Chapter 3

Analytical Frame Work for Handoff
Chapter Overview

This chapter gives us detail information about the Performance evaluation metrics of handoff algorithms and its analytical framework. It also discusses various handoff schemes and handoff models.

3.1 Performance Evaluation Metrics of Handoff Algorithms

The main objectives of a handoff procedure are, first, to minimize the number of link transfers and second, to minimize the handoff processing delay by correct choice of target BS/AP with speedy execution. This minimizes the probability of connection interruptions and reduces the switching load. If the handoff is not fast enough, the quality of the service experiences degradations. A handoff should be evaluated as to its impact on the mobile to network connection [94].

The performance of handoff algorithms is quantitatively determined by the following metrics:

1. **Number of Handoffs**: It indicates the total handoff count as the mobile terminal moves between several overlapping BSs/APs. The result determines the sensitivity of the handoff algorithm. An excessively high rate indicates that the algorithm is over sensitive to metrics fluctuations, causing high rates of radio and network signaling load, and increasing the risk of disconnection. If handoffs are too few, but the mobile crosses the boundaries of coverage of a given BS/AP we will have intrusion in providing the service and possibly a connection loss.

2. **Ping-pong Handoffs**: These are handoffs during which the mobile connection is alternating between the target and initial BS/AP several times before establishing a stable link. The ping-pong handoffs over several overlapping BS/AP coverage areas unnecessarily utilize radio and network signaling resources as explained before.
3. **Number of Handoff Attempts**: It is the number of connection attempts between the mobile and a new base station before the establishment of a reliable link. The value of such an indicator has to be minimized.

4. **Blocking Probability** ($P_b$): When a mobile request for the call connection, its request may either grant or denied. The denial of request is due to unavailability of channels. This denial of request is called as blocking and its probability is called as blocking probability.

5. **Handoff Probability** ($P_h$): The handoff is called as successful handoff if a mobile changes its connection from one BS to another BS. The probability of such successful handoff is called as handoff probability.

6. **Dropping Probability** ($P_d$): During the call duration of a mobile it may travel through several BSs and may require several successful handoffs. Failure to get a successful handoff in the path forces the network to discontinue the call and is dropped. The probability of such event is called as dropping probability.

The motivation for hybrid wireless heterogeneous networks arises from the fact that no one technology or service can provide ubiquitous coverage, and it will be necessary for a mobile terminal to employ various points of attachment to maintain connectivity to the backbone network at all times. The natural trend has been toward utilizing local-coverage high-bandwidth data networks such as IEEE 802.11 WLAN whenever available, and to switch to an overlay service such as a GPRS network with lower bandwidth when a WLAN is not available. In the inter-technology context where handoffs are envisaged between heterogeneous packet switched networks, a handoff from a WLAN to a GPRS should be performed only with a very low priority, while a handoff from a GPRS to a WLAN should be performed whenever possible.

Even though same basic principles apply to the handoffs in inter-technology networks, there are several dynamic factors that must be emphasized and considered in handoff decisions for effective network usage.
For example, information on current network conditions can help load balancing across networks; current user conditions, such as a mobile terminal’s velocity can eliminate certain networks from consideration (i.e., those networks that do not support mobility). Available hints, like user activity patterns and network coverage maps can also contribute to handoff decisions.

One way of looking at the inter-technology handoff problem is to look at the system status at the handoff instant. System status includes observed throughput, how long the user has used each network, how many bytes the user has transferred on each network, the current battery status on the mobile device, and what has been the cost that the user has spent on each network.

User policies determine the “best” network and whether handoff should take place or not. Policy parameters include:

1. Cost of the connectivity
2. Power consumption of the used network interface
3. Setup time of the connection
4. User activity history
5. Speed of the user
6. Coverage maps of the network
7. Number of users in the networks

The inter-technology handoff is two-fold; handoff from underlay network, e.g., WLAN, to overlay network, e.g., GPRS, and handoff back from GPRS to WLAN.

3.2 Analytical Framework for Handoff

3.2.1 Introduction

Mobility is the most important feature of a wireless cellular communication system. Usually, continuous service is achieved by supporting handoff (or handover) from one cell to another. Handoff is the process of changing the channel (frequency, time slot, spreading code, or combination of them) associated with the current connection while a call is in progress. It is often initiated either by crossing a cell boundary or by deterioration in quality of the signal in the current channel [95]. Handoff is divided into
two broad categories—hard and soft handoffs. They are also characterized by “break before make” and “make before break.” Hard handoffs, current resources are released before new resources are used; in soft handoffs, both existing and new resources are used during the handoff process. Poorly designed handoff schemes tend to generate very heavy signaling traffic and, thereby, a dramatic decrease in quality of service (QoS). The reason why handoffs are critical in cellular communication systems is that neighboring cells are always using a disjoint subset of frequency bands, so negotiations must take place between the mobile station (MS), the current serving base station (BS), and the next potential BS. Other related issues, such as decision making and priority strategies during overloading, might influence the overall performance [96].

3.2.2 Handoff Schemes

In urban mobile cellular radio systems, especially when the cell size becomes relatively small, the handoff procedure has a significant impact on system performance. Blocking probability of originating calls and the forced termination probability of ongoing calls are the primary criteria for indicating performance. In this section, we describe traffic model and handoff schemes.

3.2.2.1 Traffic Model

For a mobile cellular radio system, it is important to establish a traffic model before analyzing the performance of the system. Several traffic models have been established based on different assumptions about user mobility. In the following subsection, we briefly introduce these traffic models.

To develop the analytical solution for the estimation of blocking probability and handoff probability we use a generic model of tele-traffic engineering [97].

Let

\[ \lambda \]  the arrival rate of new calls,

\[ \lambda_h \]  the arrival rate of handoff calls,
average channel holding time for new calls, 

\[ \frac{1}{\mu} \]

average channel holding time for handoff calls, 

\[ \frac{1}{\mu_h} \]

the total number of channels in cell.

It is assumed that the arrival process for new calls and arrival process for handoff calls are all Poisson, and the channel holding times for new calls and handoff calls are exponentially distributed.

Fig. 3.1 indicates the transition diagram for the call threshold scheme. Let \( K \) be the threshold for the new calls. This diagram arises from the two-dimensional Markov chain with the state space,

\[ S = \{(n_1, n_2), \quad 0 \leq \pi_a \leq K, \quad n_1 + n_2 \leq C \} \]  

(3.1)
Where \( n_x \) denotes the number of new calls initiated in one cell and \( n_2 \) is the number of handoff calls in that cell. Let \( q(n_x, n_2 : n_x, n_2) \) denote the probability transition rate from state \((n_x, n_2)\) to state \((n_x, n_2)\). Then we have

\[
q(n_x, n_2 : n_x - 1, n_2) = n_x \mu \quad (0 \leq n_x \leq K, 0 \leq n_2 \leq C)
\]

\[
q(n_x, n_2 : n_x, n_2 - 1) = n_x \mu_h \quad (0 \leq n_x \leq K, 0 \leq n_2 \leq C)
\]

\[
q(n_x, n_2 : n_x + 1, n_2) = n_x \lambda \quad (0 \leq n_x \leq K, 0 \leq n_2 \leq C)
\]

where \((n_x, n_2)\) is a feasible state in \( S \). Let \( p(n_x, n_2) \) denote the steady state probability that there are \( n_x \) new calls and \( n_2 \) handoff calls in the cell. Let \( \rho = \lambda / \mu \) and \( \rho_h = \lambda_h / \mu_h \). From the detailed balance equation, the following equation is obtained

\[
p(n_x, n_2) = \frac{\rho^{n_x}}{n_x!} \frac{\rho_{h}^{n_2}}{n_2!} p(0,0) \quad 0 \leq n_x \leq K, \quad n_x + n_2 \leq C, n_2 \geq 0
\]  

(3.6)

From the normalization equation, it is esteemed that

\[
p(0,0) = \left[ \sum_{0 \leq n_x \leq K, \quad n_x + n_2 \leq C} \frac{\rho^{n_x}}{n_x!} \frac{\rho_{h}^{n_2}}{n_2!} \right]^{-1}
\]

(3.7)

\[
= \left[ \sum_{n_x = 0}^{K} \frac{\rho^{n_x}}{n_x!} \sum_{n_2 = 0}^{C-n_x} \frac{\rho_{h}^{n_2}}{n_2!} \right]^{-1}
\]

(3.8)

From this, the formulae for new call blocking probability and handoff call blocking probability as follows:

\[
P_{nb} = \frac{\sum_{n_x = 0}^{K} \frac{\rho^{n_x}}{n_x!} \frac{\rho_{h}^{n_2}}{n_2!} + \sum_{n_x = 0}^{K} \frac{\rho^{n_x}}{n_x!} \frac{\rho_{h}^{n_2}}{n_2!} \sum_{n_x = 0}^{C-n_x} \frac{\rho_{h}^{n_2}}{n_2!}}{\sum_{n_x = 0}^{K} \frac{\rho^{n_x}}{n_x!} \sum_{n_2 = 0}^{C-n_x} \frac{\rho_{h}^{n_2}}{n_2!}}
\]

51
Similarly we derived the blocking probability and handoff probability for the multi-class traffic.

We used non-preemptive and preemptive priority handoff schemes for a multiple traffic system, such as an integrated voice and data system or integrated real-time and nonreal-time system. Although we focus our attention just on integrated voice and data systems, the results can be extended to other similar systems.

### 3.2.2.2 Non-preemptive Priority Handoff Scheme

We consider a system with many cells each having $S$ channels. As the system is assumed to have homogeneous cells, we focus our attention on a single cell called the marked cell. A system model is shown in Figure 4.2. In each BS, there are two queues, $Q_v$ and $Q_d$, with capacities $M_v$ and $M_d$ for voice and data handoff requests, respectively. Newly generated calls in the marked cell are called originating calls. For voice users, there is a handoff area. For data users, the boundary is defined as the locus of points where the average received signal strength of the two neighboring cells are equal.
The process of generation for handoff request is same as in previous schemes. A voice handoff request is queued in $Q_v$ on arrival if it finds no idle channels. On the other hand, a data handoff request is queued in $Q_d$ on arrival when it finds $(S - S_d)$ or fewer available channels, where $S_d$ is the number of usable channels for data handoff users. An originating voice or an originating data call is blocked on arrival if it finds $(S - S_d)$ or fewer available channels, where $S_c$ is the number of channels for both originating calls. No queue is assumed here for originating calls. A handoff request is blocked if its own queue is full on its arrival.

The blocking probability of an originating voice call or originating data call is

$$B_o = B_{ov} = B_{od} = \sum_{j=0}^{S_d} \sum_{i=S_c-j+1}^{S} \sum_{k=0}^{M_d} P(i, j, k)$$

(3.11)

The blocking probability $B_{hv}$ of a voice handoff request is

$$B_{hv} = \sum_{j=0}^{S_d} \sum_{k=1}^{M_h} P(S + M_v - j, j, k)$$

(3.12)

The blocking probability $B_{hd}$ of a data handoff request is
The voice user in a cell is given by

\[ P_h = \frac{\mu_{h-dwell}}{\mu_{CV} + \mu_{C-dwell}} \]  

(3.14)

The forced termination probability \( P_{fr} \) of voice calls can be expressed as

\[ P_{fr} = \sum_{i=m}^{\infty} P_{fr} P_{fr} \left[ (1 - P_{fr}) P_{fr} \right]^{-1} = \frac{P_{fr} P_{fr}}{1 - P_{fr} \left( 1 - P_{fr} \right)} \]  

(3.15)

### 3.2.2.3 Preemptive Priority Handoff Scheme

This scheme is a modification of a non-preemptive priority handoff scheme, with higher priorities for voice handoff request calls. In this scheme, a handoff request call is served if there are channels available when such a voice handoff request call arrives. Otherwise, the voice handoff request can preempt the data call, when we assume there is an ongoing data call, if on arrival it finds no idle channel. The interrupted data call is returned to the data queue \( Q_D \) and waits for a channel to be available based on the FIFO rule. A voice handoff request is queued in \( Q_V \) by the system if all the channels are occupied by prior calls and the data queue \( Q_D \) is full (i.e., data calls cannot be preempted by voice handoff calls when the data queue \( Q_D \) is full). It is possible to think of another scheme where data calls in service can be preempted by voice handoff calls irrespective of whether the queue \( Q_D \) is full or not. However, the same effect can be observed if the queue capacity is increased to a relatively large value.

The same state of the marked cell is assumed and represented by a three-tuple of nonnegative integers \((i, j, k)\) as defined in the non-preemptive priority handoff scheme. In the state transition diagram for the three-dimensional Markov chain model there are states.

\[ N_{r} = (S - S_d + 1)(S_d + 1)(M_D + 1) + (S_d + M_D + 1)M_D + S_d(S_d + 1)/2 \]
Therefore, as in the non-preemptive priority handoff scheme, we can get $N_r$ balance equations through the state transition diagram. Equilibrium probabilities $P(i, j, k)$ are related to each other through the state balance equations. However, note that any one of these balance equations can be obtained from other $N_r - 1$ equations. Since the sum of all state probabilities is equal to 1, we have

$$\sum_{j=0}^{S_d} \sum_{i=0}^{S_{-j}} P(i, j, 0) + \sum_{j=0}^{S_d} \sum_{i=S_{-j}}^{M_R} P(i, j, k) + \sum_{i=2}^{S+M_r} \sum_{k=0}^{M_R} P(i, 0, k) + \sum_{j=1}^{S_d} \sum_{i=S_{-j}+1}^{S+M_r-j} P(i, j, 0) = 1$$

(3.16)

The probabilities $P(i, j, k)$ (for $i = 0, 1, 2, \ldots, S + M_r; j = 0, 1, 2, \ldots, S_d$, and $k = 0, 1, 2, \ldots, M_d$) can be obtained by using the same method of computation in the non-preemptive priority handoff scheme.

The differences are:

$$N_D = \sum_{j=0}^{S_d} \sum_{i=0}^{S_{-j}} P(i, j, k) + \sum_{j=0}^{S_d} \sum_{i=S_{-j}+1}^{S+M_r-1} P(i, j, M_D) + \sum_{k=0}^{M_R} \sum_{j=0}^{S_d} \sum_{i=S_{-j}}^{S+M_r-j} P(i, j, k)$$

$$+ \sum_{k=1}^{M_D} \sum_{j=S_{-j}}^{S+M_r-1} P(i, 0, k) + M_D \sum_{j=0}^{S_d} \sum_{i=S_{-j}+1}^{S+M_r-j} P(i, j, M_D)$$

(3.17)

$$E[C_D] = \sum_{j=1}^{S_d} \sum_{i=0}^{S_{-j}} P(i, j, k) + \sum_{j=1}^{S_d} \sum_{i=1}^{S+M_r-1} P(i, j, M_D)$$

(3.18)

This approach can quickly give a preliminary idea about the performance of some handoff algorithms for simplified handoff scenarios. This approach is valid only under specified constraints. Before the appearance of the INTERNET, tele-traffic engineering had evolved around telephone communications. With the invention of the INTERNET and explosive advances in computer networks the network traffic characteristics have changed dramatically. Now the tele-traffic comprises of packetized voice, video, images
and computer data and this does not obey Poison model well. The present scenario may consist of dissimilar traffic on a full duplex channel. This behavior of the traffic is named as self similar or impulsive traffic. We have developed an analytical model for such traffic to evaluate the performance of handoff.

### 3.2.2.4 On/Off Models

In high speed networks, packets are communicated in a packet train fashion, once a packet train is triggered, the probability that another packet will follow is very large, furthermore the length of the packet train is heavy-tail distributed. This observation led to the celebrated On/Off model, which is sometime referred to as the alternating Fractal Renewal Process (AFRP) model.

The On/Off process can be represented by

\[ V(t) = \sum_{n=0}^{\infty} 1_{(s_n, s_{n+1})}(t), \quad t \geq 0 \]  \hspace{1cm} (3.19)

where \( S_n \) denotes the time of occurrence of the \( k^{th} \) on period, and

\[ S_n = S_0 + \sum_{j=1}^{n} T_j, \quad n \geq 1 \]  \hspace{1cm} (3.20)

The expected value of \( V(t) \) equals \( \mu_1 / (\mu_0 + \mu_i) \), and its power spectral density is

\[ S(w) = E[V(t)] \delta^2(w/2\pi) + \frac{2w^2}{\mu_0 + \mu_i} + \text{Re} \left[ \frac{Q_0(-jw)Q_1(-jw)}{[1 - Q_0(-jw)Q_1(-jw)]} \right] \]  \hspace{1cm} (3.21)

where \( Q_0(-jw), Q_1(-jw) \) are the Fourier transforms of \( f_o(t) \), and \( f_i(t) \) respectively.

The AFRP is a long range dependent process with Hurst parameter

\[ H = \frac{3 - \min(\alpha_0, \alpha_1)}{2} \]  \hspace{1cm} (3.22)
Mathematically, the EAFRP can be expressed as

\[
W(t) = \sum_{n=0}^{\infty} G_n 1_{\{X_n, X_{n+1}\}}(t)
\]  

(3.23)

where \( \{G(k)\}_{k \geq 0} \) is i.i.d. heavy-tail distributed with tail index \( \alpha \), and is independent of the On/Off intervals \( X_n, Y_n \).

### 3.2.2.5 Wavelet Model

Self-similarity of long-range dependence is closely related to the concept of time/frequency scaling. In this respect, wavelet analysis emerges as a natural framework. Wavelets are complete orthonormal bases that can be used to represent signals as a function of time. Let \( \{X_k\}_{k \in \mathbb{Z}} \) be a distance time process. Then \( X_k \) can be represented by the inverse wavelet transform

\[
X_k = \sum_{j=1}^{2k+1} \sum_{m=0}^{2^k-j-1} d_j^m \varphi_j^m(k) + \varphi_0
\]  

(3.24)

where \( 0 < k < 2^k - 1 \). The discrete wavelets \( \varphi_j^m(t) \) are defined as

\[
\varphi_j^m(t) = 2^{j/2} \varphi(2^{-j} t - m), \quad \varphi < t < 2^k - 1
\]  

(3.25)

where \( j \) and \( m \) are positive integers and \( \varphi(t) \) is the so called mother wavelet. It is clear that index \( j \) characterizes timescales, while \( m \) represents the time translation. The wavelet coefficients \( d_j^m \) can be obtained through the wavelet transform.

\[
d_j^m = \sum_{k=0}^{2^k-1} X_k \varphi_j^m(k)
\]  

(3.26)
A computationally efficient method for modeling heterogeneous network traffic based on wavelet analysis is proposed. To be specific we employ the simplest Haar wavelet as the mother wavelet, i.e.,

\[
\varphi(t) = \begin{cases} 
1 & \text{if } 0 < t < 0.5, \\
-1 & \text{if } 0.5 < t < 1, \\
0 & \text{otherwise.}
\end{cases}
\] (3.27)

When applying wavelet analysis to data traffic, although the original traffic has a complicated short-and long-range temporal dependence structure, the corresponding wavelet coefficients are only short-range dependent. Thus, a low-order Markov model would suffice for traffic modeling in wavelet domain. To capture the non-Gaussian behavior of broadband heterogeneous traffic, a scheme referred to as timescale-shaping algorithm is proposed. This employs training sequences (real-traffic trace) to shape the wavelet coefficients generated by Gaussian wavelet models, and thus match their empirical distributions.

Wavelet models can be implemented with low complexity. They are behavioral models, however, and as such their parameters are not linked to network parameters.

### 3.2.2.6 Cell Dwell Time

We choose the fluid flow model as the mobility model of mobile users. However, our proposed method can be easily used to other mobility models as well. The model assumes a uniform density of users throughout the area and also assumes that a user is equally likely to move in any direction with respect to the cell boundary. Let \( f_r(V) \) be the probability density function (pdf) of the random variable \( V \) of the speed of all mobile users in the following section.

We assume that both the real-time and non real-time service subscribers have the same probability distributions of moving speed. Let \( f_r(V) \) be the probability density function (pdf) of the random variable \( V \) with mean \( E[V] \) being the speed of mobile users. For two-dimensional fluid flow model, the average outgoing rate \( \mu_{\text{dwell}} \) is given by
where $L$ is the length of the perimeter of a cell with arbitrary shape and $A$ is the area of the cell. We assume that the cell dwell time $T_{dwell}$ has a random exponential distribution with mean $1/\mu_{dwell}$, then the average cell dwell time is given by

$$E[T_{dwell}] = \frac{\pi A}{E[V]L}$$  \hspace{1cm} (3.29)

### 3.2.2.7 Handoff Area Dwell Time

A real-time service handoff request calls is going to be put in the RHRQ if the BS in the target cell can allocate channel resource to it. If a channel is assigned to the request during its stay in the handoff area, the handoff is successful. If received signal strength reaches the receiver threshold prior to the assignment of a channel (that means mobile user moves out the handoff area before it gets channel resource), the call is forced to terminate. The maximum allowable waiting time of a real-time service handoff request in RHRQ is equal to the dwell time in the handoff area. In the following, we find the average dwell time of a handoff call in the handoff area.

We consider the new random variable $V^*$, which is the speed of mobile users crossing the cell's boundary, the pdf $f_{V^*}(V^*)$ is different from the above pdfs $f_V(V)$. $f_{V^*}(V^*)$ is given by

$$f_{V^*}(V^*) = \frac{V f_{V}(V)}{E[V]}$$  \hspace{1cm} (3.30)

where $E[V]$ is the average of the random variable $V$. 

\[ \text{Development of Handoff Algorithm for Third Generation Mobile Communication System} \]
The random variable $T_h$ is defined as the time spent in the handoff area of real-time service mobile user. It is given by

$$T_h = \frac{D}{V}.$$  \hfill (3.31)

where random variable $D$ is the length of moving path of the mobile users in handoff area. Therefore,

$$\frac{1}{E[V_{ht}]} = \int_0^z \frac{1}{V} f_{V_{ht}} (V) dV = \int_0^z \frac{f_{V_{ht}} (V)}{E[V_{ht}]} dV = \frac{1}{E[V_{ht}]}.$$  \hfill (3.32)

Assuming that the path length and velocity of mobile users are independent, we can get the average dwell time $E[T_h]$

$$E[T_h] = \frac{D}{E[V^*]} = \frac{1}{E[V^*]} E[D] = \frac{E[D]}{E[V^*]}.$$ \hfill (3.33)

The random variable $T_h$ is assumed to have an exponential distribution with the above average. Therefore, we use $\mu_{h,\text{dwell}}$ to present the leaving rate of mobile user of handoff area and the average value of $\mu_{h,\text{dwell}}$ is given by

$$E[\mu_{h,\text{dwell}}] = E \left[ \frac{1}{E[T_h]} \right].$$ \hfill (3.34)

### 3.2.2.8 Channel Holding Time

A channel held by a user will be released by either of the following reasons, that is, by the completion of the conversation or by the handing off the call to a neighboring cell (or failure of it, i.e., forced termination). Thus, the channel service is decomposed
into two rates [98], [99]. We assume that the call holding time $T_{CR}$ of real-time service calls has an exponential distribution with mean, 

$$E[T_{CR}] = \frac{1}{\mu_{CR}}.$$  \hspace{1cm} (3.35)

Therefore, the channel holding time $T_r$ of a real-time service call is equal to the smaller one between $T_{dwell}$ and $T_{CR}$. Since the random variable $T_{dwell}$ and $T_{CR}$ have exponential distributions with mean $1/m_{dwell}$ and $1/m_{CR}$, the average channel holding time of real-time service call can be given by the memory-less property of the exponential pdfs.

$$E[T_r] = \frac{1}{\mu_c} = \frac{1}{\mu_{CR} + \mu_{dwell}}.$$  \hspace{1cm} (3.36)

Similarly, we assume that the call holding time of $T_{CN}$ non real-time service calls has an exponential distribution with mean

$$E[T_{CN}] = \frac{1}{\mu_{CN}}.$$  \hspace{1cm} (3.37)

and we can get the average channel holding time of non real-time service call

$$E[T_n] = \frac{1}{\mu_c} = \frac{1}{\mu_{CN} + \mu_{dwell}}.$$  \hspace{1cm} (3.38)

and the channel holding time $T_r$ and $T_n$ are exponentially distributed with mean $E[T_r]$ and $E[T_n]$, respectively.

### 3.3 Concluding Remarks
The design of handoff scheme is an important consideration for the QoS in a wireless mobile network with integrated real time and non-real time services. A handoff scheme with priority reservation and preemptive priority procedure has been proposed in this chapter. An analytical model for the system performance has been presented. A blocking probability of originating calls, forced termination probability of the real time service calls, and averaged transmission delay of non-real time service has been calculated.