SIMULATING THE EXPEDITIONARY COMBAT SUPPORT SYSTEM HELP DESK[1,2]

Michael E. Chua, Department of Operational Sciences, Air Force Institute of Technology, 2950 Hobson Way, Wright-Patterson AFB, OH 45433, 937-255-3636, michael.chua@eglin.af.mil
Jeffrey A. Ogden , Department of Operational Sciences, Air Force Institute of Technology, 2950 Hobson Way, Wright-Patterson AFB, OH 45433, 937-255-3636, jeffrey.ogden@afit.edu
Alan W. Johnson[3], Department of Operational Sciences, Air Force Institute of Technology, 2950 Hobson Way, Wright-Patterson AFB, OH 45433, 937-255-3636, alan.johnson@afit.edu

ABSTRACT

The Expeditionary Combat Support System (ECSS) is the Air Force's planned Enterprise Resource Planning (ERP) solution. Training and user support are critical components to successful ERP implementation and sustainment. Help desks are classic queuing theory problems, and the Erlang-C model is among the favorite models to use. Unfortunately, the Erlang-C model has a reputation to overestimate staffing requirements. This multi-method study applies the Erlang-A and Erlang-C queuing models analytically and through simulation to account for customer queue abandonment dynamics. Our results provide the staffing levels for the level 1 and level 2 help desks that will yield an efficient balance of customer wait time, call center agent utilization, and abandonment rate.

INTRODUCTION

The Expeditionary Combat Support System (ECSS) is an Oracle-based Enterprise Resource Planning (ERP) system that will replace 420 legacy systems and standardize the United States Air Force's logistics process from end to end to facilitate the flow of information and provide opportunities for savings and improved efficiency. While ECSS is projected to eventually have over 250,000 end users, it is projected to be initially released to 40,000 users.

The ECSS help desk will provide support for 40,000 projected users through three tiers of support. The Field Assistance Service (FAS) located at Gunter Annex, Maxwell AFB in Alabama will provide Level 1 support. Level 1 support consists of basic assistance. If the Level 1 analyst cannot close out the trouble ticket, the call will be sent up to the Level 2 help desk. Level 2 and 3 support involves issues that are functional or technical in nature. The ECSS Program Management Office (PMO) will be responsible for technical support, and the Air Force Logistics Management Agency will be responsible for functional support.

The ECSS PMO is currently analyzing the potential workload for the help desk. ECSS is a large-scale ERP implementation with very few comparable-scale established ERP systems. The scale of the implementation limits the amount of data available to conduct a top-down analysis to determine the

[1] This research is sponsored by the Air Force Logistics Transformation Office.

[2] The views expressed are those of the authors, and do not reflect the official policy or positions of the United States Air Force, Department of Defense, or the U.S. Government.

[3] Corresponding Author.

potential help desk workload. This limitation provides opportunities to conduct a bottom-up analysis of the help desk to forecast workload.

This study evaluated the ECSS help desk through a bottom-up analysis using simulation and queuing models. Simulation is the primary analysis tool due to the limited data available. We explored the following research questions: 1) What is the most probable call volume for ECSS? 2) What are the probable staffing levels to match the projected call volume? 3) What are the optimal trade-offs between service quality and cost savings?

BACKGROUND

Implementation of ERP systems has received a lot of attention, but the maintenance, sustainment, and support of ERP systems have not been as widely studied. Annual maintenance costs are approximately 25% of initial implementation costs. This has caused some to question why this subject has not received greater attention [1]. With annual maintenance and support cost averaging 25% of initial implementation and assuming a potential useful life of ERP systems at around 10 years, it is clear that a company will expend more resources on maintenance and support than on the acquisition and implementation of the ERP.

Wenrich [2] published a case study on an anonymous large company with a decade of experience and two major upgrades. In her article, Wenrich emphasizes the importance of user support. "Organizations sustaining ERP implementations can expect a larger part of their maintenance work from user support and investigation requests than a team supporting an in-house developed software package" [2]. Wenrich made the point that repeated system training counters the higher demand for user support, but this solution could be a less cost effective alternative for the Air Force when compared to establishing a help desk to supplement training. Sui Pui Ng et al. also emphasized the importance of support and specifically the help desk by incorporating the help desk in their preliminary ERP Maintenance Model [1].

Iba's research [3] emphasizes the importance of support, and breaks out the support into two categories. The first category is the initial stage where the system goes live and support teams assist users with utilizing the new system. The second category is the remaining life cycle support post implementation. Iba and other authors, like Wenrich, have discussed the support decision whether to provide support in-house or outsource support to leverage expertise [3].

Although help desk support is not the main driver of sustainment costs for most ERP systems, decision makers sometimes overlook and underestimate the direct and indirect impacts of a help desk. The direct impacts are costs from staffing, facilities, and software/hardware requirements. Failure to have an adequate formal support structure can lead to unintended consequences such as informal methods of learning and lessen the initial user buy-in [4].

Queuing theory models can be useful for studying call centers or help desks. This study analyzed three models: an M/M/S (Erlang-C) analytic model with fixed arrivals, an M/M/S (Erlang-C) simulation with time dependent arrivals, and an M/M/S + M (Erlang-A) simulation with time dependent arrivals that accounts for abandonment. The following section discusses our methodology in detail.

METHODOLOGY

We use data from legacy and analogous systems to simulate the performance metrics for the ECSS help desk. We account for utilization rates, abandonment, and varying traffic loads, and compare results from Erlang-C equations and the Erlang-C simulation with the Erlang-A simulation. Figure 1 presents our analysis approach.

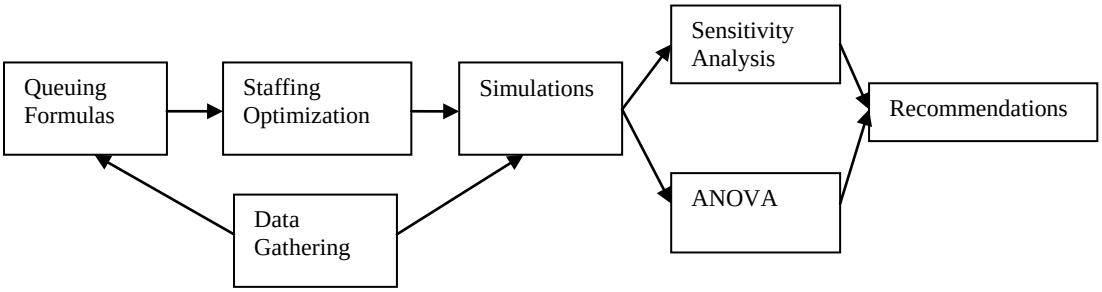

Figure 1: Methodology Summary

Help desks are classic queuing theory problems. One of the most familiar queuing models is the M/M/S model, also known as the Erlang-C model [5]. The Erlang-C model has an exponential arrival rate, exponential service rate, and *S* servers. With these assumptions, one can calculate certain characteristics like the average number of customers in the system, average time a customer spends in the system, and the utilization factor for the system. One of the many benefits of the Erlang-C model is the ease of calculations with the application of Little's Law, which states that the average number of items in a queuing system equals the average rate at which items arrive multiplied by the average time that an item spends in the system [6]. The downsides of the Erlang-C model include the assumptions made of constant exponential inter-arrival and service times and ignoring call center dynamics, such as single-type, single-skill, constant staffing, abandonment, retrials, and time-varying conditions [7]. We address these dynamics in this research.

Busy signals and abandonment add a different dimension to queuing models, and so researchers created the M/M/N/B + G and the M/M/N/B + M models to capture the trade-offs between busy-signals, delays, and abandonment. Note that B stands for the number of lines, +G stands for patience with an assigned distribution, and the final +M stands for patience with an exponential distribution. Garnett et al. discuss the M/M/N + M model, also known as the Erlang-A, in [8]. In the Erlang-A model, the A stands for abandonment. Garnett et al. argue that "The immediate effect of an abandonment is less delay for those further back in line, as well as for future arrivals … using workforce management tools that ignore abandonment would result in over-staffing as actually fewer agents are needed in order to meet most abandonment-ignorant service goals" [8]. In call center forecasting, a common safety staffing heuristic is the square root rule:

$$N = R + \beta\sqrt{R}, \tag{1}$$

where N is the number of servers, R is represents the offered load defined as the customer arrival rate divided by service rate, and β is the desired level of service. β is a safety factor in equation 1 since a system with a steady and predictable arrival rate would only need the minimum number of agents represented by R.

Customers abandon their place in the queue for many reasons, such as a preference to reenter the system at a more convenient time or the lack of patience. The question that researchers must ask when using these complex queuing models is whether their use is justified. Garnett et al. [8] show their significance

in high volume call centers, but neglect low volume help desks. The additional dimensions of abandonments, delays, and re-trials could be influential, which is why we chose simulation for our main analysis tool.

A common assumption is that call arrivals follow a Poisson process with different rates at specific blocks of time, which is empirically supported by [9]. Exponential service times are also frequently assumed. Brown et al. found that a lognormal distribution may provide a better fit of the call arrival data [9]. The challenge with using Brown et al.'s suggestion of lognormal distributions in our study is that standard deviation data were not available for the analogous and legacy systems.

The foundation of this study is the simulation of the queuing models. The simulations produced a set of data, and regression analysis was used to determine the relationships of the independent variables of number of agents, arrival rate, service rate, and customer patience with the dependent variables. The dependent variables for this study were average waiting times, agent utilization rates, and the ratio of customers that renege.

Simulation

Simulation is frequently used for call center analysis. Miller et al. [10] showed the benefits of simulation in a business case projecting return on investment. In their study, the company implemented a new call routing technology across 25 call centers. The investment costs were $17 million and operating costs were $8 million. The goal was to determine whether the cost savings would justify the investment. Their study concluded that the cost savings with the new technology was dependent on call volume, so the company adopted a strategy that incorporated two routing technologies. Saltzman and Mehrotra model two priorities for arriving customers and call abandonment [11]. They assume that the service time is exponential and abandonment is a linear function of wait time where the likelihood of abandonment increased by 3.5% for each additional two minutes of wait time, with a maximum abandonment percentage set at 40%. Mehrotra and Fama [12] discuss the reasons why call centers turn to simulations, major ways call centers use simulation, and typical inputs modeled in call center simulations. Mehrotra and Fama confirmed the challenges with actual call duration data leading to the assumption of exponential service times. They also discuss key service parameters in the call center such as agent skills, schedules, and shrinkage factors on utilization rates. For this study, agent skill was modeled since the agents from the ECSS level I and II help desk will have different service rates. Shrinkage is the factor applied to capture unscheduled lost agent time. This phenomenon is linked to scenarios like unexpected absences, longer than expected breaks, or late arrivals. Mehrotra and Fama apply a shrinkage factor of 10% [12].

We selected Rockwell's ARENA 12.0 to simulate two call center operations models: M/M/S (Erlang-C) with time dependent arrivals, and M/M/S + M (Erlang-A) with time dependent arrivals.

Equation 1 determines the appropriate call center staffing requirements for the analytic queuing models. For the square-root-rule, as β approaches zero the number of agents required becomes the ratio of the arrival and service rate. Increasing β represents an increased staffing safety factor for variability. The challenge was to determine the appropriate β while staying within the constraints of budget and resources. We chose β values of 0.7, 0.4, and 0.1 to test staffing requirements for each time interval.

The time-dependent arrival rates and shift schedule constraints makes it difficult to determine the appropriate call center staffing levels. Equation 1 is used to calculate the required number of agents for

each hour, but a call center servicing a non-uniform or multimodal call arrivals distribution would not operate efficiently due to the inability to service peak and low arrivals. To solve this problem for the simulations, we used an integer linear program to determine the number of agents to assign to each shift. We modeled six shifts: the first three shifts covered a 24-hour period in eight-hour increments. To supplement the three core shifts and enable flexibility for peaks and lows, we added three booster shifts for peak periods. The optimization minimizes the number of assigned agents, subject to preserving the required number of agents per shift.

Experiment Design

We conducted an initial pilot study examining the help desk scenarios listed in Table 1.

Table 1: Design of Experiments: Pilot Study

Staffing Regimes	Agents	Total Daily Calls			Service Rate (Minutes)			Renege Time (Minutes)		
		Low	Mid	High	Low	Mid	High	Low	Mid	High
Light Arrival, Low Service Rate	12	100	210	565	4	7.5	18.55	1.58	4	8
Light Arrival, Mid Service Rate	6	100	210	565	4	7.5	18.55	1.58	4	8
Light Arrival, High Service Rate	4	100	210	565	4	7.5	18.55	1.58	4	8
Moderate Arrival, Low Service Rate	31	210	618	1775	4	7.5	18.55	1.58	4	8
Moderate Arrival, Mid Service Rate	14	210	618	1775	4	7.5	18.55	1.58	4	8
Moderate Arrival, High Service Rate	8	210	618	1775	4	7.5	18.55	1.58	4	8
High Arrival, Low Service Rate	101	1775	2104	2893	4	7.5	18.55	1.58	4	8
High Arrival, Mid Service Rate	43	1775	2104	2893	4	7.5	18.55	1.58	4	8
Heavy Arrival, High Service Rate	23	1775	2104	2893	4	7.5	18.55	1.58	4	8

The pilot experiment yields 243 unique scenarios, and each scenario ran for 183 days with three repetitions to get a general sense of the model's performance at reasonable computational cost. The 243 staffing level, call volume, and service rate combinations resulted in a wide range of performance outcomes. This pilot study removed the staffing levels that provided unacceptable metrics for service time, utilization, and renege percentages. We then ran a second pilot simulation on the remaining staffing regimes for projected Level 1 and 2 help desk estimates, using two-year runtimes. Table 2 shows the low, middle, and high values used as inputs for the Level 1 and 2 simulations. We used longer average service rates for Level 2 support since the issues are generally more complex than Level 1 issues.

Table 2: Level 1 and 2 Simulations

Help Desk Level	Total Daily Calls			Service Rate (Minutes)	Renege Time (Minutes)
	Low	Mid	High		
Level 1	210	618	1775	7.5	4
Level 2	84	247	710	17.74	4

We analyzed the results from the second pilot simulation to show the tradeoffs between staffing levels and help desk performance. The simulations conclude with an analysis of variance and sensitivity analysis of the selected staffing levels for projected call volumes. These final simulations used 6.5-year runtimes with 20 replications per scenario.

RESULTS AND CONCLUSIONS

The ECSS help desk may experience call volumes ranging from 210 calls per day to 1,775 calls per day. We found that 31 Level 1 and 2 agents are sufficient to handle the higher projected demands for user

support when ECSS is implemented. As users become more familiar with the system and modifications have stabilized, about 12 Level 1 and 2 agents is sufficient to meet the steady state demand (we do not consider any potential effects from staffing turn-over). Further along the life cycle, ECSS can reach a long-run state where the number of calls per user will decrease. At the long-run state, 8 Level 1 and 2 agents are sufficient. Adding agents to the estimate will yield slightly better performance but at a diminishing rate. We recommend that Air Force leaders establish help desk performance goals for Levels 1 and 2 and apply the methodology used in this research to determine the staffing levels that best match their goals.

There is a positive relationship with adding agents and improving call center performance. A diminishing marginal benefit exists as managers add more agents to the system. The greatest gains in call center metrics occur when agents are added to an understaffed system. Another trade-off to consider is whether to establish the help desk for initial, steady state, or long-run states. If the ECSS PMO staffs the help desk with 8 Level 1 and 8 Level 2 agents without a contingent for the probable high call volume scenarios, the help desk will fail to provide adequate service to the ECSS users.

A fixed long-run arrival rate at implementation is unlikely, so a strategy must be in place to handle peak usage of the ECSS help desk during operations, major exercises, and major releases. Maintaining a facility to support these peaks could be inefficient since these agents would only be used during these unique situations. In this case, temporary on-site assistance and/or training might be the better alternative.

REFERENCES

[1] Sui Pui Ng, C. G., Gable, G. & Chan, T. An ERP Maintenance Model. *Hawaii International Conference on System Sciences,* 2003, 8, 234b.

[2] Wenrich, K. I. *Lessons Learned During a Decade of ERP Experience: A Case Study.* Hershey PA: IGI Publishing, 2009.

[3] Iba, B. *Towards a Whole Lifecycle Cost Model for ERP Projects.* Bedfordshire UK: Cranfield Univeristy, 2006.

[4] Boudreau, M.-C. Learning to Use ERP Technology: a Causal Model. Hawaii International Conference on System Sciences, 2003, 8, 235B.

[5] Cooper, R. B. Queuing Theory. Encyclopedia of Computer Science, 4th ed, 2003, 1496-1498.

[6] Little, J. D. Building Intuition: Insights From Basic Operations Management Models and Principles. Massachusetts Institute of Technology: Springer Science, 2008.

[7] Gans, N.G., Koole, G. & Mandelbaum, A. Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management,* 2003, 5, 79-141.

[8] Garnett, O.A., Mandelbaum, A. & Reiman, M. Designing a Call Center With Impatient Customers. *Manufacturing & Service Operations Management*, 2002, 4 (3), 208-227.

[9] Brown, L. N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. & Zhao, L.H. Statistical Analysis of a Telephone Call Center: A Queuing-Science Perspective. *Journal of the American Statistical Association*, 2005, 100, 36-50.

[10] Miller, K.V. & Bapat, V. Case Study: Simulation of The Call Center Invironment for Comparing Competing Call Routing Technologies for Business Case ROI Projection. *Proceedings of the 1999 Winter Simulation Conference*, 1999.

[11] Saltzman, R.M. & Mehrotra, V. A Call Center Uses Simulation to Drive Strategic Change. *Interfaces,* 2001, 31(3), 87-101.

[12] Mehrotra, V. J. & Fama, J. Call Center Simulation Modeling: Methods, Challenges, and Opportunities. Proceedings of the 2003 Winter Simulation Conference, 2003, 135-143.