Estimating Average Production Intervals Using Inventory Measurements: Little's Law for Partially Observable Processes

Ardavan Nozari; Ward Whitt

Operations Research, Vol. 36, No. 2, Operations Research in Manufacturing (Mar. - Apr., 1988), 308-323.

Stable URL:
http://links.jstor.org/sici?sici=0030-364X%28198803%2F04%2936%3A2%3C308%3AEAPIUI%3E2.0.CO%3B2-C

ESTIMATING AVERAGE PRODUCTION INTERVALS USING INVENTORY MEASUREMENTS: LITTLE'S LAW FOR PARTIALLY OBSERVABLE PROCESSES

ARDAVAN NOZARI
Salomon Brothers, New York, New York

WARD WHITT
AT&T Bell Laboratories, Murray Hill, New Jersey
(Received July 1986; revision received June 1987; accepted September 1987)

This paper proposes an indirect approach to estimate average production intervals (the length of time between starting and finishing work on each product) using work-in-process inventory measurements. The idea is to apply a modified version of Little's law ($L = \lambda W$) from queueing theory to cope with stochastic processes that are not directly observable. When the actual amount of completed product to be produced from the current work-in-process is not known, we suggest working with an appropriate expected amount of completed product associated with current work-in-process, taking care to properly account for such features as partial yields, changing lot sizes and reconstituted lots. This indirect estimation procedure can be applied to computer simulation as well as direct system measurement. The approach also can be used to calculate expected values of steady-state random variables in mathematical models.

To manage manufacturing systems, it is usually important to know how long it takes to manufacture each product. We call this time the *production interval*; it is sometimes called the *throughput time* or the *cycle time*, and in general queueing terminology, it is called the *sojourn time*. The production interval includes all the processing (service) time plus all the waiting time (the time spent while resources to carry out the next processing step are unavailable). In manufacturing, the total processing time is often called the *butt-to-butt time*; that is, what the production interval would be if all the processing steps could be carried out successively without any delays. Unfortunately, delays often constitute a significant portion of the production interval. Consequently, queueing theory can help to analyze manufacturing systems. In this paper we apply queueing theory to help *estimate average production intervals*. The indirect estimation methods proposed here are intended to help analyze actual production systems, i.e., to help analyze data from factories, but the indirect estimation methods can also be applied to computer simulation and mathematical models, and to other problems besides manufacturing.

When Are Production Intervals Important?

Production intervals tend to be less important in high volume production, where each item need not be identified and demand can be satisfied with inventory. Production intervals are more important in custom manufacturing, where special products are produced in response to individual orders. The production intervals determine how long it takes to meet the orders, i.e., the lead-times in inventory analysis. Production intervals are especially important in the development of complex products, such as large electronic switching systems, which are composed of many custom components and subcomponents. The production intervals are also important for production scheduling. With the trend toward more flexible manufacturing, production intervals will evidently become more important.

Defining Production Intervals

There are difficulties defining what we mean by a production interval in a specific manufacturing setting. First, we must identify the *scope of the production*

Subject classification: 345 estimating average production intervals, 681 conservation laws, 799 indirect estimation via Little's law.

0030-364X/88/3602-0308 $01.25

system. The scope might include one factory, one line in a factory or even one portion of a line, e.g., the test area or the final assembly area. On the other hand, the scope might include more than one factory because the production intervals within one factory can be dominated by the raw material delays. For example, electronic switching systems require circuit packs that are produced at other factories; in turn, the circuit packs require integrated circuits and other components produced at other factories. Experience indicates that delays caused by component shortages often contribute significantly to production intervals.

Furthermore, it is not always clear what it means to start work on a product. For example, a product may be assembled from more than one part, and work may start on these different parts at different times. Also, products may not maintain their identity from start to finish. As production proceeds, the lot sizes often change. At some stages lots are combined; at other stages lots are split. There is also the issue of yield; i.e., some of the product may be scrapped before completion. Finally, some material may require rework and different partial lots may be combined to produce reconstituted lots.

We do not try to resolve all these issues here. In fact, it is evident that these issues must be addressed in each application. However, the indirect estimation methods we propose can be applied very broadly, e.g., they are not limited to one specific scope of the production system or one specific interpretation of what it means to start work on a product. We will present a concrete example to fix these ideas.

A Statistical Problem: Time Stamping and Simple Averages

Production intervals typically fluctuate due to factors such as changes in the product mix, the labor force, machine availability and yields, so that it is necessary to address the statistical problem of summarizing somewhat diverse production interval data. In this paper we are concerned with estimating *average production intervals.*

The standard direct approach is to time stamp some or all of each product at the start and finish. For each product, this yields a sequence of observed production intervals $\{I_k\colon k \geq 1\}$. We then compute the average production interval for n observations in the usual way as $\overline{I}_n = n^{-1} \sum_{k=1}^{n} I_k$.

Because of the fluctuations, it is usually appropriate to view this average production interval $\overline{I}_n$ as an estimate of an unknown quantity obtained via a statistical experiment. Consequently, it is usually appropriate to apply statistical techniques to determine the precision of this estimate; e.g., to estimate confidence intervals. However, elementary methods to obtain confidence intervals based on independent observations are rarely applicable because the successive production interval observations usually are highly dependent. The problem caused by the dependence of successive observations and the cost of extensive time stamping can be alleviated by only occasionally time stamping products. This obviously reduces the number of observations, and so, reduces the statistical precision of the estimate based on data over a given time period. These are the same statistical issues that arise in the analysis of simulation output; see Bratley, Fox and Schrage (1983) and Law (1983).

An Alternate Indirect Estimation Method

Time stamping is often inconvenient and costly. We must record when each product starts and finishes. In factories, this typically requires special equipment and/or special operator actions. With highly sophisticated computer-controlled production systems such information may be readily available, e.g., Dunietz et al. (1986), but with less sophisticated systems, time stamping may be difficult. Even when time stamping is easy to do, it is helpful to have another method to compute average production intervals to serve as an additional consistency check.

Time stamping is also achieved with simulation, but extensive time stamping can easily violate limits on available memory on a small computer. Long runs may fail to produce any useful results because memory limits are eventually violated, and cause the program to abort.

The purpose of this paper is to show how to estimate average production intervals via an indirect method based on measurements of work-in-process inventory (WIP). The idea is to apply a modified version of Little's law, $L = \lambda W$, from queueing theory; see Little (1961), Stidham (1974) and Section 11.3 of Heyman and Sobel (1982).

However, the indirect estimation must be done with care because there are complications. In particular, we want to properly account for such factors as partial yields (defective product that is discarded), rework and changing lot sizes. For example, there may be a substantial amount of defective product that is discarded relatively early in the production process. We do not want to underestimate the average production interval of good product by counting the short times

spent in the system by defective product. *In this paper, we are concerned with estimating the average production interval for good product.* To illustrate, we give an example motivated by experience with integrated circuit manufacturing; see Burman et al. (1986).

Basic Example

Consider a production facility with five workstations where a single product (whittjits) is produced. The standard sequence of workstations visited by each unit is (1, 2, 3, 4, 2, 3, 5), but exceptions exist because of partial yields and rework. (In integrated circuit manufacturing, a thin silicon wafer is transformed into an integrated circuit using photolithography. Several layers are superimposed on the wafer, requiring basic operations to be repeated, so that the product typically visits the same workstation several times. In the basic sequence of our example some stations are visited twice.) A production flow and queueing diagram for this example appears in Figure 1.

New units enter the system before station 1 in lots of size 5 at a rate of 10 lots per hour. The lots go to a machine at station 1, where individual units are processed one at a time. Units from station 1 go next to station 2 in lots of size 4, where they are processed in batches of 20 units. Processing does not begin at station 2 until 20 units are available. After the batch of 20 units is processed, it is moved to a buffer. From this buffer, units are transported in lots of size 5 to station 3.

Before describing the rest of the operations, we point out that, in general, the units can be in one of three places at each station: waiting to be processed, being processed, or waiting after being processed. Hence, in Figure 1 there are at least two buffers at each station. (There are three buffers at station 3.) The essential point is that we should be careful to count each unit only once. For example, individual units are processed one at a time at station 1, but we must also count the other units in the same arriving lot because they are in service too. Furthermore, there is the transport system. It is often reasonable to neglect it, as we will do, but sometimes it is important to include.

Station 3 is an inspection station. Individual units are tested one at a time. Of the units not tested before, 10% fail completely and are junked, 20% fail and are sent back to station 1 for reprocessing (rework), and the remaining 70% are sent to station 4. The units passing inspection move to station 4 in their original lots, which are of size 5 if no units in the lot fail the test, but are less than 5, otherwise. For simplicity, we assume that the failure events for units within a lot and among successive lots are mutually independent. Consequently, the probability that all 5 units in a lot fail is $(0.3)^5 = 0.00243$; since it is small, we neglect it. This means that we assume for each lot tested there always is a lot of good units going from station 3 to station 4. The average size of the partial lot on the first pass is $5(0.7) = 3.5$.

Units sent back from station 3 to station 1 for rework are first set aside in a buffer. The units are reconstituted in this buffer into lots of size 5 and then transported to station 1. These units are processed at stations 1 and 2 and tested at station 3, just like new units. Of these reworked units, 20% fail their second test at station 3 and are junked, while the remaining 80% pass the test and move to station 4. (Rework is done at most once for each operation step.) Again the units leave station 3 in their arriving lots, which may be reduced by units failing the test. The average size of these reworked lots is $5(0.8) = 4.0$. The average flow rate of lots (units) per hour arriving at station 4 for the first time is $10 + 0.2(50)/5 = 12$ lots/hr. ($10(3.5) + 2(4.0) = 43$ units/hr.).

At station 4, entire lots are processed one at a time

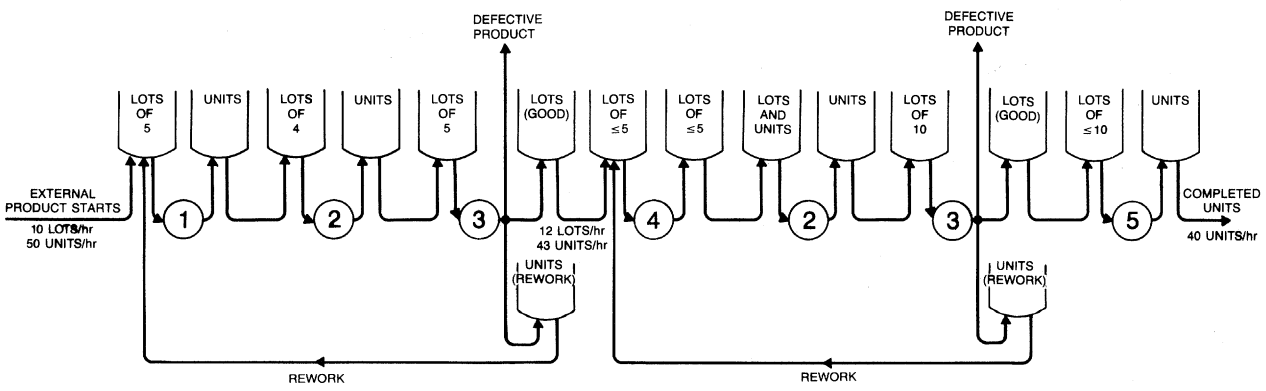

Figure 1. Production flow and queueing diagram for the basic example.

by one of three available machines and then sent to station 2. (A queueing model would have a three-server queue.) On this second visit to station 2 (not counting reworks), units are processed in batches of size 30, but sometimes processing is done with partial batches. The average service batch size is 25. After the batch is processed at station 2, it is moved to a buffer, from which units are transported to station 3 in lots of size 10, where they are tested again, one unit at a time.

We now describe the test percentages for the units processed at station 4. On the first pass, 5% of the units tested fail completely and are junked, 20% fail and are sent back to station 4 for rework (from that point on in the basic sequence), and 75% pass and are sent to station 5. Of those being tested at station 3 for the second time after being processed at station 4, 10% fail and are junked, while 90% pass and are sent to station 5. (Again, rework is done once at most.) Finally, assume that all the failure events are mutually independent. In particular, the failure probabilities after reaching station 4 are independent of the history before station 4 (whether or not rework was required before reaching station 4). Indirect estimation can be done without the independence assumption, but the test percentages must then be described in more detail.

Lots containing the units that pass this second inspection step at station 3 move directly to station 5. Since full lots of size 10 are formed after processing on this visit to station 2, the average flow rate to station 5 for products not requiring rework at station 4 is 4.3 lots/hr. and 32.25 units/hr. Units requiring rework at station 4 are collected in a buffer at station 3, from which reconstituted lots of size 5 are formed and sent back to station 4 to undergo the final processing steps (4, 2, 3, 5). At station 5, individual units are processed one at a time and sent out. The flow rate of good product out of station 5 is $32.25 + (0.86)(10)(0.9) = 40.0$ units/hr.

In this example, the production interval of good product is not difficult to define. For each good unit sent out from station 5, it is the length of time between the original product start for this unit at station 1 and the completion time. With time stamping (of units, not lots), this measurement is readily available. Without time stamping, we want to estimate this average production interval of good product using WIP measurements plus the external product start rates and the testing percentages, which might be readily available as described above. Of course, they way we use WIP measurements depends on what we can measure. At some places it may be convenient to count lots, while at other places, it may be convenient to count units. There is also the issue of identifying what stage of production each unit or lot is in. A variety of information conditions are possible. When counting WIP, we might or might not be able to distinguish between new units (or lots) and units (or lots) requiring rework; we might or might not be able to distinguish between those units (or lots) that have been to station 4 and those that have not. It may be possible to apply the indirect estimation method to all these situations, but careful accounting must be done to properly reflect the prevailing information conditions.

We apply our indirect estimation procedure to this example in Section 2. With additional specification (service-time distributions, queue disciplines, buffer sizes, etc.) and under additional conditions, we can also *predict* the average production intervals without measurements using mathematical models, either by analytic approximations (e.g., Whitt 1983, 1987) or simulation. However, note that we would have to describe the system in much greater detail. It is significant that the estimation based on measurements does not require such additional information. Even some of the information we provided is unnecessary for measurement. For example, we do not need to know the number of machines at station 4. We also do not need to know the average service batch size at station 2 if we can count the number of units in the batch in service.

Organization of the Paper

In Section 1 we review Little's law and discuss its implications for measurements. Section 2 discusses the problem of unobservable processes and the need to modify Little's law to obtain useful indirect estimates. We also analyze the basic example. In Section 3 we begin to investigate the validity of indirect estimation, and introduce a special class of indirect estimators called conditional-expectation observable estimators. In Section 4 we treat estimators based on a detailed classification of WIP, which includes the basic example. Section 5 shows that multiclass queueing networks described in terms of one-step transitions are covered by the general framework of Section 4. Finally, the conclusions are presented in Section 6.

1. Estimates Based on Little's Law

To set the stage, we briefly review Little's law. Let $\{N(t): t \geq 0\}$, $\{A(t): t \geq 0\}$ and $\{I_k: k \geq 1\}$ be stochastic processes representing, respectively, the number of units in the system at time t, the cumulative number

of product starts (arrivals) in the interval $[0, t]$ and the production interval of the kth starting unit. Little's law relates the limiting averages. (All limits are understood to be with probability one without mention.) In particular, given

$$L = \lim_{t\to\infty} t^{-1} \int_0^t N(s)\, ds, \qquad \lambda = \lim_{t\to\infty} t^{-1}A(t)$$

and

$$W = \lim_{n\to\infty} n^{-1} \sum_{k=1}^{n} I_k, \tag{1}$$

Little's law concludes that $L = \lambda W$. Hence, to know the long-run average production interval W, it suffices to know the long-run average WIP L and the long-run arrival rate λ; i.e., we can use $W = L/\lambda$. Little's law also helps determine the existence of the limits; e.g., if the limits associated with λ and W exist, then so does the limit associated with L.

In many mathematical models, the limiting averages L and W coincide with the expected values of steady-state random variables. Then Little's law provides a relationship between these expected values. Similarly, the following discussion of estimating average production intervals also applies to calculating the expected value of a production interval in steady state, given the expected values of the inventories in steady state, but we only discuss the averages.

Discrete Averages and Continuous Time Averages Based on Measurements

Applications based on measurements do not involve the infinite limits in (1). Instead, we observe the finite averages

$$\bar{N}(t) = t^{-1} \int_0^t N(s)\, ds, \qquad \bar{A}(t) = t^{-1}A(t)$$

and

$$\bar{I}_n = n^{-1} \sum_{k=1}^{n} I_k \tag{2}$$

for some fixed t or n. (In our notation, we use L, λ and W to designate the limits and N, A and I to refer to observations and finite averages.) If the averages are sufficiently close to their limits, then

$$\bar{I}_n \approx W = L/\lambda \approx \bar{N}(t)/\lambda \approx \bar{N}(t)/\bar{A}(t). \tag{3}$$

Our first idea is to invoke (3) and approximate W or $\bar{I}_n$ by $\bar{N}(t)/\lambda$ or $\bar{N}(t)/\bar{A}(t)$.

Typically, the rate of production starts, λ, or its sample average $\bar{A}(t)$ is readily available, so that to implement this indirect estimation procedure is tantamount to calculating the average WIP, $\bar{N}(t)$. The motivation for the indirect procedure is that $N(t)$ is often much easier to calculate than I_n because $N(t)$ involves counting WIP at one time t only without identifying individual products, whereas I_n involves a careful time measurement of the start and finish for an individual product. However, at first glance, this advantage may seem to be outweighed by the replacement of the relatively simple discrete average $\bar{I}_n$ by the continuous time average $\bar{N}(t)$ in (2); i.e., $\bar{N}(t)$ involves an integral instead of a sum. However, it is not too difficult to calculate this integral because $N(t)$ is a step function: there are times t_i with $0 = t_0 < t_1 < \cdots < t_m = t$ such that

$$\begin{aligned} \bar{N}(t) &= t^{-1} \int_0^t N(s)\, ds \\ &= t^{-1} \sum_{k=1}^{m} N(t_i)(t_i - t_{i-1}); \end{aligned} \tag{4}$$

i.e., the integral representation for $\bar{N}(t)$ in (2) immediately reduces to a simple sum. The times t_i in (4) are, of course, the times at which $N(t)$ changes value. Indeed, (4) is the standard transaction-based procedure to estimate the time-average queue length in a discrete-event simulation, e.g., pp. 81–83 of Bratley, Fox and Schrage, so that the integral in (2) really imposes nothing new. (Statistics such as $\bar{N}(t)$ and $\bar{A}(t)$ in (2) are also called *time-persistent statistics*; see Pritsker and Pegden 1979.)

Nevertheless, it may be inconvenient to calculate the time average $\bar{N}(t)$ in detail via (4). Then we can estimate $\bar{N}(t)$ by something easier. For example, we can simply observe $N(t)$ at k separate time points and take the sample average $k^{-1} \sum_{i=1}^{k} N(t_i)$. The special case of $t_i = it/k$, $1 \leq i \leq k$ constitutes the standard discretization procedure, which is also discussed by Bratley et al. A special case of great practical importance is $k = 1$. One can often obtain a reasonable, rough estimate of the average production interval by walking into the factory, counting WIP once and invoking Little's law!

Two Issues: Statistical Precision and the Cost of Obtaining the Estimate

We see that there are several possible estimators for the long-run average production interval W in (1); e.g., $\bar{I}_n$ in (2) and $\bar{N}(t)/\lambda$ or $\bar{N}(t)/\bar{A}(t)$ in (3), with $\bar{N}(t)$ determined via the transaction-based sum in (4) or some sampling procedure. How do we choose from among these procedures?

We suggested alternatives to $\bar{I}_n$ in (2) because they may be more convenient, i.e., overall, they may be

more cost effective. An important aspect is the statistical precision of these different estimators. It turns out that the relations among the stochastic processes that yield Little's law ($L = \lambda W$) also yield important information about the statistical precision of the different estimators. This was first discovered for the $M/G/1$ queue by Law (1975), extended to the $GI/G/s$ queue by Carson and Law (1980), and extended to the general case (covering the kind of models discussed in this paper) by Glynn and Whitt (1986a, b, c). In general, under some reasonable conditions, it is *more asymptotically efficient* to estimate W directly by $\bar{I}_n$ than indirectly by $\bar{N}(t)/\lambda$ given that the system is observed for the same length of time and that λ is known. (Similarly, it is more asymptotically efficient to estimate L indirectly by $\lambda\bar{I}_n$ than by $\bar{N}(t)$ when λ is known.) However, if λ must also be estimated from the same data, then the asymptotic efficiencies are equal. (Note that these are large-sample results.)

Thus, assuming that λ is known and that each estimate can be calculated with equal effort, we probably should prefer the direct estimator $\bar{I}_n$ in (2). However, this does not take account of the effort required to obtain the estimate. A proper notion of efficiency should account for *both* statistical precision and the cost to obtain the estimate, as suggested by Glynn and Whitt (1986d). We are motivated to consider $\bar{N}(t)/\lambda$ instead of $\bar{I}_n$ because of the cost to obtain the estimate.

We do not intend to resolve the many statistical issues here. For example, given that we are going to estimate $\bar{N}(t)$ in (2) by a finite sample average $k^{-1}\sum_{i=1}^{k} N(t_i)$, how should the time points t_i be selected to obtain good statistical precision? See Halfin (1982) and the references there for past work on this important problem.

2. The New Twist: Partially Observable Processes

What we suggested in Section 1 amounts to a direct application of Little's law to obtain a new estimator for the long-run average production interval W in (1). However, there is a fundamental difficulty in this approach, which is directly linked to the difficulty in defining a production interval, as mentioned in the introduction.

As illustrated by the Basic Example, some products produced may be defective. If defectives are discovered early in the production process and discarded, then the time spent in the system by these defective items may be significantly less than the time spent in the system by good product. Thus, we may want to stipulate that the average production interval refers only to good product. In particular, we wish to estimate the average time from start to finish, *conditional on the item turning out to be good product.*

Thus, what we want to observe for Little's law is only the WIP of product that will eventually be good. However, this good WIP is not directly observable. Some units are already defective but have not yet been inspected, while other units are still good but will become defective in some future operation. To address this problem, we suggest working with an appropriate *expected* amount of good product associated with current WIP. However, this requires some care.

There is another difficulty with the definition that is important to note. We may want to know the time from some product start until a good product is completed (for a product started at that time or later). If the given product at that epoch turns out good, then the quantity we focus on, the average production interval given that the product is good, is the correct quantity. However, if the product is destined to be defective, we have to include the delay from that product's start epoch until the next start epoch of a product that is destined to be good. Since the actual value of this additional delay cannot be observed in advance, we suggest working with a long-run average delay. This is the average interval between product starts divided by the proportion of defectives. However, henceforth, we focus on the average production interval for good products, where this is defined as the average production interval counting good products only.

If the production system produces custom products to order, then in many cases, special new items must be started at the beginning of the line to replace items that are defective. If these additional product starts are not counted separately, then we apply Little's law without making any special adjustment for defectives. We can think of these additional restarts to replace defectives as reworks, so that there are no defectives.

Whether or not we represent partial yield (defectives discarded before the end of the production process), other difficulties remain. We typically do not observe one overall WIP. The product is typically observed in different stages of production. It may only be convenient to count lots, but lot sizes may change from workstation to workstation, and the lot sizes may be unknown (random), as in the basic example (arrival at station 4).

The Proposed Procedure

What we propose, then, is a detailed classification of WIP. We define a WIP vector $[N_1(t), N_2(t), \ldots, N_n(t)]$, where $N_i(t)$ is the observable amount of WIP

at state i of production, in whatever units are convenient, and define an *expected overall WIP of good product* by

$$\tilde{N}^0(t) = \sum_{i=1}^{n} N_i(t)\alpha_i, \tag{5}$$

where α_i is the expected amount of good product produced from each item of WIP at stage i (appropriately adjusting for any differences in measuring units). The superscript 0 designates good product. We then estimate the average production interval of good product by

$$\hat{N}^0(t)/\lambda^0,$$

where

$$\hat{N}^0(t) = t^{-1}\int_0^t \tilde{N}^0(s)\,ds; \tag{6}$$

i.e., $\hat{N}^0(t)$ in (6) is defined like $\bar{N}(t)$ in (2) using $\tilde{N}^0(t)$ in (5) insead of $N(t)$. In (6), we can use either the completion rate of good product λ^0, or its estimate $\bar{A}^0(t)$, defined as $\bar{A}(t)$ in (2), where $A^0(t)$ is the amount of product started in the interval $[0, t]$ that will eventually be good. In fact, since we cannot observe this arrival process $A^0(t)$, we suggest using the departure process $D^0(t)$: the amount of good product completed in the interval $[0, t]$. With the proper measuring units,

$$\lambda^0 = \lim_{t\to\infty} t^{-1}A^0(t) = \lim_{t\to\infty} t^{-1}D^0(t). \tag{7}$$

(See Section 2 of Glynn and Whitt (1986a) for relations among the limiting averages of the arrival and the departure processes.) It is significant for applications that the weighted sum in (5) yields a single process to monitor. Moreover, the time average in (6) reduces to a simple sum, as in (4). We can now proceed by applying (5) through (7) to the basic example in the introduction.

The Basic Example Revisited

As shown in Figure 1, there are at least three places WIP can be located at each station for each operation. Since the chance of good product depends on the rework history, we should know whether or not units were tested and if they required rework. A possible classification into 43 classes appears in Table I. This classification assumes that material requiring rework can be identified. It also assumes that we can identify whether or not lots or units have reached station 4. Finally, it assumes that we can count units in batch service at station 2. We have separate classes for lots and units waiting to be processed at station 2 (classes 23 and 24, and 34 and 35), because both may be present. Note that some classes involve units and some involve lots, and some of the lots have fixed size while others have random size.

The factors α_i in (5) that give the conditional expected number of good units produced from each class-i item depend on the expected number of units per class i item and the probability or proportion of class i units that will eventually be good product. (We use "item" to refer to the class i object, which in general may be a unit, a lot, a batch or something else.) It is easy to determine the probabilities of eventually being good product from the specified test percentages. A testing event tree (like a decision tree) for this basic example appears in Figure 2. There are two standard tests, each may be repeated once because of rework. The conditional probabilities of being good product for this example are displayed in Table II. (We use the independence assumption to eliminate the rework history prior to station 4 after reaching station 4.) Finally, the overall conditional-expected-good-product factors α_i in (5) are computed by multiplying the expected number of units in each class item times the probability that each unit of class i is good. The data for this example are displayed in Table III, as are the flow rates for each class item. For example, we stipulate that lots of size 5 start at 10 per hour; thus 50 units start per hour. Since 80% of starting units ultimately become good product (Table II: 0.525 + 0.126 + 0.029 + 0.120 = 0.800), the completion rate is $\lambda^0 = 40$ units per hour. (We omit entries for classes 23 and 34; we convert lots into units in the waiting space before station 2.)

Our indirect estimate of W^0, then is (6) with $\lambda^0 = 40$ units/hr. and $\tilde{N}^0(t)$ defined in (5) with α_i expected good units per class-i item given in Table III.

Aggregation with Less Information

In some cases, it might not be convenient to classify WIP in such detail. For example, in the basic example it might not be expedient to identify material that requires rework. We can then aggregate the classes, which in this case, reduces the number from 43 to 23, as shown in Table IV; e.g., classes 1 and 11 are grouped together to form a single class.

The new class flow rate is the sum of the flow rates for the component classes. The new expected number of units per class item is then the average of the previous individual class values, weighted by the item flow rates per class. The new probabilities that units are good is the average of the previous individual class

Table I
A Definition of Classes for the Basic Example

Class Index	Class Definition
1	New lots waiting at station 1 to be processed
2	New units being processed at station 1
3	New units waiting at station 1 after being processed
4	New lots that have not been to station 4 waiting at station 2 to be processed
5	New units that have not been to station 4 being processed at station 2
6	New units that have not been to station 4 waiting at station 2 after being processed
7	New lots that have not been to station 4 waiting at station 3 to be tested
8	New units (that have not been to station 4) being tested at station 3
9	Units (that have not been to station 4) waiting at station 3 for rework
10	New lots (that have not been to station 4) waiting at station 3 after testing to go to station 4
11	Rework lots waiting at station 1 to be processed
12	Rework units being processed at station 1
13	Rework units waiting at station 1 after being processed
14	Rework lots (that have not been to station 4) waiting at station 2 to be processed
15	Rework units (that have not been to station 4) being processed at station 2
16	Rework units (that have not been to station 4) waiting at station 2 after being processed.
17	Rework lots (that have not been to station 4) waiting at station 3 to be tested
18	Rework units (that have not been to station 4) being tested at station 3
19	Rework lots (that have not been to station 4) waiting at station 3 to go to station 4 after testing
20	Lots on first pass to station 4 waiting at station 4 to be processed
21	Lots on first pass to station 4 being processed at station 4
22	Lots on first pass to station 4 waiting at station 4 after being processed
23	Lots from first pass to station 4 waiting at station 2 to be processed
24	Units from first pass to station 4 waiting at station 2 to be processed
25	Units from first pass to station 4 being processed at station 2
26	Units from first pass to station 4 waiting at station 2 after being processed
27	Lots from first pass to station 4 waiting at station 3 to be tested
28	Units from first pass to station 4 being tested at station 3
29	Units from first pass to station 4 waiting at station 3 for rework
30	Lots from first pass to station 4 waiting at station 3 to go to station 5
31	Rework lots waiting at station 4 to be processed
32	Rework lots being processed at station 4
33	Rework lots waiting at station 4 after being processed
34	Rework lots (after station 4) waiting at station 2 to be processed
35	Rework units (after station 4) waiting at station 2 to be processed
36	Rework units (after station 4) being processed at station 2
37	Rework units (after station 4) waiting at station 2 after being processed
38	Rework lots (after station 4) waiting at station 3 to be tested
39	Rework units (after station 4) being tested at station 3
40	Rework lots (after station 4) waiting at station 3 to go to station 5 after testing
41	Lots that did not require rework after station 4 waiting at station 5 to be processed
42	Lots that required rework after station 4 waiting at station 5 to be processed
43	Units being processed at station 5

probabilities weighted by the unit flow rates per class (item flow rate times expected number of units per class item).

It is important to note, however, that *there is no guarantee that this aggregation will yield the desired results*. To see this, we could aggregate everything into a single class by the scheme just specified. This would be tantamount to replacing (5) by $(\sum_{i=1}^{n} N_i(t))\alpha$ for some single parameter α. While the reduction in Table IV may be reasonable, we should not expect a total aggregation into one class to perform well. Indeed, it clearly is not reasonable because we would add the number of units at one place to the number of lots at another place.

From the discussion above, it is evident that we have not said anything about when the indirect estimation procedure is appropriate. We address this question next.

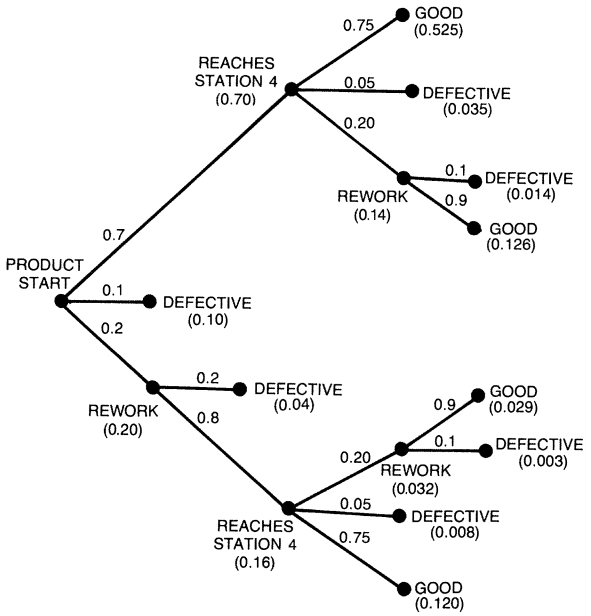

Figure 2. Testing event tree for the basic example.

3. Conditional-Expectation Observable Estimators

In this section we introduce a class of indirect estimators that are consistent (asymptotically correct). When the proposed estimator (6) takes this form, then it is consistent.

As in (5) to (7), let a superscript 0 refer to good product; let $N^0(t)$ be the number of good units eventually produced from current WIP at time t; let $A^0(t)$ be the number of good units eventually produced from product starts in the interval $[0, t]$; and let I_k^0 be the production interval for the kth unit started that is eventually good product. We index I_k^0 in the order of arrival, which for the purpose of determining long-run averages is usually equivalent to indexing in order of completion. However, this is not always the case; some products could have extraordinarily large delays that distort the averages; see Example 1 of Glynn and Whitt (1986a).

Table II
Conditional Probabilities of Being Good Product Given Testing History

Classification	Probability of Being Good
Not yet tested	0.800
Requires rework after first test	0.744
Reaches station 4, but not tested afterward	0.930
Requires rework after station 4	0.900

In addition to the limits in (1), we assume that

$$L^0 = \lim_{t\to\infty} t^{-1}\int_0^t N^0(s)\,ds, \qquad \lambda^0 = \lim_{t\to\infty} t^{-1}A^0(t)$$

and

$$W^0 = \lim_{n\to\infty} n^{-1}\sum_{k=1}^{n} I_k^0. \tag{8}$$

From Little's law, we know that $L = \lambda W$ and $L^0 = \lambda^0 W^0$. We want to estimate W^0 assuming that we know or can estimate λ^0. Our plan is to invoke Little's law in the form $W^0 = L^0/\lambda^0$ and estimate L^0. However, we cannot apply (8) directly because $N^0(t)$ is unobservable.

A consistent observable estimator obviously is any observable process $\{\tilde{N}^0(t): t \geq 0\}$ for which

$$\lim_{t\to\infty} t^{-1}\int_0^t \tilde{N}^0(s)\,ds = \lim_{t\to\infty} t^{-1}\int_0^t N^0(s)\,ds. \tag{9}$$

Given such a process, we can estimate L^0 by $t^{-1}\int_0^t \tilde{N}^0(s)\,ds$. A sufficient condition for (9) is

$$\mathrm{E}\,\tilde{N}^0(t) = \mathrm{E}\,N^0(t) \quad \text{for all } t \tag{10}$$

and both

$$\begin{aligned}\lim_{t\to\infty} t^{-1}\int_0^t \mathrm{E}\,N^0(s)\,ds &= \lim_{t\to\infty} t^{-1}\int_0^t N^0(s)\,ds\\ \lim_{t\to\infty} t^{-1}\int_0^t \mathrm{E}\,\tilde{N}^0(s)\,ds &= \lim_{t\to\infty} t^{-1}\int_0^t \tilde{N}^0(s)\,ds.\end{aligned} \tag{11}$$

We propose a general procedure to construct the observable process $\tilde{N}^0(t)$ that satisfies (10) using conditional expected values. Let H_t represent a full or partial history of the system up to time t for $t \geq 0$, i.e., an increasing family of σ-fields; see Chapter 4 of Breiman (1968) or Chapters 2 and 9 of Chung (1974) for background on sigma fields and conditional expectation. For the observable process $\tilde{N}^0(t)$, we use

$$\tilde{N}^0(t) = \mathrm{E}(N^0(t)\,|\,H_t), \quad t \geq 0. \tag{12}$$

We assume that H_t represents information that is available at time t, so that $\tilde{N}(t)$ in (12) is observable at time t. A basic property of conditional expectations is that

$$\mathrm{E}[\mathrm{E}(N^0(t)\,|\,H_t)] = \mathrm{E}\,N^0(t), \tag{13}$$

so that (10) is always satisfied by (12); see Section 9.1 of Chung. We may now verify or directly assume the relatively technical condition (11). *For applications,*

Table III
The Conditional-Expected-Good-Product Factors α_i in (5) for the Basic Example

Class Index i	Flow Rate (per hour)	Expected No. of Units	Probability Each Unit Is Good	Overall Factor α_i	Class Index i	Flow Rate (per hour)	Expected No. of Units	Probability Each Unit Is Good	Overall Factor α_i
1	10	5	0.800	4.00	23	–	3.58	0.933	3.34
2	50	1	0.800	0.80	24	43	1	0.933	0.93
3	50	1	0.800	0.80	25	43	1	0.933	0.93
4	12.5	4	0.800	3.20	26	43	1	0.933	0.93
5	50	1	0.800	0.80	27	4.3	10	0.933	9.33
6	50	1	0.800	0.80	28	43	1	0.933	0.93
7	10	5	0.800	4.00	29	8.6	1	0.900	0.90
8	50	1	0.800	0.80	30	4.3	7.50	1.000	7.50
9	10	1	0.744	0.74	31	1.72	5	0.900	4.50
10	10	3.50	0.933	3.27	32	1.72	5	0.900	4.50
11	2	5	0.744	3.72	33	1.72	5	0.900	4.50
12	10	1	0.744	0.74	34	—	5	0.900	4.50
13	10	1	0.744	0.74	35	8.6	1	0.900	0.90
14	2.5	4	0.744	2.98	36	8.6	1	0.900	0.90
15	10	1	0.744	0.74	37	8.6	1	0.900	0.90
16	10	1	0.744	0.74	38	0.86	10	0.900	9.00
17	2	5	0.744	3.72	39	8.6	1	0.900	0.90
18	10	1	0.744	0.74	40	0.86	9.00	1.000	9.00
19	2	4.00	0.933	3.73	41	4.3	7.50	1.000	7.50
20	12	3.58	0.933	3.34	42	0.86	9.00	1.000	9.00
21	12	3.58	0.933	3.34	43	40.0	1	1.000	1.00
22	12	3.58	0.933	3.34					

the problem is to compute the conditional expectation in (12).

The estimator for L^0 associated with (12) is

$$\hat{N}^0(t) = t^{-1} \int_0^t \mathrm{E}(N^0(s) \mid H_s)\, ds, \quad t \geq 0; \tag{14}$$

we call $\hat{N}^0(t)$ and $\hat{N}^0(t)/\lambda^0$ *conditional-expectation observable estimators* for L^0 and W^0.

There is a different conditional-expectation observable estimator for L^0 associated with each family of histories $\{H_t\colon t \geq 0\}$. It we could choose, we would want to pay attention to *as little information as possible.* The best situation is to observe nothing (the trivial σ-field) and use the *no-information estimator* $\mathrm{E}(N^0(t) \mid H_t) = \mathrm{E}\, N^0(t)$ for all t. However, in this case we must know $\mathrm{E}\, N^0(t)$ in advance, which invariably we would not.

More generally, conditioning on less implies lower variability. If $\{H_t^1\colon t \geq 0\}$ and $\{H_t^2\colon t \geq 0\}$ are two families of σ-fields with $H_t^1 \subseteq H_t^2$, then

$$\mathrm{E}([(\mathrm{E}\, N^0(t) \mid H_t^1) - \mathrm{E}\, N^0(t)]^2)$$
$$\leq \mathrm{E}([\mathrm{E}(N^0(t) \mid H_t^2) - \mathrm{E}\, N^0(t)]^2) \tag{15}$$

as can be seen by first conditioning on H_t^2; see Section 9.1 of Chung. So, in this sense too, we prefer to use less information.

On the other hand, the histories can be very rich; indeed there is no limit to the amount of information to include. The obvious choice is to use the least information (smallest possible σ-fields) such that it is possible to compute $\mathrm{E}(N^0(t) \mid H_t)$. (To choose appropriate histories is similar to the problem of defining appropriate states to make stochastic processes Markov.) In fact, it may be difficult to obtain a family of histories $\{H_t;\, t \geq 0\}$ for which it is possible to obtain a reasonable estimate of $\mathrm{E}(N^0(t) \mid H_t)$; then perhaps, the indirect estimation procedure should not be attempted.

From the practical perspective, we can think of *all the relevant information.* When H_t represents all the relevant information up to time t, we call (14) the *full-information observable estimator.* We usually want to work with information somewhere in between no information and full information. Indeed, for a direct application of Little's law, we want to work with the overall WIP $N(t)$. We now introduce two assumptions that make this possible.

Assumption A. For all $t \geq 0$, $\mathrm{E}(N^0(t) \mid N(t)) = \mathrm{E}(N^0(t) \mid H_t)$ where H_t represents all the relevant information up to time t; i.e., all the relevant information at time t is contained in the total WIP $N(t)$.

Assumption B. For all $t \geq 0$, $\mathrm{E}(N^0(t) \mid t)) = N(t)\alpha$ for some known α.

Note that the conditional expectations in Assumptions A and B are random variables; equality is

Table IV
The Aggregate Conditional-Expected-Good-Product Factors α_i in (5) for the Basic Example, after Reducing the Number of Classes from 43 to 23

Class Indices i	Flow Rate (per hour)	Expected No. of Units	Probability Each Unit Is Good	Overall Factor α_i
1, 11	12	5	0.791	3.96
2, 12	60	1	0.791	0.79
3, 13	60	1	0.791	0.79
4, 14	15	4	0.791	3.16
5, 15	60	1	0.791	0.79
6, 16	60	1	0.791	0.79
7, 17	12	5	0.791	3.96
8, 18	60	1	0.791	0.79
9	10	1	0.744	0.74
10, 19	12	3.58	0.933	3.34
20, 31	13.72	3.76	0.928	3.49
21, 32	13.72	3.76	0.928	3.49
22, 33	13.72	3.76	0.928	3.49
23, 34	—	3.76	0.928	3.49
24, 35	51.6	1	0.928	0.93
25, 36	51.6	1	0.928	0.93
26, 37	51.6	1	0.928	0.93
27, 38	5.16	10	0.928	9.28
28, 39	51.6	1	0.928	0.93
29	8.6	1	0.900	0.90
30, 40	5.16	7.69	1.000	7.69
41, 42	5.16	7.69	1.000	7.69
43	40.00	1	1.000	1.00

required with respect to all possible realizations (with probability one).

Under Assumption A, $E(N^0(t) | H_t) = f(N(t))$ for measurable f, see p. 299 of Chung, so that it is much easier to work with the conditional expectation, e.g., to verify (11). For example, if $\{N(t): t \geq 0\}$ is stationary and ergodic, so is $\{f(N(t)): t \geq 0\}$; see pp. 105 and 119 of Breiman. Similarly, regenerative structure for $N(t)$ carries over to $f(N(t))$. Of course, we could choose to work with $E(N^0(t) | N(t))$ even if Assumption A fails. Assumption A is not critical, but if it fails, we are likely to make an error calculating $E(N^0(t) | N(t))$; the obvious error is caused by acting as if Assumption A holds when it does not.

For applications, we need Assumption B. In fact, Assumption B makes the problem formulated in this section relatively trivial; we simply invoke Little's law for all units in the form (1). However, it is important to realize that in many manufacturing applications, Assumptions A and B do *not* hold. More general estimators such as (14) circumvent these problems. The following sections further exploit the additional structure. While Assumptions A and B often do not apply, natural modifications of the assumptions often do.

4. A More Detailed Classification

As in our analysis of the basic example in Section 2, we now classify the WIP in more detail, e.g., in accordance with its stage of production. In particular, for $1 \leq i \leq n$, let $N_i(t)$ be the number of class-i items at time t. The classes might simply represent the workstation where a unit is currently located or other relevant aspects of the history, for example, reworks (as in Section 2). Note that we have not focused on $N(t) = N_1(t) + \cdots + N_n(t)$ because the sum may be meaningless, e.g., we might be adding apples to oranges; this is why we have the α_i factors in (5). We also do not need to define $N_i^0(t)$ because we do not need to associate the eventual good product directly with the class-i WIP items.

We can apply the technique of Section 3, including the simplifications provided by Assumptions A and B, if we make some modifications.

Assumption A′. For all $t \geq 0$, $E(N^0(t) | [N_1(t), \ldots, N_n(t)]) = E(N^0(t) | H_t)$ where H_t represents all the relevant information up to time t; i.e., the general conditional expectation of the number of good units eventually produced from current WIP given that all the relevant information coincides with the conditional expectation of the number of good units eventually produced from current WIP given only the vector $[N_1(t), \ldots, N_n(t)]$.

Assumption B′. For all $t \geq 0$, $E(N^0(t) | [N_1(t), \ldots, N_n(t)]) = \sum_{i=1}^n N_i(t)\alpha_i$ for some known parameters α_i (which are independent of t).

As in Section 3, only Assumption B′ is critical. Assumption A′ is an extra property to avoid pitfalls, that is, to ensure that $E(N^0(t) | [N_1(t), \ldots, N_n(t)])$ can be calculated relatively directly. As with Assumptions A and B, Assumptions A′ and B′ provide natural simplifications, but the latter often are much more realistic than the former. Under Assumptions A′ and B′ and (11), (6) is consistent.

The discussion may seem to belabor the obvious because no difficulty is apparent in the basic example analyzed in Section 2. However, the indirect estimator $\hat{N}^0(t)/\lambda^0$ in (6) is appropriate because we *implicitly assume* A′ and B′. In fact, for the basic example, A′ and B′ are *implied* by the failure event assumptions (Figure 2).

The importance of Assumptions A′ and B′ began to emerge when we considered possible aggregation

procedures. In particular, A′ and B′ are typically violated with the aggregation. We also illustrate the importance of these assumptions in the following simple examples.

Example 1. Consider a production line with n workstations in series. Let units enter the line at rate λ, but only a proportion p_i of the units entering workstation i proceed; a proportion $1 - p_i$ of the work at station i is defective and is scrapped after processing at station i. As usual, we want to estimate the average production interval of good product, but we can only observe the WIP at each station over time, and do not know whether it will eventually be good product or bad. In this case, we apply the indirect estimator $\hat{N}_0(t)/\lambda^0$ in (6) with the parameters α_i defined by

$$\alpha_i = p_i p_{i+1} \cdots p_n, \quad 1 \leq i \leq n. \tag{16}$$

Here, obviously $\lambda^0 = \lambda\alpha_1$. However, the validity requires additional properties such as Assumption B′, as we illustrate below.

Example 2. Consider the special case of Example 1 with two workstations in series. Let units arrive deterministically at the first station at rate $\lambda = 1$. Let the processing time also be exactly 1 at each station, so that no queues form. Let the system start off empty, with the first unit arriving at station 1 at time 1. Assume that each unit at each station is scrapped with probability ½, independent of all other events, so that $p_1 = p_2 = ½$. Thus, only 1 of 4 units started turns out to be good product. These probabilistic assumptions obviously satisfy Assumptions A′ and B′, so we use the estimator $\hat{N}^0(t)/\lambda^0$ in (6) for W^0 with α_i in (16) and $\lambda^0 = \lambda\alpha_1 = ¼$, which is

$$\frac{\hat{N}^0(t)}{\lambda^0} = \frac{4}{t}\int_0^t \tilde{N}^0(s)\,ds, \tag{17}$$

where the full information observable process is

$$\begin{aligned}\tilde{N}^0(t) &= \mathrm{E}(N^0(t) \mid H_t) \\ &= \mathrm{E}(N^0(t) \mid N_1(t), N_2(t)) \\ &= N_1(t)/4 + N_2(t)/2;\end{aligned} \tag{18}$$

that is,

$$\frac{\hat{N}^0(t)}{\lambda_0} = t^{-1}\int_0^t N_1(s)\,ds + 2t^{-1}\int_0^t N_2(s)\,ds, \tag{19}$$

which converges to 2 as $t \to \infty$, as it should because the actual production interval for each unit that is good product is exactly 2. (In contrast, the average time spent in the system by *all* units is ³⁄₂.) For this system, the first station always has exactly 1 unit after the first arrival, i.e., $N_1(t) = 1$ for $t \geq 1$. The second station either has one unit or none; it has one whenever the previous processing step at station 1 was successful (was not scrapped). Therefore, by the law of large numbers, $t^{-1}\int_0^t N_2(s)\,ds \to ½$ as $t \to \infty$. Consequently, $\hat{N}^0(\lambda)/\lambda^0 \to W^0 = 2$ as $t \to \infty$, as claimed.

However, as mentioned in Section 2, it is important to be aware that we estimate the average production interval for the units that turn out to be good product, which is not the same as the average length of time from a product start until the next good product, started at this time or later, is produced. If we wanted to know the average wait until a good product is completed, then in this case, because units are processed one at a time in order of their arrival, it suffices to *add* the average wait until the next product start that will turn out to be a good product (here it is 3, using the geometric distribution) to the average production interval of a unit that turns out to be a good product (which is 2). In this case the average wait until the next good product is completed is 5, much larger than the average production interval for good product, which is 2.

Example 3. Consider the following modification of Example 2. Assume that precisely every other unit at each station is scrapped, starting with the first unit processed. The WIP at each station and the units eventually scrapped are depicted in Figure 3. Note that this is one possible realization of the scrapping sequences at the two stations in Example 2, but the modification causes Assumptions A′ and B′ to be violated. The full-information observable $\mathrm{E}(N^0(t) \mid H_t)$ is also depicted in Figure 3. In this case, we apply the estimator $\hat{N}^0(t)$ for L^0 in (14) to obtain an estimator for W^0, namely

$$\frac{\hat{N}^0(t)}{\lambda^0} = \frac{4}{t}\int_0^1 \mathrm{E}(N^0(s) \mid H_t)\,ds \tag{20}$$

which converges to $W^0 = 2$ as $t \to \infty$. In this example, however, it turns out that the simple estimator (6), α_i in (16) and $\lambda^0 = \lambda\alpha_1$ still gives the correct answer. In this case everything averages out correctly, so that Assumptions A′ and B′ were not critical. (A′ and B′ are sufficient for consistency, but are not necessary.) We use (6) because it satisfies (9). The next example shows that this is not always the case.

Example 4. Consider the following modification of Example 2. Suppose that service at station 2 is performed in batches of size 2; i.e., service at station 2 does not begin until there are at least two units ready

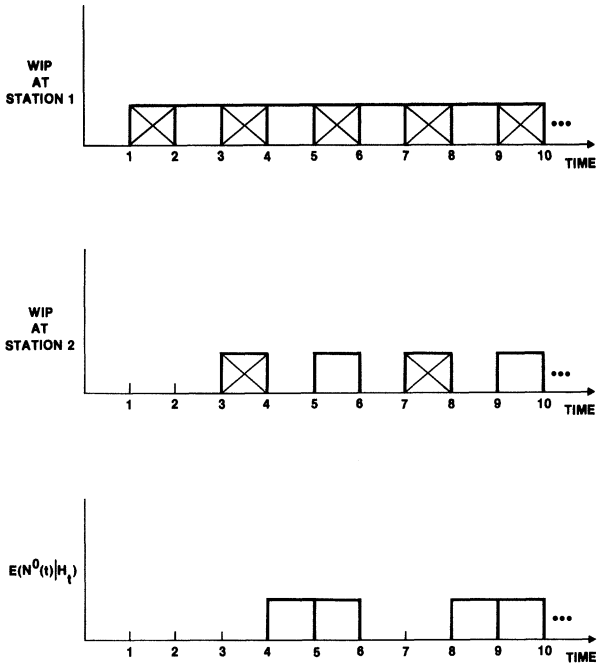

Figure 3. The WIP over time at each station and the conditional expected good product in the system for Example 3 (an × indicates that the job will eventually be scrapped).

to be processed, and then the two units are processed together in 1 time unit. A typical realization of the WIP at the two stations is shown in Figure 4. Now at station 2 the WIP at any time can be 0, 1 or 2. In fact, it is easy to see that the WIP at station 2 at successive integer time points is a discrete-time Markov chain on the state-space {0, 1, 2} with transition probabilities $P_{00} = P_{01} = P_{11} = P_{12} = P_{20} = P_{21} = ½$ and $P_{ij} = 0$, otherwise. This Markov chain is irreducible and aperiodic with stationary probability vector (¼, ½, ¼) on {0, 1, 2}. As in Example 2, each unit at each station is scrapped with probability ½, independently of all other events, Assumptions A′ and B′ still hold and we can estimate the average production intervals for good product W^0 by the simple indirect estimator $\hat{N}^0(t)/\lambda^0$ in (6) with α_i defined in (16), as given by (17) to (19), but here $t^{-1}\int_0^t N^2(s)\,ds$ converges to 1 instead of ½, so that $\hat{N}^0(t)/\lambda^0 \to W^0 = 3$ as $t \to \infty$.

Now make another modification. At station 2 assume that the last unit to arrive in each service batch is always scrapped, while the first unit in each service batch is always good product. This is depicted in Figure 4: At station 2 the units labeled 1 are good product and the units labeled 2 are scrapped. As before, exactly one half of the units processed at each station are scrapped, but as in Example 2 precisely every other unit at station 2 is scrapped.

With this second modification, Assumption A′ is satisfied, but Assumption B′ is not. In particular, the simple indirect estimator $\hat{N}^0(t)/\lambda^0$ in (6) is no longer an appropriate estimator. We apply (14) with

$$\begin{aligned} &\mathrm{E}(N^0(t) \mid H_t) \\ &\quad = 2^{-1}(1\{N_2(t) = 0 \text{ or } N_2(t) = 2\}) \\ &\qquad + \min\{N_2(t),\ 1\}. \end{aligned} \tag{21}$$

where $1A$ is the indicator function of the set A, i.e., $1A(x) = 1$ if $x \in A$ and 0, otherwise. The first term in (21) corresponds to the first station: Note that a good unit from station 1 will be the first unit in a service batch at station 2 one time unit later if and only if $N_2(t) = 0$ or $N_2(t) = 2$; an arbitrary unit at station 1 is good with probability ½.

This last modification illustrates that Assumption A′ might hold without Assumption B′, so we cannot apply (6). Moreover, unlike Example 3, $\hat{N}^0(t)/\lambda^0$ in (6) does *not* give the correct answer. From (21), we get $W^0 = 4$ as we should, whereas (6) and (16) yield 3, as in the first part of this example. We expect that the

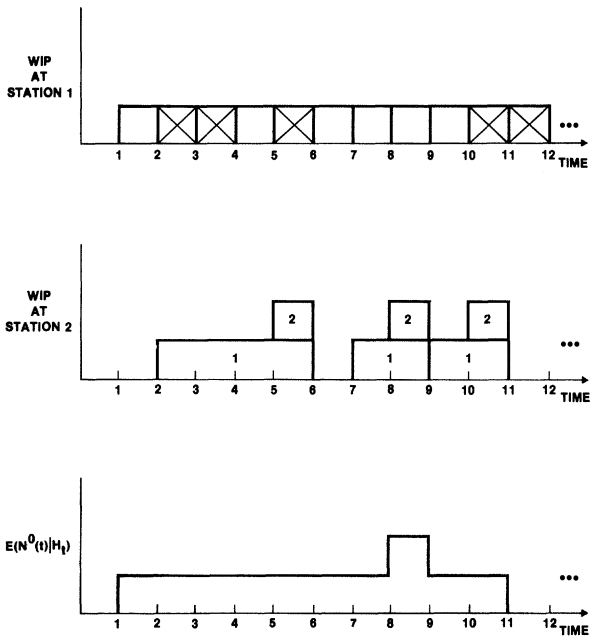

Figure 4. A possible realization of the WIP over time at each station and the conditional expected good product in the system for Example 4 (an × at station 1 and a 2 at Station 2 indicate that the job will eventually be scrapped).

simple indirect estimator $\hat{N}^0(t)/\lambda^0$ in (6) is often appropriate, but this example illustrates that we really need Assumption B′.

Example 5. Examples 1 to 4 had factors α_i defined by (16), so that $\alpha_i \leq 1$. We can also have $\alpha_i > 1$; α_i need not be a probability. In the setting of Example 1, suppose that the lot sizes are changed from station to station; that is, on average, 1 unit from station i is split into γ_i separate units for station $i + 1$ if $\gamma_i \geq 1$ or γ_i^{-1} units from station i are combined to form 1 unit or station $i + 1$ if $\gamma_i \leq 1$. After the final station n, γ_n indicates the number of units of good product produced from each completed unit at station n. We can still apply the direct estimator $\hat{N}^0(t)\lambda^0$ in (6) when we redefine α_i in (16) as

$$\alpha_i = (p_i\gamma_i)(p_{i+1}\gamma_{i+1}) \cdots (p_n\gamma_n). \tag{22}$$

As before $\lambda^0 = \lambda\alpha_1$. (The basic example in Section 2 also had $\alpha_i > 1$ for some i; see Table III.)

Example 6. This example illustrates an application outside of manufacturing, although it may be viewed as a special case of Example 5. Here we do not use the full-information observable estimator. Consider the following simple model of a packet-switched data communication network, depicted in Figure 5. Messages arrive at an initial access interface of the network at a rate λ. At the initial access interface, messages are packetized, i.e., they are divided into small packets to be transmitted over the network. The expected number of packets per message is γ. When messages arrive at the initial access interface, they wait in a message queue to be packetized; then the packets wait in an output packet queue to be transmitted. The packets are transmitted through the network to a terminal access interface, where they wait in an assembly queue until all other packets of the message arrive, so that the message can be reassembled. Afterward, the messages wait in the output message queue for transmission from the network to the ultimate destination. Inside the network, packets are occasionally retransmitted when packet errors are detected. At the access interfaces, messages are occasionally retransmitted when message errors are detected. Various congestion control schemes may also be employed to aid performance, e.g., windows and other flow control schemes. However, we assume that no messages are lost and that the system can eventually handle all messages, so that the departure rate of messages from the network is also λ.

For this system, suppose that we wish to estimate the average message transmission time, by which we

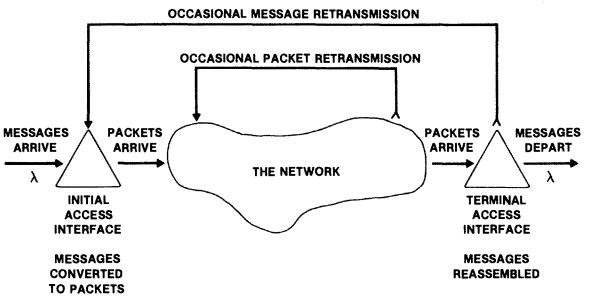

Figure 5. The packet-switched data communication network of Example 6.

mean the average length of time from when the entire message arrives at the initial access interface until the entire message departs from the terminal access interface. If we know how many messages are in the system at each time t, then we can apply Little's law directly, but suppose that we do not keep track of the messages after they have become packetized. (We simply count what we see at each time t.)

We describe how to use our indirect estimation procedure. Let $N^0(t)$ be the number of messages in the system at time t. Let $N_1(t)$ be the number of *messages* in the initial access interface message queue (or associated input buffers), counting each message as a full message as soon as the entire message arrives and until the message begins to be packetized. Let $N_2(t)$ be the number of *packets* in any message being packetized in the output packet queue of the initial access interface, in the network or in the assembly queue of the terminal access interface. Let $N_3(t)$ be the number of *messages* in the output message queue of the terminal access interface, counting the message as a full message until the complete message is transmitted out of the network. We assume that an assembled message goes immediately from the assembly queue to the output message queue of the terminal access interface (or is retransmitted if necessary) when the message is reassembled.

With these definitions, we apply the estimator (6) with $\alpha_1 = \alpha_3 = 1$ and $\alpha_2 = 1/\gamma$, i.e.,

$$\tilde{N}^0(t) = N_1(t) + \frac{N_2(t)}{\gamma} + N_3(t). \tag{23}$$

In this case, Assumptions A′ and B′ are typically not satisfied because $\mathrm{E}(N^0(t) \mid H_t) \neq \mathrm{E}(N^0(t), \mid [N_1(t), N_2(t), N_3(t)])$, where H_t is the full-information history, but nevertheless, it is easy to see that $\tilde{N}^0(t)$ in (23) satisfies (9).

5. One-Step Transitions

An important, special case in Section 4 arises in the context of an open network of queues with multiple job classes; e.g., Kelly (1979). Our general classes encompass this model because our class index can specify both the queue number and the job class. However, these network models are usually specified in terms of individual, one-step transitions, and they allow an unlimited number of transitions. For example, suppose that, on average, each class-i item in one transition produces β_i units of good product immediately and η_{ij} class-j items. Then the average number of good units eventually produced by each class-i item satisfies the system of equations

$$\alpha_i = \beta_i + \sum_{i=1}^{n} \eta_{ij}\alpha_j, \tag{24}$$

which can be viewed as a reverse-time variant of the usual system of traffic rate equations in queueing networks, generalized to allow η_{ij} not to be probabilities. Of course, we must *assume* that (24) possesses a proper solution for the α_i.

With a solution to (24), we apply the indirect estimator $\hat{N}^0(t)/\lambda^0$ in (6) to general network models. We still require Assumptions A′ and B′ after α_i is defined by (24). We also need the flow rate of good product, λ^0. We can give an expression for λ^0 if we know the external arrival rate of each class item. Let λ_i be the external arrival rate for class i. Then

$$\lambda^0 = \sum_{i=1}^{n} \lambda_i\alpha_i. \tag{25}$$

(Example 1 is the special case in which $\lambda^0 = \lambda\alpha_1$, $\beta_n = p_n$, $\eta_{ij} = p_i$, for $j = i + 1$ and $1 \leq i \leq n - 1$, and $\beta_i = \eta_{ij} = 0$, otherwise. Example 5 is the special case in which $\lambda^0 = \lambda\alpha_1$, $\beta_n = p_n\gamma_n$, $\eta_{ij} = p_i\gamma_i$ for $j = i + 1$ and $1 \leq i \leq n - 1$, and $\beta_i = \eta_{ij} = 0$, otherwise.)

Example 7. We consider a simple model of a production line with unlimited reworks. We also include reconstituted lots and partial yields as in the basic example. Let there be three stations in series; the second station is an inspection station. Units are processed at station 1 in lots of size 5. New lots of size 5 arrive at station 1 at rate λ. Lots completing processing at station 1 go immediately to station 2. Of the units (not lots) inspected for the first time at station 2, 5% must be scrapped immediately, while 10% require rework and are sent back to station 1. The rest proceed to station 3. At station 1, reconstituted lots of size 5 are formed from the units requiring rework. Of the units inspected at station 2 that previously required rework, 10% must be scrapped immediately, while 20% require rework and are sent back to station 1. (We assume that these percentages are approximately independent of the number of reworks.) Units passing inspection at station 2 proceed to station 3 for packaging. Partial lots at station 3 are reconstituted into packages of 5 units for shipping.

Suppose that we want to know the average production interval for a good unit. Let $N_1(t)$ be the number of units at either station 1 or 2 at time t that have *not* been inspected. Let $N_2(t)$ be the number of units at either station 1 or 2 at time t that previously required rework. Let $N_3(t)$ be the number of units at station 3 at time t. (We assume these are observable). As before, let $N^0(t)$ be the (unobservable) number of units in the system at time t that will eventually be good product.

We apply (24) and get a system of three equations

$$\begin{aligned} \alpha_1 &= 0.10\alpha_2 + 0.85\alpha_3 \\ \alpha_2 &= 0.20\alpha_2 + 0.70\alpha_3 \\ \alpha_3 &= 1 \end{aligned} \tag{26}$$

which has the solution

$$\alpha_1 = 0.9375, \quad \alpha_2 = 0.875 \quad \text{and} \quad \alpha_3 = 1, \tag{27}$$

so that the full-information observable estimator is $\hat{N}^0(t)/\lambda^0$, where $\lambda^0 = 5\lambda\alpha_1 = 4.69\lambda$ and

$$\begin{aligned} \tilde{N}^0(t) &= \mathrm{E}(N^0(t) \mid H_t) = \sum_{i=1}^{3} N_i(t)\alpha_i \\ &= 0.9375N_1(t) + 0.875N_2(t) + N_3(t). \end{aligned} \tag{28}$$

6. Conclusions

We suggested a way to estimate average production intervals indirectly via WIP measurements. On the positive side, Little's law provides alternative estimators, even if the WIP process is not directly observable. On the negative side, indirect estimation must be done carefully to obtain reliable estimates. In part, this is due to difficulties in defining what we mean by a production interval.

We illustrated how the indirect estimation procedure typically should be applied in practice with our basic example in Section 2. The proposed estimator is (6), which requires monitoring only the single process (5). In Sections 3 and 4 we provided conditions for the indirect estimation procedure to be valid, i.e., consistent (asymptotically correct). These conditions are primarily expressed via conditional expectations.

It remains to test the indirect estimation procedure. However, we are not certain what a good test should

be. For example, we could simulate the basic example. To do so, we would have to specify the model in much more detail. If we can construct whatever model we choose, consistent with the specification so far, then we can construct the model with natural Markov properties so that conditions A′ and B′ in Section 4 are satisfied; then we can theoretically deduce that the indirect estimator is consistent. We would only be observing what was mathematically proven. That leaves little to learn about consistency, although we would learn about statistical precision. Interpreting results about statistical precision is somewhat difficult; we need experience to put the results in perspective.

On the other hand, we could deliberately choose the model so that Assumptions A′ and B′ in Section 4 are violated. Constructing a meaningful experiment in this case is also difficult because it is not obvious how to calibrate how far the conditions are from being satisfied. The important issue seems to be whether the indirect estimation procedure works reasonably well in real systems, rather than whether or not the procedure works in a model when the conditions clearly do or do not. We think a good next step is to compare direct and indirect estimates of average production intervals using factory data (where time stamping is performed). We hope to report on such experiments in the future.

Although we did not test the indirect estimation procedure in a realistic factory setting, we have analyzed several small examples in Sections 4 and 5. These examples clearly demonstrate how the indirect estimation procedure helps and how it breaks down.

Acknowledgment

This work was begun while Ardavan Nozari was with the AT&T Bell Laboratories. The authors thank the Associate Editor and the referees for their helpful suggestions.

References

BRATLEY, P., B. L. FOX AND L. E. SCHRAGE. 1983. *A Guide to Simulation*. Springer-Verlag, New York.

BREIMAN, L. 1968. *Probability*. Addison-Wesley, Reading, Mass.

BURMAN, D. Y., F. J. GURROLA-GAL, A. NOZARI, S. SATHAYE AND J. P. SITARIK. 1986. Performance Analysis Techniques for IC Manufacturing Lines. *AT&T Tech. J.* **65,** 46–57.

CARSON, J. S., AND A. M. LAW. 1980. Conservation Equations and Variance Reduction in Queueing Simulations. *Opns. Res.* **28,** 535–546.

CHUNG, K. L.1974. *A Course in Probability Theory*. Academic Press, New York.

DUNIETZ, I. S., J. L. C. HSU, M. T. MCEACHERN, J. H. STOCKING, M. A. SWARTZ AND R. M. TROMBLY. 1986. MPCS—the manufacturing process control system. *AT&T Tech. J.* **65,** 35–45.

GLYNN, P. W., AND W. WHITT. 1986a. A Central-Limit-Theorem Version of $L = \lambda W$. *Queueing Syst. Theory Appl.* **1,** 191–215.

GLYNN, P. W., AND W. WHITT. 1986b. Indirect Estimation Via $L = \lambda W$. *Opns. Res.* (to appear).

GLYNN, P. W., AND W. WHITT. 1986c. Extensions of the Queueing Relations $L = \lambda W$ and $H = \lambda G$. *Opns. Res.* (to appear).

GLYNN, P. W., AND W. WHITT. 1986d. The Efficiency of the Simulation Estimators (submitted for publication).

HALFIN, S. 1982. Linear Estimators for a Class of Stationary Queueing Processes. *Opns. Res.* **30,** 515–529.

HEYMAN, D. P., AND M. J. SOBEL. 1982. *Stochastic Models in Operations Research*, Vol. I. McGraw-Hill, New York.

KELLY, F. P. 1979. *Reversibility and Stochastic Networks*. John Wiley & Sons, New York.

LAW, A. M. 1975. Efficient Estimators for Simulated Queueing Systems. *Mgmt. Sci.* **22,** 30–41.

LAW, A. M. 1983. Statistical Analysis of Simulation Output Data. *Opns. Res.* **31,** 983–1029.

LITTLE, J. D. C. 1961. A Proof of the Queueing Formula: $L = \lambda W$. *Opns. Res.* **9,** 383–387.

PRITSKER, A. A. B., AND C. D. PEGDEN. 1979. *Introduction to Simulation and SLAM*. Halsted Press, New York.

STIDHAM, S., JR. 1974. A Last Word on $L = \lambda W$. *Opns. Res.* **22,** 417–421.

WHITT, W. 1983. The Queueing Network Analyzer. *Bell System Tech. J.* **62,** 2779–2815.

WHITT, W. 1987. A Queueing Network Analyzer for Manufacturing (unpublished paper).