xDR Arrival Distribution

Bálint Ary
Budapest University of Technology and Economics
Department of Telecommunications
Email: ary.balint@isolation.hu

Dr. Sándor Imre
Budapest University of Technology and Economics
Department of Telecommunications
Email: imre@hit.bme.hu

***Abstract*—The knowledge of call start and call length distribution is required to design and dimension the network for telecommunication services, thus it is a well know topic for scientific publications. In order to size the charging and billing systems, we have to know the distribution of the call detail record arrivals. In this paper we will calculate the expected number of partial CDRs for long calls and introduce simulation results for the CDR arrival distributions.**

***Index Terms*—charging rating sizing mobile telecommunication**

I. Background

The price of the postpaid services is calculated from their *call detail record*, which is sent to the billing system after the call was made. These records (also known as *charging detail record*s or *event detail record*s and often abbreviated as CDRs, EDRs, or more generally xDRs) are grouped together by the Mobile Switching Centers (MSCs) or other service enabler modules of the network and sent to the billing system through an offline, file based protocol [1] [2]. Once the records arrive to the system, the appropriate module determines the price of the calls using the information stored in the records, the rating logic of the purchased tariff packages and discounts of the customers and the accumulated usage information of the subscribers in the given billing period.

The number of CDRs sent for a given service consumption depends on the type of service and the actual implementation of the system. Usually SMS generates four different CDRs (the actual message and the delivery notification for both the Mobile Originated (MO) and Mobile Terminated (MT) case), but in most cases three is filtered out by the mediation module and only one is used to determine the price of the service. Voice calls are usually generating one CDR, but if the calling and called party is connected to different MSCs, or the mobile station is moving and connected to more than one MSC during the call, then additional records are created.

With the advent of GPRS, 3G services, IMS (IP Multimedia Subsystem) and LTE (Long Term Evolution), the services are no longer priced (just) according to the length of the call. It is possible, that a rather long session costs a small amount of money, or on the contrary, a short session represents an expensive service. From business point of view, long and expensive service consumptions would have high risks, since the network operator would only be notified when the call has ended, and huge debits could be accumulated without the possibility of any intervention. To overcome this problem, standards define partial CDRs for long calls [3] [4]. Partial CDRs are generated while the call is made, and they are carrying information about the service consumption since the last partial CDR was issued. Standards define two type of partial CDR. The FQPC (Fully Qualified Partial CDR) holds all the required information, while RPCs (Reduced Partial CDRs) are containing all the mandatory fields [4] and the changes that occurred in any other field relative to the previous partial CDR. The RPCs are converted into FQPCs later on in the network elements, or in the billing system [3] [5].

On the network element level, when IMS elements (for example) are serving a long session, they are periodically sending interim messages to the corresponding elements [5] [6] [7]. After a few interim messages the network element creates partial CDR if the specified threshold is reached - this can include data volume limit, time (duration) limit, maximum number of charging condition changes or management intervention [8] [9]. Once the session is closed, a final CDR is generated.

When the partial CDRs are reaching the billing system, the network operator shall decide weather to rate the partial CDRs as unique records, or wait for the rest of information and the final CDR, aggregate the measured unit(s) and rate the whole session. This latter approach is referred as Long Duration Call (LDC) assembly. The aggregation is based on the record identifier (MSC address and Call Reference Number, PDG address and WLAN Charging ID, IMS Charging Identifier and so on) and on the partial record sequence [3] [4]. In this paper we will calculate the CDR distribution for long calls including the partial and the final CDRs. We will assume, that partial CDRs will be generated with exact time intervals, and besides these records only the final CDR is generated – thus no handover, or other behavior results in additional CDR generation. We also assume that these records will be sent directly to the billing system one-by-one without further aggregation, buffering or delay, and that the call start and call length are independent random variables.

In the next section we will calculate the expected number of partial CDRs, section three calculates the distribution of the final CDRs if the call start and call length distribution is known. Section four shows the final distribution of the partial and final CDRs, while the final section summarizes this article.

II. Number of Partial CDRs

Let us assume that the call length of the long calls is given with the probability density function $g(t)$. The corresponding cumulative distribution function is denoted with $G(t)$.

If the system generates a partial CDR after every K minutes, then the expected value of the number of partial CDRs can be calculated with the following equation:

$$N = \sum_{i=0}^{\infty} iP_i = \sum_{i=1}^{\infty} iP_i, \tag{1}$$

where P_i represents the possibility, that the call length is between iK and $(i+1)K$. Please note, that at the end of the session, an additional CDR will be generated as mentioned in Section I, thus the above equation only gives us the number of partial CDRs. The P_i probability can be calculated as follows:

$$P_i = \int_{iK}^{(i+1)K} g(t)dt = G((i+1)K) - G(iK). \tag{2}$$

Later on, we will use the following two equations in the article. Equation 3 is trivial, once the first few parts of the sum are written in brief format, while (4) is coming from the nature of cumulative density functions.

$$\sum_{i=1}^{\infty}\sum_{j=i}^{\infty} P_j = \sum_{i=1}^{\infty} iP_i \tag{3}$$

$$\sum_{i=j}^{\infty}\Big(G((i+1)K) - G(iK)\Big) = (1 - G(jK)) \tag{4}$$

Because of these, (1) can be written as follows:

$$N = \sum_{i=1}^{\infty} iP_i = \sum_{i=1}^{\infty}\sum_{j=i}^{\infty} P_j \tag{5}$$

$$N = \sum_{i=1}^{\infty} i(G((i+1)K) - G(iK)) \tag{6}$$

$$= \sum_{i=1}^{\infty}\sum_{j=i}^{\infty}(G((j+1)K) - G(jK)) \tag{7}$$

$$= \sum_{i=1}^{\infty}(1 - G(iK)) \tag{8}$$

We will prove, that the number of partial CDRs is less than the expected value of $g(t)$ divided by K, moreover, if the expected value is denoted with $E_g(t)$, than:

$$\frac{E_g(t)}{K} - 1 \leq N \leq \frac{E_g(t)}{K}. \tag{9}$$

In order to do this, let us calculate the difference between $E_g(t)/K$ and the expected number of partial CDRs:

$$\frac{E_g(t)}{K} - N = \tag{10}$$

$$\frac{\int_0^{\infty} tg(t)dt}{K} - \sum_{i=0}^{\infty} i(G((i+1)K) - G(iK)) = \tag{11}$$

$$\sum_{i=0}^{\infty}\int_{iK}^{(i+1)K}\frac{t}{K}g(t)dt - \sum_{i=0}^{\infty} i\int_{iK}^{(i+1)K} g(t)dt = \tag{12}$$

$$\sum_{i=0}^{\infty}\int_{iK}^{(i+1)K}(\frac{t}{K} - i)g(t)dt. \tag{13}$$

Since within the boundaries of the integral $iK \leq t \leq (i+1)K$ and

$$0 = \frac{iK}{K} - i \leq \frac{t}{K} - i \leq \frac{(i+1)K}{K} - i = 1, \tag{14}$$

the following relation is true for the difference:

$$0 \leq \sum_{i=0}^{\infty}\int_{iK}^{(i+1)K}(\frac{t}{K} - i)g(t)dt \leq \sum_{i=0}^{\infty}\int_{iK}^{(i+1)K} g(t)dt = 1, \tag{15}$$

thus (9) is proven.

Let us give an example if the call length distribution is a heavy tailed distribution. We've used the log-normal distribution for this purpose as it is suggested in the corresponding publications [10]. The log-normal probability density function and cumulative distribution function is

$$g(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\log_n x-\mu)^2}{2\sigma^2}} \tag{16}$$

$$G(x;\mu,\sigma) = -\frac{1}{2}\mathrm{erf}\Big(\frac{\mu - \log_n x}{\sigma\sqrt{2}}\Big) + \frac{1}{2}, \tag{17}$$

where *erf* denotes the error function:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt. \tag{18}$$

Using (8) and (17) the number of partial CDRs in this special case can be calculated as:

$$N = \sum_{i=1}^{\infty}(1 - G(iK)) \tag{19}$$

$$= \sum_{i=1}^{\infty}\Big[1 + \frac{1}{2}\mathrm{erf}\Big(\frac{\mu - \log_n iK}{\sigma\sqrt{2}}\Big) - \frac{1}{2}\Big] \tag{20}$$

$$= \frac{1}{2}\sum_{i=1}^{\infty}\Big[1 + \mathrm{erf}\Big(\frac{\mu - \log_n iK}{\sigma\sqrt{2}}\Big)\Big]. \tag{21}$$

We also created two simulations to confirm our results. We have generated 1 million calls with their length following the log-normal distribution. The used parameters were $\mu = 4$, $\sigma = 0.5$ and $K = 10$ for the first run and $\mu = 3.2$, $\sigma = 0.1$ and $K = 7$ for the second run (see Figure 1 for the call length distribution for the first run). We have used the Box-Muller algorithm to generate normal distribution which was later transformed to log-normal distribution. Once the calls were generated, we have calculated the number of partial CDRs for each call ($\lfloor length/K \rfloor$) and compared them to the

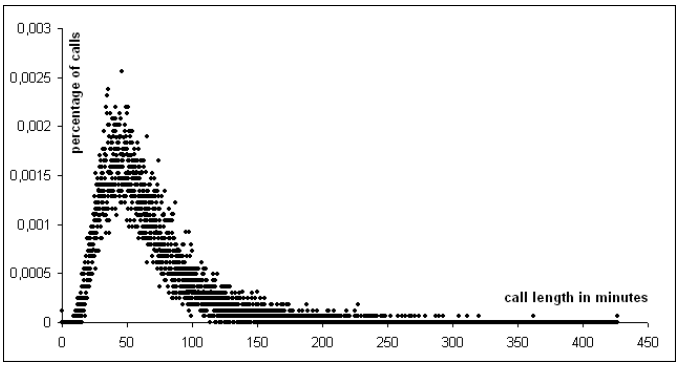

Fig. 1. Call length distribution

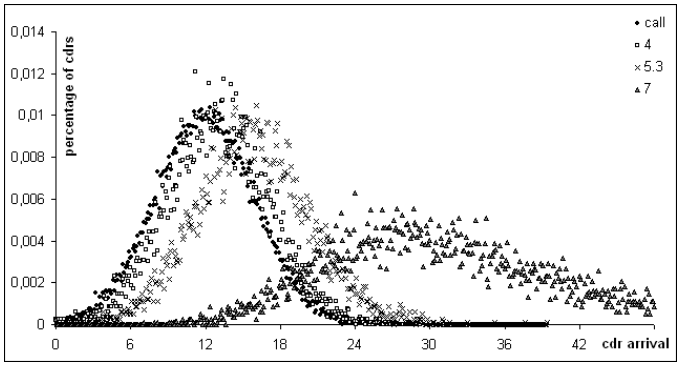

Fig. 2. Final CDR arrivals

TABLE I
EXPECTED NUMBER OF PARTIAL CDRS

parameters	$\mu = 4$; $\sigma = 0.5$; $K = 10$	$\mu = 3.2$; $\sigma = 0.1$; $K = 7$
simulation	5.68533731466	3.03301296699
approximation	5.686761151	3.03301296699
$E_g(t)/K$	6.186780925	3.522214288

TABLE II
EXPECTED VALUES

$\mu = 4$	1.031130154 hour
$\mu = 5.3$	3.783522439 hour
$\mu = 7$	20.71080278 hour

numeric approximation (summarizing from 1 to 100 – see (22) for the first run) and to the expected value ($e^{\frac{\sigma^2}{2}+\mu}$) divided by K. Table I summarizes the simulation and numeric results.

$$0.5\sum_{i=1}^{100}\Big[1+erf\Big(\frac{4-\log_n(10i)}{0.5\sqrt{2}}\Big)\Big]=5.686761151. \quad (22)$$

III. FINAL CDR ARRIVAL

Let us assume that the probability density function of calls made on the network is given with $f(\tau)$. If we would like to give a probability density function for the final CDR generation (and arrival to the billing system) we have to include the call length distribution in the equation, since the CDRs arrive to the systems once the calls were finished.

A final CDR is being generated at a given time (τ) if the call was started t minutes ago, and the call length is exactly t. This probability can be calculated with the multiplication of the two probabilities:

$$f(\tau - t)g(t). \quad (23)$$

We can calculate the probability that a CDR is generated at τ by summing up (integrating) these probabilities for all valid call length:

$$h_f(\tau) = \int_{0+}^{\infty} f(\tau - t)g(t)dt, \quad (24)$$

which gives us, that the CDR generation probability density function is the convolution of the call start distribution and the call length distribution if the call length probability distribution is zero if t is negative:

$$h_f(t) = (f * g)(t). \quad (25)$$

This result is straightforward from the assumption that the call start and call length are independent random variables.

Sadly, the call length cannot be modeled with a normal distribution because of the aforementioned restriction. The convolution of a normal and log-normal distribution cannot be given in closed format, thus we will run a simulation to observe the final distribution. Nevertheless, the convolution of the normally distributed call start and the log-normally distributed call length can be expressed as follows:

$$h(\tau) = \int_{0+}^{\infty} f(\tau - t)g(t)dt \quad (26)$$

$$= \frac{1}{2\pi\sigma\rho}\int_{0+}^{\infty}\frac{1}{t}e^{-\frac{(\tau-t-\nu)^2}{2\rho^2}-\frac{(\log_n t-\mu)^2}{2\sigma^2}}dt \quad (27)$$

In the simulation we have used the Box-Muller algorithm again to generate both the normal, and the log-normal distributions, and we have generated 10000 CDRs with normal distribution on call start and with log-normal distribution on call length. The parameters for the call start distribution was $\mu = 12$ and $\sigma = 4$ (in hours), while for the call length $\sigma = 0.5$ was used and μ varied during the test runs using the values 4, 5.3 and 7 (in minutes). Figure 2 represents the call starts and the empirical distribution for CDR arrival for the separate test runs. The expected values for the distributions are listed in Table II.

It can be observed, that for short expected lengths (comparable with the density of the call start), the CDR arrival distribution is not significantly different from the call start distribution, and only slightly shifted to the right. For dimensioning purposes, the original call start can be used, with the mean shifted with the expected length of the calls. However, for longer expected values (e.g.: one day), numerical approximation or simulation is suggested to correctly size the system.

IV. CDR DISTRIBUTION

If we would like to calculate the probability density function for CDR arrival for long calls, we have to include the final

and partial CDRs and weight the probability density function to give 1 when integrated.

It can be easily understood, that a partial CDR is generated at a given time (τ) if

- the call was started K minutes ago (at $\tau - K$), and the call length is greater than K.
- the call was started $2K$ minutes ago (at $\tau - 2K$), and the call length is greater than $2K$.
- and so on...

Let us count the partial CDRs. First of all, the total number of partial CDRs sent is equal with the number of calls (C) multiplied with the expected value of partial CDR per call (N). On the other hand, the number of partial CDRs sent at a given time ($n(t)$) can be calculated with the help of the above mentioned logic:

$$Ch_i(\tau) = n(t) = \sum_{i=1}^{\infty} Cf(t - iK) \sum_{j=i}^{\infty} P_j, \qquad (28)$$

and the integral of $n(t)$ shall give the total number of partial CDRs, thus:

$$\int_{-\infty}^{\infty} n(t)dt = \int_{-\infty}^{\infty} \sum_{i=1}^{\infty} Cf(t - iK) \sum_{j=i}^{\infty} P_j dt = CN. \qquad (29)$$

Since $f(t)$ is bounded, $\int_{-\infty}^{\infty} f(t)dt = 1$ and $f(t)$ and $g(t)$ are independent, this equation can be written as

$$\sum_{i=1}^{\infty} C \int_{-\infty}^{\infty} f(t - iK)dt \sum_{j=i}^{\infty} P_j = CN \qquad (30)$$

$$\sum_{i=1}^{\infty} C \sum_{j=i}^{\infty} P_j = CN \qquad (31)$$

$$C \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} P_j = CN, \qquad (32)$$

which is trivial since (5).

Combining this with the full CDR distribution, the final CDR probability density function for long calls can be given with

$$h(\tau) = \frac{h_f(\tau) + h_i(\tau)}{1 + N} \qquad (33)$$

$$h(\tau) = \frac{\int_{0+}^{\infty} f(\tau - t)g(t)dt}{1 + N} \qquad (34)$$

$$+ \frac{\sum_{i=1}^{\infty} f(\tau - iK) \sum_{j=i}^{\infty} P_j}{1 + N}. \qquad (35)$$

Again, we have created a simulation with different parameters to model the CDR arrivals. The approach is identical with the ones in the previous chapters. The call start distribution was a normal distribution with $\mu = 12$ and $\sigma = 4$ (in hours), and the call length was log-normal with $\sigma = 0.5$ and μ varied during the test runs (given in minutes). Figure 3 displays the results. For short expected lengths, the call start distribution is suggested, where the number of CDRs is $1 + N$. For longer $E_g(t)$, simulation or numerical approximation is suggested.

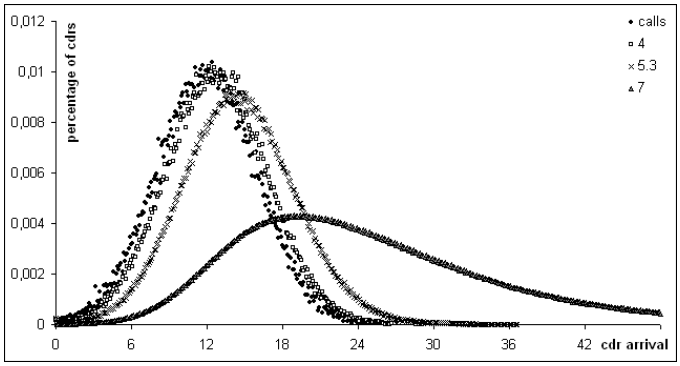

Fig. 3. CDR arrivals

V. Conclusion

In this paper we gave a mathematical model to calculate the expected number of partial CDRs, which was confirmed by simulation results. Also, we have showed the distribution for final CDR and partial CDR through mathematical formulas and simulations. Normal distribution was used to calculate the call start, while log-normal distribution was used to estimate the call length for long calls. During the simulations we have used the Box-Muller algorithm to generate these random variables.

The given model can be used to understand the incoming CDRs on a given day, and can be further used to estimate the required processing capacity for billing systems. Although, the long calls are still far more less than the regular voice calls, the trends and forecasts are showing the emerge of the data-call percentage, thus these results shall really help the experts to size the corresponding IT systems. The calculated number of CDRs and the simulation results shall be a helpful input to calculate the required database space for pairing / correlating these partial CDRs as well.

References

[1] 3GPP, "Charging Data Record (CDR) transfer," 3rd Generation Partnership Project (3GPP), TS 32.295, October 2009.

[2] ——, "Charging Data Record (CDR) file format and transfer," 3rd Generation Partnership Project (3GPP), TS 32.297, June 2009.

[3] ——, "Charging architecture and principles," 3rd Generation Partnership Project (3GPP), TS 32.240, December 2008.

[4] ——, "Charging Data Record (CDR) parameter description," 3rd Generation Partnership Project (3GPP), TS 32.298, December 2009.

[5] Y. Cai and C. G. Xu, "System and method for communicating charging data records," U. S. Patent 20 080 082 455, April 03, 2008.

[6] Y. Cai, "Charging for long duration sessions in communication networks," U. S. Patent 20 090 063 315, March 05, 2009.

[7] Y. Zou and B. Jia, "A new method for cdr processing in ip multimedia subsystem," *Business and Information Management, International Seminar on*, vol. 2, pp. 221–224, 2008.

[8] 3GPP, "IP Multimedia Subsystem (IMS) charging," 3rd Generation Partnership Project (3GPP), TS 32.260, December 2009.

[9] ——, "Packet Switched (PS) domain charging," 3rd Generation Partnership Project (3GPP), TS 32.251, December 2009.

[10] J. Guo, F. Liu, and Z. Zhu, "Estimate the call duration distribution parameters in gsm system based on k-l divergence method," *International Conference on Wireless Communications, Networking and Mobile Computing, WiCom*, pp. 2988–2991, September 2007.