

PERFORMANCE MODELING AND CONTROL OF STRUCTURALLY CONTROLLED RESOURCE ALLOCATION SYSTEMS

Jin Young Choi and Spyros Reveliotis

*School of Industrial & Systems Engineering
Georgia Institute of Technology
765 Ferst Drive, Atlanta, GA 30332*

Abstract: While the structural and performance-oriented control problems related to the real-time management of the resource allocation systems (RAS) underlying the operation of many contemporary technological applications have, both, been investigated extensively in the past, their integration in a seamless modeling and analysis framework, and the effects arising from their interaction, remain yet to be addressed. This paper undertakes these issues by investigating the class of Generalized Stochastic Petri nets (GSPN's) as a convenient and powerful analytical framework for the integrated modeling of the logical and the time-based RAS dynamics. More specifically, it is shown that GSPN's provide an exact formulation for the problem of performance control of structurally controlled RAS, in the case of systems with exponentially distributed event firing times, and effective approximations for the more general case of systems with non-exponential event timings. On the theoretical side, the aforementioned formulation can be utilized to (re-)establish various properties of the optimal scheduling policy in the considered operational context, including the existence of an optimal deterministic stationary scheduling policy. Finally, the proposed framework can support the thorough characterization of the structure of the optimal scheduling policy for small-sized RAS's, under various parameterizations, allowing, thus, for a more systematic study and a more profound understanding of the timed dynamics taking place in these environments. *Copyright©2002IFAC*

Keywords: Resource Allocation Systems, structural control, Performance control, scheduling, Generalized Stochastic Petri-nets

1. INTRODUCTION

With the migration of modern technological applications to highly automated modes of operation, the effective and efficient deployment, reconfiguration and control of the resource allocation underlying the emerging modes of these environments is an issue of ever-increasing importance. Yet, currently we are lacking an adequate methodology for the effective real-time management of these flexibly automated *resource allocation sys-*

tems (RAS), partly due to the fact that past research on the performance modeling and control of these environments has adopted a high-level perspective of their dynamics, ignoring the lower-level operational details. Characteristically, it is interesting to notice that the *hierarchical decomposition* framework (Gershwin, 1989), the most widely adopted “analytical” framework for planning and control in production environments, discerns strategic, tactical, and operational de-

cisions, all of which address performance objectives, while it presumes the logically consistent and robust system behavior. However, in an extensively automated environment, the establishment of logically correct and robust behavior to the various operational contingencies, is definitely a responsibility of the underlying control logic. This gives rise to a new set of control problems, referred to as the RAS *Structural Control (SC)*, which has been investigated extensively in the last decade, but the integration of the developed set of results with the complementary function of performance-oriented control remains still to be addressed.

The work presented in this paper undertakes the development of an analytical framework for performance modeling and control of *structurally controlled RAS*, that will support the seamless integration of logical and performance-oriented control problems. More specifically, it is shown that the class of the *Generalized Stochastic Petri nets (GSPN's)* offers a convenient and powerful analytical framework for modeling the considered problem, providing exact formulations in the case of systems with exponentially distributed event firing times, and effective approximations for the more general case of systems with non-exponential event timings. Furthermore, on the theoretical side, by providing a detailed, closed-form representation of the optimal scheduling problem for structurally controlled RAS, the aforementioned GSPN modeling and analysis framework can be used in order to (re-)establish some properties regarding the structure of the optimal scheduling policies; in particular, it can be shown that, for a large part of the considered class of problems, there will always exist an optimal *deterministic stationary* scheduling policy. Finally, for the case of small-sized systems, the proposed framework supports the thorough characterization of the structure of the optimal policies under various system parameterizations, allowing, thus, for a more systematic study and a more profound understanding of the (timed) dynamics taking place in these environments. The rest of the paper is organized as follows: Section 2 introduces the proposed GSPN-based analytical framework for performance modeling and analysis of structurally controlled RAS. Section 3 discusses the existence of a deterministic stationary optimal performance control policy for the considered problem. Section 4 elucidates the concepts and the analytical power of the proposed modeling framework, by undertaking the detailed analysis of a small re-entrant line with finite buffering capacity. Finally, Section 5 concludes the paper, and suggests directions for future work. In the following, it is assumed that the reader is familiar with the GSPN model and the basic RAS theory; an excellent introduction of the former can be found in (A. Marsan

et al., 1986), while an overview of the latter is provided at (Reveliotis, 2001).

2. GSPN-BASED PERFORMANCE MODELING AND ANALYSIS

The key idea underlying the proposed approach for modeling the time-based dynamics of the structurally controlled RAS through the class of GSPN's, is to refine the resource-process nets that typically model the qualitative/logical behavior of the considered RAS (Banaszak and Krogh, 1990), by (i) introducing a more detailed modeling of the processing, staging and transport phases experienced by the different job instances during their advancement through their various processing stages, and (ii) explicitly associating a timing distribution with the various transitions corresponding to the job active processing or transfers. Furthermore, in order to facilitate the GSPN-based modeling and analysis, it is assumed that all transition times are exponentially distributed. This last assumption can be relaxed, whenever it is deemed as too unrealistic, by substituting each timed transition in the resulting GSPN model with a GSPN subnet, modeling a phase-type distribution that approximates the original/empirical distribution of the underlying event timings; we refer the reader to (Papadopoulos *et al.*, 1993) for a detailed treatment of phase-type distributions and the relevant approximation theory. The details underlying the GSPN-based modeling of structurally controlled RAS are further elucidated through the example of Section 4.

According to the general GSPN theory (A. Marsan *et al.*, 1986), the marking process of a GSPN net, \mathcal{N} , is a semi-Markov process with a discrete state space, \mathcal{S} , given by the net reachability space $R(N, M_0)$. \mathcal{S} is partitioned to *vanishing* states / markings, \mathcal{V} , which enable at least one immediate transition of \mathcal{N} , and therefore, they have zero sojourn time, and *tangible* markings, \mathcal{T} , which enable only timed transitions, and therefore, they present positive sojourn times. Furthermore, the untimed system dynamics, defined by its transitional patterns among the various states of its reachable state space, are characterized by the, so called, *Embedded Markov Chain (EMC)*, whose branching probabilities, $Q = [q_{kl}]$ are determined by externally specified (*dynamic*) *random switches*, in case of vanishing markings, and the enabled event *exponential race*, in case of tangible markings. If this EMC is finite-state, homogeneous, and irreducible, it possesses a steady-state distribution.

In the case of GSPNs modeling RAS behavior, the underlying EMC is finite-state and homogeneous, while the imposition of a liveness-enforcing SCP

on the RAS behavior establishes also its irreducibility. Furthermore, the set of vanishing markings, \mathcal{V} , represents the decision-making points in the RAS operation, while the associated set of *dynamic* random switches, Ξ , implements the logic of the imposed scheduling policy. Hence, the set of random switching probabilities, ξ_l , defines the decision variables of the underlying RAS scheduling problem, that must be priced in a way that optimizes the performance objective under consideration. Letting $Q(\xi)$ denote the transition probability matrix (TPM) of the system EMC, the problem of optimizing a performance objective for the structurally controlled RAS can be formally expressed by the following Mathematical Programming (MP) formulation:

$$\max_{\xi} \mathcal{O}(\xi) \quad (1)$$

s. t.

$$\mathbf{y} = \mathbf{y}Q(\xi) ; \sum_{s_k \in \mathcal{S}} y_k = 1 \quad (2)$$

$$\pi_k = \begin{cases} 0, & s_k \in \mathcal{V} \\ y_k E[s_k] / \sum_{s_l \in \mathcal{T}} y_l E[s_l], & s_k \in \mathcal{T} \end{cases} \quad (3)$$

$$\forall s_k \in \mathcal{T}, E[s_k] = 1 / \sum_{T_j \text{ enabled in } s_k} r_j \quad (4)$$

$$\forall l, \xi_l \geq 0.0 \quad (5)$$

$$\forall \text{ random switch } \Xi, \sum_{l: \xi_l \in \Xi} \xi_l = 1.0 \quad (6)$$

Equation 2 computes the limiting distribution, \mathbf{y} , of the embedded Discrete Time Markov Chain. Furthermore, the steady-state probabilities, $\pi = [\pi_k]$, for the underlying continuous-time stochastic process, are obtained by Equation 3. In Equation 3, $E[s_k]$ denotes the expected sojourn time for tangible marking $s_k \in \mathcal{T}$, and it is computed by Equation 4, where r_j denotes the (firing) rate of (timed) transition T_j . Equations 5 and 6, introduce the notion of dynamic random switch in the formulation, constraining appropriately the decision variables ξ_l . Finally, once the steady-state probability vector π has been obtained, various performance measures of interest can be defined as appropriate functions of π and the other system parameters. For instance, Equation 1 can represent the RAS (steady-state) throughput function, by setting

$$\mathcal{O}(\xi) \equiv \sum_{(k,j)} \pi_k r_j I \left\{ \begin{array}{l} T_j \text{ enabled in } s_k \\ \wedge \text{ cor. to a job} \\ \text{unloading} \end{array} \right\} \quad (7)$$

3. DETERMINISTIC OPTIMAL SCHEDULING POLICIES

It can be shown that the solution space of the MP formulation employing the objective function defined by Equation 7, will always contain a *deterministic* optimal solution, i.e., an optimal solution ξ^* with $\xi_l^* \in \{0, 1\}$, $\forall l$. This result can be obtained through direct analysis of the structure of the formulation under consideration (Choi and Reveliotis, 2001), and it is consistent to some more classical results in Dynamic Programming theory (Puterman, 1994). Beyond its theoretical interest, establishing the existence of a deterministic optimal solution is of practical importance because it limits the search for an optimal solution over a discrete sub-space of the overall solution space, and therefore, it renders the considered problem amenable to enumerative approaches. We notice, however, that the aforementioned deterministic structure of the optimal solution can be lost, once some additional constraint(s) is introduced in the formulation. For instance, in a multi-item production scheduling context, such a constraint can arise from the requirement for production of the various product types at quantities that satisfy certain ratios; we refer the reader to (Choi and Reveliotis, 2001) for a more concrete example.

4. AN EXAMPLE: THROUGHPUT MAXIMIZATION OF A CAPACITATED RE-ENTRANT LINE

The considered RAS This section, elucidates the theory developed in Section 2 and 3, by applying the developed results to the detailed analysis of the small capacitated re-entrant line in Figure 2. This line produces a single item, and it possesses two stations, W_1 and W_2 , with $S_1 = S_2 = 1$ and $C_1 = 1$; $C_2 = 2$, where S_i is the number of identical servers and C_i the number of buffer slots in the workstation W_i . Furthermore, the supported production sequence is $J = \langle J_1, J_2, J_3 \rangle$, with $W(J_1) = W(J_3) = W_1$ and $W(J_2) = W_2$. Finally, stage processing times are exponentially distributed with means $m_j = 1/\mu_j > 0$, $j = 1, 2, 3$, and so are the involved transfer times, with a uniform mean $d = 1/\lambda$. Each part visiting a workstation for the execution of some processing stage is allocated one unit of buffering capacity, which it holds exclusively during its entire sojourn in the station. Once in the station local buffer, the part competes for one of the station servers, for the execution of the requested stage. A part having finished the processing of its current stage at a certain station, waits in its allocated buffer for transfer to the next requested station. This transfer is facilitated by the central (automated) material handling system, and it is authorized

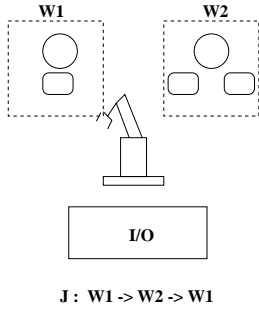


Fig. 1. Example: The capacitated re-entrant line

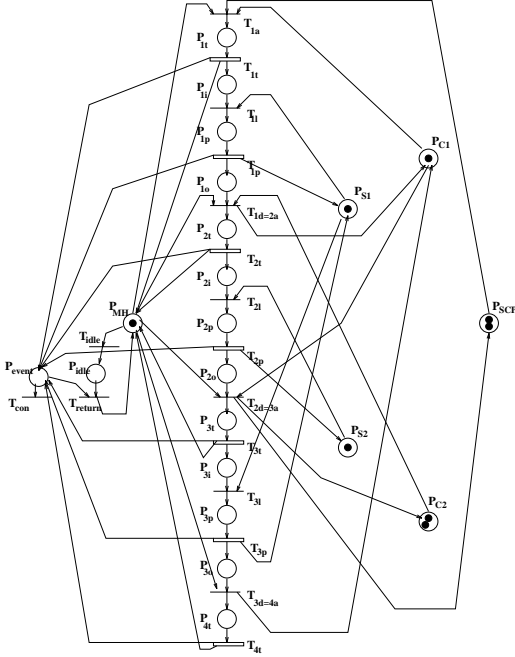


Fig. 2. Example: The GSPN model

by a supervisory control policy ensuring that (i) the destination workstation has available buffering capacity, and (ii) the transfer is *safe*, i.e., it is still physically possible from the resulting state to process all running jobs to completion. For this small configuration, it is easy to see that, under the operational assumptions outlined above, the system material flow will remain deadlock-free, as long as

$$|J_1| + |J_2| \leq C_1 + C_2 - 1 = 2 \quad (8)$$

where $|J_j|$, $j = 1, 2, 3$ denotes the number of job instances in $W(J_j)$ executing stage J_j .

GSPN-based modeling of the capacitated re-entrant line The GSPN modeling the behavior of the capacitated re-entrant line of Figure 2, under the control of the maximally permissive structural control policy (SCP) of Equation 8, is depicted in Figure 3. Specifically, in the GSPN of Figure 3, the part flow dynamics associated with each processing stage J_j , $j = 1, 2, 3$, are modeled by the corresponding net path $\langle T_{ja}, P_{jt}, T_{jt}, P_{ji}, T_{jl}, P_{jp}, T_{jp}, P_{jo}, T_{jd} \rangle$, while it also holds $T_{jd} \equiv T_{j+1,a}$, with $j = 4$ denot-

ing the last unloading step. A token in place P_{jt} represents a part in transit to the buffer of workstation $W(J_j)$; a token in place P_{ji} represents a part in the buffer of $W(J_j)$ waiting the allocation of one of the buffer servers; a token in place P_{jp} represents a part in processing of stage J_j ; finally, a token in place P_{jo} represents a part having finished processing of stage J_j , and waiting for transfer to the next requested workstation or, in case that J_j is the last processing stage, to the I/O station. On the other hand, places P_{MH} , P_{S_i} , P_{C_i} , $i = 1, 2$, and P_{SCP} model respectively the availability of the system transporter, workstation servers and buffers, and the logic of the applied SCP, according to the standard, by now, modeling practice of resource-process nets (Banaszak and Krogh, 1990). It is important to notice that transitions T_{ja} , T_{jl} and T_{jd} , that are associated with the various decisions regarding the allocation of the system buffering, processing and/or transport capacity, are untimed/ immediate transitions, while the delays experienced from the processing and/or transfer times involved with the execution of these decisions, are modeled by the timed transitions T_{jt} and T_{jp} . This separation of the net components modeling the timings of the various system events from the net structure modeling the underlying resource allocation and the associated decision making, enables the modeling of timing distributions other than exponential through the (local) substitution of the corresponding timed transitions by GSPN subnets modeling the approximating phase-type distributions (c.f. Section 2). It also allows the modeling of the required scheduling logic through a set of *dynamic random switches*, that resolve the conflicts among the immediate transitions that are simultaneously enabled at the net reachable vanishing markings. Finally, some explanation is necessary about the role of places P_{idle} , P_{event} and their associated transitions T_{idle} , T_{return} and T_{con} . This subnet essentially establishes a GSPN-compatible mechanism for representing some deliberate idleness in the underlying scheduling logic, since, in the considered operational context, the optimal scheduling policy is not necessarily non-idling. Hence, the triggering of transition T_{idle} consumes the transporter-modeling token, which remains in place P_{idle} , until the immediate transition T_{return} is enabled through the presence of a token in place P_{event} . P_{event} is marked every time that one of the system timed transitions fires, signaling the completion of some event. Notice that T_{return} will always be in conflict with transition T_{con} , but it is assumed to have priority over the latter, which is technically imposed by setting the corresponding (static) random switch to $\{\xi_{T_{return}} = 1, \xi_{T_{con}} = 0\}$. Finally, T_{con} is a sink transition that “consumes” event completion signaling tokens, in case that the transporter is not (deliberately) idling.

Synthesis and Control of Discrete Event Systems (B. Cailaud, X. Xie, P. Darondeau and L. Lavagno, Eds.). pp. -. Kluwer Academic Publishers (to appear).

Table 1. Example: The EMC markings

s_k	$P_{1t}P_{1i}P_{1p}P_{1o}$	$P_{2t}P_{2i}P_{2p}P_{2o}$	$P_{3t}P_{3i}P_{3p}P_{3o}P_{4t}$	$P_{MH}P_{idle}P_{event}$	$P_{S_1}P_{S_2}$	$P_{C_1}P_{C_2}P_{SCP}$
0	0000	0000	00000	100	11	122
1	1000	0000	00000	000	11	021
2	0000	0000	00000	010	11	122
3	0100	0000	00000	101	11	021
4	0010	0000	00000	010	01	021
5	0001	0000	00000	011	11	021
6	0001	0000	00000	100	11	021
7	0001	0000	00000	010	11	021
8	0000	1000	00000	000	11	111
9	0000	0100	00000	101	11	111
10	0000	0010	00000	100	10	111
11	0000	0010	00000	010	10	111
12	1000	0010	00000	000	10	010
13	0000	0001	00000	011	11	111
14	0000	0001	00000	100	11	111
15	0000	0001	00000	010	11	111
16	0000	0000	10000	000	11	022
17	1000	0001	00000	000	11	010
18	0000	0000	01000	101	11	022
19	0000	0000	00100	010	01	022
20	0000	0000	00010	011	11	022
21	0000	0000	00010	100	11	022
22	0000	0000	00010	010	11	022
23	0000	0000	00001	000	11	122
24	0000	0000	00000	101	11	122
25	0100	0010	00000	101	10	010
26	1000	0001	00000	001	11	010
27	0010	0010	00000	010	00	010
28	0001	0010	00000	011	10	010
29	0010	0001	00000	011	01	010
30	0001	0010	00000	100	10	010
31	0001	0010	00000	010	10	010
32	0000	1010	00000	000	10	100
33	0001	0001	00000	011	11	010
34	0001	0001	00000	100	11	010
35	0000	1001	00000	000	11	100
36	0001	0001	00000	010	11	010
37	0000	0101	00000	101	11	100
38	0000	0011	00000	010	10	100
39	0000	0002	00000	011	11	100
40	0000	0002	00000	100	11	100
41	0000	0002	00000	010	11	100
42	0000	0001	10000	000	11	011
43	0000	0001	01000	101	11	011
44	0000	0001	00100	010	01	011
45	0000	0001	00010	011	11	011
46	0000	0001	00010	100	11	011
47	0000	0001	00010	010	11	011
48	0000	0001	00001	000	11	111
49	0000	0001	00000	101	11	111
50	0100	0001	00000	101	11	010
51	0010	0001	00000	010	01	010
52	0010	0001	00000	100	01	010
53	0000	0110	00000	101	10	100
54	0000	1001	00000	001	11	100
55	0000	0110	00000	010	10	100
56	0000	0101	00000	011	11	100
57	0000	0011	00000	100	10	100
58	0000	0010	10000	000	10	011
59	0000	0001	10000	001	11	011
60	0000	0010	01000	101	10	011
61	0000	0010	00100	010	00	011
62	0000	0001	00100	011	01	011
63	0000	0010	00010	011	10	011
64	0000	0010	00010	100	10	011
65	0000	0010	00010	010	10	011
66	0000	0010	00001	000	10	111
67	0000	0001	00010	011	11	011
68	0000	0010	00000	101	10	111
69	0000	0001	00100	100	01	011
70	0000	0001	00001	001	11	111