Ep #3: Finding and Fixing Bottlenecks to Drive Law Practice Efficiency

Full Episode Transcript

With Your Host

John E. Grant

Ep #3: Finding and Fixing Bottlenecks to Drive Law Practice Efficiency

We all know that there are bottlenecks in your workflows that are creating turbulence and preventing work from flowing smoothly through your legal delivery systems. And there are a lot of products and techniques out there that claim to be able to help you tame those bottlenecks to make your system more efficient. And some of them work but often what they do is just move the bottleneck to another part of your workflow.

What if I told you that one of the best ways to get work to flow more smoothly through your overall delivery system is to create an intentional bottleneck at a key place in your workflow. In today's episode I'm going to tell you what that place is. Ready to become a more agile attorney? Let's go.

Welcome to *The Agile Attorney* podcast powered by Agile Attorney Consulting. I'm John Grant and I've spent the last decade helping lawyers and legal teams harness the tools of modern entrepreneurship to build practices that are profitable, scalable, and sustainable for themselves and their communities. Each episode I offer principles, practices, and other ideas to help legal professionals of all kinds be more agile in your legal practice.

Welcome everybody to this week's episode. Last week, you heard me talk about capacity and the importance of balancing the amount of work that you allow into your system with the delivery of work out of your system. And really making sure that you're managing things so that you are focusing as much on getting work done and out the door as you are on bringing new work in.

And in fact one of the things about agile approaches and all these different things that I do is, ultimately, it is about a systems thinking approach. And what we are looking for is a balance between the work that is coming in and the work that is going out. And if you are bringing in more work than you're getting done, that is what leads to overwhelm.

Today, I'm going to go a little bit more micro than that and I'm going to talk about capacity specifically at the most bottlenecked part of your law practice or your delivery systems. If you've heard me talk on other people's podcasts, I talk about this a lot. And so this may be a review for some of you, but it's so critical to the way that I approach this work that I think it's worth doing a whole episode myself on this topic.

So bottleneck theory. It ultimately comes from a line of thinking known as the theory of constraints. Without diving too, too deep, the best first articulation of the theory of constraints comes from a book called *The Goal* by a man named Eliyahu Goldratt. Goldratt was, as I understand it, a metallurgical expert of some sort. He was a professor who wound up working in a manufacturing space.

And he started applying some of what he knew as a scientist to solving some of the workflow problems within the manufacturing facility that he was a part of or that he was connected to. And was able to come up with some pretty interesting insights around that. And his articulation of that is known as the theory of constraints. The book *The Goal* is actually a pretty interesting read. It is one of those books that is a business novel.

If you've ever read works by Patrick Lencioni or some others that really use narrative to get their story across or get their point across as opposed to it being a typical 10 chapter business book that says do this, then do this, then do this. So I find *The Goal* to be an interesting enough read that I would recommend it. It tends to be on the syllabus for almost anyone that is getting an MBA, going to business school.

But the key takeaway from the theory of constraints is that in any multipart system. So I sometimes imagine this as a series of pipes where, as things flow through those pipes, not all the pipes are of the same diameter, they're not all the same size. So some of them are capable of carrying more work. Some of them are capable of carrying less work. In any one system there's always going to be one part of the system that is the bottleneck, that is the smallest pipe. It is capable of carrying the least amount of work.

And what the theory of constraints tells us is that the total flow of the system is limited by the flow of work through its most bottlenecked component, through its least amount of capacity. It's another way of saying that every chain has a weakest link. There might be lots of links in the chain that are rusty that look like they could use some work but under tension only one of them is going to break first and then that's going to relieve the tension on the rest of the chain.

In workflow systems, it's the same way. The work can only flow as fast as it flows through its most bottlenecked state. Now, if you accept that as true, then there are two corollaries that also have to be true. The first one is relatively easy. If you can improve the flow of work at the bottleneck, then you can improve the flow of work through your whole system.

If there's one part of the pipe that is really, really narrow, if you can open that up, if you can expand the flow through that particular phase, it's going to pay dividends for the entire enterprise. The whole thing is going to work just a little bit better. And that's the one that people usually don't have too hard a time wrapping their head around. Hopefully you're sort of nodding along as I say this.

The other corollary, and this is the one that's a little bit harder, is that any improvements you make to parts of your system that aren't the bottleneck, can't help. And I'll put a graphic in the show notes that I use when I'm doing presentations for this. If you have a seven part system and part three is where work is really getting stuck, is really piling up. If you try to improve something at part six, it doesn't matter because the work is stuck at part three.

So you can make your billing process more efficient, or you could make other things in the later stages of your workflow more efficient, but it’s not going to help make your overall practice more efficient. You’re not going to get more flow through the system overall.

The same thing goes upstream of the bottleneck. If you’ve got this workflow blockage at step three, but you go do something to improve your intake or improve your marketing, that’s upstream of your intake. That’s not going to help you get more work done, and in fact it’s actually going to probably make things worse because you’re just putting more pressure on that bottleneck. And that pressure leads to turbulence, it leads to tension, it leads to the possibility of something breaking and making a whole mess.

And so putting more work into a system that has a downstream bottleneck is actually really risky. It creates a problem with your overall system. The other ramification of that idea that improving parts of your system that aren’t the bottleneck, won’t help is that improving your pet peeve may not be helpful if it’s not also your bottleneck. And I’ll come back to that a little bit later.

But I do think it’s important when the bottleneck is an individual resource, usually a person, often you, that you do try to do some things to make that person’s life better. And I’ll talk about how to do that. But not everything that you do is necessarily going to have a systemic impact and that’s the key.

Another way to think about this is if you go back to the chain metaphor, if you shore up parts of the chain that aren’t the weak link, what you wind up doing is making the whole chain heavier, but not stronger. It still is going to break in the same place. In fact, it might be more likely to break because it’s carrying more weight now than it was before.

So, the key to workflow improvement, to process improvement is number one, to find where your bottleneck is. And there's a couple of ways to do that. My favorite way is to use a Kanban board. So I haven't done a deep episode on Kanban yet, but high level, you probably have seen them at this point. They are these vertical columns, each column represents a stage of a workflow or a phase of work. And then the work itself is represented by cards on the board.

And cards will flow through the different stages of work and get to some ultimate done stage, hopefully smoothly and regularly. But my guess is, especially if you're using a Kanban board already, there are places where work is getting stuck. And if you look on a Kanban board and you see a column where there are lots and lots of cards and that work is getting stuck. That's a pretty good indication that that is your process bottleneck. So that is the first way and frankly the easiest way to identify where the bottleneck is in your workflow.

The more technical way and actually the more accurate way is to do some measurements. And if you are able to measure the total amount of time it takes for a particular unit of work to make it through your entire system. And if it's a legal matter that is typically going to be measured in probably weeks or months, hopefully not years, but depends on the matter. It likely isn't going to be something that goes through in minutes or hours.

Then you need to look at the flow of work for each of the individual phases and measure sort of those micro measurements as well and say, "Okay, this is how long it takes for work to get through the intake phase. This is how long it takes work to get through the initial assessment phase." And notice I am saying how long it takes the work to get through, not how much time you spend on it. That's your effort.

And lawyers are really good, especially if you're an hourly biller, we measure effort really well. But I care about effort and I think that there are opportunities to improve that. But what I'm talking about now is the total time. Sometimes we call it the lead time in agile. How long does it take if you start the stopwatch the moment it enters that phase and you don't end it until the moment that it exits that phase? That's the lead time I'm talking about.

And so your bottleneck stage is typically that part of your workflow that has the longest elapsed time, the longest lead time. Whether or not it's the place where you spend the most individual effort on it or someone on your team spends the most hours working on it, it's where it gets stuck in the system overall. Once you know where your bottleneck is, then it's a matter of running some experiments to try to improve the flow of work at the bottleneck and really, the type of experiment depends on the nature of the bottleneck.

But if the bottleneck is due to a lack of human resource, the lack of a person to actually be made available to do the work. One of the things that is a little maybe counterintuitive around the theory of constraints is that it is okay to steal capacity, to borrow capacity, I should say, from parts of your workflow that aren't the bottleneck in order to make workflow more effectively at the bottleneck.

I ran into this with a firm I was working at just not too long ago where one of the attorneys was sort of the 'rainmaker', the main partner, the owner of the firm. And that attorney was spending a lot of time bringing new work into the system, but the work itself was getting stuck at this sort of quality review, quality assurance stage. Where the drafting of documents had been done but it needed someone to go through them and review them before they could get pushed out the door and delivered to the client.

And what that team came to realize in the course of our workshop is that we should borrow the capacity of that rainmaker partner and have them slow down on intakes for a while. And use their capacity to help at the quality review stage because that's where the bottleneck was and that mathematically and logistically, there was going to be no degradation in the performance of the practice. Because by allowing more work to get through that quality assurance phase and moving it further downstream into the system they actually were going to deliver on those promises more quickly.

This was a place where most of the work didn't get billed until the final delivery to the client. And so it actually was going to accelerate their recognition of revenue and help the firm make more money than if they just kept overstuffing the sausage with more and more new work. Now, the same thing would have been true if we had borrowed capacity from the drafting stage, which is just upstream of QA.

And obviously in that practice and I think in most practices you don't want the same attorney doing both the drafting and the quality assurance, if you can avoid it. If you're small enough, you maybe don't have the luxury of having multiple eyes on the thing. But either way it's better to do the quality assurance work than to do the drafting work.

And then obviously it's even more true if you're talking about something that is downstream of quality assurance, where if you have idle capacity because there's a bottleneck upstream, then you want to reallocate that capacity to the bottleneck state. The other thing though, about this firm and the borrowing of capacity from intake is that it did something else almost unintentionally although I saw it happening, I'm not sure the firm did. That was ultimately I think going to help make process improvement easier in that firm which is, it slowed down intake.

It created maybe not a full stop on intake, but it really, if you go back to the overflowing bathtub metaphor that I used in the last episode, we maybe didn't turn the tap all the way off with the overflowing bathtub, but we did turn it down. So it was more of a trickle and not a rush of new cases coming in. That in turn, helps relieve the new pressure on the bottleneck.

So by using that resource to get more of the work done at QA and also make sure that the arrival rate of new work at QA is slowing down a little bit, that gives the firm overall some breathing space. And that breathing space is really essential to being able to sort of take a step back and maybe do some process improvement work.

A lot of the stuff, if you are finding the same type of error in the quality assurance phase over and over and over again, then that's a really strong indication that we need to fix something upstream in the drafting phase. And maybe take a little time to do some better templating or to work on your document automation or do some training for the people that are responsible for the drafting. So that they don't make that mistake quite so often, or hopefully at all.

The metaphor I've heard, and I'll admit, I tried to find this on the internet and I couldn't quite find quite the story that I originally heard. But how this was told to me is that if you're trying to make the flow in a river go more smoothly, the best thing you can do is remove the biggest rocks from the river, the ones that are creating the most turbulence. But if you try to wade out into the river and address those rocks when the flow is high, spring runoff, then you're likely to get bashed against those rocks. It's really dangerous.

And so if you're going to be engaging in process improvement work, most of the time, the best thing that you can do is try to reduce the level of the water a little bit. Reduce that complete overwhelm so that you have the time and the space to actually make the changes and do the thinking and the investigation that's needed in order to make the changes that are going to allow work to flow more smoothly overall.

Another way of thinking of that, and this is one of the, I think, kind of coolest parts around theory of constraints literature is, part of what we're doing is putting an intentional bottleneck upstream of the true bottleneck so that we can limit the amount of pressure that the true bottleneck is feeling. One way it was explained to me one time is that bottlenecks are on bottles for good reason.

You want something to flow out of a ketchup bottle at a different rate than you want it to flow out of say, an olive oil bottle, which is at a different rate than you want it to flow out of a bottle of hot sauce. And each of those things has a bottleneck. And in fact the hot sauce probably has one of those little plastic things with a super tiny hole in the top of it. Because you want to control the amount of whatever condiment we're talking about, that goes on to your food. If it flows too quickly then you're going to create a problem if you're talking about the hot sauce.

The size of that bottleneck also allows you to see how much of that condiment is making it onto your food, it creates a feedback loop. And so taking this back into your law practice, one of the best things that you can do is create an intentional bottleneck at the very top of your workflow, which is usually your intake phase. And it's not that we want no intake to come in, but if you're already overwhelmed, you probably want the intake to be more like the bottle of hot sauce than the bottle of ketchup.

You need to really be intentional and then also create a balancing loop, a feedback loop. And this gets into the systems thinking elements of a lot of this work where you want to be able to understand how many cases are we closing, how much work are we delivering. And then meter our intake to match our outflow so that we're never, or we're trying to avoid putting our overall practice back into that place of overwhelm, that place of being over capacity. I'll leave it at that for now.

I am going to keep coming back to bottleneck theory because it is really one of my favorite components to this process improvement work and I think the most meaningful thing. My takeaway for you right now, I think, is to be really intentional about finding where in your practice your bottleneck is. And then don't be afraid to put intentional limiters, regulators upstream of that bottleneck. And don't be afraid to borrow resources from other parts of your practice in order to apply those resources at your bottleneck as a tool for getting the work to flow more smoothly.

Thanks for listening to *The Agile Attorney* podcast. I'm your host, John Grant. If you found today's episode interesting or useful, please share it with someone who you think would benefit from a more agile approach to their legal practice. If you have any questions, feedback or maybe a topic you'd like to hear me cover, you can reach me at john.grant@agileattorney.com.

To help other attorneys and legal professionals discover this podcast, it helps a lot if you could rate or review me on Apple Podcasts or Spotify. And of course, be sure to subscribe in your favorite podcast app. This podcast gets production support from the fantastic team at Digital Freedom Productions and our theme song is the instrumental version of Hello by Lunareh. That's it for today's episode. Thank you for listening and see you next time.