Finding Optimal Resources for IT Services

Sumit Raut, Muralidharan Somasundaram

Finding Optimal Resources for IT Services

Sumit Raut[1*] and Muralidharan Somasundaram[2]

[1]Associate Consultant, TCSL, Kolkata, India
sumit.raut@tcs.com
[2]Principal Consultant, TCSL, Chennai, India
muralidharan.somasundaram@tcs.com

Abstract. This paper studies the resource management problem in IT services, where service request arrives to resource management group (RMG) and RMG needs to allocate resources to a request for a service based on availability of resources and service requirement start date. We propose an approach to find optimal number of resources in the context of Poisson arrivals, and service times & lead-times are exponentially distributed. We provide an exact mathematical queuing model for optimal number of skilled resources needed based on waiting time cost, idle time cost and revenue from a customer service for a guaranteed service level agreement. We tested the model with real life data for various types of requirements and analytical results show the benefits of the proposed approach.

Keywords: IT services, Queuing, resource allocation, delay, waiting time.

1 Introduction

IT services business is a knowledge intensive industry where highly skilled resources are needed to meet demand for day-to-day services. Typically customers arrive with their requests to resource management group (RMG) for their service for a particular booking date and RMG needs to provide the skilled resources for the given booking date. The customer request needs to go through two stages of a process: (i) keep the request in RMG server till booking date and (ii) RMG allocates skilled resources to complete the request. Early allocation of skilled resources (i.e., before booking date) to request may lose opportunity to allocate resources of other possible future requests, i.e., paying wages while there is no realization of revenue and deterioration in service level. Delay in allocation of skilled resources indicates loss of demand resulting in revenue loss of service provider. Hence, it is important to estimate the right number of resources and active bench (safety stock) in dynamic IT services environment. The problem complexity increases when there is also possibility of substitution and multi-skilling. This paper studies the resource allocation (or optimal bench) problem and applied queuing theory approach to determine the optimal number of resources which

* Room: 4B-14, TCSL, BIPL, Sector-V, Salt lake, Kolkata - 700093

would also help to decide the active bench and number resources need to be trained for a particular skill. In past, several authors discussed the aspects of determining of optimal number of servers in a different stream of research. Some of the relevant literatures are described below:

Grassmann [6] states that: "In respect to real life, all models are approximations." He argues that it may not be practical to apply exact mathematical queuing models to real problems, and offers a number of practical methods for finding the number of servers in waiting line systems. Bassamboo et al. [4] offer a dynamic model for determining the number of agents and the assignment of arrivals to agents in large call centers using an asymptotic approximation approach, and Borst et al. [5] consider an M/M/s system with a large number of servers (s), and propose an asymptotic approximation for finding the optimum staffing levels. Radmilovic et al. [10] provide a method to determine optimum number of servers and optimum server capacity with considerations of bulk arrival rate and estimate the impact in cost with changes of number of servers, group size and bulk arrival rate. Fazel et al. [7] presents an alternative approximate method for analyzing large M/M/s systems, which is highly accurate, and considers both the cost of customer waiting and the staffing cost to develop a model that determines the optimum number of servers that would minimize the total cost. Yang et al. [12] studies the optimal resource allocation in time-reservation systems where customers arrive at a service facility and receive service in two steps; in the first step information is gathered from the customer, which is then sent to a pool of computing resources, and in the second step the information is processed after which the customer leaves the system. They used dynamic programming approach to determine near-optimal number of processors follows a step function with as extreme policy the bang-bang control. They also provide new fundamental insights in the dependence of the near-optimal policy on the distribution of the information gathering times.

All the above works consider the system where jobs are serviced in First-Come-First-Serve (FCFS) order, and how the optimal number of servers varies across workloads. In addition, these queuing models does not consider real-life constraints (except Yang et al. [12]) and impractical to use them for large service centers or industry.

There are few researchers who have been used queuing models successfully for a variety of real-life problems. Some of the real-life solutions examples are for a variety of organizations including police patrol scheduling (Kolesar el al. [8]), staffing of supermarkets (Ittig [9]), tele-retailing (Andrew et al. [2]), and call centers (Artalejo et al. [3]). These models are commonly used by service organizations to determine the number of servers through trial and error. The available queuing models, however, become exceedingly complex and cumbersome for analyzing large service centers, and the trial and error approach becomes impractical. Therefore, finding the optimum number of resources in a large service industry to provide optimal level of service for different customers at different point of time needs special attention in research.

The main difference between the existing literature and our work is that we aim to (1) optimize the resource allocation cost, idling cost and revenue from servicing customers through effectively releasing resources and estimating right number of resources and (2) satisfy SLA (Service Level Agreement) constraint on the delay of a

job, where delay is not same as waiting time. This is in contrast to the aforementioned works which have not focused on resource release time while obtaining number of skilled resources needed for a guaranteed SLA guarantee.

We organized the papers as follows: In Section 2, we describe the model formulation. Section 3 describes the simulation study and results. Conclusions and future directions of research are provided in Section 4.

2 Model Formulation

The resource allocation of a service request in IT industry has done through two stages (refer Fig. 1): (i) In the first stage, RMG receives requests and allocates the resource to service as per starting date of service; (ii) In the second stage, resources perform the service as per request/competency.

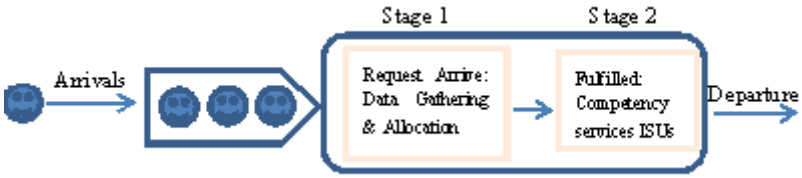

Fig. 1. Resource Allocation Process

Parameters Definition and assumptions:

- Request arrival (X) follows Poisson distribution with mean λ (per day)
- Service time (T) of a request (X) follows Exp. distribution with mean $1/\mu$ (in day)
- Request arrival date and request start date to process are A_x and B_x, respectively
- Lead time represents as $L = B_x - A_x$, and follows Exp. distribution with mean $1/\mu_1$
- Number of server is S.

The first stage of resource allocation process of RMG is represented by the queuing system where request is arriving with Poisson process (X) in infinite number of servers (since, RMG only need to store the information in this stage) with infinite capacity systems and request leaves the system once it reaches at start-date of service. The time between requests follows exponential distribution (Y) with mean $1/\lambda$. The service time / stay time of a request in first stage is followed exponential distribution with mean as average lead time ($1/\mu_1$). The requests' depart from the first stage would be allocated to second stage of RMG process, i.e., the arrival process of request to second stage is same as departure process of first stage. Therefore, the departure process is D = Y + L and the distribution is as follows:

$$f(D = d) = \int_0^d f(L = d - y) * f(Y = y)dy = \int_0^d \mu_1 * e^{-\mu_1(d-y)} * \lambda * e^{-\lambda y}dy = \frac{\lambda * \mu_1}{(\lambda - \mu_1)}\left[e^{-\mu_1 d} - e^{-\lambda d}\right] \quad (1)$$

The expected value of departure process:

$$E(D) = \frac{1}{\mu_2} = \frac{\lambda+\mu_1}{\lambda\mu_1} \tag{2}$$

The second stage of resource allocation process of RMG is represented by the queuing system where request arrival process as request departure process from first stage, i.e., follows hypo-exponential distribution and service time follows exponential distribution with mean 1/μ. Based on the queuing model, we approximate the system in G/M/1 queuing model (this can be expanded for G/M/S queuing model and for different customer priority (Adan et al. [1])). The delay time distribution of this model is given below (Adan et al. [1], White et al. [11]):

$$P(W \leq t) = 1 - \sigma * e^{-\mu(1-\sigma)t} \tag{3}$$

Where $\sigma = \widetilde{D}(\mu - \sigma\mu)$, $\widetilde{D}$ is a moment of request arrival process in second stage. Moment of request arrival process in second stage of RMG process is as follows:

$$\widetilde{D}(s) = \left(\frac{\mu_1}{\mu_1+s}\right) * \left(\frac{\lambda}{\lambda+s}\right) \tag{4}$$

Now, the equation (4) is represented as

$$\sigma = \left(\frac{\mu_1}{\mu_1+\mu-\sigma\mu}\right) * \left(\frac{\lambda}{\lambda+\mu-\sigma\mu}\right) \tag{5}$$

The expected delay time of the system is given below:

$$E(W) = \int_0^\infty t * \mu\,\sigma(1-\sigma) * e^{-\mu(1-\sigma)t} = \frac{\sigma}{\mu(1-\sigma)} \tag{6}$$

The distribution of conditional delay time W|W>0 is as follows:

$$P(W > t|W > 0) = \frac{P(W>t)}{P(W>0)} = \frac{\sigma * e^{-\mu(1-\sigma)t}}{\sigma} = e^{-\mu(1-\sigma)t} \tag{7}$$

The expected value of conditional delay time W|W>0 can be obtained as:

$$E(W|W > 0) = \int_0^\infty t * \mu\,(1-\sigma) * e^{-\mu(1-\sigma)t} = \frac{1}{\mu(1-\sigma)} \tag{8}$$

From the equation (5) and (8), the unique root σ is calculated as follows:

$$\sigma = \left(\frac{\mu_1}{\mu_1+\left(\frac{1}{E(W|W>0)}\right)}\right) * \left(\frac{\lambda}{\lambda+\left(\frac{1}{E(W|W>0)}\right)}\right) \tag{9}$$

Functional relationship of service rate (μ) with unique root (σ):
From the equation (5), we can write the equation as follows:

$$\sigma = \left\{\frac{1}{1+(1-\sigma)*\frac{\mu}{\mu_1}}\right\} * \left\{\frac{1}{1+(1-\sigma)*\frac{\mu}{\lambda}}\right\} \tag{10}$$

Suppose, $\sigma = 1 - k$, $A = \frac{\mu}{\mu_1}$ and $B = \frac{\mu}{\lambda}$

Then, the equation (10) can be written as:

$$(1 - k)(1 + kA)(1 + kB) = 1$$

$$k = \frac{-(A+B-AB) \pm \sqrt{(A+B+AB)^2 - 4AB}}{2AB} \tag{11}$$

Since at weekly level service rate (μ) $\leq$ 1, average lead-time ($1/\mu_1$) $\geq$1 and the actual arrival rate (λ) $\geq$1for IT service industry (in most project cases), then, $B \leq 1$ and the value of ($A + B$) is always greater than or equal to AB. Therefore, k can be written as

$$k = \frac{-(A+B-AB) + \sqrt{(A+B+AB)^2 - 4AB}}{2AB} \tag{12}$$

Now, the relation between service rate μ and unique root σ is as follow:

$$(1 - \sigma)\mu = \frac{-\left(\frac{\lambda}{\mu_1} + \frac{\mu_1}{\lambda} - \mu\right) + \sqrt{\left(\frac{\lambda}{\mu_1} + \frac{\mu_1}{\lambda} + \mu\right)^2 - 4\mu_1\lambda}}{2} \tag{13}$$

2.1 Determination of Optimal Number of Server or Service Rate

Allocating too many resources results in decrease waiting time but comes at high allocation costs and allocating too few skilled resources leads to long waiting times. Therefore, the problem is formulated as follows:
Minimize the expected number of customer's delays and idle resources

$$W_{cost} * E(D) * \frac{\sigma}{\mu(1-\sigma)} + R_{cost} * \left(1 - \frac{E(D)}{\mu}\right) \tag{14}$$

Maximize revenue by service (Revenue per unit * Avg. customer depart from system)

$$Rev * E(D) * \frac{1}{E(S)} = Rev * E(D) * \mu * (1 - \sigma) \tag{15}$$

Where, W_{cost} is average delaying cost, R_{cost} is average resource idle cost per day (i.e., fixed cost per day (pay bill) incurs for each resource and this pay bill need to incur after finishing their work) and Rev is the average revenue for service per day.
For priority customers, IT service providers are bounded by obligation to provide the service must. Therefore, we consider delaying time should not exceed t time-length with confidence level 95% as constraint as follows:

$$P(W \leq t) = 1 - \sigma * e^{-\mu(1-\sigma)t} > 0.95 \tag{16}$$

The above optimization problem is solved using Gradient search method and produces local optimal solution or optimal number of resources.

3 Simulation Study

We consider the resource allocation problem for different competencies in IT service industry for simulation study. We capture six months data on different competency

requests arrival (request date, start date) and allocation/service (fulfilled date) details for our analysis. From the above data set, we derive lead time (start date – request date), delay (min (fulfilled date – start date, 0), and waiting time (min (fulfilled date – request date, 0). We derive the service rate from the equation (8 & 9) using arrival rate, delay and lead-time information.

We observed that there are more than 200 competency levels in the dataset. We also observed that the variability is too high in each competency level and most of the cases data is intermittent. Hence, we group the competency as per their similarities, where one competency/skill can be substituted to other. In this way, the total number of competency groups is turned out fifteen. Some of the competency groups are: SAP-ERP, ORACLE-ERP, Programming, etc. The arrival patterns of some of the competency groups are given in Fig. 2.

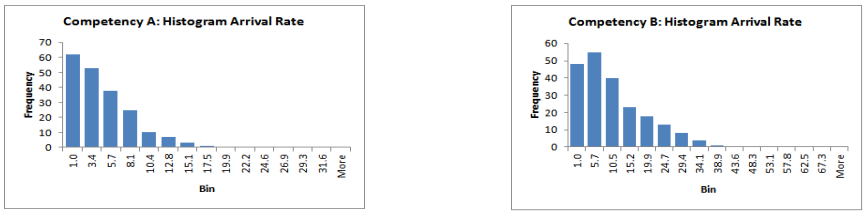

Fig. 2. Arrival patterns of different competency

The mapping of queuing model to resource management problem in IT-service is measured by error in the estimated delay with actual delay observed. W_{real} denotes the mean delay of serving of each competency group in IT-service process and W_{model} denotes the mean delay (refer equation 8) which derived from two-stage queuing model. The mapping error % ($MAPE_{comp}$) for each competency is represented as shown below:

$$MAPE_{comp} = \frac{|D_{real,comp} - D_{model,comp}|}{D_{real,comp}} * 100 \quad (17)$$

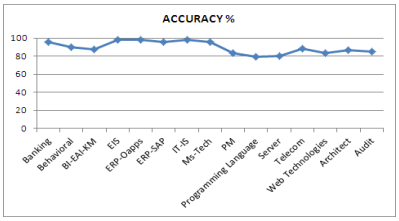

Fig. 3. Accuracy plot for queuing model accuracy

The mapping accuracy% (100 - $MAPE_{comp}$) is calculated using equation (17) and results are shown in Fig. 3. The above results showed that the estimation of delay using queuing model is highly accurate. On an average accuracy percentage is 90%. We also tested the waiting time distribution derived from the model with actual data

of each competency group. The results show that at 95% confidence level, the difference of waiting time distribution between model and actual data is not significant.

The above analysis suggests that the queuing modeling approach can accurately map the resource management problem in IT-service. Therefore, we used this approach to determine the optimal number of resources needed for IT-service industry based on different cost (holding cost of resources, delay cost) and revenue parameters (revenue for each service). This paper considers the cost ratio between idling/holding cost & delay cost is 1: 7 (ID ratio) and average revenue for service per day & idling cost is 1: 10 (RI Ratio). The total cost using equations (14 & 15) are derived for different number of resources (or service rate) for each competency group and the curve for total cost versus service rate (for ERP-Oapps competency) is shown in Fig. 4. The optimal number of resources for each competency group is determined using equations (14, 15 & 16) and the percentage of cost reduction using optimal number of resources for each competency shown in Fig. 5. The results show that the benefit of using optimal resources is on an average 10% reduction in cost.

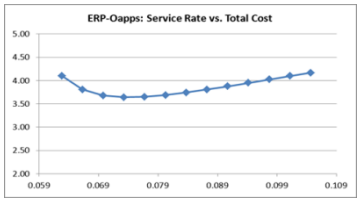

Fig. 4. Total cost versus service rate

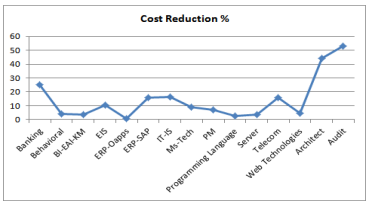

Fig. 5. Cost reduction % different competency

Overall, the performance of queuing model is found very satisfactory. We have also experimented different scenario based on different cost ratio. The results of optimal number of servers using equations (14, 15, & 16) against ID ratio are shown in **Fig. 6**.

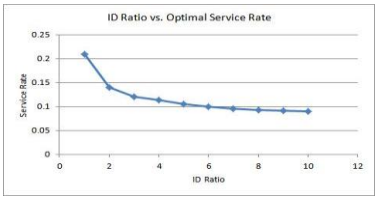

Fig. 6. Optimal service rate versus ID ratio

The above results show that the service rate needs to increase exponentially with the decrease of ID cost ratio.

4 Conclusions and Future Directions

Overall, this paper provides a queuing modeling approach to find an optimal number of resources for IT services business. An exact mathematical formulation is derived

for optimal number of resources determination. The queuing modeling approach is tested in real-life data from IT services business. The model fitness is tested and showed the accuracy of the model for mapping IT services business. The queuing approach provides the optimal number of servers and showed the effectiveness of the mathematical approach using improvement in total cost reduction. Though, this paper considers the issues of lead time and delay in allocation in IT services business, there are other issues, such as, multi-skilling servers, customer priority, need to be considered in queuing modeling. Therefore, the present approach can be extended for multi-skill resources and exact mathematical formulation need to be derived. Based on the cost of multi-skill and single-skill resources, optimal ratio of multi-skill vs. single skill determination can be considered as future directions of research. Also, the changes of resource's shape/skill level due to training need to be considered for future direction of research.

5 References

1. Adan, I. and Resing, J. 2001. Queuing Theory, Dept. of Mathematics and Computer Science, Eindhoven University of Technology, The Netherlands.
2. Andrews, B., H. Parsons. 1993. Establishing Telephone-Agent Staffing Levels through Economic Optimization. Interfaces 23: 2 14-20.
3. Artalejo, J.R., Economou, A., A. Gomez-Corral. 2007. Applications of Maximum Queue Lengths to Call Center Management. Computers and Operations Research 34 983-996.
4. Bassamboo, Achal J., Michael Harrison, and Assaf Zeevi. 2006. Design and Control of a Large Call Center: Asymptotic Analysis of an LP-Based Method. Operations Research 54:3 419-435.
5. Borst, Sem, Mandelbaum, Avi, Martin Reiman. 2004. Dimensioning Large Call Centers. Operations Research 52:1 17-34.
6. Grassmann, Winfried. 1988. Finding the Right Number of Servers in Real-World Queuing Systems. Interfaces 18:2 94-104.
7. Fazel, F., Fakheri, A. 2007. An Alternative Approach for Determining the Performance of M/M/s Queuing Model. Presented at Decision Sciences Institute Annual Meeting, November 18-21.
8. Ittig, P. T. 1994. Planning Service Capacity when Demand is Sensitive to Delay. Decision Sciences 25:4 541-559.
9. Kolesar, P. J., K. L. Rider, T. B. Crabill, W. E. Walker. 1975. A Queuing-Linear Programming Approach to Scheduling Police Patrol Cars. Operations Research 22:6 1045-1062.
10. Radmilovic, Z., B. Dragovic, R. Mestrovic. 2005. Optimal Number and Capacity of Servers in $M^{x=a}/M/c(\infty)$ Queuing Systems. Information and Management Sciences 16:3 1-16.
11. White, J. A., J.W. Schmidt, G.K. Bennett. 1975. Analysis of Queuing Systems. New York, NY: Academic Press.
12. Yang, R., S. Bhulai, R. V. D. Mei, F. Seinstra. 2011. Optimal Resource Allocation for Time-Reservation System. Performance Evaluation 6:5 414-428.