Queueing and System Modeling Notes

CMS 380 Simulation and Stochastic Modeling

Systems and Queues

A *system* is a collection of *resources*. Customers enter the system and follow a *routing* through the resources. Routings can be fixed (every customer visits the same set of resources in the same order), or probabilistic (customers may take randomly chosen paths through the system).

There are two main types of systems: open and closed.

- In an open system, customers arrive from the outside world according to some process, travel through the system, then exit back to the outside world and do not return.
- In a closed system, a fixed population of customers circulates continuously throughout the system. There are no outside arrivals or departures.

Open vs. closed. *Think about some kinds of systems that would be best modeled as open and some that would best be modeled as closed.*

The majority of our analysis will deal with resources that experience contention. The simplest such model is the *single-server queue*.

- The queue has a single waiting line and a single service center.
- A customer arrives at the queue, waits for its turn to be served, receives its required amount of service, then departs.
- We'll assume, for now, that service is always in first-come-first-served order.

The queue is a statistical model. Its behavior is governed by two distributions:

- The distribution of the time between arrivals, called the *interarrival time distribution*. In particular, it matters a great deal whether arrivals tend to be roughly evenly spaced, or tend to clump together and occur in bursts.
- The distribution of the service times. Again, it matters a lot whether service times are almost the same for every customer or if there is a large amount of variability in customer's service requirements.

The interaction of arrivals and service times determines the behavior of the queueing system.

Examples of arrival rates*. Think of some examples of systems with bursty arrival rates, where customers tend to arrive in big clumps with long gaps between clumps. Think about some systems where arrival rates are nearly deterministic.*

Important Measures

Queueing Parameters. The most important queueing parameters are

- $\overline{s}$, the average service time
- $\mu = 1/\overline{s}$, the average service rate, which is useful in some contexts
- σ_s^2, the variance in the service times
- λ, the arrival rate

Residence and Waiting Times. The most important performance measure in a queueing system is the average time spent in the system. Call this the *residence time*, denoted by $\overline{R}$. The residence time is the total amount of time an average customer spends in the system, **including both the *waiting time* and the *service time***.

Let $\overline{W}$ denote the average waiting time: the time spent waiting in line, but not being served. Let $\overline{s}$ denote the average time spent receiving service.

$$\overline{R} = \overline{W} + \overline{s}$$

In grocery store terms, $\overline{s}$ is the time required to check out your own groceries. $\overline{W}$ is the time you spend waiting before you get your turn to check out. Residence time is the combination of both.

Fundamental Laws

The Conservation Law

What goes in must (normally) come out.

The system receives arrivals at a rate λ.

Define Λ to be the *system throughput*, the rate at which customers depart from the system. Define the *service rate* to be the maximum rate at which customers can exit the system:

$$\mu = \frac{1}{\overline{s}}$$

Over the long run, if the system is stable, $\Lambda \leq \mu$, because customers cannot depart the queue faster than they can be served. In addition, $\Lambda \leq \lambda$ because customers can't depart the system faster than they arrive.

If $\Lambda = \lambda$, the system is **stable**. If $\Lambda < \lambda$, the system is **overloaded**. Due to variability in the arrival process, there will always be short bursts where $\Lambda < \lambda$, but these even out over time. We will assume that our queues always have enough buffer space to hold any customers that arrive during period of temporary overload. If the system is permanently overloaded, the number of customers in the system grows to infinity.

The interaction between throughput and residence time, Λ and $\overline{R}$, can be complex. The two parameters are not the same thing!

- As individual customers, we would like $\overline{R}$ to be as low as possible, so that we spend as little time waiting as possible.
- The owner or maintainer of a system often cares more about throughput, and having the capacity to provide acceptable service to as many customers as possible for minimum cost.

It's possible to describe systems that violate the Conservation Law, by creating new customers within the system (called *forking*) or destroying customers (*joining*), but we're not going to work with any of those models.

The Utilization Law

Define U to be the *utilization*, the fraction of time the system is busy. In a stable system, $0 < U < 1$. Utilization turns out to be an important measure, because it quantifies the load on the system independent of any particular values of λ and $\overline{s}$. Almost all queueing models have utilization as one of their key parameters.

The relationship between arrival rate, service time, and utilization is given by the *Utilization Law*:

$$U = \lambda\overline{s}$$

A new customer arrives to a system every two minutes on average. Each customer needs an average of 1.5 minutes of service. What is the utilization of the system?

The arrival rate is

$$\lambda = \frac{1 \text{ customer}}{2 \text{ minutes}}$$

and $\overline{s} = 1.5$. By the Utilization Law,

$$U = .5 \cdot 1.5 = .75$$

Note that utilization doesn't have units.

The Forced Flow Law

The Forced-Flow Law (FFL) relates throughputs at individual resources within a system to the overall system throughput.

Suppose the overall system has arrival rate λ and output rate Λ, and the kth resource has input rate λ_k and output rate Λ_k.

Let V_k be the average **visit count** of resource k, representing the average number of times each customer in the system arrives to resource k. If $V_k > 1$, then customers, on average, visit resource k multiple times. If $V_k < 1$, then only some customers visit resource k and others never visit it.

The Forced-Flow Law:

$$\lambda_k = V_k \lambda$$

The average arrival rate to resource k is the total system arrival rate times the expected number of visits made to resource k. Because of the Conservation Law, we could also state the FFL in terms of output rates, Λ and Λ_k.

Proof. The FFL is so intuitive it almost doesn't require a proof. Nonetheless, here is a measurement-based argument for its validity. Suppose we measure a system over a period of time T and record C completions at the system as whole and C_k completions at resource k. The following relationships hold:

$$\Lambda = \frac{C}{T} \qquad \Lambda_k = \frac{C_k}{T} \qquad V_k = \frac{C_k}{C}$$

Assembling everything together, we have

$$\frac{C_k}{T} = \frac{C_k}{C} \cdot \frac{C}{T}$$
$$\Lambda_k = V_k \Lambda$$

This type of measurement-based reasoning is arguably not a "real" proof because we've been very vague about the nature of the interval under measurement or how we defined completions. Nonetheless, this type of thinking is potentially more useful than a rigorous proof for understanding system dynamics.

Little's Result

Summary

Little's Result (sometimes called *Little's Law*) is the most important equation in systems analysis. What $F = ma$ is to physics, Little's Result is to queues and other systems. It's named after John D. Little, who published the first paper formally stating and proving it in the 1950's.

Little's Result relates three important system quantities:

- Λ, the system throughput
- $\overline{R}$, the average residence time of a customer in the system
- $\overline{N}$, the average number resident in the system (waiting and in service)

Here's the equation:

$$\overline{N} = \Lambda \overline{R}$$

For example, if a CS department awards 20 bachelor's degrees every year ($\Lambda = 20$) and it takes the average student four years to earn the degree ($\overline{R} = 4$), we'd expect 80 students to be active in the program.

Little's Result is powerful for two reasons.

1. It's an extremely general result. It makes no assumptions about how arrivals occur, how customers accumulate service, or any other statistical properties of the system. Little's Result applies to any system.
2. We can apply the equation to any part of a system. As we go through this course, we'll use Little's Result to analyze entire systems of queues, individual queues, groups of queues within a system, and even parts of a single queue. The same relationship between throughput, occupancy, and residence time holds in all parts of a system.

Examples

Router. *We measured a network router for 30 minutes.* 3×10^6 *packets were sent during the interval and the average number of packets waiting in the router's queue was 3, including the packet in service. Estimate the average residence time spent by a packet at this router.*

This is a straightforward application of the Result. Given two quantities, find the third. The throughput is the number of packets sent per unit time:

$$\Lambda = \frac{3 \times 10^6 \text{ packets}}{30 \cdot 60 \text{ seconds}} \approx 1666.66 \, \frac{\text{packets}}{\text{second}}$$

The average number in the system is $\overline{N} = 3$. Therefore,

$$\overline{R} = \frac{3 \text{ customers}}{1666.66 \, \frac{customers}{second}} \approx .0018 \text{ seconds}$$

Hogwarts. *How many students go to Hogwarts? J.K. Rowling could never make up her mind. Suppose 40 students enter in each class and remain at school for seven years.*

Again, a straightforward application of the law. Multiply to get 280 students.

What if 40 students enter each year, but 20% of students leave school after taking their fifth year exams? The remaining 80% stay for all seven years.

The thoughput is still $\Lambda = 40$, but the average residence time must now account for students who leave early:

$$\overline{R} = (.20)5 + (.80)7 = 6.6$$

Using the updated value yields $\overline{N} = 264$ students.

Applying Little's Result to the Waiting Line. *A system has a throughput of* $\Lambda = 1$ *customer per minute, an average service time of* $\overline{s} = .5$ *minutes and an average residence time of* $\overline{R} = 2$ *minutes. Estimate the expected number of customers waiting in the line but not being served at a random moment in time.*

This problem seems daunting, but it's really a straightforward application of Little's Result. Here, rather than applying the Result to the entire system, we're going to apply it to only *part* of the system: the waiting line, not including the server.

$$\text{average number in the waiting line} = \Lambda \cdot \text{average time spent waiting}$$

The average time spent waiting is the total time in the system, not including the service time (this is the definition of waiting time):

$$\overline{W} = \overline{R} - \overline{s}$$

Therefore,

$$\begin{aligned} \overline{N}_{\text{waiting}} &= \Lambda\overline{W} \\ &= 1\ \frac{\text{customer}}{\text{minute}} \cdot 1.5\ \text{minutes} \\ &= 1.5\ \text{customers} \end{aligned}$$

***Utilization Law**. The Utilization Law is actually a special case of Little's Result obtained by analyzing **just the server**.*

In a single-server system, the average occupancy of the server is a number between 0 and 1; that is, the server is either always empty ($U = 0$) or it's always full ($U = 1$). The "residence time" at the server is $\overline{s}$, so Little's Result tells us that

$$U = \overline{N}_{\text{server}} = \lambda \overline{s}$$

Measurement-Based Argument for the Validity of Little's Result

Suppose we observe a system for a long time period T. During this period, we measure

- C customers exiting the system
- A total of γ customer-seconds accumulated in the system. For example, if two customers are present in the system for an interval of 10 seconds, that would count for 20 customer-seconds. γ is therefore the total amount of time accumulated by all customers who passed through the system during the measurement interval.

The following relationships hold:

$$\overline{N} = \frac{\gamma}{T}$$
$$\Lambda = \frac{C}{T}$$
$$\overline{R} = \frac{\gamma}{C}$$

Use these relationships to verify that $\overline{N} = \Lambda \overline{R}$.

$$\frac{\gamma}{T} = \frac{C}{T} \cdot \frac{\gamma}{C}$$

This style of justification is called a "measurement-based argument" rather than a proof, because we're being imprecise about the nature of this measurement interval and exactly what it means for the system to be stable over the measurement period.