Vol. 5, No. 2, January 2005, pp. 17–24
ISSN 1532-0545 | 05 | 0502 | 0017
INFORMS Transactions on Education
DOI 10.1287/ited.5.2.17

The Pooling Principle

Kyle Cattani
The Kenan-Flagler Business School, UNC Chapel Hill, Chapel Hill, North Carolina 27599,
kyle_cattani@unc.edu

Glen M. Schmidt
The McDonough School of Business, Georgetown University, Washington, DC 20057,
schmidtg@msb.edu

A key insight of basic inventory models and of the G/G/m queueing model is the effect of "pooling." That is, these models suggest that pooling of customer demands, along with pooling of the resources used to fill those demands, may yield operational improvements. While there are many research papers that articulate and quantify the benefits of pooling, a review of popular texts in operations management suggests that most texts lack simple examples that illustrate this important concept. This note is intended to fill this void by articulating the "pooling principle" in an intuitive, straightforward fashion. This note is comprised of a brief instructor's note, followed by a note that can be included in a course pack or handed out to students. Earlier versions of the note have been used successfully in core operations courses at both the undergraduate business and MBA levels.

The Pooling Principle: Instructor's Note

This note fills a gap in the textbooks that are used in many core operations courses, namely a lack of discussion of the pooling principle. The lack of coverage is widespread. In a review of 13 operations management textbooks (see list in Table 1 below), we find none have "Pooling" listed in the subject index. The typical textbook in the list has discussions on resource flexibility but does not discuss the statistical benefit of pooling resources in either a queueing or inventory setting nor does it provide simple examples to illustrate the concept.

We also reviewed a number of related texts involving quantitative methods or supply chain management—these do not have "operations management" in the title (but might also be used in some settings for a core operations class). Within this broader group we found several that had explicit references to pooling along with some examples. Anupindi et al. (1998) and Bonini et al. (2000) include a treatment of pooling in a queueing model and Hopp and Spearman (2000) discusses queueing effects at some length, while Simchi-Levi et al. (2003) discusses pooling in an inventory setting and includes a computer simulation for students to gain intuition and insight about the relationships of the various parameters.

The fact that pooling is discussed only in these related texts might suggest that it is either too complex to readily be conveyed to students, or that it cannot readily be applied by MBAs and undergraduate business students, or that it is not a key operations management tool that can be used to increase profits. We believe otherwise. For example, pooling might be a key motivation behind a firm's consolidation of sales regions, its cross-training of employees, and a re-design of its products to facilitate dealer installation of certain options. Further, we believe the lessons of pooling can be conveyed in a relatively short and easy-to-understand fashion, and offer the accompanying student note as a possible aid in this effort.

In the note that follows we present a brief definition and then two examples of pooling benefits: an inventory example and a queueing example. We have successfully used this material in core operations courses in both undergraduate and MBA programs to supplement readings in our texts (which were not one of the related textbooks noted above).

For those interested in additional resources on the topic in a queueing setting, more extensive results than given in the texts listed above are offered in Whitt (1992). The literature on risk pooling in an inventory setting is extensive and includes many papers on commonality. For example, a recent paper by Van Mieghem (2004) discusses some of the value drivers of commonality including the traditional risk pooling benefit and summarizes the commonality literature. Wallace and Whitt (2004) show a little flexibility (i.e., a little pooling) can possibly "go a long

Table 1 Operations Management Textbooks

Chase, Jacobs, and Aquilano. 2004. ***Operations Management for Competitive Advantage***. Tenth Edition. McGraw Hill/Irwin. NY.

Dilworth. 1999. ***Operations Management: Providing Value in Goods and Services***. Third Edition. SouthWestern. Cincinnati OH.

Hanna and Newman. 2001. ***Integrated Operations Management: Adding Value for Customers***. Prentice Hall. Upper Saddle River NJ.

Heizer and Render. 2004. ***Operations Management***. Seventh Edition. Pearson/Prentice Hall. Upper Saddle River NJ.

Krajewski and Ritzman. 2005. ***Operations Management: Strategy and Analysis***. Seventh Edition. Pearson/Prentice Hall. Upper Saddle River NJ.

Martinich. 1997. ***Production and Operations Management: An Applied Modern Approach***. Wiley. NY.

Meredith and Shafer. 1999. ***Operations Management for MBAs***. Wiley. NY.

Morton. 1999. ***Production Operations Management***. SouthWestern. Cincinnati OH.

Reid and Sanders. 2004. ***Operations Management: An Integrated Approach***. Second Edition. Wiley. NY.

Russell and Taylor. 2003. ***Operations Management***. Fourth Edition. Prentice Hall. Upper Saddle River NJ.

Schroeder. 2000. ***Operations Management***. Irwin McGraw-Hill. Boston MA.

Starr. 2004. ***Production and Operations Management***. Atomic Dog Publishing. Cincinnati OH.

Stevenson. 2002. ***Operations Management***. Seventh Edition. Irwin McGraw-Hill. Boston MA.

way" in a queueing setting, while Tomlin and Graves (2003) discuss this in an inventory setting.

Those interested in additional classroom material on the topic of pooling might consider the Kopczak and Lee case (Hewlett-Packard Co.: DeskJet Printer Supply Chain (A)) to highlight the benefit of pooling in an inventory setting. For a good case that focuses on the benefits in a queueing setting, consider the Loch et al. case (Manzana Insurance) which deals with performance assessment and improvement of a service operation in the insurance industry. The case describes two branch offices in direct competition, and the impact of response time on performance is suggested. Pooling of geographical teams is an integral part of the case solution.

As we discuss in the student note, there are distinctions between pooling as we describe it and what students may initially associate with this word. For example, students will be familiar with the idea of pooling in such things as a car pool, where riders are grouped and "served" together. A more recent application of the vehicle pooling idea is the shared ownership of corporate jets. The demand for the jets is random, but the use by any one owner is far below capacity (see Keskinocak 1999). With this note we hope to expand students' thinking about how combining demands and the resources used to fill those demands can benefit the firm.

Table 2 Related Textbooks

Anupindi, Chopra, Desmukh, Van Mieghem, Zemel. 1998. ***Managing Business Process Flows***. Simon & Schuster. Needham Heights MA.

Bonini, Hausman, and Bierman. 1997. ***Quantitative Analysis for Management***. Ninth Edition. McGraw-Hill/Irwin. Boston MA.

Chopra and Meindl. 2004. ***Supply Chain Management: Strategy, Planning, and Operation***. Second Edition. Pearson/Prentice Hall. Upper Saddle River NJ.

Handfield and Nichols. 1999. ***Introduction to Supply Chain Management***. Prentice Hall. Upper Saddle River NJ.

Hopp and Spearman. 2000. ***Factory Physics***. Second Edition. Irwin McGraw-Hill. Boston MA.

Nahmias. 2000. ***Production and Operations Analysis***. Fourth Edition. Irwin McGraw-Hill. Boston MA.

Simchi-Levi, Kaminsky, and Simchi-Levi. 2003. ***Designing & Managing the Supply Chain***. Second Edition. Irwin McGraw-Hill. Boston MA.

References

Anupindi, Chopra, Deshmukh, Van Mieghem, Zemel. 1998. *Managing Business Process Flows*. Simon & Schuster Custom Publishing.

Hopp, Spearman. 2000. *Factory Physics*, 2nd ed. Irwin.

Keskinocak. 1999. Corporate high flyers, *ORMS Today*, **26**(6) 22–25.

Kopczak, Lee. 2001. *Hewlett-Packard Co.: DeskJet Printer Supply Chain* (*A*) case, available from Harvard Business School case number GS3A.

Loch, Grant, Harrison. 1993. *Manzana Insurance: Fruitvale Branch*. Stanford University Graduate School of Business School Case S-DS-87 (and teaching note).

Simchi-Levi, Kaminsky, Simchi-Levi. 2003. *Designing & Managing the Supply Chain* (2e), McGraw-Hill Irwin.

Tomlin, Graves. 2003. Process flexibility in supply chains. *Management Sci.* **49**(7) 907–919.

Van Mieghem. 2004. Commonality strategies: Value drivers and equivalence with flexible capacity and inventory substitution. *Management Sci.* **50** 419–424.

Wallace, Whitt. 2004. Resource pooling and staffing in call centers with skill-based routing. Working paper available at http://www.columbia.edu/~ww2040/recent.html.

Whitt. 1992. Understanding the efficiency of multi-server service systems, *Management Sci.* **38** 708–723.

The Pooling Principle: Note for Students

1. Introduction

In you form a car pool with another commuter, you use one resource (one car) to do the same job that would otherwise require two resources (two cars). This reduces your cost of commuting to work but may also cause you some inconvenience. Another way to reduce your commuting cost might be to share usage of one car with another person—Maybe you work the night shift while your spouse works the day shift. Similarly, some smaller firms who only have sporadic need for a jet might share ownership of an aircraft. This reduces the cost for each user but again, may sometimes be an inconvenience to the user (e.g., the jet may not be available on the day a firm wants it).

In each of the above two examples, one resource is asked to meet two otherwise separate demands that have been pooled together. In the first example, the user demand involves "getting to work," and pooling the two demands together reduces cost because one resource can actually process two jobs at once (each demand represents only a fraction of the resource's processing ability, so to speak). In the second example, pooling reduces cost because each demand represents only small fraction of the resource's available processing time (if demands are not pooled, a resource goes unused a large part of the time).

But what if each resource can only process one job at a time, and what if each resource is already almost fully utilized? Might it still be beneficial to pool demands and resources? For example, what if several divisions of a firm each had one aircraft that was almost always being used? Would the firm be better off by pooling all of the division's planes into one fleet? The pooling principle suggests that it might, and it is this type of pooling that we address in this note. Here we state the pooling principle as:

Pooling of customer demands, along with pooling of the resources used to fill those demands, may yield operational improvements.

It should be immediately noted that, depending on the setting, there may also be *disadvantages* to pooling. Therefore, one must weigh advantages against disadvantages to determine whether pooling should be pursued.

To help illustrate the pooling principle, this note offers two simple examples. Each example involves a comparison of a non-pooled system where there are multiple sets of customer demands and multiple sets of resources, and a pooled system where all demands and resources are pooled together. In the non-pooled system each set of customer demands can only be filled by using one specific resource set. In the pooled system, any of the available resources can be used to fill any of the customer demands. The pooling principle suggests the system of pooled demands and pooled resources may outperform the un-pooled system.

The first example that follows involves a setting where the resources are goods held in inventory, and demand comes from customers who would like to buy or consume those goods. You may have experienced this type of setting when you buy house paint: the paint store strives to have a wide variety of colors available for you to choose from and take home. In this setting the customer demand is a random variable: exact demand for specific colors and types of paint is unknown beforehand but the paint store presumably has an idea about what it will be (i.e., they may know the demand distribution). The demand is realized (it becomes known) during store hours.

In the second example demand is represented by an arriving stream of customers, and the resources are servers that perform a set of tasks for the customer. You probably experience this type of setting when you go to the bank: the bank tellers are the servers (they represent a set of resources) while you (and others) are customers (the set of demands). Demand is uncertain in that the exact timing and sequence of customer arrivals is not known beforehand, and services are also uncertain in that how long it will take to fulfill the exact service that each customer will request is unknown.

After presenting these two examples, we discuss some of the different ways firms implement pooling in the summary section.

2. Example 1: Pooling of Inventory

We first look at an example where the resource at a firm is an inventory of a specific product and the demand for that product is uncertain. We will show how pooling can help the firm deal with the demand uncertainty. (This highlights a key point—pooling can be thought of as a tool to manage uncertainty.)

Poolco has two warehouses in North Carolina one in Raleigh and one in Durham. The warehouses stock an identical item delivered to stores for purchases by customers in Durham and in Raleigh with the respective warehouses serving only stores in the same city as the warehouse. Weekly demand for this item at the warehouse in Raleigh is normally distributed with a mean of 2,000 and a standard deviation of 400. Weekly demand for the item at the warehouse in Durham is also normally distributed with a mean of 2,000 but with a standard deviation of 300. Demand at the two cities is independent. Poolco holds a safety stock of inventory to ensure they have sufficient stock in spite of the randomness of their demand, and sets safety stock equal to two times the standard deviation of weekly demand. This results in 800 units of safety stock in Raleigh and 600 units in Durham for a total of 1,400 units of safety stock. This means that on average they will hold an extra 1,400 units of inventory, and they can use these extra units to meet demand on weeks that it is higher than average.

Poolco is considering a consolidation of the warehouses: selling the warehouses in Durham and Raleigh and building a new warehouse in Chapel Hill to serve demand from stores in both Raleigh and Durham. From statistics we know that the mean and variance of a sum of random numbers is the sum of the mean and variance. (Remember that the standard deviation is the square root of the variance.) Thus, given that demands are independent, the pooled demand will be normally distributed with a mean of 4,000 $(= 2{,}000 + 2{,}000)$ and a standard deviation of 500 $(= \sqrt{300^2 + 400^2})$. The

safety stock in the pooled case necessary to achieve the same probability of a stockout in each order cycle as in the un-pooled case will again be two times the standard deviation. In the pooled case this is only 1,000 ($= 2 * 500$) units, or 400 units fewer than the original 1,400. This reduction in safety stock is the benefit offered by pooling.

The intuition behind the pooling benefit is that under the non-pooled warehouses, if demand in one city is higher than expected then it may run out of inventory even if there are extra units available in the other city. For example, this case might occur if demand in the other city is lower than expected. Thus, if the inventory for the two cities is pooled, then high demand in one city can be offset by lower demand in the other city. (If demand in both cities is high, then even the warehouse that makes use of pooling might run short.)

In consolidating its two warehouses into one, Poolco pools its customer demands, along with the resources (the inventory) used to fill those demands, and thus takes advantage of the pooling principle. Selling over the Internet is another way some companies create an opportunity to pool demands as well as resources. Consider a retailer such as Amazon.com. All customers across the nation (and perhaps beyond) are pooled into a single web site. And, as with Poolco, Amazon can also consciously pool inventory, by deciding the number and location of its various warehouses. The advantage of having warehouses in lots of cities is that warehouses will be closer on average to customers. This likely will result in lower shipping costs and quicker delivery (from the warehouse to the customer) compared to a centralized warehouse. On the other hand, a central warehouse (or small number of warehouses) will have the benefit of pooled demand with its concomitant reduction in safety stock for a desired service level. The centralized warehouse will likely have higher outgoing shipment costs (to the customer) perhaps somewhat offset by lower incoming shipment costs as orders from suppliers are consolidated into full-truck-load shipments.

Let's return to the paint retailer introduced above. Because there are many colors that a customer may want to purchase, to achieve a high service level if the paint store does not pool demands then the store would have to stock paint in every possible color, not to mention different finishes such as glossy or egg shell. But most paint retailers instead adopt an approach that allows them to pool demand for the different colors of paint. Namely they stock paint in a base color and add pigment to transform the paint into the desired color. In doing so they only have to stock to base color paint to accommodate the pooled demand for all colors. Note that they still have to stock the base color paint for each of the various finishes as well as sufficient pigment to be able to color the paint.

While our example focuses on how pooling can help deal with demand uncertainty, other benefits of pooling can be significant. For example, it may cost less to operate a single large warehouse rather than two smaller ones, due to economies of scale. In addition, other inventory savings may be possible such as might arise through order consolidation in a pooled system that might lower overall ordering costs.

3. Example 2: Pooling of Multiple Servers in a Queueing System

3.1. Place Your Vote

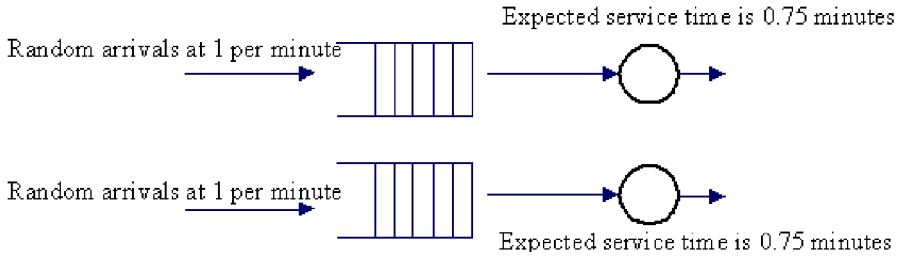
In the United States there are two dominant parties, the Republicans and the Democrats. Each party typically holds a primary election to determine who the party's candidates will be in the general election to follow. In the subsequent general election, the Republican candidates that were elected in the Republican primary run against the Democratic candidates similarly elected in the Democratic primary (candidates from other parties are also often on the ballot, and there is generally also the opportunity to write in a name). Each state, and sometimes each jurisdiction within each state, may have its own rules about how it runs its primary elections, and exactly how it runs the general election.

In Chapel Hill North Carolina the primary election process works as follows. Each party has a unique ballot, but the process and voting location (polling place) is the same for both parties: the voter identifies himself or herself and is given the appropriate ballot (Republican or Democrat) and a pen and a booth where he/she can fill out the ballot and finally submit it. Only after the voter submits the ballot is the next voter served. The polling place could choose to organize the polls based on party, with a separate line and booth for Republicans and Democrats, or the polling place could combine, or pool, the voters into one line that serves both Republicans and Democrats with two booths that can accommodate voters from either party.

Voters do not like to wait in long lines. Thus we are interested in knowing how long a voter would expect to wait in line before being served (i.e., before getting to vote), and in minimizing the time voters have to wait, if possible. Specifically, we wonder how the choice of how the lines and booths are organized affects the waiting time.

Assume that the average time between arriving Republicans is 1 minute and likewise 1 minute for the Democrats. While the average is 1 minute, the actual time between one voter arrival and the next varies.

Figure 1 Two Single-Server Systems



The time it takes to go through the voting process once a voter gets to the front of the line is 0.75 minutes, but this "service" time also varies between customers.

The process with separate polling stations is depicted in Figure 1.

In the pooled system, arriving voters form a single queue that serves both parties (see Figure 2). Voters arrive to the single queue twice per minute on average (the two streams involving arrivals of once per minute are simply added together). Each server takes the next voter in line, meaning she must be prepared to provide a Democrat or Republican ballot. That is, any arriving voter can be processed by either server, and we assume there is no difference in the time it takes to serve any one individual customer as compared to the system in Figure 1. A disadvantage of the pooled system in this case is that both servers must be flexible with the capability of handling a voter of either party, and the average service time might increase. For example, if this is a manual system where a person hands out ballots, the server must have a stock of both ballots and a process to ensure that the voter gets the correct one.

In either the un-pooled system of Figure 1 or the pooled system of Figure 2, servers are busy 75% of the time. In the un-pooled system, it takes, on average, 0.75 minutes to process a vote while voters arrive once per minute. In the pooled system, half the voters go to each server, indicating that once again, each server takes 0.75 minutes on average to process a voter, who arrives once per minute on average. In the next sections we quantify the specific benefits for the voting example and more generally the pooling benefits in queueing systems.

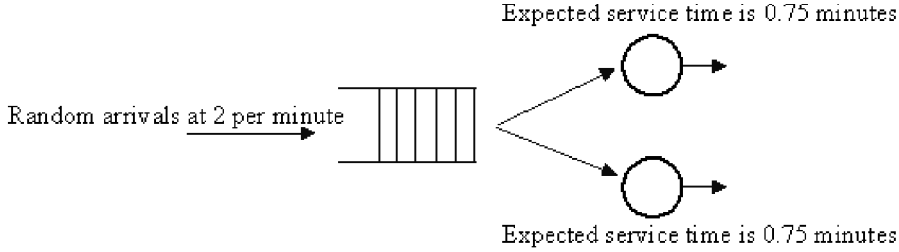
3.2. The Impact of Pooling in Queueing Systems

The pooled system reduces the time a voter expects to have to wait in line and vote. The intuition is as follows. If the servers are separate as in Figure 1, then there will be times when there are Republicans waiting but no Democrats waiting. If each server can accommodate either type of voter as in Figure 2, then instead of waiting, the next Republican in line can go to what was formerly the "Democrat's server." Similarly, in the case of Figure 1, there will be times when there are Democrats waiting but no Republicans waiting, in which case the system depicted in Figure 2 will offer shorter expected wait times.

We will calculate the magnitude of benefit that pooling offers, but first discuss four factors that determine the magnitude of the pooling benefit.

- **The average service time.** Assume you arrive to the voting station in the un-pooled system and there are two people in front of you. The longer it takes to serve each of those customers the longer you must wait for your turn—that is, the longer the average service time, the longer the expected wait time. Pooling helps mitigate this expected wait time.
- **The utilization of the servers.** If a server is almost never busy, then it will almost always be ready and waiting to process a newly arriving voter (i.e., there will be almost no waiting even in the un-pooled system), and thus pooling will be of little value.
- **The variability in arrivals and services.** The more variable these are, the greater the pooling benefit. (And the longer a customer expects to have to wait in any given system, pooled or un-pooled.) To gain some intuition as to why expected wait times go down as variability goes down, consider what

Figure 2 A Pooled System



would happen if there were no variability (e.g., voters arrived exactly one minute apart and it took exactly 45 seconds to process each person's vote). In this situation, nobody would ever have to wait.

• **The number of servers being pooled.** Say we started with three identical political parties of the type described above instead of two (assume the Green Party is equally popular in Any Town as the Republican and Democrat Parties), and then pooled the queues for the three voter streams into one queue served by three identical servers. The expected waiting time goes down as the number of pooled parties goes up, but there are diminishing returns. That is, pooling three parties into one system will further reduce the expected waiting time as compared to the system involving two pooled parties, but the reduction will not be as much as the original reduction achieved by pooling two parties into one system.

• The following formula can be used to estimate the wait time in the queue, denoted by W_q. This formula can also be used to verify the effect of the above four factors. Let s denote the average service time (in our example, $s = 0.75$ minutes), let b denote the fraction of time the server is busy, such that b is a fraction between 0 and 1 (in our example, we found the server to be busy 75% of the time), and let n denote the number of servers. (In Figure 1 we use $n = 1$ to determine the wait time for one of the two identical systems, while $n = 2$ in the system of Figure 2.) Interarrival times and service times are assumed to be variable with any distribution, and V denotes the variability factor. (V will be discussed in more detail after presenting the equation. For a more rigorous and extensive discussion of this formula, see (for example) Hopp and Spearman 2000.)

Expected Wait Time in a Queue:

$$W_q = (s)\left(\frac{b^{\sqrt{2(n+1)}-1}}{(1-b)n}\right)(V)$$

To gain further intuition as to how this formula conveys the impact of pooling, initially consider the special case where $n = 1$. Then, (you can confirm that) the formula reduces to the following.

Expected Wait Time in a Single-Server Queue:

$$W_q = (s)(b/(1-b))(v)$$

First, note that expected waiting time W_q goes up as service time s goes up. Second, note that W_q goes up as the server gets busier. To see this, note that for b close to zero, the term $b/(1-b)$, and thus the expected wait time is also close to zero while as b approaches one, the term $b/(1-b)$, and thus the expected wait time, goes to infinity.

Third, note that expected waiting time goes up as the variability factor V goes up. Recall from statistics that the coefficient of variation is the standard deviation divided by the mean. The variability factor, V, is defined as $(c_a^2 + c_s^2)/2$ where c_a and c_s are the coefficients of variation for inter-arrival times and service times, respectively. In other words, higher variability in arrivals increases wait time, as does higher variability in services, and the equation suggests they are equally "bad" (that is, c_a and c_s have the same impact on wait time).

Finally, note that as n increases, for any given value of b the term in the original equation $b^{\sqrt{2(n+1)}-1}/((1-b)n)$ decreases. This may not be obvious at first glance so we will come back to illustrate this later, in Figure 4.

Using the above formula for the expected wait time in a queue we thus have verified our earlier claim that there were four key factors that determine the magnitude of the pooling benefit. To reduce expected wait time, you must either (1) reduce the service time, (2) reduce the utilization rate, (3) reduce variability, or (4) pool resources.

3.3. Waiting Time Comparison

We now use the above formula to compare the performance of the un-pooled and pooled systems described earlier. We assume for this analysis that c_a and c_s are both equal to 1. (This assumption is consistent with a commonly used distribution called the exponential distribution.) In the first scenario of §2, each server acts independently such that $n = 1$, and the arrival rate of one customer per minute along with a service time of 0.75 minutes leads to an average utilization, $b = 0.75$.

The expected waiting time in the queue is calculated as

$$W_q = s\left(\frac{b}{1-b}\right)\left(\frac{c_a^2 + c_s^2}{2}\right) = 0.75 \text{ min.}$$

$$\left(\frac{0.75}{1-0.75}\right)\left(\frac{1^2+1^2}{2}\right) = 2.25 \text{ minutes.}$$

With pooling, $n = 2$, and utilization b remains the same. Thus,

$$W_q = s\left(\frac{b^{\sqrt{2(n+1)}-1}}{(1-b)n}\right)\left(\frac{c_a^2 + c_s^2}{2}\right) = 0.75 \text{ min.}$$

$$\left(\frac{0.75^{\sqrt{2(2+1)}-1}}{(2)(1-0.75)}\right)\left(\frac{1^2+1^2}{2}\right) = 0.99 \text{ minutes.}$$

In this example, pooling has cut average waiting time (exclusive of processing time) from 2.25 minutes to 1 minute, or less than half of the original value.

3.4. Pooling Helps Mitigate the Curse of Variability

In both the single-server and two-server systems described above, customers had to wait in the queue because of variability in arrivals and services, both of which were said to be random. But pooling helped mitigate the negative effects of this variability.

To further illustrate the how the curse of variability can be offset by pooling, first consider a single-server system without pooling, similar to the one described previously. In the section above we defined a "variability factor," denoted by V, that takes into account variability in both arrivals and in services. A variability factor equal to one is often used to represent a baseline case, where both interarrival times and service times have a coefficient of variation equal to one. In such a system, as long as the server is hardly ever busy (that is, as long as utilization is near zero), jobs will rarely have to wait for service. But if the server is almost always busy (i.e., if utilization approaches one), then there will likely be a long line of jobs waiting for service and therefore a job will likely experience a long wait time (and hence a long throughput time). Figure 3 illustrates this effect for a system with a single server—the labels on the curves denote the variability factor. (The figure shows curves for variability factors of 0.1, 0.5, 1, and 2. The curves were created using the model given above.)

As variability increases, the curves in Figure 3 shift progressively upward, indicating that we expect more jobs to be waiting and we expect a job to experience a longer waiting time (and hence a longer throughput time). We refer to this as the curse of variability.

Pooling can help mitigate the curse of variability. The mitigating effect is illustrated in Figure 4. Here we show what happens when we pool identical single-server systems, each having a variability factor of one. The curve labeled with a "1" in Figure 4 is identical to the one labeled with a "1" in Figure 3, but in Figure 4 the curve labels refer to the number of

Figure 3 **The Expected Waiting Time in the Queue Increases with Variability (Data Are for an Independent System with One Server, and Numbers on the Curves Denote the Variability Factor)**

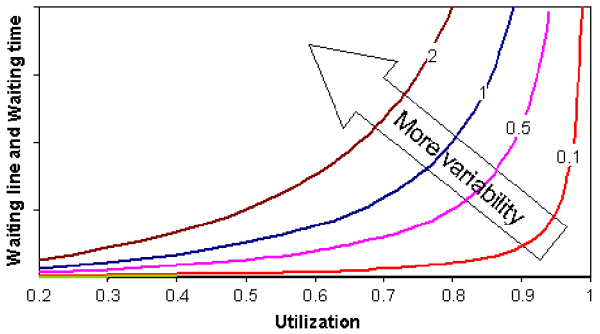
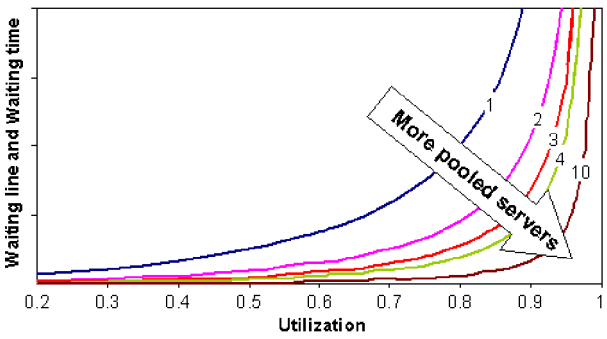


Figure 4 **The Expected Waiting Time in the Queue Is Reduced by Pooling (Data Are for a Variability Factor of One, and Numbers on the Curves Refer to the Number of Pooled Systems)**



pooled systems. For example, if ten identical systems having a variability factor of one are pooled together, then the expected system performance is as shown by the curve labeled with a "10" in Figure 4.

As more systems are pooled, the curves shift downward, indicating improved system performance. In other words, this shows how pooling has the potential to improve system performance without acquiring any additional resources (the individual resources are simply pooled together).

Note that there are diminishing returns to pooling. In other words, going from one to two servers improves system performance more dramatically than going from two to three, and so on. In other words, "a little flexibility goes a long way." In a system with many servers and many different types of demand, instead of training all servers to handle all types of demand, crosstraining some servers to handle a couple of different types of demand might achieve almost the same benefits.

Also note that, very roughly speaking, if you double the number of pooled servers in this setting, it achieves approximately the same benefit as halving the variability factor. This is evident from comparing Figures 3 and 4.

Summary

Benefits achieved from pooling represent a form of economies of scale. That is, if the single larger operation that results from pooling performs better than individual smaller operations that don't make use of pooling, then there exist economies of scale.

In some cases, the centralization of inventory may be virtual. By this we mean that the inventory need not be physically located all in the same place. For example, a group of car dealers may decide to share inventory such that each dealer can sell any car on any dealer's lot. Each dealer may have some local inventory to show to customers, but if the dealer does

not have exactly what the customer wants, then it might be able to get it from a neighboring dealer's lot.

In example 2, advantages were gained by pooling the resources of multiple servers. This is one reason why banks often funnel all arriving customers into a single queue, with the customer at the front of the queue progressing to the first available server. Another reason is that customers often perceive such a system to be fairer, in that it is first come, first served. However, other firms such as McDonalds still use individual queues in front of individual servers. At McDonalds, customers can quickly switch to a different line if they see it might be to their advantage. This mitigates the disadvantage of not using a pooled system, but may create some frustration and jockeying among customers. But some customers prefer this type of system because it gives customers the opportunity to try to choose the faster line.

There are opportunities to take advantage of pooling during product design, manufacturing, and marketing as well as during product distribution. For example, using common components across various models instead of unique components for each model can result in not only fewer part numbers to manage but also less total inventory of raw materials at the factory than in the non-pooled case. In addition to the statistical benefits which have been the focus of this note, pooling may allow for other economies of scale. In particular, if the per unit cost of a product or item decreases with volume, pooling arising from the use of a common part may allow an operation to capture these scale benefits.

Product postponement, involving postponement of the point at which a product fans out into its different varieties, is another technique that makes use of the pooling principle. The idea is to hold the product in an unfinished state (in a pre-finished state that is common to many end products) until it becomes more clear which exact end unit is demanded. The paint retailer example described above illustrates postponement in that the product is held in an unfinished (base color) state until a customer places an order, at which time the product is transformed to the desired color. On the sales side, if the salesperson can sell product version B when version A was the customer's first choice, and vice versa, then it is just as if the two product demands were pooled into one. These are just some of the many possible ways to create pooling efficiencies.

Remember, pooling isn't *always* better. With regard to resources, benefits are predicated on each resource being able to handle any demand. It may be expensive to acquire resources with this flexibility. For example, in some operations it may be costly (and sometimes virtually impossible) to cross train servers in a pooled system. On the other hand, scheduling of workers becomes easier when they are cross-trained and responding to server absences becomes easier as well. As another example noted above, pooling of inventory into a centralized warehouse may increase transportation costs. Virtual pooling of cars among dealers may create tension among dealers and possible lost sales because the customer doesn't immediately see the car offered for purchase. While our focus has been on the improvement in customer wait time, the other advantages and disadvantages should not be ignored. In each situation the advantages of pooling must be weighed against any possible disadvantages to determine the best operational policy for the firm.

Acknowledgments

The authors appreciate the many useful suggestions of the reviewers and editors, which led to significant improvements.

Reference

Hopp, Wallace, Mark Spearman. 2000. *Factory Physics*, 2nd ed. Irwin.

* * *

To reference this paper, please use:

Cattani, K. and G. M. Schmidt (2005), "The Pooling Principle," *INFORMS Transactions on Education*, Vol. 5, No 2, http://archive.ite.journal.informs.org/Vol5No2/CattaniSchmidt/.