

Practical considerations in the optimization of flow production systems

HORST TEMPELMEIER†

In this paper we consider the problems faced by an industrial planner who is responsible for the design of real-life asynchronous production lines under stochastic conditions that may be due to breakdowns or random processing times. Based on real-life system data, it is shown that a number of available algorithms for the performance evaluation of a given system configuration as well as an algorithm for determining the optimum buffer configuration can be successfully applied in industrial practice.

1. Introduction

Over the last ten years, much effort has been devoted to the development of approaches for the performance analysis and optimization of stochastic flow production systems with finite buffers between the processing stations. The current body of knowledge is presented in the papers of Dallery and Gershwin (1992), Papadopoulos and Heavey (1996) and Gershwin (2000) as well as in several monographs on stochastic models of manufacturing systems, such as Visvanadham and Narahari (1992), Buzacott and Shanthikumar (1993), Papadopoulos *et al.* (1993), Askin and Standridge (1993), Gershwin (1994), and Altiok (1997).

In spite of the significant benefits resulting from the application of analytical planning methods (an award-winning practical case study performed at Hewlett-Packard is described by Burman *et al.* 1998), many industrial planners seem to be rather reluctant to apply the available methods. As far as we know, only a very small number of companies use analytical flow line models. In many cases, planners have only limited knowledge about the existence of practically applicable evaluation methods.

If quantitative performance evaluation is carried out at all, then in almost any case simulation is the only tool used. Optimization problems, such as buffer optimization, are mainly solved through simple trial-and-error approaches, which suffer from the severe drawbacks of being both very time-consuming and providing solutions that are usually far from optimal. According to an empirical study conducted in 43 German companies in 1989, analytical planning tools were not applied at all (Schöniger and Spingler 1989). This empirical evidence has not changed much.

Several researchers provide rules of thumb and general insights acquired through the in-depth study of different configurations of several hypothetical flow production systems. See Conway *et al.* (1988), Blumenfeld (1990), Baker (1993), Hillier *et al.* (1993), Baker *et al.* (1994), Powell and Pyke (1996), Hillier and So (1996), Liu and Lin (1996), So (1997), Powell and Pyke (1998), Hillier (2000), and Enginarlar *et al.*

Revision received May 2002

†Universität zu Köln, Seminar für Produktionswirtschaft, Albertus-Magnus-Platz, D-50923 Köln, Germany. e-mail: tempelmeier@wiso.uni-koeln.de

(2002). Industrial planners often believe that an optimal flow line design can be developed based on the experience they have acquired with similar past projects.

Unfortunately, in the problem domain considered, the value of experience is rather limited. The reason is that even a small change of data or system characteristics may generate a considerably different behaviour of the system under study. For example, changing the failure characteristics of a station through the introduction of a machine based on a different technology or slightly changing the processing time at a station may shift the bottleneck of the system with the need to rearrange the buffers completely. As every production line is obviously unique, it jeopardizes the economic efficiency if a flow line planner relies completely upon experience gathered from observations of other production lines. Therefore, tools are required that can provide system-specific performance measures in a fast and reliable manner.

This paper is organized as follows. In section 2 we discuss several characteristics of real-life production systems that can be covered by analytical planning methods. It is shown, that a planner must be provided with basically three algorithms in order to find good performance approximations in a wide variety of practical situations. The quality of the algorithms is proved with the help of simulation on the basis of hypothetical as well as real-life data. In section 3, several types of optimization problems that emerge in industrial planning practice are discussed. Section 4 contains our conclusion and discusses several issues regarding the practical application of the analytical methods described.

2. Performance analysis of industrial flow production systems

Industrial factory planners who are responsible for building up sufficient production capacity in an economical way are confronted with a number of design factors that have an effect on the throughput of a planned production system subject to stochastic influences (breakdowns, random processing times). In addition to technical considerations, such as the definition of the processes and the specification of the required production and material handling resources, there are several organizational issues that have to be decided upon. In the following, we deal with the question of how the negative consequences caused by stochastic phenomena such as breakdowns or the variability of the processing times can be predicted and how the resulting loss of throughput can be regained through the introduction of buffer spaces between the stations.

Many flow production systems comprise special-purpose machines that repetitively perform a certain number of tasks on a single product type. In this case, the processing times are *deterministic*. Randomness arises from breakdown and repair processes. In the following we assume operation-dependent failures. Based on an empirical study, Inman (1999) comes to the conclusion that the widespread assumption of exponential times to failure and exponential repair times is acceptable in many cases.

There are several situations where *stochastic processing times* must be taken into account. First, if the task is repetitively performed by a human operator, then processing times will usually be random, as a certain amount of variability is inherent in human nature. Empirical studies show that task durations of human operators will have a coefficient of variation that is considerably less than one, which would be the case for an exponential distribution. Secondly, if a number of flexible automatic machines or robots assigned to a station are able to process a mixture of product variants in any sequence, then—from the point of view of an external observer—the

processing times of the variants can be considered as random. For a recent empirical analysis of processing times in automobile welding shops see Inman (1999).

Many researchers providing algorithms for the analysis of flow production systems with stochastic processing times assume that processing times are *exponential*. In addition, several approaches are applicable only if the mean processing times, failure characteristics, and buffer sizes of all stations are *identical*. It is believed that in the (deterministic) line balancing phase the system characteristics are set up such that there will be no station that is much worse than all the others (with respect to the isolated throughput). However, in industry this is not always the case. Due to technical constraints, many real-life systems have stations with non-identical mean processing times—even if automatic machines are used to process a single product type. Differences in the mean processing times may also emerge as a result of a simultaneous buffer and workload allocation, as discussed in section 3.

Figure 1 depicts the isolated station throughputs for four real-life linear flow production systems. The detailed data are given in the appendix. Observe that the data exhibit a certain amount of imbalance, which makes many algorithms for the evaluation of stochastic flow production systems inapplicable or, at best, imprecise.

From the practical data presented, it is clear that an analytical flow line model must be able to cover unequal processing times. In addition, unequal failure and repair characteristics must also be covered by such a model. In what follows, models and algorithms available for the approximate performance analysis of these types of flow production systems are tested with regard to their applicability in real-life planning environments. From a practical point of view, the algorithms must be able to analyse systems with up to 50 or 100 stations, a requirement that excludes exact algorithms.

2.1. Deterministic processing times

In a flow production system that comprises exclusively automatic stations working on a single part type, the processing times are usually deterministic but vary from station to station. This is often due to the fact that it is not possible to find combinations of subprocesses that sum to exactly the same processing time at all stations. Flow production systems of this kind can be analysed with the 'Accelerated Dallery-David-Xie' (ADDX) algorithm proposed by Burman (1995), which is based on the decomposition method developed by Gershwin (1987) (see also Gershwin 1994) and which is an extension of the DDX-algorithm of Dallery *et al.* (1988). With respect to the optimization algorithm discussed in section 3 it is noteworthy that the buffer sizes are modelled as continuous variables.

With the help of a large numerical experiment based on hypothetical system data, Burman (1995) showed that his algorithm was very accurate. In the following, this algorithm is applied to an invented system and several real-life systems.

Invented system 1

Consider a system with invented data (ten stations with identical buffer sizes; identical deterministic processing times $s = 1$; identical failure processes at all stations: failure rate $f = 0.007$, repair rate $r = 0.095$ (isolated throughput = 0.931)). Table 1 compares the system throughput found with the help of a SIMAN simulation model ($X_{\text{simulated}}$) and with the ADDX algorithm ($X_{\text{estimated}}$).

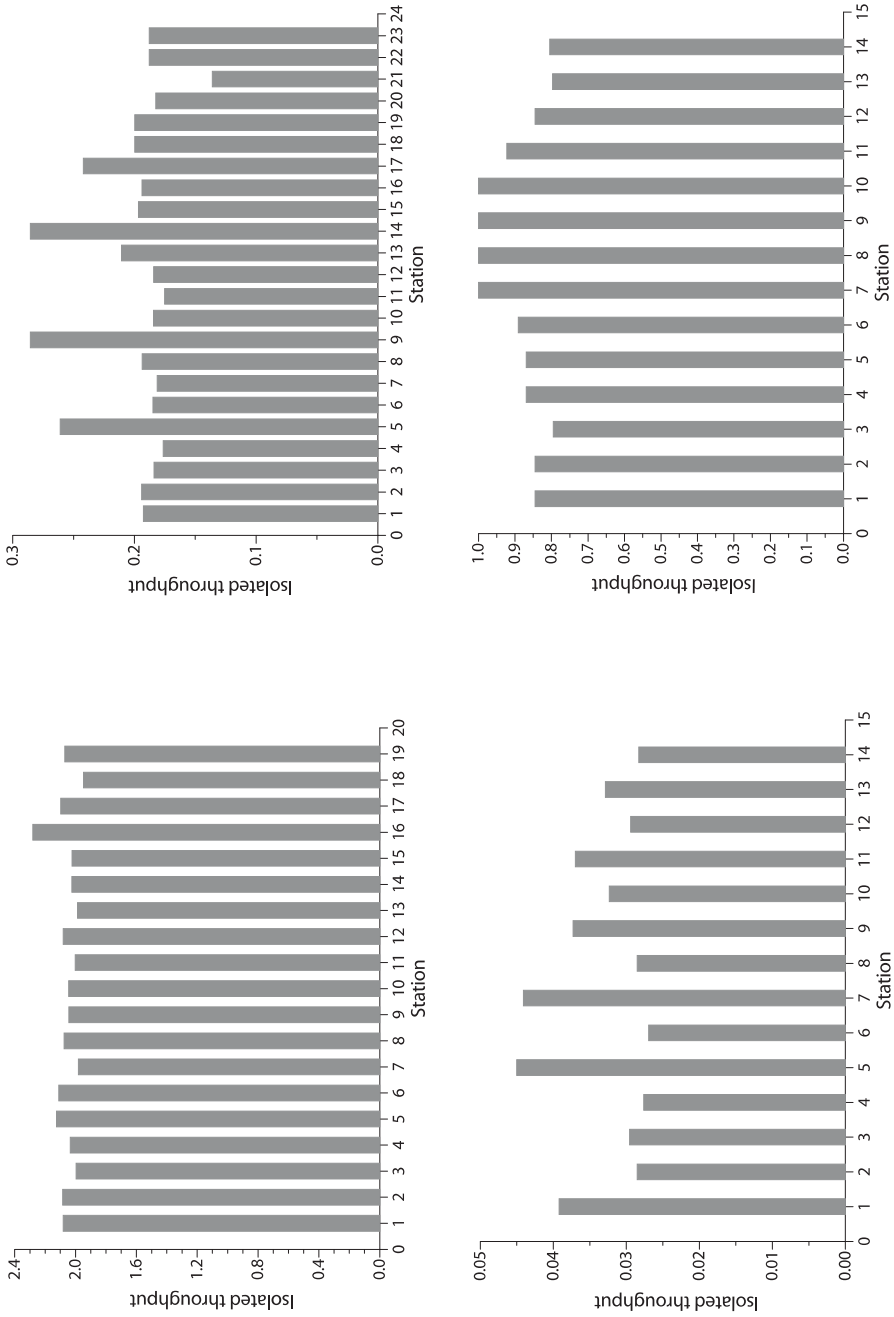


Figure 1. Isolated station throughputs of real-life systems.

Buffer size	$X_{\text{estimated}}$	$X_{\text{simulated}}$	% Deviation
0	0.57575	0.58883	-2.22%
5	0.72546	0.72310	+0.33%
10	0.78682	0.78080	+0.77%
25	0.85548	0.85018	+0.62%

Table 1. Results for an invented system.

Real-life system A

Next, we consider the data of the real-life *system A* shown in table 18 in the appendix. Table 2 presents four scenarios with different buffer allocations: (1) zero buffers; (2) allocation of 285 buffers provided by the industrial planners; (3) 285 buffers allocated optimally with the help of the algorithm discussed in section 3; (4) target throughput of 1.8 with 485 buffers allocated optimally.

In all cases considered, the simulated throughput is very well approximated. The experiment further demonstrates that distributing the given total number of 285 buffers *optimally* among the stations would result in a slightly higher throughput, which is shown both by the estimated as well as the simulated values. Moreover, if another 200 buffers were introduced into the system, the throughput could increase by several percent.

Real-life system B

In system A the availabilities of the different stations ranged between 96.7% and 99.4%. In order to examine the quality of the approximation for stations with lower availability, consider system B with the data shown in table 19 in the appendix, which comprises stations with availabilities as low as 83%. Some of the stations have availabilities of 100%. The isolated throughput of the worst station is only about 50% of the isolated throughput of the best station. Note that in addition to the large differences of the availabilities there are also very large differences in the processing times.

In table 3, the throughput estimations are compared with simulated results for the buffer configuration given by the planners and the optimum buffer sizes that were

Total number of buffers	$X_{\text{estimated}}$	$X_{\text{simulated}}$	% Deviation
0	1.52294	1.52928	-0.41%
285 (<i>given</i> allocation)	1.71403	1.68984	+1.43%
285 (<i>optimal</i> allocation)	1.73260	1.70140	+1.83%
485 (<i>optimal</i> allocation)	1.80000	1.76256	+2.12%

Table 2. Results for system A.

	$X_{\text{estimated}}$	$X_{\text{simulated}}$	% Deviation
0 buffers	0.09238	0.10344	-10.70%
1 buffer per station	0.11334	0.11503	-1.47%
459 buffers (<i>given</i> allocation)	0.13474	0.13517	-0.32%
193 buffers (<i>optimal</i> allocation)	0.13474	0.13300	1.31%

Table 3. Results for system B.

required to provide the same estimated throughput. We also simulated a system configuration without any buffers and one with a single buffer per station. For the case of zero buffers, the algorithm is too pessimistic. However, if at least one buffer is included at every station, then the approximation is very good. In addition, it is observed that the total number of buffers could be reduced from 459 to 193 if these were *allocated optimally* to the stations.

Real-life system C

The next system C, with the data shown in table 20 in the appendix, is marked by unequal processing times and unequal as well as rather low availabilities (ranging from 80% to 96%), but with identical mean repair times (see table 4).

The experiments demonstrate that the ADDX algorithm can be expected consistently to provide accurate throughput approximations not only for invented system configurations, but also for real-life systems. Only for systems with unrealistic total buffer sizes of zero does the ADDX algorithm significantly underestimate the throughput. This is not a real drawback, as no planner will consider a system with zero buffers as a design alternative under conditions of substantial randomness.

2.2. Stochastic processing times

As stated above, there are a number of situations where models of stochastic processing times may be adequate. In these cases, evaluation methods are required that are applicable for various degrees of variability of the processing times. For systems where the stations exhibit general stochastic processing times, a decomposition approach based on the stopped arrival $G/G/1/N$ queueing model (see Buzacott and Shanthikumar 1993, Buzacott *et al.* 1995) can be applied. Although the main portion of randomness is caused by the nature of the operations, in some cases failures must also be considered. Failures may arise through problems with the handling of certain workpieces by an operator or they may be caused by the tools used to perform an operation.

Although the stopped arrival $G/G/1/N$ queueing model does not take failures into account, this can be accomplished with the help of the completion time concept proposed by Gaver (1962) (see Altiok and Stidham 1983). In this case, the parameters of the processing time (mean and coefficient of variation) are adjusted such that they include the residence time of a workpiece caused by station failures. Numerical tests with a wide variety of data have shown that this approach delivers good performance approximations. Through the consideration of simultaneous starving and blocking, the quality of the approximations can be improved in some cases. See Bürger (2001), where a simple form of quality control (inspection) after processing at a station is proposed.

	$X_{\text{estimated}}$	$X_{\text{simulated}}$	% Deviation
0 buffers	0.00204	0.00248	-17.70%
1 buffer per station	0.00290	0.00297	-2.36%
131 buffers (given allocation)	0.00367	0.00366	0.27%
131 buffers (optimal allocation)	0.00371	0.00369	0.54%

Table 4. Results for system C.

CV	MTTR				
	36	72	108	144	180
0.1	0.19	0.37	0.55	0.73	0.91
0.4	0.34	0.52	0.70	0.88	1.06
0.7	0.67	0.85	1.03	1.21	1.39
1	1.18	1.36	1.54	1.72	1.90
1.3	1.87	2.05	2.23	2.41	2.59

Table 5. Squared coefficients of variation of the completion times for system 2.

CV	MTTR				
	36	72	108	144	180
0.1	0.51%	-0.10%	0.78%	-0.67%	1.48%
0.4	0.58%	1.11%	2.57%	0.86%	1.66%
0.7	0.22%	0.41%	-0.54%	0.90%	0.99%
1	-0.82%	-0.17%	1.49%	-0.89%	-0.85%
1.3	-0.35%	1.66%	0.61%	0.16%	0.47%

Table 6. Results for system 2; $(X_{\text{estimated}} - X_{\text{simulated}})/X_{\text{simulated}}$.

Invented system 2

In order to test the quality of the queueing-model based approach, we simulated an invented flow line comprising ten stations with identical mean processing times (= 36 seconds) and coefficients of variation as well as identical availabilities (= 90%), mean repair times and buffers sizes (= 5). The estimated values were computed with approximation (b) from Buzacott *et al.* 1995). Failures and repairs were accounted for through the completion time approach of Gaver (1962), as described in Tempelmeier and Bürger (2001). The mean completion time, which is identical for all stations, is 40 in all cases, and the squared coefficients of the completion times are given in table 5 with different mean times to repair and different coefficients of variation of the processing times.

In table 6, the percentage deviations of the estimated throughput from the simulated values are shown.

We also tested approximation (a) from Buzacott *et al.* (1995), which resulted in a mean absolute percentage deviation of 1.97% and performed slightly poorer than the above approximation (b), which has a mean absolute percentage deviation of 0.83%.

Real-life system D

Next, we consider the real-life *system D* with mean processing times and availabilities as given in table 21 in the appendix. For this system we varied the coefficients of variation of the processing times and the mean times to repair simultaneously for all stations. The results are presented in table 7.

For both systems considered, the quality of the approximations are quite good. Further approximation results with respect to hypothetical data are provided in Tempelmeier and Bürger (2001). With respect to the buffer optimization algorithm discussed in section 3, note that the formulas used in the evaluation of a subsystem

CV	MTTR		
	36	180	360
0.5	0.75%	-1.88%	-4.70%
1	-1.10%	-2.57%	-4.40%

Table 7. Results for system D;
 $(X_{\text{estimated}} - X_{\text{simulated}})/X_{\text{simulated}}$.

with the stopped arrival $G/G/1/N$ queueing model can also be applied for non-integer buffer sizes.

2.3. *Mixed deterministic/stochastic processing times*

In industrial practice, one will sometimes find systems where the vast majority of stations have deterministic processing times whereas only a few stations, say, up to three, have stochastic processing times. In this case, there are two options.

The *first option*—denoted ADDX(adj)—is to use the ADDX algorithm, which can only handle deterministic processing time stations, and represent every station with stochastic processing times by an equivalent station with deterministic processing times and failures. Using Gaver’s completion time approach, we modify the parameters of the stochastic stations in the deterministic model such that the first three moments of the completion time under stochastic processing times are the same as the first three moments under deterministic processing times. This is done as follows. Let S, D and ν denote the processing time, the repair time, and the failure rate, respectively for a stochastic processing time station and let s, r and f denote the deterministic processing time, the repair rate, and the failure rate, respectively for a deterministic processing time station.

With Poisson distributed failures and general distributed processing times, the moments of the completion time are

$$E\{C\}_{\text{stoch}} = E\{S\}(1 + \nu E\{D\}) \tag{1}$$

$$E\{C^2\}_{\text{stoch}} = \nu E\{D^2\}E\{S\} + E\{S^2\}(1 + \nu E\{D\})^2, \tag{2}$$

$$E\{C^3\}_{\text{stoch}} = \nu E\{D^3\}E\{S\} + 3E\{D^2\}E\{S^2\}\nu(1 + \nu E\{D\}) + E\{S^3\}(1 + \nu E\{D\})^3 \tag{3}$$

When the processing times are *deterministic* with value s , and under the assumption of Poisson distributed failures the first three moments of the completion time are

$$E\{C\}_{\text{det}} = s\left(1 + \frac{f}{r}\right), \tag{4}$$

$$E\{C^2\}_{\text{det}} = f\left(\frac{2}{r^2}\right)s + s^2\left(1 + \frac{f}{r}\right)^2, \tag{5}$$

$$E\{C^3\}_{\text{det}} = \left(f \frac{6}{r^3}s\right) + 3\frac{2}{r^2}s^2f\left(1 + \frac{f}{r}\right) + s^3\left(1 + \frac{f}{r}\right)^3. \tag{6}$$

Setting $E\{C\}_{\text{stoch}} = E\{C\}_{\text{det}}$, $E\{C^2\}_{\text{stoch}} = E\{C^2\}_{\text{det}}$, and $E\{C^3\}_{\text{stoch}} = E\{C^3\}_{\text{det}}$, we solve for the adjusted deterministic processing time s , the adjusted failure rate f , and the adjusted repair rate r .

The solution of this system of equations is as follows:

$$s = \frac{p}{2n}, \tag{7}$$

$$f = \frac{9q^3}{np}, \tag{8}$$

$$r = \frac{3q}{n}, \tag{9}$$

with

$$n = 2E\{S\}^3g^3 + 3g\nu E\{D^2\}(E\{S^2\} - E\{S\}^2) + E\{S^3\}g^3 - E\{S\}(3E\{S^2\}g^3 - \nu E\{D^3\}), \tag{10}$$

$$p = 2E\{S\}E\{S^3\}g^4 - 3E\{S^2\}^2g^4 + E\{S\}^4g^4 + E\{S\}^2(2E\{D^3\}\nu g - 3\nu^2E\{D^2\}^2), \tag{11}$$

$$q = E\{S\}\nu E\{D^2\} + (E\{S^2\} - E\{S\}^2)g^2, \tag{12}$$

$$g = 1 + \nu E\{D\}. \tag{13}$$

The *second option*—denoted GG1(CV0)—would be to apply the algorithm based on the $G/G/1/N$ stopped arrival queueing model with the coefficients of variation set to zero for all deterministic stations in the system. In order to test the quality of the approximations, we performed several simulation experiments.

Invented system 3

The first invented system comprises ten (identical) stations. Stations 1 and 10 are manual workstations with gamma-distributed processing times. Stations 2 to 9 have deterministic processing times of 1. For all stations the failure rates are 0.007, and the repair rates are 0.095 (isolated throughput = 0.931). Failures and repairs are Poisson distributed. In tables 8–10 the throughputs estimated with the approximation options are compared with simulation results for different coefficients of variation at stations 1 and 10 and for different buffer sizes.

Buffer size	$X_{\text{simulated}}$	ADDX(adj)		GG1(CV0)	
		$X_{\text{approximated}}$	% Deviation	$X_{\text{approximated}}$	% Deviation
0	0.51119	0.55072	7.7%	0.33648	-34.2%
1	0.60363	0.60464	0.2%	0.45479	-24.7%
5	0.71198	0.71966	1.1%	0.66534	-6.6%
10	0.76937	0.78388	1.9%	0.76003	-1.2%
25	0.83630	0.85429	2.2%	0.84852	1.5%

Table 8. Results for $CV = 0.6$ at stations 1 and 10.

Buffer size	$X_{\text{simulated}}$	ADDX(adj)		GG1(CV0)	
		$X_{\text{approximated}}$	% Deviation	$X_{\text{approximated}}$	% Deviation
0	0.47510	0.50562	6.4%	0.33013	-30.5%
1	0.56098	0.57603	2.7%	0.44836	-20.1%
5	0.69666	0.70954	1.8%	0.65905	-5.4%
10	0.75806	0.77841	2.7%	0.75517	-0.4%
25	0.83509	0.85182	2.0%	0.84565	1.3%

Table 9. Results for $CV = 1$ at stations 1 and 10.

Buffer size	$X_{\text{simulated}}$	ADDX(adj)		GG1(CV0)	
		$X_{\text{approximated}}$	% Deviation	$X_{\text{approximated}}$	% Deviation
0	0.41320	0.37899	-8.3%	0.30565	-26.03%
1	0.49499	0.48350	-2.3%	0.41759	-15.6%
5	0.63839	0.65699	2.9%	0.62411	-2.2%
10	0.71756	0.74131	3.3%	0.72344	0.8%
25	0.80925	0.83088	2.7%	0.82433	1.9%

Table 10. Results for $CV = 2$ at stations 1 and 10.*Invented system 4*

In addition to the assumptions made for the first system, we assume for the second system that stations 1, 5 and 10 have stochastic processing times. In tables 11–13 the throughputs estimated with both approximation approaches are compared with simulation results for different coefficients of variations at stations 1, 5 and 10 and for different X buffer sizes.

Buffer size	$X_{\text{simulated}}$	ADDX(adj)		GG1(CV0)	
		$X_{\text{approximated}}$	% Deviation	$X_{\text{approximated}}$	% Deviation
0	0.49525	0.53900	8.8%	0.33233	-32.9%
1	0.59342	0.59437	0.2%	0.44373	-25.2%
5	0.69002	0.71204	3.2%	0.65780	-4.7%
10	0.75576	0.78710	4.1%	0.75369	-0.3%
25	0.83451	0.85131	2.0%	0.84460	1.2%

Table 11. Results for $CV = 0.6$ at stations 1, 5 and 10.

Buffer size	$X_{\text{simulated}}$	ADDX(adj)		GG1(CV0)	
		$X_{\text{approximated}}$	% Deviation	$X_{\text{approximated}}$	% Deviation
0	0.44667	0.47660	6.7%	0.32133	-28.1%
1	0.53551	0.54993	2.7%	0.43473	-18.8%
5	0.67074	0.69082	3.0%	0.64222	-4.3%
10	0.73925	0.77841	5.3%	0.74077	0.2%
25	0.82823	0.84394	1.9%	0.83678	1.0%

Table 12. Results for $CV = 1$ at stations 1, 5 and 10.

Buffer size	$X_{\text{simulated}}$	ADDX(adj)		GG1(CV0)	
		$X_{\text{approximated}}$	% Deviation	$X_{\text{approximated}}$	% Deviation
0	0.37189	0.32368	-13.0%	0.28198	-24.2%
1	0.44611	0.42837	-4.0%	0.38110	-14.57%
5	0.60048	0.61188	1.9%	0.57947	-3.5%
10	0.68140	0.70539	3.5%	0.68435	0.4%
25	0.78493	0.80686	2.8%	0.79822	1.7%

Table 13. Results for $CV = 2$ at stations 1, 5 and 10.

The results show that the approach using the ADDX algorithm with adjusted parameters provides good approximations for systems with non-zero buffer sizes. The approach based on the $G/G/1/N$ model with $CV = 0$ for the deterministic stations provides good approximations only for system configurations with large buffer sizes. In these cases, the approach outperforms the ADDX(adj) approach.

2.4. Parallel stations

A further characteristic of real-life flow lines may be that, at several stations, there are parallel servers. In this case, the workpieces queue up in a *common buffer*. Downstream to the buffer there is a *switch*, which directs the workpieces to the next free machine. There is not much literature on this kind of system (see Magazine and Stecke 1996, Futamura 2000, Vidalis and Papadopoulos 2001. A recent paper with references is Patchong and Willaeyts (2001).

For the case of identical machines, which seems to be common in industry, Burman (1995) proposes, as a very simple heuristic, to multiply all parameters (in terms of rates) of a station by the number of parallel machines at that station. Through a number of simulation experiments with hypothetical systems consisting of three stages, Burman found out that this heuristic provides good approximations. See also Kalir and Arzi (1997) and Kim and Jung (2000).

Real-life system E

In order to find out how this approach works for longer flow production systems with real-life characteristics, consider the system E, comprising 14 stations with the data given in table 22 in the appendix. In table 14 the simulated and estimated throughputs are compared for different buffer configurations.

For each total number of buffers, we considered both a *balanced* buffer configuration and the *optimal* buffer allocation (solution to the dual problem, see section 3). Again, the approximation does not perform well for a total buffer size of zero. For systems with a *reasonable number of buffers* at the critical stations (which will be the result of a buffer optimization) however, the approximation is very good. Kim and Jung (2000) propose a simple data adjustment for the case of (parallel) stations performing the same operations but which are arranged in line.

2.5. Merging of material

In many companies, there are linear flow production subsystems that are combined such that a merging flow of material results. A typical example is a car body shop where parts are welded together to subassemblies, which in turn are assembled to subassemblies of higher order. This merging of material takes place in several

Total number of buffers	Buffer allocation	$X_{\text{estimated}}$	$X_{\text{simulated}}$	% Deviation
0	0	0.59221	0.70085	-15.5%
13	balanced (1 per station)	0.66125	0.72360	-8.6%
13	optimal	0.67579	0.70188	-3.7%
26	balanced (2 per station)	0.69411	0.73209	-5.2%
26	optimal	0.73182	0.74533	-1.8%
39	balanced (3 per station)	0.71455	0.73840	-3.2%
39	optimal	0.75576	0.76263	-0.9%
78	balanced (6 per station)	0.74658	0.75332	-0.9%
78	optimal	0.78280	0.78456	-0.2%
130	balanced (10 per station)	0.76508	0.76386	0.2%
130	optimal	0.79254	0.79006	0.3%

Table 14. Results for system E.

stages. In addition to ‘normal’ stations, which may be starved (blocked) by the upstream (downstream) portion of the system, flow production systems with a merging of material include assembly stations, which may suffer from starving and blocking caused by the complex interactions of the processes at all upstream and downstream stations.

In the case when a single assembly station, being the last station in the material flow, is fed by two production lines, the above-mentioned algorithms designed for linear material flow can be applied after a simple rearrangement of the system data. Due to the *reversibility property* (see Altioik 1997) it is possible to turn around one of the two linear segments of the system and connect it with the merging station. However, this transformation does not work for assembly stations with more than two merging lines and in cases where the assembly station is not the last station of the material flow under consideration. A number of algorithms have been proposed for the analysis of assembly/disassembly production systems (see Liu and Buzacott 1990, Jeong and Kim 1998, Helber 1999). Recently Gershwin and Burman (2000) proposed an algorithm applicable to non-homogeneous systems with deterministic processing times, which is a combination of the A/D algorithm described in Gershwin (1994) and the ADDX algorithm of Burman (1995). The authors presented results of a limited numerical experiment and concluded that their algorithm could be expected to provide good performance approximations.

We have tested this approach for several invented systems and found that the algorithm works very well. In addition, we considered the following portion of an existing production system with three production lines (A, B and C) feeding a single assembly station, which is followed by a ‘normal’ processing station, where the data, which are presented in table 28, are taken from a subsystem of an automobile plant.

In table 16, the simulated throughput is compared with the estimated throughput. Again, the quality of the performance approximations is very good.

3. Optimization of industrial flow production systems

In the literature, a large number of models have been presented that aim at finding the optimal system configuration under different assumptions. For a review of the literature we refer to Padopoulos *et al.* (1993), Singh and MacGregor Smith (1997), and the recent papers of Gershwin and Schor (2000),

Line A						
Station No.	1	2	3	4		
Processing time	227	236	228	216		
Availability	0.88%	0.88%	0.88%	0.875%		
MTTR	300	300	300	300		
Line B					Assembly	Processing
Station No.	5	6	7	8	13	14
Processing time	204	223	252	257	228	225
Availability	0.87%	0.80%	0.87%	0.96%	0.96%	0.87%
MTTR	300	300	300	300	300	300
Line C						
Station No.	9	10	11	12		
Processing time	227	227	219	242		
Availability	0.96%	0.88%	0.88%	0.875%		
MTTR	300	300	300	300		

Table 15. Assembly system.

Buffer size	$X_{estimated}$	$X_{simulated}$	% Deviation
1	0.0027199	0.0026766	-1.62%
2	0.0030768	0.0029905	-2.89%
3	0.0032214	0.0031634	-1.83%
4	0.0032818	0.0032678	-0.43%
5	0.0033207	0.0033117	-0.27%
6	0.0033496	0.0033431	-0.19%
7	0.0033672	0.0033785	0.33%
8	0.0033810	0.0033959	0.44%
9	0.0033921	0.0033997	0.22%
10	0.0034005	0.0034110	0.31%
11	0.0034077	0.0034173	0.28%
12	0.0034137	0.0034229	0.27%
13	0.0034187	0.0034149	-0.11%
14	0.0034230	0.0034250	0.06%
15	0.0034266	0.0034336	0.20%
20	0.0034383	0.0034390	0.02%
30	0.0034478	0.0034472	-0.02%
40	0.0034508	0.0034545	0.11%
50	0.0034517	0.0034460	-0.17%

Table 16. Results for the assembly system.

Papadopoulos and Vidalis (2001) and Helber (2001). The majority of research efforts have been devoted to the buffer allocation problem, whereas a smaller number of papers deal with the simultaneous buffer and workload allocation problem and an even smaller number of papers treat the simultaneous buffer, workload and server allocation problem.

This distribution of research effort is consistent with the occurrence of these problem types in industrial practice. In industry, it seems that the problem type

considered most frequently is the buffer allocation problem (finding the minimum number of buffers required to achieve a given throughput). Depending on the position of the planners in the factory planning workflow, sometimes a wider perspective is taken that also includes the determination of the cycle times for the stations, with an implicit consideration of the consequences for the allocation of the machines or robots among the stations.

3.1. Buffer optimization

As a single buffer unit may represent a considerable amount of investment, up to several thousand US dollars, and also requires scarce factory space, and as the planners normally must treat the throughput of the production system to be planned as a datum, they usually aim at minimizing the total number of buffers, N , with respect to a desired throughput level X_{\min} . In the literature this problem is called the ‘primal problem’ (Minimize $N = \sum b_m$ with respect to $X(b) \geq X_{\min}$, where b_m denotes the buffer size of station m). It is interrelated with the so-called ‘dual problem’ (Maximize $X(b)$ w.r.t. $\sum b_m = N$), which means to allocate a given total number of buffers such that the throughput of the system is maximized.

Figure 2 shows the development of the throughput of a flow production system as a function of total buffer size. The solid triangles mark the efficient frontier, i.e. the solutions to the different dual problems. The open triangles correspond to non-optimal buffer allocations. The vertical line marks the solution to the primal problem for a given throughput.

In the majority of solution approaches, the buffer optimization problem is treated as a combinatorial problem. Certainly in many cases the integrality requirement must be met by the final buffer configuration. However, in practice, there may also

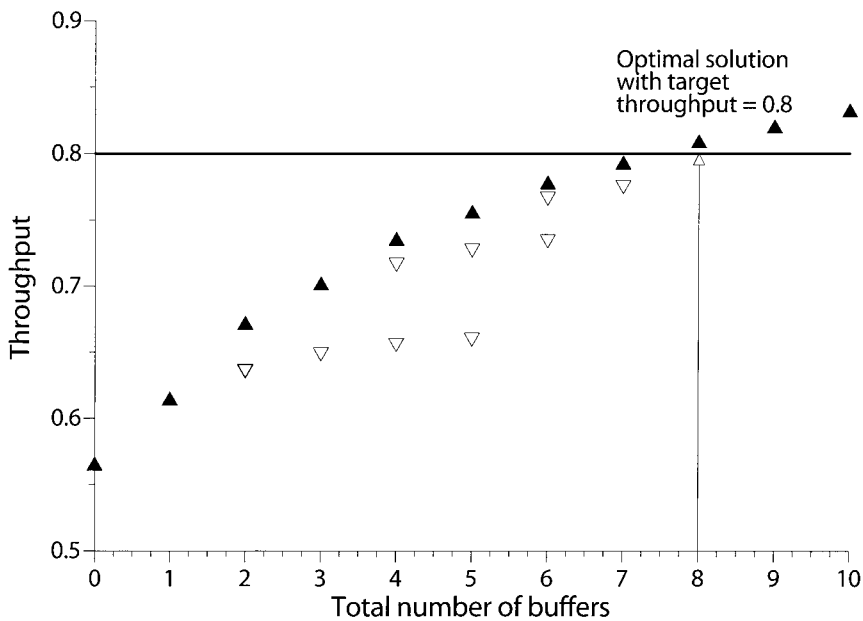


Figure 2. Throughput versus total buffer size.

be cases where buffer sizes are expressed in such a dimension that non-integer buffer sizes make some sense. A case where this is true is a filling line for ketchup bottles, where the buffer size is measured in units of 100 bottles. Physically there were buffers of size, say 2550 bottles, which would be 25.5 units. In addition, if the algorithm used for the performance evaluation of a given system configuration is also capable of analysing systems with non-integer buffer sizes, then a very efficient approach proposed by Schor (1995) (see also Gershwin and Schor 2000) can be used to solve the buffer optimization problem. As pointed out earlier, non-integer buffer sizes can be handled with all the performance evaluation algorithms that we have used in section 2. Therefore, Schor's optimization algorithm can be combined with all evaluation approaches to solve the buffer optimization problem efficiently and accurately under conditions of deterministic as well as stochastic processing times, even for deterministic processing time systems with merging material flow.

As there is no closed-form equation available that describes the development of the throughput versus total buffer size, Schor (1995) divides the problem into two parts, (a) the dual problem of finding the maximum throughput for a given total buffer size and (b) the primal problem of finding the minimum total number of buffers required to achieve a target throughput level. The dual problem is solved through a gradient-based search, while the primal problem is solved through successive linear approximations of the throughput curve, which is evaluated through the solution of several instances of the dual problem. After the algorithm has stopped, the buffers are rounded with a procedure that leaves the total buffer size unchanged.

Whenever we referred to optimal allocations for the systems considered in section 2, we used a variant of Schor's algorithm to find them. This algorithm can also easily be modified to consider the following additional practical requirements:

- Fixed buffer size at a station (defined by the planner due to technical constraints).
- Minimum buffer size at a station.
- Maximum buffer size at a station.

3.2. Buffer and cycle time optimization

If the planner has a wider perspective, then not only the buffer sizes, but also the number of servers at a station and the workload assigned to a station may be decision variables. A rather general model formulation originally proposed by Hillier and So (1995) is considered by Spinellis *et al.* (2000). They consider as variables the buffer sizes, the workloads and the numbers of workers with the objective function of maximizing the throughput of the flow production system. Each type of variable is restricted such that the total number of buffers, the total workload and the total number of servers is constant. This problem formulation is solved with the help of a simulated annealing procedure.

A similar model formulation, which is based on a practical case study at the German car manufacturer BMW, is described by Spieckermann *et al.* (2000). They consider a planning problem that arises in the early stages of the car body shop planning process. From the preceding planning phase a target throughput X_{\min} is given. The objective is to find the minimum of a function with positive coefficients for the buffer space used and with negative coefficients for cycle times at the stations. The planning problem reads as follows:

$$\text{Minimize } Z = f(b, w), \tag{14}$$

s.t.

$$X(b, w) \geq X_{\min}, \tag{15}$$

$$W_m^{\min} \leq w_m \leq W_m^{\max} \quad m = 1, 2, \dots, M, \tag{16}$$

$$b_m = \text{integer} \quad m = 1, 2, \dots, M, \tag{17}$$

$$w_m \geq 0 \quad m = 1, 2, \dots, M. \tag{18}$$

Decision variables are b_m (buffer size at station m) and w_m (cycle time at station m). The term $X(\cdot)$ denotes the throughput of the system, which depends on the buffer configuration and the workload allocation. The motivation for this problem formulation is based on the aggregate view of the car body shop, where each cell is considered as a black box. The structure of these black boxes is only fixed in subsequent planning stages. The maximum cycle time is computed on the basis of a target throughput per day and the minimum cycle time results from technical considerations with respect to the resources included in a cell.

Owing to the random nature of the failure and repair processes the target throughput will only be achieved if a large number of buffers is put between the stations, with the effect that no production is lost due to starvation or blockages. In a zero-buffer system configuration, a large amount of throughput would be lost due to starvation and blockages. This loss of throughput can be regained through an increase of the buffer sizes and/or a reduction of the cycle times.

In the practical planning environment, the cycle time of a station is considered as a decision variable that has an influence on the number of resources (robots) required at station m . For example, with a cycle time of 250 seconds, say, five welding robots may be necessary, whereas for a processing time of 220 one additional robot is needed. Obviously, in this case, a new allocation of the workload to the changed number of resources must be found. With robots this will be no problem. As a robot is costly, an important part of the objective function is to set the processing times as large as possible.

The trade-off between cycle time reduction (and consequently, the number of resources required) and the total number of buffers is illustrated in table 17, where a linear flow production system was considered.

Assume that the target throughput is 0.00356. For a given set of cycle times (100%) a total of 92 buffers are required. Now assume that the total number of

Total number of buffers	Cycle time							
	100%	99%	98%	97%	96%	95%	94%	93%
92	0.00356							
80	0.00353	0.00356						
70	0.00350	0.00353	0.00356					
60	0.00346	0.00349	0.00353	0.00356				
50	0.00341	0.00344	0.00347	0.00351	0.00354	0.00356		
40	0.00334	0.00337	0.00341	0.00344	0.00347	0.00350	0.00354	0.00356

Table 17. Throughput as a function of total buffer size and cycle time reduction.

buffers is reduced to 80, 70, 60, . . . , 40. With unmodified cycle times, the throughput would reduce to 0.00334, as shown in the '100%' -column of table 17. The remaining entries in the table show the throughput that would be achieved if the cycle times at all stations were simultaneously reduced to the percentage denoted in the column header *and* if the total number of buffers (given in the row header) were reallocated optimally among the stations (solution of the dual problem). Observe that the cycle time reduction required to regain the production loss caused by the reduction of buffer sizes depends on the number of buffers.

In the practical planning environment, Spieckermann *et al.* (2000) use a genetic algorithm that is combined with a simulation model of the body shop. The fitness function comprises the buffer sizes, deviations of the processing times from their upper bounds as well as the unbalancedness of the processing times. The genetic algorithm calls a simulation model for the performance evaluation of the aggregate car body shop. As far as we know, this approach has become standard in the planning practice of BMW and is also used by other European car manufacturers. However, the computation times required by the GA/simulation approach are extremely long, usually several days, and the solution found in this way can probably be further improved.

As an alternative, the speed of the optimization could be increased along two different lines. Instead of the genetic algorithm, a specialized optimization algorithm, at least for a subset of the problem variables could be used. Instead of simulation, an analytical algorithm for the evaluation of the system could be applied.

The availability of an algorithm for the solution of the buffer optimization problem allows the definition of the simple procedure in figure 3 to solve the problem of simultaneous buffer and cycle time optimization.

The reduction of buffer sizes will reduce the value of the objective function. The reduction in cycle times will increase the objective value due to the additional resources required, in addition to several side effects, such as increased risk of failure with the high utilization of resources. Therefore a multi-objective decision making approach is adequate, where the components of the objective function are weighted.

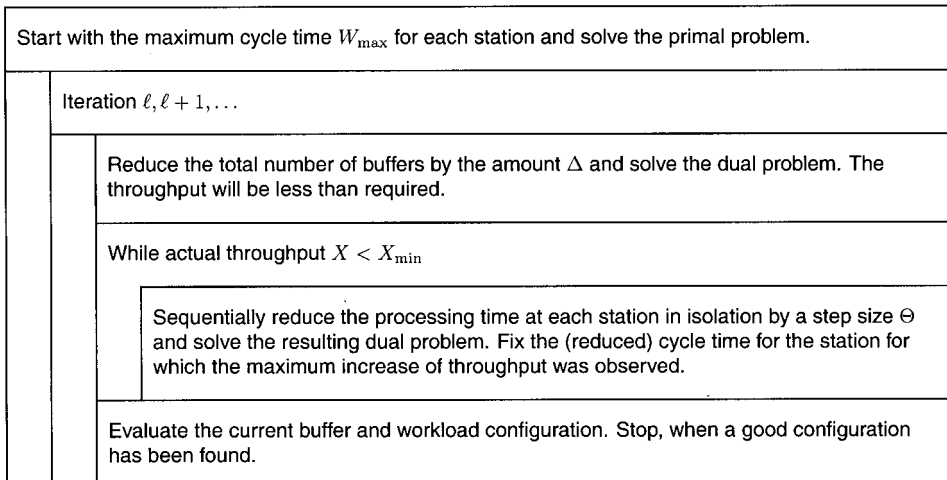


Figure 3. Procedure for simultaneous buffer and workload optimization.

4. Conclusions

We have shown that a wide variety of real-life flow production systems can be evaluated and optimized with basically three evaluation algorithms and a single optimization approach. For these algorithms to become standard tools in the system design process, several requirements must be met:

- The algorithms must be implemented with a user-friendly interface which speaks the same language as the planners.
- The planning tool must be integrated into the planners standard workflow, and it should have an interface to the databases the planners normally use.

With these requirements in mind, a software system called ‘FlowEval’ (see <http://www.pom-consult.de/indexe.html>) has been developed, which implements the algorithms discussed in this paper. The algorithms are available to the planner on his or her desktop in an interactive mode as well as via an interface to the standard overall planning environment that the planner uses. Through the use of this optimization and performance evaluation tool, we observed that the planners now provide better planning results in less time.

Acknowledgement

The author would like to thank the anonymous reviewers who provided useful suggestions that helped improve the paper’s presentation.

Appendix

Station	Buffer	Proc. time (min)	Availability	Repair time (min)	$\rho_m = (\mu_m r_m) / (r_m + p_m)$
1	–	0.470	0.978	49.00	2.081
2	14	0.470	0.980	41.00	2.085
3	9	0.493	0.984	55.00	1.996
4	9	0.488	0.993	28.00	2.035
5	16	0.463	0.984	47.00	2.125
6	28	0.463	0.977	28.00	2.110
7	23	0.498	0.987	68.00	1.982
8	27	0.467	0.970	35.00	2.075
9	8	0.473	0.967	43.00	2.044
10	12	0.487	0.996	33.00	2.045
11	28	0.498	0.997	36.00	2.002
12	6	0.478	0.996	31.00	2.082
13	9	0.488	0.970	60.00	1.988
14	12	0.488	0.988	39.00	2.025
15	8	0.482	0.975	38.00	2.023
16	24	0.430	0.989	48.00	2.281
17	29	0.473	0.981	47.00	2.097
18	10	0.500	0.992	55.00	1.948
19	13	0.467	0.974	43.00	2.071

Table 18. System A.

Station	Buffer	Proc. time (s)	Availability	Repair time (s)	$\rho_m = \frac{(\mu_m r_m)}{(r_m + p_m)}$
1	–	5.100	0.983	29.69	0.193
2	9	5.100	0.990	35.89	0.194
3	12	5.300	0.975	34.52	0.184
4	12	5.300	0.935	44.40	0.176
5	9	3.800	0.992	27.27	0.261
6	6	5.200	0.961	30.40	0.185
7	5	5.200	0.943	54.97	0.181
8	6	5.000	0.968	33.98	0.194
9	10	3.500	1.000	–	0.286
10	9	5.100	0.941	27.75	0.185
11	30	5.500	0.964	47.53	0.175
12	6	5.300	0.977	45.23	0.184
13	9	4.700	0.990	61.28	0.211
14	9	3.500	1.000	–	0.286
15	6	5.000	0.984	35.43	0.197
16	7	5.100	0.989	33.19	0.194
17	8	4.100	0.992	214.69	0.242
18	6	4.900	0.979	51.07	0.200
19	6	4.900	0.979	51.64	0.200
20	7	4.900	0.894	75.83	0.182
21	31	6.100	0.830	45.39	0.136
22	128	5.100	0.958	229.85	0.188
23	128	5.100	0.958	229.85	0.188

Table 19. System B.

Station	Buffer	Proc. time (s)	Availability	Repair time (s)	$\rho_m = \frac{(\mu_m r_m)}{(r_m + p_m)}$
1	–	233.000	0.890	300.00	0.003820
2	21	233.000	0.880	300.00	0.003777
3	31	234.000	0.875	300.00	0.003739
4	20	216.000	0.870	300.00	0.004028
5	24	212.000	0.800	300.00	0.003774
6	19	220.000	0.870	300.00	0.003955
7	4	255.000	0.960	300.00	0.003765
8	12	257.000	0.960	300.00	0.003735

Table 20. System C.

Station	Buffer	Proc. time (s)	Availability	$\rho_m = (\mu_m r_m)/(r_m + p_m)$
1	—	25.000	0.980	0.039
2	8	34.00	0.970	0.029
3	25	33.500	0.990	0.030
4	1	35.500	0.980	0.028
5	2	22.000	0.990	0.045
6	16	36.00	0.970	0.027
7	7	22.000	0.970	0.044
8	32	34.000	0.970	0.029
9	1	26.000	0.970	0.037
10	8	30.000	0.970	0.032
11	20	26.500	0.980	0.037
12	9	33.000	0.970	0.029
13	21	29.500	0.970	0.033
14	16	35.000	0.990	0.028

Table 21. System D.

Station	Servers	Proc. time (min)	Availability	Repair time (min)	$\rho_m = (\mu_m r_m)/(r_m + p_m)$
1	1	1.1	0.930	9.70	0.845
2	2	1.1	0.930	9.70	0.845
3	1	1.17	0.930	12.00	0.795
4	1	1.07	0.930	12.00	0.869
5	2	1.07	0.930	12.00	0.869
6	2	1.1	0.980	3.20	0.891
7	2	0.85	0.850	19.00	1.000
8	2	0.85	0.850	19.00	1.000
9	2	0.85	0.850	19.00	1.000
10	2	0.85	0.850	19.00	1.000
11	3	1.063	0.980	3.20	0.922
12	2	1.1	0.930	12.00	0.845
13	2	1.23	0.980	4.50	0.797
14	2	1.18	0.950	12.00	0.805

Table 22. System E.

References

- ALTIOK, T., 1997, *Performance Analysis of Manufacturing Systems* (New York: Springer).
- ALTIOK, T. and STIDHAM, S., 1983, The allocation of interstage buffer capacities in production lines. *IIE Transactions*, **15**, 292–299.
- ASKIN, R. and STANDRIDGE, C., 1993, *Modeling and Analysis of Manufacturing Systems* (New York: Wiley).
- BAKER, K. R., 1993, Tightly-coupled production systems: models, analysis, and insights. *Journal of Manufacturing Systems*, **11**(6), 385–400.
- BAKER, K. R., POWELL, S. G. and PYKE, D. F., 1994, A predictive model for the throughput of unbalanced, unbuffered three-station serial lines. *IIE Transactions*, **26**, 62–71.
- BLUMENFELD, D. E., 1990, A simple formula for estimating throughput of serial production lines with variable processing times and limited buffer capacity. *International Journal of Production Research*, **28**, 1163–1182.
- BURMAN, M. H., 1995, New results in flow line analysis. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

- BURMAN, M., GERSHWIN, S. B. and SUYEMATSU, C., 1998, Hewlett-Packard uses operations research to improve the design of a printer production line. *Interfaces*, **28**(1), 24–36.
- BUZACOTT, J. A. and SHANTHIKUMAR, J. G., 1993, *Stochastic Models of Manufacturing Systems* (Englewood Cliffs: Prentice Hall).
- BUZACOTT, J., LIU, X.-G. and SHANTHIKUMAR, G., 1995, Multistage flow line analysis with the stopped arrival queue model. *IIE Transactions*, **27**, 444–455.
- CONWAY, R., MAXWELL, W., McCLAIN, J. and THOMAS, L., 1988, The role of work-in-process inventory in serial production lines. *Operations Research*, **36**, 229–241.
- DALLERY, Y. and GERSHWIN, S. B., 1992, Manufacturing flow line systems: a review of models and analytical results. *Queueing Systems Theory and Applications*, **12**, 3–94.
- ENGINARLAR, E., LI, J., MEERKOV, S. and ZHANG, R., 2002, Buffer capacity for accommodating machine downtime in serial production lines. *International Journal of Production Research*, **40**, 601–624.
- FUTAMURA, K., 2000, The multiple server effect: optimal allocation of servers to stations with different service-time distributions in tandem queueing networks. *Annals of Operations Research*, **93**, 71–90.
- GAVER, D. P., 1962, A waiting line with interrupted service, including priorities. *Journal of the Royal Statistical Society*, **24**(1), 73–90.
- GERSHWIN, S. B., 1987, An efficient decomposition algorithm for the approximate evaluation of tandem queues with finite storage space and blocking. *Operations Research*, **35**, 291–305.
- GERSHWIN, S., 2000, Design and operation of manufacturing systems: the control-point policy. *IIE Transactions*, **32**, 891–906.
- GERSHWIN, S. B., 1994, *Manufacturing Systems Engineering* (Englewood Cliffs, NJ: Prentice Hall).
- GERSHWIN, S. B. and BURMAN, M. H., 2000, A decomposition method for analyzing inhomogeneous Assembly/Dissassembly systems. *Annals of Operations Research*, **93**, 91–115.
- GERSHWIN, S. B. and SCHOR, J. E., 2000, Efficient algorithms for buffer space allocation. *Annals of Operations Research*, **93**, 117–144.
- HELBER, S., 1999, *Performance Analysis of Flow Lines with Non-Linear Flow of Material*. Lecture Notes in Economics and Mathematical Systems (Berlin: Springer).
- HELBER, S., 2001, Cash-flow-oriented buffer allocation in stochastic flow lines. *International Journal of Production Research*, **39**, 3061–3083.
- HILLIER, M. S., 2000, Characterizing the optimal allocation of storage space in production line systems with variable processing times. *IIE Transactions*, **32**, 1–8.
- HILLIER, F. and SO, K., 1995, On the optimal design of tandem queueing systems with finite buffers. *Queueing Systems*, **21**, 245–266.
- HILLIER, F. S. and SO, K. C., 1996, On the simultaneous optimization of server and work allocations in production line systems with variable processing times. *Operations Research*, **44**, 435–443.
- HILLIER, F. S., SO, K. C. and BOLING, R. W., 1993, Notes: toward characterizing the optimum allocation of storage space in production line systems with variable processing times. *Management Science*, **39**, 126–133.
- INMAN, R. R., 1999, Empirical evaluation of exponential and independence assumptions in queueing model of manufacturing systems. *Production and Operations Management*, **8**(4), 409–432.
- JEONG, K.-C. and KIM, Y.-D., 1998, Performance analysis of assembly/dissassembly systems with unreliable machines and random processing times. *IIE Transactions*, **30**, 41–53.
- KALIR, A. and ARZI, Y., 1997, Automated production line design with flexible unreliable machines for profit maximization. *International Journal of Production Research*, **35**, 1651–1664.
- KIM, D. and JUNG, B., 2000, The equivalence of duplicate automated serial workstations and two-workstation tandem systems and its use in serial production line analysis. *International Journal of Production Research*, **38**, 1525–1538.
- LIU, C.-M. and LIN, C.-L., 1996, Predictive models for performance evaluation of serial production lines. *International Journal of Production Research*, **34**, 1279–1291.
- LIU, X.-G. and BUZACOTT, J. A., 1990, Approximate models of assembly systems with finite inventory banks. *European Journal of Operational Research*, **45**, 143–154.

- MAGAZINE, M. and STECKE, K., 1996, Throughput for production lines with serial work stations and parallel service facilities. *Performance Evaluation*, **25**, 211–232.
- PAPADOPOULOS, H. T. and HEAVEY, C., 1996, Queuing theory in manufacturing systems analysis and design: A classification of models for production and transfer lines. *European Journal of Operational Research*, **92**, 1–27.
- PAPADOPOULOS, H. and VIDALIS, M., 2001, Minimizing WIP inventory in reliable production lines. *International Journal of Production Economics*, **20**, 185–197.
- PAPADOPOULOS, H. T., HEAVEY, C. and BROWNE, J., 1993, *Queueing Theory in Manufacturing Systems Analysis and Design* (London: Chapman & Hall).
- PATCHONG, A. and WILLAEYS, D., 2001, Modeling and analysis of an unreliable flow line composed of parallel-machine stages. *IIE Transactions*, **33**, 559–568.
- POWELL, S. G. and PYKE, D. F., 1996, Allocation of buffers to serial production lines with bottlenecks. *IIE Transactions*, **28**, 18–29.
- POWELL, S. G. and PYKE, D. F., 1998, Buffering unbalanced assembly systems. *IIE Transactions*, **30**, 55–65.
- SCHÖNIGER, J. and SPINGLER, J., 1989, Planung der Montageanlage. *Technica*, **14**, 27–32 (in German).
- SCHOR, J. E., 1995, Efficient algorithms for buffer allocation. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA.
- SINGH, A. and MACGREGOR SMITH, J., 1997, Buffer allocation for an integer nonlinear network design problem. *Computers & Operations Research*, **24**, 453–472.
- SO, K. S., 1997, Optimal buffer allocation strategy for minimizing work-in-process inventory in unpaced production lines. *IIE Transactions*, **29**, 81–88.
- SPIECKERMANN, S., GUTENSWAGER, K., HEINZEL, H. and VOß, S., 2000, Simulation-based optimization in the automotive industry—a case study on body shop design. *Simulation*, **75**(5), 276–286.
- SPINELLIS, D., PAPADOPOULOS, C. and MACGREGOR SMITH, J., 2000, Large production line optimization using simulated annealing. *International Journal of Production Research*, **38**, 509–541.
- TEMPELMEIER, H. and BÜRGER, M., 2001, Performance evaluation of unbalanced flow lines with general distributed processing times, failures and imperfect production. *IIE Transactions*, **33**, 293–302.
- VIDALIS, M. and PAPADOPOULOS, H., 2001, A recursive algorithm for generating the transition matrices of multistation multiserver exponential queueing networks. *Computers & Operations Research*, **28**, 853–883.
- VISVANADHAM, N. and NARAHA, Y., 1992, *Performance Modeling of Automated Manufacturing Systems* (Englewood Cliffs, NJ: Prentice-Hall).