Throbbing Between Two Lives: Resource Pooling in Service Supply Chains

Kim van Oorschot
BI Norwegian Business School
Department of Leadership and Organizational Behavior
Nydalsveien 37
0484 Oslo, Norway
kim.v.oorschot@bi.no

Yan Wang
Delft University of Technology
Faculty of Technology Policy and Management
Jaffalaan 5
2628BX Delft, The Netherlands
y.wang-16@tudelft.nl

Henk Akkermans
Tilburg School of Economics and Management
Department of Management
Warandelaan 2
5000LE Tilburg, The Netherlands
ha@uvt.nl

Abstract

Resource pooling is known to benefit performance through reduced congestion, but primarily in settings with homogenous demand. In settings where demand is heterogeneous, pooling can be counter effective. The effects of pooling of staff when demand is heterogeneous and dependent are not known. We present a simulation model based on a service supply chain that delivers Interactive TV to customers. Customers expect high performance in terms of innovativeness and reliability. Based on the results of simulation analysis, we find that when target innovativeness of the service is increased, pooling outperforms not pooling, but the delays that are involved with pooling will make the system and hence its performance unstable. Stable and high performance can be realized through "unbalanced" hiring. This means that a target performance increase in the upstream stage of the chain (innovation), is accompanied by hiring staff in the downstream stages of the chain (QA and operation).

1. Introduction

In service supply chains, a high degree of innovativeness often needs to be combined with high reliability of the service. For example, interactive TV (ITV) is a service that is in a race with a broad array of competitors aiming to win customer preference: cable companies, content providers, YouTube, Apple TV, pirate websites offering live transmissions, Pay-tv channels, etcetera. *High innovativeness* in ITV is therefore required to offer similar functionality or lose its appeal with today's fickle customers. *High reliability* is required, because minor blips in the quality of the transmission, which would go unnoticed during an Internet browsing session or even in a telephone conversation, can already lead to customer complaints. Of all the services provided through copper wire or fiberglass, TV is arguably the one with the highest required reliability.

In most business models, high innovativeness and high reliability form an either/or decision: or you get the latest and newest, but you are bound to suffer "teething problems". The difference between leading edge and bleeding edge tends to be small. But in the service supply chain setting described above, this either/or has to be a both/and. A very innovative service which is however not reliable, in the sense that customers frequently experience performance issues, is just not acceptable to the market. Both processes (innovation and operation) require capacity which is usually constrained. Dividing limited human resources between innovation and operation poses specific challenges on the management of these service supply chains and this research zooms in on these challenges, and on how they can be overcome.

Pooling resources to serve homogeneous demand is known to reduce congestion as measured by the expected time spent by customers waiting to be served (or time needed to execute innovations, fix incidents) [1]. In practice however, demand tends to be heterogeneous, in which case, these benefits are not guaranteed [2]. There may be situations where the pooling of customers actually adds variability to the system which negatively impacts performance [3].

To avoid these negative effects, pooling can be used in emergencies only, when congestion or waiting times exceed a certain threshold. This is also known as congestion-based staffing. This policy has shown promising results in border-control stations [4] and in settings with heterogeneous but independent demand streams [5].

However, we do not know what pooling means in the context of dependent heterogeneous demand, which is encountered in the ITV case. In the ITV service supply chain, the quality of the output of the

URI: http://hdl.handle.net/10125/50024
ISBN: 978-0-9981331-1-9

upstream stage (innovation) will determine the demand and workload in the downstream stage (quality assurance and operation). High quality of innovation prevents problems later on in operation, in terms of reliability of the service. Customers of the ITV service want to have both: an innovative and reliable product.

To analyze this question, we develop a simulation model of a typical service supply chain, based on data gathered from a global player in the telecom industry. With this model we analyze different staffing scenarios to deal with trade-offs between service innovativeness and service reliability in the ITV service, and thereby contribute to theory about pooling in settings with dependent heterogeneous demand.

In section 2 we will analyze existing literature and formulate the research question. Section 3 describes the method, followed by the presentation of simulation results. Next, the results are discussed and contributions to theory and practice are presented.

2. Overview of literature

2.1. Capacity management in service supply chains

A service supply chain is defined as a network of interactive service processes [6] which often has a dynamic and nonlinear structure [7]. Capacity management in such supply chains needs to cope with the bullwhip effect that comes from the unexpected capacity assumption, the managerial and customer behavior, and the visibility and sharing of information across the entire supply chain [8]. This amplification effect adds extra pressure in capacity management. The dilemma between achieving high level innovativeness and reliability [9], [10] is especially hard to handle in ICT service enabled supply chains where the business processes are highly automated [8], [11], or when service quality is eroded due to the vicious cycle created by the heterogeneous influential factors in service supply chains [12]. Like in ambidextrous organizations, in which exploration and exploitation activities need to be balanced [13], [14], these service supply chains need to balance human resource requirements for both innovation and operation.

2.2. Pooling defined

Resource pooling refers to an arrangement in which a group of common resources or servers is used to fill demands of multiple customer streams rather than dedicated, separate resources for each individual customer stream [15]. The objective is to yield operational improvements, which implies that pooling may achieve lower congestion (shorter waiting times) than a number of decentralized (unpooled) resources that focus on a limited range of customer streams [1], [2]. This advantage is due to the portfolio effect which reduces variability [16]. With service systems working separately a customer may have to wait for a server while another server is idle —a situation that does not occur in the pooled system [15]. In other words, one large agent or resource group is more efficient than separate ones by the rationale of load balancing [3].

2.3. Boundary conditions of pooling

Pooling resources is known to improve performance when demand is homogeneous. However, when demand is heterogeneous, for example when different types of customers need to be served or different kinds of activities need to be performed, the advantages of pooling are not guaranteed [2], [17]. In fact, when faced with a mix of different types of activities pooling might not even be profitable at all because pooling increases service variability [3], [18]. By pooling two separate servers, one that was originally performing activities of type A and one dedicated to activities of type B, extra service variability may be brought in as a next service request at one, and the same server can then be either of type A or B. By assuming that the Pollaczek-Khintchine's (PK)-formula (a formula that states a relationship between queue length and service time distribution) also applies (approximately) for a two-server system, by virtue of the PK-formula this extra variability may lead to an increase of the mean waiting time [3].

Besides the mix of activities, previous research has identified other situations where pooling may actually add variability to the system and reduce performance. Pooling may decrease efficiency and increase risk when pooled servers are subject to failures (in which case the customer is preempted and placed back in the queue) [17]. Unpooled resources are preferred when the target performances of customer types differ [2]. Furthermore, these authors note that pooling requires servers to be able to accommodate various types of demand. This flexibility may be expensive and as such reduce the efficiency of the service system. Finally, pooling may also increase job setup times and/or require larger job batch sizes which may reduce the effectiveness of pooling [19].

2.4. Congestion-based staffing

To take advantage of the benefits of pooling while moderating its possible negative effects, temporary

pooling or congestion-based staffing can be a solution. This is a staffing policy where the number of servers is adjusted according to the queue length during a planning period [4]. In border-control stations between the USA and Canada, congestion-based staffing has shown to control the mean and the distribution of queue length and the expected frequency of changing staffing levels. Furthermore, it improves the server's utilization level [4]. A congestion-based switching policy has also revealed benefits for companies offering make-to-stock (MTS) and make-to-order (MTO) products through different sales channels [5]. Here, a static approach is defined, that separates a facility into two independent units, with each unit having its own distinct demand (MTS or MTO) and the responsibility for meeting that demand. This static approach is compared with a dynamic one, that consists of a hybrid MTS-MTO facility, with, in addition to machines dedicated to either MTS or MTO production, a group of flexible machines which can switch between production of MTS and MTO products. The authors find that the dynamic approach generally outperforms the static one, particularly when traffic intensity is high in both the MTS and MTO operations. So, this approach is an effective way to cope with two streams of demand: one for standardized products and the other for customized products.

2.5. Research question

The MTS-MTO setting described above resembles a service supply chain that also needs to cope with at least two streams of demand: one for innovation and the other for operation. However, the difference between the MTS-MTO system analyzed in [5] and the innovation-operation service supply chain is that the demand streams are assumed to be independent in the former, where they are dependent in the latter. The higher the quality of the innovations that are introduced to customers, the less problems customers will have with the service, and as a result, the lower the workload in operation will be to resolve customer issues or incidents. Having dedicated resources (no pooling) that focus *either* on innovation *or* on operation for a service that needs to perform high on *both* innovation *and* operation (reliability) can be a safe choice. This is because having resources solely dedicated to innovation may prevent bugs that could cause problems for customers later on. Although preventing errors is usually not an approach that pays of on the short-term, the effects on the long-term are positive [20], [21]. However, when a problem does occur, this usually requires a huge (temporal) peak in resources dedicated to fixing problems and making the service reliable again. Not having these resources in place will lead to long delays for customers (waiting for the incident to be resolved), not to mention the devastating effect on company reputation. So, having the ability to quickly move resources from innovation to operation in case of a major incident (pooling or congestion-based staffing), may also pay off.

This service supply chain setting is one in which at least one stream of demand is dependent on the other: it is influenced by the performance in which one of the other stream(s) of demand is dealt with. To our knowledge, the effect of resource pooling in such a setting has not been analyzed before, and we do not know whether or not pooling helps to improve both the innovativeness and the reliability of the service. Therefore, we formulate the following research question: *in a service supply chain in which the performance of the upstream stage (innovation) determines resource requirements in the downstream stage (operation), what is the impact of pooling of staff in these stages?*

3. Research method

3.1. Case study and simulation

The research conducted in this report fits well in the field of action research [22], which aims to gain in-depth understanding of and find solutions to the resource allocation problem in service supply chains with dependent demand streams. Having this purpose in mind, the authors followed an inductive case study approach [23] in elaborating analysis steps.

A global player in the telecom industry allowed us to analyze its service supply chain consisting of innovation and operation in Interactive TV (ITV) services. Eleven Interactive TV (ITV) incidents were selected to examine the relationship or dependency between the occurrence of these incidents and number of new innovations introduced to the ITV service. The analysis adopted the so-called triangulation [24] of multiple data collection methods. The information sources in this research include 18 semi-structured interviews with ITV experts, a model-building workshop [25], [26] with ITV managers and ITV service historical data. All the interviews and the workshop were recorded. The incident samples are firstly examined in a fact-based analysis, as it ensures to focus on the correct root cause of the problems and helps to get to the best potential solutions. In this fact-based analysis, the incident samples are analyzed according to their lifecycle phases in the incident handling process.

Simulation is a powerful methodological approach in theory building. Especially, when the inductive case method is constrained by limited data, 'simulation can mitigate the weaknesses by exploring, elaborating and extending simple theory that is produced by this theory-creating method' [pp. 495, 27]. To further analyze the root causes and seek for solutions, a system dynamics simulation model is built for scenario testing in ITV service.

3.2. Case setting

The organization offering the ITV service is a leading telecommunications and ICT service provider in Europe, offering wireline and wireless telephony, internet and IT to consumers, and end-to-end telecommunications and ICT services to business customers. Interactive television (ITV) is an innovative service solution that adds data services to traditional television technology. It provides customers with high level of interactivity with television, so that customers can order, rent, record or replay their preferred or missed programs, and also watch them online via laptops, tablets and smart phones. The programs can be chosen from 60 TV channels, 11 high-definition channels, and 90 radios in digital quality. The organization has successfully integrated internet and TV. The customer response to ITV has been very positive, and the subscriptions have been increasing steadily.

Behind the big success, one of the most complex networks for delivering the ITV service can be found. In general, the delivery network includes the content broadcast network, the ITV platform, the internet service provider infrastructure, and internet networks. All the functional components of the ITV service, such as the video on demand and network personal video recorder service, are managed in the ITV platform. The ITV signals are transmitted to the settopboxes at clients through IP routing and broadband networks.

3.3. Innovations and incidents

Incidents are unplanned interruptions or reduction in quality of IT services [28] and are often the results of customer complaints (calls), system failures or errors in the network operation. Unknown errors or undetected bugs in the software may be the root cause of incidents, and initiate the request for changes in the managed services and network. More incidents may help to discover and identify problems, while resolving the problems may help reducing the occurrence of incidents.

Incidents occur when the promised services do not function as expected, which cause disruptions or reductions in the quality of services. In the ITV case, customers are the ones who directly perceive the impaired services. Their reaction (e.g. calling) to the service provider is one of the major indicators for measuring the incident impact. The organization's network operation center provides monitoring on the status of entire service supply chain. In addition, the ITV operations also have system level monitoring on the ITV product. These monitoring systems send alarms to the operations teams when there is any disrupted service activity sensed. Incidents are usually reported via customer calling or network monitoring alarming. Once an incident is reported, the operations teams are informed and start to restore the disrupted service. The incident fixing process may include steps of analyzing the affected service samples, identifying the its possible causes, estimating and checking the impact at customer base, proposing and applying proper solutions, meanwhile maintaining the communication with other involved parties and customers. Some of these steps may be taken simultaneously, and the procedure of trying possible solutions is usually iterative during the incident fixing process, until the situation is back into control.

The causes of incidents are diverse, e.g. software bugs, human errors or accidents out of planned changes, and the discovery of these problems is very situational. Due to the iterative process and the possible change of actual customer impact, the moment of identifying the severity of the incident varies. ITV incidents were analyzed over the period 2010-2013. Starting from May 2011, serious incidents occurred at a higher rate and reached its peak in Q3 2012. Therefore, 11 of these serious incidents were chosen from this certain period as research samples, including three incidents from 2011 and eight cases from 2012. Regarding the root causes of the selected 11 incidents, 6 cases were from innovations, 3 from human error, 1 from technical error, and 1 from maintenance change. This shows that innovations are the most common cause of incidents.

3.4. Example of congestion-based staffing

As indicated, from 2010 until Q1 2011 huge amount of incidents were reported, but they did not all turn into serious cases due to the comparably small installed base at customers. From Q2 2011 onwards until Q3 2012, more and more issues occurred in ITV service, which was due to an increasing number of both installed base and changes made by innovations. Following the sharp increase of changes in August and September 2012, the call ratio also reached its peak in October 2012. Then, a revised policy was carried out in ITV, which prioritized problem management over new innovations and combined resources from both

operations and innovation in making preventive management. This gave the operation team more room with solid fulfillment in both regular administration and incident management. The impact of this policy change had immediately been reflected in terms of call ratio, which gradually declined since December 2012.

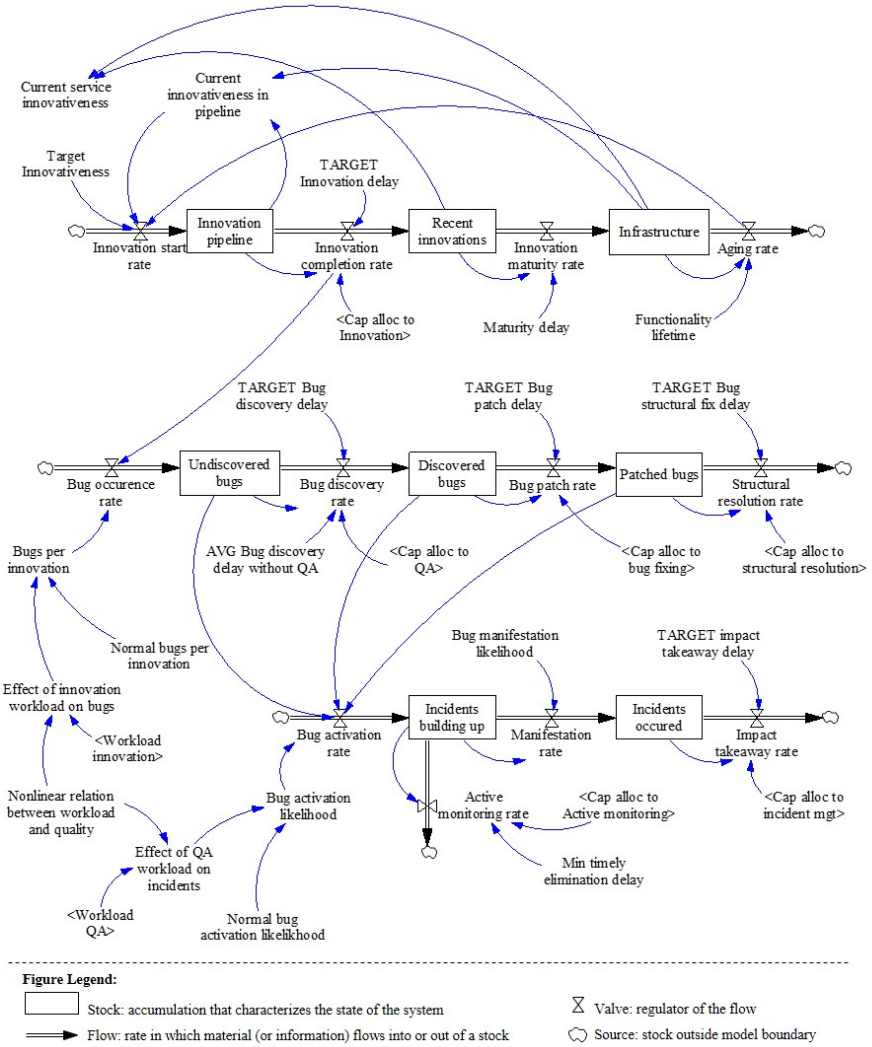
3.5. Model description

System dynamics simulation is chosen to model the ITV service supply chain. This is because, despite of the fact that the incidents are discrete events, the heterogeneous demands from both incident fixing and innovation are translated into workload which are perceived by staff continuously. In other words, the capacity provisioning is long term and continuous [29]. Therefore, it is appropriate to adopt a continuous simulation method.

The issues between innovation and incidents seem to be similar to the DevOps gap in IT development [30]. However, the ITV case contains a broader scope than normal IT service development and covers more complex and coherent factors than the ones covered by DevOps solutions [31], [32]. The activities in the ITV service supply chain are grouped into three parts, namely ITV Innovation, ITV Problem Management, and ITV Operation. In Figure 1, these three groups are represented as three horizontal groups of stocks (rectangles) and flows (double arrows with valves) that are connected with each other via causal links (single arrows) and other variables. (Note that we have not included all variables in Figure 1. More insights and the complete model can be found in [33].)

ITV Innovation is responsible for the development of ITV services that includes adding new functionality and improving existing features in ITV products. Driven by market trends, new innovation projects are initiated and carried out by innovation staff. The organization's management team sets the target level for the innovativeness of the service which drives the number of new innovation projects that are started. Innovations that the staff currently works on are gathered in the stock "Innovation pipeline". The completion rate of innovation projects depends on the time it takes to execute a project and the available capacity to do these projects. Completed projects flow into the stock "Recent innovations". After a maturity delay these recent innovations become part of the "Infrastructure".

ITV Problem Management takes care of both technical and non-technical issues in ITV services. Problems are identified from bugs or issues that can potentially influence ITV service performance. Bugs typically arise during new innovation projects. At first these bugs are undetected. Bugs can be discovered either through quality assurance (QA) or more or less spontaneously (this does not require capacity,



"Figure 1. Main stocks, flows, and variables of ITV Innovation, Problem Management and Operation"

but takes more time than with specific QA efforts). Once bugs are discovered, they need to be fixed as soon as possible to prevent these bugs from becoming incidents that will be noticed by customers. Bugs with a quick fix flow into the stock "Patched bugs". Finally, a structural solution for the bug is developed and implemented which removes the bug out of the service supply chain completely.

ITV Operation is responsible for the reliability of the service through maintenance and incident fixing. Maintenance provides monitoring at system and network level and regular maintenance of configurable items. Through active monitoring potential incidents can be discovered and resolved before the customer discovers them. These incidents come from bugs in the software of new innovations. Incidents that remain undetected by the operation staff can manifest themselves to customers. Incidents that are currently occurring require highest priority of the operation staff. The longer it takes to resolve these incidents, the longer customers have a problem with the service (low reliability) and the higher the impact will be on market reputation of the service.

Each of the three groups in the ITV service supply chain initially has dedicated resources: innovation staff, QA staff, and operation staff (note that these are not shown in Figure 1). New staff members can be hired from outside the organization. But, in case of high workload in any one of these three groups, staff can also be transferred from one group to the other. High workload means that the available staff is lower than required, which will lead to congestion (it will take more time to complete innovation, detect bugs, fix incidents, etc.). As such, using workload as an indicator for transferring staff from one group to the next, can be considered as congestion-based staffing.

3.6. Independent and dependent variables

The independent variables in our model, the variables that we will use to define and analyze different staffing scenarios, are:

- Target innovativeness: all simulation scenarios will start in equilibrium, which means that the ITV service supply chain is completely stable. There is no congestion, so no extra staff is required. This equilibrium arises when the level of target innovativeness is 0.3 (on a scale of 0 to 1). To evaluate the effect of different staffing scenarios, we will simulate a step increase of this target innovativeness at week 50, from 0.3 to 0.4 (33% increase).
- Pooling between resource groups: if pooling is allowed, the transfer rate between innovation and QA and between QA and operation is not equal to zero. When the transfer rate is positive, staff flows from left to right (e.g. from innovation to QA). When the transfer rate is negative, staff flows from right to left (e.g. from QA to innovation). The workload in the groups determines the need to transfer staff (congestion-based staffing). Highest priority is given to operation, since any congestion here will be directly noticed by customers.
- Hiring of new staff members: besides pooling as a way to increase staff for a short term, we also consider the possibility to hire new staff members in any of the three groups. If hiring is allowed, the hiring rates will be positive. In the scenarios without hiring, the hiring rates will be zero.

The dependent variables in our model, i.e. the variables that we will use to compare the performance of different staffing scenarios, are:

- Market reputation: this variable measures how the market evaluates the overall performance of the ITV service. This performance is divided in both innovation and reliability. Service innovation is determined by the number of recent innovations compared to the existing infrastructure of the ITV service, and compared to the standard innovativeness in the market. Reliability is determined by the number of incidents that are occurring compared to the number of incidents that are occurring on average in the market.
- Workload of innovation and QA staff: whereas market reputation is a variable that focuses mainly on the output of the service, the workload of these two staffing groups tells us something about the costs of realizing this output. High output combined with an extremely low workload may indicate that the organization hired too many people which will lead to high resource costs. On the other hand, workload has important side-effects. The higher the workload of innovation staff, the lower the quality of the innovation projects (more bugs in the software). Likewise, a high workload for QA staff will lead to a higher bug activation likelihood, which in turn increases the number of incidents building up (see Figure 1 for these causal effects). So both low and high workload can lead to high costs for the organization. A workload close to 1 seems therefore preferable.

4. Simulation results

Initially, we assume that the system is in equilibrium. This means that the ITV service has stable behavior, with respect to all variables. The

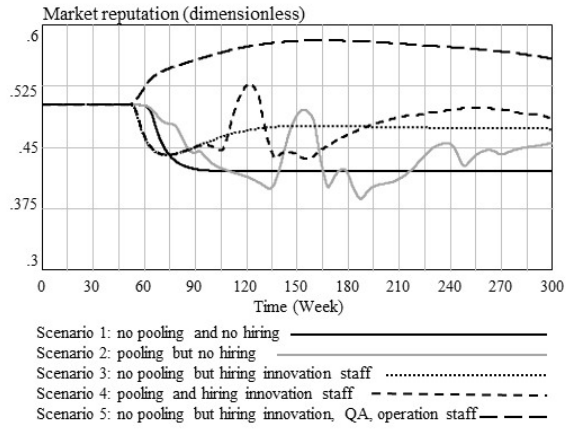
workload in all groups is stable and equal to 1, market reputation is stable and equal to 0.5 (on a scale from 0 to 1). The organization's management team wants to increase its market reputation by increasing innovation while maintaining reliability. Therefore, in week 50, the level of target innovativeness is increased from 0.3 to 0.4. The following staffing scenarios are considered:

1. No pooling and no extra hiring. In this scenario only the target innovativeness is increased in week 50. Everything else remains the same. So, no staff is transferred from one group to another. No new staff is allowed to be hired.
2. Pooling but no extra hiring. In the second scenario, also target innovativeness is increased but now pooling (based on workload) is allowed between different resource groups. No new staff members are allowed to be hired though.
3. No pooling but extra hiring of innovation staff. In this third scenario the innovation group is allowed to hire extra staff to do the extra work that is caused by the increased target innovativeness level. Because this innovativeness level is increased with 33%, also 33% extra staff is allowed to be hired in the innovation group.
4. Pooling and extra hiring of innovation staff. This scenario is like the previous one in which 33% extra staff is allowed to be hired in the innovation group to execute the extra innovation work. However, now, staff can be transferred to other groups when necessary (when workload/congestion is too high).
5. No pooling but extra hiring of innovation, QA, and/or operation staff. In this scenario 33% of total staff is allowed to be hired as a response to the increased target innovativeness. In this scenario the extra staff can be hired in any of the three groups, so not only innovation. We let the simulation model find the best hiring mix based on an objective function that maximizes market reputation over the entire simulation length (300 weeks).

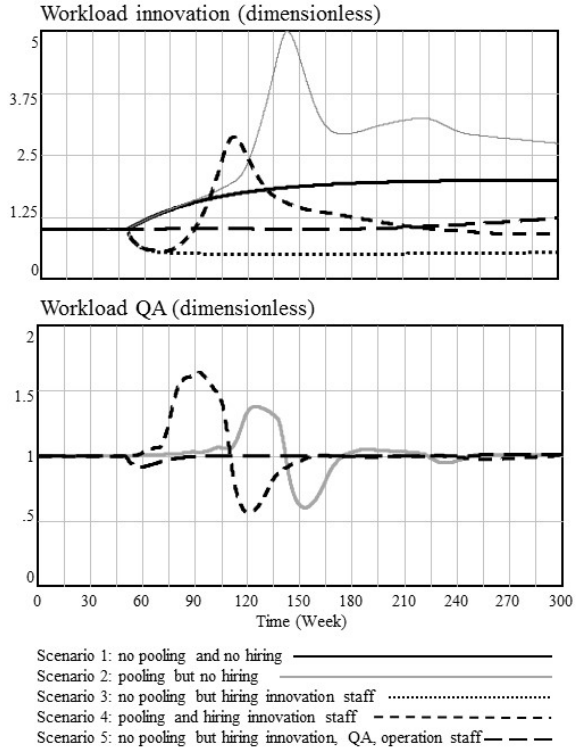
The simulation results of these five scenarios with respect to market reputation are shown in Figure 2. Furthermore, in Figure 3 the workload of the innovation and QA staff is depicted. Both market reputation and workload are modeled as dimensionless variables. Market reputation can range from 0 to 1, in which 0 reflects a very bad reputation and 1 reflects a very good reputation. Workload has a lower bound of 0. This value reflects that staff is idle. A workload of 1 means that the resource requirements are exactly equal to resource availability. A workload of, for example, 2 means that the resource requirements are twice as high as resource availability. Please note that Scenario 1 and 3 caused the QA workload to increase to such high levels that these scenarios are not included in Figure 3 (this makes it easier to compare the workload in the remaining three scenarios). In the next section we will discuss these results.

5. Discussion of results

In this section we will discuss the results of five different staffing scenarios that can be used when the service organization decides to increase its target innovativeness in order to increase its market reputation.



"Figure 2. Market reputation in 5 staffing scenarios"



"Figure 3. Workload in 5 staffing scenarios"

In Scenario 1 only the target innovativeness is increased and no changes are made with respect to staffing. The number of staff members that are present in each group (innovation, QA, operation) remain the same. These staff members do not switch between groups. Since the workload was already at level 1 before the target increased, this increase can only cause the workload for the innovation staff to accumulate further. This leads to an increase of bugs in the new innovations, which causes a huge increase in the workload for QA and finally for operation as well, because most of these bugs will cause incidents and a decrease of service reliability. Though the target innovativeness is increased, the available innovation staff can only deliver the same number of innovations as before. But, due to the huge workload, the quality of these innovations is so poor that service reliability suffers greatly. As a result, the overall market reputation goes down.

The first scenario teaches us that asking staff to do more in the same amount of time does not work. Scenario 2 tests whether it helps to move staff around (congestion-based staffing) whenever possible and required, without actually hiring new staff. This scenario performs indeed somewhat better than the first scenario and has in fact been implemented in reality in this ITV case. This meant that innovation staff was pooled with operation staff to help with service restoration after a severe incident occurred. The ITV management team confirmed the positive effect of this scenario, that it helped to reduce the pressure from the operations staff during incident fixing process. However, oscillations were also mentioned, like in our simulations, so pooling may destabilize the system for a certain time. The most important reason for these oscillations is the delays that are involved in transferring staff from one group to the next. Before the switch, staff members need to finish what they were working on, they may have to move to another building, they need to figure out how to help their colleagues in the other group, etc. So, there is a set-up time involved. By the time the staff members are fully productive in their new group, another group maybe starved for resources, and a new transfer may be required. But on average, the market reputation is higher than in Scenario 1.

Apparently, pooling staff causes instability and oscillations in innovation and reliability. Therefore, Scenario 3 tests whether it would be better to hire extra innovation staff (to deal with the extra workload that is caused by the increased target innovativeness). Pooling is switched off again in this scenario, to analyze the effect of hiring only. Right after the target is increased (in week 50), we see that market reputation decreases fast in Scenario 3. The innovativeness does increase, thanks to the new staff members, but now quality assurance suffers because of the sheer number of new innovations that need to be checked. Because the QA staff is not increased the workload for QA explodes (which is why this graph is not included in Figure 3) and as a result service reliability suffers greatly. So, although market reputation is better than in Scenario 1 and 2, it is worse than it was before the target was increased.

In Scenario 4 we analyze the effects of pooling again, but this time first new innovation staff members are hired. However, after they are hired, they can be transferred to other groups when necessary. Again, like in Scenario 2, we see that pooling causes instability and oscillations. However, on average, the market reputation is somewhat higher than in Scenario 3, so pooling does help. We also see that the workload for innovation and QA staff eventually balances out again at the level 1. Note that we have tested whether the instability and oscillations are simply caused by the length of the transfer delay of staffing in the two pooling scenarios (scenario 2 and 4). Therefore, we have simulated these scenarios with transfer delays of 1, 6, 12, and 24 weeks, in which 12 weeks is our default value. Although the oscillations are reduced for shorter delays, the market reputation on the long term is not improved.

Scenario 4 performs clearly best when compared to the other three scenarios. However, the instability and oscillations that accompany pooling are undesired. Therefore, we have tested one last scenario in which pooling is switched off again (not allowed), but here hiring is allowed at all groups. We let the simulation model find the best hiring policy by giving it the objective to maximize market reputation over the 300 weeks of the simulation run length. The maximum number of staff members that can be hired is determined by the target innovativeness. Since this is increased by 33%, the total number of staff members can be increased with 33% as well: from 38.2 to 50.9. The graphs in Figure 2 and 3 depict that Scenario 5 outperforms Scenario 4. Not only is performance (in terms of innovativeness and reliability) stable, it is also much higher than in all other scenarios. Furthermore, the workload for all groups is under control and around 1. Table 1 presents how many staff members were hired in this scenario.

"Table 1. Staffing in Scenario 5"

	initial # FTE	final # FTE	increase in %	increase in # FTE	5% conf. bound in # FTE
Innovation	8.0	10.2	27.5%	2.2	0.0 - 3.4
QA	26.0	33.2	27.5%	7.16	6.1 - 8.6
Operations	4.2	7.6	81.0%	3.4	1.9 - 4.4
Total FTE	**38.2**	**51.0**	**33.4%**	**12.76**	

The numbers presented in Table 1 reveal an interesting finding. Although the aim is to increase innovativeness (while maintaining reliability), it is not the innovation group that requires most resources after the target innovativeness is increased. Most resources are needed to deal with the "side-effects" of innovations: quality assurance, preventive maintenance (monitoring, preventing incidents from happening) and corrective maintenance (fixing incidents). Although the workload for the innovation staff is under control (which prevents an increase of bugs due to stress), the number of innovations increase which per definition will increase the number of bugs in the ITV service. More staff in the downstream stages of the service supply chain help to discover and correct these bugs before they become incidents that reduce the reliability of the service.

6. Contributions

6.1. Theoretical contributions

In this paper we have analyzed a service supply chain with dependent heterogeneous demand streams and the effects of resource pooling on the performance of this chain. Resource pooling is known to benefit performance, in terms of reduced congestion and idle time of servers, but primarily in settings with homogenous demand. In settings where demand is heterogeneous, pooling can be counter effective. The effects of pooling and other staffing policies in a service supply chain with dependent and heterogeneous demand has, to our knowledge, not been analyzed before. Yet this is a setting that can be encountered in practice quite often. We have analyzed the case of Interactive TV, a service that evaluated by customers on both its innovativeness and reliability. As such the service supply chain needs to perform well on both aspects. Based on the results of simulation analysis, we find that a target performance increase in the upstream stage of the chain (innovation), needs to be accompanied by hiring extra staff in the downstream stages of the chain (QA and operation). We label this staffing policy "unbalanced" hiring. Pooling staff to deal with congestions at any stage in the chain will help the average performance, but the delays that are involved with pooling will make the system and hence its performance unstable. As such, pooling is a less desirable staffing policy. This finding is in line with previous work [3], [17] in the sense that pooling may increase variability. These authors do not mention the instability of the entire system as a side-effect of pooling. This side-effect is primarily caused by the delays involved in pooling and transferring staff from one group to the other. Delays can be regarded as setup times which are known to reduce the effectiveness of pooling [19]. Because of these delays, staff is often not at the right place at the right time, and is continuously fighting fires. Although the total number of staff available may be sufficient to deal with the total demand, congestion may still occur and may shift from one group to the next. As such, resource shortages seem to be persistent [34]. Our finding that in settings with dependent heterogeneous demand, pooling is on average better than not pooling, but that "unbalanced" hiring (more hiring at the downstream, dependent stages, less hiring at the upstream stages) is better than pooling, contributes to this literature. Furthermore, it answers to the call for further research in MTO-MTS settings in which these two demand streams are dependent [5].

6.2. Managerial contributions

The case described in this paper shows how managerial decisions and operational performance should be 'bridged' in the context of the innovation-driven ITV service supply chain [33]. Any change brought about by the innovation group has potential impact on service performance (reliability). The operation staff is under pressure firefighting incidents, mostly caused by innovations of the service. Meanwhile the innovation staff keeps up the pace of introducing new innovations to meet their target, unaware of the resulting impact on reliability. There needs to be an effective managerial mechanism to facilitate resource allocation and understanding feedback loops in these kinds of service supply chains. This implies that managers need to understand the dynamics of the entire supply chain and set proper targets and priorities in the chain. Leaving operation staff drained of resources puts service performance (reliability) at risk. The most resource-absorbing activities in service supply chains, like the ITV chain described here, in addition to regular operations and maintenance, are incident fixing and problem solving. Incident fixing, in particular, often drains a huge amount of operation staff very quickly. The highest priority in operation is to guarantee a continuous and reliable service, as the quick service recovery is vital to maintain customer loyalty and service reputation [35], [36]. Managers should make a balanced assessment of innovation and operation performance. This means understanding the causal relationships between them, so that managers can recognize and facilitate learning between these two processes. As mentioned above, incident fixing and problem solving are the two main ways in which resources are absorbed, they should be the main focus when deciding on priority in resource allocation. A

simulation model can provide managers with a "cockpit" [37] that enables them to analyze the effects of different resource strategies.

The boundary and role of IT services in current service economy has been largely expanded [38]. The scope of DevOps in IT service management [30] is no longer sufficient to manage the unexpected demand streams from heterogeneous sources. The solutions suggested above also provide useful practices for bridging the DevOps gap in complex IT service development by expanding the scope from the connection between development and operations toward the whole ecosystem comprising the demands, development, operations, quality assurance and so on.

7. References

[1] D.R. Smith, and W. Whitt, "Resource sharing for efficiency in traffic systems", Bell System Technical Journal, 1981, 60(1), pp. 39–55.

[2] P.T. Vanberkel, R.J. Boucherie, E.W. Hans, J.L. Hurink, & N. Litvak, "Efficiency evaluation for pooling resources in health care", OR Spectrum, 2012, 34, pp, 371-390.

[3] N.M. van Dijk, and E. van der Sluis. "To pool or not to pool in call centers", Production and Operations Management, 2008, 17(3), pp. 296-305.

[4] Z.G. Zhang, "Performance analysis of a queue with congestion-based staffing policy", Management Science, 2009, 55(2), pp. 240-251.

[5] Z.G. Zhang, I. Kim, M. Springer, G. Cai, and Y. Yu, "Dynamic pooling of make-to-stock and make-to-order operations", International Journal of Production Economics, 2013, 144, pp. 44-56.

[6] S.E. Sampson, and C.M. Froehle, "Foundations and implications of a proposed unified services theory", Production and operations management, 2006, 15(2), pp. 329-343.

[7] S.E. Sampson, "Visualizing service operations", Journal of Service Research, 2012, 15(2), pp. 182-198.

[8] H. Akkermans, and C. Voss, "The service bullwhip effect", International Journal of Operations & Production Management, 2013, 33(6), pp. 765-788.

[9] Richard, N.R. and S.G. Winter, An evolutionary theory of economic change, Harvard Business School Press, Cambridge, 1982.

[10] A.K. Gupta, K.G. Smith, and C.E. Shalley, "The interplay between exploration and exploitation," Academy of Management Journal, 2006, 49(4), pp. 693-706.

[11] H. Akkermans, and B. Vos, "Amplification in service supply chains: An exploratory case study from the telecom industry", Production and Operations Management, 2003, 12(2), pp. 204-223.

[12] R. Oliva, and J.D. Sterman, "Cutting corners and working overtime: Quality erosion in the service industry", Management Science, 2001, 47(7), pp. 894-914.

[13] S. Raisch, J. Birkinshaw, G. Probst, and M. Tushman, "Organizational ambidexterity: balancing exploitation and exploration for sustained performance", Organization Science, 2009, 20(4), pp. 685–695.

[14] M.L. Tushman, and C.A. O'Reilly, "Ambidextrous organizations: managing evolutionary and revolutionary change", California Management Review, 1996, 38(4), pp. 8–31.

[15] F. Karsten, M. Slikker, and G.J. van Houtum, "Resource pooling and cost allocation among independent service providers", Operations Research, 2015, 63(2), pp. 476-488.

[16] Hopp, W.J. and M.L. Spearman, Factory physics: foundations of manufacturing management. McGraw-Hill, Boston, 2001.

[17] S. Andradóttir, H. Ayhan, and D.G. Down, "Resource pooling in the presence of failures: Efficiency versus risk", European Journal of Operational Research, 2017, 256, pp. 230-241.

[18] A. Mandelbaum, and M.I. Reiman. "On pooling in queueing networks", Management Science, 1998, 44(7), pp. 971–981.

[19] S. Benjaafar, "Performance bounds for the effectiveness of pooling in multi-processing systems", European Journal of Operational Research, 87, pp. 375-388.

[20] N.P. Repenning, and J.D. Sterman, "Nobody ever gets credit for fixing problems that never happened", California Management Review, 2001, 43(4), pp. 64-88.

[21] N.P. Repenning, and J.D. Sterman, "Capability traps and self-confirming attribution errors in the dynamics of process improvement", Administrative Science Quarterly, 2002, 47, pp. 265-295.

[22] K. Lewin, "Action research and minority problems." Journal of Social Issues, 1946, 2(4), pp. 34-46.

[23] K.M. Eisenhardt, "Building Theories from Case Study Research", Academy of Management Review, 1989, 14(4), pp. 532-550.

[24] Webb, E.J., D.T. Campbell, R.D. Schwartz, and L. Sechrest, Unobtrusive measures: Nonreactive research in the social sciences, Chicago, Rand McNally, 1966.

[25] H. Akkermans, "Renga: A systems approach to facilitating inter-organizational network development", System Dynamics Review, 1995, 17, pp. 179-193.

[26] Vennix, J.A.M., Group model-building. Facilitating team learning using system dynamics, Wiley: Chichester, 1996.

[27] J.P. Davis, K.M. Eisenhardt, and C.B. Bingham, "Developing theory through simulation methods", Academy of Management Review, 2007, 32(2), pp. 480-499.

[28] Fanning, P., ITIL Version 3 Service Operation. Buckinghamshire: Office of Government Commerce, 2008.

[29] T.T. Niranjan, and M. Weaver, "A unifying view of goods and services supply chain management", The Service Industries Journal, 2011, 31(14), pp. 2391-2410.

[30] S. Neely, and S. Stolt, "Continuous delivery? easy! just change everything (well, maybe it is not that easy)", In *Agile Conference (AGILE)*, 2013, pp. 121-128.

[31] Bass, L., I. Weber and L. Zhu, DevOps: A Software Architect's Perspective. Addison-Wesley Professional, 2015.

[32] Craig, J., DevOps and continuous delivery: Ten factors shaping the future of application delivery. Technical report, Enterprise management associates, 2014.

[33] Wang, Y., The bridge of dreams: Towards a method for operational performance alignment in IT-enabled service supply chains. Tilburg University, School of Economics and Management, 2016.

[34] B. Morrison, "The problem with workarounds is that they work: The persistence of resource shortages", Journal of Operations Management, 2015, 39-40, pp. 79-91.

[35] J.L. Miller, C.W. Craighead, and K.R. Karwan, "Service recovery: a framework and empirical investigation," Journal of operations Management, 2000, 18(4), pp. 387-400.

[36] R. Sousa, and C.A. Voss, "The effects of service failures and recovery on customer loyalty in e-services: An empirical investigation", International Journal of Operations & Production Management, 2009, 29(8), pp. 834-864.

[37] J. Vom Brocke, C. Sonnenberg, and A. Simons, "Value-oriented information systems design: The concept of potentials modeling and its application to service-oriented architectures", Business and Information Systems Engineering, 2009, 1(3), pp. 223-.

[38] K. Ramachandran, and S. Voleti, "Business process outsourcing (BPO): Emerging scenario and strategic options for IT-enabled services", Vikalpa, 2004, 29(1), pp. 49-62.