**Stability in Queuing Systems**

Major long run measures of performance of queuing systems are:
- \( L \): the long run average over time of the number of customers in the system
- \( L_Q \): the long run average over time of customers in queue \( Q \)
- \( w \): the long run average time spent in the system
- \( w_Q \): the long run average time spent in queue \( Q \)
- \( \rho \): the server utilization, the \% of time that a server is busy.

A system is said to be **stable** if its long run averages exist and are \(< \infty\). If a system is unstable, its long run measures are meaningless. A model of an unstable system must use protocols to accommodate the effects of the instability.

**Measure \( L \):**

If time units are approximated by discrete intervals \( \Delta T \), then let
- \( T_0 = \) number of intervals \( \Delta T \) that 0 customers are in the system
- \( T_1 = \) number of intervals \( \Delta T \) that exactly 1 customer is in the system
- \( T_2 = \) number of intervals \( \Delta T \) that exactly 2 customers are in the system
  
  \[ \ldots \]
- \( T_n = \) number of intervals \( \Delta T \) that exactly \( n \) customers are in the system.

Then for the time period \( T \) covered, the \( T_i \) encompass all \( \Delta T \) intervals and so

\[
T = \sum_{i=0}^{n} T_i \quad \text{and} \quad \sum_{i=0}^{n} i \cdot T_i \approx \text{the total area under the curve of } C(t),
\]

where \( C(t) \) is the number of customers in the system at time \( t \). The area represents the total time spent in the system across all customers.

For this time period, the expected value \( \hat{L} = \sum_{i=0}^{n} \left( \frac{T_i}{T} \right) \) approximates the mean number of customers in the system over the given time interval.

This leads to the form \( L = \frac{1}{T} \int_{0}^{T} C(t) dt \).

To obtain \( w \) (since customers are discrete), approximations of \( w \) are given by

\[
\hat{w} = \frac{1}{N} \sum_{i=1}^{N} w_i,
\]

where \( w_i = \) time in system for customer \( i \) and \( N \) is the customer count.

Hence, as \( N \to \infty \), \( \hat{w} \to w < \infty \) (unless the system is unstable).

In general, the **conservation equation** \( L = \lambda w \) where \( \lambda \) = the mean arrival rate holds in a stable system. This equation also provides the relationship between \( L \) and \( w \).

In other words, if we have an average time \( w \) in system per customer of 10 hours and customers arrive at the rate \( \lambda \) of 2 per hour on average, then at any time we expect to have 20 customers \( (L) \) in the system.
Server Utilization:

Viewing the server as a system in its own right, server busy means 1 customer in the system, server idle means 0 customers in the system. If the mean service rate (customers served per unit of time) is given by $\mu$ and customers arrive at the server with mean arrival rate $\lambda$, then the L statistic for this system is the server utilization $\rho$. To see this, note that $w = 1/\mu$, the average time a customer spends in the server and so by the conservation rule

$$L = \lambda w = \frac{\lambda}{\mu} = \text{arrival rate} \times \text{average time in server}$$

$$= (\text{customers per unit of time}) \times (\text{time-units of service per customer})$$

$$= (\text{time units of service})/\text{time-unit} = \% \text{ of time in use} = \rho.$$  

It is evident that for stability, we must have $\rho < 1$; i.e., $\lambda < \mu$. In particular, this confirms that the arrival rate ($\lambda$) must be less than the service rate ($\mu$) and the mean interarrival time ($1/\lambda$) must be greater than the mean service time ($1/\mu$).

Remark: For the overall system, $L$, $L_Q$, $w$, $w_Q$ are dependent on the arrival patterns, not just $\lambda$ and $\mu$; hence, analysis requires knowledge of the data distribution.

For example, as an extreme case, suppose 1000 customers arrive at time 0, another 1000 at time 1000, and so forth. Although the customer arrivals are quite dispersed, the average arrival rate is $\lambda = 1$. If the average service rate $\mu = 1$, then for $M =$ the number of blocks of 1000 people (so customer count $N = 1000M$)

$$w = \frac{1}{1000M} \cdot M \cdot \sum_{i=1}^{1000} i = \frac{1}{1000} \cdot \frac{1000 \cdot (1000 + 1)}{2} = 500.5 = w \text{ (recall that } \sum_{i=1}^{n} i = \frac{n(n + 1)}{2})$$

Essentially the same computation results in $L = 500.5$.

In contrast, if the customer arrivals distribute uniformly with $\lambda = \mu = 1$, then $w = L = 1$.

Testing Stability:

A necessary and sufficient condition for model stability is

mean interarrival time $>$ mean service time

(customers arrive on average slowly enough for the server to stay ahead).

It needs to be emphasized that this condition may be insufficient in practice (e.g., the above example where 1000 customers arrive at time 0, another 1000 at time 1000, and so forth had an uncomfortably large $w$ value of 500.5 because of the arrival pattern).
In a stable subsystem of a model where items which arrive in the subsystem must eventually depart, then the mean interarrival time (miat) at the entrance point to the subsystem is equal to the miat at the departure point from the subsystem. This is the **flow principle**. In essence items can’t depart faster than they arrive, and if the subsystem is stable, they can’t arrive faster than they depart.

**Split Rule:**

Suppose customers arrive at a point $x$ in a system with mean interarrival time (miat) $\mu$. They proceed to point A with probability $p_1$ and to point B with probability $p_2$. Then the mean iat
- at point A is $\mu/p_1$
- at point B is $\mu/p_2$

For example,

![Diagram](image)
Join Rule:

Suppose customers arrive at point x in a system from either point A or from point B. If the mean iat at point A is $\mu_A$ and at point B is $\mu_B$, then the mean iat ($\mu_x$) at point x is

$$\mu_x = \left( \frac{1}{\mu_A} + \frac{1}{\mu_B} \right)^{-1}$$

(i.e., add the arrival rates and invert to get the miat)

For the example from the split rule, if we rejoin the A and B legs, then we recover the miat as $(1/83.33 + 1/35.71)^{-1} = 25$ as expected.
Example:

Customers arrive at an average rate of 1 every 25 seconds. 30% proceed to the help desk and 70% to the checkout station. 80% of the time the help desk receives a directional question taking on average 20 seconds to answer. The remaining questions take on average 2 minutes to answer. 20% of help desk customers proceed to clearance, the remainder go to the intervention station and then exit. There are 2 servers at the checkout station, one of whom averages 60 seconds per customer and the other 90. From the checkout station, customers proceed to clearance. On average, clearance takes 33 seconds. Help desk customers proceeding to the intervention station are normally processed in 1.5 minutes.

Analysis:

The model is diagrammed to show the splits and joins as given by the problem statement. For analysis purposes, the miat on the upstream side of a server is kept the same on the downstream side. Instability is identified wherever the miat at a server is calculated to be less than its mean service time (marked with ★ on the diagram below).
Feedback Loops:

A feedback loop is a section of a model where items can loop back in the item flow; e.g., customers in a cafeteria returning for dessert, a part returned for further refinement, a job worked on for a while and then returned to queue. The section of the model containing the loop can be isolated with a single point of arrival for the section and a single point of departure.

When a model has a feedback loop, the effect is localized to the part of the model where the feedback occurs. The items feeding back increment the interarrival times that occur within the part of the model affected by the loop, but have no impact on miat values outside of the loop. To see this, consider the example below.

Restricted Feedback:

Feedback is called restricted is an item cannot loop back more than once. Consider an example where 10% of items going down a path in the model feed back to re-enter the path, but no more than once. If the miat at the top of the loop is 60, then the miat for the 10% of the items that feed back in is $60/0.1 = 600$. Beyond the loop, items continue with a miat of 60.

Intuitively, the items in the feedback loop are undergoing an extra delay, so the miat on arrival to this part of the model matches the miat on departure. Items feeding back decrement the miat for the service point (items arrive faster). Notice that the percentage at the split for feedback is slightly lower than 10%, reflecting that some items in the flow at this point have already looped back and so cannot do so again.
Unrestricted Feedback:

Feedback is called unrestricted if items can loop back whenever they reach the feedback point, whether or not they have already looped back; i.e., the item is memoryless. In this case, the miat values are determined from the bottom of the loop. Revisiting the prior example, lifting the restriction on feedback, the diagram becomes:

For another viewpoint, at the top of the loop, $p$ items recycle exactly once, $p^2$ exactly twice, $p^3$ exactly 3 times, ..., so the miat for the number recycling is given by $t \sum p_i = pt/(1-p)$

For the example, the recycle calculates as $(60 \cdot 0.09)/0.1 = 540$ which agrees with the prior calculation.
The remaining considerations are the miat into the service point and the probabilities associated with the split at the bottom of the feedback loop.

Algebraically, if $t$ is the miat of the main flow where it joins with feedback items at the top of the loop and the feedback probability is $p$, then the miat for restricted feedback at the service point is given by

$$\left(1 + \frac{1}{t} \left(\frac{t}{p}\right)\right)^{-1} = \left(\frac{1+p}{t}\right)^{-1} = \frac{t}{1+p}$$

which in the above case is $60/1.1$

By the flow principle, the miat into the main flow at split point where feedback occurs must return to $t$ ($= 60$ in the above case). This forces the probability of rejoining the main flow to be $1/(1+p)$, which in the above case is $1/(1+0.1)$. The probability of taking the feedback side at the split point is then just $1-[1/(1+p)] = p/(1+p)$, which is $0.1/(1+0.1)$ in the above case.

For unrestricted feedback, it is given by

$$\left(1 + \frac{1}{t} \left(\frac{pt}{1-p}\right)\right)^{-1} = \left(\frac{p+1-p}{pt}\right)^{-1} = pt$$

which for the prior example is $60*0.9$. Note that the probability of rejoining the main flow is just $p$ and for taking the feedback side is just $1-p$. 
Mid-stream Feedback:

Analysis is unaffected if feedback enters an interior point of the model section (defined by a single point of arrival and a single point of departure) that contains the feedback loop. Identifying the section requires moving backwards in the model until all item paths leading to the feedback point have been included. This is easily illustrated with an example:

The computations work similarly for the unrestricted case.