ABSTRACT
The continuation of an active call is one of the most important quality measurements in cellular systems. Handoff process enables a cellular system to provide such a facility by transferring an active call from one cell to another. Different approaches are proposed and applied in order to achieve better handoff service. The principal parameters used to evaluate handoff techniques are: Forced termination probability and call blocking probability. Mechanisms such as guard channels and queuing handoff calls decrease the forced termination probability while increasing the call blocking probability. In this paper we present an overview about the issues related to handoff initiation and decision and discuss about different types of handoff techniques available in the literature.

KEYWORDS Handoff, Forced Termination probability, Blocking probability, Handoff Initiation, Handoff Decision, Handoff Prioritization Schemes.

I. INTRODUCTION
CELLULAR systems deploy smaller cells in order to achieve high system capacity due to the limited spectrum. The frequency band is divided into smaller bands and those bands are reused in non interfering cells [1-3]. Smaller cells cause an active mobile station (MS) to cross several cells during an ongoing conversation. This active call should be transferred from one cell to another one in order to achieve call continuation during boundary crossings. The handoff (or handover) process is transferring an active call from one cell to another. The transfer of a current communication channel could be in terms of a time slot, frequency band, or a code word to a new base station (BS) [1-5]. If a new BS has some unoccupied channels than it assigns one of them to the handed off call. However, if all of the channels are in use at the handoff time there are two possibilities: To drop the call or to delay it for a while. Different handoff techniques are proposed in literature and two of the most important metrics for evaluating a handoff technique are forced termination probability and call blocking probability. The forced termination probability is the probability of dropping an active call due to handoff failure and the call blocking probability is probability of blocking a new call request [2, 6-7]. The aim of a handoff procedure is to decrease forced termination probability while not increasing call blocking probability significantly.

II. HANDOFF INITIATION
Handoff initiation is the process of deciding when to request a handoff. Handoff decision is based on the received signal strengths (RSS) from the current BS and neighboring BSs. In Fig. 1, we examine the RSSs of the current BS (BS1) and one neighboring BS (BS2). The RSS gets weaker as the MS moves away from BS1 and gets stronger as it gets closer to BS2 as a result of signal propagation characteristics. The received signal is averaged over time using an averaging window to remove momentary fadings due to geographical and environmental factors [1-2]. Below, we will examine the four main handoff initiation techniques mentioned in [2-3]: Relative signal strength, relative signal strength with threshold, relative signal strength with hysteresis, and relative signal strength with hysteresis and threshold.

A. Relative Signal Strength
In relative signal strength, the RSSs are measured over time and the BS with strongest signal is chosen to handoff. In Fig.1, BS2’s RSS exceeds RSS of BS1 at point A and handoff is requested. Due to signal fluctuations, several handoffs may be requested while BS1’s RSS is still sufficient to serve the MS. These unnecessary handoffs are known as the ping-pong effect. As the number of handoffs increase, forced termination Probability and network load also increases. Therefore, handoff techniques should avoid unnecessary handoffs.

B. Relative Signal Strength with Threshold
Relative signal strength with threshold introduces a threshold value (T1 in Fig. 1) to overcome the ping-pong effect. The handoff is initiated if BS1’s RSS is lower than the threshold value and BS2’s RSS is stronger than BS1’s. The handoff request is issued at point B in Fig. 1.

C. Relative Signal Strength with Hysteresis
This technique uses a hysteresis value (h in Fig. 1) to initiate handoff. Handoff is requested when the BS2’s RSS exceeds the BS1’s RSS by the hysteresis value h (point C in Fig. 1).

D. Relative Signal Strength with Hysteresis and Threshold
The last technique combines both the threshold and hysteresis value concepts to come up with a technique with minimum number of handoffs. The handoff is requested when the BS1’s RSS is below the threshold (T1 in Fig. 1) and BS2’s RSS is stronger than BS1’s by the hysteresis value h (point C in Fig. 1). If we would choose a lower threshold than T1 (but higher than T2) than the handoff initiation
would be somewhere at the right of point C. All the techniques discussed above initiate handoff before point D, which is the “receiver threshold”. The receiver threshold is the minimum acceptable RSS for call continuation (T2 in Fig. 1) [1, 5]. If the RSS drops below the receiver threshold, the ongoing call is than dropped. The time interval between the handoff request and receiver threshold enable cellular systems to delay the handoff request until the receiver threshold time is reached when the neighboring cell does not have any empty channels. This technique is known as queuing handoff calls and will be discussed in Section V.

In [8], a handoff algorithm using multi-level thresholds is proposed which assigns different threshold values to the users according to their speed. Since low speed users spend more time in handoff zone they are assigned a higher threshold to distribute high and low speed users evenly. High speed users are assigned lower thresholds. The performance results obtained by [8] shows that an 8-level threshold algorithm operates better than a single threshold algorithm in terms of forced termination and call blocking probabilities. In [9] and [10], an improved threshold-based method is introduced and compared with the basic initiation techniques such as maximum power handoff (MPH or RSS), RSS with hysteresis, RSS with threshold, and combinations of hysteresis and threshold based methods in a ten-cell structure.

III. HANDOFF DECISION

In the previous section, we discussed the time in which a handoff is requested. In this section, we will examine the handoff decision protocols used in various cellular systems.

A. Network Controlled Handoff (NCHO)

NCHO is used in first generation cellular systems such as Advanced Mobile Phone System (AMPS) where the mobile telephone switching office (MTSO) is responsible for the overall handoff decision [11]. In NCHO, the network handles the necessary RSS measurements and handoff decision. The handoff execution time is on the order of many seconds because of the high network load [12].

B. Mobile Assisted Handoff (MAHO)

In NCHO, the load of the network is high since the network handles all of the processes itself. In order to reduce the load of the network, the MS is responsible for making RSS measurements and sending them periodically to BS in MAHO. Based on the received measurements, the BS or the mobile switching center (MSC) decides when to handoff [3-4]. MAHO is used in the Global System for Mobile Communications (GSM). The handoff execution time is about 1 sec [4, 12].

C. Mobile Controlled Handoff (MCHO)

MCHO extends the role of the MS by giving overall control to it. Both, MS and BS, make the necessary measurements, and the BS sends them to the MS [3]. Then, the MS decides when to handoff based on the information gained from the BS and itself. Digital European Cordless Telephone (DECT) is a sample cellular system using MCHO with 100-500 ms handoff execution time [4, 12].

IV. HANDOFF TYPES

In this section we will mention the different types of handoffs. First, we will concentrate on channel usage. Then, we will investigate handoff in microcells and multilayered systems. Finally, we will explain handoff in homogeneous and heterogeneous systems.

A. Hard vs. Soft Handoff

The hard handoff term is used when the communication channel is released first and the new channel is acquired later from the neighboring cell. Thus, there is a service interruption when the handoff occurs reducing the quality of service. Hard handoff is used by the systems which use time division multiple access (TDMA) and frequency division multiple access (FDMA) such as GSM and General Packet Radio Service (GPRS) [13]. In contrast to hard handoff, a soft handoff can establish multiple connections with neighboring cells. Soft handoff is used by the code division multiple access (CDMA) systems where the cells use same frequency band using different code words. Each MS maintains an active set where BSs are added when the RSS exceeds a given threshold and removed when RSS drops below another threshold value for a given amount of time specified by a timer. When a presence or absence of a BS to the active set is encountered soft handoff occurs. The sample systems using soft handoff are Interim Standard 95 (IS-95) and Wideband CDMA (WCDMA) [2, 4, 13]. Bruise and Handling [14] proposed a handoff algorithm based on neighboring cells capacity instead of using the strongest RSS. The call or data connection is handed off to the cell with lower capacity and whose RSS is higher than a specified threshold.

B. Microcellular vs. Multilayer Handoff

In this section, we will first look at the handoff issues in microcellular environments. Later, we will investigate some systems that use microcells overlaid by macro cells in order to minimize number of handoffs.

1) Microcellular Handoff

The microcells are cells with small radii and employed in highly populated areas such as city buildings and streets to meet high system capacity by frequency reuse. In Fig. 2, we have two streets intersecting with three BSs employed on the streets. BS1 and BS3 have line-of-sight (LOS) with each other. The handoff between BS1 and BS3 is called LOS handoff; on the other hand the handoff between BS1 and BS2 is a non-LOS (NLOS) handoff since they don’t have LOS [2, 4, 13]. In NLOS handoffs, when a MS lose LOS (by turning the corner) with current BS, a drop in RSS (20-30 dB) occurs [4, 13]. This effect is called corner effect and needs faster handoff algorithms since the RSS can drop quickly below the receiver threshold resulting in a call drop. Two types of handoffs, LOS and NLOS, have different characteristics where LOS handoffs try to minimize the number of unnecessary handoffs between BSs and NLOS must be as quickly as possible because of the corner effect. In [13], a fast handoff algorithm for hard handoffs is proposed to remove fast fading fluctuations resulting in algorithm that reacts more quickly to corner effect.
They propose a technique called local averaging, in which the averaging time interval is smaller than averaging time interval of common handoff algorithms and improve handoff performance. The authors proposed an improved version of the algorithm by adding a drop timer to local averaging technique which decreases the unnecessary handoffs [15]. Then, they compare their proposal with a common averaging technique which uses an exponential window. A direction biased algorithm is proposed in [16] where all the BSs in handoff decision are grouped in two groups. One set of BSs are those in which MS is approaching and the other set includes the BSs in which the MS moves away. In handoff initiation an encouraging hysteresis (he) is used to first group where a discouraging hysteresis (hd) is applied to the second one. The relation between these hysteresis values are \( he \leq h \leq hd \). A signal strength based direction estimation method is used for determining the mobile positions.

2) Multilayer Handoff

Some designs used a multilayer approach in order to decrease the number of handoffs and to increase system capacity. A number of microcells are overlaid by a macrocell and the users are assigned to each layer according to their speeds. The microcells and macrocells coverage area are about 500 meters and 35 km, respectively for GSM900 in [17]. Since slow users are assigned to the microcells, the total number of handoff requests is decreased. Macrocells not only serve the fast users but also serve slow users when the microcells are congested. When a microcell allocates all of its channels, the new and handoff calls are overflowed to the macrocell layer. When the microcells load decreases it is possible to assign the slow users a channel in the microcell. This type of handoff is called take-back. So far, we have four types of handoffs: Microcell-to-microcell, microcell-to-macrocell, macrocell-to-microcell and macrocell-to-microcell [4]. In [18], a two-layer system is proposed for GSM phase 2 which uses microcells to increase the system capacity. The cell selection for fast and slow users is determined by a switching parameter and cell selection penalty. In [7], a bonus-based algorithm is proposed where it is compared with classical and macro algorithms. In the classical algorithm, in the case of new call request, a user is assigned to a microcell or overflowed to a macrocell if the capacity of the microcell is full. After the user speed estimation is done, the user is assigned to the appropriate layer using overflow and take-back. This scheme results in too many handoffs known as the ping-pong effect. The macro algorithm is similar to the classical algorithm with one exception. When a user is assigned to the macrocell it is not permitted to be taken-back to the microcell which decreases the number of handoffs. The bonus-based algorithm tries to prevent unnecessary handoffs to the microcell when fast users temporarily slow down. For each fast user, a time bonus is given and the user can use this time bonus during temporary slowdowns. If a user exceeds the timer, then it is assigned as slow user and is taken-back to the microcell layer. A speed-sensitive handoff algorithm was proposed by Vakili and Aziminejad [19] where slow users are assigned to microcells and fast users to the macro cells. The algorithm provides both overflow and take-back of a call when the MS with corresponding layer cannot find an unoccupied channel. For example, when a fast user requests a channel from a macrocell, macrocells hand-down slow user to the microcell if no free channel is available. Hu and Rappaport [20] also described and proposed a model for three-layer hierarchical network consisting of microcells, macrocells, and spot beams. Microcells and macrocells are terrestrial part of the network whereas spot beams correspond to satellite part. The users can be overflowed from lower layers to the upper layers but take back is not allowed here. In future systems, global coverage can be achieved using Hierarchical Cell Structure (HCS) where the HCS has picocells at the lowest layer for indoor communications [21] with higher data rates and the rest of the layers are as those described by [20]. Ekici and Ersoy [21] also present a probabilistic optimization technique using simulated annealing approach to determine the system parameters for achieving minimum system cost of a multi-tier cellular network. In [22], a cell selection for slow and fast users is introduced using a time offset. If a user resides in the cell at least offset time, then it is assigned as slow user. Otherwise, it is assigned as a fast user and sent to macrocell. The time offset is increased and the signal threshold for initiating the time offset is reduced in order to increase the efficiency of the cell selection mechanism.

C. Horizontal vs. Vertical Handoff

Handoff between homogenous networks where one type of network is considered is called horizontal handoff. On the other hand, handoff between different types of networks is also possible. A handoff in such a heterogeneous environment is named as vertical handoff and it is out of scope of this paper [23]. All the issues described in this paper are related to horizontal handoff.

V. PRIORITIZATION SCHEMES

In non-prioritization schemes new calls and handoff calls are treated the same way. When a BS has an idle channel, it is assigned due to first-come first-serve basis regardless of whether the call is new or handoff. But, forced termination of an active call is less desirable by the cellular users in contrast to new call blocking [1, 6, 24]. In order to provide lower forced termination probability, prioritization schemes
assigns more channels to the handoff calls. The two well-known prioritization schemes are: Guard Channels and queuing handoff calls [1, 3-4, 6]. A. Guard Channels The guard channel scheme reserves some fixed or adaptively changing number of channels for handoff calls only. The rest of the channels are used by new and handoff calls. So, the handoff calls are better served and forced termination probability is decreased. The costs of such a scheme are an increase in call blocking probability and a decrease in total carried traffic.

In [24] the number of guard channels is determined dynamically by the use of neighboring BSs. Each BS determines the number of MSSs in prehandover zone (PHZ) periodically and informs its neighbor BS related to that PHZ. PHZ is a small area located next to handoff zone and contains the possible users that will enter handoff zone in a short time. When the BS gets the number of MSSs in PHZ, it reserves that amount of guard channels for handoff calls. A new call is assigned a channel if no and off calls are queued in the queue where handoff calls are kept and the total number of free channels is greater than the number of guard channels. Zhang and Liu [25] proposed an adaptive algorithm which assigns the number of channels adaptively when forced termination probability exceeds a predefined threshold the guard channel number is increased to decrease the forced termination probability to a value below the threshold. The number of guard channels is decreased in the case where the BS does not use reserved guard channels significantly.

B. Queuing Handoff Calls

Queuing handoff calls prioritization scheme queues the handoff calls when all of the channels are occupied in a BS. When a channel is released, it is assigned to one of the handoff calls in the queue. A new call request is assigned a channel if the queue is empty and if there is at least one free channel in the BS. Also, some systems queue new calls to decrease call blocking probability [26]. The time interval between handoff initiation and receiver threshold makes it possible to use queuing handoff calls. Queuing handoff calls can be used with/without the guard channel scheme. In [3], a timer based handoff priority scheme is proposed. When a channel is released at BS, a timer is started. If a handoff request is done in that time interval it is assigned to it. Otherwise, when the timer expires, the channel can be assigned to new or handoff calls depending on the arrival order. Tekinay and Jabbari [6] introduced a new prioritization scheme called Measurement Based Prioritization Scheme (MBSP). The handoff calls are added to the queue and priorities of the calls changes dynamically based on the power level they have. The calls with power level close to the receiver threshold have the highest priorities. This scheme provides better results from the first-in-first-out (FIFO) queuing scheme where the handoff calls are served due to arrival time The Most Critical First (MCF) policy described in [24] determines the first handoff call that will be cut off and assigns the first released channel to that call. The first handoff call that will be cut off has the highest priority. The authors proposed a method to predict the first handoff call to be cut off by using simple radio measurements. In [26], a queuing scheme using guard channels is described. Both new calls and handoff calls are queued. A number of guard channels are reserved for handoff calls. When the new calls are congested, a channel from the guard channels is used if it is available. This scheme decreases the call blocking probability while increasing forced termination probability slightly. Salih and Fidanboylu [27], [28] described and modeled queuing techniques for two-tier cellular networks. In [27], a microcell/macrocell network using a FIFO queue in macrocell tier and in [28] a microcell/macrocell network using a FIFO queue in miccell tier is introduced and compared with each other. The results of both systems showed that forced termination probability for slow users is decreased when the FIFO queue is used in the microcell and forced termination for fast users is decreased when the queue is in the macrocell.

VI. CONCLUSION

In this paper, we introduced an overview on the concept of handoff and its evaluation parameters. We discussed the handoff initiation techniques based on the received signal strength and also the handoff decision protocols that are used. In addition, the handoff types based on channel usage, microcellular and multilayered systems and network characteristics are explained. Finally, we presented the handoff prioritization schemes to reduce the handoff call blocking probability, such as guard channels and queuing handoff calls.

REFERENCES

10. V. Tabataba Vakili and S. Shirvani Moghaddam, “Optimum Selection of Handoff Initiation Algorithm...