ECE459: Programming for Performance Winter 2023

Lecture 33 — More Advanced Queueing Theory

Jeff Zarnett *2023-03-27*

New Considerations

There are a few new considerations, or, if you prefer, complications to queueing theory that we haven't yet covered in the fairly simple discussion of queueing theory we have had so far. But real life is considerably more complicated than a simple model. We'll talk about two settings, one that I love and one that I hate: food halls and Service Ontario. No points for guessing which is which.

If you're not familiar with them, a quick refresher. A food hall, or food court, is a place where there are a number of counter-service restaurants that frequently offer different cuisines. Service Ontario is a place of interaction with the administrative state, specifically the services administered by the provincial government of Ontario.

Multiple Services. The way we have discussed the idea of services is that all the offered services are the same, or at the very least, every server can deliver every kind of service. That's sometimes reasonable – at Service Ontario, there are many different services available (drivers license renewal, health card renewal, vehicle registration update, etc) and any one of the staff members there can help you with any one of those services.

There are, of course, other situations where an individual member of staff cannot provide all services and therefore you must queue for the specific thing that you want, even if the other queues are shorter or empty. A food hall works like this. The Mexican restaurant may be very popular because they have excellent tacos, resulting in a long queue for that particular place. There may be no queue at the Gelato stand, but that doesn't help; the Gelato place cannot provide tacos. And while both of these are food, you will probably agree that they're not (always) interchangeable. If you really want tacos, gelato won't do.

Maintenance, Planned and Otherwise. In real life, services will potentially have downtime, planned or otherwise, which can throw a wrench in things unexpectedly and sometimes that means closing the queue and sending requesters elsewhere. If the pizza place's oven breaks, they can't make any more pizza, and everyone in line for the pizza needs to go somewhere else if they want to eat today. We don't usually account for this directly when planning and designing our systems, possibly because we expect things to always be available and for people to be paged to fix it if it's not (see our next topic for more about this). But we might account for regular maintenance or planned downtime in our estimates of how many customers we can serve in a given period.

In principle, new services can also come online also. I don't mean more workers at Service Ontario, but more like what happens if a new restaurant opens. If new a Banh Mi place opens then people who are currently in line for a different food might prefer to switch because they like that cuisine better. In reality, like a restaurant opening, the presence of a new kind of service is rarely a true surprise: people have worked hard to build, test, and deploy it and it's not as though we just wake up one day and suddenly a new thing is just... there.

Interchangeability. We already hit on the idea of interchangeability when talking about the food hall example. It's a spectrum: sometimes things are totally interchangeable, sometimes partly, and sometimes not at all. In the food hall example, total interchangeability is what happens if you're just extremely hungry and you would be completely happy whether you had tacos, pizza, or shawarma. Partial interchangeability can happen when you have preferences but would accept something else: you want tacos but would choose shawarma under certain circumstances. And no interchangeability happens when you have your heart truly set on pizza and will accept nothing else.

Whether things are interchangeable depends on the nature of the services and the needs of the people requesting them. Just as there are different services in this more complicated world, requesters also behave differently.

All of the above food hall examples are about food, so there's at least a possibility of interchangeability because every option will have some nutritional value (in the sense of, it's edible food with calories in it; a nutritionist might say the gelato has no nutritional value because it's not a "healthy" choice). If you went to Service Ontario to renew your drivers' license, there are really no alternatives that get you the same result: a new health card just isn't the same thing.

Then there are the needs of the person requesting the thing: If someone is a vegetarian, they might be okay with tacos or pizza, but not shawarma. If they are a vegan, maybe the only option for them is the taco stand, no matter how good the pizza place is or how long the line for tacos is.

Too Long. While we're on the subject, when you're in the food hall and the length of a particular queue is too long, we may experience one of three behaviours: balking, reneging, or "loss".

Balking is what happens when you look at the line and you decide it's too long and there's no point in even getting into the line. That's very common: if I want to go to Service Ontario and it's so busy that there is a long line out the door, I'm not going to bother. I'll come back another time (if I can – sometimes it's important to do something before a deadline). In the food hall, if the queue for tacos is really long, I'll line up for the shawarma.

Reneging is what happens if I enter the queue for tacos, but before I get to the front of the queue, I give up (leave the queue). If it's taking too long and I'm really hungry (or just impatient), this could happen. It could also happen if I'm at Service Ontario in between things (e.g., on a lunch break) and if time runs out I need to go back to work.

Loss is what happens when service is refused because of capacity limits. This comes from the idea of the early telephone systems, where there were k circuits and if none of the k circuits was free at the time a person wanted to make a call, that call was just dropped (that is, the call fails) [HB13]. That makes it a M/M/k/k system, in queueing theory notation. In the food hall or Service Ontario example, this is what happens if there's a capacity limit to the building and I'll be turned away by security if the building is full. I'm not allowed to queue up, no matter how badly I'd like to.

In both cases where I choose to leave, there's an implicit or explicit estimation of the waiting time provided. When I choose not to enter the queue at all, it's because I've looked at the queue length, and maybe done some assessment of the service time, and calculated that the wait time exceeds my willingness to wait. It's also possible that the establishment is kind enough to put out signs that say that the expected waiting time from this point is X minutes. Not very common in a food hall or Service Ontario, but maybe the case in other places.

Obviously, my initial estimate can be wrong if I guessed incorrectly about the service time, or if the line is so long I can't see all of it and I don't get a good estimate. But another reason why my estimate might be wrong is if people can join the line ahead of me. Wait, what?

Priority. When I was last at Service Ontario, an elderly person with mobility restrictions showed up while I was waiting in line and that person was permitted to go to the front of the queue. It's sensible that priority would be given to this person – asking them to stand in line a very long time is not nice. But of course, when they go into the line ahead of me, it increases my wait time. This increase may result in my decision to leave the line as the time has increased to the point that I can no longer, or at least no longer wish to, wait.

Priorities for some groups over others are actually really interesting, because they have interesting effects and open up questions:

- How much, if any, does giving priority to one group over another help the group being given priority?
- How much, if any, does giving priority to one group disadvantage the group not being given priority?
- Can the priority system incentivize people to choose things that are less popular?
- Recognizing that if everyone has priority, nobody has priority, how many requests can have priority before all benefit is lost?

Laboratory Study with a Mouse

This section relies on a rather informative video by the channel Defunctland, about the history of the Disney Fastpass system [Per21]. The actual video contains a lot of discussion about the history of Walt Disney World and other things that are not relevant here. But there are a few interesting things we can learn from it. In the video, there's a lot of background info but also some explanation of the simulation that was used to evaluate the situation.

Simulation? Yes – at some point the queuing theory problem becomes so complex that our ability to reason about it and do the math with spreadsheets or Wolfram Alpha or hand calculations is a limiting factor. And a simulation allows us to validate our theories. But why is this system so complex that it requires a simulation?

If we abstract away some of the details of the Mouse and his friends, we're left with a system has many details that we need to concern ourselves with:

- Every customer (requester) is an independent agent, which implies:
  - They have different times of arrival at and departure from the park. With that said, most arrive early in the day and relatively few arrive later in the day
  - They have different preferences of what they want to do while there
  - They have different willingness to wait for the things they want to do
  - They may or may not be willing to come back another day
- The park has opening and closing times which implies:
  - Requests cannot be submitted before opening time
  - Requests cannot be submitted after closing time
  - Requests submitted too close to closing time may not be served before closing
- The park has different services that each have their own service rate and any of them could be down for maintenance (indepedently of any others)
- The services have a fixed maximum capacity: you cannot make more seats on the rides or run them faster to get more people through quicker

Relevance? An important observation here is that this model is fairly different from what we are used to talking about when thinking of servicing requests. When I go to Service Ontario, I want to renew my license as fast as possible and leave as fast as possible and hope not to return any time soon. If I am at the food hall, I may want to sample some number of different cuisines, but I'll eventually be full (or at least be unable to eat any more food) and then I'll leave. But in this model, the tourist goes to the Mouse Park and stays for some period of time, trying to do as much as you want to do. That doesn't sound like most of our software service scenarios. Is anything we're learning in this model applicable?

I'll argue yes. If I'm a person waiting in line for a specific ride, it's entirely irrelevant to me whether the 400 people in front of me have just arrived at the park, or if they've been here for hours and this is their tenth ride. Rides don't have to be completed in a specific order (in the sense that nothing bad happens if someone does them out of order).

You could just imagine that this model is functionally equivalent to one in which each person who goes on a ride immediately leaves the park, never to return, and is instantly replaced by another guest who has the same preferences about what to do. With more and more complex priority systems, this way of thinking about it might not hold up, but hopefully this argument is convincing.

User types. The simulation has a few different archetypes for the users that represent their different "personalities" based on the expected type of person who would attend the park. Some of them come frequently because they have an annual pass, so they can always come back another time if they prefer. Others are here today and today only and don't want to miss out. But without getting bogged down in the archetypes, the type of user decides (1) how long they will stay at the park, (2) when they will balk (what is their willingness to wait for a particular ride), and (3) what they want to do while they're here. The third point covers both their preferences of what rides they want to go on (in what priority sequence) and if they want to do other things in the park (things that aren't rides).

Priority Systems. To establish a baseline, the simulation has an option where there's no priority system: everyone is equal. Then there are two different kinds of priority systems which the Mouse calls 'FastPass" and "FastPass+" (I'm sure the developers wanted to call it FastPass^{++}).

In the no-priority system, everyone is equal. Wait times are just based on how popular things are and nobody gets to cut in line. It helps wait times to be predictable, because the line a person in will advance at a fairly steady rate. No priority may seem fair at a glance, but is it optimal? Let's see.

In the Fastpass system, instead of waiting in line, a person could get a little ticket that specifies a return time, and when they return they can get to the front of the line. This allows the people waiting to do something useful or fun in the meantime (and maybe profitable if you buy some food). That just makes it a virtual queue system, as far as we're concerned, and not really a priority lane. You're waiting in the queue just as long as you would otherwise, but you're not having to stand in the physical one doing nothing while you wait. This does allow you to potentially get some other stuff done (imagine you can get some gelato while waiting your turn for pizza). The Fastpass system is applied only to the (few) most popular things, so for all other attractions there is no fast lane.

In the FastPass^{++} system, in advance of going to the park (or on arrival), you get three priority passes to some specific attractions well in advance. The passes can be "sold out" if there is sufficient demand. This resembles making reservations more so than waiting in a virtual queue. As you can imagine, people try really hard to get the priority passes for the most popular things.

Regardless of whether FastPass or FastPass^{++} there are two queues for any individual service: the priority queue (or fast lane) and the standby queue (that's the one for everyone else). If a ride can seat n people at a time, the ratio of priority to standby is important. Under typical circumstances the ratio of priority to standby is something like 4 priority to every 1 standby guest. In times of big backlog in the express line, it can be 20:1 or even 100:1.

Results

So let's recap the results from the simulation [Per21]. We're looking at the results in the video, but... did you want to play around with the simulation yourself? `https://github.com/TouringPlans/shapeland` – it's python, but it's not super complicated code. The results are broken down into (1) standby waits, (2) overall waits, and (3) average number of rides experienced.

Standby Waits. Standby waits increase using Fastpass, but that's not super surprising: if some people can cut to the front of the line, it delays everyone else. The FastPass^{++} approach increases the wait times on every ride except the most popular one (even though it's a small amount), often by a significant amount. Okay, so we know that it makes stuff worse in standby, but not everyone is in the standby line. So what's the overall impact?

Overall Waits. With no priority system, if a guest tried to do everything once, the average wait time would be about 41 minutes, but because people prefer to do the most popular things, the wait is about 58 minutes (on average) in reality. That's the baseline.

With the Fastpass system, the average wait in standby if doing every ride is 48 minutes, but the actual wait time is more like 40 minutes, which is about 2/3 of the wait of the no-priority system. That sounds pretty good! But keep in mind here that the benefit here is coming from pushing people to less popular attractions because they

can do those while they're waiting in the queue for a more popular thing.

The FastPass++ solution raises the average standby wait for doing everything to 67 minutes but the typical wait per attraction is 42 minutes. Again, this is because the wait times are longer on standby. But, average times are reduced, because people are encouraged to go to less popular things to make use of those.

Average Rides Ah, you've probably figured out at this point that one of the real goals of these priority systems are to incentivize people to use the under-utilized attractions. The average number of rides without any priority system is 3.31; with Fastpass it's 3.77; and with FastPass++ it's 4.23.

The distribution with FastPass++ is quite uneven though: it increases the number of people who go on many rides, and also the number of people who go on very few. So now we have winners and losers in the system, whereas FastPass does not seem to have this same effect of increasing inequality: some people do many rides in the day, but a lot more people get zero or one.

Lessons and Limitations

So the bottom line is: giving priority passes doesn't have a big impact on the wait times for the most popular attractions, but it does encourage better utilization of the less-popular things. And more usage overall.

If we didn't speed up wait times for people on rides generally, only encouraged people to do other things too, is that improvement? The Mouse is in the business of entertainment. The megacorporation wants you to spend money with them (consume, obey) and most likely a guest will return and spend more money if they have enjoyed their visit. Is going on more rides more enjoyable than more time in queues?

Maybe! That's at least what I think. It probably make some intuitive sense that at one extreme end of the spectrum where I never have to wait at all, I'm extremely happy; at the other extreme end where I wait in line all day and never ride anything, I'm very unhappy. I'm also pretty unhappy if I don't get to do my favourite things even if other things are more available. Still, it seems like getting to do more rides is good for increasing guest happiness if most guests follow a similar evaluation process.

To my knowledge, the simulations didn't account for downtime during the day. There may be some maintenance baked into the estimates of the capacity, but I mean the kind of downtime where a particular ride just goes offline and won't be back that day. In a simple model of that scenario we could just send everyone in the queue back to making a decision about what to do (go to next attraction, go home, etc). But in reality people who has a FastPass for the ride that went down might get another FastPass as compensation, which may have weird downstream effects. Also, people who were in the standby queue for a long time may get a compensatory FastPass for another attraction too, which would cause some chaos by increasing the number of passes in circulation. That might be a fun extension of the simulation.

Inequality is increased even more in real life than in the simulation by the knowledge factor. People have learned the nuances of the system and those who know the secrets (or at least learn about enough of them) get to do more things. If you have mastered the system, you get to do a lot more rides (8–9!) than someone completely unaware of it who might only get 1–2 rides in. There are countless videos out there about how to take advantage of it.

A final lesson based on the nuances of the system is that the more complex your system is, the more opportunities there may be for people with expert knowledge to exploit it and benefit themselves at the expense of others. Unexpected user behaviour like that can really mess up your capacity planning and make the user experience for everyone else worse.

References

[HB13] Mor Harchol-Balter. *Performance Modeling and Design of Computer Systems*. Cambridge University Press, 2013.

[Per21] Kevin Perjurer. Disney's FastPass: A Complicated History, November 2021. Online; accessed 2022-11-29. URL: `https://www.youtube.com/watch?v=9yjZpBq1XBE`.