TIME DEPENDENT PRIORITIES IN CALL CENTERS

LASSAAD ESSAFI and GUNTER BOLCH

Institute of Computer Science
University of Erlangen-Nuremberg
Martensstrasse 3, D-91058 Erlangen, Germany
E-mail : lassaad@essafi.de, bolch@informatik.uni-erlangen.de

Abstract:
This paper shows how time dependent priorities can be applied to providing service assurance in a call center. The main contribution of this paper is the transfer of several results derived in the field on Proportional Differentiated Services (Internet Quality of Service) to application in the research field of Call Center technology. Further, it addresses some specific aspects of application to call centers. After presentation of general concepts of call centers and priority mechanisms, service level functions in call centers is introduced and the the suitability of time dependent priorities is discussed.

Key Words: Call Centers, Time Dependent Priorities, Service Quality

1 INTRODUCTION

Call centers have become an important channel of interaction with customers for many types of businesses. Planning and optimization aspects have become an important area of investigation. Example areas of research cover for example capacity planning (e.g. staff requirement and scheduling), service level implementation and monitoring, etc. [1]. An important aspect for research is the scheduling mechanisms used to serve incoming customer calls. Different strategies can be implemented and must ideally be aligned with the service level requirements. Priority scheduling is one family of schedulers which is widely used in several application fields, e.g. network traffic scheduling, real time systems. The aim of this paper is to present how priority based schedulers can be used in call centers. This paper is structured as follows: we first give a brief overview of call center concepts and technology. We then describe how a call center can be modelled as a queueing system in section and give brief overview of priority scheduling strategies. Then we present how time dependent priorities can be applied to call centers; and conclude.

2 WHAT IS A CALL CENTER?

A call center constitutes a set of resources, typically personnel, computers and telecommunication equipment, which enable the delivery of service via the telephone [1]. Human resources, who serve the customers contacting the call center, are considered as one key resource in a call center and are usually referred to as agents. The growth of call centers has been substantial over the last years and their application is ever increasing. Examples are customer service, airline booking services, directory services, telemarketing/tele-sales, emergency response services, etc.

Call centers can be classified based on several criteria (reference is made here to [2], e.g.

- Function: e.g. booking service, inquiries, etc.
- Initiation of the contact: three types are defined. *Inbound* call centers receive calls from customers (e.g. for customer service), whereas in *outbound* call centers the calls are initiated by the agents (e.g. for market research). In a *blended* call center, agents receive and initiate calls.
- Size: This criteria is mainly dependent on the number of employees. Small to mid size call center refers to call centers with no more than 50 agents (in Germany, more than 73% of the call center are of this category [2]). Large call centers can have thousands of agents.

- Geography: either centralized of distributed operations.
- Communication Channel: In addition to contact via phone, several call centers use other channels, e.g. fax, email, internet chat to interact with their customers. This type of call centers is usually referred to as a *Contact Center*.

In the rest of this paper, only inbound call centers will be considered.

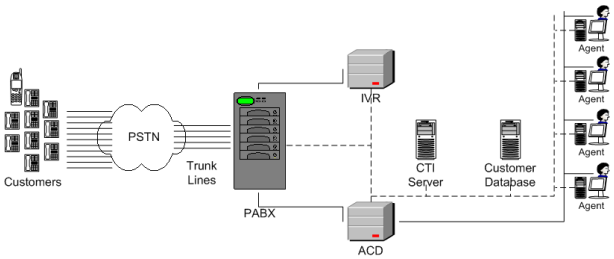

FIGURE 2.1: Call Center Technology

A typical call center consists of the following components (see also Figure 2.1): a private branch exchange (PABX), an automatic call distributor (ACD), an interactive voice response (IVR) and agent work stations, usually consisting of a PC and a telephone. The PABX is connected to the public telephone network through a number of telephone lines (trunk lines). An incoming call is connected, if one or more trunk lines is free. Usually the call is routed first to an IVR, which provides standard messages and/or guides the caller through a menu to select the requested service. If the caller requests to speak to an agent, then the call is handled by the ACD, which based on several criteria, routes the call to a free agent.

3 CALL CENTERS AS PRIORITY QUEUEING SYSTEMS

Queueing theory has been applied for a number of purposes in call center design and optimization. In fact, the (simplified) description given above about a standard call processing, enables us to directly map an inbound call center onto a queueing model. Hereby, several aspects have to be considered to define the appropriate model. In [2] page 22 and in [1] page 11, general queueing models and the criteria which have to be considered have been proposed.

[1] homogeneity refers in some papers to skill level too, e.g. in [2]

3.1 Criteria

The main criteria for the queueing model definition can be summarized as follows:

- Customer profile: describes the arrival process. Customers can belong to different classes, either on arrival (by calling specific numbers) or after classification (e.g. by IVR). Depending on whether several classes are defined or not, customers are usually referred to as *homogeneous vs. heterogeneous customers*. The arrival process is usually described as a *Poisson arrival* process. Furthermore, the customer profile characterizes the patience of customers, i.e. whether they leave the system, after being connected. It also if necessary describes the retrial behavior, in case they receive a busy signal.
- Agent characteristic: describes first the general qualification of the agent to handle an incoming call. If all agents can serve all calls, then they are called *homogeneous agents*, otherwise *heterogeneous*[1]. The agent characteristic also describes the skill level, i.e. the *service time* and the *number of agents*. The service time is usually modelled by an exponential or Weibull distribution.
- Scheduling policy: defines the policy implemented in the ACD, which chooses the customer to serve next depending on the customer classes and free agents. This can be regarded as two steps: a customer selection process and an agent selection process.
- Waiting room size: describes how many customers can wait for service, if all agents are busy. It can be defined per customer class (split buffers) or per system (shared buffer).

3.2 The Need for Priority Scheduling

The scheduling policy determines which customer to choose next. In order to achieve differentiation between the service level offered to different customer classes, priority policies can be used. Examples:

- In a booking call center, customers with different loyalty levels (e.g. regular and gold) have to be served. The call center objective is to minimize the waiting times of gold customers in comparison to regular customers.

- In a help desk call center, complaints of particular severity have to be recorded and processed before other complaints.

- A contact center with different channels (e.g. phone, fax and email), targets to provide and ensure different service levels for the different channels.

3.3 Queueing Model Definition

Based on the criteria described above, the model to be investigated can be summarized as follows (see also Figure 3.1):

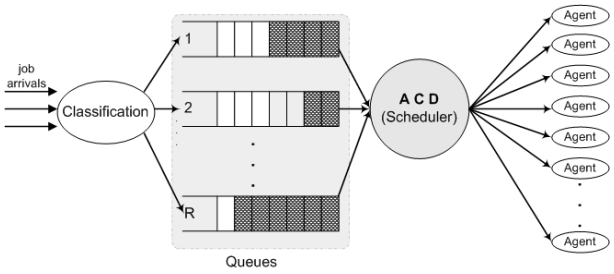

FIGURE 3.1: Call Center as Priority Queueing System

3.4 Priority Scheduling Policies

The priority disciplines can be generally classified into static and time dependent priorities. We assume in the following text a queuing model with R classes of customers, where arriving customers belong to a priority class r ($r = 1, 2, \ldots, R$). We use:

$$q_r(t) = \text{Priority of class } r \text{ at time } t$$

to denote the priority function of a class r customer.

Static Priorities In a static priority system, the priority of a customer is constant during its whole sojourn time in the system. The next customer to be served is the customer with the highest priority r. Within a priority class the queuing discipline is FCFS (First-Come-First-Served). The priority function of class r customer, can be written as:

$$q_r(t) = r,$$

where r is fixed and is independent of time t.

Time Dependent Priorities In many cases it is advantageous for a customer priority to increase with the time. Such systems are more flexible but need more expense for the administration. In this section, we introduce different types of time dependent priorities.

We refer to the same queuing model and assign each priority class a parameter b_r, which can be interpreted according to the priority function

$$q_r(t) = (t - t_0)b_r \tag{1}$$

as the increasing rate (slope) of the priority in the class r, where $0 \leq b_1 \leq b_2 \leq \ldots \leq b_R$. This means that the priority of a higher class customer increases faster that the priority of a lower class customer. A customer enters the system at time t_0 and then increases its priority at the rate b_r (see Figure 3.2).

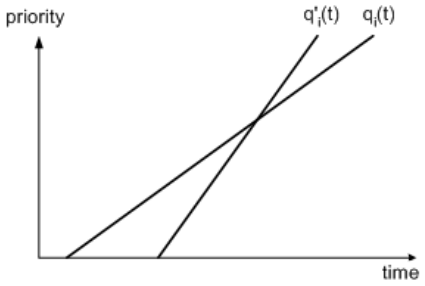

FIGURE 3.2: Priority functions with slopes b and b'

Variants of this priority function, where exponents are assigned to the time component and/or slope components have also been introduced in the literature.

Class Dependent Starting Priorities A possibility to reduce the waiting time of certain customers is to assign to each packet a class dependent starting priority r_r. With a slope 1 for the priority functions of all classes and with

$$0 \leq r_1 \leq r_2 \leq \ldots \leq r_R$$

we get the following form for the priority functions:

$$q_r(t) = r_r + t - t_0 \tag{2}$$

Earliest Deadline First Another strategy for scheduling customers is based on deadlines which are assigned to each customer. R customer classes are defined with a parameter G_i for each class, with:

$$G_1 > G_2 > \ldots > G_R$$

The parameters G_i define the time period, which may maximally be elapsed to serve a packet "in time". Packets with the lowest deadlines have the highest priority and vice versa. According to the priority function:

$$q_r(t) = \begin{cases} \frac{(t-t_0)}{(G_r - t + t_0)} & \text{if } t_0 < t \leq G_r + t_0 \\ \infty & \text{if } G_r + t_0 \leq t < \infty \end{cases}$$

the priorities increase faster, whenever the deadline of a customer nears. If the deadline is reached, the packet gets an infinite priority, which forces the system to serve the packet. Jobs with a positive infinite priority are served in a FIFO order.

4 APPLICATION OF TIME DEPENDENT PRIORITIES IN CALL CENTERS

The selection of the Scheduling Policy depends on the service level that needs to be achieved. If no customer differentiation is required and one service level is applicable to all customers, then FCFS might be the appropriate strategy to apply.
However, there are several applications, which require customer segmentation and different service levels to be achieved for each customer class/segement (see examples in Section "time dependent priorities"). Regardless of the service level to be provided, the following considerations are made:

- Static priorities are simple to implement because, to a make a scheduling decision, the scheduler needs only to determine the highest priority nonempty queue. On the other hand, a static priority scheme allows a "misbehaving" class of customers at highest priority to increase the delay and decrease the available resources of customers at all lower priority levels. This leads in extreme cases to starvation of lower priority classes. Without proper "admission control", this policy provides no means to adjust the service level provided to the customers.

- The EDF discipline is known not to handle overloads very well, causing a domino-effect of missed deadlines.

- The time dependent priorities provide appropriate means of adjusting the service level provided to different classes ("control knobs").

4.1 Service Level

Selecting a service level objective is one of the key steps in a systematic planning of a call center, as it is the critical link between resources and results. The main technical measures of service level in a call center are [2]:

1. the waiting time of customers,

2. the loss of customers, and

3. the utilization of the agents.

Service level (SL) is commonly defined as a certain percentage of calls answered in a specified time frame, measured in seconds and is referred to as "Acceptable Waiting Time" (AWT). The industry standard is - in case of no customer differentiation - that 80% of all calls should be answered in 20 seconds. Other measures related to waiting time are the expected waiting time $E[W]$, the expected queue length or the probability that a customer receives service immediately. The average waiting time for answered calls is usually referred to as "Average Speed of Answer" (ASA).
Economic considerations of service level definitions can be found in [2].

4.2 Proportional Service Level

In this section we propose a proportional service level function.
The proportional service level function can be applied in the context of waiting times of different customer classes as the service level measure. We denote $E[W_i(t, t+\tau)]$ as the average waiting time of the class i customers which were served in the interval $(t, t+\tau)$. If there are no such customers, $E[W_i(t, t+\tau)]$ is undefined. The proportional service level function states that for all pairs (i, j) of customer classes and for all time intervals $(t, t+\tau)$ for which $E[W_i(t, t+\tau)]$ and $E[W_j(t, t+\tau)]$ are defined:

$$\frac{E[W_i(t, t+\tau)]}{E[W_j(t, t+\tau)]} = \frac{\delta_i}{\delta_j}. \tag{3}$$

The parameters δ_i are known as the *Delay Differentiation Parameters (DDPs)*. As higher classes $(j > i, \forall i, j \in \{1, \ldots, R\})$ are better than lower classes with respect to delay, that is, they experience shorter delays, the DDPs satisfy the relation $\delta_1 > \delta_2 > \ldots > \delta_R$.

4.3 Examples

The proportional service level function can be applied in a variety of scenarios. For example in an airline call center, a service level can be defined so, that "normal" customers may experience up to 4 times delay loyal customers (e.g. frequent travellers or senators). In this case, the values for the delay differentiation parameters can be set to 1 and 4 for the two customer classes. The same service level function can be applied in the context of a mobile operator call center serving both high value customers with high usage and low usage customer. A similar function is also given in [7] in the context of operator directory services, where Toll and Assist services have to be better served than Directory Assistance service (2s vs. 4s waiting time).

4.4 Analysis

In addition to the discussion in the beginning of this section regarding which class of priority scheduling mechanism is best suited to fulfill service level requirements, it has been identified in several results (e.g. [8, 9]), that the time dependent priorities with linear slopes is appropriate to support the proportionality requirement at high load.
Considering two customer classes $R = 2$, the average waiting times of the two customer classes can be written as:

$$E[W_1] = \frac{E[W_{FIFO}]}{1 - \rho_2(1 - \frac{b_1}{b_2})} \tag{4}$$

and

$$\begin{aligned} E[W_2] &= E[W_{FIFO}] - \rho_1 E[W_1](1 - \frac{b_1}{b_2}) \\ &= E[W_{FIFO}] - \qquad (5) \\ &\quad \rho_1 \frac{E[W_{FIFO}]}{1 - \rho_2(1 - \frac{b_1}{b_2})}(1 - \frac{b_1}{b_2}) \qquad (6) \end{aligned}$$

ρ_1, ρ_2 represent the loads and b_1, b_2 represent the slopes of the priority functions of class 1 and class 2 respectively. $E[W_{FIFO}]$ is given for an $M/M/m$ system by the equations presented (2.9 and 2.7 and 2.6) in [5].
From (4) and (5) we calculate the ratio of the expected average waiting times:

$$\frac{E[W_1]}{E[W_2]} = \frac{1}{1 - (\rho_1 + \rho_2) \cdot (1 - \frac{b_1}{b_2})} \tag{7}$$

We solve the Eqn. (7) to finally get:

$$\frac{b_1}{b_2} = 1 - \frac{1}{\rho_1 + \rho_2}(1 - \frac{E[W_2]}{E[W_1]}) \tag{8}$$

A set of results related to this problem have been derived in [10, 11, 12, 7]:

1. A specific waiting time differentiation ratio $\frac{\delta_2}{\delta_1}$ is feasible, if and only if the system utilization ρ satisfies the condition:

   $$\rho > 1 - \frac{\delta_2}{\delta_1}. \tag{9}$$

   For example a differentiation ratio of 4 ($\frac{\delta_1}{\delta_2} = 4$), cannot be achieved at a load less than or equal to 75%. Figure 4.1 shows the maximum achievable waiting times differentiation ratios as a function of the system load.

2. In order to meet predefined delay differentiation requirements specified by δ_1 and δ_2, the increasing rates of the priority functions have to be set to:

   $$b_1 = 1 \text{ and} \tag{10}$$
   $$\frac{b_1}{b_2} = 1 - \frac{1}{\rho}(1 - \frac{\delta_2}{\delta_1}). \tag{11}$$

3. As the load approaches 100%, the ratio of the scheduler parameters b_1/b_2 tends to the inverse of the corresponding DDPs (consider limit of Eqn. (11) as the utilization ρ tends to 1).

4. The scheduler parameters do not depend on the class load distribution, the delay differentiation ratios neither. These parameters depend on the *total* utilization in the queuing system.

5. The simulation study in [9] in the context of internet traffic has shown that the time dependent priorities provide higher priority classes **consistently** get better service, even when monitoring short time frames (predictability property).

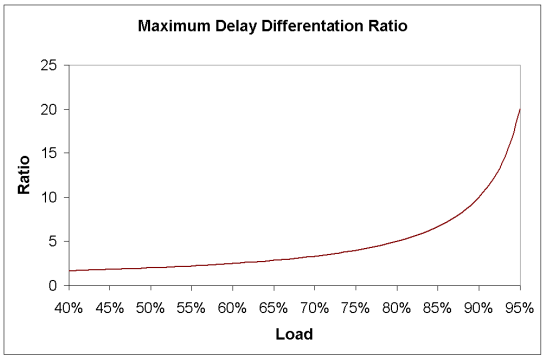

FIGURE 4.1: Maximum Delay Ratios

For a generalization to more than two customer classes, please refer to [11] for the use of genetic algorithms and lookup tables and to [12] for the use of iterative algorithms to calculate the class parameters $b_1, b_2, \ldots, b_R$ given $\delta_1 > \delta_2 > \ldots > \delta_R$ and a certain load distribution $\lambda_1, \lambda_2, \ldots, \lambda_R$.

5 CONCLUSION

In this paper we presented initial research results on how several known results in the area of internet quality of service can be mapped to call center design. We presented different priority mechanisms and used time dependent priorities to implement the proportional service level requirement. Our ongoing research includes investigation of absolute service requirements using time dependent priorities, and analyzing how this scheduling policy can be used to provide class dependent Acceptable Waiting Times (AWT), using second moments.

References

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

[2] *R.Stolletz:* "Performance Analysis and optimization of Inbound Call Centers", Springer 2003

[3] *G.Koole:* "Call Center Mathematics", Version of 24-Oct-2003

[4] *G.Bolch, L.Essafi, H. de Meer:* "Performance Evaluation of Priority based Schedulers. Proceedings des 2. MMB-Workshops: Leistungs-, Zuverlässigkeits- und Verlässlichkeitsbewertung von Kommunikationsnetzen und verteilten Systemen", Hamburg, 19./20. September 2002, Bericht 242 des Fachbereichs Informatik der Universitaet Hamburg

[5] *G.Bolch, W.Bruchner:* "Analytische Modelle symmetrischer Mehrprozessoranalagen mit dynamischen Prioritäten", Elektronische Rechenanlagen, 26(1), page 12-19, 1984

[6] *G.Bolch, S.Greiner, H.de Meer, K.S.Trivedi:* "Queueing Networks and Markov Chains : Modeling and Performance Evaluation with Computer Science Applications", Wiley, New York, 1998

[7] *M.Perry, A.Nilsson:* "Performance Modeling of automatic call distributors: Assignable Grade of Service staffing", 14. International Switching Symposium, 2:294-298, 1992

[8] *L.Kleinrock:* "Queueing Systems - Volume II" , Wiley, 1976

[9] *C.Dovrolis, D.Stiliadis, P.Ramanathan:* "Proportional Differentiated Services: Delay Differentiation and Packet Scheduling", ACM SIGCOMM '99, Cambridge, MA, Sep. 1999

[10] *L.Essafi, G.Bolch, A.Andres:* "An Adaptive Waiting Time Priority Scheduler for the Proportional Differentiation Model", ASTC HPC'01, Seattle, Apr. 2001

[11] *L.Essafi, G.Bolch, H.de Meer:* "Dynamic Priority Scheduling for Proportional Delay Differentiated Services", MMB, Aachen, Sep. 2001

[12] *M.K.H.Leung, J.C.S.Lui, d.K.Y.Yau:* "Characterization and Performance Evaluation for Proportional Delay Differentiated Services", Proceedings of IEEE International Conference on Network Protocols, Osaka Japan, Nov. 2000

AUTHOR BIOGRAPHIES

LASSAAD ESSAFI was born is Sfax, Tunisia and went to the University of Erlangen-Nuremberg, where he obtained his Master's Degree (Dipl.-Inf.Univ.) in Computer Science with highest grade. He worked for network suppliers and systems integrators especially in the mobile telecommunication area. He is also member of the the analytical modelling group in the University of Erlangen, headed by Dr. Gunter Bolch. His main area of research is analytical analysis and applications of time dependent priorities.

Dr. GUNTER BOLCH is Acad. Director at the Department of Computer Science 4 (Operating Systems) at the University of Erlangen-Nueremberg, Germany. He studied Telecommunication at the Technical Universities of Karlsruhe and Berlin and was Assistant Professor at the Department of Process Control at the University of Karlsruhe. In 1973 he completed his Ph.D. (Dr.-Ing.). After he received his Ph.D. he went to the University of Erlangen and since 1982 he is head of the Performance Modelling and Process Control research

group at the Department of Computer Science 4 (Informatik 4). He is researching and lecturing in the area of Performance Modelling (using Queueing Networks, Stochastic Petri Nets and Markov Models), Process Control and Operating Systems. He was Visiting Professor at several universities and was involved in many cooperations with industry and other universities. Dr. Bolch has written five textbooks on Performance Modelling and Process Control and more than 100 publications in these areas. He was involved in organizing national and international conferences. Dr. Bolch is a member of the GI, SCS and MMB.