A System Dynamics Approach to Web Service Capacity Management

Elena Orta, Mercedes Ruiz
Department of Computer Languages and Systems
University of Cádiz
Cádiz, Spain
{elena.orta,mercedes.ruiz@uca.es}

Miguel Toro
Department of Computer Languages and Systems
University of Seville
Seville, Spain
{migueltoro@us.es}

***Abstract*- This paper presents a dynamic simulation model applied within the field of web services capacity management. The main purpose of the model is to help manage the web services capacity that providers assign their customers accordingly so as to ensure the fulfillment of the *Service Level Agreements (SLAs)* established.Therefore, the model allows for the analysis of the effects of different web services capacity management polices on the service performance and on the penalties to be assumed by providers for non-compliance with the response times agreed with their customers. The main aims of the sensitivity analysis carried out in the case study of this research paper are as follows: (1) to assess the fulfillment of the SLAs according to the web service capacity contracted by the customer, (2) to determine the lowest web service capacity and those capacity management parameters that ensure that the actual service performance is the one that has been agreed on (3) to evaluate the penalties that the web service provider should assume for non-compliance with response times.**

Keywords; web service management; capacity management; simulation modeling.

I. INTRODUCTION

The convergence of a series of technologies and tendencies, such as high-speed Internet, web services and SOA (*Service Oriented Architecture*) has given rise to a new category of software developers with a radically different business model: *web service providers*. These software developers design their applications in such a way that they can function as technology-independent reusable services. Web services (WSs) are accessible via the Internet through open standards and a service oriented architecture.

The success of the business of these WSs providers depends more and more on the quality of service they offer their customers, and on their service levels meeting business objectives and satisfying customer expectations. For this reason, adequate service management is essential for the business they support. It also enables to meet current and future customer and organizational requirements and to improve the quality of service as well as to reduce the overall cost.

In this context, the *Information Technology Infrastructure Library (ITIL)* [1] offers a guideline of good practices in IT service management applicable to WSs management. At present, there is an increasing number of WS providers that rely on ITIL to improve the quality of their services. ITIL suggests the use of modeling and simulation techniques as tools that might complement already existing ones to solve the problems arising in service management.

The main aim of this paper is to apply dynamic simulation models within the ITIL *Capacity Management* process. In particular, the model looks at the difficulties of deciding on what WS capacity management policy is best suited to meet the *Service Level Agreements* (*SLAs*). Thus, this model allows for the analysis of the effects of different management policies regarding the performance of web services and the penalty the providers should assume if the response-time terms are not met. In this paper a case study is described and simulation outcomes through the variation of service capacity, management service and the percentages of the capacity used to respond to service requests within the agreed response times are analized. The main goal of these simulations is to study the effects these management polices have on the fulfillment of the SLAs with respect to the service capacity agreed with the customer. Also, these simulations help determine the minimum service capacity that would still ensure that the service performance is met. Furthermore, the penalty the providers would have to assume for not meeting the response-time terms established are studied.

The paper is structured as follows; section 2 offers an overview of the ITIL *Capacity Management* process. In section 3 the main WS management techniques are explained and references to some current studies dealing with related issues are made. Section 4 presents the case study in which the dynamic simulation model is applied. In section 5 the characteristics or main aspects of the model are explained. In section 6 a description of the simulated scenarios is shown, and a summary and analysis of the results obtained are presented. Finally, section 7 contains the conclusions drawn and further studies.

II. IT SERVICE CAPACITY MANAGEMENT

Web services providers must not only focus on technology and on their internal organization but also take into consideration the quality of service they offer and the relationship they maintain with their customers. With this respect, ITIL offers a best practice guideline on IT Service Management aiming at quality service.

Based on the service lifecycle, the ITIL V3 [2] divides the services management process into five modules: *Service Strategy*, *Service Design*, *Service Transition*, *Service Operation* and *Continual Service Improvement*. ITIL V3 suggests the use of modeling and simulation techniques in some of the processes such as those involved in the *Service Strategy* and *Service Design* modules. In this study, the application of a dynamic simulation model in the *Capacity Management* process within the Service Design domain is proposed.

The purpose of *Capacity Management* is to provide the necessary capacity so as to offer good quality IT services at a reasonable cost. Quality of service is defined as follows [3]: *"Quality is the totality of features of a product or service that bears on its ability to satisfy stated or implied need (ISO-8402)"*.

In order to provide quality, service providers must carry out continual evaluations on their service performance and on their customers' future expectations. In this way, the evaluation outcomes can be used to determine whether the service and prices need to be modified, or whether customers' should be provided with further information.

Some of the main benefits of efficient capacity management are: adequate capacity forecasts necessary to provide quality services, reduction in capacity-related costs, increased levels of service availability and realiability, operational functionality, customer satisfaction and fulfillment of the service level parameters established in the SLAs.

The ITIL *Capacity Management* process comprises three sub-processes that analyze capacity needs from three different perspectives [4]:

- *Business Capacity Management:* forecasts future client requirements.
- *Service Capacity Management:* analyzes service performance in order to guarantee the fulfillment of the service-level agreements established.
- *Component Capacity Management:* analyzes the use of current infrastructure to ensure the availability of necessary resources and their most effective use.

In this paper we focus on the application of dynamic simulation models to the *Service Capacity Management* sub-process. Within this context, WSs providers are able to apply different capacity management polices to the services they offer their customers. These polices have different effects on the service performance and on the fulfillment of the parameters set in the SLAs between providers and customers. The proposed model can serve as a decision-making tool to help determine what WS capacity management policy is best suited to fulfil a SLA. Thus, the model allows for the study of the actual service performance under the scope of different polices and the analysis of the effects polices have on the fulfillment of SLAs.

III. RELATED WORKS

The most important service-oriented system management techniques can also be applied within a WSs domain. Based on their problem solving function, these techniques can be classified into [5]:

- *Artificial intelligence*: decision making support.
- *Linear, nonlinear, dynamic and integer programming*: optimization and dynamic programming problems.
- *Probabilistic models*: approximation problems, predictions, inferences and decision-making support.
- *Simulation*: prediction problems and decision-making support.

The techniques applied in *artificial intelligence* and those present in *linear, nonlinear, dynamic and integer programming* are generally used for deterministic situations. *Probabilistic models* are used for uncertain situations. S*imulation models* are computational models that represent complex systems in a simplified manner. One of the main advantages of simulation models is that they allow experimentation with different decisions and an analysis of the outcomes of a system in which the costs or risks of real-life experiments are prohibitive. Additionally, simulation permits the analysis of systems which are so complex that they cannot be represented through analytical models. Simulation models provide mechanisms for experimentation, behavioral predictions, solutions to 'What if" questions and the represented system's self- learning, among other things.

In this research paper, dynamic simulation models are used. The following are some of the main advantages of dynamic simulation models over analytical models [6]:

- They are flexible and useful systems for capturing and modeling the high levels of uncertainty that other systems present. These models complement those analytical techniques that help model the risks and behavior associated to the uncertainty.
- It is a flexible technique that allows for the represention of a wide variety of dynamic structures and interactions. Analytical techniques such as dynamic programming can cause untreatable problems when the complexity of the system is high.
- They allow for system feedback modeling. There are certain behaviour patterns and decisions made at a specific point in time which can affect the evolution of the system. When the implications are complex, analytical models turn into inapplicable and useless tools.

Current studies aiming at providing solutions to IT service management as WSs, include: process modeling and technology analysis in IT service management in order to increase service efficiency and quality is applied in [7]. The application of [8] is within the process of the *Service Transition Management*. A decision-making analytical model based on deterministic and probabilistic programming techniques which help in the planning and implementation of service changes is suggested. Studies [9] and [10] focus on the *Incident Management* process. In [9] a decision-making support system based on the business objectives approach (*MBO, Management Business Objectives*) is proposed. In [10] an approach to incident management assessment and improvement is presented. The approach is based on service behavior metrics and on a methodology for guided analysis which uses data mining to find the root causes of service malfunction. Finally, the following studies apply dynamic simulation techniques: [11] puts forward a dynamic model which helps manage the delays that occur in the processes where there is a certain degree of uncertainty regarding their causes. The proposed model in [12] is applied to deal with the problem of deciding on how to allocate resources in order to guaranteee SLAs fulfillment.

Our study suggests the application of a dynamic simulation model in the ITIL *Capacity Management* domain. The model allows for the analysis of the WSs performance in terms of different management polices on the service capacity providers assign their customers, and assess the effects the aforementioned polices have on the fulfillment of the Service-Level Agreements established between providers and customers.

IV. PROBLEM DESCRIPTION

To carry out this research paper, a banking validation WSs provider and an e-commerce company that sells their products via the Internet have been used. The e-commerce company is a distributor that buys products from the supplier and sells them to their customers through its web portal. Following the SOA approach, the tasks to validate credit card details and verify that customers possess enough credit to make the purchase are handled by the credit validation WS provided by the banking validation services provider. In this context, the web service provider and the company sign a SLA in which different service level categories are set, namely hours, service availability, service continuity, performance and service capacity, customer support, failure feedback and resposibilities to be assumed by the parties, among others.

A. SLA parameters agreed on between providers and customers

The main aim of this study is to evaluate how a credit card validation service capacity management influences service performance and the penalty providers should assume for non compliance with the agreed response times. Therefore, this study focuses on the parameters most frequently used that define these aforementioned aspects [4]. These parameters are classified as follows:

Service Capacity:

- *Contracted Validation Rate: card validation service capacity contracted by the company.*

Service Response Time:

- *Expected Response Time (ERT): expected response time of a service. If this time is exceeded, the provider is penalized.*
- *Maximum Response Time (MRT): maximum response time of a service. If this time is exceeded, the provider is penalized and the validation request is abandoned.*

Service Performance:

- *Validated Request Rate within ERT:* Minimum percentage of service requests that must be validated within the expected response time.
- *Abandoned Request Rate: Maximum percentage of service requests that are permitted to be abandoned.*

Penalties for not meeting the response times:

- *MRT Penalty:* penalty the provider would have to assume for each service request validated within the maximum response time.
- *Abandonment Penalty*: penalty the provider would have to assume for each service request abandoned.

B. Service capacity management policies

In this context, service providers can apply different service capacity management policies. The ways in which providers assign the service capacity for the validation of those service requests received within the agreed response times namely, the *expected response time* and the *maximum response time,* will vary depending on the policy used and the capacity management parameters established.

The management policies applied will determine the service performance and the fulfillment of the service level agreements established. This study analyzes the effects the following polices have on the service performance and the penalties providers would have to assume in the event of non-compliance with the agreed response times:

- *Policy A*: The service provider establishes the service capacity percentages that will be used to validate the requests within the expected response time and within the maximum response time. These percentages are constant and independent of the trends displayed by the requests received.
- *Policy B*: The service provider establishes the highest service capacity percentage used to validate requests within the maximum response time and the lowest percentage to validate requests within the expected response time. At a given time, the service capacity assigned to validate the requests within the maximum response time which has not been used can be assigned to validate new requests received within the expected response time. Therefore, the percentages of the service capacity assigned to validate the requests within the agreed response times are not constant and might vary in relation to the trends displayed by the received requests.

V. MODEL CONSTRUCTION

The following is a description of the construction of the suggested model based on Kellner's proposal for describing simulation models [13] and Martinez and Richardson's methodology for model building [14].

A. Purpose and scope of the model

The dynamic simulation model suggested in this study is applied to the analysis of the effects *Policy A* and *B* regarding WS capacity management assigned by the provider have on service performance levels and on the fulfillment of the agreements set in the SLAs.

The model's main aim is to help manage the card validation service capacity that the provider assigns the customer in order to meet the service performance parameters set in the SLA. The service provider is said to meet these parameters whenever:

- The percentage of requests validated within the expected response time is higher or equal to the parameter of *Validated Request Rate within ERT* specified in the SLA.
- The percentage of requests abandoned for exceeding the maximum response time is lower or equal to that specified in the *Abandoned Request Rate* parameter set in the SLA.

Additionally, this model also allows to study the penalties to be assumed by the provider for failing to comply with the established response times.

B. *Input parameters*

Input parameters allow the configuration of different simulation scenarios. The following is a classification of the scenarios used in this study:

a) Parameters modeling the trends shown by the validation service requests.

- *Received Request Rate*: represents the trend of the validation service requests received.

b) SLA parameters agreed on between the service provider and the company:

- *Contracted Validation Rate*: service capacity contracted by the company.
- *Expected Response Time (ERT*): expected response time of service. If this time is exceeded, the provider is penalized and the request awaits validation.
- *Maximum Response Time (MRT)*: maximum response time of service. If this time is exceeded, the provider is penalized and the request is abandoned.
- *Validated Request Rate within ERT*: minimum percentage of received requests that must be validated within the expected response time.
- *Abandoned Request Rate:* maximum percentage of received requests that are permitted to be *abandoned.*
- *MRT Penalty*: penalty the provider would have to assume for each request validated within the maximum response time.
- *Abandonment Penalty*: Penalty the provider would have to assume for each validation request abandoned.

c) Validation Service Management Parameters

- *Card Validation Rate (CVR)*: card validation service capacity the provider assigns to the company.
- *CVR Percentage*: Percentage of the card validation rate that the provider uses to validate requests within the expected response time. The remaining percentage is allocated to validate the requests within the maximum response time.

C. *Output variables*

The main variables which provide information about the purpose of the model are:

- *Non Compliance within ERT*: deviation between the *Validated Request Rate within ERT* parameter set in the SLA and the rate of requests that are validated within the expected response time at any given time.
- *Abandonment Non Compliance*: deviation between the rate of requests that are abandoned at any given time and the SLA *Abandoned Request Rate* parameter.
- *RT Penalty*: Penalty the provider would have to assume for failing to comply with the response times agreed with the company, i.e., for the requests validated within the maximum response time and for the requests abandoned.

Other useful output variables for a better understanding of the system include:

- *Received Requests*: number of validation requests the service provider receives.
- *Validated Requests within ERT*: number of requests validated within the expected response time.
- *Validated Requests within MRT*: number of requests validated within the maximum response time.
- *Abandoned Requests*: number of requests abandoned.

D. *Model Conceptualization*

In order to map out the elements of the model and their relationships, *causal loop diagrams* have been used. Also, in order to model the system's behavior component and formally present its cause-effect associations, *stock and flow diagrams*, also known as Forrester diagrams [15], have been used. A stock and flow diagram represents the most important variables in a system or those variables whose behavior we wish to observe. The variables that represent how stock variables change with respect to time are known as flow variables. The aforementioned variables allow to model a system's operation policies and its decision-making rules. On a mathematical basis, stock variables are modeled through differential equations that integrate, throughout time, the difference between the input flow variables and the output flow variables of such a stock variable. Flow variables model the temporary function that regulates the change that occurs all the time on the stock variable. The resulting series of equations constitute the systems' mathematical model carried out through the simulation process.

In the proposed model, the *Validated Requests within ERT*, *Validated Requests within MRT* and *Abandoned Requests* output variables have been modeled as stock variables whose behavior pattern is controlled through flow variables namely, *Validation Rate within ERT*, *Validation Rate within MRT* and *Abandonment Rate*. *Non Compliance within ERT*, *Abandonment Non Compliance* and *RT Penalty* output variables have been modeled as auxiliary variables.

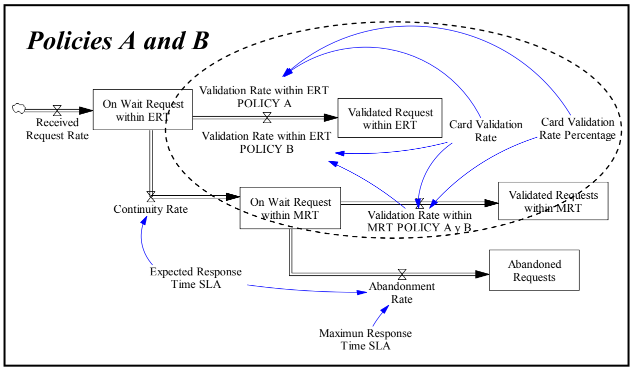
Policies A and B
Received Request Rate
On Wait Request within ERT
Validation Rate within ERT POLICY A
Validated Request within ERT
Validation Rate within ERT POLICY B
Card Validation Rate
Card Validation Rate Percentage
Continuity Rate
On Wait Request within MRT
Validation Rate within MRT POLICY A y B
Validated Requests within MRT
Expected Response Time SLA
Abandonment Rate
Abandoned Requests
Maximun Response Time SLA

Figure 1. Simplified stock and flow diagram

Fig. 1 describes the stock and flow diagram of the simplified model and details the input parameters and variables that influence the calculation of flow variables namely, *Validation Rate within ERT* and *Validation Rate within MRT* with *Policies A* and *B* of card validation rate management.

VI. Simulation Of The Model

In this section the configuration of the input parameters of the model in the case study is described. Also, sensitivity analyses are carried out, and their results analyzed and presented.

A. Configuration of input parameters for the case study

The model simulations were based on the assumption that the service provider and the e-commerce company have signed a certain SLA and that the validation service request follows a certain pattern. More specifically, the following input parameters have been considered:

1) SLA parameters agreed on between the service provider and the company.

- *Contracted Validation Rate*: 4395 requests/minute.
- *Expected Response Time:* 15 seconds.
- *Maximum Response Time:* 30 seconds.
- *Validated Request Rate within ERT*: 90% of received request rate.
- *Abandonment Rate*: 5% of received request rate.
- *MRT Penalty*: 0.2 euro per each validated request within maximum response time.
- *Abandonment Penalty*: 1.8 euro per each request abandoned.

2) Parameter that model the trend of the received validation requests.

Fig. 2 shows the *Received Request Rate* input parameter which represents the trend of the validation requests received by the server in a time frame.

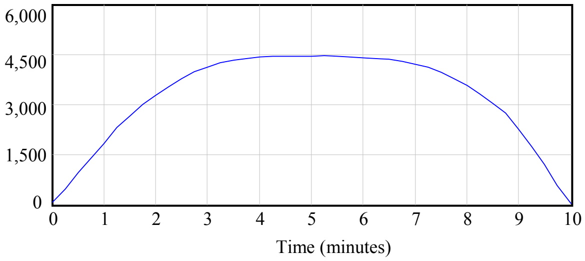

Figure 2. Behavioral pattern of *Received Request Rate* input parameter

It is considered that the received request rate experiments a gradual increase until a maximum value is reached to, then, remain constant for a certain period of time. Once finished, the aforementioned rate starts to decrease gradually again. The maximum value of this parameter corresponds to the sales register of Amazon during Christmas 2008 [16].

B. Sensitivity Analysis

The main goal of the sensitivity analyses carried out to determine the effects of *Policies A* and *B* on the fulfillment of the SLA parameters is as follows:

- *AS1:* To analyze the effects the above mentioned policies have on the SLA compliance with the card validation rate contracted by the company.
- *AS2:* To determine the lowest card validation rate which ensures the fulfillment of the performance parameters set in the SLA.
- *AS3:* To analyze the penalty for the non compliance with the response time with the card validation rate obtained in AS2.

1) AS1:To analyze the effects of the policies on SLA compliance with the card validation rate contracted by the company.

These sensitivity analyses allow to evaluate if the card validation rate contracted by the company guarantees the fulfillment of the performance parameters as stated in the SLA. Furthermore, it allows to study the penalty that the provider would assume if the agreed response time terms are not met. Additionally, the analysis of further output variables in the model helps understand the service behavior and evaluate the performance.

a) Parameter configuration of the sensitivity analysis.

Number of simulations: 200.
Input Parameters:

- *Card Validation Rate (CVR):* 4395 request/minute (SLA *Contracted Validation Rate* parameter).
- *SLA Parameters and Received Requests Rate:* case study values (see section A).

Control parameter of sensitivity analysis :

- *CVR Percentage*: varies between 1% and 100% *Distribution*: Random uniform.

b) Analysis of the fulfillment of agreed SLA parameters.

The data obtained in the sensitivity analyses indicate that in both policies the fulfillment of the SLA performance parameters with the contracted validation rate depends on the percentages of this rate used to validate requests within the agreed responses. In this case study under analysis, the following percentages need to be assigned to validate the requests within the expected response time:

- *Policy A: CVR Percentage* >= 96%
- *Policy B*: *CVR Percentage* >= 0%

c) Analysis of the penalty for non compliance with response times as set in the SLA.

Fig. 3 shows the *RT Penalty* output variable obtained with *Policy A*. The analysis of the data obtained indicates that the higher the *CVR Percentage* is, the later non compliance with response times takes place and the lower the final penalty for non compliance is. Therefore, the lowest final penalty is obtained with the *CVR Percentage* equals to 100% of the contracted validation rate.

Fig. 4 shows the *RT Penalty* output variable obtained with *Policy B*. In this case, non compliance with response times takes place at the same point in time for every simulation run and the non compliance penalty is the same with any *CVR Percentage* lower or equal to the 91% validation rate contracted. With higher percentages, the same penalty is applied until a certain point when it becomes higher. Therefore, the lowest final penalty is obtained with *CVR Percentages* which are lower or equal to 91% of the validation rate contracted.

After comparing the penalties obtained with both policies, it can be said that the lowest final *RT Penalty* is obtained with *Policy B* and with a *CVR Percentage* lower or equal to 91% of the contracted validation rate. Additionally, the dynamic feature of the model not only allows determining the final variable values, but also analyzing its evolution. Hence, it is observed that, with certain *CVR Percentages*, non compliance of response-time terms are not met occur earlier with *Policy A* than with *Policy B*. Thus, the provider is penalized earlier for such non compliance.

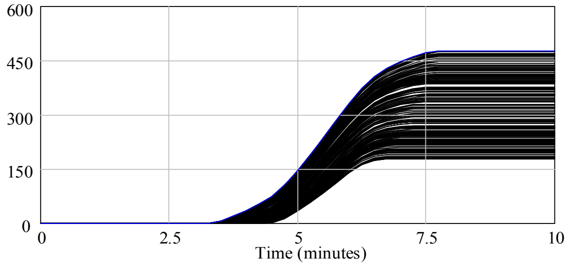

Figure 3. *RT Penalty* output variable in *AS1* (*Policy A*).

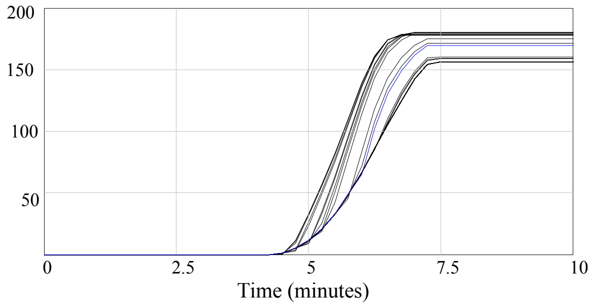

Figure 4. *RT Penalty* output variable in *AS1* (*Policy B*).

d) Analysis of other output variables in the model.

To understand the behavior of the card validation service and analyze its performance a deep understanding of the following output variables of the model is also useful: *Validated Requests within ERT*, *Validated Requests within MRT* and *Abandoned Requests*.

The following figures shows the values of these variables in the scenarios where the service performance level is the agreed one and the lowest penalty for not meeting the response times is obtained (see section c) of AS1). The values of the *CVR Percentage* parameter in these scenarios are as follows:

- *Policy A*: *CVR Percentage* = 100%
- *Policy B*: *CVR Percentage* <= 91%

Fig. 5 shows the *Validated Requests within ERT* output variable obtained with *Policy A* and *B*. The analysis of this variable indicates that at the end of the study period fewer requests are validated within the expected response time with *Policy B* than with *Policy A*. Moreover, the dynamic feature of the model allows to observe that the same number of requests are validated with both policies, though, at a certain point a larger number of requests are validated with *Policy A* rather than with *Policy B*.

Fig. 6 shows the *Validated Requests within MRT* output variable obtained with *Policies A* and *B*. No requests are validated within the maximum response time with *Policy A* and, with *Policy B*, requests within the maximum response time are validated during the time when the received request rate reaches its maximum value.

Finally, Fig. 7 displays the *Abandoned Requests* output variable obtained with *Policies A* and *B*. As it can be observed, no requests are abandoned with *Policy B* and, as to *Policy A* is concerned, requests were only abandoned when the received request rate reached its maximum value.

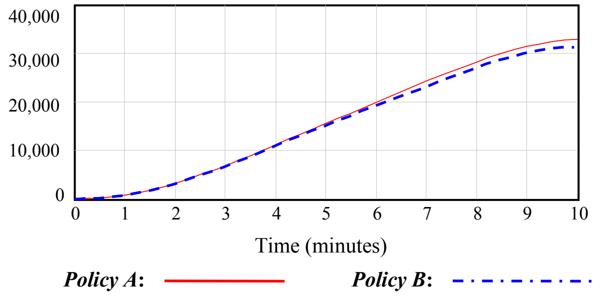

Figure 5. *Validated Requests within ERT* output variable in *AS1* (*Policies A* and *B*).

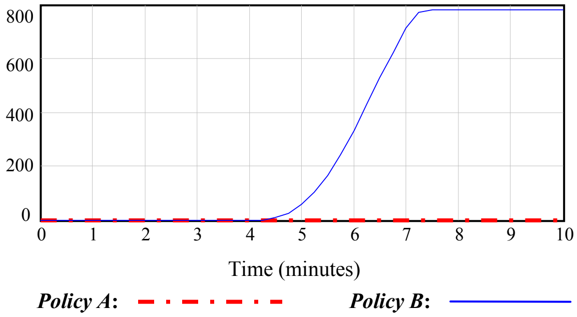

Figure 6. *Validated Requests within MRT* output variable in *AS1* (*Policies A* and *B*).

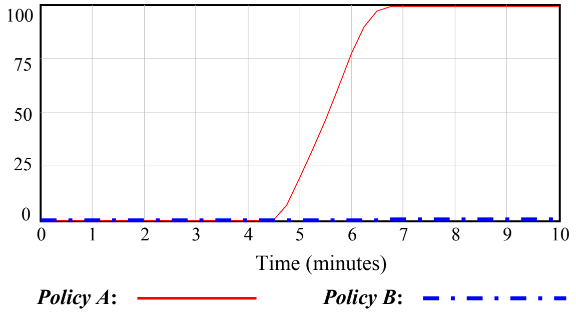

Figure 7. *Abandoned Requests* output variable in *AS1* (*Policies A* and *B*)

2) AS2:To determine the lowest card validation rate that guarantees the fulfillment of the performance parameters established in the SLA.

a) Configuration of the parameters of the sensitivity analysis.

Number of simulations: 200.
Input parameters:

- *SLA Parameters and Received Request Rate:* case study values (see *section A*).

Control parameters of the sensitivity analysis:

- *Card Validation Rate:* varies between 3500 requests/minute and the maximum value of the trend of the validation requests received (4459 requests/ minute). *Distribution:* Random uniform.
- *CVR Percentage:* varies between 1% and 100%. *Distribution:* Random uniform.

b) Determination of the lowest card validation rate

The data obtained from these sensitivity analyses indicate that the lowest card validation rate which guarantees the fulfillment of SLA performance parameters is 4239 requests/minute with both policies. Additionally, they indicate that a validation rate of the received requests within the expected response time needs to be higher or equal to 95% of such a card validation rate.

3) AS3: To analyze the penalty for not meeting the response-time terms with the card validation rate obtained in AS2.

a) Configuration of the parameters of the sensitivity analysis.

Number of simulations: 200.
Input parameters:

- *Card Validation Rate*: 4239 requests/minute, lowest credit validation rate which ensures that the service performance matches the one agreed (see *AS2*).
- *SLA Parameters* and *Received Request Rate*: case study values (see *section A).*

Control parameter of sensitivity analysis:

- *CVR Percentage*: varies between 95% and 100% (see *AS2*). *Distribution:* Random uniform.

b) Analysis of penalties for non compliance with response times as set in the SLA.

Fig. 8 shows the *RT Penalty* output variable that is obtained in the sensitivity analysis with respect to *Policy A*. Similarly, as with the card validation rate contracted (see Fig. 3), it is observed that the higher the *CVR Percentage* is, the later the response-time terms are unmet and the lower the final penalty is. Therefore, the lowest penalty is obtained with a *CVR Percentage* equal to 100% of the card validation rate.

Fig. 9 shows the *RT Penalty* output variable obtained in the sensitivity analysis with respect to *Policy B*. It can be observed that this penalty follows a different trend to that displayed with the card validation rate contracted by the company (see Fig. 4).

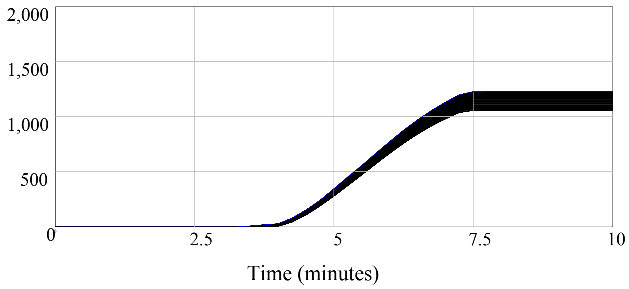

Figure 8. *RT Penalty output variable in AS3 (Policy A).*

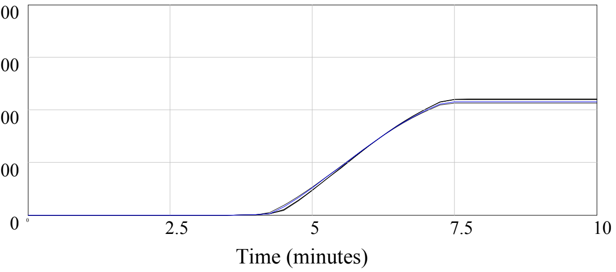

Figure 9. *RT Penalty output variable in AS3 (Policy B).*

As to the evolution in time the *RT Penalty* variable obtained from the sensitivity analysis, it must be noted that until a specific point in time (6.25 minutes) the lowest *RT Penalty* is obtained with a *CVR Percentage* equal to 95% of the card validation rate. From that moment onwards, the lowest *RT Penalty* is obtained with a *CVR Percentage* equal to 100%. Therefore, the lowest *RT Penalty* is obtained by assigning the card validation rate percentages to validate the requests within the expected response time in the following way:

- Until a specific point in time (6.25 minutes) the *CVR Percentage* is equal to 95% of the card validation rate.
- From that point in time onwards, the *CVR Percentage* is equal to 100% of the card validation rate.

From the comparison of the resulting *RT Penalty* between *Policy A* and *B* it can be observed that the lowest penalty is obtained with *Policy B* and with variations in time of the *CVR Percentage* (initially 95% and, 100% from a certain point in time). The dynamic feature of the model helped determine the point in time at which the optimum value of this percentage changed.

Thus, the model helps not only in determining the management polices that ensure the lowest *RT Penalty* but also in analyzing the evolution of this variable throughout time. Also, the model helps providers decide on when to change *CVR Percentages* so that the *RT Penalty* is the lowest possible during the whole period of study and its end.

VII. Conclusions. Futher Studies

This study suggested a dynamic simulation model applied within the field of ITIL *Capacity Management Process*. The main aim of the model is to help manage adequately the web service capacity that providers assign their customers to ensure the fulfillment of the SLAs. In order to achieve this, the model helps analyze the effects different capacity management policies have on web services performance and on the penalties that the providers would have to assume if the established response- time terms were not met.

The sensitivity analyses in this study have been based on the variation of service capacity, management criteria (*Policies A* and *B*, as described in section IV) and capacity percentages which are used to respond to customers' requirements within the response times agreed with the provider. Thus, the main goal of the aforementioned sensitivity analyses was as follows:

- To evaluate whether the actual service performance with its corresponding contracted capacity matches the agreed one in the SLA.
- To determine the lowest service capacity that guarantees the fulfillment of the SLA performance parameters established.
- To evaluate the penalties service providers would have to assume for not meeting the agreed response-time terms.

A. Findings

After a careful data analysis the following conclusions can be drawn:

- With the card validation rate contracted by the company, the service performance matched the one established in the SLA assigning the following percentages of card validation rates to validate the requests within the expected response time: *CVR Percentage* > = 96% with *Policy A* and *CVR Percentage* > = 0% with *Policy B*.
- With the card validation rate contracted by the company, the lowest penalty for not meeting the response-time terms was obtained with *Policy B* and with a validation rate of received requests within the expected response times lower or equal to 91% of the contracted validation rate.
- The lowest card validation rate that guarantees the fulfillment of SLA performance parameters, was the same with both policies (4239 requests/minute). Additionally, the necessary percentages of this capacity to validate the requests within the expected response time remained the same with both policies (C*VR Percentage* > = 95%).
- The dynamic feature of the proposed model also helped analyze the evolution of the variables throughout time. In this case, it is also observed that with a certain card validation rate, the response-time

terms were unmet at an earlier time with *Policy A* rather than with *Policy B*.

- With the lowest card validation rate that guarantees that the service performance is the one agreed (4239 requests/minute), the lowest penalty for not meeting the response-time terms is obtained with *Policy B* and varying the percentages of the card validation rate used during the study period to validate the requests within the established response times. Initially, the optimum *CVR Percentage* is 95% and, from a certain point in time this percentage reaches 100%. The dynamic feature of the model helps determine the adequate moment to change this percentage.

Therefore, the actual service performance with a certain card validation rate proved to be better with *Policy B* than with *Policy A*.

B. *Further Studies*

The main aims of further studies are as follows:

- To study the effects the different values of the SLA parameters have upon the service performance and the penalties providers would have to assume for not meeting the performance levels previously agreed. Service providers could make use of the findings of this study as a starting point to renegotiate the SLA established with their customers.
- To analyze the effects management policies have on the fulfillment of the SLAs with respect to the different trends of the received service requests.
- To broaden the model by including further capacity management policies, different to those studied in this paper.

The dynamic feature of the proposed model will help not only to determine the final values of the variables of the model under study but also to analyze their progress in time.

Acknowledgments

This research paper was partially financed by the Ministerio de Educación y Ciencia de España (Ministry of Education and Science of Spain) and the FEDER European funds, via the TIN2007-67843-C06-04 project.

References

[1] ITIL oficial site: http://www.itil-officialsite.com.

[2] Office of Government Commerce, “The Official Introduction to the ITIL Service Lefecycicle”, The Stationary Office (TSO) (2007).

[3] itSMF International, “Foundations of IT Service Management on ITIL”.,Van Haren Publishing (2006).

[4] Office of Government Commerce, “Service Design”, The Stationary Office (TSO) (2007).

[5] D. Liu, R. Deters, “Management Service-Oriented Systems”, In: Springer London (eds). Service Oriented Computing and Applications 2(2-3). pp 51-64 (2008).

[6] M. Ruiz, I. Ramos, M. Toro, “Software Process Dynamics: Modeling, Simulation and Improvement”, In Acuña, S., Sánchez-Sergura, M. (eds.) New Trends in Software Process Modeling. Series on Software Engineering and Knowledge Engineering. Vol. 18. World Scientific Publishing, 2006.

[7] B. Shen, “Support IT Service Management with Process Modeling and Analysis”, In Q.Wang, D.Pfahl, and D.M. Raffo (eds.): JCSP 2008, LNCS 5007, pp.246-256. Springer-Verlag Berlin Heidelberg (2008).

[8] T. Setzer, K. Bhattacharya, “Decision Support for Service Transition Management Enforce Change Scheduling by Performing Change Risk and Business Impact Analysis”, In Network Operations and Management Symposium, 2008. NOMS 2008. IEEE. pp.200-207 (2008).

[9] C. Bartolini, M. Sallé, D. Trastour, “IT Service Management Driven by Business Objectives: An Application to Incident Management”, In 10th IEEE/IFIP Network Operations and Management Symposium (NOMS) (2006).

[10] G. Barash, C. Bartolini, W. Liya, “Measuring and Improving the Performance of an IT Support Organization in Managing Service Incidents”, In Business-Driven IT Management (BDIM’07), pp.11-18 (2007).

[11] J.H. Lee, Y.S. Han, C.H. Kum, “IT Service Management Case based Simulation Analysis & Design: Systems Dynamics Approach”, In IEEE International Conference on Convergence Information Technology, pp 1559-1566. IEEE Computer Society, Washington, DC (2007).

[12] L. An, J.J. Jeng, ”Web Services Management Using System Dynamics”, In Proceedings of the IEEE International Conference on Web Services (ICWS’05), pp.347-354 (2005).

[13] M.I. Kellner, R.J. Madachy, D. Raffo, “Software process simulation modeling: Why? What? How?”, J. Syst. Software, 46(2-3): 91-105 (1999)

[14] I.J. Martínez, G.P. Richardson, “Best Practices in System Dynamics Modeling”, Proceedings of the 19th International Conference of the Systems Dynamics Society. Atlanta, USA (2001).

[15] J.W. Forrester, Industrial Dynamics. Pegaus Communications, Massachussets (1961).

[16] The Guardian official site http://www.guardian.co.uk/technology/blog/2008/dec/26/amazon-xmas08.