Research Center
Finance & Information Management

Fraunhofer FIT
Project Group
Business & Information
Systems Engineering

Discussion Paper

Combining Models of Capacity Supply to Handle Volatile Demand: The Economic Impact of Surplus Capacity in Cloud Service Environments

by

Christoph Dorsch, Björn Häckel

in: Decision Support Systems, 58, 2, 2014, p. 3-14

University of Augsburg, D-86135 Augsburg
Visitors: Universitätsstr. 12, 86159 Augsburg
Phone: +49 821 598-4801 (Fax: -4899)

University of Bayreuth, D-95440 Bayreuth
Visitors: F.-v.-Schiller-Str. 2a, 95444 Bayreuth
Phone: +49 921 55-4710 (Fax: -844710)

www.fim-rc.de

WI-382

Combining models of capacity supply to handle volatile demand: The economic impact of surplus capacity in cloud service environments

Christoph Dorsch, Björn Häckel
FIM Research Center, Augsburg University, Universitaetsstr. 12, 86159 Augsburg, Germany
christoph.dorsch@wiwi.uni-augsburg.de, bjoern.haeckel@wiwi.uni-augsburg.de
Decision Support Systems (2013) DOI 10.1016/j.dss.2013.01.011

Abstract

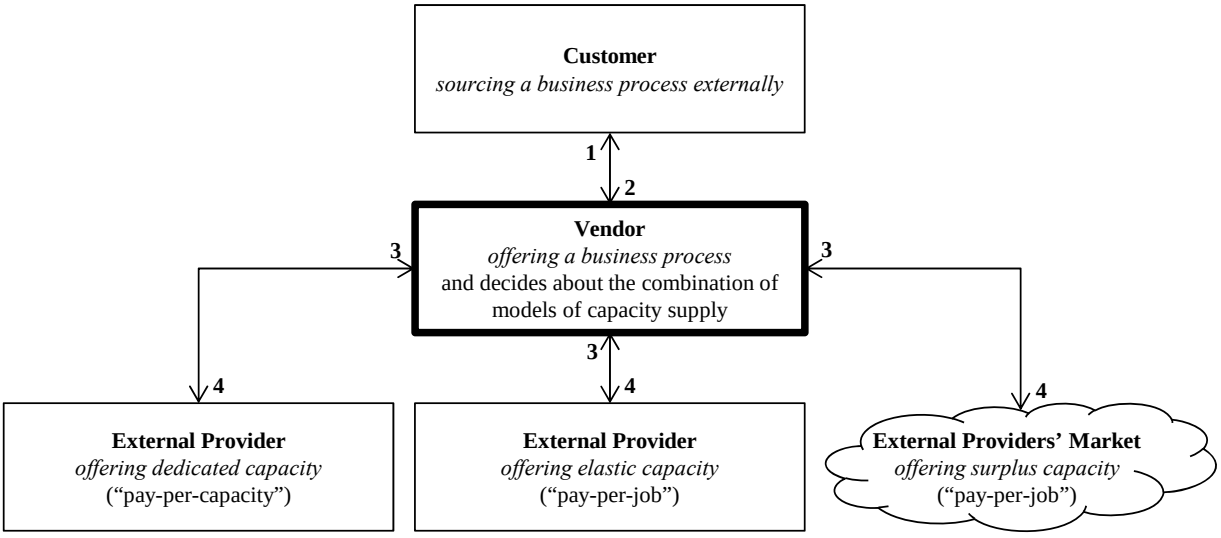
In the paper at hand we analyze the capacity planning problem of a service vendor providing a business process characterized by volatile demand to his customers. Thereby, we consider the situation that the service vendor executes certain activities by himself whereas specific parts of the business process are outsourced to external providers. For the outsourced parts, the vendor can choose between different models of capacity supply (MCS) that are offered by external providers differentiating with respect to elasticity of provided capacity and the underlying pricing model. Thereby, in addition to the two "traditional" MCS dedicated capacity and elastic capacity, recent developments in information technology enable the on-demand use of surplus capacity from an external providers' market. Since an integrated analysis of these three MCS is still missing in common literature, we develop an optimization model allowing for the simultaneous consideration of the three different MCS within an integrated queuing system. By analyzing the optimization model with help of a discrete event simulation, we study the question of how these different MCS may be combined to minimize the total operating costs of the service vendor considering volatile demand. The simulation results show that combining different MCS tends to be favorable in contrast to the stand-alone usage of a certain MCS. In particular, combining the additional option of using surplus capacity with "traditional" MCS promises cost advantages. Our optimization model therewith provides first insights in the potential economic benefits of IT-enabled MCS.

Keywords: Capacity planning, Volatile demand, Surplus capacity, Cloud Service, IT-driven service market, Simulation

1 Introduction

In recent years sourcing business processes from service vendors has established as a common practice for enterprises. Thereby, service vendors in many cases do not execute the whole business process they provide by themselves, but instead draw on specialized external providers for handling specific parts of the respective business process. This in consequence leads to the development of business process outsourcing relationships, where various companies collaborate in providing a business process for customers. For the execution of specific parts of a business process, capacity, e. g. in terms of personnel or IT-capacity is needed. Therefore external providers offer different models of capacity supply (MCS), a service vendor for one and the same part can choose between. These MCS are described by specific contractual agreements that in particular determine elasticity of the provided capacity as well as the underlying pricing model [3, 15]. Thus, within capacity planning for the respective business process a vendor is confronted with the question of how to combine the offered MCS. Combining the MCS thereby means that the vendor has the choice of which MCS to use for the execution of an incoming customer order (or bundles of incoming customer orders). Considering the case of standardized, cost-driven business processes where costs and thus achievable margins are a central competitive factor, the vendor will seek to combine the MCS such that he minimizes the total operating costs for providing the whole business process [1]. This is especially a challenge for business processes that are characterized by a volatile demand over time and a time critical execution due to the needs of the customer that requires the vendor to commit a service level (e. g. maximum execution time for the respective business process) [6, 8, 32]. Given the goal of minimizing total operating costs, in such cases the vendor's capacity planning has to be flexible in a way that it allows a preferably easy alignment to volatile customer demand: On the one hand the vendor should be able to cover peak demand to avoid high waiting costs (e. g. contractual payments due to the violation of committed service levels) and on the other hand he has to ensure acceptable execution costs with respect to the average load of demand.

Concerning this challenge in common literature the use of “traditional” volume or capacity based models as well as their combination has been widely discussed. However, recent developments in information technology (like e. g. the broad market penetration of cloud computing) enable new MCS, as they are in particular accompanied by strongly increasing on-demand integration capabilities. Due to on-demand integration capabilities external providers can be integrated much easier and faster and thus MCS that are characterized by a much more flexible usage of available capacity from external providers can emerge. To investigate the economic potentials of combining MCS that are based on on-demand integration capabilities with “traditional” MCS, in the paper at hand we analyze the optimization problem of a vendor who is part of a typical business process outsourcing relationship as outlined in Figure 1.



1 The **vendor** offers a business process with committed service level agreement (SLA) to the **customer**.

2 The **customer** sources this business process from the **vendor** sending orders characterized by **volatile demand**.

3 **External providers** offer **different models of capacity supply (MCS)** for the execution of orders to the **vendor**.

4 The **vendor** decides about a **combination of these MCS** to **minimize the total operating costs** of order execution.

Fig. 1. General setting of the three-stage business process outsourcing relationship

Within the considered business process relationship the vendor can choose between two “traditional” MCS (namely *dedicated capacity* and *elastic capacity*) and the IT-enabled MCS of *surplus capacity* for the execution of the outsourced parts of the business process. Taking a closer look at these different MCS alternatives, firstly the vendor can decide to use *dedicated capacity*. Thereby dedicated capacity means,

that the vendor reserves ex ante a certain level of capacity at an external provider and pays a fixed fee for the reserved capacity regardless of whether it is used or not ("pay-per-capacity" pricing model). If the vendor decides to assign only dedicated capacity ex ante to the respective business process he gets confronted with a trade-off [5, 35]: Assigning a high level of capacity allows the buffering of temporarily peaks in customer demand but results in idle costs in time frames of low demand. Assigning less capacity avoids idle costs but results in lost revenues or rising contractual penalties due to the violation of service level agreements (SLA) in time frames of high demand [6]. A common MCS to mitigate the described trade-off involved with dedicated capacity is the use of so called *elastic capacity* [3]. Within this MCS the external provider charges a certain fee per customer request treated ("pay-per-job" pricing model) meaning that this MCS involves payment only for capacity that is utilized. Furthermore, such MCS commonly involve committed service levels [9]. Based on the guaranteed SLA, capacity can be elastically aligned to fluctuations in customer demand, so that this MCS represents a very flexible alternative. However, these potential advantages come at a price: As this MCS means sourcing out part of the uncertainty, the external provider is now in charge to allocate capacity ex ante in such a way that he is able to fulfill the committed SLA. This means that the external provider has to schedule reserve capacities causing significant fixed costs [24]. Therefore, committed service levels most often are reflected in a higher price per customer request treated which might limit the attractiveness of this MCS for cost-driven business processes.

In addition to these two "traditional" MCS, coming along with technology-enabled on-demand integration capabilities a new possible MCS emerges: The on-demand use of *surplus capacity* from the external providers' market ("pay-per-job" pricing model). Based on developments like the growing diffusion of service-oriented infrastructures suitable for the integration of web services as well as corresponding description languages (e. g. WSDL) or standards for data exchange (e. g. XML, EDIFACT), new business relations can be established (nearly) without any loss of time by building up links fast and cheap [17, 28]. With respect to capacity planning such on-demand integration capabilities may offer substantial economic benefits for vendors, as the vendor is able to switch dynamically between various external providers depending on which service provider currently has sufficient surplus capacity available to handle the ven-

dor's peak demand. Furthermore, such a MCS promises cost advantages, as surplus capacities normally are offered at a lower price to avoid idle costs. A first approach to capture the economic potentials of this new MCS was already discussed within a former version of this paper [12]. However, the combination of surplus capacity with the two "traditional" MCS was not addressed there. In addition to its potential benefits surplus capacity involves the risk that the vendor only gets served as soon as capacity is available at the external providers' market. That might cause delays for external routed demand and thus waiting costs (e. g. due to violation of committed SLA towards customer) making this MCS riskier in contrast to (SLA backed) elastic capacity.

Due to the different characteristics of the three MCS discussed, the question of how a vendor may combine these MCS to minimize his total operating costs arises. While sourcing models based on dedicated capacity as well as elastic capacity have been widely addressed in literature in the context of capacity planning and sourcing problems for non-storable services, e. g. in Aksin et al. [3] or Gans and Zhou [15], an integrated analysis that also considers the use of surplus capacity to our best knowledge is still missing. Thus, in this paper we aim on contributing to the closure of this research gap by developing an optimization model based on queuing theory that takes into account the three MCS discussed simultaneously. In particular, we focus on the two following research questions:

1) *What are the preconditions for the use of surplus capacity? What are the characteristics of surplus capacity especially compared to "traditional" MCS? In which cases does combining different MCS promise economic benefits?*
2) *How can the different MCS – dedicated capacity, elastic capacity and surplus capacity – be combined to minimize the total operating costs of the vendor?*

The remainder of the paper is organized as follows: In section 2 we provide a brief literature review. In section 3 we address research question 1 by outlining preconditions for the use of surplus capacity and discussing economic potentials of combining different MCS. Section 4 addresses research question 2 by presenting the optimization model based on queuing systems. For analyzing the optimization model a simulation study is performed in section 5 within a case study of the securities trading and settlement

process. Finally, we summarize the key findings of the paper and give an outlook on prospective further research.

2 Related Work

Approaches for capacity planning to deal preferably flexible with uncertainty in customer demand have been widely discussed in the economics literature in the context of supply chain management (SCM) and production management. Focused only on dedicated capacity among others Bassamboo et al. [5], Tomlin and Wang [34] and Netessine et al. [29] analyze the problem of optimal mixing dedicated and flexible manufacturing capacities. For this purpose the paper of Bassamboo et al. [5] studies the basic problem of capacity and flexible technology with a newsvendor network model. The authors consider a multiproduct firm and deal with the question of whether different products should share resources or if the firm should establish dedicated resources for some of them. Tomlin and Wang [34] study unreliable supply chains that produce multiple products and like Bassamboo et al. [5] consider a firm that can invest in product-dedicated resources and totally flexible resources. Netessine et al. [29] determine the optimal mix of different types of capacity considering the effects of increasing demand correlation. Analyzing the optimal mix of different kinds of capacity, however, these papers do not consider sourcing from external business partners.

Kamien et al. [22] and Kamien and Li [21] are one of the first publications that analyze capacity constraints in the context of subcontracting production in supply chains. Thereby, their most important result is that the optimal levels of production and inventory quantities are less variable if the option to subcontract exists. In our setting we will analyze how the usage of surplus capacity as a special kind of subcontracting will affect the usage of other MCS. Furthermore, there are several papers analyzing the impacts of flexibility in external supplier markets on capacity planning. In this context Tomlin [33] studies the effect of volume flexibility of suppliers on the sourcing strategy of a firm. For this, the paper studies a single-product setting in which a firm can source either from an unreliable but cheaper or from a reliable but more expensive supplier. Furthermore the reliable supplier may possess volume flexibility. The author

shows, that contingent rerouting may constitute a possible tactic if a supplier can ramp up its processing capacity, that is, if it has volume flexibility. Dong and Durbin [11] study markets for surplus components, which allow manufacturers with excess component inventory to sell to firms with a shortage. The paper is motivated by recent developments in internet commerce, which have the potential to greatly increase the efficiency of such markets. Dong and Durbin [11] derive conditions on-demand uncertainty that determine whether a surplus market will increase or decrease supplier profits. Another paper dealing with flexibility of supplier markets is that of Lee and Whang [23]. Within this paper the authors investigate the impacts of a secondary market, where resellers can buy and sell excess inventories. For this, the authors develop a two-period model with a single manufacturer and many resellers. The authors derive optimal decisions for the resellers regarding their ordering policies and analyze the effects of the secondary market both on the sales of the manufacturer and the supply chain performance. The last-named papers are closely related to our approach regarding the basic idea of a surplus market, where firms with a shortage of capacity or inventory can buy available overcapacities or excess inventories from other firms. In our context we consider an external providers' market, were, enabled by new developments in information technology, surplus capacities can be bought on-demand.

As a fundamental difference to our approach, the papers mentioned so far are concerned with physical products. Hence, the named papers are more concerned with the possible trading of (physical) excess inventories and its implications on capacity planning. However, in our approach we focus on the capacity planning problem of vendor, who does not produce and sell physical goods but provides a service (namely providing a business process) to customers. Thereby, we understand service in a management-oriented meaning as an interaction between a service provider and a service consumer that is described by the constituting characteristics of immateriality and the simultaneity of production and consumption [10, 31]. According to this definition services are in general not storable, meaning that producing on stock and thus the building of excess inventories is not a possible strategy in this case.

The problem of capacity planning under uncertain demand for non-storable goods and in particular services has already been addressed in several papers. Especially the broad literature on call center outsourc-

ing and the capacity planning problems considered therein are closely related to our case. The two “traditional” MCS considered in our paper, namely dedicated capacity and elastic capacity are e. g. discussed in detail in Aksin et al. [3]. In their paper the authors consider a call center outsourcing analysis and choice problem faced by a contractor and a service provider. Thereby, the service provider is faced with the choice between a volume-based and a capacity-based contract offered by a contractor and based on that aims at determining the optimal capacity levels. The paper determines optimal capacity levels and partially characterizes optimal pricing conditions under each contract. In terms of our paper thereby the pay-per-capacity contract corresponds to the MCS of using dedicated capacity whereas the pay-per-job contract corresponds to using elastic capacity [3]. The paper of Gans and Zhou [15] also considers a client who can outsource some fractions of service calls to a vendor. Within their paper the authors also distinguish between volume or capacity based outsourcing contracts and analyze the centralized capacity and queuing control problem. Further papers dealing with outsourcing decisions in a service setting are e. g. Cachon and Harker [9], Allon and Federgruen [2] and Ren and Zhou [32]. Cachon and Harker [9] study the competition between two service providers with price- and time-sensitive demand by modeling this setting as a queuing game. One of their core results is that scale economies provide a strong motivation for outsourcing. In the outsourcing contract considered, the contractor charges a price per customer treated while committing a service level. This corresponds to our second MCS, namely the usage of elastic capacity. The work of Allon and Federgruen [2] is also dealing with the competition between service providers. Within their paper they analyze the situation of retailers who are locked in price and waiting-time competition and have the option to outsource their call center service to a vendor. Thereby, among others volume-based contracts and their effects on supply chain coordination are analyzed. Ren and Zhou [32] study contracting issues in an outsourcing supply chain that consists of a client and a vendor (call center), whereas the call center does outsourcing work for the client. In the model presented the call center decides on the staffing level as well as the effort to achieve a certain level of service quality. Within their paper Ren and Zhou [32] analyze contracts the client can use to induce the call center to choose staffing and effort levels that are optimal for the supply chain.

Our approach differs from the papers outlined above as we explicitly take into account a possible on-demand usage of surplus capacity from an external providers' market. Whereas the MCS of using dedicated and elastic capacity as well as their combined usage have been widely addressed in literature (e. g., in Aksin et al. [3] or Gans and Zhou [15]), their combination with the mainly IT-enabled option of using surplus capacity to our best knowledge has not gained attention in literature so far. Hence, in what follows we will study whether a combination of these three MCS can provide substantial economic benefit. For that, we develop an optimization model based on queuing systems considering the three MCS simultaneously. In doing so, the optimization model presented is a substantial extension to Dorsch and Häckel [12]. Before presenting our optimization model, in the next section we will now address research question 1.

3 Combining traditional MCS with surplus capacity

As outlined in Figure 1 (see introduction) we consider the situation of a service vendor that sources specific parts of a business process from external providers and thereby can chose between different MCS. Since the two "classical" MCS, dedicated capacity and elastic capacity are well known from literature, in this chapter we will focus on the IT-enabled MCS of *surplus capacity* in more detail. In particular, we will discuss preconditions for the use of surplus capacity as well as the economic potential of combining different MCS.

3.1 Preconditions for the use of surplus capacity

The basic idea of surplus capacity is to buy available overcapacities from external service providers on short-notice which especially might be an additional strategy to cover peak demand. In contrast to elastic capacity that is usually based on rather long-term contractual relationships, surplus capacity represents a very temporary outsourcing-relationship: the vendor can switch dynamically between various external providers depending on which provider currently has sufficient surplus capacity available to handle the vendor's peak demand. Therefore, an essential precondition for the use of surplus capacity is the existence of strong on-demand integration capabilities, meaning that business partner relationships can be established on-the-fly and thus can change frequently. Due to that, surplus capacity is gaining increasing

importance in line with the strong market penetration of concepts like dynamic business process outsourcing and “Business Process as a Service (BPaaS)”, as these concepts are mainly based on dynamic integration capabilities. BPaaS recently has gained high attention going along with innovations in IT like Cloud Computing: In a survey of Vehlow and Golkowsy [36] already 27% of the providers of cloud computing services stated to offer BPaaS. Cloud platforms that are focused on the offering of BPaaS are e. g. IBM Business Process Manager, Appian Anywhere or Process Maker Live to name but a few. The increasing market penetration of BPaaS is also substantiated by a recent study of Gartner which predicts that BPaaS will grow from $84.1 billion in 2012 to $144.7 billion in 2016, generating a global compound annual growth rate of 15% [13]. Within the concept of BPaaS services for standardized business activities such as e. g. importing high data volumes out of databases are also offered as services tailored to specific industries as e. g. life insurance origination fulfillment services. In contrast to the classical “IT Cloud” the concept of BPaaS is part of the so called “Service Cloud” as it also comprises the human part of processing work and service delivery.

As the dynamic switching between various external service providers is a basic element of the BPaaS concept, it is a rather straight-forward idea to draw on external providers with available surplus capacity on short notice in times of peak demand. However, to operationalize this approach especially some preconditions regarding the supply of information have to be fulfilled. In particular, a vendor that intends to use surplus capacity has to determine which external providers offering the required BPaaS have sufficient overcapacities at the moment respectively by when capacity will be available. Thus, the vendor’s IT-platform has to allow a continuous, mostly automated evaluation of external providers and on the other hand all relevant information has to be provided by the external providers’ market. To provide this information necessary technologies are established to a large extent. Widely recognized approaches for the support of information exchange between business partners are ebXML and RosettaNet which represent high-level frameworks. Furthermore, various product vendors focused on B2B integration like e. g. Oracle and IBM offer B2B gateways that integrate B2B protocols with internal processes. In recent years, in particular the web service paradigm coming along with service repositories and well described services

based on standardized description languages has evolved as one of the primary standards for a quick and mostly automated evaluation and integration of service providers [17]. Furthermore, in the field of on-demand services so called service marketplaces like e. g. SAP Service Marketplace, HubSpot or Zimory have developed where firms that offer or/and demand certain services can interact in a highly dynamic manner [17, 37]. The basic idea of these service marketplaces is to provide an information platform that enables a coordinated interplay of customers and providers. In this way, such service marketplaces can also be used to foster an efficient usage of capacity on the external providers' market by matching excess demand of customers with available surplus capacity of providers. A dynamic matching of capacity demand and supply can be supported by dynamic pricing mechanisms like e. g. auctions that are widely discussed in literature [4, 37, 38].

To summarize, we can state that preconditions for the use of surplus capacity especially exist with respect to on-demand integration capabilities and an adequate supply of information regarding the external providers' market. However, the necessary technologies are already established to a large extent and surplus capacity is supposed to gain increasing relevance in line with the broad establishment of business models based on concepts like BPaaS. Thus, in the following chapter we will take a closer look on the potential economic benefits of combining surplus capacity with "traditional" MCS.

3.2 Economic potential of combining different MCS

Before introducing our optimization model, in this section we will discuss in more detail, in which cases does combining the considered MCS promise economic benefits in general. Thereby, especially the specific characteristics of the different MCS have to be taken into account. In a quick overview the considered MCS can be described by the following characteristics as outlined in Table 1.

Table 1
Overview of the characteristics of MCS under review

Criterion	Dedicated capacity	Elastic capacity	Surplus capacity
General Description	Ex ante reserved capacity of external service providers	Capacity of external service providers	Capacity of external service providers
Falls due for payment	in advance for a certain time period	when used	when used
Fixed costs	for every unit of capacity	negligible	negligible
Variable costs	negligible	for every order	for every order
Availability	Fixed maximum capacity exclusively assigned (SLA backed)	Elastic amount of capacity available on short notice (SLA backed)	Capacity available only when resources of the market are underutilized (not SLA backed).
Elasticity	none	maximal (possibly restricted to given limits)	exogenous
Specifics	volume discounts (concave fixed costs)	bounds, penalties (e. g. convex costs) or reduced service level for excess usage	Variable costs expected to be lower than classic pay-per-job
Risk with regard to availability and elasticity	idle-costs of underutilization and waiting costs due to capacity shortage	waiting costs due to reduced service level or rising variable costs resp. for excess usage	waiting costs due to a lack of available capacity
***Generally* useful for**	Average Load	Average and Peak Load	Peak Load

As this characterization shows, each of the considered MCS involves specific strength and weaknesses: Dedicated capacity carries the advantage of high availability, as a fixed amount of capacity is exclusively reserved for the vendor. However, it represents a non-elastic MCS, meaning that capacity cannot be aligned to peaks or drops in customer demand. In contrast, elastic capacity allows a (fully) elastic alignment to volatile customer demand and can be regarded as mainly riskless due to the committed SLA. On the other hand, it might come along with high variable costs per job treated. Surplus capacity may allow the on-demand use of cheap overcapacities, but involves the risk that overcapacities are not available to a sufficient extent at the external providers' market. However, only based on these general strength and weaknesses, it is not yet possible to draw a valid conclusion about the general advantageousness of a certain MCS or the combination of different MCS. To be able to do so, the concrete peculiarities of each MCS (e. g. concrete level of variable or fixed costs within the different MCS) as well as the actual service level the vendor commits toward the customer have to be taken into account. Thus, in the following we will briefly discuss exemplarily constellations in which either the usage of only one MCS or else the

combination of two or more MCS tends to be dominant strategy regarding the goal to minimize total operating costs.

The use of only dedicated capacity in particular might be a dominant strategy, if a high service level is committed towards the customer, variable costs for elastic capacity are very high compared to the fixed costs of dedicated capacity and surplus capacity is regarded as very risky due to availability concerns. In such cases the use (of a rather high level) of dedicated capacity can be beneficial, as the vendor can avoid the use of either costly elastic capacity or risky surplus capacity to fulfill the committed high service level in times of peak demand. In contrast, elastic capacity can be a dominant MCS, if its variable costs are lower than the comparable fixed costs of dedicated capacity and surplus capacity is regarded as very risky. In this case it might be favorable to cover not only peak demand but also the average load with elastic capacity. And finally, surplus capacity may be the dominant MCS for the vendor, if the committed service level toward the customer is rather low, surplus capacity is "known" to be available to a sufficient extent and at the same time is comparably cheaper than dedicated and elastic capacity. As in such cases waiting times are acceptable to the greatest extent due to the low service level committed, it might be favorable to cover the whole demand with cheap surplus capacity.

However, even if there might be situations where one of the considered MCS is a dominant solution, one has to bear in mind that this is usually only the case in very specific constellations regarding peculiarities of the different MCS and the SLA toward the customer. In many cases no dominant MCS will exist and consequently a combination of different MCS might offer economic benefits. As a first possibility, dedicated and elastic capacity could be combined in such a way that dedicated capacity is used for covering the average load of demand, whereas the elastic capacity covers peak demand that exceeds dedicated capacity. The main advantage of such a combination is a significantly improved flexibility to react on temporary peaks in demand. In particular, the vendor is not forced to pay in advance for a high level of dedicated capacity to cope with potential peak demand and thus reversely reduces the risk of high idle costs in times of low demand significantly. On the other hand, the more expensive elastic capacity is only used for processing peak demand and thus to avoid costly violations of the committed SLA, while the average load

is still processed with the usually cheaper dedicated capacity. To summarize, combing these two MCS is supposed to provide economic benefits compared to their stand-alone usage [3]. Analogously the combination of dedicated capacity and surplus capacity may offer economic benefits. Once again, dedicated capacity could be used to cover the average load, while surplus capacity is used for handling peak demand. Due to the possible advantage of surplus capacity in terms of a lower price per job and its associated risk in terms of higher waiting costs compared to elastic capacity, no simple assessment is possible on whether it is more advantageous to combine dedicated capacity with elastic capacity or surplus capacity. According to these characteristics combining all three MCS can turn out to be a dominant strategy. Within such a strategy, the vendor might use dedicated capacity to cover average load. In case the peak demand exceeds dedicated capacity the vendor in a first step could try to cover capacity shortage by buying surplus capacity from the external providers' market to benefit from the more favorable prices. If capacities on the externals providers market are highly utilized and thus waiting times tend to be high, the vendor could in a second step fall back on the more expensive elastic capacity. Finally, it should be mentioned that combining elastic capacity and surplus capacity without using dedicated capacity in general is not expected to be beneficial, as this means combining an expensive with a risky MCS. So, in this case even the average load either has to be covered by usually expensive elastic capacity or by risky surplus capacity.

Having defined the different MCS as well as their basic suitability in trivial cases, we are now prepared to answer the second research question regarding a quantitative analysis of a combination of these models. For this purpose we consider all relevant characteristics of the three-stage business process outsourcing relationship outlined in the introduction along with the different MCS within a quantitative model in the following section. By analyzing this model in a subsequent step we are then able to examine different combinations of capacity supply concerning the related effects and possible economic benefits.

4 Modeling the three-stage business process outsourcing relationship

There are two essential decisions, the service vendor has to make when combining different MCS. These decisions as well as the underlying trade-offs have a significant influence on the modeling approach. So we start this section with a brief explanation of both decisions and the corresponding modeling approach before we introduce the required assumptions and cost functions of the model.

4.1 Decisions of the service vendor and modeling approach

By combining different MCS, the vendor uses the specific strength and weaknesses of each model to minimize the total operating costs arising from the execution of orders. Combining MCS thereby means that the vendor has the choice about the MCS he uses to execute an incoming order. Consequently, he decides by considering the trade-off between the specific execution costs and waiting costs associated with the different models and choosing the model with lowest costs. Facing volatile demand, this decision has to take place for every single order at the very moment it arrives as the waiting times for dedicated and surplus capacity depend on its current utilization and availability respectively.

Furthermore, when using dedicated capacity as one of the MCS an upstream decision has to be made about the amount of dedicated capacity to be reserved to execute orders. Due to the specific characteristics of this MCS, this decision has to take place ex ante and cannot be adjusted in short-term following the volatile demand. Thus, the vendor has to decide by considering the trade-off between too much capacity and excessive costs of (idle) capacity and too little capacity and excessive waiting costs.

Understandably, both decisions interact as waiting times for dedicated capacity depend on the amount of ex ante reserved capacity as well as the number of orders allocated. Consequently, along with an optimal allocation of orders, the optimal amount of dedicated capacity has to be determined within the model.

Looking at the decision about the amount of dedicated capacity first, its relevant characteristics and the resulting economic trade-off indicate the use of queuing theory as the basis for our modeling approach. Being an operation research method, queuing theory provides models to represent and analyze queuing systems handling volatile demand using a limited amount of capacity. These models enable e. g. the quan-

tification of waiting times or the number of orders in the queuing system at any point of time and help to enhance the performance of queuing systems. A central issue addressed with queuing theory is the optimization of queuing systems, e. g. of the amount of capacity with its associated costs or the maximum time an order has to wait for execution after arrival. Thus, the described trade-off related with dedicated capacity can be easily modeled relying on the basic assumptions of queuing theory as e. g. described in Gross et al. [18].

Dedicated capacity then is represented by a queuing system which has to be complemented with the two other MCS to allow the evaluation of all three models in combination. Modeled as separate queues, where all relevant parameters like waiting times etc. are exogenous, the final model results in the integrated queuing system illustrated in Figure 2 ready to be evaluated regarding the allocation of incoming orders.

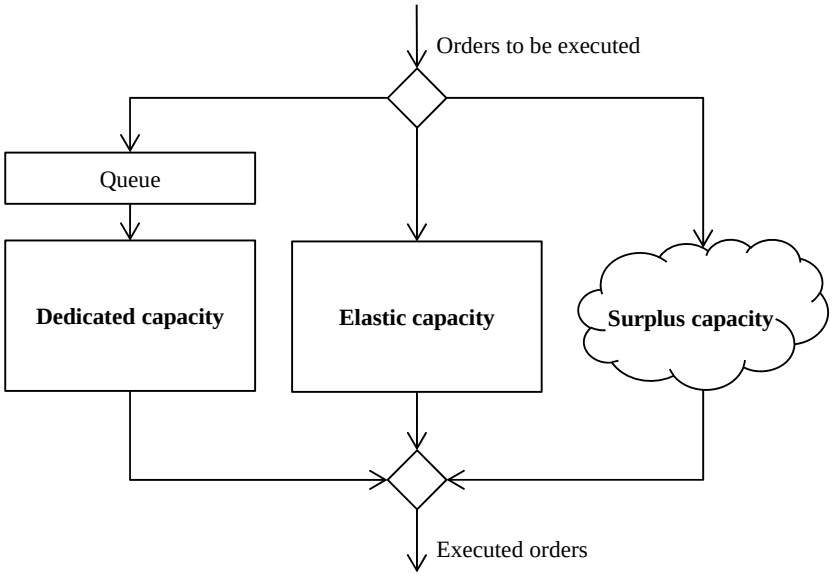


Fig. 2. Integrated queuing system representing all MCS

Regarding to the decision problems we choose this modeling approach in accordance to a wide base of literature [7, 14, 15, 20, 27] using queuing systems to model similar problems and decisions (e. g. call center or web service optimization problems). Especially when it becomes necessary to analyze detailed implementations of different MCS, queuing models are a usual approach [3, 14].

Having explained the basis for our modeling approach, we now proceed with the required assumptions to describe the model of the three-stage business outsourcing relationship in a structured way. These assumptions determine the different supply-chain-units along with their interaction, the characteristics of the different MCS and the decisions and trade-offs the vendor faces when combining MCS as described in section 3. Together they build a framework for the quantitative analysis required to answer the second research question which will be carried out with the simulation study in section 5.

4.2 Supply-Chain-Units and their interaction in the business process outsourcing relationship

The vendor offers a business process containing several process activities to his customer. Each order sent from the customer triggers this business process. Immediately after arrival of an order, the vendor starts to execute all necessary activities sequentially. After finishing the last activity, the processed order is sent back to the customer. The time between the beginning of the first activity and the end of the last activity of the business process is called *processing time*. Some process activities which are offered by external providers as standardized services are not executed by the vendor himself. For these activities the vendor uses external capacity with different contractual agreements and combines these MCS to minimize his *total operating costs c* for the business process as a whole. Figure 3 illustrates this three-stage business process outsourcing relationship in an exemplary business process view.

The arrival rate λ, i. e. the number of time-critical orders sent from the customer per unit time is random. Based on historical data and contractual agreements respectively the statistical distribution of λ can be approximated. The planning horizon is assumed to be finite and divided into equidistant time units.

A service level s is guaranteed to the customer regarding the processing time. Any order which does not keep up to this commitment causes costs c_g per order. A possible service level thereby might be a maximum processing time with monetary compensation for each time unit the order exceeds this limit or a fixed penalty for all orders of a given time frame which are not executed ahead of a final deadline.

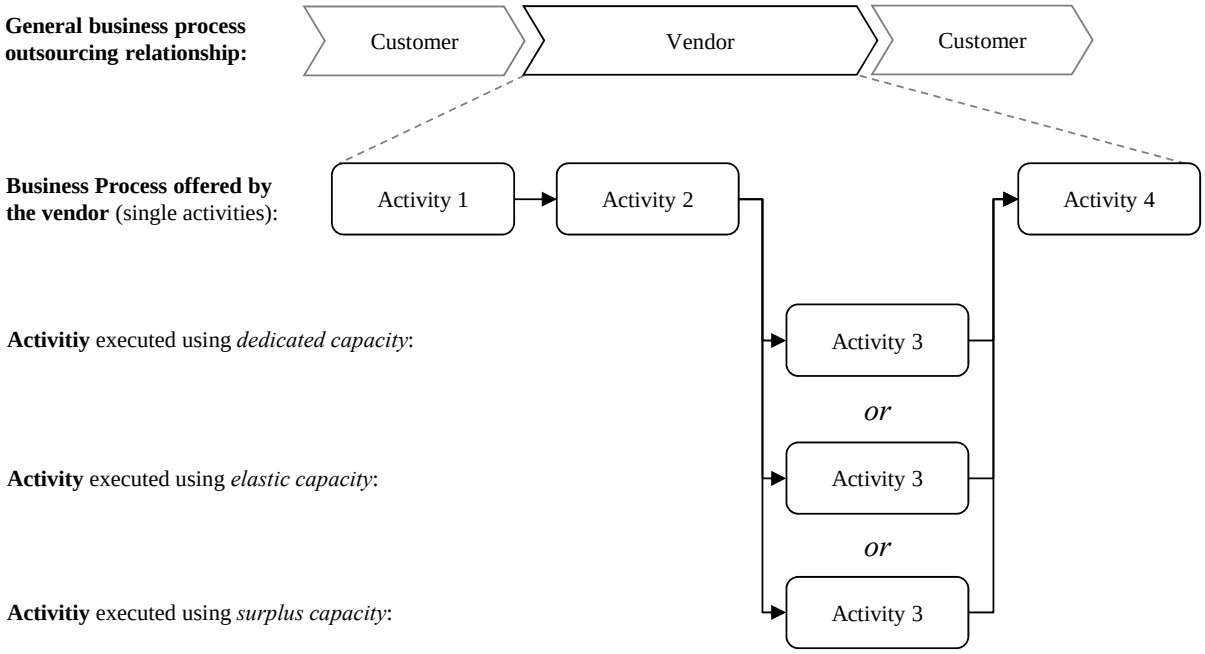


Fig. 3. The three-stage business process outsourcing relationship in an exemplary business process view

4.3 Execution of orders with dedicated capacity and basic optimization problem of the vendor

Using dedicated capacity, the vendor has to decide ex ante about the number of orders y which can be handled simultaneously, which minimizes the total operating costs:

$$\min_{y} c(\lambda, y, s)$$

Thereby the dedicated capacity works as a queuing system with the following characteristics: Unless all units of dedicated capacity are busy, the execution of orders starts immediately with the arrival of an order. Otherwise each order lines up in an infinite waiting queue. The queued orders will be executed immediately after dedicated capacity is available according to the first in/first out principle. The time frame the order stays in the queue is called *waiting time* and extends the processing time. One order uses at least one unit of capacity. Free units of capacity are idle.

The *execution time* t_d for one order depends on the individual characteristics of the order. Based on historical data the statistical distribution of t_d can be estimated. The total number of orders executed with

dedicated capacity is denoted with o_d. There are fixed costs c_f per unit capacity. The execution itself might cause additional variable costs c_d per order.

4.4 Execution of orders with elastic capacity

Elastic capacity is available with infinite amount on a pay-per-job basis. With this MCS the provider commits to start with the execution of orders immediately after it arrives. Because this applies for any amount of orders sent to this external provider, there are no waiting times. The execution time t_e for one order depends on the individual characteristics of the order. Based on historical data the statistical distribution of t_e can be estimated. The total number of orders executed with elastic capacity is denoted with o_e. There are no fixed costs but variable costs c_e which come up with the execution of an order.

4.5 Execution of orders with surplus capacity offered by the external providers' market

In addition to dedicated and elastic capacity external providers which are able to execute orders offer surplus capacity building the external providers' market. Regarding the availability of this MCS no service level is agreed. An order therefore can be executed with surplus capacity only if one or more external providers are (temporarily) underutilized and surplus capacity is offered. As this might not be the case to that very moment an incoming order has to be executed, there may be an endogenous waiting time for an order sent to the external providers' market.

According to the real world situation, the relevant parameters of dedicated and elastic capacity are known in advance, can be approximated as a statistical distribution based on historical data or can be calculated evaluating the queues (e. g. the waiting time for dedicated capacity). This does not hold true for the waiting time of an order sent to the external providers' market and the corresponding execution costs c_s. These parameters are endogenous and depend on the point in time an order has to be executed. For this purpose, we pick up our discussion in section 3.1 and assume for the external providers' market:

The vendor's IT-platform allows a continuous evaluation and integration of the external providers' market where all relevant information is provided. The necessary technologies (e. g. service repositories and well described services based on standardized description languages) for a quick and mostly automated

evaluation and integration of service providers operating at this market are established. For this reason, the endogenous time frame a until the next unit of surplus capacity is available on the external providers' market and the corresponding execution costs c_s can be determined at any point of time. The execution time t_s for one order depends on its individual characteristics. Based on historical data the statistical distribution of t_s can be estimated. The total number of externally routed orders is denoted with o_s. There are no additional fixed costs.

The absence of a service level agreement for surplus capacity therefore carries risk. With $a > 0$ orders cannot be executed immediately and this exogenous waiting time might be too long to meet up with the service level agreed to the customer. This risk has to be considered to make an appropriate decision about relying on this sourcing model.

4.6 Extended optimization problem and detailed objective function

Having defined the characteristics of each MCS determining the trade-off to be considered when deciding about how incoming orders have to be allocated to minimize the total operating costs, we are now able to state the extended optimization problem. This extended optimization problem incorporates both decisions, the vendor has to make when combining MCS by considering all corresponding trade-offs within one objective function. This can be done by adding all costs of the MCS described above. The detailed objective function minimizing the total operating costs then reads (see Table 2 for an overview of the notation used):

$$\min_{y} c = c_f\, y + c_d\, o_d + c_e\, o_e + c_s\, o_s + c_g(\lambda, y, o_d, o_e, o_s, s, t_d, t_e, t_s, a)$$

The total operation costs including all three MCS consists of three major parts: the fixed costs of dedicated capacity, the different variable execution costs of each MCS and the costs associated with the service level guaranteed to the customer. Thus the detailed objective function represents the *integrated queuing system* outlined in Figure 2, which has to be evaluated to answer the second research questions of this paper.

Table 2
Notation overview

Category	Notation	Description
Costs	c	Total operation costs
	c_f	Fixed costs of one unit of dedicated capacity
	c_d	Variable costs of one order executed with dedicated capacity
	c_e	Variable costs of one order executed with elastic capacity
	c_s	Variable costs of one order executed with surplus capacity
	c_g	Costs associated with the service level guaranteed to the customer
Orders	o	All orders sent from the customer to the vendor
	o_d	All orders executed with dedicated capacity
	o_e	All orders executed with elastic capacity
	o_s	All orders executed with surplus capacity
Execution time	t_d	Execution time of one order executed with dedicated capacity
	t_e	Execution time of one order executed with elastic capacity
	t_s	Execution time of one order executed with surplus capacity
Other	y	Amount of dedicated capacity (optimization variable)
	λ	Arrival rate of orders sent from the customer to the vendor
	a	Waiting time for an order sent to the external providers' market
	s	Service level guaranteed to the customer

4.7 Evaluating the extended decision problem: introducing the routing algorithm

By integrating the MCS within the extended optimization problem, we are now able to determine the efficient combination of these MCS. Although queuing theory provides a strong mathematical foundation to evaluate queues and several queuing systems, the necessary evaluation for the queuing system presented cannot be done analytically.

However, to derive interpretable results, a discrete-event simulation is a typical approach to evaluate queuing systems [18] and moreover often used to simulate business process related topics [16, 19, 25, 30]. Therefore we will also rely on a simulation based evaluation of the queuing system to derive results for different scenarios capturing the relevant influencing factors and to answer the second research questions. Using a simulation approach furthermore has the advantage that a wide range of possible settings of the three-stage business process outsourcing relationship (e. g. limited business hours, day and night operation, overtime work, alternating processing times) can be considered easier as in an analytical model, making the model applicable for many different scenarios found in practice.

The central component of the simulation model is an order routing algorithm. This algorithm decides for every incoming order which MCS should be used. As mentioned above this decision is made on the complete processing costs an order raises. These costs subsume all characteristics which have to be taken into account, e. g. the current processing time with regard to the service level agreed to our customer, fixed and variable costs, quantity discounts or minimum purchasing quantity as described in the previous sections.

The routing algorithm determines the processing costs and works as follows: With each arrival of an order the processing costs of the MCS are evaluated and the one with lower processing costs is chosen. Therefore, the algorithm first determines the execution time for each MCS. For dedicated capacity it is determinable as the state of the system is known: It depends on the capacity available, the arrival rate of orders and the execution time. For elastic capacity, the execution time is determined within the negotiated contract. And for surplus capacity, the time frame a until free capacity is available on the external providers' market, has to be retrieved to determine the execution time. Second, along with variable execution costs and the costs possibly incurring from the service level agreed, the execution costs can be calculated for each MCS.

Running the simulation for different amounts of dedicated capacity the corresponding total operation costs can be determined. The simulation run with lowest costs reveals the optimal amount of dedicated capacity as well as the efficient allocation of orders to MCS.

For further analysis and to derive interpretable results, we implemented a discrete-event simulation along with the routing algorithm and performed a simulation study presented in the following section.

5 Evaluating the effect of surplus capacity

For the simulation study we rely on a real world example of the securities trading and settlement process. The business process outsourcing relationship of this real world example is described before introducing two scenarios analyzed to examine the effect of combining MCS and especially the usage of surplus capacity. Then we describe the set up for the discrete event simulation and present the results subsequently.

5.1 Real world example: The securities trading and settlement process

The securities trading and settlement process contains all necessary activities to be executed when securities are sold or bought e. g. via the stock exchange. This process is a typical case addressed with our model. It is a business process most financial service providers source from a specialized business partner called “transaction bank”. A large number of orders have to be processed in time to meet regulatory standards and to avoid penalties or losses of interest when payments are not executed in time. Therefore detailed service levels are agreed. With few exceptions this process is fully digitalized and standardized through regulations and cross-company agreements. Nevertheless some manual interventions are necessary, especially after an order is placed and the corresponding transaction is closed. Within the settlement process for example, digitalized documents have to be checked, files and reports have to be completed or fees must be calculated. Figure 4 illustrates the exemplary business process outsourcing relationship with the financial service provider as customer and the transaction bank as vendor for the securities trading and settlement process.

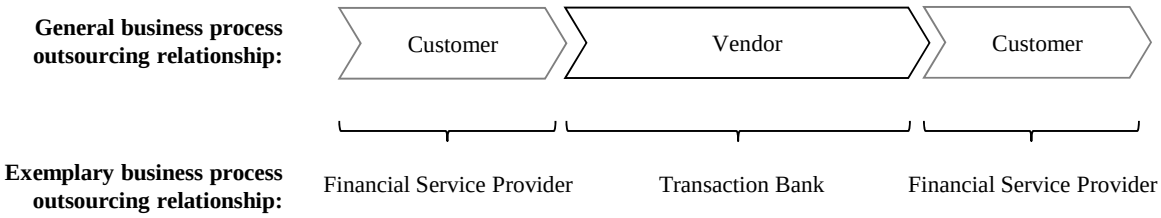


Fig. 4. Exemplary business process outsourcing relationship: the securities trading and settlement process

Optimizing capacity for manual interventions is usually an important problem for the transaction bank. The margins for this business process are small and therefore the total operating costs should be kept as small as possible. However the limited time for execution has to be taken into account. Along with the volatile arrival rates of incoming orders there is a trade-off between idle times or delayed execution respectively. Therefore these manual interventions performed on orders are the starting point for capacity optimization. They represent the parts of the business process, for which the vendor combines the different MCS offered by external service providers.

Based on this real world example we define two scenarios for the simulation study, useful to examine the economic impact of combining MCS and especially the role of using surplus capacity. The following sections describe the parameter settings used in these scenarios. Concerning the parameters used we designed scenarios which are not representing trivial cases with obvious dominant MCS as discussed in section 3 as a combination of MCS then never would lead to an economic benefit.

5.2 Scenario 1: Examining the combination of MCS

The input data for our basic scenario is determined as follows: The vendor accepts orders every working day between 7 a.m. and 10 p.m. Analyzing historical data reveals different peaks concerning the arrival rate of orders depending on exogenous factors. Dividing the 15 hours of order acceptance in seven time frames, the arrival rate within each time frame is approximated by an exponential distribution as summarized in Table 3.

Table 3
Arrival rates within a working day (mean number of orders per minute)

7:00 a.m. – 8:30 a.m.	8:30 a.m. – 2:00 p.m.	2:00 p.m. – 3:00 p.m.	3:00 p.m. – 6:30 p.m.	6:30 p.m. – 8:00 p.m.	8:00 p.m. – 9:30 p.m.	9:30 p.m. – 10:00 p.m.
60	3	30	20	4	50	3

The manual interventions performed on an order take 4 minutes in average, no matter what MCS is used. One unit of capacity y causes fixed costs amounting to EUR 240 a working day. There are no additional variable costs.

It is necessary to execute orders in time and it is necessary that no order is left unexecuted. The service level agreement between the transaction bank and the financial service provider therefore consists of two deadlines: First, each order has to be processed and sent back to the financial service provider within a fixed time frame. With regard to all other activities within the whole business process the execution time of all supporting activities on one order then must not exceed 12 minutes. For each minute an order exceeds this time frame, a compensational payment of EUR 0.033 is due. Second, there is a final processing

deadline at 12:00 p.m. for each working day. For each order which is not processed within this deadline the compensation rises to a penalty of EUR 51.

For elastic capacity, the price of executing supporting activities on one order is fixed by EUR 2.60. The external providers' market was evaluated revealing an average price of EUR 2.40 for an order executed with surplus capacity. Based on historical data, the waiting time for surplus capacity of the external providers' market is approximated (being a single customer to a set of service providers we assume no substantial effects on this waiting times routing orders externally): During a working day three time frames with different utilization are identified. Each time frame shows different waiting times for surplus capacity which can be approximated by a normal distribution as outlined in Table 4 (to avoid negative values we used a truncated normal distribution within the simulation).

Table 4
Distribution parameters of the exogenous waiting time until an order can be executed at the external provider's market in scenario 1 (mean and standard deviation in minutes)

7:00 a.m. – 12:00 noon	12:00 noon – 6:00 p.m.	after 6:00 p.m.
μ = 55:00; σ = 12:00	μ = 12:00; σ = 2:10	μ = 16:40; σ = 4:00

Orders sent to the external provider's market have to wait for surplus capacity according to these time frames. The routing decision is made immediately after the orders arrived depending on the processing costs. Due to operational reasons subsequent changes to this decision are not possible.

5.3 Scenario 2: Examining the role of surplus capacity when combining MCS

To examine the role of surplus capacity when combining MCS a second scenario is defined picking up a characteristic of surplus capacity the vendor has to consider carefully: Using surplus capacity, the absence of a service level agreement carries risk. With waiting time $a > 0$ orders cannot be executed immediately and this exogenous waiting time might be too long to meet up with the service level agreed to the customer. On the external providers' market, waiting time is exogenous and cannot be affected by any of the vendors decisions. Therefore it is interesting how a change in availability of surplus capacity affect the

simulation results. To examine this influence, we define this scenario 2, where different distribution parameters apply to the waiting time as outlined in Table 5.

Table 5
Distribution parameters of the exogenous waiting time until an order can be executed at the external provider's market in scenario 2 (mean and standard deviation in minutes)

7:00 a.m. – 12:00 noon	12:00 noon – 6:00 p.m.	after 6:00 p.m.
μ = 16:40; σ = 7:00	μ = 45:00; σ = 6:40	μ = 5:00; σ = 4:00

In scenario 1, the waiting times are high in the morning and in the evening while scenario 2 shows long waiting times occurring in the afternoon. This contrary behavior is chosen as it reflects two opposite settings compared to the arrival rate: In the scenario 1, waiting times are high when the arrival rate of orders sent to the vendor is high. In scenario 2, long waiting times occur mostly in correspondence with a low arrival rate.

5.4 Discrete event simulation set up

To evaluate the extended optimization problem, the following procedure is applied: We perform multiple simulation experiments with increasing integer values for dedicated capacity. Each experiment consists of 100 independent simulation runs (samples) to ensure the validity of simulation results. This number was chosen as several tests showed that an increasing number of samples did not affect the results significantly. For each run the total operating costs are determined. Starting the experiments with one unit of dedicated capacity, we increase the value by one unit before the next experiment is started, which ensures the solution to be an integer value. This is done until the results of an experiment show that no waiting occurs for dedicated capacity for all runs. From this it follows that a further increase of capacity does not have any positive effect concerning the total operating costs. Finally, comparing the average total operating costs for each experiment and choosing the one with the lowest costs then lead to the optimal amount of dedicated capacity and in combination with the recordings of the number of orders routed to the different MCS, the efficient combination is found.

With regard to the simulation time it is convenient that all working days of the real world example are independently of each other (e. g. no unexecuted orders left due to the processing deadline) and the relevant events which determine the optimal amount of dedicated capacity are recurrent each bank working day. Thereby it is sufficient to simulate a single working day to determine the optimum.

For each simulation run, incoming orders are generated randomly following their statistical distributions. Whenever a new time frame is reached, the arrival rate is adapted. Concerning the availability of surplus capacity a random value is generated from the corresponding statistical distribution each time an order arrives. This random value represents the time frame the respective order has to wait until it can be executed on the external providers' market. It is used by the routing algorithm to determine the corresponding processing costs.

The routing algorithm determines the current execution costs of the different MCS each time an order arrives. Then it routes the order to the path with lowest expected costs. Thereby the execution costs of dedicated capacity result from the service level agreement with the customer only. There are no variable costs and all fixed costs are sunk costs which must not be taken into account. From the service level agreement with the customer costs can occur in two different ways: If an order cannot be processed ahead of the final processing deadline, the penalty has to be considered within the processing costs. Otherwise, if the agreed processing time per order is exceeded costs per minute are charged. For elastic capacity only the variable costs exists as no waiting time has to be concerned. For surplus capacity finally, the execution costs consist of the variable cost per order and the costs resulting from waiting.

5.5 Numerical results and analysis

For both scenarios, the simulation was performed for any combination of the three sourcing models under review. The numerical results are summarized in Table 6. Analyzing the results leads to the following statements.

Table 6

Numerical results of the simulation study

	dedicated capacity only	elastic capacity only	dedicated and elastic capacity	dedicated and surplus capacity	elastic and surplus capcity	dedicated and elastic and surplus capacity
SCENARIO 1						
optimal amount of dedicated capacity y [units]	**120**	**0**	**93**	**120**	**0**	**93**
number of orders executed with dedicated capacity	17,313	0	15,630	17,313	0	15,630
number of orders executed with elastic capacity	0	17,313	1,683	0	11,614	1,683
number of orders executed with surplus capacity	0	0	0	0	5,699	0
total numer of orders executed	17,313	17,313	17,313	17,313	17,313	17,313
total operatig costs [EUR]	**39,041.30**	**45,013.80**	**38,690.75**	**39,041.30**	**44,549.37**	**38,690.75**
costs of dedicated capacity [EUR]	28,800.00	0	22,320.00	28,800	0	22,320.00
costs of order execution with elastic capacity [EUR]	0	45,013.80	4,375.80	0	30,196.40	4,375.80
costs of order execution with surplus capacity [EUR]	0	0	0	0	13,677.60	0
costs associated with service level guarated to customer [EUR]	10,241.30	0	11,994.95	10,241.30	675.37	11,994.95
SCENARIO 2						
optimal amount of dedicated capacity y [units]	**120**	**0**	**93**	**92**	**0**	**93**
number of orders executed with dedicated capacity	17,313	0	15,630	15,462	0	15,462
number of orders executed with elastic capacity	0	17,313	1,683	0	5,698	0
number of orders executed with surplus capacity	0	0	0	1,851	11,615	1,851
total numer of orders executed	17,313	17,313	17,313	17,313	17,313	17,313
total operatig costs [EUR]	**39,041.30**	**45,013.80**	**38,690.75**	**38,462.11**	**44,288.11**	**38,650.23**
costs of dedicated capacity [EUR]	28,800.00	0	22,320.00	22,080.00	0	22,320.00
costs of order execution with elastic capacity [EUR]	0	45,013.80	4,375.80	0	14,814.80	0
costs of order execution with surplus capacity [EUR]	0	0	0	4,442.40	27,876.00	4,044.00
costs associated with service level guarated to customer [EUR]	10,241.30	0	11,994.95	11,939.71	1,597.31	12,286.23

Concerning scenario 1, especially the combination of dedicated and elastic capacity leads to a reduction of total operating costs. This is due to the fact that all orders formerly causing very high waiting costs are now executed with elastic capacity reducing the waiting time for dedicated capacity. Compared to this combination, the combination of all three MCS does not lead to any changes within the total operating costs as the external providers' market apparently cannot provide surplus capacity which is cheapest regarding the execution costs for any order requiring supporting activities. Only in combination with elastic capacity several orders are routed to the market but without positive effects on the total operating costs.

Scenario 2 leads to different findings regarding an efficient combination of MCS: As the change in availability of surplus capacity only affects the results of combinations which rely on surplus capacity, the first three columns do not change compared to scenario 1. From the following columns, however, the use of surplus capacity has to be evaluated more positively: In combination with elastic capacity, surplus capacity now executes the main part of orders. Within the combination of all three MCS, surplus capacity replaces the elastic capacity in support of dedicated capacity as the execution costs are constantly lower then within the elastic capacity. Furthermore, comparing both scenarios is interesting as it reveals that surplus capacity can be used particularly when the utilization of the external providers' market does not follow the peak loads of the vendor.

Summarizing, the simulation reveals, that there are positive effects by combining dedicated and elastic capacity, as even with high execution costs the latter one can support dedicated capacity in executing orders otherwise would have been causing excessive waiting costs. The external providers' market is valuable especially when utilization of capacity on the external providers' market does not follow the peak loads of the counterpart buying this capacity for his excess demand. This holds true particularly for different industry sectors, where peak demand might not occur simultaneously and the surplus capacity provides by other industries is suitable to work on the specific demanded tasks.

6 Summary and further research

Enabled by new developments in information technology, like e. g. the growing diffusion of service-oriented infrastructures suitable for the integration of web services as well as corresponding description languages, the dynamic integration of business partners has become considerably easier. This development is e. g. reflected in the rapidly increasing market penetration of business paradigms like dynamic business process outsourcing or the closely related BPaaS. The possibility to establish new business relationships simple and fast also enables new and more flexible MCS for service providers. Based on these developments in the paper at hand we focused on the capacity planning problem of a service vendor who can choose between three different MCS. Thereby, in addition to the two "traditional" MCS dedicated capacity and elastic capacity we took into account the option to use surplus capacity from the external providers' market. While the two MCS dedicated capacity and elastic capacity have been widely addressed in literature in the context of capacity planning and sourcing problems for non-storable services (e. g. in Aksin et al. [3] or Gans and Zhou [15]), an integrated analysis that also considers the use of surplus capacity is still missing. Thus, our paper contributes to literature in particular in the following ways: First, we provide an optimization model that allows for the simultaneous analysis of the three different MCS by modeling them as one integrated queuing system. Second, based on our optimization model we analyze how the different MCS can be combined to minimize the total operating costs of a service vendor. Third, we show the economic potentials of using the IT-enabled MCS surplus capacity and how it affects the usage of "traditional" MCS. Coming along with these contributions some managerial implications arise. First of all, the optimization model developed in our paper might serve as theoretical base for developing a decision support system that allows service vendor to optimize their capacity planning decisions. As the optimization model due to its flexibility allows the consideration of a wide range of different settings, a decision support systems based thereupon especially could be helpful to support management decisions in dynamically changing market environments. Furthermore, the results of our simulation study give rise to the assumption that the high upfront-investments in on-demand integration capabilities might

pay off in the mid to long-run due to the economic potentials of using surplus capacity. Finally, service providers might consider to offer non SLA backed surplus capacity as a distinct business model, as it allows selling available overcapacities at least at a cheap price and thus helps to avoid high idle costs. This implies that service providers should consider joining service marketplaces where capacity demand and capacity supply can be matched dynamically.

Regarding the applicability of our model we would like to highlight the following: Within the case study only a small part of possible settings which can be evaluated using the model are considered. However, based on the optimization model and the simulation approach various different settings can be analyzed. In fact, our approach can easily be customized to the characteristics of various business process outsourcing relationships that can be modeled as a queuing system and thus enables determining the optimal sourcing and capacity planning policy for a wide range of business process outsourcing relationships. On the other hand, the main challenges for the applicability of the model are: Various historical data is needed as input for the simulation. As internal data about incoming orders or execution times are traceable or can be derived from contractual agreements, gathering data from external service providers could be difficult. As buying surplus capacities from the external providers' market supports these companies to ensure their capacity utilization this might be an incentive to provide the relevant information. Furthermore, there has to be a market providing all necessary tasks which have to be outsourced. However, following recent developments e. g. discussed under the labels of "cloud computing" and "BPaaS" shows that a wide range of very different services is supplied already.

Regarding potential extensions of the optimization model presented, in particular the two following aspects promise further insights in the mechanisms of combining MCS in a cloud service environment: Firstly, the model presented leaves room for improvement as it currently applies only to scenarios where one activity or consecutive activities of a business process are executed by external service providers. Modeling more complex business processes would require a queuing network considering e. g. different arrival times, processing times and a more complex layout of activities suitable for external execution. Secondly, a more detailed modeling of external service providers and their behavior regarding price set-

ting and capacity decisions would be a next logical step in analyzing the peculiarities of business process outsourcing relationships. Thus, e. g. the capacity and price setting decisions of an external service provider offering elastic capacity could be considered depending on his committed SLA. This would allow for modeling a set of different pay-per-job contracts that differ on their ratio of charged price and offered service quality. Consequently, the vendor now could decide - in addition to the two other MCS - on a set of different SLA backed pay-per-job contracts that may enables a more fine-grained capacity planning. Furthermore, the pricing processes on the market for surplus capacity could be model explicitly. In doing so, e. g. the relationship between the general level of utilization of market capacity and the resulting (market) price for surplus capacity could be modeled. Therewith, e. g. the risk of price jumps due to a temporary high utilization of market capacity could be considered within the model. In addition, considering correlations between the volatile demand of the service vendor and the level of capacity utilization of the external providers' market could be a further step for analyzing the risks involved with using surplus capacity in more detail.

References

[1] B. Adenso-Diaz, P. Gonzalez-Torre, V. Garcia, A capacity management model in service industries, International Journal of Service Management 13 (2002) 286-302.

[2] G. Allon, A. Federgruen, Outsourcing Service Processes to a Common Service Provider under Price and Time Competition, Working Paper, Kellogg School of Management, Northwestern University, Evanston, IL (2006).

[3] Z. Aksin, F. de Vericourt, F. Karaesmen, Call Center Outsourcing Contract Analysis and Choice, Management Science 54 (2008) 354-368.

[4] A. Anandasivam, M. Premm, Bid Price Control and Dynamic Pricing in Clouds, Proceedings of the 17th European Conference on Information Systems, ECIS, Verona (2009).

[5] A. Bassamboo, S. R. Ramandeep, J. A. van Mieghem, Optimal Flexibility Configurations in Newsvendor Networks: Going Beyond Chaining and Pairing, Management Science 56 (2010) 1285-1303.

[6] A. Bassamboo, R. S. Randhawa, A. Zeevi, Capacity Sizing Under Parameter Uncertainty: Safety Staffing Principles Revisited, Management Science 56 (2010) 1668-1686.

[7] M. Bondareva, A. Seidmann, Peaker Outsourcing for Service Systems with Time-Varying Arrival Rates, Proceedings of the 45th Hawaii International Conference on System Sciences, HICSS, Hawaii (2012) 4806-4813.

[8] K. Braunwarth, C. Ullrich, Valuating Business Process Flexibility achieved through an alternative Execution Path, Proceedings of the 18th European Conference on Information Systems, ECIS, Pretoria, South Africa (2010).

[9] G. P. Cachon, P. T. Harker, Competition and Outsourcing with Scale Economies, Management Science 48 (2002) 1314-1334.

[10] H. Chesbrough, J. Spohrer, A Research Manifesto for Service Science, Communications of the ACM 49 (2006) 35- 40.

[11] L. Dong, E. Durbin, Markets for Surplus Components with a Strategic Supplier, Naval Research Logistics 52 (2005) 734-753.

[12] C. Dorsch, B. Häckel, Integrating Business Partners On Demand: The Effect on Capacity Planning for Cost Driven Support Processes, Proceedings of the 45th Hawaii International Conference on System Sciences, HICSS, Hawaii (2012) 4796-4805.

[13] Gartner, Forecast: Public Cloud Services, Worldwide, 2010-2016, 2Q12 Update, 2012.

[14] N. Gans, G. Koole, A. Mandelbaum, Telephone call centers: Tutorial, review and research prospects, Manufacturing Service Operations Management 4 (2003) 97-141.

[15] N. Gans, Y.-P. Zhou, Call-Routing Schemes for Call-Center Outsourcing, Manufacturing & Service Operations Management 9 (2007) 30-50.

[16] A. Greasley, Using Business-Process Simulation within a Business-Process Reengeneering Approach, Business Process Management Journal 9 (2003) 408-420.

[17] P. Grefen, H. Ludwig, A. Dan, S. Angelov, An Analysis of Web Services Support for Dynamic Business Process Outsourcing, Information and Software Technology 48 (2006) 1115-1134.

[18] D. Gross, J. F. Shortle, J. M. Thompson, C. M. Harris, Fundamentals of Queuing Theory, Wiley, Hoboken, NJ, 2008.

[19] V. Hlupic, G.-J. De Vreede, Business process modelling using discrete-event simulation: current opportunities and future challenges, International Journal of Simulation and Process Modelling (2005) 72-81.

[20] O. Hühn, C. Markl, M. Bichler, On the Predictive Performance of Queuing Network Models for Large-Scale Distributed Transaction Processing Systems, Information Technology and Management 10 (2009) 135-149.

[21] M. Kamien, L. Li, Subcontracting, Coordination, Flexibility, and Production Smoothing in Aggregate Planning, Management Science 36 (1990) 1352-1363.

[22] M. Kamien, L. Li, D. Samet, Bertrand Competition with Subcontracting, RAND Journal of Economics 20 (1989) 553-567.

[23] H. Lee, S. Whang, The Impact of a Secondary Market on the Supply Chain, Management Science 48 (2002) 719-731.

[24] T. Liu, Revenue Management model for on-demand IT services, European Journal of Operations Research 207 (2010) 401-408.

[25] N. Melao, M. Pidd, Use of Business Process Simulation: A Survey of Practitioners, Journal of the Operational Research Society 54 (2003) 2-10.

[26] J. R. Meredith, A. Raturi, K. Amoako-Gyampah, B. Kaplan, Alternative Research Paradigms in Operations, Journal of Operations Management 8 (1989) 297-326.

[27] J. M. Milner, T. L. Olsen, Service-level agreements in call centers: Perils and prescriptions, Management Science 54 (2008) 369-383.

[28] D. Moitra, J. Ganesh, Web Services and Flexible Business Processes: Towards the Adaptive Enterprise, Information & Management 42 (2005) 921-933.

[29] S. Netessine, G. Dobson, R. Shumsky, Flexible Service Capacity: Optimal Investment and the Impact of Demand Correlation, Operations Research 50 (2002), 375-388.

[30] S. Nidumolu, N. Menon, B. Zeigler, Object-Oriented Business Process Modeling and Simulation: A Discrete Event System Specification Framework, Simulation Practice and Theory 6 (1998) 533-571.

[31] A. Rai, V. Sambamurthy, V, Editorial Notes – The Growth of Interest in Service Management: Opportunities for Information Systems Scholars, Information Systems Research 17 (2006) 327-331.

[32] Z. J. Ren, Y.-P. Zhou, Call Center Outsourcing: Coordinating Staffing Level and Service Quality, Management Science 54 (2008) 369-383.

[33] B. Tomlin, On the Value of Mitigation and Contingency Strategies for Managing Supply Chain Disruption Risks, Management Science 52 (2006) 639-657.

[34] B. Tomlin, Y. Wang, On the Value of Mix Flexibility and Dual Sourcing in Unreliable Newsvendor Networks, Manufacturing & Service Operations Management 7 (2005) 37-57.

[35] J. van Mieghem, N. Rudi, Newsvendor Networks: Inventory Management and Capacity Investment with Discretionary Activities, Manufacturing & Service Operations Management 4 (2002) 313-335.

[36] PriceWaterhouseCoopers, Cloud Computing: Navigating the Cloud, 2010.

[37] C. Weinhardt, A. Anandasivam, B. Blau, N. Borissov, T. Meinl, W. Michalk, J. Stößer, Cloud Computing – A Classification, Business Models, and Research Direction, Business & Information Systems Engineering 5 (2009) 391-399.

[38] P. Wurman, Dynamic Pricing in the Virtual Marketplace, IEEE Internet Computing 5 (2001) 36-42.