An Analytical Investigation of the Deadlock Avoidance vs. Detection & Recovery Problem in Buffer-space Allocation of Flexibly Automated Production Systems*

Spyros A. Reveliotis
School of Industrial & Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332

Abstract

Motivated by recent developments in the semiconductor manufacturing industry, this paper undertakes an analytical investigation of the problem of selecting optimally the deadlock resolution strategy for buffer space allocation in flexibly automated production systems. In the process, it extends the behavioral models for the aforementioned systems currently considered in the literature, to account for probabilistic uncontrollable effects like the requirement for extra finishing steps and/or rework, and it introduces a new deadlock resolution scheme, characterized as *randomized* deadlock avoidance. The combination of these two extensions brings the considered system behavior(s) to the realm of probabilistic automata, an area of increasing academic interest. For the resulting model, and under the assumption of Markovian timings, it develops an analytical methodology for selecting the optimal deadlock resolution strategy that maximizes the steady-state system throughput, and it demonstrates its effectiveness through application to a "prototype" system configuration. The obtained results provide an interesting analytical expression of the need to assess the gains obtained by the increased concurrency supported by the deadlock detection & recovery strategy vs. the productivity losses experienced under this approach due to increased system blocking, and/or additional material handling overheads. It turns out that, for the considered system configuration, the optimal selection scheme switches between detection & recovery and pure deadlock avoidance, every time that the time cost of deadlock recovery, τ_d, crosses a threshold Θ, which is a function of the remaining system behavioral and timing parameters. Beyond its own theoretical merit, this last result raises also the question of whether the policy randomization introduced in this work will ever enhance the performance of *any* configuration in the considered class of resource allocation systems; this issue will be investigated in a sequel paper.

*This work was partially supported by NSF grant ECS-9979693

1 Introduction

The effective resolution of deadlock arising in the buffer-space allocation of flexibly automated production systems is a well-established problem in the current academic literature. According to the prevailing problem formulation(s) (e.g., [1, 20]), the production system is represented by a set of *resources* – modeling the system workstations and/or material handling system (MHS) components – each of which possesses finite buffering capacity, i.e., it can accommodate only a certain number of job instances at a time. Jobs processed through the system are characterized by an associated *job route / process plan*, i.e., a sequence of processing and/or transport operations executed at one of the system resources. Collectively, the system resources and the supported job routes are said to define a *resource allocation system (RAS)*. In this RAS context, *deadlock* arises whenever a set of the system resources is allocated to capacity to a set of jobs requiring for their further advancement another resource in the considered resource set. Unless corrective action is taken, deadlocked jobs are permanently blocked, and the utilization of the involved resource set is driven to zero. Hence, effective deadlock resolution is a key issue for the robust and efficient operation of flexibly automated production systems.

One of the first realizations in the deadlock-related research was the fact that from a strategic/conceptual standpoint, deadlocks can be either *prevented / avoided* or *detected and recovered* [4, 17]. Correct prevention / avoidance strategies never allow deadlock to arise by controlling the part flow and the underlying resource allocation according to some *deadlock avoidance policy (DAP)* that, furthermore, guarantees the *liveness* of the controlled system [1, 20].[1] On the other hand, the detection & recovery strategy is a reactive strategy, i.e., it allows deadlocks to occur, but it effectively detects them, and subsequently resolves them by staging some of the deadlocked jobs to auxiliary buffers dedicated to deadlock resolution, or by (temporarily) unloading them from the system.

Most of the academic research on deadlock resolution in flexibly automated production systems has taken the prevention / avoidance approach. Some representative results of this research can be found in [1, 5, 20, 24, 21, 11, 6]. In these works, justification for the preference of the prevention/avoidance paradigm over its detection & recovery counterpart has been typically based on the following two arguments:[2]

i. Deadlock recovery in the considered operational context involves the physical transfer of parts, which tends to be time-consuming, and occasionally dangerous to the quality of the transferred items. These effects are further aggravated by the fact that, under

[1] An alternative characterization of the requirement that the applied DAP guarantees the liveness of the controlled system is that the policy is *induced-deadlock free* [1, 20].

[2] We also believe that an additional reason that has turned the interest of the academic community to the prevention / avoidance approach is the fact that, in the considered operational context, the resulting research problem is much more challenging than its counterpart based on the detection & recovery strategy.

the detection & recovery regimen, a highly utilized production system will tend to be "clogged" with jobs most of the time, and therefore, the probability for deadlock formation will always be high.

ii. The fact that the job process plans supported by a given system configuration are stable and known a priori, allows the synthesis of deadlock avoidance policies which are *efficient*, in the sense that the restrictions imposed by these policies to the system part flow are kept to a minimum. In fact, under certain circumstances, the policy imposed constraints can be proved to be the *minimally* restrictive set of constraints establishing the liveness of the controlled system. The efficiency of the developed policies, in turn, implies increased operational concurrency and resource utilization, and under proper scheduling, effective exploitation of the system production capacity.

Recently, the issue of the dominance of deadlock prevention/avoidance over detection & recovery has been revived with the emergence of *cluster tools* [22] in the semiconductor manufacturing industry. This type of equipment essentially integrates a number of semiconductor manufacturing processes around a robotic manipulator, reducing, thus, the material handling required for the wafer transfer from process to process, while enhancing their protection from contamination, by maintaining a clean environment. On the other hand, this technology is very expensive, and therefore, it is imperative that the system productivity is maintained at its maximum. Given that the typical processing times required for the operations supported by these environments are considerably lengthy, whereas the part transfers - facilitated by the robotic manipulator - can be quite fast and safe, it is legitimate to question the extent to which a prevention/avoidance based deadlock resolution scheme is preferable over a deadlock detection & recovery approach. Specifically, the free part flow and the ensuing *maximal* operational concurrency established by the detection & recovery approach might lead to a better utilization of the system bottleneck process(es), without overloading too much the tool material handling system.

Yet, to the best of our knowledge, the problem of selecting an appropriate deadlock resolution strategy for a given RAS configuration has been given very limited consideration in the academic literature. Furthermore, existing results, e.g., [23, 12, 10], are of an empirical nature, primarily performance comparisons of the two approaches based on a number of performance indices which are evaluated through simulated experimentation on a series of randomly generated system configurations. Motivated by these remarks, in this work, we undertake a more theoretical investigation of the problem. Some of the questions that we attempt to address are as follows:

1. What formal framework is appropriate for the modeling and analysis of the effect of (the structural issue of) deadlock resolution on the system performance?

2. In the context of this framework, how can the problem of selecting the *optimal* deadlock resolution strategy be formally posed and analyzed?

3. Are prevention/avoidance and detection & recovery the only viable deadlock resolution strategies? If not, what other options exist?

4. What is the structure of the *optimal* strategy selection criterion?

In order to provide a more concrete exposition of the paper results, the paper addresses the above four questions in the operational context of the resource allocation system presented in Figure 1. The entire set of assumptions defining the system structure and operation is detailed in the next section. These assumptions have been designed such that they capture all the essential characteristics of the system operation under the considered deadlock resolution strategies, while maintaining the tractability of the pursued analysis. Furthermore, they extend the standard modeling and analysis paradigm for the deadlock resolution problem, by introducing a probabilistic dimension in the underlying RAS behavior, and an alternative deadlock resolution strategy, to be known as *randomized deadlock avoidance*. Having defined the system structure and behavior in Section 2, Section 3 addresses Questions (1) and (2) above, providing a performance evaluation framework for the operation of the considered system under the three deadlock resolution strategies, and a formal characterization of the optimal strategy selection problem. Section 4 identifies the conditions under which each of the two strategies – i.e., prevention/avoidance and detection & recovery – is the optimal selection for the considered system. Finally, Section 5 concludes the paper presentation and discusses future work.

2 The Considered Resource Allocation System and Randomized Deadlock Avoidance

2.1 The considered RAS

The production system to be considered in this work consists of a load/unload (L/U) station, where jobs are released into and retrieved from the system, two processing stations, W_1 and W_2, and a robotic manipulator that facilitates the part transfer among the system stations. To provide a more concrete application context, one can think of this system as a simple cluster tool (alt., robotic cell), where the two processing stations correspond to the tool *chambers* (alt., cell workstations), and the L/U station corresponds to the access port of the tool *stocker* (alt., cell automatic storage / retrieval system – AS/RS). Each of the two processing stations possesses unit buffering capacity, while the capacity of the L/U station is arbitrarily large, enabling free flow of jobs to and from the system. In its current configuration, the system supports the production of a single job type, with the following process plan: A job entering

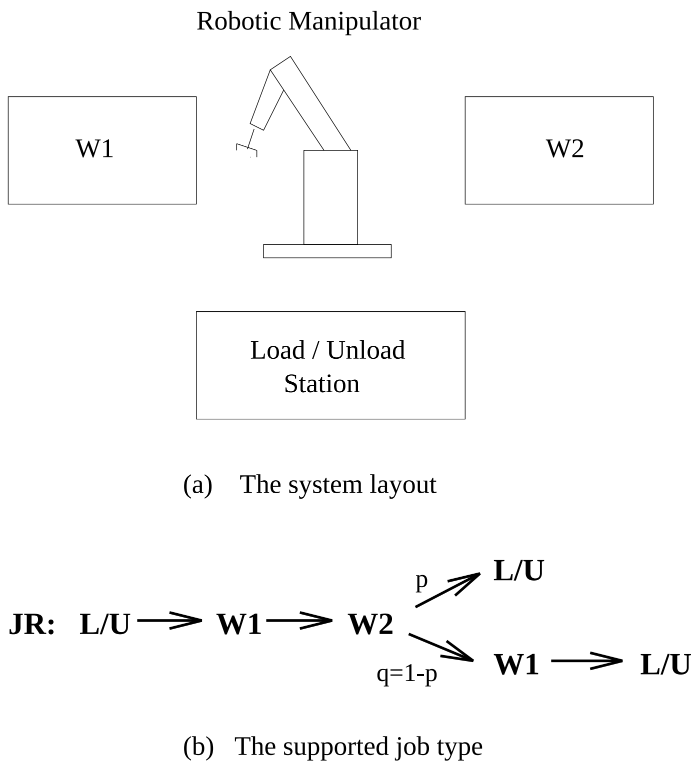


Figure 1: The considered production system

the system executes sequentially two processing stages, J_1 and J_2, supported respectively by workstations W_1 and W_2, in that order. Upon completion of stage J_2, the job either exits the system, with probability p, or it returns to workstation W_1, for an additional processing stage, J_3, with probability $q = 1 - p$.

The performance objective of interest in the undertaken analysis is the *maximization of the system throughput*, one of the most typical concerns in the semiconductor manufacturing industry [9, 2]. To facilitate the quantitative analysis of the system performance with respect to (w.r.t) this index, it is assumed that the processing times associated with job stages $J_i,\ i = 1, 2, 3$, are exponentially distributed, with respective means $\tau_i,\ i = 1, 2, 3$. Furthermore, loading a new job into the system requires an exponentially distributed time, with mean ϵ; this time includes interarrival delays, in case of random arrivals, and/or the time spent in fixturing/set-up operations that might be necessary before releasing a job into the system. On the other hand, the times associated with the job transfer among the different stations, and their unloading from the system, are assumed to be negligible. For notational convenience, we also define: $\mu_i \equiv 1/\tau_i,\ i = 1, 2, 3$, and $\lambda \equiv 1/\epsilon$. Finally, we notice that the *exponentiality* assumption regarding the various job processing and/or transferring times facilitates the modeling of the system behavior and the analysis of its performance through the well-developed theory of Continuous

Time Markov Chains (CTMC) [14]. However, the methodology developed in this paper can be generalized to arbitrary job processing and/or transferring time distributions by approximating these distributions through an appropriate *phase-type* distribution. The relevant theory and discussion of some standard techniques that can be used for developing these approximations can be found in ([16], Chpt. 1).

Under the natural assumption that a part is picked for transferring by the robotic manipulator only when the destination station has free buffering capacity, one can study the logical properties / structure of the system part flow by abstracting it to a RAS of two unit-capacity resource types, R_1 and R_2 – corresponding to the buffering capacity of workstations W_1 and W_2, respectively – and a single job type with the job route annotated in Figure 1. In this RAS, *deadlock* arises whenever resources R_1 and R_2 are loaded with jobs executing stages J_1 and J_2, respectively, and furthermore, the job in resource R_2, upon finishing its current stage, requires additional processing on resource R_1. Given the probabilistic nature / uncertainty characterizing job routes past stage J_2, it is easy to see that, in order to prevent this deadlock development, jobs must be loaded and processed through the system one at time.[3] On the other hand, the detection & recovery approach will allow deadlock to occur, and then, it will resolve it by *swapping* the job instances in resources R_1 and R_2. The time associated with the swapping operation models the underlying material handling delays, as well as the elapsed time until the deadlock is detected and addressed, and it is assumed to be exponentially distributed, with mean equal to τ_d; let us also denote $r \equiv 1/\tau_d$.

2.2 Randomized Deadlock Avoidance

In this section we discuss the repercussions of the probabilistic nature / structure of the considered job route for the modeling and analysis of the underlying system behavior, and for the development of an efficient deadlock resolution strategy. This novel modeling element extends the standard RAS model that has been typically used in the study of the deadlock resolution problem in production environments (cf., e.g., [1, 20, 6]), and intends to capture effects like the requirement for extra job finishing steps or rework arising in many contemporary production systems. From a theoretical standpoint, introducing a level of uncertainty in the evolution of the RAS job requests, brings the system behavior in the domain of *probabilistic automata* [18, 15], an area of resurgent interest in DES theory [13, 7]. In this modeling framework, an alternative strategy beyond the deadlock prevention / avoidance and detection & recovery, is to try to control the probability of deadlock formation, by *randomizing* appropriately the RAS transition to the unsafe territory of its behavioral space. In the context of the RAS considered

[3] It can be formally shown that this operational rule constitutes the *optimal* – i.e., minimally restrictive – correct DAP for the considered RAS (cf. Section 3).

in Figure 1, this randomization is materialized by allowing a new job to enter the system on workstation W_1 only with some probability ω, every time that another job is executing stage J_2 on workstation W_2. Hence, by varying the job acceptance probability ω over the interval $[0, 1]$, this strategy – which will be called *randomized deadlock avoidance* and denoted by R-DAP(ω) – establishes a continuum between the deadlock prevention / avoidance and the detection & recovery strategies – which correspond to the extreme cases of R-DAP(0.0) and R-DAP(1.0), respectively. Conceptually, randomized deadlock avoidance tries to establish a trade-off between the possibly excessive conservatism of the deadlock prevention / avoidance approach – e.g., in the case that the re-entrance probability q is fairly low – and the overhead costs experienced during the deadlock recovery process.

In the next two sections we proceed with the formal analysis of the system behavior and performance under these three alternative strategies – deadlock prevention / avoidance, randomized deadlock avoidance, and detection and recovery – and we establish the conditions under which each of these three strategies is the optimal selection for the RAS of Figure 1.

3 An Analytical Framework for the Performance Evaluation of the Considered RAS, and the Selection of the Optimal Deadlock Resolution Strategy

3.1 Outlining the overall methodology

In this section we develop an analytical framework that supports: (i) the performance evaluation of the production system in Figure 1, under the control of any R-DAP(ω), $\omega \in [0, 1]$, and (ii) the selection of the optimal value of the policy parameter ω, for various system parameterizations w.r.t. p, ϵ, τ_d, and τ_i, $i = 1, 2, 3$. Our general plan of work is as follows:

1. First, by exploiting the exponentiality assumption of the job processing and/or transfer times, we model the controlled system behavior as a *Continuous Time Markov Chain (CTMC)* [14]. The state space of this CTMC is determined by the operational assumptions incorporated in the RAS modeling the (logical) behavior of the considered system, and the control logic of the applied DAP. The rates associated with the various transitions/events of the system state space are determined by the first order moments of the various job processing and/or transfer times. Once the CTMC modeling the controlled system behavior has been developed, the system throughput can be straightforwardly obtained through standard CTMC steady-state analysis techniques.

2. To formulate the optimal deadlock resolution strategy selection problem, we recognize that the variation of the policy parameter ω over the interval [0,1] reflects in the CTMC

model of the controlled system behavior as a variation of a transition rate in the chain infinitesimal generator, between zero and a maximal value corresponding to the uncontrolled – i.e., non-randomized – occurrence of the associated job loading event. Hence, we convert the system of equations providing the CTMC performance evaluation – obtained in Step (1) – to an optimization problem, by allowing certain entries of the CTMC infinitesimal generator matrix, Φ, to vary between zero and some maximal values determined by the timing parameters of the original system definition.

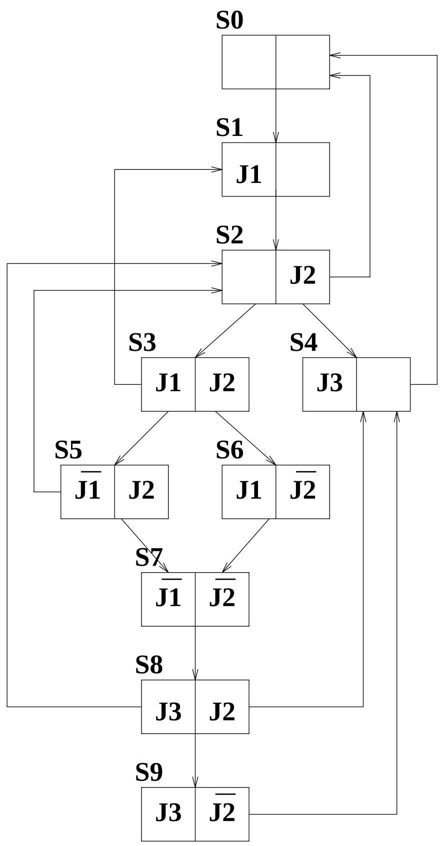
3. It turns out that the performance-optimizing formulation of Step (2) can be transformed to a *Linear Program (LP)* [3]. In this context, the optimality of the different deadlock resolution strategies – i.e., (i) deadlock prevention / avoidance: $\omega \equiv 0.0$, (ii) detection & recovery: $\omega \equiv 1.0$, and (iii) randomized deadlock avoidance: $\omega \in (0.0,\ 1.0)$ – can be translated to special structure of the optimal basic feasible solution(s) (bfs), and therefore, analyzed by the LP *duality theory / complementary slackness conditions* [3].

Steps (1) and (2) above are detailed in the rest of this section, while Step (3) is the content of Section 4.

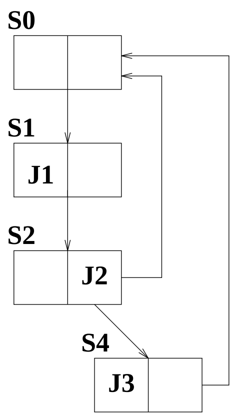
3.2 CTMC-based modeling and performance evaluation of the considered production system

The RAS state space As it has been previously established (cf. [20]), the logical behavior of the RAS modeling the buffer space allocation of a flexibly automated production system can be represented by a *Finite State Automaton (FSA)* [8]. The *states* of this automaton correspond to all possible allocation patterns of the system resources to the supported job types, arising during the (controlled) system operation; in particular, the FSA *initial* and *final* states correspond to the RAS idle and empty state. In the case that the deadlock prevention / avoidance strategy is applied, the *FSA state transitions / event types* correspond to the *loading*, *advancement* and *unloading* of single job instances running through the system. To facilitate the modeling of the system behavior under the broader class of randomized DAP's – which includes the detection & recovery case – and its subsequent performance evaluation through CTMC-based analysis, the previous event set is expanded by two additional transition / event types, corresponding to: (i) the *blockage* of a job finishing its currently executed stage due to the unavailability of the next requested resource, and (ii) the *resolution* of a developed deadlock though the swapping of an appropriately selected subset of the deadlocked jobs.

Given the operational assumptions for the production system of Figure 1 introduced in Section 2, the RAS state space modeling the controlled system behavior under R-DAP(ω), $\omega \in (0.0,\ 1.0]$, is depicted by the *state transition diagram (STD)* [8] of Figure (2,a). Each RAS



(a) $\omega \in (0.0,\ 1.0]$



(b) $\omega = 0.0$

Figure 2: The state space of the RAS modeling the production system of Figure 1, when controlled by R-DAP(ω), $\omega \in [0.0,\ 1.0]$

state in this STD is represented by two cells, indicating the allocation of the (unit) buffering capacity of resources R_1 and R_2, in that order. More specifically, the allocation of the buffering capacity of resource $R_i,\ i = 1, 2$, to a job instance executing stage $J_j,\ j = 1, 2, 3$, is depicted by marking / annotating the corresponding cell by that stage. Furthermore, the blockage of a job instance in its currently allocated resource is represented by *over-scoring* the corresponding cell annotation. Notice that the resulting state space consists of ten states, $s_0, \ldots, s_9$, with the *initial* and *final* FSA state corresponding to state s_0, and the arising *deadlock* state corresponding to state s_7. The transition that must be disabled in order to prevent the formation of the deadlock state s_7, is transition $(s_2,\ s_3)$, which corresponds to the loading of a new job in resource R_1 while another job instance is processed in resource R_2.[4] As a result, the logical behavior of the considered RAS when controlled according to the (optimized) deadlock prevention / avoidance strategy, is represented by the STD of Figure (2,b), which is a subgraph of that in Figure (2,a). In the following, we provide a unified treatment of the two cases, since the restriction of the RAS behavior to the subgraph of Figure (2,b), when controlled by the deadlock prevention avoidance strategy, can be obtained in the CTMC-based analysis by setting the rate of transition $(s_2,\ s_3)$ equal to zero.

The CTMC infinitesimal generator Under the exponentiality assumption of the various job processing, transferring, and deadlock resolution times,[5] and considering the notation introduced in Section 2, the dynamics of the system behavior when controlled by R-DAP(ω), $\omega \in [0.0,\ 1.0]$, – i.e., the *timed* pattern underlying the RAS transition among the various states depicted in the STD of Figure (2,a) – can be modeled by a *Continuous Time Markov Chain (CTMC)*, with the following *infinitesimal generator matrix*, $\bar{\Phi}$ [14]:

[4] All other transitions on the paths from state s_3 to s_7 are *uncontrollable* transitions [19], since they correspond to the blockage of certain job instances after finishing processing on their currently allocated resource.

[5] As it was mentioned in Section 2, the analytical framework proposed in this work can be effectively generalized to arbitrary timing distributions, through their approximation by appropriate *phase-type* distributions; cf. [16] for the technical details of this modeling mechanism.

$$
\bar{\Phi} = \begin{bmatrix}
-\lambda & \lambda & & & & & & & & \\
& -\mu_1 & \mu_1 & & & & & & & \\
p\ \mu_2 & & -(\omega\ \lambda + \mu_2) & \omega\ \lambda & q\ \mu_2 & & & & & \\
& p\ \mu_2 & & -(\mu_1 + \mu_2) & & \mu_1 & q\ \mu_2 & & & \\
\mu_3 & & & & -\mu_3 & & & & & \\
& & p\ \mu_2 & & & -\mu_2 & & q\ \mu_2 & & \\
& & & & & & -\mu_1 & \mu_1 & & \\
& & & & & & & -r & r & \\
& & \mu_3 & & p\ \mu_2 & & & & -(\mu_2 + \mu_3) & q\ \mu_2 \\
& & & & \mu_3 & & & & & -\mu_3
\end{bmatrix} \tag{1}
$$

CTMC-based performance evaluation It can be easily checked – cf. Figure (2,a) – that the CTMC defined by Equation 1 is irreducible, aperiodic (i.e., primitive), finite-state, and therefore, ergodic [14]. Thus, it has a *unique limiting stationary distribution / steady-state probability vector*, π, obtained by the following system of equations [14]:

$$
\pi^T \bar{\Phi} = 0.0 \tag{2}
$$

$$
\sum_{i=0}^{9} \pi_i = 1.0 \tag{3}
$$

Furthermore, given the stability implied by the ergodic nature of the system behavior, a characterization of the steady-state system throughput can be obtained by considering the cumulative rate according to which new jobs are loaded into the system. Recognizing that the steady state probabilities $\pi_i,\ i = 0, \ldots, 9$, can be interpreted as the percentage of time that the RAS spends in state $s_i,\ i = 0, \ldots, 9$, while the element $\bar{\phi}_{ij},\ i, j = 0, \ldots, 9, i \neq j$, denotes the rate according to which the system transitions from state i to state j, once in state i [14], this cumulative job loading rate, expressing the system throughput, can be computed as follows:

$$
TH(\bar{\Phi}) = \pi_0 \bar{\phi}_{01} + \pi_2 \bar{\phi}_{23} \tag{4}
$$

3.3 A Linear Programming Formulation for selecting the Optimal Deadlock Resolution Strategy

The initial mathematical programming formulation of the optimal strategy selection problem Equations 2, 3, and 4, combined, allow the steady-state performance (throughput) evaluation of any parameterization of the system of Figure 1, when operated under the control of R-DAP(ω), for any certain value of $\omega \in [0.0,\ 1.0]$. It follows, then, that the problem of selecting optimally the policy parameter ω for a given parameterization of the system of Figure 1, in terms of parameters $p,\ \epsilon,\ \tau_d$ and $\tau_i,\ i=1,2,3$, can be expressed by:

$$\max TH(\omega;\ p,\ \epsilon,\ \tau_d,\ \tau_i,\ i=1,2,3) \tag{5}$$

s.t.

$$\begin{aligned}
\pi^T\bar{\Phi}(\omega;\ p,\ \epsilon,\ \tau_d,\ \tau_i,\ i=1,2,3) &= 0.0 &(6)\\
\sum_{i=0}^{9}\pi_i &= 1.0 &(7)\\
TH &= \pi_0\bar{\phi}_{01}+\pi_2\bar{\phi}_{23} &(8)\\
\omega &\in [0.0,\ 1.0] &(9)
\end{aligned}$$

Notice that the variation of the policy parameter ω over the interval $[0.0, 1.0]$, essentially moderates over the interval $[0.0,\ \lambda]$, the rate according to which the underlying CTMC model transitions from state s_2 to state s_3, once in state s_2. In the following developments, we address the optimization problem defined by Equations 5 – 9, by considering a broader optimization problem that allows all the non-zero entries, ϕ_{ij}, of the infinitesimal generator matrix Φ to vary between the value of zero, and a maximal value $\bar{\phi}_{ij}$, defined by the system parameterization w.r.t. to parameters $p,\ \epsilon,\ \tau_d$, and $\tau_i,\ \imath=1,2,3$, according to the logic of Equation 1. This generalization of the original problem formulation, except from broadening the scope of – and therefore, strengthening – the obtained results, indicates also how the proposed framework can incorporate *scheduling* decisions, by modeling them as *regulations* of the occurrence rates of the various events taking place in the system. In its generalized version, the problem of selecting the optimal deadlock resolution strategy is stated as follows:

$$\max TH(\Phi;\ p,\ \epsilon,\ \tau_d,\ \tau_i,\ i=1,2,3) \tag{10}$$

s.t.

$$\begin{aligned}
\pi^T\Phi &= 0.0 &(11)\\
\sum_{i=0}^{9}\pi_i &= 1.0 &(12)\\
\forall i,\ \ \phi_{ii} &= -\sum_{j\neq i}\phi_{ij} &(13)\\
TH &= \pi_0\phi_{01}+\pi_2\phi_{23} &(14)\\
\forall i,j,\ i\neq j,\ \ \phi_{ij} &\in [0.0,\ \bar{\phi}_{ij}(p,\ \epsilon,\ \tau_d,\ \tau_i,\ i=1,2,3)] &(15)
\end{aligned}$$

Problem linearization The mathematical programming formulation of Equations 10 – 15 is a *non-linear* formulation, since it involves cross-products of the steady-state probability auxiliary variables, π_i, with the main decision variables, ϕ_{ij}. In general, these *bi-linear* formulations are characterized by a large number of local optima, and as a result, they are hard to solve to optimality. Fortunately, however, the formulation of Equations 10 – 15 can be *linearized* by:

1. eliminating Equation 13 from the set of constraints by direct substitution of ϕ_{ii}'s in Equation 11,

2. replacing the remaining product terms $\pi_i \phi_{ij},\ i \neq j$, with a dummy variable u_{ij},

3. rewriting the constraints of Equation 15 as:

$$\forall i, j\ i \neq j,\ u_{ij} \in [0.0,\ \pi_i \bar{\phi}_{ij}(p,\ \epsilon,\ \tau_d,\ \tau_i,\ i = 1, 2, 3)] \tag{16}$$

4. and imposing the additional requirement that:

$$q\ u_{ij} - p\ u_{ik} = 0.0 \tag{17}$$

   for every pair of transition rates $u_{ij},\ u_{ik}$ where transition (i, j) (resp., (i, k)) corresponds to the transfer of a job instance finishing the execution of stage J_2 in state s_i, out of the system (resp., to resource R_1 for the execution of stage J_3).

The resulting LP can be written as follows:[6]

$$\max u_{01} + u_{23} \tag{18}$$

s.t.

$$\forall i \in \{0, \ldots, 9\},\ \sum_{j \neq i} u_{ij} - \sum_{j \neq i} u_{ji} = 0.0 \tag{19}$$

$$\sum_{i=0}^{9} \pi_i = 1.0 \tag{20}$$

$$\forall i, j \in \{0, \ldots, 9\}, i \neq j, \quad u_{ij} - \bar{\phi}_{ij}(p,\ \epsilon,\ \tau_d,\ \tau_i,\ i = 1, 2, 3)\ \pi_i \leq 0.0 \tag{21}$$

$$\begin{aligned} q\ u_{20} - p\ u_{24} &= 0.0 \\ q\ u_{31} - p\ u_{36} &= 0.0 \\ q\ u_{52} - p\ u_{57} &= 0.0 \\ q\ u_{84} - p\ u_{89} &= 0.0 \end{aligned} \tag{22}$$

$$\forall i, j \in \{0, \ldots, 9\},\ i \neq j, \quad u_{ij} \geq 0.0 \tag{23}$$

[6] Notice that in the following LP formulation, Eq. number (22) characterizes the entire set of the four equations contained between this equation number and the previous one. This is a convention to be followed throughout the paper development, i.e., an equation number characterizes the equation in its own line, as well as any other *unnumbered* equations immediately preceding it. Beyond introducing some economies in the numbering scheme, this convention intends also to highlight the *logical* grouping of the paper equations during the development of the various arguments.

LP simplification In its basic formulation, the LP of Equations 18 – 23, has 10+1+90+4 = 105 *technological* constraints – i.e., all constraints excluding the variable non-negativity constraints – and 90 ($u_{ij},\ i \neq j$) + 10 (π_i) = 100 decision variables. In order to facilitate the analysis of this LP along the lines of Step (3) described in the general work plan of Section 3.1, we proceed with its simplification to an equivalent formulation containing the minimum possible number of variables and constraints. This simplification is achieved according to the following steps:

1. First, it should be recognized that variables u_{ij} corresponding to a zero entry of the matrix $\bar{\Phi}$, will be forced to zero by the corresponding constraint in Equation 21, and therefore, they can be omitted in the final formulation.

2. Further reduction of the remaining set of constraints and potentially non-zero u_{ij} variables can be achieved by solving Equations 22 for variables $u_{24},\ u_{36}, u_{57}$ and u_{89}, and substituting the obtained expressions in Equations 19 and 21.

3. Substitution of the expressions derived in Step (2) in Equation 21 indicates that the resulting constraints are redundant, and therefore, they can be eliminated from the final formulation. As a more concrete example, consider the expression $u_{24} = (q/p)u_{20}$, obtained from the first of Equations 22. Substitution of this expression in the constraint of Equation 21 corresponding to variable u_{24} leads to $(q/p)u_{20} - q\mu_2\pi_2 \leq 0.0$, which is equivalent to $u_{20} - p\mu_2\pi_2 \leq 0.0$, i.e., the constraint of Equation 21 corresponding to variable u_{20}. (cf. Eq. 1)

4. On the other hand, the system of the state balance Equations 19 contains 9 independent equations, which, after the reductions of Steps (1) and (2), involves 13 variables u_{ij}. These equations can be solved for 9 variables, giving:

$$\begin{aligned}
u_{52} &= pu_{35} \\
u_{31} &= p(u_{23} - u_{35}) \\
u_{12} &= u_{01} + p(u_{23} - u_{35}) \\
u_{67} &= q(u_{23} - u_{35}) \\
u_{78} &= qu_{23} \\
u_{82} &= qu_{23} - (1/p)u_{84} \\
u_{94} &= (q/p)u_{84} \\
u_{20} &= pu_{01} - u_{84} \\
u_{40} &= qu_{01} + u_{84}
\end{aligned} \tag{24}$$

5. Solving Equation 20 for π_0, and substituting the resulting expression, together with the expressions of Equation 24, in the remaining set of constraints of Equation 21, leads to the following, simplified version of the original LP formulation:

$$\max u_{01} + u_{23} \tag{25}$$

s.t.

$$\begin{array}{rcll}
u_{01} + \lambda \sum_{i=1}^{9} \pi_i & \leq & \lambda & (0,1) \\
u_{23} - \lambda \pi_2 & \leq & 0.0 & (2,3) \\
u_{35} - \mu_1 \pi_3 & \leq & 0.0 & (3,5) \\
u_{84} - p\mu_2 \pi_8 & \leq & 0.0 & (8,4) \\
u_{35} - \mu_2 \pi_5 & \leq & 0.0 & (5,2) \\
u_{23} - u_{35} - \mu_2 \pi_3 & \leq & 0.0 & (3,1) \\
u_{01} + pu_{23} - pu_{35} - \mu_1 \pi_1 & \leq & 0.0 & (1,2) \\
qu_{23} - qu_{35} - \mu_1 \pi_6 & \leq & 0.0 & (6,7) \\
qu_{23} - r\pi_7 & \leq & 0.0 & (7,8) \\
pqu_{23} - u_{84} - p\mu_3 \pi_8 & \leq & 0.0 & (8,2) \\
qu_{84} - p\mu_3 \pi_9 & \leq & 0.0 & (9,4) \\
pu_{01} - u_{84} - p\mu_2 \pi_2 & \leq & 0.0 & (2,0) \\
qu_{01} + u_{84} - \mu_3 \pi_4 & \leq & 0.0 & (4,0) \qquad (26) \\
u_{01},\ u_{23},\ u_{35},\ u_{84},\ \pi_i,\ i = 1, \ldots, 9 & \geq & 0.0 & \qquad\qquad (27)
\end{array}$$

Notice that the simplified LP has 13 technological constraints and 13 (nonnegative) variables. Furthermore, since the technological constraints of the above LP were finally derived through a series of substitutions to the original constraints of Eq. 21, each of them has been labeled by the indexing combination (i, j) that defines its counterpart in the original constraint set (cf. Eq. 21).

The dual LP To facilitate the development of the subsequent results, we conclude this section by providing also the *dual* of the LP formulation defined by Equations 25 – 27. Noticing that the primal LP is in *"normal form"* , its dual is derived by direct application of the basic definition of the "dual LP" concept (cf. [3], pg. 56), as follows:

$$\min \lambda y_{01} \tag{28}$$

s.t.

$$\begin{aligned}
y_{01} + y_{12} + py_{20} + qy_{40} &\geq 1.0 \\
y_{23} + y_{31} + py_{12} + qy_{67} + qy_{78} + pqy_{82} &\geq 1.0 \\
y_{35} + y_{52} - y_{31} - py_{12} - qy_{67} &\geq 0.0 \\
y_{84} - y_{82} + qy_{94} - y_{20} + y_{40} &\geq 0.0 \qquad (29) \\
\lambda y_{01} - \mu_1 y_{12} &\geq 0.0 \\
\lambda y_{01} - \lambda y_{23} - p\mu_2 y_{20} &\geq 0.0 \\
\lambda y_{01} - \mu_1 y_{35} - \mu_2 y_{31} &\geq 0.0 \\
\lambda y_{01} - \mu_3 y_{40} &\geq 0.0 \\
\lambda y_{01} - \mu_2 y_{52} &\geq 0.0 \\
\lambda y_{01} - \mu_1 y_{67} &\geq 0.0 \\
\lambda y_{01} - r y_{78} &\geq 0.0 \\
\lambda y_{01} - p\mu_2 y_{84} - p\mu_3 y_{82} &\geq 0.0 \\
\lambda y_{01} - p\mu_3 y_{94} &\geq 0.0 \qquad (30) \\
\forall i,j \quad y_{ij} &\geq 0.0 \qquad (31)
\end{aligned}$$

Constraints 29 correspond respectively to the primal variables u_{01}, u_{23}, u_{35} and u_{84}, while Constraints 30 correspond to the primal variables π_i, $i = 1, \ldots, 9$. Finally, the dual variables y_{ij} are in one-to-one correspondence with the constraints of Equation 26 in the primal LP formulation, and their indexing follows the constraint labeling introduced in Equation 26.

4 The Optimal Strategy Selection Criterion

Having rigorously defined the problem of selecting the optimal deadlock resolution strategy for the RAS of Figure 1 as an LP formulation, in this section we proceed to the derivation of an analytical (closed-form) selection criterion, that provides the optimal deadlock resolution strategy for all possible parameterizations of the RAS of Figure 1, w.r.t. parameters p, ϵ, τ_d, and τ_i, $i = 1, 2, 3$. From a methodological standpoint, the development of this section is as follows: Recognizing that the deadlock prevention / avoidance (resp., detection & recovery) strategy corresponds to a solution of the LP formulation of Eqs 25 – 27 with $u_{23} = 0.0$ (resp., $u_{23} - \lambda\pi_2 = 0.0$), we employ the LP *Complementary Slackness* theorem (cf. [3], Theorem 5.3) to derive the parameter subspaces in which each of these two strategies is optimal. It turns out that, for the considered RAS, the two parameter subspaces derived as indicated above, cover exhaustively the entire configuration parameter space, and therefore, the optimal selection strategy is a switching policy between the two extreme options of deadlock avoidance, and detection & recovery. Characteristically, the optimal selection scheme switches from deadlock detection & recovery to deadlock avoidance when the expected deadlock resolution time, τ_d, exceeds a threshold $\Theta = f(p, \epsilon, \tau_i, i = 1,2,3)$. Hence, at least for the considered RAS configuration, the policy randomization introduced in this work, will never lead to enhanced performance. An interpretation of this result is provided at the end of this section. The broader research issue of whether randomized deadlock avoidance will ever be *the* preferred strategy for *any* RAS configuration, under the considered modeling assumptions, is a research topic currently under investigation.

In the following, the optimality of each of the two deadlock resolution strategies is investigated in a separate subsection, while the obtained results are combined in a concluding theorem at the end of the section.

4.1 The optimality of the Deadlock Avoidance strategy

As it was stated above, under the LP formulation of the optimal deadlock resolution strategy selection problem, deadlock avoidance is expressed by the condition:

$$u_{23} = 0.0 \tag{32}$$

The reader should be able to verify that Equation 32 implies also (cf. Figure (2,b)):

$$\begin{aligned} \pi_3 = \pi_5 = \pi_6 = \pi_7 = \pi_8 = \pi_9 &= 0.0 & (33)\\ u_{35} = u_{31} = u_{52} = u_{67} = u_{78} = u_{82} = u_{84} = u_{94} &= 0.0 & (34) \end{aligned}$$

For the remaining variables, we have the following proposition:

Proposition 1 *Consider the RAS of Figure 1, operated under the prevention / avoidance strategy, i.e., under the condition expressed by Equation 32. Then,*

1.

$$\max TH = u^*_{01} = [\epsilon + \tau_1 + \tau_2 + q\tau_3]^{-1} \equiv W_0^{-1}$$

2.

$$\begin{aligned} u^*_{01} - \lambda\pi^*_0 &= 0.0\\ u^*_{12} - \mu_1\pi^*_1 &= 0.0\\ u^*_{20} - p\mu_2\pi^*_2 &= 0.0\\ u^*_{40} - \mu_3\pi^*_4 &= 0.0 \end{aligned}$$

Proof: In the light of Eqs. 32 – 34, application of the nodal balance equation 19 on the active states s_1, s_2 and s_4, gives:

$$\begin{aligned} u_{12} &= u_{01} \\ u_{20} &= pu_{12} = pu_{01} \\ u_{40} &= u_{24} = qu_{12} = qu_{01} \end{aligned} \tag{35}$$

Furthermore, from Eq. 21, we have:

$$\begin{aligned} u_{01} &\leq \lambda\pi_0 \\ u_{12} &\leq \mu_1\pi_1 \\ u_{20} &\leq p\mu_2\pi_2 \\ u_{40} &\leq \mu_3\pi_4 \end{aligned} \tag{36}$$

while, in the light of Eq. 33, Eq. 20 becomes:

$$\pi_0 + \pi_1 + \pi_2 + \pi_4 = 1.0 \tag{37}$$

Combining Eqs 35, 36,and 37, it follows that the maximal throughput obtained under the deadlock avoidance strategy, is provided by the following LP formulation:

$$\max u_{01} \tag{38}$$

s.t.

$$u_{01} + \lambda\pi_1 + \lambda\pi_2 + \lambda\pi_4 \leq \lambda \tag{39}$$

$$u_{01} - \mu_1\pi_1 \leq 0.0 \tag{40}$$

$$u_{01} - \mu_2\pi_2 \leq 0.0 \tag{41}$$

$$u_{01} - (\mu_3/q)\pi_4 \leq 0.0 \tag{42}$$

$$u_{01},\ \pi_1,\ \pi_2,\ \pi_4 \geq 0.0 \tag{43}$$

Then, the results of Proposition 1 can be obtained by showing that an optimal basis for the above LP is $B = \{u_{01}, \pi_1, \pi_2, \pi_4\}$.[7] $\square$

Having identified the structure of the solution of the LP formulation that maximizes the system throughput under the deadlock prevention / avoidance strategy, we proceed to the identification of the condition on the system parameterization that renders this solution a global optimum for the LP formulation of Eqs 25 – 27. We state and prove this result in the following proposition:

[7] An intuitive argument for the optimality of basis B can be based on the fact that, under the assumption of non-zero, finite, loading / processing times and branching probabilities, states s_1, s_2, and s_4 should have a non-zero steady state probability, and the system should present a strictly positive throughput.

Proposition 2 *Deadlock prevention / avoidance is an* optimal *deadlock resolution strategy for the RAS configuration of Figure 1, if and only if (iff)*

$$\tau_d \;\geq\; \frac{\epsilon}{q} + [\frac{\tau_1}{q(\tau_1+\tau_2)} + \frac{\tau_3}{\tau_2+\tau_3}]\tau_2 \tag{44}$$

Proof: As it was pointed in the introductory discussion of this section, the above result is obtained by applying the *complementary slackness (CS)* theorem (cf. [3], Theorem 5.3) on the LP formulation of Eqs 25 – 27, for the basic feasible solution (bfs) structure identified in Proposition 1. Specifically, the CS theorem implies that the bfs of Proposition 1 is optimal for the LP of Eqs 25 – 27 *iff* the following system of inequalities, in variables y_{ij}, is feasible:

$$\begin{aligned}
y_{01} + y_{12} + py_{20} + qy_{40} &= 1.0 \quad \text{(since } u_{01} > 0\text{)} && (45)\\
y_{23} + y_{31} + py_{12} + qy_{67} + qy_{78} + pqy_{82} &\geq 1.0 && (46)\\
y_{35} + y_{52} - y_{31} - py_{12} - qy_{67} &\geq 0.0 && (47)\\
y_{84} - y_{82} + qy_{94} - y_{20} + y_{40} &\geq 0.0 && (48)\\
\lambda y_{01} - \mu_1 y_{12} &= 0.0 \quad \text{(since } \pi_1 > 0\text{)} && (49)\\
\lambda y_{01} - \lambda y_{23} - p\mu_2 y_{20} &= 0.0 \quad \text{(since } \pi_2 > 0\text{)} && (50)\\
\lambda y_{01} - \mu_1 y_{35} - \mu_2 y_{31} &\geq 0.0 && (51)\\
\lambda y_{01} - \mu_3 y_{40} &= 0.0 \quad \text{(since } \pi_4 > 0\text{)} && (52)\\
\lambda y_{01} - \mu_2 y_{52} &\geq 0.0 && (53)\\
\lambda y_{01} - \mu_1 y_{67} &\geq 0.0 && (54)\\
\lambda y_{01} - r y_{78} &\geq 0.0 && (55)\\
\lambda y_{01} - p\mu_2 y_{84} - p\mu_3 y_{82} &\geq 0.0 && (56)\\
\lambda y_{01} - p\mu_3 y_{94} &\geq 0.0 && (57)\\
\forall i,j \quad y_{ij} &\geq 0.0 && (58)\\
y_{23} &= 0.0 \quad \text{(since } u_{23} = 0 < \lambda\pi_2 = \lambda\tau_2 W_0^{-1}\text{)} && (59)
\end{aligned}$$

Notice that the above system of inequalities is essentially defined by the constraint set of the *dual* LP formulation of Eqs 28 – 31, where some of the constraints have been tightened to equalities, and some of the (dual) variables have been forced to zero. Furthermore, in the light of the above discussion, the problem of identifying the optimality condition for the deadlock prevention / avoidance strategy has been reduced to the problem of identifying the conditions on the RAS parameterization space under which the system of inequalities defined by Eqs 45 – 59 is feasible. Next, we proceed with a series of transformations of the above system of inequalities that will allow us eventually to derive the desired condition.

Eqs 49, 50 and 52, combined with the fact that $y_{23} = 0$ (cf. Eq. 59), respectively imply that:

$$
\begin{aligned}
y_{12} &= \frac{\lambda}{\mu_1} y_{01} && (60)\\
y_{20} &= \frac{\lambda}{p\mu_2} y_{01} && (61)\\
y_{40} &= \frac{\lambda}{\mu_3} y_{01} && (62)
\end{aligned}
$$

which, combined with Eq. 45, give:[8]

$$
y_{01} = [1 + \frac{\lambda}{\mu_1} + \frac{\lambda}{\mu_2} + q\frac{\lambda}{\mu_3}]^{-1} = (\lambda W_0)^{-1} \qquad (63)
$$

Furthermore, Eqs 53, 54, 55 and 57, combined with Eqs 58 and 63, give:

$$
\begin{aligned}
0 \le y_{52} &\le \frac{\lambda}{\mu_2} y_{01} = \frac{\tau_2}{W_0} && (64)\\
0 \le y_{67} &\le \frac{\lambda}{\mu_1} y_{01} = \frac{\tau_1}{W_0} && (65)\\
0 \le y_{78} &\le \frac{\lambda}{r} y_{01} = \frac{\tau_d}{W_0} && (66)\\
0 \le y_{94} &\le \frac{\lambda}{p\mu_3} y_{01} = \frac{1}{p}\frac{\tau_3}{W_0} && (67)
\end{aligned}
$$

Finally, based on Eqs 59 – 63, the remaining equations 46, 47, 48, 51 and 56, can be respectively rewritten as:

$$
\begin{aligned}
y_{31} + qy_{67} + qy_{78} + pqy_{82} &\ge 1 - p\frac{\tau_1}{W_0} && (68)\\
y_{35} + y_{52} - y_{31} - qy_{67} &\ge p\frac{\tau_1}{W_0} && (69)\\
y_{84} - y_{82} + qy_{94} &\ge \frac{1}{p}\frac{\tau_2}{W_0} - \frac{\tau_3}{W_0} && (70)\\
\frac{1}{W_0} &\ge \mu_1 y_{35} + \mu_2 y_{31} && (71)\\
\frac{1}{W_0} &\ge p\mu_2 y_{84} + p\mu_3 y_{82} && (72)\\
y_{31},\ y_{35},\ y_{84},\ y_{82} &\ge 0.0 && (73)
\end{aligned}
$$

Notice that in the last set of inequalities, variables y_{52}, y_{78} and y_{94} appear only in a single position, and in a way that setting them to their upper value will only enhance the solution space of this system of inequalities w.r.t. the remaining variables. Based on this remark, and Eqs 64 – 67, above, the original CS condition reduces to:

$$
y_{31} + qy_{67} + pqy_{82} \ge 1 - p\frac{\tau_1}{W_0} - q\frac{\tau_d}{W_0} \qquad (74)
$$

[8] Notice that Eq. 63, combined with Eq. 28, imply that, as expected, the optimal value for the dual LP is equal to W_0^{-1}, i.e., equal to the optimal value of the primal LP (cf. Prop. 1).

$$
\begin{aligned}
y_{35} - y_{31} - q y_{67} &\geq p\frac{\tau_1}{W_0} - \frac{\tau_2}{W_0} && (75)\\
y_{84} - y_{82} &\geq \frac{1}{p}(\frac{\tau_2}{W_0} - \frac{\tau_3}{W_0}) && (76)\\
\mu_1 y_{35} + \mu_2 y_{31} &\leq \frac{1}{W_0} && (77)\\
\mu_2 y_{84} + \mu_3 y_{82} &\leq \frac{1}{p}\frac{1}{W_0} && (78)\\
y_{67} &\leq \frac{\tau_1}{W_0} && (79)\\
y_{31},\ y_{35},\ y_{84},\ y_{82},\ y_{67} &\geq 0.0 && (80)
\end{aligned}
$$

In a similar fashion, the feasibility of the above system of inequalities w.r.t. variables y_{31}, y_{82} and y_{67} is enhanced by setting:

$$
\begin{aligned}
y_{35} &= \frac{\tau_1}{W_0} - \tau_1\mu_2 y_{31} \geq 0 && (81)\\
y_{84} &= \frac{1}{p}\frac{\tau_2}{W_0} - \tau_2\mu_3 y_{82} \geq 0 && (82)
\end{aligned}
$$

which, by substitution to Eqs 75, 76, gives the transformed CS condition:

$$
\begin{aligned}
y_{31} + q y_{67} + pq y_{82} &\geq 1 - p\frac{\tau_1}{W_0} - q\frac{\tau_d}{W_0} && (83)\\
(1 + \frac{\tau_1}{\tau_2}) y_{31} + q y_{67} &\leq \frac{\tau_2 + q\tau_1}{W_0} && (84)\\
(1 + \frac{\tau_2}{\tau_3}) y_{82} &\leq \frac{1}{p}\frac{\tau_3}{W_0} && (85)\\
y_{67} &\leq \frac{\tau_1}{W_0} && (86)\\
y_{31} &\leq \frac{\tau_2}{W_0} && (87)\\
y_{82} &\leq \frac{1}{p}\frac{\tau_3}{W_0} && (88)\\
y_{31},\ y_{82},\ y_{67} &\geq 0.0 && (89)
\end{aligned}
$$

Finally, setting

$$
y_{82} = \frac{1}{p}\frac{\tau_3}{\tau_2 + \tau_3}\frac{\tau_3}{W_0} \qquad (90)
$$

satisfies Eq. 88, above, while it enhances the feasibility of the system of inequalities 83 – 89 w.r.t. the remaining variables y_{31} and y_{67}. The resulting system of inequalities is:

$$
\begin{aligned}
y_{31} + q y_{67} &\geq 1 - \frac{p\tau_1}{W_0} - \frac{q\tau_d}{W_0} - \frac{q\tau_3^2}{(\tau_2 + \tau_3) W_0} && (91)\\
\frac{\tau_1 + \tau_2}{\tau_2} y_{31} + q y_{67} &\leq \frac{\tau_2 + q\tau_1}{W_0} && (92)\\
y_{67} &\leq \frac{\tau_1}{W_0} && (93)\\
y_{31} &\leq \frac{\tau_2}{W_0} && (94)\\
y_{31},\ y_{67} &\geq 0.0 && (95)
\end{aligned}
$$

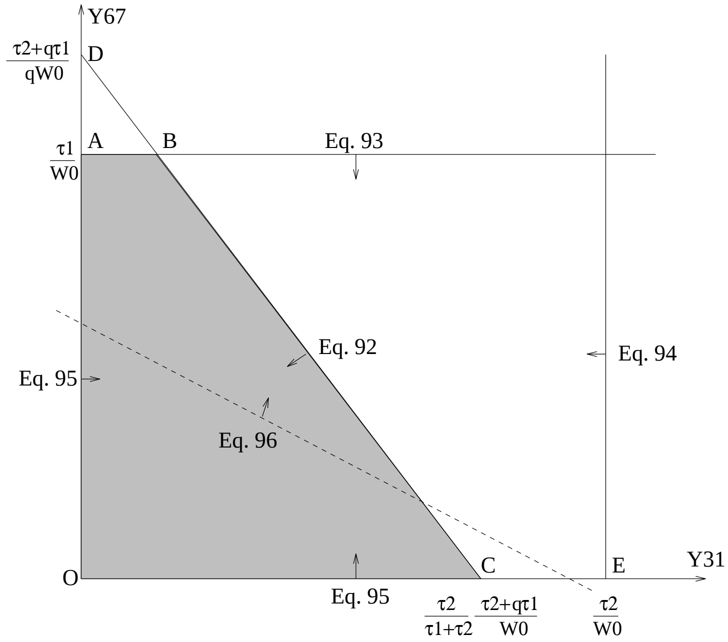


Figure 3: The topology of Eqs 92 – 96

Since the last system of inequalities involves only two variables, it is analyzed graphically, by means of Figure 3. Specifically, Figure 3 depicts the topology of the solution space defined by Eqs 92 – 95, which is delineated by the trapezoid $ABCO$, and furthermore, it indicates the relative positioning of the *"iso-profit"* lines corresponding to the equation

$$y_{31} + qy_{67} = \alpha \tag{96}$$

together with their direction of improvement. Notice that since the slope of these iso-profit lines is $s_x = -\frac{1}{q} \geq -\frac{\tau_1+\tau_2}{q\tau_2} = s_{BC}$, as α increases, they will leave the region delineated by trapezoid $ABCO$ through the extreme point B. Furthermore, since B is defined as the intersecting point of the lines corresponding to the binding of constraints 93 and 92, its coordinates are

$$y_{31}^B = \frac{\tau_2^2}{(\tau_1+\tau_2)W_0} \quad ; \quad y_{67}^B = \frac{\tau_1}{W_0} \tag{97}$$

which, when plugged into Eq. 96, give:

$$\alpha^B = \frac{\tau_2^2}{(\tau_1+\tau_2)W_0} + \frac{q\tau_1}{W_0} \tag{98}$$

It follows, then, that in the transformed CS condition expressed by Eqs 91 – 95, Eq. 91 will be compatible with the remaining system of inequalities 92 – 95, *iff*

$$1 - \frac{p\tau_1}{W_0} - \frac{q\tau_d}{W_0} - \frac{q\tau_3^2}{(\tau_2+\tau_3)W_0} \leq \frac{\tau_2^2}{(\tau_1+\tau_2)W_0} + \frac{q\tau_1}{W_0} \tag{99}$$

which after some algebraic manipulation results in the condition of Proposition 2. □

4.2 The optimality of the Detection & Recovery strategy

Under the LP formulation of the optimal deadlock resolution strategy selection problem for the RAS configuration of Figure 1, the deadlock Detection & Recovery strategy is expressed by the condition:

$$u_{23} - \lambda \pi_2 = 0.0 \tag{100}$$

This section identifies a condition on the RAS parameterization space under which the Detection & Recovery strategy, expressed by Eq. 100, is optimal w.r.t. throughput maximization. Specifically, we state and prove the following proposition:

Proposition 3 *Deadlock detection & recovery is an* optimal *deadlock resolution strategy for the RAS configuration of Figure 1, if*

$$\tau_d \leq \frac{\epsilon}{q} + [\frac{\tau_1}{q(\tau_1+\tau_2)} + \frac{\tau_3}{\tau_2+\tau_3}]\tau_2 \tag{101}$$

Furthermore, under the condition expressed by Eq. 101, an optimal solution for the throughput maximizing LP of Eqs 25 – 27, is obtained by binding all the constraints of Eq. 26.

Proof: The proof consists of two parts: First, we establish that the solution obtained by binding all the constraints of Eq. 26 is a basic feasible solution (bfs) for the LP of Eqs 25 – 27, and subsequently, using complementary slackness, we show that this bfs is optimal *iff* the condition of Eq. 101 is satisfied by the system parameters.

Part I: Establishing the feasibility of the considered bfs

Since Eq. 26 contains 13 linearly independent constraints in 13 variables, the solution obtained by their binding defines a basic solution for the LP of Eqs 25 – 27. To show that the resulting solution is also feasible, we proceed as follows:

Employing the constraint labeling introduced in Eq. 26, and recognizing that $1 - \sum_{i=1}^{9} \pi_i = \pi_0$, the binding of constraints (01), (23), (35), (84), (52), (12), (67), (78), (94) and (40) respectively gives:

$$\pi_0 = \epsilon u_{01} \tag{102}$$
$$\pi_2 = \epsilon u_{23} \tag{103}$$
$$\pi_3 = \tau_1 u_{35} \tag{104}$$
$$\pi_8 = \frac{1}{p}\tau_2 u_{84} \tag{105}$$
$$\pi_5 = \tau_2 u_{35} \tag{106}$$
$$\pi_1 = \tau_1 u_{01} + p\tau_1 u_{23} - p\tau_1 u_{35} \tag{107}$$
$$\pi_6 = q\tau_1 u_{23} - q\tau_1 u_{35} \tag{108}$$

$$\pi_7 = q\tau_d u_{23} \quad (109)$$
$$\pi_9 = \frac{q}{p}\tau_3 u_{84} \quad (110)$$
$$\pi_4 = q\tau_3 u_{01} + \tau_3 u_{84} \quad (111)$$

Furthermore, the binding of constraints (31) and (82) combined respectively with Eqs 104 and 105, above, gives:

$$u_{35} = \frac{\tau_2}{\tau_1+\tau_2} u_{23} \quad (112)$$
$$u_{84} = \frac{pq\tau_3}{\tau_2+\tau_3} u_{23} \quad (113)$$

while, the binding of constraint (20), combined with Eqs 103 and 113, gives:

$$u_{01} = (\frac{\epsilon}{\tau_2} + \frac{q\tau_3}{\tau_2+\tau_3}) u_{23} \quad (114)$$

Finally, Eqs 102 – 114, combined with the fact that $\sum_{i=0}^{9} \pi_i = 1.0$, give:

$$[(\epsilon+\tau_1+q\tau_3)(\frac{\epsilon}{\tau_2} + \frac{q\tau_3}{\tau_2+\tau_3}) + \epsilon + \tau_1 + q\tau_d + \frac{\tau_2^2}{\tau_1+\tau_2} + q\tau_3] u_{23} = 1.0 \quad (115)$$

Eq 115 implies that $u_{23} \geq 0$, while the nonnegativity of the remaining variables is inferred by Eqs 102 – 114, by noticing also that Eq. 112 implies that $u_{35} \leq u_{23}$.

Part II: Deriving the optimality condition for the considered bfs

Since the considered bfs contains all the primal variables and binds all the primal technological constraints, the LP CS theorem implies that it will be optimal *iff* the following system of equations, in variables y_{ij}, is feasible (cf. [3], Theorem 5.3, and the dual LP of Eqs 28 – 31):

$$y_{01} + y_{12} + py_{20} + qy_{40} = 1.0 \quad (116)$$
$$y_{23} + y_{31} + py_{12} + qy_{67} + qy_{78} + pqy_{82} = 1.0 \quad (117)$$
$$y_{35} + y_{52} - y_{31} - py_{12} - qy_{67} = 0.0 \quad (118)$$
$$y_{84} - y_{82} + qy_{94} - y_{20} + y_{40} = 0.0 \quad (119)$$
$$\lambda y_{01} - \mu_1 y_{12} = 0.0 \quad (120)$$
$$\lambda y_{01} - \lambda y_{23} - p\mu_2 y_{20} = 0.0 \quad (121)$$
$$\lambda y_{01} - \mu_1 y_{35} - \mu_2 y_{31} = 0.0 \quad (122)$$
$$\lambda y_{01} - \mu_3 y_{40} = 0.0 \quad (123)$$
$$\lambda y_{01} - \mu_2 y_{52} = 0.0 \quad (124)$$
$$\lambda y_{01} - \mu_1 y_{67} = 0.0 \quad (125)$$
$$\lambda y_{01} - r y_{78} = 0.0 \quad (126)$$
$$\lambda y_{01} - p\mu_2 y_{84} - p\mu_3 y_{82} = 0.0 \quad (127)$$

$$\lambda y_{01} - p\mu_3 y_{94} = 0.0 \tag{128}$$

$$\forall i, j \quad y_{ij} \geq 0.0 \tag{129}$$

Next we show that the above set of inequalities is feasible *iff* the condition of Equation 101, in Proposition 3, is satisfied. To this end, we first solve Eqs 120, 123, 124, 125, 126 and 128 respectively for variables y_{12}, y_{40}, y_{52}, y_{67}, y_{78} and y_{94}, obtaining:

$$\begin{aligned}
y_{12} &= \lambda\tau_1 y_{01} && (130)\\
y_{40} &= \lambda\tau_3 y_{01} && (131)\\
y_{52} &= \lambda\tau_2 y_{01} && (132)\\
y_{67} &= \lambda\tau_1 y_{01} && (133)\\
y_{78} &= \lambda\tau_d y_{01} && (134)\\
y_{94} &= \frac{1}{p}\lambda\tau_3 y_{01} && (135)
\end{aligned}$$

Next, combining Eq. 116 with Eqs 130 and 131, we obtain:

$$y_{20} = \frac{1}{p}[1 - (1 + \lambda\tau_1 + q\lambda\tau_3)y_{01}] \tag{136}$$

Eqs 121 and 136 give:

$$y_{23} = \mu_2\epsilon[(1 + \lambda\tau_1 + \lambda\tau_2 + q\lambda\tau_3)y_{01} - 1] \tag{137}$$

Eqs 118 and 122, combined with Eqs 130, 132 and 133, give:

$$\begin{aligned}
y_{31} &= \frac{\lambda\tau_2^2}{\tau_1 + \tau_2} y_{01} && (138)\\
y_{35} &= \frac{\lambda\tau_1^2}{\tau_1 + \tau_2} y_{01} && (139)
\end{aligned}$$

Eqs 119 and 127, combined with Eqs 131, 135 and 136, give:

$$\begin{aligned}
y_{82} &= \frac{\tau_3}{p(\tau_2 + \tau_3)}[(1 + \lambda\tau_1 + \lambda\tau_2 + (1+q)\lambda\tau_3)y_{01} - 1] && (140)\\
y_{84} &= \frac{\tau_2}{p(\tau_2 + \tau_3)}[1 - (1 + \lambda\tau_1 + q\lambda\tau_3)y_{01}] && (141)\\
&= \frac{\tau_2}{\tau_2 + \tau_3} y_{20} && (142)
\end{aligned}$$

Finally, Eq. 117, combined with Eqs 130, 133, 134, 137, 138 and 140, give:

$$\begin{aligned}
y_{01} = {} & [\tau_2^2 + (1+q)\tau_2\tau_3 + (\tau_2 + \tau_3)\epsilon](\tau_1 + \tau_2)\epsilon \cdot \\
& \{(\epsilon\tau_2 + \epsilon^2 + \tau_1\epsilon + q\tau_3\epsilon + \tau_1\tau_2 + q\tau_d\tau_2)(\tau_1 + \tau_2)(\tau_2 + \tau_3) + \\
& \tau_2^3(\tau_2 + \tau_3) + q\tau_2\tau_3(\tau_1 + \tau_2)[\epsilon + \tau_1 + \tau_2 + (1+q)\tau_3]\}^{-1} \qquad (143)
\end{aligned}$$

Hence, $y_{01} \geq 0$, which also implies immediately the nonnegativity of variables y_{12}, y_{40}, y_{52}, y_{67}, y_{78}, y_{94}, y_{31} and y_{35}. Furthermore, substitution of the result of Eq. 143 in Eq. 136 gives:

$$\begin{aligned} y_{20} \; = \; & (\tau_2^4 + \tau_2^3\tau_3 + q\tau_1\tau_2^2\tau_d + q\tau_2^3\tau_d + q\tau_1\tau_2\tau_3\tau_d + q\tau_2^2\tau_3\tau_d) \cdot \\ & \{p(\epsilon\tau_2 + \epsilon^2 + \tau_1\epsilon + q\tau_3\epsilon + \tau_1\tau_2 + q\tau_d\tau_2)(\tau_1 + \tau_2)(\tau_2 + \tau_3) + \\ & p\tau_2^3(\tau_2 + \tau_3) + pq\tau_2\tau_3(\tau_1 + \tau_2)[\epsilon + \tau_1 + \tau_2 + (1 + q)\tau_3]\}^{-1} \end{aligned} \tag{144}$$

which implies that $y_{20} \geq 0$. But then, Eq. 142 implies that $y_{84} \geq 0$, as well. Finally, noticing that Eqs 137 and 140 imply that $y_{23} \geq 0 \Longrightarrow y_{82} \geq 0$, it can be concluded that the system of equations 116 – 129 is feasible *iff*

$$y_{23} \geq 0 \tag{145}$$

Eq. 145, combined with Eqs 137, 143 and some algebraic manipulation, leads to Eq. 101 of Proposition 3. □

4.3 An Optimal Strategy Selection Criterion

Eventually, an optimal deadlock resolution strategy selection criterion for the RAS of Figure 1, can be obtained by combining the results of Propositions 2 and 3. This criterion is stated in the following theorem:

Theorem 1 *Consider the steady-state operation of the RAS of Figure 1, under the structural control of some policy R-DAP(ω), $\omega \in [0,1]$. Then, an optimal selection of the policy parameter ω that maximizes the system throughput is as follows:*

$$\omega_{opt} := \begin{cases} 0.0 & \textit{if} \;\; \tau_d \;\geq\; \frac{\epsilon}{q} + [\frac{\tau_1}{q(\tau_1+\tau_2)} + \frac{\tau_3}{\tau_2+\tau_3}]\tau_2 \\ 1.0 & \textit{o.w.} \end{cases} \tag{146}$$

Proof: This theorem is an immediate consequence of Propositions 2 and 3. □

Notice that, from a strictly theoretical standpoint, the criterion of Theorem 146 is *an* optimal selection criterion and not *the* optimal selection criterion. This is because the condition of the CS theorem, used in the derivation of this criterion, guarantees the *optimality* of a given primal feasible solution, but not its *uniqueness* (cf [3], Theorem 5.3). An interesting practical implication, though, of Theorem 1 is that, at least for the RAS configuration considered in this work, randomization of the system transitions to its unsafe operational space does not enhance its performance w.r.t. the attained throughput. An intuitive interpretation of this qualitative result can be obtained by rewriting the inequality defining the selection criterion threshold in the following form:

$$\omega_{opt} := \begin{cases} 0.0 & \text{if} \;\; q\tau_d \;\geq\; \epsilon + \frac{1}{\mu_1+\mu_2} + q\frac{1}{\mu_2+\mu_3} \\ 1.0 & \text{o.w.} \end{cases} \tag{147}$$

In this new expression, the right-hand-side (rhs) of the thresholding inequality corresponds to the expected work gain that will be achieved by the system every time that it ventures in its unsafe region – i.e., it performs transition $(2,3)$ in the STD of Figure (2,a) – compared to the workload fulfilled by the system during a pure deadlock avoidance operational cycle: specifically, the first term reflects the work gain obtained by loading a new job into the system while another job is in process on workstation W_2, while the next two terms express the work gains obtained through the concurrency supported by states s_3 and s_8. On the other hand, the left-hand-side (lhs) of the inequality in Eq. 147 expresses the expected idleness that will be experienced by the system during a visit to the unsafe operational region, due to deadlock recovery. Hence, under this hindsight, the optimal selection criterion merely says that, in the operational context of the considered RAS, *we should be willing to take the risk and burdens of venturing into the unsafe region if and only if the expected benefit - measured in terms of extra fulfilled work - is higher than the expected loss - measured in terms of system extra idleness*; in that case, we should risk 100%. It remains an open research question whether there exist RAS configurations for which the optimal deadlock resolution strategy is obtained for $\omega \in (0,1)$.

5 Conclusions

Motivated by recent developments in the semiconductor manufacturing industry, this paper has undertaken an analytical investigation of the problem of selecting the optimal deadlock resolution strategy for buffer space allocation in flexibly automated production systems. In the process, we have extended the behavioral models for the aforementioned systems currently considered in the literature, to account for probabilistic uncontrollable effects like the requirement for extra finishing steps and/or rework, and we have introduced a new deadlock resolution scheme, characterized as randomized deadlock avoidance. The combination of these two extensions brings the considered system behavior(s) to the realm of probabilistic automata, an area of increasing academic interest. For the resulting model, and under the assumption of Markovian timings, we have developed an analytical formulation for selecting the optimal deadlock resolution strategy that maximizes the steady-state system throughput, and we have demonstrated its effectiveness by application to the "prototype" RAS configuration of Figure 1. The obtained results provide an interesting analytical expression of the need to assess the gains obtained by the increased concurrency supported by the deadlock detection & recovery strategy vs. the costs of the idleness caused by the increased system blocking, and the overheads of the increased material handling requirements. Finally, given that for the considered system the optimal selection criterion switches between detection & recovery and *pure* deadlock avoidance, every time that the time cost of deadlock recovery, τ_d, crosses a threshold Θ, defined as a function of the remaining system parameters p, ϵ and τ_i, $i = 1, 2, 3$, it remains to be seen whether the policy randomization introduced in this work enhances the performance of *any* configuration in the considered RAS class; this is part of our current investigations.

References

[1] Z. A. Banaszak and B. H. Krogh. Deadlock avoidance in flexible manufacturing systems with concurrently competing process flows. *IEEE Trans. on Robotics and Automation*, 6:724–734, 1990.

[2] D. A. Bodner and S. A. Reveliotis. Flexible semiconductor manufacturing. In J. G. Webster, editor, *Encyclopedia of Electrical and Electronics Engineering, Vol. 7*, pages 596–607. John Wiley & Sons, 1999.

[3] V. Chvátal. *Linear Programming.* W. H. Freeman & Co., N.Y., N.Y., 1983.

[4] E. G. Coffman, M. J. Elphick, and A. Shoshani. System deadlocks. *Computing Surveys*, 3:67–78, 1971.

[5] J. Ezpeleta, J. M. Colom, and J. Martinez. A petri net based deadlock prevention policy for flexible manufacturing systems. *IEEE Trans. on R&A*, 11:173–184, 1995.

[6] M. P. Fanti, B. Maione, S. Mascolo, and B. Turchiano. Event-based feedback control for deadlock avoidance in flexible production systems. *IEEE Trans. on Robotics and Automation*, 13:347–363, 1997.

[7] V. Garg, R. Kumar, and S. Marcus. A probabilistic language formalism for stochastic discrete-event systems. *IEEE Trans. Autom. Control*, 44:280–293, 1999.

[8] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages and Computation.* Addison-Wesley, Reading, MA, 1979.

[9] P. R. Kumar. Scheduling semiconductor manufacturing plants. *IEEE Control Systems Magazine*, 14–6:33–40, 1994.

[10] T. Kumaran, W. Chang, H. Cho, and R. Wysk. A structured approach to deadlock detection, avoidance and resolution in flexible manufacturing systems. *IJPR*, 32:2361–2379, 1994.

[11] M. Lawley, S. Reveliotis, and P. Ferreira. A correct and scalable deadlock avoidance policy for flexible manufacturing systems. *IEEE Trans. on Robotics & Automation*, 14:796–809, 1998.

[12] Y. T. Leung and G. J. Sheen. Resolving deadlocks in flexible manufacturing cells. *Journal of Manufacturing Systems*, 12:291–304, 1994.

[13] Y. Li, F. Lin, and Z-H. Lin. Supervisory control of probabilistic discrete-event systems with recovery. *IEEE Trans. Autom. Control*, 44:1971–1975, 1999.

[14] J. Medhi. *Stochastic Models in Queueing Theory.* Academic Press, San Diego, CA, 1991.

[15] A. Paz. *Introduction to Probabilistic Automata.* New York: Academic, New York, NY, 1971.

[16] H. G. Perros. *Queueing Networks with Blocking: Exact and Approximate Solutions.* Oxford University Press, NY, NY, 1994.

[17] J. L. Peterson. *Operating System Concepts.* Addison-Wesley, 1981.

[18] M. O. Rabin. Probabilistic automata. *Inform. Contr.*, 6:230–245, 1963.

[19] P. J. G. Ramadge and W. M. Wonham. The control of discrete event systems. *Proceedings of the IEEE*, 77:81–98, 1989.

[20] S. A. Reveliotis and P. M. Ferreira. Deadlock avoidance policies for automated manufacturing cells. *IEEE Trans. on Robotics & Automation*, 12:845–857, 1996.

[21] S. A. Reveliotis, M. A. Lawley, and P. M. Ferreira. Polynomial complexity deadlock avoidance policies for sequential resource allocation systems. *IEEE Trans. on Automatic Control*, 42:1344–1357, 1997.

[22] P. Singer. The driving forces in cluster tool development. *Semiconductor International*, July '95:113–118, 1995.

[23] R. A. Wysk, N. S. Yang, and S. Joshi. Resolution of deadlocks in flexible manufacturing systems: Avoidance and recovery approaches. *Journal of Manufacturing Systems*, 13:128–138, 1994.

[24] K. Y. Xing, B. S. Hu, and H. X. Chen. Deadlock avoidance policy for petri net modeling of flexible manufacturing systems with shared resources. *IEEE Trans. on Aut. Control*, 41:289–295, 1996.