda} & \mathbf{R}_{T,V}(\mathbf{I} - \mathbf{U}_{V,V})^{-1} & \mathbf{0} \\ \hline \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet \end{array}\right] \\
&= \left[\begin{array}{c|c|c} \mathbf{Q} & \mathbf{F} & \mathbf{0} \\ \hline \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet \end{array}\right] \qquad \Box
\end{aligned}
$$

If the number of tokens in p_i is always at least $m_i > 0$, we can, of course, define $\mathcal{S}^i = \{m_i, \ldots, n_i - 1\}$ or, equivalently, change the definition of the GSPN so that the range of tokens in $\mathbf{p}_i$ becomes $\mathcal{S}^i = \{0, \ldots, n_i - m_i - 1\}$. This would not affect the proofs in this paper, but could improve the efficiency of the implementation.

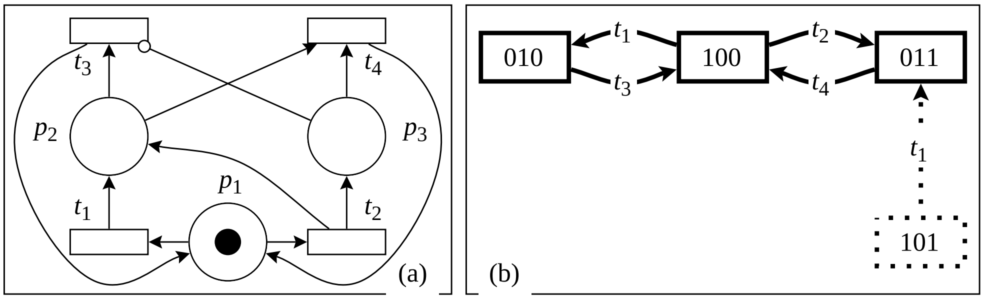

Figure 2: A case where $\mathbf{m}\xrightarrow{t_j}\mathbf{n}$, $\mathbf{n} \in \mathcal{S}$, but $\mathbf{m} \notin \mathcal{S}$.

It is important to stress that $\mathbf{R}'_{\mathcal{S}'\setminus\mathcal{S},\mathcal{S}'}$ and $\mathbf{U}'_{\mathcal{S}'\setminus\mathcal{S},\mathcal{S}'}$ cannot be guaranteed to be zero (nor can $\mathbf{R}'_{\mathcal{S}_V,\mathcal{S}'}$, but this is because we allow timed transitions to have concession in vanishing markings). We now formalize this observation, already implicit in previous works using the Kronecker approach, since it is fundamental for a better understanding of the nature of the matrices $\mathbf{R}'$ and $\mathbf{U}'$.

Lemma 3.1 The matrices $\mathbf{R}'$ and $\mathbf{U}'$ defined in Theorem 3.1 satisfy the following "forward reachability condition":

$$\mathbf{m} \in \mathcal{S}_T \;\wedge\; \mathbf{R}'_{\mathbf{m},\mathbf{n}} > 0 \;\Rightarrow\; \mathbf{n} \in \mathcal{S} \qquad \text{and} \qquad \mathbf{m} \in \mathcal{S}_V \;\wedge\; \mathbf{U}'_{\mathbf{m},\mathbf{n}} > 0 \;\Rightarrow\; \mathbf{n} \in \mathcal{S} \quad (15)$$

However, the analogous "backward reachability condition" does not hold:

$$\mathbf{n} \in \mathcal{S} \;\wedge \mathbf{R}'_{\mathbf{m},\mathbf{n}} > 0 \;\;\not\Rightarrow\;\; \mathbf{m} \in \mathcal{S}_T \qquad \text{and} \qquad \mathbf{n} \in \mathcal{S} \;\wedge \mathbf{U}'_{\mathbf{m},\mathbf{n}} > 0 \;\;\not\Rightarrow\;\; \mathbf{m} \in \mathcal{S}_V \quad (16)$$

Proof: It is straightforward to show that Eq. (15) holds when the premises of Theorem 3.1 are satisfied. We simply need to observe that, given the definition of $\mathbf{R}'$, $\mathbf{R}'_{\mathbf{m},\mathbf{n}} > 0$ and $\mathbf{m} \in \mathcal{S}_T$ imply that there is a transition $t_j \in \mathcal{E}(\mathbf{m})$ and that its firing leads to $\mathbf{n}$; thus, $\mathbf{n}$ is reachable. An analogous reasoning holds for $\mathbf{U}'$ and $\mathbf{m} \in \mathcal{S}_V$. To show that Eq. (16) holds, it is sufficient to give an example; we do so for $\mathbf{R}'$. Consider the GSPN in Fig. 2, having positive finite transition rates. The firing of two transitions t_2, t_3 could lead to $\mathbf{n} = \langle 0,1,1 \rangle$: by t_2, from marking $\langle 1,0,0 \rangle$, and by t_1, from marking $\langle 1,0,1 \rangle$. In our notation, $n_1 = n_2 = n_3 = 2$,

$$\mathbf{W}^{1,1} = \left[\begin{array}{c|c} 0 & 0 \\ \hline 1 & 0 \end{array}\right] \quad \mathbf{W}^{2,1} = \left[\begin{array}{c|c} 0 & 1 \\ \hline 0 & 0 \end{array}\right] \quad \mathbf{W}^{3,1} = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & 1 \end{array}\right].$$

Thus, the contribution of t_1 to $\mathbf{R}'_{101,011}$ is $w_1^* \cdot \mathbf{W}^{1,1}_{1,0} \cdot \mathbf{W}^{2,1}_{0,1} \cdot \mathbf{W}^{3,1}_{1,1} = w_1^* > 0$, hence $\mathbf{R}'_{101,011} > 0$. However, given the initial marking, $\mathbf{m}^0 = \langle 1,0,0 \rangle$, the marking $\mathbf{m} = \langle 1,0,1 \rangle$ is not reachable. □

Theorem 3.1 gives a characterization of the infinitesimal generator $\mathbf{Q}$ and of the matrix $\mathbf{F}$ of a GSPN by focusing on the effect of each transition on each place. An alternative statement of this theorem, is,

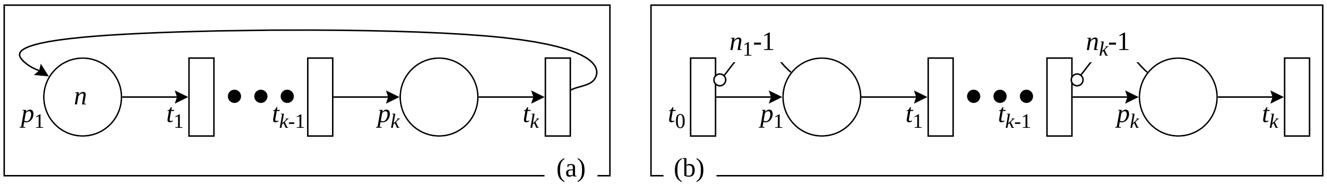

Figure 3: A GSPN where $|\mathcal{S}_T| \ll |\mathcal{S}'|$, and one where $|\mathcal{S}| = |\mathcal{S}'|$.

of course,

$$\begin{aligned}
\mathbf{Q} &= \mathbf{R}'_{\mathcal{S}_T,\mathcal{S}_T} + \mathbf{R}'_{\mathcal{S}_T,\mathcal{S}_V} \left(\mathbf{I}_{|\mathcal{S}_V|} - \boldsymbol{\Gamma}'^{-1}_{\mathcal{S}_V,\mathcal{S}_V} \cdot \mathbf{U}'_{\mathcal{S}_V,\mathcal{S}_V} \right)^{-1} \mathbf{U}'_{\mathcal{S}_V,\mathcal{S}_T} - \boldsymbol{\Lambda}'_{\mathcal{S}_T,\mathcal{S}_T} \\
\mathbf{F} &= \mathbf{R}'_{\mathcal{S}_T,\mathcal{S}_V} \left(\mathbf{I}_{|\mathcal{S}_V|} - \boldsymbol{\Gamma}'^{-1}_{\mathcal{S}_V,\mathcal{S}_V} \cdot \mathbf{U}'_{\mathcal{S}_V,\mathcal{S}_V} \right)^{-1} .
\end{aligned}$$

In either form, however, this result has little practical value in itself, since both expressions contain an inverse which cannot be expressed using Kronecker operators on smaller matrices. One case where Theorem 3.1 has a direct application is when there are no immediate transitions. Then, $\mathcal{S}_T = \mathcal{S}$, $\mathcal{S}_V = \emptyset$, and Eq. (2) simplifies to $\mathbf{Q} = \mathbf{R} - \boldsymbol{\Lambda} = \mathbf{R}_{T,T} - \boldsymbol{\Lambda} = \mathbf{R}'_{\mathcal{S}_T,\mathcal{S}_T} - \boldsymbol{\Lambda}'_{\mathcal{S}_T,\mathcal{S}_T}$.

However, a solution approach based on this idea alone has limitations, due to the restrictions that the GSPN must satisfy. Even more importantly, though, the size of $\mathbf{Q}'$ is enormous, potentially leading to inefficiencies. Consider for example the GSPN in Figure 3(a), having positive finite transitions rates. If the initial marking contains a total of n tokens,

$$|\mathcal{S}| = \begin{pmatrix} n+k-1 \\ n \end{pmatrix} \quad \ll \quad |\mathcal{S}'| = (n+1)^k.$$

From the simple case when $n = 1$, it is apparent that the difference, k vs. 2^k, can be enormous. For this type of closed networks, Buchholz [4] suggested a solution method based on Kronecker algebra that does not create unreachable states, applicable when the interaction between submodels is of the asynchronous type described in the introduction.

On the other hand, it is possible for $\mathcal{S}$ to equal $\mathcal{S}'$. This happens, for example, in a live free-choice GSPN with capacities whose undirected graph obtained by ignoring arc directions is acyclic (this is a generalization of [13, Property 3], which refers, however, to unbounded marked graphs). Another example is that of open acyclic queueing networks with communication blocking due to bounded buffers [17] which could be named "open state machines with capacities" in Petri net terminology. The transitions in these nets have *at most* one input and one output place and, if capacities were removed, every place would become unbounded. See the GSPN in Figure 3(b) for a simple example of a tandem network. Indeed, when $\mathcal{S}' = \mathcal{S}_T$, $\Psi_T(\mathbf{m})$ is simply the mixed-base value of $\mathbf{m}$, hence $\Psi_T^{-1}(k)$ does not have to be stored explicitly. Unfortunately, such a situation is rare.

3.2 Merging places into macroplaces

The type of marking dependence expressed by Eq. (6) and (7) is quite general, but for example, it does not let us specify a firing rate proportional to a nonlinear function of several places (e.g., $\min\{\mathbf{m}_1, \mathbf{m}_2\}$). We now show how this limitation can be overcome in practice by merging places (p_1 and p_2, in our example).

Consider a GSPN $A = (\mathcal{P}, \mathcal{T}, \mathcal{I}, \mathbf{C}^-, \mathbf{C}^+, \mathbf{m}^0, \mathbf{w})$ with finite reachability set $\mathcal{S}$, and partition $\mathcal{P}$ into $\hat{\mathcal{P}} = \{\hat{\mathcal{P}}_1, \dots \hat{\mathcal{P}}_{|\hat{\mathcal{P}}|}\}$, where $\hat{\mathcal{P}}_i = \{p_{i_1}, \dots p_{i_{|\hat{\mathcal{P}}_i|}}\}$. Then, define an order-preserving bijection $\gamma : \mathcal{S} \to \hat{\mathcal{S}} \subseteq \mathbb{N}^{|\hat{\mathcal{P}}|}$

$$\gamma\left(\mathbf{m}_1, \dots, \mathbf{m}_{|\mathcal{P}|}\right) = \left(\hat{\mathbf{m}}_1, \dots, \hat{\mathbf{m}}_{|\hat{\mathcal{P}}|}\right) \qquad \text{satisfying} \qquad \gamma(\mathbf{m}) \succ \gamma(\mathbf{n}) \iff \mathbf{m} \succ \mathbf{n}, \tag{17}$$

where $\hat{\mathbf{m}}_i$ is the position, in lexicographic order, of $(\mathbf{m}_{i_1}, \dots, \mathbf{m}_{i_{|\hat{\mathcal{P}}_i|}})$ in the set obtained by projecting $\mathcal{S}$ over $\hat{\mathcal{P}}_i$.

Lemma 3.2 Given A, $\hat{\mathcal{P}}$, and γ defined as above, consider the "compacted" GSPN

$$\hat{A} = (\hat{\mathcal{P}}, \hat{\mathcal{T}} = \mathcal{T}, \hat{\mathcal{I}} = \mathcal{I}, \hat{\mathbf{C}}^-, \hat{\mathbf{C}}^+, \hat{\mathbf{m}}^0 = \gamma(\mathbf{m}^0), \hat{\mathbf{w}}),$$

where the input and output arc cardinalities are defined to ensure that, in corresponding markings $\mathbf{m}$ and $\gamma(\mathbf{m}) = \hat{\mathbf{m}}$, $t_j \in \mathcal{T}$ has concession in $\hat{A}$ iff it has concession in A and that, in this case, $\hat{\mathbf{m}} \xrightarrow{t_j} \hat{\mathbf{n}} = \gamma(\mathbf{n})$ in $\hat{A}$ iff $\mathbf{m} \xrightarrow{t_j} \mathbf{n}$ in A , while the weights for t_j are defined to have the same value in corresponding markings:

- If $t_j \in \mathcal{E}(\mathbf{m})$ and its firing does not change the marking of any place in $\hat{\mathcal{P}}_i$, that is, if $\forall p_l \in \hat{\mathcal{P}}_i : \ \mathbf{C}^-_{l,j}(\mathbf{m}) \leq \mathbf{m}_l \wedge \mathbf{C}^-_{l,j}(\mathbf{m}) = \mathbf{C}^+_{l,j}(\mathbf{m})$, define $\hat{\mathbf{C}}^-_{i,j}(\hat{\mathbf{m}}) = \hat{\mathbf{C}}^+_{i,j}(\hat{\mathbf{m}}) = 0$.
- If $t_j \in \mathcal{E}(\mathbf{m})$ and its firing changes the marking of some place(s) in $\hat{\mathcal{P}}_i$, that is, if $\forall p_l \in \hat{\mathcal{P}}_i : \ \mathbf{C}^-_{l,j}(\mathbf{m}) \leq \mathbf{m}_l \wedge \exists p_l \in \hat{p}_i, \ \mathbf{C}^-_{l,j}(\mathbf{m}) \neq \mathbf{C}^+_{l,j}(\mathbf{m})$, define $\hat{\mathbf{C}}^-_{i,j}(\hat{\mathbf{m}}) = \hat{\mathbf{m}}_i$ and $\hat{\mathbf{C}}^+_{i,j}(\hat{\mathbf{m}}) = \hat{\mathbf{n}}_i$.
- Otherwise, t_j is disabled in $\mathbf{m}$, that is, $\exists p_l \in \hat{\mathcal{P}}_i, \ \mathbf{C}^-_{l,j}(\mathbf{m}) > \mathbf{m}_l$; then, define $\hat{\mathbf{C}}^-_{i,j}(\hat{\mathbf{m}}) = \hat{\mathbf{m}}_i + 1$, while the value of $\hat{\mathbf{C}}^+_{i,j}(\hat{\mathbf{m}})$ is irrelevant.
- Define $\hat{\mathbf{w}}_j(\hat{\mathbf{m}}) = \mathbf{w}_j(\mathbf{m})$.

Then, the transition rate matrices $\mathbf{R}$ and $\hat{\mathbf{R}}$, defined by A and $\hat{A}$, respectively, are identical.

Proof: Omitted for brevity (it is sufficient to show that the stochastic processes described by A and $\hat{A}$ are identical). □

Lemma 3.2 allows us to compact an arbitrary set of places into a single place, which, together with the transitions connected to it, corresponds to a sub-GSPN of [11, 12]. This operation *must* be performed when the marking dependencies in the GSPN are not of the type allowed by Theorem 3.1. It *might* be performed, even when the theorem is applicable, to reduce the number of matrices involved in the description of $\mathbf{R}$ at the cost of increasing their size.

From now on, macroplaces are indicated as dashed boxes surrounding sets of places; the compacted GSPNs are not shown explicitly, since they would not add to the comprehension of the model.

4 Timed synchronizing transitions

The contributions in [4, 5, 11, 12] have assumed that the GSPN is decomposed in such a way that each iteration of the solution method performs Kronecker products of few large (but manageable) matrices, while Theorem 3.1 uses many ($|\mathcal{T}| \cdot |\mathcal{P}|$) small ($n_i \times n_i$) matrices.

Lemma 3.2 addresses the size issue: we can merge places, thus obtaining $|\mathcal{T}| \cdot |\hat{\mathcal{P}}|$ larger matrices. In this section, we show how the number of matrices involved can be further reduced by merging transitions, or rather, the corresponding matrices. The results are similar to those derived by previous authors, who assumed all synchronizing transitions are timed, but we present them here for three reasons. First, we exhibit substantially different proofs for these results; [4, 5, 11, 12] consider a set of sub-GSPNs and combine them using synchronizing transitions, thus Kronecker operators are introduced only at the last step. Instead, we start from the Kronecker expression of Theorem 3.1 for the entire GSPN and derive our results by exploiting the properties of Kronecker operators. Second, our results include the management of immediate transitions and vanishing markings, while previous works have simply assumed that these are eliminated locally using the traditional approach. Finally and most importantly, our result apply to a larger class of GSPNs, since a more general marking-dependent behavior is allowed by Theorem 3.1.

4.1 Partitioning the set of transitions

Without loss of generality, we assume from now on that each transition in $\mathcal{T}$ has at least one input or one output place: $\forall t_j \in \mathcal{T}, \exists p_i \in \mathcal{P},\ \mathbf{C}^-_{i,j} \not\equiv 0\ \vee \mathbf{C}^+_{i,j} \not\equiv 0$. Then, let $\mathcal{T}^i \subseteq \mathcal{T}$ be the set of "local" transitions which affect, or are affected by, only a single place p_i:

$$\forall p_i \in \mathcal{P} :\ \mathcal{T}^i = \left\{ t_j \in \mathcal{T} \mid \forall p_l \in \mathcal{P}, p_l \neq p_i,\ \mathbf{C}^-_{l,j} \equiv \mathbf{C}^+_{l,j} \equiv 0\ \wedge\ w_{l,j} \equiv 1 \right\}, \tag{18}$$

and $\mathcal{T}^\bullet = \mathcal{T} \setminus \bigcup_{p_i \in \mathcal{P}} \mathcal{T}^i$ be the set of synchronizing transitions which instead affect or are affected by at least two places. Clearly, these sets constitute a partition of $\mathcal{T}$. Also, let $\mathcal{X}^\bullet = \mathcal{T}^\bullet \cap \mathcal{X}$, $\mathcal{I}^\bullet = \mathcal{T}^\bullet \cap \mathcal{I}$, $\mathcal{X}^i = \mathcal{T}^i \cap \mathcal{X}$, and $\mathcal{I}^i = \mathcal{T}^i \cap \mathcal{I}$.

Lemma 4.1 Consider a GSPN satisfying the requirements of Theorem 3.1. Then,

$$\mathbf{R}' = \sum_{t_j \in \mathcal{X}^\bullet} w^*_j \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{W}^{i,j} + \bigoplus_{p_i \in \mathcal{P}} \mathbf{R}^i \quad , \quad \mathbf{U}' = \sum_{t_j \in \mathcal{I}^\bullet} w^*_j \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{W}^{i,j} + \bigoplus_{p_i \in \mathcal{P}} \mathbf{U}^i, \tag{19}$$

$$\boldsymbol{\Lambda}' = \sum_{t_j \in \mathcal{X}^\bullet} w^*_j \cdot \bigotimes_{p_i \in \mathcal{P}} \boldsymbol{\Lambda}^{i,j} + \bigoplus_{p_i \in \mathcal{P}} \boldsymbol{\Lambda}^i \quad , \quad \boldsymbol{\Gamma}' = \sum_{t_j \in \mathcal{I}^\bullet} w^*_j \cdot \bigotimes_{p_i \in \mathcal{P}} \boldsymbol{\Gamma}^{i,j} + \bigoplus_{p_i \in \mathcal{P}} \boldsymbol{\Gamma}^i, \tag{20}$$

where $\mathbf{R}^i$, $\mathbf{U}^i$, $\boldsymbol{\Lambda}^i$, and $\boldsymbol{\Gamma}^i$ are square matrices of size $n_i \times n_i$ defined as

$$\mathbf{R}^i = \sum_{t_j \in \mathcal{X}^i} w^*_j \cdot \mathbf{W}^{i,j} \quad , \quad \mathbf{U}^i = \sum_{t_j \in \mathcal{I}^i} w^*_j \cdot \mathbf{W}^{i,j} \quad , \quad \boldsymbol{\Lambda}^i = \sum_{t_j \in \mathcal{X}^i} w^*_j \cdot \boldsymbol{\Lambda}^{i,j} \quad , \quad \boldsymbol{\Gamma}^i = \sum_{t_j \in \mathcal{I}^i} w^*_j \cdot \boldsymbol{\Gamma}^{i,j}.$$

Hence, as special cases, $\mathbf{R}^i = \boldsymbol{\Lambda}^i = \mathbf{0}$ if $\mathcal{X}^i = \emptyset$ and $\mathbf{U}^i = \boldsymbol{\Gamma}^i = \mathbf{0}$ if $\mathcal{I}^i = \emptyset$.

Proof: We only prove the result for $\mathbf{R}'$, the proof for the other matrices is analogous. Given the condition specified by Eq. (18), we know that $\mathbf{W}^{l,j} = \mathbf{I}_{n_l}$ if $p_i \neq p_l$ and $t_j \in \mathcal{T}^i$. Then, the proof is a simple matter of matrix manipulation within the Kronecker expressions:

$$\begin{aligned}
\mathbf{R}' &= \sum_{t_j \in \mathcal{X}} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{W}^{i,j} \\
&= \sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{W}^{i,j} + \sum_{p_i \in \mathcal{P}} \sum_{t_j \in \mathcal{X}^i} w_j^* \cdot \bigotimes_{p_l \in \mathcal{P}} \mathbf{W}^{l,j} \\
&= \sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{W}^{i,j} + \sum_{p_i \in \mathcal{P}} \sum_{t_j \in \mathcal{X}^i} w_j^* \cdot \mathbf{I}_{n_1} \otimes \cdots \otimes \mathbf{I}_{n_{i-1}} \otimes \mathbf{W}^{i,j} \otimes \mathbf{I}_{n_{i+1}} \otimes \cdots \otimes \mathbf{I}_{n_{|\mathcal{P}|}} \\
&= \sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{W}^{i,j} + \sum_{p_i \in \mathcal{P}} \mathbf{I}_{n_1} \otimes \cdots \otimes \mathbf{I}_{n_{i-1}} \otimes \left(\sum_{t_j \in \mathcal{X}^i} w_j^* \cdot \mathbf{W}^{i,j} \right) \otimes \mathbf{I}_{n_{i+1}} \otimes \cdots \otimes \mathbf{I}_{n_{|\mathcal{P}|}} \\
&= \sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{W}^{i,j} + \sum_{p_i \in \mathcal{P}} \mathbf{I}_{n_1} \otimes \cdots \otimes \mathbf{I}_{n_{i-1}} \otimes \mathbf{R}^i \otimes \mathbf{I}_{n_{i+1}} \otimes \cdots \otimes \mathbf{I}_{n_{|\mathcal{P}|}} \\
&= \sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{W}^{i,j} + \bigoplus_{p_i \in \mathcal{P}} \mathbf{R}^i. \qquad \square
\end{aligned}$$

This partition reduces the number of Kronecker product terms from $|\mathcal{X}|$ to $|\mathcal{X}^\bullet|$ in $\mathbf{R}'$ and from $|\mathcal{I}|$ to $|\mathcal{I}^\bullet|$ in $\mathbf{U}'$, respectively, and adds one Kronecker sum to both. Transitions satisfying Eq. (18) arise after applying Lemma 3.2, that is, after "decomposing" a large GSPN into several smaller sub-GSPNs. Each sub-GSPN corresponds, in our terminology, to a (macro)place, plus the set of transitions local to it. This transformation does not have to be explicitly performed in practice, only its result, the matrices corresponding to the set of macroplaces, need to be computed. A good partition of the places results in a compacted GSPN where most transitions are local and the number of tokens in each compacted place (number of markings in the sub-GSPN, in the terminology of [11, 12]), is manageable. Methods to determine a good partition are beyond the scope of this paper and are left for future research.

4.2 An efficient Kronecker expression for the CTMC

Given any GSPN, we can always apply Lemma 3.2, resulting in a compacted GSPN satisfying the requirements of Theorem 3.1, then apply Lemma 4.1. If the partition is such that all immediate transitions are local, we can then restate the main results of [11, 12] in a more general setting.

Theorem 4.1 Consider a GSPN satisfying the same requirements as for Theorem 3.1, and such that all immediate transitions are local:

$$\mathcal{I}^\bullet = \emptyset \quad \Rightarrow \quad \mathbf{U}' = \bigoplus_{p_i \in \mathcal{P}} \mathbf{U}^i, \qquad \mathcal{T}^\bullet = \mathcal{X}^\bullet. \tag{21}$$

Then

$$\mathbf{Q}'' = \sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \left(\mathbf{W}^{i,j} \cdot \mathbf{X}^i \right) + \bigoplus_{p_i \in \mathcal{P}} \left(\mathbf{R}^i \cdot \mathbf{X}^i \right) - \mathit{\Lambda}',$$

where

$$\mathbf{X}^i = \left(\mathbf{I}_{n_i} - (\mathbf{T}^i + \boldsymbol{\Gamma}^i)^{-1}\mathbf{U}^i\right)^{-1} \qquad \boldsymbol{\Gamma}^i = \mathrm{rowsum}(\mathbf{U}^i) \qquad \mathbf{T}^i = \mathbf{I}_{n_i} - \delta(\boldsymbol{\Gamma}^i),$$

satisfies $\mathbf{Q}''_{\mathcal{S}_T,\mathcal{S}_T} = \mathbf{Q}$, with the meaning defined in Eq. (2).

Proof: First, we observe that condition Eq. (21) implies $\boldsymbol{\Gamma}' = \bigoplus_{p_i \in \mathcal{P}} \boldsymbol{\Gamma}^i$ and $\mathbf{T}' = \bigotimes_{p_i \in \mathcal{P}} \mathbf{T}^i$. In other words, a "global" marking $\mathbf{m}$ is tangible iff all its "local" components are tangible. We can then manipulate $\mathbf{Q}''$ as follows:

$$\begin{aligned}
\mathbf{Q}'' &= \sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \left(\mathbf{W}^{i,j} \cdot \mathbf{X}^i\right) + \bigoplus_{p_i \in \mathcal{P}} \left(\mathbf{R}^i \cdot \mathbf{X}^i\right) - \boldsymbol{\Lambda}' \\
&= \sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{W}^{i,j} \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{X}^i + \bigoplus_{p_i \in \mathcal{P}} \mathbf{R}^i \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{X}^i + \underbrace{\bigoplus_{p_i \in \mathcal{P}} \left(\mathbf{R}^i \cdot \mathbf{X}^i\right) - \bigoplus_{p_i \in \mathcal{P}} \mathbf{R}^i \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{X}^i}_{\mathbf{D}'} - \boldsymbol{\Lambda}' \\
&= \left(\sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{W}^{i,j} + \bigoplus_{p_i \in \mathcal{P}} \mathbf{R}^i\right) \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{X}^i + \mathbf{D}' - \boldsymbol{\Lambda}' \\
&= \mathbf{R}' \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{X}^i + \mathbf{D}' - \boldsymbol{\Lambda}'
\end{aligned}$$

Partition $\mathcal{S}^i$ into $\mathcal{S}^i_T$ and $\mathcal{S}^i_V$, corresponding to local markings enabling only timed local transitions, or some immediate local transition, respectively, and rearrange the rows and columns of X^i accordingly:

$$\mathbf{X}^i = \left(\mathbf{I}_{n_i} - (\mathbf{T}^i + \boldsymbol{\Gamma}^i)^{-1}\mathbf{U}^i\right)^{-1} = \left[\begin{array}{c|c} \mathbf{I}^i_{T,T} & \mathbf{0} \\ \hline \mathbf{P}^i_{V,T} & \mathbf{N}^i_{V,V} \end{array}\right],$$

where the subscripts "T,T", "V,T", and "V,V" have the usual meaning, but applied to the local matrix for place i. $\mathbf{N}^i_{V,V} = \left(\mathbf{I}^i_{\mathcal{S}_V,\mathcal{S}_V} - (\boldsymbol{\Gamma}^i_{\mathcal{S}_V,\mathcal{S}_V})^{-1} \cdot \mathbf{U}^i_{\mathcal{S}_V,\mathcal{S}_V}\right)^{-1}$ describes the expected number of visits to each local vanishing marking before reaching a local tangible marking, starting from each each local vanishing marking, while $\mathbf{P}^i_{V,T} = \mathbf{N}^i_{\mathcal{S}_V,\mathcal{S}_V} \cdot (\boldsymbol{\Gamma}^i_{\mathcal{S}_V,\mathcal{S}_V})^{-1} \cdot \mathbf{U}^i_{\mathcal{S}_V,\mathcal{S}_T}$ describes the probability of reaching each local tangible marking, starting from each local vanishing marking.

We continue assuming that $|\mathcal{P}| = 2$, the general proof follows exactly the same idea. Local matrices are partitioned according to whether the corresponding local markings are tangible or vanishing (regardless of whether the global markings are reachable or not). Global matrices are partitioned according to the following order: tangible states, vanishing states enabling only immediate transitions in $\mathcal{T}^2$, vanishing states enabling only immediate transitions in $\mathcal{T}^1$, and vanishing states enabling immediate transitions in both $\mathcal{T}^1$ and $\mathcal{T}^2$. First, we show that the tangible rows of $\mathbf{D}'$ are zero:

$$\begin{aligned}
\mathbf{D}' &= (\mathbf{R}^1\mathbf{X}^1 \oplus \mathbf{R}^2\mathbf{X}^2) - (\mathbf{R}^1 \oplus \mathbf{R}^2)(\mathbf{X}^1 \otimes \mathbf{X}^2) \\
&= \mathbf{R}^1\mathbf{X}^1 \otimes \mathbf{I}_{n_2} + \mathbf{I}_{n_1} \otimes \mathbf{R}^2\mathbf{X}^2 - \mathbf{R}^1\mathbf{X}^1 \otimes \mathbf{X}^2 - \mathbf{X}^1 \otimes \mathbf{R}^2\mathbf{X}^2
\end{aligned}$$

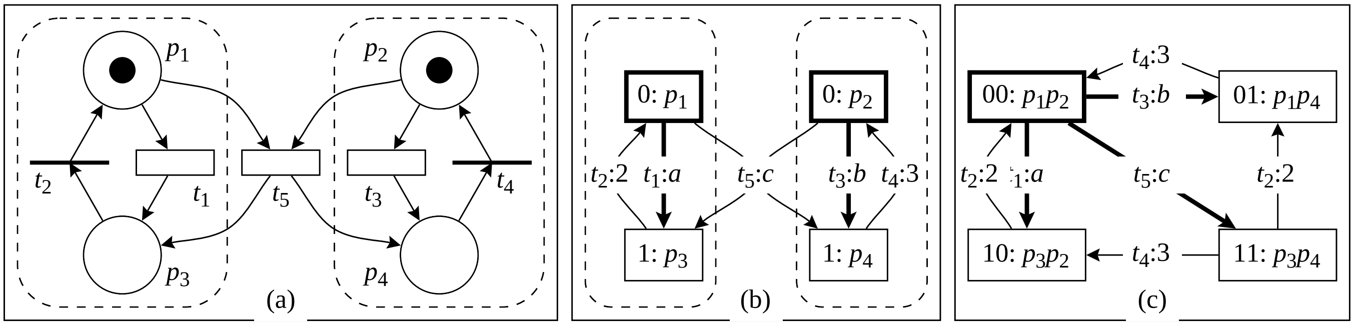

Figure 4: A GSPN to illustrate the difference between $\mathbf{Q}'$ and $\mathbf{Q}''$.

$$
\begin{aligned}
&= \mathbf{R}^1\mathbf{X}^1 \otimes (\mathbf{I}_{n_2} - \mathbf{X}^2) + (\mathbf{I}_{n_1} - \mathbf{X}^1) \otimes \mathbf{R}^2\mathbf{X}^2 \\
&= \mathbf{R}^1\mathbf{X}^1 \otimes \left[\begin{array}{c|c} \mathbf{0} & \mathbf{0} \\ \hline \bullet & \bullet \end{array}\right] + \left[\begin{array}{c|c} \mathbf{0} & \mathbf{0} \\ \hline \bullet & \bullet \end{array}\right] \otimes \mathbf{R}^2\mathbf{X}^2 \\
&= \left[\begin{array}{c|c|c|c} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline \bullet & \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet & \bullet \\ \hline \bullet & \bullet & \bullet & \bullet \end{array}\right],
\end{aligned}
\tag{22}
$$

since $\mathbf{X}^i$ and $\mathbf{I}_{n_i}$ have the same tangible rows. Then, we show that the tangible columns of $\bigotimes_{p_i \in \mathcal{P}} \mathbf{X}^i$ and $(\mathbf{I} - (\mathbf{T}' + \boldsymbol{\Gamma}')^{-1}\mathbf{U}')^{-1}$ coincide:

$$
\begin{aligned}
\mathbf{X}^1 \otimes \mathbf{X}^2 &= \left[\begin{array}{c|c} \mathbf{I}^1_{T,T} & \mathbf{0} \\ \hline \mathbf{P}^1_{V,T} & \mathbf{N}^1_{V,V} \end{array}\right] \otimes \left[\begin{array}{c|c} \mathbf{I}^2_{T,T} & \mathbf{0} \\ \hline \mathbf{P}^2_{V,T} & \mathbf{N}^2_{V,V} \end{array}\right] \\
&= \left[\begin{array}{c|c|c|c} \mathbf{I}^1_{T,T} \otimes \mathbf{I}^2_{T,T} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{I}^1_{T,T} \otimes \mathbf{P}^2_{V,T} & \mathbf{I}^1_{T,T} \otimes \mathbf{N}^2_{V,V} & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{P}^1_{V,T} \otimes \mathbf{I}^2_{T,T} & \mathbf{0} & \mathbf{N}^1_{V,V} \otimes \mathbf{I}^2_{T,T} & \mathbf{0} \\ \hline \mathbf{P}^1_{V,T} \otimes \mathbf{P}^2_{V,T} & \mathbf{P}^1_{V,T} \otimes \mathbf{N}^2_{V,V} & \mathbf{N}^1_{V,V} \otimes \mathbf{P}^2_{V,T} & \mathbf{N}^1_{V,V} \otimes \mathbf{N}^2_{V,V} \end{array}\right].
\end{aligned}
\tag{23}
$$

The blocks in the first (tangible) column of this last matrix contain the correct values, since the top left block is simply the identity and the other blocks correctly describe the probabilities of reaching tangible markings from vanishing markings, which are the values in the corresponding blocks of $(\mathbf{I} - (\mathbf{T}' + \boldsymbol{\Gamma}')^{-1}\mathbf{U}')^{-1}$. From Eq. (22) and (23) we can then conclude that $\mathbf{Q}''_{\mathcal{S}_T,\mathcal{S}_T} = \mathbf{Q}$, as in the proof of Theorem 3.1. □.

In practice, Theorem 4.1 is used to generate only the relevant portion of $\mathbf{Q}''$ in the numerical solution method. In other words, we eliminate the vanishing markings "on the fly" (as in [4, 5, 10, 11, 12])

$$
\begin{aligned}
\mathbf{Q} = \mathbf{Q}''_{\mathcal{S}_T,\mathcal{S}_T} &= \sum_{t_j \in \mathcal{X}^\bullet} w^*_j \cdot \bigotimes_{p_i \in \mathcal{P}} \left(\mathbf{W}^{i,j} \cdot \mathbf{X}^i\right)_{T,T} + \bigoplus_{p_i \in \mathcal{P}} \left(\mathbf{R}^i \cdot \mathbf{X}^i\right)_{T,T} \\
&\quad - \left(\sum_{t_j \in \mathcal{X}^\bullet} w^*_j \cdot \bigotimes_{p_i \in \mathcal{P}} \boldsymbol{\Lambda}^{i,j}_{T,T} + \bigoplus_{p_i \in \mathcal{P}} \boldsymbol{\Lambda}^i_{T,T}\right).
\end{aligned}
$$

We observe that both $\mathbf{Q}'$ and $\mathbf{Q}''$ describe $\mathbf{Q}$, but the two normally differ. For example, consider the GSPN in Fig. 4, having a single tangible state and three vanishing states (the example is trivial, but it is sufficient to illustrate the point). Assuming that the rates of timed transitions t_1, t_3, and t_5 are a, b, and c, respectively, and that the weights of transitions t_2 and t_4 are 2 and 3:

$$\mathbf{R}^1 = \begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix} \quad \mathbf{R}^2 = \begin{bmatrix} 0 & b \\ 0 & 0 \end{bmatrix} \quad \mathbf{W}^{1,5} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \mathbf{W}^{2,5} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \mathbf{U}^1 = \begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix} \quad \mathbf{U}^2 = \begin{bmatrix} 0 & 0 \\ 3 & 0 \end{bmatrix}.$$

From which we obtain

$$\mathbf{W}^1 = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \quad \mathbf{W}^2 = \begin{bmatrix} 0 & 0 \\ 0 & 3 \end{bmatrix} \quad \mathbf{T}^1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \mathbf{T}^2 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \mathbf{X}^1 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \quad \mathbf{X}^2 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$$

and

$$\mathbf{R}' = \begin{bmatrix} 0 & b & a & c \\ 0 & 0 & 0 & a \\ 0 & 0 & 0 & b \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{T}' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{U}' = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 2 & 3 & 0 \end{bmatrix} \qquad \boldsymbol{\mathit{\Gamma}}' = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}.$$

The resulting $\mathbf{Q}'$ and $\mathbf{Q}''$ matrices are then different:

$$\mathbf{Q}' = \begin{bmatrix} a+b+c & b+\frac{2}{5}c & a+\frac{3}{5}c & c \\ a & \frac{2}{5}a & \frac{3}{5}a & a \\ b & \frac{2}{5}b & \frac{3}{5}b & b \\ 0 & 0 & 0 & 0 \end{bmatrix} - \boldsymbol{\mathit{\Lambda}}' \qquad \mathbf{Q}'' = \begin{bmatrix} a+b+c & b+c & a+c & c \\ 0 & a & 0 & a \\ 0 & 0 & b & b \\ 0 & 0 & 0 & 0 \end{bmatrix} - \boldsymbol{\mathit{\Lambda}}'.$$

Indeed, the difference between $\mathbf{Q}'$ and $\mathbf{Q}''$ is already apparent from Eq. (23). The diagonal blocks $\mathbf{I}^1_{T,T} \otimes \mathbf{N}^2_{V,V}$ and $\mathbf{N}^1_{V,V} \otimes \mathbf{I}^2_{T,T}$ correctly describe the expected number of times the corresponding global vanishing markings (enabling only local immediate transitions in $\mathcal{T}^2$ or $\mathcal{T}^1$, respectively) are entered, given that a timed transition firing leads to the corresponding diagonal block. However, the last three blocks on the bottom row do not reflect the same quantities when a timed transition firing leads to vanishing markings enabling immediate transitions in both $\mathcal{T}^1$ and $\mathcal{T}^2$. In particular, $\mathbf{P}^1_{V,T} \otimes \mathbf{N}^2_{V,V}$ describes the correct quantity only if we could assume that all enabled immediate transitions in $\mathcal{T}^1$ keep firing before any of those in $\mathcal{T}^2$ do, which is not necessarily the case, while $\mathbf{N}^1_{V,V} \otimes \mathbf{P}^2_{V,T}$ assumes the opposite. Finally, $\mathbf{N}^1_{V,V} \otimes \mathbf{N}^2_{V,V}$ does not reflect the number of times global markings are entered at all. This leads us to the following observation.

Corollary 4.1 If the GPSN of Theorem 4.1 is such that the firing of any timed (synchronizing) transition $t_j \in \mathcal{X}^\bullet$ enables (local) immediate transitions in at most one set $\mathcal{T}^i$, for some $p_i \in \mathcal{P}$, then $\mathbf{Q}''_{\mathcal{S}_T,\mathcal{S}_V} = \mathbf{F}$, as defined in Eq. (5).

Proof: The condition for this corollary implies

$$\mathcal{S}_V \subseteq \bigcup_{p_i \in \mathcal{P}} \mathcal{S}^1_T \times \cdots \times \mathcal{S}^{i-1}_T \times \mathcal{S}^i_V \times \mathcal{S}^{i+1}_T \times \cdots \times \mathcal{S}^{|\mathcal{P}|}_T.$$

As discussed in Theorem 4.1, the blocks on the columns corresponding to this type of vanishing markings are computed correctly in $\bigotimes_{p_i \in \mathcal{P}} \mathbf{X}^i$, see Eq. (23). The bottom rows of Eq. 23, corresponding to any (unreachable) vanishing markings $\mathbf{m}$ enabling immediate transitions in more than one set $\mathcal{T}^i$, are irrelevant, since Lemma 3.1 guarantees that $\mathbf{R}'_{\mathcal{S}_T,\mathbf{m}} = \mathbf{0}$ in this case. □

5 Immediate synchronizing transitions

If some of the synchronizing transitions are immediate, Theorem 3.1 still applies, but Theorem 4.1, which allows the efficient computation of the solution in practice, does not. In this section, we show how "preservation of the vanishing markings" [8] can be used to remove this limitation, also present in [11, 12].

5.1 Embedding a DTMC

First, we summarize the main ideas in [8], which examines an alternate method to compute $\boldsymbol{\pi}$:

- Define the transition probability matrix $\mathbf{P}$ of the *embedded* DTMC, expressing the probability of going, in one firing, from any marking $\mathbf{m} \in \mathcal{S}$ to any other marking $\mathbf{n} \in \mathcal{S}$, regardless of whether they are tangible or vanishing:

$$\mathbf{P} = \left[\begin{array}{c|c} \boldsymbol{\Lambda}^{-1}\mathbf{R}_{T,T} & \boldsymbol{\Lambda}^{-1}\mathbf{R}_{T,V} \\ \hline \mathbf{U}_{V,T} & \mathbf{U}_{V,V} \end{array} \right]. \tag{24}$$

- Compute the steady-state probability vector $\boldsymbol{\gamma} \in \mathbb{R}^{|\mathcal{S}|}$ of the embedded DTMC:

$$\boldsymbol{\gamma} \cdot \mathbf{P} = \boldsymbol{\gamma} \qquad \text{subject to the normalization} \qquad \boldsymbol{\gamma} \cdot \mathbf{1}_{|\mathcal{S}|\times 1} = 1. \tag{25}$$

- Obtain both $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ from $\boldsymbol{\gamma}$, using the holding times in the tangible markings as weights:

$$\forall \mathbf{m} \in \mathcal{S}_T,\ \boldsymbol{\pi}_{\mathbf{m}} = \frac{\gamma_{\mathbf{m}} \cdot \mathbf{h}_{\mathbf{m}}}{\sum_{\mathbf{n}\in\mathcal{S}_T} \gamma_{\mathbf{n}} \cdot \mathbf{h}_{\mathbf{n}}} \qquad \text{and} \qquad \forall \mathbf{m} \in \mathcal{S},\ \boldsymbol{\phi}_{\mathbf{m}} = \frac{\gamma_{\mathbf{m}}}{\sum_{\mathbf{n}\in\mathcal{S}_T} \gamma_{\mathbf{n}} \cdot \mathbf{h}_{\mathbf{n}}}. \tag{26}$$

The correctness of the method can be verified by observing that, from

$$[\boldsymbol{\gamma}_T \mid \boldsymbol{\gamma}_V] \cdot \left[\begin{array}{c|c} \boldsymbol{\Lambda}^{-1}\mathbf{R}_{T,T} & \boldsymbol{\Lambda}^{-1}\mathbf{R}_{T,V} \\ \hline \mathbf{U}_{V,T} & \mathbf{U}_{V,V} \end{array} \right] = [\boldsymbol{\gamma}_T \mid \boldsymbol{\gamma}_V],$$

we can obtain $\boldsymbol{\gamma}_V = \boldsymbol{\gamma}_T \boldsymbol{\Lambda}^{-1}\mathbf{R}_{T,V}\left(\mathbf{I} - \mathbf{U}_{V,V}\right)^{-1}$, and, substituting it in the above equation,

$$\underbrace{\boldsymbol{\gamma}_T \boldsymbol{\Lambda}^{-1}}_{\boldsymbol{\pi}\cdot\left(\boldsymbol{\gamma}_T\cdot\boldsymbol{\Lambda}^{-1}\cdot\mathbf{1}_{|\mathcal{S}|\times 1}\right)} \cdot \underbrace{\left(\mathbf{R}_{T,T} + \mathbf{R}_{T,V}\left(\mathbf{I} - \mathbf{U}_{V,V}\right)^{-1}\mathbf{U}_{V,T}\right)}_{\mathbf{Q}} = \mathbf{0}.$$

We can then divide both sides of the equation by the constant $\boldsymbol{\gamma}_T \cdot \boldsymbol{\Lambda}^{-1} \cdot \mathbf{1}_{|\mathcal{S}|\times 1}$, resulting in $\boldsymbol{\pi} \cdot \mathbf{Q} = \mathbf{0}$.

In [8], it was found that the solution time is often greater than with the "elimination" approach based on Eq. (3). This is due to the number of nonzero entries in $\mathbf{P}$, normally larger than in $\mathbf{Q}$, and, frequently, to a slower numerical convergence. However, pathological cases where $\mathbf{P}$ has substantially fewer entries than $\mathbf{Q}$ arise when N tangible markings can reach a small set of vanishing markings, which can, in turn reach M tangible markings. This "N-to-M switch" behavior corresponds to $O(N+M)$ arcs in $\mathbf{P}$ and $O(N \cdot M)$ in $\mathbf{Q}$, hence it affects the elimination approach negatively both in terms of storage and execution time, although the number of iterations in the numerical solution might still be smaller with elimination. The use of preservation results in the following analog of Theorem 3.1.

Theorem 5.1 Under the same conditions of Theorem 3.1, the matrix

$$\mathbf{P}' = (\mathbf{T}' \cdot \boldsymbol{\Lambda}' + \boldsymbol{\Gamma}')^{-1} \cdot (\mathbf{T}' \cdot \mathbf{R}' + \mathbf{U}') \tag{27}$$

satisfies $\mathbf{P}'_{\mathcal{S},\mathcal{S}} = \mathbf{P}$, as defined in Eq. (24).

Proof: The pre-multiplication of $\boldsymbol{\Lambda}'$ and $\mathbf{R}'$ by $\mathbf{T}'$ eliminates the effect of timed transitions having concession in vanishing markings. However, if we focus on the reachable states, the statement of the theorem is equivalent to saying that

$$\mathbf{P} = \left[\frac{{\boldsymbol{\Lambda}'_{\mathcal{S}_T,\mathcal{S}_T}}^{-1} \cdot \mathbf{R}'_{\mathcal{S}_T,\mathcal{S}}}{{\boldsymbol{\Gamma}'_{\mathcal{S}_V,\mathcal{S}_V}}^{-1} \cdot \mathbf{U}'_{\mathcal{S}_V,\mathcal{S}}} \right]. \tag{28}$$

This equality then follows from the definition of $\mathbf{P}$ and the meaning of $\mathbf{R}'$ and $\mathbf{U}'$ already established in Theorem 3.1. Eq. 28 is, of course, the expression used in practice for a numerical solution. □

The efficiency of a solution based on Eq. (28) is improved by exploiting the existence of local transitions (Lemma 4.1):

$$\mathbf{P} = \left[\frac{\left(\sum_{t_j \in \mathcal{X}^\bullet} w_j^* \bigotimes_{P \in \mathcal{P}} \boldsymbol{\Lambda}^{i,j} + \bigoplus_{p_i \in \mathcal{P}} \boldsymbol{\Lambda}^i \right)^{-1}_{\mathcal{S}_T,\mathcal{S}_T} \cdot \left(\sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{W}^{i,j} + \bigoplus_{p_i \in \mathcal{P}} \mathbf{R}^i \right)_{\mathcal{S}_T,\mathcal{S}}}{\left(\sum_{t_j \in \mathcal{I}^\bullet} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \boldsymbol{\Gamma}^{i,j} + \bigoplus_{p_i \in \mathcal{P}} \boldsymbol{\Gamma}^i \right)^{-1}_{\mathcal{S}_V,\mathcal{S}_V} \cdot \left(\sum_{t_j \in \mathcal{I}^\bullet} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \mathbf{W}^{i,j} + \bigoplus_{p_i \in \mathcal{P}} \mathbf{U}^i \right)_{\mathcal{S}_V,\mathcal{S}}} \right], \tag{29}$$

since this expression for $\mathbf{P}$ reduces the number of Kronecker products to be performed at each iteration.

5.2 Using partial elimination to improve solution efficiency

A disadvantage of the approach just described is that the size of the probability vector $\boldsymbol{\gamma}$ for the DTMC is now $|\mathcal{S}|$, considerably larger than $|\mathcal{S}_T|$ in many practical models. Analogously, the size of the matrices for place i is given by the projection of $\mathcal{S}$ onto its i-th component, $\mathcal{S}^i = \{l : \exists \mathbf{m} \in \mathcal{S}, \mathbf{m}_i = l\}$, regardless of whether markings satisfying $\mathbf{m}_i = l$ are vanishing or tangible.

It was already observed in [8] that it is possible to eliminate a subset of the vanishing markings, preserving only those involved in large switches, in the hope of achieving the best memory–execution tradeoff. We can exploit the same idea in our approach, but for a different purpose. Partition the local state space for place i, $\mathcal{S}^i$, into

- $\mathcal{S}^i_T = \{l : \forall \mathbf{m} \in \mathcal{S}_T, \mathbf{m}_i = l\}$, the set of local tangible markings.
- $\mathcal{S}^i_S = \{l : \exists \mathbf{m} \in \mathcal{S}_V, \mathbf{m}_i = l \ \wedge \ \mathcal{I}^\bullet \cap \mathcal{E}(\mathbf{m}) \neq \emptyset\}$, the set of possibly synchronized local vanishing markings.
- $\mathcal{S}^i_L = \{l : \forall \mathbf{m} \in \mathcal{S}_V, \mathbf{m}_i = l \Rightarrow \mathcal{I}^\bullet \cap \mathcal{E}(\mathbf{m}) = \emptyset\}$, the set of non-synchronized local vanishing markings.

Any immediate synchronizing transition can be enabled only in markings having components in $\mathcal{S}^i_T \cup \mathcal{S}^i_S$, that is, in $\mathcal{S}_S = \mathcal{S}_V \cap \left((\mathcal{S}^1_T \cup \mathcal{S}^1_S) \times \cdots \times (\mathcal{S}^{|\mathcal{P}|}_T \cup \mathcal{S}^{|\mathcal{P}|}_S)\right)$.

We now define a "partially eliminated" (or "partially preserved") DTMC with transition probability matrix $\tilde{\mathbf{P}}$ and state space $\mathcal{S}_K = \mathcal{S}_T \cup \mathcal{S}_S$, ($K$ stands for 'keep") which can be used to compute $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ more efficiently than from $\mathbf{P}$. Partition the matrices $\mathbf{W}^{i,j}$, $\mathbf{R}^i$, $\mathbf{U}^i$, $\boldsymbol{\Lambda}^{i,j}$, $\boldsymbol{\Gamma}^{i,j}$, $\boldsymbol{\Lambda}^i$, and $\boldsymbol{\Gamma}^i$, according to the sets $\mathcal{S}^i_K = \mathcal{S}^i_T \cup \mathcal{S}^i_S$ and $\mathcal{S}^i_L$. For example,

$$\mathbf{R}^i = \left[\begin{array}{c|c} \mathbf{R}^i_{K,K} & \mathbf{R}^i_{K,L} \\ \hline \mathbf{R}^i_{L,K} & \mathbf{R}^i_{L,L} \end{array}\right].$$

Then, assuming that the rows $\left[\mathbf{U}^i_{L,K} | \mathbf{U}^i_{L,L}\right]$ are already normalized (this can be easily enforced since each $\mathbf{U}^i$ is built before starting the overall solution), define the matrices

$$\begin{aligned}
\tilde{\mathbf{W}}^{i,j} &= \mathbf{W}^{i,j}_{K,K} + \mathbf{W}^{i,j}_{K,L} \cdot \left(\mathbf{I} - \mathbf{U}^i_{L,L}\right)^{-1} \cdot \mathbf{U}^i_{L,K}, \\
\tilde{\mathbf{R}}^i &= \mathbf{R}^i_{K,K} + \mathbf{R}^i_{K,L} \cdot \left(\mathbf{I} - \mathbf{U}^i_{L,L}\right)^{-1} \cdot \mathbf{U}^i_{L,K}, \text{and} \\
\tilde{\mathbf{U}}^i &= \mathbf{U}^i_{K,K} + \mathbf{U}^i_{K,L} \cdot \left(\mathbf{I} - \mathbf{U}^i_{L,L}\right)^{-1} \cdot \mathbf{U}^i_{L,K}.
\end{aligned}$$

Given this definition, the blocks for the rows and columns of $\mathcal{S}_K$ in $\boldsymbol{\Lambda}^{i,j}$, $\boldsymbol{\Gamma}^{i,j}$, $\boldsymbol{\Lambda}^i$, and $\boldsymbol{\Gamma}^i$ still contain the correct row sums for the corresponding "~" matrices. Then, we can state our final theorem, which allows the efficient solution of a structured GSPN with immediate synchronizing transitions.

Theorem 5.2 Under the same conditions of Theorem 3.1, define the transition probability matrix

$$\tilde{\mathbf{P}} = \left[\begin{array}{c}
\left(\sum_{t_j \in \mathcal{X}^\bullet} w^*_j \bigotimes_{p_i \in \mathcal{P}} \boldsymbol{\Lambda}^{i,j}_{K,K} + \bigoplus_{p_i \in \mathcal{P}} \boldsymbol{\Lambda}^i_{K,K}\right)^{-1}_{\mathcal{S}_T,\mathcal{S}_T} \cdot \left(\sum_{t_j \in \mathcal{X}^\bullet} w^*_j \cdot \bigotimes_{p_i \in \mathcal{P}} \tilde{\mathbf{W}}^{i,j} + \bigoplus_{p_i \in \mathcal{P}} \tilde{\mathbf{R}}^i\right)_{\mathcal{S}_T,\mathcal{S}_K} \\ \hline
\left(\sum_{t_j \in \mathcal{I}^\bullet} w^*_j \cdot \bigotimes_{p_i \in \mathcal{P}} \boldsymbol{\Gamma}^{i,j}_{K,K} + \bigoplus_{p_i \in \mathcal{P}} \boldsymbol{\Gamma}^i_{K,K}\right)^{-1}_{\mathcal{S}_S,\mathcal{S}_S} \cdot \left(\sum_{t_j \in \mathcal{I}^\bullet} w^*_j \cdot \bigotimes_{p_i \in \mathcal{P}} \tilde{\mathbf{W}}^{i,j} + \bigoplus_{p_i \in \mathcal{P}} \tilde{\mathbf{U}}^i\right)_{\mathcal{S}_S,\mathcal{S}_K}
\end{array}\right], \tag{30}$$

and solve for $\tilde{\boldsymbol{\gamma}} \cdot \tilde{\mathbf{P}} = \tilde{\boldsymbol{\gamma}}$ subject to the normalization $\tilde{\boldsymbol{\gamma}} \cdot \mathbf{1}_{|\mathcal{S}_K| \times 1} = 1$. Then, the steady-state probability and the rate of entering the markings in $\mathcal{S}_K$ are

$$\forall \mathbf{m} \in \mathcal{S}_T,\ \boldsymbol{\pi}_{\mathbf{m}} = \frac{\tilde{\gamma}_{\mathbf{m}} \cdot \mathbf{h}_{\mathbf{m}}}{\sum_{\mathbf{n} \in \mathcal{S}_T} \tilde{\gamma}_{\mathbf{n}} \cdot \mathbf{h}_{\mathbf{n}}}, \quad \forall \mathbf{m} \in \mathcal{S}_K,\ \boldsymbol{\phi}_{\mathbf{m}} = \frac{\tilde{\gamma}_{\mathbf{m}}}{\sum_{\mathbf{n} \in \mathcal{S}_T} \tilde{\gamma}_{\mathbf{n}} \cdot \mathbf{h}_{\mathbf{n}}}.$$

Proof: We only need to show that $\tilde{\mathbf{P}}$ correctly describes the DTMC obtained when embedding the GSPN at the times when markings in $\mathcal{S}_T \cup \mathcal{S}_S$, but not those in $\mathcal{S}_L = \mathcal{S} \setminus (\mathcal{S}_T \cup \mathcal{S}_S)$, are entered. In other words, $\tilde{\boldsymbol{\gamma}}$ should differ from $\boldsymbol{\gamma}_K$ only by a multiplicative constant, where $[\boldsymbol{\gamma}_K \mid \boldsymbol{\gamma}_L] = \boldsymbol{\gamma}$ is the solution to Eq. 25, which, in block form, is

$$[\boldsymbol{\gamma}_K | \boldsymbol{\gamma}_L] \left[\begin{array}{c|c} \mathbf{P}_{K,K} & \mathbf{P}_{K,L} \\ \hline \mathbf{P}_{L,K} & \mathbf{P}_{L,L} \end{array} \right] = [\boldsymbol{\gamma}_K | \boldsymbol{\gamma}_L].$$

We can then obtain $\boldsymbol{\gamma}_L = \boldsymbol{\gamma}_K \cdot \mathbf{P}_{K,L} \cdot (\mathbf{I} - \mathbf{P}_{L,L})^{-1}$ and, by substitution,

$$\boldsymbol{\gamma}_K \cdot \left(\mathbf{P}_{K,L} \cdot (\mathbf{I} - \mathbf{P}_{L,L})^{-1} \cdot \mathbf{P}_{L,K} + \mathbf{P}_{K,K} \right) = \boldsymbol{\gamma}_K.$$

Then, it is sufficient to show that $\left(\mathbf{P}_{K,L} \cdot (\mathbf{I} - \mathbf{P}_{L,L})^{-1} \cdot \mathbf{P}_{L,K} + \mathbf{P}_{K,K} \right)$ and $\tilde{\mathbf{P}}$ coincide. For any $\mathbf{m}, \mathbf{n} \in \mathcal{S}_K$, $(\mathbf{P}_{K,K})_{\mathbf{m},\mathbf{n}}$ represents the probability of going from marking $\mathbf{m} \in \mathcal{S}_K$ to marking $\mathbf{n} \in \mathcal{S}_K$ in a single firing, while $(\mathbf{P}_{K,L} \cdot (\mathbf{I} - \mathbf{P}_{K,K})^{-1} \cdot \mathbf{P}_{L,K})_{\mathbf{m},\mathbf{n}}$ represents the probability of going from $\mathbf{m}$ to any marking $\mathbf{m}^1 \in \mathcal{S}_L$, visiting any number of markings in $\mathcal{S}_L$, and finally leaving $\mathcal{S}_L$ from some marking $\mathbf{m}^2$ (possibly the same as $\mathbf{m}^1$) to reach $\mathbf{n}$ in one firing.

Assuming that $\mathbf{m} \in \mathcal{S}_T$ and $\mathbf{n}$ differs from $\mathbf{m}$ in at most the position for place p_l, the corresponding entry $\tilde{\mathbf{P}}_{\mathbf{m},\mathbf{n}}$ is

$$\begin{aligned}
\tilde{\mathbf{P}}_{\mathbf{m},\mathbf{n}} &= \left(\sum_{t_j \in \mathcal{X}^\bullet} w_j^* \bigotimes_{p_i \in \mathcal{P}} \boldsymbol{\Lambda}^{i,j}_{K,K} + \bigoplus_{p_i \in \mathcal{P}} \boldsymbol{\Lambda}^{i}_{K,K} \right)^{-1}_{\mathbf{m},\mathbf{m}} \cdot \left(\sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \bigotimes_{p_i \in \mathcal{P}} \tilde{\mathbf{W}}^{i,j} + \bigoplus_{p_i \in \mathcal{P}} \tilde{\mathbf{R}}^{i} \right)_{\mathbf{m},\mathbf{n}} \\
&= \boldsymbol{\Lambda}^{-1}_{\mathbf{m},\mathbf{m}} \cdot \left(\sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \prod_{p_i \in \mathcal{P}} \tilde{\mathbf{W}}^{i,j}_{\mathbf{m}_i,\mathbf{n}_i} + \tilde{\mathbf{R}}^{l}_{\mathbf{m}_l,\mathbf{n}_l} \right) \\
&= \boldsymbol{\Lambda}^{-1}_{\mathbf{m},\mathbf{m}} \cdot \left(\sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \prod_{p_i \in \mathcal{P}} \left(\mathbf{W}^{i,j}_{\mathbf{m}_i,\mathbf{n}_i} + \sum_{\mathbf{m}_i^1 \in \mathcal{S}^i_L} \mathbf{W}^{i,j}_{\mathbf{m}_i,\mathbf{m}_i^1} \sum_{\mathbf{m}_i^2 \in \mathcal{S}^i_L} \left(\mathbf{I} - \mathbf{U}^i_{L,L} \right)^{-1}_{\mathbf{m}_i^1,\mathbf{m}_i^2} \mathbf{U}^i_{\mathbf{m}_i^2,\mathbf{n}_i} \right) \right. \\
&\quad \left. + \left(\mathbf{R}^l_{\mathbf{m}_l,\mathbf{n}_l} + \sum_{\mathbf{m}_l^1 \in \mathcal{S}^l_L} \mathbf{R}^l_{\mathbf{m}_l,\mathbf{m}_l^1} \sum_{\mathbf{m}_l^2 \in \mathcal{S}^l_L} \left(\mathbf{I} - \mathbf{U}^l_{L,L} \right)^{-1}_{\mathbf{m}_l^1,\mathbf{m}_l^2} \mathbf{U}^l_{\mathbf{m}_l^2,\mathbf{n}_l} \right) \right)
\end{aligned}$$

(if $\mathbf{n}$ and $\mathbf{m}$ differ in more than one position, the cause must be the firing of a synchronizing transition, so the "local term" for place p_l in the last expression is absent). In any case, $\boldsymbol{\Lambda}^{-1}_{\mathbf{m},\mathbf{m}}$ is just a normalization factor (if $\mathbf{m} \in \mathcal{S}_S$, $\boldsymbol{\Gamma}^{-1}_{\mathbf{m},\mathbf{m}}$ would be used instead), so the expression indicates the required probability. The key issue is that the order of firing of

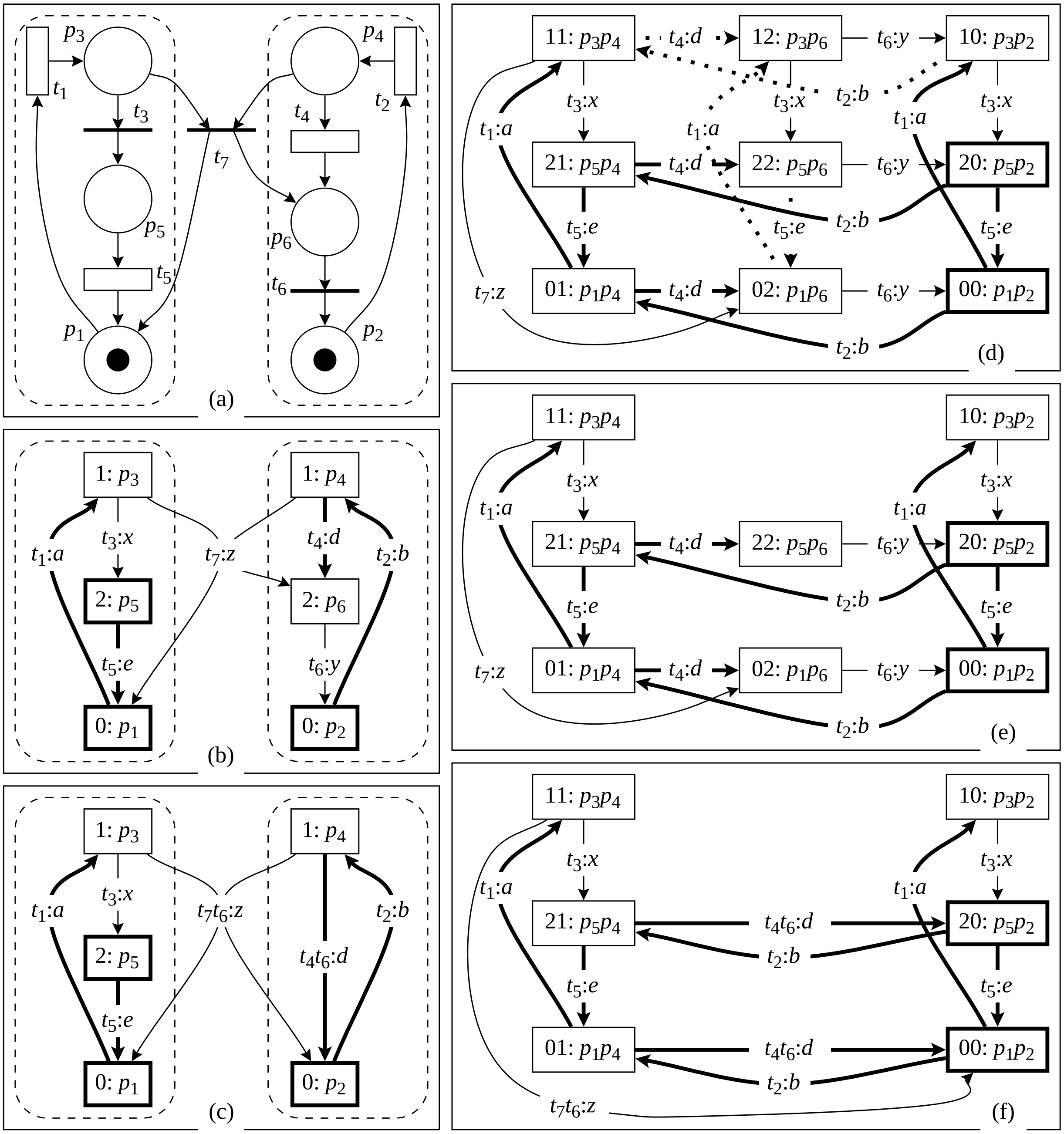

Figure 5: A GSPN with an immediate synchronizing transition.

local immediate transitions does not affect the probability of reaching a given $n \in \mathcal{S}_K$, since their weights and disabling are decided locally: the events "going from $\mathbf{m}_i^1$ to $\mathbf{n}_i$" for each p_i are independent, so their product correctly describes the overall probability of going from $\mathbf{m}^1$ to $\mathbf{n}$. □

We stress that our approach might not eliminate every vanishing marking enabling only local immediate transitions. For example, consider the GSPN in Fig. 5(a). We have $\mathcal{S}_T^1 = \{\langle 0\rangle, \langle 2\rangle\}$, $\mathcal{S}_S^1 = \{\langle 1\rangle\}$, $\mathcal{S}_L^1 = \emptyset$, $\mathcal{S}_T^2 = \{\langle 0\rangle\}$, $\mathcal{S}_S^2 = \{\langle 1\rangle\}$, and $\mathcal{S}_L^2 = \{\langle 2\rangle\}$, as shown in Fig. 5(b). After eliminating the local markings $\mathcal{S}_L^2$ from the second sub-GSPN, we obtain the "local reachability graphs" in Fig. 5(c). Fig. 5(d), 5(e), and 5(f) show the graphs describing $\mathbf{P}'$, $\mathbf{P}$, and $\tilde{\mathbf{P}}$, respectively. The dotted arcs in Fig. 5(d) correspond to timed transitions with concession in vanishing markings, which lead to markings in $\mathcal{S}'$. These are not present in $\mathbf{P}'$. Hence, (global) marking $\langle 1, 2\rangle$ is unreachable and is absent in $\mathbf{P}$ and $\tilde{\mathbf{P}}$. Furthermore, markings $\langle 0, 2\rangle$ and $\langle 2, 2\rangle$ are absent from $\tilde{\mathbf{P}}$, because their second component, 2, enables only local immediate transitions in the second sub-GSPN, $\langle 2\rangle \notin \mathcal{S}_S^2$. However, marking $\langle 1, 0\rangle$ is still present in $\tilde{\mathbf{P}}$, even if it enables only t_3, a local immediate transition. This is because its first component, 1, corresponds to having a token in p_3, which is a condition for the enabling of the synchronizing immediate transition t_7. In other words, we cannot eliminate the local marking $\langle 1\rangle$ from the local state space for the first sub-GSPN, because this would eliminate both global markings $\langle 1, 0\rangle$ and $\langle 1, 1\rangle$, and eliminating $\langle 1, 1\rangle$ would make it impossible to capture the effect of synchronizing transition t_7 in the Kronecker products of Eq. (30).

6 Numerical solution

Applying Theorem 4.1, we use the generator matrix $\mathbf{Q} = \mathbf{Q}''_{\mathcal{S}_T,\mathcal{S}_T}$ to compute the stationary distribution $\boldsymbol{\pi} \in \mathbb{R}^{|\mathcal{S}_T|}$ of the CTMC underlying the GSPN according to Eq. (3). We use an approach based on Kronecker algebra to avoid storing $\mathbf{Q}$ explicitly, so only iterative methods which do not require the modification of $\mathbf{Q}$ itself can be used effectively. Adopting the Jacobi method with overrelaxation (JOR), we transform $\mathbf{Q}$ into the iteration matrix $\mathbf{M} = (1-\omega)\cdot\mathbf{I} + \omega\cdot\mathbf{R}\cdot\mathrm{diag}(\mathbf{h})$ and solve the eigenvector problem:

$$\boldsymbol{\pi}\cdot\mathbf{M} = \boldsymbol{\pi} \qquad \text{subject to} \qquad \boldsymbol{\pi}\cdot\mathbf{1}_{|\mathcal{S}_T|\times 1} = 1.$$

Successive approximations of $\boldsymbol{\pi}$ are obtained iteratively as

$$\boldsymbol{\pi}^{[m+1]} \leftarrow \boldsymbol{\pi}^{[m]}\cdot\mathbf{M}, \tag{31}$$

starting from an initial probability vector $\boldsymbol{\pi}^{[0]}$ satisfying $\boldsymbol{\pi}^{[0]} \geq 0$ and $\boldsymbol{\pi}^{[0]}\mathbf{1}_{|\mathcal{S}_T|\times 1} = 1$. If the CTMC is ergodic and if the iterations converge, JOR is guaranteed to result in the correct solution, regardless of the value of $\boldsymbol{\pi}^{[0]}$, that is, $\boldsymbol{\pi} = \lim_{m\to\infty}\boldsymbol{\pi}^{[m]}$, if this limit exists. We do not discuss the choice of the relaxation parameter ω, $0 < \omega \leq 2$, which affects the convergence rate [16]; for a detailed analysis of numerical techniques for the solution of Markov chains, see [15].

We then need to multiply a full vector, $\boldsymbol{\pi}^{[m]}$, by a matrix, $\mathbf{R}$, which is described as a (submatrix of a) Kronecker expression of smaller matrices,

$$\mathbf{R} = \left(\sum_{t_j\in\mathcal{X}^\bullet} w_j^*\cdot\bigotimes_{p_i\in\mathcal{P}}\overline{\mathbf{W}}^{i,j} + \bigoplus_{p_i\in\mathcal{P}}\overline{\mathbf{R}}^{i}\right)_{\mathcal{S}_T,\mathcal{S}_T},$$

where $\overline{\mathbf{R}}^i = \mathbf{R}^i\cdot\mathbf{X}^i$ and $\overline{\mathbf{W}}^{i,j} = \mathbf{W}^{i,j}\cdot\mathbf{X}^i$. A similar discussion applies when using Corollary 4.1 (to obtain $\boldsymbol{\phi}_\mathbf{m}$ for $\mathbf{m}\in\mathcal{S}_V$ after $\boldsymbol{\pi}$ has been obtained, we compute $\boldsymbol{\pi}\cdot\mathbf{Q}''_{\mathcal{S}_T,\mathcal{S}_V}$) or Theorem 5.2 (to obtain $\tilde{\boldsymbol{\gamma}}^{[m+1]}$, we perform the product $\tilde{\boldsymbol{\gamma}}^{[m]}\cdot\tilde{\mathbf{P}}$).

First, we precompute $\mathcal{S}_T$, Ψ_T^{-1}, and $\mathbf{h}$, for each tangible marking:

- During the reachability set construction, $\mathcal{S}_T$ is normally stored in lexicographic order in a balanced search tree. However, we can assume from now on that is accessed exclusively using Ψ_T^{-1}.
- After the reachability graph construction, Ψ_T^{-1} is stored as an array of size $|\mathcal{S}_T|$ of pointers to the markings. $\Psi_T^{-1}(k)$ points the k-th marking, that is, the marking $\mathbf{m}$ satisfying $\Psi(\mathbf{m}) = k$. Several methods can be used to store a marking. In principle, a single integer describing the position of $\mathbf{m}$ in $\mathcal{S}'$ according to lexicographic order is sufficient, but, in practice, this integer might require more than 32 bits.
- $\mathbf{h} \in I\!\!R^{|\mathcal{S}_T|}$ carries the same information as $\boldsymbol{\Lambda}$. Instead of storing this vector explicitly, we could recompute $\boldsymbol{\Lambda}$ at each iteration, to further reduce the memory requirements. This is a memory-execution trade-off.

It is important to note that an alternative method suggested by Kemper [12] for the state space exploration has the advantage of allowing the determination of whether a marking is reachable in $O(1)$ instead of $O(\log|\mathcal{S}_T|)$ operations. However, it requires a bit vector of size $|\mathcal{S}'|$, which might be a problem when $|\mathcal{S}'| \gg |\mathcal{S}_T|$.

If we define

$$\forall p_i \in \mathcal{P},\ low(i) = \prod_{l=1}^{i-1} n_l \qquad \text{and} \qquad up(i) = \prod_{l=i+1}^{|\mathcal{P}|} n_l,$$

the contribution of the Kronecker sum to the new iteration vector $\boldsymbol{\pi}^{[m+1]}$ in Eq. (31) can be rewritten as

$$\boldsymbol{\pi}^{[m]} \cdot \bigoplus_{p_i \in \mathcal{P}} \overline{\mathbf{R}}^i = \sum_{p_i \in \mathcal{P}} \boldsymbol{\pi}^{[m]} \cdot \left(\mathbf{I}_{low(i)} \otimes \mathbf{Q}_i \otimes \mathbf{I}_{up(i)}\right) = \sum_{p_i \in \mathcal{P}} \boldsymbol{\pi}_{p_i}^{[m+1]},$$

while the contribution of the Kronecker products is

$$\sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \boldsymbol{\pi}^{[m]} \cdot \bigotimes_{p_i \in \mathcal{P}} \overline{\mathbf{W}}^{i,j} = \sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \boldsymbol{\pi}^{[m]} \cdot \prod_{i=1}^{|\mathcal{P}|} \left(\mathbf{I}_{low(i)} \otimes \overline{\mathbf{W}}^{i,j} \otimes \mathbf{I}_{up(i)}\right) = \sum_{t_j \in \mathcal{X}^\bullet} w_j^* \cdot \boldsymbol{\pi}_{t_j}^{[m+1]}.$$

We implemented the basic operation $\mathbf{a} \cdot (\mathbf{I}_{low(i)} \otimes \mathbf{A} \otimes \mathbf{I}_{up(i)})$ in a function $mult(\mathbf{a}, \mathbf{A}, i)$, where $\mathbf{a}$ is a vector of size $|\mathcal{S}_T|$ and $\mathbf{A}$ is a $n_i \times n_i$ matrix. Each term $\boldsymbol{\pi}_{p_i}^{[m+1]}$ is computed with a single call to $mult$, while each term $\boldsymbol{\pi}_{t_j}^{[m+1]}$ requires $|\mathcal{P}|$ successive calls to $mult$. During this iteration, some states (possibly unreachable) are passed and temporarily stored. However, if all the states of $\mathcal{S}'$ are reachable and tangible, that is, if $|\mathcal{S}'| = |\mathcal{S}_T|$, there is no need to explicitly identify the reachable states. We stress that this definition allows any number of Kronecker products to be performed, so it does not restrict the possible number of sub-GSPNs involved in a synchronization.

For the numerical computation we consider a fork/join kanban network of four sub-GSPNs as given in Figure 6(b). Each sub-GSPN $i = 1, \ldots, 4$, of the network is modelled by the GSPN shown in Fig. 6(a) on the left. Multiplying $\mathbf{R}^i$ and $\mathbf{W}^{i,j}$ by $\mathbf{X}^i$ corresponds, de facto, to the automatic elimination of the immediate transitions t_{redo_i} and t_{ok_i}, resulting in the GSPN shown in Fig. 6(a) on the right, where the rates of $t_{m_i redo_i}$ and $t_{m_i ok_i}$ are the the rate of t_m multiplied by the firing probabilities of t_{redo_i}

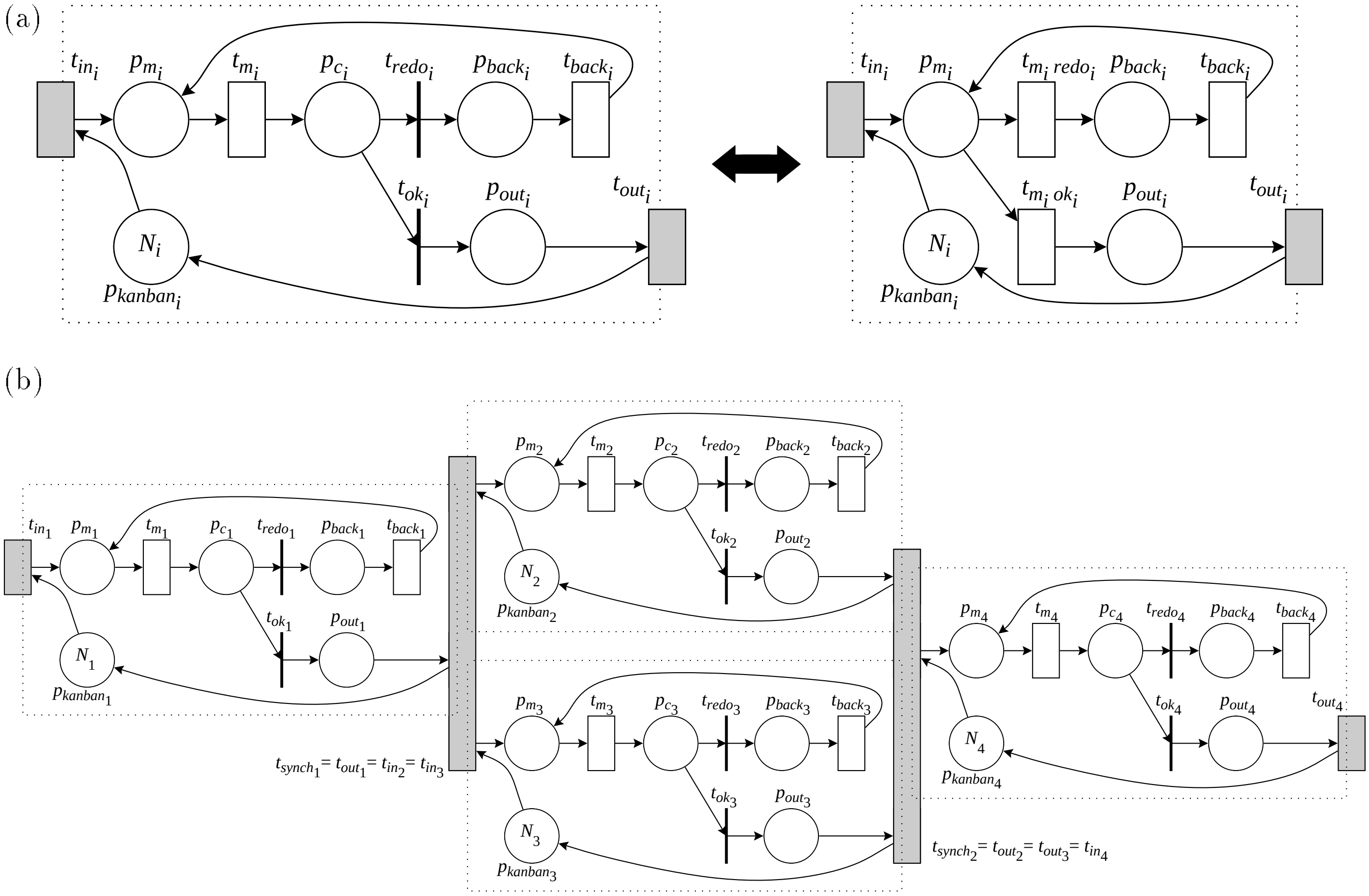

Figure 6: A sub-GSPN and a fork/join arrangement of four sub-GSPNs.

and t_{ok_i}, respectively. The parameters specifying the stochastic behavior of the sub-GSPNs are the rates $\mathbf{w}_{in_i}$, $\mathbf{w}_{m_i}$, and $\mathbf{w}_{back_i}$ and the probabilities $\mathbf{w}_{ok_i}$ and $\mathbf{w}_{redo_i}$. We assume that these quantities are constant, but they could depend on the local marking (or even on the global marking, in the case of synchronizing transitions) without affecting the feasibility or the complexity of the solution algorithms. Pallets enter sub-GSPN i of the kanban network through transition t_{in_i}, which requires the availability of a kanban ticket in place p_{kanban_i}. Then, the pallet proceeds to the machine, in place p_{m_i}. After being worked by t_{m_i}, a part is checked for quality and it is either transported back to p_{m_i} by t_{back_i} for further rework, or moved out of the machine by t_{out_i}. The numerical values of the parameters for the model are: $\mathbf{w}_{m_1} = 1.2$, $\mathbf{w}_{m_2} = 1.4$, $\mathbf{w}_{m_3} = 1.3$, and $\mathbf{w}_{m_4} = 1.1$, while, for each sub-GSPN i, $\mathbf{w}_{ok_i} = 0.7$, $\mathbf{w}_{redo_i} = 0.3$, and $\mathbf{w}_{back_i} = 0.3$. All transitions have single-server semantics. The input and output rates for the entire kanban network are set by assigning $\mathbf{w}_{in_1} = 1.0$ and $\mathbf{w}_{out_4} = 0.9$. The synchronizing transition t_{synch_1} corresponds to the merging of transitions $\{t_{out_1}, t_{in_2}, t_{in_3}\}$ and has rate 0.4, while t_{synch_2} corresponds to transitions $\{t_{out_2}, t_{out_3}, t_{in_4}\}$ and has rate 0.5. In the computations, we vary the number $N = N_i$ of tokens initially in each place p_{kanban_i}.

We consider two alternative decompositions of the model. In Case 1, each node of the kanban network corresponds to a sub-GSPN, or one macroplace, in our terminology. Hence, $|\hat{\mathcal{P}}| = 4$ and $|\mathcal{X}^\bullet| = 2$. In this case, the fork-join synchronization creates unreachable markings (any marking where the number of tokens in p_{kanban_2} and p_{kanban_3} differ), so that $|\mathcal{S}_T| < |\mathcal{S}'|$. Instead, in Case 2, we merge sub-GSPNs 2 and 3 into a single macroplace, hence $|\hat{\mathcal{P}}| = 3$. Now, however, the constraint on the number of tokens in p_{kanban_2} and p_{kanban_3} is taken explicitly into account when generating the local state space for the corresponding sub-GSPN. Thus, the overall state space becomes the exact cross-product of the three local state spaces, $|\mathcal{S}_T| = |\mathcal{S}'|$, and the computation is more efficient, since there is no need to distinguish the reachable states from the unreachable ones. This shows how an intelligent decomposition of the model can greatly affect the computational requirements of the solution.

Finally, we make the two synchronizing transitions immediate and apply Theorem 5.2 to a decomposition into either four or three sub-GSPNs, resulting in Case 3 and 4, respectively. We observe that the presence of synchronizing transitions leads to unreachable vanishing markings in this model.

N	Case	e_1	e_2	e_3	e_4	τ
1	1,2	0.907	0.671	0.671	0.355	0.132
2		1.810	1.328	1.328	0.764	0.248
3		2.722	1.944	1.944	1.154	0.332
4		3.646	2.515	2.515	1.510	0.393
5		4.588	3.053	3.053	1.876	0.439
1	3,4	0.860	0.810	0.810	0.534	0.139
2		1.691	1.671	1.671	1.268	0.263
3		2.529	2.545	2.545	2.037	0.348
4		3.379	3.425	3.425	2.807	0.409
5		3.789	4.454	4.454	3.914	0.529

Table 1: Performance results as a function of N

We compute the expected number e_i of tokens in places p_{m_i}, p_{back_i}, and p_{out_i} for each sub-GSPN of the kanban system. For a given $i = 1, \ldots, 4$, this corresponds to a reward structure where the reward rate $\rho(\mathbf{m})$ of a marking $\mathbf{m}$ is given by $\mathbf{m}_{m_i} + \mathbf{m}_{back_i} + \mathbf{m}_{out_i}$, and the reward impulses $\mathbf{r}$ are identically zero. We also compute the throughput τ of the system, defined as the expected firing rate of t_{synch_1}. This corresponds to a reward structure where the reward rates are identically zero and the reward impulses are one if they correspond to the firing of t_{synch_1}, zero otherwise (the same quantity could be computed by observing t_{in_1}, t_{synch_2}, or t_{out_4} instead). We stress that, in general, reward rates and impulses could be marking-dependent. Table 1, shows the resulting numerical values; as expected, $e_2 = e_3$. It is apparent that the first sub-GSPN is the most loaded for Case 1 and 2, while the second and third sub-GSPNs are the bottleneck in Case 3 and 4. This would suggest that, the synchronization behavior is an important performance factor, in addition to the actual machining rate of each station.

Table 2 shows the size of the state spaces $\mathcal{S}_T$, $\mathcal{S}_V$, and $\mathcal{S}'$ as a function of N. The overall number of nonzero elements for the sparse matrices used by the Kronecker approach ("local") can be compared

N	Case	$\lvert\mathcal{S}_T\rvert$	$\lvert\mathcal{S}_V\rvert$	$\lvert\mathcal{S}'\rvert$	local	nonzero($\mathbf{Q}$)	nonzero($\tilde{\mathbf{P}}$)	"explore"	"solve"
1	1	160	0	256	20	616	616	0.017	0.267
2		4,600	0	10,000	80	28,120	28,120	0.967	12.550
3		58,400	0	160,000	200	446,400	446,400	16.267	309.533
4		454,475	0	1,500,625	400	3,979,850	3,979,850	190.500	4,721.217
5		2,546,432	0	9,834,496	700	24,460,016	24,460,016	1,759.000	22,215.257
1	2	160	0	160	40	616	616	not req.	0.017
2		4,600	0	4,600	186	28,120	28,120	not req.	1.517
3		58,400	0	58,400	678	446,400	446,400	not req.	50.483
4		454,475	0	454,475	1878	3,979,850	3,979,850	not req.	855.917
5		2,546,432	0	2,546,432	4368	24,460,016	24,460,016	not req.	6,054.783
1	3	152	8	256	20	600	608	0.017	0.167
2		3,816	697	10,000	80	23,832	24,529	0.517	20.450
3		41,000	13,656	160,000	200	316,360	330,016	5.967	84.600
4		268,475	128,000	1,500,625	400	2,343,050	2,471,050	75.783	719.050
5		1,270,962	769,480	9,834,496	700	12,025,566	12,795,046	707.483	4,111.681
1	4	152	8	160	40	600	608	0.013	0.150
2		3,816	697	4,600	186	23,832	24,529	0.333	18.550
3		41,000	13,656	58,400	678	316,360	330,016	5.550	76.250
4		268,475	128,000	454,475	1878	2,343,050	2,471,050	67.750	647.100
5		1,270,962	769,480	2,546,432	4368	12,025,566	12,795,046	725.333	3,562.531

Table 2: Computational and storage requirements as a function of N.

to that of conventional solution methods that explicitly generate $\mathbf{Q}$. For comparison, we also list the number of nonzero elements in $\tilde{\mathbf{P}}$, that is after eliminating the local vanishing markings. For Case 1 and 2, this makes no difference (there are no synchronizing vanishing markings), while, for Case 3 and 4, the memory requirements are somewhat larger. It is clear that the only practical limitation of our approach is the memory required by the two iteration vectors, $\boldsymbol{\pi}^{[m]}$ and $\boldsymbol{\pi}^{[m+1]}$, and the addressing vector Ψ_T (of sizes $|\mathcal{S}_T|$), or by $\tilde{\boldsymbol{\gamma}}^{[m]}$, $\tilde{\boldsymbol{\gamma}}^{[m+1]}$, and Ψ (of sizes $|\mathcal{S}_T| + |\mathcal{S}_V|$), plus the vector $\mathbf{h}$ (of size $|\mathcal{S}_T|$), if not computed at each iteration. It is also apparent how the approach allows the solution of problems one order of magnitude larger than with conventional methods in an acceptable amount of time.

The computation times "explore", to explore the reachability set, and "solve", for the Jacobi method with relaxation parameter $\omega = 0.9$, are given in seconds. The vector $\mathbf{h}$ is stored explicitly. The convergence criterion is set to $||\boldsymbol{\pi}^{[m]} - \boldsymbol{\pi}^{[m+1]}||_\infty < 10^{-6}$ (or $||\tilde{\boldsymbol{\gamma}}^{[m]} - \tilde{\boldsymbol{\gamma}}^{[m+1]}||_\infty < 10^{-6}$). The uniform distribution was used as the initial guess for $\boldsymbol{\pi}^{[0]}$ (or $\tilde{\boldsymbol{\gamma}}^{[0]}$). The program was run on a Sony NWS-5000 workstation with 90 Mbyte of main memory.

We stress that the elimination approach (Theorem 4.1) should be used whenever possible, and that preservation of the non-local vanishing markings (Theorem 5.2) should be used only when there are immediate synchronizing transitions. Preservation of all vanishing markings (Theorem 5.1) is probably never appropriate, since memory usage is the paramount consideration, and this approach increases the already critical size of the probability vectors. In few pathological cases, it could reduce the size of the "local" data structures, but these are not critical.

7 Conclusion

We rigorously formalized and implemented an approach based on Kronecker algebra for the solution of the CTMC underlying a GSPN. The results extend previous works by Donatelli, Buchholz, and Kemper [4, 5, 10, 11, 12], to include immediate synchronizing transitions, quite general marking-dependent behavior, and a reward structure allowing reward impulses associated with immediate transition firings. The restrictions imposed on the GSPN are minimal, thus the approach has obvious practical applications. Furthermore, the structure of the GSPN itself gives strong hints on its decomposition. For example, if an "elimination-based" solution is desired, the GSPN must be decomposed so that immediate transitions are local to a sub-GSPN; if the marking-dependency of a transition does not satisfy our requirements, all the places responsible for this behavior should also be merged in the same sub-GSPN.

Memory requirements are still the main limitation to the solution, but these have now been reduced from the size of the transition rate matrix to that of the steady-state probability vector, for a very general class of GSPNs. Even for highly sparse matrices, this corresponds to the ability to solve problems whose state space is one order of magnitude larger than with a traditional solution approach.

To further increase the size of models that can be solved, we can use the distributed state-space generation algorithm described in [7], which allows one to partition the memory and execution requirements to generate the state space over a set of workstations, and which exhibits excellent speedups for large problems. The Jacobi method we employed can also be parallelized using a set of workstations, so that the entire solution process is performed in a distributed fashion and uses the available memory. In particular, the "local" matrices occupy a negligible amount of memory and can be duplicated on each processor, while only the probability vector needs to be distributed.

Since our results are complementary to those in [7], we can reasonably hope for a two-orders of magnitude increase in the size of manageable models, assuming the availability of a network of workstations. As a target for the near future, we will attempt to solve CTMCs with 10^8 states and a transition rate matrix with 10^9 non-zeros in a matter of hours using a dozen workstations with 128Mbytes of memory each.

Eventually, even CTMCs of this size might not be enough to satisfy the needs of a serious modeler. Our results, combined with the distributed state-space generation algorithm, are directly applicable to the family of approximations suggested in [17]. We can simply consider each sub-GSPN separately for the coarsest model, two adjacent sub-GSPNs for the next more accurate model, then three adjacent sub-GSPNs and so on. The comparison of the results from successive approximate models can suggest an estimate of the quality of the results. This allows trading off lower approximation errors against higher computational effort. We are currently investigating the requirements to assure an accurate solution (e.g., approximation errors below 5%) on a single workstation for a large class of models, where the CTMCs may have more than 10^{30} states.

A prototype version of the program used to compute the results we presented is available and can be obtained by contacting the second author. The input to the program has a SPNP-style syntax [9].

References

[1] Ajmone Marsan, M., Balbo, G., Bobbio, A., Chiola, G., Conte, G., and Cumani, A. The effect of execution policies on the semantics and analyis of Stochastic Petri Nets. *IEEE Trans. Softw. Eng. 15*, 7 (July 1989), 832–846.

[2] Ajmone Marsan, M., Balbo, G., and Conte, G. A class of Generalized Stochastic Petri Nets for the performance evaluation of multiprocessor systems. *ACM Trans. Comp. Syst. 2*, 2 (May 1984), 93–122.

[3] Ajmone Marsan, M., Balbo, G., Conte, G., Donatelli, S., and Franceschinis, G. *Modelling with generalized stochastic Petri nets.* John Wiley & Sons, 1995.

[4] Buchholz, P. Numerical solution methods based on structured descriptions of Markovian models. In *Computer performance evaluation* (1991), G. Balbo and G. Serazzi, Eds., Elsevier Science Publishers B.V. (North-Holland), pp. 251–267.

[5] Buchholz, P., and Kemper, P. Numerical analysis of stochastic marked graphs. In *Proc. Int. Workshop on Petri Nets and Performance Models (PNPM'95)* (Durham, NC, Oct. 1995), IEEE Comp. Soc. Press, pp. 32–41.

[6] Ciardo, G., Blakemore, A., Chimento, P. F. J., Muppala, J. K., and Trivedi, K. S. Automated generation and analysis of Markov reward models using Stochastic Reward Nets. In *Linear Algebra, Markov Chains, and Queueing Models*, C. Meyer and R. J. Plemmons, Eds., vol. 48 of *IMA Volumes in Mathematics and its Applications.* Springer-Verlag, 1993, pp. 145–191.

[7] Ciardo, G., Gluckman, J., and Nicol, D. Distributed state-space generation of discrete-state stochastic models. *ORSA J. Comp..* Submitted.

[8] Ciardo, G., Muppala, J. K., and Trivedi, K. S. On the solution of GSPN reward models. *Perf. Eval. 12*, 4 (1991), 237–253.

[9] Ciardo, G., Trivedi, K. S., and Muppala, J. K. SPNP: Stochastic Petri net package. In *Proc. 3rd Int. Workshop on Petri Nets and Performance Models (PNPM'89)* (Kyoto, Japan, Dec. 1989), IEEE Comp. Soc. Press, pp. 142–151.

[10] Donatelli, S. Superposed Stochastic Automata: a class of stochastic Petri nets amenable to parallel solution. In *Proc. 4th Int. Workshop on Petri Nets and Performance Models (PNPM'91)* (Melbourne, Australia, Dec. 1991), IEEE Comp. Soc. Press, pp. 54–63.

[11] Donatelli, S. Superposed generalized stochastic Petri nets: definition and efficient solution. In *Application and Theory of Petri Nets 1994, Lecture Notes in Computer Science 815 (Proc. 15th Int. Conf. on Applications and Theory of Petri Nets, Zaragoza, Spain)* (June 1994), R. Valette, Ed., Springer-Verlag, pp. 258–277.

[12] Kemper, P. Numerical analysis of superposed GSPNs. In *Proc. Int. Workshop on Petri Nets and Performance Models (PNPM'95)* (Durham, NC, Oct. 1995), IEEE Comp. Soc. Press, pp. 52–61.

[13] Murata, T. Circuit theoretic analysis and synthesis of marked graphs. *IEEE Trans. Circ. and Syst. CAS-24*, 7 (July 1977), 400–405.

[14] Plateau, B., and Atif, K. Stochastic Automata Network for modeling parallel systems. *IEEE Trans. Softw. Eng. 17*, 10 (Oct. 1991), 1093–1108.

[15] Stewart, W. J. *Introduction to the Numerical Solution of Markov Chains.* Princeton University Press, 1994.

[16] Stewart, W. J., and Goyal, A. Matrix methods in large dependability models. Tech. Rep. RC-11485, IBM T.J. Watson Res. Center, Yorktown Heights, NY, Nov. 1985.

[17] Takahashi, Y. Aggregate approximation for acyclic queuing networks with communication blocking. In *Queueing Networks with Blocking*, H. G. Perros and T. Altiok, Eds. Elsevier Science Publishers B.V., 1989, pp. 33–47.

[18] Wirth, N. *Algorithm + Data Structures = Programs.* Prentice-Hall, 1976.