*Proceedings of the 2001 Winter Simulation Conference*
*B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer, eds.*

# HOW "OVERSTAFFING" AT BOTTLENECK MACHINES CAN UNLEASH EXTRA CAPACITY

Robert C. Kotcher

Headway Technologies
A TDK Group Company
497 South Hillview Dr.
Milpitas, CA 95035 U.S.A.

## ABSTRACT

Using simulation, Headway Technologies predicted that increasing staffing among a group of already lightly loaded machine operators—"overstaffing"—would significantly improve throughput of its factory. This was counterintuitive since the operators already had significant idle time. Yet time studies confirmed that bottleneck equipment for which these operators were responsible was spending over 22% of its uptime idle solely due to lack of an operator. Analysis showed how this could be so: production equipment has a frequent and unpredictable need for operators, yet the operators must spend time away from the equipment tending to other demands of their jobs. A method of estimating the cost of this operator-induced throughput loss is described. The result shows how extremely profitable the hiring of extra operators is in such situations. A means of estimating the most profitable level of staffing is also described, along with several alternative solutions for reducing operator absences.

## 1 INTRODUCTION

Headway Technologies is a maker of read-write heads for hard disk drives. Read-write heads are tiny integrated circuits about as big as grains of pepper that magnetically read and write data onto and off of disks within a disk drive. The heads are fabricated inside cleanrooms on wafers just like computer chips, with about 20,000 identical heads per 6" wafer. The production process consists of over 400 visits to approximately 80 different pieces of production equipment and takes several weeks.

The factory is run 24 hours a day, seven days a week. This is typical of wafer fabs (computer-chip factories) because the bulk of their cost is in fixed facilities and equipment, meaning that the gross margin on each wafer is very high. This makes it cost effective to operate on nights and weekends, despite the higher operating costs. Lot size at Headway is 1, which is unusual for a wafer fab. The Headway fab employs about 60 equipment operators per shift and produces about 20 completed wafers per day.

Wafer fabs are among the most complex manufacturing operations in existence, for the following reasons:

1. Reentrant flow (i.e., most production equipment is visited several times by each wafer) precludes a neat, orderly, "production-line" arrangement of equipment; equipment is instead usually arranged by type, and wafers criss-cross the fab as they move from operation to operation.
2. High variability:
   a. Production equipment, being on the leading edge of technology, goes down more frequently and unpredictably than equipment in typical manufacturing operations
   b. High rework and scrap rates, due to the leading-edge nature of the technology and extremely tight product specs
   c. Almost weekly introduction of new products into the line
   d. A great variety of products in the line at any one time
   e. Short product life (a few months)
   f. High percentage of Engineering wafers (top priority, unique) in the line

Because of all of the above, capacity at any area varies quite a bit from hour to hour or even minute to minute, as does the volume and mix of incoming WIP.

Though engineers are constantly working to reduce the above variability, the leading-edge nature of chip technology will always cause more variability in wafer fabs than exists in almost any other type of production facility. In this naturally chaotic and complex environment, simulation is an ideal tool for developing improved operating methods.

## 2 HEADWAY'S PROBLEM

In early 2001, orders at Headway were ramping up and the fab was producing all it could. With demand exceeding capacity, and the high gross margin inherent in any product made in a

wafer fab, the profit increase possible if capacity could be increased was significant. The fab bottleneck was Headway's steppers. Steppers are machines that optically transfer the images of microscopic electrical circuitry onto each of the 20,000 heads in each wafer. Each stepper costs several million dollars and there is approximately a one-year lead time from order to production readiness, so buying additional steppers was not a solution to the immediate problem. The company needed to find a way to improve its steppers' throughput—and therefore its fab throughput—immediately.

Headway's Industrial Engineering department used its Factory Explorer™ fab capacity analysis and simulation model to look for such a solution.

## 3 USING STATIC CAPACITY ANALYSIS AS THE FIRST STEP IN SOLVING THE PROBLEM

Static capacity analysis is often the first step in using a factory model to find solutions to a problem. Compared to simulation, a static capacity analysis takes a fraction of the time to set up, run, and interpret. It can also be useful in pointing toward specific scenarios to simulate.

The fab model was set up to emulate the fab's current situation. A static capacity analysis was performed that showed that stepper "capacity loading" at the current throughput of 100 WGR ("weekly going rate"—good finished wafers per week) was a modest 56%, and photo operator capacity loading was only 42% ("photo" operators operate Headway's four steppers plus six other pieces of photo equipment). "Capacity loading" as used by Factory Explorer and throughout this paper is defined in Figure 1.

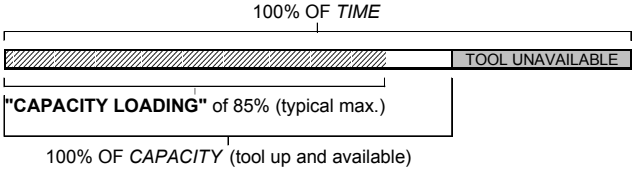


Figure 1: Definition of "Capacity Loading"

As you can see in Figure 1, "capacity" as used here is defined as the amount of product that a resource (machine or operator) could produce if, during its *available* time, it operated full time (with full batches each time if a batch tool). Because of all the variability in a wafer fab, a typical fab would never be run at close to 100% capacity loading because doing so would cause horrendously long queues and cycle times and massive WIP levels—with high associated costs. As a result, wafer fabs are usually not run at more than 80-90% capacity loading—this is deemed as the approximate level that optimizes profitability. On either side of this "optimum" loading, profits fall: if we scrimp on production equipment and allow the fab to be loaded more highly, we save on equipment but the higher cycle times and WIP levels increase our production costs in a variety of ways and can reduce our performance as a supplier, reducing sales; on the other hand, if we buy more equipment to reduce loading, we enjoy reduced costs of cycle time and WIP and perhaps increasing sales, but of course the extra equipment and cleanroom space is very expensive. Because of the intangibility of the cost of WIP and cycle time, no one can say for certain what the profit-maximizing capacity loading is for a fab (though Leachman et al. (1999) have come up with a clever way of estimating this), but 80-90% loading seems to be the consensus for most fabs. Fowler and Robinson (1995) define the capacity of a factory, given a maximum acceptable cycle time, as "cycle-time-constrained" capacity.

Headway set a goal of pumping up throughput 30%, to 130 WGR, without increasing cycle time. Figure 2 shows the capacity loading of photo operators at different staffing levels for this throughput target.

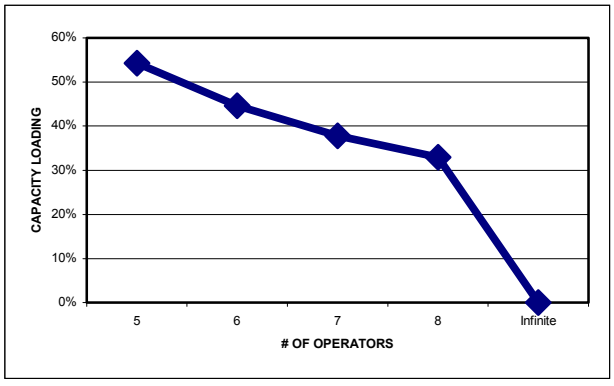


Figure 2: Capacity Loading of Photo Operators at 130 WGR

Figure 2 shows modest operator loading at 130 WGR even with the current staffing level of five. With steppers only loaded 73% at 130 WGR, and operators, 54%, the static capacity analysis was implying that 130 WGR should be obtainable without unreasonably long queue times (because, typically, machines loaded below 80-90% of capacity do not experience uncomfortably long queue times). However, the reality was that queue times were already uncomfortably high at only 100 WGR. It was obvious that things were happening in the fab that were causing queue times relative to capacity loading to be much higher than expected. A simulation was conducted to shed light on this.

## 4 USING SIMULATION TO ZERO IN ON THE PROBLEM AND FIND SOLUTIONS

A simulation was run with current machines, product mix, and staffing but at 130 WGR. The simulation began with WIP in the fab from a previous similar run, in order to reduce the duration of initialization bias. Initialization bias is the bias caused by the fab's starting WIP being different from the steady-state level that results from that run; starting the run with zero WIP creates an initialization bias that lasts for a long time.

The model was run for a ten-month period. The Factory Explorer™ software then plotted fab-wide average WIP and cycle time for each month. The end of the initialization bias could be seen as the point at which the cycle time and WIP remained almost stable for the remainder of the run—from month seven on in this run. So the last four months were deemed unaffected by initialization bias and their data were used in this analysis. Calculations were made of average cumulative queue time in front of the steppers per wafer out (i.e., total queue time per wafer for all 30+ visits to the steppers), as well as average % time in which the steppers were available and had WIP but sat idle because no operator was available.

The results showed that, at 130 WGR, despite the operators having 46% surplus capacity (100% - 54%), 19% of stepper capacity would be lost *solely due to no operator being present at the moment of need.* How could this be? This was because the operators had to spend a fair amount of time away from the steppers to tend to other equipment. And this of course was on top of operator "allowances"—accommodation in the model for the usual things that keep personnel in a factory from being 100% efficient: breaks, discussions with supervisors and engineers, etc. Allowances are counted as "unavailable" time so they are deducted before capacity loading is calculated. Nevertheless they sometimes cause machines with WIP in front of them to be idle, even when operator loading is modest.

While these factors make operators frequently unavailable for the steppers, the steppers require frequent attention to be kept running at full capacity:

1. They frequently stop and alarm until an operator comes and performs a manual alignment. This only takes a few seconds but the wait for operator response can be significantly longer.
2. With their short processing times and inability to easily form large batches (due to the great variety of products in the line, most of which require a unique reticle and/or exposure level for each of their 30+ visits), the steppers have a frequent need for loading and unloading.

These frequent—if brief—semi-random needs for operator attention, combined with the fact that operators were frequently—if briefly—required to be away from their machines, led Headway I.E.s to feel that increased staffing might boost throughput despite the modest overall current loading. To test this hypothesis, the simulation was re-run using an additional photo operator per shift, then two additional operators, and so on. The runs were conducted as described previously, except that each run was nine to twelve months in length (the longer ones being those with ultimate WIP and cycle times farther from the model's starting level, thus having a longer-lasting initialization bias). Each run yielded 3-6 "unbiased" months of data. Averages across the unbiased months were used to generate a single data point per run that was used to construct operating curves—curves which show the correlation between throughput and cycle time. These are shown in Figure 3 below.

Figure 3 shows how the model seems to confirm the I.E. Department's hypothesis. At 130 WGR, going from five operators to six operators per shift reduces collective queue time for the 30+ visits from about 24 days to ten days. Additional operators bring further benefit, though we start seeing diminishing returns.

Another way of looking at Figure 3 is to see what would happen if, using today's five operators, we increased stepper throughput from 100 WGR to 130 WGR. Collective queue time jumps from about four days to about 24. By adding two operators per shift, however, we could achieve this 30% increase with little increase in stepper queue time.

Figure 4 graphically displays an interesting statistic from the simulation that shows, of the time that steppers are up and WIP is present, what % of time the steppers are kept from processing by the lack of an available operator. In other words, this is the percent of tool capacity that would be lost solely due to a lack operators, at, say, 130 WGR.

Fab management had a hard time believing that, with operators loaded so modestly, machine idleness could be so high. Actual fab data were then obtained as part of a project by a team of U.C. Berkeley Industrial Engineering students who performed an analysis of Headway's Photo area as their senior project. They conducted 90 hours of time studies and calculated, among other things, the percent of time in which the steppers were available and had WIP but were idle due to no operator. The figure they came up with was 22.5% (Han, Le, and Yan, 2001). Thus the model seemed optimistic in predicting a 19% number if production were increased to 130 WGR. This is not surprising given that there are nuances of the real-world operation that are not captured in the computer model. For example, in the real world, operators often need to leave their machines to get reticles for upcoming operations (reticles are what the steppers shine light through—like picture slides—to project the image of the electronic circuitry onto a wafer). Also, operators are frequently required for manual alignments. The *overall loading* on operators for manual alignments *is* captured in the model in that, after loading and starting a stepper, operators are "held" there for 13% of the processing time. This is the percent of time, based on analysis of stepper computer logs, which processing time is extended by waits for manual alignments. This is modeled pessimistically in that, in real life, not all of this time uses the operators—some of it is merely wait ing time. But the model is *optimistic* in the way that the Factory Explorer™ software bunches all of this time at the beginning of processing, when the operator is already there, having just finished loading. In real life, this operator assistance time is sprinkled sporadically through processing, thus creating the possibility that an operator will not be available at that time.

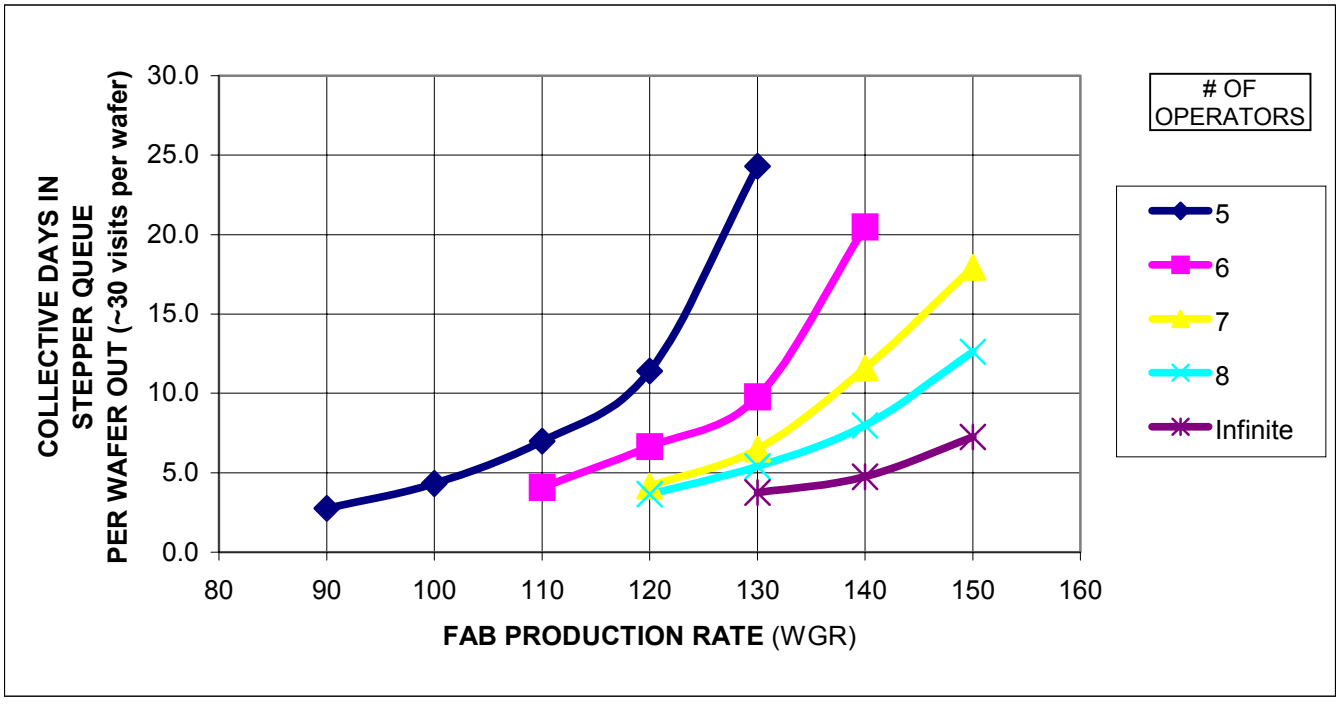


Figure 3: Operating Curves for Photo Operators

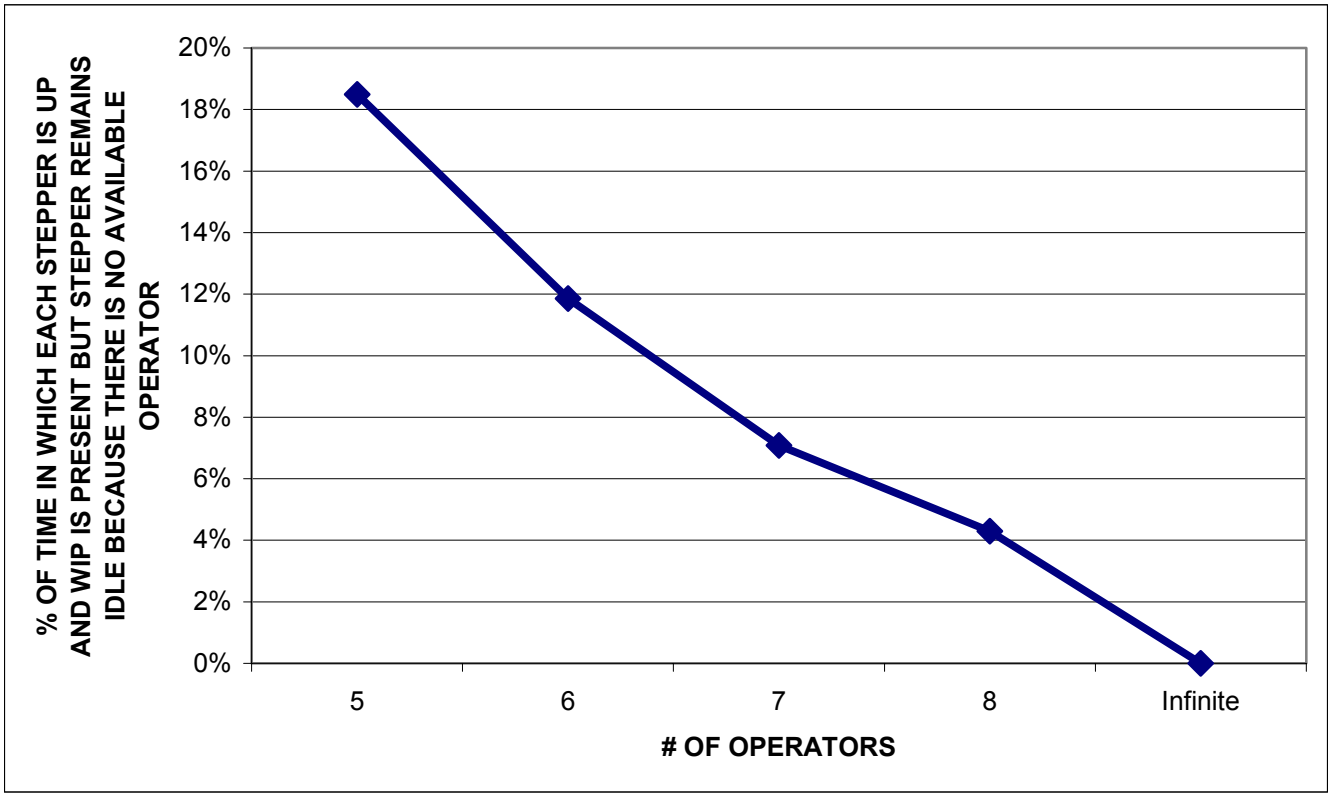


Figure 4: At 130 WGR, Stepper Capacity Lost Due to No Operator

Comparing Figures 2 and 4 shows how, even at low operator loadings, the steppers still lose many percentage points of capacity due to a lack of available operators. Why? This can be explained by our old friend (actually, enemy…): variability. For example, even though there will be slow periods with much operator idle time, during peak periods all ten tools for which these operators are responsible (the four steppers plus six other pieces of photo equipment) will have WIP in front of them and will compete for the five operators' attention. This is exacerbated during meal and snack breaks, which the model staggers as in real life but which still leave us with only half a staff for about a quarter of each twelve-hour shift. These factors explain how, even at light operator loadings, we can still lose 19+% of stepper capacity solely due to lack of an operator at the time of need.

## 5 ESTIMATING THE MOST PROFITABLE STAFFING LEVEL

Figure 3 shows how simulation runs predicted that adding a sixth photo operator in each shift would allow stepper throughput to be increased by 10 WGR without increasing cycle time. Adding a seventh operator each shift would allow throughput to be increased by about another 10 WGR without increasing cycle time. Additional operators beyond seven would provide steadily less improvement.

But…would it be *profitable* to invest in these additional operators? One way of looking at this is to compare the throughput increase obtainable through the addition of operators and compare that to the throughput improvement obtainable through the purchase of an additional stepper (keeping the operator-to-tool ratio constant). Using this approach, if it costs $30,000 to employ an operator for a year, two operators for each of the four weekly shifts for a year comes to $240,000 to increase the capacity of the fab bottleneck—and therefore, the entire fab—by about 20%. This is almost as much capacity as would be gained by the purchase of an additional stepper, but for several million dollars less! Even allocating the costs of a stepper over, say, a five-year useful life, the annual operator expense is only a fraction of the annual contribution of stepper cost.

Another perhaps more direct way of estimating whether it's profitable to add a certain number of operators is to look at what this does to the output of the fab, and then add up the revenues that the increased output will bring us. In a situation such as Headway's—in which demand exceeds capacity and the fab can sell any additional wafer it can produce—extra throughput at the bottleneck machine translates directly to additional sales. In a wafer fab, as discussed earlier, the low direct costs per wafer make each additional wafer produced highly profitable. *The Goal* (Goldratt & Cox, 1984) vividly portrayed how, since the bottleneck machine dictates the output of the entire factory, any increase in its throughput translates directly to an increase in *factory* throughput. When the machine is not processing, it's like the entire factory is *frozen*. Conversely, every additional minute in processing we can gain from it translates directly to increased output for the factory *as a whole.* Headway was in the perfect situation to maximally benefit from an increase in throughput in its bottleneck machine: (1) it could sell all additional output it could produce; (2) its product had a huge gross margin. These two factors dramatically increase the profitability of producing one additional wafer. Say, for example, that Headway, by reducing operator-induced idleness, increases the amount of hours that each stepper is processing by 10%, from 15.3 to 16.8 per day, and this increases throughput by 10%, from 100 WGR to 110 WGR, or 1.43 additional wafers per day. Assuming a gross profit per wafer of $20,000 (for proprietary reasons, this is not the actual number), our increased revenues per day are 1.43 x $20,000 = $30,629. This is for 6.0 hours of additional processing time ([16.8 – 1.53] x 4 steppers), so we can calculate that each of the four steppers, when operating, is generating revenue at the rate of $7657 an hour ($30,629/4). This means that, when a stepper has WIP but is idle, it is costing us $128 *per minute* in the form of lost revenue! Eliminating these $128 minutes can justify a lot of $15-an-hour operators!

The above calculation assumes that wafer starts are directly related to stepper throughput. Using Goldratt & Cox's terminology, the steppers are the "Herbie," the wafer-start process is the lead hiker, and the steppers and wafer-start process are connected by a "rope." In real life, however, Headway's "Herbie" shifts around from day to day and hour to hour, and also, stepper queue time isn't the *sole* influencer of start rate. However, whether the $128 number is actually $150 or $50 an hour, it is still gargantuan compared to the 25-cent-a-minute cost of an operator ($15/hour divided by 60).

Looking at the impact over a full year is even more impressive: the additional ten wafers per week bring the company $200,000 a week in additional revenue, or $10.4 million a year! For $10.4 million a year we can justify hiring *quite a few* more operators!

This is a vivid example of the power of Goldratt & Cox's theory of constraints.

## 6 ALTERNATE SOLUTIONS

The Headway model showed how increasing the number of operators can be extremely profitable, even in operator groups that are modestly loaded. But better yet would be solutions that attack the root cause—eliminating the *frequency* and *unpredictability* in need for operators in the first place—or use operators more efficiently to get the benefits of greater staffing without having to hire *so many* additional operators. Headway industrial engineers and the

U.C. Berkeley students came up with the following recommendations.

1 REDUCE THE NEED FOR MANUAL ALIGNMENTS
These alignments are especially harmful to throughput since, unlike loading and unloading, they occur *unpredictably*, increasing the probability that no operator will be present at the time of occurrence.

2. REDUCE THE FREQUENCY WITH WHICH OPERATORS ARE NEEDED FOR LOADING/ UNLOADING
In their newest products (which currently comprise only a small percentage of production), Headway photo engineers have been successful in reducing the variety of reticles and exposure levels. This increases the average batch size, thereby increasing run time per batch and decreasing the frequency with which operators are needed for loading and unloading.

3. REDUCE NEED FOR OPERATORS TO BE AWAY FROM THEIR MACHINES
Only operators at their machines can respond *immediately* to the need for manual alignment or loading/unloading. Headway is looking into:
   a. Employing reticle stockers and other means of making reticles quickly and locally accessible to all operators.
   b. Invoking the following operating rules:
      A. If a stepper is up and there's WIP, it must *always* have an operator physically in front of it, 24 hours a day, seven days a week (if for no other reason than to respond *instantly* when manual alignment is needed).
      B. Each stepper should have one "secondary operator." This person's duties are as follows:
         1. Do everything involving the stepper that cannot be done *at* the stepper, including:
            - Getting reticles.
            - Transporting WIP
            - Going away from the tool to get needed help or information (from lead, engineer, maintenance, etc.)
         2. When everything is taken care of *away* from the stepper, help the operator *at* the stepper.
         3. Fill in for the primary operator when he/she is on break.
      C. Before any batch finishes, there *must* be, at the machine:
         1. The next batch of wafers, swabbed, inspected, and ready to load.
         2. The reticle (if it's not in the machine already).
      D. Use a pre-break checklist. During breaks, we will only have *one* operator for each stepper. To keep from losing fab capacity during breaks, the primary and secondary operators must make sure of the following before either goes on break:
         1. The stepper has WIP and reticles in front of it (or in it) to last to the end of the break.
         2. The remaining operator is ready to spend the entire break in front of the machine.

## 7 CONCLUSION

The most profitable level of staffing can be significantly higher than intuition and static capacity analysis suggest if the following conditions are in effect:

1. The machine is the factory bottleneck
2. The machine has *frequent* (if short) and *unpredictable* need for operators
3. Operators must frequently be away from machines (even if for short periods of time)

Good responses to this situation are:

1. Increasing staffing levels, perhaps significantly (simulation can help estimate optimum levels)
2. Reducing the *frequency* of need for operators.
3. Reducing the *unpredictability* of need for operators.
4. Adopting operating rules for operators to ensure that, when WIP is present and a machine is up, there is always an operator in front of it. [NOTE: *Implementing* this is relatively easy, but it's only the beginning. The difficult part *is maintaining* the discipline in operators to strictly adhere to procedures that, to the average person, don't make a lot of sense ("I only let the machine sit idle for a couple minutes an hour—what's the big deal?"). This can be achieved through recurring "refresher" courses on the importance of keeping the bottleneck machines processing and the *huge* cost ($128 per minute?) of having them idle. It may also be profitable to put your sharpest, most disciplined (and probably most highly paid) operators at the bottleneck machine.]

## 8 IMPLEMENTATION

The Berkeley students' independent reaching of the model's conclusion—that additional staffing at photo would be highly profitable—began to persuade fab management of the benefit of more operators. But at that time Engineering saw it fit to relax an important test wafer requirement, which freed up enough capacity to eliminate steppers as the fab bottleneck. Therefore, a staffing increase was not implemented. The company did succeed in virtually eliminating the need for manual alignment, and it also increased average batch size by greatly reducing the variety of reticles and stepper exposure levels used by Headway products. Nevertheless, capacity analyses using the latest demand projections and Headway's most leading-edge products project steppers becoming the bottleneck again in 6-12 months. Fab management has committed to meeting throughput goals by incorporating improved operating rules and "overstaffing" the steppers as necessary to reduce idle time.

## ACKNOWLEDGMENTS

The author would like to thank the following for their assistance: Headway industrial engineer Sook-Tying Choong, who ran many of the simulations and compiled the results, Drs. Frank Chance and Jennifer Robinson of FabTime, Steven Brown of Arizona State University, Assistant Professor Scott Mason of the University of Arkansas, and U.C. Berkeley students Steven Han, Robert (Hoang) Le, and Sunny Yan.

## REFERENCES

Fowler, J. W., and Robinson, J. K. 1995. Measurement and Improvement of Manufacturing Capacity (MIMAC) Designed Experiment Report. SEMATECH Technology Transfer #95062860A-TR.

Goldratt, E. M., and Cox, J. 1984. *The Goal.* Croton-on-Hudson: North River Press.

Han, S., Le, R. H., and Yan, S. 2001. Improving Photolithography Throughput Without Increasing Cycle Time. Student project for U.C. Berkeley IEOR Department class, University of California, Berkeley.

Leachman, R. C., Plummer, J., and Sato-Misawa, N. June 1999. *Understanding Fab Economics.* Available online: <http://esrc.berkeley.edu/csm/csmab.html>

## AUTHOR BIOGRAPHY

**ROBERT C. KOTCHER** received his B.S. in Industrial & Systems Engineering from San Jose State University, San Jose, California, and his MBA from Santa Clara University, Santa Clara, California. He was recently Manager of Industrial Engineering at Headway Technologies. His e-mail address is <BKotcher@Compuserve.com>.