Proceedings of the 1986 Winter Simulation Conference
J. Wilson, J. Henriksen, S. Roberts (eds.)

ISSUES IN THE DESIGN AND MODELING OF AUTOMATIC ASSEMBLY SYSTEMS

by

J.L. Sanders, Ph.D.
Dept. of Industrial Engineering
University of Wisconsin - Madison
1513 University Ave.
Madison, WI 53706

ABSTRACT

In this paper we explore three research areas in the development of design and analysis tools required for asynchronous or free transport automatic assembly systems. These areas include (1) the use of simulation analysis for the investigation of subsystems for a special class of automatic assembly stations - the "tunnel-gated" or "High-Lift" station, (2) the development of a theory of analysis for assembly systems based on the theory of closed networks of GI/G/1 queues and (3) methods for the optimization of assembly systems.

1. INTRODUCTION

Since the introduction of assembly line concepts into automobile assembly in the early 1900's, the organization of the production flow shop has become increasingly automated and sophisticated. In the earliest forms of assembly systems there was little automation except a tow line that pulled the undercarriage of the item to be assembled (the automobile) along at a constant rate and forced the manual assembly tasks to be executed at a predetermined rate. By the 1940's, one saw the emergence of more automated forms of assembly including totally automated assembly machines for the assembly of such items as artillery shell fuses. These systems were mechanical marvels but were totally specialized in the sense that they could produce only one type of assembly per machine albeit at a very high rate. More recently one sees an increasing sophistication in both the variety and technology of assembly systems concepts.

Boothroyd, Poli, and Murch (1982) classify assembly systems or machines into two primary classes. These include synchronous and asynchronous systems. By their very nature the synchronous systems are totally automated and often are "hard-automated" i.e. capable of assembling only a single product type. These systems are currently used in both mechanical and electrical assembly work with outstanding success in tasks which involve components with low mass and high production volume requirements. Assembly machines currently exist in electronic circuit board assembly which permit the exact placement of up to 600,000 surface mounted electronic devices per hour. The major disadvantages of synchronous systems are (1) that if a single station in the machine jams or malfunctions, however briefly, the entire system is forced to a halt and (2) especially in the area of mechanical assembly, most of these systems are not flexible or reprogrammable. A product change usually requires a complete redesign and modification of the machine.

Within the general category of asynchronous machines, we may distinguish several subtypes according to the types of assembly stations that are installed and also according to the type of transfer mechanism used. The individual assembly stations may consist of a "hard-automated" machine, programmable automated station or human assembler. The transfer mechanism on the other hand may be of the type where the transfer occurs at a constant rate and the work is removed from the transfer device to perform the assembly task as in the case of circulating conveyer baskets on manual assembly systems. A second type of transfer mechanism is often used which is referred to a "free transfer" or "power free" and occurs on assembly machines where the assembly circulates on a pallet to be queued up at an assembly station and which waits for the assembly station to finish all of the assemblies ahead of the last assembly in the waiting line. This type of transfer permits stations to work at randomly varying rates and permits individual stations to be jammed or broken for brief periods without requiring the complete shutdown of the remainder of the line.

While asynchronous machines seldom approach the very high production rates of the synchronous machines, still this class of assembly systems can often reach production rates of 100 to 1000 assembliers per hour on assemblies with considerable mass and complexity. Machines of this type with from 20 to 100 assembly stations are relatively common today.

2. STUDY OF THE PERFORMANCE OF PARALLEL PROCESSING CONCEPTS IN POWER-AND-FREE AUTOMATIC ASSEMBLY SYSTEMS

Traditionally the design of AAS has centered on configurations consisting of a circular or oval shaped transfer line with assembly stations placed at intervals around the oval, with each of the stations operating asynchronously. As the applications of such systems has expanded outside the automotive industry it has been found that various forms of parallel processing are desirable or required.

Multiple In-Line Station vs. High-Lift Stations

It is clear that in some sense no AAS can run faster than its slowest station. As a consequence, if in the design process, the design engineer finds one or two stations with much longer cycle times than the remainder of the stations he is faced with the problem of how to speed up these slow units. He may either try to decompose the original assembly tasks into shorter subtasks that can be broken into two or more new stations with shorter cycle times or he can install two or more identical stations in series to form a type of parallel processing system. If the parts feeding systems for these new stations work with few or no problems and the input parts quality is high and if the stations themselves are extremely reliable then this arrangement may work well. However, whenever the first station in this series arrangement is busy or down then all of the downstream stations are blocked from receiving additional fixtures and eventually will go into a starved down condition. If this is a problem, a highlift or tunnel gated station has the considerable

advantage that it permits the same parallelism that is inherent in the multiple in-line station concept, but the work on the highlift stations is done above the transfer line thus permitting fixtures to flow beneath the station and to reach downstream stations. Consequently busy or down time on the highlift station need not shut down the remainder of the line.

In fact several tunnel-gated stations can be installed in series to accommodate assembly tasks that have cycle times that are much longer than the average cycle time for the rest of the line.

There are a number of analysis tasks that accompany the insertion of a series of tunnel-gated stations in an assembly system. First of all, one would like to be able to predict the productivity of a series of tunnel-gated stations as a function of the (common) cycle time, the mean jam rate, the distribution of the time required to clear a jam, the transport speed, the total buffer space allocated to the set of stations and the relative spacing of the stations within the total buffer capacity. In addition to just being able to predict the throughput, one would like to be able to optimize the performance of the sector by determining the optimal buffer requirements and the optimal buffer spacing for each of the stations within the total buffer capacity.

In an extensive series of simulation experiments, Leung and Sanders (1986) have found that the behavior of a series of tunnel-gated stations is quite different than a similar series of standard assembly stations. The productivity of a series of standard independent stations is governed primarily by reductions brought on by blocking and starvation effects induced by the jam rate and clear time effects. On the other hand, a series of tunnel-gated stations exhibit a type of synchronization behavior which is not seen in a series of standard stations.

This synchronization in turn implies some unusual properties for buffer allocation. Leung and Sanders have found the optimal placement for tunnel-gated stations in a combined buffer space with one buffer space between the previous (standard) station and the group of tunnel-gated stations. Perhaps the most interesting result is that the productivity of tunnel-gated section is not a monotone function of the total buffer space allocated to the section. This result can be seen in the plots of figures 1 and 2. Synchronization effects make certain buffer sizes less productive than those that are smaller as well as those that are larger. In addition they have developed some simple mathematical models that predict these effects and can be used by the machine designer to predict the productivity of the tunnel-gated sectors of the machine

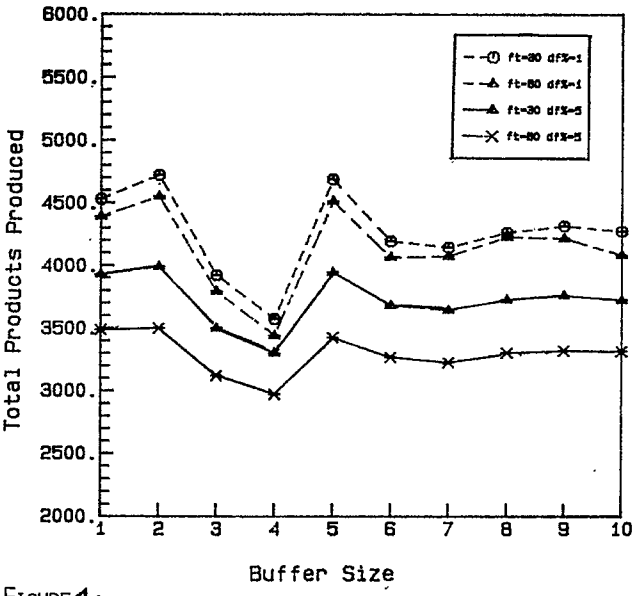

FIGURE 1:
Open-loop with 2 high-lift stations (Upstream)

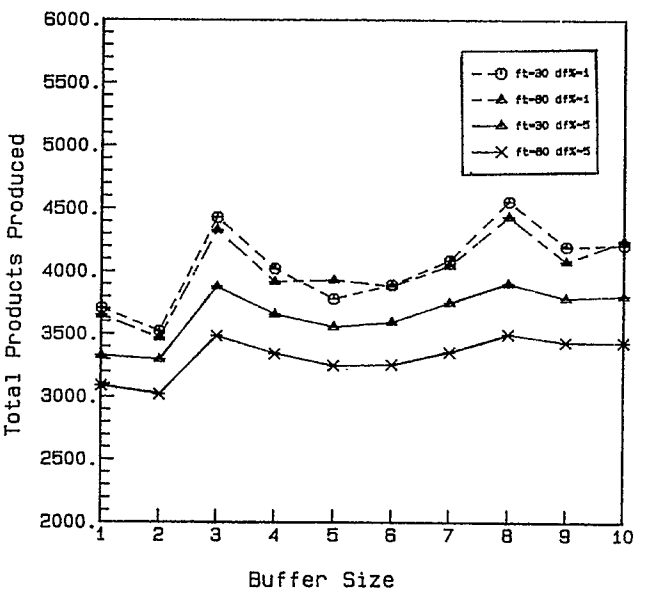

FIGURE 2:
Open-loop with 3 high-lift stations (Upstream)

3. MODELING OF ASYNCHRONOUS ASSEMBLY SYSTEMS

Modern manufacturing systems pose a number of very difficult problems for mathematical modeling. Major classes of systems that are in use that present special problems are flexible assembly systems (FAS) and flexible manufacturing systems (FMS). These systems consist of a closed network of assembly or fabrication stations which are linked by an automated transport system. Each processing stations is subject to random jams and may take a random time to clear. The transport systems introduces delays between the time a part is finished by one station and the time it is available for processing at the next station. This property plus the problems introduced by the finite buffers, finite number of pallets and the randomness of the individual stations makes the modeling process especially challenging. In addition multiple types of parts may be circulating in the system at any given time and each part type will have its own characteristic cycle time on each station.

Our goal here is the development of a set of analysis tools which can be used to accurately predict the productivity of a proposed or existing assembly machine. Historically discrete event digital simulation has been the only tool available for such

analysis. While this class of techniques continues to be a valuable set of tools it has some major disadvantages. Development costs for simulators are substantial, modification of existing models is often a time consuming and expensive task and analysis of existing systems is often slow and expensive. Evaluating tens or hundreds of alternative machine configurations, as one is likely to want to do in design analysis, may become impossible due to the quantity of computer time required to perform the analysis.

Consequently while we want to keep discrete event simulation (DES) in the analysis tool kit, we really need a set of methods that permit very rapid and inexpensive (although perhaps necessarily approximate) analysis of alternative designs. What we find is that methods developed for the performance analysis of computer systems and for the analysis of networks of queues have excellent potential to provide just the sort of fast approximate analysis that we desire in many cases. These tools including DES we will refer to as general purpose analysis tools. These tools on the other hand require a number of assumptions that are violated in certain specific problem areas. In these contexts we need to develop special analytic tools. A number of these special areas are enumerated in the section on special purpose tools below.

Analysis Tools

A. General Purpose Tools

The most promising general purpose analysis tools available for the study of asychronous assembly machines are (in addition to digital simulation) the MVAQ analysis (Suri and Hildebrant 1984) and a new class of models based on second order approximations to networks of GI/G/1 queues (Whitt 1985).

a. MVAQ

The MVAQ (Mean Value Analysis of Queues) is a method for the analysis of queueing networks based on exponential service time assumptions. It uses a variety of mean value techniques from the queueing literature to derive the mean number of each queue in the system, the main waiting time for customers at each facility and it can be used in our problem to derive the mean throughput for the systems. It is quite simple to use but it does have the drawback that it is based on exponential service time assumptions and hence seriously overestimates the variance of the service time for most assembly system problems. In addition, it does not provide a method of estimating the second moments of the important parameters of the system.

b. Second-Moment Models

One of the difficulties with the MVAQ analysis is that it assumes that each station in the system has exponential service time or more exactly it assumes that the coefficient of variation service time is 1.0. This creates problems for the analysis of assembly systems since certain stations have essentially zero jam rates and consequently have a coefficient of variation near zero where others have significant jam rates and hence significant coefficients of variation. Clearly to provide an analysis that takes the service time variation of individual stations into account in the analysis of the machine it is necessary to introduce the possibility of differing station variances into the analysis. Just such an analytic method was developed by Whitt (1985) for the analysis of telecommunication systems. He has developed a second order approximation for the analysis of station delays for GI/G/1 queues. He then proposes an approximate analysis for networks of GI/G/1 queues. Recently, Kamath and Sanders (1986) have applied these methods to the analysis of assembly machines. The results are given below.

These approximations assume knowledge of mean station cycle time and the variance of the station cycle time. In open queueing network examples it is also assumed that each external customer arrival process is a renewal process and that the mean and variance of each the interarrival processes is known. All third order and higher moments of the arrival and service distributions are ignored for the sake of simplicity in the approximations. In the case of an assembly machine there are no external arrival processes since the arrivals to each station come form the discharge of the preceding station in the line.

We shall briefly review some relevant approximations in a stable GI/G/1 queue. Let λ represent the mean arrival time and τ the mean service time. The utilization rate is $\rho = \lambda\tau < 1$. We use the squared coefficients of variation (variance divided by the square of the mean) c_a^2 and c_s^2 to approximately characterize the variability of the general interarrival-time and service-time distributions. Let EW denote the expected equilibrium waiting time (delay excluding service time). The departure process is approximated by a renewal process partially characterized by the first two moments of the renewal interval. The mean of the renewal interval in the approximating renewal process is just the mean of the interarrival time, so that the departure rate equals the arrival rate. The squared coefficient of variation of the interdeparture time, c_d^2, is given by

$$c_d^2 = c_a^2 + 2\rho c_s^2 - 2(1-\rho)EW/\lambda. \qquad (1)$$

This expression is originally due to Marshall (1968); also see Whitt (1985).

Next, we proceed to analyze closed tandem networks of queues.

Consider symmetric closed tandem networks of GI/G/1 queues with unlimited waiting room at each queue. The service discipline is FIFO (first-in first-out). In particular, we assume a closed tandem network with 'M' service centers and a fixed number, 'N' of customers who circulate in the system. In the following i = 1,2,...,M. We adopt the following notation.

- τ : mean service time at a service center (same at each service center)
- c_s : coefficient of variation of service time at a service center (same at each service center)
- λ : throughput (equilibrium) of the network (equilibrium mean arrival rate at each service station)
- c_{ai} : coefficient of variation of inter-arrival time at service center i
- c_{di} : coefficient of variation of inter-departure time at service center i
- ρ : utilization of a service center (all centers have the same utilization)
- EW_i : expected equilibrium waiting time (delay excluding service time) at service center i

It is easy to see that the following equalities hold because of the symmetric nature of the tandem network:

$$c_{a1}^2 = c_{a2}^2 = \ldots = c_{aM}^2 = c_a^2 \quad (2)$$

$$c_{d1}^2 = c_{d2}^2 = \ldots = c_{dM}^2 = c_d^2, \text{ and} \quad (3)$$

$$EW_1 = EW_2 = \ldots = EW_m = EW.$$

In closed tandem queueing networks since each arrival process is the departure process from the previous queue and because of equalities (2) and (3), we have

$$c_a^2 = c_d^2 \quad (4)$$

Focusing on a particular service center and using formula (1) and equality (4) we get

$$EW = \mathcal{T}\rho c_s^2/(1-\rho). \quad (5)$$

Using Little's formula and formula (5) for the expected equilibrium waiting time in queue we have

$$N = \lambda\{M(\mathcal{T}+\mathcal{T}\rho c_s^2/(1-\rho))\} \quad (6)$$

Rearranging equation (6) we get:

Case (a) $c_s^2 < 1$

$$\rho^2(M-Mc_s^2) - \rho(M+N) + N = 0.$$

It can be shown that the value of ρ from the solutions to the above quadratic equation is

$$\rho = [(M+N) - \sqrt{\{(M+N)^2 - 4(M-Mc_s^2)N\}}] \;/\; 2(M-Mc_s^2) \quad (7)$$

Case (b) $c_s^2 = 1$

$$\rho(M+N) - N = 0.$$

or

$$\rho = N/(M+N).$$

Case (c) $c_s^2 > 1$

$$\rho^2(Mc_s^2-M) + \rho(M+N) - N = 0.$$

From the solutions to the above quadratic equation it can be easily shown that the value of ρ is

$$\rho = [\sqrt{\{(M+N)^2 + 4(Mc_s^2-M)N\}} - (M+N)] \;/\; 2(Mc_s^2-M) \quad (8)$$

Noting that equations (7) and (8) are the same, we summarize the results as follows:

Service Center Utilization Rate

Case (a) $c_s^2 = 1$

$$\rho = N/(N+M)$$

Case (b) $c_s^2 \neq 1$

$$\rho = [\sqrt{\{(M+N)^2 + 4(Mc_s^2-M)N\}} - (M+N)] \;/\; 2(Mc_s^2-M)$$

Throughput Rate

$$\lambda = \rho/\mathcal{T}$$

Next, we apply the above formulas to predict the performance of completely balanced AASs.

Consider an asynchronous AAS with M assembly stations and a fixed number, N, of pallets circulating in the AAS. The AAS is totally balanced, that is, all assembly stations have identical characteristics. Let us briefly analyze the dynamics of a typical assembly station. The time to finish an assembly operation is fixed and is usually known as cycle time. Let s represent the cycle time. Whenever a station receives a defective assembly we say that the station is jammed. The occurrence of jams is infrequent, but these events have a significant influence on the performance of an AAs. The percentage of total assemblies processed by an assembly station that are defective is called the percent defective and is denoted by p. It takes a random amount of time for the station to be cleared of a defective assembly. Let r and v represent the mean and variance respectively, of the clear time.

If D (mean = s, variance = 0) represents the deterministic component and R (mean = r, variance = v) the stochastic component of the total processing time T of a station then we have

$$T = D + \alpha R,$$

where, α is a Bernoulli random variable with parameter p/100 (percent defective/100).

The random variables and R are stochastically independent. Hence, the mean and variance of the random variable T can be easily derived and are

$$E[T] = E[D] + E[\alpha]E[R] \quad \text{or}$$

$$E[T] = s + (p/100)r$$

$$Var[T] = Var[D] + Var[\alpha R] \quad \text{or}$$

$$Var[T] = 0 + \{Var[\alpha]Var[R] + Var[\alpha](E[R])^2 + (E[\alpha])^2 Var[R]\} \quad \text{or}$$

$$Var[T] = (p/100)\{v + r^2(1-(p/100))\}$$

For our example we assume that the time to clear an assembly station of defective assembly is geometrically distributed with parameter equal to 1/r. It is worth noting that the clear times can follow any probability distribution provided we are able to estimate the required parameters. In fact, the mean and variance of the clear times estimated for operating data would suffice. For the special case of geometric clear times, since $v = r(r-1)$ we have

$$Var[T] = (p/100)r\{r(2 - (p/100)) - 1\}.$$

Results

The results of comparisons of the class of models to both simulation and MVAQ models is shown in Table 1 for 10 station machines and in Table 2 for 100 station examples for varying pallet loading, jam rates and clear time assumptions.

TABLE 1. 10-Station AAS Example for the Balanced Case

Mean Clear Time	% Defective	Pallets	SIMULATION		MVAQ **			RENA (Renewal Approximations)		
			Station Utilization	Thruput Rate	Station Utilization	Thruput Rate	% Error in Thruput	Station Utilization	Thruput Rate	% Error in Thruput
6	0.5	10	0.963 ± 0.010	0.1596 ± 0.0016	0.526	0.0873	- 45.30	0.913	0.1514	- 5.14
6	0.5	20	0.993 ± 0.005	0.1646 ± 0.0008	0.690	0.1144	- 30 50	0.991	0.1644	- 0.12
6	3.0	10	0.861 ± 0.006	0.1393 ± 0.0014	0.526	0.0852	- 38.84	0.816	0.1320	- 5.24
6	3.0	20	0.963 ± 0 007	0.1557 ± 0.0013	0.690	0.1116	- 28.32	0.955	0.1546	- 0.71
36	0.5	10	0.814 ± 0 042	0.1315 ± 0.0070	0.526	0.0852	- 35.21	0 634	0.1026	- 21.98
36	0.5	20	0.882 ± 0 050	0.1425 ± 0.0082	0.690	0.1116	- 21.68	0.814	0.1317	- 7.58
36	3.0	10	0.538 ± 0 038	0.0754 ± 0.0048	0.526	0.0743	- 1.46	0.449	0.0634	- 15.92
36	3.0	20	0.668 ± 0 024	0.0935 ± 0.0026	0.69	0.0974	4.17	0.605	0.0854	- 8.66

** exact solution for the symmetric exponential cyclic network

TABLE 2. 100-Station AAS Example for the Balanced Case

Mean Clear Time	% Defective	Pallets	SIMULATION		MVAQ **			RENA (Renewal Approximations)		
			Station Utilization	Thruput Rate	Station Utilization	Thruput Rate	% Error in Thruput	Station Utilization	Thruput Rate	% Error in Thruput
6	0.5	100	0.907 ± 0.005	0.1504 ± 0.0009	0.503	0.0833	- 44.61	0.913	0.1514	0.66
6	0.5	200	0.984 ± 0.007	0.1632 ± 0 0010	0.669	0.1109	- 32.05	0.991	0.1644	0.74
6	3.0	100	0.806 ± 0.010	0.1305 ± 0.0015	0.503	0.0813	- 37.70	0 816	0.1320	1.15
6	3.0	200	0.949 ± 0.010	0.1536 ± 0.0010	0.669	0.1082	- 29.56	0.955	0.1536	0.65
36	0.5	100	0.619 ± 0.015	0.0999 ± 0.0026	0.503	0.0813	- 18.62	0 634	0.1026	2.70
36	0.5	200	0.796 ± 0 018	0.1282 ± 0.0028	0.669	0.1082	- 15.60	0.814	0.1317	2.73
36	3.0	100	0.441 ± 0.032	0.0612 ± 0.0015	0.503	0.0710	16 01	0.449	0.0634	3 59
36	3.0	200	0.595 ± 0 045	0.0832 ± 0.0032	0.669	0.0945	13.58	0 605	0.0854	2 64

** exact solution for the symmetric exponential cyclic network

4. STOCHASTIC DESIGN OPTIMIZATION OF ASYNCHRONOUS ASSEMBLY SYSTEMS

The Optimization Problem

Our purpose in this section is to investigate the application of stochastic optimization to the improvement of assembly systems. The problem is formulated as a discrete Monte-Carlo optimization problem and is solved using stochastic quasigradient methods (Ermoliev 1983). The objective functional we will use is the expected rate of production of completed assemblies. Our decision variables will be the buffer sizes between each pair of stations in the systems as well as the number of pallets loaded on the line.

The only general analytical techniques for lines with more than two unreliable stations and finite buffers are for special systems. In these systems the probability that a machine is under repair during a given cycle is independent of the state of that machine during the previous cycle. There are relatively few optimization studies in this area but in a paper whose approach is similar in spirit to the methods presented here, Ho, Eyler, and Chien (1979) use perturbation analysis and gradient methods to study the effect of buffer sizes, cycle times and clear times on the production rate of open transfer lines.

In this section we examine a closed AAS which has both finite buffers and transport delays included explicitly in the model. Our primary goal is to find the optimal number of buffers between each pair of stations and the optimal number of pallets to be loaded on the system. Since there are no general analytical methods available for this complex problem we attempt the application of a general stochastic optimization procedure (stochastic quasigradient method) to obtain the solution. Since the model of the system is obtainable only as a discrete event simulation we are involved in a Monte Carlo optimization problem. In addition, a further complication is present here since the SQG method was not originally intended for problems with discrete decision variables. As a consequence, the original convergence proofs for the method do not apply in our situation.

A wide range of engineering optimization problems cannot be solved by using deterministic optimization methods. The SQG method is a generalization of standard gradient methods under conditions where direct calculation of the gradient is not possible. The SQG method substitutes estimate of the gradient for the (unobservable) true value. In applications such as the optimization of design parameters for manufacturing systems both the values of the objective function itself as well as the gradient must be obtained from multiple observations from a discrete event simulation model. While under certain conditions (Glynn and Sanders 1986) convergence proofs can be obtained, matching the method for obtaining estimates of the gradient to the problem type, determining the step size and other parameters of the method appropriately for the specific problems remains something of an art form. Liu and Sanders (1986) have adapted the SQG method for application to assembly system optimization.

The Model

The problem we are attempting to solve is to find the optimum production rate of an AFAS with the buffer size of each station and the number of pallets on the entire system as decision variables. We can write:

$$\max: \text{ production rate} = F(x); \quad x \text{ in } X$$

$$F(x) = E_w\, f(x,w)$$

The production rate is the ratio of the expected number of finished assemblies per unit time multiplied by the common (station) cycle time. x is a vector of the decision variables to be optimized. We will consider a case with 5 assembly stations. In our case x represents the capacities of the buffers in front of stations 1,2,3,4 and 5 plus the number of assembly pallets circulating on the entire system. X is a set of constraints. In our case, it includes the maximum and minimum buffer sizes for each station. The w is a random variable belonging to the appropriate probability space. In our case the randomness comes from the random time between station jams (geometric distribution) and the random time required to clear the jam from the station (geometric distribution). For example, for a particular set of buffer sizes and the number of pallets, we can simulate the expected value of f(x,w) given the parameters of the jam and clear distributions. Based on the value of f(x,w), and the quasigradient obtained at each iteration, we can drive the value of x toward the optimum solution using the standard constrained gradient procedures detailed below.

3. The Algorithm

The stochastic quasigradient algorithm moves from one feasible position to another as follows:

$$x^{s+1} = \Pi_X(x^s - P_s V^s)$$

where X^s is the current approximation to the optimal solution, P_s is the step size, and V^s is a random step direction i.e. an estimate of the gradient direction at the current point x^s value within the constraint set. The projection operator simply finds the closest point inside X to the new point arrived at from moving from x^s to x^{s+1}.

3.1 Estimation of the production rate

A discrete event simulation model was developed in Pascal to run on multiuser micro-computer running an Intel 80286 microporcessor with a 80287 numeric coprocessor. The problem is capable of simulating assembly machines with from two to over a hundred assembly stations with or without explicit consideration of transport delays on the machine. We assume 5 stations with transport delays of one time unit per buffer space. Each station is assumed to have a constant assembly cycle time of 5 units. The first station in the line can be assumed to be a load-unload station where new assembly bases are loaded on to pallets and where completed assemblies are removed from the line. We assume as stated above that the time between station jams and the station clear time are random variables with geometric distributions. Unless otherwise specified we assume that the jam rates are (0,5%,0,5%,0) and the mean clear times are (0,15,0,15,0) for the five stations. In summary, stations 2 and 4 are "bottleneck" stations and the rest are jam free. The constraint set X is represented by upper and lower bound constraints on the buffer sizes and the number of pallets. We assume that each buffer size is no smaller than one and no larger than 15. The number of pallets on the system is assumed to be between 1 and 30.

3.2 Choice of step direction

The step direction may be a statistical estimate of the gradient of function F(x): then $v^s \equiv \xi^s$ such that

$$E(\xi^s | x^1, x^2, \ldots, x^5) = F_X(x^s) + a^s$$

Where ξ^s is a statistical estimate of v^s, a^s decreases as the number of iterations increases. In this case, v^s is called a stochastic quasigradient of function F(x).

There are several methods which can be used to find the estimate of gradient direction: finite difference approximations, analogues of random search methods, etc. In this paper, we'll use forward finite differences (FFD) and central finite differences, (CFD) to obtain the step direction.

(i) Forward finite differences approximation. The forward finite differences method can be written in the form:

$$\xi^s = \sum_{i=1}^{n} \frac{f(x^s + \delta_s e_i, w^s_{i,1}) - f(x^s, w^s_{i,2})}{\delta_s} e_i$$

where the e_i are unit basis vectors form R^n, ξ^s is the estimate of step direction at iteration s, δ_s is the step i finite-difference approximation, $w^s_{i,1}$ and $w^s_{i,2}$ are stochastic random values generated for iteration s.

(ii) Central finite differences approximation. The central finite differences has the form:

$$\xi^s = \sum_{i=1}^{n} \frac{f(x^s + \delta_s e_i, w^s_{i,1}) - f(x^s - \delta_s e_i, w^s_{i,2})}{2\delta_s} e_i$$

where the notations are the same as forward finite differences approximation.

3.3 Choice of step size

We start with a reasonable step size and then modify it during the iteration process. The criterion for modification is the ratio of the improvement of function value to the path length. That is

$$\Phi^1(x^s, u^s) = \frac{F^{s-M} - F^s}{q(s, M_s)}$$

where

$$q(s, M_s) = \sum_{i=s-M_s}^{s-1} \left\| x^{i+1} - x^i \right\|$$

and M_s is the number of iterations before every performance evaluation.

3.4 Projection

In our example, we take on as the minimum value for buffer sizes and the number of pallets. The upper bounds for buffer sizes and the number of pallets are set to 15 and 30.

3.5 Stopping criteria

We use two stopping criteria: the maximum number of iterations and the minimum step size. The algorithm will stop when either condition is satisfied.

4.0 Results of an Example Optimization

4.1 Performance of the algorithm

Table 3 shows the results of 10 iterations of the algorithm starting from an initial condition with 1 buffer space between each of the five stations and 4 pallets on the entire machine. The "observation" column represents the current estimated machine performance as a fraction of the maximum possible performance if no machine jams were to occur. This estimate is based on 4000 time units of observation with the first 400 omitted to reduce initialization bias. The "performance" column represents the ratio of cumulative length of the optimization path traveled to that point. The "stepsize" column indicates the length of the current move to be made in the decision variables.

The results of the process show an orderly and in fact nearly monotonic convergence of the decision variables to the final position so that the algorithm appears to work reasonably reliably. However, further experimentation has demonstrated that a number of problems must be overcome with this approach in order to attack machine optimization problems involving 50 to

TABLE 3 Information for iterations using the CFDA method.

iter	b1	b2	b3	b4	b5	pal	observation	performance	stepsize
0	1	1	1	1	1	4	0.6100	0.0000	4.0000
1	1	1	1	1	1	8	0.8522	0.0000	4.0000
2	4	2	2	1	4	8	0.8578	0.0007	3.6000
3	3	2	2	1	3	12	0.8644	0.0017	3.6000
4	4	4	4	1	4	11	0.8756	0.0025	3.2400
5	4	4	4	1	4	14	0.8867	0.0034	3.2400
6	6	6	6	1	6	15	0.8889	0.0021	2.9160
7	6	6	6	1	6	17	0.8889	0.0004	2.9160
8	6	6	6	1	6	17	0.8889	-1.0000	2.6244
9	6	6	6	1	6	17	0.8889		

a. Initial point = (1 1 1 1 1 4).
b. Step in finite difference approximation = 2.00.
c. Initial stepsize = 4.00.
d. Stepsize multiplier = 0.90.
e. Frequency of stepsize change = 2.
f. Lower bond on function increase = 0.00400
g. 'starting point = (1 1 1 1 1 4) - denotes the values for buffers numbered 1 to 5 and the total number of pallets for the whole line.
h. 'b1, b2, b3, b4, b5' denotes the ratio of the improvement of the function value to the path traveled.
k. 'stepsize' denotes that step size for each iteration.
l. 'observation' denotes the objective value from single simulation run

100 assembly stations. Machines this size are well within the practical realm of machines being built today. The major problem is that in order to estimate the quasi-gradient at each point it is necessary to run 2N simulations where N is the number of decision variables. Further the run size of each simulation needs to increase as the number of decision variables increases. The results is an extremely slow optimization process with immense computational requirements at each stage. This comparatively simple example required nearly 30 minutes for the 10 iterations on IBM-PC/AT class computer. Further problems arise from the usually sources of difficulty in any gradient based method. Convergence (or lack thereof) is determined by the choice of starting point, step size multiplier and a host of other factors. Convergence is often very slow and erratic when the algorithm get close to the optimum point.

In conclusion, while this method does show some promise it clearly needs major work to become a practical tool for large assembly machine optimization. Some progress has been made in adapting a very different optimization approach based on homotopy or imbedding methods to this difficult class of problems. It may be that quasi-gradient methods can be used in conjunction with the homotopy formulation to arrive at the "neighborhood" of the optimal solutions where the method can be switched to a Monte-Carlo version of what is known in the homotopy literature as "path following" to arrive at the final optimal solution. Research in this area is reported in Glynn and Sanders (1986).

REFERENCES

Boothroyd, G., Poli, C., and Murch, L. (1982). Automatic Assembly Systems, M. Dekker Inc., New York.

Ermoliev, Y.M. (1983). Stochastic quasigradient methods and their application to systems optimization. Stochastics, 9, 1-36.

Glynn, P. W. and Sanders, J. L. (1986). Monte Carlo optimization of Stochastic Systems: Two New Approaches. Proceedings of the 1986 ASME International Conference on Computers in Engineering, Chicago, Illinois.

Ho, Y. C., Eyler, M. A., and Chien, T. T. (1983). A new approach to determine parameter sensitivities of transfer lines. Management Science, 29(6), 700-713.

Kamath, M. and Sanders, J. (1986). RENA: A new approach to the analysis of asynchronous automatic assembly systems. Submitted to Journal of Manufacturing Systems.

Leung, W. K. and Sanders, J. (1986). Technical Report I.E. 86-3, Department of Industrial Engineering, University of Wisconsin-Madison, Madison, Wisconsin.

Liu, C. M. and Sanders, J. L. (1986). Stochastic design optimization of Asynchronous Flexible Assembly Systems. In: Proceedings of the Second ORSA/TIMS Conference on Flexible Manufacturing Systems, Ann Arbor, Michigan.

Marshall, K. T. (1968). Some inequalities in queueing. Operations Research, 16, 651-665.

Suri, R. and Hildebrant, R. R. (1984). Modeling flexible manufacturing systems using mean value analysis. Journal of Manufacturing Systems 3(1), 27-38.

Whitt, W. (1985). The best order of queues in series. Management Science, 31(4), 475-487.

AUTHOR'S BIOGRAPHY

JERRY L. SANDERS is a Professor of Industrial Engineering at the University of Wiconsin-Madison. Prior to 1974, he was Professor of Systems and Industrial Engineering at the University of Arizona-Tucson. He received his B.S. and M.S. degrees at Montana State University and his Ph.D. in Operations Research in 1963 at Case Western Reserve University. His current research interests are in computer simulation, approximate stochastic models of manufacturing systems, and in development of new methods for Monte Carlo optimization. He is a member of TIMS and the ASME.

Jerry L. Sanders
Dept. of Industrial Engineering
University of Wisconsin - Madison
1513 University Avenue, Rm. 459
Madison, WI 53706
(608) 263-5784