SLA Impact Modeling for Service Engagement

Yixin Diao, Linh Lam, Larisa Shwar
IBM Thomas J. Watson Research Cen
Yorktown Heights, NY 10598, US/

***Abstract*— During the customer engagement phase it is c for the service providers to estimate the impact of servic constraints on service personnel needs. However, it is difficult due to the implication from customer workload. I paper we propose an SLA impact evaluation methodolog uses queueing models to quantitatively evaluate the imp SLAs to the engagement cost model.**

I. Introduction

In today's world, the technological environment ch rapidly and spontaneously, thus creating an unpredi environment for businesses that attempt to ride the latest nological wave. Environments that acquire little predicta of the customer's needs in 3-5 years must have suff agility to respond with quality offerings that are base the customers' predictions of their own future needs. areas, like Cloud Computing, depend on standardizatic delivering the benefit of low cost services to clients. As standardization is a great driver of progress for such and periods of time. While clouds provide the flexibil accommodate requests for infrastructure changes, they re how services are delivered, locking customers into sta service level agreements.

Standard level agreements are designed to optimiz delivery of services on cost, which is one of the key pro of IT strategic outsourcing. However, the ability to re flexible service levels from a service provider could be c a business. This raises a critical question from the po view of the engagement and delivery teams: how cai quantify and predict the necessary delivery staffing that will deliver promised SLAs?

One common strategy used in engagement at predicting the labor cost in IT service management is through the use of the engagement cost model. However, most of the the strategic outsourcing applications (e.g., Unix and Intel platform support, console monitoring) are more sophisticated than call centers applications, because of the existence of multiple ticket classes as well as the non-ticket workload.

In previous work [1] we have discussed a modeling framework based on discrete event simulation. It works well in service delivery. However, we cannot apply it directly to engagement because it has the challenge of getting data at engagement time.

To address the above issue, in this paper we propose an SLA impact evaluation methodology that uses queueing models to quantitatively evaluate the impact of SLAs to the engagement cost model. The proposed approach calculates the SLA impact

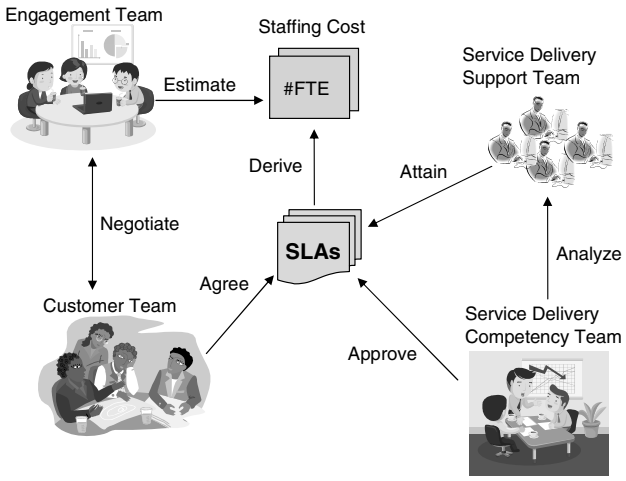

Fig. 1. SLA lifecycle management.

while considering delivery workload and effort. It extends the Erlang-C model commonly used in help desk modeling to fit into the complex environment of service delivery.

The remainder of this paper is organized as follows. Section II discusses the background for service delivery systems and engagement as well as the challenges for SLA impact modeling. Section III presents the proposed SLA impact model. Section IV reviews related work. Our conclusions are contained in Section V.

II. Service Delivery Systems and Engagement

A. SLA Lifecycle Management

In today's challenging business environment, many businesses make the decision to focus on their core business and outsource their ever complex IT needs to service delivery providers to help keep their current business running and to continue to meet their IT demands as their business grows. In order to ensure there is no unambiguous expectation between the parties, a written agreement called Service Level Agreement (SLA) detailing all the targeted services and their level of service delivery quality is put in place between the customer and the IT service delivery so that there will be no misunderstanding or any later dispute about key service targets and responsibilities of both parties [2]. From a high level glance and for the scope of this paper, we will focus our discussion around the involvement of SLA lifecycle management around four main parties as depicted in Figure 1.

In Figure 1, during engagement time, the Engagement Team of a selected service delivery provider works with the Customer Team to collect their IT and business requirements either through interviews or questionnaires. With the specified requirements, the two teams works together to draft a contract and start to negotiate on terms, conditions, service to be delivered along with the prices. As the negotiation is going on, the Engagement Team estimates the cost case model for this contract which is then used to drive the SLAs. There are various parameters that are important inputs to the cost case model such as the numbers of different servers to be supported, the types of applications and whether they fit the service delivery provider's standard offerings or not. However, the essence of the cost case model is to estimate the staffing cost (the number of FTEs needed for this contract) so that the SLA attainment targets could be met and the service delivery provider remains profitable during the lifecycle of this contract.

As part of the contract negotiation, objective quantitative measurements of availability and performance of IT system as well as business process specified in SLAs will also be negotiated. For example, to express acceptable service downtime (e.g. "monthly availability of Individual Web Server will be no less than 99.7 percent") or refund policies for missing service level targets downtime (e.g. "credit customer two thousand dollars if a monthly average network latency across the provider ISP access links to the backbone is higher than 95 milliseconds") [3]. Before the SLAs are finalized and agreed by the Customer Team, the SLAs have to be reviewed and approved by the Service Delivery Competency Team to make sure the targets promised by the Engagement Team to the customers are actually deliverable and attainable by their Service Delivery Support Teams. Since SLA negotiation happens during the engagement time and there is no customer data from the customer to validate the proposed SLAs, the Service Delivery Competency Team approves the proposed SLAs largely based their experience and on similar SLAs of customers from the same industry.

Once the SLAs are live, they will be monitored and tracked in to ensure service quality and performance targets are met. Analysis can be done to see if the estimates predicted by the cost case model match the actual cost of the Service Delivery Support Team. If the discrepancy is too big, the allocation of the staffing may need to be adjusted and remediation or recommendation will need to feedback to the Engagement Team for future SLAs negotiation.

B. Modeling Service Delivery

Service delivery involves customers contracting with a services provider on a menu of IT services such as security patch management, network management, and data backup and restores management [4]. In order for the service delivery provider to fulfill such services, they usually have teams of service agents serving multiple customers on their work requests. These agents typically are grouped into teams according to the depth and breadth of the skills they have, where the breadth of skill (i.e. knowledgeable in a range of IT areas) and the depth of skills (i.e. master the skill in specific in those IT areas). Agents are usually assigned to multiple customers whose service support requires the agents' skill sets. As service requests arrive with the severity specified by the customers, the service delivery center will still prioritize the requests based on the customer's SLA in relative to its scope, target time and percentage attainment for the time frame that being measured. It has always been a challenge of the service delivery center to balance the right number of staffing at different shifts so that even with the random arriving workload, the handful of agents at any shift can help to resolve the requests without missing SLAs targets and avoid the associated penalties.

Facing the constant challenge to stay competitive in cost, performance and customer satisfaction, the service delivery providers use various modeling techniques to help them balance workload request, service agents service level targets. However, there are a number of complexities in trying to model a service delivery center properly. First, in reality requests with different severity come in at a random rate over the course of business/calendar and days of the week, they could be non-uniform. Second, the center could be staffed with agents with different level skill sets. Third, it is difficult to model the exact timing of breaks agents take randomly during their shift. Lastly, the service level attainment level is not measured on per request basis but instead measured through out a period of time like a month. In addition, the service target time can be either in the unit of calendar hours or business hours; in the latter case, a business calendar is required [4].

To address these modeling complexities, simulation approach which allows users to accurately capture the complexity of the reality tends to work better than the analytical models. As in [1], we chose simulation approach since we were interested in modeling a services delivery environment. The solutions obtained must be relevant for the real world services delivery environment that was being modeled because the solutions would be implemented in practice. Further more, simulation modeling allows for what-if analysis and can support the delivery manager to quantitatively measure the cost related to service delivery decisions such as (i) changing the mix of customers assigned to a team of agents, (ii) unskilled agents (improving agent knowledge or increasing agent service time), or (iii) including unreasonable contract terms in customer contracts.

However, during engagement time because of the lack of data availability and the necessity of fast turn around negotiation time, we choose to use the analytical model to capture the key queuing effect and apply approximation techniques to get closer to realistic. Such an analytical model is simple, fast and can provide a good estimation on SLA attainment with the right level of staffing.

III. SLA Impact Model

In this section we describe the modeling framework for calculating the SLA impact while considering the interaction

ISBN 978-3-901882-53-1, 9th CNSM and Workshops ©2013 IFIP

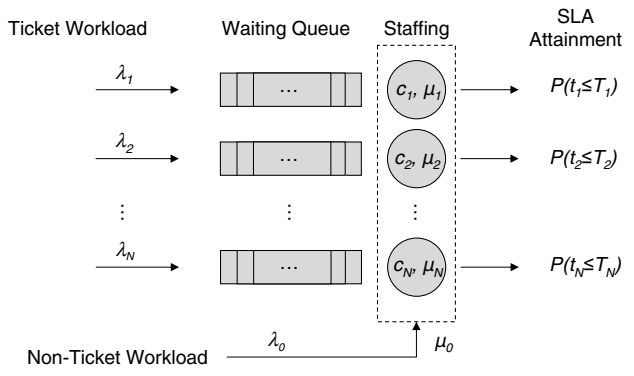

Fig. 2. SLA impact model.

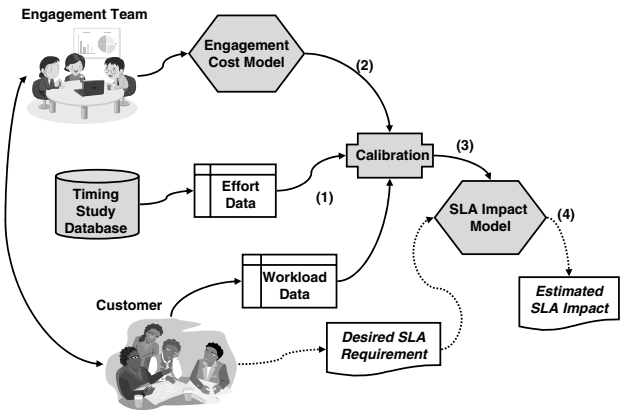

Fig. 3. Approach to constructing and applying the SLA impact model.

with service workload and delivery effort.

The overall model structure is first discussed in detail. Subsequently, we summarize the model usage scenario.

Queueing theory is the study of "the phenomena of standing, waiting, and serving" [5]. Many forms of queueing models have been studied in a vast of literature and applied to various real world problems. In this paper, we choose to build our model based on the Erlang-C formula, which captures the multiple servers (a.k.a., service agents) nature of service delivery and has the closed form solution for tail probabilities.

The Erlang-C formula is also known as the $M/M/c/\infty$ or simply $M/M/c$ queueing model in Kendall notation. An $M/M/c$ model assumes Poisson arrivals, exponential service times, c number of servers, infinitely many waiting positions, and first-come first-served queueing discipline. The Erlang-C model has been widely used in telecommunication systems and especially call center applications to calculate the required number of operators subject to given call volume and desired waiting time. However, we cannot directly apply the Erlang-C model to service delivery applications. because of the existence of multiple ticket classes as well as the non-ticket workload.

In this paper we propose a mixed multi-queue Erlang-C model as an extension of the standard single-queue Erlang-C. As shown in Figure 2, the overall ticket workload is divided into multiple sub-workload where λ_i indicates the arrival rate for class $i, i = 1, 2, \ldots, N$. Each queue is servicing an single ticket service class with c_i service agents and μ_i per-agent service rate. The SLA attainment level can be calculated as the probability that the ticket resolution time t_i will be less than or equal to the target time T_i

$$P(t_i \leq T_i) = \frac{c_i - x_i - 1 + C(c_i, x_i)}{c_i - x_i - 1}\left(1 - e^{-\mu_i T_i}\right) - \frac{C(c_i, x_i)}{c_i - x_i - 1}\left(1 - e^{-(c_i - x_i)\mu_i T_i}\right) \quad (1)$$

$$C(c_i, x_i) = \frac{\frac{x_i^{c_i}}{c_i!}}{\frac{x_i^{c_i}}{c_i!} + \left(1 - \frac{x_i}{c_i}\right)\sum_{k=0}^{c_i - 1}\frac{x_i^k}{k!}} \quad (2)$$

where $x_i = \lambda_i/\mu_i$ denotes the traffic intensity and $C(c_i, x)$ denotes the Erlang-C function [6].

In addition to the ticket workload, we specify the service agents to work on non-ticket work if the corresponding ticket work queue is empty. This specification allows the work conserving behavior of each queue and keeps the above SLA attainment level calculation intact. Given non-ticket workload with arrival rate λ_0 and service rate μ_0, the following inequality defines the stability condition to be satisfied for maintaining the overall queueing stability

$$\sum_{i=0}^{N}\frac{\lambda_i}{\mu_i} < \sum_{i=1}^{N} c_i \quad (3)$$

in addition to the per-queue ticket workload stability condition $\lambda_i/\mu_i < c_i, i = 1, 2, \ldots, N$.

The main scenario of using the SLA impact model is to estimate the labor cost (or FTEs) under different SLA terms. The estimation procedure consists of four distinct steps, which is depicted in Figure 3 and summarized as follows:

1) Collecting Workload and Effort Data: There are two sets of data to be collected for constructing the SLA impact model: workload data and effort data. The workload data captures the incoming ticket workload. It includes the workload volume and severity percentage. Both can be obtained from the ticketing system. The effort data defines the working time that the service agents spend on the ticket. Since this data is usually unavailable for new customers, we resort to the historic effort database which stores the effort time from existing customers collecting through timing studies.

2) Defining Baseline Staffing: While the SLA impact model quantifies the SLA impact through the queueing-based delivery operation, it does not capture all engagement factors that are used to estimate the labor cost. In contrast, the engagement team typically uses the so-called engagement costing model to quantify the impact of a large number of engagement factors but not the SLAs in most cases. Due to the complementary nature of the two models, we resort to the engagement costing model first to come up with the baseline staffing from the standard SLAs.

3) Calibrating Model Parameters: We calibrate the SLA impact model to ensure that under the standard SLAs the

ISBN 978-3-901882-53-1, 9th CNSM and Workshops ©2013 IFIP

model will give the same staffing recommendation as the engagement costing model. This is conducted by calibrating the non-ticket workload volume and effort time. Also note that the non-ticket workload volume and effort time are typically not available from the customer as the data input.

4) Conducting SLA What-if Analysis: Given the calibrated SLA model, we will be able to conduct two types of what-if analysis. The first is to calculate the SLA attainment level subject to the staffing and SLA requirements. The second is to calculate the staffing subject to meeting the SLA requirements.

IV. Related Work

As it is important to avoid missing SLA target during runtime, there are quite a few publications touched on how to avoid SLA violation from different architecture angles. [7] proposed an approach for SLA violation prediction model based on machine learning regression techniques and trained using historical process instance at defined checkpoints during runtime for composite services. [8] proposed an approach using analysis on historical workload and applied predictive and reactive provisioning to minimize SLA violations for their data center in power consumption and resource allocation. [9] presented a generic ontology-based approach to monitor SLA for violation.

From a slightly different angle [10] proposed a model to maximizing profit for service delivery provider while minimizing the SLA violation probability using a combination of efficient SLA-aware routing and intelligent admission control. Authors from [11] investigated profit maximization model using mathematical model with optimization-based profit maximization strategy for data center to voluntary load reduce while still meeting the commitment to their SLA customers. [12] proposed to use simulation with stochastic approximation to reach optimal staffing and while maintaining service level attainment. In [13], the authors proposed using fixed-queue-ratio rules for routing their workload in order to keep staffing cost optimal while achieving SLA targets. [14] modeled a single-class calling center using a queuing model with Poisson arrivals to balance the number of staffing including hiring temporary operators without missing the SLA target.

[1] described a simulation model using real historical data from large service delivery center to provide a baseline as-it and what-if analysis for their system performance with their SLA attainment. [15] focused on modeling the dispatching aspect of the delivery service center where incoming requests were prioritized according to the up-to-date SLA attainment targets. The works discussed above were SLA modeling related either for avoiding violations or for aiming to maximizing profit, they all focus on modeling SLAs during runtime. In contrast, our paper is looking at the whole lifecycle of SLAs and focuses on the SLA engagement/negotiation phase, with the challenge of limited data availability.

V. Conclusions and Future Work

During the customer engagement phase it is critical for the service providers to estimate the impact of service level constraints on service personnel needs. However, it is often difficult due to the implication from customer workload. In this paper we proposed an SLA impact evaluation methodology that uses queueing models to quantitatively evaluate the impact of SLAs to the engagement cost model.

The model can answer many questions that our current tools cannot, especially with respect to the impact of SLAs on resource requirements. Is the current staffing sufficient given the expected workload (both incident and non-incident) without backlog growth? Are the SLA target times feasible given the expected service times? Is the current staffing sufficient to meet the SLA target times?

While the initial results are encouraging, there are several areas for further research. First, we would like to extend the model structure to capture more realistic features of service delivery. Second, we would like to have complete evaluation studies to help us bring the model to its full capability.

References

[1] Y. Diao, A. Heching, D. Northcutt, and G. Stark, "Modeling a complex global service delivery system," in *Proceedings of 2011 Winter Simulation Conference, Phoenix, Arizona*, 2011, pp. 690–702.

[2] Office of Government Commerce, "IT Infrastructure Library. ITIL Service Support, version 3," 2007.

[3] M. J. Buco, R. N. Chang, L. Z. Luan, C. Ward, J. L. Wolf, and P. S. Yu, "Utility computing SLA management based upon business objectives," *IBM System Journal*, vol. 43, pp. 159–179, 2004.

[4] Y. Diao and A. Heching, "Staffing optimization in complex service delivery systems," in *Proceedings of 7th International Conference on Network and Service Management, Paris, France*, 2011.

[5] L. Kleinrock, *Queueing Systems. Volume 1: Theory*. Wiley-Interscience, 1975.

[6] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*. John Wiley and Sons, 1998.

[7] P. Leitner, B. Wetzstein, F. Rosenberg, A. Michlmayr, S. Dustdar, and F. Leymann, "Runtime prediction of service level agreement violations for composite services," in *Proceedings of 2009 international conference on Service-Oriented Computing*, 2009, pp. 176–186.

[8] A. Gandhi, Y. Chen, D. Gmach, M. Arlitt, and M. Marwah, "Minimizing data center SLA violations and power consumption via hybrid resource provisioning," in *International Green Computing Conference and Workshops*, 2011, pp. 1–8.

[9] K. Fakhfakh, T. Chaari, S. Tazi, K. Drira, and M. Jmaiel, "A comprehensive ontology-based approach for SLA obligations monitoring," in *Proceedings of the 2nd International Conference on Advanced Engineering Computing and Applications in Sciences*, 2008, pp. 217–222.

[10] A. Das, "Maximizing profit using SLA-aware provisioning," in *Proceedings of IFIP/IEEE Network Operations and Management Symposium*, 2012, pp. 393–400.

[11] M. Ghamkhari, "Data centers to offer ancillary services," in *Proceedings of Third International Conference on Smart Grid Communications*, 2012, pp. 436–441.

[12] Z. Feldman and A. Mandelbaum, "Using simulation based stochastic approximation to optimize staffing of systems with skills based routing," in *Proceedings of the 2010 Winter Simulation Conference*, 2010, pp. 3307–3317.

[13] I. Gurvich and W. Whitt, "Service-level differentiation in many-server service systems via queue-ratio routing," *Operations Research*, vol. 58, no. 2, pp. 316–328, 2010.

[14] C. Feng and H. Jia-zhen, "Dynamic operator staffing problem for call centers with uncertain arrival rate," in *Proceedings of the 2011 International Conference on Business Computing and Global Informatization*, 2011, pp. 619–621.

[15] Y. Diao and A. Heching, "Closed loop performance management for service delivery systems," in *Proceedings of IFIP/IEEE Network Operations and Management Symposium*, 2012.