A Simulated Annealing Approach for Buffer Allocation in Reliable Production Lines [*†]

D. Spinellis and H. T. Papadopoulos
Department of Mathematics
University of the Aegean
GR-832 00 Karlovasi, Samos
Greece
e-mail: {dspin,hpap}@aegean.gr

April, 1997

Abstract

We describe a simulated annealing approach for solving the buffer allocation problem in reliable production lines. The problem entails the determination of near optimal buffer allocation plans in large production lines with the objective of maximising their average throughput. The latter is calculated utilising a decomposition method. The allocation plan is calculated subject to a given amount of total buffer slots in a computationally efficient way.

1 Introduction and Literature Review

Buffer allocation is a major optimisation problem faced by manufacturing systems designers. It has to do with devising an allocation plan for distributing a certain amount of buffer space among the intermediate buffers of a production line. This problem is a very complex task that must account for the random fluctuations in mean production rates of the individual workstations of the lines. To solve this problem there is a need of two different tools. The first is a tool that calculates the performance measure of the line which has to be optimised (e.g., the average throughput or the mean work-in-process). This may be an *evaluative* method such as simulation or a decomposition method or the traditional

[*] *International Workshop on Performance Evaluation and Optimization of Production Lines*, pages 365–375, Samos, Greece, May 1997. University of the Aegean, Department of Mathematics.

Markovian state model. The second tool is a search (*generative*) method that tries to determine an optimal or near optimal value for the decision variables, which in our case are the buffer capacities of the intermediate buffer locations of the line. Examples of such methods are the classical search methods such as the well-known Hooke-Jeeves method, various heuristic methods, knowledge based methods, and genetic algorithms.

There are some advantages and disadvantages of the various evaluative methods used for modelling production lines. For example, simulation can handle production lines with more realistic processing times but it is a time-consuming method that cannot be utilised in conjunction with a search method that needs to apply it many times. The Markovian state model, on the other hand, provides with an exact solution of only short lines (with up to twelve stations). Besides, the processing time distribution has to be of phase-type. This is very a restrictive and unrealistic assumption, however, the method has been used to provide some insights into the problem under investigation. Finally, the decomposition method, works well for large systems (with well over than 50 stations), and it is also applicable to exponentially distributed service times. On the other hand, this method is inaccurate for short lines (with up to 10 stations).

As far as the search methods are concerned, the traditional segmentation methods (such as the well-known Hooke-Jeeves method) fail to provide with an optimal solution in all cases, whereas, the various heuristic methods reported in the literature have the advantage of being very fast but are in general also inaccurate. Lately, genetic algorithms have been applied to solve the buffer allocation problem. The accuracy and the efficiency of these algorithms have to be tested to proceed with an assessment.

For a systematic review of the existing literature in the area of evaluative and generative models of manufacturing systems, the interested reader is addressed, respectively, to two review papers by [DG92] and [PH96] and to the books by [PHB93], [AS93], [BS93], [Ger94], [Per94] and [Alt97], among others.

Although several researchers have studied the problem of optimising buffer allocation to maximise the efficiency of a reliable production line, there is no method that can handle this problem for large production lines, in a computationally efficient way (see for example, [HS91], and [HSB93]). These methods are based on comprehensive studies to characterise the optimal buffer allocation pattern. Authors have provided extensive numerical results for balanced lines with up to 6 stations and limited results for lines with up to 9 stations.

Other relevant studies are: [CMMT88], who used simulation to investigate the buffer allocation problem. [Pow92], who studied the buffer allocation problem for *unbalanced* production lines. [So97], who presented a heuristic method for determining a near optimal buffer allocation in production lines. The differentiation of So's work from the others was that the objective was to minimise the average work-in-process, provided a minimum required throughput is attained.

Recently, [PV97] developed a knowledge based system, called ASBA, for solving the buffer allocation problem in reliable production lines.

Furthermore, [BDI95] applied genetic algorithms for the buffer allocation in asynchronous assembly systems.

The objective of this paper is to present a search method for solving the buffer allocation problem in large reliable, balanced and unbalanced, production lines with computational efficiency. The proposed method is a simulated annealing approach that works in

close cooperation with a decomposition method as given in [DF93].

Simulated annealing is an adaptation of the simulation of physical thermodynamic annealing principles described by [MRR$^+$53] to the minimisation of combinatorial optimisation problems [KGV83, Cer85]. In common with genetic algorithms [Hol75] and tabu search techniques [Glo90] it follows the "local improvement" paradigm for harnessing the exponential complexity of the solution space.

The algorithm is based on randomisation techniques. An overview of algorithms based on such techniques can be found in [GSB94]. A complete presentation of the method and its applications can be found in [LA87] and accessible algorithms for its implementation are presented by [CMMR87, PFTV88]. A critical evaluation of different approaches to annealing schedules and other method optimisations are given by [Ing93].

As a tool for operational research simulated annealing is presented by [Egl90], while [KAJ94] provide a complete survey of simulated annealing applications to operations research problems.

This paper is organised as follows. Section 2 states the problem and the assumptions of the model, whereas, section 3 describes the proposed simulated annealing approach. In section 4, we provide numerical results obtained from the algorithm. Finally, section 5 concludes the paper and suggests some future research directions.

2 Assumptions of the Model and the Buffer Allocation Problem

In asynchronous production lines, each part enters the system from the first station, passes in order from all stations and the intermediate buffer locations and exits the line from the last station. The flow of the parts works as follows: in case a station has completed its processing and the next buffer has space available, the processed part is passed on. Then, the station starts processing a new part that is taken from its input buffer. In case the buffer has no parts, the station remains empty until a new part is placed in the buffer. This type of line is subject to manufacturing blocking (or blocking after service) and starving.

Assumptions of the model: It is assumed that the first station is never starved and the last station is never blocked. The processing (service) times at each station are assumed to be independent random variables following the exponential distribution, with mean service rates, μ_i, $i = 1, 2, \ldots, K$. In our model, the stations of the line are assumed to be perfectly reliable, that is, breakdowns are not allowed.

Figure 1 depicts a K-station line that has $K - 1$ intermediate locations for buffers, labelled $B_2, B_3, \ldots, B_K$.

The basic performance measures in the analysis of production lines are the average throughput (or mean production rate) and the average work-in-process (WIP) or equivalently the average production (sojourn) time.

The object of the present work is the buffering of asynchronous, reliable production lines with the assumptions given above. The objective is the maximisation of the line's throughput, subject to a given total buffer space.

The buffer allocation problem: In mathematical terms, our problem can be stated as follows:

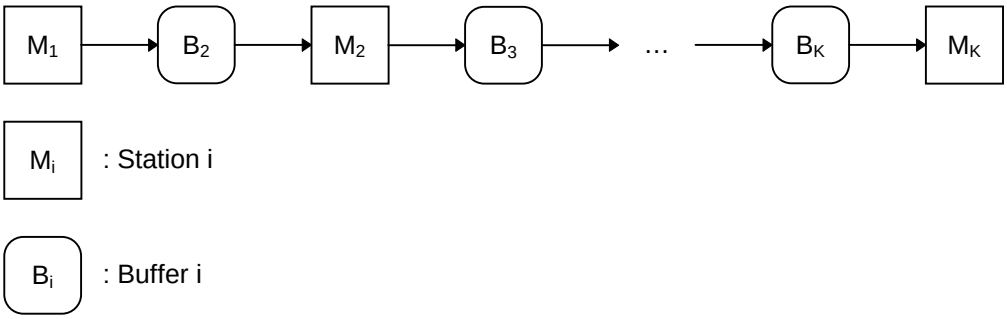

Figure 1: A K-station production line with $K-1$ intermediate buffers

P Find $B = (B_2, B_3, \ldots, B_K)$ so as to

$$\max X_K(B) \tag{1}$$

subject to:

$$\begin{array}{l} \sum_{i=2}^{K} B_i = N \\ B_i \geq 0 \\ B_i \text{ integer } (i = 2, 3, \ldots, K) \end{array} \tag{2}$$

where: N is a fixed nonnegative integer, denoting the total buffer space available in the production line.

$B = (B_2, B_3, \ldots, B_K)$ is the 'buffer vector', i.e., a vector with elements the buffer capacities of the $K-1$ buffers.

X_K, denotes the average throughput of the K-station line. This is a function of the mean service rates of the K stations, $\mu_i, \quad (i = 1, 2, \ldots, K)$, of the coefficients of variation, CV_i, of the service times and the buffer capacities, B_i.

Methodology of investigation: To solve the optimal buffer allocation problem (P), we have performed the following steps:

S1 We utilised the decomposition method given by [DF93], as an evaluative tool, to determine the mean throughput of the lines. The algorithm gives the throughput for any K-station line with finite intermediate buffers and exponentially distributed processing times.

The number of feasible allocations of N buffer slots among the $K-1$ intermediate buffer locations increases dramatically with N and K and is given by the formula:

$$\left(\begin{array}{c} n+K-2 \\ K-2 \end{array} \right) = \frac{(N+1)(N+2)\cdots(N+K-2)}{(K-2)!} \tag{3}$$

S2 To find the buffer allocation that maximises the throughput of the line, we utilised the *simulated annealing* method specifically adapted for solving this problem. Our approach is described in detail in the following section.

3 The Simulated Annealing Approach

Simulated annealing is an optimisation method suitable for combinatorial minimisation problems. Such problems exhibit a discrete, factorially large configuration space. In common with all paradigms based on "local improvements" the simulated annealing method starts with a non-optimal initial configuration (which may be chosen at random) and works on improving it by selecting a new configuration using a suitable mechanism (at random in the simulated annealing case) and calculating the corresponding cost differential (ΔX_K). If the cost is reduced, then the new configuration is accepted and the process repeats until a termination criterion is satisfied. Unfortunately, such methods can become "trapped" in a local optimum that is far from the global optimum. Simulated annealing avoids this problem by allowing "uphill" moves based on a model of the annealing process in the physical world.

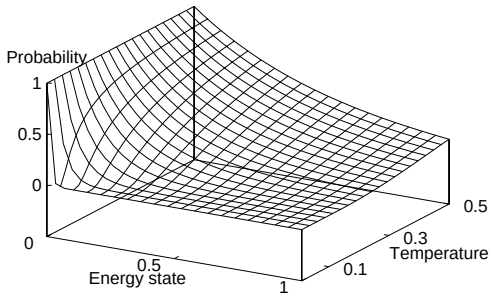

Figure 2: Probability distribution $w \sim \exp(\frac{-E}{kT})$ of energy states according to temperature.

When metals slowly cool and anneal their atoms are often ordered in the minimal energy crystalline state for distances billions of times their diameter in all directions i.e. the solid is in a state of a global minimum. During the cooling process the system can escape local minima by moving to a thermal equilibrium of a higher energy potential based on the probabilistic distribution w of entropy S

$$S = k \ln w \tag{4}$$

where k is Boltzmann's constant and w the probability that the system will exist in the state it is in relative to all possible states it could be in. Thus given entropy's relation to energy E and temperature T

$$dS = \frac{đE}{T} \tag{5}$$

we arrive at the probabilistic expression w of energy distribution for a temperature T

$$w \sim \exp(\frac{-E}{kT}) \tag{6}$$

This so-called Boltzmann probability distribution is illustrated in figure 2. The probabilistic "uphill" energy movement that is made possible avoids the entrapment of local minima and can provide a globally optimal solution.

select an initial line configuration C_0 and an initial temperature T_0
repeat until no better configurations can be found
    repeat for a number of optimisation steps for the given temperature
        Configure a new line C_n by moving a random ammount of buffer space from one randomly selected buffer to another
        Calculate the energy differential ΔE between the current line configuration C and the new one C_n
        If the new line C_n is more efficient ($\Delta E < 0$) or it satisfies the Metropolis criterion $R < \exp(\frac{-\Delta E}{T})$ for a random number $R, 0 < R < 1$ and an annealing temperature T then
            Make the new configuration C_n the current configuration C
    Lower the annealing temperature T following the cooling schedule (equation 7).

Figure 3: Simulated annealing pseudocode for production line buffer allocation.

The application of the annealing optimisation method to other processes works by repeatedly changing the problem configuration and gradually lowering the temperature until a minimum is reached.

This can be expressed for the production line buffer allocation using the pseudocode listed in figure 3.

4 Numerical Results

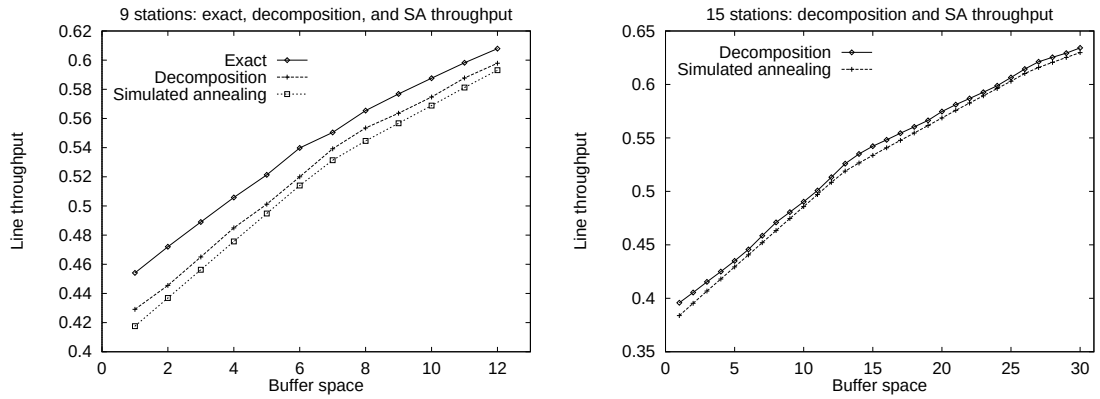

Figure 4: Computed throughput of simulated annealing compared with full and reduced enumerations for 9 stations (left); compared with reduced enumeration for 15 stations (right).

In order to evaluate the applicability of the simulated annealing method to the buffer allocation problem we designed and implemented a system to calculate the optimum buffer configuration for a given reliable production line using a simulated annealing algorithm. The system takes as input:

- the number of production line stations K,
- the available buffer space N, and
- the station mean service rates, μ_i, $i = 1, 2, \ldots, K$.

Based on the above input the system calculates the buffer allocations $B = (B_2, B_3, \ldots, B_K)$ for the maximal line throughput. Furthermore, the system is instrumented to provide as part of the solution the throughput of the suggested configuration, as well as the number of different configurations that were tried. The line throughput is used to evaluate the *quality* of the suggested configuration when compared with the throughput calculated by other methods. The number of different configurations tried, is used as an objective *performance criterion*, because the configuration evaluation step is the dominant execution time factor and the basic building block of all optimisation methods.

The system is based on the simulated annealing algorithm as described in [PFTV88]. The authors do not clarify that, strictly described, their implementation is a *simulated quenching* [Ing93] algorithm as it uses an exponential cooling schedule. Our implementation, for efficiency reasons, uses the same exponential cooling schedule namely:

$$\begin{array}{ll} T_{k+1} = cT_k & 0 < c < 1 \\ \frac{\Delta T}{T_k} = (c-1)\Delta k & k \gg 1 \\ T(k) = T_0 \exp((k(c-1)) & \end{array} \tag{7}$$

instead of the standard logarithmic schedule consistent with the Boltzmann algorithm

$$T(k) = T_0 \frac{\ln k_0}{\ln k} \tag{8}$$

The random floating point numbers $0 < R < 1$ used for selecting energy differentials based on the annealing temperature $R < \exp(\frac{-\Delta E}{T})$ are produced using the *substractive method* algorithm described in [Knu81]. Finally, the evaluative function that we used for calculating ΔE is based on the decomposition method [DF93].

In order to evaluate our method's applicability in selecting line configurations we run a number of tests on both balanced and unbalanced lines and compared the simulated annealing results against the results obtained by other methods. For short lines and limited buffer space a full enumeration of all configurations provided an accurate measure when comparing with the simulated annealing results. For larger configurations we used a reduced enumeration in order to provide the comparative measure. Both methods are subject to the reduced evaluative accuracy of the decomposition method compared to the Markovian model. In figure 4 we present the optimum throughput configurations for balanced lines found using the simulated annealing method against the throughput found using full (for 9 stations) and reduced enumeration techniques. It is apparent that the simulated annealing results follow closely the results obtained by the other methods.

In addition to the balanced line evaluation we compared the simulated annealing method against unbalanced line enumeration using the Markovian evaluative procedure for a variety of line sizes, service time configurations, and available buffer space. The results are summarised using error bars in figure 5. It is apparent that the simulated annealing configurations are not always optimal for limited available buffer space, but they

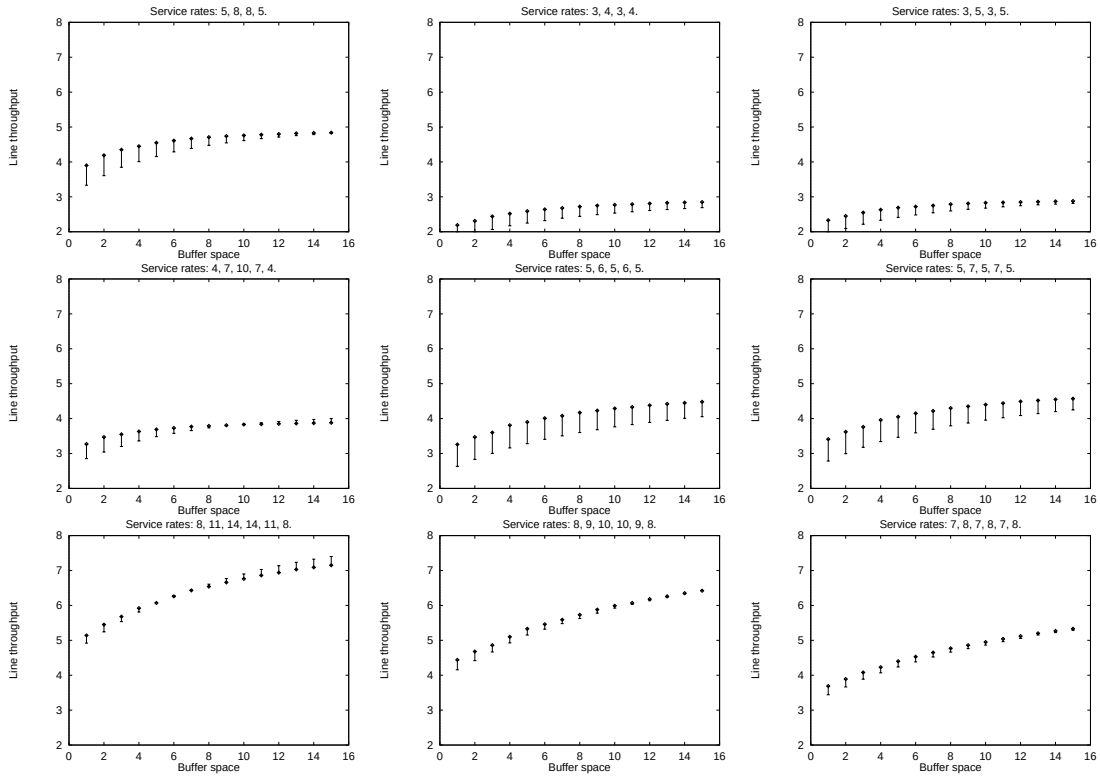

Figure 5: Simulated annealing with decomposition evaluation (dash ticks) versus enumerated Markovian (dot ticks) throughputs for unbalanced lines with 4–6 stations.

quickly converge with the optimal configurations as buffer space increases. This difference can be accounted by the use of the fast decomposition evaluative procedure in the similated annealing method against the use of the Markovian evaluative procedure for the enumeration method. As the decomposition method is not accurate for small sets of unbalanced lines this is an expected outcome and could be corrected by using the Markovian evaluation in the simulated annealing optimisation of small unbalanced production lines.

Our goal for using the simulated annealing method was to test its applicability to large production line problems where the cost of other methods was prohibitevely expensive. As an example the *reduced* enumeration method when run on a 15 station line with a buffer capacity of 30 units took more than 10 hours to complete on a 100MHz processor. As shown in figure 6 the cost of the simulated annealing method is higher than the cost of the full and reduced enumeration methods for small lines and buffer allocations. However, it quickly becomes competitive as the number of stations and the available buffer size increase. Notice that — in contrast to the other methods — the simulated annealing cost does not increase together with the available buffer space and that it increases only linearly with the number of stations.

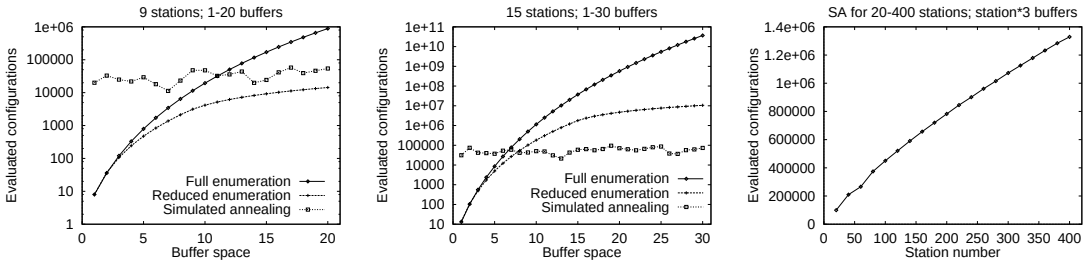

Figure 6: Performance of simulated annealing compared with full and reduced enumerations for 9 stations (left); compared with reduced enumeration for 15 stations (middle); run for up to 400 stations (right). Note the log_{10} scale on the ordinate axis.

5 Conclusions and Further Work

The results obtained using the simulated annealing method to the reliable line optimal buffer allocation problem are clearly encouraging. The performance and the accuracy of the method, although inferior for optimising small lines with limited buffer space, seem to indicate clearly that it becomes the method of choice as the problem size increases. These characteristics suggest that the method fits nicely *together* with the existing optimisation tools satisfying the need for a large production line optimisation tool.

Further investigation is needed in order to fully evaluate the method's potential. The annealing schedule that we used can clearly be optimised potentially increasing both the method's accuracy and its performance. Methods such as adaptive simulated annealing [Ing89] can be tried on the problem set in order to test their applicability. The use of heuristics in setting up the initial buffer configuration can decrease the number of steps needed for reaching the optimal. Other evaluative methods such as the Markovian model can be used in place of the decomposition algorithm for determining the change differentials. Furthermore, the small nature of changes made to the configuration during the annealing process could be taken into account for optimising the evaluative procedure. Finally, we would like to test the method's potential on similar problems especially involving parallel station production lines.

Acknowledgments

Michael Vidalis implemented the decomposition algorithm and provided us with the decomposition and reduced decomposition evaluative results.

References

[Alt97] T. Altiok. *Performance Analysis of Manufacturing Systems*. Springer-Verlag, 1997.

[AS93] R. G. Askin and C. R. Standridge. *Modeling and Analysis of Manufacturing Systems*. Wiley, 1993.

[BDI95] A. A. Bulgak, P. D. Diwan, and B. Inozu. Buffer size optimization in asynchronous assembly systems using genetic algorithms. *Computers ind. Engng*, 28(2):309–322, 1995.

[BS93] J. A. Buzacott and J. G. Shanthikumar. *Stochastic Models of Manufacturing Systems*. Prentice Hall, 1993.

[Cer85] V. Cerny. Thermodynamical approach to the traveling salesman problem: an efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45:41–51, 1985.

[CMMR87] A. Corana, M. Marchesi, C. Martini, and S. Ridella. Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm. *ACM Transactions on Mathematical Software*, 13(3):262–280, September 1987.

[CMMT88] R. W. Conway, W. L. Maxwell, J. O. McClain, and L. J. Thomas. The role of work-in-process inventories in series production lines. *Operations Research*, 36:229–241, 1988.

[DF93] Y. Dallery and Y. Frein. On decomposition methods for tandem queueing networks with blocking. *Operations Research*, 41(2):386–399, 1993.

[DG92] Y. Dallery and S. B. Gershwin. Manufacturing flow line systems: A review of models and analytical results. *Queueing Systems: Theory and Applications*, 12:3–94, 1992.

[Egl90] R. W. Eglese. Simulated annealing: A tool for operational research. *European Journal of Operational Research*, 46:271–281, 1990.

[Ger94] S. B. Gershwin. *Manufacturing Systems Engineering*. Prentice Hall, 1994.

[Glo90] F. Glover. Tabu search — part I. *ORSA Journal on Computing*, 1:190–206, 1990.

[GSB94] R. Gupta, S. A. Smolka, and S. Bhaskar. On randomization in sequential and distributed algorithms. *ACM Computing Surveys*, 26(1):7–86, 1994.

[Hol75] J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, Michigan, 1975.

[HS91] F. S. Hillier and K. C. So. The effect of the coefficient of variation of operation times on the allocation of storage space in production line system. *IIE Transactions*, 23:198–206, 1991.

[HSB93] F. S. Hillier, K. C. So, and R. W. Boling. Notes: Toward characterizing the optimal allocation of storage space in production line systems with variable processing times. *Management Science*, 39(1):126–133, 1993.

[Ing89] L. Ingber. Very fast simulated re-annealing. *Journal of Mathematical Computation Modelling*, 12:967–973, 1989.

[Ing93] L. Ingber. Simulated annealing: Practice versus theory. *Journal of Mathematical Computation Modelling*, 18(11):29–57, 1993.

[KAJ94] C. Koulamas, S. R. Antony, and R. Jaen. A survey of simulated annealing applications to operations research problems. *Omega International Journal of Management Science*, 22(1):41–56, 1994.

[KGV83] S Kirkpatrick, C. Gelatt, and P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–679, 1983.

[Knu81] D. E. Knuth. *The Art of Computer Programming*, volume 2 / Seminumerical Algorithms, pages 171–173. Addison-Wesley, second edition, 1981.

[LA87] P. J. Van Laarhoven and E. H. L. Aarts. *Simulated Annealing: Theory and Applications*. Dordrecht, 1987.

[MRR+53] N. Metropolis, A. N. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and H. Teller. Equation of state calculation by fast computing machines. *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[Per94] H. Perros. *Queueing Networks with Blocking*. Oxford University Press, 1994.

[PFTV88] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C*, pages 343–352. Cambridge University Press, 1988.

[PH96] H. T. Papadopoulos and C. Heavey. Queueing theory in manufacturing systems analysis and design: A classification of models for production and transfer lines. *European Journal of Operational Research*, 92:1–27, 1996.

[PHB93] H. T. Papadopoulos, C. Heavey, and J. Browne. *Queueing Theory in Manufacturing Systems Analysis and Design*. Chapman and Hall, 1993.

[Pow92] S. G. Powell. Buffer allocation in unbalanced serial lines. Working Paper 289, The Amos Tuck School of Business Administration, Dartmouth College, 1992.

[PV97] H. T. Papadopoulos and G. A. Vouros. A model management system (MMS) for the design and operation of production lines. International Journal of Production Research, 1997. (to be published).

[So97] K. C. So. Optimal buffer allocation strategy for minimizing work-in-process inventory in unpaced production lines. *IIE Transactions*, 29:81–88, 1997.