A New Algorithm for Buffer Allocation in Production Lines

R. Levantesi (1), A. Matta (1), T. Tolio (1)

(1) Politecnico di Milano, Dipartimento di Meccanica
Via Bonardi, 9, 20133 Milano, Italy

raniero.levantesi@mecc.polimi.it
andrea.matta@mecc.polimi.it
tullio.tolio@mecc.polimi.it

Abstract

In the paper a new efficient algorithm for the allocation of storage capacity in serial production lines is derived. The proposed method aims to find out the distribution of storage capacity between machines that minimizes the total buffer space assigned to the line satisfying a target production requirement. The algorithm relies on an iterative scheme that, starting from the minimal capacity required in each buffer to meet the production requirement, proceeds increasing the capacity of buffers until the target production rate is reached. At each iteration the buffer corresponding to the largest component of the gradient of production rate with respect to each buffer capacity is selected and its capacity is increased by a small amount of space. The effect of such an increasing is then estimated by applying the decomposition technique to the resulting line. Preliminary numerical results shows that the buffers distribution provided by the algorithm is optimal or near optimal and that the convergence is rapidly reached.

Keywords: *Buffer Allocation, System Design, Performance Evaluation, Decomposition.*

1. Introduction

The paper presents an efficient method for the allocation of storage capacity in serial production lines. The proposed method can be applied to the configuration of buffer capacity in Flow Line Systems where machines are de-coupled by means of in-process inventories. Flow Line Systems are quite common in shop floors being adopted in environments characterized by large production volumes and long product life cycles. Their configuration is a well-established topic in literature due to both the economical and the strategical relevance of the investment required.

More specifically, the design phase of a transfer line involves a large number of variables that have to be defined. Indeed, in order to find out the final configuration of the line, decisions on the number of stations in the system, the efficiency of the stations and the inter-storage capacity of buffers have to be taken. Therefore, in this context, once the number of production stages has been identified and the machines characteristics have been selected (namely processing rates and reliability parameters, i.e. *Mean Time To Failures* and *Mean Time To Repair*), a tool to determine the buffer distribution that achieves the production requirement and, at the same time, minimizes the total buffer space is highly recommended. Indeed, the availability of the most suitable distribution of buffers with respect to a target throughput rate allows establishing a good compromise in the trade-off production rate/inventory, that is in the relationship between the need to increase the buffer storage to enhance the system performance and the need to reduce it to limit space and in-process inventory costs.

The objective of this paper is therefore to make a further step toward one of the major goal of the literature on analytical methods, that is to provide the system designers a set of tools to support them in the definition of the design variables.

The proposed method is able to find out the optimal or near optimal distribution of buffers in transfer lines with finite capacity buffers and unreliable stations relying on efficient approximated techniques for the performance evaluation of this type of systems [7]. In this paper the continuous model of homogeneous production lines has been considered, but the extension of the approach to the case of deterministic, exponential and non homogeneous lines appears rather straightforward.

The approach proposed in this paper can be referred as a gradient approach [3,4,8]. The algorithm relies on an iterative procedure that, starting from the minimal capacity required in each buffer to meet the production requirement, proceeds increasing the capacity of buffers until the target production rate is achieved. At each iteration, the buffer corresponding to the largest component of the gradient of production rate with respect to each buffer capacity is increased of a fixed small amount of space. The effect of such an increasing is then estimated by applying the decomposition technique on the resulting line.

Differently from the previous works based on the use of the gradient, in the single step of the proposed iterative method the global direction to follow in the solution space is not searched; on the contrary, the method tries to get closer to the optimal solution following one dimension of the space solution at time. The dimension is identified by the largest derivative value of the throughput with respect to each dimension.

2. Brief Literary Review

A large amount of literature has been developed in the last decades on the analysis of transfer lines or flow line systems, see [1] for a survey on this topic. However, this research is mainly devoted to the performance estimation of such systems by assuming known machine parameters and buffer capacities. Several papers deal with the optimal buffer allocation and with the assessment of qualitative properties of transfer lines, but very few of them aims to provide effective tools for the system design. In this area, two main research streams can be identified: the first one is based on the evaluation of system performance by means of simulation technique while the other one makes use of analytical methods to obtain the performance measures of the system. A very detailed and exhaustive review of these topics in literature can be found in the work of Gershwin and Goldis [3].

Different papers address the problem of the buffer space configuration between machines in serial production lines. In particular, Park [8] developed a two-phase tree search branch and bound method to solve the problem under examination for discrete model of transfer lines and similar problems closely related. Gershwin and Goldis [3] derived a gradient method to solve the problem on the basis of the intuition that the first order expansion of the production rate gives the possibility to formulate the problem in terms of integer linear problem. The results provided by the authors demonstrate that this method can lead to optimal or near optimal solutions. Gershwin and Schor [4] improved the Goldis algorithm in terms of accuracy and, mainly, of convergence speed: in particular, they develop an iterative technique that solves the problem under examination, referred as *primal*, on the basis of the solution obtained for the *dual* problem (i.e. the maximization of production rate given a total amount of buffer space available) by means of a gradient approach.

3. Description of the method

3.1. Problem Formulation

As pointed out in the introduction, the design phase of a transfer line involves different system variables to be defined such as machine speed, machine reliability and buffer capacities. In this paper we assume that the characteristics of machines have already been selected. Also the processing cycle of the product has already been defined and, as a consequence, the number of stages building-up the line. Parts always follow a linear path in the system starting from the first machine, where they undergo the first operation, to the last machine in which they complete their process cycle. Buffers are located between machines so that the behavior of machines is partially de-coupled by the rest of the line. In such a way, it is possible to adsorb the disruption of flow due to processing times and failure occurrences, thus reducing the interdependencies among machines and improving the system performance. Buffers can be neither infinitely large, given the high costs of the floor space and work in process inventories, nor too small to avoid unsatisfactory levels of production rate. Therefore, a certain level of capacity for each buffer has to be defined in the design phase of a transfer line by relating them to the desired performance of the line.

The problem we deal with in this paper is that of minimizing the total buffer capacities so that the production rate of the line is greater than or equal to a certain desired value. This problem, already known in literature as the *primal problem* [3], is well suited and motivated when the production rate of the system cannot be lower than a certain value and the floor space has a cost. In addition, the cost of work in process inventories does not change through the different stages of the system.

Let us indicate with K the number of machines of the flow line that has to be designed, with M_i (with $i=1,\ldots,K$) and B_i (with $i=1,\ldots,K-1$) respectively the machines and the buffers of the system. Machines are unreliable and can fail in different F_i (with $i=1,\ldots,K$) modes. Time to failures and time to repairs are exponentially distributed for each mode. We denote with $p_{i,j}$ (with $j=1,\ldots,F_i$) the failure rate of machine M_i going down in mode j and with $r_{i,j}$ the repair rate of machine M_i going up from failure mode j.

We assume that failures are *ODFs* (Operation Dependent Failures) and that machines cannot be down in more than one mode. Machine M_i process material provided that buffer B_{i-1} is not empty and buffer B_i is not full. The sizes of each buffer B_i, denoted with N_i, are the decision variables that have to be determined. In particular, we address the continuous model of homogeneous production lines; for a more detailed description of continuous material flow lines refer to [1,2,6,7].

The production rate of the system, denoted with P, is a function of machine parameters and of buffer capacities. Since it is assumed that machine parameters have already been defined, buffer capacities are the only variables that have to be chosen for reaching the target production rate. Thus, the primal problem can be formulated as follows:

$$\begin{aligned}&\min \quad N^T = \sum_{i=1}^{K-1} N_i \\ &s.t. \\ &P(N_1,\ldots,N_i,\ldots,N_{K-1}) \geq P^*\end{aligned} \tag{1}$$

where P^* represents the target production rate while N^T is the total amount of allocated buffer capacity.

Note that the above problem has not a unique optimal solution since it is possible to find more combinations of vectors $\mathbf{N}=(N_1,\ldots,N_K)$ satisfying the production rate constraint. The solution vectors differ in the distribution of the single buffer spaces even if they are characterized by the same total buffer capacity N^T.

A further assumption is that the production rate P is a concave function monotonically increasing in N^T. That is, the production rate of the line increases when the total buffer capacity N^T increases, keeping constant the machine parameters. Concavity implies that gains in production rate, obtained by increasing the total buffer capacity, decrease as N^T increases. Also the function P is assumed to be continuous in the variables N_i. These assumptions are reasonable and motivated by several numerical results appeared in literature [4,5,9].

3.2. Solution methodology

In this Section a simple heuristic method is proposed for solving the problem formulated in (1). The method is mainly based on the decomposition technique that is used to evaluate the performance, i.e. average production rate and buffer levels, of transfer lines. The main idea of the decomposition [2] is to represent the behavior of complex transfer lines by means of smaller systems or building blocks that can be easily analyzed with exact analytical methods. Thus, by assessing the performance of building blocks, it is possible to approximately predict the performance of the whole system.

In the decomposition method a transfer line with K machines is decomposed into a set of K-1 building blocks that model aggregately the whole line. Each building block is a line composed of two pseudo-machines and an inter-stage buffer having the same capacity of the corresponding buffer of the original line. The first pseudo-machine represents the portion of the original system upstream the corresponding buffer while the second pseudo-machine models the portion of the original system downstream the corresponding buffer. In order to model the behavior of the original line by means of the single two-machine lines, proper parameters have to be calculated for the pseudo-machines of each building block. Failure rates, repair rates and, eventually, speed of the pseudo-machines of each building block have to be calculated by taking into account the whole system. In particular, the pseudo-machine parameters are chosen so that the material flow in the buffers of the two-machine lines closely matches that of the corresponding buffers in the original line.

Let us face the buffer allocation problem. The method we propose to solve the primal problem starts from a minimal capacity allocation which does not allow the system to satisfy the desired value P^*. Then, it incrementally provides the line additional buffer capacity until the target value P^* is reached. This approach is based on an iterative procedure leading to optimal or near optimal solutions if buffers to be enlarged at each step are properly selected.

The main idea of the proposed method is to increase the capacity of a selected buffer B_s so that the additional production rate thus obtained is greater or at least equal than it would be gained by increasing any other buffer B_i of the line (with $i=1,\ldots,K\text{-}1$ and i¹s). In order to establish the rule for selecting the most appropriate buffer to be increased, we analyze the production rate as a concave function monotonically increasing in N^T. If we also consider that the function P is continuous and differentiable, it is possible to calculate the gradient g_i:

$$g_i = \frac{\partial P}{\partial N_i}\left(N_1,...,N_i,...,N_{K-1}\right) \qquad \text{for } i = 1,...,K-1 \tag{2}$$

The buffer corresponding to the largest gradient $g_s = max\{g_i\}_{i=1,\ldots,K-1}$ represents the buffer that determines the major enhancement of P for a small increment of storage capacity. Since the function P is not explicitly known in a closed relation, it is not straightforward to calculate the gradients g_i and a numerical approach has to be adopted.

We calculate the gradient of production rate with respect to each buffer capacities on the basis of the decomposition technique applied to the whole line:

$$g_i = \frac{\partial P}{\partial N_i} = \frac{P\left(N_1,...,N_i + ?N,...,N_{k\text{-}1}\right) - P\left(N_1,...,N_i,...,N_{k\text{-}1}\right)}{?N} \qquad \text{for } i = 1,...,K-1 \tag{3}$$

After this evaluation, the capacity of the buffer related to the largest g_i is increased and the production rate of the whole line is improved; all the process is repeated until the target P^* is reached.

Summarizing, the method is iterative: starting from an initial distribution of buffer capacities, it evaluates the performance of the line by using the decomposition technique [7]. Then, the gradients g_i are approximately calculated respect to each buffer by using the equation (3). At this point, the buffer related to the largest gradient is increased by a fixed incremental quantity and the performance of the line are again evaluated by taking into account the new buffer capacity. The method continues in such a way until the production rate P is at least equal or greater than the target value. It is worthwhile to note that the results of the last line evaluation are used as initial parameters for the next decomposition thus strongly reducing the required computational effort.

As in all iterative methods, an initial condition from which to start has to defined. We propose to set the initial conditions of the algorithm equal to the lower bounds that buffer capacities N_i have to assume for reaching the target value P^*. These lower bounds for each buffer N_i are equals to the minimal capacity that the two-machine lines, related to each buffer, need to meet in isolation (that is, ignoring the propagation of blocking and starvation phenomena inside the line) the production requirement. Therefore, by imposing P^* as production rate on each building block, the minimal buffer capacities N_i^{min} can be easily calculated thus obtaining a good lower bound to solve the problem (1).

The major improvement of the proposed method over existing techniques can be found in a conceptual and computational simplification of the overall gradient approach that is obtained preserving the accuracy of the results established in the previous papers.

Indeed, the gradient based techniques use the gradient of the production rate as a whole to identify the best direction to follow in order to determine new solutions for all the variables N_i. However, only the main derivatives (i.e. variations of the function P with respect to the single buffer capacities) are represented by the gradient and, therefore, the interdependencies between the variables N_i and their combined effect on the production rate are not properly considered when the best direction is selected.

Finally, it is worthwhile to point out the following aspects concerning the proposed method:

- the introduction of a good starting point or lower bound of buffer distribution, which consists in the capacity that each two machine line would require to meet in isolation the production requirement;
- the achievement of the optimal or near optimal buffer configuration starting from the lower bound and proceeding by means of a series of incremental steps that does never lead to exceed the target point;
- given a current buffer distribution, only one buffer is updated/increased at time and its effect is immediately evaluated by means of a decomposition on the resulting buffer distribution without introducing any further modifications.

The algorithm used for solving the addressed problem is reported the Appendix.

4. Numerical Results

On the basis of the numerical results obtained in a large number of experiments carried out, it is possible to state that the algorithm always converges to a solution. The proposed method cannot provide any guaranty to reach the optimal solution; however, in all the experiments performed the algorithm converges to a buffer configuration that is optimal or near-optimal.

In order to test the accuracy of the method, an exhaustive exploration has been carried out to find out, for each considered test case, the buffer distribution that allows to reach the production target minimizing the total buffer space allocated. Also in this case the decomposition technique has been used to evaluate the system performance. The exhaustive research of the optimal buffer distribution has been also motivated by a substantially lack of reference cases in the literature above mentioned concerning the continuous model of homogeneous lines.

Results on two different transfer lines are reported in the following tables. The first analyzed system is a line composed of three machines. Machines can fail in only one mode and the failure and repair rates are reported in Table 1 with also the efficiencies in isolation mode. The same system has been analyzed by Gershwin and Schor [4] in the discrete case. The two methods have been applied for different values of target production rates; these values have been chosen in order to get incrementally closer to the maximum throughput achievable by the system from case 1 to case 5.

	M_1	**M_2**	**M_3**
p	0.037	0.015	0.020
r	0.350	0.150	0.400
e	0.9044	0.9091	0.9524

Table 1: Three Machine Line Parameters

Table 2 reports the results of the proposed method (LMT) and the exhaustive research (ER). LMT results are obtained by setting the increment *DN* equal to 0.1. The same value corresponds also to the precision of the exhaustive research. It is possible to notice from Table 2 that the solution provided by the proposed method is always near to the optimal solution founded by the exhaustive research.

Case	**P***	**Method**	**P**	N_1	N_2	N^T
1	0.8700	N^{min}		10.56	0.06	10.62
		LMT	0.8700	14.56	5.86	20.42
		ER	0.8700	13.97	6.50	20.47
2	0.8800	N^{min}		18.12	1.89	20.01
		LMT	0.8801	22.72	8.99	31.71
		ER	0.8801	22.44	9.20	31.64
3	0.8900	N^{min}		34.85	4.87	39.72
		LMT	0.8900	40.45	14.47	54.92
		ER	0.8900	39.69	15.30	54.99
4	0.9000	N^{min}		104.98	11.10	116.08
		LMT	0.9000	112.48	28.20	140.68
		ER	0.9000	113.41	27.29	140.70
5	0.9040	N^{min}		398.44	16.11	414.55
		LMT	0.9040	402.04	47.61	449.65
		ER	0.9040	401.60	48.05	449.65

Table 2: Three Machine Line Buffer Allocation

The second analyzed system is a line composed of four machines with parameters reported in Table 3. Also in this case the method has been applied for different values of target production rates, moving toward the maximum production rate.

	M_1	M_2	M_3	M_4
p	0.050	0.0060	0.0454	0.0454
r	0.091	0.0526	0.0833	0.1429
e	0.645	0.8978	0.6471	0.7583

Table 3: Four Machine Line Parameters

Table 4 reports the results in this second case. The solution provided by the proposed method is always near to the optimal solution founded out by the exhaustive research excepted in the last case in which the target production rate is very close to the maximum throughput of the line. Indeed, in this particular case, in which the production rate is very near to the maximum throughput achievable by the system, the different gradients calculated by the algorithm are very close. Since the way of evaluating the gradient is approximated, it is difficult to select the proper buffer to be increased; however, the error on the allocated total buffer capacity is lower than 1.4% calculated on the optimal solution provided by the exhaustive research.

Case	**P***	**Method**	**P**	N_1	N_2	N_3	N^T
6	0.4950	N^{min}		0.01	0.01	0.01	0.03
		LMT	0.4953	5.81	7.51	4.71	18.03
		ER	0.4950	4.70	8.20	5.10	18.00
7	0.530	N^{min}		0.01	0.01	0.01	0.03
		LMT	0.5301	9.91	12.11	8.31	30.33
		ER	0.5300	9.20	12.70	8.40	30.30
8	0.565	N^{min}		0.01	0.01	3.84	3.86
		LMT	0.5651	16.61	19.41	14.14	50.16
		ER	0.5650	16.00	19.00	15.20	50.20
9	0.600	N^{min}		0.01	0.01	12.38	12.40
		LMT	0.6000	29.91	33.01	24.68	87.60
		ER	0.6000	29.20	33.20	25.10	87.50
10	0.640	N^{min}		38.69	33.51	50.87	123.07
		LMT	0.6400	106.10	93.61	62.77	262.48
		ER	0.6400	100.00	89.00	69.90	258.90

Table 4: Four Machine Line Buffer Allocation

5. Conclusions

A new efficient algorithm for the allocation of storage capacity in serial production lines has been developed. The heuristic proposed method is able to find out the distribution of storage capacity between machines that minimizes the total buffer space assigned to the line satisfying a desired production rate.

The main improvement of the proposed method over existing techniques can be found in a conceptual and computational simplification of the gradient approach for buffer space selection that is obtained preserving the accuracy of results.

In the paper the continuos model of homogeneous production lines have been considered, but the extension of the approach to the case of deterministic, exponential and non homogeneous line appears rather straightforward as suggested by some pilot tests carried out. Further research is needed. Indeed, in order to evaluate the accuracy of the proposed method more numerical results are needed.

Appendix

Algorithm

<u>*Step 1*</u> **Set initial conditions**.

For $i=1,\ldots,K\text{-}1$

Starting values of N_i^{min} are calculated from the single two-machine lines defined by setting their pseudo-machine parameters equal to those of the corresponding machines of the original line.

Set $k=0$.

<u>*Step 2*</u> **Evaluate the performance of the line**.

Evaluate the performance of the line and the machine parameters of each building block by using the decomposition technique described in [7].

Set $k=k+1$.

<u>*Step 3*</u> **Calculate the gradient**.

For $i=1,\ldots,K\text{-}1$

Calculate the gradient g_i by using:

$$g_i = \frac{\partial P}{\partial N_i} = \frac{P(N_1,\ldots,N_i + ?N,\ldots,N_{K\text{-}1}) - P(N_1,\ldots,N_i + ?N,\ldots,N_{K\text{-}1})}{?N}$$

<u>*Step 4*</u> **Select the buffer to be increased**.

Increase the capacity of buffer B_s related to largest gradient value:

$$g_s = \max_{i=1,\ldots,K-1}\{g_i\} = \max_{i=1,\ldots,K-1}\left\{\frac{\partial P}{\partial N_i}(N_1,\ldots, N_{K-1})\right\}$$

Increase the capacity for the selected buffer:

$$N_s^{(k)} = N_s^{(k-1)} + \Delta N$$

where the suffix k indicates the number of the iteration.

<u>*Step 5*</u> **Exit condition.**

If $P^{(k)}<P^*$ then go to step 2.

Else exit.

Bibliography

[1] Dallery Y., Gershwin S.B. 1992: *"Manufacturing Flow Line Systems: A Review of Models and Analytical Results"*, Queuing System Theory and Applications, Special Issue on Queuing Model of Manufacturing System, 12 (1-2):3-94.

[2] S.B. Gershwin: *"Manufacturing Systems Engineering"*, Prentice Hall, 1994.

[3] S.B. Gershwin and Y. Goldis: *"Efficient Algorithms for Transfer Line Design"*, MIT Laboratory for Manufacturing and Productivity Report LMP-95-005, 1995.

[4] S.B. Gershwin and J. E. Schor: *"Efficient Algorithms for Buffer Space Allocation"*, Annals of Operations Research}, Volume 93, pp 117-144, 2000

[5] P. Glasserman and D. D. Yao: *"Structured bugger-allocation problems"*, Journal of Discrete Event Dynamic Systems, Vol.6, pp. 9-42, 1996.

[6] R. Levantesi, A. Matta and T. Tolio: *"Continuous Two-Machine Lines with Multiple Failure Modes and Finite Buffer Capacity"*. IV° Convegno AITEM, Brescia, Italia, 13-15 September 1999.

[7] R. Levantesi, A. Matta and T. Tolio: *"Performance Evaluation of Continuous Production Lines with Deterministic Processing Times, Multiple Failure Modes and Finite Buffer Capacity"*. Fourth International Workshop on Queuing Network with Finite Capacity, Ilkley, Yorkshire, UK, 20-21 July 2000.

[8] T. Park: *"A Two-Phase Heuristic Algorithm for Determining Buffers Sizes of Production Lines"*, International Journal of Production Research, Vol. 31, No.3, pp. 613-631, 1993.

[9] L.E. Meester and J.G. Shantikumar: *"Concavity of the throughput of tandem queueing systems with finite buffer storage capacity"*, Advances in Applied Probability, Vol. 22, pp. 764-767, 1990.