Never the twain shall meet?

Simulating Sales & Operations Planning ramp-up dynamics in IT-enabled service supply chains

Henk Akkermans, Chris Voss, Roeland van Oers, and Quan Zhu (corresponding author)

Tilburg University
Warandelaan 2, 5037AB, Tilburg, the Netherlands
Tel: +31 13 466 3334
Email: qzhu@uvt.nl

Oh, East is East and West is West, and never the twain shall meet,
Till Earth and Sky stand presently at God's great Judgment Seat;
But there is neither East nor West, Border, nor Breed, nor Birth,
When two strong men stand face to face, though they come from the ends of the earth!

Rudyard Kipling, The Ballad of East and West (1889)

Abstract
In IT-enabled service supply chains offering services through a mix of manual and automated activities, volume ramp-ups are notoriously difficult. Especially difficult is the attempt to effectively balance the inherently conflicting objectives of the functions Sales, who want to increase output as fast as possible to capture the market, and Operations, where the emphasis is on effective utilization of scarce resources. This paper presents findings from a system dynamics model that simulates the dynamics of IT-enabled service supply chains during ramp-ups. The model is based on three real-world cases of service ramp-ups in the telecom industry and calibrated against real-world behavior in these empirical settings. Simulation analysis suggests (1) a typical "service ramp-up syndrome" in IT-enabled service supply chains, (2) root causes for this syndrome, and (3) policy options to prevent this syndrome from occurring. We put forward six propositions concerning these findings and the corresponding managerial policy to improve ramp-up performance.

Introduction

The increasing innovation speed of information technology as well as customer demand have put great pressure on companies to develop new services quickly and get them to the market as fast as possible through rapid deployment and ramp-ups. Telecom providers, cable service providers, and banks are examples of such companies with

complex service supply chains. We label their supply chains as IT-enabled service supply chains that offer services through a mix of manual and automated activities. The volume ramp-ups are notoriously difficult in IT-enabled service supply chains. According to Akkermans & Vos (2003), order backlog, rather than inventory, is the key element in IT-enabled service supply chains. Compared to inventory, order backlog is invisible and easier to accumulate. Especially when Sales and Operations Planning (S&OP) are not in the same pace, the problem of order fulfillment would become severer. S&OP conflicts would be largely amplified during service ramp-ups, leading to severe ramp-downs. We term this phenomenon as "service ramp-up syndrome". Although S&OP strategies have been widely discussed in the context of product supply chains (cf. the literature review of Tuomikangas & Kaipia (2014)), little research has looked at this issue in the context of IT-enabled service supply chains. Hence, our motivation is to investigate S&OP challenges for IT-enabled service supply chains. Specifically, we will address the three following research questions:

RQ1: Can a typical "service ramp-up syndrome" be distilled from these cases?
RQ2: What are the root causes of this service ramp-up syndrome that amplify S&OP conflicts?
RQ3: What are feasible policies to improve ramp-up performance in conditions where the service ramp-up syndrome may occur?

This paper presents a system dynamics model that is based on three real-world cases of service ramp-ups in the telecom industry and is calibrated against real-world behavior in these empirical settings. The paper is organized as follows. Three cases are described in the next section. This is followed by the illustration of the simulation model in section three. Model analysis methods and results are shown in section four. The paper concludes with answers to three research questions in the last section.

Case description

The telecom industry is an appropriate area to study of managing service ramp-ups. Among the factors that contribute to an increased need for improved coordination mechanisms in the telecom industry are heavy ramp-up of demand early in product life cycles and an increased demand for on-time deliveries in shorter time frames, and generally shorter lead times (Agrella et al., 2004). Our three cases were from a medium sized European telecommunications provider, ETEL. At the outset of our research, ETEL, a formerly state-owned company who owned the national phone grid, was suffering at the hands of cable operators who were able to offer higher speeds using new technology. ETEL, therefore, sought to upgrade its now old-fashioned copper infrastructure to a new fiber infrastructure and sell the new service to as many people as possible. Building on this, it sought to develop and ramp-up voice over internet protocol (VOIP) services. A key policy was to seek to provide superior services over its existing infrastructure and to keep its market-leading customer base intact.

Case 1 – Consumer Voice over Internet Protocol (VOIP)

The company had taken a strategic decision to move proactively into offering consumer VOIP. Once the service had been publicly launched there were ten times the expected orders, leading to a high level of outstanding orders early on, but the supply network, after some delay, was able to adapt to the higher volume and installations kept pace with still moderate level of sales. However, during this period there were many technical problems with the service connection. For example, customers would lose the use of their previous service before the new one had yet to be installed, and there were complaints of availability due to intermittent service.

The customer groups were segmented into various levels of technical complexity. As a precaution, marketing had initially focused only the 'easy' customer segments. Around week 20, this approach was abandoned and a major ramp-up started (Figure 1). The number of customers with connection problems grew rapidly and this led to an ever increasing number of customers contacting the helpdesk and complaints. At the same time, the company became increasingly worried about its fulfillment capability, which eventually led to recruitment of 300 extra technicians to help customers with installation in their homes. In addition, an extra team was formed to deal with complex complaints on an individual basis. A new installation package was developed that was easier for customers to use. In week 42, as the helpdesk calls peaked, the top management team shared Champagne for reaching record sales. Soon afterwards, the problems became publicized on TV and in social media, and by week 60 both complaints and negative media coverage peaked. Sales were scaled down to allow for existing consumer problems to be addressed effectively, and by week 65 were running at pre ramp-up levels.

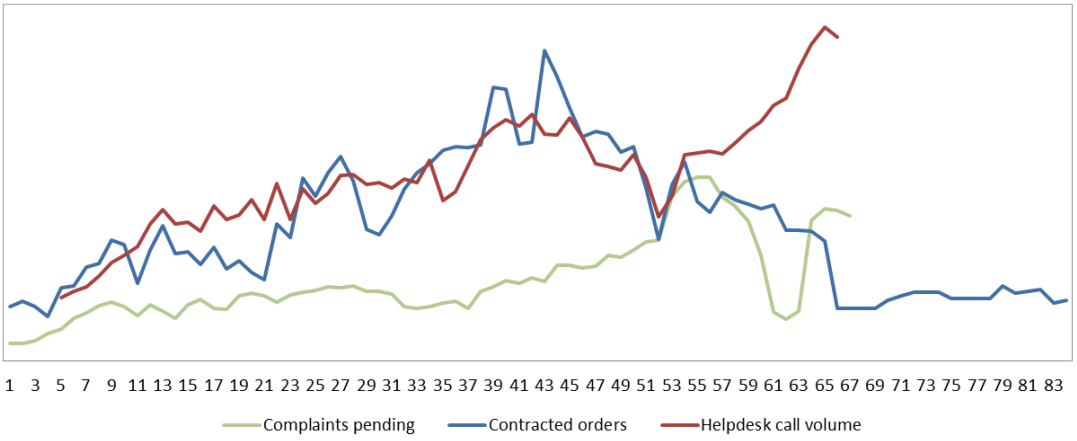

Figure 1. Timeline of Case 1

Case 2 - Fiber based broadband services

Reflecting on the problems in Case 1, the management team decided to create a more thorough S&OP process that would steer the speed of the roll-out by taking both central

capacity and local constraints, such as a shortage of engineers, into account. At the beginning, there was an explicit strategy to increase sales, but slowly and with some restraint. However, even in the early stages, there were both pressures for and an expectation of rapid ramp-up once the initial technical problems had been addressed. Despite growing evidence of technical delays, marketing continued to contract new customers. At the same time, the pressure to deliver ambitious ramp-up goals remained and even intensified. As can be seen in Figure 2, during the first 15 weeks, the level of orders contracted remained high, but the level of orders activated remained very low causing the backlog to escalate rapidly. Around week 15, the ramp-up accelerated, and sales efforts were substantially increased. The rate of contracts with customers per week strongly increased, yet the customer activation rate continued to lag behind considerably. During the rapid ramp-up, technical problems remained, and capacity issues continued to emerge. In week 23, a project to "bash quality problems out of the way" was initiated. By week 35, management had come to realize that the backlog of contracted customers waiting to be activated had become too large, lead times were too long, and quality problems and complaint levels were at too high levels, so the decision was made to scale down drastically the intended sales volume. All order intakes were stopped except for one customer group. Following this, the technical problems in the service were stabilized and a steadier and more effective ramp-up followed.

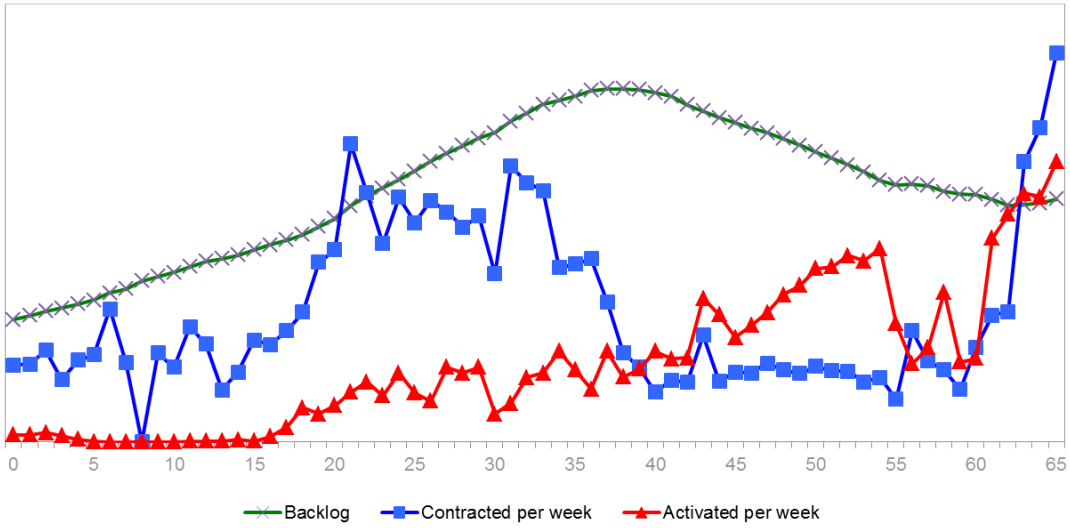

Figure 2. Timeline of Case 2

Case 3 – Business VOIP

Following consumer VOIP, the company set out to launch a business VOIP service. Management had sought to learn from the problems with the preceding launches and held a clear view that a cautious roll-out should be planned. Learning from the launches in cases 1 and 2, there was an extended initial testing period with 100 "friendly users" (Figure 3). This extended testing found IT system problems and every order required manual rework and numerous flaws in the IT system were found, resulting in a new IT

system being installed. This was followed by an evaluation of the service robustness. Once the initial technical problems had been addressed, orders began to being taken on in advance, allowing a small backlog to develop. In week 40, the ramp-up was launched in full. The ramp-up was very ambitious.

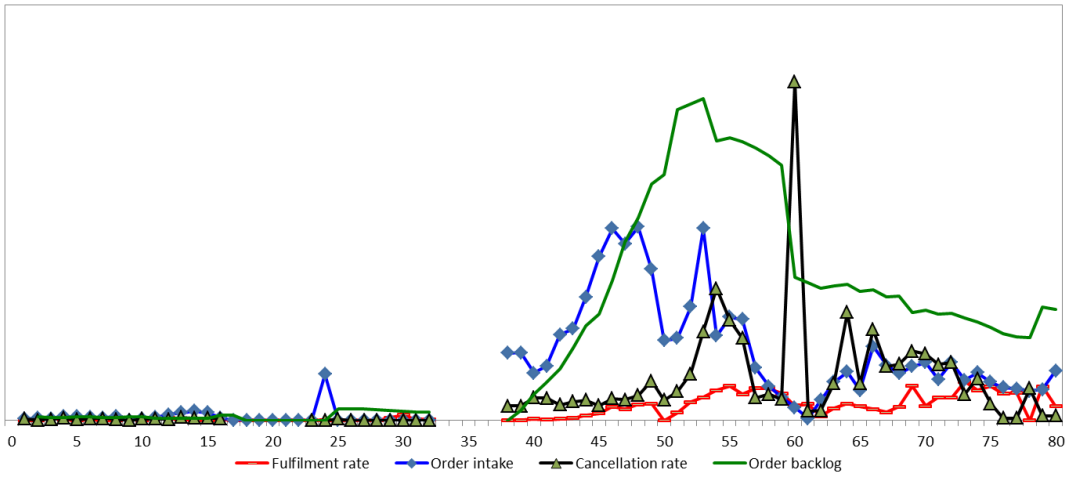

Figure 3. Timeline of Case 3

However, although the initial problems had been solved by manual fixes and workarounds, these did not transfer to high order volumes. In addition, problems surfaced in parts of the supply chain that had not been involved in the early careful testing, such as external partners, billing and technical operations. Having to do workarounds for large numbers of orders proved very difficult both in terms of resources required and in terms of fundamental problems not being fixed. As a result, the activation rate remained very low, and within ten weeks a massive backlog had developed. This, from week 52 onwards, led to a substantial rise in order cancellations. In response, the sales volume was drastically cut back until the problems could be solved. Problematic orders, amounting to one third of the backlogged orders, were cancelled. Once the problems had been effectively addressed, the launch continued with a more modest ramp-up.

The simulation model

In this section, we generate a computational representation of the conceptual system dynamics model shown in Figure 4 that forms the synthesis of our case-based analysis model. In this model, two key managerial decisions are specified: the decision on the height of the target sales rate and the decision on the desired order fulfillment capacity size. Each of these decisions drives a regulating negative feedback loop, as visualized in Figure 4. In addition, there is also a third feedback loop in the interactions between backlog, capacity, workload, fallout percentage, and rework; not balancing but reinforcing and therefore an escalating one. We next describe our formalization of these three feedback loops.

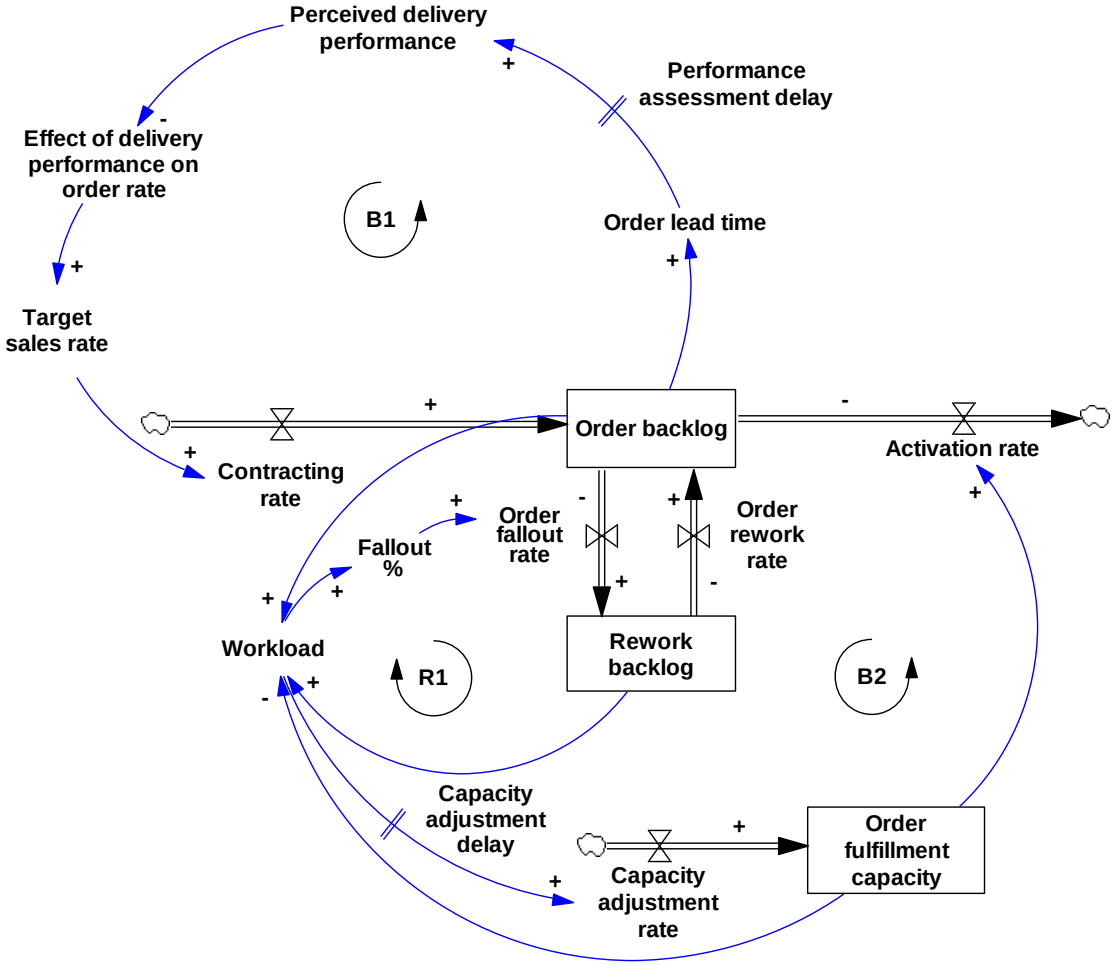

Figure 4. The conceptual system dynamics model

The target sales rate decision feedback loop (B1)

Managers cannot influence the contracting rate C_t directly, but they can set the target sales rate C^* which will influence sales activity and thereby, smoothed over a delay τ_c, the contracting delay, achieve this sales rate, certainly in a new and favorable market:

$$(1) \qquad C_t = \frac{\sum(C^*_{t-\tau_c} + C^*_{t+1-\tau_c} + \cdots + C^*_t)}{\tau_c}$$

A crucial equation is how C^* is determined. Here we assume that sales managers have no direct observation of the actual performance P_t of the service supply chain at time *t*, nor of the actual maximum capacity C_t that the supply chain can take. What they can do is infer, albeit it with some delay τ_p, a perceived delivery performance Π_t, in a smoothing process called an information delay (Sterman, 2000):

$$(2) \qquad \Pi_t = \int_0^t (P_t - \Pi_{t-1})dt + P_0$$

The perception of the performance then determines the desired sales rate. If management perceives the performance to be *below* a certain threshold T_{RD}, then management will want to *decrease* the sales rate that has been achieved recently, Ω_t with a certain factor, F_{RD}. If management perceives performance to be *over* a certain threshold T_{RU}, then it will want to *increase* the current sales rate by a certain percentage, F_{RU}.

$$(3) \quad \begin{Bmatrix} \Pi_t \leq T_{RD} & \rightarrow C^* = F_{RD} \times \Omega_t \\ \Pi_t \geq T_{RU} & \rightarrow C^* = F_{RU} \times \Omega_t \\ T_{RD} > \Pi_t > T_{RU} & \rightarrow C^* = 1 \times \Omega_t \end{Bmatrix}$$

In our base case setting, the ramp-up factor F_{RU} is set at a 5.0, so a 500% increase of the recently observed sales rate. In a manufacturing setting, such a growth rate would be an extreme, but in an IT-enabled service supply chain where capacity is in theory unlimited, this is not uncommon indeed. Similarly, the ramp-down factor F_{RD} is set at one-fifth of the recent order rate, so 0.2.

Similar to Forrester (1968), order fulfillment performance P_t is approximated by monitoring the average order lead time. This lead time is calculated, in accordance with Little's Law, as the order backlog B_{O_t} divided by the outflow of the backlog, the activation rate A_t:

$$(4) \quad P_t = \frac{B_{Ot}}{A_t}$$

The backlog B_t (including the order backlog B_{O_t} and the rework backlog B_{R_t} that we will describe in the next section) is determined also by the inflow into the backlog which results from the contracting rate C_t:

$$(5) \quad B_t = \int_0^t (C_t - A_t)dt + B_0$$

This equation closes a regulating negative feedback loop between contracting rate C_t, performance P_t, and target sales rate C^*, as visualized in Figure 4.

The capacity adjustment decision feedback loop (B2)

We assume, based on the empirical settings in the three cases, that in the base case the decision on how much capacity will be made available to serve the orders contracted is made independently from the decision on how many orders are to be contracted. Moreover, we assume that this decision is made without precise and direct knowledge of the actual sales and the target sales rate. Rather, the decision-making process that

governs capacity adjustments is what Forrester (1991) calls an *implicit* decision, rather than the overt, conscious decision such decision to set the target sales rate. One decision that management will take, albeit not adjusted continuously but rather as a fixed parameter, that is taken for granted over time, is how to set target workload W^*. We define workload, in analogy with (Akkermans & Vos, 2003), as the ratio of staff required S_t^* and staff available S_t:

$$W_t = \frac{S_t^*}{S_t} \quad (6)$$

The staff required at any time is the accumulation of all the past capacity adjustments S_t:

$$S_t = \int_0^t \sigma_t dt + S_0 \quad (7)$$

The rate of change in the actual staff σ_t is then the adjustment of the current capacity towards the level at which the workload would equal the target workload, so where $W_t = W^*$, over an adjustment period τ_s (cf. Akkermans & Vos, 2003):

$$\sigma_t = \frac{(W_t - W^*) \times C_t}{\tau_s} \quad (8)$$

So, if the workload is equal to the target workload, no capacity adjustment takes place. If the current workload is higher than the target workload, say it is set at 1.0 while the target workload is set at 0.8, then there will be a 20% shortage in capacity, and the implicit decision will be to make a capacity adjustment of 1.2-1.0 = 0.2 of the current capacity.

How much capacity will be required depends on two accumulations in the model, the order backlog B_{O_t} and the rework backlog B_{R_t}. A certain percentage of orders ε "falls out" of the normal ordering process and has to be reworked separately and manually (Akkermans & Voss, 2013). The productivity for these rework orders is much lower than the productivity for regular orders. In the base settings for our model, $\rho_O = 5\rho_R$. The total staff required then becomes the sum of the regular staff required $S_{O_t}^*$ and the rework staff $S_{R_t}^*$ required:

$$S_t^* = S_{O_t}^* + S_{R_t}^* = \frac{B_{O_t}/p_O^*}{\rho_O} + \frac{B_{R_t}/p_R^*}{\rho_R} \quad (9)$$

With p_O^* being the target lead time for the normal order processing process and p_R^* the target rework lead time, which in the cases studied was normally equally long. The two backlogs B_{O_t} and B_{R_t} both depend on the size of the capacity allocated to them, because their outflows A_t and O_{R_t} depend on these capacities, as we will describe in the next sub-section.

$$(10) \qquad \mathrm{B_{O_t}} = \int_0^{\mathrm{t}} (\mathrm{C_t} + \mathrm{O_{R_t}} - \mathrm{O_{\epsilon_t}} - \mathrm{A_t})\mathrm{dt} + \mathrm{B_{O_0}}$$

$$(11) \qquad \mathrm{B_{R_t}} = \int_0^{\mathrm{t}} (\mathrm{O_{\epsilon_t}} - \mathrm{O_{R_t}})\mathrm{dt} + \mathrm{B_{R_0}}$$

Backlog is determined by capacity, and capacity determines backlog over time, in a regulating negative feedback loop that was previously visualized in Figure 4 and described informally in the section Case description.

In the next sub-section, we will describe the equations that govern the behavior of orders in this service supply chain.

The rework cycle feedback loop (R1)
The distinction between regular backlog and rework backlog in this IT-enabled service supply chain is necessary because of the vastly different behavior these two types of work exhibit and their complex dynamic interactions. Let us start with the outflow of regular orders, the activation rate A_t:

$$(12) \qquad \mathrm{A_t} = \mathrm{C_{O_t}}(1 - \epsilon_\mathrm{t})$$

This outflow out of the backlog is determined by how much capacity is required to process them timely, $C_{O_t}^*$, provided that sufficient activation capacity C_{O_t} is available:

$$(13) \qquad \mathrm{C_{O_t}} = \mathrm{MIN}\,(\mathrm{C_{O_t}^*}, \mathrm{S_t}\rho_\mathrm{N}) = \mathrm{MIN}\,(\frac{\mathrm{B_{O_t}}}{\mathrm{p_O^*}}, \mathrm{S_t}\rho_\mathrm{N})$$

This then has to be corrected for that fraction of orders that will fall out as a result of quality issues $(1 - \epsilon_\mathrm{t})$.

The fallout percentage ϵ_t is not a constant. It is determined by the normal fallout percentage ϵ and the relationship between workload W_t and the fallout percentage ϵ_t, which is a concave function where error rates are at their lowest at some 80-90% of maximum workload, but are somewhat higher for very low levels of workload and

substantially higher for very low levels of workload. This relationship between mental state and operational performance is referred to as the Yerkes-Dobson Law (Sterman, 2000). So, for high workloads which occur during steep ramp-ups, the activation rate will not rise as fast as the available capacity, and certainly not as fast as the inflow from the contracting rate. Moreover, the orders that fallout will have to be processed in the normal work flow once more, thereby further increasing the workload, in a dynamic phenomenon known as the rework cycle (Akkermans & Vos, 2003; Oliva & Sterman, 2001).

Those orders that are processed in the normal backlog and that "fall out" for some reason or another becomes part of the order fallout rate O_{ϵ_t}:

$$(14) \qquad \mathrm{O}_{\epsilon_t} = \frac{\epsilon_t}{(1-\epsilon_t)}\,\mathrm{A}_t$$

This equation implies that O_{ϵ_t} will increase disproportionally when, during the ramp-up, the order backlog becomes so high that there is not enough capacity to activate all orders timely. First, the activation rate increases and so O_{ϵ_t} will increase proportionally. Second, workload will become greater than one and so ϵ_t will increase as well, the second multiplicative term in equation (14).

How quickly these orders will leave the rework backlog B_{R_t} will depend on the order rework rate O_{R_t}:

$$(15) \qquad \mathrm{O}_{\mathrm{R}_t} = \mathrm{MIN}\left(\frac{\mathrm{B}_{\mathrm{R}_t}}{\mathrm{p}_{\mathrm{R}}^*}, \mathrm{C}_{\mathrm{R}_t}\right)$$

In line with actual capacity allocation policies at the company studied, priority is given to processing regular orders and then the remaining capacity is allocated to improvement activities, so to rework (cf. Repenning & Sterman, 2002). So:

$$(16) \qquad \mathrm{C}_{\mathrm{R}_t} = \left(\mathrm{S}_t - \frac{\mathrm{C}_{\mathrm{O}_t}}{\rho_{\mathrm{N}}}\right)\rho_{\mathrm{R}}$$

This flow of rework then becomes input again for the regular order processing, or, as it was referred to in the company studied, "put back on the conveyor belt". So, this flow feeds back into B_{O_t}, as was specified before in Equation (10). This completes the

reinforcing feedback loop known as the rework cycle between regular backlog, fallout percentage, rework backlog and back into regular backlog.

Model analysis

In this section, we will analyze our model in three steps: first, we present a history-friendly simulation that broadly replicates the historical behavior of our three cases. This single model based on three different real-world cases represents our dynamic hypothesis (Sterman, 2000) of what the typical service ramp-up syndrome entails. We aim for history-friendly behavior rather than for statistical fitting of the model parameters to the historical behavior (Malerba et al., 1999). This dynamic hypothesis also shows what we believe to be the root causes of this dynamic phenomenon. Secondly, we develop two scenarios based on two kinds of S&OP strategies to identify sensible policy to prevent this "ramp-up syndrome" from occurring. Thirdly, we conduct sensitivity analyses of two scenarios to analyze under what conditions these policies apply.

History-friendly simulation

The model structure that we specified in the previous section is a formalization of the conceptual causal model that we developed as a synthesis of our case analysis in the second section. However, such a formal model can still exhibit a very high variety of dynamic behaviors, depending on the choice of specific parameter values. Fortunately, many of these parameter values can be observed in the real world. For instance, how long it takes to process an order or what is a "normal" percentage of orders that need rework. Other parameter values are very hard to observe, especially if they refer to inner thoughts of people, such as at what managers consider an unacceptable lead time.

In the present section, we use a parameterization of the formal model from the previous section that exhibits behavior similar to what ETEL experienced in the three cases we described. This historical behavior, or "reference behavior" (Sterman, 2000), is visualized in Figures 1, 2, and 3. As mentioned, we do not attempt to fit our model statistically to any of these historical phenomena, as we would partly be fitting to noise, rather than to the fundamental underlying pattern of behavior (Forrester, 1991), as the reader can observe from Figures 5 to 8. This we call our "base case behavior", or Scenario 0.

In Figure 5, we see a first period of service *initialization*, in which a relatively small ramp-up is attempted. Small in terms of its impact on the target population, not small in terms of increase in contracting rate, as this is already an increase of 500 percent compared to the previous contracting rate. Such early ramp-up behavior was found in all three cases, only in Case 3 do we actually have quantitative data on the size of this proto-ramp-up, as shown in Figure 3.

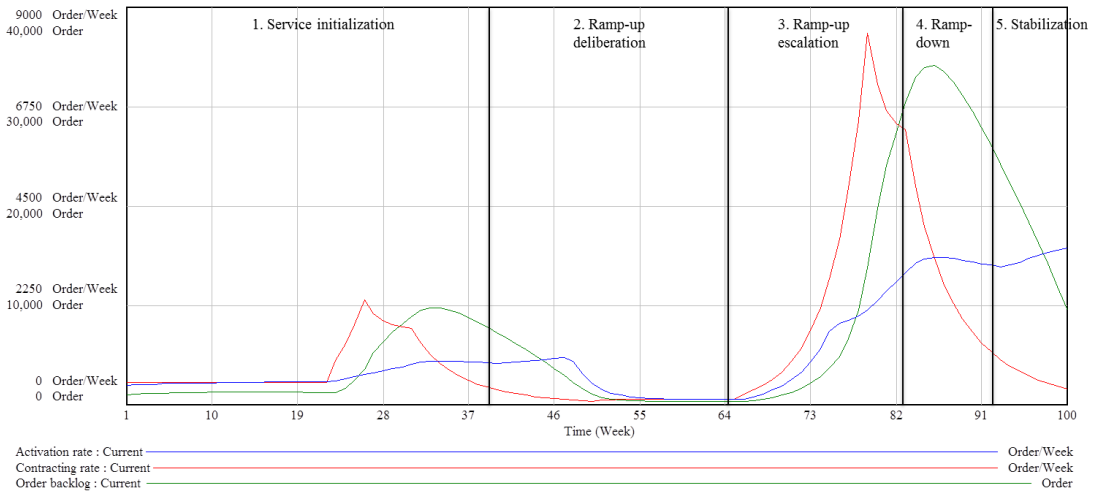

Figure 5. Simulated "history-friendly" behavior of service ramp-up dynamics

Next there is a relatively quiet period of *deliberation*, where the contracting rate remains stable, this is Phase 2. Then Phase 3, the ramp-up, starts with its period of *escalation*, during which the contracting rate grows fivefold and the backlog begins to grow as well. Notice that the activation rate lags behind considerably to this growth in the order inflow. As a result, the backlog grows exponentially, but with again some delay. By the time the performance as perceived by sales management has become such that ramp-down is inevitable in Phase 4, the backlog still keeps growing, up to the point where the workload has reached an acceptable level once more and the *stabilization* starts. Figure 6 visualizes the various triggers and consequences of the sales ramp-up / ramp-down decision in some more details. It shows the performance indicator that is most closely observed by management, i.e. the perceived delivery performance. Management does not change its preference as long as behavior stays within an upper and a lower control limit (i.e. ramp-down threshold and ramp-up threshold, respectively). As soon as it becomes higher than the ramp-down threshold (i.e. lead time becomes too long to remain acceptable), management decides to ramp-down on the contracting rate. As a result, the target sales rate goes down to 0.2 of the base rate. Over time, this policy is implemented and leads to lower and lower lead times, which then during two relatively short periods (week 21-31 and week 65-83) results in a desire to ramp-up the sales rate fivefold. This then rapidly leads to an overshoot (i.e. the perceived delivery performance again gets a value higher than acceptable) and again the target sales rate is set at a lower level. This decision logic is consistent with how our stakeholders described how they work. The behavior that results is also consistent with what we observed historically in the three cases.

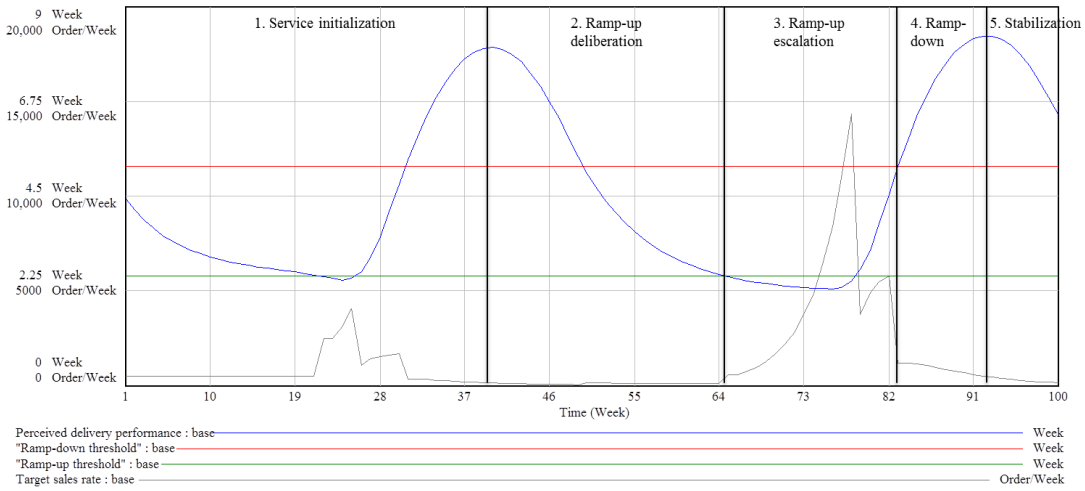

Figure 6. Simulated dynamics of ramp-up / ramp-down decision triggers

On the operations side, decisions to ramp-up and ramp-down are made as well. Here the trigger is the workload level. Once that goes over 80%, capacity is increased; else it is decreased, proportionally to the over- or under-shoots. This becomes apparent from Figure 7. After week 22 the workload starts to increase rapidly and the order fulfillment capacity grows.

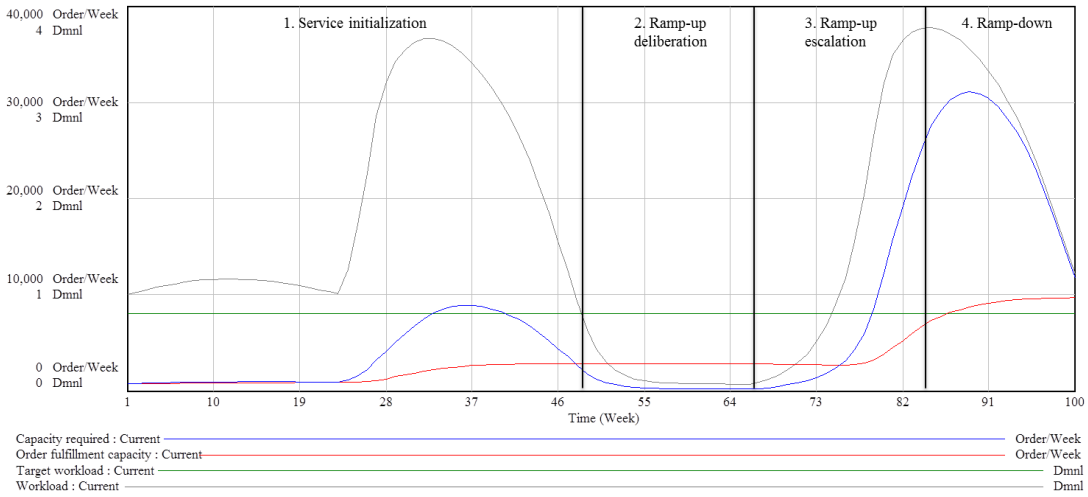

Figure 7. Simulated dynamics of workload and capacity

We reproduced in this history-friendly simulation model the rework cycle dynamics whose underlying causal structure we discussed as the second key feedback loop in the previous section. Shortly after the ramp-up escalation starts, the workload grows exponentially, and so the capacity required grows very fast as well. The available order fulfillment capacity responds in a delayed manner to this much higher capacity required, because of the long hiring delay. The capacity required peaks at around week 87 and then drops off again, as the ramp-down phase has been entered some time ago and workload is rapidly decreasing. However, because of the hiring delay involved, the

order fulfillment capacity continues to grow. A similar phenomenon occurs in the first 48 weeks, but which less dramatic results, because of the very small order rate that the supply chain starts from. However, it is clear that we are looking at a cyclical process of oscillatory behavior, in combination with some form of exponential growth, which reminds one of the classical Market Growth model (Forrester, 1968).

Figure 8 shows the rework cycle dynamics for which we described the structure in previous section as the third regulating feedback loop.

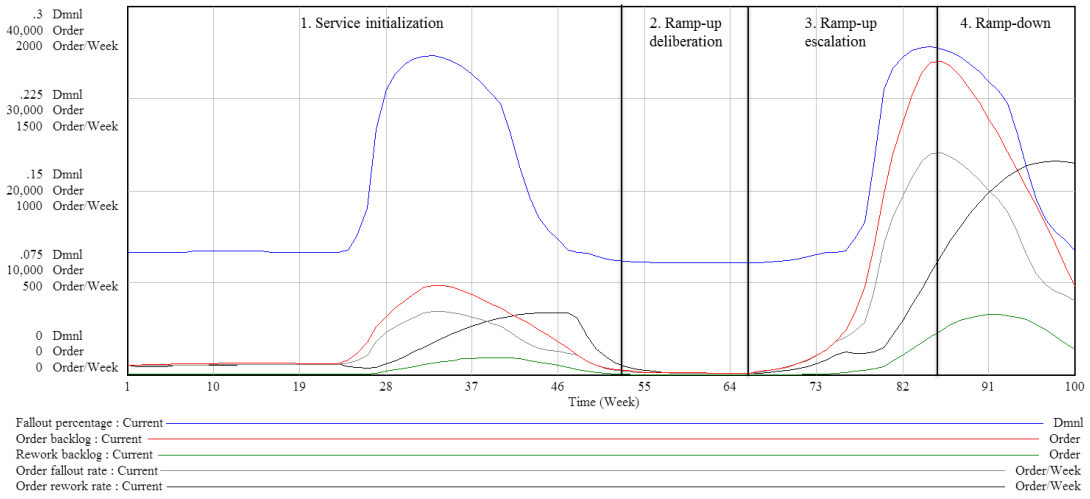

Figure 8. Simulated dynamics of order fallout and rework

It shows how, triggered by higher workloads, fallout percentages will grow, leading to higher rework backlogs, and those leading to higher order rework rates, all with some delays. Again, this process occurs twice over the course of the simulation, once very visible in weeks 67-86, once considerably more modest but mathematically similar in weeks 24-33. Also this simulated behavior is consistent with what was observed historically.

Scenario analysis results

We have seen how the current structure of IT-enabled service supply chain and its management will generate overshoots during ramp-ups, with the undesirable effects on volume, cost, quality, and time. What policy choices could improve performance on all four dimensions? Fundamentally, there are two types of choices that correspond with the notion of loosely and tightly coupled systems (Perrow, 1984). Under a tightly coupled regime, there is a direct relation between the target sales rate and the target capacity. There then is only one control loop, for the supply chain as a whole. This corresponds to some popular S&OP policies in manufacturing supply chains (e.g., Feng et al., 2010; Oliva & Watson, 2011). Under a loosely coupled regime, sales and operations retain their own, independent control loops, but make these loops more responsive or "agile".

We use our simulation model to evaluate performance of these two regimes by changing key parameter values and establishing the effect on system performance, with our other assumptions unchanged. We will first look at the scenarios that represent these two fundamentally different kinds of S&OP strategies. Then in the next sub-section, we will test the robustness of our results through sensitivity analysis. Our objective here is not to optimize according to some objective function, but rather to increase our understanding of what different policies do to the performance of the system and why that is.

Our base case scenario is labeled Scenario 0. All performance indicators are observed over 2 years (100 time periods). Volume is measured by size of the customer base at this time V_t (the more, the better), costs by the cumulated workforce cost over this period *K* (the fewer, the better). Quality is measured by the average order fallout percentage $\overline{\epsilon_t}$ (the lower, the better) and time by the average order lead-time $\overline{P_t}$ (the shorter, the better). To facilitate comparison of the scenarios, we propose a composite performance indicator that incorporates all four performance measures of volume, costs, quality, and time into one. This performance indicator *I* is calculated as:

$$\text{(17)} \qquad \mathrm{I} = \frac{V_t}{\sqrt[3]{K \times \overline{\epsilon_t} \times \overline{P_t}}} \times r$$

To give equal weight to all four performance measures, we first note that for V_t, it holds that more means better performance, while for *K*, $\overline{\epsilon_t}$, and $\overline{P_t}$, the lower the value the better the performance. Secondly, to give all indicators equal weight, we calculate the average value of the three performance measures of the second kind. Considering their different numeric ranges, we decide to use their geometric mean as the average value. Thirdly, by adding a specific multiplicative factor *r* (=4.84) to the calculation, we normalize the performance indicator of the base case (Scenario 0) to 100 to facilitate comparison.

We can now easily compare the performance effects of the two different types of S&OP policies. The simulation results are shown in Table 1.

Table 1. Scenario analysis results

NO.	Scenario	Customer base V_t	Workforce cost *K*	Average fallout % $\overline{\epsilon_t}$	Average lead-time $\overline{P_t}$	Performance indicator *I*
0	History-friendly base case	278	4470	0.1468	3.713	100
1	Tightly coupled S&OP	1672	30800	0.1868	9.36	214
2	Loosely coupled S&OP	6799	71120	0.1318	3.224	1056

In Scenario 1, the objective of tightly coupled S&OP processes is to "achieve alignment in the execution of plans" (Oliva & Watson, 2011). We model such an alignment by linking the signal from sales that decides on the height of the ramp-up with the rate at which capacity is targeted to grow. In many industrial firms that have adopted S&OP, such calculations are performed on the basis of the existing Enterprise Resource Planning (ERP) systems and their add-on Advanced Planning Systems (APS). The results of this scenario are mixed. The customer base grows impressively compared to Scenario 0 (+501%), but all other performance indicators are significantly down. Costs are 589% up, fallout is 27% higher and the average lead-time increases by 152%. The total effect is a 114% increase in its performance indicator compared to the base case. The immediate root cause for these mixed results is that the workload in this scenario is much greater in the second half of the simulation than it is in the base case (Scenario 0), due to the combined impact of the fallout effect and the rework cycle effect. Why? This is because in IT-enabled service supply chains, there is simply *no* clear link between the volume of orders coming in and the amount of work to be done in the system, caused by the compounded nonlinear effects of this reinforcing feedback loop.

Surprisingly, since this runs counter to intuition as it seems less sophisticated, the loosely coupled S&OP policy shown in Scenario 2 performs much better. In this scenario, both the sales loop and the capacity adjustment loop are speeded up independently of each other. For the sales loop, we reduce the performance assessment delay τ_p by 75% from 8 to 2 weeks, so that Sales become more quickly aware of recent performance of the service supply chain under the current order backlog. Much greater transparency of Operational performance for Sales could establish this in practice. For the capacity adjustment loop, we assume that the capacity adjustment delay τ_s is reduced from 25 to 12.5 weeks, so that faster hiring and firing can be achieved.

This scenario leads to improvement over the base case (Scenario 0) on three out of four performance dimensions: there is a 2345% increase in customer base, a 10% reduction in average order fallout percentage, and a 13% decrease in average order lead time. Workforce cost is almost 15 times higher (+1491%). However, this is also because far more consumers are converted into paying customers. In the base case (Scenario 0), it takes some 16 cost units to convert 1 customer (4470 / 278). In the tightly coupled scenario (Scenario 1), this factor goes up to 18.4 (30800 / 1672), while in the loosely coupled scenario (Scenario 2), this goes down significantly to 10.5 (71120 / 6799).

Sensitivity analysis results

How robust are the findings listed above? To address this question, we have conducted a sensitivity analysis for key parameters that we kept unchanged so far but now want to check for "tipping points". First, we look at the value of the ramp-up factor F_{RU}, which is similar to "an explicit strategy to increase sales, but slowly and with some restraint" at the beginning of Case 2. The range of the ramp-up factor F_{RU} is now reduced from 2

to 5 (with an increase of 0.5 each scenario run). Sensitivity analysis results are shown in Table 2 and Figure 9.

Table 2. Sensitivity analysis results for the ramp-up factor F_{RU}

Ramp-up factor F_{RU}	Performance indicator (base)	Performance indicator (tightly coupled)	Performance indicator (loosely coupled)
2	99	960	334
2.5	106	301	502
3	113	223	603
3.5	119	198	703
4	125	193	830
4.5	117	198	953
5	100	214	1056

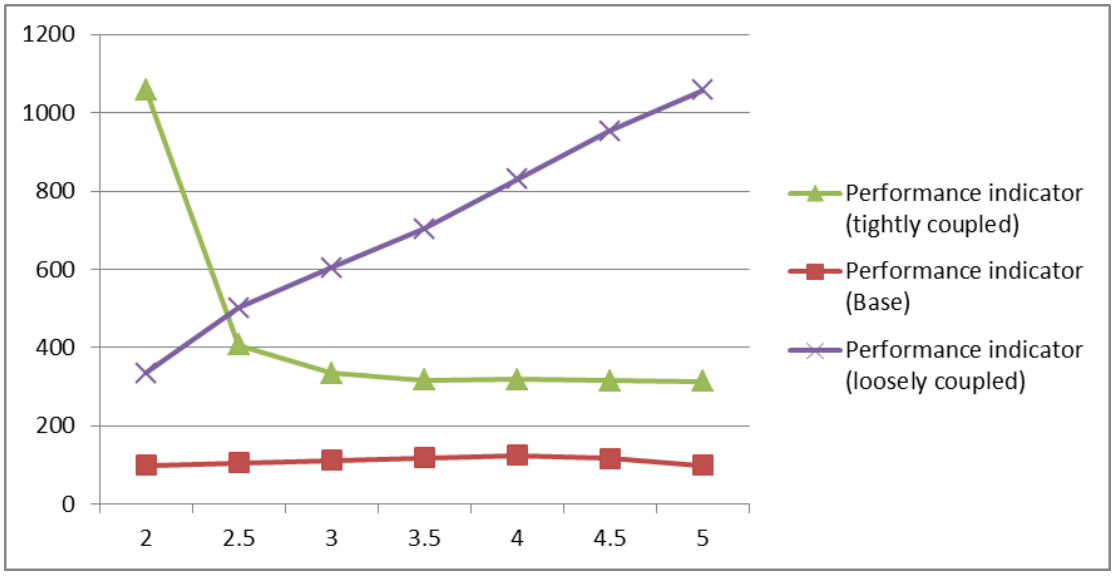

Figure 9. Sensitivity analysis results for the ramp-up factor F_{RU}

Figure 9 clearly shows that both tightly coupled (Scenario 1) and loosely coupled S&OP (Scenario 2) perform better than the base case (Scenario 0) in the sensitivity analysis for the ramp-up factor F_{RU}. But which scenario is the best depends on the value of the ramp-up factor F_{RU}: when it is set as 2 (which is similar to strategy applied at the beginning of Case 2), tightly coupled S&OP (Scenario 1) significantly outperforms the other two scenarios. It would seem that this is because the "error" in extreme front loading of the supply chain becomes lower as the sales rate grows less steep. This becomes clear as we increase the ramp-up step from 2 to 5. Then, the advantage of the tightly coupled S&OP (Scenario 1) drops dramatically, and the performance of loosely coupled S&OP (Scenario 2) improved steadily.

Second, we conduct a sensitivity analysis for the normal fallout percentage ϵ. This range is varied from 0.02 to 0.1 (with an increase of 0.02 each run). The reason for a

sensitivity analysis for this parameter is that it simulates the strategy applied at the beginning of Case 3: "an extended initial testing" to increase "the service robustness". The sensitivity analysis results are demonstrated in Table 3 and Figure 10.

Table 3. Sensitivity analysis results for the normal fallout percentage ϵ

Normal fallout % ϵ	Performance indicator (base)	Performance indicator (tightly coupled)	Performance indicator (loosely coupled)
0.02	167	1755	313
0.04	151	473	760
0.06	147	327	967
0.08	141	264	1009
0.1	100	214	1056

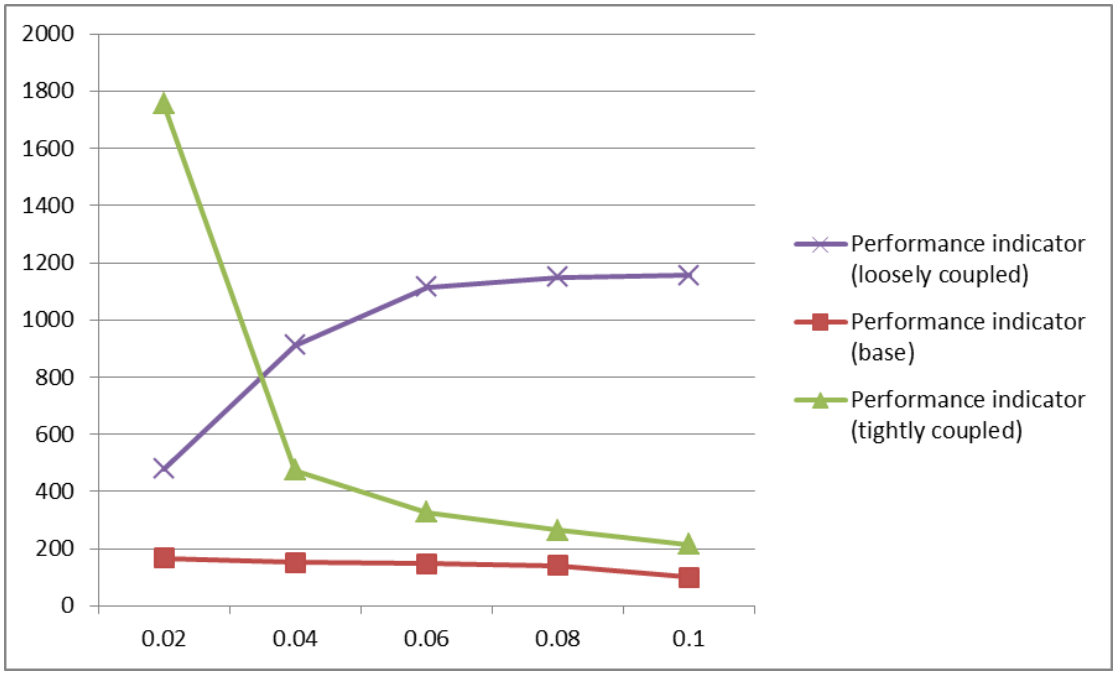

Figure 10. Sensitivity analysis results for the normal fallout percentage ϵ

Similar to the sensitivity analysis for the ramp-up factor F_{RU}, Figure 10 demonstrates that both tightly coupled (Scenario 1) and loosely coupled S&OP (Scenario 2) perform better than the base case (Scenario 0) in the sensitivity analysis for the normal fallout percentage ϵ. This makes sense, since it is this error factor that drives the destabilizing rework cycle that makes effective alignment of Sales and Operations so difficult. The lower this ϵ, the more the service supply chain resembles the planning setting for a manufacturing supply chain. Scenario 1 is the superior policy for manufacturing supply chains, so it comes as no surprise that, for a very low fallout percentage ϵ of 0.02, this tightly coupled S&OP significantly outperforms the other two scenarios. However, as the normal fallout percentage increases, this advantage radically drops. Conversely, the advantage of loosely coupled S&OP (Scenario 2) grows steadily, as the error rate comes

closer to the error range observed empirically.

Besides the two parameters that we mentioned above, all the other parameters that can potentially influence our policy choice are tested through a sensitivity analysis, which results are summarized in Appendix A. We find no significant difference between their sensitivity analysis results and the scenario analysis results: loosely coupled S&OP (Scenario 2) outperforms tightly coupled S&OP (Scenario 1) and is the best strategy to deal with the "service ramp-up syndrome" in IT-enabled service supply chains.

Discussion

We can clearly see the "service ramp-up syndrome" -- the pattern of ramping up too fast, followed eventually by ramping down -- in Figures 1 to 3 and 5 to 8. Based on our case description as well as scenario analysis results, we put forward three propositions to describe what the " service ramp-up syndrome" is from an S&OP perspective (P1) and the root causes of the "service ramp-up syndrome" in IT-enabled service supply chains that amplify S&OP conflicts (P2 and P3). Furthermore, we propose two improvement options for the "service ramp-up syndrome" and their best application scope (P4 to P6), in the light of our sensitivity analysis results.

The "service ramp-up syndrome" is a typical phenomenon observed in S&OP settings, which we find also applies to IT-enabled service supply chains, but in a severer manner. We have witnessed its characteristic behavior in both Figures 2 and 3 of the cases and Figure 5 of our simulation. In response to RQ1, we propose:

P1: The service ramp-up syndrome is characterized by a very fast ramp-up of Sales, which leads to high workload and rework levels in Operations, which is then followed by a ramp-down of the target sales rate until acceptable work levels are achieved, and then this cycle repeats at a higher level of installed base.

On the one hand, the objective of Sales is to increase volume as fast as possible to capture the market. In their mind, IT-enabled services are mostly automated, thus should not be capacity constrained. Therefore, they always set an over-optimistic sales target to boost a sales ramp-up. However, because of the capacity adjustment delay, managers of Operations do not perceive the sales plan immediately. In Figure 7, this is illustrated by the smoothed rather than step-shape increase in capacity required. Moreover, due to a lack of operational transparency, managers of Sales do not perceive the current performance instantly, either. Therefore, what they really do is to follow the decision logic which we used to explain Figure 6: they correct, but after a rather long performance assessment delay. When they correct, it is too late that they have to follow a ramp-down of the target sales rate.

On the other hand, Operations planning emphasizes effective utilization of scarce resources, i.e. manual capacities to deal with installation, activation, and rework, etc.

Due to a mix of manual and automated activities, IT-enabled service supply chains have more serious S&OP challenges than manufacturing supply chains. Akkermans & Vos (2003) argued that *quality* problems, such as rework that results from the order fallout effect are important in service supply chain dynamics. This is also what we observe in the backlog buildup in both Figure 1 of the empirical case and Figure 8 of our simulation model. In the presence of the rework cycle and the fallout effect, given that the capacity requirements for rework are very different from the capacity requirements for regular work, and that the amounts of rework vary dynamically and nonlinearly, quality issues will quickly accumulate, largely slowing down Operations capacity adjustments.

In response to RQ2, we propose:

P2: The first root cause for the service ramp-up syndrome is the decision biases by Sales to set over-optimistic sales targets;

P3: The second root cause for the service ramp-up syndrome is the rework cycle in which poor quality and technical problems leading to higher order fallout effects, leading to higher workloads which results in more quality issues.

Conventional wisdom, derived from studies of manufacturing supply chains, strongly supports the improvement of information synchronization, which is also proved by our scenario analysis results of tightly coupled S&OP (Scenario 1). In manufacturing supply chains, this goes so far as that the sales plan offset in time is the direct input of the manufacturing plan. In IT-enabled service supply chains, such a tightly coupled process can be problematic. Over-reliance on information synchronization in service supply chains can be a dangerous policy, especially when it is not carefully coordinated with capacity adjustment (Anderson et al., 2005). This is illustrated by our sensitivity analysis results that an ambitious ramp-up plan and/or a complex rework cycle can cause such incoordination. Therefore in this study, we also look at a loosely coupled S&OP strategy (Scenario 2), where both Sales and Operations improve the efficiency independently. What Sales does is to shorten the information feedback from Operations performance to Sales. In this way, managers of Sales not only focus on generating customer demand, but also pay attention to meeting customer demand (Ellram et al., 2004). Meanwhile, what Operations does is to focus on fast capacity adjustments. In a manufacturing supply chain, capacity utilization can be used as a substitute for product inventory (Helo, 2000). In a service supply chain, capacity agility helps the company quickly cover order backlog (instead of inventory) and is therefore also an important enabling factor to meet sales ramp-up. The twin improvements of Sales and Operations form an effective decentralized mechanism to maximize the overall performance, which is similar to Demirkan & Cheng's (2008) findings after they studied an application services supply chain consisting of one application service provider (ASP) to sell service to the market and one application infrastructure provider (AIP) to supply the computer capacity to the ASP. According to our sensitivity analysis results, especially

when launching a rapid ramp-up and/or facing a complex rework cycle, the IT-enabled service supply chain benefits most by applying loosely coupled S&OP strategy (Scenario 2). Fast information sharing helps reduce the risks of a rapid ramp-up as well as those of a ramp-down (Lee, 2002). Meanwhile, decentralized units are better able to handle the continual stream of small failures, forestalling the widespread multiple failures (Perrow, 1999). In summary, in response to RQ3, we propose:

P4: The first improvement policy that will eliminate the service ramp-up syndrome is to create a tightly coupled S&OP process in which the signal from Sales on the height of the ramp-up is connected with the rate that Operations capacity is targeted to grow;

P5: The second improvement policy that will eliminate the service ramp-up syndrome is to have a loosely coupled S&OP process in which Sales is alerted timely of changes in Operations performance, and Operations capacity adjustments are speeded up;

P6: The first improvement policy is more suitable to deal with slow ramp-ups and/or low product and service complexity, while the second improvement policy is more proper to cope with rapid ramp-ups and/or high product and service complexity.

Conclusion

In this study, we have looked at the dynamics of IT-enabled service supply chains during ramp-ups. This is an important topic, because IT-enabled service supply chains are partly similar to manufacturing supply chains, but also partly different precisely because of the nature of service. Through three case studies as well as system dynamics modeling, we have found that IT-enabled service supply chains are similar to manufacturing supply chains in the sense of S&OP challenges, but due to the nature of services, IT-enabled service supply chain will experience a rapid ramp-up and the corresponding ramp-down, rather than a stair-shape growth in manufacturing supply chains (cf. Senge's (1990) systems archetype: Growth and Underinvestment).

In a manufacturing setting, based on a form of MRP-logic, one might aspire to translate a specific value for a higher target sales rate into a specific value for a target capacity at a specific period in time later. This would result in tightly coupled S&OP activities. However, in an IT-enabled service supply chain, this may not be feasible. In the presence of a rapid ramp-up plan and a rework cycle with order fallout effect, the correlation between a specific incoming order rate and a specific level of capacity requires is highly tenuous.

Thus, in facing an ambitious ramp-up plan and a complex rework cycle, the "best" policy can be categorized as "loosely coupled policy" (Weick, 1976) between Sales and Operations. To paraphrase the poet here: *Sales is Sales and Ops is Ops and never the twain will meet*.

Under such a loosely coupled regime, Sales and Operations manage by accommodating very *quickly* to each other's actions, while indirectly improving base quality of Operations significantly, in line with the observations and recommendations in Oliva & Watson (2011). These synchronized policies do not fully take away the fluctuations in demand rates, workload, order backlog, or capacity. So, oscillatory behavior remains and still the IT-enabled service supply chain exhibits some instability. The temptation from manufacturing supply chains is to try and eliminate such instabilities with some form of tightly coupled S&OP. However, organizations may be prone to accidents under conditions of high interactive complexity and tight coupling (Perrow, 1984). Given the inherent instabilities of IT-enabled service supply chains, any mechanical tight coupling of the sales loop with the capacity loop may well be bound to fail.

References

Agrella PJ, Lindroth R, Norrman A. 2004. Risk, information and incentives in telecom supply chains. *International Journal of Production Economics* **90**(1): 1–16.

Akkermans H, Vos B. 2003. Amplification in service supply chains: An exploratory case study from the telecom industry. *Production and Operations Management* **12**(2): 204–223.

Akkermans H, Voss C. 2013. The service bullwhip effect. *International Journal of Operations & Production Management* **33**(6): 765–788.

Anderson EG, Morrice DJ, Lundeen G. 2005. The "physics" of capacity and backlog management in service and custom manufacturing supply chains. *System Dynamics Review* **21**(3): 217–247.

Demirkan H, Cheng HK. 2008. The risk and information sharing of application services supply chain. *European Journal of Operational Research* **187**(3): 765–784.

Ellram LM, Tate WL, Billington C. 2004. Understanding and managing the services supply chain. *Journal of Supply Chain Management* **40**(4): 17–32.

Feng Y, D'Amours S, Beauregard R. 2010. Simulation and performance evaluation of partially and fully integrated sales and operations planning. *International Journal of Production Research* **48**(19): 5859–5883.

Forrester JW. 1968. Market growth as influenced by capital investment. *Industrial Management Review* **9**(2): 83–105.

Forrester JW. 1991. System dynamics and the lessons of 35 years. In De Greene KB. (ed.) *The System Basis of Policy Making in the 1990s*. Sloan School of Management, MIT, Boston, MA.

Helo PT. 2000. Dynamic modelling of surge effect and capacity limitation in supply chains. *International Journal of Production Research* **38**(17): 4521–4533.

Lee HL. 2002. Aligning supply chain strategies with product uncertainties. *California Management Review* **44**(3): 105–119.

Malerba F, Nelson R, Orsenigo L, Winter S. 1999. “History-friendly” models of industry evolution: The computer industry. *Industrial & Corporate Change* **8**(1): 3–40.

Oliva R, Sterman JD. 2001. Cutting corners and working overtime: Quality erosion in the service industry. *Management Science* **47**(7): 894–914.

Oliva R, Watson N. 2011. Cross-functional alignment in supply chain planning: A case study of sales and operations planning. *Journal of Operations Management* **29**(5): 434–448.

Perrow C. 1984. *Normal Accident: Living with High-Risk Technologies*. Princeton University Press, Princeton, NJ.

Perrow C. 1999. Organizing to reduce the vulnerabilities of complexity. *Journal of Contingencies and Crisis Management* **7**(3): 150–155.

Repenning NP, Sterman JD. 2002. Capacity traps and self-confirming attribution errors in the dynamics of process improvement. *Administrative Science Quarterly* **47**(2): 265–295.

Senge PM. 1990. *The Fifth Discipline: The Art and Practice of The Learning Organization*. Doubleday, New York, NY.

Sterman JD. 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Irwin/McGraw-Hill, Boston, MA.

Tuomikangas N, Kaipia R. 2014. A coordination framework for sales and operations planning (S&OP): Synthesis from the literature. *International Journal of Production Economics* **154**: 243–262.

Weick KE. 1976. Educational organizations as loosely coupled systems. *Administrative Science Quarterly* **21**(1): 1–19.

Appendix

Appendix A. Sensitivity analysis results for parameters that can potentially influence our policy choice

		Minimum value (for what specific value) - Maximum value (for what specific value)		
Parameter (range, interval)	Parameter illustration	Performance indicator (base)	Performance indicator (tightly coupled)	Performance indicator (loosely coupled)
Productivity ratio (1 - 5, 1)	Productivity for regular orders: productivity for rework orders	69 (1) - 100 (5)	190 (2) - 242 (1)	423 (1) - 1056 (5)
Target lead time (1 - 3, 0.5)	For both regular and rework capacity adjustment, in weeks	74 (1) - 135 (3)	128 (1) - 292 (3)	981 (3) - 1695 (1.5)
Ramp-up threshold (0.6 - 1.4, 0.2)	T_{RU} in the simulation model	100 (1) - 482 (0.6, 0.8)	77 (1.4) - 217 (0.6, 0.8)	480 (0.6, 0.8) - 11070 (1.4)
Ramp-down threshold (1.6 - 2.4, 0.2)	T_{RD} in the simulation model	39 (1.6) - 135 (2.2, 2.4)	214 (2) - 617 (1.6)	698 (1.6) - 1056 (2, 2.2, 2.4)
Ramp-down factor (0.1 - 0.5, 0.1)	F_{RD} in the simulation model	94 (0.1) - 114 (0.5)	203 (0.1) - 269 (0.5)	1056 (all)
Observation period (1 - 5, 1)	Time to form the target sales rate, in weeks	68 (2) - 119 (5)	58 (1) - 437 (5)	956 (5) - 1331 (3)
Contracting delay (1 - 5, 1)	τ_c in the simulation model, in weeks	46 (1) - 115 (5)	37 (1) - 556 (5)	974 (5) - 1636 (1)
Target workload (0.75 - 0.95, 0.05)	W^* in the simulation model	69 (0.95) - 112 (0.75)	153 (0.95) - 240 (0.75)	746 (0.95) - 1263 (0.75)