A Simple and Scalable Handoff Prioritization Scheme

Jörg Diederich [a,*]

[a] *Institute of Operating Systems and Computer Networks, Technische Universität Braunschweig, Mühlenpfordtstr. 23, D-38106 Braunschweig, Germany*

Martina Zitterbart [b]

[b] *Institute of Telematics, Universität Karlsruhe (T.H.), Postfach 6980, Zirkel 2, D-76128 Karlsruhe, Germany*

Abstract

Cellular networks, e.g., Universal Mobile Telecommunications System networks, will be based on the Internet Protocol (IP) to provide an efficient support for applications with bursty traffic characteristics, as, for example, issued by Web browsers or streaming applications. Such IP-based networks must include Quality of Service mechanisms to enable the usage of real-time applications, for example, mobile telephony. In wireless mobile networks this especially challenges handoff functions: handoffs should not lead to significant interruptions even though resource shortages after a handoff cannot be avoided completely.

To overcome this problem, a simple and scalable handoff prioritization scheme called *SiS-HoP* is proposed. It dynamically reserves resources for handoff purposes. This way, *SiS-HoP* limits the number of sessions, which are interrupted in case of a handoff, to less than 1%. This is achieved without compromising the resource utilization in scenarios with many non-mobile terminals. In contrast to existing schemes, *SiS-HoP* is well-suited for future IP-based mobile networks with small cells where a high number of handoffs may occur during the lifetime of a session.

Key words: Mobile networks, handoff, quality of service, handoff prioritization

* Corresponding author. Present address: L3S Research Center, Deutscher Pavillon, Expo Plaza 1, 30539 Hannover, Germany, Email: diederich@l3s.de. Tel.: +49 511 762 9749. Fax: +49 511 762 9779

Email addresses: dieder@ibr.cs.tu-bs.de (Jörg Diederich), zit@tm.uka.de (Martina Zitterbart).

1 Introduction

One difference of cellular mobile networks compared to fixed networks is that a mobile terminal can change its point of attachment to the network during an ongoing communication session. This phenomenon, known as a *handoff*, can lead to a resource shortage which means that the negotiated bandwidth for a session is no longer available after a handoff. In such a case of *handoff resource shortage*, the communication session must be terminated or an adaptation/re-negotiation with the application could take place. A termination of a session constitutes a major problem because the user of a mobile network, in general, expects a Quality of Service (QoS) enabled application to work over the lifetime of the communication session. An example for a handoff resource shortage in a cellular mobile network is shown in Figure 1. The

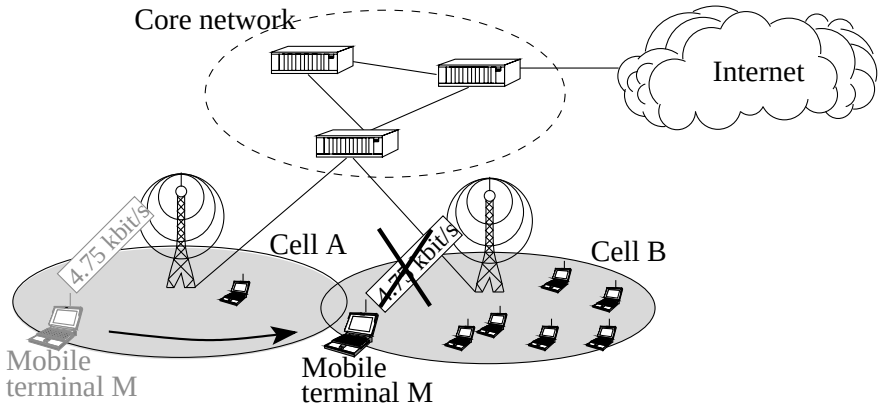

Fig. 1. Handoff resource shortage in a cellular mobile network

mobile terminal M has successfully requested a telephony session with a bandwidth of $4.75\,\mathrm{kbit}/s$ from the mobile network provider and initially resides in the lightly loaded cell A. If this mobile terminal M performs a handoff to the heavily loaded cell B, the network cannot continue to provide the negotiated bandwidth since the available bandwidth in cell B is lower than the currently used bandwidth.

To accommodate handoff resource shortages, a mobility-specific QoS parameter must be considered when providing QoS: the handoff success probability. Providing *assurances on the handoff success probability* is crucial, especially for future cellular mobile networks where the cell size is decreased to accommodate more mobile terminals in a given geographical area. In this case, the number of handoffs per session and, thus, the probability for a handoff resource shortage can become high even if the mobile terminal moves with only a moderate speed. As an example, the cell size in a densely populated area may be $700\,m$, e.g., in the downtown area of a city. If the mobile terminal resides within a vehicle, such as a car, a moderate speed is $17\,m/s$ (about $60\,km/h$). In this example, a mobile terminal performs on average 4.4 hand-

offs per session already in case of a session duration of three minutes (the average duration reported for telephony sessions [1]).

An appropriate solution for the problem of handoff resource shortage needs to comply with the following requirements in order to be applicable to future mobile networks [2–7]:

- Scalability, i.e., the ability to provide Quality of Service even if the number of mobile terminals or the number of handoffs varies over several orders of magnitude.
- Support for assurances on the handoff success probability.
- Easy administrability [8] with regard to the configuration of the necessary QoS components.
- Robustness against failure of components [9,10] or mis-configuration.
- Incremental deployment.
- Efficiency regarding the utilization of network resources (e.g., bandwidth which is typically a rather scarce resource in wireless mobile networks compared to fixed networks).

Handoff prioritization schemes are intended to solve the problem of handoff resource shortages [11–13]. They reserve a certain amount of resources for handoff purposes, the so-called handoff resources. These handoff resources cannot be used by newly emerging sessions, so that these new sessions are blocked early. This way, it becomes less likely that ongoing sessions are terminated before their desired ending.

This article presents a new proposal for a handoff prioritization scheme named *SiS-HoP* (Simple and Scalable Handoff Prioritization scheme) which is especially designed to meet the above mentioned requirements. The article is organized as follows: Section 2 describes *SiS-HoP* and its components in detail. Section 3 presents related work, for example, existing state-of-the-art handoff prioritization schemes and why they do not fulfill the above mentioned requirements. *SiS-HoP* has been evaluated using the network simulator ns2. Section 4 describes the simulation environment including the simulation model and those handoff prioritization schemes which are compared to *SiS-HoP*. Simulation results are presented in Section 5. Finally, Section 6 summarizes the results and discusses future work.

2 SiS-HoP: Simple and Scalable Handoff Prioritization

This section contains a detailed description of *SiS-HoP* including a comprehensive discussion on the necessity and the design of each component. At first, a short summary describes the main contributions of *SiS-HoP*.

2.1 SiS-HoP in a Nutshell

SiS-HoP incorporates two main components:

(1) An aggregated mobility prediction based on a small mobility cache, which stores the history of handoffs in each base station.
(2) A handoff resource reservation, which pre-reserves handoff resources aggregately between neighboring cells according to the outcome of the mobility prediction.

The basic function is as follows: After a mobile terminal has performed a handoff towards a neighboring cell, an entry is stored in a mobility cache of the old cell. This entry contains an identifier of the neighboring cell and optionally further information, such as the duration, for which the mobile terminal has consumed resources in the old cell (the so-called resource holding time). As an example, the mobility cache in cell X may contain ten entries describing handoff events, from which three were to the neighboring cell A, three to cell B and four to cell C. In this case, it is estimated that 30% of all mobile terminals in cell X will handoff to cell A, 30% to cell B and 40% to cell C. On the basis of this mobility prediction, *SiS-HoP* reserves handoff resources aggregately between neighboring cells. Continuing the above example, 30% of the resources currently consumed in cell X are pre-reserved as handoff resources in cell A, another 30% in cell B, and 40% in cell C.

The mobility prediction of *SiS-HoP* includes two major innovations with regard to the described basic cache-based scheme:

(1) *'Current-cell' entries*
In case of a session termination, an entry is stored in the cache in the same way as in case of a handoff. However, this entry contains the identifier of the current cell instead of the neighboring cell and is, thus, called 'current-cell' entry. Such entries are very useful, for example, in cells where the majority of mobile terminals terminates their sessions so that only a small amount of handoff resources has to be pre-reserved in neighboring cells.
(2) *Resource holding time normalization*
In the above described basic cache-based scheme, mobile terminals with a short resource holding time (e.g., moving at high speed) tend to push out those entries from the mobility cache which have a long resource holding time (e.g., moving at low speed). This leads to wrong aggregated estimates of the mobility pattern because too few entries of mobile terminals with a long resource holding time are used for the mobility prediction. In *SiS-HoP* the resource holding time is additionally stored in each cache entry in order to be able to normalize the influence of each cache entry on the final estimation.

Thus, *SiS-HoP* uses only information about the next cell (which can be the current cell in case of a session termination) and the time a mobile terminal consumes resources in a cell, so that the *complexity* of the mobility prediction is limited. Mobility prediction and handoff resource reservation of *SiS-HoP* are also *scalable* because no per-flow or per-mobile state information is added to the mobile network. To avoid per-mobile signaling, the amount of handoff resources to reserve in neighboring cells as well as the handoff probabilities are performed periodically. Furthermore, *SiS-HoP* provides a high *assurance on the handoff success probability* for a wide variety of mobility patterns without compromising the resource utilization in scenarios with many non-mobile terminals. The scope of *SiS-HoP* lies within the access network where bottleneck links of mobile networks are located. If *SiS-HoP* is coordinated with a legacy DiffServ resource management in the core network, *incremental deployment* of QoS in mobile networks becomes possible.

One of the highlights of *SiS-HoP* is its single parameter to tune the *efficiency* of the scheme similar to over-reservations in airline reservation systems. This single parameter is simple to configure and enables an *easy administrability*, also because it is not necessary to reconfigure it as a response to a change in the mobility pattern. Additionally, this single parameter is *robust* against mis-configuration which is different for existing schemes, for example, in case of the single parameter of static handoff prioritization schemes such as the Guard Channel scheme [11].

SiS-HoP will be explained in detail in the following sections.

2.2 SiS-HoP Architecture

SiS-HoP extends the architecture of a legacy QoS-enabled network by a *handoff resource reservation* component and a *mobility prediction* component (cf., Fig. 2). The legacy control plane comprises admission control and resource management which are necessary even in case there is no handoff prioritization scheme implemented in the mobile network. The handoff resource reservation of *SiS-HoP* determines the necessary amount of handoff resources which is necessary to ensure that the QoS of already admitted sessions can be maintained even in case of future handoffs. Thus, *SiS-HoP* enhances the legacy admission control component to include handoff resources in the admission decision. To determine the amount of handoff resources, the mobility prediction of *SiS-HoP* provides estimates on the handoff probabilities for each neighboring cell based on aggregated mobility patterns. These handoff probabilities are used by the handoff resource reservation to calculate the necessary amount of handoff resources for each neighboring cell. This amount is forwarded to the corresponding handoff resource reservation component in that cell.

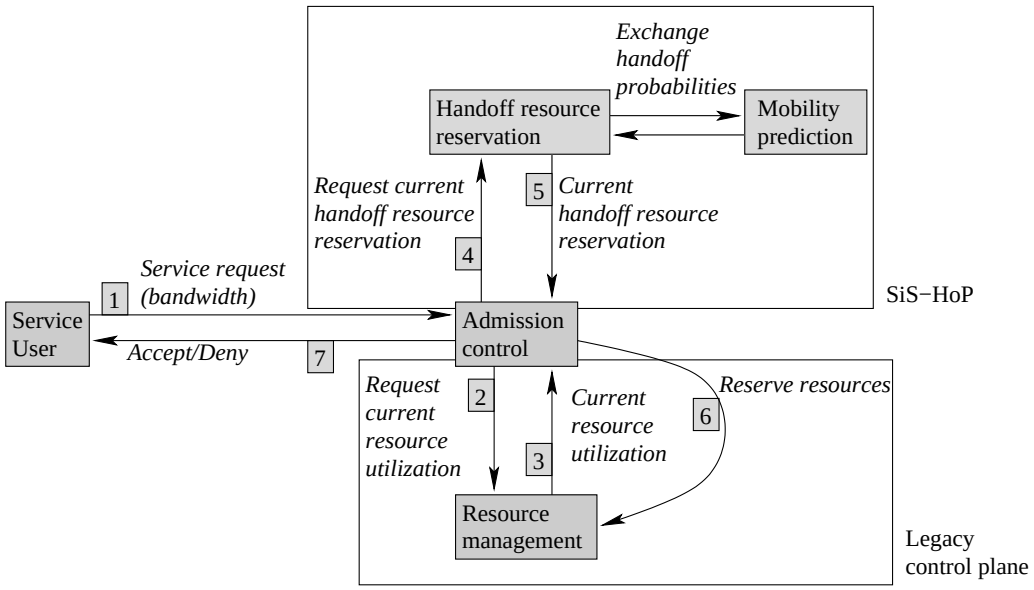

Fig. 2. *SiS-HoP* as an extension of the legacy control plane

The first phase of the legacy control plane procedure (arrows 1–3) comprises sending the resource request from the service user to admission control which contacts the resource management to determine whether the requested resources are available or not. Thereafter, the enhanced procedure of *SiS-HoP* begins (arrows 4+5): Admission control consults the handoff resource reservation component in order to receive information about the handoff resources needed to satisfy future handoff requests from neighboring cells. Finally, the legacy control plane procedure can continue (arrows 6–7) with reserving resources for the service request and sending an accept message (or a deny message) to the service user.

The key components of *SiS-HoP*, mobility prediction, handoff resource reservation, as well as the admission control are described in detail in the following Sections 2.3, 2.4, and 2.5.

2.3 Mobility Prediction

Determining the necessary amount of handoff resources is essential in mobile networks with dynamic mobility patterns changing over time. These patterns are typical, for example, within a city, where the mobility patterns of vehicles are different in the morning rush hours compared to the evening rush hours or to the weekend hours. In networks with static mobility patterns, it is possible to measure the traffic pattern once off-line and use this data to statically configure the necessary amount of handoff resources. However, networks with dynamic mobility pattern require on-line measurements of the mobility pattern so that the necessary handoff resources can be adapted over time. Thus,

the objective of a mobility prediction component within a handoff prioritization scheme is to support the handoff resource reservation component in dynamically determining the necessary amount of handoff resources.

In *SiS-HoP* the mobility prediction calculates the *aggregated* handoff probabilities for each neighboring cell. This means, all mobile terminals moving to a neighboring cell are aggregated, individual mobile terminals are not considered. An example is depicted in Figure 3. The mobility prediction in cell A

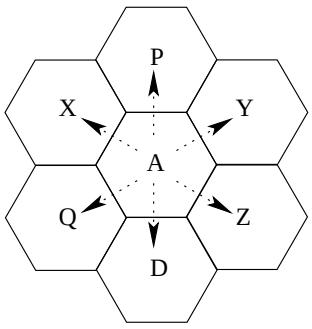

Fig. 3. Mobility prediction: Determination of handoff probabilities

determines the probabilities, with which any mobile terminal, being currently in cell A, performs a handoff to the neighboring cells X, P, Y, Z, D, and Q. Since the accuracy of the predicted mobility determines the quality of the handoff prioritization scheme, the mobility prediction component constitutes an important part within *SiS-HoP*.

The design of the mobility prediction component involves two decisions as discussed below:

(1) Where should the mobility prediction be located (on the mobile terminal or in the mobile network)?
(2) Should it be based on counters or a history-cache (which are the two main possibilities for predicting mobility aggregately between neighboring cells)?

With regard to the location of the mobility prediction, *SiS-HoP* deploys a mobility prediction component on each base station which keeps the additional complexity away from the mobile terminals. In the alternative case, i.e., if a mobility prediction component is used on each single mobile terminal, the complexity of each mobile terminal would be increased significantly because each mobile terminal would need to keep track of many visited cells to achieve good prediction results.

Furthermore, it is assumed that the mobility of *all* mobile terminals visiting this base station is sufficiently correlated to achieve an accurate mobility prediction. This is true, for example, if there is a main street crossing the cell so

that a large percentage of mobile terminals moves along this street (although, for example, a single mobile terminal may drive along this street only once a week). Such a per-cell mobility prediction captures also regular mobility patterns of single mobile terminals, for example, in the surrounding of meeting rooms or coffee machines. In contrast to the alternative case of a mobility prediction on each single mobile terminal, there is no need for a correlation within the mobility of a *single* terminal to achieve good prediction results.

With regard to the choice between a counter-based scheme or a history-cache-based scheme, the main requirement on the mobility prediction component is that is must not contain additional per-flow/per-mobile state keeping or signaling to be simple and scalable. Although there is per-flow state keeping in the legacy control plane components for storing QoS parameters of resource requirements, additional per-flow state information, for example, per-flow handoff resource reservations from neighboring cells, should be avoided. For this reason, mobility is estimated aggregately within each cell in *SiS-HoP*. Two different mobility prediction schemes incorporating aggregation can be distinguished

- aggregation based on counters, or
- aggregation based on a history cache.

Counter-based mobility prediction schemes [14] count the number of handoff events, for example, separately for each next cell. They cannot react sufficiently fast to changes in the mobility pattern.

In contrast, history-based schemes [15,12] store a certain amount of information for each handoff event (e.g., the next cell), the so-called *handoff information unit*. Each handoff information unit constitutes an entry in the cache, the sum of the entries form the *mobility cache* C. The size of the mobility cache is limited by the cache size $s(C)$, which is a design parameter of history-cache-based approaches. It denotes the number of handoff information units to be stored in the cache and has two purposes:

(1) It limits the memory requirements of the mobility prediction component.
(2) It determines the adaptability and the accuracy of the approach.

In contrast to counter-based schemes, a history-cache-based mobility prediction can forget 'old' handoff events. New handoff information units replace aged units if the cache is completely filled. The time to react to changes depends on the cache size $s(C)$. If $s(C)$ is small, the scheme can react very fast to changing mobility patterns. However, a too small cache size lowers the accuracy of the mobility prediction and can lead to wrong predictions. For example, for a cache size $s(C) = 10$, the arrival of ten 'unusual' handoff events can lead to a completely different result of the mobility prediction.

Therefore, history-based schemes are better suited to react to changing mobility patterns and to deal with heterogeneous bandwidth demands. For these reasons, the mobility prediction component of *SiS-HoP* is built on a history-based approach which calculates the aggregated mobility per-cell.

The Next-Cell-Based Mobility Prediction

The main task of the mobility prediction is to calculate separate *handoff probabilities* for each neighboring cell. The handoff resource reservation component uses the handoff probabilities to estimate how many resources from the currently utilized resources in the current cell are to be reserved in the neighboring cells.

To provide separate handoff probabilities for each neighboring cell, the next cell, to which a mobile terminal has performed a handoff, must be stored in a handoff information unit.

$$\text{Handoff information unit } = <\text{next-cell}> \tag{1}$$

This is the simplest possible handoff information unit, on which the so-called *Next-cell-based mobility prediction* is based.

The prediction function f_{next} is defined as follows:

$$f_{next}(h) = \frac{|\{h_1 \in C \wedge h.\,\textit{next-cell} == h_1.\,\textit{next-cell}\}|}{|\{h_1 \in C \wedge h_1 \neq \textit{empty}\}|} \tag{2}$$

h and h_1 are handoff information units, containing a next cell only, and C is the mobility cache. The probability to handoff into a specified next cell $h.\,\textit{next-cell}$ is the number of all handoff information units in the cache with the same next cell divided by the total number of handoff information units in the cache (i.e., all entries which are not empty). To retrieve this probability in a computational complexity O(1), the mobility prediction should keep a counter for each neighboring cell. Such a counter for a neighboring cell i is increased in case a new handoff information unit containing i as next-cell is added to the cache and it is decreased by one if a handoff information unit containing i as next-cell is deleted from the cache. Therefore, only one additional counter for each neighboring cell is necessary, which slightly increases the memory consumption of the mobility prediction.

Session Terminations

A remaining problem of the mobility prediction is that it does not adapt to different degrees of mobility within different cells. A high degree of mobility in a cell means that most terminals move and perform handoff eventually. An example is a 'highway cell' where all terminals move normally (if there is no traffic jam). A low degree of mobility can have two reasons:

(1) Many portable (i.e., non-moving) terminals.
(2) Many session terminations in a cell.

Portable terminals may constitute a large percentage of all terminals in a cell, for example, located at an airport. The degree of mobility is low if many terminals wait for the arrival of an air plane. In contrast, the mobility can become high if the air plane has arrived so that all mobile terminals move towards the exit gate while continuing their sessions. A high number of session terminations may occur in a cell with a popular target, for example, a cinema. In this case, many terminals perform handoff into the cell and terminate their session, for example, when the film starts.

Scenarios with a low degree of mobility lead to an unnecessarily high handoff resource reservation in neighboring cells because the sum of the handoff probabilities is always equal to one if there is any handoff information unit in the mobility cache. In the extreme, no handoffs to neighboring cells might occur at all for an extended period of time, but handoff resources are still reserved as long as there are (old) mobility patterns in the cache.

To avoid such an over-reservation of handoff resources in scenarios with a low degree of mobility, *SiS-HoP* considers *session terminations* in the mobility cache which is a unique feature of this approach. Instead of estimating the cell residence time for each mobile terminal within the current cell, for example, as proposed in the Choi scheme [12], *SiS-HoP* introduces additional entries into the cache for next-cell = current cell, the so-called *'current-cell' entries*. That means, each time a mobile terminal terminates its session, a handoff information unit is created with the current cell stored as 'next-cell'. In this case, session terminations can populate the cache the same way as handoff events, so the sum of the handoff probabilities to all neighboring cells can become smaller than 100% and can even go down to 0% if the cache is completely populated with handoff information units from session terminations.

An example for using current-cell entries in the next-cell-based scheme is shown in Figure 4. The cell topology consists of a center cell A, which is surrounded by six other cells. For this center cell, the right part of the figure depicts an example for the values in a mobility cache. It contains ten handoff information units: Two mobile terminals have performed handoff to cell Z, cell Y,

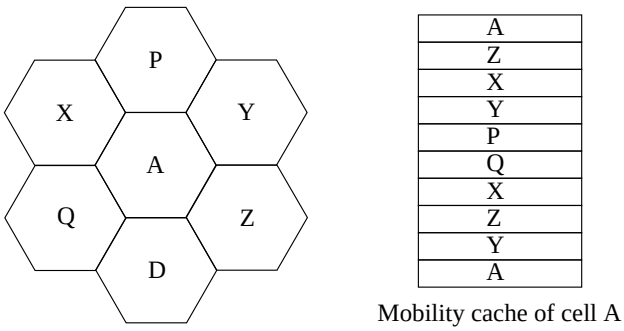

Fig. 4. Session terminations in a next-cell-based mobility prediction

and to cell X, respectively. A single handoff occurred to the cells P and Q, and no handoff to cell D. In this case, $f_{next}(Y) = f_{next}(Z) = f_{next}(X) = \frac{2}{10} = 20\%$, $f_{next}(P) = f_{next}(Q) = \frac{1}{10} = 10\%$ and $f_{next}(D) = 0\%$. Since two sessions were terminated in cell A (= 20% of all sessions), the sum of the handoff probabilities is only 80%. If a mobile terminal in cell A performs a handoff, for example, to cell Z, the oldest handoff information unit in the mobility cache will be replaced as in a FIFO queuing discipline (i.e., the bottom-most entry 'A' in the example).

Impact of Mobility Patterns with Different Speeds

The next-cell-based approach is well-suited for mobility patterns with no correlation between the next-cell and the speed of the mobile terminals. However, it can produce wrong estimates if this assumption is not true as explained in the following.

In general, a mobility cache with next-cell-based handoff information units provides the information, that $x\%$ of the last $s(C)$ handoffs moved to cell X, $y\%$ to cell Y etc. These percentage values do not necessarily reflect the handoff probabilities as shown in the following example. Figure 5 depicts a scenario with three cells around the current cell A. In cell A, there is a correlation

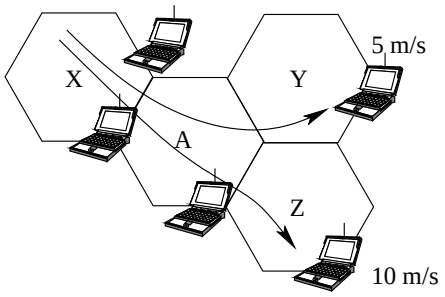

Fig. 5. Mobility scenario with heterogeneous speeds

between the next cell and the speed of the mobile terminals, which move with

$5\,m/s$ to cell Y (so-called 'low-speed terminals') and $10\,m/s$ to cell Z (denoted as 'high-speed terminals'). For example, there may be a speed limitation to cell Y owing to a construction area or the average speed may be lower because of traffic lights on the way from cell A to cell Y. If the actually utilized resources are the same for both speed classes so that always the same number of mobile terminals is active in a cell for both speed classes, the high-speed terminals perform twice as many handoffs into cell Z than the low-speed terminals into cell Y within a time window. This can be seen from the example in Figure 6 which depicts a series of three parts of the mobile terminals in the cell A for such a heterogeneous speed scenario. At time t=$0\,s$ one low-

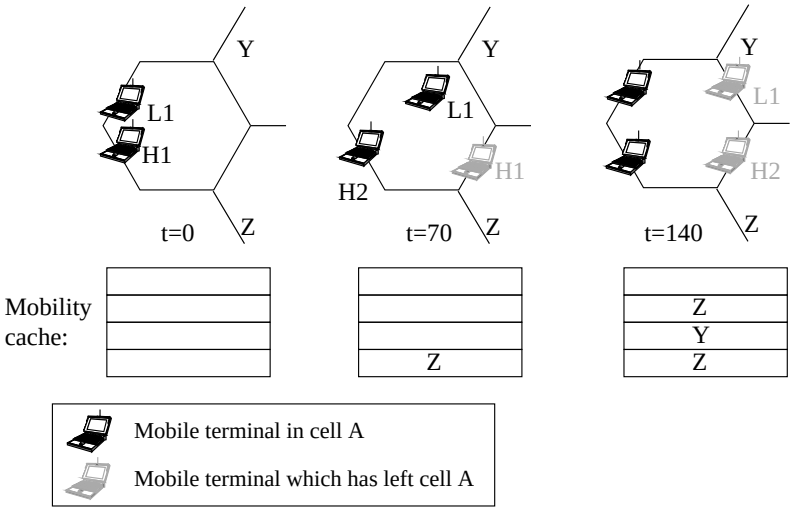

Fig. 6. Heterogeneous speeds: Resource utilization vs. number of handoffs

speed terminal L1 and one high-speed terminal H1 enter the current cell, the mobility cache is empty. H1 moves with $10\,m/s$ through the cell which has a diameter of $700\,m$, so H1 leaves the cell after t=$70\,s$. This leads to the first entry in the mobility cache. At the same time L1 reaches the center of the cell. When H1 leaves the current cell, another high-speed terminal H2 enters the cell so there is still one high-speed terminal and one low-speed terminals within the cell. After $140\,s$, both L1 and H2 reach the cell boundary and two further entries in the mobility cache are created. Two new terminals enter the current cell so a new cycle of this periodical mobility pattern starts.

In this scenario, there are one handoff for the low-speed terminals to cell Y and two handoffs for the high-speed terminals to cell Z. However, the handoff probability is 50% to cell Y and 50% to cell Z as there is always one high-speed terminal and one low-speed terminal in the cell simultaneously. For this reason, the same amount of handoff resources should be reserved in cell Y and cell Z in this case, not $f_{next}(Z) = \frac{2}{3} = 0.66 = 66\%$ and $f_{next}(Y) = \frac{1}{3} = 0.33 = 33\%$ as obtained from the next-cell-based mobility prediction.

To deal with correlations between the speed and the next cell as explained above, *SiS-HoP* takes the resource holding time into account to normalize the purely next-cell-based handoff probabilities. In this case, a handoff information unit in the cache contains the resource holding time (RHT) in addition to the next cell.

$$\text{Handoff information unit } = <\text{next-cell}, \text{RHT}> \tag{3}$$

With this resource holding time, the mobility prediction can differentiate the 'high-speed' terminals from the 'low-speed' terminals and can adapt the handoff probabilities as follows.

In the first step, the mobility prediction calculates the average resource holding time for each next-cell (*avg_RHT(next_cell)*) and the average resource holding time for all cache entries (*avg_RHT*). These values are used in the prediction function f_{norm_next} as follows:

$$f_{norm\_next}(h) = f_{next}(h) \cdot \frac{avg\_RHT(h.next\text{-}cell)}{avg\_RHT} \tag{4}$$

Again, h is a handoff information unit. As described previously for the next-cell-based mobility prediction, it is possible to achieve a computational complexity of O(1): In this case, it is necessary to store the sum of the resource holding times for all those handoff information units in the cache, which contain the same next-cell. Together with the one counter per neighboring cell and $f_{next}(h)$, as described in the next-cell-based mobility prediction, it is possible to gain *avg_RHT(next-cell)* and *avg_RHT*. A difference to the previous scheme is that more memory is consumed for storing one sum of the resource holding times per neighboring cell and because the resource holding time must be stored additionally in each handoff information unit in a floating point variable. It is furthermore necessary to store the time, when a mobile terminal starts to consume resources in the current cell, so the resource holding time can be calculated when the terminal performs handoff. However, this additional amount is not significant because it only increases the amount of local state information, which is still limited by the cache size and the number of neighboring cells, and does not lead to additional state information in neighboring cells or to further signaling between neighboring cells.

Continuing the previous example, Figure 7 depicts a mobility cache where the resource holding time has been added to each handoff information unit. The average resource holding time is $(2 \cdot 70s + 140s)/3 = 93.3s$ and the handoff probabilities are calculated as $f_{norm\_next}(Z) = 0.66 \cdot \frac{70s}{93.3s} = 0.5 = 50\%$ and

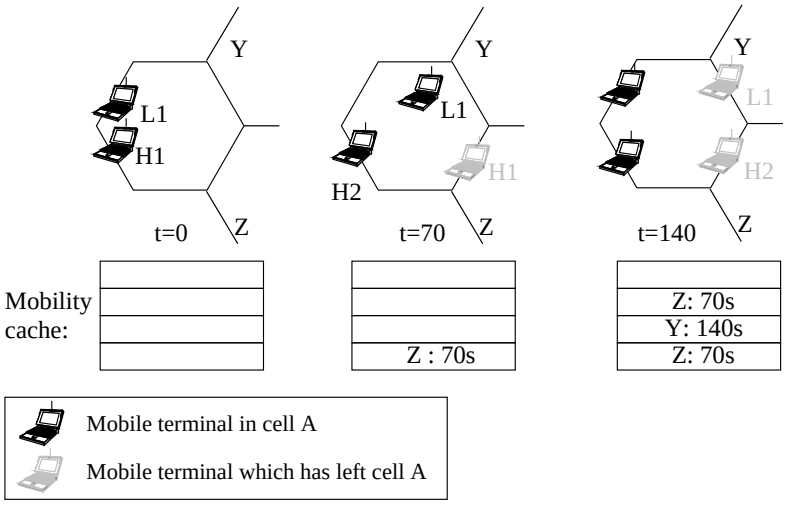

Fig. 7. Heterogeneous speeds: Mobility cache for the normalized next-cell-based prediction

$f_{norm\_next}(Y) = 0.33 \cdot \frac{140s}{93.3s} = 0.5 = 50\%$. This corresponds to the actual handoff probabilities.

In the same way as for the normalized resource holding times, it is possible to handle correlations between the next-cell and heterogeneous bandwidth demands per mobile terminal (i.e., low-demand vs. high-demand sessions). This is important for multimedia applications where the bandwidth demand of single sessions can differ significantly. As an example, the bandwidth demand may be higher for those sessions moving to a next-cell which is located at the entrance to the highway. In this case, the passengers in the back may tend to start high-bandwidth consuming video streaming sessions because of the upcoming long highway travel. In such a case with heterogeneous bandwidth demands, the demand of each mobile terminal has to be stored in each handoff information unit. This way, a calculation of normalized bandwidth demands becomes possible so that the actual bandwidth demand in the current cell can be weighted appropriately.

2.4 Handoff Resource Reservation

The handoff resource reservation component is responsible for calculating the necessary handoff resources to reserve in neighboring cells using the mobility prediction component. It is collaborative, i.e., it performs reservations in neighboring cells, because a local scheme cannot provide sufficiently high assurances on the handoff success probability. There are few approaches [16–18], which also reserve resources in cells which are several hops away from the current cell (herein referred to as 'distant-cell-based schemes'). These schemes are basically intended for scenarios, where mobile terminals move very fast

or cells are very small. In this case, neighboring-cell-based schemes may not be able to provide handoff resources in time. For example, a mobile terminal M originates in cell A and will move to cell B and cell C. If resources are reserved in neighboring cells only, M will be admitted in A and resources are reserved in cell B. However, resources in cell C will be reserved when M moves to cell B. If cell C is now very busy, it might be that no resources become available in cell C during the time M crosses cell B. Thus, a handoff drop would occur when M finally performs a handoff into cell C. Such a handoff drop could be avoided by the above mentioned distant-cell-based approaches, which would reserve handoff resources in cell C earlier, e.g., before M performs a handoff to cell B. However, these approaches have a rather high complexity and a high communication overhead to signal the reservation information between the cells. Furthermore, it is questionable why these approaches should provide a higher assurance on the handoff success probability compared to the neighboring-cell-based schemes. First, neighboring-cell-based schemes also support mobility across multiple cells implicitly as long as sufficient resources become available (e.g., because of session terminations) in the predicted next cell while a mobile terminal crosses the current cell (there are no per-mobile handoff reservations but a common pool of reserved handoff resources in each base station). This, however, depends on a concrete network scenario and/or mobility pattern. Second, it is rather difficult to predict the mobility of a terminal for more than one cell (except for special cases, e.g., highways). Therefore, *SiS-HoP* adopts a neighboring-cell-based approach as described in the following.

2.4.1 Functional Overview

The function of the handoff resource reservation component in *SiS-HoP* in a particular cell consists of two parts:

a) It collects the handoff resource reservation requests from neighboring cells and adjusts the amount of handoff resources in the current cell accordingly.
b) It calculates the necessary handoff resources to be reserved in each neighboring cell using the mobility prediction component and propagates the results to the neighboring cells.

For scalability reasons, handoff resources are reserved aggregated between neighboring cells to avoid per-mobile pre-reservations in the neighboring cells.

2.4.2 The Algorithm for Handoff Resource Reservation

The handoff resource reservation in a cell X is performed as shown in Figure 8: The handoff resource reservation for a particular neighboring cell i is deter-

```
totalUsed = 'total amount of used bw in cell X';
for each neighboring cell i
   handoff_resv[i] = f_norm_next(i) · totalUsed · CUR;
   send handoff_resv[i] to neighboring cell i;
done
```

Fig. 8. Handoff resource reservation

mined by multiplying the currently used bandwidth '`totalUsed`' in cell X with the probability to handoff into cell i. The latter can be calculated using the normalized next-cell-based mobility prediction scheme.

The calculation of the handoff resource reservation amount is performed periodically to avoid per-mobile signaling between neighboring cells, so *scalability* is ensured. Using an additional set of simulations, it was shown for a particular simulation scenario (cf., Sect. 4) that exchanging the data once per second is a good compromise between achieving a high accuracy while maintaining a low signaling overhead [19].

The currently used bandwidth '`totalUsed`' is obtained from the legacy resource management component. The calculation of '`totalUsed`' depends on the provided service: For a service with strict bandwidth assurances (e.g., telephony), the peak rates of all currently active sessions are summed up.

For the handoff resource reservation, it is necessary that cell X is aware of those neighboring cells, where mobile terminals have performed handoff to. This information can be obtained from the mobility cache in which all next cells are assumed to be neighboring cells. If a neighboring cell has currently no entry in the cache, no handoff resource reservation takes place for this cell.

Furthermore, it is recommended to use another parameter for the calculation of the handoff resources, the so-called *Controlled Under-Reservation* (CUR) parameter.

2.4.3 The Controlled Under-Reservation (CUR) Parameter

The intention of the CUR parameter is to decrease the amount of handoff resources manually in order to increase the *efficiency* of the scheme. Thus, the CUR parameter enables a tuning of the system performance comparable with over-reservations in airline reservation systems: The probability of handoff drop increases while the amount of handoff resource decreases when the CUR parameter is decreased below its default value, which is 100%. It can be decreased by the network operator if the resource utilization is very low and if there are no or only few handoff drops. A similar design parameter has already been proposed in other schemes [20–22].

The CUR parameter is useful since *SiS-HoP* is a rather conservative handoff prioritization scheme in that it overestimates the necessary handoff resources. This is because handoff resources are already reserved in the neighboring cells if a mobile terminal performs a handoff into a cell. These resources are unused until the mobile terminal actually performs a handoff. A high assurance on the handoff success probability can be achieved in this way although it leads to a rather low resource utilization at the same time: If, for example, the above described mobility prediction is used without considering session terminations, the amount of reserved handoff resources would be the same as the actually used resources, so that the network utilization would not exceed 50% on average. The consideration of session terminations with 'current-cell' entries in the mobility cache reduces the overall handoff reservation level below 100% of the current load: The handoff resources are even reduced to zero in case terminals do not perform handoff at all. The normalization factor in the normalized next-cell-based mobility prediction only redistributes the amount of handoff resources to be reserved in which neighboring cell, but does not lower the overall amount of handoff resources reserved in the network. Thus, there is still room to improve the resource utilization, for example, if the resource holding time in a cell is rather long. This can occur in case of a low-speed mobile terminal or in case of large cells. In such scenarios, it would in principle be better to reserve handoff resources not directly after a mobile terminal has entered a cell, but, e.g., after the mobile terminal has crossed half of the cell. Available schemes, for example, the one proposed by Choi and Shin [12], try to reserve handoff resources 'just-in-time', i.e., almost exactly before the mobile terminal performs the handoff. The main problem is that these schemes require several parameters to achieve a good estimation of the exact handoff time which makes the administrability of the approach difficult. In contrast, *SiS-HoP* implements a simpler approach by the introduction of the CUR parameter.

The CUR parameter is intended to be changed by the system administrator to *tune* the performance whereas the value of all other design parameters of *SiS-HoP* (e.g., the mobility cache size) are normally set only once on system startup.

Partial Deployment

The handoff resource reservation component can be deployed partially, for example, in highly loaded areas only (e.g., cell A in Fig. 9). In this case, the neighboring cells around cell A must at least perform mobility prediction and signal the expected handoff resource demand to cell A. These cells themselves, however, do not need to reserve handoff resources.

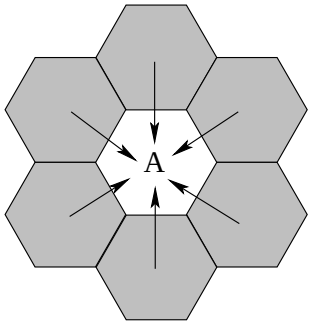

Fig. 9. *SiS-HoP*: Partial deployment of handoff resource reservation

2.5 Admission Control

Currently, *SiS-HoP* is intended to support CBR traffic as, for example, generated by IP telephony applications (cf., Fig. 1). Therefore, the admission control algorithm within *SiS-HoP* is currently based on peak rates.

For a service with strict bandwidth assurances, a new session request in cell X from a mobile terminal M with a peak bandwidth demand $M.bw$ is admitted if the following two conditions are met:

(1) Local bandwidth test:
There is sufficient bandwidth (bw) available in cell X:

$$X.\,total\_bw - X.\,used\_bw - X.\,reserved\_bw \geq M.bw \tag{5}$$

This considers both, the bandwidth currently in use (*used_bw*) and the bandwidth necessary to satisfy future handoff requests (*reserved_bw*). The latter is determined by the handoff resource reservation components of the neighboring cells. In case the request can be admitted, the resource management component is instructed to increase *used_bw* by $M.bw$.

(2) Bandwidth test in the neighboring cells:
There is sufficient bandwidth available in all neighboring cells, i.e., the following condition must hold for all neighboring cells $c_i, i = 1, \ldots, n$:

$$c_i.\,total\_bw - c_i.\,used\_bw - c_i.\,reserved\_bw \geq f_{next}(c_i) \cdot M.bw \cdot \mathrm{CUR} \tag{6}$$

This test is based on the handoff probabilities obtained by the mobility prediction component. Furthermore, the CUR parameter allows that only a percentage of the bandwidth demand as determined by CUR must be available in the neighboring cells.

This test may lead to the case that a new session request is denied in a cell A because a neighboring cell X is busy, even though the mobile terminal would not move to cell X. This basically means that the mobility prediction has failed to predict '$f_{next}(X) = 0$' for all mobile terminals

in cell A, which may happen in schemes, where mobility is predicted aggregately and not per-mobile in order to achieve a scalable scheme. Furthermore, the basic idea of handoff prioritization schemes is to reduce handoff drops by increasing the blocking of new session requests. So this test is necessary to avoid that mobile terminals are admitted in one cell and move to a congested neighboring cell immediately where they would be dropped. As the variable *reserved_bw* depends on the value of CUR, the administrator of the network can also tune the system behavior indirectly by changing the CUR parameter.

Admission control on a handoff arrival in cell X considers only if there is sufficient bandwidth available to support the mobile terminal locally:

$$X.\,total\_bw - X.\,used\_bw \geq M.bw \tag{7}$$

The amount of used resources (*used_bw*) is also increased by $M.bw$ in this case to satisfy the handoff resource request. A reservation of handoff resources for further handoffs of the mobile terminal M into neighboring cells of cell X is performed by the handoff resource reservation component implicitly. This is because of the increase of *used_bw* which is the basis for the decision, how many handoff resources cell X should reserve in the neighboring cells at the end of the next signaling period.

3 Related Work

Many handoff prioritization schemes have been proposed during the previous years. However, existing schemes are not suited for future mobile networks since they suffer at least from one of the following problems with regard to the requirements, mentioned initially:

- They provide only a low *efficiency*, i.e., a too high number of new sessions is blocked in order to reduce the probability of a handoff resource shortage. For example, Oliveira et al. [22] proposed a scheme without any mobility prediction where the full amount of resources for a single mobile terminal is reserved in all neighboring cells. Jayaram et al. [23] propose to reserve resources for a mobile terminal in those three neighboring cells where the mobile terminal will move to with the highest probability which also leads to the fact that the handoff resources are always three times higher than the actually used resources. Another problem with regard to resource efficiency is that handoff resources are reserved even for portable terminals which do not move at all [22].
- They cannot provide a sufficiently high *assurance on the handoff success probability* in mobile networks with small cells for a wide variety of mobility

patterns. The Guard Channel scheme [11] reserves a manually configurable (i.e., static) amount of handoff resources, which requires manual intervention in case of dynamically changing mobility patterns. Local handoff prioritization schemes adapt the amount of handoff resources dynamically depending on information available from the local cell. However, they require that the mobility pattern do not change too fast [24–26] because they cannot use information from neighboring cells to pro-actively adapt the handoff resources according to the future needs. Thus, they are not suited for mobile networks where mobility patterns are highly dynamic.

- They distribute per-flow / per-mobile state information [13,27] within the mobile network which leads to *scalability* problems in the presence of a high number of handoffs. For example, the MRSVP scheme requires the mobile terminal to specify in advance all the cells to be visited in the future, so that handoff resources can be specified accordingly in advance. The Shadow Cluster approach [28] also reserves resources in a region around the current cell of a mobile terminal depending on the predicted mobility of that terminal. However, keeping per-mobile state information to pre-reserve resources per-mobile in one or several neighboring cells does not scale well in presence of a high number of handoffs. This is because such state information has to be updated permanently, e.g., if the mobility prediction changes.
- The *administrability* of the schemes is comparatively difficult. For example, the Adaptive-bandwidth Reservation Mechanism [14] requires five design parameters to be handled by the network administrator. The mobility-dependent call admission control scheme [29,12] proposes separate admission control thresholds for each of the different classes of service, supported by the network. There are separate thresholds for new session requests and handoff session requests.
- They require large-scale changes in the mobile terminals which, to some extent, impairs the *deployment* of the scheme.
- In general, handoff prioritization schemes provide assurances on the handoff success probability either end-to-end or for the bottleneck links only, i.e., the wireless links or the last wired mile, which connects the base station to the backbone of the mobile network. Many schemes provide end-to-end assurances on the handoff success probability in a single deployment step [13], which increases the *complexity* of the scheme and hinders *deployment*. In contrast, bottleneck-centered approaches limit the *complexity* of the handoff prioritization scheme and simplify their initial *deployment* in wireless mobile networks. Furthermore, they can achieve a high gain even in the first stage of deployment since most of the resource shortages occur on the bottleneck links. Thus, *SiS-HoP* is initially intended to be used for the resources on the bottleneck link.

There is a further set of approaches, called fast handoff, seamless handoff, or low-latency handoff schemes, which are also focused on providing QoS for handoffs. In contrast to handoff prioritization schemes, these fast handoff ap-

proaches try to minimize the necessary delay for signalling a handoff. This way, disruptions of connectivity and packet losses during the handoff signaling can be avoided. Example approaches are currently discussed in the IETF [30] as an extension to the Mobile IP protocol [31] or the many micro-mobility approaches [32,33] which are intended to complement macro-mobility approaches such as Mobile IP. There are also approaches which combine such a mobility management with resource management [34].

In mobile networks with overlapping cell areas or in multitier cellular networks [35], a mobile terminal may have several possible next cells to which it can perform a handoff to. In this case, it may become reasonable to exchange data between the mobile terminal and the possible next cell candidates on what amount of resources are available at the potentially next cell. This resource signaling might be incorporated into a general candidate discovery protocol (cf., Liebsch et al. [36], a proposal for IP routers, but a layer-2 approach might be necessary, as well). This way, a resource shortage can be detected in advance and be avoided if another alternative cell has sufficient resources to support the handoff. However, this leads to per-mobile signaling on each handoff which is why this approach is currently not integrated in *SiS-HoP*.

4 The Simulation Environment

SiS-HoP has been evaluated extensively using the network simulator ns2 [37]. The most interesting results are presented in this section.

4.1 Simulation Model

The network model consists of two parts. The wireless part is composed of a variable number of base stations (either nine or sixteen) which are placed onto a rectangular grid. The distance between two base stations is $700\,m$ horizontally and vertically which is a typical distance for mobile networks in a densely populated city area. The cell size is $800\,m$ so the coverage areas of two neighboring base stations overlap up to $100\,m$ to enable soft handoffs without interruptions of connectivity. The handoff control algorithm is based on a hysteresis [38] which can avoid subsequent handoffs between two base stations within a short period of time (the so-called *flip-flop effect*). The wireless network is based on the IEEE 802.11 standard to simulate realistic effects such as collisions on the air interface. It is important to model these effects to achieve realistic simulation results [39].

The base stations are interconnected via a tree-like topology (cf., Fig. 10)

leading towards the root node of the backbone which itself may be connected to the Internet in a real-world scenario. Apart from the last wired mile, the links

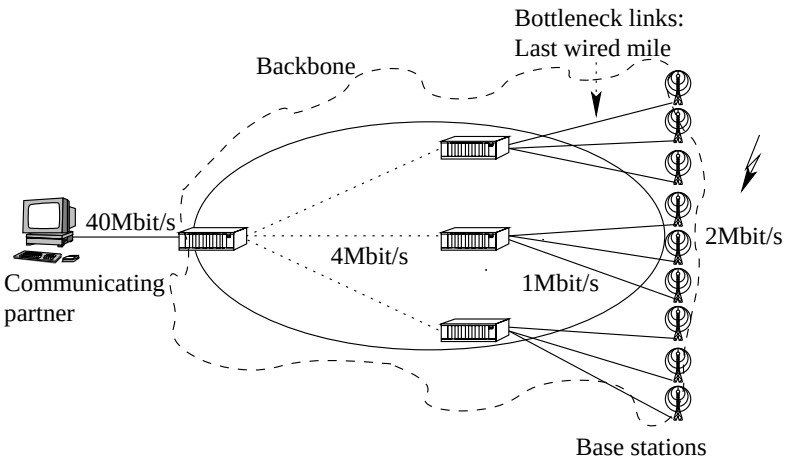

Fig. 10. The network model

are over-provisioned to limit bottleneck effects to the link directly attached to the base station. The node connected to the root-node represents a node in the Internet and is the communicating partner of all mobile terminals.

The network is designed such that each base station can carry up to 100 sessions simultaneously. The session duration is modeled as an exponential distribution with a mean of 180 s, which models telephony sessions quite accurately [1].

Two different mobility models are used in the following simulations: A scenario without directional traffic (the so-called Random-Move scenario) and a highly-directional scenario with large differences between the actual resource utilization (the so-called Directional-Move scenario). Such different schemes can confirm the applicability of a handoff prioritization scheme for a wide area of traffic patterns [40].

The Random-Move scheme consists of a 3x3 mobility cell scenario (with one base station in the center of each cell) where eight cells (2–9) are in use (cf., Fig. 11). In each cell, the probability of a mobile terminal to change to one of

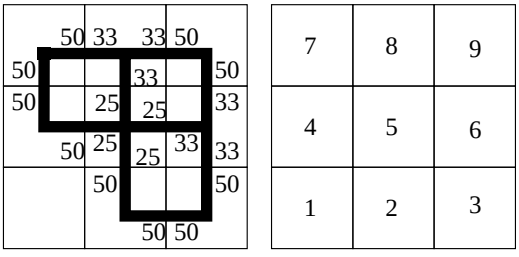

Fig. 11. The Random-Move mobility pattern

the neighboring cells is equal. Mobile terminals have the same probability to start in any of these eight cells, the average speed is a truncated Gaussian distribution [38] with different means and a maximum deviation of ±20%. Since this paper considers mobile terminals in vehicles and because the simulated network topology is appropriate for city areas, the Random-Move scenario is simulated with two average speed values: The simulations with an average speed of $17\,m/s$ (about $60\,km/h$) represent scenarios with no traffic jams where the terminals can move fluently, for example, on a main-street with coordinated traffic lights. The simulations with an average speed of $5\,m/s$ (about 20 km/h) represent scenarios with traffic jams such as in the morning or the evening rush hours. Higher speeds than $17\,m/s$ are not reasonable in this scenario because using such small cells is questionable in this case. Furthermore, this scenario is used for a rather static scenario, where only 25% of the mobile terminals move, the remaining 75% do not move at all (the so-called 'Static Random-Move scenario').

The Directional-Move scenario consists of a 4x4 mobility cell topology where 15 cells are used (cf., Fig. 12). In contrast to the Random-Move scenario,

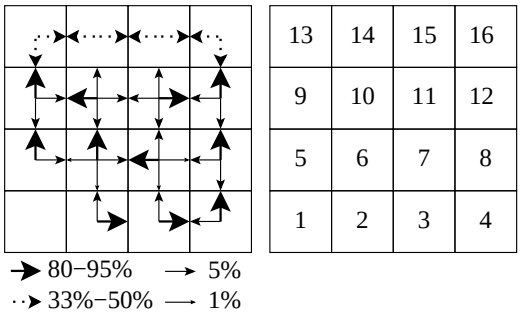

Fig. 12. The Directional-Move mobility pattern

traffic is highly directional for many cells. Furthermore, the probability where a session starts is not the same for all cells, it is 68% for cell 7, 19% for cell 6 and 1% for the remaining cells. This way, the resource utilization shows large differences between the cells. The speed of a mobile terminal is according to a truncated Gaussian distribution with a mean of $17\,m/s$ or $5\,m/s$.

4.2 Performance Metrics

Several different metrics are used in the literature to evaluate handoff prioritization schemes. In this article, the forced termination rate is used, which is defined as the number of handoff drops divided by the number of successfully admitted sessions. However, this metric alone is not sufficient, since avoiding handoff drops in general leads to an increase of the number of blocked new session requests. Therefore, the so-called *Cost of Service* metric [11] is

used additionally, which combines both, the probability of forced termination (P_{forced_term}) and the probability of new session blocks (P_{block}) into a single number using the following cost function:

$$CF_{forced\_term} = (1 - \alpha) \cdot P_{block} + \alpha \cdot P_{forced\_term} \quad (8)$$

The weight α was set to 0.84 so that five blocked sessions have the same influence on the cost of service as a single forced termination. Additional simulations not shown here [19] have confirmed that this value for the weight α is reasonable so that the Cost of Service metric for a mobile network without handoff prioritization differs significantly from the Cost of Service for a mobile network which includes a handoff prioritization scheme.

4.3 Schemes Used for Comparison

To evaluate *SiS-HoP*, it is necessary to choose different handoff prioritization schemes for comparison. At first, a scheme without handoff prioritization (NO-PRIO) is used to show the gain of prioritizing handoffs. Furthermore, a comparison to a theoretically optimal scheme (OPT) [40] makes it possible to show the gain any handoff prioritization scheme can achieve at maximum. Additionally, most schemes in the literature are compared to the Guard Channel scheme [11] since it is a simple scheme which is easy to implement. Although it has already been shown to be unsuited for mobile networks with dynamic traffic, it is used as a reference here. Further schemes from the literature are not used here because:

- they are not fully specified so they cannot be remodeled exactly in our simulation scenario [28,41], or
- they have such a high complexity or so many parameters to configure that rebuilding them in a different simulation environment than the original one is not feasible to achieve a fair comparison (e.g., [42]). This would most likely require to ask the authors of the scheme for proper values for the design parameters suited to our simulation scenario.

Therefore, a different comparison procedure is proposed in this paper. Since the main part of *SiS-HoP* is its mobility prediction, it is important to evaluate the mobility prediction component of *SiS-HoP* explicitly. Therefore, *SiS-HoP* is compared to a simplified version of itself, the so-called 'Load-dependent equal-probability virtual distributed admission control (LEP-DAC)' scheme. In contrast to *SiS-HoP*, LEP-DAC does not incorporate a mobility prediction scheme, so its complexity is even lower than the one of *SiS-HoP*. That means, in LEP-DAC handoff resources are reserved in the neighboring cells depending on the current bandwidth demand in the current cell. As there is no mobility

prediction, LEP-DAC can only assume an equally distributed probability to handoff into a neighboring cell. Hence, admission control checks whether the bandwidth demand divided by the number of neighboring cells is available in all neighboring cells or not. Resources are also reserved according to the number of neighbors for a cell. For example, if a cell has three neighboring cells, one third of the current bandwidth demand is reserved for handoff purposes in each neighboring cell. To accommodate session terminations, the current cell is included as a neighbor so that session terminations are treated as handoffs into the current cell (the same as the 'current-cell' feature of *SiS-HoP*). To be comparable to the *SiS-HoP* proposal, the CUR parameter is available in LEP-DAC as well, which decreases the amount of handoff resources reserved in neighboring cells to a certain percentage.

5 Simulation Results

For a comparison of *SiS-HoP* and LEP-DAC, it is important to use a single value of the CUR parameter for all considered mobility patterns. This models a real-life situation where the network administration will not change the CUR parameter each time the mobility pattern changes, for example, from highly directional to mainly static. Therefore, it is necessary to find a single value for the CUR parameter which provides a good performance in all considered mobility patterns.

5.1 Directional-Move, Speed=17 m/s

Figure 13 depicts the simulation results with regard to the forced termination rate in the Directional-Move scenario with speed=17 m/s. The offered load is a simulation parameter to vary the load in the network [12]. It is intuitively defined as follows: At an offered load of 100%, the new session arrival rate is such that no new session request has to be blocked and all resources of all cells are busy if all terminals are static, all sessions start simultaneously, having a constant session duration of 180 s.

SiS-HoP achieves a very low forced termination rate even for a CUR value of 70%. In contrast, the Guard Channel scheme with a reservation of 30% (the best value for all examined simulation scenarios) leads to up to 15% forced terminations. The simulations of the LEP-DAC scheme result in 4% forced terminations already for a CUR value of 100%. This is because of too small handoff resource reservations in case of highly directional traffic patterns. For example, mobile terminals move with a 95% probability to cell 10 from cell 6 (cf., Fig. 12). However, LEP-DAC reserves only 20% of the actually used resources

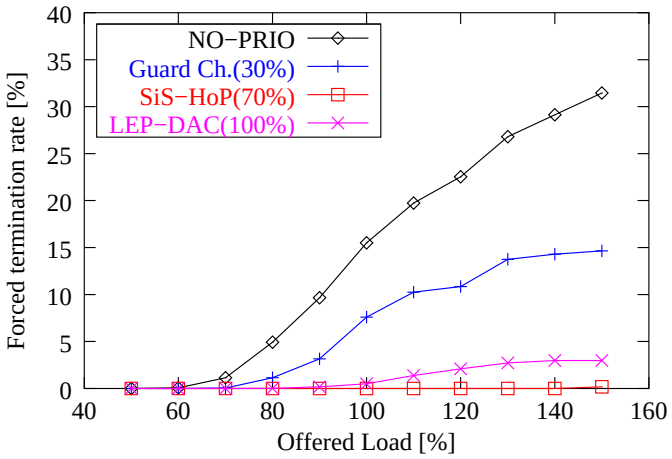

Fig. 13. Directional-Move: Forced termination rate

in a cell with four neighbors. This is because it assumes an equally distributed mobility due to the lack of a mobility prediction component. This 'LEP-DAC under-reservation effect' occurs even if the CUR parameter of LEP-DAC is set to 100%. Thus, *SiS-HoP* can provide a *higher assurance on the handoff success probability* in this scenario compared to LEP-DAC. Additionally, it is possible to set the CUR parameter of *SiS-HoP* to 70% while keeping the forced termination rate below 1% in the considered offered load range. This is not possible for the LEP-DAC scheme, for which the CUR parameter is set to 100% in the following simulations.

The Cost of Service (cf., Fig. 14) of *SiS-HoP* is lower than for LEP-DAC at offered loads of 100% and higher because of the increase in forced terminations for LEP-DAC. The shape of the Cost of Service curve of *SiS-HoP* is similar to the one of the optimal approach. In this scenario, the simulation results

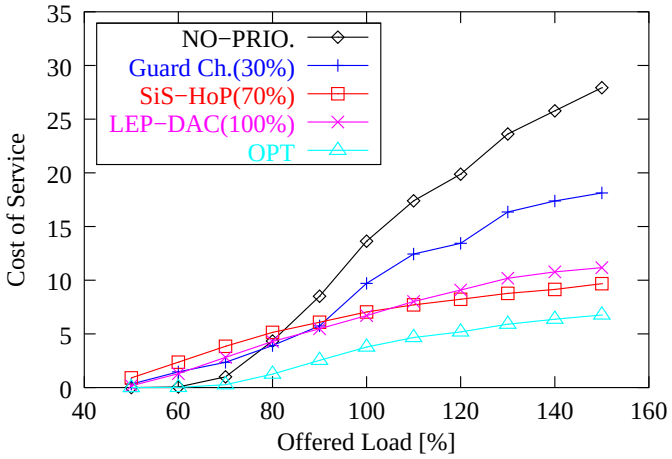

Fig. 14. Directional-Move: Cost of service

confirm that *SiS-HoP* performs better than LEP-DAC for offered load levels above 100%. In contrast to LEP-DAC, *SiS-HoP* can avoid forced terminations

entirely even for very high offered loads because it reserves handoff resources according to the actual mobility pattern. For offered loads below 100%, *SiS-HoP* has a higher Cost of Service because there are no forced terminations for both schemes and LEP-DAC reserves less handoff resources than *SiS-HoP* owing to the LEP-DAC under-reservation effect. Therefore, *SiS-HoP* blocks more new session requests for offered loads below 100% than LEP-DAC.

Comparing the forced termination rate of *SiS-HoP* for several values of the CUR parameter, it can be seen that the difference between the curves is almost proportional to the extent of the decrease of the CUR parameter (cf., Fig. 15). At the same time, the forced termination rate increases only slowly

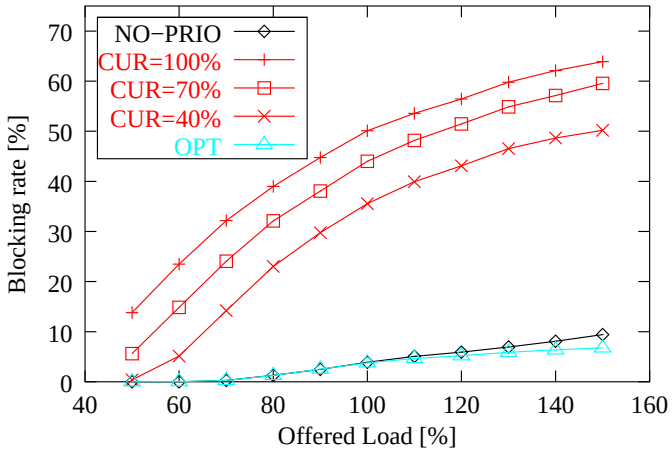

Fig. 15. Directional-Move: Blocking rate and the CUR parameter

(cf., Fig. 16). This underlines the robustness of the CUR parameter against

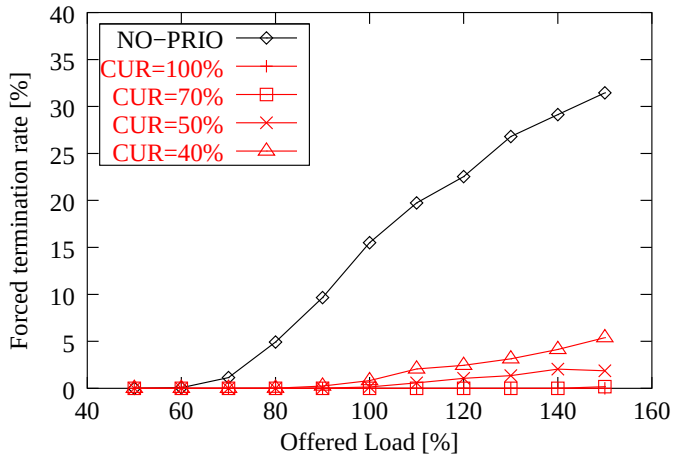

Fig. 16. Directional-Move: Forced termination rate and the CUR parameter

mis-configuration and the easy administrability: The CUR parameter can be decreased step-by-step until the system performance reflects the desired compromise between a low forced termination rate and a high resource utilization.

Furthermore, a slight mis-configuration of the CUR parameter does not influence the system performance badly which underlines the robustness of the CUR parameter.

5.2 Static Random-Move

In the static Random-Move scenario, 75% of the terminals do not move at all, the remaining terminals move with a speed of 17 m/s. In these simulations, the forced termination rate of *SiS-HoP* is zero for CUR parameter values between 20% and 100%. LEP-DAC can also achieve a forced termination rate of zero, but the Cost of Service is higher than for *SiS-HoP*: This is because the CUR parameter is set to 100% for the LEP-DAC scheme to ensure that LEP-DAC can at least provide a forced termination rate of 4% in the previously considered Directional-Move scenario. If the CUR parameter is set below 100%, the simulations of LEP-DAC would lead to an even higher forced termination rate in that scenario. In contrast, a value of 70% for the CUR parameter for *SiS-HoP* achieves a very low forced termination rate in the Directional-Move scenario and in this static Random-Move scenario.

As a result, the blocking rate of LEP-DAC with CUR = 100% is higher than the blocking rate of *SiS-HoP* with CUR = 70% as shown in Figure 17. *SiS-HoP*

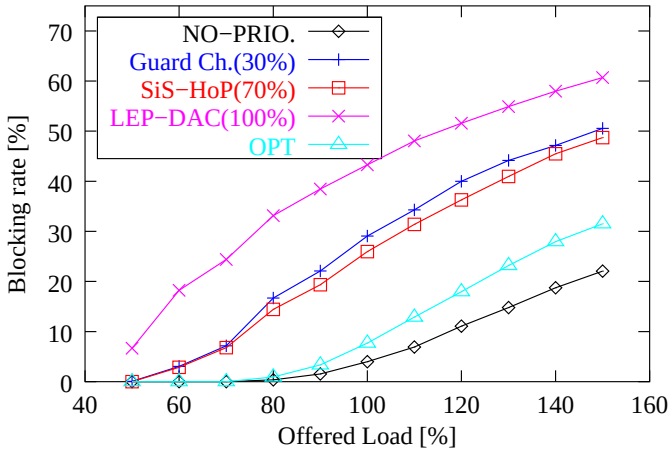

Fig. 17. Static Random-Move: Blocking rate

can adapt its handoff resource amount to the actual demand which is rather low in this static scenario: *SiS-HoP* reserves 27% of the resource utilization as handoff resources which approximates the 22% necessary in the optimal scheme quite accurately compared to LEP-DAC which reserves 72% of the utilized resources for handoff purposes. The high reservation of LEP-DAC is because LEP-DAC does not include a mobility prediction component but assumes an equally distributed mobility pattern instead. As a result, *SiS-HoP* can admit up to 25% more new session requests which reduces the probability

of blocking and leads to a lower Cost of Service than for LEP-DAC in this scenario. In this scenario, *SiS-HoP* can explicitly take advantage of the normalized mobility prediction based on resource holding times: The 75% static terminals have a considerably longer resource holding time than the remaining 25% mobile terminals. Without the normalization, *SiS-HoP* would reserve too many handoff resources and would accept less new session requests (only 15% more than LEP-DAC according to additional simulations not shown here). This is because from the network-wide view, a mobile terminal creates more events to be stored in the mobility caches compared to a static terminal: Static terminals create only a single 'current-cell' entry (when the session terminates) while mobile terminals create additional handoff entries when they perform handoff.

5.3 Random-Move, Speed=17 m/s

For the Random-Move scenario, *SiS-HoP* and LEP-DAC perform similar as expected. This is because the actual mobility pattern is very close to LEP-DAC's assumption of an equally distributed mobility pattern. It has been confirmed by the simulations: *SiS-HoP* can only admit about 3% more new sessions than LEP-DAC because of the different values for the CUR parameter (70% for *SiS-HoP* vs. 100% for LEP-DAC).

5.4 Scenarios with Speed=5 m/s

In the Random-Move scenario with speed=5 m/s, *SiS-HoP* has again a higher performance than LEP-DAC because of its higher accuracy in the resource reservation. As a result, *SiS-HoP* can admit about 8% more new session requests than LEP-DAC. In the Directional-Move scenario, LEP-DAC can admit about 6% more new session requests than *SiS-HoP* because of the LEP-DAC under-reservation effect: At a speed of 17 m/s, LEP-DAC achieves a lower blocking rate at the cost of a higher forced termination rate compared to *SiS-HoP*. In this low-speed scenario, however, the under-reservation of handoff resources does not lead to forced terminations because the number of handoffs per session is rather low (about 1.3 handoffs per session). Therefore, LEP-DAC can achieve a slightly higher performance than *SiS-HoP* in this case because it reserves too few handoff resources and *SiS-HoP* overestimates the necessary handoff resources in this low-speed case.

LEP-DAC cannot reserve a handoff resource amount which corresponds to the actual demand since it does not include a mobility prediction component to adapt to non-uniform mobility pattern. For example, LEP-DAC is not able to avoid forced terminations in the Directional-Move scenario with speed=17 m/s where it cannot provide a high assurance on the handoff success probability for high offered loads. Furthermore, LEP-DAC is not able to adapt to low-mobility scenarios if the CUR parameter remains unchanged. This holds for scenarios such as the static Random-Move scenario or the Random-Move scenario with speed=5 m/s. As a result, *SiS-HoP* can achieve a significantly better performance than LEP-DAC in these scenarios (e.g., 25% additional sessions in the static Random-Move scenario).

Therefore, *SiS-HoP* can either provide a higher assurance on the handoff success probability than LEP-DAC or a lower blocking rate owing to its mobility prediction. It can adapt even to dynamic mobility patterns which vary over time since the CUR parameter does not have to be changed as a response to a change in the mobility pattern. An example for such a change in the mobility pattern is the transition from the Directional-Move scenario with speed=17 m/s to a static Random-Move scenario owing to a traffic jam.

6 Conclusions and Future Work

SiS-HoP is new proposal for a handoff prioritization scheme which can provide a high *assurance on the handoff success probability* for a large variety of mobility patterns. This is ensured by an aggregated mobility prediction between neighboring cells based on a mobility cache. The mobility prediction based on normalized resource holding times and the usage of 'current-cell' entries in the mobility cache enable a high resource efficiency and an easy adaptation to those mobility patterns where many mobile terminals terminate their session in a single cell. The CUR parameter enables an easy tuning of the system performance with regard to the trade-off between the assurance on the handoff success probability and the *efficiency* of resource utilization. Using this CUR parameter is comparable to over-reservations in airline reservation systems, its configuration is intuitive and robust against mis-configuration. *SiS-HoP* has furthermore a high *scalability* because it does not require per-flow state keeping or per-flow signaling to exchange handoff resource reservations between neighboring cells. The *robustness* of *SiS-HoP* is high since there is no per-flow state information to restore after, for example, a failure of a mobile terminal. Additionally, *deployment* of *SiS-HoP* can start incrementally in heavily-loaded areas first in which the base stations can be enhanced with

SiS-HoP-functionality initially.

For future work, the following extension to *SiS-HoP* could be considered. At first, *SiS-HoP* is currently limited to reserve handoff resources between neighboring cells. In principle, it is possible to reserve handoff resources also between distant cells but at the cost of a higher signaling overhead and more state information to be stored in each base station. Second, the CUR parameter could be replaced by more complex schemes to optimize the handoff resources reservation. Third, *SiS-HoP* could be combined with legacy QoS mechanisms such as Differentiated Services [9,43], e.g., in the backbone of the mobile network, to incrementally deploy QoS to further links apart from the bottleneck links. This way, an end-to-end QoS can be achieved between the communicating partners. However, *SiS-HoP* is currently designed focusing on a low complexity, a high scalability, and an easy administrability. Therefore, it has to be verified thoroughly if each of these extensions would lead to a significant performance gain which justifies the necessary additional complexity. For example, an end-to-end scheme might not be appropriate as long as wired backbone networks have capacities which are an order of magnitude higher than the ones of access networks.

A further issue remaining for future work is that a different (i.e., non peak-rate-based) admission control scheme is needed for applications with bursty data traffic, where the average rate and the peak rate differ significantly. In the current scheme, the resource utilization is rather low for these applications. Further mechanism to gain the current bandwidth consumption should be integrated into the admission control scheme in this case, for example, based on measurements.

Information on the Authors

Jörg Diederich received the diploma degree in computer science in 1998 from the Technical University of Braunschweig, Germany, and the doctoral degree from University of Karlsruhe (TH), Germany, in 2002. He was a member of the Institute of Operating Systems and Computer Networks, Technical University of Braunschweig, from February 1999 to September 2003. From October 2003 to June 2004, he was a visiting professor at the Department of Telematic Engineering, Carlos III University of Madrid, Spain. Since July 2004, he is working at the L3S Research Center in Hanover, Germany. His research interests include mobile communication networks, quality of service and trust in communication networks, and eLearning. He is a member of IEEE.

Martina Zitterbart is full professor in computer science at the University of Karlsruhe (TH), Germany. She received her doctoral degree from the University of Karlsruhe in 1990. From 1987 to 1995 she was Research Assistant at the University of Karlsruhe. From 1991–1992 she was on leave of absence as a visiting scientist at the IBM T.J. Watson Research Center, Yorktown-Height, NY. She was visiting professor at the University of Magdeburg and the University of Mannheim and full professor at the Technical University of Braunschweig (1995–2001). Her primary research interests are in the areas of multimedia communication systems, mobile and ubiquitous computing, ambient technologies as well as Elearning. She is member of the IEEE (served on the Board of Governors of the communication society 1995–1998), ACM (treasurer of ACM SIGCOMM) and the German Gesellschaft für Informatik. In 2002 Martina Zitterbart received the Alcatel SEL research award on technical communication.

References

[1] BAKOM: Bundesamt für Kommunikation, Abteilung Telecomdienste, Average voice call durations in Switzerland (1998–2000), Report on Fernmeldestatistik 2000, Biel, Schweiz, in German. (Nov. 2001).

[2] J.-C. Cheng, A. Caro, A. McAuley, S. Baba, Y. Ohba, P. Ramanathan, A QoS Architecture for Future Wireless IP Networks, in: Proceedings of the IASTED International Conference on Parallel and Distributed Computing and Systems (PDCS), Las Vegas, Nevada, USA, 2000.

[3] J. Manner, A. Lopez, A. Mihailovic, H. Velayos, E. Hepworth, Y. Khouaja, Evaluation of Mobility and QoS Interaction, Computer Networks 38 (2) (2002) 137–163.

[4] B. Teitelbaum, J. Sikora, T. Hanss, Quality of Service for Internet2, in: First Internet2 Joint Applications/Engineering QoS Workshop, Santa Clara, CA, USA, 1998, pp. 5–16,
URL: http://www.internet2.edu/qos/may98Workshop/9805-Proceedings.pdf.

[5] R. Braden, D. Clark, S. Shenker, Integrated Services in the Internet Architecture: an Overview, Request for Comments (Informational) 1633, Internet Engineering Task Force (Jun. 1994).

[6] R. Bush, T. Griffin, D. Meyer, Some Internet Architecture Guidelines and Philosophy, Request for Comments (Informational) 3439, Internet Engineering Task Force (Dec. 2002).

[7] S. Floyd, Internet Research: Comments on Formulating the Problem, unpublished manuscript in progress
URL: ftp://ftp.ee.lbl.gov/papers/assumptions.ps (Jan. 1998).

[8] C. Dovrolis, D. Stiliadis, P. Ramanathan, Proportional Differentiated Services: Delay Differentiation and Packet Scheduling, ACM Computer Communication Review 29 (4) (1999) 109–120.

[9] K. Nichols, S. Blake, F. Baker, D. Black, Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers, Request for Comments (Proposed Standard) 2474, Internet Engineering Task Force (Dec. 1998).

[10] D. Clark, The Design Philosophy of the DARPA Internet Protocols, in: SIGCOMM Symposium on Communications Architectures and Protocols, ACM, Stanford, California, USA, 1988, pp. 106–114, also in *ACM Computer Communication Review* 18(4), Aug. 1988 and *ACM Computer Communication Review* 25(1), Jan. 1995, pp. 102–111.

[11] D. Hong, S. Rappaport, Traffic Model and Performance Analysis for Cellular Mobile Radio Telephone Systems with Prioritized and Nonprioritized Handoff Procedures, IEEE Transactions on Vehicular Technology 35 (3) (1986) 77–92, see also: CEAS Technical Report No. 773, June 1, 1999, College of Engineering and Applied Sciences, State University of New York, Stony Brook, NY 11794, USA.

[12] S. Choi, K. Shin, A comparative study of bandwidth reservation and admission control schemes in QoS-sensitive cellular networks, Wireless Networks 6 (4) (2000) 289–305.

[13] C. Leong, W. Zhuang, Call admission control for wireless personal communications, Computer Communications 26 (2003) 522–541.

[14] C. Choi, M. Kim, T. Kim, S. Kim, Adaptive Bandwidth Reservation Mechanism Using Mobility Probability in Mobile Multimedia Computing Environment, in: IEEE Conference on Local Computer Networks, Tampa, Florida, USA, 2000, pp. 76–85.

[15] S. Choi, K. Shin, Predictive and Adaptive Bandwidth Reservation for Hand-Offs in QoS-Sensitive Cellular Networks, in: SIGCOMM Symposium on Communications Architectures and Protocols, Vancouver, British Columbia, Canada, 1998, pp. 155–166.

[16] S. Ganguly, D. Niculescu, B. Vickers, Dynamic QoS Provisioning in Wireless Data Networks, in: Proc. of the IEEE Semiannual Vehicular Technology Conference (VTC), Rhodos, Greece, 2001, pp. 2172–2176.

[17] B. Sadeghi, E. Knightly, Architecture and Algorithms for Scalable Mobile QoS, Wireless Networks 9 (1) (2003) 7–20.

[18] J. Peha, A. Sutivong, Admission Control Algorithms for Cellular Systems, Wireless Networks 7 (2) (2001) 117–125.

[19] J. Diederich, Simple and Scalable Quality of Service for Wireless Mobile Networks, Shaker Verlag, Aachen, Germany, 2003, doctoral thesis, University of Karlsruhe.

[20] W. Soh, H. Kim, Dynamic Guard Bandwidth Scheme for Wireless Broadband Networks, in: Proceedings of the Conference on Computer Communications (IEEE Infocom), Anchorage, Alaska, USA, 2001, pp. 572–581.

[21] K. Lee, Supporting mobile multimedia in integrated services networks, Wireless Networks 2 (2) (1996) 205–217.

[22] C. Oliveira, J. Kim, T. Suda, An Adaptive Bandwidth Reservation Scheme for High-Speed Multimedia Wireless Networks, IEEE Journal on Selected Areas in Communications 16 (6) (1998) 858–871.

[23] R. Jayaram, N. Kakani, S. Das, S. Sen, A Call Admission and Control Scheme for Quality-of-Service (QoS) Provisioning in Next Generation Wireless Networks, Wireless Networks 6 (1) (2000) 17–30.

[24] B. Li, L. Yin, K. Wong, S. Wu, An Efficient and Adaptive Bandwidth Allocation Scheme for Mobile Wireless Networks Using an On-line Local Estimation Technique, Wireless Networks 7 (2) (2001) 107–116.

[25] H. Zeng, I. Chlamtac, Adaptive Guard Channel Allocation and Blocking Probability Estimation in PCS Networks, Computer Networks 43 (2003) 163–176.

[26] Y. Zhang, D. Liu, An Adaptive Algorithm for Call Admission Control in Wireless Networks, in: Proceedings of the IEEE Conference on Global Communications (GLOBECOM), San Antonio, Texas, USA, 2001, pp. 3628–3632.

[27] A. Talukdar, B. Badrinath, A. Acharya, Integrated Services Packet Networks with Mobile Hosts: Architecture and Performance, Wireless Networks 5 (2) (1999) 111–124.

[28] D. Levine, I. Akyildiz, M. Naghshineh, A Resource Estimation and Call Admission Algorithm for Wireless Multimedia Networks Using the Shadow Cluster Concept, IEEE/ACM Transactions on Networking 5 (1) (1997) 1–12.

[29] S. Chung, J. Lee, Mobility-dependent call admission control in hierarchical cellular networks, Computer Communications 25 (2002) 700–713.

[30] K. E. M. (ed), Low Latency Handoffs in Mobile IPv4, Internet Draft, Internet Engineering Task Force, work in progress. (Jan. 2004).

[31] C. Perkins, IP Mobility Support for IPv4, Request for Comments (Proposed Standard) 3344, Internet Engineering Task Force (Aug. 2002).

[32] F. Chiussi, D. Khotimsky, S. Krishnan, Mobility Management in Third-Generation All-IP Networks, IEEE Communications Magazine 40 (2002) 124–135.

[33] Y. Cheng, W. Zhuang, DiffServ Resource Allocation for Fast Handoff in Wireless Mobile Internet , IEEE Communications Magazine (2002) 130–136.

[34] J. Hillebrand, C. Prehofer, R. Bless, M. Zitterbart, Quality-of-Service Signaling for Next-Generation IP-Based Mobile Networks, IEEE Communications Magazine (2004) 72–79.

[35] K. Pahlavan, P. Krishnamurty, A. Hatami, M. Ylianttila, J.-P. Makela, R. Pichna, J. Vallström, Handoff in hybrid mobile data networks, IEEE Personal Communications Magazine 7 (2000) 34–47.

[36] M. Liebsch, A. Singh, H. Chaskar, D. Funato, E. Shim, Candidate Access Router Discovery, Internet Draft, Internet Engineering Task Force, work in progress. (Dec. 2003).

[37] The UCB/LBNL/VINT Network Simulator - ns (version 2), UC Berkeley, LBL, USC/ISI, and Xerox PARC, URL: http://www.isi.edu/nsnam/ns/.

[38] N. Tripathi, J. Reed, H. VanLandingham, Handoff in Cellular Systems, IEEE Personal Communications Magazine 5 (6) (1998) 26–37.

[39] J. Heidemann, N. Bulusu, J. Elson, C. Intanagonwiwat, K. Lan, Y. Xu, W. Ye, D. Estrin, R. Govindan, Effects of detail in wireless network simulation, in: Proc. of the SCS Multiconference on Distributed Simulation, Phoenix, USA, 2001, pp. 3–11.

[40] R. Jain, E. Knightly, A Framework for Design and Evaluation of Admission Control Algorithms in Multi-Service Mobile Networks, in: Proceedings of the Conference on Computer Communications (IEEE Infocom), New York, 1999.

[41] O. Yu, V. Leung, Adaptive Resource Allocation for Prioritized Call Admission over an ATM-Based Wireless PCN, IEEE Journal on Selected Areas in Communications 15 (7) (1997) 1208–1225.

[42] T. Zhang, E. van den Berg, J. Chennikara, P. Agrawal, J.-C. Chen, T. Kodama, Local Predictive Resource Reservation for Handoff in Multimedia Wireless IP Networks, IEEE Journal on Selected Areas in Communications 19 (10) (2001) 1931–1941.

[43] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, W. Weiss, An Architecture for Differentiated Service, Request for Comments (Informational) 2475, Internet Engineering Task Force (Dec. 1998).