Identification of the Constraint

What are we trying to do with any system measurement? We are trying to evaluate, assess and gauge how well our system performs. System performance is optimized when demand and supply have a close relationship - when the delays are minimized.

Our first step is to identify the streams of demand that are distinctly different, and have markedly different requirements for successful completion. We thus need to measure the performance for all distinct streams independently.

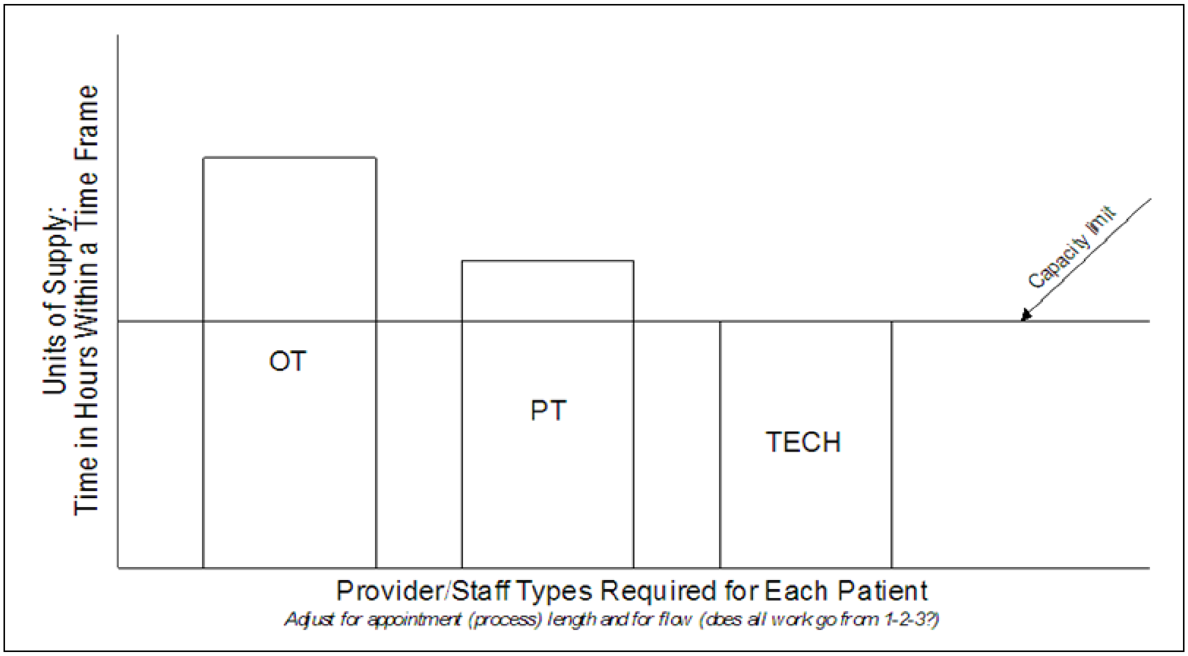
How long is the delay for each of these demand streams?

One common problem is having work that appears to be unscheduled. This can be somewhat deceiving though since the pieces are often gathered spontaneously out of an undifferentiated "schedule," and the parts are put together for a last-minute "schedule." So what appears on the surface to be undifferentiated work and unscheduled work does contain under the surface some form of spontaneous schedule.

Honestly, I would convert this spontaneous scheduling into an explicit schedule. You ought to be able to predict/anticipate the requirements to meet demand, and to be able to plan your supply in advance. I suspect that this spontaneity grows from the variation in demand (i.e. not knowing the needs for the distinct demand streams).

Thus, over time, I would convert unscheduled work to scheduled work, and schedule for the demand streams. You will have to use historical data to predict the relative allocation you need for each of these visit types. I would measure the demand for each type - "count" the number of appointments made on each day for each of the demand types. Then compare this number to the supply allocated for each, and analyze retrospectively just how well you used that supply (activity).

I would measure delay as third next available appointment (TNA). There is sometimes concern regarding no-shows, but this is not a concern. The decision about using TNA or FNA (first next available) is not a decision based on no-shows for patients. We would use TNA because of no-shows for providers. Let me walk you through this. If the first next available for the triad is this week, this sounds and "measures" well, but the problem with multiple provider visits (e.g. triad and dyad visits) is that you have to make sure that ALL providers are present. So any provider "no-shows" (vacations, time off, etc.) will have a huge effect. You will miss that effect with the FNA and see it much better with TNA. We are looking at this performance through the customer's eyes, not ours.



This graph may be difficult to comprehend, but it is a critical view. You can use this analysis by appointment type or in aggregate.

On the vertical axis we have provider time in hours (or any unit of supply). This is the number of hours (or supply units) within a specific time frame (like a week, for example). The amount of time this week is affected by multiple factors: is the provider here or not? How much time is spent doing other things and how much time is spent doing work in the office? If you are looking at this graph for a specific appointment type, say the triad visit, then the time in hours or units reflects only the time devoted to or allocated for the triad visits. This is why scheduling is very important.

Unscheduled work is very difficult to measure. Unscheduled work may make us feel good and feel useful, but we cannot measure how well the system is performing except by "feelings," and you know what I think about that.

On the horizontal axis, we have each of the provider types. In this example, we are talking about a triad visit - OT, PT and TECH.

We can see that the provider type with the least number of hours will determine just how fast the system performs, the delay, and the performance. We can only go as fast as the slowest (least available) of each of these providers.

I assumed in this graph that the patient needs to have the same number of time units for each of the providers (i.e. that the visits are the same length). In the case of the triad visit, this visit is the same length. If the visits are not the same length, for example if every patient needs to see OT, PT and TECH in some sequence but the OT visit is 60 minutes, the PT is 30 minutes and the TECH is 15 minutes, then we need four times as much OT person as we need TECH because for every hour of patient work that goes through TECH we need four hours of OT. The horizontal component of the graph may have to be normalized to account for that.

The graph works well for the triad visit. We can only have a triad visit if all three parts are in place with the same number of hours. If one of the parts is missing this week, then it doesn't matter if the others are there for 40 hours - there can be no visit.

The graph can be altered to see the same dynamic for the dyad, and can be altered to see visits in sequence. The appointment length plays a huge role in the "in sequence" visit types. The essential issue in the in sequence visits is as mentioned on the graph - do all patients need to see all providers?

This is pretty complicated and I don't expect that you will get it right away. But here are the main lessons:

1. Measure third next available appointment, not first next available appointment.

2. Convert unscheduled to scheduled work so you can measure it

3. Analyze for the constraint. One of these supply persons, particularly in the triad visit, will determine the delay. You already organically know that - we expect that this is the main driver for unscheduled, spontaneous, put-it-all-together-at-the-last-minute type of planning. We understand that. But when you spontaneously schedule you are blind to how well the system is performing.