THE WALL STREET JOURNAL.

http://www.wsj.com/articles/the-science-of-standing-in-line-1475850601

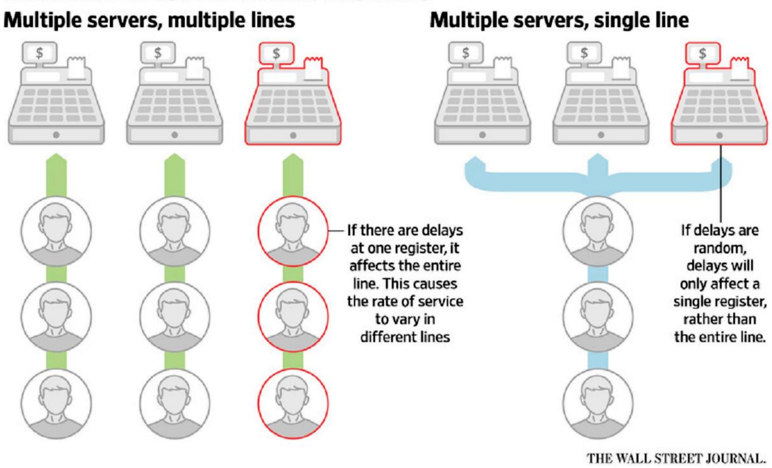
U.S. | THE NUMBERS

The Science of Standing in Line

Many service providers use queuing theory, or the mathematical study of lines, to manage their customers' waiting times.

The Line Dance

Queuing theory, the mathematical study of lines, helps businesses, call centers, computer networks and others figure out how to keep things moving.



By JO CRAVEN MCGINTY

Oct. 7, 2016 10:30 a.m. ET

You've probably participated in this familiar dance: Given a choice of checkout lines, you've somehow picked the slowest.

You could wait it out. You could chassé to another queue. Or you could bail out altogether. After all, no one likes to wait. But are the other lines really faster?

When parallel lines feed multiple cashiers, you may not be in the slowest one, but chances are, you also are not in the fastest.

Bill Hammack, a professor at the University of Illinois at Urbana-Champaign and YouTube's "Engineer Guy," explained it like this:

Imagine three lines feeding three cash registers. Some shoppers will have more items than others, or there may be a delay for something like a price check. The rate of service in the different lines will tend to vary. If the delays are random, there are six ways three lines could be ordered from fastest to slowest—1-2-3, 1-3-2, 2-1-3, 2-3-1, 3-1-2 or 3-2-1. Any one of the three (including the one you picked) is quickest in only two of the permutations, or one-third of the time.

There are two sources of variability, according to Linda V. Green, a professor of health-care management at the Columbia University Business School who specializes in mathematical models of service systems.

"When the demands for service come in and how long it takes the servers to process them," she said. "That inevitably causes temporary mismatches between supply and demand and hence backups, delays and congestion."

Luckily, most service providers take steps to manage the wait. A supermarket with parallel lines reserves some registers for customers with fewer items. Airline security uses a single serpentine line to feed multiple agents to mitigate bottlenecks at individual checkpoints. Emergency rooms and 911 dispatchers give priority to those whose needs are most urgent.

Each approach is based on queuing theory, or the mathematical study of lines.

Queues can be trivial, like a line at an ATM, or they can be serious, like a list of people waiting for an organ transplant, said Richard Larson, director of Massachusetts Institute of Technology's Center for Engineering Systems and an expert on queuing, but the fundamentals are the same: A basic queue funnels clients demanding service to one or more servers who respond. If the servers are busy, other demands must wait.

The clients may include a line of people, a series of 911 calls, or a string of commands issued over a computer network (think of a printer queue). The servers are the cashiers, the dispatchers or the devices that respond.

Queuing theory helps untangle the mess of requests, or at least smooth it out, by estimating the number of servers needed to meet demand over a given period and designing rules for advancing the queue.

The best system depends on the situation. “First come, first served” is most familiar, and people often prefer it because it seems fair. But most also accept that a heart attack should take precedence over a sprained ankle or someone with five items shouldn’t have to wait behind a procession of brimming shopping carts.

Queuing theory originated in 1908 when a Danish scientist named Agner Krarup Erlang went to work for the Copenhagen Telephone Company and set about trying to determine how many telephone trunks, or lines, were needed to serve callers.

The company could have provided a trunk for each telephone, Dr. Hammack said, but that would have been wasteful because not everyone calls at the same time. It could have provided enough trunks to handle the average number of calls, but too many would be blocked when the average was exceeded.

“Think of it like a bike wheel,” said Dr. Larson. “The switchboard is the hub. Each spoke is a wire coming in from a home to a human operator who can connect your wire to some other spoke in the wheel. What should the capacity be? If it’s too big, you’ll spend too much money on capital investments. If it’s too small customers won’t be able to be connected reliably.”

Customers waited in line to chec out at an Iea store in New Yor City in September. *PHOTO: MICHAEL NAGLE/BLOOMBERG NEWS*

Erlang's basic formula included three parameters: the number of trunks, the number of calls per hour and their average length. Adjusting the number of trunks in the formula raised or lowered the probability that a call would be blocked during the busiest hour.

Since then, retail stores, banks, call centers, emergency rooms, manufacturing plants, computer networks and all variety of queuing environments have used Erlang's formulas or related models to figure out how to manage their lines.

Meanwhile, clients—at least the human variety—can't help but wonder if there isn't a way to game the system.

Raj Jain, a computer-science and engineering professor at Washington University in St. Louis takes an analytical approach: When he's in line, he times the service. "Then I count the number of people ahead of me, and I know how much time I am to wait."

Dr. Larson takes a different approach. "I just strike up a conversation with an adjacent queue dweller," he said, "and wait."

Write to Jo Craven McGinty at Jo.McGinty@wsj.com