Dynamic Allocation of Reconfigurable Resources in a Two-Stage Tandem Queueing System with Reliability Considerations

Cheng-Hung Wu, Mark E. Lewis and Michael Veatch

Abstract

Consider a two stage tandem queueing system, with dedicated machines in each stage. Additional reconfigurable resources can be assigned to one of these two stations without setup cost and time. In a clearing system (without external arrivals) both with and without machine failures, we show the existence of an optimal monotone policy. Moreover, when all of the machines are reliable, the switching curve defined by this policy has slope greater than or equal to -1. This continues to hold true when the holding cost rate is higher at the first stage and machine failures are considered.

I. Introduction

In recent years, competition and market changes have caused customized products to become more popular. To meet the challenge, many factory managers choose reconfigurable machines to improve robustness under uncertain demand [1]. When the demand variation is high, Narongwanich et al. [2] suggest that reconfigurable capacity can be particularly valuable. This value may be amplified when machine failures are considered. In this paper we consider a tandem queueing system where there exists an external, potentially expensive, reconfigurable resource. Machines are subject to failure and the resource can be configured to work at either station. We show for a clearing system (without external arrivals) that the optimal policy is characterized by

Cheng-Hung Wu is a Ph.D. candidate in the Department of Industrial and Operations Engineering at The University of Michigan, 1205 Beal Avenue, Ann Arbor, MI 48109, wuch@umich.edu

Professor Lewis is a member of the School of Operations Research and Industrial Engineering at Cornell University, 226 Rhodes Hall, Ithaca, NY 14853, melewis@orie.cornell.edu

Professor Veatch is a member of the Department of Mathematics at Gordon College, Wenham, MA 01984, veatch@gordon.edu

a switching curve. Without reliability considerations this curve is shown to have slope greater than or equal to -1. In the case with machine failures and the added assumption that the holding cost rate at station 1 is higher than station 2, the same bound on the slope holds.

Previous clearing system analyses of flexible systems (where servers can perform several tasks) include [3], [4], [5]. Ahn et al. [3] examined a clearing system problem with two-stage tandem queues and two flexible servers. They show that an exhaustive policy for the upstream or downstream queue is optimal. Farrar [5] shows the existence of an optimal policy that is *monotone* in a tandem queueing system with one fixed server at each stage and a controllable service rate at the first station. We extend this result to cover the case when the server can serve at either station and also to include machine breakdowns. Our approach establishes properties of the value function directly, rather than showing that properties are preserved by value iteration as in Hajek [4] and Veatch and Wein [6].

In contrast, when external arrivals are allowed, Hajek [4] shows that the optimal reconfigurable machine assignment in a parallel queueing system follows a monotone switching curve. Ahn et al. [7] show that the results of [3] extend under the average cost criterion. Duenyas et al. [8] and Iravani [9] consider the optimal control of a single flexible server in a tandem queueing system. The control of two interconnected queues with identical machines is discussed by Javidi et al. [10]. The allocation of flexible workers with regard to maximizing throughput is considered in [11], [12], [13] and [14]. None of the previous work considers machine failures in a reconfigurable manufacturing system with regards to minimizing total expected holding costs.

II. Model Formulation and Statement of Main Results

Consider a two-station tandem queueing system. Jobs move from station 1 to station 2 and then exit the system. There are N_1 and N_2 dedicated servers in station 1 and station 2, respectively. In addition, there are N_r reconfigurable servers that can be configured at any time to serve at either station. There are no external arrivals. All jobs require an exponentially distributed amount of service with mean 1. Let the service rate of the k^{th} dedicated server at station ℓ be denoted $\mu_{k,\ell}$, for $\ell = 1, 2$. When there are fewer jobs at a station than available servers, we assume that the service rates are additive; servers can *collaborate* on a single job. Similarly, the k^{th} reconfigurable resource, which serves at rate $\mu_{k,r}$, can also collaborate. Although this assumption has always been reasonable for large scale operations (automotive assembly, parallel processing in computer systems), it is becoming more common even in small scale operations such as the drilling of several bores simultaneously rather than in series. Let the failure (repair) time distribution of the k^{th} reconfigurable server and k^{th} dedicated server at station ℓ be exponential with rates $\alpha_{k,r}$ ($\beta_{k,r}$) and $\alpha_{k,\ell}$ ($\beta_{k,\ell}$), respectively. Without loss of generality, assume $\sum_k \left(\mu_{k,r} + \alpha_{k,r} + \beta_{k,r}\right) + \sum_{k,\ell} \left(\mu_{k,\ell} + \alpha_{k,\ell} + \beta_{k,\ell}\right) = 1$.

For each job at station 1 (2), a holding cost rate of h_1 (h_2) per unit time is accrued. Let $\mathbb{X} = \{(i,j,m,n,r) | i,j \in \mathbb{Z}^+, m = (m_1, m_2, ..., m_{N_1}), n = (n_1, n_2, ..., n_{N_2}), r = (r_1, r_2, ..., r_{N_r}), m_k \in \{0,1\}, n_k \in \{0,1\}, r_k \in \{0,1\}\}$ be the state space, where i and j represent the number of customers (including that in service) at stations 1 and 2, respectively, m_k and n_k denote the machine status (0 = failed, 1 = online) of the dedicated servers at each station, and r_k denotes the machine status of the k^{th} reconfigurable machine. Suppose that (X_t, Y_t) denote the number of customers at stations 1 and 2 at time t. For a policy π that describes where to allocate the reconfigurable resource for all time and $x \in \mathbb{X}$, define $v^{\pi}(x) := \mathbb{E}_x^{\pi} \int_0^{\infty} (h_1 X_t + h_2 Y_t) dt$. The optimal cost say $v(x)$ minimizes v^{π} over all non-anticipating, non-idling policies π. We note here that although it might be optimal to idle some servers, the non-idling assumption seems congruent with current practice. The following example illustrates the usefulness of the reconfigurable resource.

Example 2.1: Suppose there is one reconfigurable server, only one dedicated server in each station, and the reconfigurable server never fails. Given the following inputs: $\mu_1 = 1; \mu_2 = 3; \mu_r = 1; \alpha_1 = \alpha_2 = 0.001; \beta_1 = \beta_2 = 0.01; h_1 = h_2 = 1$. The expected total cost at state $(i,j,m,n,r) = (10,10,1,1,1)$ of a system that does not have a reconfigurable resource is 87.7 while that with access to this capacity (used optimally) is 63.0; a 28.2% improvement. Also, note that if the allocation policy of the resource is simply to only use the reconfigurable resource at station 1 or station 2, the total costs are 71.5 and 76.5, respectively; still 9.7% and 15.4% away from optimal.

In order to simplify notation, extend v to $v(i,-1,m,n,r) = v(i,0,m,n,r)$ and $v(-1,j,m,n,r) = v(0,j-1,m,n,r)$. Denote the failure and repair transitions, e.g., $F_k(m) = (m_1, m_2, ..., m_k 0, ..., m_{N_1})$ and $R_k(m) = (m_1, m_2, ..., m_k = 1, ..., m_{N_1})$. Furthermore, let $< a,b >= \sum a_i b_i$ and $\mathbf{1}$ be a vector of all ones. Then v satisfies the dynamic programming (DP) optimality equations.

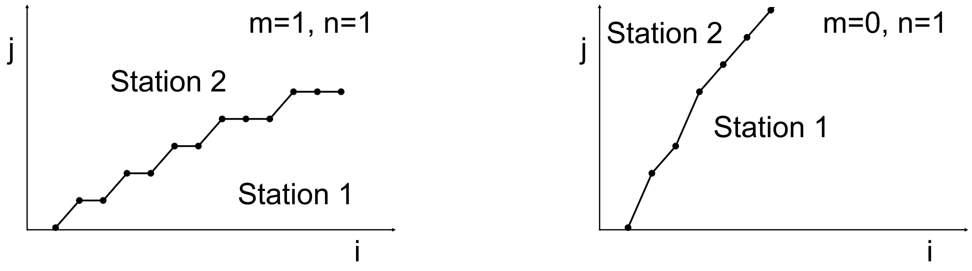
$$v(i,j,m,n,r) =< r, \mu_r > \min\{v(i-1,j+1,m,n,r), v(i,j-1,m,n,r)\} + u(i,j,m,n,r), \qquad \text{(II.1)}$$

where

$$\begin{aligned} u(i,j,m,n,r) = ih_1 + jh_2 &+ \sum_k \mu_{k,1} m_k v(i-1,j+1,m,n,r) + \sum_k \mu_{k,2} n_k v(i,j-1,m,n,r) \\ &+ \sum_k \alpha_{k,1} v(i,j,F_k(m),n,r) + \sum_k \alpha_{k,2} v(i,j,m,F_k(n),r) + \sum_k \alpha_{k,r} v(i,j,m,n,F_k(r)) \\ &+ \sum_k \beta_{k,1} v(i,j,R_k(m),n,r) + \sum_k \beta_{k,2} v(i,j,m,R_k(n),r) + \sum_k \beta_{k,r} v(i,j,m,n,R_k(r)) \\ &+ \Big(< \mathbf{1} - m, \mu_m > + < \mathbf{1} - n, \mu_n > + < \mathbf{1} - r, \mu_r > \Big) v(i,j,m,n,r). \end{aligned}$$

It is clear from the minimum in (II.1) that it is optimal to place all reconfigurable servers at station 1(2) if $v(i-1,j+1,m,n,r) \leq (\geq) v(i,j-1,m,n,r)$. We say that a policy is *transition monotone* (cf. [5]) if for fixed $i,j,m,n,$ and r, $v(i,j-1,m,n,r) \leq v(i-1,j+1,m,n,r)$ implies that $v(i,j+k-1,m,n,r) \leq v(i-1,j+k+1,m,n,r)$ for all $k \geq 0$.

It can be inferred that for a transition monotone policy for each fixed i, m, n and r there is a threshold level, say $L(i)$, such that for $j > L(i)$ it is optimal to use the reconfigurable machine at station 2, otherwise use it at station 1. The function $L(i)$ is called a *switching curve*. Figure 1 depicts optimal switching curves for $(m,n,r) = (0,1,1)$ and $(1,1,1)$. The inputs are the same as Example 2.1. Unfortunately, we were unable to prove that the switching curve is monotone in general although our numerical work seems to confirm it.



(a) (m,n,r) = (1,1,1)

(b) (m,n,r) = (0,1,1)

Fig. 1. An example of (optimal) switching curves.

The remainder of the paper is dedicated to proving the following theorem.

Theorem 2.2: The following hold

1) there exists an optimal policy that is transition monotone;

2) there exists a switching curve defined by an optimal transition monotone policy that has slope greater than or equal to -1 when $h_1 \geq h_2$ or when the machines are reliable.

We conclude this section with some results that are quite intuitive but will simplify the analysis.

Proposition 2.3: The following hold:

1) the optimal cost function, $v(i, j, m, n, r)$ is non-decreasing in both i and j;
2) $v(0, j+2, m, n, r) - v(0, j+1, m, n, r) \geq v(0, j+1, m, n, r) - v(0, j, m, n, r)$ for all $j \geq 0$;
3) if $h_1 \geq h_2$, $v(i, j, m, n, r) \geq v(i-1, j+1, m, n, r)$.

Proof. To show the first assertion above we prove $v(i+1, j, m, n, r) \geq v(i, j, m, n, r)$. The fact that v is non-decreasing in j is analogous. Consider two processes defined on the same probability space. Process 1 starts in state $(i+1, j, m, n, r)$ and uses the optimal policy, while Process 2 starts in state (i, j, m, n, r) and uses the policy ϕ that mimics the policy followed by Process 1 at each time point with the caveat that if Process 1 allocates the server to station 1 when the station is empty for Process 2, Process 2 avoids idling by allocating the server to station 2. Since the processes are defined on the same space, they see the same service times, failures and repairs. The first time the allocations might differ is when Process 2 empties station 1. If the dedicated servers of Process 1 complete the remaining job at station 1, the processes continue to have a difference of one customer (at station 2) until Process 1 empties. If the remaining job in station 1 is completed by the reconfigurable resources, Process 2 sees this completion in station 2 and the difference in the number of jobs is 2. That is to say that Process 1 has (at least) one more customer in the system than Process 2 and at least as many at each station until such time that Process 1 empties. Hence, $v(i+1, j, m, n, r) - v(i, j, m, n, r) \geq v(i+1, j, m, n, r) - v^{\phi}(i, j, m, n, r) \geq 0$, and the result is proven.

The second result is proved similarly. Define 3 processes on the same probability space starting in states $(0, j+2, m, n, r)$, $(0, j+1, m, n, r)$, and $(0, j, m, n, r)$, respectively and proceeding optimally. Since we have assumed the policies are non-idling, the processes maintain their relative queuelength positions (Process 1, one more customer than Process 2 and Process 2 one more than Process 3) until Process 3 empties and Processes 1 and 2 see a service that is not seen by Process 3. Before this occurs, the difference in the holding costs per unit time is zero ($h_2 - h_2$). After this time Process 1 has one customer and the others are empty; the differential cost rate is then h_2. That is

$$v(0, j+2, m, n, r) - v(0, j+1, m, n, r) - v(0, j+1, m, n, r) + v(0, j, m, n, r) \geq h_2 \, \mathbb{E} \, X \geq 0,$$

where X is the service time of this last customer.

To show the third result, again consider two processes defined on the same probability space with Process 1 starting in (i, j, m, n, r) and using the optimal policy. Process 2 starts in $(i - 1, j+1, m, n, r)$ and mimics Process 1 as previously indicated. There are two important events to consider. In the first, Process 2 empties the first station followed (perhaps after several other events) by a service at station 1 by Process 1 (not seen by Process 2). Since prior to this event, both processes have seen the same services, failures and repairs, Process 1 has one more customer at station 1 and one less at station 2. After the extra service is seen by Process 1, the processes couple and accrue the same costs. The difference in the holding costs up until this time is $h_1 - h_2 \geq 0$ (per unit time). The second event of interest is if Process 1 empties the second queue followed by a service at station 2 in Process 2 (not seen by Process 1). After this event, both processes have the same number of customers at station 2, but Process 1 has one more customer at station 1. Assume that both processes then follow the optimal policy. The difference in the holding costs up until this time is $h_1 - h_2 \geq 0$ (per unit time) while after this time the difference is $v(i'+1, j', m', n', r') - v(i', j', m', n', r')$ for some (i', j', m', n', r'). This difference is non-negative from the first assertion. Letting ϕ denote the policy used by Process 2 we again have that $v(i, j, m, n, r) - v(i-1, j+1, m, n, r) \geq v(i, j, m, n, r) - v^{\phi}(i-1, j+1, m, n, r) \geq 0$, and the result is proven. ■

III. Resource Allocation Without Reliability Considerations

In this section, we show the existence of a transition monotone optimal policy for the problem without reliability considerations. The method forms a baseline and will be extended to the case with machine failures. However, we view the results as interesting without this extension. We also note that in this case the non-idling assumption (on the reconfigurable servers) is not restrictive; there exists optimal non-idling policies within the broader class of potentially idling policies. The proof of this fact has been omitted for brevity.

Assume that all machines are reliable (no failures). Since in this case $m_k = n_k = r_k = 1$ and $\alpha_{k,r} = \alpha_{k,\ell} = 0$ for all k and ℓ, this is equivalent to considering a single dedicated server at each station with service rate $\mu_\ell = \sum_k \mu_{k,\ell}$ for $\ell = 1,2$ and a single reconfigurable server with service rate $\mu_r = \sum_k \mu_{k,r}$. The uniformization rate is $\mu_1 + \mu_2 + \mu_r = 1$. The DP equations $v(i,j)$ now simplify considerably:

$$v(i,j) = (Tv)(i,j), \tag{III.1}$$

where

$$(Tv)(i,j) = ih_1 + jh_2 + \mu_1 v(i-1,j+1) + \mu_2 v(i,j-1) + \mu_r\big(v(i-1,j+1) - [v(i-1,j+1) - v(i,j-1)]^+\big).$$

It is optimal to place the reconfigurable server at station 1 if and only if $v(i-1,j+1) - v(i,j-1) \geq 0$. We establish the following submodularity condition for all $i \geq 0$ and $j \geq 0$

$$v(i-1,j+1) - v(i,j-1) \geq v(i-1,j) - v(i,j-2), \tag{III.2}$$

which implies transition monotonicity. Note that Proposition 2.3 also applies to $v(i,j)$. We will prove (III.2) using only Proposition 2.3 and (III.1). This is done by nested induction by showing that (III.2) holds along each strip of the form (i,j) where $i \geq 0$ is fixed and j ranges over the positive integers.

When $i = j = 1$, (III.2) holds since $v(0,2) - v(1,0)v(0,2) - v(1,-1) \geq v(0,1) - v(1,-1)$. Assume (III.2) for $(1,j-1)$ and $j \geq 2$. For $(1,j)$ note

$$\begin{aligned} v(0,j+1) - v(0,j) = h_2 &+ \mu_1(v(0,j+1) - v(0,j)) + \mu_2(v(0,j) - v(0,j-1)) + \mu_r(v(0,j+1) - v(0,j)) \\ &- \mu_r([v(0,j+1) - v(0,j)]^+ - [v(0,j) - v(0,j-1)]^+). \end{aligned} \tag{III.3}$$

and

$$\begin{aligned} v(1,j-1) - v(1,j-2) = h_2 &+ \mu_1(v(0,j) - v(0,j-1)) + \mu_2(v(1,j-2) - v(1,j-3)) + \mu_r(v(0,j) - v(0,j-1)) \\ &- \mu_r([v(0,j) - v(1,j-2)]^+ - [v(0,j-1) - v(1,j-3)]^+), \end{aligned} \tag{III.4}$$

Compare (III.3) and (III.4) term by term. The h_2 terms cancel and the difference in the second and third terms is non-negative by the convexity of $v(0,j)$ with respect to j and the inductive hypothesis, respectively. Consider now the difference in the terms with coefficient μ_r. Since v is nondecreasing, we can drop both "+" operators in (III.3). Thus, since $[v(0,j) - v(1,j-2)]^+ - [v(0,j-1) - v(1,j-3)]^+ \geq 0$ holds by the inductive hypothesis, (III.2) holds for $i = 1$.

Consider now $i > 1$. First note that (III.2) holds for $(i,1)$ since $v(i-1,2) - v(i,0) = v(i-1,2) - v(i,-1) \geq (v(i-1,1) - v(i,-1))$. Using this result and the result for $(1,j)$ (for all $j > 0$), we make two inductive assumptions; first that (III.2) holds for all (m,n) with $m < i$

and $n > 0$ and second that (III.2) holds for all (i, n) such that $n \leq j$. It remains to show that (III.2) holds at $(i, j+1)$. A little algebra yields

$$\begin{aligned} v(i-1,j+2)-v(i-1,j+1) = h_2 &+ \mu_1(v(i-2,j+3)-v(i-2,j+2)) + \mu_2(v(i-1,j+1)-v(i-1,j)) \\ &+ \mu_r(v(i-2,j+3)-v(i-2,j+2)) - \mu_r\big([v(i-2,j+3)-v(i-1,j+1)]^+ \\ &- [v(i-2,j+2)-v(i-1,j)]^+\big), \end{aligned} \tag{III.5}$$

and

$$\begin{aligned} v(i,j)-v(i,j-1) = h_2 &+ \mu_1(v(i-1,j+1)-v(i-1,j)) + \mu_2(v(i,j-1)-v(i,j-2)) \\ &+ \mu_r(v(i-1,j+1)-v(i-1,j)) \\ &- \mu_r\big([v(i-1,j+1)-v(i,j-1)]^+ - [v(i-1,j)-v(i,j-2)]^+\big). \end{aligned} \tag{III.6}$$

The first three terms in (III.5) are greater than the respective term in (III.6) by simple algebra and the inductive assumptions at $(i-1, j+2)$ and (i, j), respectively. By using the hypothesis at (i, j) note, $[v(i-1,j+1)-v(i,j-1)]^+ - [v(i-1,j)-v(i,j-2)]^+ \geq 0$. Thus,

$$\begin{aligned} &(v(i-1,j+1)-v(i-1,j)) - \big([v(i-1,j+1)-v(i,j-1)]^+ - [v(i-1,j)-v(i,j-2)]^+\big) \\ &\quad\leq (v(i-2,j+3)-v(i-2,j+2)) - \big(v(i-2,j+3)-v(i-1,j+1)) - (v(i-2,j+2)-v(i-1,j))\big) \\ &\quad\leq (v(i-2,j+3)-v(i-2,j+2)) - \big([v(i-2,j+3)-v(i-1,j+1)]^+ - [v(i-2,j+2)-v(i-1,j)]^+\big), \end{aligned}$$

where the second inequality follows from the inductive hypothesis applied at $(i-1, j+2)$ and the proof is complete. This implies the existence of an optimal transition monotone policy.

We next show that the switching curve has a slope ≥ -1. This holds if the following inequality holds (refer to (III.2)),

$$v(i-1,j+1)-v(i,j-1) \geq v(i,j)-v(i+1,j-2) \tag{III.7}$$

for all (i, j) such that it is optimal to allocate the reconfigurable resource to station 1. Intuitively, this implies that the reconfigurable machine tends to stay at station 1 if i is increased and the workload for station 2 is constant. First, we establish some useful properties of $v(i, j)$.

Lemma 3.1: For $i \geq 1$

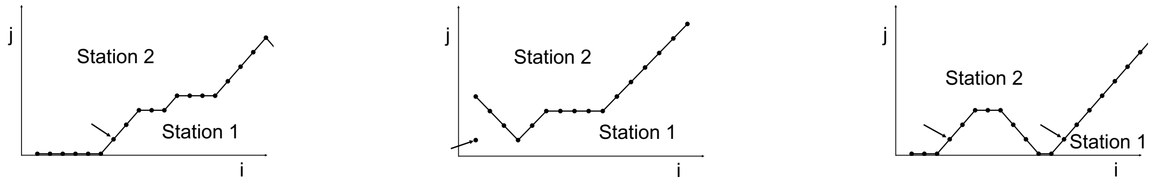
1) $v(i-1,1)-v(i,0) \geq v(i,1)-v(i+1,0)$
2) $v(i-1,1) \leq v(i,0)$

Proof. Rearranging the optimality equation at $(i, 0)$ shows (2) and yields $v(i-1,1)-v(i,0) = \frac{-ih_1}{\mu_1+\mu_r} \geq \frac{-(i+1)h_1}{\mu_1+\mu_r} = v(i,1)-v(i+1,0)$. ■

Let $A_K := \{(i,j) \in (\mathbf{Z}^+ \times \mathbf{Z}^+) | i+j = K, i \geq 1, j \geq 1\}$ be the set of states for which there are K service completions required at station 2 before emptying the system, where $\mathbf{Z}^+$ is the set of positive integers. Define $B_K := \{(i,j) \in A_K | j \leq L(i)\}$. Note that B_K represents the subset of A_K such that it is optimal for the reconfigurable machine to work at station 1. Using the fact that B_{k+1} is not accessible from B_k, we show inductively that (III.7) holds for all $(i,j) \in B_K$ and $K \geq 2$. Let

$$S_K: \quad \text{If } (i,j) \in B_K, \text{ then (III.7) holds.} \tag{III.8}$$

Observe that for fixed K, the hypothesis of S_K may not hold for any $(i,j) \in A_K$. In this case, we say that S_K holds *vacuously* and search for the minimum K such that B_K is non-empty to establish the induction basis. Denote this minimum by k and note that k depends on the optimal switching curve (see Figure 2). In Figure 2(a) it is optimal to serve at station 2 for $i \leq 6$ and $j = 1$ while it is optimal to serve at station 1 for $i = 7$, $j = 1$. The induction basis would be established at $k = 8$. In Figure 2(b), it is optimal to serve at station 1 at $i = j = 1$; the induction basis is established at $k = 2$. Moreover, note that there may exist points $K > k$ such that B_K is empty. Suppose there exists $K > k$ for which B_K is empty. Define ℓ to be the minimum point larger than K such that B_ℓ is non-empty. That is to say, the switching curve reaches zero for a second time. The arguments that follow hold for $k \leq K < \ell$ such that B_K is non-empty and can be repeated at ℓ and so on (see Figure 2(c)).



(a) k=8 (b) k=2 (c) k=5, ℓ = 14

Fig. 2. Possible (optimal) switching curves. The leftmost (rightmost) arrow represents the point that defines k (ℓ, when necessary).

The existence of an optimal transition monotone policy implies that $B_k = \{(k-1, 1)\}$. Also using Lemma 3.1 (part 2) for $k < \infty$,

$$\begin{aligned} v(k-n-1,n+1) - v(k-n,n-1) &\geq 0, \quad \text{for } n = 2, \text{ and} \\ v(k-n-1,n+1) - v(k-n,n-1) &\leq 0, \quad \text{for } n = 0,1. \end{aligned} \tag{III.9}$$

That is to say that it is optimal for the reconfigurable resource to be allocated to station 2 in state $(k-2,2)$ while it is optimal to allocate it to station 1 in states $(k,0)$ and $(k-1,1)$. Note that $(k-1,1)$ is the point that the arrow points to in Figure 2.

Assume first that state $(1,1) \in B_2$ so that $k = 2$ may be used as the inductive basis. Note,

$$\begin{aligned} v(0,2) - v(1,0) = &- h_1 + 2h_2 + \mu_1\big(v(0,2) - v(0,1)\big) + \mu_2\big(v(0,1) - v(1,0)\big) \\ &+ \mu_r\big(v(0,2) - v(0,1)\big) - \mu_r\big([v(0,2) - v(0,1)]^+ - [v(0,1) - v(1,0)]^+\big) \end{aligned} \tag{III.10}$$

$$\begin{aligned} v(1,1) - v(2,0) = &- h_1 + h_2 + \mu_1\big(v(0,2) - v(1,1)\big) + \mu_2\big(v(1,0) - v(2,0)\big) \\ &+ \mu_r\big(v(0,2) - v(1,1)\big) - \mu_r\big([v(0,2) - v(1,0)]^+ - [v(1,1) - v(2,0)]^+\big) \end{aligned} \tag{III.11}$$

Comparing (III.10) and (III.11) term by term we note that the difference in the constant terms is clearly positive while $v(0,2) - v(0,1) \geq v(0,2) - v(1,1)$ follows from the fact that $v(1,1) \geq v(0,1)$. The inequality holds for the terms with coefficient μ_2 since by (1) of Lemma 3.1, $v(0,1) - v(1,0) \geq v(1,1) - v(2,0) \geq v(1,0) - v(2,0)$. Moreover, for the terms with coefficient μ_r in (III.10) note that since $v(i,j)$ is non-decreasing in j the first two of these terms cancel. The last term is zero since $v(0,1) \leq v(1,0)$. For the analogous terms in (III.11), the fact that $v(1,0) \leq v(1,1)$ implies that $v(0,2) - v(1,1) \leq [v(0,2) - v(1,0)]^+$. Finally, (2) of Lemma 3.1 implies that the last term is zero; and the result holds for $k = 2$.

If $k \geq 3$, the DP equations yield

$$\begin{aligned} v(k-2,2) - v(k-1,0) = &-h_1 + 2h_2 + \mu_1(v(k-3,3) - v(k-2,1)) \\ &+ \mu_2(v(k-2,1) - v(k-1,0)) + \mu_r(v(k-3,3) - v(k-2,1)) \\ &- \mu_r([v(k-3,3) - v(k-2,1)]^+ - [v(k-2,1) - v(k-1,0)]^+) \end{aligned} \tag{III.12}$$

$$\begin{aligned} v(k-1,1) - v(k,0) = &-h_1 + h_2 + \mu_1(v(k-2,2) - v(k-1,1)) + \mu_2(v(k-1,0) - v(k,0)) \\ &+ \mu_r(v(k-2,2) - v(k-1,1)) \\ &- \mu_r([v(k-2,2) - v(k-1,0)]^+ - [v(k-1,1) - v(k,0)]^+) \end{aligned} \tag{III.13}$$

In order to show that (III.7) holds, note again that comparing the first terms of (III.12) and (III.13) yields $-h_1 + 2h_2 \geq -h_1 + h_2$. The difference in the terms associated with μ_1 is non-negative by (III.9) and the fact that $v(k-1,1) \geq v(k-1,0)$. Consider the terms associated with μ_r. By (1) of Lemma 3.1, the last terms in (III.12) and (III.13) are zero. Furthermore,

$$\begin{aligned}
&\mu_r\big(v(k-3,3) - v(k-2,1) - [v(k-3,3) - v(k-2,1)]^+\big) \\
&\quad \geq \mu_r\big(v(k-2,2) - v(k-1,0) - [v(k-2,2) - v(k-1,0)]^+\big) \\
&\quad \geq \mu_r\big(v(k-2,2) - v(k-1,1) - [v(k-2,2) - v(k-1,0)]^+\big), \qquad \text{(III.14)}
\end{aligned}$$

where the inequalities follow from (1) of Lemma 3.1 and the fact that $v(i,j)$ is non-decreasing in i and j. We remark that the proof of the inductive basis is valid not only for the basis but also for all states $(i,1) \in B_{i+1}$ with the only change needed being that we invoke the inductive hypothesis for the terms associated with μ_1. Thus, we need only consider $j \geq 2$ in the following induction process.

Assume now for $K > k$ that S_{K-1} holds. Since the optimal reconfigurable machine location is station 1 at state $(K,0)$ and station 2 at state $(0,K)$, there exists a state $(i,j) \in A_K$, in which using the reconfigurable machine at station 1 is optimal for state (i,j) and at station 2 for $(i-1,j+1)$. This (coupled with the monotonicity in j for $i=1$) implies $v(i-2,j+2) - v(i-1,j) \geq 0 \geq v(i-1,j+1) - v(i,j-1)$. Our inductive proof for K starts from that specific state. This proof can extended to all states in B_K. The DP equations yield

$$\begin{aligned}
v(i-1,j+1) - v(i,j-1) = &-h_1 + 2h_2 + \mu_1(v(i-2,j+2) - v(i-1,j)) \\
&+ \mu_2(v(i-1,j) - v(i,j-2)) + \mu_r(v(i-2,j+2) - v(i-1,j)) \\
&- \mu_r\big([v(i-2,j+2) - v(i-1,j)]^+ - [v(i-1,j) - v(i,j-2)]^+\big), \qquad \text{(III.15)}
\end{aligned}$$

$$\begin{aligned}
v(i,j) - v(i+1,j-2) = &-h_1 + 2h_2 + \mu_1(v(i-1,j+1) - v(i,j-1)) \\
&+ \mu_2(v(i,j-1) - v(i+1,j-3)) + \mu_r(v(i-1,j+1) - v(i,j-1)) \\
&- \mu_r\big([v(i-1,j+1) - v(i,j-1)]^+ - [v(i,j-1) - v(i+1,j-3)]^+\big) \qquad \text{(III.16)}
\end{aligned}$$

The difference in the terms with coefficient μ_1 is non-negative by our previous discussion about $(i-1,j+1)$. This implies

$$\begin{aligned}
&(v(i-2,j+2) - v(i-1,j)) - [v(i-2,j+2) - v(i-1,j)]^+ \\
&\quad \geq v(i-1,j+1) - v(i,j-1) - [v(i-1,j+1) - v(i,j-1)]^+.
\end{aligned}$$

Since we have the existence of an optimal transition monotone policy, using the reconfigurable resource at station 1 is also optimal in state $(i, j-1)$. The inductive assumption, S_{K-1}, implies that (III.7) holds for $(i, j-1)$ so that $[v(i-1, j) - v(i, j-2)]^+ \geq [v(i, j-1) - v(i+1, j-3)]^+$. This yields that the difference in the terms associated with μ_2 and μ_r are non-negative and the result is proven for (i, j).

Since (i, j) can be any state in A_K for which it is optimal to serve at station 1 while it is optimal to serve at station 2 in $(i-1, j+1)$, S_K is true for all $K \geq 2$. This guarantees that the slope of the switching curve $L(i)$ is greater than or equal to -1.

IV. Resource Allocation with Reliability Considerations

In this section, we extend the results of the previous section to the case with reliability considerations. For now, we assume there is only one reconfigurable server and one dedicated server in each station. Let $\psi := \mu_1 + \mu_2 + \mu_r + \alpha_1 + \alpha_2 + \alpha_r$. For $(m, n, r) \in \{0, 1\}^3$, let

$$\begin{aligned}(T_{mnr}v)(i, j) &:= ih_1 + jh_2 + m\mu_1 v(i-1, j+1, m, n, r) + n\mu_2 v(i, j-1, m, n, r)\\ &\quad + r\mu_r\big(v(i-1, j+1, m, n, r) - [v(i-1, j+1, m, n, r) - v(i, j-1, m, n), r]^+\big).\end{aligned}$$

We make the important observation that T_{mnr} has the same form as T in (III.1), except that the coefficients of v on the right hand side no longer sum to one. Recall that the proof of (III.2) for reliable systems did not require the coefficient to sum to 1 and only used Proposition 2.3, and (III.1). Hence the same proof shows that (III.2) holds for $(T_{mnr}v)(i, j)$.

Consider $v(i, j, 1, 1, 1)$ and rearrange terms in (II.1) to get

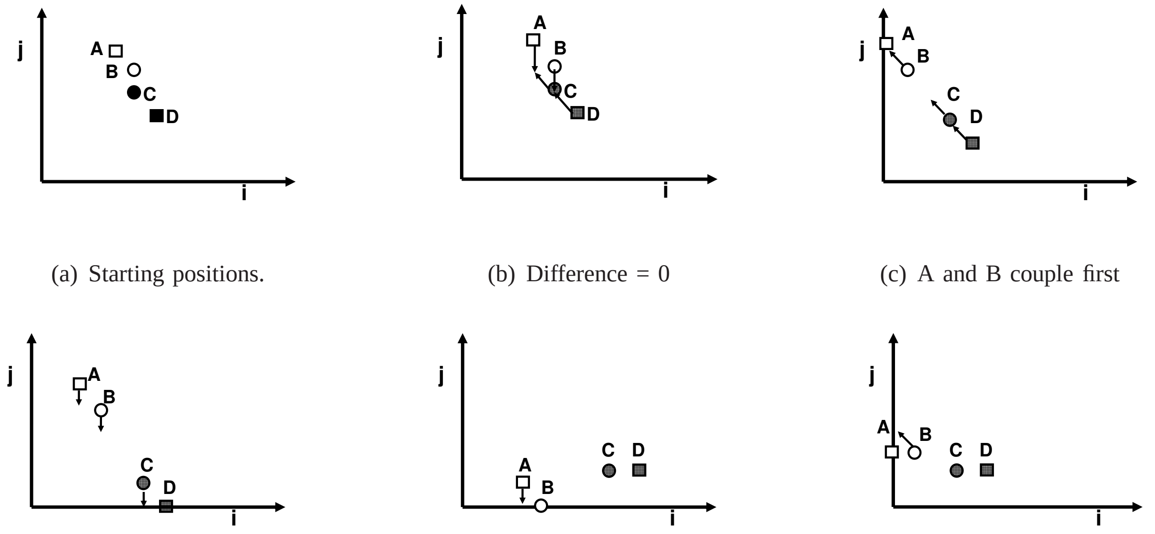
$$\psi v(i, j, 1, 1, 1) - \alpha_1 v(i, j, 0, 1, 1) - \alpha_2 v(i, j, 1, 0, 1) - \alpha_r v(i, j, 1, 1, 0) = (T_{111}v)(i, j). \quad \text{(IV.1)}$$

Repeating the same procedure for all $(m, n, r) \in \{0, 1\}^3$ the DP equations for a given (i, j) can be written in matrix form $\mathbf{AV} = \mathbf{T}$. Here $\mathbf{A}$ is an 8×8 matrix with, e.g., $A_{11} = \psi$; $A_{12} = -\alpha_r$; $A_{13} - \alpha_2$; $A_{14} = A_{16} = A_{17} = A_{18} = 0$; $A_{15} = -\alpha_1$. Let $\mathbf{Q} \equiv [\mathbf{I} - \mathbf{A}]$. Since $\mathbf{Q} \geq 0$ and some of the rows of $\mathbf{Q}$ do not sum to 1, it can be interpreted as the transitions between transient states of a discrete-time Markov chain. That is to say, $\mathbf{I} - \mathbf{QA}$ is invertible. Moreover, all elements of $(\mathbf{I} - \mathbf{Q})^{-1}$ are non-negative. Thus, $\mathbf{VA}^{-1}\mathbf{T}$ and each element of $\mathbf{V}$ can be written as a nonnegative linear combination of $(T_{mnr}v)(i, j)$ for $(m, n, r) \in \{0, 1\}^3$. Since (III.2) holds for $(T_{mnr}v)(i, j)$ it also holds for $v(i, j, m, n, r)$, i.e., an optimal transition monotone policy exists with reliability considerations.

Now we extend to the case with multiple dedicated and reconfigurable servers. Suppose again that there are N_r reconfigurable machines and N_1 and N_2 dedicated machines at station 1 and 2, respectively. Define $\mathbf{A}$ to be the ($2^{N_1+N_2+N_r} \times 2^{N_1+N_2+N_r}$) matrix that encodes the probabilities (times -1) of next entering each server status along with the diagonal that holds the sum of all possible probabilities out of each state. Let $\mathbf{V}$ be the ($2^{N_1+N_2+N_r} \times 1$) vector that represents the value functions for all possible states of the servers (for fixed i and j). Finally let $\mathbf{T}$ be the ($2^{N_1+N_2+N_r} \times 1$) vector that encodes the holding costs plus the probabilities of type of service multiplied by the appropriate value function. Similar to the previous argument, the components of $\mathbf{T}$ satisfy (III.2). Again, some of the rows of $\mathbf{Q} := \mathbf{I} - \mathbf{A}$ do not sum to 1 and $\mathbf{A^{-1}} \geq 0$ exists. Hence, the components of $\mathbf{V} = \mathbf{A^{-1}T}$ also satisfy (III.2) and the proof of the first assertion of Theorem 2.2 is complete.

It remains to show that when $h_1 \geq h_2$ the switching curve defined by an optimal transition monotone policy has slope ≥ -1. We cannot use the approach above because Lemma 3.1 does not hold for $v(i, j, m, n, r)$. Instead, we show that this holds by showing that (III.7) holds for all $i \geq 1$ and $j \geq 2$. Consider 4 processes, A, B, C, and D, defined on the same probability space,and starting in states $(i-1, j+1, m, n, r)$, (i, j, m, n, r), $(i, j-1, m, n, r)$, and $(i+2, j-2, m, n, r)$, respectively (see Figure 3(a)). Notice that as long as the relative queuelength positions of A to B and C to D are the same, the difference in the holding cost rates is $h_2 - h_1 - (h_2 - h_1)0$. Thus, suppose process B (C) uses the policy ϕ_1 (ϕ_2) that mimics the policy followed by A (D) with the caveat that it does not idle. Since the reconfigurable resources serve at the same rate for either station, the services seen by that resource are seen by all of the processes (as long as the queuelengths are positive). Through this fact (and Theorem 2.2, part 1) we note that the relative position of processes A and B are always to the left of processes C and D.

There are several cases to consider. First suppose that the relative positions of all processes have remained in tact, but that Process A serves at station 2 while Process D serves at station 1. Eventually, Processes A and C will couple as will Processes B and D; see Figure 3(b) and the cost rate difference thereafter is zero. Suppose now that Process A empties the first station followed by a service at station 1 in Process B. Since station 1 is empty for Process A, the two processes couple leaving processes C and D in their same relative positions; see 3(c). Since the difference in Processes C and D is of the form $v(i', j', m', n', r') - v(i'-1, j'+1, m', n', r')$, the third result of Proposition 2.3 yields that the difference is non-negative.



(a) Starting positions.

(b) Difference = 0

(c) A and B couple first

(d) C empties station 2

(e) A and B empty station 2

(f) A and B empty station 1

Fig. 3. Possibilities for Coupling.

If Process D empties station 2 followed by a service at station 2 by Process C before any of the previously described events occur, the relative position of the processes changes as depicted in Figure 3(d). However, the difference in costs becomes $h_1 - h_1 + h_2 = h_2 \geq 0$. Note it is still possible that Processes A and B couple leaving Processes C and D in their (new) relative positions, but the difference is now of the form $-v(i', j', m', n', r') + v(i', j'+1, m', n', r') \geq 0$.

After the relative position of C and D changes there are two more cases to consider. In the first, Processes A and B empty station 2, but Processes C and D may not have station 2 empty; see Figure 3(e). The difference due to the relative positions of each Process becomes $h_2 - h_2 = 0$. In this case, Processes A and B will clear the first queue before C and D. Process A will empty station 1 followed (perhaps after several services at station 2 seen by both processes) by a service in station 1 by Process B (see Figure 3(f)). At this point the relative difference in the holding costs is $-h_2 + h_1 \geq 0$. It then remains to wait until Process B is empty at both stations (so A is also empty) or Process C and D empty station 1. In the prior subcase the remaining difference in the costs between Processes C and D is of the form $-v(i', j', m', n', r') + v(i'+1, j', m', n', r') \geq 0$. In the latter $v(0, j', m', n', r') - v(0, j'+1, m', n', r') - v(0, j'+k', m', n', r') + v(0, j'+k'+1, m', n', r') \geq 0$, for some $k' \geq 0$. The last inequality follows from the convexity of $v(0, j, m, n, r)$ in the second assertion of Proposition 2.3.

Since in each case, the cost difference between processes is non-negative, (III.7) holds for the whole state space (not just when it is optimal to serve at station 1) and the result follows.

V. Conclusion

This paper investigates the optimal allocation of reconfigurable resources in a tandem queueing system. The optimality of transition monotone policies defines an optimal switching curve while the lower bound on the slope of this curve further reduces the computation of optimal policies. With respect to transition monotone policies, our findings indicate that with or without reliability considerations it will tend to be optimal to use the additional resource at the downstream station when the queue length is increased at that workstation. There is significant potential for further research in this area. We have a strong belief that the optimal switching curve is actually monotone. Indeed in over 30000 cases with randomly chosen parameters in **every** case the monotonicity held. Moreover, it is unknown whether these properties still hold in a system with external arrivals or setup costs/time. Again, our numerical studies suggest that the results extend. We would also like to see if the ideas here can be used to develop heuristics for problems with several stations.

Acknowledgment

This research was supported in part by the NSF grant DMI-0132811 and by the NSF Engineering Research Center for Reconfigurable Manufacturing Systems part of the Engineering Research Centers Program of the National Science Foundation under NSF Award Number EEC 95-29125. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of NSF.

References

[1] J. DeGaspari, “All in the family: Flexible machining systems give manufacturers a hedge on their bets,” *Mechanical Engineering*, vol. 124, no. 2, pp. 56–58, February 2002.

[2] W. Narongwanich, I. Duenyas, and J. Birge, “Optimal portfolio of reconfigurable and dedicated capacity under uncertainty,” 2003, preprint.

[3] H.-S. Ahn, I. Duenyas, and R. Zhang, "Optimal stochastic scheduling of a two-stage tandem queue with parallel servers," *Advances in Applied Probability*, vol. 31, pp. 1095–1117, 1999.

[4] B. Hajek, "Optimal control of two interacting service stations," *IEEE Transactions on Automatic Control*, vol. AC-29, no. 6, pp. 491–499, 1984.

[5] T. Farrar, "Optimal use of an extra server in a two station tandem queueing network," *IEEE Transactions on Automatic Control*, vol. 38, pp. 1296–1299, August 1993.

[6] M. Veatch and L. Wein, "Monotone control of queueing networks," *Queueing Systems*, vol. 12, pp. 391–408, 1992.

[7] H.-S. Ahn, I. Duenyas, and M. E. Lewis, "The optimal control of a two-stage tandem queueing system with flexible servers," *Probability in the Engineering and Informational Sciences*, vol. 16, no. 4, pp. 453–469, 2002.

[8] I. Duenyas, D. Gupta, and T. Olsen, "Control of a single server tandem queueing system with setups," *Operations Research*, vol. 46, no. 2, pp. 218–230, March-April 1998.

[9] S. Iravani, M. Posner, and J. Buzacott, "Two-stage tandem queue attended by a moving server with holding and switching costs; static and semi-dynamic policy," *Queueing Systems*, vol. 26, no. 3-4, pp. 203–228, 1997.

[10] T. Javidi, N. Song, and D. Teneketzis, "Expected makespan minimization of identical machines in two interconnected queues," *Probability In The Engineering And Informational Sciences*, vol. 15, no. 4, pp. 409–443, 2001.

[11] S. Andradottir, H. Ayhan, and D. G. Down, "Dynamic server allocation for queueing networks with flexible servers," *Operations Research*, vol. 51, no. 6, pp. 952–968, 2003.

[12] ——, "Server assignment policies for maximizing the steady-state throghput of finite queueing sytems," *Management Science*, vol. 47, pp. 1421–1439, 2001.

[13] S. Andradottir and H. Ayhan, "Throughput maximization for tandem lines with two stations and flexible servers," preprint.

[14] H.-S. Ahn and R. Righter, "Dynamic load balancing with flexible workers," 2004, preprint.