BANDWIDTH RESERVATION POLICY PERFORMANCE ANALYSIS IN A WIRELESS CELLULAR NETWORK UNDER NON-EXPONENTIAL DISTRIBUTIONS

R. Nandhini
School of Computing Science Engineering, VIT University, Chennai campus, Chennai, India
E-Mail: nandhini.rohini@gmail.com

ABSTRACT
Past studies in bandwidth reservation strategies in wireless multimedia cellular networks use the exponential distribution for service time and inter-arrival time distributions. But in real world (example like GSM/GPRS), inter-arrival, service time and cell residence time in cellular systems have been shown to be non-exponential. As a result, network performance could be worse than that reported in the literature. The main objective of this paper is to investigate the impact of the network performance under various non-exponential distributions on a specific bandwidth reservation policy. The threshold-based bandwidth reservation policy has been taken into consideration and its performance is observed under different inter-arrival time distributions, and channel holding time distributions. The network performance is measured in terms of new call blocking probability and handoff call dropping probability.

Keywords: threshold-based bandwidth reservation, channel holding time, blocking probability, dropping probability.

INTRODUCTION

Background
Wireless cellular network infrastructures consist of a wired backbone and a number of Base Stations (BSs) or Access Points (APs). The geographical area controlled by a BS is called a cell. Several base stations are connected to a Mobile Switching Center, (MSC) that acts as a gateway from the cellular network to existing wireline networks. The user, while staying in a cell, communicates with another user, who may be in the same cell or may be in other cell through the BS in the same cell. When mobile moves into an adjacent cell in the middle of a communication session, a hand-off will enable the mobile to maintain connectivity to its communication partner, i.e., the mobile will start to communicate through the new BS, without noticing any difference. Figure-1 depicts the architecture of a wireless cellular network.

The next generation of Wireless Cellular Networks (WCNs) (e.g., Universal Mobile Telecommunication System - UMTS) is expected to support real-time multimedia applications with different classes of traffic (data as well as voice) and diverse bandwidth and QoS requirements. Nowadays the demand for broadband multimedia communication involving digital audio and video has increased. The increasing demand for mobile communications services will soon require the addition of multimedia access for their users. Providing multimedia services with Quality of Service (QoS) guarantees in WCNs presents great challenges due to the limited bandwidth and the high rate of handoff events. The mobiles require different amounts of bandwidth, depending on the nature of the applications that are running. Since bandwidth is a scarce resource in wireless networking, it is necessary to allocate it carefully amongst competing connections. The users expect good QoS from the system, e.g., low delay, low call-dropping and blocking probabilities.

As mentioned earlier a connection level QoS are expressed in terms of New Call Blocking Probability (NCBP) and Handoff Call Dropping Probability (HCDP). Other metrics like probability of successful call completion, probability of unsuccessful call completion can be derived from above. A new call results when a user requests for a new connection. A handoff call occurs when the user moves one cell to another during the session. NCBP is the probability of a new arriving call being blocked while HCDP is the probability of an ongoing call is forced to terminate before the completion of its service. The probability of a handoff failure is an important criterion in performance evaluation of cellular networks. Figure-2 represents the occurrence of a handoff event.
The objectives of this work are three-fold.

- To identify various non-exponential distributions proposed for inter-arrival time, and channel holding time in a wireless multimedia network.
- To study a threshold-based bandwidth reservation policy that improves the handoff call drop probability.
- To investigate the impacts of various non-exponential distributions on the network performance under the threshold-based bandwidth reservation policy by means of simulation.

LITERATURE REVIEW

Threshold-based bandwidth reservation policy

In wireless multimedia cellular networks handoff calls have a great impact on the overall network performance. The existing bandwidth reservation policies addressing the impact of various reservation schemes which reduce the dropping probability. The effects of having a variable number of reserved channels being adaptive, user mobility behavior, and location have been evaluated by Miquel Oliver [22]. They provide advantages of a dynamic reservation scheme over a fixed reservation scheme in terms of call blocking probability.

Various types of reservation strategies are examined by Brocha Epstein, et al. [2]. Their resource allocation methods are based on complete sharing; complete partitioning under various arrival rates of traffic and load conditions. In this paper, the cost measure is derived, which gives easy comparison of different policies. The resource allocation scheme based on the max-min fairness protocol by using bandwidth borrowing to lower NCBP and HCDP is considered in [3]. Multiple types of handoff prioritizing schemes are given in [4] such as call admission control channel reservation schemes [36], guard channel scheme etc. Different queuing disciplines are also suggested there.

The traffic arriving into cell considered as either homogeneous or heterogeneous. For simplification of model construction homogeneous traffic is considered in which all cells have the same mobility and traffic conditions. Two types of calls such as new calls and handoff calls are assumed as traffic coming to the cell. Fixed Channel Allocation (FCA) is most commonly used allocation scheme.

In the threshold-based bandwidth reservation policy, the total bandwidth of the cell is shared by both new calls and handoff calls. Calls arriving at the cell are partitioned into K classes based on bandwidth requirements. The bandwidth of a call is nothing but number of basic bandwidth units (bbu) that is adequate for guaranteeing desired QoS for a call with certain traffic characteristics. The bandwidth of a class-i, i.e., the number of basic bandwidth units required to accommodate the call, is given by bi. The classes are indexed in an increasing order according to their bandwidth requirements,

\[ b_1 \leq \ldots \leq b_i \leq b_{i-1} \leq \ldots \leq b_K \]  

The main idea of this bandwidth reservation policy is based on reserving bandwidth for aggregate handoff connections, thus giving them a higher priority over new connections and lower HCDP. In addition, the policy prioritizes between different classes of handoff connections.
connections according to their QoS constraints by assigning a series of bandwidth thresholds:

\[ t_0, t_1, \ldots, t_K, \text{ where } t_0 \leq \ldots \leq t_i \leq t_{i+1} \leq \ldots \leq t_K \]  

(2)

where, \( t_0 \) denotes the maximum number of total bbu’s that can be allocated to new connections, and \( t_i, 1 \leq i \leq K \), denotes the maximum number of total bbu’s that can be allocated to class-\( i \) handoff connections. However, if many handoff connections were allowed to completely share the bandwidth, then connections require lower bandwidth will have a better chance of occupying the bandwidth than those with higher bandwidth requirements.

As soon as the call is terminated, or the call moves out of this cell, the neighbors are informed to remove the amount of bandwidth that was reserved for this particular call. A call that requests a large chunk of bandwidth is more likely to be blocked. It may be better to reject a number of connections at initialization rather than have to drop them at a later stage.

Some of the bandwidth reservation strategies implements queue for incoming calls. Call arriving to the cell is not admitted immediately the can be buffered provided very minimum waiting time [14]. The simulated model have buffer for call arrivals. They use buffer for the following reason:

a. Data traffic is generally more tolerant to delay than voice traffic.

b. Buffering can effectively mitigate the variability in the data call arrival process.

c. Buffering data calls temporarily rather than immediately blocking them provides these calls a better opportunity to enter the system later.

More details of the considered threshold-based bandwidth reservation policy will be discussed later.

Inter-arrival time distributions

Although many existing works use the exponential distribution to model inter-arrival time and call holding time in cellular networks [9] some studies on GPRS/GSM networks proposed the use of non-exponential distributions to better model the characteristics of data and multimedia traffic. Ming Zhang, et al [11] develops a system model with shared voice and packet data channel. When they are considering voice traffic, they suggested Poisson arrival with inter-arrival and exponential call holding times. However, the observation from Common Channel Signaling [17] shows that if the cell traffic is smooth then inter-arrival time cannot be modeled as Poisson process.

The performance of GPRS data source model has been observed under different probability distributions [8]. The time between downloading two consecutive web page requests have been analyzed under triangular, uniform, truncated exponential and truncated gamma distributions and the results were compared. The triangular distribution introduces the highest level of delay, while the gamma and the exponential distribution caused the smallest delays.

The uniform distribution lies between the two. Their simulation results show there are only small load differences between applying Gamma and exponential distributions. The triangular one generated slightly heavier load and uniform distribution lies in between them. UMTS recommendation [23] is to use an exponential distribution for the inter-arrival time for requests for consecutive web pages. As the web browsing process is similar in nature to the call hand-off process and because of the small differences between the investigated distributions, they propose the Gamma distribution because of its good statistics.

To meet the requirement of wireless networks which carries multimedia traffic (voice, video, data, and image), it becomes necessary to provide efficient and better bandwidth reservation schemes [24]. Dynamic-grouping bandwidth reservation scheme discussed in this paper presents a geometric arrival to reduce the connection blocking rate and connection dropping rate, while increasing the bandwidth utilization. The simulation result show that less connection-blocking rate and less connection-dropping rate and achieves high bandwidth utilization.

In [5], new call arrivals and handoff call arrivals are assumed to follow heavy tailed Weibull distribution. The accuracy of network dimensioning with respect the distributions investigated Pareto distributed inter-arrival time of traffic [7]. Several works mentioned in [10] shows that inter-arrival time is modeled by heavy-tailed distributions like log-normal, Weibull and Pareto.

Channel holding time distributions

The channel holding-time distribution describes the distribution of the time spent by a mobile subscriber making use of the resources (channels) within a cell. The channel holding time is the minimum of the call holding time and the cell residence time. The cell residence time is the amount of time during which a mobile terminal stays in a single cell during a single visit. While it is possible to directly obtain the minimum of two separate distributions, the analysis is cumbersome and intractable. Therefore, the channel holding time is typically modeled as a single distribution by fitting field data.

The channel holding time in cellular systems depends on many factors such as the mobility of the customers, speed, cell size, the geographic situations, and the channel allocation schemes [17]. Previous analyses used exponential distributions to model channel holding time [13], [28]. But experimental data showed that actual channel occupancy distributions are significantly different from exponential distributions used in these analyses [32]. Based on simulations, Guerin [34] showed that for some cases the channel occupancy time distribution is quite close to exponential distribution but for the low rate of change of direction the channel occupancy time distribution shows rather poor agreement with the exponential distribution. In [15-20], the channel holding distribution has been modeled as the exponential distribution, the lognormal distribution, the (mixed) Erlang distribution, and the (generalized) Gamma distribution.
Even though the exponential and Erlang distributions have good properties for queuing analysis, they are not enough to fit the field data. The (generalized) Gamma [7], [33] and log-normal distributions [16] are shown to be more appropriate. User mobility modeling and characterization of mobility patterns, by Zonoozi and Dassanayake [33] used the generalized Gamma distribution to model the cell residence time. For instance, series of experiments conducted by Barceló et al. [29-32] for mobile radio and cellular systems concluded that channel holding times and related time variables are not exponentially distributed. They further showed that the lognormal distribution and the mixture of Erlang distributions provided better statistical fitting to the experimental data.

To model the channel holding time for cellular systems with mixed platforms and various mobility, the sum of Hyper-exponential (SOHYP) distributions has been suggested [21]. Phase-type distributions [12], [27] of Generalized Erlang form used to model channel holding time in a mobile environment. However, the complexity of the analysis has increased considerably with these techniques. The hyper-Erlang distribution is proposed in [6],[15] for the channel holding time to maintain tractable queuing analysis while providing good fit to field data.

PROPOSED WORK

Simulation model description

The bandwidth reservation policy can be modeled as a multidimensional Markov chain in which each chain is modeled as M/M/∞ system. In order to evaluate the performance effectively I use the same simulation model described in [1]. The system uses Fixed Channel Allocation (FCA) which means the cell has fixed amount of channel capacity. Call arrivals entered in to the cell can be new calls or handoff calls. So the total bandwidth of the cell is shared by both new and handoff calls. The call arrivals are generated by equal probability \( \Lambda_{nc1} = \Lambda_{nc2} = \Lambda_{hc1} = \Lambda_{hc2} = \Lambda \). The basic assumption of simulation model is given below:

a) Bandwidth is determined at call initialization, and is fixed for the duration of the call.

b) Arriving connection (new or handoff) that is not admitted immediately is blocked or dropped, i.e., a call is never buffered.

c) Assume that the traffic offered to the cellular network belongs to one of two classes:

Class 1-real-time multimedia traffic, such as interactive audio and video (voice traffic)
Class 2-non-real-time data traffic, such as email and web applications (data traffic)

The minimum bandwidth for required to accommodate each class of calls is denoted by \( b_i \). Since two classes of calls are considered, the bandwidth of class 1 is represented as \( b_1 \) and bandwidth of class 2 calls is represented as \( b_2 \). \( b_1 \leq b_2 \)

The main idea of threshold reservation policy is to assigning series of threshold for each class of calls in order to prioritize the calls. The threshold value assigned as mentioned in reference paper [1]. If the admission of a new call exceeds this threshold value, the call is blocked.

Threshold bandwidth for new calls is denoted by \( t_0 \), \( t_1 \) and assumed to be 15 bbu’s, \( t_2 \) is the maximum number of total basic bandwidth units that can be allocated to class-1 handoff connections and \( t_3 \) is the maximum number of total basic bandwidth units that can be allocated to class-2 handoff connections. In simulation \( t_2 \) and \( t_3 \) are assumed to be equal as 30 bbu’s. Figure-3 depicts the accessible bandwidth for 2-class along with thresholds.

\[
\text{B-Accessible Bandwidth}
\]

Class 1 and Class 2 Handoff calls

\[
\text{Class 1 and 2 new calls}
\]

Figure-3. Accessible bandwidth system for 2-class system with its threshold.
Simulation input and output

The simulation model investigates the influence of four different distributions that can be applied in a wireless multimedia cellular network. The distributions used for arrival and call holding times have been discussed in literature section. The experimental set up consists of two different scenarios. The first scenario is to fix the inter-arrival time distribution to exponential and vary the channel holding time distribution to log-normal [25], [26], mixed Erlang, generalized Gamma [16], sum of Hyper-Exponential [21], [35], and Pareto [35].

The second scenario is to fix the channel holding time distribution to the one that give worst performance from the above scenario and vary the inter-arrival time distribution to Pareto [35], Weibull [5], gamma [8] and log-normal [7].

The performance of the cellular network has been studied under these two scenarios. A pair of inter-arrival and channel holding time distributions will be obtained and its performance will be compared to the case of exponential inter-arrival and channel holding time distributions in terms of call blocking probability and call dropping probability.

Simulation software

Our simulation model is implemented in OMNet++. OMNeT++ is an object-oriented modular discrete event network simulator. An OMNeT++ simulation model consists of hierarchically nested modules. Modules communicate through message passing. Modules at the lowest level of the module hierarchy encapsulate behavior. These modules are termed simple modules, and they are programmed in C++ using the simulation library. OMNeT++ models are often referred to as networks. Model structure is described in OMNeT++’s NED language.

Model flow

The simulation model consists of 3 modules (C++ Objects) HostCell, OtherCell and CallProcessingServer. The HostCell and OtherCell send the new calls and handoff calls based on the specified distribution with arrival rate $\Lambda$ to the CallProcessingServer. The CallProcessingServer receives the calls and allocates bandwidth depending on the type of call (voice or data) and also with respect to restricted threshold ($t_0$, $t_1$, $t_2$ or $t_3$). Each call will reside in the CallProcessingServer according to the delay specified in Channel holding time distribution. The CallProcessingServer will register a timeout call for the delay specified. During the timeout call back the Call ProcessingServer will de-allocate the bandwidth. Figure-4a depicts the flow diagram of simulation model.
The Call Processing Server module consists of 3 C++ objects Call Processing Server, Bandwidth Manager and Bandwidth Threshold. During startup the Call Processing Server creates an instance of Bandwidth Manager object and reads the threshold details ($t_0$, $t_1$, $t_2$ and $t_3$) from the OMNet++ ini file and for each threshold ($t_0$, $t_1$, $t_2$ and $t_3$) a Bandwidth Threshold object is created and stored in a list. When a call (either new call or handoff call) is received the Call Processing Server calls the Bandwidth Manager function to allocate bandwidth for the call. The Bandwidth Manager then identifies the type of call and searches the list of threshold and selects the appropriate threshold object (Bandwidth Threshold). It then calls the Bandwidth Threshold function to allocate the bandwidth. Figure-4b represents the sequence diagram of call processing server.

After the call is allocated the Call Processing Server schedules a timer call back for the delay specified by the Channel holding time distribution. During this time timer call back the Call Processing Server calls the Bandwidth Manager function to free the bandwidth. The Bandwidth Manager in turn calls the appropriate Bandwidth Threshold function to de-allocate the bandwidth.

SIMULATION RESULTS

Simulation Parameters

We consider a cellular network in which the cell has a total capacity of 30 bbu (B) [1]. It is assumed that each data service (b1) requires 3 bbu and voice service (b2) requires only 1 bbu. QoS metrics such as New Call Blocking Probability, Handoff Call Dropping Probability are evaluated. Since the distribution used for call arrival rate, channel holding time effect these metrics, two sets of numerical results are shown under the restricted threshold. The restriction threshold ranges from 1 to 30 in each analysis. New calls threshold limited to 15 units of bandwidth ($t_0$, $t_1$). Remaining 15 units of bandwidth is completely shared by both classes of handoff calls ($t_2$, $t_3$). As for traffic characterization, new call arrivals and handoff call arrivals of class-i connections are assumed to follow exponential distribution with rates $\Lambda_{nci}$ and $\Lambda_{hci}$, respectively. The total numbers of calls generated from each cell are 10 million.

Effects of channel-holding time distributions

Figures-5 shows the effect of varying Exponential call arrival rate on the New Call Blocking Probability (NCBP) and the Handoff Call Dropping Probability (HCDP) for the following system parameters under different Channel Holding Time distribution such as Pareto, Gamma, Hyper–exponential and lognormal. The variance of the channel holding time distribution has been fixed as same as exponential variance. The entire channel holding time distributions has been set as their mean =120 sec and variance =14706 sec. Since Erlang distribution behavior is same as exponential distribution, the Erlang has not taken into consideration. Pareto, Gamma and Hyper-exponential show similar blocking and dropping probability. Lognormal shows higher degree of blocking and dropping probability. The
observation reveals that the performance of wireless cellular network will be worse in case of lognormal CHT distribution. I observe that the NCBP, HCDP of both classes increases as the call arrival rate increases. However, the HCDP is always lower than the NCBP as result of the 15 bandwidth units (B-t0) reserved exclusively for handoff connections. Moreover, the HCDP of class-2 connections is higher than that of class-1 connections. This is due to complete bandwidth sharing between class-1 and class-2 connections (t2=t3=30) which results in a higher dropping probability for higher bandwidth class. The input parameters of the various channel holding time distributions used in simulation run is represented by Table-1.

### Table-1. Input parameters of the various channel holding time distributions.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>API</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>Exponential(Mean)</td>
<td>Mean=120, ( \lambda = \frac{1}{120} )</td>
</tr>
<tr>
<td>Gamma</td>
<td>Gamma(( \alpha, \theta ))</td>
<td>( \alpha = 0.9792 ), ( \theta = 122.549 )</td>
</tr>
<tr>
<td>Pareto</td>
<td>Pareto(k, ( X_m ))</td>
<td>k=2.4068, ( X_m = 70.1421 )</td>
</tr>
<tr>
<td>Lognormal</td>
<td>Lognormal(( \mu, \sigma ))</td>
<td>( \mu = 0.8 ), ( \sigma = 4.5328 )</td>
</tr>
<tr>
<td>Hyper-Exponential</td>
<td>Hyper-Exponential(( \mu, \sigma ))</td>
<td>( \mu = 120 ), ( \sigma = 121.26 )</td>
</tr>
</tbody>
</table>

### Effects of inter-arrival time distributions

As we mentioned in proposed section, the worst performance Channel Holding time distribution Lognormal from section above have been selected and II phase of experiment has been continued with various inter arrival time distribution such as Weibull, Pareto, and Gamma.

The effect of varying the call arrival rate of different distributions such as Gamma, Pareto, and Weibull on the New Call Blocking Probability and the Handoff Call Dropping Probability for the simulation parameters under lognormal CHT is represented by Figure-6. I observe that the NCBP, HCDP of both classes increases as the call arrival rate increases. However, the HCDP is always lower than the NCBP as result of the bandwidth units (B-t0) reserved exclusively for the handoff connections. From figure 6, Gamma shows higher Blocking and Dropping probability than Weibull and Pareto. Pareto and Weibull have same range of blocking and dropping probability. The dropping probability of Weibull, Pareto distribution was initially 0. However when arrival rate increases dropping probability also increases. The observation reveals that the network performance will be worse in case of Gamma inter-arrival time distribution.

Figure-5. Performance under exponential inter-arrival time distribution.
Performance comparison

The worst performance channel holding time distribution from the scenario 1 (Lognormal) and worst performance inter arrival time distribution (Gamma) from scenario 2 has been selected and new experiment has been made with this pair of distribution. This new pair distribution’s performance has been compared with Exponential arrival-Exponential channel holding time distribution.

Figure-7 shows the significant difference between Lognormal CHT and Exponential CHT. The new experiment with pair (Gamma arrival-Lognormal CHT) shows better performance than classical assumption (Exponential-Exponential).

CONCLUSIONS

In this paper four different probability distributions are applied for inter-arrival time and Channel holding time to investigate the performance of wireless cellular network in terms of blocking and dropping probability. A simulation model was built and used for the tests. Our simulation result shows the worst performance channel holding time distribution as lognormal and worst inter-arrival time distribution as Gamma. The new experiment has been made with the worst pair distributions. This worst pair distribution has been compared with classical assumption (exponential arrival with exponential channel holding time). It is important to notice that worst pair distribution (gamma inter-arrival-lognormal CHT distribution) exhibit better performance than exponential assumption.

REFERENCES


