Improving wireless network capacity by introducing logical discontinuous coverage

Antoine Luu, Marie-Line Alberi-Morel, Sylvaine Kerboeuf, Roman Ménard, Mazen Tlais, Frédéric Weis

I R I S A
INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES

PUBLICATION
INTERNE
N° 1877

IMPROVING WIRELESS NETWORK CAPACITY BY INTRODUCING LOGICAL DISCONTINUOUS COVERAGE

ANTOINE LUU , MARIE-LINE ALBERI-MOREL , SYLVAINE KERBOEUF , RONAN MÉNARD , MAZEN TLAIS , FRÉDÉRIC WEIS

ISSN 1166-8687

IRISA
CAMPUS UNIVERSITAIRE DE BEAULIEU - 35042 RENNES CEDEX - FRANCE

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES
Campus de Beaulieu – 35042 Rennes Cedex – France
Tél. : (33) 02 99 84 71 00 – Fax : (33) 02 99 84 71 71
http://www.irisa.fr

Improving wireless network capacity by introducing logical discontinuous coverage

Antoine Luu , Marie-Line Alberi-Morel , Sylvaine Kerboeuf , Ronan Ménard , Mazen Tlais , Frédéric Weis

Systèmes communicants
Projet Aces

Publication interne n1877 — January 2008 — 21 pages

Abstract: Mobile cellular networks are now largely deployed, and mobile customers are used to place and to receive calls at any location. The network coverage is continuous, but radio conditions met by users are not homogeneous. Thus, data rate is varying according to user mobility. In this paper, we propose to discriminate terminals by their radio conditions in order to select which data has to be sent. More precisely, terminals are supplied with data only when they are in zones of the cellular network offering good radio quality. With such an approach, continuous cellular networks are exploited in a logical discontinuous way.

In order to hide this *logical discontinuity*, we propose to introduce a *network cache* between the mobile terminal and the content server. Its role is to store incoming data sent by the server even when the terminal is not under good radio conditions. Then, as the terminal enters into a zone of the highest radio throughput (*a transfer area*), it starts storing data in its cache, as fast as the radio conditions allow. Then data are consumed to continue running the service up to the next transfer area.

Our approach is based on a smart scheduling solution (1) that favors data transfer when users have a good radio link, (2) and that regulates the flow of data between the network cache and the terminal according to the available bandwidth. Simulation results highlights the improvements with QoS policies guided by radio conditions for streaming services. Results are promising, a cache-based architecture contributes to a good service quality with no interruption. Furthermore, simulations show that this approach can increase significantly users density supported by the network.

Key-words: Mobile networks, logical discontinuous coverage, caching, scheduling

(Résumé : tsvp)

This paper presents results from a collaboration between the INRIA ACES research team and the MAG project from Alcatel Lucent Bell Labs

Centre National de la Recherche Scientifique
(UMR 6074) Université de Rennes 1 – Insa de Rennes

Institut National de Recherche en Informatique
et en Automatique – unité de recherche de Rennes

Amélioration de la capacité d'un réseau cellulaire via une gestion logique et discontinue de la couverture

Résumé : Les réseaux cellulaires sont maintenant très largement déployés. Ils permettent à des utilisateurs mobiles de recevoir et de lancer des appels, indépendamment de leur localisation. La couverture des réseaux est continue, mais les conditions radio rencontrées par les utilisateurs ne sont pas homogènes. Par voie de conséquence, le débit offert varie en fonction de la mobilité des utilisateurs. Dans cet article, nous proposons de discriminer les terminaux en fonction des conditions radio qu'ils rencontrent, ceci afin de déterminer si des données peuvent leur être transmises. Ainsi, nous proposons d'exploiter des réseaux cellulaires offrant une couverture continue, en introduisant une *discontinuité logique*.

Afin de masquer cette dernière, un *cache réseau* est introduit entre le terminal et un serveur délivrant des données. Le rôle de ce cache intermédiaire est de stocker les données envoyées par le serveur, même lorsque le terminal dispose d'un lien radio de mauvaise qualité. Quand un terminal entre dans une zone offrant un débit élevé (ce que nous appelons une *zone de transfert*), il charge efficacement les données reçues dans un cache interne. Puis il consomme ces données de manière continue, jusqu'à rencontrer à nouveau une zone de transfert.

Au final, l'approche proposée s'appuie sur une politique d'ordonnancement, (1) qui favorise l'envoi des données en direction des utilisateurs rencontrant de bonnes conditions radio, et (2) qui régule l'émission des données entre le cache réseau et le cache terminal en fonction de la bande passante disponible. Nos simulations mettent en évidence les améliorations apportées par des mécanismes de qualité de service utilisés pour des services de type *streaming*, et gouvernés par les conditions radio rencontrées par les terminaux. Les résultats obtenus sont prometteurs, une architecture s'appuyant sur une distribution de caches permet d'obtenir une délivrance de données sans interruption. Enfin, les simulations montrent que notre approche permet au réseau de servir efficacement une densité accrue de terminaux.

Mots clés : Réseaux mobiles, couverture discontinue «logique», gestion de caches, ordonnancement

This paper presents results from a collaboration between INRIA ACES research team and the MAG project from Alcatel Lucent Bell Labs

1 Introduction

Mobile cellular networks are now largely deployed, and mobile customers are used to place and to receive calls at any location. This requires continuous coverage, which in turn requires significant infrastructure. Today, Internet with all its throughput consuming services comes to customer's mobile terminal. So to provide a wireless coverage everywhere without exploding the deployment cost, the coverage of each access point is extended with detriment to the mean throughput of the access point. The network coverage is continuous, but radio conditions met by users are not homogeneous. Thus, data rate is varying according to user mobility. In this paper, we propose to discriminate terminals by their radio conditions in order to select which data has to be sent. More precisely, terminals are supplied with data only when they are in zones of the cellular network offering good radio quality. With such an approach, continuous cellular networks are exploited in a discontinuous way.

This paper argues in section 2 that providing services in a logical discontinuous way may enhance the throughput used over a continuous coverage. Then section 3 describes an architecture to provide services over a logical discontinuous coverage wireless network. It focuses on describing an infrastructure to transport data from the wired network to the mobile terminal. Finally section 4 suggests a QoS policy to distribute data in spite of the logical discontinuous coverage, and shows simulation results.

2 Improve Quality of Experience through user mobility

This section introduces a proposal to improve the *Quality of Experience* (QoE) for a mobile user. Here we present the theory basis of our work. The main principle is to discriminate terminals by their radio conditions in order to select which data will be sent.

Main hypothesis Figure 1 shows the access point radio basic model that will be used for calculus and simulations. The cell radio coverage is assumed to be represented by a set of concentric circles around the access point, associated to different data transfer rates. The throughput of each zone decreases non-uniformly from the cell center to the cell edge. The access point provides packets to terminals by using a round-robin distributing policy. When the users are assumed to be uniformly distributed in the radio cell, the mean throughput of the access point can be calculated by the formula 2(a).

A simple mobility model can be considered (see Figure 3): terminals move straight with a fixed speed v from the center of a cell to another one. The access points deliver data to all users whatever their position. The cell's radio coverage is divided in two areas: a first zone, which covers higher

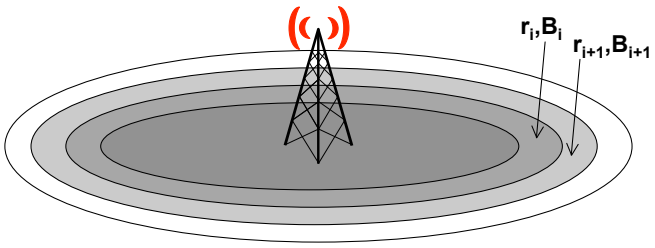


Figure 1: Radio Model

B: Cell mean throughput
B_i: throughput in the area i
r_i: width of the area i

$S_i = \pi.(r_i^2 - r_{i-1}^2)$: surface of the area i $\quad$ $d_i = r_i - r_{i-1}$: width of the ring i

$$B = \frac{\sum_{i=0}^{n} B_i.S_i}{\sum_{i=0}^{n} S_i}$$

(a) Area distribution

$$B = \frac{\sum_{i=0}^{n} B_i.d_i}{\sum_{i=0}^{n} d_i}$$

(b) Linear distribution

Figure 2: Mean Radio Bandwidth of an access point

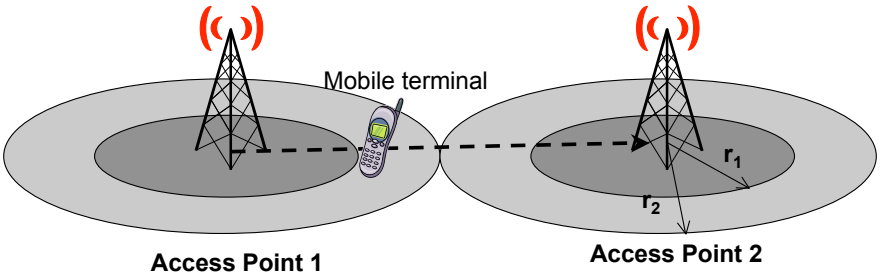


Figure 3: Simple mobility model

radio rate areas; and a second zone, which covers lower radio rate areas[1]. The linear density (number of users/m^2) is noted λ.

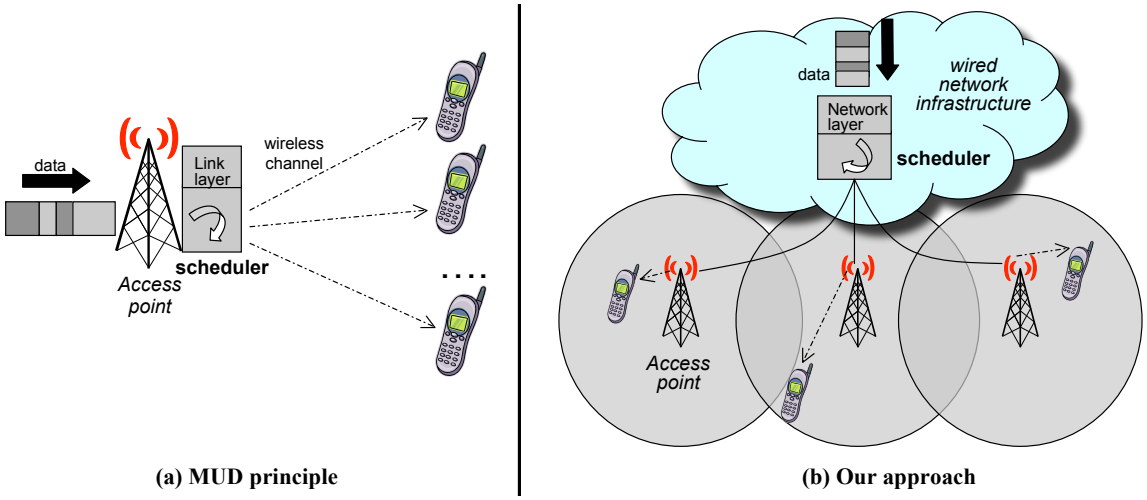
In such a model, the capacity can be evaluated thanks to the formula 2(b) because users are uniformly distributed on a line. So during its travel, the user experiences a throughput that is equal to the cell mean throughput divided by the number of users inside the cell: $bw = \frac{B}{r_2.\lambda}$.

[1]By using the mean bandwidth formula, any n-ring access point model can be converted into a two-ring access point model.

Virtual throughput concept We now propose to consider the case where data are transmitted only to terminals located in the zone of the highest radio rate zones (*i.e.* the nearest zones from the access point). The principle is to concentrate data delivery to terminals placed in the highest throughput areas. So in the worst case when the network is over high load, only terminals that are in these areas receive data. Then a terminal goes through a radio cell during $\frac{r_2}{v}$, but it is provided with data during a limited time equal to $\frac{r_1}{v}$. So the bandwidth «seen» by a user can be represented by *a virtual bandwidth*: $vbw = \frac{B_1}{r_1.\lambda}\frac{\frac{r_1}{v}}{\frac{r_2}{v}} = \frac{B_1}{r_2.\lambda}$.

$\frac{vbw}{bw} = \frac{B_1}{B} > 1$ when when user moves with constant speed along the line, so we can conclude that the proposed distribution policy always enhances the throughput.

Discussion In recent years, new MAC[2] scheduling policies based on the Multi User Diversity[3] principle [5, 8] have been proposed (for example HSDPA technology using scheduler combining both channel state information and service requirements) to mitigate efficiently the radio throughput limitations due to the radio link quality variations. They exploit efficiently the time diversity generated by the radio channel variations and the multiple users. The access point allocates radio resources only to the cell users which have the best radio link quality, that thereby the overall capacity of the cell is increased (see Figure 4(a)).



(a) MUD principle **(b) Our approach**

Figure 4: Data distribution policies

In this paper, we use similar mechanisms, but scheduling policies are implemented *higher* into the wired network infrastructure, in order to distribute data to multiple cells (see Figure 4(b)). However our scheduling policies are be intended to address only long term variations of the radio channel

[2]Media Access Control
[3]MUD

(typically the Path Loss). We propose to deliver data *discontinuously*, only into highest throughput rate areas of the cells, and to avoid transferring data to terminals in lowest radio rate areas. As a consequence the mean throughput offered to a user is increased when he moves accross the network. With such an approach, a wireless continuous network can be seen as a *logical discontinuous network*.

In the next section we describe the network architecture to provide services over a logical discontinuous coverage.

3 Exploiting a logical discontinuous coverage network

Global wireless networks are more and more connected to the Internet. Now such networks use IP protocol, that means a terminal can address directly any server on the Internet without any translation as it is done when using the WAP[4] stack. To provide seamless and no degraded services over a logical discontinuous network, the way the data are transported from the content server to the mobile terminal (MT) needs to be modified. While the MT is under good radio conditions, the objective is to supply data to the terminal as effectively as possible. The approach relies on cache mechanisms that enable data storage in advance, and data delivery at very high rates, subject to bandwidth availability.

3.1 A cache-based network architecture

In order to hide the logical discontinuity, an entity must be placed between the mobile terminal and the content server. Its role is to store incoming data sent by the server even when the MT is not under good radio conditions. Then, as the terminal enters into a zone of the highest radio throughput (*a transfer area*), it starts storing data in its cache, as fast as the radio conditions allow. Then data are consumed to continue running the service up to the next transfer area.

Introduction of an intermediate entity The content server, located into the wired side of the Internet, remains unchanged. The mobile terminal communicates with other equipments of the network through a wireless connection to the nearest access point (AP). The latter is the last equipment of the wired network before the wireless link. It could be candidate for implementing all these tasks (data storage and distribution). It knows the available bandwidth of any terminal it manages. So it can feed the terminal as fast as the radio conditions allow. Nevertheless setting such mechanisms into the access point creates many problems. One of them is the deployment of such an architecture, that requires to implement new functionalities in all access points. Another problem is related to the mobility management. Indeed a user moves from one access point to another. So data sent by the server and not consumed by the user, have to follow the terminal, and as a consequence must be moved from the first access point to the next one. Current cellular APs are not designed to perform that. For all these reasons the access point is not the good candidate to store and distribute data sent by the content server.

[4] Wireless Application Protocol

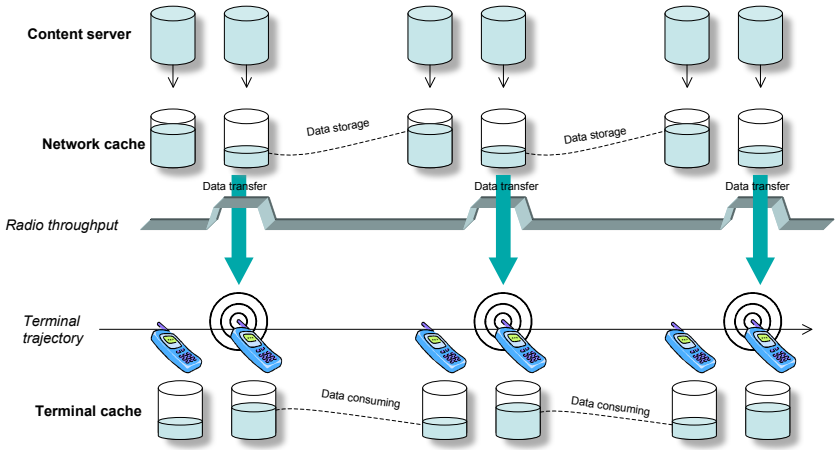


Figure 5: Cache solution elements

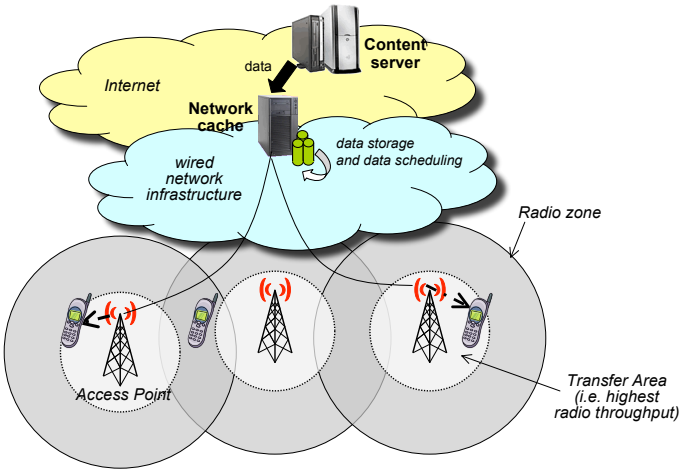


Figure 6: Network Architecture

Since there is no existing entities suitable to provide all conditions to store and to distribute efficiently data, a new equipment must be inserted into the wired network infrastructure. In the rest of this paper this equipment is called the *network cache*. In our view it is able (1) to store data from the content server when the terminal is out-of-coverage and (2) to provide data to MTs when they enter in transfer areas. This cache-based approach is illustrated by Figure 5.

Position of the network cache Moreover, mobile terminals can move from one IP subnet to another one. Then a layer 3 mobility system must be used in the network to avoid disconnections or service disruptions. To be able to reach MTs whatever their locations in the network, the network cache must be co-located with (or located above) the layer 3 device in charge of the mobility system. For example, in WiMax 802.16e architecture, this device is the MAP (Mobility Anchor Point) [2]. Figure 6 gives an overwiew of this network architecture.

Cache-based layer In such an architecture, the delay induced by storing data may introduce some problems like timeout if the content server is waiting for acknowledgements from the client. To solve this problem, the network cache acts as an *application proxy*, so it can be seen as a *virtual client* of the content server.

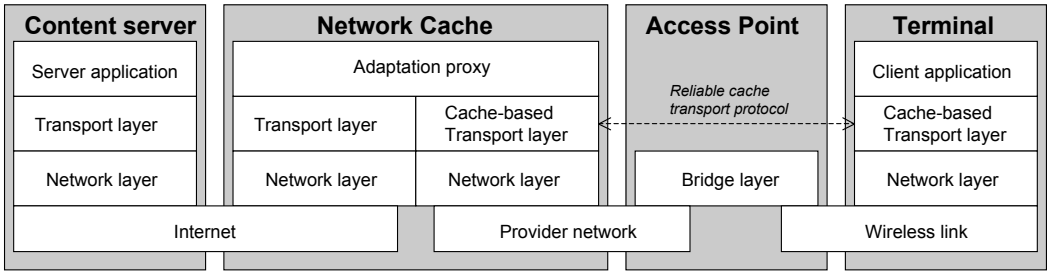


Figure 7: Proposed protocol stack

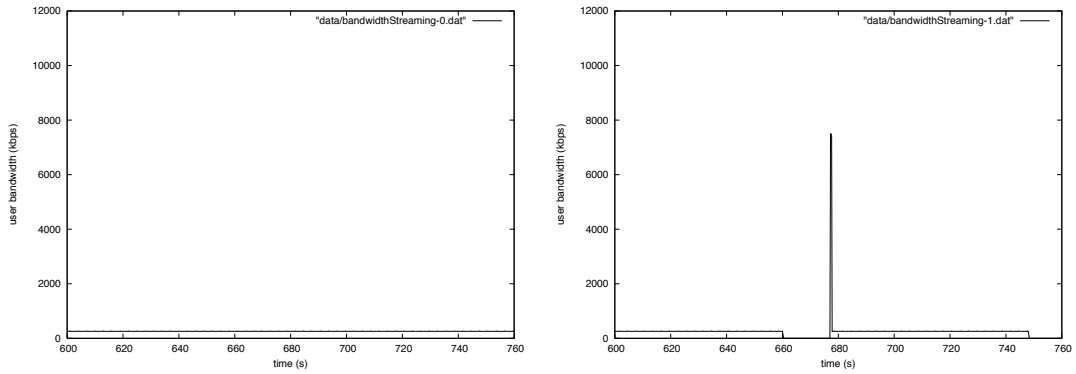
Optionally the network cache can provide some adaptation mechanisms such as changing the media resolution to fit the terminal screen or performing data prefetch to reduce the access time to a service.

As it is shown in Figure 7, a network cache and each terminal implement a specific transport layer protocol to provide seamless services. The main features of the protocol are discussed in this section. Moreover, the main goal of this architecture is to enhance the usage of the access point bandwidth. We show that it requires flow control, bandwidth estimation and quality of services mechanisms. For instance since all data go through a network cache which has the knowledge of the user location, the latter has all information required to organize (*i.e.* to schedule) data flows to be sent to each access point. So it implements a specific component to manage the bandwidth of each access point independently. By this way the network cache is able to avoid congestion into the access points.

Finally the proposed architecture requires a few modification of existing infrastructure to provide service over a logical discontinuous coverage network. Only the transport layer of terminals need to be modified, and a new equipment (the network cache) must be inserted into the wired network. Now we focus on the communication protocol between the network cache and the mobile terminal, and on the mechanisms to manage the access point bandwith.

3.2 A discontinuous compliant transport protocol

The discontinuity of the coverage modifies the traditional traffic shape of a service. For instance, a data transfer can be assumed to be characterized by a fixed mean throughput during a period of time as it is shown by the figure 8(a).



(a) With continuous coverage

(b) With logical discontinuous coverage

Figure 8: Bandwidth evolution of user during its travel

But in a discontinuous context (see Figure 8(b)), the user throughput is null while the mobile user is out-of-cover. When the terminal arrives into a communication area, a burst occurs to provide all data delayed during it was out-of-cover. Then the mean throughput becomes constant, at the same value as the server can provide until the terminal goes out.

As a consequence, services which used to be implemented over an unreliable transport protocol, such as video streaming, must at least be built on a partial-reliable one to avoid congestion after a discontinuity. Indeed the burst of data may create a high packet loss rate due to congestions. This problem is discussed in the following of the section.

3.3 Managing the bandwidth of an access point

The cache-based transport protocol regulates the flow of data between the network cache and the cache in the terminal according to the available bandwidth. It must be *reliable enough* to ensure that required data are received by the remote host. For this reason, it must inherited mechanisms implemented in other transport protocols like TCP or SCTP [10, 3]: delayed acknowledgement timer, retransmission timer, bandwidth estimation and flow control through a sliding window management. PR-SCTP, an extension of SCTP implements partial-reliability [11, 4] which allows to drop out-of-date packets, so it is more suited than TCP for transport time-dependent services like video streaming.

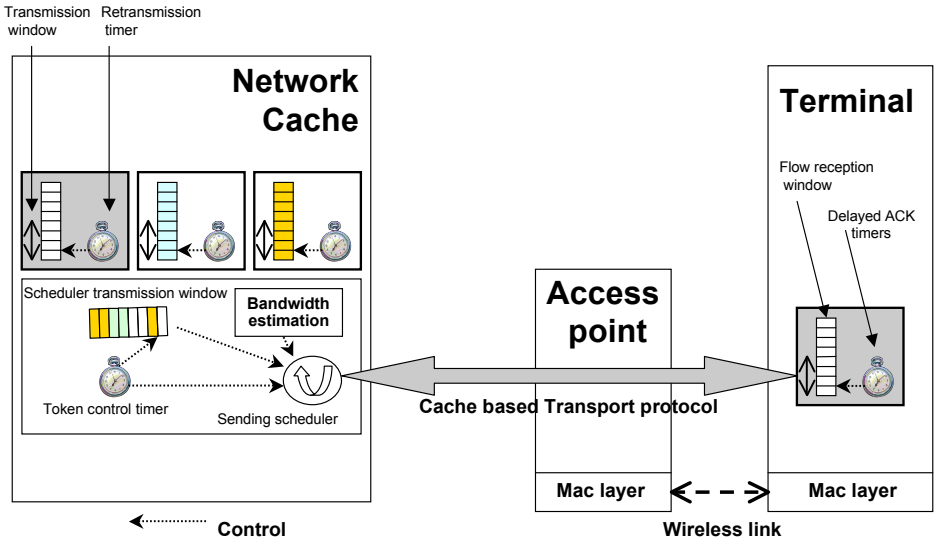


Figure 9: Flow control functionality

Much work has been done about bandwidth evaluation using TCP standard acknowledgements in the specific context of wireless communications (Westwood [6], TIBET [1]). Our model of transport layer uses some added information into the acknowledgement. In TCP, acknowledgements do not carry the exact size of the received data; some packets are not reported, especially duplicated packets or corrupted ones. Our cache-based transport protocol uses specific acknowledgments, which gives the size of the data received from the network layer. When the network cache receives an acknowledgement, it knows exactly the amount of data received by the terminal. If the estimator evaluates the amount of data received between two timestamps, then it can estimate the bandwidth accurately.

As it is shown in Figure 9, an access point is attached to a given scheduler in the network cache. Data scheduled for terminals in the transmission window are sent at the estimated cell bandwidth. This prevents any data overflow in the access point, which is the throughput bottleneck. Each scheduler evaluates the bandwidth capability of the access point and manages the bandwidth used to avoid congestion into the access point. Moreover, due to the control of the evaluator over the scheduler, a low frequency filter must be applied into the evaluator to avoid instability of the sending mechanism. Here we used a *Westwood* [6] like evaluation. The latter is computed according to the following formula:

$$Bw_i = \frac{19}{21} Bw_{i-1} + \frac{2}{21} \frac{\sum sizeof(Ack)}{\delta t}$$

where *sizeof(Ack)* is the amount of data that has been acknowledged during the evaluation period δt. The low-pass filter is used to average the evaluation in order to avoid drastic changes due to burst

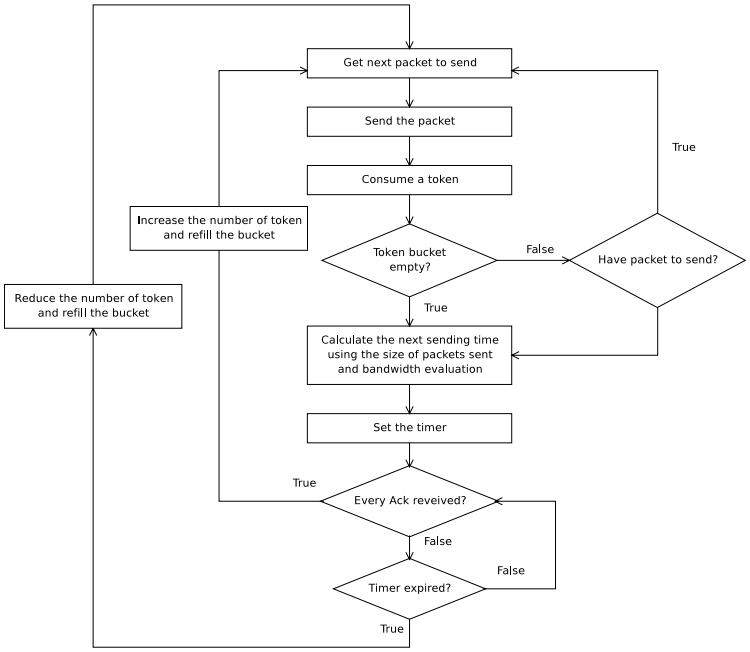


Figure 10: Flow control using token bucket

arrival of acknowledgements, or the lack of acknowledgements during one δt due to delayed acks (lower bandwidth), which could falsify the evaluation.

3.3.1 Token bucket like scheduler and retransmission timer

A token bucket is used to control the transmission window associated with the scheduler (see Figure 9). As shown in Figure 10, its size automatically grows up when every sent packet is acknowledged and it decreases when the idle timer is expired. The token bucket is fed when an acknowledgement has been received or when the idle timer has expired. The bucket size and the timer duration must be limited to avoid deny of service by setting a too long timer.

3.3.2 Flow window, retransmission timer and acknowledgement timer

As in TCP, the two agents communicating must provide mechanisms to improve reliability for the data transport. The receiver has a receiving window which defines the maximal amount of packets it can store. This window must be larger or equal to the sending one. To maintain this constraint the receiver must notify the sender of its window size inside the acknowledgement packet. Moreover the sender can reduce the sending window size if it detects a packet loss.

The sending agent detects a packet loss when the static retransmission timer expires, or some packets, which should be acknowledged before this timer expires, were not acknowledged. The acknowledgment timer is used to reduce the number of acknowledgements to be transmitted to the sender.

4 Design and evaluation of Quality of Service mechanisms

As it is highlighted in section 3.1, to enhance the cell's capacity, the architecture must use a scheduling policy based on the radio link quality of the terminal. This problem is discussed in this section.

4.1 QoS Policy for data transfer service

Policy overview The scheduling decision is based on the *radio conditions* encountered by the terminal. So an equipment in the architecture must send this information to the network cache. Both access point and terminal are able to characterize the radio conditions. The terminal was chosen, because we want to let unchanged as many devices as possible. For our simulations, we use a «WiMax-like» radio coverage model represented in Figure 11.

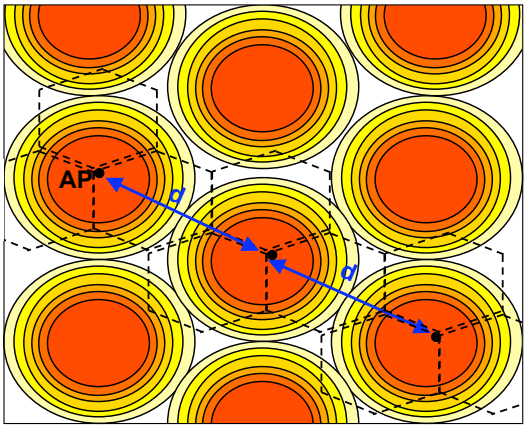


Figure 11: Radio coverage model

The radio coverage (*i.e.* distribution statistic of the peak radio throughput) was determined using a WiMAX radio coverage tool for a given deployment in a dense urban and outdoor context for the different Physical transmission modes assumed for Wimax (30.0 Mbps-26.67 Mbps-20.00 Mbps-13.33 Mbps-10.0 Mbps-6.67 Mbps). In our study, each cell is modeled in a basic way by concentric circles of 6 radio zones. The APs ensure the different radio throughput over different distances as illustrated in Figure 12.

The data transfer classifing policy (see Figure 13) is designed as follows: data flows are classified into two queues. The first one, with the highest priority, contains flows for terminals which are into

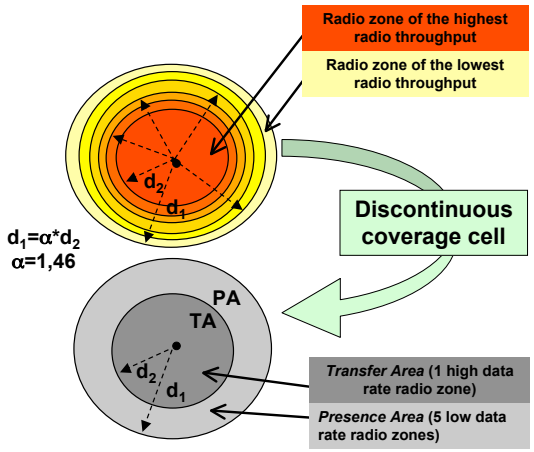


Figure 12: Non uniform coverage cell

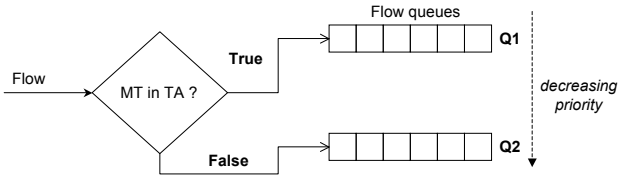


Figure 13: Data transfer QoS classifier

the *transfer area* (TA); the second one manages flows for terminals located in the *presence area* (PA). The scheduling policy provides packets of flows registered in the first queue until there is no more data to be sent. Then it transmits data to terminals located into the PA. To be fair into a given queue, the scheduler uses a round-robin on every registered flow.

Simulation and evaluation We have used the DesmoJ (Discrete-Event Simulation and MOdelling in Java) environment [7] to implement and execute our simulations. The latter is a Java Framework for discrete-event modelling. It offers a set of ready-to-use classes for model components like queues, data collectors, and a simulation infrastructure comprising scheduler, event list, and simulation time clock. We have developed all the needed components, network cache, AP, MT, mobility models, wireless links and wired links, according to the topology presented in Figures 11 and 12.

The features of simulation are as following:

- The test-bed spans an area of 1558 x 900 meters,
- the simulated time is 4200 seconds,

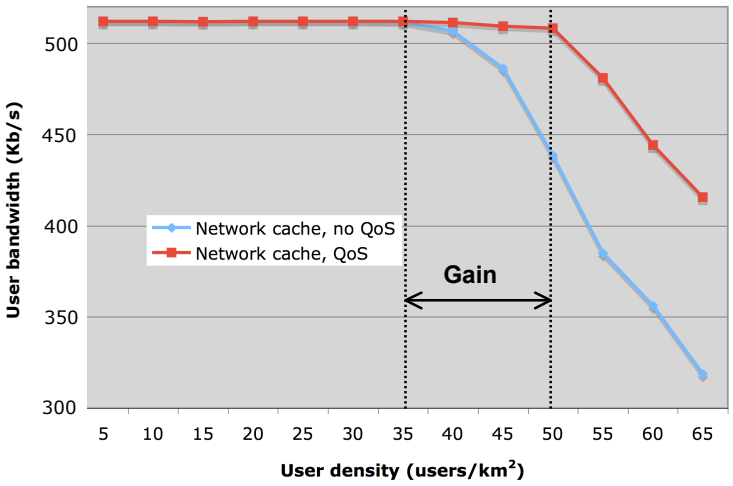


Figure 14: Bandwidth evolution with the user density

- the server content rate is 512 Kb/s,
- the user mobility profile is individual and random, with a fixed speed of 0.83 m/s,
- finally, in the AP, the MAC layer is implemented as a Frame based round-robin.

Our goal is to study the impact of the scheduling policy. Figure 14 shows the evolution of the bandwidth «seen» by users when the users density increases. We can determine the user density value from which the service is degraded (*i.e.* the value from which data cannot be delivered to the user at the rate imposed by the content server). Two cases are considered, the first one uses the proposed architecture but without any QoS policy. This case can be considered as a «witness policy». The second one uses the proposed QoS policy. The main result is that the QoS policy increases the network performance when the network is loaded. The gain between the both cases (with and without QoS) is about **150%** of user/km^2.

So the way data are scheduled and distributed improves significantly the bandwidth used when the network is over high load. But some applications which require bandwidth and regular data providing (for example streaming applications) need a more sophisticated distribution policy to provide a good User Experience. This point is discussed in the next section.

4.2 Extension of the policy to deal with streaming applications

Applications, such as data transfers (file transfer for instance), can work in spite of jitter and delayed transmissions introduced by the coverage discontinuity. But it is not the case for time-constraint services like video streaming. Streaming data must be received continuously and with a fixed deadline.

RTP (Real Time Protocol) [9] was developed to hide the network jitter in best effort networks, like IP infrastructures. It uses buffers and timestamps.

Discontinuous coverage introduces a similar jitter at a higher scale. Indeed the mobile terminal is not able to receive data continuously. That is why it must exploit the coverage to store as much data as possible, and to be able to read them offline. For these reasons, two levels of data storage are required.

The first one is into the network cache to hide discontinuity to the content server and to avoid data losses when a terminal is out-of-cover. The second one, like in existing streaming players, is implemented into the terminal with a storage capacity (*i.e.* an internal cache).

Definition of the ToC parameter The first impact of the coverage discontinuity is that the user cannot get any data when he is out-of-cover. Moreover, depending on the access points deployment and on the mobility model, it might not be possible to know *a priori* the value of the out-of-cover duration. We define the *Time Out-of-Cover* (ToC) as the longest time for which the service is satisfied. This time corresponds to an amount of data[5] that the terminal must store in its cache to guarantee a continuous data delivery to the displayer. So if the user remains out-of-cover more than the ToC value, service disruptions may appear. Using the same mobility model and the same access point topology as in the section 4.1, we evaluate that 99,9% of the «out-of-cover times» are under of 240 s. So the value of the ToC parameter is set to 240 s.

Starting of the streaming service The first requirement of the caching mechanism is to ensure no interruption during the service duration. To garantee a continuous service, mobile terminal caches should be filled to *full level* with data when a user starts the service. *Full level* is the amount of data required for a mobile user to be autonomous when he is out-of cover. It corresponds to ToC = 240 s of data in its cache. This approach is not realistic because it leads to significant service access time.

So in a first step, we consider that an application starts when a threshold of data equivalent to 30 s of streaming is reached in the terminal (*start level*), which is a typical size of an internal buffer for a streaming service. The principle of these two levels (*start* and *full*) is illustrated in Figure 15.

We evaluate the impact of this delayed start by counting the number of service disruptions. We consider that a service disruption occurs when a packet is delivered to the application layer 200 ms later, or when 200 ms of RTP packet are lost. Our simulation results are presented in Figure 16. Using the QoS policy presented in the section 4.1, we evaluate the performances of (1) a unreliable *UDP like* transport layer, and (2) the cache-based transport layer (presented in the sections 3.2 and 3.3) associated with an application starting level equivalent to 30 s of streaming.

Results highlight that when an unreliable way to deliver data is used, the first service disruption appears for a density of 20 users/km^2 (curve 1). When the cache-based protocol is introduced and if at the start of the service, the terminal is fed with 30 s of prefetch data, the network capacity is improved of about 100% (curve 2).

Introduction of a burst The network cache needs to provide as quickly as possible the amount of data required at first to start the applications (*i.e. internal cache level* = 30 s * *Service Data Rate*)

[5] *Data Size = Service data rate * ToC*

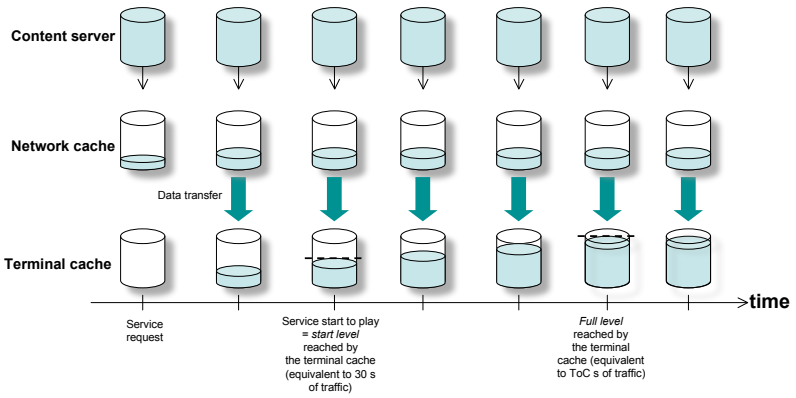


Figure 15: Starting of the streaming service

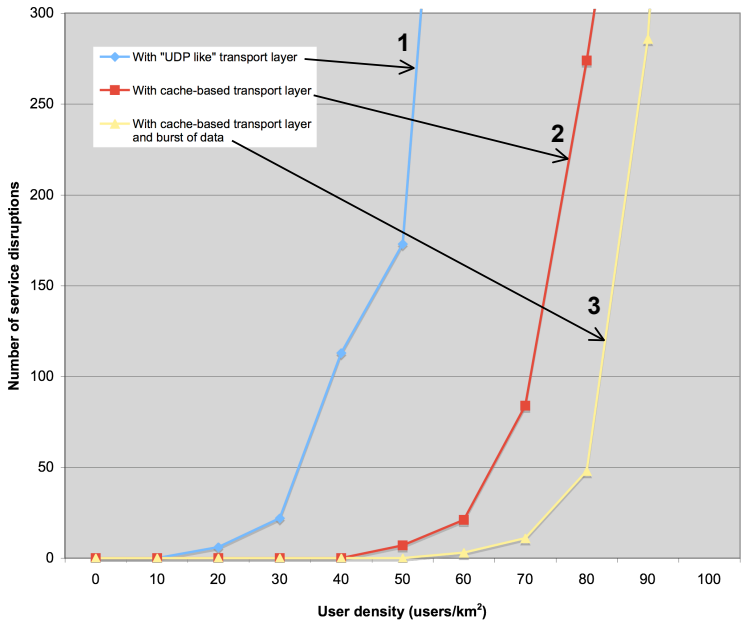


Figure 16: Evolution of service disruptions using a transport protocol and delayed service start

and secondly to minimize service disruptions (*i.e. internal cache level* = ToC * *Service Data Rate*) as soon as the application has begun to consume data in the terminal. To reach these goals we propose to introduce a long data burst. For the streaming service, the network cache is able to retrieve the beginning of the stream two times faster than the normal service rate. As the terminal is provided

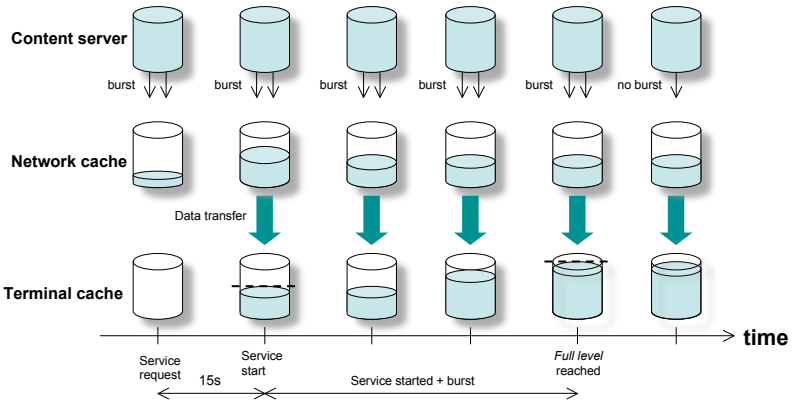


Figure 17: Introduction of a data burst

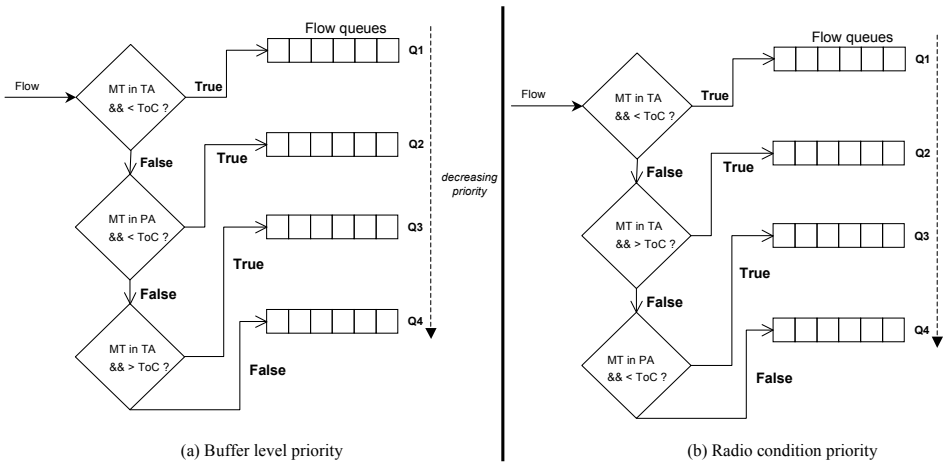


Figure 18: QoS Classifers rules for streaming application

with a doubled data rate, the time necessary to start the streaming application decreases from 30 s to 15 s. Moreover this burst period is extended after the start of the service. While consuming data, the terminal continues to store its internal cache, until the latter reaches the «ToC value». After that, the burst phase is stopped, and data are sent by the network cache to the terminal at a normal throughput rate. This principle is illustrated by Figure 17.

Results of the *burst policy* are given in Figure 16 When the service start is delayed of 30 s, the first service disruption appears for a density of 40 users/km^2 (curve number 2). If the mechanism is extended with a burst that allows feeding the terminal with 240 s of prefetch data, the network

capacity is improved of about 50% (curve number 3). Finally, a streaming application is more robust if it uses smarter transport layer than UDP and a burst of data sent by the content server.

QoS policies for streaming applications Introducing the ToC parameter, we now propose to extend the previous QoS policy presented in the section 4.1. Two classification policies are defined (see Figures 18) with four priority queues on each of them.The ST_a policy (see Figure 18(a)) puts in the highest priority queue $Q1$ flows for which the duration of data inside the internal storage of targeted terminal[6] is lower than the ToC and for which the terminal is in the TA. In the second priority queue $Q2$, this policy puts flows for terminals in the PA and where the duration of data available in the terminal is under ToC. In the third priority queue $Q3$, it pushes flows of other terminals in the TA. And finally $Q4$ contains the rest of terminals.

The ST_b policy (see Figure 18(b)) is the same policy as the previous one except that the rules to feed $Q2$ and $Q3$ are switched. The ST_a policy favors the level of the internal cache (characterized by the condition $DataSize\ in\ cache > ToC$), whereas the$ST_b$ policy gives highest priority to optimal radio conditions (data are transfered to MTs in TAs).

To distribute data, the scheduler is working in the same way as for a data transfer service, but using the four priority queues.

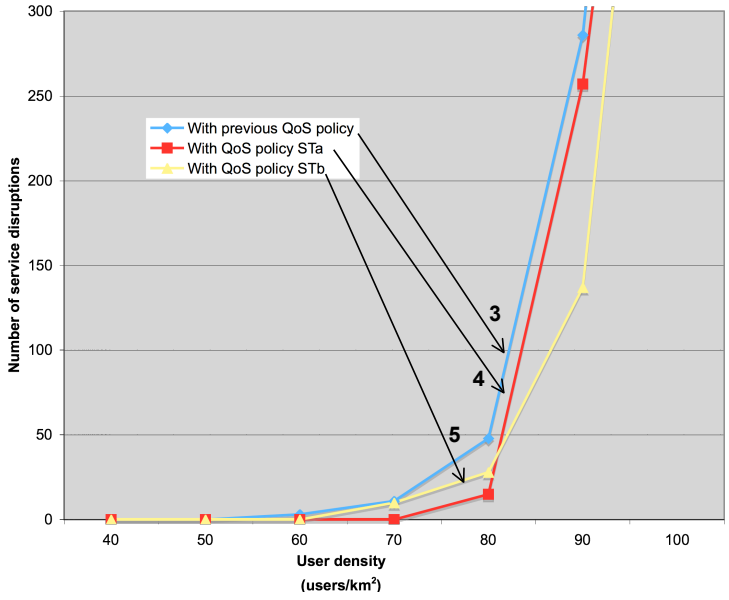


Figure 19: Evolution of service disruptions using QoS policies based on the ToC value

[6] $DataSize/ServiceDataRate$

Figure 19 illustrates the evolution of service disruptions in function of the users density for QoS policies ST_a and ST_b. It shows that both policies allow to increase the number of users into the network of 20 users/km^2 for ST_a and 10 users/km^2 for ST_b. Using the second policy ST_b reduces the number of service disruptions when the network is loaded. But it is less efficient than the first policy ST_a on the edge of the network capacity. So different QoS policies must be used for data transfer and for video streaming. For data-oriented services, it is better to transfer data only in the TAs of the network. An for streaming services, it is important to play the video as quick as possible, a policy based on the size of data available in the terminal cache must be used to avoid service disruptions.

5 Conclusion

The main idea of this paper is to argue that providing services through a logical discontinuous view to mobile terminals may enhance the global capacity of a cellular network. First we have presented the theorical reasons to improve the wireless network in a mobile environment by introducing logical discontinuity. Then we have presented the required mechanisms to provide service into a logical discontinuous coverage network. We have proposed to concentrate data delivery to mobile users with good radio conditions.

Our approach is based on a smart caching solution that favors data transfer when users have a good radio link. Simulation results have highlighted the improvements with QoS policies guided by radio conditions for streaming services. Results are promising, a cache-based architecture contributes to a good service quality with no interruption. Furthermore, simulations have shown that this approach can increase significantly users density supported by the network.

Our future works will be to evaluate a real transport cache protocole based on SCTP protocol.

References

[1] A. Capone, L. Fratta, and F. Martignon. Bandwidth estimation schemes for tcp over wireless networks. *IEEE Transaction on Mobile Computing*, 3(2):129–143, April 2004.

[2] C. Chang. A mobile-ip based mobility system for wireless metropolitan area network. In *Proceedings of the IEEE International Conference on Parallel Processing Workshops (ICPPW'05)*, June 2005.

[3] S. Fu and M. Atiquzzaman. Sctp: State of the art in research, products, and technical challenges. *IEEE Communication Magazine*, April 2004.

[4] H. Huang, J. Ou, and D. Zhang. Efficient multimedia transmission in mobile network by using pr-sctp. In *Proceedings of the IASTED Communications and Computer Networks conference CCN 2005*, October 2005.

[5] R. Knopp and P. A. Humblet. Information capacity and power control in single-cell multiuser communications. In *Proceedings of the IEEE ICC'95*, June 1995.

[6] S. Mascolo, M.Y. Sanadidi, C. Casetti, M. Gerla, and R. Wang. Tcp westwood: End-to-end congestion control for wired wireless networks. *Wireless Networks*, 8:467–479, 2002.

[7] University of Hamburg. A framework for discrete-event modelling and simulation. available at http://asi-www.informatik.uni-hamburg.de/desmoj, 2006.

[8] X. Qin and R. Berry. Exploiting multiuser diversity for medium access control in wireless networks. In *Proceedings of the 22nd IEEE Infocom conference*, March 2003.

[9] H. Schulzrinne and al. Rtp: A transport protocol for real-time applications. RFC 1889, January 1996.

[10] R. Stewart and al. Stream control transport protocol. RFC 2960, October 2000.

[11] R. Stewart and al. Streamcontrol transport protocol (sctp) - partial reliability extension. RFC 3578, May 2004.

Contents