Quantitative Module D

Queuing Models

**Background**

Waiting lines build because arrival and service rates are variable, and they *are not synchronized*. In particular, lines grow when short inter-arrival times are matched with long service times; however, conditions of long inter-arrival times matched with short service times produce idle time *that cannot be stored*. Queuing theory can be fun because we all can relate to the pain of waiting in lines. The models presented in this module are all manageable enough to provide good insight rather easily. Social psychologists have also studied the psychology of waiting. See Other Supplementary Material below for examples of how to manage the waiting process.

A nice primer for this topic could be a discussion of the burgeoning competition among hospitals over emergency room (ER) waiting times, and even the introduction of virtual ER queues! From Yoshino, Kimi, “Advertising emergency room wait times gains popularity,” *Los Angeles Times*, Dec. 21, 2009:

“... a growing number of suburban emergency rooms around the country are advertising wait times. Some post the times on their websites. Others tweet, send text messages or display the times on huge highway billboards. A few are testing a service by a start-up company, InQuickER, that allows patients to register online, pay a small fee and hold their place in line while they wait at home.”

**Class Discussion Ideas**

1. A good way to start the lecture can be to ask the students to share any particularly bad experiences that they have had waiting in lines. What should managers of those organizations have done differently? Then ask the students if any have had particularly “good” waiting experiences. In other words, was there something that management did to make the wait enjoyable or at least less annoying?

**Active Classroom Learning Exercises**

1. “Fast pass” at Disneyworld creates virtual lines by allowing customers to reserve a spot within a short window of time (maybe 10 minutes) with which they can show up at their attraction and essentially cut to the front of the actual line. Split the class into groups and have students in each group put themselves in the Disney managers’ shoes. How are the reservation times determined? (Groups should feel free to estimate arrival rates and service rates for a popular attraction.) What should the arrival rate in the fast-pass line be limited to in order to keep that actual wait very short? Qualitatively, what changes need to be made to implement such a system (e.g., building a second line, having a computer-based ticket reservation system, placing appropriate signs and reservation systems around the park)? What happens (or should happen) if someone shows up for his or her reservation 60 seconds late? Finally, (at last check), Disney only allows a guest to hold a reservation for one ride at a time. Why not more than one? (A guest that really planned this out well could ride every ride in the park, some more than once, without ever waiting in line.) Should Disney move to more like two or three reservations held simultaneously, as long as they are spaced out at least an hour apart? Have the groups share selected findings with the class.

**Cinematic Ticklers**

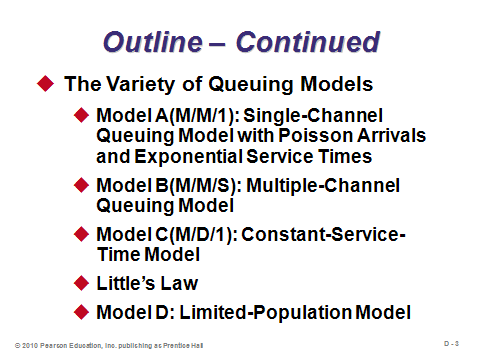
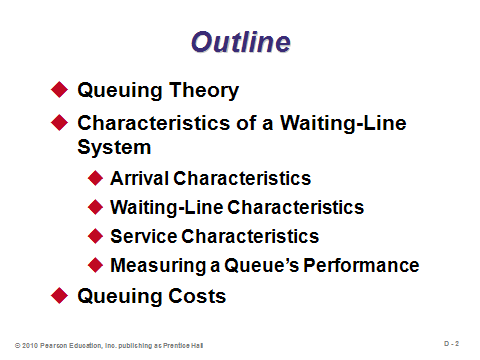
1. *The Simpsons, Season 2: “Brush with Greatness,”20th Century Fox Video, 2002 (1990-1991)*

At the beginning of the episode, there’s a huge line at one of the slides in the water park. Lisa and Bart cut to the front of the line when Lisa starts to fake a crying attack. Homer gets to the front of the line by claiming to be the “line inspector.” He subsequently gets stuck in the water tube due to his extreme girth, and a queue of riders builds up behind him inside the tube as kids keep crashing into each other.

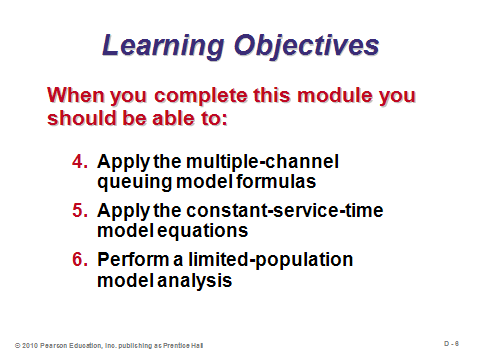
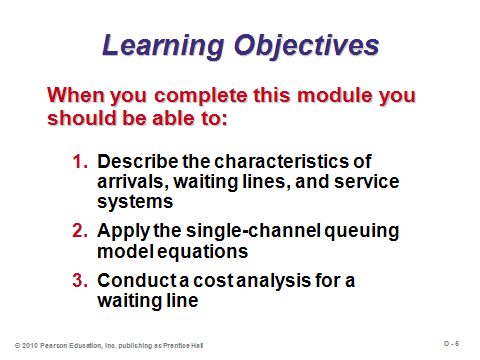
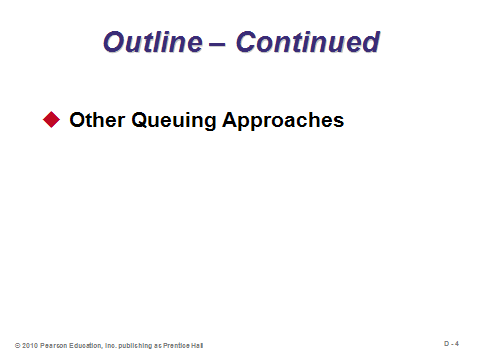
**Presentation Slides**

INTRODUCTION (D-1 through D-6)





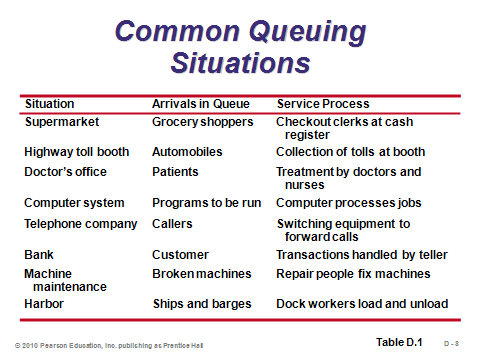
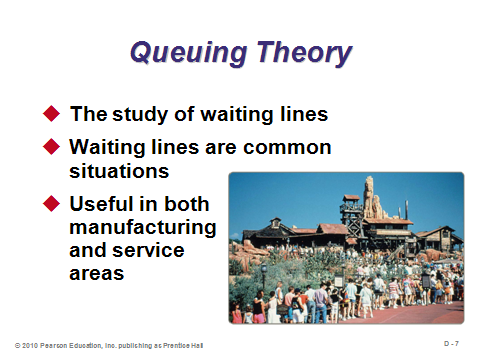
**D-1 D-2 D-3**



**D-4 D-5 D-6**

QUEUING THEORY (D-7 through D-8)

Slides 8-9: The English call a waiting line a *queue*, and queuing theory represents a significant body of knowledge, both quantitative and qualitative, about waiting lines. Slide 8 (Table D.1) identifies a few of the main potential applications of queuing models. *Most* operations must manage queues of some kind.



**D-7 D-8**

CHARACTERISTICS OF A WAITING-LINE SYSTEM (D-9 through D-20)

Slides 9, 11: Slide 9 identifies the three parts of a queuing system: (1) how inputs arrive, (2) how the queue works, and (3) how service is provided. These are illustrated in Slide 11 (Figure D.1).

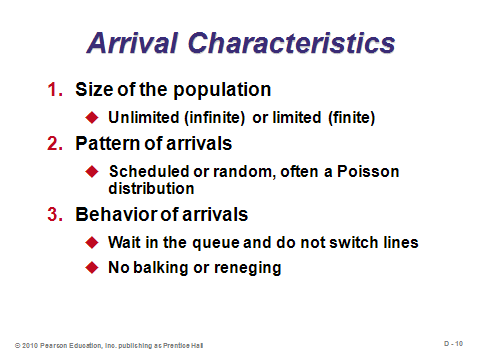
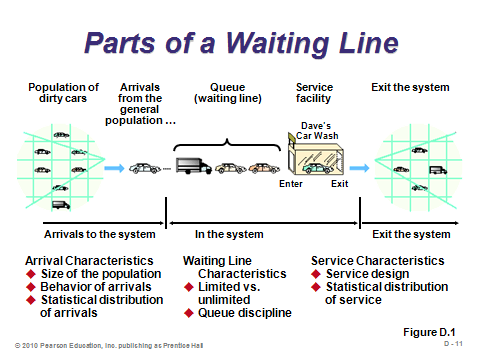
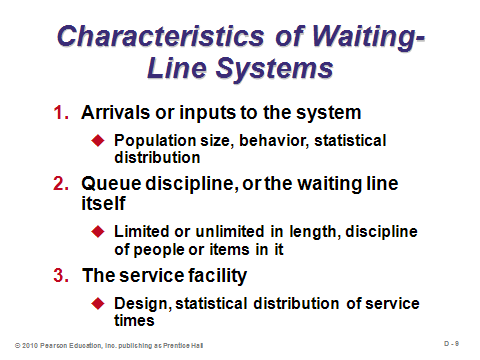
Slides 10-13: Slide 11 describes the arrival characteristics. When the number of customers or arrivals on hand at any given moment is just a small portion of all potential arrivals, the arrival population is considered *unlimited*, or *infinite*. Most queuing models make this assumption.

The behavior of arrivals can add interesting twists. *Balking* customers never join the line because it appears too long, and *reneging* customers leave a line because it is moving too slowly. Plus, some customers switch lines when they think that another line is moving faster (and we all know from experience that this tactic often backfires). These real-world behaviors are ignored by most queuing models, but they can be rather easily incorporated into computer simulation programs. Concerning arrival patterns, s*cheduled arrivals* would occur, say, in a doctor’s office where patients make appointments. Often, however, arrivals are considered to be random. The *Poisson* distribution (Slides 12-13) frequently describes the arrival pattern well. (If students have ever worked at a reception desk or at a checkout stand, they may have noticed that significant downtime occurs, followed by a bunch of customers arriving at the same time. Arrival patterns tend to be very lumpy (if they were smooth, queues would seldom form), and the Poisson distribution can represent such a process.)

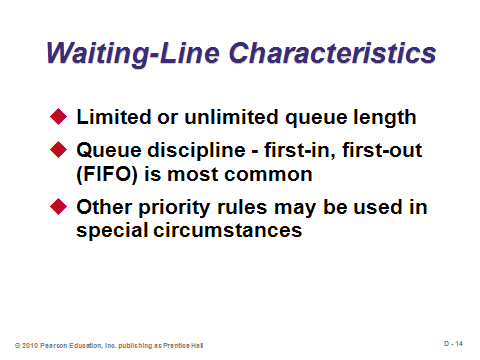
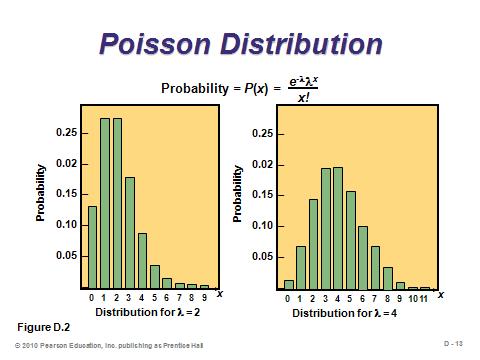
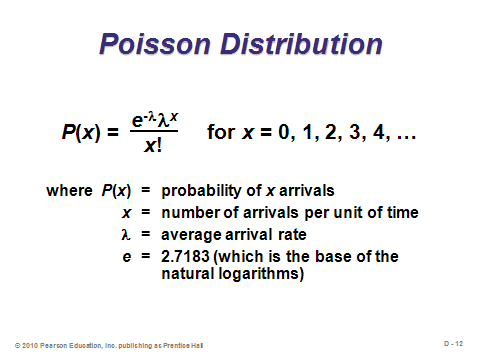
Slide 14: Usually, the most important consideration about waiting-line characteristics with respect to the queuing formulas is whether or not the size of the queue is limited.

Slides 15-19: The design of queuing systems specifies the number of channels (i.e. servers) and the number of phases (service stops). Slides 16-18 (Figure D.2) present four basic queuing system designs. Service times may be considered constant, but they are often random depending on the needs of the customer and/or the abilities of the server. The negative exponential distribution often represents service times well (Slide 19). (For example, the times at a bank teller window might average about 60 seconds for simple deposits and withdrawals, but every now and then a customer arrives who has a problem with a bank statement that will take time to resolve or who brings in bags of pennies that must be counted, etc.)

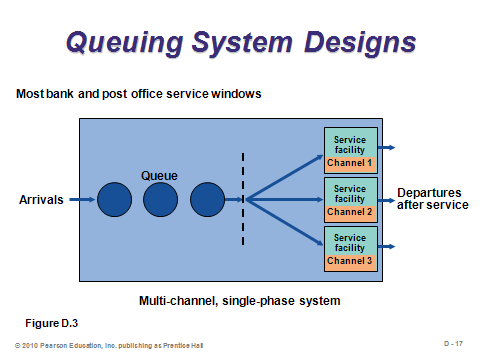
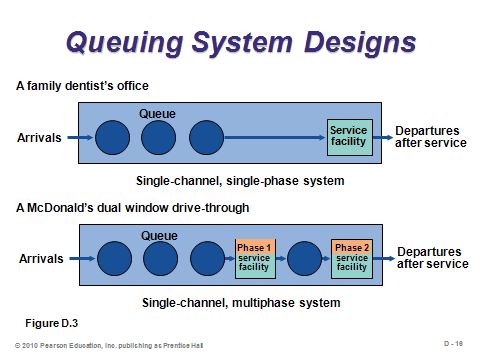
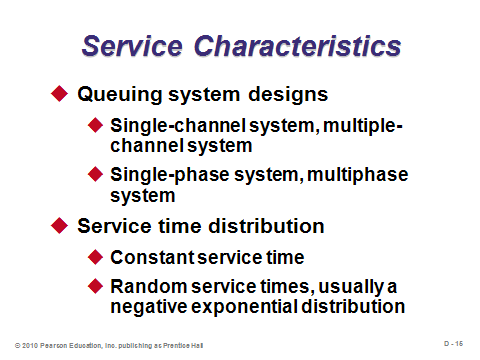
Slide 20: This slide identifies common performance measures for queues. These can all be computed by hand for the models presented in this module.



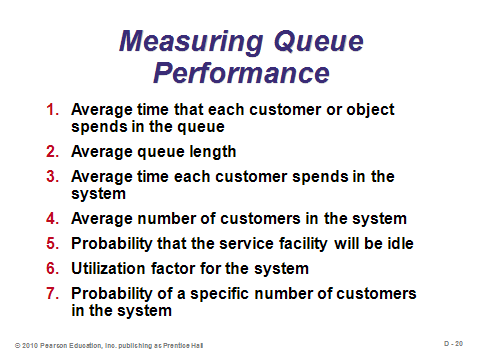
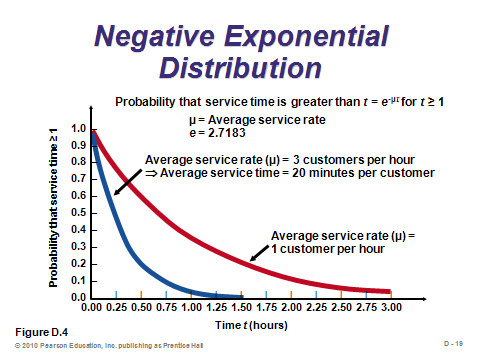
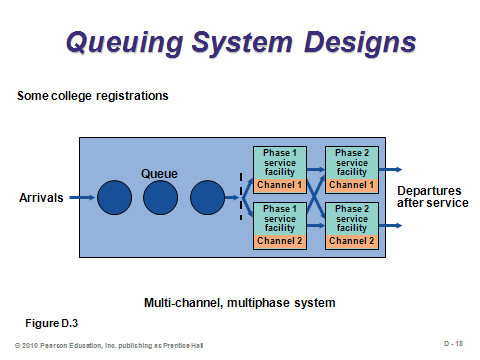
**D-9 D-10 D-11**



**D-12 D-13 D-14**



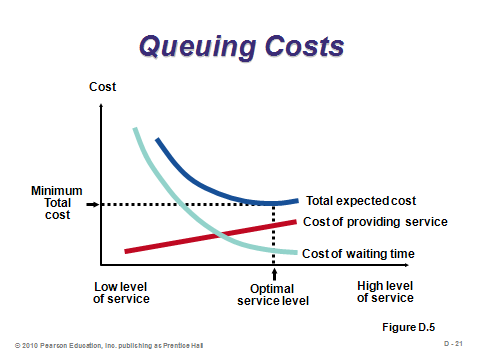
**D-15 D-16 D-17**



**D-18 D-19 D-20**

QUEUING COSTS (D-21)

Slide 21: This slide (Figure D.5) represents the trade-off between keeping customers happy by having short lines vs. the higher costs incurred trying to do so. Sometimes firms can use standby personnel and machines to pitch in during peaks in demand. To elaborate, instructors can ask the students what they think about the lines at the local Department of Motor Vehicles (DMV) (which are quite long in many towns). What value do they think the DMV manager places on their waiting time? In this case, it’s close to zero, so we can see how the graph in Slide 21 changes.

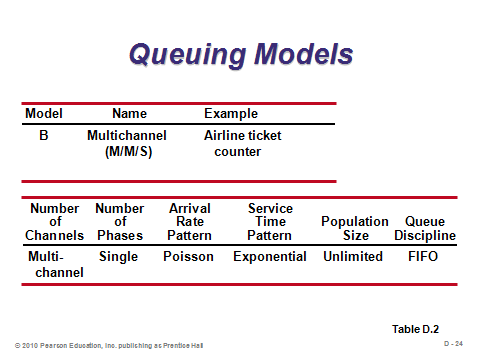
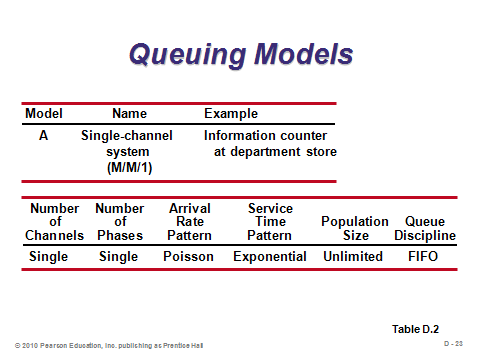
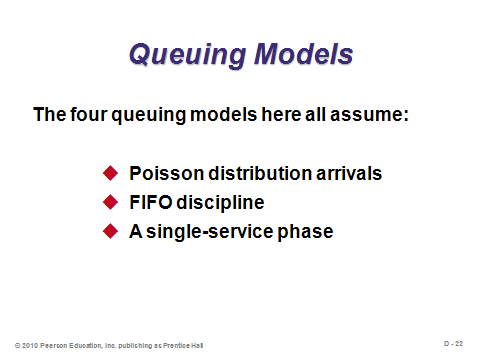


**D-21**

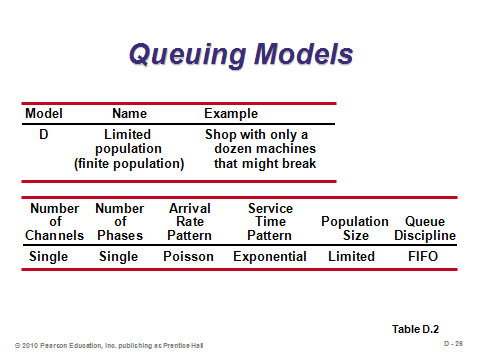
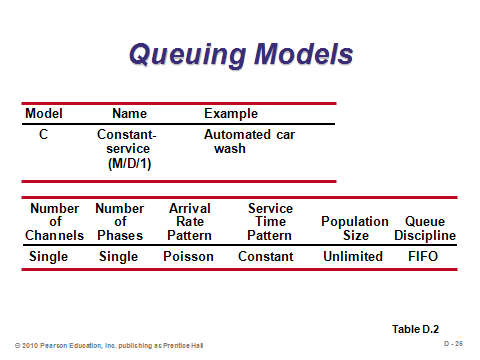
THE VARIETY OF QUEUING MODELS (D-22 through D-49)

Introductory Section (D-22 through D-26)

Slides 22-26: This module describes four important queuing models. Slide 22 identifies the assumptions common to each one. (They also assume patient customers, with no line switching, balking, or reneging. Also, the queue size is not limited.) Each of Slides 23-26 (Table D.2) describes the characteristics of one of the four models.



**D-22 D-23 D-24**



**D-25 D-26**

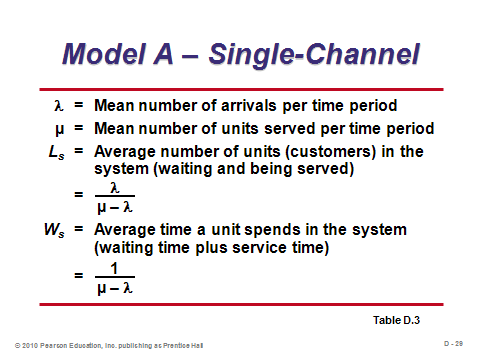
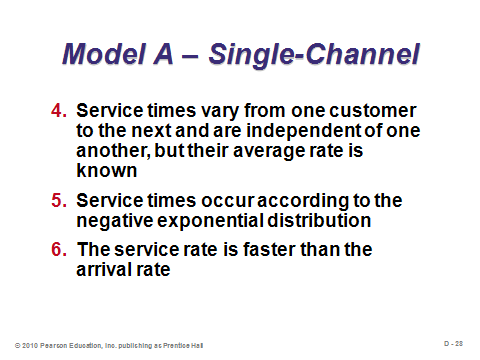
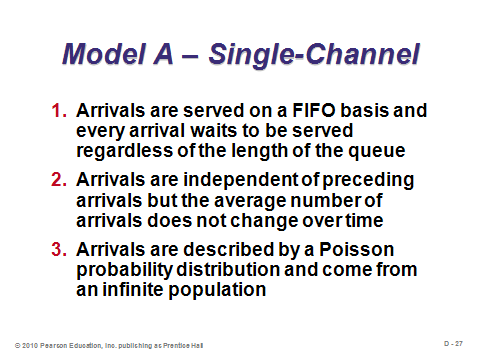
Model A (M/M/1): Single-Channel Queuing Model with Poisson Arrivals and Exponential Service Times (D-27 through D-35)

Slides 27-28: Model A represents the most common queuing problem with one line, one server, and random arrival and service times. These slides identify the specific assumptions. Note that assumption 6 is vital for ensuring that the system does not “blow up,” i.e., that the line does not keep growing forever.

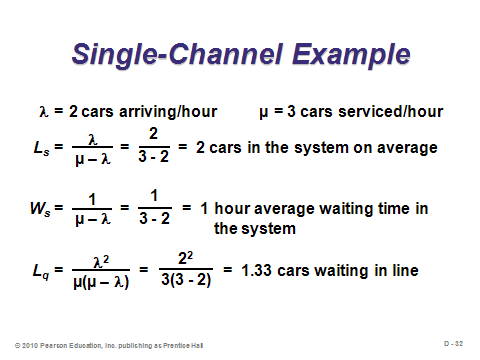
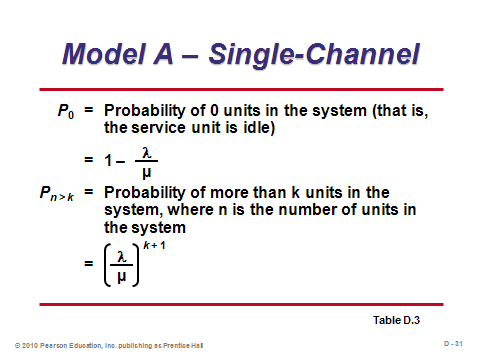
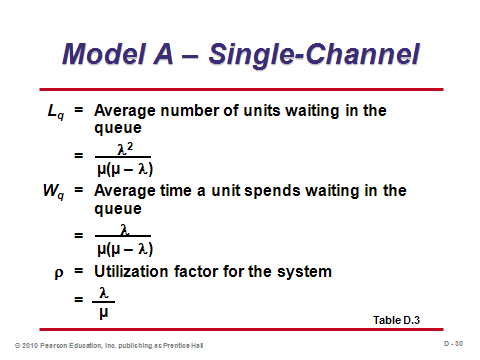
Slides 29-31: These slides present the Model A formulas, which can all be easily computed by hand. Notice that λ and μ are expressed as *rates*, not times. Some software programs ask for inputs as times, so students need to make sure that they express these parameters appropriately.

Slides 32-34: These slides present a Model A example (Example D1). Instructors may note a couple of relationships. For example, the average time in the system equals the average time in the queue plus the average service time 1/μ. Similarly, the average number of units in the system equals the average number of units in the queue plus the average number of people being served (which equals the utilization rate ρ).

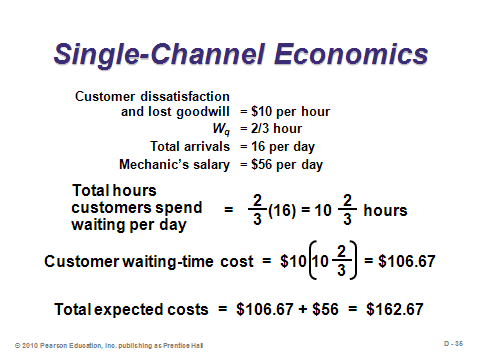
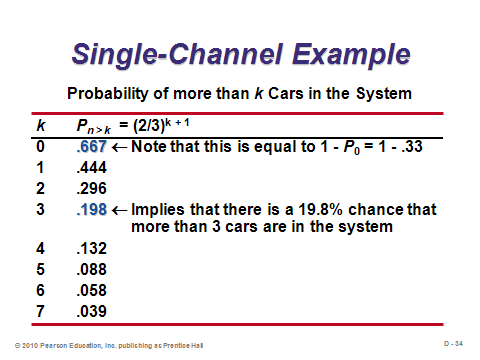
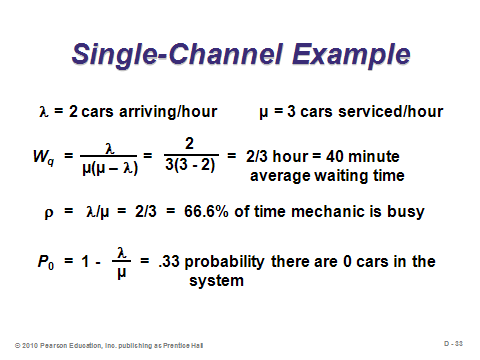
Slide 35: This slide (Example D2) extends Example D1 to compute the total costs of customer waiting plus labor. The most difficult part of such a calculation is trying to get a good estimate of what each hour (or minute) of waiting costs in terms of dissatisfaction and lost *goodwill*.



**D-27 D-28 D-29**



**D-30 D-31 D-32**



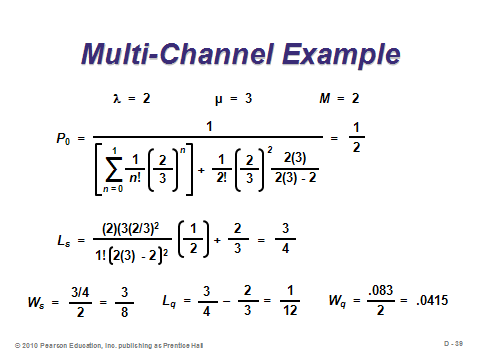
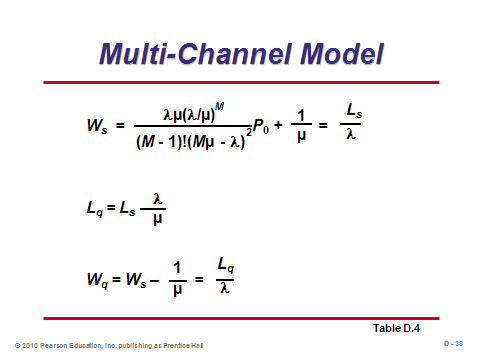
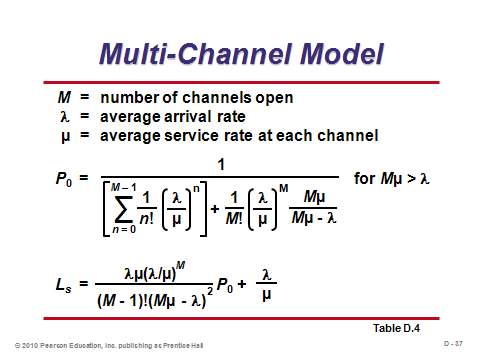
**D-33 D-34 D-35**

Model B (M/M/S): Multiple-Channel Queuing Model (D-36 through D-41)

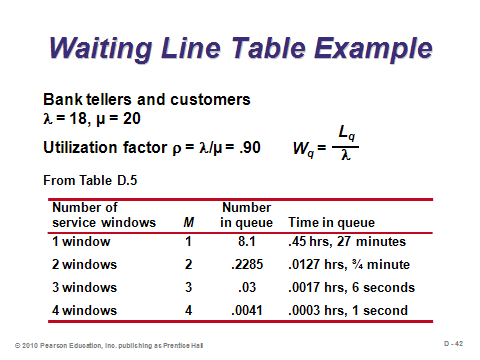
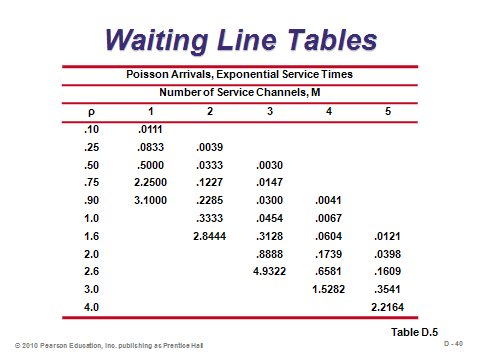
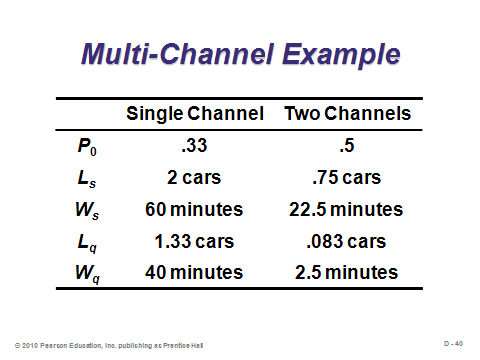
Slides 36-37: The multiple-channel model describes many queues at airline check-in counters, banks, and post offices, where customers stand in one long line that feeds into multiple servers (as opposed to most grocery stores, which have multiple servers but a line in front of each server). Queuing theory helps us understand that the single-line design (airline) performs much better than the multi-line design (grocery) for the same number of servers. It also seems fairer to customers. The formulas are presented in Slides 36-37. While the calculations can be done manually, they become rather onerous when the number of servers exceeds two. Software such as Excel OM and POM for Windows can be quite helpful for analyzing Model B.

Slides 38-39: Slide 38 (Example D3) extends Example D1 to add a second server. Slide 39 compares the single-line example with two servers (Slide 38) to the same system with one server (Slides 32-33). Obviously, all of the queuing measures improve when more servers are added. (To compare Slide 38 to a system with two servers but a separate line in front of each server, use Model A dividing λ by 2 for each line and adding the length of both lines together. Solution: *P*0 = .33, *Ls* = 1 car, *Ws* = 30 minutes, *Lq* = .33 cars, and *Wq* = 10 minutes. So the waiting and line length measures are all better than with just one server, but worse than if both servers were fed by one line.)

Slides 40-41: Because Model B can be so onerous to compute by hand, a table (D.5) can be used to retrieve the value of *Lq* given ρ and the number of channels. Slide 40 presents excerpts from that table. Slide 41 applies that table in an example (Example D4).



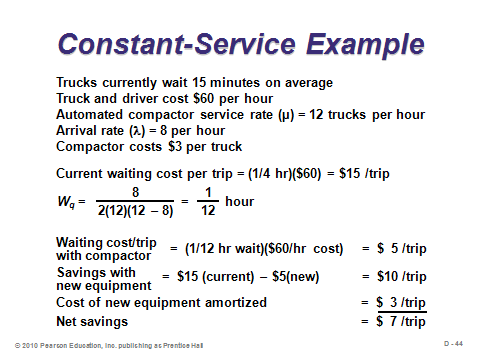
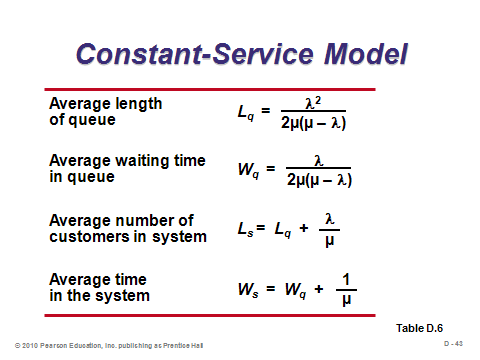
**D-36 D-37 D-38**



**D-39 D-40 D-41**

Model C (M/D/1): Constant-Service-Time Model (D-42 through D-43)

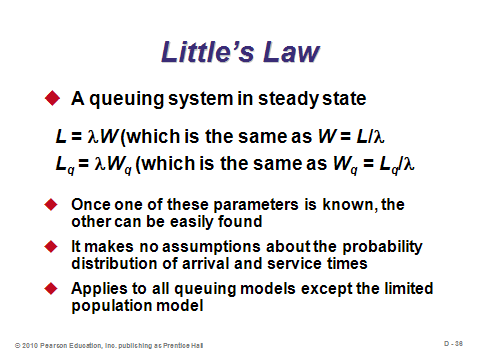
Slides 42-43: When service times are constant, the output measures improve (as do most phenomena when variability is reduced). In fact, both the average queue length and the average waiting time in the queue are halved as compared to Model A. Slide 42 presents the formulas, and Slide 43 applies them to an example (Example D5).



**D-42 D-43**

Little’s Law (D-44)

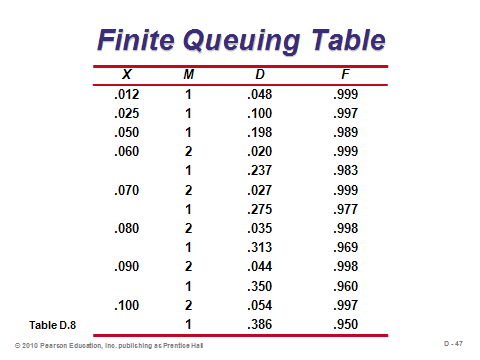
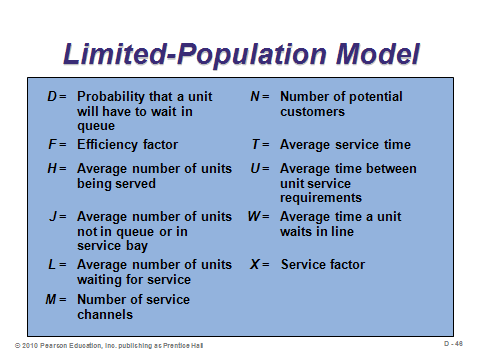
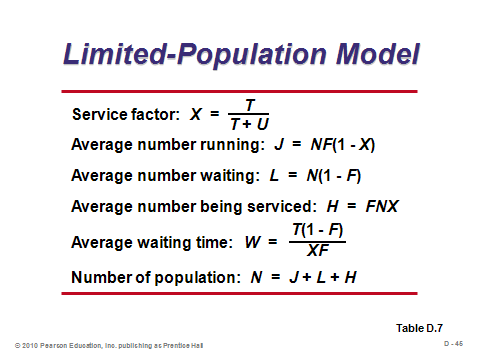
Slide 44: Little’s Law states that in a steady-state system, inventory equals flow rate multiplied by flow time. It represents an excellent way to analyze process flows in an operation. As it relates to queuing theory, number in the queue is inventory, arrival rate is flow rate, and waiting time is flow time. Thus, *L* = λ*W*, and *Lq* = λ*Wq*. (This can be verified by manipulating the associated Model A formulas.) Little’s Law is particularly helpful in queuing theory because it applies for any type of probability distribution.



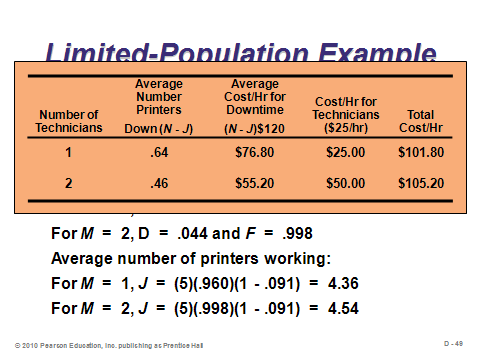
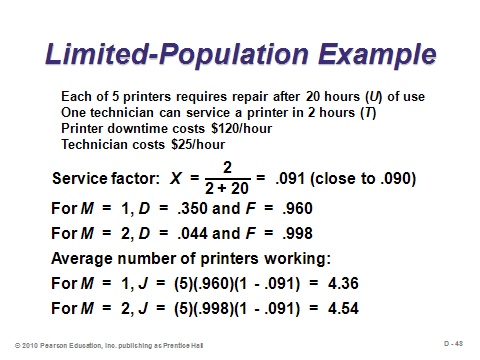
**D-44**

Model D: Limited-Population Model (D-45 through D-49)

Slides 45-49: Model D is somewhat more involved because there is now a dependent relationship between the length of the queue and the arrival rate. In particular, as the waiting line lengthens, the arrival rate drops. Different notation is used for Model D (Slide 46). Slide 45 presents the formulas. Slide 47 presents an excerpt from Table D.8, which helps to determine the values of *D* and *F*, given *X* (based on *T* and *U*) and the number of channels. Slide 48 presents an example of Model D (Example D6), and Slide 49 presents the associated cost analysis.



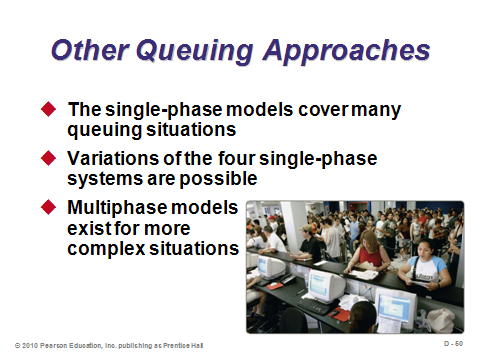
**D-45 D-46 D-47**



**D-48 D-49**

OTHER QUEUING APPROACHES (D-50)

Slide 50: Queuing theory models can become very complicated for analyzing systems with different assumptions than those of the four models described in this module. Computer simulation (Module F) is often used in place of mathematical queuing models when the situation becomes complicated enough.



**D-50**

**Additional Assignment Ideas**

1. Go to a grocery store with multiple checkout lines and observe them for one hour. For one of the regular (not “express”) lines, record the arrival rate as the service times. Also, across all of the lines, make note of any reneging, balking, or line jumping that you observe. First, write a short report about what you have seen. Qualitatively, how did these lines behave? How long did the lines grow? Did any customers seem dissatisfied with the length of their respective wait? Did the store seem to have enough servers available for this time of day? Second, use Model A in this module to estimate output measures for the line that you timed. Third, multiply the arrival rate for your line by the number of regular lines that were in operation at that time. Use Model B to estimate the average waiting time if the store had been set up with one big feeder line instead of a separate line for each register. The waiting time estimates should improve. Why do you think that the store has not set up the checkout process with one big feeder line?

2. Visit the Web sites of any simulation software product and describe how that product can be used to model queuing systems. Note that some software vendors have included what used to be discrete queuing applications in their larger simulation applications. Here one example:

* Queuing Model from Bizpep Business Support Software: [http://www.bizpep.com/queuingmodel.html](http://bizpep.com/queuingmodel.html)

**Additional Case Studies**

Internet Case Study (www.pearsonglobaleditions.com/heizer)

* *Pantry Shopper*: This case requires the redesign of a checkout system for a supermarket.

Richard Ivey School of Business (http://cases.ivey.uwo.ca/cases/pages/home.aspx)

* *Café D. Pownd* (9B00D007): An assistant manager of a university student residence is aware that there are capacity and service problems in the cafeteria. Long waits in line were common, and he hoped to propose some improvements to residence management, preferably ones with no major investments or disbursements involved.

**Other Supplementary Material**

Managing the Waiting Process

Once management has determined how to set up the queues and service stations, lines with certain lengths will form. Given that people have to wait, social psychologists have helped to identify ways to reduce the perceived costs of waiting. Examples are provided below.

* Make waiting more comfortable (chairs; air conditioning; access to a beverage bar; pagers that notify customers when their table is ready so they can wander around in the meantime).
* Distract customers’ attention (play a video that can be viewed from the line; place mirrors on the wall—one famous studied found that people believe that they wait for a shorter period of time for elevators when mirrors are placed next to the elevators, appealing to human vanity to some extent).
* Start the service early (fulfill customer drink orders while they wait; let them look at the menu while in line—this also may speed up the actual service, which may produce faster customer turnover).
* Explain the reasons for the wait (uncertainty creates anxiety—e.g., “there is ice on the wings of the airplane”; “the pilots are arriving on another plane”).
* Willingness to wait is somewhat proportional to service time (doctors should not conduct a three-minute examination for a patient who has waited for an hour).
* Provide a pessimistic estimate so that customers become pleasantly surprised (look at your watch when reaching one of those time estimate signs at Disneyworld and see if it actually takes that long; *Star Trek* fans can recall Chief Engineer Scotty’s dire predictions that he somehow always beat—he even admitted to this tactic in one of the movies).
* Don’t make promises that you can’t keep (forget about optimistic predictions).
* Compensate for any extraordinary waiting (free drinks, coupons).
* Perhaps most importantly, be fair! (waiting is sometimes a fact of life, but we get quite upset when somebody else cuts in line).