PERFORMANCE ANALYSIS MODELING FOR MANUFACTURING SYSTEMS

Albert A. Fredericks

AT&T TECHNICAL JOURNAL

Albert A. Fredericks is head of the Performance Analysis department at AT&T Bell Laboratories in Holmdel, New Jersey. He is currently responsible for providing AT&T with needed performance analysis theory, methods, and tools for computer, communications, information, and manufacturing systems. He joined the company in 1961. Mr. Fredericks received a B.S. in mathematics from Fairleigh Dickinson University and an M.S. and Ph.D. in mathematics from the Courant Institute, New York University.

Performance analysis modeling can provide the means of ensuring cost-effective engineering and operation of manufacturing lines. This article presents the basic concepts of performance analysis modeling and illustrates its application to the design, engineering and operation of manufacturing systems. A particularly versatile tool for performance analysis of manufacturing systems, the Performance Analysis Workstation, is also discussed. Finally, comments on future trends in performance analysis modeling are given.

Performance Analysis Modeling

Performance analysis modeling is the use of abstract mathematical or simulation models to quantify certain aspects of a system's performance. These models are abstract or mathematical models, not, e.g., the building of system prototypes. Prototyping is, however, often an essential part of the design and development process and can contribute significantly to the characterization of system performance. Our models are abstractions of the system under consideration, generally accompanied by some mathematical expressions or realization scenario that allows us to simulate the behavior of the abstract model.

Another key term is quantification. Performance analysis models allow one to make quantitative statements about system performance (e.g., system throughput in circuit packs per hour, or expected work in progress). They are also extremely useful in comparing the relative performance of design or operating alternatives.

It is generally not possible to build one performance model that will answer all performance questions. These models can examine only certain aspects of the system at a time. Indeed, a precondition to constructing an appropriate performance analysis model is the specification of the performance questions it is expected to address.

Performance models are useful because they are cheaper and easier to build than actual systems, they are cheaper and easier to "perturb" than actual systems, and sometimes they can yield important insights into system performance that would not be readily obtained even by studying the system itself. The importance of these concepts to the design, development, engineering and operation of computer and communications systems has been recognized for some time. They are of equal importance to manufacturing systems as well.

When To Use Performance Analysis Models. Performance analysis models are used throughout a system's life cycle. Models should be introduced early in the design phase so that they can help determine system viability. The following is an example. A planned line may be required, for economic viability, to produce circuit packs at a rate of 20,000 packs per week on average with an interval of two weeks and also with the capability of responding to peak throughput demands of 30,000 packs per week for a limited time. Can the planned line meet such requirements? Performance analysis models can answer such a question.

During development and implementation of the line, performance analysis models can be used to address various tradeoff questions. Questions such as: Is it preferable to use cheaper insertion machines that may have longer down times or to use more expensive but more reliable machines? Or, what is the *quantitative* tradeoff between work-in-progress costs and investment in line capacity?

Performance analysis models can be used for the engineering and operation of the line to help determine such factors as the buffer space needed for storage between machines, the optimal allocation of common buffer space, or optimal loading and code scheduling. Other analyses could include: the most economical method of meeting special code mix requests or of increasing line capacity either temporarily or permanently.

Tools of the Trade

Performance analysts use a variety of sophisticated mathematical methods and tools as the fundamental building blocks for macro as well as micro models. These methods are borrowed from various disciplines including applied probability, stochastic processes, queueing theory, scheduling theory, control theory, optimization theory and various tools common to operations research, such as mathematical programming. Often the performance analyst must be creative in selecting the abstract model from these disciplines that will fit a given physical system. In addition, it may be necessary to develop new tools when those available do not adequately characterize system performance.

Fortunately, these disciplines have been used to construct macro modeling tools of quite broad applicability. Such tools have made it considerably easier for engineers to develop needed performance analysis models. Perhaps some of the most useful macro tools are those based on the concept of queueing networks.

Queueing Network Models. The nature of manufacturing systems often means that they naturally fall into the modeling scope of queueing networks. In addition, there are many good software packages that can readily provide solutions once the analyst has constructed an appropriate model. From a macro view, a queueing network model consists of a set of work centers or service centers where servers (e.g., a set of insertion machines or test sets) provide service to arriving jobs (e.g., insert components in arriving boards) together with a connectivity that specifies the allowable flow of jobs (Figure 1).

The micro level adds needed details that characterize each service center. These details could include: the number of machines at the center, the various kinds of jobs that can arrive, the time needed to process a job of a given type, where each job is to be sent after service, or, the amount of buffer space available to store arriving jobs.

Perhaps the best way to illustrate these concepts is to closely examine the model in Figure 1. The figure shows the topology of a macro model of a maufacturing line with seven work centers:

- A Label center, where boards are labeled,
- Insertion center 1, where certain components are inserted on some boards,
- Insertion center 2, where all boards have additional compounds inserted,
- A Solder Machine center,

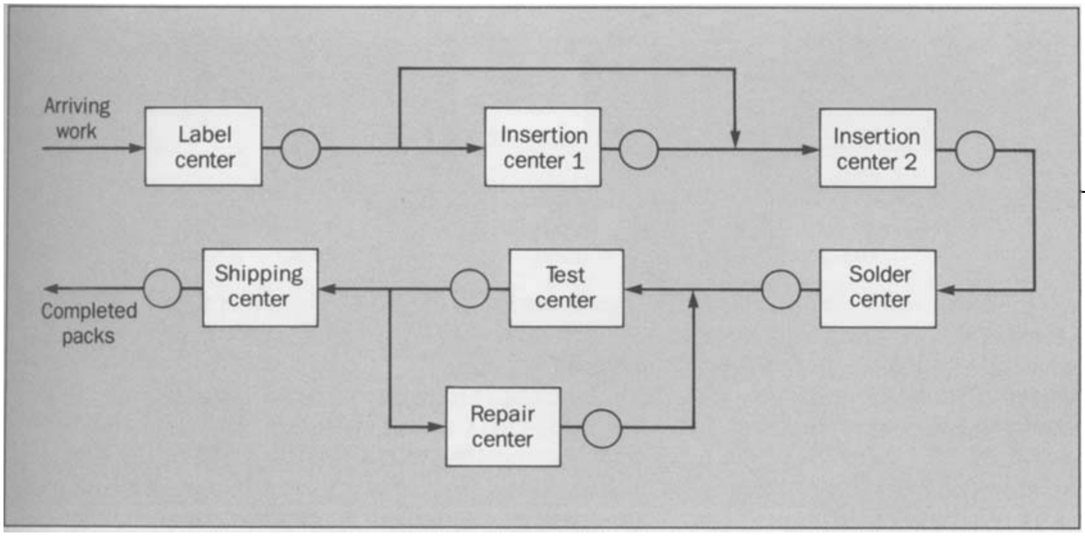


Figure 1. A macro model of a manufacturing line with seven work centers.

- A Test center, where completed packs are tested,
- A Repair center, where defective packs are repaired,
- A Shipping center, where all packs that have passed testing are sent for packing and shipping.

Within this model the performance analyst can further specify what kind of jobs arrive at each center, what facilities are available to process them, how they are to be processed, where they are to go after processing at each center, etc. For example, at the Label center, the analyst may specify that there is a single labeling machine and that the time it takes to label a board is the same for all boards. There might be two classes of circuit packs (codes) to be made. Some would go to Insertion center 1 after labeling while others would go directly to Insertion center 2 after labeling. In addition, the model would need to specify an arrival rate for boards of each type. Perhaps, this would be 2,000 boards of each type loaded at the beginning of each day.

Other specifications might be: how much room there is to store arriving boards, and what is to be done if either the storage area for boards arriving to the Label center is full or the storage areas at the Insertion centers are full. (In this latter case, the resulting action might be to stop labeling new boards.) Finally, other items can be specified for the label machine, including how often the machine might breakdown, how long it would take to repair, and what the setup time would be. Such items are somewhat random and need to be specified via probability distributions.

Similar items for other work centers would also be specified, as well as additional items that might be unique to various centers. For example, at the Test center, the performance analyst might need to specify the defective rate for each pack. This would define the probability that a pack would need to be sent to repair. At the Shipping center, it might be necessary to accumulate a given number of each circuit pack type before packaging and shipping, etc.

Obtaining a Solution. While the example shown in Figure 1 may seem simple from a manufacturing viewpoint, it already represents a level of model sophistication for which exact analytic solutions cannot be obtained. Indeed, even obtaining a good approximate solution for such a model is a formidable task. There are, however, simulation modeling tools that can be used to obtain performance measures for such models.

First, it is important to examine the three potential solution methods: exact analytic, approximate analytic, and simulation. It is only in a limited number of cases that queueing network models (and most other performance models) can be specified via exact mathematical expressions. However, often the simplifying assumptions that are needed to obtain exact analytic solution, while not strictly valid, are nevertheless reasonable to make during early system analysis. During this time, only rough estimates of factors such as throughput and interval might be needed. Moreover, during early system design, one often has only rough estimates of many of the quantities needed for modeling. Hence, it may not be cost effective from a performance modeling viewpoint to make more detailed modeling assumptions.

In some cases, even during early design, certain assumptions needed to yield exact solutions are known to be invalid yet critical to system performance. In this case, it may be possible to use some of the well-known approximation methods for queueing models. Indeed, even when simulation methods are employed, analytical solutions and approximation methods may be valuable supplements to narrow the scope of the analysis.

For example, the Repair center of Figure 1 may be a simplified version of a more complex (and accurate) model of the Repair center. In Figure 2, the Repair center is expanded to include an Analyze center, an Integrated Circuit Repair center, and a Touch-up center as well as an alternate Bench Repair center. Statistics could be gathered from the solution of this model and then used to obtain a "typical" mean and variance for the time in this complex repair center. These items would then be used as inputs to characterize the simple repair center model of Figure 1. This powerful approximation technique is often referred to as *decomposition.*

Analytic methods can serve another important purpose. Often it is useful to evaluate many engineering alternatives via quickly solved analytic models (perhaps as part of an optimization scheme). The performance analyst can then use more accurate simulation models to evaluate a few of the seemingly best alternatives.

Thus, to provide appropriate performance analysis for even a single line, it may be appropriate to use a combination of various analytic and simulation models. This could mean that several modeling tools and analyses might be needed. Fortunately, there is at least one available tool that can simplify the modeling process: the Performance Analysis Workstation.

The Performance Analysis Workstation

The Performance Analysis Workstation(PAW)[1,2] is a growing suite of interactive tools for carrying out the performance analysis of manufacturing as well as computer and communications systems. Developed by the Performance Analysis department at AT&T Bell Laboratories, it is currently in wide use among AT&T development and systems engineers as well as engineers at many of AT&T's manufacturing locations. (A version of this workstation is available to non-AT&T organizations through the UNIX® System Toolchest.)

The Performance Analysis Workstation runs under the UNIX operating systems and requires a 3B or VAX host together with a Teletype® DMD 5620 terminal. Visual models are built on the screen using a mouse. Once the topology is specified, appropriate parameters for each of the workcenters are entered via a user friendly form entry system. On-screen guidance is provided to the user at all times.

A host of useful statistics-gathering modules can readily be invoked and their results displayed graphically as simulation time evolves. The animated simulation provided is extremely useful for debugging and for gauging the performance of the system. At any time, the simulation can be stopped and parameters or structural changes added (e.g., a breakdown) and then the effect can be observed.

The simulation capabilities are quite sophisticated. New features, responding to the need for modeling manufacturing systems, are always being added. The system is well documented and easy to learn to use. In addition, training courses are available through the Performance Analysis department, although only within AT&T at this time.

Although PAW's modeling capabilities are extensive, there are many times when the use of supplemental analytic models can be very useful. For this purpose, PAW also provides a common interface to two useful analytic packages developed at Bell Laboratories for solving queueing networks. One is PANACEA,[3] which can provide exact analytical solutions to a wide class of "product form" networks. The other analysis package, QNA,[4] provides approximate analytic solutions to an important class of queueing networks for which exact solutions are not available.

In addition to PAW, there are, of course, many other simulation packages that are widely used for analyzing models of manufacturing systems. Most notable are IDSS[5] and SIMAN.[6] More general-purpose simulation lan-

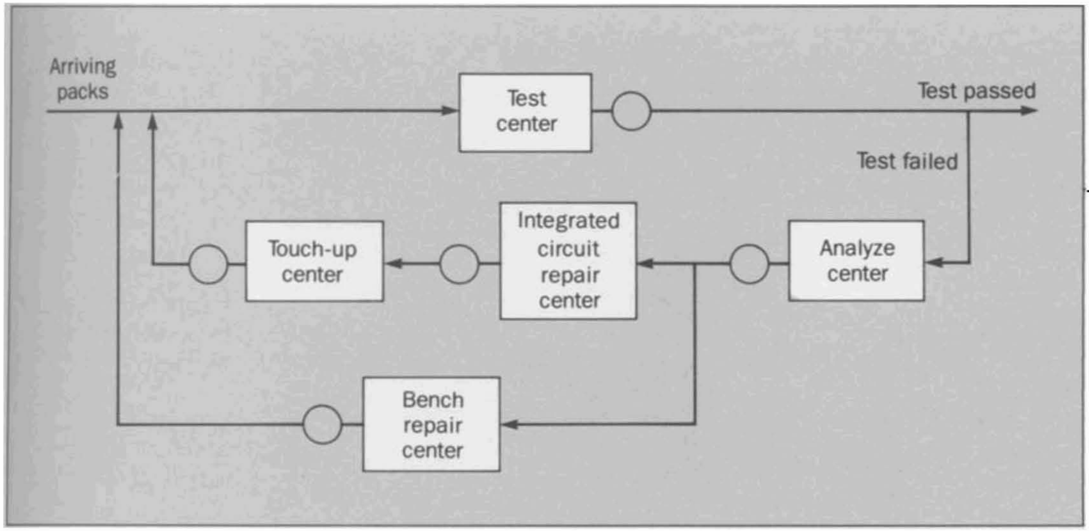


Figure 2. Expanded view of test and repair centers.

guages such as GPSS and SIMSCRIPT are also widely used.

Examples

The following examples have been greatly simplified, but they are based on actual performance analysis modeling efforts undertaken to support existing and planned manufacturing systems at various AT&T manufacturing locations.

System Design and Capacity Planning. Many lines built in recent years by AT&T and many other corporations are based on the in-line manufacturing concept. These lines tend to use highly automated insertion machines, material handling, and shop floor control. In an effort to enhance interval control, most provide for only a relatively small buffer area between service centers. While these lines are ideally suited for high-volume production, it is also desirable to make use of any capacity not needed by high-volume codes to process lower volume runners that might be needed to form the final assembled product.

Figure 3 is a reproduction of a computer screen showing a performance analysis model built using PAW. The analysis addresses several performance questions during the capacity planning process for such a line. (For simplicity, the model for the test area is not included.)

Method 1. This particular model corresponds to one of several methods studied for sharing line resources between high- and low-volume codes. In this case, one of each of the high-speed insertion machines (shown in the figure as "sing" for single, through "quad," for quadruple) is dedicated to low-volume codes (shown as "singlv," for example), while all others are used for high-volume code ("hv") production. (The terms single, double, triple, etc. are purely for ease of identification and do not refer to any specific feature of these machines.)

The robot loader at the front end of the solder machine gives high priority to high-volume codes. Shown in the screen inset of Figure 3 are a histogram and time series for the total queue length at the solder machines. These were used to determine if the buffer space provided was sufficient. In the figure, note the visual cues available to the analyst: actual location of each code type (lv, hv), indication of jobs being processed (highlighted by reverse characters), jobs waiting (plain label), jobs blocked (dimpled background).

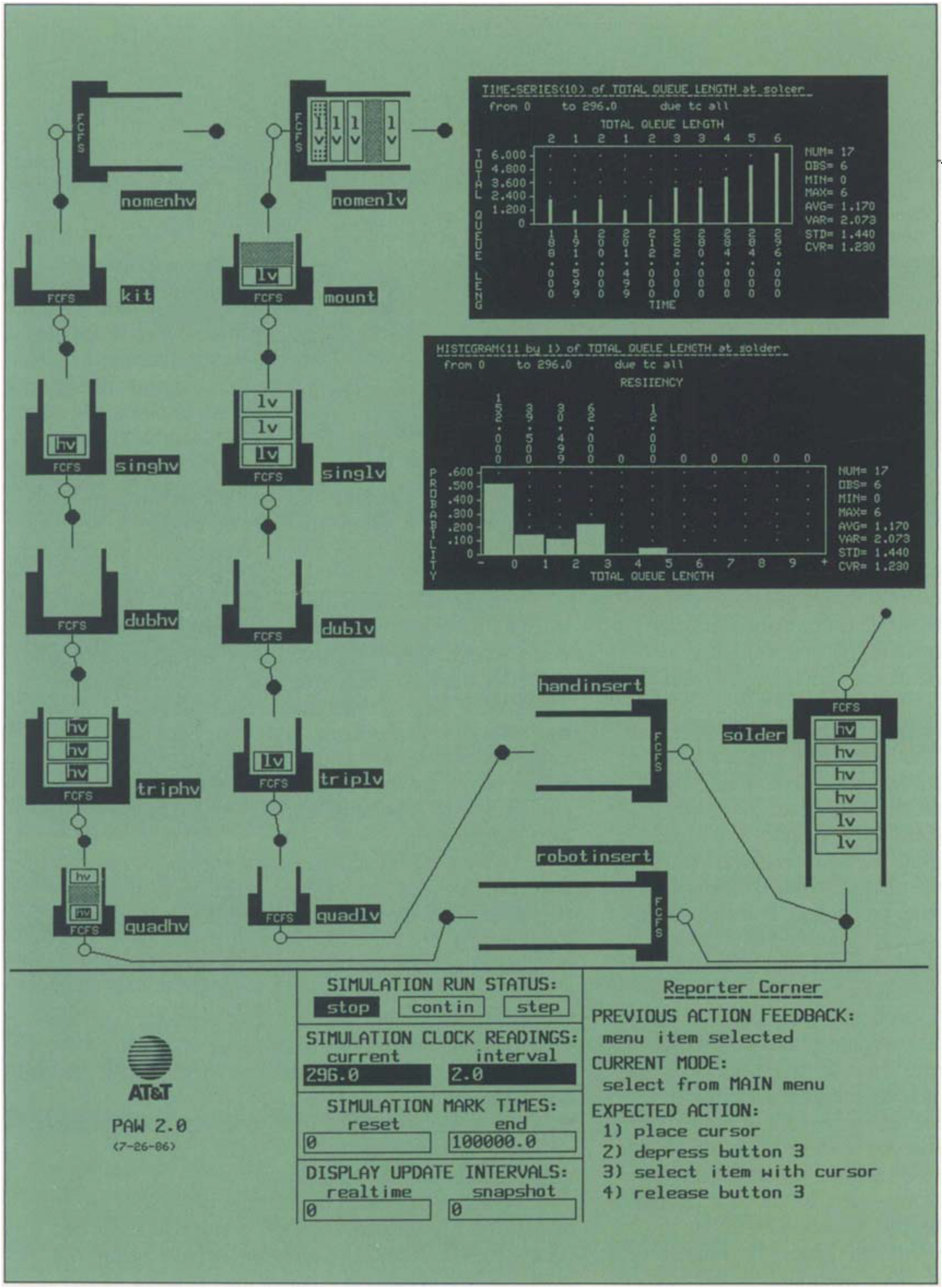
Method 2. In addition to the method of mixing production of high- and low-volume codes noted in Method 1, two other methods were considered. Method 2 specified the following:

- Allocate all insertion machines to high-volume codes to maximize their insertion rate.
- After completion of a given high-volume run, produce the appropriate number of low-volume codes.

Method 3. Method 3 differed from Method 2 in that it specified:

- While giving preference to high-volume code production, dynamically "mix" low-volume codes into production as capacity dictates.

The first two methods listed can readily be implemented via PAW models and their performance can be compared. For example, the results in Table I show the length of time needed to produce the expected weekly pro-



TIME-SERIES(10) of TOTAL QUEUE LENGTH at solcer
from 0 to 296.0 due tc all

NUM= 17
OBS= 6
MIN= 0
MAX= 6
AVG= 1.170
VAR= 2.073
STD= 1.440
CVR= 1.230

HISTCGRAM(11 by 1) of TOTAL QUELE LENCTH at solder
from 0 to 296.0 due tc all

NUM= 17
OBS= 6
MIN= 0
MAX= 6
AVG= 1.170
VAR= 2.073
STD= 1.440
CVR= 1.230

SIMULATION RUN STATUS:

SIMULATION CLOCK READINGS:
current 296.0
interval 2.0

SIMULATION MARK TIMES:
reset 0
end 100000.0

DISPLAY UPDATE INTERVALS:
realtime 0
snapshot 0

Reporter Corner

PREVIOUS ACTION FEEDBACK:
menu item selected

CURRENT MODE:
select from MAIN menu

EXPECTED ACTION:
1) place cursor
2) depress button 3
3) select item with cursor
4) release button 3

Figure 3. PAW model of circuit pack assembly line.

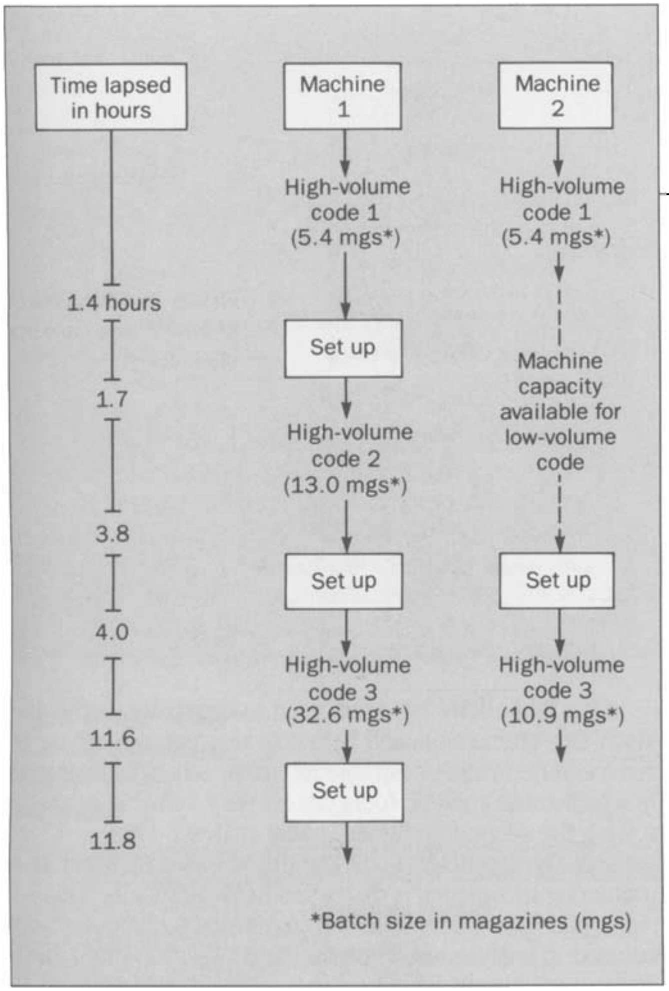

Time lapsed in hours
Machine 1
Machine 2
High-volume code 1 (5.4 mgs*)
High-volume code 1 (5.4 mgs*)
1.4 hours
Set up
Machine capacity available for low-volume code
1.7
High-volume code 2 (13.0 mgs*)
3.8
Set up
Set up
4.0
High-volume code 3 (32.6 mgs*)
High-volume code 3 (10.9 mgs*)
11.6
Set up
11.8
*Batch size in magazines (mgs)

Figure 4. Feasible scheduling for an insertion machine.

Table I. Machine Availability

	100%	95%	90%
Method 1	12.5	13.3	14
Method 2	14	14.5	15

Time (in shifts) to complete weekly production of codes

duction quotas at a particular point in ramp up.

While there is potentially more capacity to be gained with Method 3, its implementation requires more complex scheduling. Given the code production goals, machine set-up times, code processing times, etc., one must schedule high-volume code production so that the excess machine capacity is available for low-volume code production in reasonably large chunks. Otherwise, the overhead of additional set ups could destroy any potential for increased capacity.

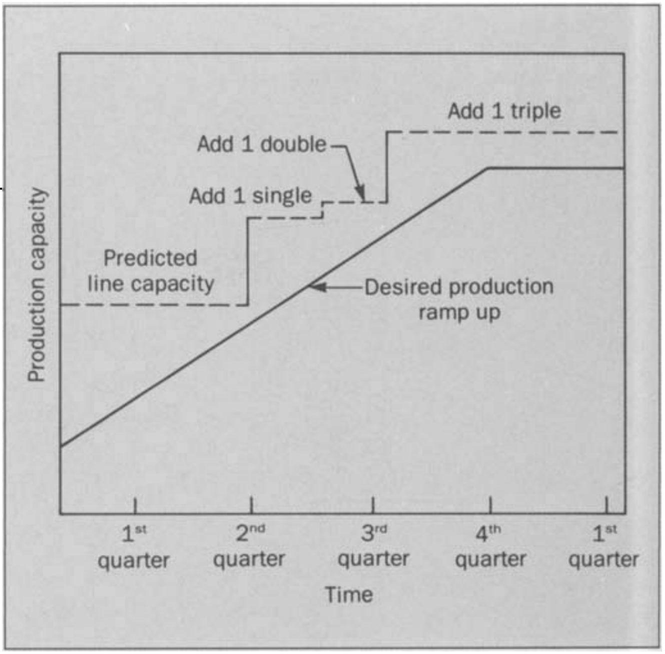
Accordingly, a software package, Scheduling Evaluation System for Automated Manufacturing Environments (SESAME) was developed. Figure 4 shows one of the outputs of this package and the information it provides at each work center. As indicated in the figure, a feasible scheduling algorithm of this type could provide up to a 20-percent increase in system production capacity. However, from a capacity planning viewpoint, it was decided to assume the simple schedule of Method 1 because it could provide the same capacity as Method 3 with a very modest addition of capital investment. Method 3 does provide an important solution to any unexpected demand for quickly increasing system production capabilities.

Having chosen Method 1, a PAW model was used to ensure that the planned system would meet the desired production ramp up. Results obtained from this model are summarized in Figure 5. Shown is the ramp up in capacity desired as well as the predicted line capacity. The first line change, resulting in the desired capacity increase at the start of second quarter, is accomplished by adding an additional single insertion machine. The addition of a double insertion machine at the start of the third quarter does not significantly increase capacity (because it is not the bottleneck). It does, however, enhance availability and is needed in addition to another triple insertion machine in order to reach the desired year-end capacity.

As noted earlier, when a wide range of parameters are to be studied for a line, it is desirable to have at least an approximate analytic method available in order to do some preliminary performance analysis. An approximation technique was developed for this line. For a discussion of that approximation, as well as its application to in-line manufacturing lines, see Reference 7.

Line Engineering and Operation. One of the key methods being employed to help control manufacturing intervals is the provisioning of relatively small storage areas for work in progress. This allows lines to obtain some of the important benefits of pure pull systems and (if well

Figure 5. Capacity planning curve.



designed) not suffer excess capacity loss because of idling machines. However, finite buffers between work centers can greatly complicate the performance analysis of lines. While this is true for in-line systems (see Reference 8), it is even more crucial for flexible manufacturing lines in which many different codes of small lot sizes need to be processed.

Figure 6a shows three Service centers that process two different codes. Code *O* takes 0.1 time units at Service center 1, 0.9 time units at Service center 2 and 0.75 time units at Service center 3. The corresponding values for Code *X* are 0.9, 0.1, 0.75. These differences are not atypical, particularly in a flexible manufacturing environment. This also assumes two buffers are available at the input to each Service center for storage.

In Figure 6a, it is assumed that a lot size of 100 is used. As shown in the figure, this results in a system throughput of one production unit per second and a maximum machine use of 75 percent at Service center 3. Figure 6b shows that if the lot size is reduced to one, the capacity will increase by one third. It is true that some of this capacity would be lost because of increased setup time associated with smaller lot sizes. However, there is probably a lot size smaller than 100 that will result in an increase in capacity.

Now consider the same three Service centers when required to produce 200 codes. For simplicity, assume that 100 of these codes have processing times like those of Code *O* and the other 100 like Code *X*. If the production cycle consists of producing the first 100 codes in lot sizes of one followed by the latter codes in lot sizes of one, the capacity will be 1/0.9. If on the other hand, we produce one lot of the first type followed by one lot of the second type, the capacity increases by one third. In this case, the former strategy does not even provide the advantages of reduced set-up penalty.

The problem is that if these centers were part of a larger, more complex line processing a large number of codes with considerable diversity in processing times, the excess capacity in the line might never have been detected.

While PAW has been used to construct models to study this phenomena and help engineer appropriate buffer sizes and determine desirable lot sizes, this is another case in which some analytic tools can be very valuable in narrowing the scope for refined model analysis. One particularly important analytic method based on linear programming techniques is discussed in Reference 9. This reference shows how linear programming techniques can be used to analyze and improve the production rate of in-line manufacturing systems.

In flexible manufacturing lines involving hundreds of codes, simple analytic methods are harder to find. However, the insights developed from performance modeling can be used to derive heuristics for code grouping and mixing to achieve high throughput and small intervals. One such heuristic algorithm is discussed in Reference 10. This algorithm has been tested using simulation of a real, large-variety assembly shop and has been shown to yield capacity increases of 15 to 20 percent.

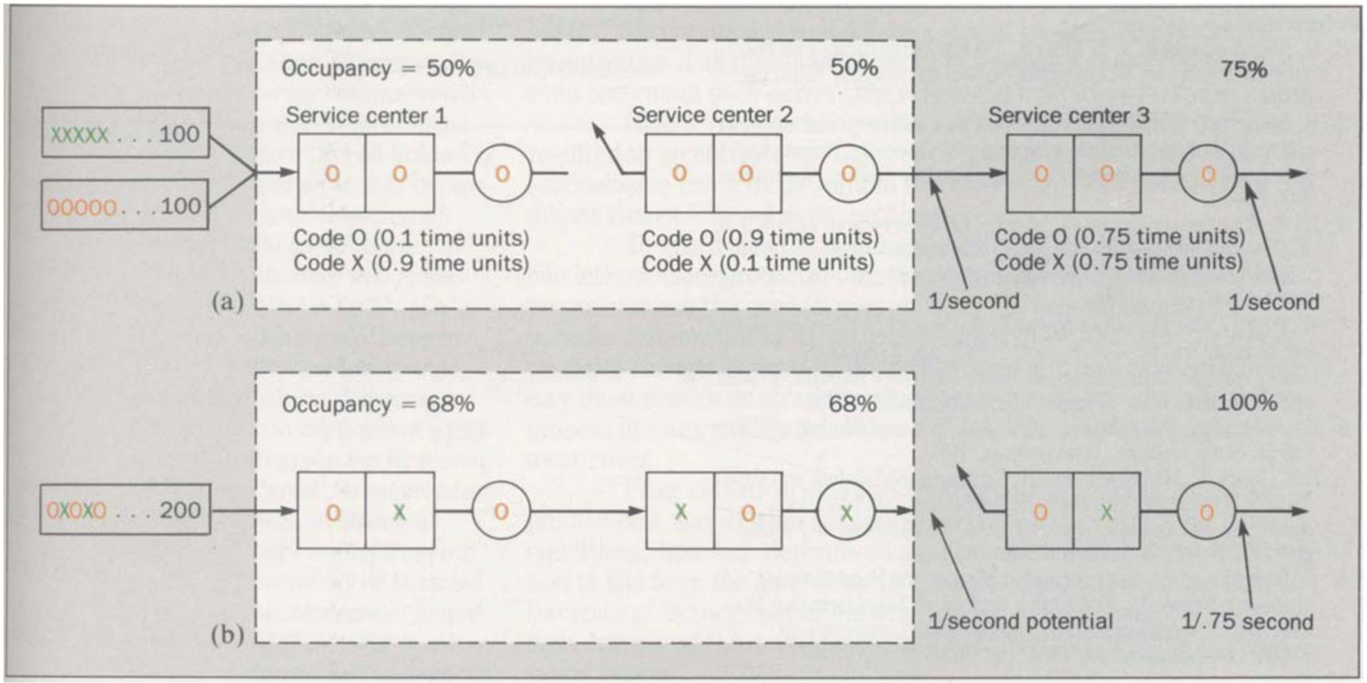
Small buffer areas provide an effective means of controlling interval and in-process inventory, although, as noted, performance analysis models may be essential to ensure the effective use of valuable system resources. For lines that have not been designed with small buffer areas, controlled loading can provide an alternative way of reducing interval and in-process inventory. For example, Figure 7 shows an in-line manufacturing system in which the required average loading is 200 units per week. If this is accomplished by loading precisely 200 units every week, we see that the average interval (determined via a PAW model) will be 300 hours.

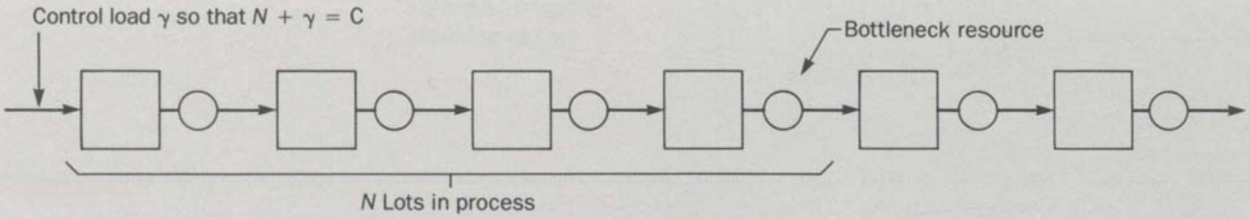
This rather large interval is the result of inherent variability in line production capability caused by, for example, machine down times. Another important source of potential variability is the availability of components. (Keeping a cost-effective inventory supply in the face of uncertain lead times can be a challenge. See Reference 11.) This interval, and hence the work-in-process inventory, can be significantly reduced by using controlled loading based on the state of the line. In the simplest case illustrated in Figure 7, the bottleneck work center is identified and the amount of work in process in front of this center is used to control the input. Evaluation (again via PAW) of the impact that this control can have on interval (work in process) is shown in Figure 7.

The Future

The development of sophisticated performance analysis tools is becoming increasingly vital to ensure cost-effective performance engineering and operation of today's modern manufacturing lines. There is an effort to not only increase the technical scope of available tools but also to enlarge the user community by providing user friendly interfaces for such tools. As more lines become fully integrated via computer-based information and control systems, the tendency will be toward integrating perform-

Figure 6. Interaction between scheduling, bottlenecks and capacity.





	Control variable, C (lots)	Throughput (lots)	Interval (hours)	Work in progress (lots)
Controlled	230	196	100	245
	260	200	118	295
	290	200	136	340
Uncontrolled		200	300	750

Figure 7. Controlled loading.

ance analysis tools into the computer systems supporting the lines. This would provide the flexibility of obtaining needed model inputs directly from the line's information support systems. Thus, workload characterization and even model building could be essentially automated and perhaps integrated into the control processes.

References

1. B. Melamed and R. J. T. Morris, "Visual Simulation: The Performance Analysis Workstation," *Computer,* August 1985, pp. 87-94.
2. B. Melamed, "The Performance Analysis Workstation: An Interactive Animated Simulation Package for Queueing Networks," to appear, Fall Joint Computer Conference '86, Dallas, November 2-6, 1986.
3. K. G. Ramakrishnan and D. Mitra, "An Overview of PANACEA: A Software Package for Analyzing Markovian Queueing Networks," *Bell System Technical Journal,* Vol. 61, No. 10, 1982, pp. 2849-2872.
4. W. Whitt, "The Queueing Network Analyzer," *Bell System Technical Journal,* Vol. 62, No. 9, Nov., 1983, pp. 2779-2815.
5. *IDSS Prototype (2.0), Version 4, User's Reference Manual,* Pritshr and Associates, Inc. Albuquerque, New Mexico, 1983.
6. C. D. Pegden, *Introduction to SIMAN,* Systems Modeling Corporation, State College, Pennsylvania, 1982.
7. P. K. Johri, E. H. Lipper, and B. Sengupta, "Modeling and Analysis of a Production Line with Finite Buffers and Machines Subject to Breakdown," R.A.I.P.O. APII, *Systemes de Production,* Vol 19, No. 5, 1985, pp. 471-483.
8. C. Buyukkoc, "An Approximation Method for Feed Forward Queueing Networks with Finite Buffers, A Manufacturing Perspective," *Proceedings IEEE International Conference on Robotics and Automation,* 1986, pp. 965-972.
9. P. K. Johri, "A Linear Programming Approach to Capacity Estimation of Automated Production Lines with Finite Buffers," *International Journal of Production Research* (accepted for publication).
10. K. Rege, "A Loading Algorithm for a Job Shop Assembly Line Run in a Pull Mode," to appear in *ORSA/TIMS,* Miami, October, 1986.
11. A. Kumar, "Model for Stockroom Inventory When Supply Lead Times are Uncertain," to appear in *ORSA/TIMS,* Miami, October, 1986.

(Manuscript received February 19, 1986)