Estimating the Lead Time Distribution of Priority Lots in a Semiconductor Factory *

Asbjoern M. Bonvik†

July 10, 1995

Abstract

We consider the lead time of priority lots in a semiconductor factory. The main cause of delays is waiting for repairs of failed machines, and the failures are mainly caused by power disturbances. This can be modeled as an exogenous time-dependent failure process, and we develop a probabilistic model of the system based on this. Using this model, a convolution algorithm for finding the lead time distribution is described.

We describe a method of creating synthetic samples of the lead time from historical failure and repair data collected in the factory. Based on such a sample from the MIT Integrated Circuits Laboratory, we approximate the distribution of lead times by a gamma probability density function shifted by the smallest possible lead time for the product type. The parameters of the gamma distribution are found by using a maximum likelihood estimator. The resulting distribution gives good agreement with the synthetic data for values less than two standard deviations above the mean lead time.

Since our procedure only depends on a description of the process and the failure and repair history of the factory, it can also be used to obtain lower bounds on the lead time for new product types.

*This work was supported by the Advanced Research Projects Agency under contract N00174-93-C-0035.

†Massachusetts Institute of Technology, Operations Research Center, One Amherst Street, Room E40-194, Cambridge, MA 02139.

1 Introduction

1.1 Background

We define the *lead time* of a lot as the time from when an order is released to the factory until the finished product is available to the customer. Firms that produce to order will have competitive advantages if their lead time is low, but the lead time is also important in volume production. Semiconductor products have short lifecycles from design to obsolescence, so shorter lead times give more opportunities to improve quality and yield using lessons learned from early lots before the product is obsolete. In the queuing literature, the term *sojourn time* is often used for this concept. A related measure is the *cycle time*, defined as the time from when the lot is loaded on the first machine until it is unloaded from the last.

The MIT Integrated Circuits Laboratory (ICL) is an experimental semiconductor factory, whose main purpose is to produce lots where either the chip design or the production process is experimental. Such lots constitute a light load on the system with lot releases separated by several weeks. For experimental lots, the lead time is a much more important performance measure than the throughput. The laboratory also produces lots of what is known as the *baseline process* to ensure that all machines in the laboratory are working and correctly adjusted. These lots are relatively complex products that exercise the machines and operators in the laboratory. Baseline lots have much lower priority than experimental ones.

Observations by laboratory personnel lead to the conclusion that the most important cause of delays in processing is the failures and repairs of machines.

When a machine fails, any lot that uses this machine at a later process step may have to wait until the machine is repaired. These delays can be significant compared to the overall operation time required for processing the lot. It is also observed that the most important cause of machine failures is problems with the power supply. Spikes or brownouts in the power can cause sensitive machines to shut down or fail completely. Recovery can involve anything from restarting the machine to ordering spare parts from the manufacturer.

This failure process has the important characteristic that the power disturbances are exogenous to the laboratory itself. The different machines may have different sensitivities to power disturbances, but they suffer failures independent of whether they are operating or not. In the classification of Buzacott and Hanifin [1] this type of failure is called *time dependent*. The defining characteristic of such failures is that their occurrence only depends on the time since the last repair through some exogenous stochastic process. This is contrasted with *operation dependent* failures that only can occur when an operation is taking place and depend on the accumulated operating time since the previous repair.

ICL uses CAFE, a computer integrated environment for semiconductor manufacturing [2]. Parts of the system are still under development, but it has collected data on failures and repairs of machines since about 1990. The database also contains descriptions of the processes used in the laboratory and the production histories of individual lots. The description of all processing steps for a particular product type is called a *process flow*. The production data are less complete, but the available data indicate that three machines have utilization between 20 and 30 percent, six have utilization between 10 and 20 percent, and the remaining 39 machines have utilization less than 10 percent. There are some data available on historical cycle times in the laboratory, but the number of lots produced of each product type is quite low, so it is hard to make reliable statistical inferences from these data.

1.2 Literature Review

Relatively little attention has been given to the lead time as a performance measure in manufacturing. Karmarkar [3] states that this is true both for the academic literature and for the current industry practice. The lack of publications and need for research in this field is also stated by Gershwin [4].

The most basic relation between the lead time and the volume of work in progress is the well-known Little's Law $L = \lambda W$, relating the mean number of parts in the system to the production rate and the mean time spent in the system [5]. This law holds for a wide range of queuing systems. There are also extensions of this law to higher moments [6] and to the distributions of lead time and number of parts in the system [7, 8].

There has been some work on soujourn times in queuing networks, see [9]. In this paper, we only consider the highest priority class of lots in the system. These lots do not suffer congestion delays, so we can simplify the problem considerably.

Seidmann [10] studies the lead time distribution of a single flexible manufacturing cell where the product is inspected after processing is completed, and routed back for rework if the inspection is failed. As a consequence, the number of passes through the system for all members of a batch has a negative binomial distribution. The Laplace-Stieltjes transform of the lead time is derived from the transform of the processing time and the transform of the number of passes through the system. A numerical inversion of the transform is then used to find the lead time distribution for entire batches of various sizes.

Karmarkar [11] also studies the effect of product batching on the lead time. For small batch sizes, many lots are queuing in the system, creating congestion delays. As the lot sizes increase, the queuing delays go down, but the time to process a lot increases. No particular item will be finished before the entire lot it belongs to is finished. We do not consider batching, but assume that all lots have a common standard size.

Wein [12] considers the due-date setting problem, in which a firm has to decide a due date for a product. Setting an earlier due date is a competitive advantage for the firm, but increases the probability of failing to meet the due date because of randomness in the production process. A multiclass M/G/1 model is used to derive the Laplace-Stieltjes transform of the lead time. Our work provides another way of finding the lead time distribution for the due-date setting problem.

One approximation scheme is to use an Erlang or a gamma distribution to represent the lead time distribution. This is recommended by Buzacott and Shantikumar [13] for systems where the coefficient of variation is empirically determined to be less than unity, but this reference does not describe in detail how to find the best approximation. This is the approach we adopt in later sections of this paper.

Other authors, e.g. Jacobs and Meerkov [14], use a normal approximation to the lead time distribution. We will show that this may be asymptotically true for the model we analyze, but that the process we consider (with 174 operations) is too short to closely conform to a normal distribution.

Another approximation scheme for heavily loaded systems is based on a Brownian motion system approximation. The scheduling problem for the system is then solved as a control problem in the approximated system, see [15, 16].

1.3 Outline of the paper

In section 2, we develop and analyze a probabilistic model of the system. The purpose of this section is to gain some intuition about the general behavior of the lead time distribution rather than to derive an exact expression.

Section 3 describes a sampling procedure for inferring lead time data from historical machine failure and repair data and the process description. This procedure is used to generate a lead time sample for a particular process flow in the ICL, and we discuss some data quality issues raised by the procedure.

Based on the probabilistic model and the generated sample, we argue in section 4 that a gamma distribution shifted to the right by the smallest possible lead time is an appropriate approximation and derive a minimum likelihood estimator for this distribution. We then compute the parameters of the gamma distribution and compare with the sample from the laboratory.

2 Probabilistic analysis

In this section, we develop and analyze a probabilistic model of the system. We find the lead time distribution at a single machine. We then discuss the extension to a longer process flow and simplify by assuming independence. Under this assumption, we can use a convolution algorithm to find the lead time distribution.

2.1 Preliminaries

The purpose of this section is to gain intuition about what type of distribution lead times can reasonably be expected to have. Therefore, we make the following assumptions to simplify the analysis.

We assume that all failures are due to the external power disturbances, and that these disturbances are identical and occur as a Poisson process. We assume that a disturbance affects each machine independently, so that the machine failures can be represented as independent Bernoulli trials at the disturbance arrival times. We also assume that the machine repair times are exponentially distributed.

The lot in question is assumed to be the single member of the highest priority class in the factory. The priority scheme is assumed to be preemptive, so the lot never suffers any congestion delays caused by machines being busy or queues forming. Note that it is possible to schedule all other operations in the factory to avoid interference with the priority lot, so actual preemption need not occur.

We define these parameters:

λ	Disturbance arrival rate
γ_i	Repair rate for machine i
p_i	Sensitivity (failure probability) for machine i
d_j	Duration of operation j in the process flow
k	Number of individual operations in the process flow.

The process flows may be re-entrant. That is, machine i may be visited more than once during the process. We therefore define m_j to be the machine on which operation j takes place, and will refer to parameters γ_{m_j} and p_{m_j} as a description of the machine used for this operation.

We define these random variables:

Y	Time since the last disturbance
W'_j	Time the lot spent waiting for machine m_j to be repaired before operation j could start
W''_j	Time spent waiting for repairs of machine m_j during interruptions of operation j
W_j	Total time spent at operation j, $W_j = d_j + W'_j + W''_j$
W	Total lead time of the lot, $W = \sum_j W_j$

When it is necessary to distinguish between different realizations of the lead time variables, we use subscripts in parentheses.

2.2 Time spent at the first machine

We now derive the distribution of W_1, the time spent at the first operation in the process flow. We first find the probability of finding the machine up, then the probability of suffering interruptions during the operation. We find the conditional distributions for the various combinations of these events, and combine these distributions to find the distribution of W_1.

The lot can be delayed in two ways at a machine; it can find the machine down when it arrives, and it can be interrupted by a failure during processing. Since we assume Poisson disturbances and Bernoulli trials at each disturbance, the machine will stay up for a exponentially distributed period with mean $1/\lambda p_{m_1}$. The time from failure to repair is exponentially distributed with mean $1/\gamma_{m_1}$ by assumption. Assuming steady state, the probability that a lot arriving at the first machine will find it up is the time average availability of the machine. This is [4, 13, 17]:

$$P\{m_1 \text{ up}\} = \frac{\gamma_{m_1}}{\lambda p_{m_1} + \gamma_{m_1}} \tag{1}$$

Now consider what occurs after the lot arrives at the machine. If the machine is up, W_1', the wait before the operation can commence, equals zero with probability 1. If the lot finds the machine down, it will wait for a repair. Since we assume exponentially distributed repair times, the remaining time to repair observed by the lot is also exponentially distributed. Then

$$f_{W_1'|m_1 \text{ down}}(w) = \gamma_{m_1} e^{-\gamma_{m_1} w}, \qquad w > 0 \tag{2}$$

Consider what occurs after the operation commences. It may complete without any delays, making W_1'', the wait during interruptions of the operation equal to zero, but it may also be interrupted one or more times during processing. For each such interruption, the lot suffers a delay equal to the time to repair the machine, since we assume that the operations are resumable after a failure. The failures occur as a Poisson process with rate λp_{m_1}, so the number of failures suffered during an operation of duration d_1 has the Poisson distribution

$$p(m) = \frac{(\lambda p_{m_1} d_1)^m}{m!} e^{-\lambda p_{m_1} d_1}, \qquad m = 0, 1, 2, \ldots \tag{3}$$

Since a typical operation has duration on the order of an hour and the time between failures is on the order of weeks, we make the simplifying approximation that at most one interruption can occur during a single operation. The probability of an interruption is then

$$P\{\text{interruption during operation } j\} = \lambda p_{m_1} d_1 \tag{4}$$

and the delay given an interruption is again distributed as the repair time of the machine.

It is possible that the machine is found down and the operation is interrupted. These are events on non-overlapping intervals, so the events are independent, and the combined event has probability

$$\begin{aligned} P\{\text{found down and interruption}\} &= P\{\text{found down}\}\, P\{\text{interruption}\} \\ &= \frac{\lambda p_{m_1}}{\lambda p_{m_1} + \gamma_{m_1}} \lambda p_{m_1} d_1 \end{aligned} \tag{5}$$

Given that this event occurs, the lot will be delayed for two identically distributed repair times. Since the repair time is assumed to have an exponential distribution, the sum of these random variables have the Erlang distribution

$$f_{W_1' + W_1'' | \text{down and interruption}}(w) = \gamma_{m_1}^2 w e^{-\gamma_{m_1} w}, \qquad w > 0 \tag{6}$$

Similarly, the probability of having no delays is

$$\begin{aligned} P\{\text{up and no interruption}\} &= P\{\text{found up}\}\, P\{\text{no interruption}\} \\ &= \frac{\gamma_{m_1}}{\lambda p_{m_1} + \gamma_{m_1}} (1 - \lambda p_{m_1} d_1) \end{aligned} \tag{7}$$

and the probability of having exactly one delay is

$$\begin{aligned} P &= P\{\text{down and no interruption } \textbf{or} \text{ up and interruption}\} \\ &= P\{\text{found down}\}P\{\text{no interruption}\} + P\{\text{found up}\}P\{\text{interruption}\} \\ &= \frac{\lambda p_{m_1}}{\lambda p_{m_1} + \gamma_{m_1}}(1 - \lambda p_{m_1} d_1) + \frac{\gamma_{m_1}}{\lambda p_{m_1} + \gamma_{m_1}} \lambda p_{m_1} d_1 \qquad (8) \end{aligned}$$

The conditional densities for W_1' and W_1'' can be converted to conditional densities for W_1 by recalling that d_1 is the nominal duration of the operation, so W_1' and W_1'' are delays beyond that duration. Therefore, the conditional density for W_1 given that the machine was found down and no interruption occured, say, is found by replacing w by $w - d_1$ in equation (2):

$$f_{W_1|\text{down and no interruption}}(w) = \gamma_{m_1} e^{-\gamma_{m_1}(w-d_1)}, \qquad w > d_1 \qquad (9)$$

We can now get the distribution of time spent at the first machine by combining the conditional densities:

$$\begin{aligned} f_{W_1}(w) &= P\{\text{no delay}\} f_{W_1|\text{no delay}}(w) \\ &+ P\{\text{1 delay}\} f_{W_1|\text{1 delay}}(w) \\ &+ P\{\text{2 delays}\} f_{W_1|\text{2 delays}}(w) \\ &= \frac{\gamma_{m_1}}{\lambda p_{m_1} + \gamma_{m_1}}(1 - \lambda p_{m_1} d_1)\, \delta(w - d_1) \\ &+ \left(\frac{\lambda p_{m_1}(1 - \lambda p_{m_1} d_1)}{\lambda p_{m_1} + \gamma_{m_1}} + \frac{\gamma_{m_1} \lambda p_{m_1} d_1}{\lambda p_{m_1} + \gamma_{m_1}} \right) \gamma_{m_1} \exp\Big(- \gamma_{m_1}(w - d_1)\Big) \\ &+ \frac{\lambda p_{m_1}}{\lambda p_{m_1} + \gamma_{m_1}} \lambda p_{m_1} d_1\, \gamma_{m_1}^2 (w - d_1) \exp\Big(- \gamma_{m_1}(w - d_1)\Big) \\ &= \frac{\gamma_{m_1}}{\lambda p_{m_1} + \gamma_{m_1}} \Big[(1 - \lambda p_{m_1} d_1)\, \delta(w - d_1) \\ &\quad + \lambda p_{m_1} \Big(1 + d_1(\gamma_{m_1} - \lambda p_{m_1}) + d_1 \lambda p_{m_1} \gamma_{m_1} (w - d_1) \Big) \\ &\quad\quad \exp\Big(- \gamma_{m_1}(w - d_1)\Big) \Big], \qquad w \geq d_1 \qquad (10) \end{aligned}$$

where $\delta(w - d_1)$ is the Dirac delta function, indicating an impulse at the nominal operation duration d_1. This represents the probability of completing the operation without delays.

The simplification of ignoring the possibility of more than one interruption during processing prevents the distribution from containing an infinite series of higher order Erlang terms scaled by the probabilities of experiencing the corresponding number of failures. Figure 1 shows the appearance of this density for an operation of five hours duration with machine failure probability of 0.8, mean time between disturbances of twenty hours, and mean time to repair of ten hours. This gives a machine availability of 0.71 and a probability of no delays of 0.57.

An alternate procedure to get the distribution in equation (10) is to convolve the densities of the components d_1, W_1', and W_1'', where the densities contain impulses. This gives the same result, and provides a partial check on the derivation.

2.3 State of the second machine visited by the lot

The times spent at successive machines are not independent random variables. If a lot has to wait for a repair at one machine, the probability of finding the next machine down increases, since the same disturbances affect both machines. We will not derive the joint distribution of time spent at several machines, but will discuss the complications involved, and simplify by assuming independence. In this section we discuss the probability of finding the second or later machines up when the lot arrives for an operation.

Suppose that the lot found the first machine up and that no disturbances

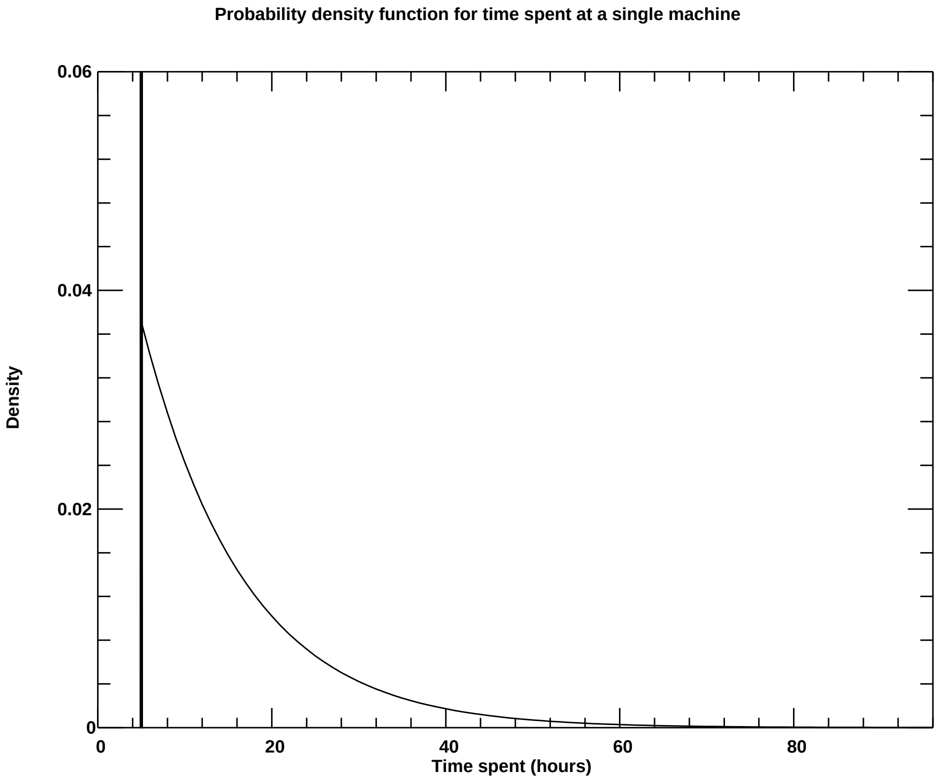

Figure 1: Probability density for time spent at first machine

occurred during processing. Then the lot arrives at the second machine d_1 time units after it arrived at the first machine. Since we assume Poisson disturbances, the time Y between the last disturbance and the lot arrival at the first machine is distributed with

$$f_Y(y) = \lambda e^{-\lambda y}, \qquad y > 0 \tag{11}$$

The time between the last disturbance and the lot arrival at the second machine is now the sum of Y and the constant d_1. We can find the probability of the lot arriving while the machine is down in two steps. First, we find the probability of the second machine being down immediately after the last disturbance. This is the steady state probability of the machine being down plus the product of the steady state probability of the machine being up and p_{m_2}, the probability of its failing as a result of a disturbance. Second, we compute the probability of the time to repair the machine being larger than $Y + d_1$ by integrating the joint distribution of these two (independent) random variables. The probability of the lot finding the machine down is then the product of these two probabilities.

This is not the only case, however. Suppose the lot found the first machine down. Several disturbances may have occurred while the lot was waiting for the first machine to be repaired. These disturbances will not affect m_1, since it is down already, but may affect m_2. We therefore need to find the distribution of Y, the time since the last disturbance, conditioned on finding the first machine down. A similar problem occurs if the lot is interrupted during processing. The lot may also spend some time in transit between machines, and disturbances may occur during that time. Treating all the possible cases in this fashion, we can derive the probability of finding m_2 up.

A further complication is the fact that the factory may not be operating continuously, but can close overnight and for weekends. During these periods no repairs or operations take place, but disturbances will occur. Such idle periods cover a significant fraction of time, so this effect cannot be ignored. One way of modeling these disturbances is to use working time rather than clock time as the time scale, and to introduce impulses in the disturbance process representing the non-zero probabilities of finding equipment down at the end of idle periods.

We note that the time spent at m_2, given that the machine is found in a particular state, does not depend on the previous history of the lot. This is because we assume Poisson disturbances and exponential repairs, so any events on non-overlapping time intervals will be independent. Any dependencies between machines will then result in changes to the probability of finding the machine in a particular state, while the shape of the distribution remains unchanged.

We also note that the introduction of transfer times between machines and periods when the factory is not operating weakens the correlation between machines, since repairs may take place while no processing is being done on the lot, and disturbances may occur while the factory is closed. We have seen that such non-processing times cover a significant fraction of the lead time of a lot.

2.4 A convolution algorithm for the lead time distribution

For our present purposes, we make the simplifying approximation that the times spent at successive machines are independent random variables. In this section, we describe a convolution algorithm to find the lead time distribution over several operations and discuss the implications of the algorithm.

Assuming independence, the time spent at operation i follows the distribution given in equation (10) with the appropriate parameters for machine m_i. We initialize the lead time distribution to an impulse with probability 1 at zero. This represents the lead time for no operations at all. Then we find the approximate distribution of time spent over several operations by convolving the individual distributions of time spent at an operation into the lead time distribution. This can be formulated as a convolution algorithm:

Let $f_W^0(w) = \delta(w)$.

For $i = 1, \ldots, k$:

Find f_{W_i} from equation (10).

Let $f_W^i(w) = \int_0^w f_W^{i-1}(w-t) f_{W_i}(t)\, dt$.

Since the density in equation (10) is a sum of several terms, we can convolve term by term. Notice that the impulses at the operation durations d_i will convolve with each other, giving a smaller impulse at the sum of the individual operation durations. As the length of the process grows, the size of the impulse goes to zero geometrically fast. Intuitively, this is because the probability of having no delays goes down as the number of process steps increases.

Notice also that the exponential and Erlang terms will convolve with each other, giving higher order hyperexponential or Erlang terms in the result. The first order terms in the lead time distribution result from the convolution of a single exponential delay at some machine with impulses at all other machines. As the process gets longer, the scaling factors of lower order terms will decrease. This reflects the fact that the probability of having exactly one delay also goes down as the process becomes longer. The same reasoning can be applied to second order terms, third order terms, etc.

By careful analysis (see [18], chapter XV.6), it may be possible to prove that this lead time distribution converges to a normal distribution as the number of processing steps goes to infinity, although we will not pursue this result here. If we impose the additional restrictions of identical machines and equal length operations, the time distributions at the individual machines will be identically distributed and the lead time will converge to a normal distribution according to the Central Limit Theorem.

We conclude that the lead time distribution consists of a small impulse at the sum of the processing times and a sum of Erlang and hyperexponential terms, and that higher order terms dominate the lead time distribution for long process flows. This distribution will resemble a gamma distribution, shifted to the right by the sum of all operation durations in the process flow. As the process becomes longer, the distribution will become less skewed.

3 Synthetic sampling of lead times

To avoid some of the restrictions of the probabilistic model presented in section 2, we have used a different approach that generates synthetic lead time samples from the failure and repair history of the factory. By synthetic samples, we mean lead time data for hypothetical lots that can be inferred from the known history of the factory. In this section, we describe the procedure used and present data generated from the ICL by this procedure.

It is possible to infer the state of machines in the laboratory at any time during the last few years from the data stored in CAFE. We can therefore introduce a hypothetical lot to the laboratory at some time in the past and observe its progress. A program has been written to do this, using the CAFE database as input. The program assumes that:

- The lot is the single member of the highest priority class in the laboratory, so it never suffers any congestion delays.

- The lot has standard size.
- All operations take the time stated in the process description. By convention, the operation durations in CAFE include the time required to do any necessary setups etc.
- The lot takes a fixed time to transfer from one operation to the next. This transfer time is added between every pair of successive operations.
- An operation can not be started unless there is sufficient time left in the working day to finish it.
- Any failures that occur during operations are resumable without any additional penalty.

The output of the program is a specification of how much time the lot would have spent processing, in transfer, waiting for repairs, and waiting for any other reason (such as holidays), if it were a real lot in the laboratory. The program also records the progress of the lot by plotting the remaining total operation time as a function of time spent in the laboratory. Figure 2 shows such plots for two lots started at different times. The horizontal lines in the plots indicate periods when the lot was waiting for some reason, and the downward sloping lines indicate times when an operation was in progress. The shorter delays are nights and weekends, and some failures cause long delays. The point where the plot reaches the horizontal axis is the lead time of the lot.

By performing this procedure systematically, we can get an impression of how the lead time varies over time and how it is distributed. As an exper-

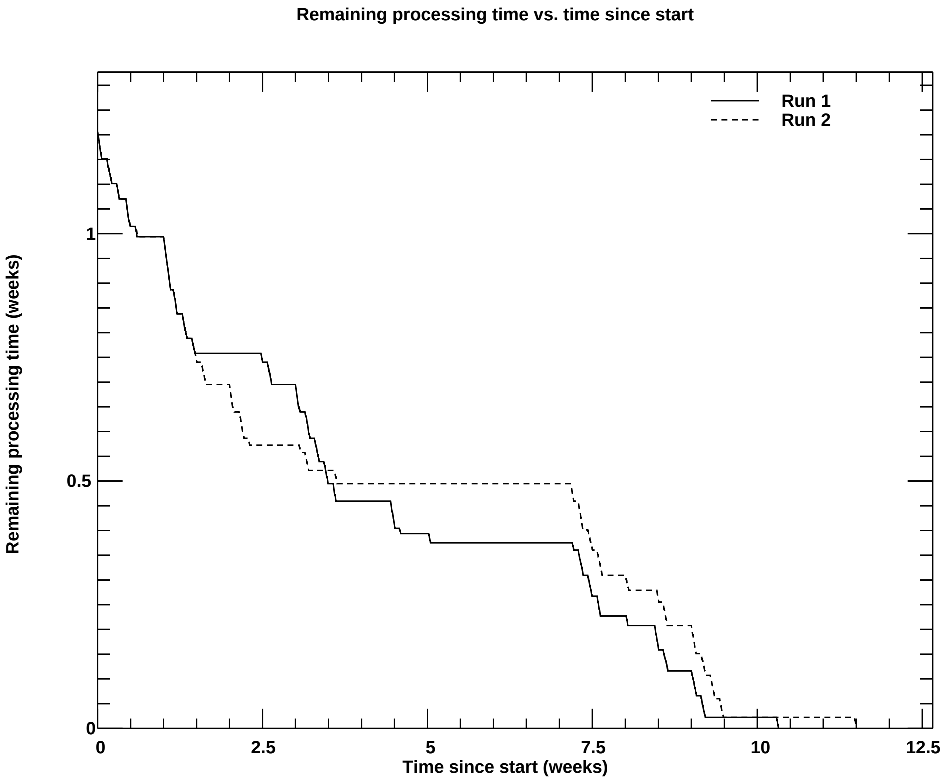

Figure 2: Two sample paths through the laboratory

iment we considered the process flow DA, which consists of 174 operations. We started a hypothetical lot of this process in the Integrated Circuits Laboratory 8am every working day of 1991, 1992, and most of 1993. This interval was chosen to avoid some inconsistent data for 1990 and to ensure that all lots started would finish within the stored history. The laboratory was assumed to work 1.5 shifts a day, 5 days a week, and follow the MIT academic calendar with regard to holidays. We set the transfer time to 30 minutes for this experiment.

If no failures occurred, the lots would spend 203 hours processing, 89 hours transferring, and at least 463 hours waiting during nights and weekends. This gives about 27% processing of the time spent in the laboratory and a minimum lead time of 4.5 weeks. These numbers include maximal length setups in the processing time, as described above. Figure 3 shows the resulting lead time for each data point. It is clear that the data points are strongly correlated.

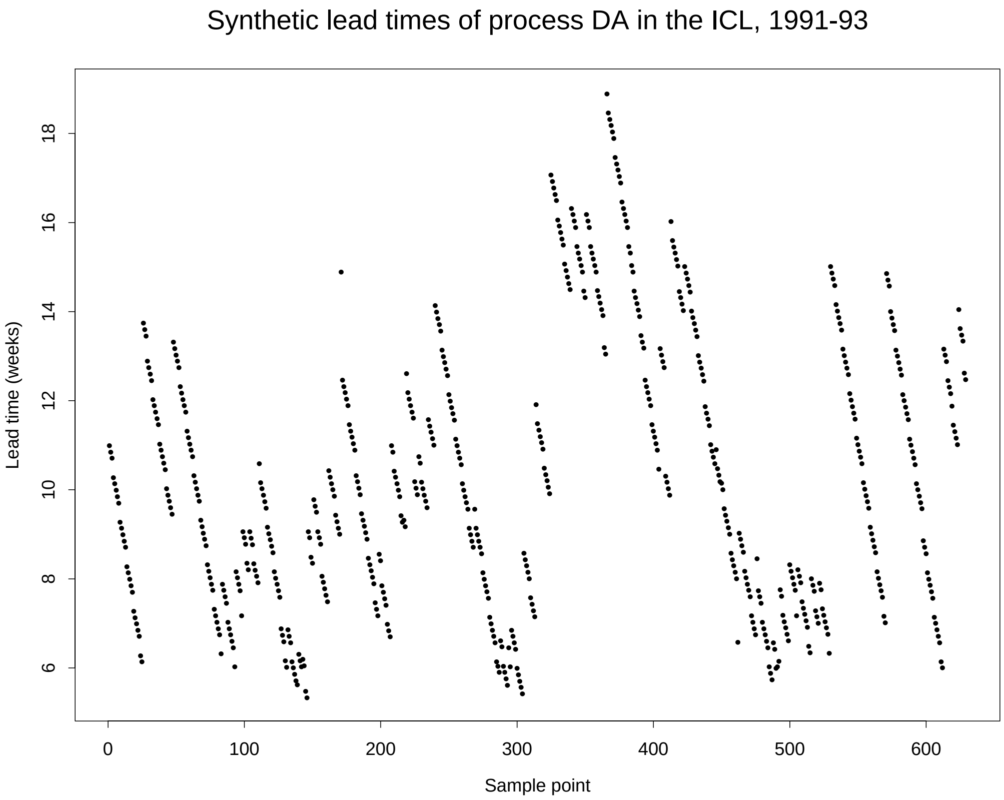

Figure 3: An experiment in the ICL, lead time in weeks

The mechanism that gives the downward sloping runs in figure 3 is that by releasing a lot one day later, the lot is one day closer to the repair of the machine that is delaying it. The upward jumps are caused by lots encountering failures that they would narrowly avoid by starting one day earlier. In the plots of figure 2, the new delay will typically occur at an operation close to the end of the process flow. As the lot is released later and later, earlier operations in the process flow will be affected. It is possible to find the times where such transitions take place and base the analysis of the lead times on this. This approach is known as *perturbation analysis* [19].

The correlation between sample points effectively reduces our sample size for statistical purposes. A less correlated sample can be taken by increasing the spacing between lot releases to a week or a month, which would give fewer data points of presumably better quality, but also give a less detailed picture of the system dynamics. We will use the entire sample in the remainder of this paper.

Figure 4 shows the experimental lead time data as a histogram. Note that the histogram is skewed towards lower values, and does not resemble a normal distribution. This violates the assumption of normally distributed lead times of [14], and suggests that the convergence to a normal distribution is slow, at best. A possible explanation for this is the strong correlation between data points. If a certain sample point has a high lead time, it is quite likely that the next sample point will have a lead time that is precisely one day shorter. The reentrant nature of the processes also contributes to the correlation. For example, if a certain machine is used five times during the process flow, a failure of the machine will cause delays at each of the five operations for different lot release dates. This increases the number of sample points affected by a failure.

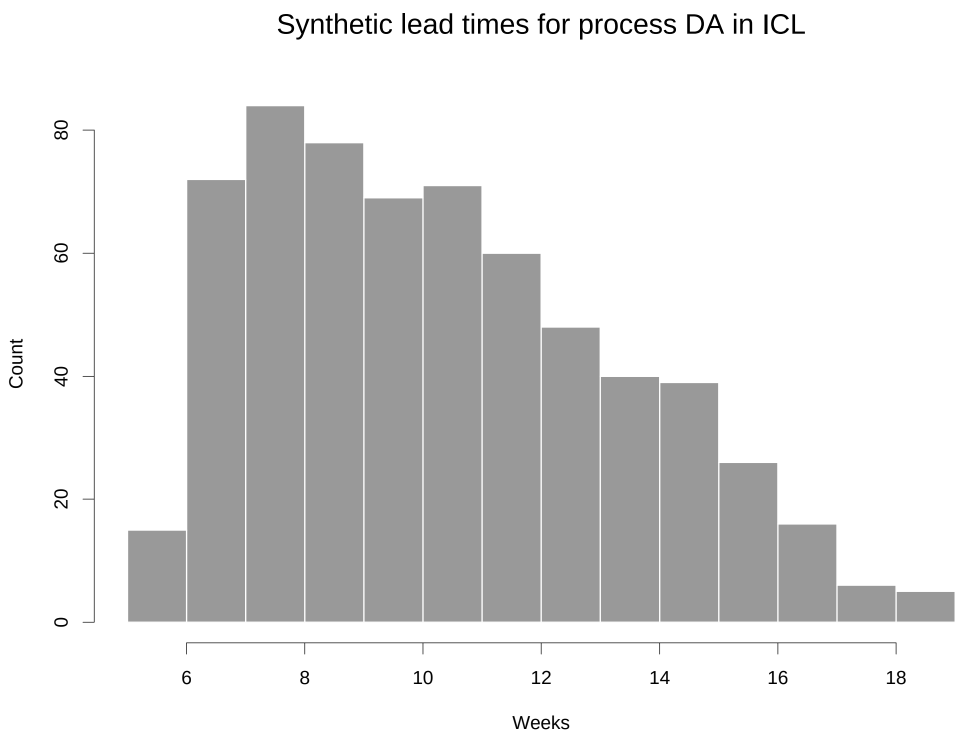

Figure 4: Histogram of the synthetic lead time sample

The correspondence with real-world results is quite good. An actual experiment was made in the ICL to see how quickly these products can be manufactured, and a cycle time of eight weeks was achieved for a single lot by running the laboratory at 2.5 shifts. The laboratory staff also reports day-to-day results that seem to agree with our data.

4 Approximating the Lead Time Distribution

As discussed in section 2.4, the lead time distribution is likely to resemble a shifted gamma distribution. The histogram in figure 4 also indicates the same general shape. The gamma distribution contains the exponential and Erlang distributions as special cases, and converges to a normal distribution as one of the parameters goes to infinity. It therefore seems natural to approximate the lead time distribution with a gamma distribution shifted to the right by the smallest possible lead time.

The total processing time for the DA process is 203 hours. As discussed in section 3, this translates to a minimum lead time of 4.5 weeks when transfer times, nights, and weekends are considered. We therefore define the random variable $X = W - 4.5$ weeks, representing the delay suffered by the lot.

We use a statistical estimation approach, where we use a maximum likelihood estimator for the gamma distribution to simultaneously estimate the two parameters of the distribution. An alternative procedure is to use the method of moments, where one will choose the parameters of the distribution so as to replicate the moments of the observed data, but the maximum likelihood estimator has better properties [20].

In the following, we first derive the maximum likelihood estimator for the gamma distribution. We then apply the estimator to the synthetic data and find an approximate distribution of the delays suffered by lots. Finally, we compare the approximate distribution with the synthetic data.

4.1 A Maximum Likelihood Estimator for the Gamma Distribution

Suppose we observe a random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Assume X_i is independently and identically distributed with the gamma density

$$f_{X_i}(x; a, b) = \frac{x^{a-1}e^{-\frac{x}{b}}}{b^a\Gamma(a)}, \qquad x > 0 \tag{12}$$

where

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}\,dt \tag{13}$$

Then the likelihood function of $\mathbf{X}$ is

$$\begin{aligned} L_{\mathbf{X}}(a,b) &= \prod_{i=1}^n f_{X_i}(x_i; a, b) \\ &= \frac{1}{b^{na}\Gamma^n(a)} \prod_{i=1}^n x_i^{a-1} \exp\left(-\frac{1}{b}\sum_{i=1}^n x_i\right) \end{aligned} \tag{14}$$

and the logarithm of the likelihood function is

$$\log L_{\mathbf{X}}(a,b) = -na\log b - n\log\Gamma(a) + (a-1)\sum_{i=1}^n \log x_i - \frac{1}{b}\sum_{i=1}^n x_i \tag{15}$$

To find the values of a and b that maximize the likelihood function, we differentiate with respect to each of these parameters and set the result equal to zero:

$$\frac{\partial}{\partial a}\log L_{\mathbf{X}}(a,b) = -n\log b - n\frac{\Gamma'(a)}{\Gamma(a)} + \sum_{i=1}^n \log x_i = 0 \tag{16}$$

$$\Rightarrow \log b + \frac{\Gamma'(a)}{\Gamma(a)} = \frac{1}{n}\sum_{i=1}^n \log x_i \tag{17}$$

$$\frac{\partial}{\partial b}\log L_{\mathbf{X}}(a,b) = -\frac{na}{b} + \frac{1}{b^2}\sum_{i=1}^n x_i = 0 \tag{18}$$

$$\Rightarrow ab = \frac{1}{n}\sum_{i=1}^n x_i \tag{19}$$

If we solve equation (19) for b and substitute the result into equation (17), we get the equation

$$\frac{\Gamma'(a)}{\Gamma(a)} - \log a = \frac{1}{n}\sum_{i=1}^{n} \log x_i - \log\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \tag{20}$$

This equation can be solved for a numerically, and we can then find b from equation (19). We observe that the resulting distribution will have the same mean as the experimental data, since $E[X] = ab$ for a gamma distribution, and equation (19) fixes ab to the sample mean.

4.2 Experimental Data

The experimental data yield the following statistics:

$$\begin{aligned} n &= 629 \\ \frac{1}{n}\sum_{i=1}^{n} x_{(i)} &= 5.813998 \\ \frac{1}{n}\sum_{i=1}^{n} \log x_{(i)} &= 1.608886 \end{aligned}$$

where $x_{(i)}$ indicates realization i of the random variable X (delay suffered by the lot). By inserting the experimental data in equation (20), we get the equation

$$\frac{\Gamma'(a)}{\Gamma(a)} - \log(a) = -0.15138 \tag{21}$$

and the solution

$$a = 3.46073$$

$$b = 1.67999$$

by using Mathematica [21].

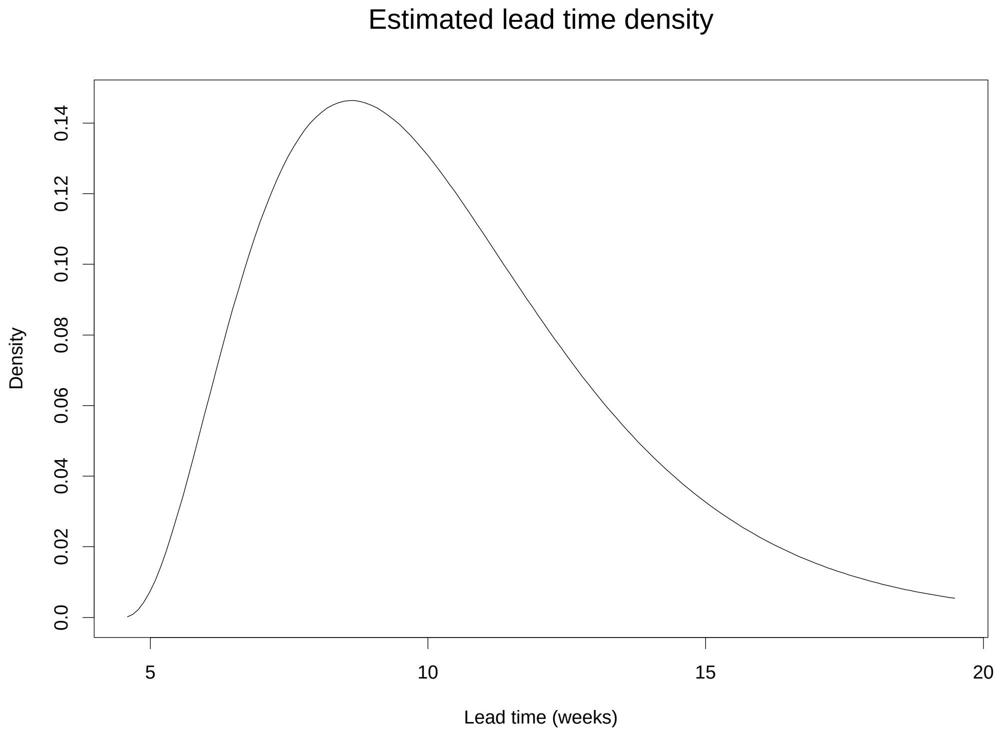

Figure 5: The estimated distribution

Figure 5 shows the probability density function of the estimated distribution, and figure 6 compares the experimental lead time data with the estimated distribution. Each point has an experimental lead time value as its X-coordinate and the corresponding quantile of the shifted gamma distribution as the Y-coordinate. The straight line in the figure goes through the origin with slope 1 and represents perfect agreement between the distributions. This was done in the statistical package Splus [22].

We see that the correspondence is quite good up to about 16 weeks lead

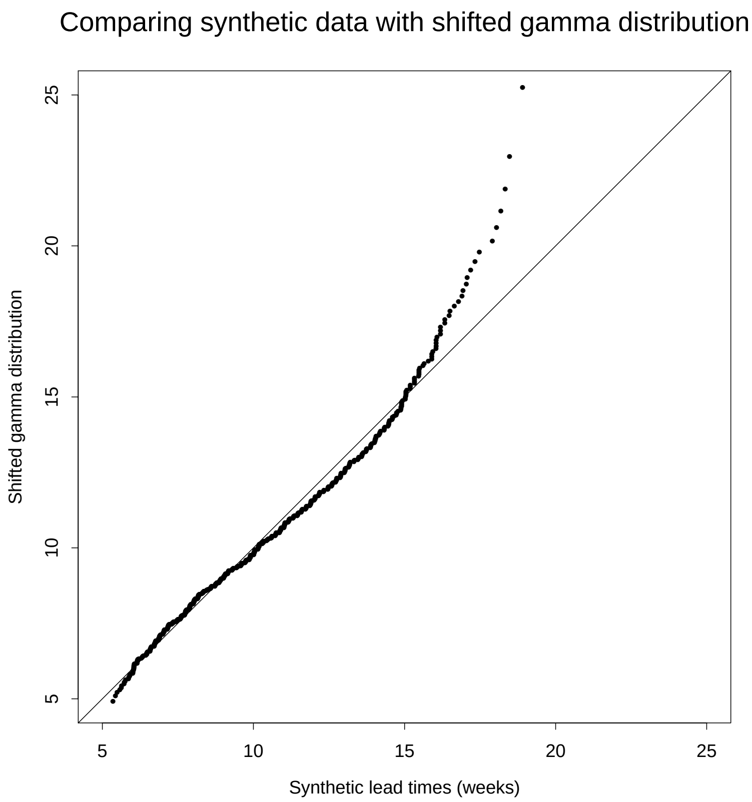

Figure 6: Comparing experimental data and estimated distribution

time, and that the gamma distribution becomes more pessimistic than the experimental data thereafter. The transition point is about two standard deviations from the mean of the distributions. This behavior is to be expected, since the gamma distribution has an infinite tail, while our sample stems from a finite period of time.

We can provide an additional check on the goodness of this approximation by comparing the sample variance of the experimental data with the variance of the gamma distribution. Our data have a sample variance of 9.00 and the gamma distribution has a variance of 9.77. This again indicates good, but less than perfect agreement.

5 Conclusions

We have argued that in one particular model, the lead times of priority lots in a semiconductor factory have a distribution consisting of an impulse at the minimum possible lead time and a sum of higher order hyperexponential and Erlang terms. The lead time distribution can be approximated by a gamma distribution shifted to the right by the shortest possible lead time.

We have developed a sampling procedure to construct synthetic lead time data from the failure and repair history of the facility. This procedure is used to generate experimental data, and we fit a gamma distribution to the data by using a maximum likelihood estimator. The resulting gamma distribution gives a good fit for lead times less than two standard deviations above the mean, and overestimates the probability of higher lead times compared to the experimental data.

This procedure can be useful to estimate the lead time of rarely made or completely new products in an established factory, for instance one-off products or experimental lots. It will not give the lead times in volume production, but it can be argued that the lead time is the most important performance measure for "hot" lots that pass through the factory with high priority. Our model is appropriate for such lots.

6 Acknowledgements

Thanks to Linus F. Cordes of the MIT Microsystems Technology Laboratories for bringing this problem to our attention, for proposing the sampling procedure in section 3, for supplying data from the laboratory, and for patiently explaining how the laboratory operates. Thanks to Tomoyuki Masui of Hitachi for insightful comments on the connection between lead time and quality improvement. Thanks to Stanley B. Gershwin for supplying a constant stream of good advice and for proposing the plot format of figure 2.

References

[1] J. A. Buzacott and L. E. Hanifin, "Models of automatic transfer lines with inventory banks — a review and comparison," *AIIE Transactions*, vol. 10, no. 2, pp. 197–207, 1978.

[2] M. B. McIlrath, D. E. Troxel, D. S. Boning, M. L. Heytens, P. Penfield, Jr., and R. Jayavant, "CAFE — the MIT computer aided fabrication environment," in *Proceedings 1990 IEMT symposium*, pp. 297–305, IEEE/CHMT, 1990.

[3] U. S. Karmarkar, "Manufacturing lead times, order release and capacity loading," in *Logistics of production and inventory* (S. C. Graves, A. H. G. Rinnooy Kan, and P. H. Zipkin, eds.), vol. 4 of *Handbooks in OR and MS*, Amsterdam: Elsevier Science Publishers, 1993.

[4] S. B. Gershwin, *Manufacturing Systems Engineering.* Englewood Cliffs, New Jersey: PTR Prentice Hall, 1994.

[5] J. D. C. Little, "A proof of the queueing formula $L = \lambda W$," *Operations Research*, vol. 9, pp. 383–387, 1961.

[6] S. L. Brumelle, "A generalization of $L = \lambda W$ to moments of queue length and waiting times," *Operations Research*, vol. 20, pp. 1127–1136, 1972.

[7] D. Bertsimas and D. Nakazato, "The distributional Little's law and its applications." To be published, 1992.

[8] D. Nakazato, *Transient Distributional Results in Queues with Applications to Queuing Networks.* PhD thesis, Massachusetts Institute of Technology, 1990.

[9] J. Walrand, *An Introduction to Queuing Networks.* Englewood Cliffs, New Jersey: Prentice Hall, 1988.

[10] A. Seidmann, P. J. Schweitzer, and S. Y. Nof, "Performance evaluation of a flexible manufacturing cell with random multiproduct feedback flow," *International Journal of Production Research*, vol. 23, no. 6, pp. 1171–1184, 1985.

[11] U. S. Karmarkar, "Lot sizes, lead times and in-process inventory," *Management Science*, vol. 33, no. 3, pp. 409–417, 1987.

[12] L. M. Wein, "Due-date setting and priority sequencing in a multiclass M/G/1 queue," *Management Science*, vol. 37, no. 7, pp. 834–850, 1991.

[13] J. A. Buzacott and J. G. Shantikumar, *Stochastic Models of Manufacturing Systems.* Englewood Cliffs, New Jersey: Prentice Hall, 1993.

[14] D. Jacobs and S. Meerkov, "Due time performance in lean and mass manufacturing environments," Tech. Rep. CGR-93-5, University of Michigan, 1993.

[15] P. B. Chevalier and L. M. Wein, "Scheduling networks of queues: Heavy traffic analysis of a multistation closed network," *Operations Research*, vol. 41, no. 4, pp. 743–758, 1993.

[16] L. M. Wein, "Scheduling networks of queues: Heavy traffic analysis of a multistation network with controllable inputs," *Operations Research*, vol. 40, no. Supp. No. 2, pp. S312–S334, 1992.

[17] S. M. Ross, *Stochastic Processes.* New York: John Wiley and Sons, 1983.

[18] W. Feller, *An introduction to probability theory and its applications*, vol. 2. New York: John Wiley and Sons, 1971.

[19] C. G. Cassandras, *Discrete Event Systems : Modeling and Performance Analysis.* Homewood, Illinois: Aksen Associates, 1993.

[20] S. F. Arnold, *Mathematical Statistics.* Englewood Cliffs, New Jersey: Prentice Hall, 1990.

[21] S. Wolfram, *Mathematica : A system for doing mathematics by computer.* Redwood City, California: Addison Wesley, 2 ed., 1991.

[22] R. A. Becker, J. A. Chambers, and A. R. Wilks, *The new S language : A programming environment for data analysis and graphics.* Pacific Grove, California: Wadsworth & Brooks/Cole, 1988.