A Quantitative Approach to the Performance of Internet Telephony to E-business Sites

Prathiusha Chinnusamy
TransSolutions
Fort Worth, TX 76155, USA

Natarajan Gautam
Harold and Inge Marcus Department of Industrial & Manufacturing Engineering
Pennsylvania State University
University Park, PA 16802, USA

Abstract

We consider a case study of an airline reservation system that employs web-based reservations as well as conventional telephone reservations. For this system we use analytical models based on queueing theory to evaluate the enhancement in performance when this system has an Internet telephony component added. We observe that the revenue and the customer satisfaction increases however with a caveat that the load in the system increases as well.

Keywords System performance, capacity planning, queueing networks, simulation, customer service

1 Introduction

The phenomenal growth of the Internet in last few years has provided a new medium through which consumers can purchase goods and services. In December of 1999, NewsNow reported that estimates show that Internet shopping had tripled since the previous year [1]. This fast rate of increase in on-line purchases is garnering exceptional revenue for e-commerce sites; however, the profits of these businesses are far from being capitalized to their fullest extent. Studies show that a majority (67%) of on-line purchases never go through to completion due to a lack of real-time customer service and support [2]. A fear of out-of-stock items, late delivery, and customer service difficulties causes customers to become hesitant during their on-line buying experience, which results in the abandonment of purchases midway through the sales process. We consider the airline industry and perform a case study of the airline reservation system. Presently there are two modes by which a customer can purchase a ticket directly from the airline: via the telephone and via the World Wide Web. On one hand although most airlines offer discounts to purchase tickets via the web, there are still several customers who prefer to use the telephone because they would like to speak to a reservations agent. On the other hand since each reservations agent spends a lot of time taking down information about the customer, it results in very large waiting times at the agent queue.

The technology that could alleviate customer-service concerns for e-businesses and industries is Internet Telephony. Using Internet Telephony (based on voice-over-IP, i.e. VoIP technology), it will be possible to interact with a customer service agent simultaneously via two forms: data and voice. For the airline reservations case study, using Internet Telephony it will be possible for a customer to show the reservations agent what information he/she has entered (such as their name, origin, destination, dates, etc) and also talk to the agent simultaneously. This way not only do the service times of the reservations agents reduce (since they do not have to enter customer information) but also the customers using the Web who wish to talk to an agent could do so using Internet Telephony. In general, with the addition of an Internet telephony option that directly connects users to customer service representatives from e-commerce sites, customer questions and concerns can be addressed in a timely manner, thus potentially gaining the confidence of consumers and significantly increasing the completion rate of on-line purchases. Also, users will no longer be inconvenienced by having to place a phone call from a separate line or by disconnecting from the line they are currently using to contact customer service agents. The ability to browse and talk simultaneously from a single connection will prevent customers from prematurely ending their shopping due to a frustration with the limited real-time customer support available.

As mentioned above, a case study of an airline reservation system is considered in this paper. The current system, which will be referred to as W/P, accepts reservations from the web and the traditional telephone; the proposed system, which will be referred to as W/P/IT, will provide better customer service by enabling web customers to have

the additional option of using Internet telephony to place calls to customer service representatives. We focus on comparing the revenues obtained by modeling the W/P system and the W/P/IT system using queuing network models with reneging. There is a considerable amount of research available on reneging in a single queue [3], but there is very limited literature that addresses reneging in a queuing network. The primary reason for this lack of available literature is that the product form structure [5] is lost and hence this makes analysis intractable. We develop suitable approximations in this paper.

The analysis of these systems will first be done as three individual units (a web-based reservation system, a traditional telephone reservation system, and a reservation system with an Internet telephony component) to study the raw performance of the individual systems. Section 2 contains the analysis of the individual reservation systems. Approximations are developed to determine the output rate of customers at the web server and the total service time of the customer service agents. The approximations are validated using simulation. In Section 3, we combine the three separate systems into the current W/P system and the proposed W/P/IT system and compare their performances. The comparisons illustrate that adding an Internet telephony component results in increased system performance. Section 4 presents conclusions along with recommendations for expanding this research.

One of the key findings of this paper is that although Internet Telephony intuitively seems like a very attractive option for e-businesses, it could come at the price of increased load in the system that could degrade the performance unless the system capacity is planned appropriately. Other contributions of this paper include the models developed that will serve as tools for companies who want to evaluate the effect on the system performance by adding Internet telephony. The main quantitative contribution presented in this research is the derivation of approximations that can be used to compute the performance of these systems.

2 Analysis of Individual Systems

In this section, the architecture of three different airline reservation systems is introduced: a web-based reservation system, a traditional telephone reservation system, and a reservation system with an Internet telephony component. The modeling and analysis of these three systems will be done individually in this section and as integrated units in Section 3. These systems are studied individually to analyze and compare their undiluted performances and to develop approximations at the most basic level.

2.1 Web-Based Reservation System

In this system, customers log on to a website that sells airline tickets. After spending some time browsing and entering information, customers will either make a request to the reservation server or idle for a while and browse more. We model the idling using a dummy station called the think station where a customer spends thinking before making their next request, without tying up either one of the server resources. From the think station, the customer will either decide to go back to the web server or exit the system. There is a possibility that a customer will loop between the think station and the web server multiple times before exiting the system. Just before exiting the system, there is a probability associated with whether or not a sale is made. A schematic representation of the flow of customers through this system is included below, with queues denoted as rectangles and servers denoted as circles. Probabilities are assigned to which path a customer will follow after departing one of the servers.

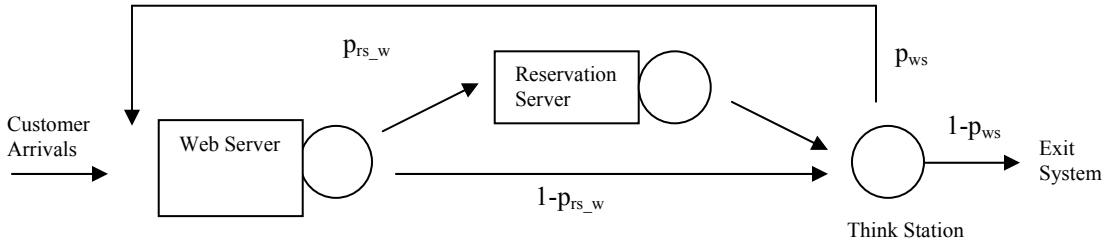

Figure 1 Schematic Diagram of Web System

This system is modeled as an open queuing network of 3 queues/stations. Inter-arrival times into the system are assumed to be iid exponential random variables with a mean of $1/\lambda_w$. All service times are assumed to be iid exponential random variables with mean $1/\mu_{ws}$, $1/\mu_{ts}$, and $1/\mu_{rs}$ for the web server, the think station, and the reservation server respectively. Reneging can occur only at the web server. It is assumed that the time a customer waits before reneging at the web server queue is an iid random variable with a mean of $1/\theta_w$. Since there is reneging at the web server queue, standard Jackson Network results [4] cannot be applied, as the output rate from this server

is no longer Poisson. Therefore, in order to compute the overall renege rate of customers in this system, an approximation is made that the output process from the web server is Poisson with mean β_w. Also, the effective arrival rate to the web server is α_w. The quantities α_w, and β_w can be computed recursively using:

$$\beta_w = \frac{\alpha_w \mu_{ws}}{\mu_{ws} p_0 + \alpha_w}; \quad \alpha_w = \lambda_w + p_{ws}\beta_w; \quad p_0 = \frac{1}{\sum_{i=1}^{\infty}\prod_{j=0}^{i-1}\frac{\lambda_w}{\mu_{ws}+j\theta_w}+1} \qquad (1)$$

The derivation of the expressions in Equation (1) will appear in a forthcoming paper. Based on the queueing network results, the system revenue is computed. The analytical approximations yielded very good results as indicated by Figure 2 where the results are compared with simulations for 3 different input parameters (or indices).

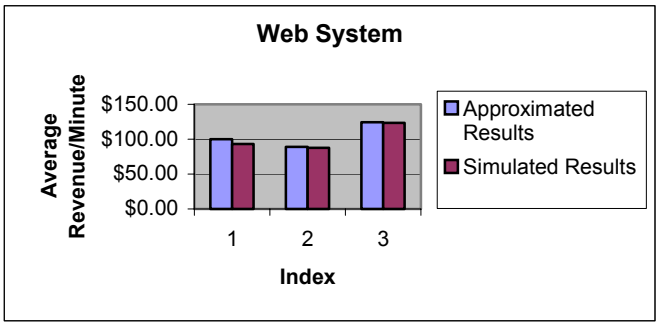

Figure 2 Comparison of Approximated Results vs. Simulated Results for Web System

2.2 Traditional Phone Reservation System

In this system, a customer places a telephone call to a human service agent. The agent works exclusively with this customer only and has the ability to interact with the reservation server to make requests. Thus, while an agent is processing a customer, there is a possibility of looping back and forth from the human agent to the reservation server multiple times before service is completed at the agent and the customer exits the system. Just before exiting the system, there is a probability associated with whether or not a sale is made. The schematic representation of the flow of customers through this system is included below.

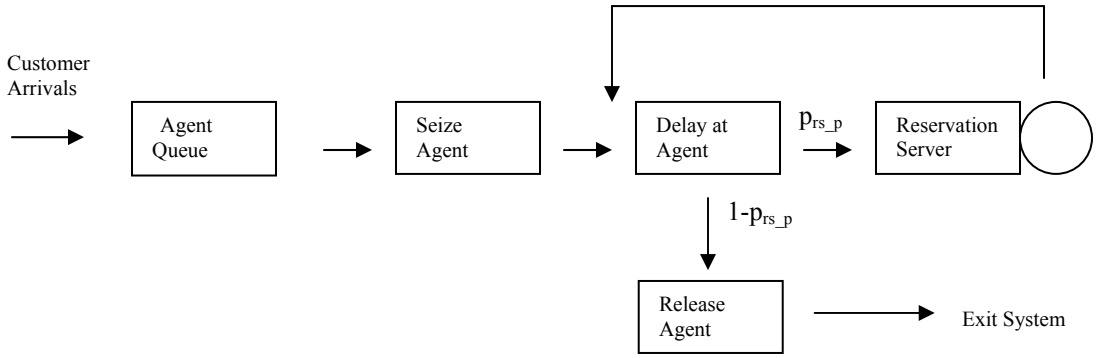

Figure 3 Schematic Diagram of Telephone System

This system is also modeled as an open queuing network of 2 queues. Inter-arrival times into the system are assumed to be iid exponential random variables with a mean of $1/\lambda_p$. In this system, reneging can occur only at the human agent. It is assumed that the time a customer waits before reneging at the human agent queue is an iid exponential random variable with a mean of $1/\theta_p$. All service times are assumed to be iid exponential random variables with mean $1/\mu_p$ and $1/\mu_{rs}$ for the human agent and the reservation server respectively. The number of human agents is denoted by s. Therefore, in order to determine, $1/\mu_{p'}$, the overall service time at the agent, an approximation is developed to compute the total waiting and service time at the reservation server, W by modeling the reservation server using an M/M/s queueing system. We observed that approximation works very well by comparing against the simulation. The details of the approximation and the figure similar to Figure 2 will appear in a forthcoming paper.

2.3 Reservation System with Internet Telephony Component

This system can be thought of as a combination of the two systems mentioned in Sections 2.1 and 2.2, with all customers sharing one reservation server. In this system, customers initially enter at the web server. They follow they same flow as described in Section 2.1, with one exception. Instead of eventually exiting the system, the customer will deflect to the human agent. At the human agent, the customer will proceed as described in Section 2.2 and can only exit the system through the phone portion. The schematic representation of customers flowing through this system is represented in Figure 4. The solid arrows represent customers who are solely using the web and the dashed arrows represent customers who are using the Internet telephony component and have deflected from the web portion to the human agent portion of the system. The probabilities whose values have remained the same as the respective portions described above have the same names, while the values that have changed are denoted by new names. It is assumed that $p_{ITws} > p_{ws}$ and $p_{rs\_IT} < p_{rs\_w}$.

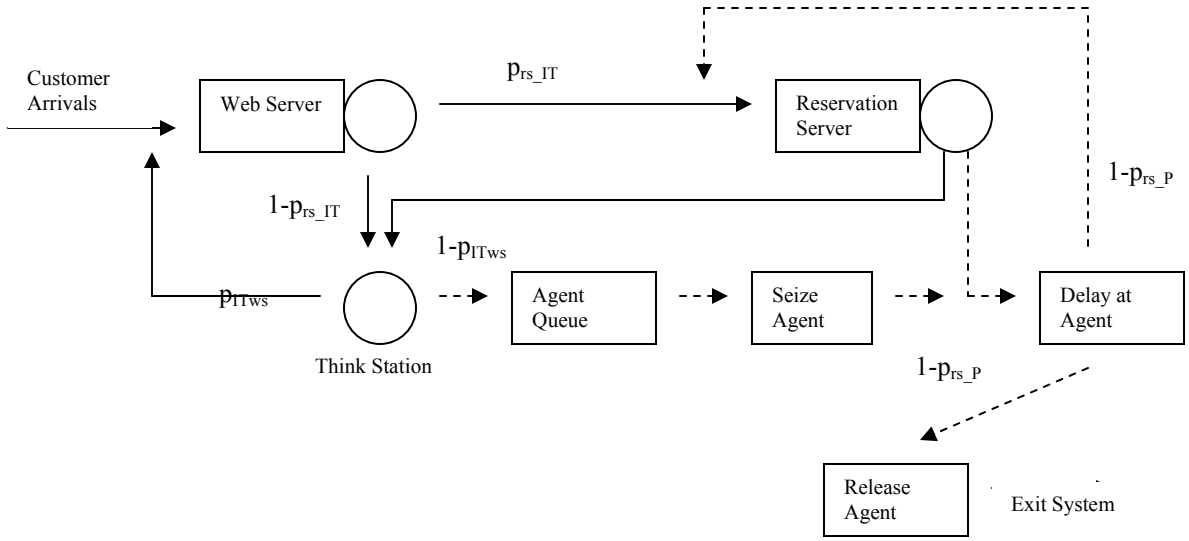

Figure 4 Schematic Diagram of Internet Telephony System

The initial inter-arrival times to the web portion are assumed to be iid exponential random variables with a mean of $1/\lambda_{ITw}$. Service times are assumed to be iid exponential random variables with mean $1/\mu_{ws}$, $1/\mu_{ITts}$, and $1/\mu_{rs}$ for the web server, the think station, and the reservation server respectively. It is assumed that $1/\mu_{ITts} < 1/\mu_{ts}$. This assumption is due to the fact that customers that do not have an Internet telephony option available will spend more time thinking and browsing than customers who have that option. The time a customer waits before reneging at the web server is assumed to be iid exponential random variable with a mean of $1/\theta_w$. Interarrival times to the phone portion are assumed to be iid exponential random variables with a mean of $1/\lambda_{ITp}$. The arrival rate to this portion, λ_{ITp}, is equal to the initial arrival rate to the web portion minus the renege rate at the web portion. The service times in the phone portion are assumed to be iid exponential random variables with mean $1/\mu_{ITp}$ for the human agent. The overall service time at the human agent can be denoted as $1/\mu_{ITp'}$. It is assumed that the time a customer waits before reneging at the human agent is an iid exponential random variable with a mean of $1/\theta_{ITp}$. In this system it is assumed that $1/\mu_{Itp} < 1/\mu_p$ and $1/\theta_{Itp} > 1/\theta_p$. These assumptions exist because customers who have Internet telephony will require less processing time at the human agent since they have already entered in some preliminary information through the computer. These customers will also be willing to wait longer before reneging because of the time they have already invested in entering their information. The number of human agents is denoted by s. The analysis of this system is completed by partitioning the entire unit into a web portion and a traditional telephone portion (by making suitable substitutions). The analytical model with its approximations matched the simulations.

3 Analysis of Combined Systems

In this section, the airline reservation system will be presented and analyzed as two, combined systems: 1) the current W/P system: the entirely web-based system combined with the traditional telephone system and 2) the proposed W/P/IT system: the W/P system with an added Internet telephony component.

3.1 W/P System

The architecture of this system is very similar to the individual units seen in Figures 1, 3 and 4. Both the web and the phone customer types will still follow the same flow described in Figures 1 and 3 from Section 2, except now they share a common reservation server. The modeling assumptions and parameters discussed in the previous section hold for this system because the analysis will be done by partitioning the combined system into their individual units. In this analysis, instead of using the same arrival rate λ in all cases, each customer type will have a different λ value. The attempt is to keep the total λ value going into both portions the same, but have different fractions of this total representing the arrival rate to each portion. The analysis of this system is almost identical to the analysis of the individual units discussed in Section 2. The only change is in finding the value of W (waiting time) across the reservation server. Since a common reservation server is used to process both customer types, the arrival rate to calculate W has to account for the two different types of arrivals. After making this adjustment, the expected revenue of each system is computed using the renege rates and the total expected revenue is obtained by summing the individual values.

3.2 W/P/IT System

This system's architecture is the same as the W/P System discussed in Section 3.1. However, now the Internet telephony customer type must be incorporated into the overall flow. The Internet telephony customer's path flow is the same as described Figure 4, with the web, phone, and Internet telephony customers all sharing one reservation server. The modeling assumptions and parameters discussed in Section 2 also hold for this system because the analysis will be done by partitioning the combined system into their individual units. Approximations are developed to determine some of the overall processing and renege times used in the analysis. The analysis of this system can be performed using the previous approximations from Section 2 along with some additional approximations of the input parameters. These additional approximations require scaling some of the parameters by a ratio of λ to account for the two different customer types that will now be flowing through each portion. The total expected revenue for the W/P/IT system is obtained by adding the individual expected revenues calculated for the web and the phone portions.

3.3 Results

The analytical approximations based on queueing networks used to calculate the total revenue of the W/P/IT system are validated using simulation as shown in Figure 5. As indicated by the figure, the error between the simulated and the approximated results is less than three percent in all cases. Figure 6 contains a graphical representation of the total expected revenue of the two systems using different fractions of the web and phone customers from the W/P system to obtain the Internet Telephony customer arrival rate in the W/P/IT system. The figure indicates that the latter system results in improved performance and greater revenue. However, there is a tradeoff between improving the performance of the system by adding an Internet telephony component and increasing the arrival load to the system. Internet telephony customers from both the web portion and the phone portion are now accessing the reservation server, along with the traditional web and telephone customers. The traffic from all directions increases the wait at the reservation server, which in turn increases the wait at the web server and the human agents. The renege rates at both the web server and the human agent will then significantly increase as a result of longer waiting times. As more customers renege, there is less chance of potentially completing the sale of an airline ticket.

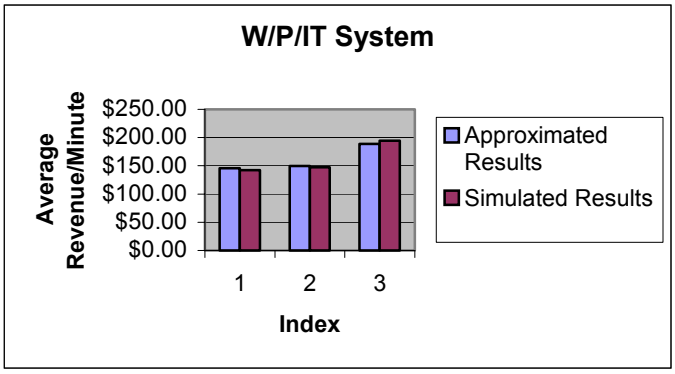

Figure 5 Comparison of Approximated Results vs. Simulated Results for the W/P/IT System

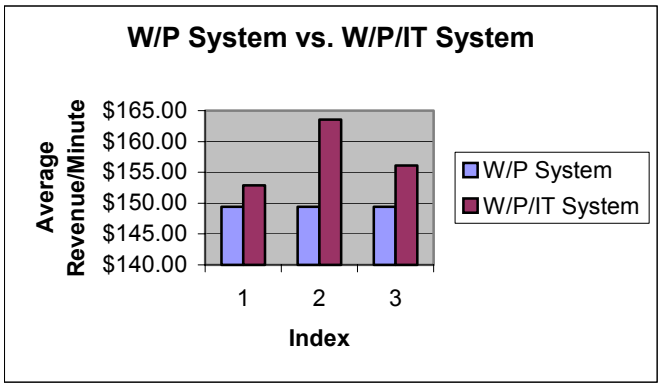

Figure 6 Comparison of W/P System vs. W/P/IT System

4 Conclusions and Extensions

In this research we showed that implementing an Internet telephony component to an airline reservation system enhances the performance of the system. However, a tradeoff exists between the benefits of better customer service provided by Internet telephony and an increased traffic load to the system. In order to maximize the benefits of this feature, capacity-planning issues must be addressed (which will be done in a forthcoming paper) to determine the adjustments that should be made to the system to account for the increase in traffic.

Although this research was based on a case study of an airline reservation system, the potential impact of this study is for all e-businesses to consider Internet telephony as a viable option to improve customer service. These businesses are losing substantial amounts of revenue because of the lack of real-time customer service and support. Gaining the confidence of customers through Internet telephony and easy access to customer service representatives can significantly reduce this loss. It must be noted that the benefits in this research were shown for chosen values of the input parameters (indices). An organization that is interested in evaluating the effect that adding Internet telephony will have on their business should go through the tools presented in this paper with their set of parameters.

The queuing models used in this performance study assumed exponential times and infinite buffer capacities. Further research will consider different distribution of times and finite buffer capacities. The reservation server queue served as the bottleneck of this system as traffic from many different customer types utilized a common reservation server. Thus, it might be helpful to prevent lengthy delays by limiting the size of the reservation server queue. Another direction of future research is to evaluate different priority service disciplines at each queue based on customer type.

Acknowledgements

The authors are grateful to the reviewers for their insightful and valuable suggestions towards improving the quality of this paper.

References

1. Thomsen, G., and Jani, Y., 2000, "Internet Telephony: Going Like Crazy," *IEEE Spectrum*, 37, 52-58.
2. Alvord, C., 2000, "Advance Planning and Optimization", <http://www.ie.psu.edu/mts2000/TRW_APO.htm> (9 June, 2000)
3. Prabhu, N.U., 1997, "Foundations of Queuing Theory", Kluwer Academic, Boston.
4. Kulkarni, V. G., 1995, "Modeling and Analysis of Stochastic Systems", Chapman & Hall, New York .
5. Bolch, G., Greiner, S., de Meer, H., and Trivedi, K.S., 1998, "Queuing Networks and Markov Chains", John Wiley & Sons, New York.
6. Arena, 2000, <http://www.systemsmodeling.com>
7. MATLAB, 2000, <http://www.mathworks.com >