Chapter 5
Little's Law

John D.C. Little and Stephen C. Graves
Massachusetts Institute of Technology

The average waiting time and the average number of items waiting for a service in a service system are important measurements for a manager. Little's Law relates these two metrics via the average rate of arrivals to the system. This fundamental law has found numerous uses in operations management and managerial decision making.

Introduction

Caroline is a wine buff and bon vivant. She likes to stop at her local wine store, *Transcendental Tastings*, on the way home from work. She browses the aisles looking for the latest releases from her favorite vineyards. Occasionally she picks up a few bottles. She stores these in a rack in a cool corner of her cellar. She and her partner eat out frequently but when they are at home they usually split a bottle of wine at dinner. Sometimes they have friends over and that puts a bigger dent in the wine inventory.

They have been doing this for some time. Her wine rack holds 240 bottles. She notices that she seldom fills the rack to the top but sometimes after a good party the rack is empty. On average it seems to be about 2/3rds full, which would equate to 160 bottles.

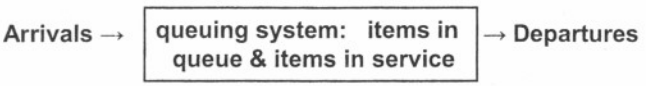
Many wines improve with age. After reading an article about this, Caroline starts to wonder how long, on average, she has been keeping her wines. She went back through a few months of wine invoices from *Transcendental* and estimates that she has bought, on average, about eight bottles per month. But she certainly doesn't know when she drank which bottle and so there seems to be no way she can find out, even approximately, the average age of the bottles she has been drinking.

This is a good task for Little's Law.

D. Chhajed and T.J. Lowe (eds.) *Building Intuition: Insights From Basic Operations Management Models and Principles.*
doi:10.1007/978-0-387-73699-0,

Little's Law Deals with Queuing Systems

A "queuing system" consists of discrete objects we shall call "items" that "arrive" at some rate to the "system." Within the system the items may form one or more queues and eventually receive "service" and exit. Figure 5.1 shows this schematically.



Flow of items through a queuing system

Fig. 5.1 Schematic view of a queuing system

While items are in the system, they may be in queues or may be in service or some in queue and some in service. The interpretation will depend on the application and the goals of the modeler. For example in the case of the wine cellar, we say that a bottle (an "item") arrives to the system when it is first placed into the wine cellar. Each bottle remains in the system until Caroline selects it and removes it from the cellar for consumption. If we view the wine rack as a single channel server, the service time is the time between successive removals. It is interesting to note, however, that we do not know which bottle Caroline will pick and there is no particular reason to believe that she will pick according to a first-in, first-out (FIFO) rule. In any case, to deal with the average number of bottles in the cellar or average time spent by a bottle in the cellar, we need to consider the complete system consisting of queue plus service.

Little's Law says that, under steady state conditions, the average number of items in a queuing system equals the average rate at which items arrive multiplied by the average time that an item spends in the system. Letting

L = average number of items in the queuing system,
W = average waiting time in the system for an item, and
λ = average number of items arriving per unit time, the law is

$$L = \lambda W. \tag{1}$$

This relationship is remarkably simple and general. We require stationarity assumptions about the underlying stochastic processes, but it is quite surprising what we do *not* require. We have not mentioned how many servers there are, whether each server has its own queue or a single queue feeds all servers, what the service time distributions are, or what the distribution of inter-arrival times is, or what is the order of service of items, etc.

In good part because of its simplicity and generality, the equation (1) is extremely useful. It is especially handy for "back of the envelope" calculations. The reason is that two of the terms in (1) may be easy to estimate and not the third. Then Little's Law quickly provides the missing value.

Thus for Caroline, the average number of bottles in the system is $L = (240)*(2/3) = 160$ bottles and the average arrival rate is $\lambda = (12)*(8) = 96$ bottles/year. Without ever collecting individual data on how long each bottle remains in her cellar, she can calculate the average amount of time a bottle stays in her cellar as $W = (160)/(96) \cong 1.67$ years. That's not very old. She needs a bigger rack and more patience, or, alternatively, she should develop selection rules to favor holding special bottles longer than the others. This wouldn't affect the average but might give her some fine old wines.

Arguing Little's Law with a Picture

Figure 5.2 shows one possible realization of a particular queuing system. We can make a heuristic argument for Little's Law by interpreting the area under the curve in Fig. 5.2 in two different ways. Let

$n(t)$ = the number of items in the queuing system at time t;
T = a long period of time;
$A(T)$ = the area under the curve $n(t)$ over the time period T;
$N(T)$ = the number of arrivals in the time period T.

On the one hand, an item in the queuing system is simply there. The number of items can be counted at any instant of time t to give $n(t)$. Its average value over T is the integral of $n(t)$ over T (i.e., $A(T)$) divided by T. On the other hand, at time t each of the items is waiting and so is accumulating waiting time. By integrating $n(t)$ over the time period T, we obtain a cumulative measure of the waiting time, again equal to $A(T)$. Furthermore, the arrivals are countable too, and given by $N(T)$. Therefore, inspecting the figure, we define

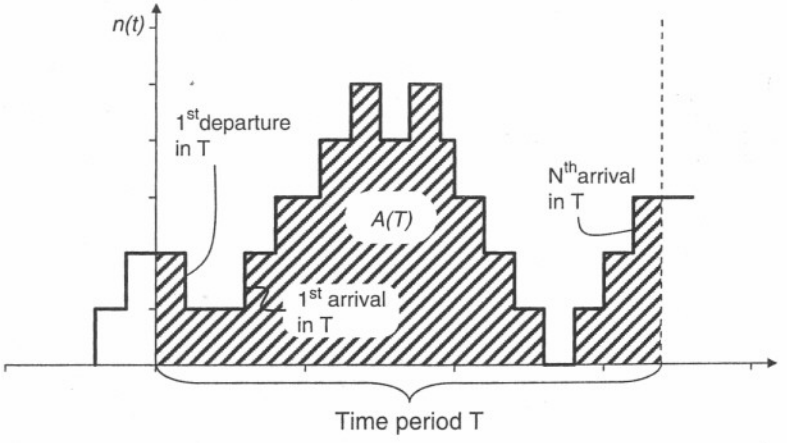


Fig. 5.2 Number of items in a queuing system versus time

$\lambda(T) = N(T)/T$ = arrival rate during time period T,
$L(T) = A(T)/T$ = average queue length during time period T,
$W(T) = A(T)/N(T)$ = average waiting time in the system per arrival during T.
A slight manipulation gives $L(T) = \lambda(T)W(T)$.

All of these quantities wiggle around a little as T increases because of the stochastic nature of the queuing process and because of end effects. End effects refer to the inclusion in $W(T)$ of some waiting by items which joined the system prior to the start of T and the exclusion of some waiting by items who arrived during T but have not left yet. As T increases, $L(T)$ and $\lambda(T)$ go up and down somewhat as items arrive and later leave.

Under appropriate mathematical assumptions about the stationarity of the underlying stochastic processes, the end effects at the start and finish of T become negligible compared to the main area under the curve. Thus, as T increases, these stochastic "wiggles" in $L(T)$, $\lambda(T)$, and $W(T)$ become smaller and smaller percentages of their eventual values so that $L(T)$, $\lambda(T)$, and $W(T)$ each go to a limit as we increase T to infinity. Then, using the obvious symbols for the limits, we have:

$$\lim_{T\to\infty} L(T) = L; \quad \lim_{T\to\infty} \lambda(T) = \lambda; \quad \lim_{T\to\infty} W(T)$$

from which we get the desired result (1).

It is interesting and important to note that the formula holds for *each* realization of the queuing system over time. This was argued by Little, in his original paper (Little 1961), noting that the relationship (1) held for each evolution of the time series of a particular queuing system. In other words, if we watch a specific case or realization designated, say, by ω, as it develops over time, then we will find that

Building Intuition Little's Law provides a fundamental relationship between three key parameters in a queuing (or waiting line/service) system: the average number of items in the system, the average waiting time (or flow time) in the system for an item, and the average arrival rate of items to the system. The system can be very general. For example, it might include both the service facility and the waiting line, or it might be only the waiting line. An important feature of Little's Law is that by knowing, perhaps via direct measurement, two of the three parameters, the third can be calculated. This is an extremely useful property since measurement of all three parameters may be difficult in certain applications.

Little's Law is applicable in many environments including manufacturing and service industries as well as everyday decision making by individuals.

$L(\omega) = \lambda(\omega)\, W(\omega)$, given the steady state and other assumptions made. Averaging cross-sectionally across the many possible realizations of a particular system gives (1), but it is a useful insight to know that the formula holds for each evolving time series as it is observed over a long time period.

Law or Tautology?

Equation (1) is commonly called Little's Law and we have cheerfully adopted that terminology. However, as pointed out by various people, including Little (1992), Eq. (1) is a mathematical theorem and therefore a tautology. The relationship turns out to be useful in practice, but there is no need to go out on the factory floor and collect data to test it. This would be required in the case of a physical law such as Newton's Law of Gravitation. Each side of Newton's equation has to be measured and it is an empirical question whether they are equal within the measurement error. For a mathematical theorem, if the assumptions are satisfied by the application, the result will hold. Note that calling a mathematical theorem a law is not without precedent. The Law of Large Numbers would be another instance.

Usefulness of Little's Law in Practice

In this section we try to convey the generality of the result and its usefulness in different contexts by means of simple examples. In each case we see how the observation of two of the three measures provides the third. We try to bring out why such back-of-the-envelope analyses are of interest and value in different situations.

Semiconductor Factory: Semiconductor devices are manufactured in extremely capital-intensive fabrication facilities. The manufacturing process entails starting with a silicon wafer and then building the electronic circuitry for multiple identical devices through hundreds of process steps. Suppose that the semiconductor factory starts 1,000 wafers per day, on average; this is the input rate. The start rate has remained fairly stable over the past 9 months. We track the amount of work-in-process (*WIP*) inventory. The *WIP* varies between 40,000 and 50,000 wafers; the average *WIP* is 45,000 wafers.

Then we can infer the average flow time in the factory. The arrival rate to the factory is the wafer start rate: λ = 1,000 wafers per day. The *WIP* is the system queue length: L = 45,000 wafers. Thus the wait time or expected time in the system is W = 45 days. In a manufacturing context, we often refer to this as the flow time, the time between when a job starts and finishes in a factory. For instance, if we think of one wafer as being a job, then it takes the factory on average 45 days to process it, that is, to convert it from a blank wafer into a finished wafer comprised

of electronic devices. Knowing the flow time is critical for planning and scheduling the factory, and for making delivery commitments to customers. We shall return later to the connection between Little's Law and operations management.

E-Mail: Managing our e-mail is a common and time-consuming daily activity. For many it is hard to keep up with the volume of messages, let alone provide timely responses. A student Sue might receive 50 messages each day to which she must generate a response. Can we easily assess how well this student handles her e-mail duties?

Indeed we can apply Little's Law to get a quick sense of how promptly Sue responds to messages. Suppose that she receives about 50 messages every day; then this is the arrival rate: $\lambda = 50$ messages/day. Suppose we can also track how many messages have yet to be answered. For instance, suppose that Sue removes a message from her InBox once she has responded to it. Then the remaining messages in her InBox are the messages that are waiting to be answered. Over the last semester, the size of the InBox has varied between one and two hundred messages with an average of 150 messages. Then we can regard this to be the system queue length: $L = 150$ messages. From Little's Law we immediately have an estimate of how long it takes Sue to answer a message, on average: $W = 3$ days.

Hospital Ward: We wish to determine the size and staffing levels for the maternity ward for a local hospital. From historical records we know that the birth rate for the local community is about five births per day. We also know that most women stay in the maternity ward for 2 days before going home with child; however occasionally, there are complications with the birth that require much longer stays. Over the past 6 months, we find that 90% of the births have resulted in 2-day stays; for the remaining 10% of the cases, the average length of time in the maternity ward is 7 days. Thus, on average, the length of stay is $0.9 \times 2 + 0.1 \times 7 = 2.5$ days.

We can use Little's Law to predict the average number of mothers in the maternity ward. The arrival process corresponds to the women arriving to deliver their babies; the arrival rate is $\lambda = 5$ mothers per day. The relevant waiting time in the system is the length of stay in the maternity ward: $W = 2.5$ days. Thus, the expected queue length or number in the system is $L = 12.5$ mothers. This would be useful in determining the size of the maternity ward (e.g., beds) and the staffing requirements. However, the law only provides the average requirements, and one would need to design the maternity ward to accommodate its peak requirements. For instance, we would certainly want more than 12.5 or 13 beds in order to handle the variability in the occupancy of the ward. One needs to use queuing models and/or simulation to explore the trade-offs between the utilization of the beds and the likelihood of not having a bed for an expectant mother. Nevertheless, Little's Law provides a starting point for this investigation, since we know the average number of beds that are needed.

Toll Booths: The Ted Williams Tunnel travels under the Boston harbor, connecting East Boston to South Boston. During the course of a day, about 50,000 vehicles go through the tolls at the entry point to the Tunnel in East Boston. The Massachusetts

Transit Authority (MTA) tries to modulate the number of toll booths that are open at any point in time so that the average number of vehicles waiting at the tolls (including those at the booths) never exceeds 20 vehicles. For instance, all six booths are open during the peak time in the morning from 6:00 AM to 10:00 AM. During this morning rush hour, the tunnel handles up to 4,000 cars per hour, and the MTA estimates that the average number of vehicles waiting at any point of time is near the target maximum of 20 vehicles.

With the assumption that the arrivals occur at a relatively stable rate over the morning rush hour, we can then use Little's Law to ask what quality of service is being delivered in terms of average waiting time per vehicle. Suppose that the arrival rate to the toll booths is $\lambda = 3{,}600$ vehicles per hour (or 1 vehicle per second), and the expected number of vehicles in the system is $L = 20$ vehicles. Thus, on average, the time a vehicle spends at the toll booths is $W = 20/3{,}600$ h = 20 s.

Housing Market: The local real estate agent in your community estimates that it takes 120 days on average to sell a house; whereas this number changes some with the economy and season, it has been fairly stable over the past decade. You observe from monitoring the classified ads that over the past year the number of houses for sale has ranged from 20 to 30 at any point in time, with an average of 25. What can we say about the number of transactions in the past year?

From Little's Law we can estimate this by viewing the real estate market as a queuing system. We regard a house being put up for sale as an arrival to the system. We assume that an unsold house remains on the market until it is sold. Thus, when a house "completes its service" and departs from the market, we infer that it has been sold. We have estimates of the average time in the system and the average number in the system, namely, $W = 120$ days and $L = 25$ houses. From this, we can estimate the arrival rate to the system, $\lambda = 25/120$ houses per day $\cong$ 75 houses per year.

Doughnut Shop: From your daily morning trip to the doughnut shop, you know they have a healthy business, at least financially speaking. As you might want to invest in a franchise, you wonder what amount of revenue they generate. Over the course of several months, you visit the shop at random times between 6:00 AM and 9:00 AM; you observe that the queue averages about 10 customers, and that it takes you about 3 min to get in and out of the shop.

If you assume your experience is typical, then you can apply Little's Law to estimate what the throughput rate is for the enterprise for the morning peak period. The expected number in the system is $L = 10$ customers and the expected time in the system is $W = 3$ min. We can then estimate the arrival rate to the system, namely $\lambda = 10/3$ customers per minute = 200 customers per hour. We also term this the throughput rate as arriving customers become throughput once served. To get an estimate for the revenue potential from this shop, we need to estimate how much each customer spends. If you typically spend \$5 per visit, then with the assumption that you are a typical customer, we have a rough estimate of the shop's revenue during these morning hours, i.e., \$1,000 per hour.

The Robustness and Generality of Little's Law in Certain Systems

So far we have developed and discussed Little's Law as a relationship among steady-state stochastic processes. The contexts we have examined have been well-behaved, stable, and on-going. In particular we assume that the characteristics of the arrival and service processes are stationary over time. For example, in the case of the maternity ward, we assume that the average arrival rate of mothers has been steady at five per day for some time, and that this rate does not vary with day of week or season of the year. Similarly, we have regarded the service process as being stationary; for instance, we read and process our e-mail at roughly the same average rate, day in and day out, independent of the backlog of unread messages. For some of our examples, we have focused on an interval of time, e.g., the morning rush hour through the toll booths. However, in these instances due to the huge volume of arrivals, we contend that the system behavior is virtually equivalent to that of a steady-state system.

The purpose of this section is to show the great robustness and generality of Little's Law under certain circumstances. Indeed Little's Law is exact in these cases even though arrival and service process may be nonstationary. The essential condition is to have a finite window of observation that starts and stops when the system is empty. We use an example to motivate and illustrate the validity of Little's Law in this situation. Consider the *Sweet & Sour* supermarket, which opens every day at 7:00 AM and closes 16 h later at 11:00 PM. When *S&S* opens at 7 AM, there are no customers in the store. When it closes at 11 PM, all of the customers depart. Between opening and closing, customers arrive to the store, do their shopping and leave. The arrivals over the course of the day are quite varied. They include several customer segments, each with quite distinct shopping habits. Families with school-age children will shop between 9AM and 2 PM, and tend to have fairly lengthy shopping forays as they stock up for a week at a time. Seniors will tend to shop at quiet times of the day, like first thing in the morning, and will also be fairly leisurely in their shopping, taking up to an hour to complete a visit. Working couples will shop at night after work or on the weekends; their evening visits are often to run in, grab something and run out.

We propose to model *S&S* as a queuing system with the arrivals being the customers as they enter the store and service being the duration of their time in the store selecting and purchasing their groceries. However, from the above discussion, we see that this is anything but a stable system. The supermarket is never in a steady-state. It starts and ends each day with zero customers. Over the course of the day, customers arrive at varying rates, and the nature of their shopping trips also varies over the day, due to the different clienteles. Nevertheless, we will show next that Little's Law applies each and every day to this supermarket in an exact way.

An Analytic Interlude

Let N denote the number of customers that shop on a particular day. Suppose that we keep track of when customers arrive and when they depart from the store. Then we can define and create two processes, one for the arrivals and the other for the departures. We define time $t = 0$ to correspond to the opening at 7 AM, and time $t = 16$ to be the store closing at 11 PM, 16 h later.

We let $N(t)$ denote the cumulative number of arrivals to the store by time t. Thus, as we start the day with zero customers, we have $N(0) = 0$; as we assume a total of N customers arrive during the day, we have $N(16) = N$. The cumulative arrivals increase in a stair-case fashion, as shown in Fig. 5.3, over $0 < t < 16$.

In similar manner we define $D(t)$ to denote the cumulative number of departures from the store by time t. Again, we have $D(0) = 0$, $D(16) = N$, and the cumulative departures increase in a stair-case fashion over the time interval $0 < t < 16$; see Fig. 5.3.

We note that at all time instants we have $N(t) \geq D(t)$, as the number of departures can never exceed the number of arrivals. Indeed, the difference between the two cumulative processes is the number of customers in the supermarket at time t:

$$L(t) = N(t) - D(t).$$

With this observation we can determine the average number in the supermarket over the course of the day from the following integral:

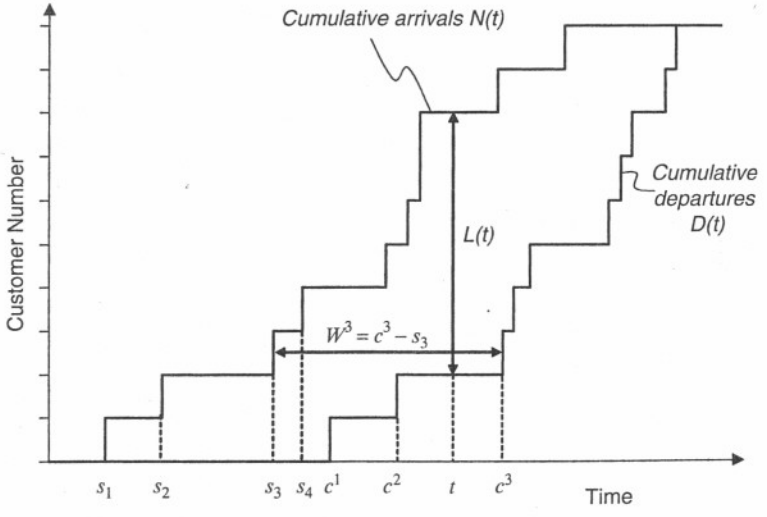


Fig. 5.3 Cumulative arrivals to and departures from a system, for example, the supermarket

$$L = \frac{1}{16} \times \int_{t=0}^{t=16} (N(t) - D(t))dt. \tag{2}$$

To model the average time in the supermarket for each customer is a bit more involved. We define $\{s_1, s_2, \ldots s_N\}$ to be the sequence of arrival or start times for the N customers, where s_j denotes the start time for the jth arriving customer. We define $\{c_1, c_2, \ldots c_N\}$ to be the sequence of departure or completion times for the N customers, where c_j denotes the completion time for the jth arriving customer. Thus, the time in the supermarket for the jth arriving customer is $W_j = c_j - s_j$. Averaging this over all the customers gives:

$$W = \frac{1}{N} \times \left(\sum_{j=1}^{N} c_j - \sum_{j=1}^{N} s_j \right). \tag{3}$$

To compute (3) we shall develop an equivalent expression based on the geometry in Fig. 5.3. Let us define $\{c^1, c^2, \ldots c^N\}$ to be the sequence of departure or completion times for the N customers, where c^j denotes the completion time for the jth departing customer. Since customers need not exit the store in the order that they arrive, we shall often have $c_j \neq c^j$. However, the sequence $\{c^1, c^2, \ldots, c^N\}$ is just a permutation or reordering of the sequence $\{c_1, c_2, \ldots c_N\}$ as the departure time for each customer must appear exactly once in each sequence.

As shown in Fig. 5.3, we can define the jth wait time as $W^j = c^j - s_j$, equal to the difference between the departure time for the jth departing customer and the start time for the jth arriving customer. Now let us consider the average of these wait times:

$$\frac{1}{N} \times \sum_{j=1}^{N} W^j = \frac{1}{N} \times \left(\sum_{j=1}^{N} c^j - \sum_{j=1}^{N} s_j \right).$$

But this will equal W given by (3), since $\sum_{j=1}^{N} c_j = \sum_{j=1}^{N} c^j$. Hence we conclude that

$$W = \frac{1}{N} \times \sum_{j=1}^{N} W^j = \frac{1}{N} \times \left(\sum_{j=1}^{N} c^j - \sum_{j=1}^{N} s_j \right). \tag{4}$$

Now we need to relate our expression for L, given by (2) to our expression for W, given by (4). From the geometry in Fig. 5.3, we observe the following equivalence:

$$\int_{t=0}^{t=16} (N(t) - D(t))dt = \sum_{j=1}^{N} W^j. \tag{5}$$

That is, on each side of the equation, we have an expression for the area between the cumulative arrivals and the cumulative departures. On the left side we compute the area by integration over time of the function that tracks the number in the system; on the right side we compute the area by summing up the time in system for N customers.

From (2), (4) and (5), we can now write Little's Law for the supermarket:

$$L = \frac{1}{16} \times \int_{t=0}^{t=16} \left(N(t) - D(t)\right) dt = \frac{1}{16} \times \sum_{j=1}^{N} W^j = \frac{N}{16} \times W. \tag{6}$$

We recognize $\frac{N}{16}$ to be the arrival rate in customers per hour for the particular day, and we define $\lambda = \frac{N}{16}$; thus we have (6) in its familiar form, $L = \lambda W$.

With this simple example we have shown that Little's Law can be true over a finite time window (16 h) with nonstationary arrivals and with no notion of any steady state for the system in question. On reflection, there were two essential conditions for this result:

- Boundary conditions—we specify the finite time window to start and end with an empty system. This was a natural condition for the supermarket, and indeed, would be common for many service systems.
- Conservation of customers—we assume that all arriving customers will eventually complete service and exit from the system; there are no lost customers, so that the number of arrivals equals the number of departures. Again, this is a valid assumption for many systems of interest.

We really needed nothing else beyond these conditions in order to establish the law in our case. We have no assumptions about N, the number of customers; indeed, all the equations hold true for any N, e.g., for $N = 1$. We have no assumptions about the process for arrivals, or about how customers are serviced within the store. There might be a long period of no arrivals followed by the arrival of several busloads of customers; there might be periods of no service completions, say, if all the cash registers stopped working for an hour. The only conditions are as stated above: we need to start and end with an empty system and we need to conserve customers.

Notice that our formula is exact, but after the fact. In other words, we cannot complete our calculation until the supermarket door shuts. This is not a complaint. It merely says that we are observing and measuring not forecasting. Another point to mention is that the numbers will be different each day because of different sets of shoppers on different days of the week, the weather, holidays, and other changes in the store's internal and external environment. Nevertheless, the relationship $L = \lambda W$, as measured for that day, will be exact and the ability to measure two of the parameters and deduce the third still holds.

Further Discussion of "Average"

Little's Law holds exactly, but let us examine further what we mean by "average" wait, queue, and arrival rate. We have no probability distributions and so these are not expected values. Looking at the derivation of the result, we see that we are talking about everyday sample averages in the case of waits and arrival rates and finite time averages in the case of queues. So Little's Law here shows us an exact relation among *sample and time averages*. Next, consider a customer segment at the *Sweet & Sour* supermarket that consists of men with children in strollers. We can compute sample or time averages for their arrival rate, time in store, and number in store. Little's Law will hold exactly. Therefore, Little's Law is true for these averages for any identifiable segment. To use the relationship in practice, it will be necessary to collect data observing how many people of the target segment enter the store during the day.

To summarize, Little's Law is robust and remarkably general for queuing systems for which a finite window of observation starts and stops when the system is empty. Interpretation of the area between cumulative arrivals and cumulative departures permits an analytic argument that Little's Law is exact despite possibly nonstationary arrival and service processes. What we have discussed here turns out to be the tip of a fascinating mathematical iceberg that has been developed in recent years, called sample path analysis of queuing systems. An adequate discussion of it is beyond the scope of this chapter but the interested reader is referred to the book of El-Taha and Stidham (1999).

Evolution of Little's Law in Operations Management

Over the past 15 years or so, Little's Law has played an increasingly important role in the teaching and practice of operations management. However, the law is usually stated in a modified format to emphasize its applicability to operations. For instance, we cite as an example the very successful textbook of Hopp and Spearman (2000) who refer to Little's Law as a "...an interesting and fundamental, relationship between *WIP*, cycle time and throughput." They go on to state the law as

$$TH = \frac{WIP}{CT}, \tag{7}$$

where they define throughput (*TH*) as "the average output of a production process (machine, workstation, line, plant) per unit time," work in process (*WIP*) as "the inventory between the start and end points of a product routing," and cycle time (*CT*) as "the average time from release of a job at the beginning of the routing until it reaches an inventory point at the end of the routing (that is, the time the part spends as *WIP*)." They note that cycle time is also referred to as flow time, throughput time, and sojourn time, depending on the context.

We easily see that (7) is equivalent to Little's Law with $TH = \lambda$, $WIP = L$ and $CT = W$. However, there is a more fundamental difference in that the law is stated in terms of the average output or departure rate for the system, rather than the arrival rate. This reflects the perspective of a typical operating system, especially a manufacturing operation. Output is a primary attribute of any manufacturing system, since it is nominally its *raison d'etre*. As stated, we see that any increase in output requires either an increase in work-in-process inventory or a reduction in cycle time or both.

Furthermore in many contexts, the output rate is determined exogenously and is given to the manufacturing system; it reflects actual sales and/or a forecast of sales. The manufacturing system must then manage its operations to achieve this output rate. It will need to determine how to release work to the operation so as to meet the output target. In effect, the arrival process is endogenous. The operations manager decides the arrivals to the system based on the desired outputs. There is extensive research in the operations literature on how best to set the work release (or arrival process) to achieve the output targets. The best policies are dynamic policies that depend on the state of the manufacturing shop, e.g., depend on the work-in-process.

Our original development of Little's Law assumes a stable system with a stationary arrival process; as discussed above, we cannot assume a stationary arrival process for the typical context in which we might apply (7). Thus, we ask what conditions are necessary for (7) to be valid. At a minimum we must have conservation of flow. Thus, the average output or departure rate (TH) equals the average input or arrival rate (λ). Furthermore, we need to assume that all jobs that enter the shop will eventually be completed and will exit the shop; there are no jobs that get lost or never depart from the shop. In addition, we need some notion of system stability. We consider two possibilities, as this issue raises another important consideration.

First, we might assume that the shop will occasionally empty, i.e., $WIP = 0$. Then, as with the supermarket example, we will see that Little's Law holds exactly between any two time instances at which the shop is empty.

However, in many manufacturing systems, the WIP never drops to zero. In some contexts, this occurs for behavioral reasons; as the WIP decreases, the shop naturally slows down so as to not run out of work. The shop adjusts its service rate dynamically so as to keep from driving the WIP to zero. In other contexts, there might be an explicit control rule that maintains some target level or range of WIP. For instance, a very effective control policy is the so-called CONWIP policy that maintains an absolutely constant level of WIP (Hopp and Spearman, 2000); that is, the control rule releases one unit of new work to the system whenever one unit of work completes processing and exits the system. In either case, Little's Law applies, at least as an approximation, as long as we select a time interval that is long enough for two conditions to hold.

First, we need the size of the WIP to be roughly the same at the beginning and end of the time interval so that there is neither significant growth nor decline in the size of the WIP.

Second, we need some assurance that the average age or latency of the *WIP* is neither growing nor declining. We have previously assumed that all jobs that enter the shop will eventually be completed and will exit the shop. But if the *WIP* never drops to zero, it is possible for the jobs to be getting older or younger, in which case the law does not hold.

To illustrate the problem, consider a doll store that offers a line of international folk dolls. Each doll is adorned with traditional folk clothes from a particular country. The artist who produces the dolls gives each doll a distinctive hat in one of 10 different colors, which span the rainbow; the choice of color is completely at random. So, sometimes the Irish folk doll has a green hat, and sometimes a red hat. The store stocks 100 of these dolls so as to have a rich assortment of dolls from which to choose; whenever it sells a doll, the store immediately obtains a replacement from the supplier so as to maintain its in-store stock at 100 dolls. The supplier chooses the replacement doll at random from its supply.

Demand for the dolls has been quite good, as customers appreciate the artistry and novelty of the dolls. Indeed, the dolls sell consistently at a rate of about two per week. However, customers have a subtle but strong subconscious dislike for dolls with mauve hats. Dolls with mauve hats seldom sell; indeed, these mauve-hat dolls sell at a rate of about one every 2 years.

A naïve application of Little's Law might assume a doll arrival rate, equal to the demand rate, $\lambda = 2$ per week and an average number in system $L = 100$ dolls, and then conclude that the average time in the store for a doll is $W = 50$ weeks. However, this is not likely to be an accurate estimate over any moderate time interval, like a few years. Suppose we start with none of the hundred dolls having mauve hats. Every time we sell a doll there is a 1 in 10 chance that it will be replaced with a doll with a mauve hat, as the artist has 10 colors from which to choose. Since these mauve-hat dolls sell much, much less frequently than any other doll, the mix of dolls will gradually change over time. Indeed, the number of mauve-hat dolls in the store grows by about 10 dolls per year. Dolls without mauve hats make up a smaller percentage, but continue to sell at a rate of two per week; the time in system for these dolls actually shrinks as they make up a smaller percentage of the store assortment. However, the dolls with mauve hats do not sell and just accumulate more and more waiting time. Hence, the average age of the dolls in the store's assortment continues to grow older until at some point, the entire assortment has mauve hats. As a consequence, we cannot apply Little's Law during this transient period.

For instance, suppose we observe the system for 5 years, and then use Little's Law to estimate $W = L/\lambda = 100/2 = 50$ weeks; this estimate overstates the actual time in system for those dolls that have sold. Over the first 5 years, the number of mauve-hat dolls in the store grows from zero to about 50. As a consequence the active inventory, namely the dolls without mauve hats, drops from 100 to about 50. These dolls stay in the system, on average, less than 50 weeks, whereas the mauve-hat dolls just sit and get older. Of course, if we extend the time interval, in 10 years the entire assortment becomes mauve and the demand rate falls to one doll every 2 years; eventually (about 200 years) the mauve inventory turns over and Little's Law

will now apply. Presumably, the store would recognize the trend a bit earlier and do something about it! Nevertheless, the example shows how Little's Law might not hold over some time interval during which the average age of the *WIP* or queue is changing.

A quite different approach to the problem of nonzero *WIP* is motivated by what we learned in the supermarket problem. There we noted that Little's Law applies independently to each customer segment, but we have to be able to identify the customers in each segment and collect data on them.

So, in the case of non-zero *WIP* (or, actually, any existing *WIP*!), we can ignore all of it and focus on a group of new items, which, for example, might be colored blue. The system starts empty of blue items, even though it may be cluttered with others. We count blues as they enter the system (the rate may be controlled if we wish—stationarity is not required). The observations we make depend on what we want to learn and what is easy to measure. Suppose we want to observe and process N blue items. And suppose we want to know the cycle time $CT = W$, the average time a blue item spends in processing. At regular intervals we take an inventory of blues so that we can estimate $WIP = L$ (for blues only) by simple averaging. Eventually all N leave, say at T and so the average arrival rate is $\lambda = TH = N/T$. Then $CT = W$ can be calculated by Little's Law. The underlying theory is exact, although we have introduced some sampling error in estimating the blue *WIP*. However, this is something we can control by putting whatever resources we think are worthwhile to reduce it.

Alternatively, we might have a way of determining blue cycle time $CT = W$ exactly and wish to know the average inventory of blue items, i.e., the blue $WIP = L$. This is the case in the following example.

Consider a toy manufacturer that contracts with a third party logistics firm, 3PL, to handle its on-line business. The toy manufacturer will supply inventory to 3PL to fill orders. When the toy manufacturer receives an order on-line, it will instruct 3PL to fill it.

The toy manufacturer pays 3PL to provide this service. The contract terms depend on two factors: the number of orders shipped and the amount of inventory space occupied by the toys at 3PL. The toy manufacturer pays 3PL \$10 for every order that is shipped and \$0.03 per day for each unit in inventory.

At the start of each month, the toy manufacturer ships a batch of toys to 3PL. These toys typically sell out within the month, but not always. For accounting reasons, the toy manufacturer insists that 3PL submit an invoice for its services for each batch of toys. Hence, 3PL waits until the entire batch has been sold before it can finalize an invoice. In preparing the invoice 3PL can easily determine the shipping-cost component, as it just depends on the number of toys in the batch. However, 3PL is less clear about how to account for how much inventory space is attributable to a batch of toys. One approach would be to count its inventory each day. However, it would be quite difficult to track the inventory associated with a particular batch since the toys from all batches, as well as from other suppliers, are stored together in one section of the warehouse. Thus, it would be cost prohibitive to do an inventory count each day.

A much simpler approach is to use Little's Law. Suppose that whenever 3PL receives a batch of toys, it records the arrival date for each toy. 3PL can also record the date at which the toy is shipped, and hence obtain the time in system for each toy.

For instance, if there were an RFID tag attached to each toy, then 3PL can easily record these transactions with RFID readers at its shipping docks. Thus, 3PL knows the wait times W_i for $i = 1,2, \ldots N$, where W_i is the difference between the ship time and receipt time for the i^{th} toy and N is the number of toys in a batch. The average wait time for the batch is

$$W = \frac{1}{N}\sum_{i=1}^{N} W_i.$$

If it takes T days to sell the batch of N toys, then the average arrival rate for the batch is

$$\lambda = \frac{N}{T}.$$

By Little's Law we now can find the average inventory attributable to this batch:

$$L = \lambda W = \frac{1}{T}\sum_{i=1}^{N} W_i.$$

Since L represents the average inventory over a period of T days, 3PL will charge the toy manufacturer for L × T unit-days of inventory storage, at \$0.03 per unit-day.

Concluding Remarks

We have given a variety of examples showing the kinds of situations where Little's Law can usefully convert an estimate of an average queue into an estimate of average waiting time and vice versa when one may be relatively easy to measure and the other not.

We have also briefly examined how Little's Law has been used in operations management. Here we observe that a different terminology and set of symbols is usually adopted. Operations managers are concerned with their throughput rate rather than an arrival rate; their queue length is usually *WIP*, their wait time is termed cycle or flow time. We also discussed two differences in orientations between the use of the law in operations management and its original derivation and application:

- For operations management the law is often expressed in terms of output rates rather than arrival rates. Furthermore, the arrival process to most operating systems is not a stationary process, and, indeed, may be controlled.
- Many operating systems are never empty. That is, the number in the system or *WIP* is always positive.

In each case we see that Little's Law can apply, albeit with some required conditions and thoughtful attention to the goals of the application.

Historical Background

We trace the evolution of Little's Law from the early days of queuing theory in operations research and management science to our own chapter in this book.

The earliest paper we have found that makes use of Little's Law simply assumes it to be true. Cobham (1954), in an article on priority queues, writes, "... it is sufficient to observe that the *expected number* of units *of priority k* waiting to be serviced is $\lambda_k W_k$, where W_k *is the expected wait* for a unit of priority k." (λ_k is the arrival rate of priority k items.)

We attribute the explicit formula "$L = \lambda W$" to Philip Morse and his book, *Queues, Inventories and Maintenance*, (Morse, 1958). He does not give the law a name but simply talks about "the relation between mean number and mean delay." In Chap. 7 he proves the relationship for a single channel queuing system having Poisson arrivals and a service time distribution of the general class known as k-Erlang. The proof of $L = \lambda W$ is, in a certain sense, incidental to his main task of solving for the steady-state joint probabilities of the number of items in the system and the stage of item in service. Using the fairly standard approach of describing the system with a set of differential equations, Morse solves the system by building a two variable generating function of the desired joint probabilities. One variable relates to the number of items in the system and the other to the stage of the item in service. The generating function can then be used rather easily to find both L and W. Examination shows that L and W differ only by the constant of proportionality λ. This is a nice piece of analysis of the particular system, but it does not readily suggest a route to greater generality.

Morse was clearly interested in the general case. After the above analysis he wrote, "we have now shown that... the relation between the mean number and mean delay is via the factor λ, the arrival rate: $L = \lambda W$ and..... We will find, in *all* the examples encountered in this chapter and the next, for a wide variety of service and arrival distributions, for one or for several channels, that this same relationship holds. Those readers who would like to experience for themselves the slipperiness of fundamental concepts in this field and the intractability of really general theorems, might try their hand at showing under what circumstances this simple relationship between L and W does *not* hold."

The next paper to appear was Little (1961) and had the title "A Proof of the Queuing Formula: $L = \lambda W$." Little essentially analyzes the picture shown in Fig. 5.2 of our chapter. His argument uses ergodic theorems for strictly stationary stochastic processes, as drawn from Doob (1953). The basic analytic task is to get rid of "end effects" as discussed below Fig. 5.2 in this chapter. However, Little does something else not done previously. His proof argues that the law holds in any specific realization of the queuing system when observed over a long period of time. This is different from finding steady state probabilities for the number in the system and calculating L and performing a corresponding analysis of the distribution of waiting times to find W. The idea that the theorem is true for each evolution of the system provides a deeper understanding of the importance of the relationship $L = \lambda W$ in the queuing process itself. It also lays a basis for sample path proofs of the relationship that are to come.

Not too long afterward, Jewell (1967) came along with "A Simple Proof of $L = \lambda W$." He draws a picture very similar to our Fig. 5.3 and makes the assumption that the event when the system becomes empty is a recurrent event. By definition the times between such events are mutually independent random variables having the same distribution. Thus the time line alternates between intervals during which the system is empty and ones during which the system is busy. Jewell assumes that arrival and waiting mechanisms are reset at the start of each busy period. In other words, in each busy period, the random variables in those mechanisms have the same joint distribution, and their values are new random draws, independent of previous busy periods. An advantage of Jewell's paper was that it used a vocabulary more familiar to queuing system analysts than the measure-theoretic arguments of stationary stochastic processes invoked by Little (1961).

Jewell's paper was followed by Eilon (1969), which had the title "A Simpler Proof of $L = \lambda W$." He shows essentially the same picture as Jewell and as our Fig. 5.3 and makes the same heuristic analysis that we do under Fig. 5.2. He notes that, if the limits of $L(T)$, $W(T)$, and $\lambda(T)$ exist as T goes to infinity, the result is proven. Most writers on the subject, however, consider that a further argument is required to be certain that end-effects go away in the limit.

As time went on, the number and importance of applications of queuing systems increased and with them the applications of Little's Law. One major area lay in the design and analysis of computer systems. Kleinrock (1975, 1976) develops and summarizes a set of tools to assist this. One of the present authors (Little) recalls receiving a telephone call out of the blue from a computer engineer on the West Coast during the 1970s. The caller asked, "Do you have any more laws? I use Little's Law all the time and find it really helpful." But the response from the East Coast was "Sorry, we're fresh out of new laws today."

Although researchers in the field had no doubt about the remarkable generality of Little's Law, there developed among some of them the belief that its validity could be proven deterministically by sample path analysis. Such an approach would produce a reader-friendly proof without recourse to probability, somewhat in the manner that this chapter argues the deterministic validity of Little's Law in the supermarket example. The net result was Stidham (1974), "A Last Word on $L = \lambda W$."

However, as he pointed out in a later review article (Stidham, 2002), it was not the last word at all because the research potential of sample path analysis for queuing and other applications has ramifications far beyond Little's Law. Many results of this sort appear in a book by Stidham and a colleague: (El-Taha and Stidham, 1999): *Sample-Path Analysis of Queueing Systems*.

Over the past few years, Little's Law has become increasingly important in operations management. The notation and type of thinking are different, but applications are growing in number and importance. The goal of this chapter has been to facilitate such applications by providing illustrative examples, especially ones that explore new territory.

A Note of Personal History (Little)

How did a sensible young PhD like me get involved in a crazy field like this? From 1957–1962, I taught operations research at the Case Institute of Technology in Cleveland (now Case Western Reserve University). I was asked to teach a course on queuing. OK. Initially I used my own notes, but when Morse (1958) came out, I used his book extensively. Queuing was taken by most of the OR graduate students and, indeed, one of these, Ron Wolff, went on to become a first class queuing theorist (Wolff 1989). One year we were at the point when we had done the basic Poisson-exponential queue and moved through multi-server queues, and some other general cases. I remarked, as many before and after me probably have (and Morse does), that the often reappearing formula $L = \lambda W$ seemed very general. In addition I gave the heuristic proof that is essentially Fig. 5.2 at the beginning of this chapter. After class I was talking to a number of students and one of them (Sid Hess) asked, "How hard would it be to prove it in general?" On the spur of the moment, I obligingly said, "I guess it shouldn't be too hard." Famous last words. Sid replied, "Then you should do it!"

The remark stuck in my mind and I started to think about the question from time to time. Clearly there was something fundamental going on, since, when you draw the picture you do not really seem to need any detailed assumptions about interarrival times, service times, number of servers, order of service, and all the other ingredients that go into the panoply of queuing models. You only seemed to need a process that goes up and down in unit amounts and some guarantee of steady state and conservation of items. In addition, because I could see there were end effects in the picture, there needed to be a way to get rid of them in the limit. It seemed to me I was in the general arena of stationary stochastic processes. I am not a mathematician by training, and so I bought copies of measure theoretic stochastic process books like Doob (1953), which mentioned stationary processes and ergodic theorems.

My family's habit at the time was to go to Nantucket in the summer where my wife's family had a small summer house. We would load our children in a station wagon, drive to Woods Hole, take the ferry, and spend a couple months away from

the world. Since the beach was the baby-sitter, I was able to split off solid blocks of time to work on research as a good assistant professor should. (I wish I could do that today!) I always brought a pile of books and projects with me. $L = \lambda W$ was one of them. I soon ran into problems that required more than looking up theorems in my new books, but I worked out approaches to the road blocks and eventually wrote everything up, giving it my best shot. I sent the paper off to *Operations Research.* It was accepted on the first round.

Nevertheless, I had learned my lesson. I decided that Borel fields and metric transitivity were not going to be my career and retired from queuing. That was in 1961. My retirement held until 2004 when I was accosted by email and in person at an INFORMS meeting by Tim Lowe. Even then, as he will tell you, I resisted re-entrapment, saying, "I don't know anything about OM and I haven't looked at $L = \lambda W$ for 40 years." Being always susceptible to a new challenge and, more importantly, thanks to much help from Steve Graves, who really does know OM, I took a run at holding up my end of the chapter. It has been a wonderful experience and I have learned much. Now I need to find ways to use it.

References

Cobham, A. (1954) "Priority Assignment in Waiting Line Problems," *Operations Research*, 2, (1) (Feb.), 70–76.

Doob, J. L. (1953) *Stochastic Processes*, John Wiley, New York.

Eilon, S. (1969) "A Simpler Proof of L = λW," *Operations Research*, 17, (5) 915–917.

El-Taha, M. and S. Stidham (1999) *Sample-Path Analysis of Queueing Systems*, Kluwer Academic, Boston MA.

Hopp, W. J. and M. L. Spearman (2000) *Factory Physics: Foundations of Manufacturing Management*, 2nd (ed.), Irwin/McGraw Hill, New York, NY.

Jewell, W. S. (1967), "A Simple Proof of L = λW," *Operations Research*, 15, (6) 1109–1116.

Kleinrock, L. (1975) *Queueing Systems, Volume I: Theory*, A Wiley-Interscience Publication, New York.

Kleinrock, L. (1976) *Queueing Systems, Volume II: Computer Applications*, A Wiley-Interscience Publication, New York.

Little, J. D. C. (1961) "A Proof of the Queuing Formula: L = λW," *Operations Research*, 9, (3) 383–387.

Little, J. D. C. (1992) "Are there 'Laws' of Manufacturing," *Manufacturing Systems: Foundations of World-Class Practice*, edited by J. A. Heim and W. D. Compton, National Academy Press, Washington, D.C., 180–188.

Morse, P. M. (1958) *Queues, Inventories and Maintenance*, Publications in Operations Research No. 1, John Wiley, New York.

Stidham, S., Jr. (1974) "A Last Word on L = λW," *Operations Research*, 22, (2) 417–421.

Stidham, S., Jr. (2002) "Analysis, Design, and Control of Queueing Systems," *Operations Research*, 50, (1) 197–216.

Wolff, R. W. (1989) *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.