The Engineering Management of Speed

Robert C. Leachman
Dept. of Industrial Engineering and Operations Research
University of California, Berkeley

Abstract

Analytical formulas are introduced for quantifying the revenue gains associated with local, incremental improvements in the speed of product development, supply chain development, or supply chain execution. The formulas provide practical means of imputing an overall economic value to engineering projects that impact lead time. Practical analytical queuing formulas for estimating changes in supply chain speed resulting from engineering changes to product or process also are discussed. An overall approach to engineering management in manufacturing companies characterized by rapid technological evolution is proposed, emphasizing disciplined, sophisticated management of speed.

[This paper was part of the *Proceedings of the 2012 Industry Studies Association Annual Conference,* Industry Studies Association, Univ. of Pittsburgh, Pittsburgh, PA, May 30 – June 1, 2012.]

1. Introduction and Overview

We have all heard expressions like "Time is money" or "We must make the market window" in reference to the importance of the speed of product and supply chain development or the speed of supply chain execution. In many technology-based companies, the urgency of product development, process development, supply chain development and supply chain execution is keenly felt. But the true economic value of incremental improvements in supply speed is rarely quantified. Worse, job descriptions and performance evaluations of engineers and managers rarely measure the total economic impact of improvements made to the speed of development or execution.

Going back to the Industrial Revolution, the traditional organization of management emphasizes a focus on product cost for managers and engineers developing and operating the supply chain. Supply chain managers and engineers typically are evaluated in terms metrics of cost, throughput and quality. In many industries this works fine. But in businesses subject to Moore's Law rates of technological evolution, it does not work so well. Prices for any given product decline relentlessly and quickly as its obsolescence steadily grows. It has become very difficult to make profits in electronics hardware businesses; only the swiftest can do so. While many and perhaps most improvements in speed result in cost reduction, throughput enhancement and/or yield improvement, many have a more profound impact on sales revenues than they do on these factors. As a result, their true economic worth is undervalued. In this paper I propose changes to engineering management to more properly value improvements in speed, to measure the impact on speed from changes to process or product, and to make the engineering organization more proficient at managing speed.

"Cycle time" is semiconductor industry jargon for the elapsed time to pass manufacturing lots through the manufacturing process, from lot creation until lot completion. The term is also applied to individual manufacturing steps, measuring the elapsed time from completion of the preceding step until completion of the step in question, or to a series of manufacturing steps (the sum of the cycle times of the subject steps).

In produce-to-order environments, cycle time is part of the product/service apparent to the customer and is therefore an important competitive issue. Suppliers able to offer shorter cycle times will be preferred. In the case of goods experiencing a rapid pace of technological obsolescence such as semiconductors, cycle time has a very strong influence on realized average selling prices. Firms with shorter cycle times are able make sales at earlier times when prevailing prices are higher. And by making early sales, such firms drive prices down and thereby diminish revenue available to competitors. Thus cycle time is very important even in a make-to-stock environment.

In this paper we will first impute economic value to cycle time reduction in terms of the revenue gain resulting from increased selling prices enabled by shorter time to market. Next, we review and adapt models from queuing theory as a practical means for computing entitlement cycle times. Finally, we propose organizational policies promoting disciplined and sophisticated cycle time management and improvement.

2. Empirical Evidence of Price Decline

Driven by relentless progress following Moore's Law, products produced in a planar fabrication process experience a very rapid decline in selling prices over the product lifetime. These products include semiconductors, liquid crystal displays, solar panels, light emitting diodes, read-write heads for disk drives, nanotechnology products, etc., as well as higher level products utilizing these products as components, such as computers, smart phones, tablets and televisions.

Examples of this are displayed in Figures 1, 2 and 3. Figure 1 depicts average selling prices for five generations of dynamic random-access memory chips. Prices are normalized so that 100% represents the price at product introduction. The time scale shows the number of months since the product was first introduced into the market. A log scale is used to display prices so that the percentage rate of decline in prices may be more easily discerned. The various product generations were introduced during quite different conditions in the DRAM industry, e.g., the 4MB DRAM was produced in volume during a period of tight product supply, whereas the 16MB DRAM was produced during a period of generous product supply. Nonetheless, on a percentage basis, prices track remarkably closely in every generation. About 12 months after product introduction, prices have declined to 20% of the initial price. A year later, prices are down to about 10% of the introductory price, and end-of-life is less than three years after first introduction.

Figure 2 displays a similar graph for the prices of four generations of Intel microprocessors, Pentium through Pentium 4. Here the slopes of the price decline curves are increasing every generation. By the Pentium 4 generation, price has declined to 10% of introductory price only 9 months after introduction. Product lives for all four generations are less than two years.

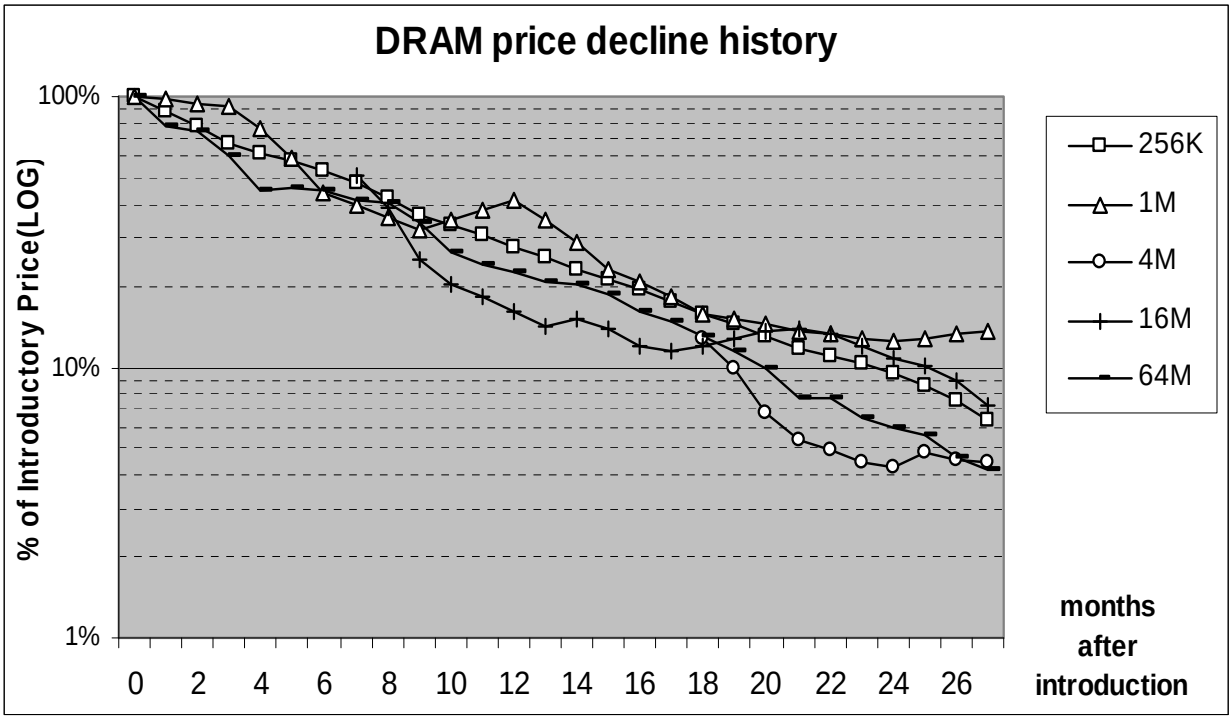

Figure 1. DRAM Average Selling Prices

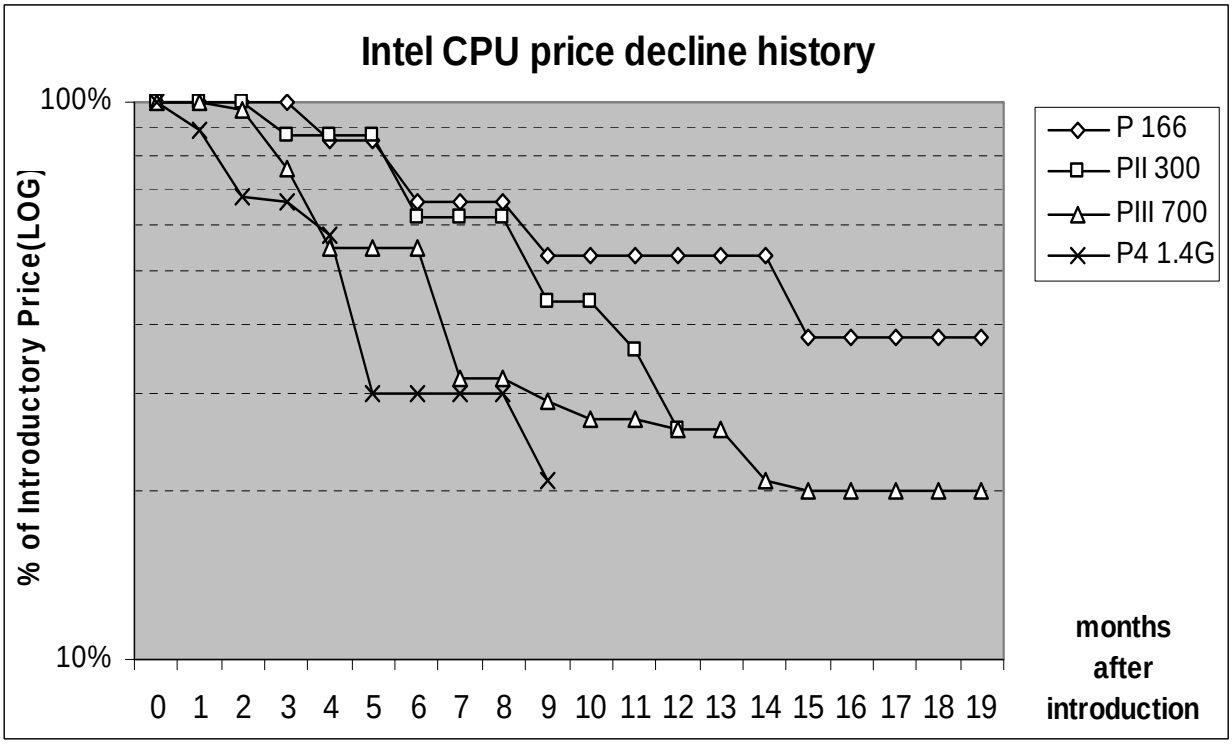

Figure 2. Intel Microprocessor Average Selling Prices

Figure 3 displays quoted prices from semiconductor foundries for wafer fabrication in various CMOS digital process technologies. Prices for foundry service do not decline as fast as they do for DRAMs or microprocessors, but they nonetheless decline with time. Thus even a contract manufacturer providing generic manufacturing services faces declining prices for fabrication in any particular technology. This decline reflects the introduction and improvement of newer process technologies.

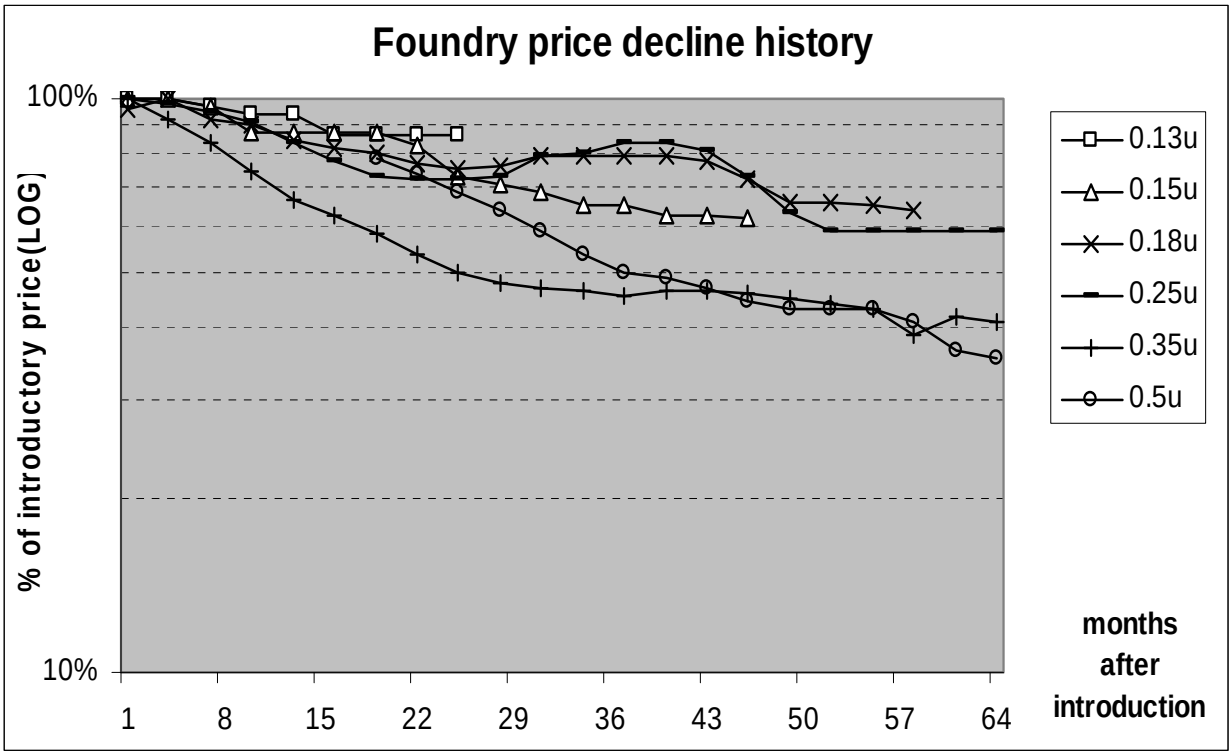

Figure 3. Semiconductor Foundry Wafer Prices

These graphs suggest there must be large economic values associated with compressing the time until volume production capability is realized. That is, there must be large economic values stemming from compression of product development time, time to install and qualify production equipment, time to ramp up yield and volume, and the manufacturing cycle time itself. Quite apart from cost reduction, these values stem from the opportunity to realize higher selling prices by being earlier to market.

3. Imputing the Economic Value of Cycle Time

The nearly straight-line curves in Figures 1, 2 and 3 suggest price history is well-modeled as a negative exponential. Let P_0 denote the market price at the time a product is conceived, and let $t = 0$ denote the epoch when the product is conceived. We assume product price is continuously declining at a constant rate, and let α denote the rate of price decline. The product price at time t is expressed as

$$P(t) = P_0 e^{-\alpha t}.$$

We assume there are multiple suppliers in a well-established commodity market whereby the market can absorb any output from a given supplier at time *t* and the output can be immediately sold at price *P*(*t*).

We assume there is a period of length *VT* termed the development time from the time the product is conceived until the manufacturing process is qualified and blank silicon wafers or other substrates started into the manufacturing process (hereafter called wafer starts) may be sold. During this period, the process technology to fabricate the product is developed and qualified, and the manufacturing equipment necessary to process the product is procured, installed and qualified. Let *H* denote the lifetime for the manufacturing process. Wafer starts that will be sold are made from time *VT* until time *VT* + *H*, at which time the product becomes obsolete and the wafer starts in the technology are terminated. Let *W*(*t*) denote the wafer starts made at time *t*. Let *Y*(*t*) denote the yield of wafers started at time *t*. Let *CT*(*t*) denote the manufacturing cycle time for wafers started at time *t*. We assume wafer starts made at time *t* will be sold at time *t* + *CT*(*t*).

Let β denote a suitable discount rate for computing a present value of lifetime revenues of the product. In that case, the present value of the lifetime revenue from the given product for that supplier may be expressed as

$$\int_{VT}^{VT+H} P_0 e^{-(\alpha+\beta)[t+CT(t)]} W(t)Y(t)dt\,. \tag{1}$$

Note the rate of price decline α acts like an add-on discount factor, as if the overall discount factor were α + β. If we re-set time 0 to be the epoch development is completed, and redefine the functions *CT*(*t*), *Y*(*t*) and *W*(*t*) such that the argument of the functions denotes how long after this epoch that the wafer start is made, then (1) may be re-written as

$$P_0 e^{-(\alpha+\beta)VT} \int_{VT}^{VT+H} e^{-(\alpha+\beta)[t+CT(t)]} W(t)Y(t)dt\,. \tag{2}$$

In principle, the lifetime revenue integral (2) may be calculated for status quo conditions and for changed histories of *CT*, *Y*, and *W* resulting from a proposed engineering project. This should be done for all products affected by the project. Taking the difference between the two integrals expresses the present value of estimated revenue gains from the project. If we let the subscript *i* denote product and if we let the superscripts *B* denote status quo values (*B* standing for before project) and *A* denote values if the project is pursued (*A* for after project), the present value of revenue gains attributable to the project may be expressed as

$$\begin{aligned}\sum_i P_0 e^{-(\alpha+\beta)VT_i^A} \int_{VT_i^A}^{VT_i^A+H_i^A} e^{-(\alpha+\beta)\left[t+CT_i^A(t)\right]} W_i^A(t)Y_i^A(t)dt - \\ \sum_i P_0 e^{-(\alpha+\beta)VT_i^B} \int_{VT_i^B}^{VT_i^B+H_i^B} e^{-(\alpha+\beta)\left[t+CT_i^B(t)\right]} W_i^B(t)Y_i^B(t)dt\end{aligned}\,. \tag{3}$$

We note in the foregoing equations that the price decline rate α and the discount factor β for computing present values always appear together as a sum. Without loss of generality, we shall assume for the rest of this paper that the parameter α represents the sum of the price decline rate and the discount factor.

Yield Ramp

In industrial practice there typically is a yield learning curve. At time of process qualification, yield is relatively low; we shall call this value Y_0. After a period of engineering detective work and problem solving, yield is gradually raised to a mature value we shall call Y_F. The period during which yield is improved is called the *ramp time*, denoted by *RT*. After time *RT*, the product is produced at mature yield Y_F until wafer starts are terminated at time *H*. Figure 4 illustrates the typical yield history, albeit depicting a smoother profile than in a real case. The start of product and process development is set to be time *–VT;* the first time wafers are started that will be sold is set to be time 0.

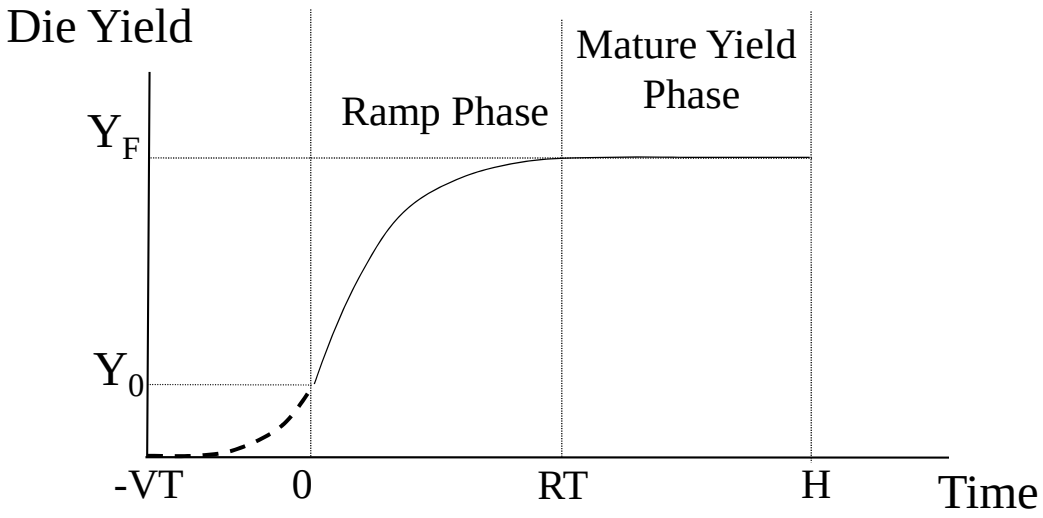

Figure 4. General Pattern of Yield vs. Product Life

The shape of the yield curve in Figure 4 suggests the yield ramp can be modeled by one minus a negative exponential. We posit the following model for the yield curve:

$$Y(t) = Y_0 + (Y_F - Y_0)\frac{1 - e^{-bt}}{1 - e^{-bRT}}, \quad 0 \le t \le RT . \tag{4}$$

Here, *b* is a shape factor for yield learning during the ramp phase. For example, suppose *RT* = 180 days and 2/3rds of the yield learning is completed halfway through the ramp phase. In that case, *b* should be set to a value of about 0.0077.

If *CT* and *W* are constant, the integral (2) is practical to calculate when the yield history *Y* is expressed as in (4). In that case, it can be shown that (3) reduces to

$$P_0 W Y_F e^{-\alpha(VT+CT)}\left\{\left[\frac{e^{-\alpha RT}-e^{-\alpha H}}{\alpha}+\left(\frac{Y_0}{Y_F}\right)\left[\frac{1-e^{-\alpha RT}}{\alpha}\right]+\left(\frac{1-\frac{Y_0}{Y_F}}{1-e^{-bRT}}\right)\left[\frac{1-e^{-\alpha RT}}{\alpha}-\frac{1-e^{-(\alpha+b)RT}}{\alpha+b}\right]\right]\right\} \qquad (5)$$

In the case $RT = VT = CT = 0$, (5) reduces to

$$P_0 W Y_F \frac{1-e^{-\alpha H}}{\alpha}. \qquad (6)$$

Equation (6) expresses the present value of the *ideal lifetime revenue* if there were no delays for process and product development, yield and volume ramps, and zero manufacturing cycle time. The development time *VT* and the manufacturing cycle time *CT* result in discount factors $e^{-\alpha VT}$ and $e^{-\alpha CT}$ applied to the lifetime revenue. The long expression within square brackets involving the ramp time *RT* and the learning curve shape factor *b* is much more complicated, but numerically it also acts as a discount factor, just a weaker discount factor than caused by *VT* and *CT*. That is, while waiting for *VT* or *CT* there is *zero* output; on the other hand, during *RT* there is *reduced* output associated with inferior yield, but not zero output.

Value of Development Time

To illustrate the value of development time, consider the following example. Suppose the selling price at start of development is $10,000 for a perfect-yielding wafer. Assume the selling price is declining 50% per year, and the discount rate is 25% per year. Suppose the process life *H* is three years, the manufacturing cycle time is a constant 50 days over the process life, and the wafer starts volume is 10,000 wafers per week, also constant of the process life. Suppose the yield ramp time *RT* is 180 days, and the yield learning curve shape factor *b* is set such that 2/3rds of yield learning is completed halfway through the ramp. In Table 1 below, we examine the present value of lifetime revenue for four different durations for *VT*.

Table 1. Economic Return from Compressing Development Time

VT **(days)**	**Present Value of Lifetime Revenue**	**PV Gain Compared to *VT* =120 days**
120	$1.279 billion	$0
105	$1.356 billion	$76 million
90	$1.436 billion	$157 million
75	$1.522 billion	$242 million

As may be seen, compressing development time is worth about $15 million per day in this example. Note that there are slightly increasing returns to scale with respect to cycle time, i.e., the incremental value of compressing one more day is slightly larger than the value of the previous day of compression.

Value of Manufacturing Cycle Time

To illustrate the value of manufacturing cycle time, consider the following simplified situation. Suppose yield *Y*, wafer start volume *W*, and cycle time *CT* are constant over the entire process life *H*. At time *t*, the average selling price is

$$P(t) = P_0 e^{-\alpha t}$$

where, in this case, P_0 denotes the average selling price when saleable wafer starts commence. Then the lifetime revenue integral (3) may be simplified in this case as

$$\int_0^H WYP_0 e^{-\alpha(t+CT)}\,dt = WYP_0 e^{-\alpha CT}\int_0^H e^{-\alpha t}\,dt = WYP_0 e^{-\alpha CT}\left(\frac{1-e^{-\alpha H}}{\alpha}\right).$$

Now suppose cycle time is permanently shortened by one day, i.e., $CT \to CT - 1$. Then the lifetime revenue becomes

$$WYP_0 e^{-\alpha(CT-1)}\left(\frac{1-e^{-\alpha H}}{\alpha}\right).$$

The revenue gain from reducing cycle time by one day is therefore

$$\Delta R = WYP_0 e^{-\alpha(CT-1)}\left(\frac{1-e^{-\alpha H}}{\alpha}\right) - WYP_0 e^{-\alpha CT}\left(\frac{1-e^{-\alpha H}}{\alpha}\right)$$

or

$$\Delta R = \left(e^{\alpha} - 1\right)WYP_0 e^{-\alpha CT}\left(\frac{1-e^{-\alpha H}}{\alpha}\right). \qquad (7)$$

We illustrate (7) with 2006 data from a fabrication plant producing image sensors. The fabrication plant made 10,000 wafers per week yielding an average of 420 good die per wafer. The cycle time was 50 days. The selling price at time 0 was $4.50 per die, declining 35% per year. The product life was two years. Assume a discount rate of 25%.

If time is expressed in days, then α (accounting for both the price decline rate and the discount rate) satisfies

$$(0.65)(0.75) = e^{-\alpha 365}$$

or $\alpha = 0.0019684$.

Applying (7), the value of one day of cycle time is then

$$\Delta R = \left(e^{0.0019684} - 1\right)\frac{10{,}000}{7}(420)(4.50)e^{-(0.0019684)(50)}\,\frac{1 - e^{-(0.0019684)(730)}}{0.0019684}$$

or

$$\Delta R = \$1{,}867{,}235,$$

i.e., the present value of revenue gains from permanent cycle time reductions achieved at time 0 without diminution of yield or wafer volume in this fab was worth about $1.9 million per day of reduction.[1]

Not included in the above analysis is the reduction in product cost associated with reducing cycle time. The most important element of such cost reduction in semiconductor manufacturing concerns the positive impact on yield from cycle reduction. The positive impact on yield not only reduces product cost, it also provides further revenue gain associated with shortened yield ramp time and/or higher mature yield. This occurs for two reasons: (1) Some yield loss mechanisms involve equipment or process "excursions" in which the process or equipment shifts out of control, but the presence of this excursion is not detected until the first lot processed after the excursion commenced reaches the end of the production line and is tested. All lots that had passed the out-of-control point before the first lot is tested also will have poor yield. When cycle time is reduced, the work-in-process in the manufacturing line is reduced, and so the number of lots with exposure to excursion loss is reduced, and therefore total excursion losses are reduced. (2) A process change that will improve yield must be justified on the basis of a successful in-line experiment. Typically, a portion wafers in a selected manufacturing lot are processed the old way, while the other wafers from the same lot are processed the new way. The lot then must travel through the rest of the fabrication process to the end of the line where all wafers are tested, and where it must be demonstrated statistically that the process change indeed improves yield. The shorter the cycle time, the less time is required to complete the experiment and implement the process change, and therefore the yield learning curve can be improved. A quantification of such benefits is described in Leachman and Ding (2011).

4. Cycle Time Metrics and Cycle Time Analysis

Given the large economic values associated with cycle time reduction, it is meaningful to consider metrics for managing and engineering cycle time. *Actual cycle time* (*ACT*) is a statistic measuring the time from the creation or arrival of an unprocessed manufacturing lot until its completion. *ACT* may be measured for the entire manufacturing process to fabricate a given

[1] Similar to the case of development time, each succeeding day of cycle time reduction is worth slightly more than the previous day of reduction.

product as well as for an individual process step in the flow. A weighted-average *ACT* may be computed for all process steps on all products under the responsibility of a particular engineering section as well as across all products to obtain a factory-level metric. *Dynamic cycle time* for a given process step is computed based on averaging actual cycle times for many lots passing through a particular process step within a given time frame such as a shift, a day or a week. *Static cycle time* for a given product is computed based on averaging the actual total elapsed time from lot entry into the process flow until lot completion at the last step of the flow over a series of lots completing the flow. An estimate of total-flow cycle time also may be derived by summing up step dynamic cycle times. This derived figure typically will be a different number than the static cycle time statistic because of the different time frames involved.

For the purposes of engineering improvement of particular processes and equipment, the dynamic cycle time metric on actual lot cycle time to pass through individual steps or groups of steps performed by the same equipment is of interest.

We define *entitlement cycle time* as the average manufacturing cycle time if manufacturing execution were perfect, taking into account wafer start volumes for all products, the process specifications for all steps, equipment qualified by process engineering to perform each step, statistics on lot inter-arrival times, and statistics on process and equipment trouble. Such trouble includes temporary disqualification of equipment from performing certain process steps, lots placed on hold because of out-of-spec or out-of-control process parameters, and equipment unavailable for production work because of maintenance or engineering reasons.

ECT may be computed for individual steps on individual products, for the entire process flow to make a given product, and for collections of steps within one process section. ECT is not an observable statistic in the data collected at a factory. Nevertheless, ECT can be analytically estimated, either conducting discrete-event simulations or by exercising formulas from analytical queuing theory. Either analytical approach uses statistical data collected from the factory. The information includes data from equipment logs on processing times, data from equipment tracking on equipment non-available times, process specifications as to which process tools are qualified to perform which steps, production volumes, etc.

The importance of ECT is its usefulness for separating cycle time issues into engineering problems versus execution problems. ECT may be viewed as the cycle time report card for process engineering. If ECT is too long, no amount of execution improvement by the manufacturing department can overcome this weakness. Excessive ECT is an engineering problem: ECT can only be reduced by engineering changes to equipment maintenance, to process control, to equipment qualifications, to process specifications, or by modifications to the equipment itself.

The gap between actual cycle time and entitlement cycle time may be viewed as the report card for manufacturing management and supporting staff, including information systems, automation and industrial engineering. Where there is a large gap between actual manufacturing cycle time and entitlement cycle time, it is an indication that execution could be and should be improved. Closing the gap could entail improvements to scheduling, information, automation, training or administration.

In many companies, industrial engineers have developed simulation models of the manufacturing line. These models can be helpful for evaluating cycle time gains associated with engineering or operational changes. But they tend to be large, complex models that can be exercised only by a handful of expert users. In a typical high-technology factory, every process engineering section faces technical issues affecting cycle time every day. Asking the expert simulation modelers to address all of these issues would amount to a hopeless organizational bottleneck; as a result, there is no time to analyze most issues. What is needed instead are distributed tools, tailored for the issues faced by each section, and deployed in parallel. These could be simulation models, but there need to be separate or duplicate models for each section and proficient modelers/users within each engineering section so that parallel analyses can be pursued.

Analytical queuing theory generally provides more approximate estimations of cycle time that can be achieved using simulation. But the virtues of queuing theory are (1) the formulas can be housed in Excel spreadsheets, making the tools accessible and practical for all engineers, (2) cycle time analysis is essentially instantaneous, and can be integrated with other spreadsheet analyses engineers do, and (3) while there is some bias in queuing-theoretic formulas, if one is simply computing cycle time *differences* associated with engineering changes, the amount of bias in the difference will be small, making the estimates reasonably accurate. For this reason, application of queuing theoretic formulas is attractive, if reasonably accurate formulas can be developed. The next section discusses the application of queuing theory for entitlement cycle time analysis in high-technology manufacturing.

5. Estimating Entitlement Cycle Times Using Queuing Theory

ECT has several components: time manufacturing lots are placed on hold; time lots are actually engaged in material handling transport or processing; and time lots are available for processing but waiting for selection by operators or automation and/or waiting for a qualified process tool.

Actual statistics on hold times may be collected. Actual statistics on transport times and process times also may be collected; in modern advanced technology factories, these statistics are collected automatically from machine and transport equipment logs. The analytical challenge for estimating ECT largely rests with estimation of lot waiting times.

In a manufacturing environment, if there is no lot hold time, the observed cycle time for a particular process step performed on a particular type of processing machine is

$$CT = WT + SCT \tag{8}$$

where *SCT* denotes the *standard cycle time* and *WT* denotes the lot waiting time. Standard cycle time is an observable statistic, it is the average time required for a lot from when it is accepted by the machine for performance of a process step until the lot is completed and ready for transport to the next step. *SCT* is the irreducible portion of cycle time (irreducible without engineering changes to process specifications or equipment). Another parameter about the process step is the *process time PT*, the average time between starts of consecutive lots on the machine when the lots are tendered to the machine as quickly as possible and there are no interruptions of

processing activity. For some equipment, *PT* and *SCT* are identical, but for many they are not. On many types of equipment, lots can be "pipelined" whereby the next lot is started before the previous lot is ready for departure. In such cases, *SCT* is larger than *PT.* On other types, there may be some necessary delay for cleaning or reconditioning of the machine chamber between consecutive lots. In such cases, *SCT* may be smaller than *PT.*

The queue-theoretic view of the world is that servers are engaged in providing processing service without interruption, and the queue-theoretic cycle time is the sum of queue time and service time:

$$CT = QT + t_s$$

where *QT* is the queuing-theoretic waiting time (hereafter called the queue time) and t_s is the average service time, i.e., the reciprocal of the processing rate. The queue-theoretic cycle time does not account for the discrepancy between standard cycle time and process time. Thus if we apply queuing theory to manufacturing, we must account for this difference, i.e.,

$$CT = QT + t_s + (SCT - PT) . \quad (9)$$

In any manufacturing environment, machines are sometimes down for maintenance. Let *A* denote the average fraction of time a machine is available for processing service, i.e., the average fraction of time the machine is not down for maintenance or engineering work *A* is termed the *availability* of the machine. To adapt queuing theory for manufacturing, the nominal processing rate of the machine is de-rated by the availability. If *PT* denotes the average process time per lot, then

$$t_s = PT/A . \quad (10)$$

Substituting (10) into (9) we have

$$CT = QT + PT/A + (SCT - PT) = QT + (1/A - 1)PT + SCT . \quad (11)$$

Comparing (11) to (8), we see that the observed waiting time in manufacturing is estimated using queuing theory as

$$WT = QT + (1/A - 1)PT . \quad (12)$$

i.e., waiting time equals the queue-theoretic queue time plus a correction factor for non-availability of the machine.

Queuing theory provides exact analytical expressions for queue-theoretic waiting time in the case of exponential distributions for independent lot inter-arrival times and process times. For manufacturing applications, we are interested in generalizations modeling machine non-availability, multiple machines performing in parallel a certain kind of process step, and lot inter-

arrival and process time distributions that are different from exponential. Such a general case defies an exact expression. But useful and reasonably accurate approximate formulas for this case have been progressively developed by Kingman (1961), Sakasegawa (1977) and Hopp and Spearman (1999). Formula (13) below applies to the case where *m* machines are qualified to perform a given manufacturing step. Other notation is as follows:

u – the utilization of availability of the work station (i.e., of the bank of *m* parallel machines)
PT – mean lot process time
c_0^2 – the squared coefficient of variation in lot process time (i.e., the ratio of the variance of lot process time to the square of the mean process time)
A – the average availability of the machines in the work station
MTTR – mean length of a machine down time event
cr^2 – the squared coefficient of variation in the length of a down time event
ce^2 - short-hand notation for the squared coefficient variation of the effective service time, a function of *A*, *MTTR*, cr^2, *PT*, and c_0^2 defined below.
ca^2 - the squared coefficient of variation in the lot inter-arrival time; for Poisson arrivals, ca^2 =1.

The expected (average) queue time is then

$$QT = \left(\frac{ca^2 + ce^2}{2}\right)\left(\frac{u^{\sqrt{2(m+1)}-1}}{m(1-u)}\right)\left(\frac{PT}{A}\right) \qquad (13)$$

where

$$ce^2 = c_0^2 + (1+cr^2)A(1-A)\left(\frac{MTTR}{PT}\right). \qquad (14)$$

Let us examine this general queue time formula (13) qualitatively. It is the product of three terms. The first term is a term involving the variability of lot arrivals and the variability of service time. More variability increases cycle time.

The second term in (13) concerns the utilization of the work station and the number of qualified machines. Note the (1-*u*) in the denominator; as utilization is pushed to 100% of the availability, queuing theory predicts wait time will explode. The behavior of this term is graphed in Figure 5 for various values of *m* and for *u* ranging from 0 up to 100% of availability. Note that wait time is reduced as *m* is increased, even for the same utilization. At 90% utilization of availability, average wait time when only one machine is qualified is more than nine times higher than when eight machines are qualified.

The third term in (13) expresses a ratio of process time to availability. All else being equal, a longer process time also means a longer average wait time. All else being equal, a lower availability means a longer average wait time (even if utilization is reduced to the same utilization of availability).

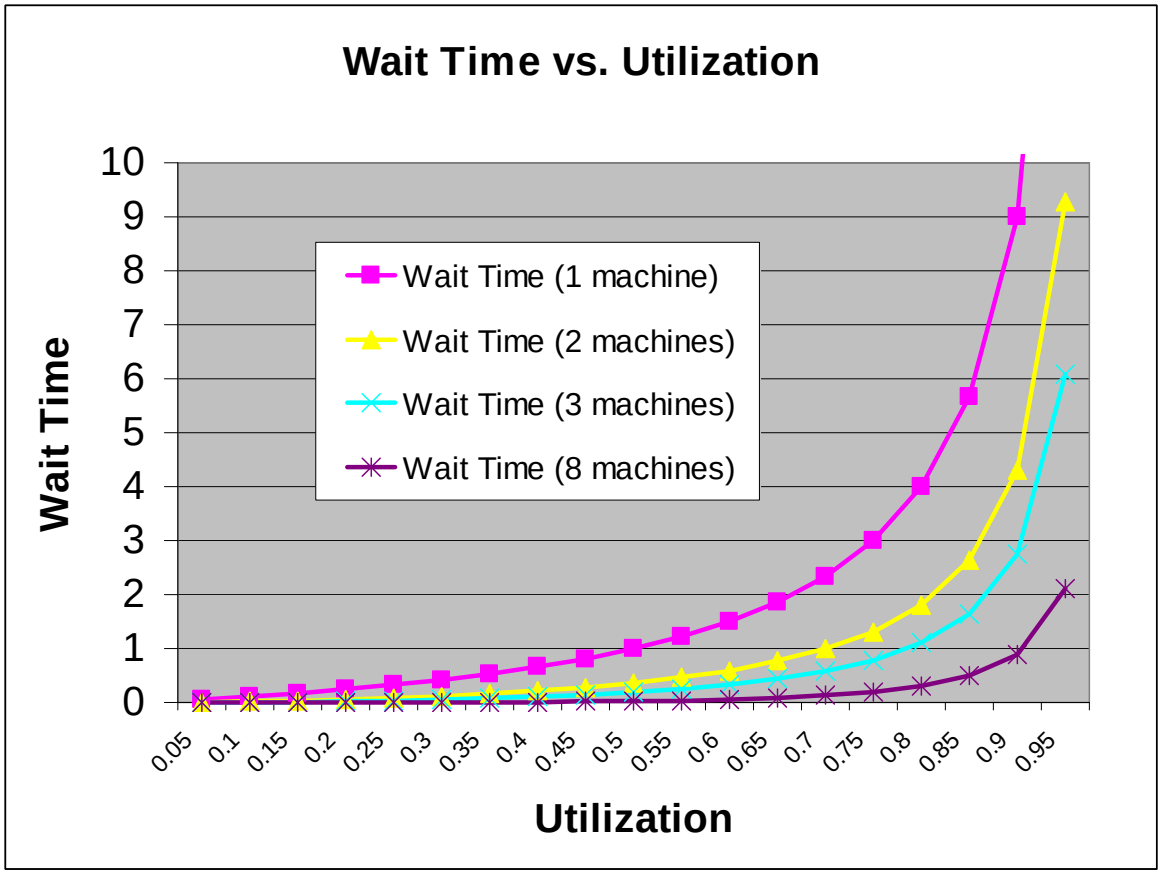

Figure 5. Wait Time as a Function of the Number of Qualified Machines and the Utilization of Availability

Queue time may be reduced if any of the three terms in (13) is reduced. This suggests the general avenues for cycle time reduction: (1) reducing variability (i.e., reducing *ca* or reducing *ce*), (2) increasing *m* or reducing *u*, and (3) reducing *PT* or increasing *A*.

The basic queuing formula (13) needs to be modified for the cases of machines whose operation requires the lots to be grouped into batches that are sequenced through the machine. Batching may be performed because of a large load size for the equipment (e.g., diffusion furnaces), or because of significant setups applicable to classes of manufacturing steps (e.g., species setups on ion implanters). For details, see Hopp and Spearman (2001).

6. Engineering Organizational Policies to Better Engineer and Manage Speed

Given the large economic values associated with cycle time reduction and given the analytical means to estimate entitlement cycle time described in preceding sections, it is meaningful to consider the engineering organization and how it should deal with supply-chain speed in a high-technology company. Typically, a high-technology company has a product design organization and a process engineering organization. Process engineering may be further divided into an

organization dealing with R&D of new process technology and another dealing with improvement of existing process technologies. The process engineering organizations are further subdivided into *sections* dealing with particular types of process technology. For example, in companies manufacturing using planar fabrication process technologies, there will be sections for photolithography, etch, ion implant, chemical vapor deposition, metal deposition, diffusion, chemical mechanical polish, etc. Each section is responsible for the *process specifications* precisely defining how each manufacturing step is to be performed and the determination or verification of quality of output of each step. Typically, the sections are responsible for resource planning, staffing, and costs of the operations to perform the steps within their domain. In a typical high-technology company, on nearly every working day every process section faces issues and makes decisions that influence cycle time. Supporting the sections are staff departments such as information systems, industrial engineering, automation and human resources, as well as a manufacturing department executing the process and maintaining the equipment. They too face issues and make decisions affecting cycle time on a near-daily basis.

For a high-technology manufacturing company to be proficient at managing cycle time, the following managerial and organizational policies are proposed:

(1) Management should impute an economic value to cycle time for each product and declare this value to the engineering and operations organizations. Management should require that any proposals for changes to product development plans, to the manufacturing process or to operational policies that would change cycle times must be justified by quantifying the overall economic impact, including the gain or loss in future revenues associated with changes in cycle time.

(2) Management should establish cycle time goals for each product. These goals may be dynamic, anticipating learning curve improvements. Where entitlement cycle time is larger than the cycle time goal, the engineering organization must devise changes to equipment or process enabling the goal to be met.

(2) Entitlement cycle times should be calculated for every product and process. Each engineering section has tailored queuing formulas or simulation models embedded in spreadsheet tools and knowledgeable staff enabling it to routinely calculate its entitlement cycle times, and to routinely compute the predicted change in its ECTs as a function of proposed changes to product, process, equipment, operational policies, product mix or factory volume.

(3) Every proposal for an engineering project or an operational change quantifies the change in ECTs as well as changes in product cost anticipated from the project. Every proposal for an engineering project or an operational change quantifies the true discounted cash flow to the company from the project, including changes in product lifetime revenues as well as changes in product cost. Management does not sign off on any project proposal unless the cycle time impact and consequent speed dollars are included in the evaluation.

(4) Where actual cycle time is larger than entitlement cycle time, the manufacturing organization and supporting engineering staff needs to improve execution. Improved execution tools may be required, i.e., more advanced planning and scheduling systems.

(5) Engineering sections and operational management are evaluated on the basis of their overall discounted cash flow contribution to the company, including revenue gains as well as cost reduction. Engineering or operational improvements that increase speed are appropriately recognized and rewarded.

Table 1 displays the 2006 cycle time results at an image sensor fabrication plant embracing this approach. ECT analysis tools (Excel spreadsheets housing customized analytical queuing formulas) were developed and implemented in every process engineering section during the spring of 2006. There are rows in the table for each section (e.g., "photo" for photolithography, "CFA" for color filter array, "CMP" for chemical mechanical polish, "Wet" for wet-bench etching and cleaning, etc.). There are columns for entitlement cycle time in June and September, 2006 and for actual cycle time in the same months. Viewing the totals, as of June, 2006, the entitlement cycle time for fabricating image sensors was almost 39 days and the actual cycle time was almost 48 days. After three months of engineering effort, entitlement cycle time had been dropped to a little less than 29 days, i.e., a 10-day reduction. ECT in only one section (CFA) got worse; this was because this portion of the process was being removed from the factory, and therefore there was no value to improving it.

Table 1. Entitlement and Actual Cycle Time Improvements in 2006 at an Image Sensor Fabrication Plant

Section	June Entitlement	Sept Entitlement	June Actual	Sept Actual
Photo	3.31	2.43	6.16	3.25
CFA	1.72	1.83	2.96	2.35
Metrology	4.91	1.96	4.47	3.03
Strip	0.90	0.74	2.17	1.84
Scrub	1.37	0.93	2.19	1.34
Implant	4.14	3.62	4.28	3.96
CMP	1.23	1.09	1.28	1.14
CVD	1.33	1.11	3.31	2.31
Metals	1.03	0.67	0.85	0.75
Etch	2.10	1.43	3.14	2.36
Probe	0.06	0.02	0.11	0.06
Diffusion	9.31	6.65	6.88	4.75
Wet	7.37	6.43	10.11	8.46
Total	**38.78**	**28.92**	**47.91**	**35.60**

As may be seen, actual cycle time dropped to less than 36 days, i.e., a more than 12-day reduction. During the summer months there was considerable improvement in ECT as well as execution improvements reducing the gap between actual and entitlement cycle times.[2] The 12-day reduction in actual cycle time was worth about $24 million in increased lifetime revenues for the image sensors in production at that time. Moreover, the improved competitiveness enabled the company to capture major new accounts.

[2] Reported actual cycle times in the Diffusion section were less than the entitlement cycle times because of an accounting convention: Most of the waiting time for diffusion steps was credited to the Wet section's actual cycle time. The sum of Diffusion + Wet actual and entitlement cycle times may be compared.

References

Hopp, Wallace and Mark Spearman, *Factory Physics*, McGraw-Hill, New York (2001).

Kingman, J. F. C., "The Single-Server Queue in Heavy Traffic," *Proceedings of the Cambridge Philosophical Society*, **57**, p. 902-904 (1961).

Leachman, Robert C. and Shengwei Ding, 2011. "Excursion Yield Loss and Cycle Time Reduction in Semiconductor Manufacturing," *IEEE Transactions on Automation Science and Engineering*, **8** (1), p. 112-117 (January, 2011).

Sakasegawa, H. "An Approximation Formula $L_q \cong \alpha \cdot \rho^{\beta} / (1-\rho)$," *Annals of the Institute of Statistical Mathematics*, **29** (1A), p. 67-75.