In most processes, the demand for finished goods or services varies over time, as do raw material supply rates and flow rates at different process steps. This note discusses the effects of random variability on basic process performance measures such as flow time, inventory levels, and flow rate. The concepts presented in this note are drawn from the field of queueing, the study of how variability influences flow time and flow rate in processes. After a brief discussion of terminology, we begin by introducing two simple congestion models for single-step processes with unlimited buffer space. The note then discusses different managerial approaches that can be taken in response to variability, as well as some economic consequences of each approach.

**Wait Time for a Simple Case**

To provide insight into the relationship between variability and performance, consider a simple operating system consisting of a single step (Figure 1). Here work arrives from an upstream step and is processed by a single worker or machine. Work that arrives to find the system empty is processed immediately, otherwise it joins a queue and is processed on a first-in-first-out (FIFO) basis. We would like to determine the average amount of time a unit of work spends in a buffer waiting for service. For simplicity, we assume there is sufficient buffer space before the process step for placing work (i.e., work is never blocked from entering the system).

![Figure 1: A One-Step Operation](image)

---

1 This is a revision by Marshall Fischer, Kathy Pearson, Christian Terwiesch, and Karl Ullrich of a note prepared by Karen Donohue for OPIM 631. The original note was prepared by the Operations teaching group at the Stanford University Graduate School of Business.
**Intuitive Example**

To illustrate a simple system and the effects of variability in interarrival time and service time on a queue or buffer, consider a scenario familiar to many of you. At Lippincott Library, the student clerk at the front desk seems to be idle a fair amount of time, yet often has a waiting line at the very moment you need assistance. Why does this happen? Suppose the clerk has the capacity to process an average of 20 requests each hour and the average number of requests is 10 per hour. On the surface, these times seem to ensure that no student will ever wait for assistance. However, both processing time and time between student arrivals have some variability. For example, although the average request takes 3 minutes to complete (60 minutes per hour / 20 requests per hour), each individual request can vary significantly. Similarly, the average time between arrivals is 6 minutes (60 minutes per hour / 10 requests per hour), but the individual interarrival times vary greatly.

If we examine a typical day at the library front desk, we can quickly see the effect of this variability. Suppose the doors open at 8:00 a.m. The first student, Jane, approaches the desk at 8:30. Her request takes 3 minutes and she leaves the desk at 8:33. The next student, Joe, arrives 6 minutes after Jane, at 8:36. His request takes only 2 minutes, and Joe leaves at 8:38. In the meantime, Jack arrives only 1 minute after Joe, at 8:37. The clerk is busy with Joe for another minute then gets to Jack at 8:38, after Jack has waited for 1 minute. Jack has a very complicated request that takes 7 minutes, so the clerk is not available again until 8:45. In the meantime, Jill has arrived with a question at 8:39, just 2 minutes after Jack, and must listen to Jack and the clerk for a total of 6 minutes before having her own lengthy question answered. Table 2 summarizes the front desk activities for the first nine students that ask for help. This table shows that a total of four students out of the nine must wait, and the clerk is idle when the other five students approach the desk.

<table>
<thead>
<tr>
<th>Time Until Next Arrival</th>
<th>Service Time</th>
<th>Student</th>
<th>Time Student Reaches Desk</th>
<th>Time Service Begins</th>
<th>Time Service Ends</th>
<th>Student Wait Time</th>
<th>Server Idle Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>3</td>
<td>Jane</td>
<td>8:30</td>
<td>8:30</td>
<td>8:33</td>
<td>0 min</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Joe</td>
<td>8:36</td>
<td>8:36</td>
<td>8:38</td>
<td>0 min</td>
<td>3 min</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>Jack</td>
<td>8:37</td>
<td>8:38</td>
<td>8:45</td>
<td>1 min</td>
<td>0 min</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>Jill</td>
<td>8:39</td>
<td>8:45</td>
<td>8:50</td>
<td>6 min</td>
<td>0 min</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>John</td>
<td>8:51</td>
<td>8:51</td>
<td>8:52</td>
<td>0 min</td>
<td>1 min</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>Jim</td>
<td>9:09</td>
<td>9:09</td>
<td>9:13</td>
<td>0 min</td>
<td>17 min</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>Jennifer</td>
<td>9:10</td>
<td>9:13</td>
<td>9:16</td>
<td>3 min</td>
<td>0 min</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>Jessica</td>
<td>9:11</td>
<td>9:16</td>
<td>9:18</td>
<td>5 min</td>
<td>0 min</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>Joshua</td>
<td>9:21</td>
<td>9:21</td>
<td>9:22</td>
<td>0 min</td>
<td>3 min</td>
</tr>
</tbody>
</table>

**Table 2: Front Desk Activities at Lippincott Library**
Queueing Formulas for Calculating Performance Measures

The scenario depicted in the previous example can be represented mathematically, using formulas developed in queueing theory. We will concentrate on one of these formulas in this note, and then revisit Little’s Law.

The average wait, designated by \( W_q \), is defined as the average amount of time that work must wait in the queue before being processed. (Note that work arriving to find an empty system suffers no wait at all but those zeros are still taken into account when computing average wait.) Before proceeding it is necessary to define some notation for the single-process system:

- \( a \): average interarrival time (i.e., average time between arrivals of units of work or \( 1/D \), where \( D \) is the demand rate),
- \( p \): average processing time (activity time) per work unit,
- \( u \): capacity utilization = (demand rate) / (capacity) = (1/a) / (1/p),
- \( c_a \): coefficient of variation of interarrival time = (standard deviation of interarrival time) / (average interarrival time),
- \( c_p \): coefficient of variation of processing time = (standard deviation of processing time) / (average processing time).

The coefficients of variation \( c_a \) and \( c_p \) are used to reflect the level of variability in processing times and interarrival times. A coefficient of variation of less than 1 implies that the time’s standard deviation is smaller than its mean. From a process efficiency point of view, the lower the coefficient of variation the better.

Using these variables, the waiting time of work in the queue can be approximated by the following formula when utilization is relatively high (e.g., \( u \geq 0.5 \)):

\[
W_q \approx p \times \left( \frac{u}{1-u} \right) \times \left( \frac{c_a^2 + c_p^2}{2} \right)
\]

Returning to our example, a time study revealed that the standard deviation of process time and interarrival time at the clerk’s station are both 6 minutes. Using these inputs, we can calculate the following:
If we plug these values into equation (1), we obtain:

\[ W_q = 3 \times \left( \frac{0.5}{1 - 0.5} \right) \times \left( \frac{1^2 + 2^2}{2} \right) = 7.5 \text{ minutes}. \]

Customers will spend an average of seven and a half minutes waiting for the clerk. Of course, some customers will not wait at all while others wait much longer than this.

Once \( W_q \) is known, the average number of units of work in queue (e.g., the inventory level or number of customers) is also computable using a variation of Little’s Law. In Little’s Law, we see that inventory is equal to the flow rate multiplied by the flow time. Similarly, we can build a relationship between the average number of units of work in a queue, the flow rate, and the average waiting time in the queue:

\[ \text{Average Units of Work in Queue} = L_q = W_q \times \text{Flow Rate}. \quad (2) \]

In our example, the flow rate is simply the rate of customer requests, or 10 requests per hour (which also can be expressed as 1 request every 6 minutes). We also know that the average wait time is 7.5 minutes. Using these two pieces of information, we can calculate the average number of students waiting in line by applying equation (2):

\[ \text{Average number of students waiting} = (7.5 \text{ minutes}) \times (1 \text{ request / 6 minutes}) = 1.25 \text{ requests or students waiting}. \]

**Significant Implications**

The approximation given in Equation (1) becomes more accurate as utilization grows. A derivation of this formula can be found in Appendix II of Wagner (1975). This expression has the following implications:

- Variability in work arrivals and in processing times causes congestion and delay.
- As the variability of either arrival or processing times increases, waiting time and inventory levels increase.
• For a given level of variability, increasing utilization increases waiting time at an accelerating rate. As utilization approaches 100%, waiting time becomes infinite!

The last point is surprising and probably counterintuitive. However, we can use a simple example to illustrate this implication. Suppose we have a manufacturing station for which we want high utilization. At this station, we have control over the arrival rate of parts to be processed. Processing time at our station averages 5 minutes, with a variance of +/- 1 minute. In order to insure high utilization, we choose parts to arrive every 5 minutes. We soon discover that when the processing time is less than 5 minutes, our station is idle, which decreases utilization. We then choose an arrival rate of 4 minutes to ensure that there will always be a part waiting to be processed. We now have 100% utilization at our station but our work-in-process inventory, and subsequently waiting times, are increasing steadily as the process continues.

Note that in reality we will not approach infinite wait times. In our discussion thus far we have assumed that work always joins the queue and awaits processing, regardless of how many other units of work it finds waiting ahead of it. In real manufacturing processes, there is often limited space for holding inventory between steps. If the inventory buffer fills up, no additional work is allowed to “arrive”. In service processes, buffer limitations are less common but real customers may refuse to join a queue if it grows too long. This causes the overall system to behave as if it had a finite inventory buffer. Congestion and delay constitute a degradation of product quality for service systems. Once quality gets low enough, customers refuse to buy.

**Effect of Variability on Capacities**

Now, suppose we are designing a production or service system in which there will be some variability. Should the capacity at each step be equal to the average demand rate? The preceding model shows that the answer is no. With stochastic variability, it is not optimal to balance capacities exactly.

For example, Figure 2 depicts a simple manufacturing process where Step 1 represents fabrication of the product and Step 2 represents final assembly. Here fabrication capacity is expensive, assembly capacity is cheap, raw materials are relatively cheap, but high energy requirements make for a large value added in fabrication. This means that WIP inventory is expensive to carry, and the firm should build excess capacity at the assembly step to keep WIP inventory low. Just exactly how much excess capacity is called for depends on many factors including the desired manufacturing lead time, magnitude of demand fluctuations, storage area available for raw materials, and variability within the production process itself.
Figure 2: A Two-Step Manufacturing Process.

Wait Times for Parallel Processors

Imagine, now, that the front desk at Lippincott library is staffed by several clerks. People walk up to the desk and wait in line—a single queue, like the “snake” you may encounter at a bank branch or a post office. Whenever a clerk frees up he or she serves the person who is waiting at the front of the line. Figure 3 shows the manufacturing analog of this system.

Figure 3: A One-Step Operation with Multiple, Parallel Processors.

The following approximation for average wait in this system’s queue, can be found in Chapter 8 of Hopp and Spearman (1996). Again, it should only be used when utilization is fairly high (e.g., $u \geq 0.5$):
\[ W_q = \left( \frac{p}{m} \right) \times \left( \frac{u^{\frac{1}{m+1}} - 1}{1 - u} \right) \times \left( \frac{c_a^2 + c_p^2}{2} \right), \quad (3) \]

where

\( a = \) average interarrival time to the (shared) input buffer,

\( c_a = \) coefficient of variation of inter-arrival times to the (shared) input buffer,

\( p = \) average processing time of each processor,

\( c_p = \) coefficient of variation of processing time at each processor,

\( m = \) number of processors working in parallel,

\( u = \) system utilization

\[ = \text{(demand rate) / (capacity)} = \frac{1}{a} / \left[ m \times \frac{1}{p} \right]. \]

Note that for a one-server system, \( m \) equals 1, and approximations (1) and (3) become identical. Furthermore, when designing systems, we can use (3) as a guide to roughly determine the number of processors we wish to operate in parallel at a given processing step.

For example, suppose we had the choice of buying one fast machine of processing speed \( p \), or \( m \) slow machines, each of speed \( \frac{p}{m} \). Clearly the average processing rates of the two are the same. Suppose also that the choice of 1 or of \( m \) machines does not affect \( c_a \) or \( c_p \). Then approximation (3) shows that increasing parallelism \( m \) decreases the average wait in queue \( (W_q) \). Figure 4 summarizes these results for a variety of system utilizations.

**Delay and Congestion in More Complex Systems**

Generally, one would like to know about congestion and delay in more complex systems like the process pictured in Figure 5. The customers in this process may be actual human beings, as in the case of a hospital, or simply paper orders, as in the case of Manzana Insurance (which we will discuss in a later class). Note that work may begin its journey through the system at any of several process steps; multiple arrows emanating from step indicate that all work does not follow the same route.
Figure 4: $W_q$ decreases as $m$ increases, for a given level of capacity utilization, $u$, and assuming constant variability, $c_a$ and $c_p$.

Figure 5: Network Structure of a Complex System.

For simpler systems, there exist simple mathematical formulas, such as (1) and (3), that approximate (or, sometimes, exactly describe) system performance. For more complex situations, simulation can be used model system performance. We will demonstrate a series of simulation models in a later class. All of these models offer the same basic insights into the relationship between utilization and average delay that we saw in (1). Remembering that higher utilization of the resources results in a higher throughput rate, the following generalizations can be made:
• In the presence of variability, there is a trade-off between high throughput rates and low manufacturing lead times.

• This trade-off becomes more severe as process variability increases.

• In the presence of variability, queue time can become many times larger than processing time as utilization approaches 100%.

These generalizations are graphically summarized in Figure 6.

*Empirical Observation from the Retail Industry*

In general, these predictions are consistent with empirical observation. Consider, for example, the processing of suits at Joseph Bank Clothiers. The average raw process time for one suit is only a few hours, and yet it takes five to six weeks for a suit jacket to complete its journey through the plant. Almost all of the time that the suit jacket spends in the factory is spent having nothing done to it. Some of that time is “queueing”, which is time spent in line while other batches of suit jackets (that are ahead in the line) are processed.
Figure 6: Increased Utilization and Increased Variability Adversely Affect Lead Time

The low ratio of processing time to manufacturing lead time at Joseph Bank should come as no surprise. High utilization of operators is an important goal at Joseph Bank, whose customers are price sensitive. At the same time there is significant variation in the time it takes one operator to do different tasks, as well as in the time it takes different operators to do a given task. To achieve this high utilization of operators in a highly variable environment, Joseph Bank requires both a high inventory and a high manufacturing lead time.

Responses to Variability

As the variability in a process becomes more severe, it becomes more and more difficult to achieve high utilization and low inventory simultaneously. Possible responses to this situation include turning away business and foregoing potential production, building extra capacity, or simply allowing queues and inventories to be large. However, there are generally also opportunities to decrease the variability effectively, smooth out the flow of materials, or decrease the inventory requirements needed to achieve a given throughput rate. These opportunities will be divided into three categories: external policies, internal policies, and technology.

External Policies

External policies are policies that affect orders or material outside the manufacturing or service system. Such policies can be used to decrease the variability inside the system. Examples of such policies are:

- Work with vendors to provide a smooth, continuous supply of materials, and with customers to accept a smooth, continuous supply of products.
- Limit the variability and complexity of work accepted for processing. For example, McDonald’s does not normally accept customized sandwich orders. This significantly reduces variability in processing times and enables McDonald’s to fill orders in less time than Wendy’s or Burger King.

Internal Policies

Internal policies are policies that affect customers or materials that are waiting to enter or have already entered the operating system. Examples of such policies are:

- Release work into the system in a way that smoothes the flow of material through the process. For example, work may be released as a steady stream, rather than in bursts of many units of work at once. As another example, work may be released only when buffer inventories between resources become too low.
• Sequence work to decrease average queueing time. For example, the “shortest processing time first” (SPT) rule can be shown to achieve shorter average throughput times than a “first-in-first-out” (FIFO) rule. Queued work may also be sequenced to reduce the number of set ups, or even the length of sequence-dependent set ups. If due-date integrity is an issue, due-dates should also be incorporated into the sequencing rule.

• Work may be routed through the process in ways that decrease variability at specific resources. For example, one variety of similarly milled parts may be routed to one mill operator, and another variety may be routed to a different mill operator.

Technology
Technology, either applied to resources or to information processing, can also be used to decrease variability, or the impact of variability. For example:

• Equipment can be made more flexible or versatile. For example, increasingly flexible equipment often implies reduced or eliminated set ups to switch between types of work. (Longer set up times contribute to overall process time variability and motivate larger batch sizes to maintain a given throughput rate.) In addition, if multiple copies of the same type of versatile equipment can replace many different types of dedicated equipment, then pooling efficiencies may result.

• Greater automation can decrease variability. For example, automation can significantly reduce the standard deviation of the times to perform such operations as loading, unloading, or processing on machines.

• Long, unexpected breakdowns can be a particularly severe contributor to variability in manufacturing systems. Increased equipment reliability can reduce this source of variability.

Conclusion
The big message is that variability is costly. It is the source of many fundamental problems in operations management and requires trade-offs among utilization, inventory levels, and lead time, as well as a concerted effort to manage and reduce variability as much as possible.

References

Study Questions

1. A new branch of the First National Franklin (FNF) Bank is opening in Doylestown, where one branch is already in existence. The current branch has one drive-up ATM for the convenience of its customers. First National Franklin is debating whether to include a drive-up ATM at its new facility, or whether the current machine is servicing the customers satisfactorily. During peak hours, the average arrival time between cars is 1.25 minutes, with each car spending an average time of 1 minute at the ATM. The time between arrivals varies with a standard deviation of 0.4 minutes, while the time spent at the machine varies with a standard deviation of 0.3 minutes.
   (a) What is the utilization time for the ATM during peak hours?
   (b) Given that FNF wants to keep the average wait times at the ATMs to less than half a minute, should the new branch have an ATM? Specifically, what is the average wait time for each car, and on average, how many cars are waiting at any given time during the peak hours?

2. M.M. Sprout, a catalog mail order retailer, has one customer service representative (CSR) taking orders on an 800 telephone number. If the CSR is busy, the caller is put on hold. For simplicity, assume that any number of incoming calls can be put on hold, and that nobody hangs up in frustration due to a long wait. Suppose, on average, one call comes in every four minutes with a standard deviation of two minutes, and it takes the CSR an average of three minutes to take the order with a standard deviation of three minutes. The CSR is paid $20 an hour, the telephone company charges $5 an hour of time that the line is used, and it is estimated that each minute that a customer is kept on hold costs Sprout $2 due to customer dissatisfaction and loss of future business.
   (a) Estimate the proportion of time the CSR will be busy, the average time that a customer will be on hold, and the average number of customers on line.
   (b) Estimate the total hourly cost of service and waiting.

3. M.M. Sprout is now considering adding another line to his customer service department. If he does, he can assume the same average processing time and the same costs.
   (a) Provide new estimates for system utilization, the average time that a customer will be on hold, and the average number of customers on hold.
   (b) Estimate the new total hourly cost of service and waiting.
4. Since the deregulation of the airline industry, the increased traffic and fierce competition have forced Global Airlines to reexamine their operations for efficiency and economy. As a part of their campaign to improve customer service in a cost effective manner, they have focused on the passenger check in operation at the airline terminal. For best utilization of their check-in facilities, Global operates a common check-in system wherein passengers for all of Global’s flights line up in a single “snake line”, and each can be checked in at any one of the counters as the clerks become available. The arrival rate is estimated to be 52 passengers per hour, with an interarrival time standard deviation of one minute. During the check-in process, an agent checks the reservation, assigns a seat, issues a boarding pass, and weighs, labels, and dispatches the baggage. The entire process takes an average of two minutes, with a standard deviation of two minutes. Currently, Global Airlines staffs two check-in ticket agents at any given time.

(a) Draw a process flow diagram for this process.

(b) Estimate the average customer wait before being served and the average number of passengers waiting.

(c) Should Global Airlines consider adding another ticket agent? What is the estimated difference in wait times?

(d) Consider a separate queue system, where passengers make a decision to line up at one of two different ticket agents. Draw a process flow diagram for this process. How does it differ from the “snake” queue? Intuitively, which is more efficient?