PERGAMON

Computers & Industrial Engineering 37 (1999) 691–710

computers & industrial engineering

www.elsevier.com/locate/dsw

Optimal buffer allocation in short μ-balanced unreliable production lines

H.T. Papadopoulos*, M.I. Vidalis

Department of Mathematics, University of the Aegean, GR-832 00 Karlovassi, Samos, Greece

Accepted 7 December 1999

Abstract

In this work, we investigate the optimal buffer allocation in short μ-balanced production lines consisting of machines that are subject to breakdown. Repair times and times to failure are assumed exponential, whereas service times are allowed to follow the Erlang-k distribution (with $k = 1$, 2, 4 and 8). By an improved enumeration procedure and applying the evaluative algorithm of Heavey et al. (European Journal of Operational Research 1993;68:69–89) for the calculation of throughput, we have examined in a systematic way several systems with 3, 4, 5, 6 and 7 stations and with a different total number of buffer slots. We have been able to give answers to some critical questions. These include the effect of the distribution of the service and repair times, the availability of the stations and the repair rates on the optimal buffer allocation and the throughput of the lines.

Keywords: Stochastic modeling of production lines; Buffer allocation problem

1. Introduction and literature review

Over the years a large amount of research has been devoted to the analysis of production lines. Much of this research has concerned the design of these manufacturing systems when there is considerable inherent variability in the processing times at the various stations, a common situation with human operators/assemblers.

* Corresponding author.
E-mail address: hpap@aegean.gr (H.T. Papadopoulos).

0360-8352/99/$ - see front matter
PII: S0360-8352(00)00004-8

One of the key questions that the designers face in a serial production line is the buffer allocation problem, i.e., how much buffer storage to allow and where to place it within the line. This is an important question because buffers can have a great impact on the efficiency of the production line. They compensate for the blocking and the starving of the line's stations. Unfortunately, buffer storage is expensive both due to its direct cost and due to the increase of the work-in-process (WIP) inventories. Also, the requirement to limit the buffer storage can be a result of space limitations in the shop floor.

The literature on production lines is vast, allowing us to review only the most directly relevant studies here. For a systematic classification of the relevant works on the stochastic modelling of these and other types of manufacturing systems (e.g., transfer lines, flexible manufacturing systems (FMS) and flexible assembly systems (FAS)), the interested reader is addressed to a review paper by Papadopoulos and Heavey [14] and some recently published books, such as Ref. [15], Buzacott and Shanthikumar [3], Gershwin [8], Viswanadham and Narahari [19] and Altiok [1]. In Ref. [15], both evaluative and optimization models are given for modelling the various types of manufacturing systems. The former are concerned with the evaluation of the various performance measures of the systems (see, for example, [21,22]), whereas the latter try to optimize these measures by determining the optimal values of the decision variables involved. This work falls into the second category. More specifically, for a given number of buffer slots available in a certain K-station production line with $K-1$ intermediate buffers, we are trying to find the optimal values of the buffer capacities that maximize the throughput of the line. The literature on this problem again is vast. We are not going to give all the relevant material here. Instead, the reader is referred to [15] and to the few papers that we mention, below. Many other references may be found therein.

There are basically two main classes of models concerning the buffer allocation problem of production lines: (a) for balanced lines and (b) for unbalanced lines. (Un)balanced lines are further classified as (i) μ-(un)balanced, (ii) CV-(un)balanced, and (iii) (fully) perfectly (un)balanced lines. The definition of these lines is self explanatory. For example, a CV-unbalanced line is a line with unequal coefficients of variation of the service time distribution and so forth. The majority of the research works on the buffer allocation problem in production lines have been devoted to reliable balanced lines. This review concentrates on studies which develop rules of thumb for the use of buffers in unreliable production lines.

We begin with an important paper by Conway et al. [6] which reviews and extends much of the previous literature. Conway et al. report a number of useful generalizations about the effect of buffers on production lines that are: (a) balanced and unbuffered, (b) balanced and buffered, (c) unbalanced and (d) unreliable. Conway et al.'s method of investigation was experimental and they used computer models to simulate production lines. They obtained some useful results.

Hillier and So [10,11] have studied the effect of machine breakdowns, interstage storage and the coefficient of variation of service times, respectively, on the performance of production lines and the optimal allocation of buffers in balanced production lines.

Carnall and Wild [4] have examined the effect of buffer capacity, service time variability and the order of constant and variable workstations on the throughput and the average idle time.

Altiok and Stidham [2] have dealt with the optimal allocation of buffer capacities so that the average profit in the long run is maximized.

Seong, Chang and Hong [17] developed heuristic algorithms for the buffer allocation in production lines with unreliable machines. Powell [16] examined the effect of service time variability on the optimal buffer allocation while So [18] studied the characteristics of the optimal buffer allocation in order to minimize the average WIP.

In this study, we examine the optimal buffer allocation in short μ-balanced unreliable production lines. We have been able to give answers to some critical problems. These include the effect of the distribution of the service and repair times, the availability of the stations and the repair rates on the optimal buffer allocation and the throughput of these types of lines. This is the main contribution of the present work. Our investigation also confirms that the reversibility property is valid to the unreliable lines, however this does not always help in reducing the search (buffer allocation) space for calculating the optimal buffer allocation. The latter is applicable only to the case of fully unreliable and fully balanced lines.

The remainder of the paper is organized as follows. Section 2 describes the production line model and the methodology of our investigation. Section 3 gives the numerical experimentation. Finally, Section 4 gives the findings of our study and some future research directions. The Appendix gives the derivation of some formulas for the expected value, the variance and the coefficient of variation of the effective service (or service completion) time.

2. The model and methodology of investigation

An asynchronous line (or flow line or production line) is one in which each workstation can pass parts on when its processing is complete as long as a buffer space is available (or if no buffer exists when the downstream workstation is idle). This type of line is subject to manufacturing blocking and starving. Material is not 'pulled' by demand, instead, models are operated in a 'push' mode, i.e., it is assumed that a part is always available when needed at the first workstation and space is always available after the last workstation to dispose of a complete part. A critical design choice is the selection of system throughput (mean production rate) as the primary measure of performance. Throughput is the most often used measure of efficiency in the literature, although alternatives are occasionally used, including the average work-in-process (WIP) and the average sojourn time.

Other typical assumptions of our model are: all the random variables (processing or service times, uptimes, downtimes) are independent random variables. The transfer through the buffers takes zero time. The failures are single-machine operation dependent failures, as opposed to total line failures (see [7,15] among others). When a failure occurs, the part stays on the machine and it is reworked when the machine is up again, that is, there is no scrapping of parts and the work resumes exactly at the point it stops. Times to failure and repair times are assumed to be exponentially distributed, whereas service times are allowed to follow the Erlang-k distribution (with $k \geq 1$, in general).

Next, the basic quantities of our model are defined.

S_i	a random variable representing service (processing) time of station i.
$\overline{S_i}$ or $E[S_i]$	average service (processing) time of station i.

μ_i	average service (processing) rate or speed of station i.
G_i	a random variable representing the time to failure of station i.
$MTTE_i$	average time to failure of station i.
β_i	average failure rate of station i.
R_i	a random variable representing the repair time of station i.
$MTTR_i$	average time to repair of station i.
r_i	average repair rate of station i.
$B = (B_2, B_3, \ldots, B_K)$	is the 'buffer vector', i.e., a vector with elements equal to the buffer capacities of the $K-1$ intermediate buffers. In an analogous way, we define $\mu = (\mu_1, \mu_2, \ldots, \mu_K)$ the mean service rates vector, or for simplicity, the 'rate vector', with elements equal to the mean service rates of the K stations of the line.
A_i	availability of station i. A_i is the average fraction of the time that station i is operational if it is operated in isolation, i.e., never starved or blocked. This quantity is also referred to as the isolated efficiency of station i and is also denoted by e_i. For a reliable station $A_i = 1$ and for an unreliable station $A_i < 1$. In our model, we assume that all the unreliable stations, denoted by U_i, have the same availability $A_i = A$.
T_i	a random variable representing the effective service (or service completion) time of station i. T_i equals the sum of the service time and the repair times. (The number of failures during one service period is a random variable following the *geometric* distribution, see the Appendix.)
$\overline{T_i}$ or $E[T_i]$	average effective service (or service completion) time of station i.
ρ_i	average effective service (or service completion) rate or production rate of station i in isolation.
X_K	the throughput or mean production rate of the K-station line.

The following relationships hold (see [7]):

$$\mu_i = \frac{1}{\overline{S_i}} = \frac{1}{E[S_i]}, \tag{1}$$

$$\beta_i = \frac{1}{MTTF_i}, \tag{2}$$

$$r_i = \frac{1}{MTTR_i}, \tag{3}$$

$$A_i = e_i = \frac{MTTF_i}{MTTF_i + MTTR_i} = \frac{r_i}{\beta_i + r_i} = \frac{r_i/\beta_i}{1 + r_i/\beta_i}, \tag{4}$$

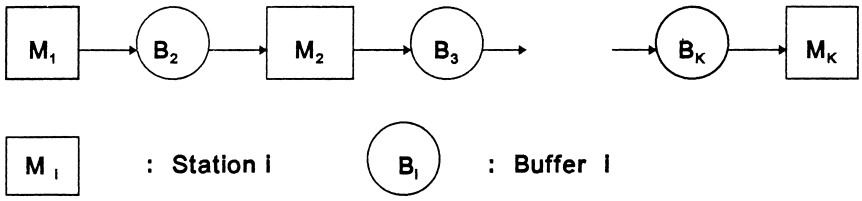

Fig. 1. A K-station unreliable production line with $K-1$ intermediate buffers.

$$\rho_i = \mu_i A_i = \frac{\mu_i r_i}{\beta_i + r_i}. \tag{5}$$

The K-station line has $K-1$ locations for buffers, labelled $B_2, B_3, \ldots, B_K$, in Fig. 1.

The object of our study is the buffering of μ-balanced asynchronous production lines with breakdowns, i.e., some or all the workstations are subject to failures with all the unreliable stations having the same availability.

2.1. The optimal buffer allocation (OBA) problem

In mathematical terms, our problem could be stated as follows:

Find $B_N=(B_2^N, B_3^N, \ldots, B_K^N)$ so as to
max $X_K(B)$
subject to $\sum_{i=2}^{K} B_i^N = N$, $B_i^N \geq 0$ B_i^N integer $(i = 2, 3, \ldots, K)$,

N is a fixed nonnegative integer denoting the total buffer space available in the system, which has to be allocated among the $K-1$ buffer locations so as to maximize the throughout of the production line.

In the unreliable case, X_K, the throughput of the K-station line is a function of the moments of the effective service (or service completion) time distribution and the buffer capacities. As higher moments have a minor effect on the throughput of the production lines, only the mean effective service (or service completion) rates are utilized.

Table 1
Production lines analysed in this study

K	Exponential	Erlang-2	Erlang-4	Erlang-8
	Range of N	Range of N	Range of N	Range of N
3	1–20	1–20	1–16	1–12
4	1–15	1–15	1–10	
5	1–16	1–10		
6	1–15			
7	1–16			

We have analysed systematically a number of short production lines consisting of $K = 3$, 4, up to 7 stations for the μ-balanced case and different values of N (see Table 1). We could not analyze much longer lines due to the limitations imposed by the evaluative algorithm of Heavey et al. [9]. Increasing either N or K, the number of iterations needed to solve the systems increase greatly. However, we believe this sample of production lines was enough to establish a pattern for the optimal buffer allocation (OBA) problem.

In order to determine the optimal allocation of a given total number of N buffer slots, an enumeration approach was primarily used to evaluate the throughput of the systems. This was done in order to find the maximum throughput and therefore to determine the optimal buffer allocation, for the various system configurations. However, the number of iterations we had to run to obtain the OBA of the systems were huge. The number of distinct ways for allocating N buffer slots in $K-1$ buffer locations in a K-station line is equal to

$$\binom{N+K-2}{K-2}. \tag{6}$$

Note, that although the reversibility property as mentioned by Hillier and So [11] is valid here too (see Section 3), it does not help in reducing the search (buffer allocation) space. To find the optimal buffer allocation, we are obliged to search all the possible allocations given by Eq. (6). For the calculation of the throughput of the production lines considered in this study, we used the Markovian state method and the software developed by Heavey et al. [9]. This gives the exact throughput of short K-station production lines with exponential times to failure and Erlang-$k(\ell)$ (for any k, $\ell \geq 1$ and in general $k \neq \ell$) service times and repair times, respectively. By the Markovian method, throughput evaluation involves formulation of the underlying queueing process as a finite state, continuous time Markov chain and then using an appropriate numerical solution procedure (the successive over-relaxation (SOR) method) to solve the resulting system of linear equations and obtain the stationary distribution of the Markov chain. Throughput then is easily calculated by summing the probabilities of those states in which the last station is busy. Unfortunately, the number of states of the Markov chain and thus the number of equations to be solved grows very rapidly with the number of stations K and the total buffer capacity N. For many cases examined in this study, this number is in the thousands ranging upwards to well over 200,000 (in many cases 300,000–500,000). In [15,9], a formula is provided which calculates the number of states. This rapid growth imposes limits on the size of problem that are computationally tractable.

At a second stage, after having applied full enumeration for some systems, we exploited the knowledge acquired from the investigation of these systems and we succeeded in reducing considerably the number of iterations needed to find the OBA, as may be seen in Table 2. To do so, we used the experimental observation that the absolute difference of the respective elements of the OBA vectors with N and $N+1$ buffer slots is less than or equal to 1, i.e., it holds:

$$|B_i^{N+1} - B_i^N| \leq 1, \quad \forall_i: 2 \leq i \leq K.$$

In this way, we have been able to derive the OBA, by induction, for any number N of buffer

Table 2
Number of iterations needed for the OBA

N	$K = 4$		$K = 5$		$K = 6$	
	Full enumeration	Improved enumeration	Full enumeration	Improved enumeration	Full enumeration	Improved enumeration
1	3	3	4	4	5	5
2	6	9	10	14	15	20
3	10	16	20	30	35	51
4	15	22	35	49	70	101
5	21	28	56	65	126	167
6	28	34	84	81	210	243
7	36	40	120	97	330	319
8	45	46	165	113	495	395
9	55	52	220	129	715	471
10	66	58	286	145	1001	547
11	78	64	364	161	1365	623
12	91	70	455	177	1820	699
13	105	76	560	193	2380	775
14	120	82	680	209	3060	851
15	136	88	816	225	3876	927
16	153	94	969	241	4845	1003
17	171	100	1140	257	5985	1079
18	190	106	1330	273	7315	1155
19	210	112	1540	289	8855	1231
20	231	118	1771	305	10,626	1307

slots that are to be allocated among the $K-1$ buffer locations of the line. The reduction works as follows. When N^* and K are given, one needs to determine all the OBA vectors for $N=1,2,\ldots,N^*$ and then for $N=N^*+1$, by searching only the values of $B_i^N=1$, B_i^N and B_i^N+1. Furthermore, this reduction starts after a number of total buffer slots, N ($N=9$, 6, 7 for $K=4$, 5, 6, respectively, as may be seen from Table 2). To quantify the reduction, by applying the improved enumeration, it has been experimentally observed that the number of iterations were reduced by at least 60%. For example, for $K=5$ stations and $N=12$ slots (see Table 2), the reduction is $[(455-177)/455]\times 100=61.1\%$

3. Numerical experimentation

In this section, we give the questions we tried to answer in this study and the numerical experimentation.

The search method, as explained in the previous section, is the improved enumeration procedure. In some cases, the reversibility property helps reducing the number of buffer allocations to determine the optimal solution. Hillier and So [11] applied that property in the case of reliable balanced production lines and they reduced the buffer allocation space by 50%. In the unreliable case, the reversibility property holds but it does not always contribute to the reduction of the buffer allocation space. This is explained below.

The following notation and definition are used:

Notation 1 By convention, we denote a K-station line with m ($m\le K$) unreliable stations as a vector where we indicate only the unreliable stations by U_i, $i=1,2,\ldots,m$. All the remaining $K-m$ stations that are not indicated in the vector are assumed to be reliable. For example, when we write (U_2, U_5, U_6) for a six-station line and $m=3$ unreliable stations, we denote a line with stations 2, 5 and 6 being unreliable and the remaining three stations (1, 3 and 4) being reliable.

Definition 1 Given a μ-balanced unreliable line denoted by the vector $(U_i,\ldots,U_j,\ldots,U_\ell)$ $(1\le i,j,\ell\le K)$ (this will be called the *original line*), with the assumptions given in the previous section, we define as its symmetric line, the one that is denoted by the vector $(U_{K+1-\ell},\ldots,U_{K+1-j},\ldots,U_{K+1-i})$, e.g., the symmetric line of the six-station (U_2, U_5, U_6) line, is the line (U_1, U_2, U_5). In other words, in the original six-station line, stations 2, 5 and 6 are unreliable whereas in the symmetric line stations 1, 2 and 5 are unreliable.

The *reversibility property* (see for its definition [13,20], among others) is applicable to the unreliable balanced case ($\mu_i=\mu$, $r_i=r$, $\beta_i=\beta$, for all $i=1,\ldots,K$) as follows. The (original) line

$$(\mu_1, r_1, \beta_1, B_2, \mu_2, \beta_2, B_3,\ldots,B_K, \mu_K, \mu_K, r_K, \beta_K)=(U_1, B_2, U_2, B_3,\ldots,B_K, U_K)$$

for a given N (number of buffer slots optimally allocated) has the same throughput as the line

$$(\mu_K, r_K, \beta_K, B_K, \mu_{K-1}, r_{K-1}, B_{K-1},\ldots,B_2, \mu_1, r_1, \beta_1)=(U_K, B_K, U_{K-1}, B_{K-1},\ldots,B_2, U_1)$$

As a result of this event, looking for the optimal location of the m unreliable stations among the K stations of the line, the search space is reduced by almost half and definitely the number of searches is less than or equal to

$$\left[\frac{\binom{K}{m}}{2}\right] + 2. \tag{7}$$

However, as far as the OBA is concerned, the reversibility property does not always reduce the allocation space which is given by Eq. (6). This is valid only when all K stations of the line are unreliable with the same parameters, i.e., in the case of a *fully unreliable and fully balanced line*. In that case, the search space is reduced by half. If the number of the unreliable stations is less than K, this is not valid. Then, we just get information about the symmetric line which is of no interest to the buffer allocation problem, as the object is the optimal buffering of the original line.

Example. Consider a five-station exponential line with $m = 3$ unreliable stations, the U_1, U_2, U_5 ($\mu_i = 1$, $r_i = 0.10$, $\beta_i = 0.05$, $A_i = 0.67$). The original line is $(U_1, U_2, \ldots, U_5)$ and its symmetric line is $(U_{5+1-5}, U_{5+1-2}, \ldots, U_{5+1-1}) = (U_1, U_4, \ldots, U_5)$. Both these two lines give the same throughput and thus for the optimal placement of the three unreliable stations within the line we have two options. By finding the first we don't need to search for the other. This would be its symmetric placement. As far as the OBA is concerned in the original line, the whole buffer allocation space has to be searched.

In this study, we tried to give answers to the following three questions:

1. The effect of the distribution of the service times on the throughput and the OBA.
2. The effect of the availability of the (unreliable) stations on the throughput and the OBA.
3. The effect of the repair rate on the throughput and the OBA, when the availability of the (unreliable) stations is maintained constant.

Below, we give some representative numerical results supporting our findings for the three problems. We split them in three subsections, one per each problem. In all cases examined, both the repair times and the times to failure follow the exponential distribution with mean rates r and β, respectively. Service times are allowed to follow the Erlang-k distribution with $k = 1, 2, 4, 8$, in general. The observations are summarized at the next section. More numerical results confirming the observations are available from the authors. Due to space considerations these are not given here.

3.1. Problem 1 (P1)

To answer the first question, i.e., the effect of the distribution of the service time (or indirectly of the service completion time) on the throughput and the OBA, we studied systems with $K = 3$, 4 and 5 stations. For each one of these cases we further considered systematically various sub-cases taking 1 out of K, 2 out of K and so on to K out of K unreliable stations. More specifically, we examined three-, four- and five-station lines with the service times

Table 3
OBA and throughput of three-station lines with $m = 1$

U_1 unrel.		Exponential		Erlang-2		Erlang-4		Erlang-8	
N	r/β	OBA	X_3	OBA	X_3	OBA	X_3	OBA	X_3
12	1	(9–3)	0.4879	(9–3)	0.4971	(10–2)	0.4996	(10–2)	0.5000
	2	(8–4)	0.6271	(8–4)	0.6523	(9–3)	0.6633	(9–3)	0.6662
	5	(7–5)	0.7466	(7–5)	0.7913	(8–4)	0.8172	(8–4)	0.8288
	10	(7–5)	0.7945	(7–5)	0.8482	(7–5)	0.8809	(8–4)	0.8984
	20	(6–6)	0.8210	(6–6)	0.8787	(7–5)	0.9150	(7–5)	0.9356

following the Erlang-2 distribution. The Erlang-4 distribution was considered in the cases of three- and four-station lines, whereas the Erlang-8 distribution was examined only in three-station lines.

Tables 3–5 give the throughput and the optimal buffer allocation for different values of the parameters N, r/β, K and $m = 1$ (station 1 being the unreliable station, denoted by U_1). We allowed the ratio r/β to take the values 1, 2, 5, 10, 20 (equivalently, the availability of the unreliable stations takes on the values 0.50, 0.67, 0.83, 0.91, 0.95, respectively).

Fig. 2 gives the OBAs for $K = 4$, $m = 1$ (U_1 the unreliable station), $N = 12$ buffer slots and $A = 50\%$, for exponential, Erlang-2, Erlang-4 and Erlang-8 service times, respectively, which are (7–3–2), (8–2–2), (8–2–2) and (9–2–1).

Fig. 3 shows the effect of the service time distribution on the throughput of a three-station line with one unreliable station (the U_1). The availability of the unreliable station is assumed to be 83% and the $N = 10$ buffer slots allocated among the two buffers contribute, respectively, to the recovery of 87.67%, 93.64%, 97.24% and 99.11% of the maximum throughput of the exponential, Erlang-2, Erlang-4 and Erlang-8 lines.

Table 6 shows the effect of the service (or indirectly of the service completion) time distribution on the efficiency recovery of a three-station line with the following parameters: $m = 1$ (the first station being the unreliable), $A = 0.91$ ($r = 0.1$, $\beta = 0.01$). The extra 10 buffer

Table 4
OBA and throughput of four-station lines with $m = 1$

U_1 unrel.		Exponential		Erlang-2		Erlang-4	
N	r/β	OBA	X_4	OBA	X_4	OBA	X_4
12	1	(7–3–2)	0.4768	(8–2–2)	0.4931	(8–2–2)	0.4988
	2	(6–4–2)	0.5982	(7–3–2)	0.6376	(7–3–2)	0.6580
	5	(5–4–3)	0.6982	(5–4–3)	0.7603	(6–3–3)	0.8000
	10	(4–4–4)	0.7372	(5–4–3)	0.8091	(5–4–3)	0.8580
	20	(4–4–4)	0.7591	(4–4–4)	0.8362	(4–4–4)	0.8886

Table 5
OBA and throughput of five-station lines with $m = 1$

U_1 unrel.		Exponential		Erlang-2	
N	r/β	OBA	X_5	OBA	X_5
10	1	(5–2–2–1)	0.4559	(6–2–1–1)	0.4824
	2	(4–3–2–1)	0.5571	(5–2–2–1)	0.6093
	5	(3–3–2–2)	0.6391	(3–3–2–2)	0.7153
	10	(2–3–3–2)	0.6713	(3–3–2–2)	0.7573
	20	(2–3–3–2)	0.6894	(2–3–3–2)	0.7812

slots allocated among the two buffers of the line increase its efficiency by a percentage that decreases according to the number of the service phases.

In the same three-station line, we observed that when the service times follow the exponential distribution, the line needs 16 buffer slots to recover 90% of its maximum efficiency. To recover approximately the same percentage of its maximum throughput, the line needs eight, four and two buffer slots, respectively, when the service times follow the Erlang-2, Erlang-4 and Erlang-8 distribution. This may be seen in Fig. 4. This means that the following linear fit approximates well the relationship between the CVs of the service time distributions and the number of buffer slots needed to obtain a given throughput level:

$$\frac{CV^2_{E_\nu}}{CV^2_{E_{2\nu}}} = 2, \quad \frac{N_{E_\nu}}{N_{E_{2\nu}}} \doteq 2, \quad \nu = 1, 2, 4, \tag{8}$$

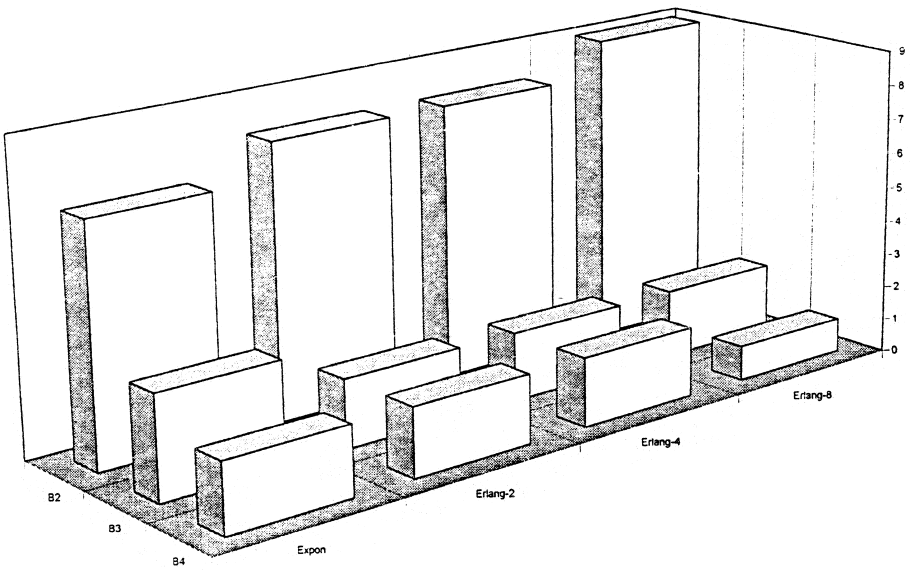

Fig. 2. The effect of the service time distribution on the OBA ($K = 4$, $m = 1$, $A = 0.50$, $N = 12$).

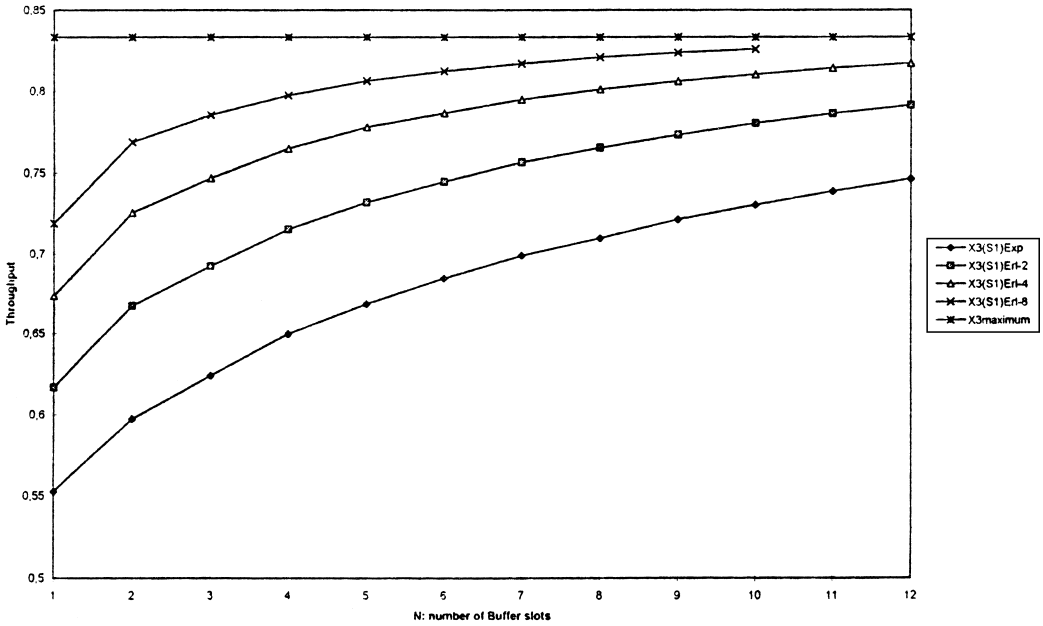

Fig. 3. The effect of the service time distribution on the throughput ($A = 0.83$, $N = 10$).

where, N_{E_ℓ} denotes the number of buffer slots required to obtain a given throughput level when the service times follow the Erlang-ℓ distribution.

3.2. Problem 2 (P2)

To answer the second question, i.e., the effect of the ratio (r/β or, indirectly the availability of the unreliable stations on the throughput and the OBA, we studied systems with $K = 3$, 4, 5, 6 and 7 stations. More specifically, we investigated systems with a given number of stations, K, a given number of unreliable stations, m, a specific location of the unreliable stations among all the possible $\binom{K}{m}$ locations and specific distribution of the service times. As in Problem 1, we

Table 6
The effect of service time distribution on the efficiency recovery

	Throughput and its percentage recovery (PER)			
	Exponential	Erlang-2	Erlang-4	Erlang-8
$N = 0$	0.5356	0.6036	0.6681	0.7244
$N = 10$	0.7775	0.8356	0.8730	0.8939
PER	53.6%	38.4%	30.7%	23.4%

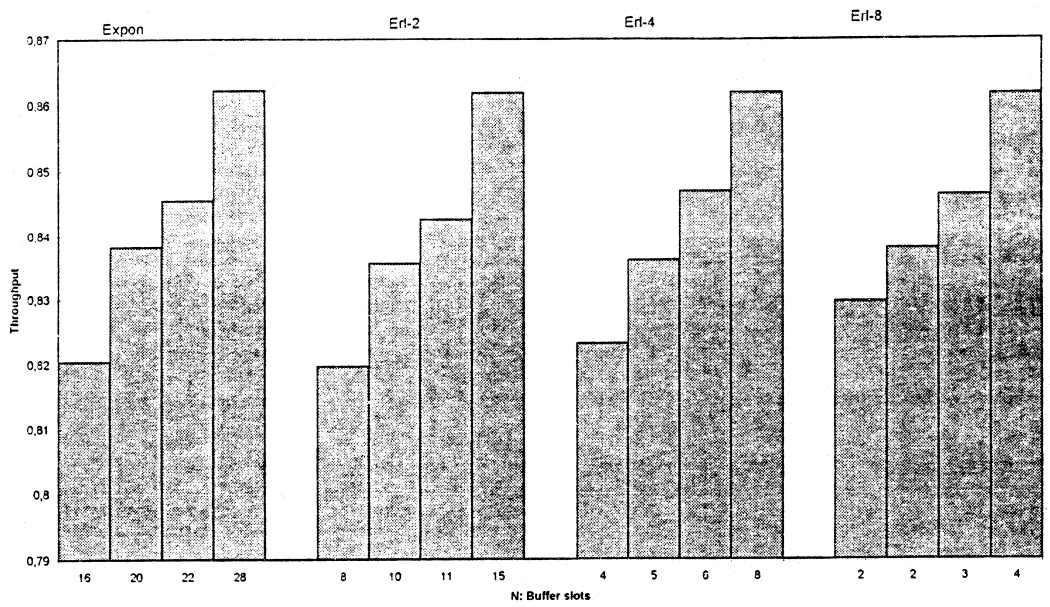

Fig. 4. Number of buffer slots needed to recover a certain percentage of the efficiency loss as a function of the number of service phases ($K = 3$, $m = 1$, $A = 0.91$).

allowed the ratio r/β to take the values 1, 2, 5, 10, 20 (equivalently, the availability of the unreliable stations takes on the values 0.50, 0.67, 0.83, 0.91, 0.95, respectively). Here, only some representative numerical results are given.

Tables 7–11 give the effect of the stations' availability on the optimal buffer allocation in a few lines with different number K of stations and m, unreliable stations. Moreover, Tables 7 and 8 provide also with the throughput of the respective lines.

3.3. Problem 3 (P3)

In this problem, the effect of the mean repair rate, r, on the throughput and the OBA with all the other parameters (K, m, A and N) kept constant is examined. More specifically, for a

Table 7
OBA and throughput of 4-station lines with $m = 2$

U_1, U_4 unrel.		Exponential		Erlang-4	
N	r/β	OBA	X_4	OBA	X_4
15	1	(6–3–6)	0.3930	(6–3–6)	0.4125
	2	(5–4–6)	0.5208	(6–3–6)	0.5613
	5	(5–5–5)	0.6573	(6–4–5)	0.7315
	10	(5–5–5)	0.7237	(5–5–5)	0.8187
	20	(5–5–5)	0.7630	(5–5–5)	0.8723

Table 8
OBA and throughput of five-station lines with $m = 2$

U_1, U_5 unrel.		Exponential		Erlang-2	
N	r/β	OBA	X_5	OBA	X_5
10	1	(3–2–2–3)	0.3637	(3–2–2–3)	0.3826
	2	(3–2–2–3)	0.4733	(3–2–2–3)	0.5096
	5	(2–3–3–2)	0.5857	(3–2–2–3)	0.6472
	10	(2–3–3–2)	0.6402	(2–3–3–2)	0.7163
	20	(2–3–3–2)	0.6722	(2–3–3–2)	0.7583

certain value of the ratio r/β, we considered four sub-cases where r takes on the values: 0.50, 0.20, 0.10, 0.05, 0.02, with a simultaneous equal decrease of β so as to keep the ratio r/β constant.

Table 9
OBA of six-station exponential lines with $m = 1$ and 3

N	r/β	OBA (U_1 unrel.)	OBA (U_1, U_2, U_6 unrel.)
15	1	(7–3–2–2–1)	(5–4–2–2–2)
	2	(5–3–3–2–2)	(4–4–2–2–3)
	5	(4–3–3–2–2)	(4–3–3–2–3)
	10	(3–3–3–3–3)	(3–3–3–3–3)
	20	(3–3–3–3–3)	(3–3–3–3–3)

Table 12 gives some numerical data to show the effect of the mean repair rate on the throughput of five-station exponential lines with different number of unreliable stations (m), $N = 16$ slots and $A = 91\%$.

Table 10
OBA of six- and seven-station exponential lines with $m = 2$

	$K = 6$		$K = 7$	
r/β	N	OBA (U_1, U_6 unrel.)	N	OBA (U_1, U_7 unrel.)
1	15	(4–3–2–2–4)	18	(5–2–2–2–2–5)
2		(4–3–2–2–4)		(4–3–2–2–3–4)
5		(3–3–3–3–3)		(3–3–3–3–3–3)
10		(3–3–3–3–3)		(3–3–3–3–3–3)
20		(3–3–3–3–3)		(3–3–3–3–3–3)

Table 11
OBA of six-station exponential lines with $m = 6$

N	r/β	OBA (all U_i unrel.)
15	1	(1–4–5–4–1)
	2	(2–3–4–4–2)
	5	(2–4–3–4–2)
	10	(3–3–3–3–3)
	20	(3–3–3–3–3)

Table 12
Throughput of five-station exponential lines with $m = 1,\ldots,5$, ($N = 16$, $A = 91\%$, $r/\beta = 10$)

r	X_5 (rel.)	X_5 (U_1)	X_5 (U_1, U_5)	X_5 (U_i, $i = 1, 2, 5$)	X_5 (U_i, $i = 1, 2, 4, 5$)	X_5 ($m = 5$)
0.50	0.762	0.7437	0.7242	0.6993	0.6791	0.6581
0.20	0.762	0.7348	0.7078	0.6771	0.6512	0.6258
0.10	0.762	0.7270	0.6936	0.6589	0.6290	0.6009
0.05	0.762	0.7205	0.6818	0.6443	0.6115	0.5814

Table 13
OBA of five-station expon. lines with a varying m ($A = 50\%$)

N	r	OBA (U_1)	OBA (U_1, U_5)	OBA ($m = 5$)
20	0.50	(13–3–2–2)	(8–2–2–8)	(4–6–6–4)
	0.20	(12–4–2–2)	(7–3–3–7)	(4–6–6–4)
	0.10	(11–4–3–2)	(7–3–3–7)	(3–7–7–3)
	0.05	(11–4–3–2)	(7–3–4–6)	(3–7–7–3)
	0.02	(8–5–4–3)	(6–4–4–6)	(3–7–7–3)

Table 14
OBA of four-station exponential lines with $m = 1,\ldots,4$ ($A = 50\%$)

N	r	OBA (U_1)	OBA (U_1, U_4)	OBA (U_i, $i = 1, 2, 4$)	OBA ($m = 4$)
24	0.50	(18–4–2)	(10–4–10)	(12–6–6)	(8–9–7)
	0.20	(17–4–3)	(10–4–10)	(12–6–6)	(8–9–7)
	0.10	(16–5–3)	(10–4–10)	(11–7–6)	(7–10–7)
	0.05	(15–5–4)	(9–5–10)	(11–7–6)	(6–11–7)
	0.02	(13–6–5)	(9–6–9)	(10–8–6)	(6–12–6)

Tables 13 and 14 give some numerical results concerning the optimal buffer allocation of five- and four-station exponential lines, respectively, with a varying number of unreliable stations, m and availability $A = 50\%$ (or equivalently, $r/\beta = 1$).

From Table 14, for $m = 2$ (two unreliable stations, U_1 and U_4) and $m = K$, respectively, the OBA forms the shape of a bowl and an inverse bowl, respectively, whereas, for $m = 1$ (U_1 being unreliable), the elements of the optimal buffer vector are arranged in descending order.

4. Findings of the study and further research

In this study, we examined the problem of the optimal buffer allocation in short μ-balanced unreliable production lines with all the unreliable stations having the same availabilities. These lines are not fully balanced, due to the random failures of 1 or $m < K$ stations.

The reversibility property holds in the case of these unreliable lines; however this property does not always help in reducing the buffer allocation space. The latter is applicable only to the case of fully unreliable and fully balanced lines. In this study, we succeeded in improving the enumeration process and we reduced considerably the search (allocation) space in order to determine the OBA of the μ-balanced unreliable lines. For the calculation of the throughput of the lines we used the evaluative algorithm of Heavey et al. [9]. This gives the exact throughput of short K-station unreliable lines with exponential times to failure and Erlang-k and Erlang-ℓ ($k \neq \ell$, in general) service and repair times, respectively. From the study of many systems we have been able to draw some conclusions that may help practitioners answering their buffer design questions. These are concerned with:

- the effect of the distribution of the service times on the throughput and the optimal buffer allocation;
- the effect of the availability of the (unreliable) stations on the throughput and the optimal buffer allocation and
- the effect of the repair rate on the throughput and the optimal buffer allocation when the availability of the (unreliable) stations is kept constant.

The findings derived from the numerical experimentation, given in the previous section, may be summarized as follows:

1. As far as the OBA is concerned, there are three cases. For small values of the availability of the unreliable stations:
   - (i) when $m < K$ and even, the OBA presents the shape of a bowl;
   - (ii) when $m < K$ and odd, the OBA presents the shape of a non-symmetric bowl and
   - (iii) when $m = K$, the OBA presents the shape of an inverse bowl. This observation violates the well-known result about the uniformity of the optimal buffer allocation prevailing in a balanced line.

In all three cases, as the (common) availability of the unreliable stations tends to unity (i.e., the stations become near perfectly reliable), at the optimal situation, all the buffers are evenly allocated the buffer slots.

2. As the number of service phases increase (from exponential to Erlang-k ($k > 1$) distribution) then
   - (i) the coefficient of variation of the effective service time, CV_T, decreases and this results in an increase in the throughput of the line;
   - (ii) it is more difficult to justify economically the presence of more buffer spaces. That is, for the same number, N, of buffer slots we get less percentage increase of throughput;
   - (iii) the OBA characteristics given in conclusion 1(i,ii,iii), above, become more pronounced and
   - (iv) A linear fit approximates well the relationship between the CVs of the service time distribution and the number of buffer slots needed to obtain a given throughput level.

3. For exponential or Erlang-k service times, an increase in the ratio r/β affects the availability of the stations. Also the coefficient of variation of the effective service time, CV_T decreases, leading the result to an increase in the line efficiency.
4. As the cycle time (service time plus repair time) increases, that is, as r and β decrease simultaneously by the same percentage, to keep the availability, A constant, the throughput of the line decreases. As far as the optimal buffer allocation is concerned, where bowl shape was formed, this tends to become smooth and finally to be eliminated as the cycle time increases. On the other hand, in the cases where the inverse bowl was the shape of the OBA, this shape becomes sharper as the cycle time increases. Finally, in the cases where the elements of the optimal buffer vector are arranged in a descending order, this characteristic becomes less apparent as the cycle time increases (i.e., the differences among the buffer allocations of subsequent buffers are reduced).

An interesting area for further research would be the development of an effective optimization algorithm that would give the OBA, in an accurate and fast way, for large production lines operating under more general assumptions. These include the totally unbalanced and unreliable lines, i.e., lines with different mean service rates and different availabilities of the unreliable stations. This would involve (i) the use of an effective decomposition algorithm (such as that developed by Choong and Gershwin [5]) to calculate the throughput of large unreliable lines very fast and (ii) the development of a fast and accurate search algorithm to determine the optimal buffer allocation.

Acknowledgements

The authors would like to express their sincere thanks to the anonymous referees. Their valuable and penetrating comments improved significantly the appearance of the paper.

Appendix. Derivation of $E[T]$, Var[T], CV_T

We give here the formulas for the expected value ($E[T]$), the variance (Var[T]) and the coefficient of variation (CV_T) of the effective service (or service completion) time (T).

The effective service time consists of the service time and the repair time. This, in general, is

a random sum of the service time and M repair times, if the respective station breaks down M times during the service completion period. In mathematical terms, we write:

$$T = S + \sum_{i=1}^{M} R_i, \tag{A1}$$

where, R_i denotes the repair time following the ith failure and S is the service time. The probability that exactly ℓ breakdowns occur before a service completion is given by (see [2])

$$p^{\ell} q, \quad \ell = 0, 1, 2, \ldots \tag{A2}$$

where p ($q = 1 - p$) is given by the following expressions depending on the distribution of the service, repair and failure times. This means that the number of breakdowns before a service completion is geometrically distributed.

When the service, repair and failure times follow the exponential distribution with mean rates μ, r and β, respectively, p is given by

$$p = \frac{\beta}{\beta + \mu} = 1 - \frac{\mu}{\beta + \mu}, \tag{A3}$$

whereas when the service times follow the Erlang-k distribution and both the repair times and the failure times are exponentially distributed with the same rates, as above, p is given by

$$p = 1 - \left(\frac{\mu}{\beta + \mu}\right)^{k}. \tag{A4}$$

The expected value of M is the expected value of the geometric distribution:

$$E[M] = \frac{p}{q}. \tag{A5}$$

$E[T]$, $\mathrm{Var}[T]$ and CV_T when S follows the Erlang-k and R, G follow the exponential distribution: it is assumed that the Erlang-k distribution of the service times has mean rate μ at each of its k phases and the exponential repair and failure times have mean rates, respectively, r and β at each station of the line. We derived the respective formulas both directly from the Laplace transform of the effective service time distribution (see also [1]) and via the analysis of the first passage time in Markov chains (see [12]).

$$E[T] = \frac{k(r + \beta)}{\mu r}, \tag{A6}$$

$$\mathrm{Var}[T] = \frac{k(r + \beta)^2 + 2k\beta\mu}{(\mu r)^2}, \tag{A7}$$

$$CV_T = \frac{1}{k} + \frac{2\beta\mu}{k(r + \beta)^2}. \tag{A8}$$

Special case: $E[T]$, $\mathrm{Var}[T]$ and CV_T when S, R, G are all exponentially distributed

$$E[T] = \frac{r+\beta}{\mu r}, \tag{A9}$$

$$\mathrm{Var}[T] = \frac{(r+\beta)^2 + 2\beta\mu}{(\mu r)^2}, \tag{A10}$$

$$CV_T = 1 + \frac{2\beta\mu}{(r+\beta)^2}. \tag{A11}$$

References

[1] Altiok T. Performance analysis of manufacturing systems. New York: Springer-Verlag, 1997.
[2] Altiok T, Stidham Jr S. The allocation of interstage buffer capacities in production lines. IIE Transactions 1983;15/4:292–9.
[3] Buzacott JA, Shanthikumar JG. Stochastic models of manufacturing systems. NJ: Prentice-Hall, 1993.
[4] Carnall CA, Wild R. The location of variable work stations and the performance of production flow lines. International Journal of Production Research 1976;14/6:703–10.
[5] Choong YF, Gershwin SB. A decomposition method for the appropriate evaluation of capacitated transfer lines with unreliable machines and random processing times. IIE Transactions 1987;19(2):150–9.
[6] Conway R, Maxwell W, McClain JO, Thomas LJ. The role of work in process inventory in serial production lines. Operations Research 1988;36/2:229–41.
[7] Dallery Y, Gershwin SB. Manufacturing flow line systems: a review of models and analytical results. Queueing Systems 1992;12:3–94.
[8] Gershwin SB. Manufacturing systems engineering. NJ: Prentice-Hall, 1994.
[9] Heavey C, Papadopoulos HT, Browne J. The throughput rate of multistation unreliable production lines. European Journal of Operational Research 1993;68:69–89.
[10] Hillier FS, So KC. The effect of machine breakdowns and interstage storage on the performance of production line systems. International Journal of Production Research 1991;29(10):2043–55.
[11] Hillier FS, So KC. The effect of the coefficient of variation of operation times on the allocation of storage space in production line systems. IIE Transactions 1991;23(2):198–206.
[12] Kemeny JG, Snell JL. Finite Markov chains, 2nd ed. New York: Springer-Verlag, 1976.
[13] Muth EJ. The reversibility property of production lines. Management Science 1979;25/2:152–8.
[14] Papadopoulos HT, Heavey C. Queueing theory in manufacturing systems analysis and design: a classification of models for production and transfer lines. European Journal of Operational Research 1996;92:1–27.
[15] Papadopoulos HT, Heavey C, Browne J. Queueing theory in manufacturing systems analysis and design. London: Chapman and Hall, 1993.
[16] Powell SG. Buffer allocation in unbalanced three-station serial lines. International Journal of Production Research 1994;32/9:2201–17.
[17] Seong D, Chang SY, Hong Y. Heuristic algorithms for buffer allocation in a production line with unreliable machines. International Journal of Production Research 1995;33/7:1989–2005.
[18] So KC. Optimal buffer allocation strategy for minimizing work-in-process inventory in unpaced production lines. IIE Transactions 1997;29(1):81–8.
[19] Viswanadham N, Narahari Y. Performance modeling of automated manufacturing systems. NJ: Prentice-Hall, 1992.
[20] Yamazaki G, Kawashima T, Sakasegawa H. Reversibility of tandem blocking systems. Management Science 1985;31/5:78–83.

[21] Yeralan S, Tan B. Analysis of multistation production systems with limited buffer capacity. Part I: the subsystem model. Mathematical and Computer Modelling 1997;25(7):109–22.
[22] Yeralan S, Tan B. Analysis of multistation production systems with limited buffer capacity. Part II: the decomposition method. Mathematical and Computer Modelling 1997;25(11):109–23.