FabTime Cycle Time Management Newsletter

Volume 1, No 6, September 2000

Information

Mission: To discuss issues relating to proactive wafer fab cycle time management.

Publisher: FabTime Inc.

Editor: Jennifer Robinson

Table of Contents

- Welcome
- Main Topic – Performance Measures
- Community News/Announcements
- Recommendations and Resources
- Current Subscribers

Welcome

Welcome to Issue #6 of the FabTime Cycle Time Management Newsletter. It's hard to believe that it's been half a year already since we started the newsletter. We're now up to 200 subscribers, from 72 companies and universities, plus several independent consultants. Motorola continues to have the largest number of subscribers (24), with Intel and AMD also in double-digits. This month's topic is performance measurement in wafer fabs. Moves, utilization, cycle time, OEE, and turns, to name a few, are used in various fabs. We've observed that different people sometimes have slightly different definitions for these measures. We also discuss when, and why, these metrics sometimes conflict, and our opinions of which measures are most valuable.

Performance Measures Used in Wafer Fabs

Introduction

Let's begin with the obvious: wafer fabs cost a lot of money. Fab managers, therefore, are constantly under pressure to run them well, so that the huge investment in capital equipment is not wasted. But what does it mean to run a wafer fab "well"? In an ideal world, we would be able to keep all of that expensive equipment highly utilized, with the utilization dedicated completely to productive work. At the same time, we would have low and predictable cycle times, and a minimal amount of capital tied up in WIP. We would keep

FabTime

325M Sharon Park Dr.
#219
Menlo Park, CA 94025
Tel: 408 549 9932
Fax: 408 549 9941
www.FabTime.com

our operators busy and effective all of the time, so that we weren't wasting salary on having people stand around the fab. We would constantly improve our products, yet always maintain 100% line yield. We would keep costs down, but be able to charge high prices by having speedy time to market.

Of course this combination of circumstances is impossible for many reasons. A wafer fab, as we discussed in the early issues of this newsletter, is a highly variable environment. In the presence of variability, high utilizations lead inevitably to high cycle time and WIP. You can load your operators and your tools heavily, or you can have low cycle time and WIP. You can't do both, unless you stamp out variability.

So the question is, what performance metrics should a fab manager use to make sure things are on track? Utilization, cycle time, OEE, turns, moves? And after deciding which to use, what are the correct definitions to use for these metric? We have observed, during our years of consulting, that different people often define the same metric differently. This is a source of confusion when comparing performance between or within companies. When people talk about utilization, for example, there are several things that they might mean. Similarly for turns. We therefore are proposing some definitions to apply within our niche of cycle time management.

Definitions

Here are some performance measures commonly used in wafer fabs:

Starts: Number of wafers started into the fab during a particular time period (e.g. 5000 wafer starts per week). Starts are often used as an indicator of the size of a fab. Is it a 5000 WSPW fab or a 500 WSPW fab?

Utilization: The percentage of time that a piece of equipment is busy (Utilization #1). "Busy" in this context is usually taken to include both productive and non-productive time. So a tool with a utilization of 85% is idle 15% of the time, and in some way busy the other 85% of the time, whether it is down, in setup, or processing wafers. Factory utilization is typically reported as the utilization of the bottleneck tool (the tool with the highest utilization).

Another way of defining utilization (Utilization #2) is to divide actual throughput during a time period by the maximum throughput that the tool can sustain, where the latter is adjusted to account for downtime. This is also referred to as capacity loading. To understand the difference between these two ways of computing utilization, consider a tool that spends 12 hours a day down for maintenance, six hours processing, and 6 hours idle (in a fab that operates 24 hours per day).

By our first definition, the tool is busy 12 hours (down for maintenance) plus six hours (processing) for a total of 18 hours. The utilization is:

Utilization #1 = busy time/total time = 18 hours/24 hours = 75%

However, if the 12 hours a day of maintenance are unavoidable, then the most time we could spend processing on the tool is 12 hours per day. Suppose that we can process one wafer per hour on the tool, when the tool is up. Then, if the tool is processing for six hours, it processes six wafers. And if the most time it can spend processing is 12 hours (because it's down for the other 12), then the most wafers it can process per day is 12. Then the utilization is:

FabTime

Cycle Time Management Newsletter

Volume 1, No. 6

Utilization #2 = current throughput per day/maximum throughput per day = 6 wafers/12 wafers = 50%.

So, is this tool 50% utilized or 75% utilized? You begin to see the problem. Because utilization can be calculated either way, some standards such as the Semi E10 equipment guidelines do not use utilization at all. The Semi E35 Cost of Ownership standard uses even a slightly different calculation from those described here. Utilization is a commonly reported metric - we just advise that you understand how it is calculated.

Overall Equipment Effectiveness (OEE): OEE is a measurement of equipment productivity defined by SEMATECH. It separates equipment productivity into three categories of corrective actions: availability, performance, and quality, according to the formula: OEE% = (availability) x (performance efficiency) x (rate of quality) x 100. Availability is the percentage of total time left after accounting for both scheduled and unscheduled downtime ((total time - downtime)/(total time)), where downtime includes non-scheduled time such as shutdowns. Performance efficiency is a factor consisting of rate efficiency (ideal process time over actual process time) times operational efficiency (time spent processing vs. time available for processing). Rate of quality is simply good wafers processed divided by total wafers processed. The idea behind OEE is to clearly show the reasons why equipment is not fully effective, so that they can be tackled via improvement projects.

Turns: Operation moves (for a factory, area, toolgroup, or operation) during a time period, divided by the WIP at the beginning of the time period. This is analogous to operation throughput divided by WIP, which (from Little's Law) is the inverse of operation cycle time. For example, if the factory turns measure is four per day, this means that on average, each lot passes through four operations per day (operations take six hours). As a more detailed example, suppose that we have a toolgroup that starts a 12-hour shift with 24 wafers in queue. During the shift, say it processes 72 wafers (the 24 that were in queue to start, plus 48 others that arrived during the shift, and were processed). Then the toolgroup turns ratio is:

Turns = operations moves / starting WIP = 72 wafers / 24 wafers = 3 turns.

To understand the need for a metric like turns, suppose that the above toolgroup normally processes 240 wafers per shift. We can see that during a shift when it only processes 72 wafers, the throughput is way down. Does that mean we should go speak to the operators, to see why they've been slacking off? Not if the throughput is down because only 72 wafers came to the toolgroup. Turns will likely be high for the next shift, because starting WIP (the denominator) will be very low. But if we had 240 wafers sitting there at the start of the shift, and only processed 72, then the turns would be 72 / 240 = 0.3 turns. Decreasing turns alert management to problems, lower than expected moves and/or higher than expected WIP. If throughput remains constant, but WIP is building, for example, turns will be decreasing.

> *"Decreasing turns alert management to problems"*

One reason why we're being very careful about the definition of turns is that there can be confusion between this operational definition of turns and the common accounting term "inventory turns", which is defined as annual sales dollars divided by on-hand inventory dollars. Different types of industries turn over their inventory at different rates.

Throughput: At the factory level, throughput is recorded as number of good wafers out during some time interval. For example, 500 wafers per week. For a piece of equipment, or an operation, throughput is usually just the number of wafers that complete processing on that tool or operation during a given time period. Equipment-level throughputs are often expressed in units per hour, or UPH. Factory throughput is equal to start rate multiplied by line yield for a particular time period. Some fabs track the number of wafers that complete processing in, or move from, a particular area or toolgroup, and call that measure either moves or activities. Moves are much like throughput, but are generally used as a more aggregate measure. For example, we might speak of an individual stepper's throughput, but of the entire photo area's number of moves per day.

Line Yield: (Wafers started - wafers scrapped)/(wafers started).

Line yield is a very common and well-known performance measure. Note, however, the difficulty of defining line yield on a current-condition basis... if you look backwards in time to a week when all the wafers started during that week have either been shipped or scrapped, you can calculate line yield as above. But that is a trailing measure of what is actually happening in the fab, because it takes many weeks before all of the wafers that start are scrapped or shipped. Other possibilities for a leading measure of yield include:

(weekly ships) / (weekly ships + weekly scraps)

(weekly wafer moves that did not result in a scrapped wafer) / (total weekly wafer moves)

Cycle Time: Usually reported only for good wafers shipped, cycle time is the time from when the wafer is released into the fab until it completes processing. Cycle time includes both process time and queue time.

Cycle Time/Raw Process Time: Cycle time is often reported in terms of the ratio of total cycle time to theoretical (also called raw) process time. Theoretical process time is the time that it would take to process a single wafer if it experienced no delays. This ratio of cycle time to raw process time is often called an X-factor. For example, a lot with a cycle time of three times the theoretical cycle time is referred to as having a cycle time of 3X.

Cycle Time Per Layer: Usually calculated as total cycle time for a product, divided by the number of layers for that product. This gives a metric that can be used to make comparisons across products of different levels of complexity.

"the time from when the wafer is released into the fab until it completes processing"

Discussion

Here are some of our thoughts on these performance measures. We would love to hear your thoughts, too. Send them to Jennifer.Robinson@FabTime.com.

Starts: Starts are generally a good thing to watch because they are one of the few places in the fab where management is totally in control of variability (or at least the starts planner is!). Thus starts as a metric is fundamentally different from things like equipment downtime or yield busts, which you don't directly control (although you can take measures to lessen their likelihood). However, striving to increase starts to a level that is unrealistic leads to problems. By increasing starts beyond what the factory can handle, all that you accomplish is increasing WIP and cycle time. A fab has a maximum sustainable throughput rate, determined by the capacity of the bottleneck. If the start rate is more than the bottleneck can handle (after accounting for pre-bottleneck yield loss), WIP just piles up in front of the bottleneck. Throughput will not increase. However, linearity (smoothness) of starts can be a useful measure. This is because variability in arrivals is known to increase average cycle times (See, for example, L. M. Wein, "Scheduling Semiconductor Wafer Fabrication," IEEE Transactions on Semiconductor Manufacturing, Vol. 1, No. 3, 115-126, 1988.)

Utilization: Fabs are often driven to increase utilization across all of the tools. The reason for this is obvious. When you spend $5 Million for a tool, every minute that it spends idle feels like lost money. If you have a 1000 wafer start per week fab, and you need a particular tool that has a per-tool capacity of 1000 wafers per week, you don't want to have to buy a second expensive tool, and have both of them sitting idle for half the time. From a pure cost-accounting perspective, this is crazy. The situation is made worse by the fact that different types of tools have different per-tool capacities. So, even if you only want to process 100 wafers per week, there is bound to be some tool that you need that is designed to process 5000 wafers per week. You still need to buy one for process reasons, but it's going to sit idle most of the time. This is referred to as non-granularity of equipment.

But even putting granularity aside, we know from Frank's discussion last month on JIT and TOC that you wouldn't want to run a fab in which all of the tools were

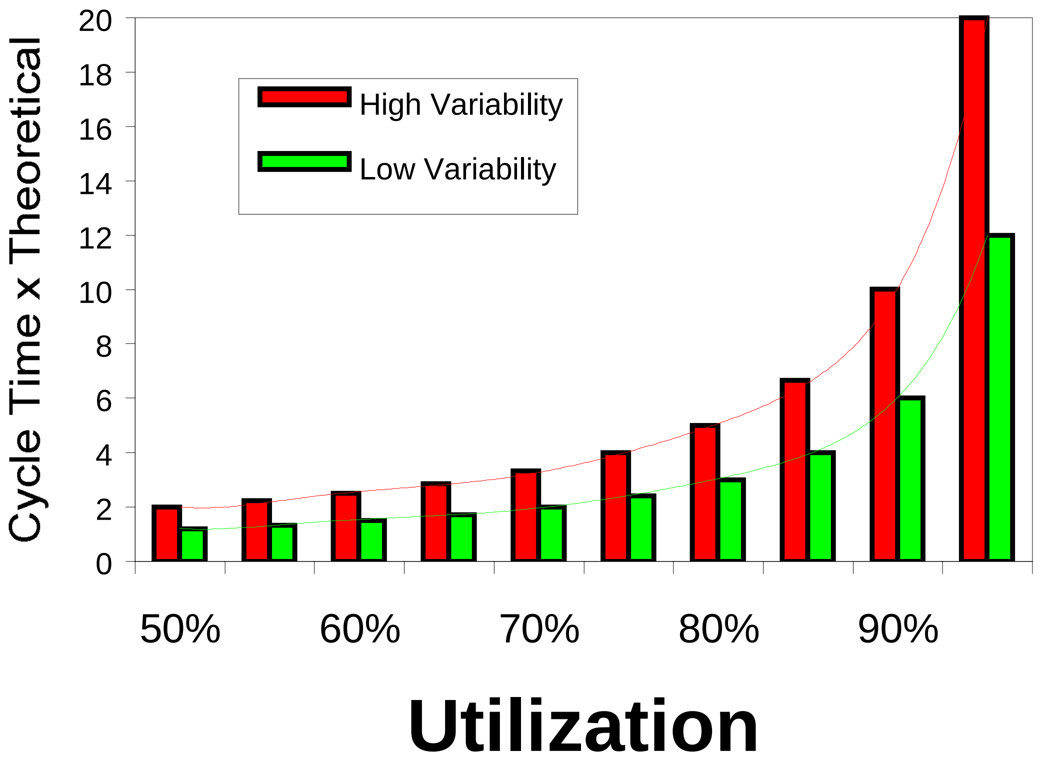

Figure 1: Illustration of the relationship between cycle time, utilization, and variability. By reducing variability, we can move from the red curve to the green curve.

FabTime
Cycle Time Management Newsletter
Volume 1, No. 6

loaded to 99% of capacity. With all the variability in the fab, you would have essentially infinite cycle time and WIP. The quickest way to reduce cycle time and WIP is to increase capacity at individual toolgroups, which means decreasing their utilization. It's basically a catch-22 -- high utilizations and low cycle times are in direct opposition in an environment with variability. This is due to physical laws, and no amount of threats on the part of accounting people can change the basic relationship. Some additional observations on this subject:

- No matter how low their utilization relative to the other tools in the fab, one-of-a-kind tools tend to have a disproportionately adverse effect on cycle times. The reason is simple. If the tool goes down, everything stops. And if everything stops for a time, bad things happen. The bottleneck can be starved, and time lost on the bottleneck can never be recovered. Then when the one-of-a-kind tool goes back up, it quickly goes through the waiting pile of WIP, thus sending a big WIP bubble through the line. All this variability drives up cycle time and WIP, sometimes for much longer than you would imagine.

> *"you cannot increase throughput beyond the capacity of the bottleneck"*

- You can sometimes get away with higher utilizations on toolgroups that have lower variability. If a tool has exceptional uptime, and no setups or rework, and if all of the wafers that go through the tool have about the same process time, then you can probably operate the tool at a higher utilization than that of other tools before running into cycle time problems.

- In general, reducing variability will always help you to reduce cycle times. Because the relationship between cycle time and utilization is highly non-linear (cycle time goes up much more rapidly when utilization is high), you will see the most overall cycle time benefit from variability reduction programs applied to high utilization tools.

- It makes sense, where possible, to relate planned utilization to per-tool cost. Fabs commonly plan for higher utilizations only on the more expensive tools, while allowing for more buffer capacity on cheaper tools such as metrology tools.

OEE: OEE is a useful performance metric, because it generally tries to get you to do the right things. Improve equipment uptime, reduce setups, reduce rework, improve yields, etc. We do, however, question one element of the Operational Efficiency factor that's part of Performance Efficiency. The idea is to discourage minor stoppages and idling, which certainly makes sense. But as we interpret it, standby/idle time is always considered bad, always hurts your OEE. This includes time that the tool is starved due to having no WIP to process. While this is a reasonable thing to guard against for bottleneck tools, other tools (for reasons just discussed) benefit from some planned idle time. Or rather, the overall factory cycle time benefits from allowing this idle time at certain tools. Perhaps for non-bottlenecks, a simplified variant of OEE might be proposed. Something like:

(Actual throughput of non-scrapped, non-rework wafers per time period)

(Theoretical maximum throughput, if the tool had no non-productive time (no downtime, setups, etc))

This would still drive management to

improve uptime, reduce setups, reduce rework, and improve yields, but would not penalize tools that were idle due to lack of WIP, because of lower planned utilization.

Turns: Turns can be a nice metric to look at for two reasons. One, turns give us an idea of what the fab cycle time will be in the future, by estimating the rate at which we are currently processing WIP. For example, if we have a fab with a current turns rate of 8 operations/day, and a single product with 400 operations, then we can expect the cycle time of lots currently in the fab to be about 50 days. During a ramp, this turns-based estimate can be more reliable than simply looking at the cycle time of lots that recently exited, because the fab is moving into a higher-utilization, higher-cycle time region.

The second useful thing about turns is that they can highlight situations in which we have good cycle time, but still have a problem. For example, if a tool is having a throughput problem, it might be building WIP while only processing the most critical hot lots. The hot lots have low cycle times. Since the hot lots are the only ones getting through, the average cycle time looks good. However, the turns will be poor, because the WIP is increasingly high. If turns are lower than expected, either WIP is building, or operation moves (activities) are less than expected. Either situation should be brought to management's attention.

"any fab that can produce lots at two times theoretical process time is doing really well."

Throughput: Clearly, increasing throughput is better than increasing starts, because throughput implies finished wafers that can be sold. You cannot increase throughput beyond the capacity of the bottleneck (maybe briefly in the short term, especially if you are flushing WIP from the end of the line, but not in a sustainable manner). Therefore attempting to maximize the throughput of the factory tends to make sense. Maximizing the throughput of all of the individual tools, however, doesn't make sense. Ideally, you want the throughput of each tool group to be exactly what is required so that the bottleneck (or bottlenecks) gets the WIP that it needs when it needs it, and is never starved.

The best thing that you can do for overall cycle time is to keep the variability in the throughput of each toolgroup as low as possible. For example, we worked on a study with Infineon Technologies (see http://www.fabtime.com/abs_Siem98.htm) in which a particular back-end operation had a much higher throughput than the overall capability of the assembly factory. This operation was thus run during only two of the factory's three shifts. Spreading the same volume across all three shifts led to an 8% overall cycle time decrease in a simulation model, because the change significantly reduced variability to downstream operations. This recommended change was later implemented into the factory, and contributed to reduced actual cycle times.

Line Yield: Line yield is outside of our area of expertise, except insofar as it impacts cycle time. If your line yield is poor, or yield is variable, that can negatively impact cycle time in a number of ways. First off, poor line yield means more starts, higher utilization of equipment than would otherwise be required, and thus increased cycle times. Variability in yield (yield busts) can lead to WIP bubbles as batches of wafers are started from the beginning of the line (or some intermediate staging point) and rushed through the line to make up for the lost wafers. These expedited wafers cause additional variability (increased setups, queueing for non-expedited wafers), which in turn adds to overall cycle time.

Cycle Time: The purpose of this newsletter is to discuss issues related to cycle time management. Obviously, we think that managing cycle times is important. There are many benefits to reducing cycle time, including decreased WIP, reduced likelihood of obsolescence losses, and improved cycles of learning. Some have argued that lower cycle times improve line yields - the longer things are in the fab, the greater the possibility of contamination. (See, for example, K. Srinivasan, R. Sandell, and S. Brown, "Correlation Between Yield And Waiting Time: A Quantitative Study," Proceedings of the Seventeenth IEEE/CPMT Symposium, Austin, TX, 65-69, 1995).

Cycle time over raw process time and cycle time per layer are both particularly useful cycle time metrics, because they let you make comparisons across products. No matter how complex the technology is, or how many layers are included, any fab that can produce lots at two times theoretical process time, for example, is doing really well.

One final point about improving cycle times is that you add cycle time throughout the process flow, and thus you can improve overall cycle times by making improvements at both bottleneck and non-bottleneck toolgroups. This is one area in which we think that Theory of Constraints is somewhat oversimplified. TOC would have you focus only on improvements at the bottleneck. However, any cycle time improvement at operations that take place after the bottleneck lead to a direct reduction in lot cycle times. Improvements before the bottleneck can also help by smoothing flow to the bottleneck, and reducing the possibility of starvation.

Conclusions

Some of the performance metrics defined here are trailing performance measures that tell you what your performance was like in the recent past. For example, shipped lot cycle time reflects the status of the factory throughout the weeks during which the recently shipped lot was processed. Other metrics, such as turns and daily throughput, are leading measures that give you a more clear idea of where problems are right now. Both types are important. Trailing performance measures help you to figure out what happened, so that you can learn from your experience. Leading measures define more immediate opportunities for solving problems, without giving you the same historical perspective.

A fab is a complex environment, with many things happening at the same time. As such, it makes sense to have different metrics for different situations. And it's important in applying performance measures to start out with definitions that are clear (e.g. utilization). It's also important to understand that sometimes metrics can conflict with one another (e.g. cycle time and utilization). It's up to the fab manager to understand how these different metrics relate to one another, and how they are defined, so that the best balance between them can be achieved. That's our idea of what a "balanced line" should be. We'd love to hear your thoughts, too.

Community News

Call for Papers - SMOMS '01

We recently received a call for papers for the 2001 International Conference on Semiconductor Manufacturing, Operational Modeling, and Simulation (SMOMS '01), to be held April 24-25, 2001 at the Renaissance Madison Hotel in Seattle, Washington. The target audience for this conference is very similar to the audience for the MASM 2000 conference, which we wrote about in an earlier issue. MASM will return in 2002. The conference announcement reads:

"The electronics industry is now the second largest basic industry (behind agriculture) and the fastest growing manufacturing industry in the world. At the heart of this industry is the manufacturing of integrated circuits, or semiconductor devices. In the past, all that was necessary for a semiconductor company to make money was for them to design a good product. However, over the last decade, increased competition had led to the need for semiconductor companies to also be able to manufacture their products in an efficient and cost effective manner. Increasingly, these companies have turned to data intensive operational modeling and analysis tools and techniques because of their potential to significantly improve the bottom line. This conference intends to be a forum for international efforts to meet those needs."

The program chair for SMOMS '01 is John Fowler, of Arizona State University. The program co-chair is Tae-Eog Lee of KAIST. Other program committee members include Joerg Domaschke from Infineon Technologies, Mani Janakiram from Intel Corp., Robert Leachman from UC Berkeley, and D. B. Perng from Natl. Chung Tung University of Taiwan. If you are interested in submitting a paper for this conference, send an abstract to John Fowler (john.fowler@asu.edu) by October 31st. Acceptance notification will be by November 30th, with final papers due January 11th.

■ FabTime welcomes the opportunity to publish announcements for individuals of companies. Simply send them to Jennifer.Robinson@FabTime.com

Recommendations

■ If you're interested in Theory of Constraints, you might want to check out the Goldratt Institute's quarterly newsletter, TOC Times. The most recent issue is available for viewing at http://www.goldratt.com/toctquarterly/september2000.htm, or you can download a PDF version. It's a bit of a sales pitch for the Goldratt Institute, but they have customers contributing comments and stories, too. I should also point out that Eli Goldratt has a new book coming out in October called "Necessary but not Sufficient." You can order it from Amazon.

■ We've recommended the book Factory Physics since the earliest newsletter issues. We recently discovered that there's a Factory Physics website (www.factory-physics.com). Here you can find explanations of the various principles identified in the book. For example, in a system like a wafer fab, one principle is that "WIP and flow time (cycle time) increase non-linearly in utilization." The website includes an explanation for this principle, including an illustrative chart. There is also a list of improvement strategies, in order of expected impact, with specific examples of how they might be implemented (increase capacity at the bottleneck, which you can do by improving uptime, etc). Finally, there are spreadsheets that you can download to use as analysis tools to better understand your factory. If you're interested in the book, but not ready to spend $100 on a manufacturing textbook, you might want to check out this website.

■ FabTime's book of the month for October is "The Tipping Point" by Malcolm Gladwell. You can find our review of the book at http://www.fabtime.com/tippingpoint.htm. We also recommend the author's website, at http://gladwell.com.

Subscriber List

Total Subscribers: 200

Advanced Energy Industries (1)
Advanced Micro Devices (10)
Agilent Technologies (1)
Amkor (2)
Analog Devices (2)
Applied Materials Corporation (3)
Arizona State University (2)
Artest Corporation (1)
AT & S India Limited (1)
BP Solarex (3)
Carsem M Sdn Bhd (1)
Chartered Semiconductor Mfg (3)
Clarkson University (1)
Cofer Corporation (1)
Dallas Semiconductor (4)
Dick Williams and Associates (1)
Durham ATS Group (2)
Etec Systems (1)
FabTime (2)
Gintic Institute of Mfg. Technology (1)
Headway Technologies (4)
Hewlett-Packard Company (2)
Hyundai Semiconductor America (2)
IBM (2)
Infineon Technologies (9)
Intarsia Corporation (2)
Integrated Technologies Company (2)
Intel Corporation (14)
James Nagel Associates (1)
Ken Rich Associates (1)
Lockheed Martin Fairchild Systems (1)
LSI Logic (4)
Lucent Technologies (1)
Macronix International Co. (1)
Micrel Semiconductor (1)
MicroVision-Engineering GmbH (1)
Mitel (5)
Motorola Corporation (24)
MTE Associates (1)
Multimedia University (1)
Nanyang Technological University (1)
National Semiconductor (6)
Nortel Networks (3)
Oklahoma State University (1)
ON Semiconductor (4)
Penn State University (1)
Philips Semiconductors (2)
Powerex, Inc. (1)
Productivity Partners Ltd (1)
Raytheon (1)
Read-Rite Corporation (1)
RTRON Corporation (2)
SAMES (1)
Samsung Semiconductor (2)
Seagate Technology (9)
SEMATECH (7)
Solectron Corporation (1)
SSMC (1)
STMicroelectronics (6)
Synquest (4)
Takvorian Consulting (1)
TDK (1)
TECH Semiconductors (1)
Texas A&M University (1)
Texas Instruments (5)
TRW (1)
University of Arkansas (1)
University of Virginia (1)
University of Wuerzburg (Germany) (2)
White Oak Semiconductor (2)
Unlisted Companies (2)

Independent Consultants:
Stuart Carr
Alison Cohen
Doreen Erickson
Ted Forsman
Dan Theodore
Craig Volonoski

Note: Inclusion in the subscriber profile for this newsletter indicates an interest, on the part of individual subscribers, in cycle time management. It does not imply any endorsement of FabTime or its products by any individual or his or her company. To protect the privacy of our subscribers, email addresses are not printed in the newsletter. If you wish contact the subscribers from a particular company directly, simply email your request to the editor at Jennifer.Robinson@FabTime.com. To subscribe to the newsletter, send email to the same address. We will not, under any circumstances, give your personal information to anyone outside of FabTime.