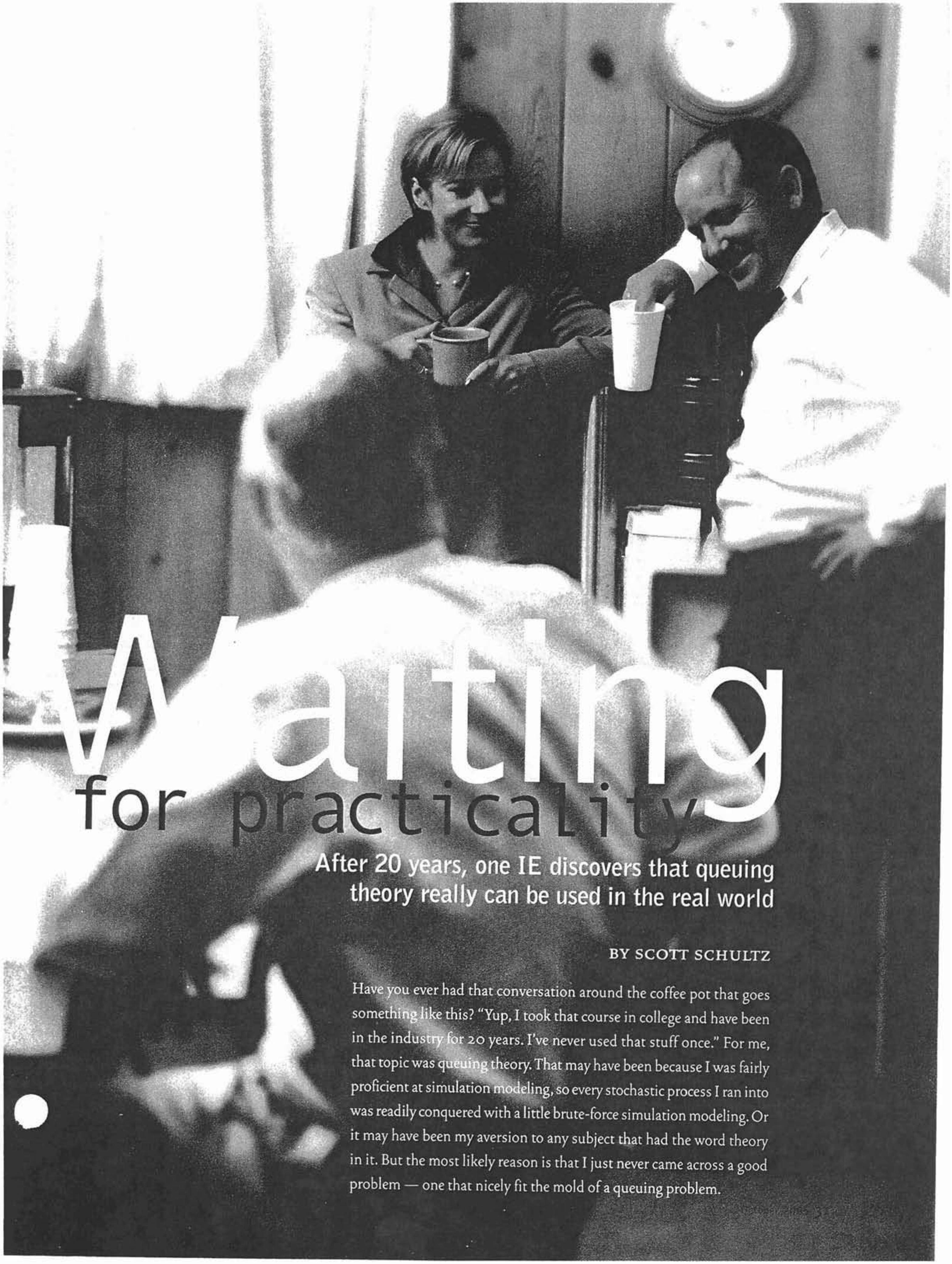

Waiting for practicality

After 20 years, one IE discovers that queuing theory really can be used in the real world

BY SCOTT SCHULTZ

Have you ever had that conversation around the coffee pot that goes something like this? "Yup, I took that course in college and have been in the industry for 20 years. I've never used that stuff once." For me, that topic was queuing theory. That may have been because I was fairly proficient at simulation modeling, so every stochastic process I ran into was readily conquered with a little brute-force simulation modeling. Or it may have been my aversion to any subject that had the word theory in it. But the most likely reason is that I just never came across a good problem — one that nicely fit the mold of a queuing problem.

Having never applied queuing theory on the job is not a tragedy in itself. I could happily have gone to my grave without once determining λ or μ. Alas, that was not to be. I recently took a job teaching industrial engineering. I was having a great time teaching all those subjects for which I had firsthand experience: simulation, linear programming, engineering economy, production control. But wouldn't you know it, right there in the course catalogue description for my recently acquired operations research course were those dreaded words: queuing theory. It wasn't so much the fear of wiping the drool off students passed out during my eloquent formulation of the MM1 steady state equations, but that I had no stories of applications in which I had used QT. So I did what every good QT instructor does: I compared McDonald's to Wendy's.

ORIGINAL LAYOUT

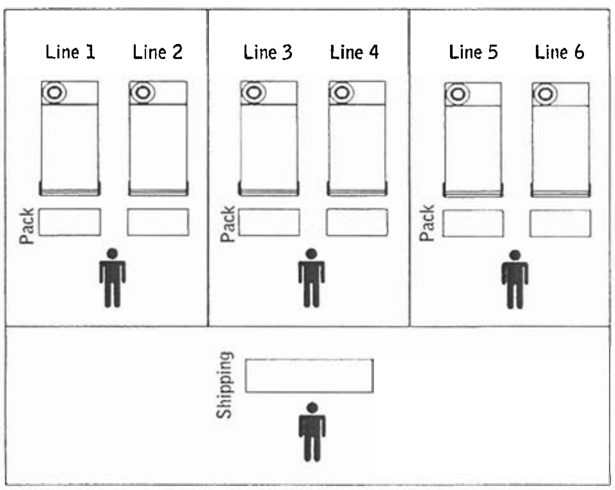

Figure 1. Six lines, three packers, fourth operator assigned to shipping

It came as a great relief to find a problem with real beef. Here's the story.

The story unfolds

It all began with a couple of bright students venturing naively forth on that quest called Senior Design, a rite of passage during a senior's last two semesters. Often the students will seek out a local company with a not-so-neatly defined problem. In this case, it was a company that manufactured supplies for the apparel industry. These supplies, or parts, will hereafter be referred to as widgets.

Laura Mock and Amanda Parish Malcom, industrial engineering and industrial management students respectively, were asked to look at the final assembly and packing process for the widgets, their goal being to increase production. The manufacturing plant is owned by a company headquartered on a small island nation in the Pacific. The final assembly machines that produce the widgets are designed and built by the mother company, which sets the standard at 4,000 parts per day. The local guys were turning out only 3,050, and Mama wasn't happy.

So the students began their quest for the silver bullet by spending time on the floor observing the process. They found that the final assembly and packing area consisted of 14 parallel lines, with each operator responsible for two lines. The lines were fully automated except that once a batch of 25 widgets was produced, the operator would fan through the widgets for a quick quality inspection and then pack the widgets into small boxes. The operator was also responsible for keeping two lines running, unjamming the machines when necessary, and performing setups when changing to a different type of widget.

While there were 14 parallel lines, the lines were readily enabled or disabled based on scheduled production volume. During the observation, six lines were being run by three operators. This area also had a fourth operator (the shipping operator), who was responsible for accumulating the packages into a larger shipping box, placing labels, and sealing the box for shipment to the customer (Figure 1).

After spending several hours over various shifts observing the process, the students began to get a good feel for the process and noticed that the operators weren't all performing at the same rate. Furthermore, they didn't all perform the job in the same way. This steered the students to pull out their stopwatches and time each operator. They collected a significant amount of observations and produced some nice charts showing cycle times for various operators.

The operators were not all operating at the same cycle time. Where were the standards? Why were some workers slower than others? Was the training insufficient? Did they lack motivation? Is this why the production standard couldn't be met?

The students related these findings to me and we explored the process in more detail. They casually noted that frequently the machines were not running. (Light bulb illuminates.)

Me: "Why do the machines stop?"

Student: "Sometimes the machines get jammed or run out of parts."

Me: "How often does this happen?"

Student: "We're not exactly sure, but a lot."

Me: "How long are the machines down?"

Student: "Sometimes it is a few minutes, but at other times it is much longer."

After some minor cajoling (i.e., threatening a course grade of B- vs. A), the students marched back to the plant to begin watching the machines rather than the operators. Here's what they found.

Data and solution

The data is fairly simple. The students needed to observe how often the line stopped running and how long the lines were down. Fortunately, the machine controller for each line automatically and accurately calculated the line efficiency. This measure is essentially the time the machine is running and producing parts divided by the total time during the shift. A wide variety of line efficiencies was found, ranging from as high as 86 percent to as low as 30 percent. However, for 30 observations, the average efficiency was 57 percent with a standard deviation of 16 percent.

The second measure of importance was the rate at which machines stopped. The students observed on average about two stops per hour.

At this point, the students had determined the following about an average machine:

- Line efficiency: 57 percent
- Stoppage rate: two stops per hour
- Average throughput: 3,050 parts per day

Armed with this data, the students sat down with the plant folks and devised a plan: Dedicate one of the three operators to keeping the machines running. This operator would perform the setups, unjam the machines, and replenish materials. The other two operators would each take on three lines and perform inspection and packing.

The rationale for this design, which was the students' idea, was that when operators were busy packing, it was difficult and seemed inefficient to stop in the middle of the process to unjam a machine. With this new setup, the line operator's only duty would be to keep the lines running. The packing operators would be focused on packing only and not be pulled away for other tasks (Figure 2).

Queuing comes in

I asked the students, "Do you think this re-allocation of job functions will help? If so, how much will it help?" Here's where queuing theory comes in. Let me begin with a quick review.

Queuing theory deals with processes in which objects arrive, are serviced, and then leave. The theory helps determine measures such as process throughput, the average number of objects waiting to be serviced, and the average time objects will spend in the system. If you can make a gross assumption that the time between arrivals and service times are exponentially distributed, then you can approach the problem as a birth-death process, which makes the analysis pretty straightforward. The first step is to define the state of the system. The second step is to develop a state diagram, including the arrival rates (λ) and service rates (μ). The third step is to set up and solve the balance equations. The fourth step is to apply steady state probabilities to the problem.

Step 1. State definition. What are we interested in for this problem? Management wants to increase throughput. When the lines are running, widgets are being produced. The problem is that lines are not running half the time. This happens because the operators are busy packing widgets and let a machine stay

QUEUING THEORY DEFINED

The theory involving the use of mathematical models, theorems, and algorithms in the analysis of systems in which some service is to be performed under conditions of randomly varying demand, and where waiting lines or queues may form due to lack of control over either the demand for service or the amount of service required or both. Utilization of the theory extends to process, operation, and work studies.

Source: *Industrial Engineering Terminology*

PROPOSED LAYOUT

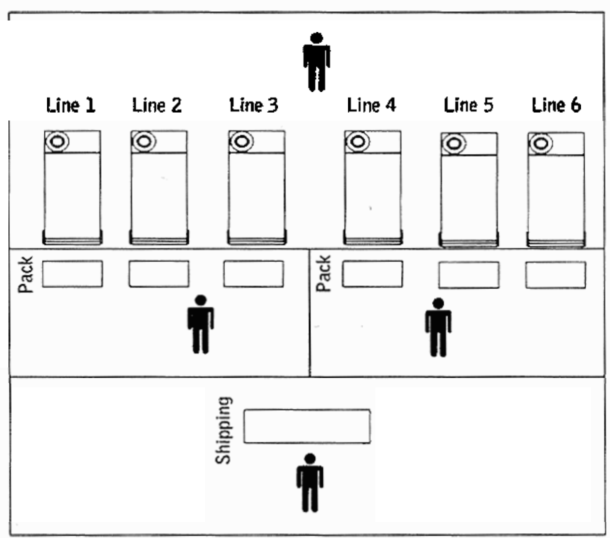

Figure 2. A re-allocation of job functions where one operator focuses on keeping the lines running, two operators focus strictly on packing, and one operator focuses on shipping

STATE DEFINITION AND DIAGRAMS

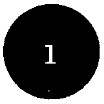

Figure 3. The first step in queuing theory is to define the state of the system. This is a diagram of the state of the system when one line is not producing.

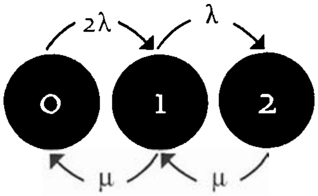

Figure 4. State diagram for the orginal layout (for a single operator and two lines)

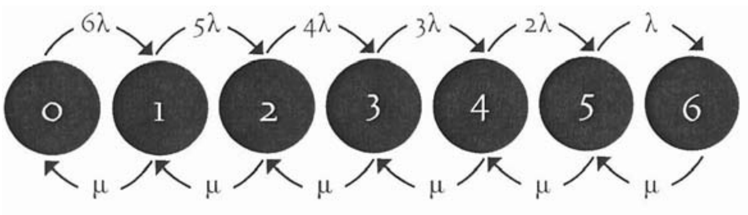

Figure 5. State diagram for the proposed layout

jammed longer than necessary or they are packing on one line when the other line needs to be set up. To determine the throughput of the process, we need a model that describes how many lines are running. We could choose to show either the number of lines running or the number of lines not running. In this case, the state definition will describe the number of lines not running (Figure 3).

The arrivals to this queuing system are when a line stops running and needs to be serviced. Services are when a machine is tended to by the operator in order to get the line running again. This service could be a changeover to a new product or fixing a machine stoppage.

Exponential distributions. Can we make the gross assumption that the time between arrivals and the service times are exponentially distributed? We can assume it is or we could plug the raw data into an input analyzer software package and evaluate the goodness of fit. However, when looking at this process and noticing that machines tend to jam at random intervals, the exponential distribution doesn't seem to be a bad assumption for the time between arrivals. Where you usually have to stretch the exponential assumptions is on the service times. However, in this case where frequently the operator will immediately respond and get the line running and in other cases where the line remains down for extended periods, the exponential fits rather nicely.

Step 2. State diagram. To model the current process shown in Figure 1, we need to consider only the throughput of a single operator. If more than one operator is in the process, throughput would just be a multiple of the number of operators. Since a single operator operates two lines, our state space would be no lines down, one line down, or two lines down.

The arrival rate (λ) represents the rate at which a line will stop running. For this process, the observed rate for a line is two stops per hour.

The service rate (μ) represents the rate at which a line will be restarted. Observing the service rate in this case was difficult since the person responsible for restarting the machine was performing other duties as well. However, since we know from historical records the probability that a single line is running, we could calculate the service rate.

Figure 4 fully describes the state space for a single operator and her two lines. The state diagram has a transition rate of 2λ between state 0 and state 1 because for state 0, two lines are running, each with a rate of λ for transitioning to a down state. Therefore, the transition rate from two machines running to one machine running is 2 times λ.

For the proposed method shown in Figure 2, a single operator is manning six lines. Since no action is being taken to improve the lines, the rate of line stoppages should remain the same (λ = 2 stops/hour). However, since an operator is dedicated to keeping the lines running and is not responsible for inspection and packing, the service rate should increase significantly to some unknown rate μ (Figure 5).

Step 3. Balance equations. Once the state diagrams have been created, we can use the concept of the conservation of flow to determine the steady state probabilities. The steady state probabilities (p_i) represent the probability of being in a given state. In other words, if you went to observe an operator and her two lines, what is the probability you would see both lines running (p_0), one line running (p_1), or no lines running (p_2)? The conservation of flow is simply that the flow in equals the flow out. Specifically, the flow is the probability of being in a state times the transition rate.

Balance equations for Figure 4	Balance equations for Figure 5
Node 0) $2\lambda p_0 = \mu p_1$	Node 0) $6\lambda p_0 = \mu p_1$
Node 1) $\lambda p_1 + \mu p_1 = 2\lambda p_0 + \mu p_2$	Node 1) $5\lambda p_1 + \mu p_1 = 6\lambda p_0 + \mu p_2$
Node 2) $\mu p_2 = \lambda p_1$	Node 2) $4\lambda p_2 + \mu p_2 = 5\lambda p_1 + \mu p_3$
	Node 3) $3\lambda p_3 + \mu p_3 = 4\lambda p_2 + \mu p_4$
	Node 4) $2\lambda p_4 + \mu p_4 = 3\lambda p_3 + \mu p_5$
	Node 5) $1\lambda p_5 + \mu p_5 = 2\lambda p_4 + \mu p_6$
	Node 6) $\mu p_6 = \lambda p_5$

To solve for the probability of being in a given state, replace one of the seven node equations above with $p_0 + p_1 + p_2 + p_3 + p_4 + p_5 + p_6 = 1$, substitute the value of 2 for λ, and solve the series of simultaneous equations.

$$p_0 = \mu^6/(\mu^6 + 13.2\mu^5 + 145.2\mu^4 + 1277.76\mu^3 + 8433.216\mu^2 + 37106.15\mu + 81633.53)$$

$$p_1 = 13.2\, p_0/\mu$$

$$p_2 = 145.2\, p_0/\mu^2$$

$$p_3 = 1277.76\, p_0/\mu^3$$

$$p_4 = 8433.216\, p_0/\mu^4$$

$$p_5 = 37106.15\, p_0/\mu^5$$

$$p_6 = 81633.53\, p_0/\mu^6$$

Step 4. Application to the problem. Now that we have the steady state probabilities, we can begin predicting the impact of the proposal. Recall that the primary measure of interest is line efficiency. The average efficiency collected from the machine controllers was 57 percent for the original layout. To obtain the average efficiency for the proposed layout, all you need is the steady state probabilities and the following equation:

Average line efficiency $= (6p_0 + 5p_1 + 4p_2 + 3p_3 + 2p_4 + 1p_5 + 0p_6)/6$

Since for each probability p_i in the efficiency equation, the service rate μ is unknown, we can perform a sensitivity analysis by calculating the throughput over a range of service rates. Figure 6 displays this sensitivity analysis. We see that with the proposed setup, the single line operator will have to have a service rate of about 8.75 services per hour (6.8 min./service) just to get the same throughput as the current configuration. But the students and plant personnel believe the service rate will be much better. With service rates approaching 17 services per hour (3.5 min./service), line efficiencies could exceed 80 percent.

SENSITIVITY ANALYSIS

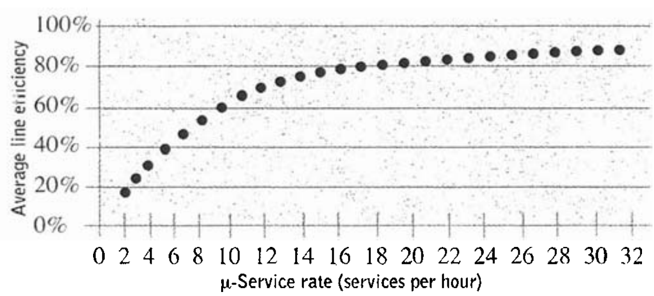

Figure 6. Sensitivity analysis depicting line efficiency vs. service rate

The rest of the story

The students' idea of re-allocation of the job functions sounded like a good one, and I anxiously awaited the results of their implementation. I was hoping they would become heroes with a 30 percent improvement in line efficiency, but I still would have been ecstatic if they achieved 20 percent.

The numbers started trickling in. After reorganizing the operators' responsibilities as proposed and without spending a dime, three operators (two packers and one machine tender) working six lines produced 4,036 parts per line the first day — well above the former rate of 3,050 and meeting the company standard of 4,000.

On day two, plant personnel added an additional operator (totaling three packers and one machine operator) and two production lines (totaling eight), resulting in 3,729 parts per line. On day three, they stayed with eight lines but added a fourth packer (totaling four packers and one machine operator), producing 4,838 parts per line. And on day four, staying with this same eight-line, four-packer, one-machine operator layout, 6,205 parts per line were produced.

Taking an average over 10 days, the production rate was 5,263 per line per day — a 72 percent increase in productivity over the earlier average of 3,050 parts per line! (Go to *www.iienet.org/magazine/oct05/schultz* to view the layouts for days one and three).

Because the production numbers are a function of the type of product being produced, a more accurate measure of improvement is the line efficiency captured by the machine's controller. The original average machine efficiency was 57 percent. Using the new layout, the efficiency jumped to 84 percent, a 47 percent improvement in machine efficiency.

When the dust settled, everyone lived happily ever after. The operators were proud of their results and happy that they were no longer taking heat for poor performance. Management was ecstatic over the productivity improvement, line efficiency numbers, and reduced cost per widget. The students became heroes and got their A's. And, most importantly, I now have a real live application of queuing theory.

Scott R. Schultz, Ph.D., is an assistant professor in the department of mechanical and industrial engineering at Mercer University in Macon, Ga., and consults for the Mercer Engineering Research Center. Schultz worked for Ford Motor Co. for 13 years and the Furniture Manufacturing and Management Center at North Carolina State University for two years. He is a member of IIE and ASEE.