Sizing of xDR Processing Systems

Bálint Ary and Sándor Imre

Budapest University of Technology and Economics,
Department of Telecommunications,
Magyar Tudósok körútja 2., 1117 Budapest, Hungary
ary.balint@isolation.hu, imre@hit.bme.hu

Abstract. Postpaid billing systems in most cases are using offline charging methods to rate the calls. Since latency is an accepted property, the throughput can be lower than the capacity required to process peak-hour traffic in a real-time manner. In this paper we will give an efficient mathematical model to calculate the processing power while taking the maximum queue size and maximum record age constraints into consideration.

1 Background

Billing systems in the telecommunication industry have different modules to fulfill the business processes from call pricing to the settlement of the bill. The module, which is responsible to rate the calls is often called *rater*. The mobile telecommunication companies generally have two different payment methods (prepaid and postpaid), and two different rating approach (offline and online). Usually the method determines the approach and the IT system underneath: online charging requires real-time computation thus requiring more processing power, while offline rating has softer constraints on the processing time and on the capacity of the supporting IT infrastructure.

Online charging is done while the call is made via online, socket based interfaces, while offline rating is based on files. Sizing of the real-time (online) system can be done with the help of queuing theory and since the system shall be capable to process all the calls real-time (even in peak period), the sizing must take these busy periods into consideration. The records in offline rating are called *call detail records*, *charging detail records* or *event detail records* and often referred as CDRs, EDRs, or more generally xDRs. The price of the call made in the offline system is derived from the corresponding CDR and since these records are sent to the billing system using some non-real-time protocol (FTP for instance), there is no real-time requirement against these modules. This allows some latency during the processing and we can undersize the systems according to peak periods. Even though queuing theory can be applied here with changing incoming probability over time, in most cases the business is not interested in a few minutes difference between processing times. This simplification allows us to observe and calculate the required processing power with a greater scale and using functions that represents the incoming CDR number and the processing power over time.

R. Szabó et al. (Eds.): AccessNets 2010, LNICST 63, pp. 62–70, 2011.

Although we made some simplification, offline rating must comply with several requirements. As such, usually the business is interested in the maximum age of the unprocessed CDRs to calculate the fraud possibility and the operational team is interested in the maximum queue size to calculate the required disk space. In the next chapters we will represent the required mathematical formula to compute the minimum processing power if the maximum queue size and latency is given. Finding the proper, not oversized processing power is beneficial, and should lower the cost of IT infrastructure as well as software licensing fees.

In this paper we will discuss the incoming CDRs versus the computing capacity, however, the same equations and results can be used to size call centers to the incoming calls, as in most cases the same business requirements (with different values and functions) should apply. Sadly, the call centers are far more sensitive for processing time differences, and the maximum age of a request in a queue shall kept low. Since processing time differences resulted from the call arrival distribution is significant, our model shall be circumspectly used.

The available literature mainly deals with queuing theory while calculating the appropriate sizing for telecommunication and other queue based processing systems [2][6]. In many case [7] the estimated waiting time is calculated during peak hour, or a constraint is given for the maximum waiting time [3][4] but the job or record is vanishing from the queue after a certain amount of time, thus these models cannot be applied for telecommunication networks and call detail records. Some literature is dealing with call center sizing [1] and scheduling [5], which – as mentioned above – is more sensitive to processing time jitters, and as so, these models shall be used instead in these cases.

In chapter two we will clarify the used model and simplifications as well as the possible business requirements. In section three and four we will detail the queue size and record age constraints respectively, while in section five we will represent a simple case with simplified functions as an example for the calculations. Section six summarizes the results of this paper and outlines possible future works.

2 Assumptions and Requirements

The queue size and the maximum item age in a processing queue cannot be given or calculated in a closed mathematical form in general. Since we are calculating the aforementioned values in a specific system, we can make some assumptions in order to simplify the complexity of the required formulas.

We will use two different functions to represent the main characteristics of the system. We will denote the number of incoming CDRs over time with $c(t)$, and we will represent the processing capacity of the system with $p(t)$. The later one is measured with the number of processable CDRs. Thus, if $p(t) \geq c(t)$ for every t, then the system will process every CDR immediately, which (taking the real-life examples into consideration) is a rude waste of resources and a beautiful example of system oversizing.

The terminology of charging and billing systems define the *processing window* as a daily time period, where the rating system is up and running and capable to process and rate incoming CDRs. This processing window normally starts when the nightly activities (often referred as *end of day* or EOD) are finished and ends around midnight (where EOD activities are started again). The rating module must be turned off during the EOD period to assure consistent rating and billing, since the different reference modification and billing activities taking place at this time.

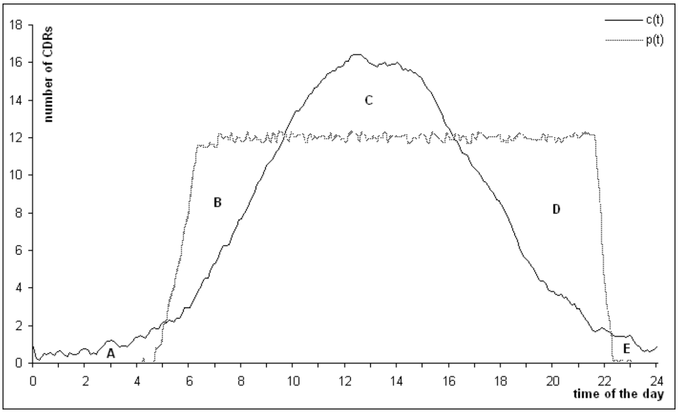

Fig. 1. General incoming CDR and processing power functions

The incoming number of CDRs can be represented with a general bell-curve: the number of phone calls, sent SMSs and GPRS activities are low during the night and peaks during the mid-day. The price of the call (or other services) is often different in this two (peak and off-peak) periods. Generally, the processing window starts, when the number of incoming CDRs is low but rising, and ends when the number of records is decreasing. The maximum processing power generally does not exceeds the number of incoming CDRs during peak hour, thus the two functions intersects four times as represented on Figure 1. The number of unprocessed CDRs in the processing queue increasing as long as the processing power is less then the number of incoming CDRs and decreasing in every other case. In our paper, we will assume the followings:

AS1. The function representing the incoming CDRs ($c(t)$), and the function representing the processing power of the rating system ($p(t)$) resemble the functions represented in Figure 1. At least, the intersections and positions of the functions can be related to the displayed functions.

AS2. Both functions are day-independent. We do not distinguish between weekdays, holidays and working days, and we do not calculate or care the differences between consecutive days.

AS3. The scheduling of the CDRs in the processing queue is FIFO (first in, first out), which complies with the implementation of the available commercial telecommunication billing systems.

Generally, these rating systems shall comply with different business requirements as mentioned in Section 1. Some of them are mandatory from engineering point of view, some of them are purely business, financial or security requirements. In this paper we will represent a sizing model, where the following three requirements are taken into consideration.

R1. The system shall be capable to process the daily CDRs in one day. Moreover, the system shall have some spare capacity to process additional CDRs (taking Christmas or New Years Eve into consideration for example).

R2. The maximum number of unprocessed CDRs should not exceed Q (a given IT parameter).

R3. The oldest unprocessed CDR during the normal period shall not be older then K (a given business requirement) during the normal period. The system shall catch-up (lower the oldest record age below this level) shortly after it is started.

To ease the further computations, please let us distinguish five different areas (A, B, C, D and E) and five different moments (m_1, m_2, m_3, m_4 and m_5) as displayed in Figure 1 as follows:

A Early morning area. The processing is not yet started, or the processing capacity is less then the number of incoming CDRs. The size of this area is equal with the number CDRs increasing the processing queue during this period.

B Morning area. The rater is up and running and the processing capacity is more than the number of incoming CDRs. The size of this area is equal with the number CDRs vanishing from the queue during this period.

C Peak area. The processing has started, but the number of incoming CDRs exceeds the processing capacity again. The processing queue is increasing, and the increment is equal with the size of this area.

D Afternoon area. The number of incoming CDRs is below the processing capacity. The processing queue is decreasing.

E Night area. The system shut down, but CDRs are still coming in. The processing queue is increased with the size of this area.

m_1 Start time. The moment, when the processing power exceeds the number of CDRs in the morning. This is the end of area A and the start of area B.

m_2 Peak start time. The moment, when the number of incoming CDRs exceeds the processing power. This moment is around the start of the peak hour before noon. This is the end of area B and the start of area C.

m_3 Offpeak start time. The moment, when the processing power exceeds the number of CDRs in the afternoon. This is the end of area C and the start of area D.

m_4 Shutdown time. The number of incoming CDRs exceeds the processing power again. EOD will start shortly. This is the end of area D and the start of area E.

m_5 Midnight. This is the end of the day, and the end of area E.

We can calculate the above defined areas with the help of $c(t)$, $p(t)$ and the above defined moments as follows:

$$A = \int_0^{m_1} c(t)dt - \int_0^{m_1} p(t)dt \tag{1}$$

$$B = \int_{m_1}^{m_2} p(t)dt - \int_{m_1}^{m_2} c(t)dt \tag{2}$$

$$C = \int_{m_2}^{m_3} c(t)dt - \int_{m_2}^{m_3} p(t)dt \tag{3}$$

$$D = \int_{m_3}^{m_4} p(t)dt - \int_{m_3}^{m_4} c(t)dt \tag{4}$$

$$E = \int_{m_4}^{m_5} c(t)dt - \int_{m_4}^{m_5} p(t)dt. \tag{5}$$

3 Queue Size

In this section we will give mathematical formulas for the first two requirements mentioned in Section 2. In order to process the proper amount of CDR in one day, we have to determine the processing capability to satisfy to following inequality:

$$\int p(t)dt > \int c(t)dt. \tag{6}$$

Using the areas defined in the requirements section, the following statement must comply:

$$D = -A + B - C + D - E > 0. \tag{7}$$

where D denotes the additional CDR processing power in one day if it is greater then 0. Otherwise the first requirement ($R1$) is not met.

We will prove, that if $D > 0$, then there is no unprocessed CDR at m_2 or m_4. In order to do this, let us denote the number of unprocessed CDRs at the end of the day with R. Since the queue size is increasing before m_1, during m_2 to m_3 and after m_4, the queue size cannot be negative and due to assumption $A2$ the value of R on the previous day shall be equal with the current value, we can calculate R as:

$$R = max(0, max(0, R + A - B) + C - D) + E. \tag{8}$$

If $R + A > B$, then the queue is not empty at m_2, since the unprocessed CDRs from the previous day, plus the morning CDRs are not processed till this moment, thus

$$R = max(0, R + A - B + C - D) + E. \tag{9}$$

If $R+A-B+C-D>0$, then $R=R+A-B+C-D+E$, but the condition of $A-B+C-D+E<0$ (see equation 7) out rules this possibility, leaving us only with the $R+A-B+C-D\leq 0$ option. In such case $R=E$, thus the processing queue is empty at m_4.

If $R+A<B$, then the queue is empty at m_2, and we have the following equation for the queue size at the end of the day:

$$R = max(0, C-D) + E. \tag{10}$$

Thus, the queue size is either $R=E$ if $C\leq D$ (which makes the processing queue empty at m_4 as well), or $R=C-D+E$ otherwise.

It can be easily understood, that the maximum queue size can be calculated as follows:

$$Q_{max} = max(E+A, C, E+A-B+C, C-D+E+A). \tag{11}$$

4 Constraint on Record Ages

According to the requirements, the system shall catch-up with the CDRs early in the morning. More precisely, the system shall process all the CDRs which are older then K between m_1 and m_2. If the system queue is empty in m_2, then this requirement is straightforward. Otherwise (if the processing queue is empty only at m_4), this requirement can be modeled with the following integral function:

$$E + \int_0^{min(0,x-K)} c(t)dt \leq \int_0^x p(t)dt. \tag{12}$$

The processing function ($p(t)$) shall be capable to fulfill this inequality with the condition of $m_1\leq x\leq m_2$. Let us denote the result of this equation (the minimal x, which fulfills this inequality) with G (as grace period).

Taking the mid-day ageing requirement into consideration, we have to differentiate two cases. If the queue does not clear out till m_2, then the above equation can be used with the condition of $G\leq x\leq m_4$, otherwise the requirement is fulfilled trivially until $x\leq m_2+K$, for the rest of the time, the following equation can be used where $m_2+K\leq x\leq m_4$:

$$\int_{m_2}^{(x-K)} c(t)dt \leq \int_{m_2}^x p(t)dt. \tag{13}$$

If the queue is empty at m_4, then the requirement is trivially fulfilled after m_4 until m_4+K, moreover it is fulfilled until N (extension period), where N is the solution of the following integral function if $x\geq m_4+K$

$$\int_{m_4}^{x-K} c(t)dt = \int_{m_4}^x p(t)dt. \tag{14}$$

5 Example

Let us give an example for these calculations. We will simplify both the incoming CDR and the processing functions as represented on Figure 2. The processing power will be denoted with P as follows:

$$c(t) = \begin{cases} 2 & \text{if } t < 6 \text{ or } t > 18 \\ 10 & \text{if } 6 \le t \le 18 \end{cases} \tag{15}$$

$$p(t) = \begin{cases} 0 & \text{if } t < 4.5 \text{ or } t > 23 \\ P & \text{if } 4.5 \le t \le 23. \end{cases} \tag{16}$$

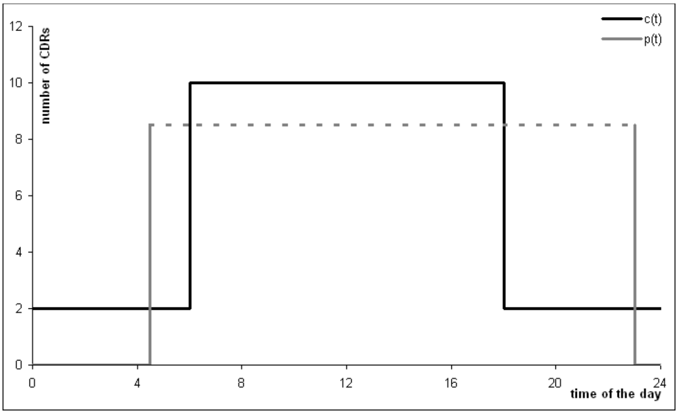

Fig. 2. Incomming CDRs and processing power

Our task is to calculate the value of P so it fulfills the different requirements. The minimum processing power (P_{min}) can be calculated from the main queuing theory requirement, thus solving the inequality of $\int p(t)dt > \int c(t)dt$ gives us, that

$$P_{min} > \frac{144}{18.5} \approx 7.78. \tag{17}$$

The maximum queue size can be calculated with (1)-(5) and (11), and represented on Figure 3 as a function of P. We have drawn all four values from the max function, but only the function with the highest value with a given P shall be used. It can be seen, that we cannot decrease the queue size below 11 but for smaller P values the $Q = A + E - B + C$ is dominant. With the minimum required power the system will operate with a queue size around 28.9.

Taking the CDR age into consideration, we have to fulfill two different requirements. The system shall catch-up between 4.5 and 6 o'clock and for the rest of the day, the maximum age in the queue shall not exceed K. It is obvious, that K is a function of P (and vice-versa), and it is represented on Figure 4. As it can be seen, the mid-day requirement is stronger, and it requires more power capacity.

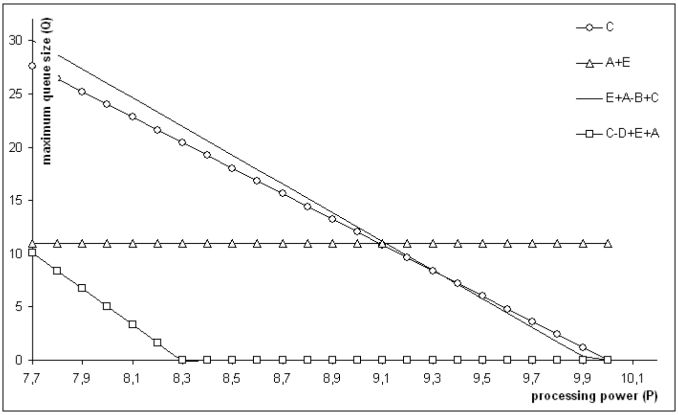

Fig. 3. Queue size

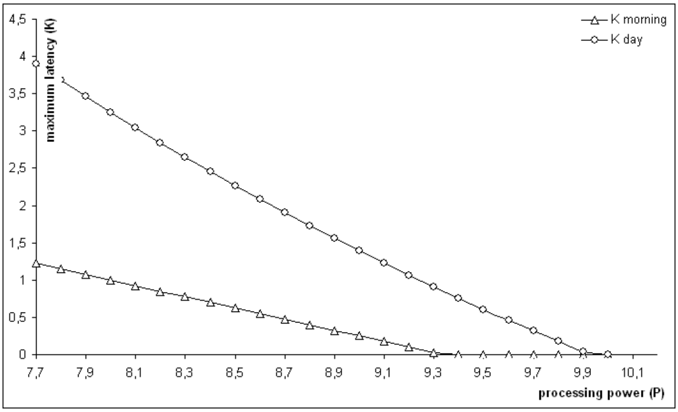

Fig. 4. Latency

Let us do some calculation with the following given requirements: The maximum queue size should not exceed 15, and every CDR should be processed within 1.5 hours. Thus, we have the following equations for queue size:

$$Q = A + E - B + C \tag{18}$$

$$Q = 9 + 2 - 1.5(P_Q - 2) + 12(10 - P_Q) \tag{19}$$

$$P_Q = \frac{134 - Q}{13.5} = \frac{134 - 15}{13.5} \approx 8.814, \tag{20}$$

and for the age requirement we get the strongest constraint if $6+K < x \leq 18+K$ when using (12) since the queue is not empty at m_2:

$$P_K(x-4.5) = 2+12+10(x-K-6) \quad (21)$$

$$P_K = 10 + \frac{-10K-1}{x-4.5}, \quad (22)$$

and we need the highest P if $x = 18 + K$, thus:

$$P_K = 10 + \frac{-10K-1}{13.5+K} \quad (23)$$

$$= \frac{134}{13.5+K} = \frac{134}{15} \approx 8.933 \quad (24)$$

The required power is the maximum of the above calculated powers, thus

$$P_{total} = max(P_{min}, P_Q, P_K) = P_K \approx 8.933. \quad (25)$$

6 Summary

In this paper we have summarized the nature of offline billing systems from sizing point of view. We gave mathematical formulas for the possible business requirements and guidelines to calculate the required processing capacity. If the number of incoming CDRs is known over time, and we have constraints on the start and end time of the processing window we can calculate the required processing capacity that fulfills the business requirements. This model can also be used to size the call centers of a telecommunication system if the number of incoming calls is known over time.

The model can be further refined, if the required background capacity (processing capacity, that is not used to process CDRs, but for other required activities) is known, and different that 0. Also, in most cases the number of CDRs on weekdays and national holidays are different from the regular working days. This fact, and additional business requirements may affect the required processing power and the calculation can be refined accordingly in future researches.

References

1. Anderson, G.L., Flockhart, A.D., Foster, R.H., Mathews, E.P.: Queue waiting time estimation (EP0899673) (August 2003)
2. Daigle, J.N.: Queueing Theory with Applications to Packet Telecommunication. Springer Science, University of Mississippi (2005)
3. Graves, S.C.: The application of queueing theory to continuous perishable inventory systems. Management Science 28(4) (April 1982)
4. Rajabi, A., Hormozdiari, F.: Time constraint m/m/1 queue (2006)
5. Rottembourg, B.: Call center scheduling (2002)
6. Schoenmeyr, T., Graves, S.C.: Strategic safety stocks in supply chains with capacity constraints (2009)
7. Shtivelman, Y.: Method for estimating telephony system-queue waiting time in an agent level routing environment (6898190) (May 2005)