Disruption Tolerant Networking by Smart Caching

Journal:	*International Journal of Communication Systems*
Manuscript ID:	IJCS-08-0075.R2
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	
Complete List of Authors:	Goebbels, Stephan; RWTH Aachen, Faculty 6
Keywords:	Disruption Tolerance, Smart Caching, Intermittent Network Coverage, Extensive Buffering, Pre-fetching

Disruption Tolerant Networking by Smart Caching

Stephan Goebbels
Chair of Communication Networks
Faculty 6, RWTH Aachen University
sgs@comnets.rwth-aachen.de

Abstract: Future mobile radio networks will aim to achieve "broadband access for all", anywhere. The performance of a radio network vitally depends on the characteristics of the transmission path between the user terminal and the Access Point and the degree of network coverage. In urban areas, full broadband radio coverage is difficult to provide, causing a high variation in the link quality and making broadband services hard to realize. In rural regions, massive deployment costs prevent a full broadband coverage. Most of the time users have to settle for UMTS-like wide area networks. For mobile users accessing services such as video streaming, which require continuous broadband connectivity, it virtually results in intermittent network connectivity. The frequent disruption of the broadband link and its replacement with no or only low-performance connections is a problem that should be addressed. This article introduces a new technique called Smart Caching, which is able to mitigate variations in the network performance so that non-real-time and non-interactive services' quality is substantially improved.

Smart Caching supports pre-fetching from a server and buffering data at the edge of the core network, in the so called Smart Cache. It transmits data with extremely high speed to be buffered in the mobile terminal when it is in the service range of an Access Point. This allows for the provisioning of data-intensive services even in the case of patchy wireless broadband network coverage and intermittent connectivity.

The performance of the Smart Caching service is evaluated with two different sophisticated queuing models, both based on the Markov Arrival Process. The benefit of the new technique is discussed and dimensioning issues are outlined. Furthermore, a comparison with legacy network setups is given.

Keywords:

Disruption Tolerance, Smart Caching, Intermittent Network Coverage, Extensive Buffering, and Pre-fetching.

1. Introduction

Nowadays ubiquitous telecommunication access is no longer a dream. Mobile users have Internet access wherever they are. But the connectivity is mainly provided by wide area networks like UMTS which offer a peak data rate of only up to 384 kbit/s. In contrast to that in so-called Hot Spot zones broadband Internet is provided by wireless networks like WiMAX. These allow link throughput capacities of several tens of Mbit/s. In addition, the provided data rate always depends on the distance between the access node and the user terminal. Thus, peak data rates are only provided close to the access node – but at the edge of the coverage area the achievable throughput drops to a much lower level. All this results in a highly heterogeneous network coverage. In urban environments, the attenuation of radio waves and shadowing due to obstacles further reduce the coverage range of broadband networks. A mobile user moving in a heterogeneous network such as that shown in the left picture of Figure 1 perceives a continuously changing link performance. Especially if the user tries to access broadband services the data rate in the UMTS network is insufficient. Therefore, the coverage for services with high data rate requirements is virtually reduced to areas of wireless broadband coverage, as shown in the right picture of Figure 1.

In rural areas, the realization of ubiquitous wireless broadband network coverage fails because of high deployment costs. However, for intensively frequented regions such as motorways at least a partial broadband coverage is achievable, e.g., by mounting access nodes at bridges or traffic signs. But all this underscores the fact that in the future, only intermittent broadband connectivity can be provided for mobile users.

For mobile users, frequent disruptions of the broadband network connection have to be compensated to provide sophisticated data-intensive services such as high-quality video streaming. The connection interrupts create a dramatic increase in packet delays compared to legacy networks. Such packet delays have to be countervailed somehow. Nevertheless, for mobile users the delay is inherent to the coverage structure of the network and to user's mobility so that a reduction is impossible. If a packet delay is unavoidable the introduction of delay tolerance to broadband services can solve the problem for adequate service provisioning. A service is only interrupted when the user detects an odd behavior. For example, as long as the playback of a video continues, the service is not regarded as disrupted by the user, although the network connectivity might be lost for a long time. This obviously requires the build-up of a stock of video data in the user terminal. The same holds for push services of personalized (emails, RSS feeds, etc.) or commonly demanded content (newscasts, video streams, etc.). By filling the user terminal with a huge stock of such information, which is of interest to the user, it might be possible that he may not even notice connection interruptions because of consuming the already buffered content. He can, for example, peruse his favorite news page even during connectivity gaps, since the data was pre-fetched in advance. But such an approach requires determining how much data has to be pre-fetched in the end device in order to continue the service even during connectivity gaps. In the case of video streaming, this question is strongly correlated to the packet delay. The longer one has to wait for the next packet, the more data needs to be buffered in the terminal to achieve an uninterrupted service. Therefore, in the following the analysis is focused on the determination of the packet delay under intermittent broadband connectivity. From the packet delay the necessary parameters for the dimensioning of Smart Caching (SC) enabled networks are determined.

SC has to take care of an additional challenge: the accumulation of big reservoirs of user data in the terminal requires a massive transfer of data in periods of good connectivity. Together with the frequent connectivity losses, it is of vital importance that periods of good link performance have to be optimally used. Then the utilization of the wireless system is brought to its optimum. If it is aimed to fully use the wireless capacity backbone links can have a direct impact on the end-to-end performance.

Therefore SC integrates two fundamental paradigms: first, the optimal utilization of available network resources if broadband connectivity can be provided and second, the build-up of stocks in the end device in the Terminal Buffer (TB) to virtually continue services even in the case of numerous connection interruptions.

Section 2 summarizes related work. It also shows that the SC approach is a logical extension of existing research topics. The SC concept is further elaborated in the Section 3. It specifically illustrates how the performance of wireless broadband networks can be optimized. The main focus is on the impact of the backbone on the achievable performance. A performance analysis of SC is presented in Sections 4 – 8. Section 9 concludes.

2. Related Work

The use of the Smart Caching approach requires that several aspects have to be taken into account. The cooperation of SC with the existing network infrastructure has to be considered. Cache replacement strategies as well as content and user path, respectively next visited cell, prediction are strongly related to the SC technique.

A problem associated with temporarily storing user data is that it might interfere with network protocols like TCP as TCP is an end-to-end connection oriented protocol. Therefore it is necessary to decouple the connection into two subsections – one between the server and Smart Cache (SCache) and another between the SCache and Terminal Buffer. For TCP a solution, called SNOOP, was already presented in [1]. It was developed for the use of TCP over wireless links, which suffers from similar problems. Nevertheless the SNOOP approach fails to further extend the separation between wired and wireless links. The introduction of a buffer like proposed in SC allows a much better performance enhancement then the simple interception of the end-to-end protocol connection known from SNOOP.

In [2] a special extended RTP protocol is proposed, which is focused on streaming traffic including caching devices and allows reliable data delivery through caches. It does not require any changes in the end user software so that it is transparent to it. Video streaming, investigated in the later analysis, can be supported by such protocols.

Although due to the clustering of APs it is not required to predict the next visited cell in most of the cases it can further boost the performance of SC. In [3] a good survey about next cell prediction techniques for wireless networks is given. It uses long term log files of pedestrian user mobility as training sequences for different analytical models. It is shown that even with low complexity models a good mobility prediction is possible. Hence, further enhancement of the SC approach is possible. However, the long initial monitoring intervals might cause a lower performance at the deployment phase.

The pre-fetching of data in the end device makes it necessary that the content of the cache is continuously replaced by more recent information. In [4] updating strategies for caches are discussed. Different strategies are compared which take into account content recency or access frequency. In [5] personal profiles of web surfing behavior are used to predict the next accessed content. Both approaches allow a most accurate pre-fetching of data which is of further benefit for SC.

The SC approach is highly related to Delay Tolerant Networking (DTN). SC also tries to handle communication in network setups with long packet delays. DTN, started from interplanetary communication, noted that terrestrial communication also suffers under specific circumstances from similar problems. DTN tries to overcome connectivity interruptions by using the terminals of other mobile users for data transport between the access node and the destination terminal. SC instead follows the direct way. The mobile user himself reaches the next coverage zone and accesses the data there. The intermittent connectivity is compensated by extensive buffering. In both approaches, long packet delays are expected. But the success of DTN already shows that in the future such long delays are acceptable for mobile users. While sensor networks can accept longer packet delays without for human users compensation techniques have to be found – extensive buffering in the case of SC.

In [6] the general architecture, the design, and the state of the art of DTNs are outlined. The article does not consider the performance but focuses on the protocol and design aspects. The proposed protocols can be used to make SC working. In [7] the store and forward principle of DTNs is discussed. The analysis in this article focuses on regions with rudimentary network coverage which should be enhanced by an intermittent broadband connectivity. Such areas should be supplied by so called data mules which physically transport the data. For example, consider a small village connected via one wireless link for basic access provisioning, and additionally a frequently operating bus piggybacks the data between an access gateway in the next bigger city and the village. From this approach the organization of the store and forward principle can be applied for the current concept.

In [8] and [9] performance results of DTN are presented. Both articles show that packet delays of several tens of seconds are bearable. In the latter article it is shown that the buffer size has to be taken into account as it is a finite resource and can impact the performance of the approach. This shows that the later analysis of the TB size considers most relevant parameters. The impact of the mobility model on the performance evaluation of intermittent networks is self-evident. As used in the later analysis in [10] a Manhattan like scenario is taken to evaluate the performance of SC. Although DTN and SC are much related the former simply tries to exploit the delay resistance of specific services, e.g., data gathering within sensor networks. SC goes further and tries to mobilize (broadband) services which are not supportable in legacy networks due to their delay bounds. Therefore not the service is adapted to the network but the networks adapts to the needs of the service.

The idea to transfer data only in regions where high bit rates can be provided was already followed by [11]. This approach also inspired the later presented motorway scenario. In the article only general ideas and concepts are outlined so that in the following analysis the focus is put on the achievable improvement in system performance and network capacity.

3. The Smart Caching Concept

The architecture of Smart Caching entails two new network devices: the Smart Cache (SCache) and the Terminal Buffer (TB), as sketched in Figure 2. Caching in the SCache, as described in this article, serves for the reduction or even removal of the "inner resistance" (resulting from flow control algorithms or from capacity constraints) of a fixed telecommunication network, as seen by a user terminal. To achieve this, data from a server is pre-fetched and stored in the SCache so that the mobile terminal can access user data with the maximum transmission speed its wireless link is able to support. So instead of fine tuning transmission protocols to slightly increase the throughput SC introduces a completely new network setup so that the end-to-end data rate can be brought to its optimum.

Which impact SC can have on the data rate is shown in Figure 3. The graphs do not show real world measurements but instead the general characteristics which wireless and wired networks reveal for mobile users. In each graph the data rate of the wireless network, the backbone network and the resulting end-to-end data rate is displayed for a terminal moving from left to right through the center of the cell. The data rate of the wireless network as seen by the user terminal is of pyramid shape. Periods of broadband wireless network coverage may be followed by periods without any broadband connectivity (no connectivity or only service by a cellular wide-area network). Accordingly the radio data rate is set (close) to zero between two phases of AP connectivity in the example shown. The percentage of time where network connectivity is provided is denoted by the coverage ratio (*cov*).

Contrary to wireless networks in the backbone the achievable data rate is relatively constant. This data rate is in the following called the Backbone Limit (BL) as it might prevent the wireless network from achieving its best performance. The backbone data rate will usually not achieve the peak rates of the wireless network. In legacy networks, the resulting end-to-end data rate is always the minimum rate of all involved links, here the wireless and the backbone connection. This behavior “cuts of” a substantial part of the possible performance of the wireless network, see upper graph of Figure 3.

By buffering data in the AP, the initial phase of low radio data rates might be used to build up a stock, which then is forwarded when enough capacity in the wireless network becomes available (second graph of Figure 3). But this has only a minor impact on the overall performance.

Only by caching packets which arrive during the gap period it is possible to gather a sufficient amount of data which makes it possible to substantially utilize the otherwise unused capacity of the wireless network. To make use of the cached data, it is necessary that it is stored close to the next visited AP. Since the prediction of which AP will be visited next is usually difficult, all APs of a closer proximity have to be clustered and connected to one SCache.

Such an operation involving massive buffering is unsuitable for interactive communication services like speech and video conference owing to strict delay bounds specified for these services’ data packets. But these applications may benefit from the service provided by a wide area radio network, since they are not bound to wireless broadband networks.

Non-real-time services like unidirectional video streaming, FTP, email push/pull, and similar services tolerate high packet delays. SC exploits this property by pre-fetching contents and buffering it to transfer it at maximum possible speed to the user terminal. The corresponding storage is the second new network device introduced by SC the TB.

From the perspective of communication partners, the application server and the service user, the end-to-end connection is enhanced with two buffers. One is the SCache in front of the wireless link and one is the TB directly behind the wireless link. These two buffers together create a black box which is fed with data from the server, see Figure 4. This data is consumed in the end device by the user.

The TB is a storage which keeps a substantial data reservoir in the end device. Owing to the connection interruptions, a mobile user perceives the delivery of data to be much delayed. In Figure 5 it is shown how a coverage gap interrupts the data delivery and the packets are delayed. The time axis is oriented in the negative y direction. Packets are reflected by horizontal arrows. SCache and AP are regarded as colocated since the data exchange between these nodes is so fast that it has no impact on the performance of the end-to-end connection. Packets arrived in the SCache/AP are usually forwarded to the terminal with a short delay. In the terminal the data packets are buffered in the TB until the included content is displayed on the screen.

But in case of a coverage gap, the packets have to be buffered until the connection is regained. This is done in the SCache. For video streaming, therefore, it is necessary to determine how big the reservoir in the TB has to be to compensate the packet delays caused by the connection interruptions.

The right hand side of Figure 5 illustrates that the delay can be compensated by shifting the playback of the video in time. This time shift has to be big enough to compensate the packet delay. In the case of a coverage gap, the data in the TB is consumed but the new packets should arrive early enough to continue the playback before the TB is drained. At the outset it is only necessary to store an amount of data in the TB that corresponds to the required time shift before the actual playback starts. Therefore SC is the first technique which allows mobilization of broadband services which are currently limited to, so called, Hot Spot zones. The network configuration is changed in such a way that service interrupts are so much reduced that the user perceives a virtually continuous broadband connection. Since the time shift vitally depends on the packet delay, in the following analyses a special focus is put on the determination of the end-to-end packet delay.

The SCache works as a mediator between the wired and the wireless world by introducing a buffer between them so that the end-to-end data flow is separated into two subsections. The SCache terminates the first subsection commencing at the service provider. Concurrently, it is the starting point of the second subsection which terminates in the TB.

The actual communication partners are the service provider offering, e.g., a video stream, and the user terminal by which the service is requested. The SCache resides in the middle between the two communication partners. For the service provider the SCache adopts the role of the data flow end point. At this point the end-to-end connection between the service provider and the client is interrupted. The ingress data leaves the transport layer and moves up to a data buffer where it is cached until it can be finally forwarded to the user terminal. The first subsection exclusively consists out of wired Internet connections. Therefore TCP or similar protocols are suitable for the transfer of data in this subsection.

Pre-fetching and buffering is only reasonable if the data actually reaches its final destination, the user terminal. Data which is pre-fetched in the SCache but does not arrive at the end user wastes network resources. To locate the SCache directly in an AP, like in the SNOOP attempt [1], would imply that all pre-fetched data is useless, if the user leaves the coverage range of this AP and never returns. Making use of already buffered data requires that it must be available also from other access nodes which the user possibly roams to in the near future.

If all APs of a certain area are clustered and connected to one SCache it is possible to benefit from the cached data as long as the user terminal is supplied by one of these APs. Starting a communication session in cell **A** the data stream is routed through the SCache which is responsible for the cluster that AP cell is assigned to. If the data rate on the wireless link decreases because either the user moves away from the AP or completely leaves the cell, the data transmission on the first subsection continues. But as no forwarding through the second subsection is possible, the data is buffered in the SCache. Entering a new AP cell **B** the communication session will be re-established. If the new AP is included in the same cluster as cell **A** it is sufficient to reopen the connection between SCache and user terminal while the first subsection between service provider and SCache remains unchanged.

The analysis in the Sections 4 – 8 proves that SC-supported networks allow continuous video streaming service even under intermittent network coverage with long connectivity gaps. Furthermore, the advantage of the new approach over existing techniques is shown.

4. Urban Application Scenario

To analyze the performance of Smart Caching, especially its ability to compensate connectivity gaps in urban environments, the well-known UMTS 30.03 Manhattan scenario is chosen. All the APs in the scenario are connected to and served by one Smart Cache. The street grid consists of 200m-sized blocks and 30m-wide streets. WiMAX APs are placed on every second crossover. They are operated with a transmission power of 100 mW. WiMAX uses different modulation and coding schemes (PHY modes) to adapt the radio transmission to the current radio link quality. Each of them is assigned with a minimum signal quality necessary to decode the information and an available data rate (network parameters for WiMAX are taken from [12]). The pathloss between sender and receiver is given by an adapted free space model with an attenuation factor γ of 3 (2 stands for free space – up to 5 for indoor). Transmission power, pathloss and minimum signal level allow the mapping of a PHY mode to each terminal position in the scenario. As shown in Figure 6 the streets are divided in annulus shaped zones served by a specific PHY mode and zones without WiMAX connectivity. The network coverage ratio (for streets) is given by 82%.

To get representative input parameters for the later analysis in a mobility simulation the behavior of users in such a scenario is derived. Pedestrian users are traversing the streets of the scenario with a velocity equally distributed between 0.5 and 1.5 m/s. Their mobility is subject to an adapted Brownian motion which includes velocity updates and direction changes of up to 45 degrees on average every 10 seconds. The outcome (the user's pathway) of the first 10000 seconds of such a simulation is shown in Figure 6 (right). Out of much longer simulation runs reliable values for the average residence time in each PHY mode zone and the transition probability of users between different PHY modes are gained. Additionally, a large sample set of gap period durations can be derived.

For later use, a pure sample set is insufficient; thus the complete Cumulated Distribution Function (CDF) of the gap duration sample set is approximated with the support of the EM algorithm [13]. The outcome is a phase-type distribution which provides a closed-form of the CDF. The results for the gap duration distribution function and its approximation are shown in Figure 7.

Furthermore, the parameters of the different PHY mode zones can be employed to set up models for arrival and service processes, see Section 5. Out of the coverage ratios of each PHY mode and the corresponding data rates, it is possible to calculate the average data rate a user perceives. While for the backbone limited setups the performance of WiMAX is cut to a certain boundary the SC-enabled setups allow an optimal utilization of the network resources. In Figure 8 the available average data rate a user perceives for a test service depending on the Backbone Limit with the capacity share as a parameter is shown. The dotted red lines display SC-enabled setups and the solid black lines display the legacy setups without SC. The capacity share reflects the portion of radio resources that the test service can use and is not occupied by the other services (irrespective of whether they use SC or not). If network resources are bounded to other services the wireless network can only be used partially for the data transport of the test service. This can be reflected by a reduction of the available data rate. For example, if 75% of the resources are occupied by other services (corresponds to 25% capacity share), the data rate available for the SC service is automatically cut to one-fourth.

Assuming that the mobile user streams and watches a high quality video an average data rate of 3.37 Mbit/s can be assumed (MPEG4 Codec, see [14] page 108). Considering the curves of Figure 8, it can be seen that the advantage of the SC-enabled network for such a data rate is not directly evident. Especially for large capacity shares, the difference between the two curves (red dashes and solid black) is marginal for a wide range of the BL. But for a capacity share of only 20%, the advantage increases. While without SC the BL must be above 7.6 Mbit/s this value can be decreased to the actual streaming rate of 3.37 Mbit/s which is less than one half. Later on it will be shown that the advantage of SC is that the achievable data rate drastically increases compared to the legacy case if the coverage ratio gets lower (between 20 and 50%).

A capacity share of 20% implies that with full capacity at least five mobile video users can be served simultaneously per cell. For approximately 0.022 km^2 covered street area per WiMAX cell it means one user per 4000 m^2 which in a densely populated urban environment is not an unlikely value. Due to statistical multiplexing of users residing in a coverage zone or being outside of it the number of acceptable users could be further increased. Thus, depending on the BL the service might not be realizable without SC as the average data rate would be insufficient. But however there still exists the problem of the duration the packets spend in the SCache. Therefore in the following a sophisticated queuing model is used to develop the delay.

5. *MMAP/G/1* Queuing System

As stated earlier, the packet delay is essential for the dimensioning and performance evaluation of Smart Caching. Modern queuing theory allows a detailed modeling of different arrival streams competing for the limited radio resources. As a direct output the packet delay of the different streams becomes available.

The queue itself can be seen as a combination of the Smart Cache and Access Point, compare Figure 2. The SCache buffers the packets until they can be forwarded via the wireless link to the user terminal. The wireless link is the server of the model which works off the queued packets. The Terminal Buffer and the actual display of the user terminal can be regarded as a second queuing system. The arrival process represents packets which are transferred via the wireless link. The TB is reflected by the queue. Packets are worked off by the display when the corresponding packet content is shown on the screen. The next analysis is focused on the first queuing system. The second system is discussed later on and the results of both analyses are compared.

To show the benefit the employment of SC implicates the approach has to be compared with legacy network setups. Without SC the Backbone Limit reduces the achievable data rates between server and client. By adapting the service process of the queuing system both approaches can be compared.

In [15] new results of queuing theory are introduced, which allow the modeling of queuing systems with different incoming packet streams, each of them represented by a Markov Arrival Process (MAP) and can be subject to a separate service time distribution. The queuing delay of each arrival stream can be calculated separately. Since it can be distinguished between the arrival streams the process is called marked MAP (MMAP) and the resulting queuing system has the notation *MMAP(i)/G(i)/1*. This allows a very detailed modeling and analysis of the investigated scenario setup.

A MAP can be described by a Markov chain which allows transitions between all states. But contrary to a normal Markov chain two types of state transitions can occur. The first type is the transition between two internal states witch is not noticeable from the outside and the other is a state transition with a concurrent arrival. The transition rates of the two types are summarized in the matrices D_0 and D_1. D_0 contains simple state transitions; D_1 covers transitions with arrivals. For each arrival process i a separate pair of matrices $D_{0,i}$ and $D_{1,i}$

exists. The integration of the different streams can be done by composing the Kronecker sum of the different matrices $D_{0,i}$ and $D_{1,i}$, which is described in [16]. At the end there is only one matrix D_0 and i different matrices D'_i according to the separate traffic streams. The dimension of the resulting matrices arises from the product of the size of the input matrices $D_{0,i}$.

Two types of arrival processes have to be considered. First, a pure Poisson process ($D_{0,i}$=-λ and $D_{1,i}$=λ, λ = average arrival rate) to reflect background traffic which consumes network resources that are unavailable for the SC-supported traffic. This can be used to vary the capacity share. And secondly, the actual video streaming process which is supported by SC has to be modeled.

Already in [17] it was proposed to model the Variable Bit Rate (VBR) video traffic by using a certain number M of so called mini sources. To do so the video streaming rate is quantized in chunks of size λ_q. Each quantization step is modeled by a mini source which produces data with an average rate of λ_q. Depending on how many mini sources are active, the accumulated data rate varies as that by VBR traffic. The number of active sources is steered by an underlying birth death process as shown in Figure 9.

The above modeling of the video streaming process corresponds to a Markov Modulated Poisson Process (MMPP). An underlying Markov chain steers the state changes and to each state a data rate is assigned. A MMPP consists of two matrices Γ and Λ. The first contains the state transitions and the second the arrival rates per state. The translation to a Markov Arrival Process is achieved by

$$\begin{aligned} D_0 &= \Gamma - \Lambda, \\ D_1 &= \Lambda. \end{aligned} \quad \textbf{(1)}$$

The transition rates α and β as well as λ_q of Figure 9 can be matched to the parameters of the video streaming process. In [17] the average streaming rate $E[\lambda]$, the variance $C(0)$ and the auto-covariance function $C(a,\tau)$ are used for that. With the following equations the necessary parameters can be deduced.

$$\alpha = a\left(1 + \frac{E^2[\lambda]}{MC(0)}\right)$$
$$\beta = \mathrm{a} - \alpha \qquad \textbf{(2)}$$
$$\lambda_q = \frac{C(0)}{E[\lambda]} + \frac{E[\lambda]}{M}$$

The actual parameters for MPEG4-coded video streaming are taken from [14], see Table 1.

The service process of the queuing model has to reflect the different PHY modes by which a mobile user is served. To reflect the different data rates of each PHY mode in the queuing model a hyperexponential service time distribution is taken, see Figure 10. According to path probabilities p_i packets are processed with different service rates μ_i. Both parameters can be derived out of the probabilities that a user resides in PHY mode area i and the corresponding data rate r_i of the WiMAX system. These values are provided by the mobility simulation and the wireless network characteristics.

To fully understand the benefit of SC, it is important to compare its performance with the legacy case. To emulate the “cutoff” behavior of legacy network setups as shown in the upper graph of Figure 3 the service rates μ_i have to be reduced accordingly.

Up to now the service process does not cover areas without connectivity. As the transmission rate is zero in such regions, it is not possible to directly include it in the above model. If a customer traverses a coverage gap the incoming packets of the, e.g., video stream are buffered in the SCache. Due to this buffering the end-to-end delay of packets is drastically increased. Which influence the gap has on the actual packet delay is depicted in Figure 11. The scene is separated in three periods. In the first period, denoted by x, the customer leaves a coverage zone and the SCache starts the buffering so that the fill level increases. After re-entering a new coverage zone the incoming traffic is buffered as there are still earlier arrived packets left in the cache which have to be served first. Since the arrival rate for new packets is smaller than the transmission rate of the air interface the buffer starts to get drained and the fill level decreases. The period until the whole buffer is emptied is called y. And finally the period z

just denotes the normal operation of the wireless network. Packets which arrive have to wait a short period until radio resources become available to be transmitted.

In the lower part of Figure 11 the additional delay depending on the arrival time of the packet in the SCache is shown. It is caused by the gap in the coverage. The maximum delay is suffered from the first packet which arrives after the user terminal has left the coverage zone. It is delayed for the whole period of no coverage and is instantly transmitted when the next coverage zone is reached. Therefore the additional delay vitally depends on the duration of the gap period. The packet delay linearly decreases until the point is reached where the complete SCache is drained and the operation migrates to the normal behavior. For later evaluations it is necessary to determine the ratio of packets which are affected by the gap and the portion which is transferred without additional delay. Clearly the first portion is reflected by (x+y)/(x+y+z).

The amount of traffic which has been aggregated in the SCache during a gap is given by $x\lambda_{video}$, where λ_{video} is the arrival rate of the video process. During the y period this amount of stored data together with the currently arriving traffic has to be worked off. The current traffic loads the system with the utilization ρ. The utilization is defined as the quotient out of the average arrival rate λ of a queuing system and the average service rate μ. The arrival rate is predefined by the data rate of the regarded user services (e.g., video streaming rate). The service rate in the regarded system model reflects the ability of the end-to-end connection to transport the data packets. This is influenced by two aspects the capacity of the wireless link but as well the BL, which might prevent an optimal performance of the radio connection. The capacity share which is available to work off the aggregated data heap is $(1\text{-}\rho)$. All this results to

$$x\lambda_{video} = y(1-\rho)\mu_{video} \qquad \textbf{(3)}$$

where μ_{video} is the video service rate. Together with the fact that the coverage ratio is given by the quotient out of x and the overall size x+y+z, it can be concluded that

$$\frac{x+y}{x+y+z} = (1-\text{cov})(1+\frac{\rho_{video}}{1-\rho}). \quad \textbf{(4)}$$

The additional delay the packets perceive due to the gaps in the connectivity is equally distributed between gap duration and zero. Since this delay is independent from the waiting time in the queuing model both values can be summed up. For their probability densities this implies a convolution. Since the outcome of the *MMAP/G/1* queuing system is the Laplace Stieltjes Transform (LST) of the waiting time distribution it is natural to perform the convolution by a simple multiplication of the LSTs. Therefore it is also necessary to get a closed form expression of the LST of the gap duration. As packets during the period z out of Figure 11 are not affected by the gap it is necessary to separate the distribution of the additional delay in two parts. With the probability $z/(x+y+z)$ the additional delay is zero and with the probability $(x+y)/(x+y+z)$ an additional delay exists. The first part can be reflected by a Dirac impulse in the origin $\delta(t)$. For the second part the probability density function (pdf) of the additional delay *pdf(t)* (compare the proof at the end of the article) is required. *pdf(t)* and the resulting *pdf(delay=t)* for all packets are given in Equation (5).

$$\begin{aligned} pdf(t) &= \int_t^\infty \frac{p(x)}{E[x]}dx = \frac{1}{E[t]}(1-cdf(t)) \\ pdf(delay=t) &= \frac{z}{x+y+z}\delta(t) + \frac{x+y}{x+y+z}pdf(t) \end{aligned} \quad \textbf{(5)}$$

What is still missing for a final solution of the packet waiting time is the waiting time expression of the *MMAP/G/1* system. The LST of the waiting time distribution per packet type i is given by

$$\begin{aligned} W_k(s) &= \frac{V(s)D_k}{\lambda_k} \\ V(s) &= (1-\rho)s\mathbf{g}(sI + D0 + D(s))^{-1} \\ D(s) &= \sum_k D_k(s) = \sum_k D_k H_k(s) = \sum_k \int_0^\infty e^{-st} D_k(t)dt \end{aligned} \quad \textbf{(6)}$$

where λ_i is the arrival rate of type i packets and ρ is the above mentioned utilization of the system. The vector $\mathbf{g}$ is the stationary vector of the matrix Q ($\mathbf{g}Q=1$ and $\mathbf{g}1=1$) which is given by the recursive formula

$$Q = D_0 + \int_0^{\infty} D(x) e^{Qx} dx\,. \tag{7}$$

D(x) is similarly defined to *D(s)* but instead of using the LST the probability densities have to be used. The matrix Q can be iteratively calculated by starting with $Q=D_0$ and continuously substituting it in Equation (7) until the differences between concurrent results gets small enough.

With all this preparation it is possible to derive the CDF of the video streaming packet delay as depicted in Figure 12 (please note that the y-axis ranges from 0.5 to 1.0) for the transport between server and TB. The 95th percentile of the packet delay is a good measure for the required time shift. The scenario is still the urban environment as introduced above. The capacity share is reduced by varying the arrival rate of the additional Poisson process. From the curves as well as from the 95th percentile of the waiting time (listed in the legend) it can be seen that the capacity share has an impact on the waiting time. However, even the provisioning of the full network capacity for one video streaming process can not reduce the waiting time below a certain limit.

If all wireless network resources are reserved for the observed video streaming service (100% capacity share) the 95th percentile of the packet delay and therefore the required time shift accounts to 102 seconds. This means that 102 seconds of video data have to be buffered in the TB before the video playback should start. Under this condition the risk of a playback interruption is much reduced. Later analyses will show that it is around 5%. Since at the beginning of a streaming process the transfer rate can be drastically increased compared to the actual streaming rate it is possible to download the time shift reservoir faster than the 102 seconds. If the wireless network provides in the beginning a data rate which is twice the streaming rate it implies that the required data can be downloaded in around 50 seconds. Although this value is higher than used from legacy streaming services SC allows the service provisioning of high quality video streaming under intermittent network connectivity, which is definitely not possible with legacy setups.

Even for a capacity share of 25% the SC-enabled scenario setup still performs satisfactorily. The 95th percentile of the packet delay only increases to 142 seconds so that still a reasonable time shift is reached. In the throughput analysis above it was stated that a legacy network setup under this condition might get into problems. Therefore the performance of the SC approach is compared with a legacy setup. For the legacy case the service rate of the WiMAX APs is bounded by the BL. For different BLs the CDF is shown in Figure 13 (capacity share = 25%). The curves get even lower and the 95 th percentile further increases.

Although the BL has some impact on the packet waiting time the changes are not dramatic. The resulting time shift is not influenced much so that the service can still be supported. Only the initial download phase is increased by 80% compared to a full capacity share and no BL. But the ability of SC to compensate the BL has another much more important relevance. It is an enabler for broadband services. Without SC, many services cannot be supported at all.

If the BL becomes less than 8 Mbit/s in the given scenario the packet delay goes to infinity as the service can no longer be supported by the network. The reason is that the available average data rate drops below the playback rate of the video.

Nevertheless, even for the SC-enabled network setup the problem on how the delay can be handled exists. For applications like video streaming the delay can be compensated by extensively buffering video data in the end device. If data storage is provided in the end device (namely the Terminal Buffer) the playback of the video can continue when the mobile user leaves the zone of wireless network connectivity – as long as there is still buffered video data at hand. The question is, how big the TB has to be? And secondly, what is the required initial fill level before the playback of the video can start? A reference value for the initial fill level of the buffer could be again the 95 percentile of the packet waiting time. It has to be guaranteed that new packets arrive early enough to continue the video playback before the SCache is drained. If an amount of video data is stored in the end device which corresponds to the 95 th percentile of the CDF it means that with 95% probability the next packet will arrive within that time limit.

6. *MAP/G/1/N* Queuing System

Now, the second queuing system is regarded. The input process reflects the amount of data transferred via the wireless link. As with Smart Caching always enough data is provided in the Smart Cache the capacity of the wireless link is always fully utilized. Therefore the arrival rate basically reflects the available wireless data rate. The Terminal Buffer is reflected by the queue and the service process models the playback of the video. To allow good service, the TB size has to be determined. And to identify the necessary TB size, the SC-enabled system is considered from the end device's perspective. The TB tries to build up a stock of video data that can be consumed during phases of non-existent or insufficient network connectivity to continue the playback of the video stream. To achieve this goal, it is necessary to fill up the buffer in periods of perfect network connectivity to its maximum. The question is, which buffer size is required to provide a predefined Quality of Service (QoS) level? An obvious QoS criterion is the probability that a video stops, or in other words, the TB is drained.

Such an end device together with a limited TB can be modeled by a queuing system with finite queue size. The *MAP/G/1/N* system presented in [18] allows the thorough modeling of the arrival process as well as the monitoring of the queue size and the idle probability of the queue.

The key factor for the analysis of such a queuing system is given by the transition probability matrix P of the counting process which monitors the number of packets in the finite queue. The matrix is given by

$$P = \begin{matrix} \overline{D}A_0 & \overline{D}A_1 & \cdots & \overline{D}A_{N-1} & \overline{D}A'_N \\ A_0 & A_1 & \cdots & A_{N-1} & A'_N \\ 0 & A_0 & \cdots & A_{N-2} & A'_{N-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & A_0 & A'_1 \end{matrix}. \qquad \textbf{(8)}$$

For the matrices A_i and A'_i in [18] easy calculation formulas are provide in case the service time distribution is subject to a phase type distribution. Since this type includes a wide area of possible distributions the modeling flexibility is not significantly restricted by that condition. The matrix $\bar{D}$ can be derived from the arrival process. $\bar{D}$ results to $(-D_0)^{-1} D_1$.

The steady state vector x' of the matrix P ($x'P= x'$ and $x'1=1$) provides the queue size distribution at departure epochs. Departure epochs are events when a packet leaves the queue and is handed over to the server. The whole vector can be separated into subvectors. The entries of one of these subvectors represent situations with the same buffer fill size but different states of the arrival process; $x' = (x'_0,\ldots, x'_N)$.

The queue size distribution at arbitrary times ($x_0,\ldots, x_N$) and the probability that the whole system is empty (including the server) $\tilde{x}$ can be derived from the following formulas.

$$\begin{aligned}
\tilde{\mathbf{x}} &= \frac{1}{E^*}\mathbf{x}'_0\overline{D}\ (-D_0)^{-1} \\
\mathbf{x_0} &= \left(\frac{1}{E^*}(\mathbf{x}'_0 - \mathbf{x}'_1) - \tilde{\mathbf{x}}D_1\right)D_0^{-1} \\
\mathbf{x_n} &= \left(\frac{1}{E^*}(\mathbf{x}'_n - \mathbf{x}'_{n+1}) - \mathbf{x}_{n-1}D_1\right)D_0^{-1}, \quad 1 \le n \le N-1 \qquad \textbf{(9)} \\
\mathbf{x_N} &= \left(\frac{1}{E^*}\mathbf{x}'_N - \mathbf{x}_{N-1}D_1\right)D_0^{-1}, \text{ with} \\
E^* &= \frac{1}{\mu} + \mathbf{x}'_0(-D_0)^{-1}\mathbf{1}
\end{aligned}$$

The most interesting vector is $\tilde{x}$ as it contains the probability the system is idle which means the playback of a video has stopped ($P_{idle}= \tilde{x}1$).

The size N of the finite buffer has a direct impact on the size of the matrix P. Due to memory and processing speed restrictions the number of entries in the matrix should be limited. A buffer size of $N>500$ storage units is not recommended. If one data packet would be used as a storage unit the queue could only hold up to 500 packets. This is much too less for the considered approach. Therefore it is necessary to quantize the amount of buffered data. The most significant expression for the buffer size is the number of seconds the video can be continued with the data stored in the buffer. Therefore the quantization is done on data chunks which are enough to playback one second of the video stream (in the case of 3.37 Mbit/s playback rate one second corresponds to 0.42 MByte of video data).

For the service time of the queuing system a negative exponential distribution is chosen. Since it only reflects the consumption of the buffer by the video playback, no more specific distribution is taken here. Owing to the quantization of the buffer content, the rate μ of the service process is simply given by 1/s. This means that per second, exactly one chunk of buffered data is consumed by the playback of the video.

Contrary to the service process, the arrival process of the second queuing system requires much more modeling effort. The arrival process is not so much influenced by the actual traffic stream it carries, but to a large extent by the data rate that the wireless network can provide. Due to the SC approach, always the entire available network resources are consumed. For a mobile user traversing an urban scenario like shown in Figure 6 this means that the arrival rate varies depending on the PHY mode zone he currently resides in. This behavior is illustrated in Figure 3. The amount of stored data in the SCache during the gap period is used to “fill up” the otherwise unused resources, reflected by the grey area.

To model such an arrival process for each PHY mode a state has to be defined. Assigned to this state is a data rate λ_i and from the before described mobility simulations transition rates p_{ij} from state i to state j can be derived. This perfectly fits to a MMPP arrival process. Such a MMPP process can be translated to a MAP as shown in Equation (1).

$$\Gamma = \begin{pmatrix} -p_1 & p_{12} & \cdots & p_{1N} \\ p_{21} & -p_2 & \cdots & p_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ p_{N1} & p_{N2} & \cdots & -p_N \end{pmatrix}, \; p_i = \sum_{j \neq i} p_{ij}, \; \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{pmatrix} \qquad \textbf{(10)}$$

But as shown in Figure 14 the data rate might drop to the backbone rate. Two reasons can be cited: either the SCache is drained so that only packets can be transmitted which arrive in the SCache from the video server or the TB is full. Although this does not mean that the data rate decreases the amount of packets accepted from the queue drops to the playback rate and simultaneously the loss ratio of the system increases. Only if a packet leaves the TB a new packet can be stored. Otherwise it has to be discarded. But in such a case the arrival process does not have to be adapted since it has no influence on the idle probability of the queue whether the arrival rate drops down or packets are discarded. Although the loss rate of the queuing system increases, the idle probability remains unchanged.

So the question is, what happens if the SCache is drained? As shown in Figure 4, the combination of SCache and TB can be seen as a black box system with an incoming data rate λ and an outgoing data rate which is as well equal to λ. This means that after the playback of the video has started the amount of buffered data in the black box does not (or only slightly) change.

For later analysis now it is assumed that the maximum TB size is equal to the (initial) fill level of this black box (a bigger TB would even improve the performance so that the later results are a lower limit). Under this condition the occurrence of an empty SCache coincides with a full TB. Therefore the aforementioned statement holds that the arrival process can be modeled without any drop of the data rate. The situation that the data rate drops to the BL is implicitly included in the queuing model by discarding packets when the queue is full.

If the full capacity share is not available for the video streaming process then the data rates in each PHY mode state have to be adapted accordingly. If only 50% of the network resources can be used by the streaming process than the data rate in each PHY mode has to be halved as well.

To model a backbone limited network, the same MMPP based approach can be used as for the SC-enabled system. It is enough to cut in each state the data rate to the value of the BL. But the adaptation due to limited network resources for the SC enabled service has to take place ahead.

7. Terminal Buffer Size

With the help of the before described finite buffer queue it is possible to derive the idle probability of the end user device depending on its buffer size. While the analysis with the MMAP/G/1 queuing system has allowed an estimation of the initial buffer fill level now the overall Terminal Buffer size is discussed.

In Figure 15 the idle probability depending on the capacity share is shown. The idle probability reflects the likelihood that the TB is drained and the last chunk of video data in the queue is played back. On the *x*-axis the available capacity share of the wireless network is shown. The parameter of the set of curves is the TB size. The last parameter is measured in seconds the video playback can be continued with a completely filled TB and no additionally arriving packets.

It can be observed that the available capacity share has a deep impact on the performance of the SC enhanced network setup. For a realistic service provisioning the idle probability should be below 5%. In that case video playbacks are seldom interrupted and the user has the impression of continuous network connectivity.

Only for a very large TB size of 300 seconds, the system remains within reasonable boundaries even if only a capacity share of 25% is taken for the video streaming process. For smaller dimensioned TBs the impact on the idle probability is clearly visible. While for capacity shares of more than 80% all curves are below 7% idle probability, and therefore at least close to the prior defined quality level, the situation changes in case of lower capacity values. The idle probability increases substantially so that the service quality is no longer sufficient. Especially for a TB size of 100 seconds an idle probability of 5% can only be reached if the full capacity share of 100% is reserved for the regarded service.

If these results are compared with the outcomes of the *MMAP/G/1* queuing system it can be shown that they match very well. In Table 2, the 95th percentile of the *MMAP/G/1* and the idle probability of the *MAP/G/1/N* systems are compared. In the latter a TB size is assumed which corresponds to the 95 th percentile of the first analysis.

It can be seen that a buffer size which is derived from the 95 th percentile produces an idle probability of around 5%. Only for heavy loaded systems (where only one-fourth of the capacity remains for the video streaming) do the results diverge. But the reason is that in the *MMAP/G/1* case it was assumed that during coverage the system always succeeds to drain the SCache. In an overload situation this assumption might be wrong. Gap and coverage period's size differ around an average value so that in case of a long gap and a following short coverage period it might occur that some packets remain in the SCache when the connection breaks again. Then these packets have to wait until the next coverage period before getting delivered. Therefore, the waiting time is extended further and the 95 percentile is pushed to even higher dimensions.

The earlier results show a good match between the two analytical approaches. However, a simulative evaluation could strengthen the results. For the considered scenario, very long simulation runs are necessary to cover enough coverage gaps and reach statistically significant results. Together with the sophisticated modeling of the arrival process, the simulation effort would be too complex to reach results within a reasonable amount of time, so the analytical approach has to be taken. Nevertheless for simplistic scenario setups simulations were conducted. A very good match of the packet delay CDFs between the simulation and the analytical results could be shown. As the results are of no further relevance for the here regarded scenario, they are omitted.

To dimension a SC network from the analysis of the *MMAP/G/1* queuing system the initial fill level of the TB can be deduced. This guarantees that the video playback does not stop in 95% of the cases during the first gap period. Such a value is already a first indication for the required time shift. Nevertheless, the finite buffer analysis delivers the exact results as it includes not only the first gap but also all subsequent gaps. Thus, the long term probability can be derived so that the video playback does not stop even during longer video streams. It is important to note that these two values, at least for low loaded networks, match so that the M/G/1/N based analysis is not necessary in each case.

8. Motorway Application Scenario

To further illustrate the applicability of Smart Caching and its capability to increase performance in delay tolerant networks a second scenario is investigated. In this motorway scenario users go by car and access again a high quality video stream. Due to the massive deployment costs not the complete motorway is covered with broadband wireless access, but only in certain intervals the APs are mounted, see Figure 16. The AP sites are chosen depending on the motorway infrastructure. To provide power supply and backbone connectivity usually places like bridges which cross the motorway are the best places. While moving from an AP site to another AP site data is downloaded and stored in the Terminal Buffer when enough network capacity is available – hence, close to the node. This data is then consumed during the idle periods between two AP sites. In such phases the streaming of data is continued between video server and Smart Cache.

In such a scenario setup it can be assumed that the attenuation between sender and receiver is reduced to a minimum. Therefore the attenuation factor γ is set to a free space value of 2. The average distance between the APs is chosen in such a way that the wireless network coverage is equal to 50% in a first setup and reduced to 20% in a second scenario.

The probability density of the average car velocity is taken from [19] and corresponds to a normal distribution. The density of the gap distance between two concurrent coverage zones is as well normally distributed. All parameters are listed in Table 3.

From the distributions of the car velocity and the gap distance, the probability density of the duration of a gap period can be derived by applying the following formula.

$$f_t(t) = \int_0^\infty v f_l(vt) f_v(v) dv \qquad \textbf{(11)}$$

In Figure 17 the average data rate with (dot-dashed) and without SC (solid) functionality for such a scenario setup depending on the Backbone Limit is shown (capacity share varied from 10 to 100% in 15% steps). On the left, the coverage ratio is 50% and on the right, it is 20%. Now it is obvious that especially for low coverage ratios the performance gain of the SC-enabled networks is substantial. For 20% network coverage the required average data rate of 3.37 Mbit/s can be provided for an available capacity share of 90 – 100%. With SC the BL does not have to be much above the average value of 3.37 Mbit/s while with a legacy network setup this value dramatically increases to more than 40 Mbit/s. As expected, SC shows its capability for substantial performance enhancement in low coverage scenarios.

But before the TB size is discussed an additional improvement of SC is introduced. It is termed “Making a virtue out of necessity”. Delay tolerant networks have to deal with connection interruptions. It has been shown that with SC, connectivity gap periods of more than 100 seconds can be bridged for services like video streaming. But if such long periods have to be handled it might be useful to even lengthen this period. Why should we do so? The outer part of the coverage region of a WiMAX AP is served by a very robust PHY mode to give the receiver a chance despite the long distance to decode the signal. But this also means that the data rate is simultaneously decreased. The transfer of 1 MB of data requires in this robust PHY mode more than 1.3 seconds while close to the AP this can be completed in less than 0.15 seconds. So from a resource allocation point of view it is useful to wait until a mobile user reaches the most inner PHY mode region before massive data transfers take place.

Such a behavior can be obtained by deactivating the most robust PHY modes which cover the outer annuli of a coverage region. The more PHY modes are deactivated the longer the gap between concurrent AP zones becomes. But simultaneously the efficiency of the network increases.

In Figure 18 the CDF of the packet waiting time is shown depending on how much PHY modes are deactivated. It can be seen that the 95^{th} percentile which reflects the initial buffer fill level – but is as well a good approximation of the required TB size – increases the more PHY modes are deactivated. But the increase is only moderate from 220 to 380 seconds. The impact this has on the network efficiency can be seen on Figure 18 (right). For different values of traffic intensity the ratio of possible video users per 100 cars is shown depending on the number of deactivated PHY modes. The reason for the efficiency gain is not that more users can be served by one AP, but that less users have to download data simultaneously. If only the inner PHY mode zones are used for the transfer of data then the amount of cars residing in the coverage area decreases. This means that the ratio of supported users can be increased accordingly.

For such a setup there has to be found a tradeoff between the packet waiting time, its resulting initial buffer fill level and the network performance gain. But of course it is possible to switch between the different configurations. At the beginning a mobile user starts with a fill level of 220 seconds and all PHY modes are used. As soon as the fill level of the buffer could be increased above 280 seconds (e.g., due to very good network performance) the most robust PHY mode is deactivated. This can continue until 5 PHY modes are switched off. Of course the reverse is also possible. If there is a severe threat that the TB drains then the outer PHY modes can be reactivated. The exclusive usage of only the inner PHY mode (mode 1-6 deactivated) is not possible as the network resources are insufficient to provide the necessary average data rate in case of such a reduced coverage ratio.

9. Summary and Conclusions

Smart Caching is introduced as a method for making networks delay tolerant and handling connectivity disruptions. The technique allows the optimization of throughput capacity in broadband wireless networks. Moreover, it supports and elaborates the extensive buffering of service data in the end user device. This allows the virtual continuation of broadband services even with intermittent network connectivity.

The complete communication path between server and client is included so that it is possible to assess the potential of Smart Caching for an end-to-end service provisioning. In the analysis, a video streaming service is regarded as it has the highest demands to communication networks. Nevertheless, the content to be cached is easily predictable so that taking into account deviations is not required. For further evaluations this constraint might be excluded as other services do not allow for a precise prediction.

The presented results allow a good dimensioning of streaming services as the delay analysis gives the initial time shift, and the analysis of the buffer size allows the dimensioning of the Terminal Buffer in the user terminal. For services that are delay tolerant by default (such as file transfer), the delay evaluations are not of much interest but the achievable data rate in Figure 8 and Figure 17 can be perfectly used to derive the necessary transfer durations. So this article covers a wide range of services. In the future, it is necessary to investigate services which are not 100% predictable.

Two application scenarios are presented – a pedestrian user in an urban environment and a vehicular user on a motorway. In both scenarios the applicability of Smart Caching is proven. The analysis with the support of two sophisticated queuing models allows a very accurate dimensioning of the Terminal Buffer size. Furthermore the requirements of an initial fill level for the Terminal Buffer is derived and verified by the analysis.

In the motorway scenario a further improvement of Smart Caching is given. It allows an enhancement of the network capacity so that more users can be supported simultaneously.

The additional delay which arises can be absorbed by the Smart Caching technology. All this proves that Smart Caching is a perfect way to circumvent problems occurring in wireless networks with patchy coverage.

References

[1] H. Balakrishnan, S. Seshan, E. Amir, and R.H. Katz, *Improving TCI/IP performance over wireless networks*, In Proc. 1st ACM Int'l Conf. on Mobile Computing and Networking (Mobicom),pp. 2—11 , Nov. 1995.

[2] M. Zink, A. Jonas, C. Gridwodz, and R. Steinmetz, *LC-RTP (loss collection RTP): reliability for video caching in the Internet}*, In Proc. of Seventh International Conference on Parallel and Distributed Systems, pp. 281—286 , 2000.

[3] L. Song, D. Kotz, R. Jain, and X. He, *Evaluating Next-Cell Predictors with Extensive Wi-Fi Mobility Data*, In IEEE Transactions on Mobile Computing, pp. 1633—1649, 2006.

[4] S. Podlipnig and L. Böszörmenyi, *A survey of Web cache replacement strategies*, In ACM Computing Surveys (CSUR), Volume 35, Number 4, pp. 374—398 , 2003.

[5] Z. Su, Q. Yang, and H.J. Zhang , *A prediction system for multimedia pre-fetching in Internet*, In Proceedings of the eighth ACM international conference on Multimedia, pp. 3—11 , 2000.

[6] K. Fall and S. Farrell, *DTN: an architectural retrospective*, In IEEE Journal on Selected Areas in Communications, Volume 26, Number 5, pp. 828—836 , 2008.

[7] M. Demmer, E. Brewer, K. Fall, S. Jain, M. Ho, and R. Patra, *Implementing Delay-Tolerant Networking*, Intel Research, Berkeley, Technical Report, IRB-TR-04-020, Dec. 2004.

[8] Y. Lin, B. Li, and B. Liang, *Stochastic analysis of network coding in epidemic routing*, In IEEE Journal on Selected Areas in Communications, Volume 26, Number 5, pp. 794—808 , 2008.

[9] X. Zhang, J.K. Kurose, B.N. Levine, D. Towsley, and H. Zhang, *Study of a bus-based disruption-tolerant network: mobility modeling and impact on routing*, In Proceedings of the 13th annual ACM international conference on Mobile computing and networking, pp. 195—206 , 2007.

[10] R.A. Nichols and A.R. Hammons, *Performance of DTN-Based Free-Space Optical Networks with Mobility*, In Proceedings of IEEE Military Communications Conference 2007, MILCOM 2007, pp. 1—6 , 2007.

[11] B. Walke, *On the importance of WLANs for 3G cellular radio to become a success*, In Proc. 10th Aachen Symposium on Signal Theory - Mobile Communications, 2001.

[12] C. Hoymann, *Analysis and performance evaluation of the OFDM-based metropolitan area network IEEE 802.16*, Computer Networks, 2005.

[13] S. Asmussen, O. Nerman, and M. Olsson, *Fitting phase-type distributions via the EM algorithm, Scandinavian Journal of Statistics*, 1996, 23, 419-441.

[14] P. Seeling, F. Fitzek, and M. Reisslein, *Video Traces for Network Performance Evaluation*, Springer, 2007.

[15] T. Takine, *Queue Length Distribution in a FIFO Single-Server Queue with Multiple Arrival Streams Having Different Service Time Distributions*, Queueing Systems, Vol. 39, No. 4, pp. 349—375, 2001.

[16] T. Takine, *The Nonpreemptive Priority MAP/G/1 Queue*, Operations Research, Vol. 47, No. 6, 1999, pp. 917-927.

[17] B. Maglaris, D. Anastassiou, P. Sen,G. Karlsson, and J.D. Robbins, *Performance models of statistical multiplexing in packet videocommunications*, IEEE Transactions on Communications, Vol. 36, No. 7, pp. 834—844, 1988.

[18] U. Gupta and P. Laxmi, *Analysis of the MAP/G(a, b)/1/N Queue*, Queueing Systems, Springer, 2001, 38, 109-124.

[19] C.-H. Rokitanski, Editor, *Validation of the Mobility Model of SIMCO2 with Dutch Motorway Measurements*, Prometheus Workshop on Simulation, December 1992.

Proof of Equation (5)

$pdf(t) = P(W = t)$: Probability that the additional delay of an arbitrary packet is t

$p(x) = P(gap\_duration = x)$: Probability that gap duration is x

$P(Packet\_in\_gap)$: Probability that the arrival of the packet occurs when the user is in a gap

Three factors have to occur in order to perceive an additional delay *t*. First there must be a connectivity gap of size *x*. Second the arrival of the packet has to occur while the terminal is in the gap. And finally the condition has to hold that the gap duration *x* must be larger than the delay *t*.

$$P(W = t) = P(Packet\_in\_gap, gap = x \mid t \le x)$$

Since the condition $t \le x$ is independent of the other probabilities it can be stated that

$$P(W = t) = P(Packet\_in\_gap, gap = x, t \le x)$$
$$= \int_{y}^{\infty} P(Packet\_in\_gap, gap = x, t = x)dx$$
$$= \int_{x'}^{\infty} P(t = x \mid Packet\_in\_gap, gap = x)P(Packet\_in\_gap \mid gap = x)P(gap = x)dx$$

Two terms have to be further evaluated *P(Packet_in_a_gap|gap=x)* and *P(t=x|Packet_in_gap,gap=x)*. The second one describes the probability that if the packet gets definitely an additional delay due to a gap of duration *x* that the delay *t* is equal to *x*. Since it can be assumed that packets arrive equally distributed over the whole gap interval *x* the probability density is as well equally distributed with:

$$P(t = x \mid Packet\_in\_gap, gap = x) = \frac{1}{x}.$$

The other required term *P(Packet_in_a_gap|gap=x)* describe the probability density that a packet arrives during the gap given the fact the size of the gap is *x*. This cannot be directly evaluated. But with the same reasoning as before it is clear that the probability that the packet arrives during the gap increases linearly with the gap duration *x*.

$$P(Packet\_in\_gap \mid gap = x) = cx$$

where *c* is just a constant factor. In order to determine *c* it can be used that the integral of the unconditional probability *P(Packet_in_a_gap,gap=x)* must be equal to *1*.

$$\int_{0}^{\infty} P(Packet\_in\_gap, gap = x)dx = \int_{0}^{\infty} P(Packet\_in\_gap \mid gap = x)P(gap = x)dx$$

$$= \int_{0}^{\infty} cxP(gap = x)dx = c\int_{0}^{\infty} xP(gap = x)dx = cE[x] = 1$$

The last transformation is just the expected value of the gap duration probability distribution. So it follows that:

$$P(Packet\_in\_gap \mid gap = x) = \frac{1}{E[x]} x .$$

With all this *P(W=t)* results to

$$P(W = t) = \int_{x'}^{\infty} P(t = x \mid Packet\_in\_gap, gap = x) P(Packet\_in\_gap \mid gap = x) P(gap = x) dx$$

$$= \int_{x'}^{\infty} \frac{1}{x} \frac{x}{E[x]} p(x) dx = \int_{x'}^{\infty} \frac{p(x)}{E[x]} dx \text{ q.e.d.}$$

	C(0)	a	Data Rate [10^6/s]	M
Video (MPEG4)	0.509	0.09	3.37	20

Table 1

Net ratio/buffer size (based on 95 percentile) [s]	Idle probability
1.0 / 102	4.37%
0.5 / 112	6.92%
0.25 / 142	11.78%

Table 2

Car velocity	Gap size (Scenario type 1)	Gap size (Scenario type 2)
$f_v(v)$	$f_l(l)$	$f_l(l)$
	Coverage=50%	Coverage=20%
Normally distributed	Normally distributed	Normally distributed
μ=104 km/h	μ=5100 m	μ=20.5 km
σ=12.5 km/h	σ=500 m	σ=5 km

Table 3

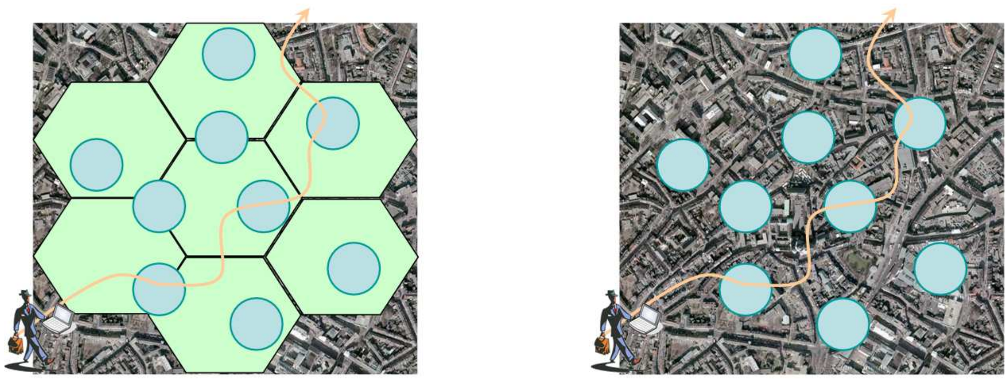

Figure 1: Coverage of heterogeneous wireless networks
(Green hexagons represent UMTS coverage and blue circles WiMAX Hot Spots)

213x86mm (150 x 150 DPI)

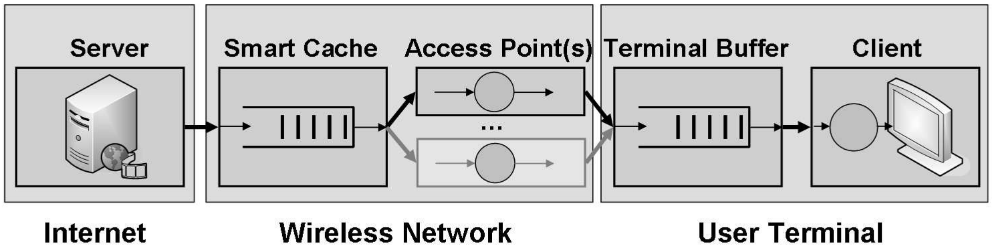

Figure 2: Architecture of Smart Caching
253x71mm (150 x 150 DPI)

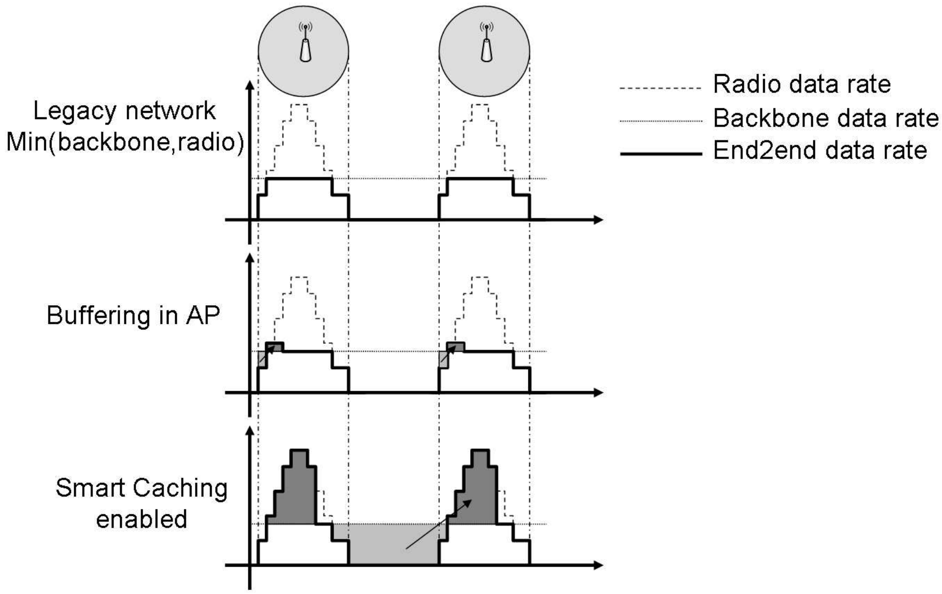

Figure 3: Impact of Smart Caching on Data Rate
243x154mm (150 x 150 DPI)

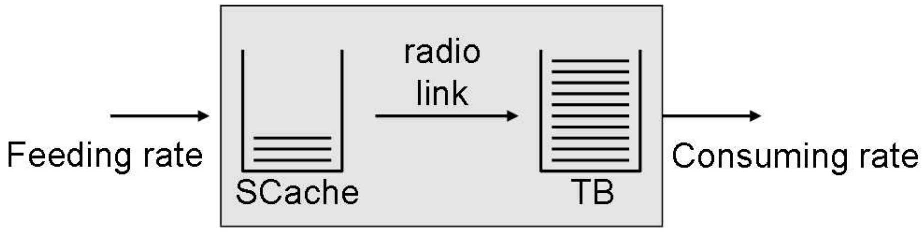

Figure 4: Buffering Concept of Smart Caching
181x47mm (150 x 150 DPI)

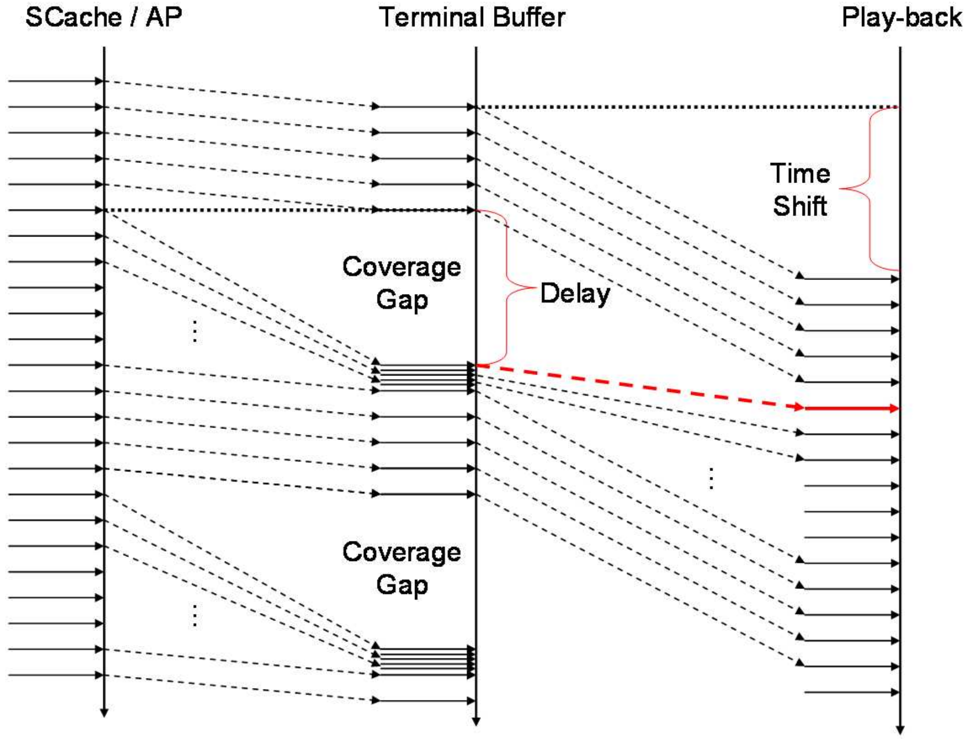

Figure 5: Relation between Time Shift and Packet Delay
185x140mm (150 x 150 DPI)

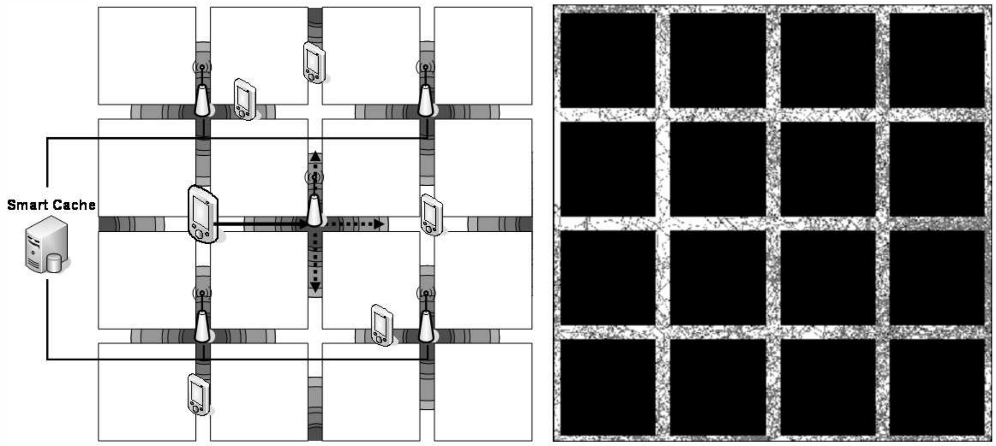

Figure 6: Urban Scenario and Mobility Simulation
197x90mm (150 x 150 DPI)

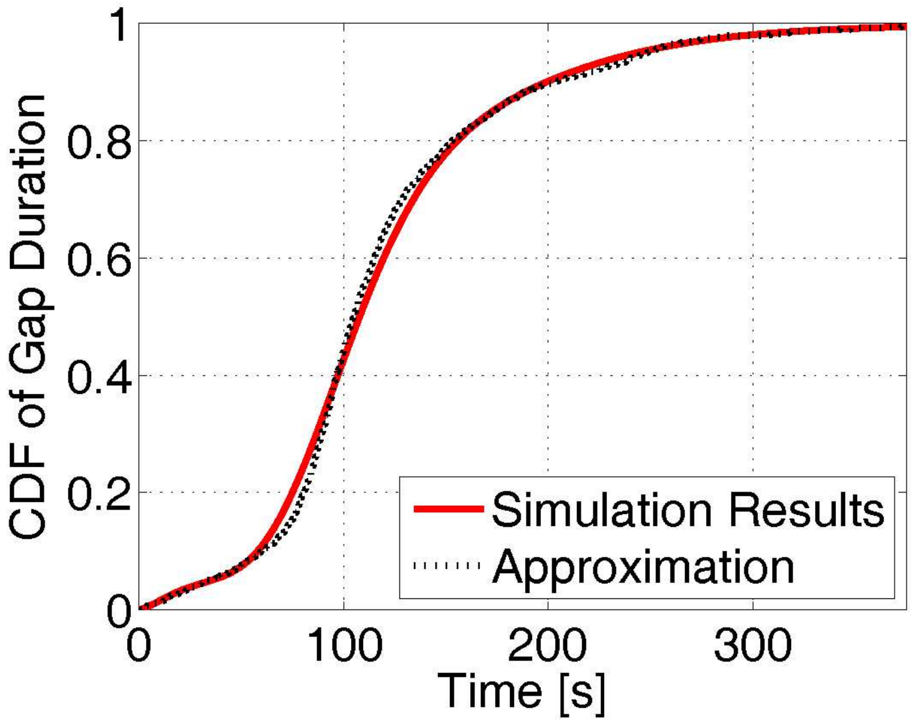

Figure 7: Gap Duration Distribution
148x111mm (200 x 200 DPI)

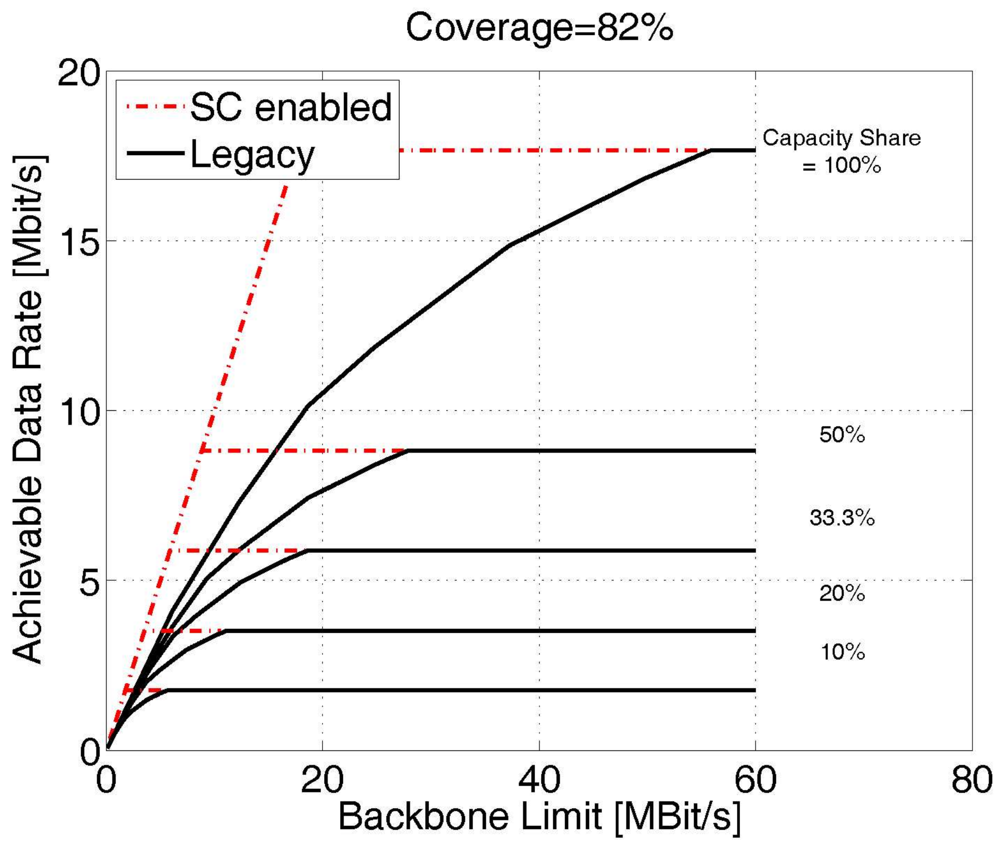

Figure 8: Data Rate depending on Backbone Limit
182x153mm (200 x 200 DPI)

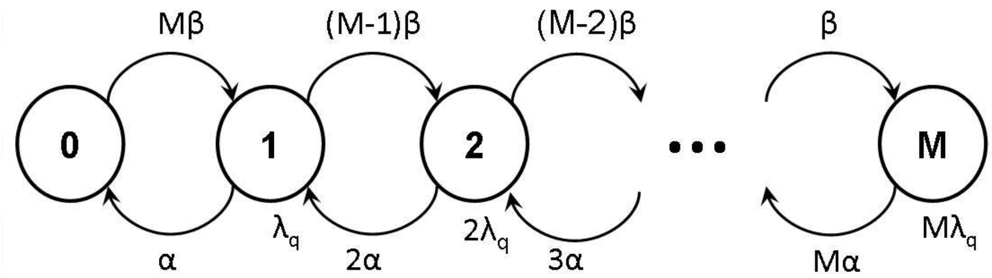

Figure 9: MMPP model for Video Streaming
188x70mm (150 x 150 DPI)

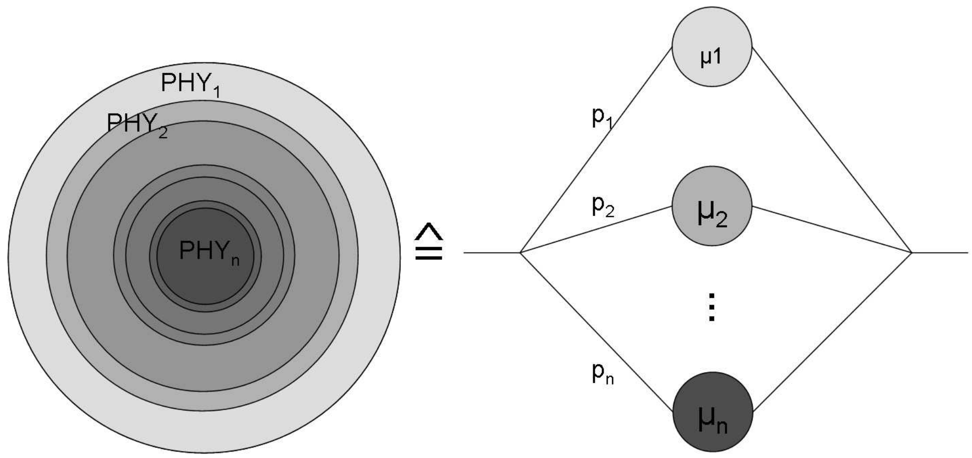

Figure 10: Service Process Modeling
250x123mm (150 x 150 DPI)

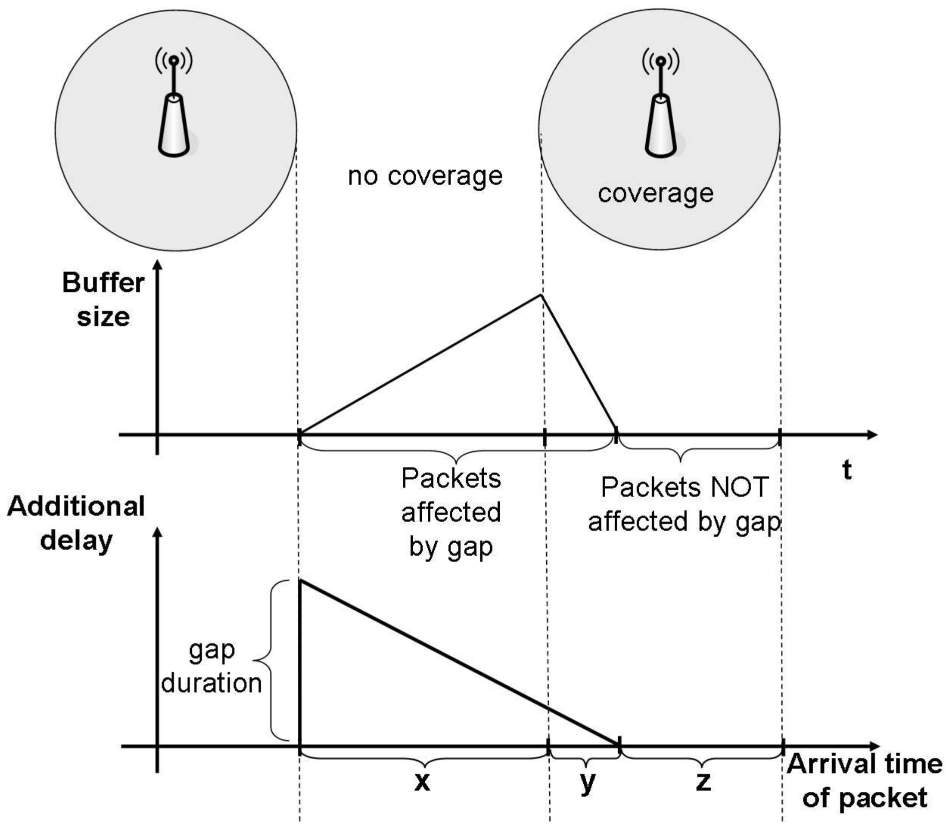

Figure 11: Influence of Coverage Gap on Buffer Size and Packet Delay
224x191mm (150 x 150 DPI)

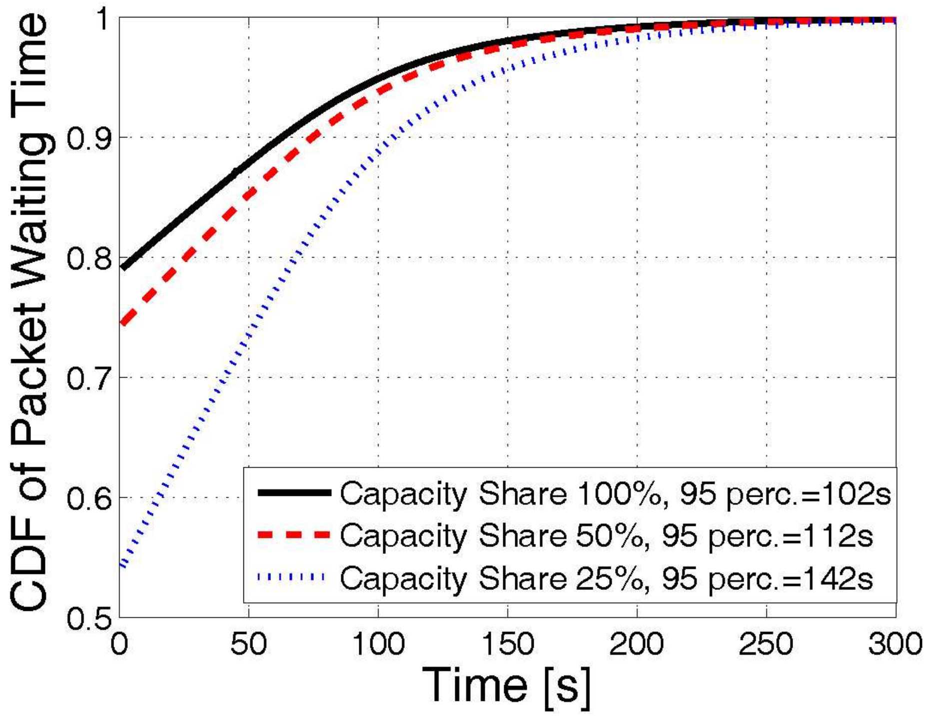

Figure 12: Packet Waiting Time in Smart Caching-enabled Urban Scenario
148x111mm (200 x 200 DPI)

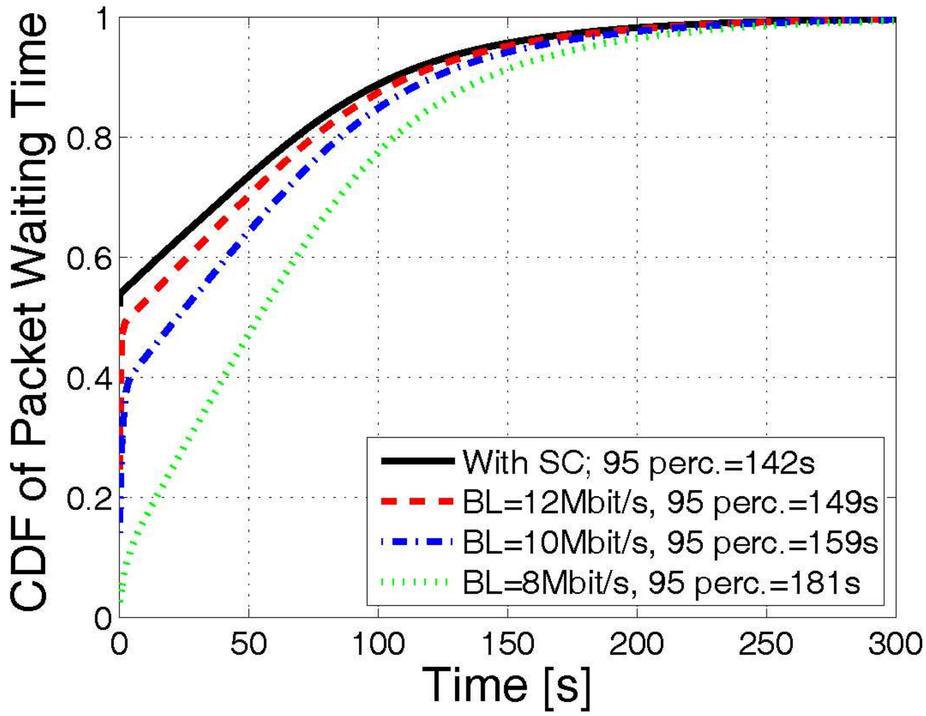

Figure 13: Impact of BL on CDF
148x111mm (200 x 200 DPI)

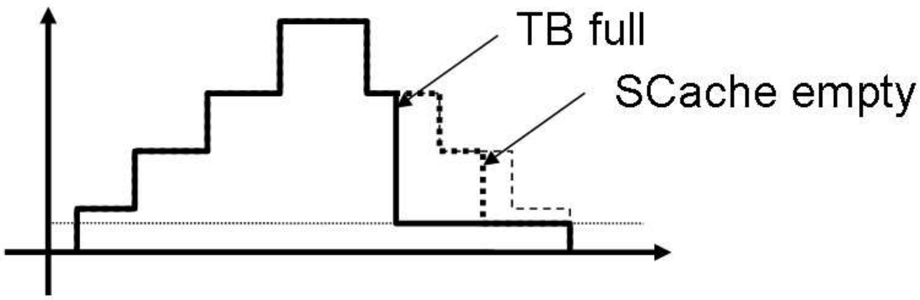

Figure 14: Impact of full TB and empty SCache
138x50mm (150 x 150 DPI)

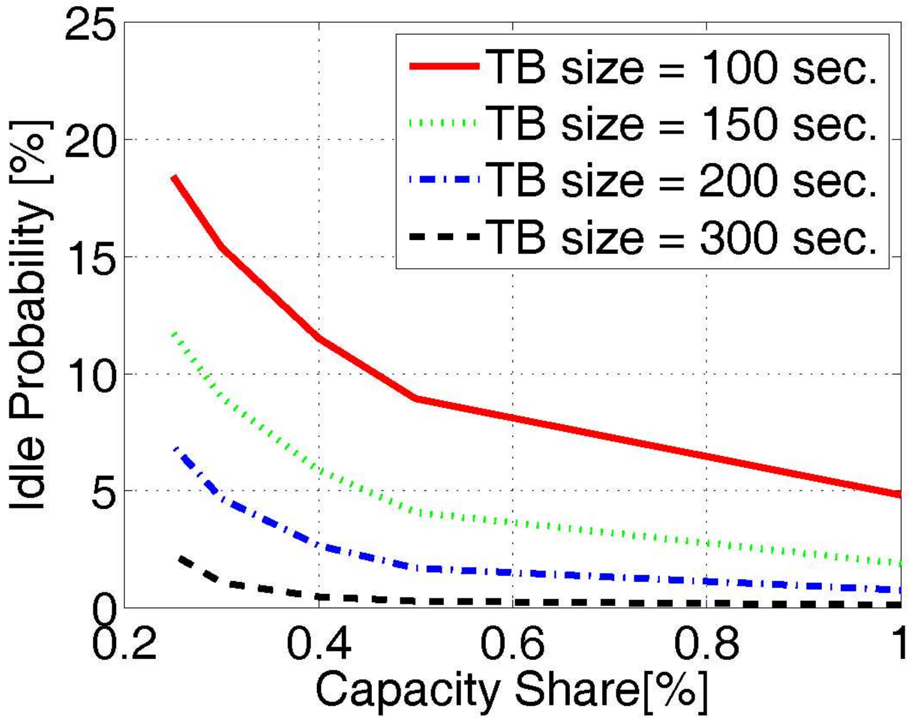

Figure 15: Impact of TB Size on Idle Probability
148x111mm (200 x 200 DPI)

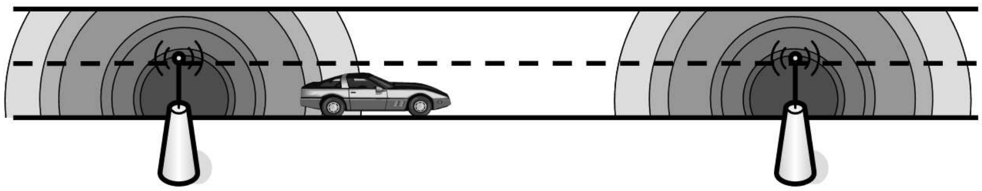

Figure 16: Highway Scenario
219x47mm (150 x 150 DPI)

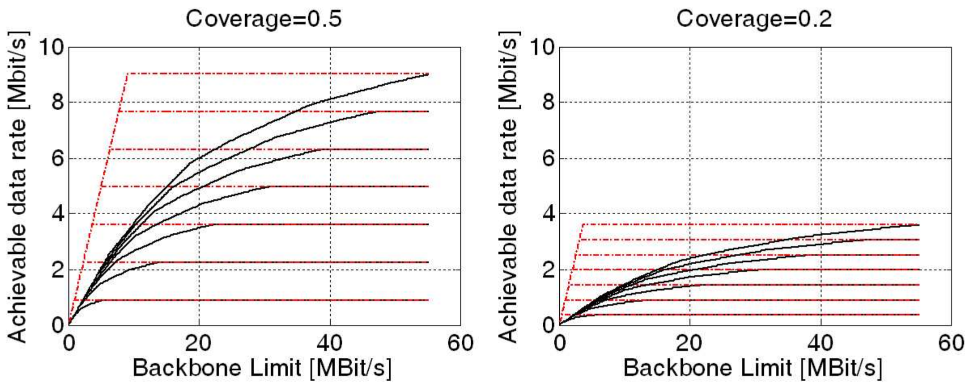

Figure 17: Data Rate depending on Network Coverage
209x83mm (150 x 150 DPI)

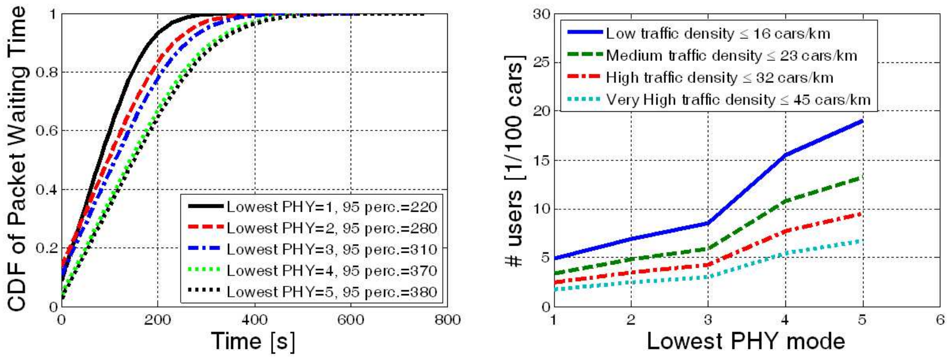

Figure 18: Capacity Gain by PHY Mode Deactivation
214x79mm (150 x 150 DPI)