Ep #5: Close the Closable: Finding Balance in Your Caseload

Full Episode Transcript

With Your Host

John E. Grant

Ep #5: Close the Closable: Finding Balance in Your Caseload

I'm going to cover two important and related topics in today's episode. One of them has to do with utilization rate, which is a really common law practice management metric. And the conventional wisdom around utilization rate is this idea that you should be trying to make sure that the people and resources that you're paying good money for, are working as close to 100% capacity as possible. Because that's how you make sure you're getting a good return on your investment, and it makes intuitive sense.

But I'm going to argue that that conventional wisdom is not just wrong, it is actively working against your goal of making your law practice more productive. The other concept has to do with how to think about prioritizing which matter in your system you should be working on when you and your team are already working at or over your ideal carrying capacity. In other words, how do you decide where to apply your and your team's finite time and attention when the universe of options is really big?

If you're like a lot of firms, you've been overlooking a golden opportunity to reduce the total amount of work that is in your system. And by reducing that work in progress, you'll also reduce the stress and anxiety that comes from feeling overwhelmed by so much work. 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.

We're still in the early part of the year. And it's a great time to think about maybe a little early spring cleaning or maybe making a strong start on that resolution to get more organized. Although there's really never a bad time

to get more organized. To help you do that, this episode is going to build on a few concepts we've been talking about over the last few weeks. In episode two, I talked about an honest reckoning with capacity, specifically reckoning with the fact that your capacity is finite.

And while there are ways we'll talk about in future episodes to grow your capacity, one of the core principles of the Kanban method is, start with what you do now. Which means we need to start by accepting your current capacity as it is today, not fantasize about what we want it to be. So in this episode we're going to talk about some things that are using up significant chunks of your capacity but in a low value way. And how reducing or eliminating those things will help the higher value work flow more smoothly through your systems.

In episode three, we talked about theory of constraints or bottleneck theory, high level. There's always some place in your workflow where work is getting stuck. No, duh. Probably there's a lot of places, but one of the things that the theory of constraints teaches us is that only one of those sticking points really matters. Only one of them is constraining the flow of work through your entire system.

But the thing I want you to keep in mind for today's episode is that those sticking points in your workflow can be expensive from a capacity standpoint. Your capacity is finite therefore, anything that takes up your capacity in non-productive ways is a bad thing. You want to avoid it as much as possible. I said non-productive ways. So let me backup for a minute and define productivity. And specifically I want to define the unit of production in a legal workflow.

If you ask any economist to define productivity, they'll tell you it's the number of units produced in a given time period. Expressed as a fraction, it would be widgets in the numerator and time in the denominator. There's sort of a tacit assumption that those units actually get purchased by and

delivered to a customer, but because the relative term is productivity, we're talking about production, not consumption.

The key is if the units are complete, there's something that someone would actually be able to buy and make use of. Then there's the way a lot of lawyers define productivity, which is the number of hours worked or billed or collected against a matter. The thing is, as Dan Lear and I discussed on his *Financially Legal* podcast a while back. Your typical economist would view that measure of productivity as nonsense because you've got time in both the numerator and the denominator. So the units essentially cancel out and you're left with just a number instead of a measure.

The other problem is that those hours in the numerator aren't consumable. They don't represent a complete unit of value that someone can buy and use. They're inputs. They're a measure of effort needed to produce the thing that ultimately has value to your client, which is the finished case or the closed matter. And my point from this tangent is that the better thing to use as a measure of productivity in your law practice is the number of matters completed in a given time period.

Your client doesn't retain you because they enjoy the process of working through a legal issue. They retain you to try to achieve some legal outcome. And they won't always get the outcome they're hoping for, but there's value in the certainty of achieving a result or having an answer and being able to move on. I'm going to talk about some models for understanding what your clients truly value in the next podcast episode, so stay tuned for that.

So if the true unit of value of productivity in your law practice is completed matters, then the Kanban board I described in the last episode, and specifically I'm talking about the slow flow or the matter level board. That board is a visual representation of your production line. It's a visual fiction for knowledge work that lets you better track it across the various stages

necessary to get your matter to the point where both you and the client consider it to be done, to be complete.

If you have a matter level Kanban board already, then the thing to recognize is that every card on your board, at least until it hits that done column at the far right of the board, represents an incomplete or an unfinished unit of production. And yes, you may have delivered some value against that case already, but you've also got kind of a debt, a promise to deliver more value, more work on that matter, until the point where it's actually 100% completely done.

And once you look at it that way, that each card, each matter represents a commitment you've made, therefore, you're kind of in this form of delivery debt until you've fulfilled that commitment. Then there are two ways that the cards on your board, the in-flight matters in your law practice consume your finite capacity. The good way, the first way is when you're engaged in productive work on that matter. Meaning you're doing things that will pull that matter forward through your workflow and move it closer to that place where both you and your client consider it to be 100% done.

And that done state may be way out in the future, but you've still made tangible progress towards it. And my guess is that if you think about the times when you feel truly productive in the sense of your personal fulfillment, that you've made good use of your capacity in a given day. It probably has a strong correlation to days when you can point to a matter or two and feel that tangible sense of progress on that case.

The other draw on your capacity, though, is the administrative labor of tracking all that work that's in your system. It's the carrying cost of coordinating your resources, your case status meetings, your chasing down client homework, or bugging opposing counsel for a response. It's the stuff that you have to do to manage the work in your practice. But it doesn't directly translate to progress on any one matter, which means it doesn't directly contribute to the productivity of your workflow overall.

An economist might see it as general overhead more than the production cost for any one unit of delivery. Looping back to bottleneck theory for a minute. When work gets stuck behind a bottleneck, that work tends to require more administrative overhead. One of the most expensive things I see about work that is stuck in a legal workflow is that the longer a matter is stagnant, the longer it will take someone to ramp back up when they actually do get a chance to work on it.

The amount of working information your brain can hold about any one matter has a half-life. It decays over time, which means the longer you go without working on it, the longer it will take you to reboot your brain and remind yourself what's really going on with this thing before you can engage in the actual productive work.

Let me digress further for a moment and revisit that notion of billable hours as productivity. And yes, if you're an hourly biller, you can probably bill for that ramp up time, which means you don't really have an economic incentive to minimize it. In fact, it's the opposite. It's no secret that efficiency is not the friend of the billable hour. But if you're listening to this podcast or you've connected with me somehow, my guess is that isn't really the way you're going to feel good about making your money.

Let me also be clear. I am not saying you need to rush out and ditch the billable hour tomorrow. Changing your economic model is complex. It requires a good deal of consideration and preparation. But what I am saying is, even if you're billing by the hour, you're going to be a lot more aligned with your clients' interests when you're billing for actual productive work than when you're billing for ramp up time or administrative overhead.

So one of my goals for my clients, with the Kanban method or with the Kanban system. Is to help them find ways to use as much of their finite capacity as possible on productive tasks that are actually moving a matter forward as opposed to administrative or overhead or ramp up tasks that are

necessary to keep the plate spinning but don't directly translate to productivity.

Now, when we talk about reducing administrative overhead, there are two main levers we can pull. One is to make the administrative work itself more efficient and a Kanban board is a tool for this. By making work visible, our brains tend to process it more efficiently. And by making it visible across a team, where everyone can see it, then that team winds up needing to spend less time coordinating around the status of a matter because the board shows everyone what that status is.

And there are other tools and systems that you're probably already using to lower your administrative costs. One is a form of labor arbitrage where you push as many of the administrative tasks as you can onto lower cost resources like paralegals or assistants or a virtual assistant. Another is using technology to cover some of that administrative work. Maybe you can use some form of automation to send client homework reminders instead of having a human do it.

And these are the kinds of things I see a lot of law practices investing a lot of time and attention and money trying to do. But there's another way to reduce administrative overhead that is a whole lot easier and less expensive but for any number of reasons, most people don't consider it as an option: Work on fewer matters at once.

And as I said in episode one, the number one reason people reach out to me is that they or their teams are overwhelmed by the total amount of work that's jammed into their system. Yet one of the hardest jobs I have is convincing those overwhelmed people to reduce the amount of work they commit to doing. Now, we're already a fair ways into this episode, but I'm going to introduce a new concept that I hope will help drive this point home.

There's a component of operations known as queuing theory, which deals with the amount of time that things spend waiting in line. As you might imagine, this is something a lot of businesses in the physical world pay

attention to. Grocery stores, banks, restaurants, amusement parks, they all care about queues because those businesses know that when a queue gets too long, people will start to leave before they spend money in that establishment. And of course, that's not good for business.

Interestingly, queuing theory also plays a big role in technology and telecommunications businesses. All of the data and bits and bytes and streaming video requests that come across the internet every second are using actual finite resources in the form of the cables and routers and switches that are carrying packets of digital information around the world.

And technologists use queuing theory to help them understand how to prioritize information requests, how to balance the loads on those physical machines. And maybe how to determine when they need to add additional capacity in the form of more or newer components.

But the thing I want to put in your head around queuing theory is a relationship known as Kingman's formula. And this is named after a mathematician John Kingman, who first articulated it. I'm not going to give you the actual formula because it's a little complicated to try to describe verbally. I'll put a link to it in the show notes. But it involves the relationship between the length of time it takes to complete a process, any process, and the degree to which the resources required to perform that process are utilized.

The fundamental relationship is consistent with which you've probably observed in the real world. The more a resource is utilized, in other words, the busier it is, the longer it takes that resource to deliver any single unit of work on average through its particular process. And when the resource itself is highly utilized, which is to say it's delivering the work more slowly, then that means longer and longer queues are going to form in front of that resource, while new requests for that resource's finite capacity keep coming in.

The easiest real world example you've surely encountered is the utilization of the finite road space of a freeway or highway. When that highway has really low utilization at 3:00am, then the rate of flow of any one vehicle, which is our unit of value, is pretty speedy. Take away legal speed limits for a minute and a car can travel really fast down that highway when there's nobody else on it. Once you get to around 40 or 50% utilization, which that's actually kind of high if drivers are keeping plenty of space between them and the car ahead of them, you'll notice a definite slowdown on that highway.

And the Kingman's formula says it should be right around 50%. So if the speed limit is 60 mph or 100 kilometers per hour, then when you get to a true 50% utilization of the highway then the rate of flow is going to drop to around 30 miles an hour or 50 kilometers an hour. Here's the thing with the Kingman's formula though, the relationship between utilization rate and the rate of flow is not linear, it's logarithmic. After you get past about 50% utilization the exponential nature of the slowdowns becomes noticeable.

All systems are different, but at 75% utilization the systems flow typically shrinks by around a fifth to a sixth. So if you're on that highway now you're down to about 10 miles an hour or 15 kilometers an hour, although even that maybe is fast for a highway that's three-quarters covered in cars. And it gets worse from there because 100% utilization is an asymptote and remember that word from trigonometry. It means that the delays get longer and longer, spiking to infinity the closer the resource gets to actual 100% utilization.

And on the freeway, that's when you see cars backing up onto the on-ramps and the side streets because with lower throughput you get those longer and longer queues as well. The same thing is true for the resources, both the people and the tools in an office environment. If everyone in a large office tries to print out a 100 page document at the same time, it's going to cause a spike in utilization of the printer. And that means the poor

person who hit control P slightly later than everyone else is going to have to wait a long time to get their TPS report.

Now, the other problem with high utilization is that it increases the likelihood of error or failure. The more cars on the freeway, the more likely they are to bump into one another, causing a distraction that takes away even more of that road's capacity. A highly utilized printer is more likely to run out of paper or toner or maybe have a paper jam causing unexpected downtime right at the moment when downtime is most disruptive.

And this introduces yet another concept called failure demand, which I'm going to talk about in more detail in a future episode. But high level, failure demand is something that forces an unexpected demand on your already finite capacity due to the system's failure to get something right or to do something on time in the first place.

If a driver fails to keep adequate distance or fails to ignore the distractions of their cell phone and they bump into another car, now you've got a whole new set of demands on the already busy highway while those drivers sort out their accident and maybe need other systems like police or an ambulance or a tow truck, that's failure demand.

Now, if a lawyer or a paralegal or any other knowledge worker is working at 80 or 90% utilization and that means that they don't get something delivered to a client on time or when they said they would, they're probably going to experience some form of failure demand too. And if they're proactive, maybe they'll shoot off an email to the client explaining the delay. But that explanation distracts the worker from getting the actual work done. It uses capacity without creating progress or value for the client.

And of course, if they're late on a deliverable and they're not proactive with their communication, then they risk a different form of failure demand, which is when the client calls or emails to say, "Hey, where's my stuff?" And now you've got to react to that and that's not productive value either.

Another form of failure demand in knowledge work, we've talked about already, which is the amount of time that can elapse between touches on a particular matter, causing ramp up time. If it's been a week or more since you last picked up a file, you're going to have to spend some amount of time rebooting your brain around what the heck is going on with that case. The information that you can carry in your active memory has a half-life. The longer the delay, the longer it's going to take you to ramp back up.

And as the Kingman's formula tells us, the more active cases you're trying to manage, the longer the delays will be between the times your brain gets to focus on any one case. Which means the longer it's going to need to use your already overtaxed capacity for the non-productive work of getting yourself back up to speed.

I'm going to loop back to hourly billers for a minute. I consider it one of the dirty little secrets of our industry that our conventional wisdom says that you should bill for that ramp up time and all the other administrative work because it's just the cost of doing business as a lawyer. I call BS on that. There is no way a client should have to pay a premium because you keep putting their case on the back burner while you fight other fires. And then you need to spend a whole big chunk of time getting yourself back up to speed when you finally do pick their case up again.

And even worse is if the client makes that where's my stuff request and the lawyer bills a .2 or a .3 to provide the status update. Man, if it is part of your business model to profit from your own inefficiency, you are not being part of the solution. And if you're doing it intentionally or even if you're putting off doing the efficiency work because your business model doesn't reward it, I think you've got to take a long look in the mirror.

But in reality, I think a lot of lawyers agree with me and write off that time. But that just means you've put the cost of being over-capacity on yourself or on your firm instead of on your client and that's not ideal either. And if

you don't bill hourly, then you're absorbing the cost of that failure demand no matter what.

Okay, that's a lot of background, but hopefully you're with me on my main thesis that number one, the closer you get to operating at 100% capacity, the slower all of the work is going to move through your systems. And as you approach 100%, it's going to get exponentially worse.

Number two, as you approach that 100% capacity, your likelihood of having to contend with failure demand increases. And that threatens to put you in this doom spiral where you're using more and more of your finite resources to put out fires, meaning more and more unexpected demands on your capacity which drives your utilization even higher. Which means higher likelihood of failure demand and so on and so on and so on. And that's where the feeling of being on the hamster wheel where you're running at full speed but not actually going anywhere comes from.

So what's the solution? Well, the metaphor I use with my clients all the time is when your bathtub is overflowing the first thing you need to do is turn off the tap. And the next thing you need to do is open the drain. And I get it, turning off the tap in a law practice, what I often refer to as an intake pause. It can be controversial, so I'm going to address it more fully in a future episode.

The main thing is if you can get yourself or your team to stop or even just slow down the rate of arrival of new work into your system, I promise it will help a lot. Think of it as metering the on-ramps on your freeway. Yeah, you delay some cars who are trying to access the road, but by doing so, you make it so that all the cars that are already on there can keep moving at a reasonable speed. And ultimately that means more of them will hit the off-ramp faster, which means the system will actually work better for everyone.

The thing I want to leave you with in this episode is another phrase I use with my clients all the time, which is close the closable. Every open case in your system has a carrying cost. If you're done with all the active work and

you just need to close the matter in your practice management tool or maybe send a disengagement letter, you're probably tempted to leave that non-urgent work for another day and then another day and then another day.

Probably because you're using the members of your team who would be responsible for those close out tasks to do intake instead. Which means you're actually blocking the drain and turning up the tap, exactly the opposite of what's needed. Tracking those almost closed matters exacts a toll on your finite capacity. There's a carrying cost. From a systems perspective, you are better off buckling down and closing those darned files so that you never have to think of them again.

The same thing is true for work that's stuck in the middle of your workflows. It's a little different, but if you're doing an estate plan or handling an immigration matter, for example, and you're waiting on the client to do some piece of homework before you can do something on your end. I highly recommend you put a time box around that homework period. And that means setting up a fish or cut bait decision on whether you'll continue with that matter if the client doesn't get their homework done. You shouldn't be carrying it forever in an unfinished state.

And another one of the episodes I have planned for this podcast is my proven hack for getting clients to do their homework faster, so stay tuned for that. But for purposes of this episode, what we're looking for is opportunities to close cases that either only require a small amount of work before you're done with them forever or cases that look like they're roadblocked for some reason, and whoever is responsible for the blockage isn't doing what they need to do to get through it.

In other words, close the closable, up or out, get it done. Because the reason you're feeling overwhelmed is that you've got too many open projects. And maybe not all of them are cases, but most of them probably are. And if you're like many, many of the law practices I've worked with,

you've got a lot of cases that are so close to done, but they're not urgent, so they just sit there. I'm begging you, just get them done. Schedule a few hours or half a day with your team and just bang them out.

Maybe close your back office projects too, but that's probably a different chunk of time. Now, here's the hard part. After you close the closable, don't replace those cases with new work. Let yourself and your team live in that lower percentage of utilization, because if you're getting yourself down to 70% utilized, that's the place where traffic is finally flowing a little better on the highway of your internal workflows.

The last thing you want to do is turn off the meters on your on-ramps or block the off-ramps and fill the darned thing back up again because it's going to take you right back to gridlock. Going back to episode two, you need to have that honest reckoning with capacity. And the reality is, your capacity probably isn't as big as you want it to be, and that's okay. We're going to talk about ways to grow your capacity, but I need a slice of that capacity in order to do that work.

And I can tell you from my long experience that a team that is overloaded or overwhelmed with work will never have the time or the brain space to engage in meaningful process improvement because they're too busy putting out fires. And what's worse is, they're probably going to be really susceptible to chasing rainbows and magical solutions that promise some one-click fix to their problems.

Or maybe they're going to be more likely to fall prey to some of the charlatans out there, the coaches or membership programs that promise to give you hyper growth or teach you how to 11x your business. And not all coaches are that. There's great coaches out there, but I think you know the ones I'm talking about. They're the ones that have that whiff of an ick factor or want to lock you into some multi-year contract. Because if you're not willing to step up and invest in my magic beans then you're probably not serious about growing your law practice. Alright, enough of that.

Ep #5: Close the Closable: Finding Balance in Your Caseload

As I said in the last episode, if you're using a Kanban board already, you can probably see where the work is getting stuck. For a lot of my clients in the early phases of working with me, one of the most obvious places where work gets stuck is right at that final close out stage. And it's stuck, like I said, because the people who are responsible for closing the files are the ones who are responsible for opening them, and the firm prioritizes intake over close out. I want you to flip those priorities.

Focus on closing cases. Focus on creating space, not filling it. Give yourself and your team some breathing room so that your commitments stop crashing into one another and causing the little emergencies that take up even more of your finite capacity. Because when you need to swoop in and clean up the mess, you're diverting precious time and attention away from making actual progress on somebody's case.

Close the closable and then give it some time to feel how much more smoothly the rest of the work flows when you don't have a million balls in the air that you're trying to juggle. I promise you, you're going to like it. It may feel weird at first because your brain has probably gotten so used to managing emergencies that it's going to almost crave them. But after a few weeks you can retrain it to get an even bigger reward out of the sense of real flow of actual progress on the work that matters.

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.