The Dynamics of a Judicial Service Supply Chain: A case study

Luis López
INCAE Business School
Alajuela, Costa Rica
Tel: +506 2437-2061
Luis.Lopez@incae.edu

Roy Zúñiga
INCAE Business School
Alajuela, Costa Rica
Tel: +506 2437-2256
Roy.Zuniga@incae.edu

Abstract

In this paper we explore behavioral issues, coupled with temporary capacity imbalances, which could influence the characteristics that a service supply chain may assume in the long run. We look at a service chain in which processing times by human agents are endogenously determined by what constitutes an acceptable and credible backlog; but implicit incentives, particularly within a formal hierarchy, may also impinge upon throughput rates at certain stages of the supply chain when agents are trying not to overwhelm downstream stations with excess work. We explore these issues in the context of a managerial intervention in a judicial service supply chain. Using data from a detailed case study we develop a preliminary model and discuss some results.

INTRODUCTION AND LITERATURE REVIEW

Service rates in service supply chains are often determined by looking at the system's past performance data. It is common to look at distributions of past processing times and then use those data for estimation. Such managerial approach is generally found in many and different types of service supply chains. For instance Gans, Krole and Mandelbaum (2003) in a review of common practices in telephone call centers found that staffing is typically done, as these authors put it, with "grand averages for historical service rates, productivity rates, and turnover." (Gans et al., 2003:96).

Though this approach may very well serve the purposes of supply chains whose pace is dominated by machinery, it may prove ineffectual in some service supply chains. This stems from the particular characteristics of services. The intangible nature of services makes it difficult to measure the termination of a service process, or how the service experience is shaping customer expectations as he or she enters a service system. Most importantly, in many instances services are performed entirely by human beings. Service pace is, thus, dictated not by equipment, but by people. As such, service rates within service supply chains are prone to be full of behavioral effects. Services rates for human agents may, unlike for instance the service rates of a piece of equipment, vary as agents may adapt output rates to a given set of circumstances. These circumstances may stem from incentives, explicit or implicit, present in the service supply chain, or from incentive-like perceptions. Adaptive behavior may occur over the long run, as when, for instance, servers within a service supply chain settle into a pace of work that is socially acceptable, or in the short term, as when people react to circumstances. In call centers for instance, it has been shown that as server utilization goes up, then people start taking breaks that take several forms, like extending calls with easy customers unnecessarily, or using extra time to fill out paperwork in between calls.

People's tendency to change the rate at which they work when engaging in a task has long been demonstrated in the literature. One well-known aphorism that relates to this is the one first suggested by C. Northcote Parkinson (1966). His dictum has now become accepted as a behavioral law: "Work expands so as to fill the time available for its completion" (Parkinson, 1955). Although Parkinson was referring to government bureaucracies, it is now accepted that at the individual level people tend to use all time available to complete a task. There has been empirical support to this law (Beate, 2009), and some authors have developed staffing and capacity estimations that take into account Parkinson´s Law (Hasija, Pinker, Shumsky, 2010), (Gutiérrez, G. and Ponagrotis, K., 1991) From a service supply chain standpoint, what we can derive from this law is that people will tend to slow down or otherwise change the speed at which work is performed in reaction to some external cue. If, for instance, goals are too slack, people, other things being equal, might tend to slow down the pace of work instead of taking up more work.

The opposite might also be true. Some authors argue that challenging goals lead to higher performance than easy goals (Locke, et al., 1981). Some authors show that as deadlines approach, people might increase the rate at which work is performed, although with varied results in terms of quality, anxiety, and overall results (Rothblum, et al., 1986).

Unlike in Locke's findings, the System Dynamics literature has provided several examples that show how chronic understaffing and work pressure —which lead to permanently challenging working environments in service systems— may result in long term degradation of service quality as service providers "cut corners" to cope with the increase in work load. Oliva (2001) and Oliva and Sterman (2001) show that attempts to boost throughput and reduce costs in service organizations may lead to unintended consequences in terms of employee burnout and quality decay.

In all, what these studies show is that, unlike processes whose pace is dictated by machines, service operations are staffed by humans who adapt to incentives and circumstances. Such adaptive behavior may take many forms. Human servers in service supply chains may slow down in order to justify the need for some particular job, creating the impression that more time than what is actually needed is necessary to complete a job, as in Parkinson 's Law. Or it may also be true that if a challenging goal is provided, then humans will speed up their pace in order to fulfill a given task within a certain challenging time frame. Or it may also be that if goals are to challenging, pushing people beyond their sustainable physical capacity over the long term, may result in the system experiencing detrimental effects. The important point we make is that people within supply chains will alter the pace at which they work, or use some other equivalent mechanism to leverage work rate to some desired state dictated by environmental circumstances.

Yet, altering work speed is not the only mechanism used to bring work pace to some desired level. Other mechanisms are also used. Some of these mechanisms have been previously reported in the SD literature. For instance, López and Guevara (2009) found that in a criminal justice system judges facing an unexpected growth in the intake of cases to be handled, tended to sharply increase the number of cases dismissed (as a proportion of total cases). In this case, case backlogs were managed by simply taking cases out of circulation, as opposed to the more time consuming task of prosecuting, which entailed additional work and the possibility of the file being returned for clarification or an appeal after its actual dispatching.

In a similar vein Lansing (2001), argued that conviction capacity in courts was adjusted by varying the criteria for prosecution. Her writing is telling. When referring to conviction capacity she says (Lansing, 2001: 3):

> *The mechanism through which conviction rates are adjusted is the threshold for upper court prosecution. Thresholds, which vary based largely on offense type and, seriousness, are determined by prosecutors and achieved through the creation of guidelines (generally informal) that specify the legal criteria that must be met for a case to be prosecuted in the upper court. When capacity is*

strained, thresholds are raised to reduce the number of cases prosecuted. Conversely, when there is excess capacity, thresholds are lowered.

Thus, it appears that human service providers adjust their output capacity to some goal that is set by incentives and some other factors. This capacity adjustment may go both ways. It may be increased upwards, temporarily or not, to cope with an increase in demand. The mechanisms may vary from cutting corners (Oliva) to rejecting cases (López and Guevara) to changing standards (Lansing). It may also be decreased, following Parkinson's Law in the face of falling demand. As servers perceive a falling backlog, capacity may be decreased as they engage in activities that had been postponed during activity upsurges. Thus servers strive to maintain a backlog that is considered to be acceptable. Backlog size is not instantaneously perceived, however, and as backlog grows, they continue to work at accelerated rates until they perceive the backlog under their responsibility is falling to an acceptable standard.

In addition to adaptive human agent behavior (to a localized backlog) human agents may also adjust their effort in accordance to some capacity unbalance present in the system. Lansing (2001) argues that the previously mentioned thresholds, hence capacity, within a particular echelon of the court system are adjusted primarily in response to capacity shortfalls stemming from capacity imbalances.

Despite these behavioral effects that plague the functioning of service supply chains, capacity is often modeled as the number of agents in the system. When considered thus, then all behavioral aspects might be misjudged. Anderson (2001) studied staffing for a one-stage supply chain with experienced employees and apprentices. Capacity, however, was simply measured in terms of the number of experienced workers. In a related paper Anderson et al. (2006) explore staffing policies for a two-stage service supply chain. Capacity here is measured in terms of a continuous function of number of employees, with adjustments for overhead, inefficiency or uncertainty losses, but no allowances are made for behavioral losses or gains. These behavioral issues may exacerbate, within service supply chains, the effects that have been so well documented in tangible goods supply chains, such as the bullwhip effect. Anderson et al. (2005), for instance, show that lead-time reduction in an inventory-less service supply chain, under certain conditions, may actually worsen the bullwhip effect if no information coordination is present. These authors argue that (Anderson et al., 2005: 217): " ...the natural tendency to pursue system-wide process improvement by imposing uniform parameter targets across the supply chain exacerbates demand, capacity, and backlog variances at higher stages. " Simply put, they indicate that in service supply chains it may be preferable to have different parameter targets, in terms, for instance, of acceptable backlogs, according to the location of the backlog within the chain.

We argue that behavioral issues, coupled with temporary capacity imbalances, may dictate the characteristics that a service supply chain may assume in the long run. We look at a service chain in which processing times by human agents are endogenously determined by what constitutes an acceptable and credible backlog, but implicit incentives, particularly within a formal hierarchy,

may also impinge upon throughput rates at certain stages of the supply chain when agents are trying not to overwhelm downstream stations with excess work. Thus, we are interested in the effects of perceived backlogs within service supply chains and also capacity unbalances stemming from such incentives.

Our guiding research questions concern both types of behavioral effects in a two-echelon service supply chain. These guiding research questions are:

1. Is there an effect of perceived backlog upon effort in a service supply chain?
2. What is the system-wide effect of imposing certain parameter targets in some portion of the system?
3. Do imposition of localized parameter targets result in overall system underperformance?

We will argue that during slack demand, human agents will adjust their level of effort according to the perceived backlog. The imposition of certain parameter targets on one stage of the service supply chain may prove to be ineffective unless accompanied by appropriate coordination between stages, and it is highly dependent on the magnitude of demand. We will see that the imposition of one such parameter target, intended to reduce overall backlogs and throughput time in a service supply chain disregarding the rate of growth in demand does not reduce the backlog system-wide. We expected that the imposition of a more strict and rigid parameter target in the upstream stage of the service supply chain, coupled to an increase in capacity at this stage, would lead to backlog growth in the downstream station. Although this happens, it is much less than expected due to the behavioral capacity adjustment that takes place in the upstream stage and the fact that capacity downstream cannot be increased in the short term.

We explore these issues from a purely empirical standpoint by drawing from a case study in a judicial supply chain. The case can be considered a quasi-experiment as, over a long period, we made two interventions in the system; one to establish performance parameter targets in terms of time to complete a task, and one to increase capacity by simply adding bodies. We were able to observe and track the results thereupon.

The paper is organized as follows: in the next section we describe the case study in detail and establish a reference mode; in a subsequent section we develop a dynamic hypothesis; then we model the system. The final section compiles the results of the analysis as we conclude.

Our results are particularly relevant to service supply chains in judicial systems. In many countries the rising crime rates have naturally placed the judicial system under close public scrutiny. The public feels, perhaps rightly so, that a large portion of the problem of rising crime rates can be traced to inefficient courts. Long delays in the resolution of cases and mounting case backlogs add to the generally negative perception that common folk have about courts. If prompt and fair justice is to be attained, courts must improve their performance significantly, but, as we will see, unilateral increases in capacity in some stages of the system, or the imposition of stricter performance targets to other portions of the system, may not be sufficient. As these lines tend to

self-coordinate through implicit incentives and observed outcomes in other parts of the system, the effects of parameter targets may be counterintuitive.

A CASE STUDY IN A JUDICIAL SYSTEM

Justice systems can be understood as service supply chains. Cases flow through these chains, as they are examined by judicial personnel. The justice system is not, however, a simple supply chain. It stands at the core of any democratic system because its mission is to guarantee a reasonable coexistence in society. Its importance cannot be overestimated. In these supply chains users have the right to receive –and justice providers have the responsibility to provide– high levels of quality, efficiency, and transparency.

Long processing times, increased costs, difficulties in access to justice, absence of set organizational and operational patterns for jurisdictional and administrative offices, and process complexity itself, are just some examples of the most common complaints from users about the justice systems. Justice system users have set their perceptions of service delivery based on their personal experiences, information gathered from their surrounding environment, and public reactions to justice system dysfunctions.

As part of an effort to improve performance of judicial systems, we had the opportunity to observe and work with, for an extended period, one particular court within a judicial system. This court, the Second Court of Appeal, was attached as an appeal court to a country's Supreme Court.

The Second Court of Appeal comprised 5 justices, and each one was entitled to work along 2 assistant lawyers, chosen by the Justice himself, with the approval of the Judiciary Superior Council. These lawyers were called advisor/assistant attorneys and were responsible for writing sentence project drafts (tentative sentences to the appeals filed before the court) for the justices.

We compiled information about this court using observation and also monthly reports of data between 2004 and 2010. We were interested in the process performance scenario, and for that matter we collected process input and output rates, work in process build up data, and other figures.

We performed extensive interviews with all personnel involved, Justices and assistant attorneys, and walked through several appeals to understand how the process worked.

From this work we could establish that:

a- The process had performed unevenly during about half of the period under study. In fact, between 2004 and 2007, average WIP had remained fairly constant, but experiencing wild variations around the mean WIP.

b- The rate of incoming work started to steadily increase after 2007, although not entirely beyond intake levels that had been experienced previously. These arrivals were entirely exogenous. The Court could not influence the rate of arrival.

c- After work started to increase after 2007, the court came under increasing pressure to get cases out the door.

d- Several interventions were made to, in principle, expedite the flow of cases. The first intervention was to increase the number of assistant attorneys.

e- The second intervention was to set strict standards to assistant attorneys regarding the dimensions of their legal reasoning, effectively limiting the time they could spend in each case.

f- This created a capacity unbalance, as upstream service stages started to send pro-forma sentences to judges at a higher rate than they could actually process them downstream.

g- Despite this capacity unbalance, backlogs at both stages remained fairly high. Although one would have expected cases to accumulate downstream, they did so at a less than expected pace.

h- We noticed that this occurred because attorneys upstream would try to balance both their backlogs and their bosses. They did not want their own backlogs to fall far from an implicit target level (above or below), but they did not want to send cases down and overwhelm their bosses with too many drafts, so as not to make evident the lack of capacity downstream. Their ability to know exactly the size of the backlogs was, however, limited, and mediated by an important delay in perception formation.

i- As a result, attorneys seemed to regulate the time they devoted to every case, either slowing down (speeding up) the process when their own backlog was too small (large). But at the same time they would regulate their speed according to the size of the backlog downstream, either speeding up (slowing down) when the backlog downstream was too small (large). Notice that the incentives worked in reverse.

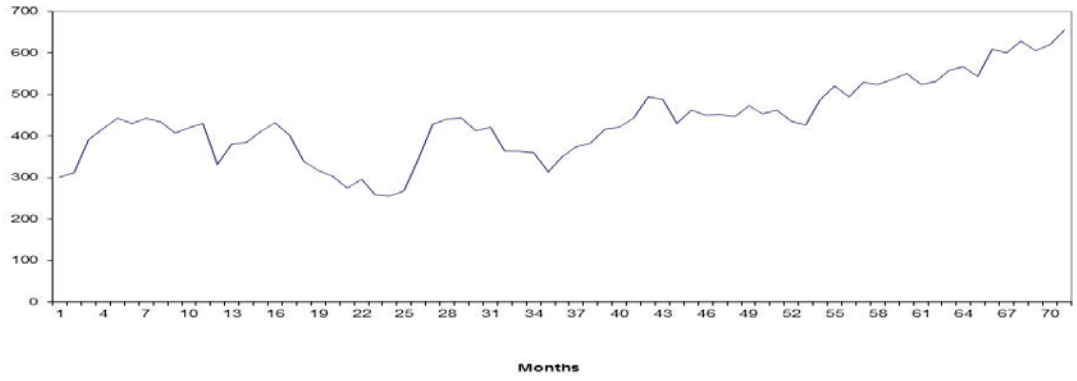
In a regular supply chain, such self-regulation would not have occurred except through the phenomenon of blocking. As no more inventories can be shipped downstream, due to sheer lack of space or some other restriction, then upstream machines become blocked and stopped. In this service supply chain a sort of self-regulating behavior is present, by which upstream providers attempt to balance upstream and downstream backlogs. The reason they do this is because having too small a backlog upstream would signal overcapacity (and, in the long run, the probability of personnel downsizing), whereas a too large backlog would signal ineptitude or laziness, as lack of capacity is rarely accepted as a reason for judicial inefficiency by the press or general public. The downstream backlog was also regulated somewhat because it belonged to the

assistant attorney bosses. Should they load it too much, it would put the Justice in evidence or at least signal lack of effort downstream. It became clear through our interviews that bosses did not like being overloaded with cases, at all.

j- We noticed, finally, that as case intake into the system went beyond some level, then attorneys lost the ability to adapt. This was accentuated by the policy interventions that took place in the system.

Figure 1 presents our reference mode. This figure depicts total cases in circulation, as work in process between the years of 2005 and 2010.

Figure 1. Historical Second Court of Appeal WIP build-up (2005 to 2010)

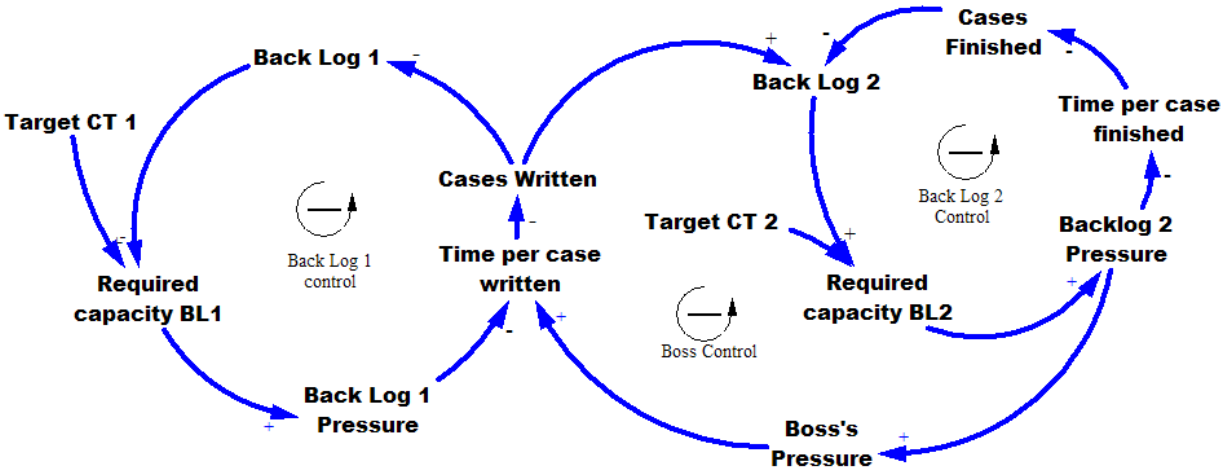


Dynamic Hypothesis

The causal structure implicit in the explanation above is shown in Figure 2. The structure is comprised of one balancing loop that attempts to maintain the initial backlog, called Backlog 1, within some limits. Attorneys are assumed to have a goal for the Target Backlog 1 level. This goal was explicitly stated during interviews. As attorneys solve cases, they also attempt to maintain, through a balancing loop, the downstream backlog within some acceptable level, which, again, was explicitly stated during the interviews that we carried out. Thus, the structure seeks to adjust the state of the system, Backlog 1, as a function of some desired state. When a discrepancy exists between the perceived state and the target state, a corrective action occurs. In this case the connecting action takes place through the adjustment time, as assistant attorneys adjust the time employed per case. The time per case is subject to pressures from downstream the chain. As the perceived Backlog 2 increases relative to some target, there is an incentive to slow down (increase

the time per case) so as not to overload the bosses downstream. This is a balancing loop also, but notice that the effect on time per case operates in the opposite direction to that of Backlog 1. We deemed the combined effects to be multiplicative, and the loop is termed "Boss Control." Justices downstream also adjust the state of the system as a function of the desired state (loop Backlog 2 control).

Fig 2 Causal loop diagram.



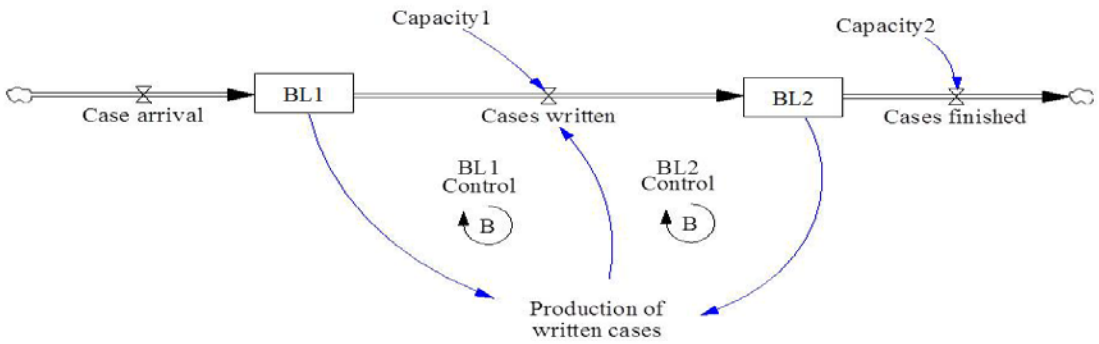
Dynamically, from the standpoint of the agents working upstream, the situation is one in which they want to maintain the system somewhat balanced. If Backlog 1 grows out of control, then they will be tagged as inefficient bureaucrats. If, though, Backlog 1 decreases before a certain level, then the claims will be that they are not doing anything. They, thus, fear both marked positive and negative deviation from a target level.

If, however, they work too fast, their bosses downstream will be overwhelmed with cases. They, hence, regulate time spent per case so as to maintain acceptable levels both upstream and

downstream. This means that at times they will write simple and concise reasonings and at some other times they will write, even for the simplest and most obvious case, reasonings that resemble treatises or ample dissertations akin to legal theses, which, of course, strictly speaking, were unnecessary.

From the causal loop diagram we then built a simple model of this judicial service supply chain. The policy structure looks as in Figure 3. Although from our case study we had a wealth of data, in this preliminary paper we are concentrating on building the simplest possible model that will capture some of the dynamics of interest. We think that there might be some other feedback loops that can plausibly be added to the system, but at this stage we just want to understand the basic dynamics and the effects of the interventions performed.

Figure 3. Policy structure



At the general level of the policy structure shown in Figure 3, several observations are in order:

First, case arrival is treated as exogenous. This is because no control was exerted upon the arrival of cases by the members of this court, as was already mentioned.

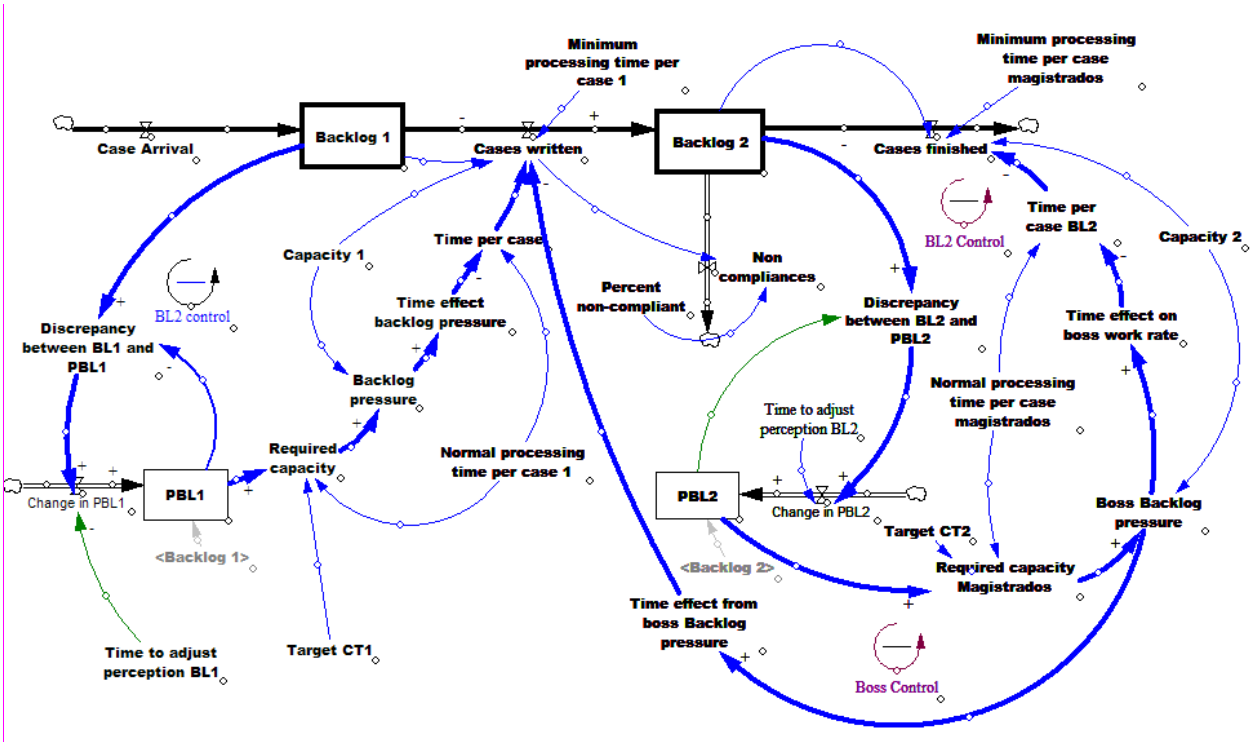
Secondly, the capacity (in terms of number of people) is considered constant and exogenous. This is because, within the time frame of interest, those variables change little. In order to change the number of assistant attorneys (Capacity 1), it is necessary to demonstrate the need. Additional assistant attorneys are rarely assigned to justices. In the Second Court of Appeals there was a sole instance that allows for exploring the dynamics of changing capacity in the upstream portion of the system. We will look at this in the context of our policy interventions.

The number of justices (Capacity 2) is a variable that only changes in the very long term. Justices need to have come through the ranks of the judiciary. They must have a great deal of experience.

Thus, qualified justices are hard to find. In our case study the number of justices remained constant throughout the period. Although a sustained work pressure may induce actions toward increasing the number of justices, these dynamics are likely to happen beyond the time frame we are focusing our study on.

We now look in more detail at the structure associated to the production of written cases. We modeled this as shown in Figure 4.

Figure 4 Structure associated to the production of written cases



Equations are as follows:

Case arrival: an exogenous variable. Cases per month.

BL1: Backglog1. Incoming cases are not processed immediately and accumulate in BL1. BL1 is reduced by the Cases written.

(d/dt)BL1) = Case arrival - Cases written

The Cases written rate is the Capacity1 divided by the actual Time per case1 allocated. In the case of excess capacity, the Cases written are limited by the cases in BL1.

Cases written = MIN((Backlog 1/Minimum processing time per case 1),(Capacity 1/(Time per case * Time effect from boss Backlog pressure)))

Capacity1 is measured in capacity hours provided by assistant attorneys.

Time per case is determined by adjusting the normal time per case (t normal) by the effects of backlog pressure.

Time per case = Normal processing time per case 1*Time effect backlog pressure

Normal processing time per case 1; a parameter.

Time effect backlog pressure: Table function assumed linear. We normalized using (Required capacity-Capacity 1)/Capacity 1 and calculated effects accordingly; for instance:

Effect of Perceived BL2 on time: Table function assumed linear. Constant and increasing with (Required capacity Magistrados-Capacity 2)/Capacity 2.

PBL1: Perceived backlog1.

PBL2: Perceived backlog2-

Target CT1: Parameter. Reasonable expected cycle time for a case upstream. 2 months.

Target CT@> : Parameter. Reasonable expected cycle time for a case downstream, 2 months

(d/dt)PBL1 = Change in PBL1

Change in PBL1 is the Desired BL1 divided by the Time to Adjust Perception of BL2.

Time to adjust perceptions of BL1 and BL2 : Parameter. Time to adjust perceptions. One month.

Change in PBL2 = Discrepancy /Time to adjust

Required capacity = Normal processing time per case 1*PBL1/Target CT1
(similar formulation downstream)

BL2: Backglog2. Cases written are not processed immediately and accumulate in BL2. BL2 is reduced by the Cases finished.

The Cases finished rate is equal to: MIN((Backlog 2/Minimum processing time per case magistrados),(Capacity 2/(Time per case BL2)))

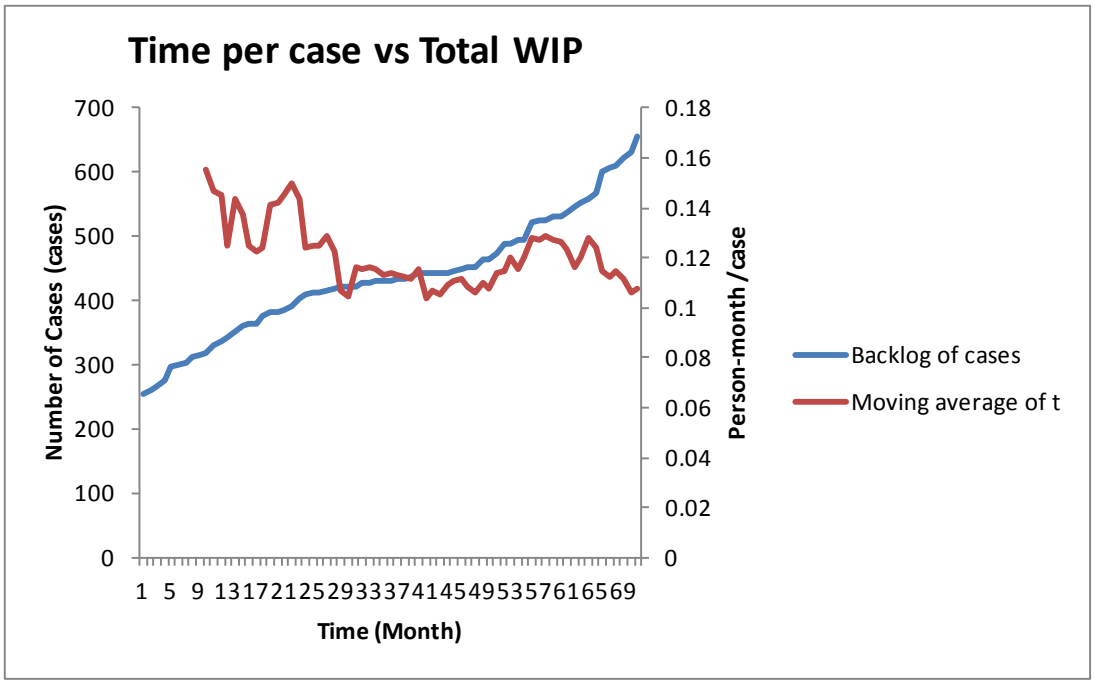
Capacity2: Parameter. It is measured in capacity hours provided by justices.

Some preliminary results

After checking for dimensional consistency, we calibrated the model to actual system parameters, when possible. Parameters were estimated from our directly from our data. The model was thereupon shown to one Justice who was familiar with the situation, as to assess whether the assumptions that had gone into the formulation of our dynamic hypothesis were reasonable. Moreover, we were able to share some of the results of the model which seemed to adjust to behavior, at least qualitatively and to a large extent quantitatively, to the behavior they were experiencing at the Court. We are in the process of building confidence in the model through more validation procedures.

Our data clearly showed that service providers within the system were not operating at a constant speed. They seemed to adjust the rate at which work was performed as a function of the observed backlog. In Figure 5 we show in the same graph the time per case and the total WIP in the system. We observe a declining trend. As there are more cases in the system, it appears that less time is devoted, on average, to each case. Agents, as posited in our dynamic hypothesis, seem to adjust their rate of work with the number of cases.

Figure 5. Time per case variation



We used our model to test this by adding progressively the feedback loops. We show this in a panel of plots in Figure 6.

Figure 6: Model behavior. Solid line is actual data. Dashed line represents model run.

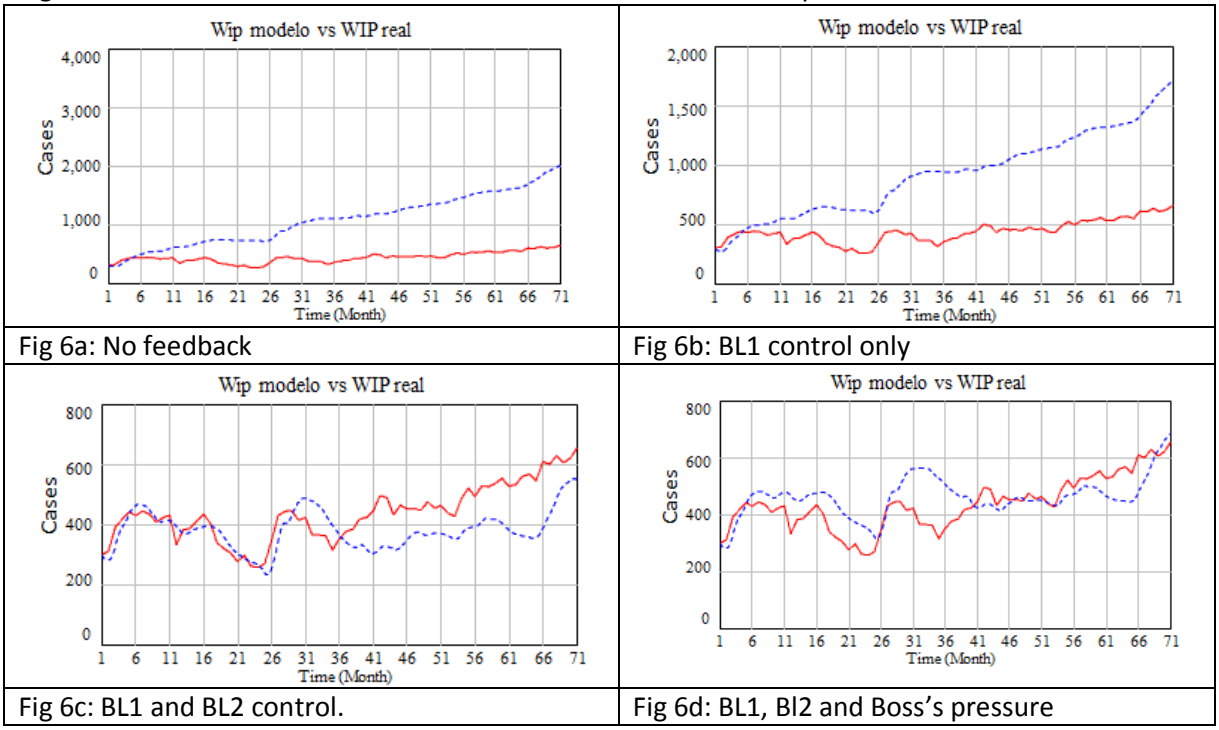


Fig 6a: No feedback

Fig 6b: BL1 control only

Fig 6c: BL1 and BL2 control.

Fig 6d: BL1, Bl2 and Boss's pressure

The model was thereupon used to test policy initiatives. The first policy initiative was actively promoted by one of the Justices. It had to do with restricting the total time that assistant attorneys could devote to a particular case. This meant that the long standing practice of trying to balance the backlog upstream by varying the time devoted per case was, for all practical purposes, eliminated. This measure was actually implemented at the Court of Appeals under the impression that speeding up this process would reduce overall caseload in the system. In practice, the measure appeared not to have any impact whatsoever upon caseload system-wide. The model shows why. By essentially eliminating the feedback loop between BL1, Perceived BL1, and time per case, the assistant attorneys started to work, essentially, at full capacity. They had to adhere to a new standard in practically all cases they reviewed, except for very few exceptions that required more detailed study, were written in a relatively short period of time. The result was that BL1 remained constant or started to decrease, but BL2 started to increase, and the backlog was moved to BL2.

Justices, however, still believing that more capacity was needed, negotiated with the Judicial Ministry the hiring of two additional attorneys. This happened about 60 months into the time series of our reference mode. The result was that cases started to actually leave BL1 at a rate faster than new cases arrived, effectively concentrating practically all inventory of cases in BL2.

This notwithstanding, the addition of two assistant attorneys did not actually result in an increase in output at the lower stage. Output remained the same overall, as the assistant attorneys were only working to maintain some balance in the downstream backlog. If one looks at the average cases per assistant attorney after this intervention, there is a decreasing trend. We show this in Figure 7, which contains actual system data.

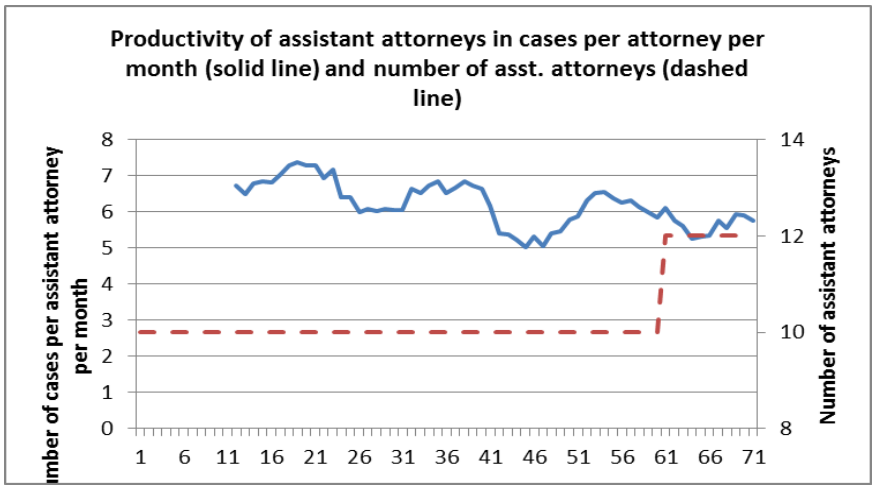


Figure 7. Productivity of assistant attorneys and number of attorneys over time.

The graph in Figure 7 shows the moment of the policy interventions. First two additional assistant attorneys were brought in. At around the same time, a new policy limiting the time to be spent per case was implemented. We can see that productivity per attorney never recovers to its historical maxima. On a per attorney basis productivity only starts increasing after the policy of limiting the time per case goes into place. However, total cases in circulation never go down.

Of course there is an exogenous increase in intake cases that the system tries to compensate through these policy interventions, unsuccessfully. Look at inflows and outflows in Figure 8. The number of cases in circulation does increase in the system, but its balance is altered, as most end up downstream.

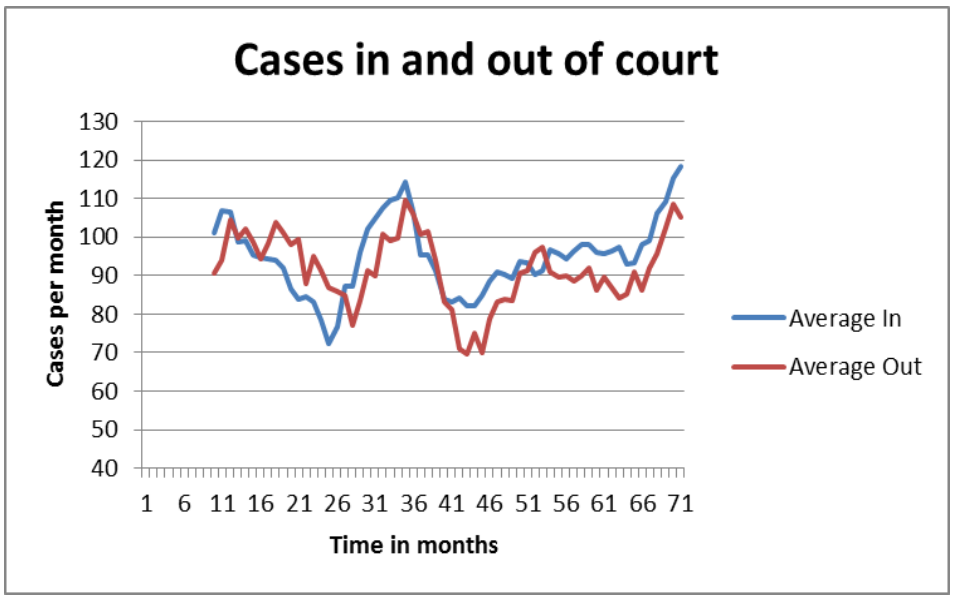


Figure 8: Case inflows and outflows from the system.

The sustained intake of new cases exceeding the total number of cases out causes the number of cases in circulation to increase. Notice however that this increase does not come necessarily from an average increase in cases that exceeds historical maxima. In fact, inventories tend to grow when new cases in is near its historical average. It is toward the end of the series that the intake grows significantly, but, again, not unlike a similar growth already observed in months 26 to 35 approximately. The fact that the number of cases that go out seem to follow the intake line would suggest that people in the system are just adapting, and are able to increase the speed of work as backlogs grow. They, unfortunately, do it with an incorporated delay. The late reaction causes, particularly in the later part of the series, the inventory to grow relentlessly.

A run of the model showed that a combination of eliminating BL1 pressure, eliminating Boss' pressure coupled with a 20% decrease in the average normal processing time per case upstream could reduce the work in process significantly. Figure 9 shows the model run. This policy was implemented in one circuit, as a pilot project, with good results. Overall, WIP went down and delivery delay fell from 8 months on average to an average of 1.5 to 2 months per case. Other measures of quality did not deteriorate.

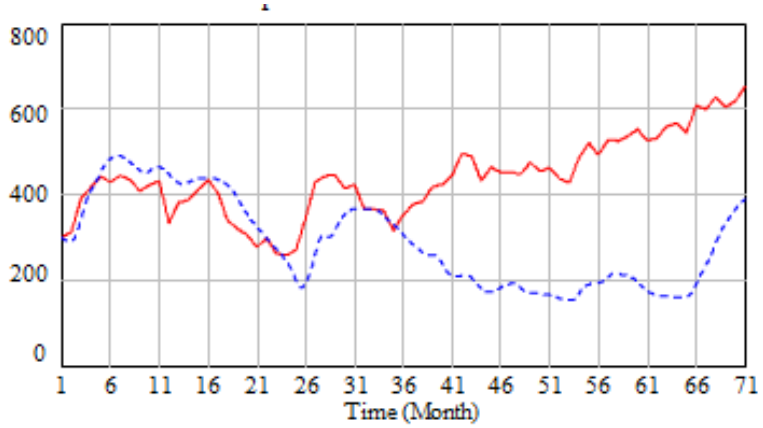


Figure 9: The dashed line shows a model run in which BL1, and Boss's pressure has been eliminated and the time per case has been reduced by 20% upstream.

Some behavioral implications

The results here shown have prompted the Second Court of Appeals to work on a program to attain a system-wide improvement in justice quality. It has been realized that processes must be revised at the Justices' stage, and not just upstream, where most problems were presumed to be occurring.

From a supply chain management perspective, this case serves to illustrate behavioral issues that are commonly present in service supply chains. As we have stated before, capacity in service supply chains is prone to be summarily altered when agents, acting on their own accord or moved by some environmental issue or incentive, may decide to regulate the speed at which work is being performed. This phenomenon has not been widely studied in the context of service supply chains. We believe that human agents will goal-seek, moving towards an implicitly stated goal utilizing for that purpose any means available. Judicial service chains are very good laboratories to observe some of these phenomena, as human servers not only work at varying speeds as a function of some perceived amount of work, but also use other mechanisms, like rejecting cases, changing thresholds for prosecution, varying acquittal rates, and other similar measures.

From a public policy perspective, our findings have an implication upon the notion of prompt and fair justice; as we see that the concept is driven primarily by agents within judicial systems who are just following their own agendas and behaving opportunistically. If extrapolated system-wide, there are about 800 hundred courts similar to this one in the country. The solution to the ever increasing time per case has traditionally been to add people to the system. The number of people in the system doubled in the last 9 years. This has only served to exacerbate the opportunism of agents and to increase costs, as no reductions in the time per case have been observable at the system level. Our research shows that there might be opportunities to improve the system without making any important investment.

Here we have presented just some preliminary findings of what is an ongoing research project. We look forward to work some more in formalizing and refining our simulation model. As we explore the data we have collected, we have started to unveil additional feedback loops that will hopefully shed some light on how humans work within service supply chains.

References

Aksin, Z., Armony, M., Mehrotra,V.(2007). The Modern Call Center: A Multi-Disciplinary Perspective on Operations Management. Research in Production and Operations Management. 16, 6, 665-688.

Anderson. E. (2001). The nonstationary staff-planning problem with business cycle and learning effects. Management Science, 47, 6, 817-832.

Anderson, E. G., Morrice, D. J.,Lundeen, G. (2005). The "physics" of capacity and backlog management in service and custom manufacturing supply chains. System Dynamics Review, 21, 3, 217-247.

Gutierrez, G. and Kouvelis, P. (1991). Parkinson's Law and its implications for project management. Management Science, 37, 8, 990-1001.

Hasija, S., Pinker, E., Shumsky, R.A. (2010). Work Expands to Fill the Time Available: Capacity Estimation and Staffing under Parkinson's Law. Manufacturing and Service Operations Management, 12, 1,1-18.

Jochimsen, B. (2009). Service Quality in Modern Bureaucracy: Parkinson's Theory at Work. Kyklos, 62, 1, 44-64.

Lansing, S. E. (2001). Estimating the Effect of Targeted Enforcement Strategies on Conviction and Imprisonment Rates for new York City. Proceedings of the 19th. International Conference of the System Dynamics Society. Atlanta.

Locke, E. A.; Show, K. N.; Saari, L..M., La Cham, G. P.; (1981). Goal Setting and Task Performance. Psychological Bulletin, 90, 1, 125-152.

López, L. and Guevara, P. (2009). Judicial Process Dynamics, Conference Proceedings, The 27th International Conference of the System Dynamics Society

Oliva, R. (2001). Tradeoffs in responses to work pressure in the service industry. California Management Review, 43, 4, 26-43.

Oliva R. and Sterman, J. D. (2001). Cutting corners and working overtime: Quality erosion in the service industry. Management Science47, 7, 894-914.

Parkinson, C. N. (1955). Parkinson's Law. The Economist. November.

Rothblum, E. D., Salomon, L.J., Murakami, J. (1986). Affective, Cognitive, and Behavioral Differences Between High and Low Procastinators. Journal of Counseling Psychology,33, 387-394.