Performance evaluation of backhaul bandwidth aggregation using a partial sharing scheme

Valentin Burger [a,*], Michael Seufert [a], Tobias Hoßfeld [b,a,1], Phuoc Tran-Gia [a]

[a] University of Würzburg, Chair of Communication Networks, Würzburg, Germany
[b] University of Duisburg-Essen, Chair of Modeling of Adaptive Systems, Essen, Germany

1. Introduction

According to [1], in 2014, mobile networks carried nearly 30 exabytes of traffic, which is expected to increase nearly 10-fold towards 2019. To handle the growth and reduce the load on the mobile networks, offloading to WiFi has come to the center of industry thinking [2]. In 2014, 46% of total mobile data traffic was offloaded onto the fixed network through WiFi or femtocells, and it is forecast that, by 2016, there will be more traffic offloaded than remaining on cellular networks.

In contrast to strict offloading, in which the Internet access link is switched completely (e.g., from cellular to WiFi), current concepts (e.g., BeWifi[2]) also consider multiple connections to the Internet, thereby sharing and aggregating available backhaul access link capacities. The problem arises what sharing policy to apply for which system characteristics. In the case of BeWifi, which considers access link sharing among neighboring users, each user should only share his access link when having spare capacity in order to avoid negatively affecting his own Internet connections. Therefore, two thresholds were introduced, (i) a support threshold until which utilization a user will offer bandwidth to other users, and (ii) an offloading threshold indicating from which utilization a user can offload to supporting neighbors.

In this work, we model and evaluate the performance of such a system using basic Markov chain methodology from queuing theory. In particular, we consider a scenario with two access links and investigate the impact of the thresholds on the bandwidth aggregation. Partitioned

* Corresponding author.
E-mail addresses: valentin.burger@informatik.uni-wuerzburg.de (V. Burger), seufert@informatik.uni-wuerzburg.de (M. Seufert), tobias.hossfeld@uni-due.de (T. Hoßfeld), trangia@informatik.uni-wuerzburg.de (P. Tran-Gia).
1 Now at University of Duisburg-Essen.

2 http://www.bewifi.es/.

access links and completely shared access links will be used as reference systems. We provide analytic and simulation results for the probability of available bandwidth excess, the utilization of each access link, and the received bandwidth for each user.

We show that the Markov model can be used to seamlessly evaluate the performance of systems between partitioning and complete sharing dependent on the threshold settings. We find that a system can receive less bandwidth and higher blocking probabilities than with partitioning if the cooperating system is highly loaded. However, a highly loaded system can benefit from offloading to a cooperating system by receiving considerably more bandwidth than its own capacity. Simulation with different service time distributions shows that the Markov model also holds in more general conditions.

The paper is structured as follows. Section 2 summarizes offloading and bandwidth sharing systems and technologies. In Section 3, the model of a bandwidth aggregation system is described in detail. Results of the performance evaluation are reported in Section 4, while Section 5 lays out the conclusions derived from the entire study.

2. Background and related work

The principle of sharing or offloading between multiple Internet access links is already widely used by commercial services as well as research work. WiFi-sharing communities like Fon,[3] Karma,[4] WeFi,[5] and Boingo[6] offer access to an alternative Internet link (WiFi instead of mobile), which provides a faster access bandwidth and reduces the load on stressed mobile networks. With respect to this so called WiFi offloading, the research community investigated incentives and algorithms for access sharing [3], and ubiquitous WiFi access architectures for deployment in metropolitan areas [4,5]. Moreover, [6–8] describe systems for trust-based WiFi password sharing via an online social network (OSN) app. WiFi sharing is not a legal vacuum and a first exemplary overview on Swiss and French rights and obligations was given in [9] but must be treated with caution due to international differences and interim law revisions. The opposite concept to Wifi offloading, i.e., WiFi onloading, is presented in [10]. The idea is to utilize different peaks in mobile and fixed networks to onload data to the mobile network to support applications on short time scales (e.g., prebuffering of videos, asymmetric data uploads).

An access link sharing concept, which goes beyond pure offloading, is BeWifi, which was developed by Telefonica [11] and builds on previous works about backhaul capacity aggregation [12,13]. BeWifi uses modified access points, which act as normal access points until their clients saturate more than 80% of the backhaul capacity. Then, the access point will scan for close access points, which will provide additional bandwidth if their utilization is below 70%. Backhaul capacity and utilization are announced by each access point via beacon frames. Instead of introducing a secondary WiFi radio, BeWifi uses time-division multiple access (TDMA) and the 802.11 network allocation vector (NAV) to connect to neighboring access points for bandwidth aggregation in a round robin fashion with a weighted proportional fairness schedule.

From a technical perspective, bandwidth sharing and offloading are enabled by implementing handovers and/or multipath connections, which are well covered in research. [14–16] show the feasibility of multipath TCP for handovers between mobile and WiFi networks in the current Internet and [17] describes available features for mobile traffic offloading. Furthermore, [18] gives an overview on approaches that enable mobility and multihoming. In [19] a collaborative token bucket algorithm, which implements an effective distribution of the transmission rates is analyzed to evaluate the performance of wireless ad-hoc and mesh networks.

Theoretically, bandwidth sharing between WiFi access points can be considered as load sharing among systems. Generally load sharing systems can be classified in partitioning, partial sharing and complete sharing systems. Partitioning systems work completely independent from each other. Each system has its own queue and buffer space and processes only requests arriving at its queue. Complete sharing systems have a shared queue and buffer space. When processed, a request in the shared queue is assigned to the system which is currently least loaded. Partial sharing systems have their own queues, but may offload requests to other systems if they are overloaded, or process requests from other overloaded systems. Different partial sharing or complete sharing models have been investigated in literature. In [20] the bandwidth usage by different services in a broadband system in complete sharing and partial sharing mode with trunk reservation is investigated. Multidimensional Markov chains are used in [21–23] to evaluate the performance of cellular network systems with different service categories. The blocking probability of a complete sharing system has been approximated in [24]. This approximation is used in [25] to evaluate the performance of mobile networks with code division multiplexing supporting elastic services. However, none of the models can be used to seamlessly evaluate the performance of systems between partitioning and complete sharing. In this work a model based on a two dimensional Markov chain with thresholds is developed that allows to study the transition of blocking probabilities of partitioned, partial sharing and complete sharing systems.

3. Model and analysis

In the following, we first describe the system model and the considered scenario in detail covering the notation used for parameters throughout this work. Finally, we present analytic approaches that are used to derive the resulting performance metrics like blocking probability and link utilization.

[3] http://www.fon.com.
[4] https://yourkarma.com/.
[5] http://wefi.com/.
[6] http://www.boingo.com/.

3.1. General model

In our scenario, we look at loaded access links on a short time scale. The throughput of each Internet connection is limited by a bottleneck (either on application side, on server side, or in the core Internet), such that the capacity of an access link cannot be fully utilized by a single connection. This means, each Internet connection will utilize a certain share of the access link bandwidth. The available capacity of a link c is divided into a number n of small atomic bandwidth fractions of equal size. This means, $c = n \cdot \xi$ with ξ resembling the granularity of bandwidth allocation. For example, a $c = 10$ Mbps link can be modeled as $n = 20$ bandwidth fractions of $\xi = 500$ kbps each, or also as $n = 100$ bandwidth fractions of $\xi = 100$ kbps each. For the remainder of this paper, we will consider ξ as a global constant in the given scenario and model different capacities c_i by assigning different n_i to the links.

We model an access link as a multi-server blocking system and each available bandwidth fraction of the link as a server in the corresponding system. For mathematical tractability, the overall model of an access link will be an M/M/n loss system [26] and its utilization variations will be modeled as a stationary process of singular and independent arrivals of traffic, i.e., bandwidth fraction requests. Thus, the number of occupied bandwidth fractions on each Internet link X is a random variable modeled by a birth–death-process, in which bandwidth fractions are requested with Poisson arrivals at rate λ and occupied for an negative-exponentially distributed service time with globally normalized rate $\mu = 1$. Consequently, $\rho = \frac{\lambda}{n \cdot \mu} = \frac{\lambda}{n}$ represents the load on the link. The state probability in the considered M/M/n queue is $x(k) = P(X = k)$, i.e., the probability that k bandwidth fractions are occupied.

Following the approach of BeWifi (see Section 2), two thresholds are introduced, which define the bandwidth aggregation/offloading policy. First, we use a support threshold α, which indicates at which percentage of utilization (i.e., number of own occupied bandwidth fractions) the system will stop offering bandwidth fractions to other systems. Second, we use an offloading threshold β with $\alpha \leq \beta$. If the percentage of utilization is at the offloading threshold β or higher, the system will try to use bandwidth of other systems. Thus, a system can be in one of the following three macro states:

(1) *support* ($0 \leq X < \lfloor \alpha \cdot n \rfloor$):
low utilization and offering bandwidth

(2) *normal* ($\lfloor \alpha \cdot n \rfloor \leq X < \lfloor \beta \cdot n \rfloor$):
normal operation

(3) *offloading* ($\lfloor \beta \cdot n \rfloor \leq X \leq n$):
high utilization and offloading to other systems

By applying these offloading thresholds, different Internet access links will collaborate and share traffic. More details on the bandwidth aggregation and its model are presented in the following section.

3.2. Bandwidth aggregation scenario with two access links

In this work, we consider a scenario with two different Internet access links. Fig. 1 shows a schematic view of the model as described above and highlights the most important system characteristics. In the case of two links, the actual system state can be described by two random variables X_1 and X_2, which represent the number of occupied bandwidth fractions in the respective access link. As the model components comprise the memoryless property, a two-dimensional Markov process can be analyzed using standard techniques of queueing theory.

With the state probabilities

$$x(i, j) = P(X_1 = i, X_2 = j), \quad 0 \leq i \leq n_1, 0 \leq j \leq n_2, \quad (1)$$

i.e., the probability that i bandwidth fractions are occupied in system 1 and j bandwidth fractions are occupied in system 2, the two-dimensional state transition diagram, presented in Fig. 2, can be arranged. Two major areas are visible. In the upper left part and the lower right part (white background), each system operates independently in such way that all arriving requests are served locally by this system. In the top-right and bottom-left parts (shaded in gray), one of the links is in offloading state and the other link is in support state. In these cases, all traffic arriving at the offloading link will be served by the supporting link. Thus, blocking only occurs when the other link cannot help, i.e., in states $\{(n_1, j) : \lfloor \alpha_2 n_2 \rfloor \leq j \leq n_2\}$ and $\{(i, n_2) : \lfloor \alpha_1 n_1 \rfloor \leq i \leq n_1\}$.

3.3. Model limitations

The model has limitations. A critical part of the model is the negative exponential service times, which may in reality be more deterministic, since the link throughput has low variations. However, as will be shown in Section 4.3 the model also provides good approximations for the received bandwidth and blocking probability for different service time distributions. There are different effects in real systems, which are not considered in the model. For example signaling among the cooperating access points is necessary to report the current load and the offloading state. The messages exchanged produce a signaling overhead which can limit the performance of the system. Interference can limit the capacity of the wireless links, which is not considered in our model. Finally, switching to another access link might add delays when setting up the connection and redirecting the traffic. This delay can also slightly decrease the effective throughput of the system. As these effects have only marginal impact on the system performance they are neglected in the model. The aim of the model is to evaluate the performance of bandwidth aggregation systems and to identify critical parameters.

3.4. Analysis

In order to evaluate our model and to compare it with reference systems, we analyze related systems from literature. To investigate the performance gain of bandwidth aggregation, we further analyze the bandwidth received by cooperating systems.

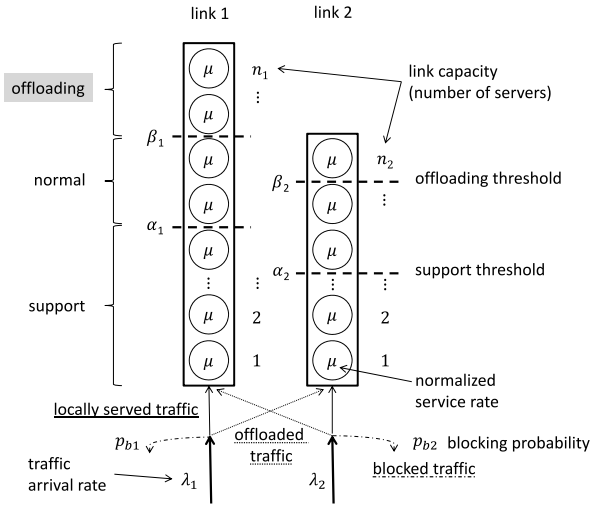


Fig. 1. System model.

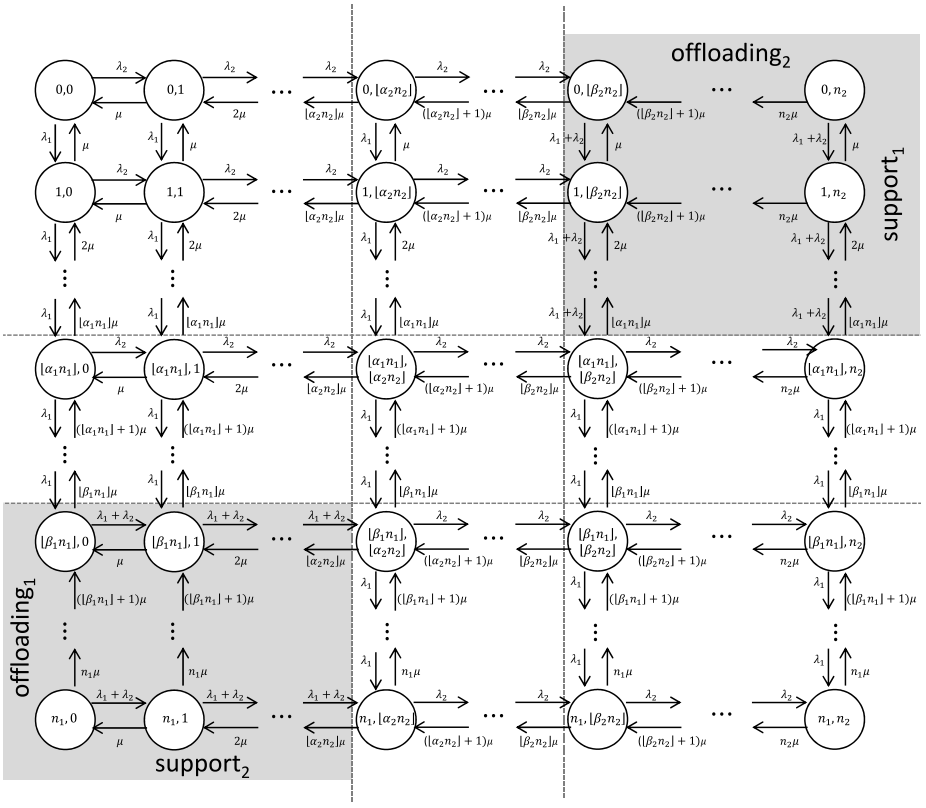


Fig. 2. The state transition diagram.

3.4.1. Reference systems

As references for the bandwidth aggregation gain with two Internet access links, both partitioned systems (i.e., without offloading) and a complete sharing system (i.e., economies of scale) are considered, although it has to be noted that in many practical cases complete sharing is physically not possible. We investigate the blocking probability p_{b_i} of each system i, which relates to the probability that the available bandwidth of the access link i is exceeded. Thus, in our scenario, blocking means that a bandwidth request of an application cannot be entirely satisfied because the link is fully utilized. In practice, if TCP is used on the access link, the Internet connections throttle themselves and share the link equally. Depending on the used application and its characteristics, the application performance can then suffer, which can result in user

dissatisfaction. Moreover, the received bandwidth of each access link is important.

For completely partitioned systems, two different M/M/n_i loss systems with arrival rates λ_i, $i = \{1, 2\}$, the received bandwidths $E_0[X_1]$ and $E_0[X_2]$ can be computed individually for each access link by Little's Theorem as

$$E_0[X_i] = \frac{\lambda_i}{\mu} \cdot (1 - p_{b_i}), \tag{2}$$

in which we use the rate of accepted arrivals $\lambda_i \cdot (1 - p_{b_i})$ and the globally normalized service rate $\mu = 1$, and p_{b_i} follows from the Erlang-B formula [26]

$$p_{b_i} = \frac{\frac{\frac{\lambda_i}{\mu}^{n_i}}{n_i!}}{\sum_{k=0}^{n_i} \frac{\frac{\lambda_i}{\mu}^{k}}{k!}}. \tag{3}$$

The performance $E_s[X]$ of a complete sharing system, i.e., a single M/M/n loss system with $n = n_1 + n_2$ servers and an arrival rate of $\lambda = \lambda_1 + \lambda_2$, can be computed by the same formulae.

3.4.2. Modeled bandwidth aggregation system

For our modeled bandwidth aggregation system with two Internet access links, we consider the blocking probability p_{b_i} of each system i and the total blocking probability p_b, which is calculated by the sum of blocking probabilities of each system weighted by the probability that a request arrives at each respective system.

$$p_{b_1} = \sum_{k=\lfloor \alpha_2 \cdot n_2 \rfloor}^{n_2} x(n_1, k), \qquad p_{b_2} = \sum_{k=\lfloor \alpha_1 \cdot n_1 \rfloor}^{n_1} x(k, n_2) \tag{4}$$

$$p_b = \frac{\lambda_1}{\lambda_1 + \lambda_2} \cdot p_{b_1} + \frac{\lambda_2}{\lambda_1 + \lambda_2} \cdot p_{b_2}. \tag{5}$$

Since requests can be offloaded from system 1 to system 2 in states (n_1, k) for $k < \lfloor \alpha_2 \cdot n_2 \rfloor$, the requests are not blocked and the state probabilities are not added to the blocking probability p_{b_1}. The same holds for states (k, n_2) with $k < \lfloor \alpha_1 \cdot n_1 \rfloor$ and p_{b_2}.

An approximation $\tilde{p}_b$ of the blocking probability p_b can be calculated by the joint probability of a single system being fully occupied, while a separate single system is above the support threshold α, i.e. could not help. If X_1 and X_2 are random variables for the number of jobs in system 1 and system 2, the joint probability is

$$\begin{aligned} \tilde{p}_b &= P(X_1 = n_1, X_2 \geq \alpha \cdot n_2) \\ &= P(X_1 = n_1) \cdot P(X_2 \geq \alpha \cdot n_2). \end{aligned} \tag{6}$$

Moreover, we analyze the mean total number of occupied bandwidth fractions $E[X]$, which corresponds to the mean of total aggregated bandwidth. Following the same argumentation as above, $E[X]$ can be computed by Little's Theorem as

$$\begin{aligned} E[X] &= \frac{\lambda_1 + \lambda_2}{\mu} \cdot (1 - p_b) \\ &= \frac{\lambda_1}{\mu} \cdot (1 - p_{b_1}) + \frac{\lambda_2}{\mu} \cdot (1 - p_{b_2}). \end{aligned} \tag{7}$$

Finally, we take a look at the received bandwidth at each access link $E[X_{A_i}]$. Thereby, X_{A_i} is a random variable for the number of bandwidth fractions (in all systems), which are occupied by arrivals from system i. It is obvious that $E[X_{A_i}] = E[X_i] = E_0[X_i]$ for the partitioned system. In case of offloading, $E[X_{A_i}]$ can be calculated from the mean total number of occupied bandwidth fractions by taking into account the share of accepted requests from each system.

$$\begin{aligned} E[X_{A_i}] &= \frac{\lambda_i(1 - p_{b_i})}{\lambda_1(1 - p_{b_1}) + \lambda_2(1 - p_{b_2})} \cdot E[X] \\ &= \frac{\lambda_i}{\mu} \cdot (1 - p_{b_i}). \end{aligned} \tag{8}$$

Nevertheless, it is the goal of bandwidth aggregation to cooperate in order to use spare capacity on access links to increase the received bandwidth where needed. Therefore, we can quantify the percentage of bandwidth gain for each system as

$$\omega_i = \frac{E[X_{A_i}] - E_0[X_i]}{E_0[X_i]}. \tag{9}$$

3.5. Simulation description

In order to validate the analytic model and to assess the system performance in more general cases, we use a discrete-event based simulation. The simulation is implemented using arrival and departure events. Each of the m systems has an arrival process with rate according to its load. The average service time of bandwidth fractions is one and the service time distribution can be specified. The simulation state holds the requests being processed and the number of occupied bandwidth fractions for each system. Offloading decisions are made according to the available bandwidth fractions in the systems.

4. Numerical examples

Using the model we aim to calculate numerical examples to evaluate the performance of the system in different scenarios. As parameters we study the load on the reference system ρ_1 and the load on the cooperating system ρ_2. We consider the blocking probability of the reference system p_{b_1} and the normalized received bandwidth of the reference system $E[X_{A_1}]/n_1$. To validate our model and to get a first assessment, we analyze the performance of systems with equal thresholds and compare the analytic results with the results obtained from simulation and those of simple reference systems. We consider the symmetric case with even load $\rho_1 = \rho_2$ to investigate the impact of the offloading thresholds and to optimize them. We then consider the asymmetric case to analyze the performance of systems with imbalanced load. We conduct parameter studies to find system configurations where one of the systems can highly benefit from offloading, e.g. by being prioritized. Finally we run simulations with different service time distributions to assess the system performance in more general cases.

Fig. 3 shows the blocking probability dependent on the system load of two server groups with equal arrival

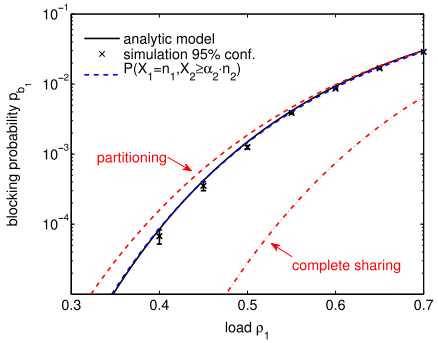


Fig. 3. Blocking probability of two systems with equal load.

processes. In this case the blocking probability is equal for both systems. Both systems have $n = 20$ bandwidth fractions, and the thresholds are set to $\alpha = 40\%$ and $\beta = 80\%$. The black line shows the result based on the analytic model for a composite system as described in Section 3.4. The markers show the mean of 8 simulation runs with 95% confidence intervals. The blocking probability increases with the load on the system as expected. The results of the simulation match the analytical model with high confidence.

For comparison the analytic result for the approximation $\tilde{p}_b$, for partitioning and for complete sharing, i.e., with combined arrival process and bandwidth fractions, is plotted. The latter equals a system with a single server group, double arrival rate and double number of bandwidth fractions. Compared to partitioning the composite system performs slightly better for low loads. For low system loads the probability is high, that one of the two systems has less than $\alpha \cdot n$ active jobs and can help if the other system is in an offloading state. The load is taken from the highly loaded system and the blocking probability is decreased. This effect is negated for higher loads on the system, since the probability to be in a support state, with less than $\alpha \cdot n$ jobs, diminishes. If the systems cannot help each other, their performance equals partitioning the systems.

To investigate the potential of the system, it is compared to a complete sharing system. The red dash dotted line shows the result of a system with double arrival rate and $n' = 2 \cdot n = 40$ combined bandwidth fractions. The blocking probability is reduced by a magnitude. This effect is also known as the economy of scale.

4.1. Offloading thresholds

In the following we investigate the setting of the thresholds α and β to optimize the performance of the system. Therefore we analyze the symmetric case with $\rho_1 = \rho_2$ and vary the thresholds α and β. The number of bandwidth fractions per system is again set to $n = 20$.

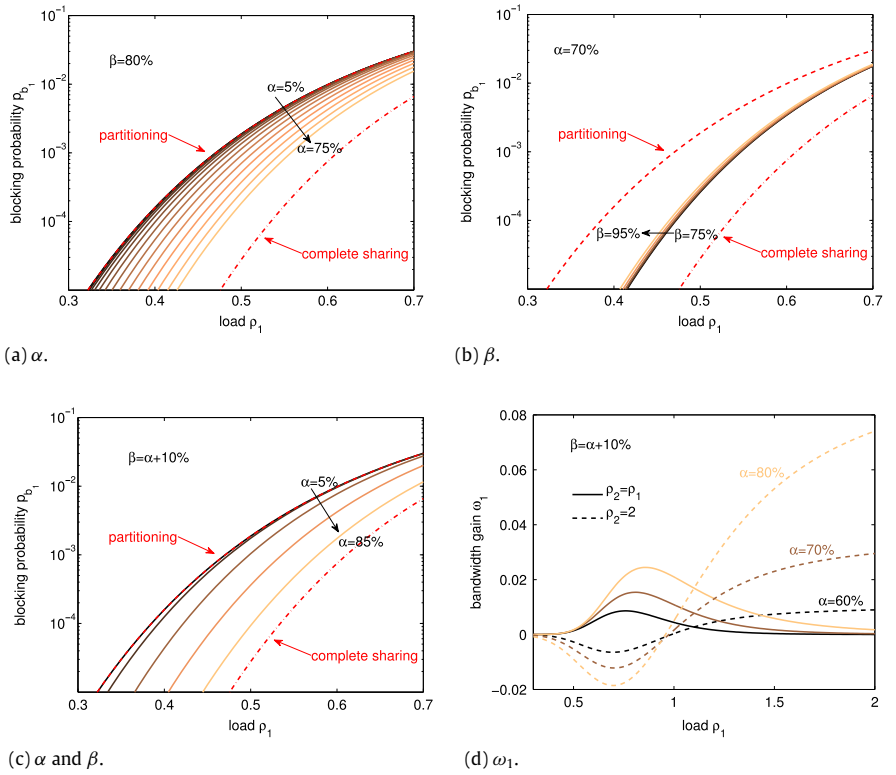
Fig. 4(a) shows the blocking probability of the reference system p_{b_1} dependent on the load ρ_1 for different support thresholds α. The offloading threshold β is constant at 80% of the system capacity. For $\alpha = 5\%$ a system only helps if it is empty and is not processing jobs. The systems work almost isolated from each other and thus the performance is equal to the performance of a single system. By increasing the support threshold α the systems can offer more help when one of the systems is overloaded and decrease the blocking probability. The support threshold α determines the amount of jobs that can be offloaded.

Fig. 4(b) shows the blocking probability of the reference system p_{b_1} dependent on the load ρ_1 for different offloading thresholds β. The support threshold α is constant at 70% of the system capacity. The offloading threshold β is increased from 75% to 95%. Increasing the offloading threshold has almost no impact on the blocking probability. The effect on the blocking probability is small, since the threshold β just shifts the point of time at which the system starts offloading. The amount of jobs that can be offloaded is not dependent on β. The reason for the slight increase of the blocking probability with β is that there are less chances to find the cooperating system in support state when β is high.

We have seen that the performance of the system depends on the amount of jobs that can be offloaded, so the support threshold α needs to be set as high as possible. Theoretically, the support threshold could be set to the offloading threshold $\alpha = \beta$, so that a system would switch directly from support to offloading mode. However, in practice this may lead to problems, since the systems could switch unnecessarily frequently among the modes. This is especially the case if mode switches result in a high signaling overhead or imply expensive context switches. Therefore, a gap is left among the thresholds. Hence, in order to prevent frequent mode switches, we set $\beta - \alpha$ to 10%. In order to maximize the available bandwidth we can increase the support threshold α. Fig. 4(c) shows the blocking probability of the reference system p_{b_1} dependent on the load ρ_1 with fixed gap $\beta - \alpha$ for increasing support thresholds α from 5% to 85%. The blocking probability decreases with increasing α, since more bandwidth fractions are shared among the systems. However, the performance of the system can also drop if the support threshold α is too high, which can be seen in Fig. 4(d). Fig. 4(d) shows the bandwidth gain ω_1, c.f. Eq. (8), of the reference system for an equally loaded cooperating system with $\rho_2 = \rho_1$ and an overloaded cooperating system with $\rho_2 = 2$. If the cooperating system is equally loaded the bandwidth gain is always positive. If the cooperating system is overloaded, the bandwidth gain is negative, if the reference system is underutilized. In this case an increasing α has a negative effect on the bandwidth gain, because less bandwidth fractions are left for arrivals in the own system. To prevent the system from being overloaded, we leave 30% of the capacity as buffer for peak periods and set the support threshold α to 70%. Hence, we set the default values of the support threshold α and the offloading threshold β to 70% and 80% respectively.

4.2. Imbalanced system load

The composite system can benefit if the load is heterogeneously distributed among the systems, such that a system which is currently busy can offload to an idle system. To investigate the performance in heterogeneous load conditions we calculate the blocking probability p_{b_1} of the reference system dependent on its load ρ_1 and the load



(a) α. (b) β.

(c) α and β. (d) ω_1.

Fig. 4. Blocking probability p_{b_1} dependent on thresholds (a) α, (b) β and (c) α and β, and (d) bandwidth gain ω_1.

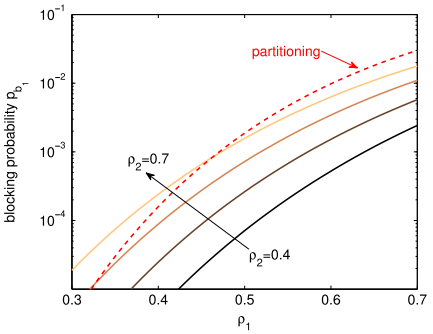


Fig. 5. Blocking probability p_{b_1} dependent on load and load on cooperating system.

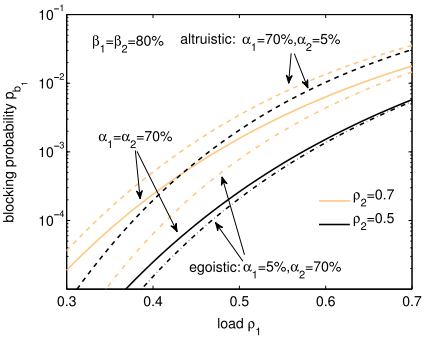


Fig. 6. Blocking probability of selfish system and altruistic system.

on the cooperating system ρ_2. Fig. 5 shows the blocking probability of the reference system for different loads on the cooperating system.

If the load on the cooperating system is low the blocking probability of the reference system is decreased. That confirms that the system can benefit from a heterogeneous load distribution. If the cooperating system is under high load ($\rho_2 = 0.7$) the blocking probability is even increased compared to partitioning if the load on the reference system is low. This depends on the fact that the traffic that is offloaded to the reference system produces a slightly higher load and increases the blocking probability.

To prevent a system from being congested from an overloaded cooperating system it can be prioritized. One possibility of prioritizing is to decrease the offloading threshold α, so that it still can get support from other systems, but shares less bandwidth fractions to help. Fig. 6 shows the blocking probability for three cases. The solid lines show the blocking probability if reference and cooperating system have equal support threshold $\alpha_1 = \alpha_2$. The dashed lines show the case where the reference system is altruistic and does not change its threshold, but interacts with an egoistic cooperating system with support threshold $\alpha_2 = 5\%$. The dash-dotted line shows the egoistic case where the reference system decreased its support threshold to $\alpha_1 = 5\%$. The altruistic system suffers from an egoistic cooperating system by receiving higher blocking probabilities. Compared to that, the blocking probability of an egoistic system is only reduced slightly. The gain of the egoistic system decreases with the load of

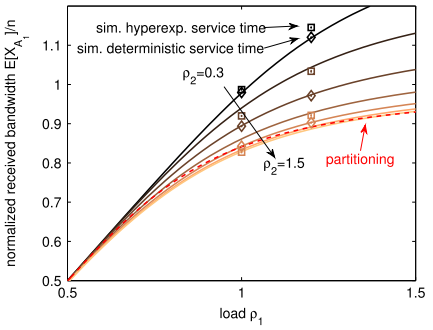


Fig. 7. Normalized received bandwidth dependent on the load of reference and cooperating system.

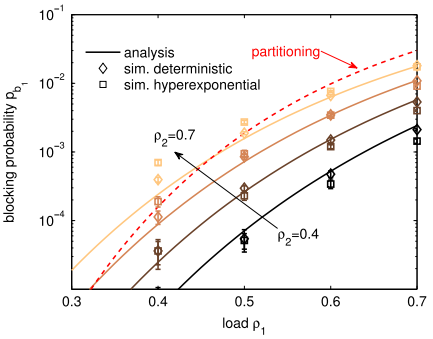


Fig. 8. Blocking probability p_{b_1} dependent on load. Simulation with different service time distributions.

the cooperating system. Hence, prioritizing is only viable if the cooperating system is highly loaded.

The performance of the system accelerates compared to partitioning if the load on one system exceeds its capacity. The system can then allocate available resources of neighboring systems and receive a higher bandwidth than its capacity. Fig. 7 shows the normalized received bandwidth dependent on the load of reference and cooperating system. For $\rho_1 < 1$ the reference system receives only slightly more bandwidth than an isolated system, if the load on the cooperating system is low. If the load on the cooperating system is high the reference system receives even less bandwidth than an isolated system. If the reference system is highly loaded it can benefit a lot from an underutilized cooperating system. If the load on the cooperating system is $\rho_2 = 0.5$ the reference system receives 20% more bandwidth if its load is $\rho_1 = 1.5$.

4.3. Simulation with general service times

To assess the system performance in more general cases we run simulations with different service time distributions. Fig. 8 shows the blocking probability of the reference system dependent on the load of the systems. The mean values with 95% confidence intervals of 8 simulation runs are plotted for the service time distributions Deterministic and Hyper-exponential. For constant service times the blocking probability does not differ from the analytic model for high system loads. The blocking probability differs slightly from the analytic model for deterministic service times in low system loads, showing higher blocking probabilities if the load on the cooperating system is high. The reason for this has to be investigated and is part of future work. In case of the Hyper-exponential distribution the service times are highly variant. Here the system which is highly loaded benefits from lower blocking probabilities compared to the analytic model.

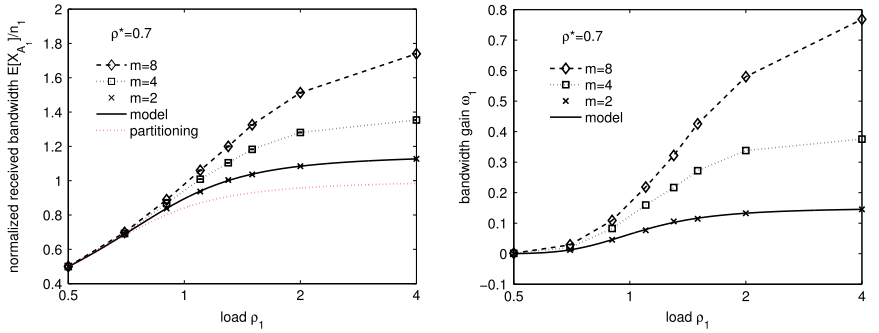
In Fig. 7, which shows the available bandwidth of the reference system dependent on the load, simulation results are plotted for Deterministic distributed and highly variant Hyper-exponential distributed service times. The service times in the Deterministic process are constant. In the Hyper-exponential process we use two branches with probabilities 10% and 90%. For deterministic service times the analytic model fits the simulation results. If the service times are highly variant the reference system receives only slightly more bandwidth than in the model if it is overloaded. Hence, considering the available bandwidth the analytic model can be used to assess the system performance with general service time distributions.

4.4. Simulation with m systems

In order to assess the potential of bandwidth aggregation of more than 2 systems, we evaluate the performance of $m = 4$ and $m = 8$ systems by the implemented simulation. We study the load of the reference system ρ_1 and set the load of the other $m - 1$ systems to the same value ρ^*, i.e., $\rho_i = \rho^*, \forall i \in 2, \ldots, m$. As performance metric we consider the normalized received bandwidth of the reference system $E[X_{A_1}]/n_1$ and the bandwidth gain of the reference system ω_1, c.f. Eq. (8). We first investigate the impact of the number of cooperating systems m for a fixed load $\rho^* = 0.7$, then we investigate the impact of ρ^*.

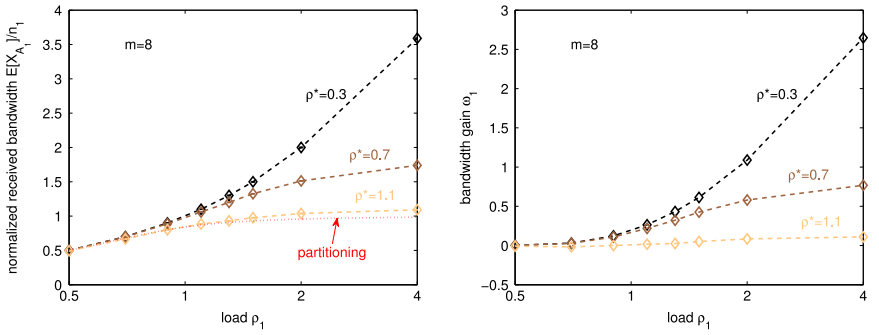
Fig. 9(a) shows the normalized received bandwidth of the reference system $E[X_{A_1}]/n_1$ dependent on the number of cooperating systems m for $\rho^* = 0.7$. The analytic model for $m = 2$ fits exactly with the simulation results. The received bandwidth increases with the number of cooperating systems m. This behavior is expected, since the amount of spare bandwidth increases with the number of cooperating systems. The reference system profits from offloading by receiving more bandwidth. For a load of more than 100% the received bandwidth exceeds its throughput n_1. This is also reflected by the bandwidth gain of the reference system ω_1, which is depicted in Fig. 9(b). The bandwidth gain is close to zero, if ρ_1 is lower than one. If the reference system is overloaded the bandwidth gain increases. Especially if the number of cooperating systems is high, an overloaded system gains a lot of bandwidth.

In the following we investigate how the load on the cooperating systems ρ^* affects the throughput of the reference system for $m = 8$ cooperating systems. Fig. 10(a) shows the normalized received bandwidth of the reference system dependent on the throughput of the cooperating systems ρ^* for $m = 8$ cooperating systems. In case of $\rho^* = 0.3$ a lot of spare bandwidth is available for offloading. If the reference system is overloaded it can use the spare bandwidth and receives almost 400% of its throughput if its load is 400%. If the load ρ^* on the



(a) Received bandwidth $E[X_{A_1}]/n_1$. (b) Bandwidth gain ω_1.

Fig. 9. Received bandwidth and bandwidth gain dependent on number of cooperating systems m for $\rho^* = 0.7$.



(a) Received bandwidth $E[X_{A_1}]/n_1$. (b) Bandwidth gain ω_1.

Fig. 10. Received bandwidth and bandwidth gain dependent on throughput of cooperating systems ρ^* for $m = 8$.

cooperating systems is higher, less bandwidth is available, which limits the received bandwidth. Still, the received bandwidth is above partitioning although the cooperating systems are overloaded with $\rho^* = 1.1$, if the reference system is even more overloaded. This can also be seen in the bandwidth gain ω_1 depicted in Fig. 10(b), which is positive if the reference system is overloaded with $\rho_1 > 1$. The bandwidth gain is only marginally negative, if the load on reference system is low, which is manageable in off-peak periods. In busy periods the reference system benefits a lot by gaining more than 2.5 times more bandwidth if $\rho^* = 0.3$.

5. Conclusion

To cope with the increasing demand of traffic carried by mobile networks, offloading to WiFi networks is considered to ease cellular networks. Recent concepts consider aggregating backhaul access link capacities to increase the available bandwidth for customers. In this work a Markov chain is developed to analyze the performance of a system with two access links that share their bandwidth. In parameter studies we investigate the impact of thresholds that decide when a system offloads to a helping system or shares bandwidth to support depending on its load. Our results show that cooperating systems benefit from aggregating bandwidth and can get close to the performance of complete sharing if thresholds are set accordingly. The received bandwidth of a system can exceed its capacity significantly if the cooperating system is underutilized. This effect is multiplied if a high number of cooperating systems are available. Part of future work is to extend the analytic model for more than two cooperating systems.

Acknowledgments

The authors would like to thank Prof. Onno Boxma for his valuable input.

This work was funded in the framework of the EU ICT Project SmartenIT (FP7-2012-ICT-317846). The authors alone are responsible for the content.

References

[1] Cisco, Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2014-2019, Tech. rep., Cisco (2015).
[2] Wireless Broadband Alliance, WBA Industry Report 2011: Global Developments in Public Wi-Fi, Tech. rep., 2011.
[3] L. Mamatas, I. Psaras, G. Pavlou, Incentives and algorithms for broadband access sharing, in: Proceedings of the ACM SIGCOMM Workshop on Home Networks, New Delhi, India, 2010.
[4] N. Sastry, J. Crowcroft, K. Sollins, Architecting citywide ubiquitous Wi-Fi access, in: Proceedings of the 6th Workshop on Hot Topics in Networks (HotNets), Atlanta, GA, USA, 2007.
[5] P. Vidales, A. Manecke, M. Solarski, Metropolitan public WiFi access based on broadband sharing, in: Proceedings of the Mexican International Conference on Computer Science, ENC 2009, Mexico City, Mexico, 2009.
[6] C.B. Lafuente, X. Titi, J.-M. Seigneur, Flexible communication: A secure and trust-based free Wi-Fi password sharing service, in: Proceedings of the 10th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, TrustCom, Changsha, China, 2011.

[7] L.J. Donelson, C.W. Sweet, Method, Apparatus and System for Wireless Network Authentication Through Social Networking, US Patent App. 13/287,931, 2012.
[8] M. Seufert, V. Burger, T. Hoßfeld, HORST—Home router sharing based on trust, in: Proceedings of the Workshop on Social-aware Economic Traffic Management for Overlay and Cloud Applications, SETM 2013, Zurich, Switzerland, 2013.
[9] G. Camponovo, D. Cerutti, WLAN communities and Internet access sharing: a regulatory overview, in: Proceedings of the International Conference on Mobile Business, ICMB, Sydney, Australia, 2005.
[10] C. Rossi, N. Vallina-Rodriguez, V. Erramilli, Y. Grunenberger, L. Gyarmati, N. Laoutaris, R. Stanojevic, K. Papagiannaki, P. Rodriguez, 3GOL: Power-boosting ADSL using 3G onloading, in: Proceedings of the 9th Conference on Emerging Networking Experiments and Technologies, CoNEXT, Santa Barbara, CA, USA, 2013.
[11] E. Goma Llairo, K. Papagiannaki, Y. Grunenberger, A Method and a System for Bandwidth Aggregation in an Access Point, WO Patent App. PCT/EP2012/064,179 (2013).
[12] S. Kandula, K.C.-J. Lin, T. Badirkhanli, D. Katabi, FatVAP: Aggregating AP backhaul capacity to maximize throughput., in: Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation, NSDI, San Francisco, CA, USA, 2008.
[13] D. Giustiniano, E. Goma, A. Lopez Toledo, I. Dangerfield, J. Morillo, P. Rodriguez, Fair WLAN backhaul aggregation, in: Proceedings of the 16th Annual International Conference on Mobile Computing and Networking, MobiCom, Chicago, IL, USA, 2010.
[14] M. Gonzalez, T. Higashino, M. Okada, Radio access considerations for data offloading with multipath TCP in cellular/WiFi networks, in: Proceedings of the International Conference on Information Networking, ICOIN, Bangkok, Thailand, 2013.
[15] C. Paasch, G. Detal, F. Duchene, C. Raiciu, O. Bonaventure, Exploring Mobile/WiFi handover with multipath TCP, in: Proceedings of the ACM SIGCOMM Workshop on Cellular Networks: Operations, Challenges, and Future Design, Helsinki, Finland, 2012.
[16] S. Chen, Z. Yuan, G.-M. Muntean, An Energy-aware multipath-TCP-based content delivery scheme in heterogeneous wireless networks, in: Proceedings of the IEEE Wireless Communications and Networking Conference, WCNC, Shanghai, China, 2013.
[17] Y. Khadraoui, X. Lagrange, A. Gravey, A Survey of available features for mobile traffic offload, in: Proceedings of the 20th European Wireless Conference, Barcelona, Spain, 2014.
[18] A. Gladisch, R. Daher, D. Tavangarian, Survey on mobility and multihoming in future Internet, Wirel. Pers. Commun. 74 (1).
[19] M. Shifrin, I. Cidon, C3: collective congestion control in multi-hop wireless networks, in: Wireless On-demand Network Systems and Services, WONS, 2010 Seventh International Conference on, IEEE, 2010, pp. 31–38.
[20] P. Tran-Gia, F. Hübner, An analysis of trunk reservation and grade-of-service balancing mechanisms in multiservice broadband networks, in: IFIP-TC6 Workshop on Modelling and Performance Evaluation of ATM Technology, La Martinique, 1993.
[21] W.-Y. Chen, J.-L. Wu, H.-H. Liu, Performance analysis of radio resource allocation in GSM/GPRS networks, in: Vehicular Technology Conference, 2002, Proceedings, VTC 2002-Fal, 2002 IEEE 56th, vol. 3, IEEE, 2002, pp. 1461–1465.
[22] Y. Zhang, B.-H. Soong, M. Ma, A dynamic channel assignment scheme for voice/data integration in GPRS networks, Comput. Commun. 29 (8) (2006) 1163–1173.
[23] K.-W. Ke, C.-N. Tsai, H.-T. Wu, Performance analysis for hierarchical resource allocation in multiplexed mobile packet data networks, Comput. Netw. 54 (10) (2010) 1707–1725.
[24] J. Kaufman, Blocking in a completely shared resource environment with state dependent resource and residency requirements, in: INFOCOM'92, Eleventh Annual Joint Conference of the IEEE Computer and Communications Societies, IEEE, IEEE, 1992, pp. 2224–2232.
[25] G. Fodor, M. Telek, Bounding the blocking probabilities in multirate CDMA networks supporting elastic services, IEEE/ACM Trans. Netw. 15 (4) (2007) 944–956.
[26] L. Kleinrock, Queuing Systems, Wiley, 1975.

Valentin Burger studied at the University of Würzburg and received his diploma degree in Computer Science in 2011. During the diploma thesis, he was working on QoE models and monitoring of voice applications. Since then he is a research assistant at the Chair of Communication Networks at the University of Würzburg. His current research focus is performance evaluation by analysis and simulation of caching in information centric networks and the utilization of edge resources.

Michael Seufert studied Computer Science, Mathematics, and Education at the University of Würzburg. In 2011, he received his diploma degree in Computer Science, and additionally passed the state examinations which are prerequisites for teaching Mathematics and Computer science in secondary schools. From 2012–2013, he has been with FTW Telecommunication Research Center Vienna where he has been working in the area of user-centered interaction and communication economics. Currently, he is a researcher at the University of Würzburg and pursuing his Ph.D. His research mainly focuses on QoE of Internet applications, social networks, performance modeling and analysis, and traffic management solutions.

Tobias Hoßfeld is professor and head of the Chair "Modeling of Adaptive Systems" at the University of Duisburg-Essen, Germany, since 2014. During the time of this work, he was heading the "Future Internet Applications & Overlays" research group at the Chair of Communication Networks in Würzburg. He finished his Ph.D. in 2009 and his professorial thesis (habilitation) "Modeling and Analysis of Internet Applications and Services" in 2013. He has been visiting senior researcher at FTW in Vienna with a focus on Quality of Experience research. He has published more than 100 research papers in major conferences and journals, receiving 5 best conference paper awards, 3 awards for his Ph.D. thesis, and the Fred W. Ellersick Prize 2013 (IEEE Communications Society) for one of his articles on QoE.

Phuoc Tran-Gia is professor at the Institute of Computer Science and head of the Chair of Communication Networks at the University of Würzburg, Germany. His current research areas include architecture and performance analysis of communication systems, and planning and optimization of communication networks. He has published more than 100 research papers in major conferences and journals, and recently received the Fred W. Ellersick Prize 2013.