ACCEPTED FROM OPEN CALL

User-in-the-Loop: Spatial and Temporal Demand Shaping for Sustainable Wireless Networks

Rainer Schoenen and Halim Yanikomeroglu, Carleton University

ABSTRACT

The demand for wireless access data rates is growing exponentially at a pace where supply cannot keep up with. Wireless resources (spectrum, time, space) are limited and shared, and transmission rates cannot be improved anymore solely with physical layer innovations. On the consumer side, flat rate type tariffs have established unnecessarily high expectations and often wasteful consumption. Dealing with congestion is unavoidable as a consequence of operating in a regime where demand is close to, equal to, or exceeding the supply. We can no longer assume that the current over-provisioning approach continues to be feasible.

Complementary to the engineering for the growth of the supply side, this article focuses on the engineering for the control of the demand side. An approach referred to as the "user-in-the-loop" (UIL) is therefore motivated here.

This article proposes spatial control, in which the user is encouraged to move to a less congested location, and temporal control, in which incentives (e.g., dynamic pricing) ensure that the user reduces (or postpones) his current data demand in case the network is congested. Results from a survey, which measures how willing a user is to respond to such control, are also presented.

As users are modeled by a system-theoretic box in a closed-loop (control) system, they feature an input handle for incentives and an output handle for the reaction.

Incentives can be progressive tariffs, reward programs, higher access rates, or even environmental (green) indicators. Incentives are tailored to the major Quality-of-Service (QoS) classes and help to shape the demand at the application layer-7 as well as at the user ("layer-8"). UIL can safely be applied in addition to other technologies, which are mainly for increasing the supplied capacity.

INTRODUCTION AND MOTIVATION

Exponential growth is observed in many indicators of human, social and economic sciences [1]. Similarly, there is an enormous growth rate of telecommunications data rates. The number of smart phones is increasing very fast and laptop computers come with cellular interfaces built in. In addition, user behavior is changing towards using more features and apps on their user terminal (UT). Many of them are not necessary or urgent at the moment, but users have built up habits and expectations that they must work everywhere with the same quality. Commercial stakeholders want users to consume more as long as the market is young and dynamic. As an inevitable consequence, this wireless traffic demand is growing faster than the supply side. Current analyses estimate that the almost 100 percent per year growth rate will continue for the next decade [2–5].

On the other side, scientists and engineers are working hard to push technology to the limit. Due to limitations by physics (like the Shannon bound, pathloss, number of antennas), computation (processing power), energy (electrical power for transmitters and computational hardware), and economics (financial limits for the capital expenditures to deploy more macro-, micro-, or pico-basestations), there is no simple and economic way of providing the same rate increase on the supply side.

In the worst case, technical solutions which aim to increase the wireless network capacity are doomed to fail, because any capacity gained by more efficiency will be used up soon by even more UT initiated traffic. We can learn from experiences also made in the computer/processor sector, where the blessings of Moore's law are eaten up by more and more hungry operating systems and applications. In wireless networks, the situation is exacerbated by the inappropriate tariff systems in the market which promotes the flat-rate abuser [6] and leaves many users frustrated due to various concealedly blocked/choked services by operators. Recent reports clearly state this dilemma of demand approaching and exceeding capacity in the near future [4, 5]. It is worth investigating solutions to improve the capacity by more spectral efficiency, such as relay nodes, femtocells, short-range communications, and multiuser access through multiple-input-multiple-output (MIMO) antennas on orthogonal-frequency-division-multiple-access (OFDMA). The better use and acquisition of new spectrum, e.g., with cognitive radio technologies, may also buy some time. However, all these new technologies will only postpone the

We thank Huawei Technologies Inc., Canada, for their support.

0163-6804/14/$25.00 © 2014 IEEE

UIL dynamic tariff interface and download manager

By temporal control, the aim is to motivate users to postpone their non-urgent traffic (updates, streaming videos) out of the busy hours. When user intends to download or open a web page, song, video, or app, a popup may appear telling the price of the transaction. If user agrees with the price then it continues the process, otherwise it may cancel the job or it may ask the download manager in which favorable circumstances to operate. A download manager in the UT can be designed in a sophisticated manner so that it knows the user's preferences, such that it can start to download automatically as soon as the current price drops below the stated threshold price, and that it knows the maximum acceptable limit by the user. Such a system needs a maximum price guarantee for the full duration of the operation. Therefore the instantaneous dynamic price is assumed constant for a flow, as soon as it is started, and up to a certain time T_c (we call it the price coherence time). Flows exceeding T_c in duration must be re-negotiated, otherwise it can be misused either by the user (extending into the busy hour) or the operator (increasing the price without telling the user). Note that "bill shock management" is not required when prices are transparent, known in advance, and the accumulated costs can always be checked online by the customer. The customer must as well be able to control financial limits per transaction and total per month. This is in accordance with parental control issues by having the parents set up the limits.

The number of popups must be minimal in order to keep the GUI use convenient and not annoying. Therefore a transaction would simply go through without questioning, if the tariff or transaction price is below the limits set by the customer.

Box 1.

time at which demand exceeds supply. By having more heterogeneity in the future deployments, the gap between hotspots and rural provisioning (spatial), and between busy hours and off-peak times (temporal), will only drift further apart.

This article describes the new user-in-the-loop (UIL) approach as one orthogonal solution to handle this dilemma of increasing demand without increasing revenue. Questioning the users' demand and traffic ingress is not necessarily a taboo. Often the appetite, e.g., for watching a video, is only triggered by minor factors like an instantaneous thought, or being bored, and allowed by the care-free promise of all-you-can-use flat-rates. On the other hand, there are important applications, such as making phone calls and writing text messages, which became part of our daily life and must be ubiquitously supported. At the same time this heterogeneous demand in time and space causes unpredictable congestion as well as imbalance problems between cell-center and cell-edge performance.

This article further defines the model for UIL, which is system-theoretic, with mixed technical, human, and economic blocks. The user is controlling a subset of the technical system, i.e., his demand, by starting and stopping the application sessions. The user activity is itself controlled by suggestions and incentives calculated by a controller block based on the current location, signal-to-interference-plus-noise ratio (SINR) and traffic situation.

We introduce the UIL system model, its spatial and temporal control domain, and define the control system model. We comment on initial results of two surveys demonstrating the user acceptance of this novel idea. The economic aspects are discussed, and some ecological aspects are addressed.

The System Model for User-in-the-Loop

The UIL concept aims at controlling the user ("layer-8") behavior in a wireless network in order to obtain a better spectral efficiency by convincing the users to move from one location to a better one or to avoid traffic congestion by postponing session traffic out of the busy hours. Depending on this impact dimension, the approach is called spatial or temporal UIL control. In both cases the user is within, and part of, a closed-loop control system.

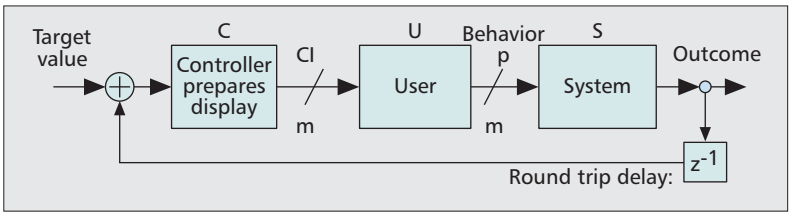

Figure 1. *The UIL system theoretic model is a closed loop with the user as the system to control. The user's output is a probability of conforming to the suggestions, which is called* p_S *in the spatial and* p_T *in the temporal domain. In a real cell we have multiple users, so the arcs between controller, user, and system are vectorized for* m *users.*

UIL extends the past assumption of the user being a traffic generating and consuming black box only, in nature similar to the noise input into a system. Instead, the system-theoretic framework allows a control input into the user block, on which the user receives suggestions and incentives (and eventually penalties) in order to convince him to diverge from the default behavior (which is uncontrolled, i.e., open loop), so that the traffic can be shaped. A user within a closed control loop receives this control information (CI) in form of suggestions on the graphical user interface, e.g., a map and directions towards a better location, a better time to start his session (out of the busy hour), or a color indicator (e.g., green, yellow, red). Once the difference between the target value (vector of goals) and the system state (global and user-specific) is computed, the controller translates this difference into user-perceivable CI and expects that user and wireless system block react according to the control, so that the system output (performance) becomes as expected and determined, which in turn gives a low dif-

ference at the controller input (note that not all users are expected or required to cooperate). The general perspective of UIL is shown in Fig. 1. Next, we describe the model in more detail for the spatial and temporal dimensions.

Spatial Control

In the UIL concept, a controller gives necessary information to the user, so it is expected that the user sometimes voluntarily changes its current location from point A to B. The current signal quality (SINR) at point A and the spectral efficiency γ_A are assumed to be known by the controller. The assumption is that the average signal quality and the corresponding spectral efficiency will be known for all relevant locations of the network from a database of previous measurements (e.g, by filtered channel quality measurements, even over past months). If γ_A is low compared to the minimum (target) spectral efficiency threshold γ_Θ, the best location B is determined by searching in the database for $\gamma_B \geq \gamma_\Theta$ within a reasonable distance d_{AB}. After that the network provides the necessary information and suggests the better position B to the user. Instead of one threshold level, there can easily be L levels with suitable new locations B_L ordered by increasing $\gamma_{\Theta L}$ and monotonously increasing distance d_{ABL}.

Before the movement, a user knows his utility advantage between points B and A, δu_{AB}. The network can determine the reasonable incentive by its own utility difference $\Delta U = U(\gamma_B) - U(\gamma_A)$. This utility advantage (incentive) can be financial (discount for real-time traffic) and/or an increased data rate (for best effort data traffic). In addition to the incentive, the network is providing the information where (in which direction to which location B) to move, e.g., by showing it on a map, compass directions or walking/driving instructions. An example for such a display can be found in [7] (note that it can by assumed in the near future the geo-location of UTs will be known almost everywhere, even inside buildings).

Before making a decision, the user has all the necessary information (such as discount rate, increased data rate, how far the next improved step is). We define the probability p_S to quantify the decision output of the user; a certain proportion of users, p_S, participates in moving and the rest of them, $(1 - p_S)$, stays in place, which includes all users that cannot move, do not want to move, or do not have enough incentive to move. The user block in Fig. 1 outputs the new location B, if the user decides to move. This probability p_S depends on factors such as the distance, the given incentive utility, and the importance of the application. Early work on UIL assumed a constant p_S [7], but recent survey results are able to better quantify this number.

The target spectral efficiency, γ_Θ, is the minimum spectral efficiency that the user should achieve after the movement. This value is not fixed and it is set by the operator. Initial studies suggest that typically half of the maximum spectral efficiency (e.g., 2.5 b/s/Hz for single antenna assumption) is a good operating point [7].

Spatial control becomes even more relevant when it comes to pico/femtocells and indoor navigation, where it is highly recommended to guide the user closer to the next hotspot. In the near future indoor navigation will be as convenient as outdoor navigation with GPS at the moment. Industry leaders are improving on this technology and providing maps for more and more public indoor environments. Performance results in [7] indicate substantial potential gains and the recent surveys endorse the users' acceptance and interest [8].

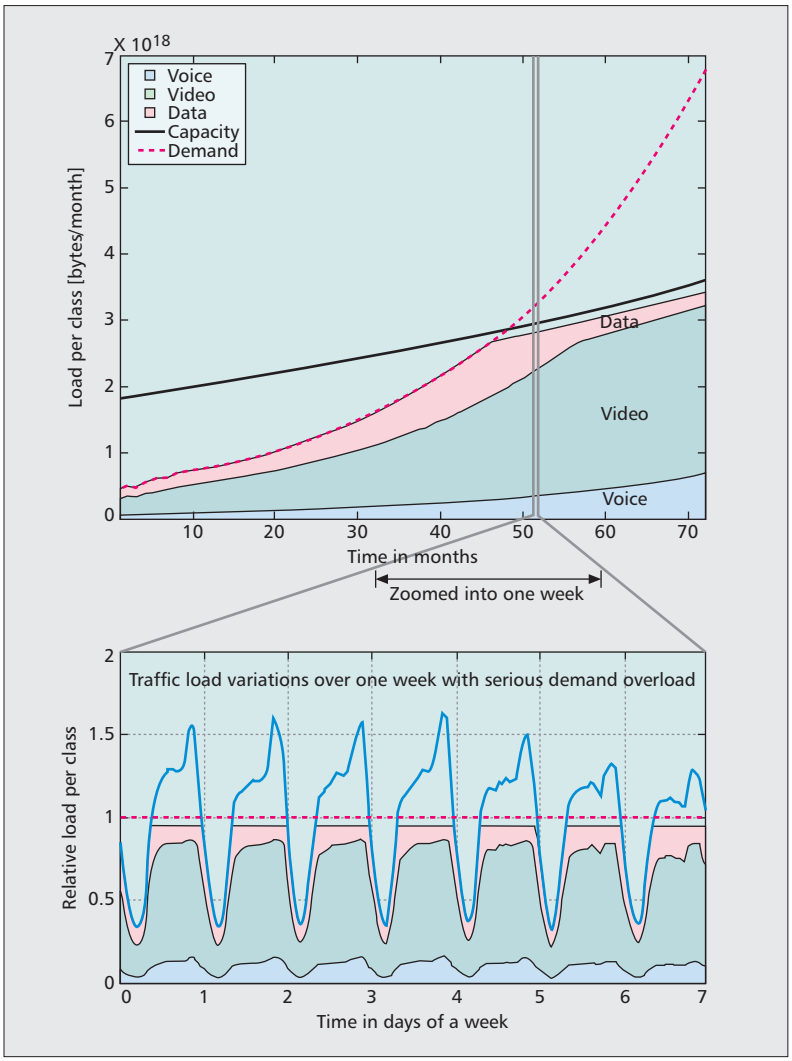

Figure 2. *Top: Aggregate mobile unconstrained traffic prediction* $r^{(u)}(t)$ *("demand"), "capacity" limitation* $\hat{R}(t)$ *(which is an upper bound of the target rate* $r^{(Target)}(t)$*) and predicted proportions of traffic classes in the controlled load region (demand overload right of the crossing lines). Bottom: zoomed in: UIL temporal control in times of predicted congestion during the busy hours. The three classes voice (blue, 10 percent), video (green, 59 percent) and data (red, 31 percent) are put together by the ratio which results from individual controlled price rates. Voice is rarely changed, but the other classes are because of their demand elasticity. The traffic sum never exceeds the capacity. Short-term packet generation processes are not modeled here.*

Temporal Control

The demand increase in cellular networks is fueled by the flat-rate pricing policy which dominates the market at the present time. Flat-rates are favored by some users in order to minimize the risk of surprisingly exploding bills (bill shock). Unfortunately, the top 20 percent of the users often account for 80 percent of the traffic

[4]. Flat-rates (especially the unbounded flat-rates) promote heavy-tailed traffic distributions and in the long term the spiral of growth leads to unbounded demand increase [4, 5]. Figure 2 shows the total estimated traffic over the coming years according to these forecasts. There will be a congestion problem at some point in the future because physical wireless capacity supply cannot grow adequately with the demand increase.

The pricing policy is expected to change because of the unbounded demand increase. Some operators have already started charging flat-rates with different caps and a variety of plans for each traffic category. It should be noted that this is only a transitional solution, because of the fact that neither the users are happy with high prices and the risk of failing service nor the operators due to the risk of running a network in congestion. Furthermore, a cap can only provide *open-loop* control on a monthly basis.

Usage-based pricing [9, 10] and dynamic (also known as congestion-) pricing [11] have been proposed before. It has been studied in the era where ATM was believed to replace the Internet protocol. The previous work on dynamic pricing has never been applied to cellular networks, closed-loop control was not considered back then, and the user behavior was not studied. Pure usage-based pricing does not solve the congestion problem in the busy hours, but dynamic pricing can. In UIL, a closed loop control [12] determines the dynamic price and communicates this to the UT. In order to be non-intrusive, in the default mode the user can optionally select to see the dynamic price on the UT. It is then displayed before each (financially significant) application session transaction. The user can then decide to use or not to use the service at the current time, location, and price. In an advanced mode, an agent or manager software on the UT can act on behalf of the user, knowing the preferences. Preferences can be taken from static settings or can be generated by machine learning. See the separate framed text "download manager" (Box 1) for a discussion. The main idea is clear - the user will generate less traffic when the session price goes up above the normal. As a result the pricing method will change the user behavior and the traffic. Similar ideas are considered in future electricity tariffs and smart-grid applications [13]. The situation in wireless is more severe though, due to the heavy-tailed consumption statistics. The user behavior is explained in more detail later.

Control Model

According to Fig. 1, the classical control system consists of the three blocks C, U, and S, in the forward path, representing the controller, user, and system, respectively. A delay block symbolizes the round-trip delay (mainly extracted from the user block and assumed to be in the order of seconds). The target value can be, e.g., $\gamma_\Theta = 2.5$ b/s/Hz for spatial control, or the tolerable network traffic load of $\rho^{(t)} = 95$ percent for temporal control. This is a simplified view, in order to help understanding the big picture by abstraction.

For control engineering, all blocks and signals in Fig. 1 have to be expanded from scalars to vectors, for m users. Each flow (user) is controlled by a separate loop, but the controller acts upon the cumulative state. Another property in the picture is the service class k of a flow, because decisions (prices, incentives) depend on the type of service (ToS, e.g., voice, video, data). Note that all users get to see the same price rate π_k (unit is \$/Mbit, but displayed in a user-understandable way, including a highly simplified approach with color indication). The system block S models the input/output behavior given the user location and new ToS requests. The controller knows the system model, but it does not know the individual user reaction.

For spatial UIL, a database in the controller knows the statistics of spectral efficiency at each location point based on the complete history of channel-quality-indication (CQI) measurements, including uncertainty intervals and some basic time-dependent process knowledge. This was not possible in the past, but computation capabilities and data storage capacity enable it now.

There is no additional signaling overhead for the measurements, as CQI is performed anyway. Signaling to the user occurs less than once per second, and only if a new and significant transaction starts. A significant transaction is the one with substantial volume (and thus, price). For temporal UIL, only a price vector is signaled, which can be broadcasted because it is the same for all users in a cell. For spatial UIL, this is at most a map of roughly 80KB size ($320 \cdot 240$ pixel $\cdot 8$ bit/pixel $\approx$ 77KB), in comparison to a transaction of several tens of MB.

For temporal UIL control, a price has to be calculated (in /Mbit). There is a different specific price for each QoS class k (e.g., nominally 3 ct/min for voice, 0.5 ct/MB for data). The price is valid for all users who are starting a new transaction (session) at the current time. After a session starts, the cell-wide global price is assigned to the session price. The session price has to stay constant for the duration of a session, in order to be transparent to customers.

The following variables denote aggregates over all users and services. The controller compares a target value (maximum utilization $u^{(Target)}$ equals to load $\rho^{(Target)}$, or absolute rate $r^{(Target)}$) with network measurements of the controlled rate $r^{(c)}$. The comparison error is $\varepsilon(\tau) = r^{(Target)}(\tau) - r(\tau)$, for each time τ, where $r^{(Target)}$ denotes the target sum rate and $r(\tau)$ is the measured rate. Note that the loop contains all the users in a cell, and all users are supposed to see the same price π_k for new sessions.

The output of the user box is the control ratio, defined as $p_T = r^{(c)}/r^{(u)}$, where $r^{(c)}$ is the controlled output rate and $r^{(u)}$ is the uncontrolled output rate (assuming a regular price level). An alternative interpretation for p_T is the proportion of users that do not change their original demand, while $1 - p_T$ of the users react and do not trigger the data transmission.

To reduce the traffic load from the uncontrolled rate $r^{(u)}$ to the controlled $r^{(c)}$ (which should approach $r^{(Target)}$), the control ratio p_T must be enforced as $p_T = \min(r^{(Target)}/r^{(u)}, 1)$. The controller can only tell the tariff levels π_k, so it must determine π_k from p_T by a reverse operation of the user transfer function.

In the near future indoor navigation will be as convenient as outdoor navigation with GPS at the moment. Industry leaders are improving on this technology and providing maps for more and more public indoor environments.

For the user's convenience, the tariff π_k is multiplied with the expected session volume (e.g., download size) to display the price per transaction in advance (e.g., $0.50 for a video).

Simulation studies revealed that the closed loop control converges to the target value in all cases, due to the controller's integral component and is thus not sensitive to modeling errors in the inverse user function. The model imprecision has an effect only on the price level and the balance between QoS classes. At this point, machine learning of the user response can dynamically improve the knowledge. The stochastic nature of the users' reaction does not have a significant impact, due to the fact that in large numbers the aggregate converges to the measured quantitative behavior. In a real implementation the UT behavior is assessed and the model is improved (for instance, by a Kalman filter) by gathering systemwide statistics on its conditional accept/deny pattern. Figure 2 (bottom) displays the resulting load situation with and without UIL control for three ToS classes, whereas Fig. 2 (top) displays the long term benefit.

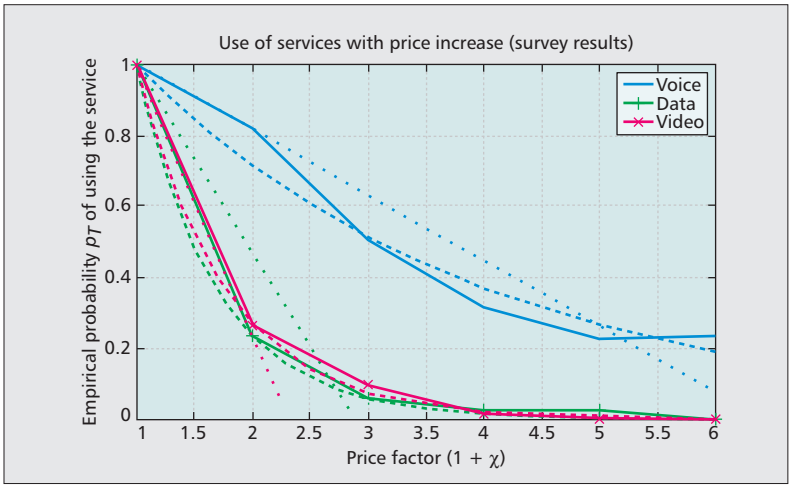

Figure 3. *Reaction to a dynamic price increase* p_T *(χ) differentiated by service [8]. The linear and exponential fits are added as dotted and dashed lines, respectively. Voice is less elastic than data or video. The wide range of the usage probability* p_T *makes a closed-loop control feasible in a wide range of demand overload.*

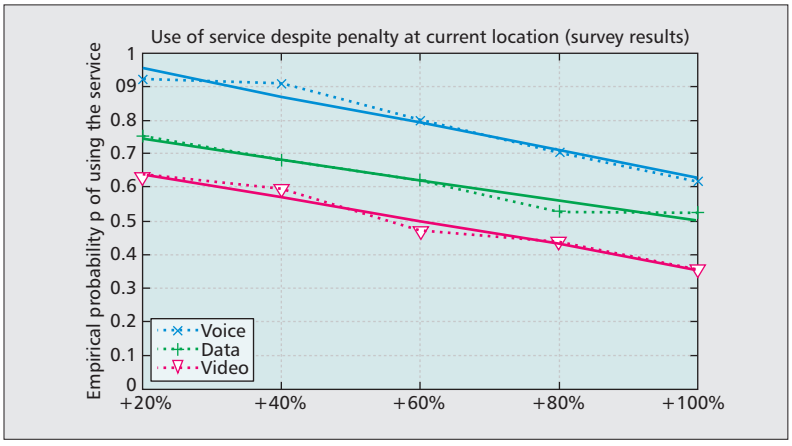

Figure 4. *This graph shows the fraction* p *of users which still want to use the service despite a penalty* ϕ *(surcharge during the busy hour). This includes users willing to relocate first and then use the service (which takes off the penalty). Voice is least elastic, but people are not bothered giving up on video consumption.*

About The User Box: A Behavioral Survey

The user (box) and it's behavior is naturally not subject to a precise science. In order to get some usable properties of the input/output system response, two surveys have been conducted [8]. In total more than 160 students at Carleton University in Canada participated in these surveys in 2011. Quantitative user behavior data is available separated by three service classes: D = data, V = video, and S = voice. The survey asks for the tolerated spatial and temporal reaction given different incentives or penalties.

Additionally the green consciousness was assessed in the above mentioned survey; the favorable response obtained rises the hope that a motivation for conforming to UIL suggestions is possible even without an explicit incentive (see Box 2).

Such a survey can only be a first step to understanding the user behavior and it is not perfectly representative for the whole population of cellular users. However, as the first study of its kind, it provides valuable insight and a better basis for quantitative studies than just assuming numbers without knowing the order of magnitude.

Survey Results and Model

The detailed survey results are discussed in [8, 14]. A fitting mathematical model derived from the empirical distributions follows an exponential shape for the function between incentive and the probability p of using the service. This p, over all flows, is exactly the same as the control ratio which the controller wants to establish. Figure 3 and Fig. 4 show the typical behavior for exemplary cases of data, voice, and video services in the temporal UIL scenario, where $p = p_T$. For the spatial UIL case Fig. 5 shows the typical behavior. The cited papers provide many more results especially for all service classes of data, voice, and video, positive and negative incentives (discount or penalty) of various strengths, as well as spatial or temporal reaction (UIL relocated or UIL postponed use).

The tolerated spatial reaction is described by the complementary cumulative distribution function (CCDF) $f_S(d)$, which quantifies the probability of a user to move d meters or more from his default position in order to get the discount δ_i (incentive index i) or to avoid the penalty ϕ_i if the service is used at the current location. Likewise, the CCDF $f_T(t)$ tells the probability of a user to postpone the service use for a time t or longer given the same set of incentive options. For example, the CCDF for a data session postponed by t can be expressed as $f_T(t) = e^{-0.0325 \cdot t/minute}$ for a discount of 40 percent [8].

For the temporal control we are also interested in the likelihood of using the service, p_T, if

the price is higher than the reference price. A reaction is only expected for an influential cost, e.g., an order of magnitude below and above $1 per transaction. Figure 3 shows the user reaction to a dynamic price increase at the current time (this could be the busy hour or an extended congestion situation). The demand response (relative price χ to control ratio p_T [12]) can be fitted well with exponential functions or a linear approximation, i.e., $p_{T,voice}(\chi) = e^{-0.300\chi}$.

Positive incentives are always embraced by users in contrast to penalties. Nonetheless, the study also reveals the behavior of how many users would still want to use a service despite a penalty ϕ. Results in Fig. 4 include the options of either paying the penalty or relocating spatially. Remarkably, video services are not considered very important compared to voice calls. Data services range in the middle field. It can be concluded that the use of video is very elastic, i.e., easy to control with slight monetary incentives or penalties. With the additional fact that video applications easily consume more data rate by one or two orders of magnitude (i.e., dominating the future traffic), it is clear that the total consumed rate can be easily controlled as well. Recent results also reveal that a control of voice traffic is most likely not necessary in the future, assuming its negligible share of the total demand. More results can be found in [8] and [14], including separate quantities for the voice, video and data service.

An additional survey finding is that the young generation easily embraces the UIL idea, as it is more accustomed to contributing towards greener goals (see Box 2).

Economic Aspects

Presently, the operators spend $50 billion per year into cellular infrastructure because of the increasing demand [5]. This is in line with the current practice of over-provisioning the capacity, in order to minimize the service outage probability. Assuming that demand will exceed the capacity soon, this approach is not only very costly but also at the end of its feasibility. In the resulting congestion, applications on layer-7 will see severe degradations, e.g., freezing of videos and near stalls due to unwanted TCP reactions for data traffic. As thus the quality-of-experience (QoE) drops to zero in congestion state, it is no longer useful to assume the revenue as proportional to the carried traffic. Instead imagine a fair payment being charged only for the satisfied part of the traffic, in order to account for the many unsatisfactory sessions. We call this *virtual refund policy* by assuming operators have to give refunds for unsatisfied sessions, without really doing so. It is obvious that the virtual revenue (proportional to carried traffic minus unsatisfied traffic) drops beyond the point of 100 percent load.

By demand shaping (in congestion), new investments into infrastructure can be avoided, postponed, or reduced. This results in less over-provisioning and saves infrastructure investment CAPEX and OPEX. Future network planning can also be done in a more flexible manner.

UIL does not only reduce demand to the sustainable level of just below the capacity. Rather, with UIL it is now feasible to operate in demand overload. By its nature, it also leads to increasing tariffs during the times of demand congestion. The effect on revenue is that for demand levels above 100 percent of the capacity, the carried traffic is still constant, but the revenue rises with the price. Thus we observe that operating beyond 100 percent demand load could even be considered desireable by operators, as it provides better revenue than without UIL. Figure 6 shows the revenue comparison for these cases.

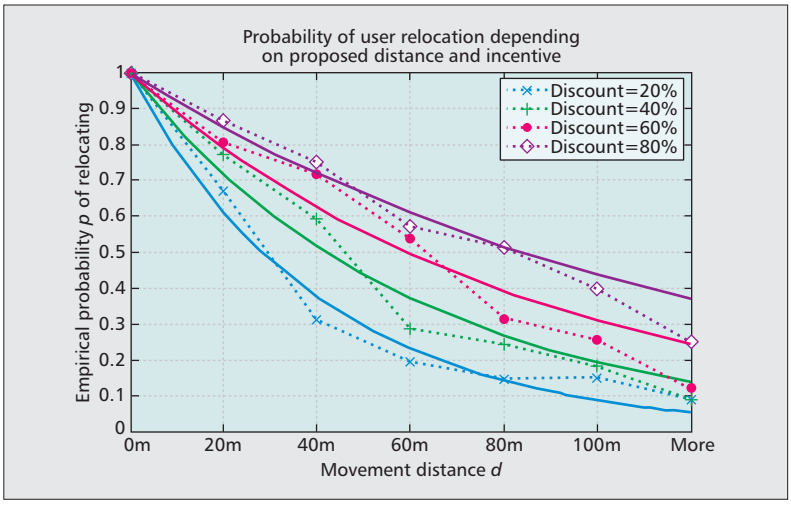

Figure 5. *This figure shows the reaction* (p_S) *to relocation suggestions (proposed distance* d*) with different discount incentives for data traffic [8]. The solid lines are exponential fits. Moving within a small distance is easily accepted. Moving more requires an incentive or a good reason.*

We can even account for some unsatisfied clients in the case when using UIL, as for some unsatisfied clients (share $1 - p_T$) the tariff is considered "too expensive." This dissatisfation is multiplied to the revenue to account for unhappy (and virtually refunded) clients in Fig. 6. As the graph shows, the virtual revenue with UIL is still increasing despite the dissatisfaction and it is high above the virtual revenue of the case without UIL.

The difference between the "no limits" revenue in Fig. 6 and the virtual revenue with dissatisfied customers is proportional to the incentive of an operator to invest into new infrastructure. Thus this metric can be derived for each cell individually and indicates the point in time when to upgrade the infrastructure.

The spatial UIL helps into the same direction, as more sessions can be carried when users are guided towards locations of better spectral efficiency. As a result, the operator can expect more revenue. As shown in [7], the gains can be up to 200 percent.

Abandoning the flat-rate tariff model turns out to be beneficial for the operator, and even for the majority of customers, namely those who are not extreme users. Even if some stakeholders don't want to dry out flat-rates, there is a way to emulate usage based pricing by implementing a refund policy proportional to the unused volume at the end of the billing period. Thus, UIL incentives can work by promising refunds proportional to the saved dynamic costs.

Green Aspects and Outlook on Sustainability

While discussions on campaigns like "one less bit" (one less car) will have limited impact, it is well worth focusing on high volume usage. The green index of wireless communications is the emission of 34kg of CO_2 per 1 GB of wireless data (Box 2). Spatial UIL helps reducing this impact, because less energy is needed per bit for those who relocate to a better spectral efficiency.

The spiral of growth is a big concern. Similar to the race between CPU, RAM, hard disk space on one side and hungry (inefficient) software on the other side, a race between capacity and traffic will also happen. All technical solutions for providing more capacity for uncontrolled traffic are doomed to fail, because any capacity gained will be filled up soon by more traffic. By deploying more macro, micro, or femto cells, the total energy consumption and CO_2 emission will continue to increase.

By using UIL, this spiral can be tamed. The user puts pure demand into relation with cost and benefit. Thus UIL enables networks to become "sustainable," i.e., without the need to upgrade for a longer time. The users' satisfaction with prices will define when to upgrade.

The long term effect is even better, as very low-latency feedback enables a learning or training effect for the user. For example, for the user it will become subconscious when and how to best watch an online video after experiencing the price for this transaction in different locations and times.

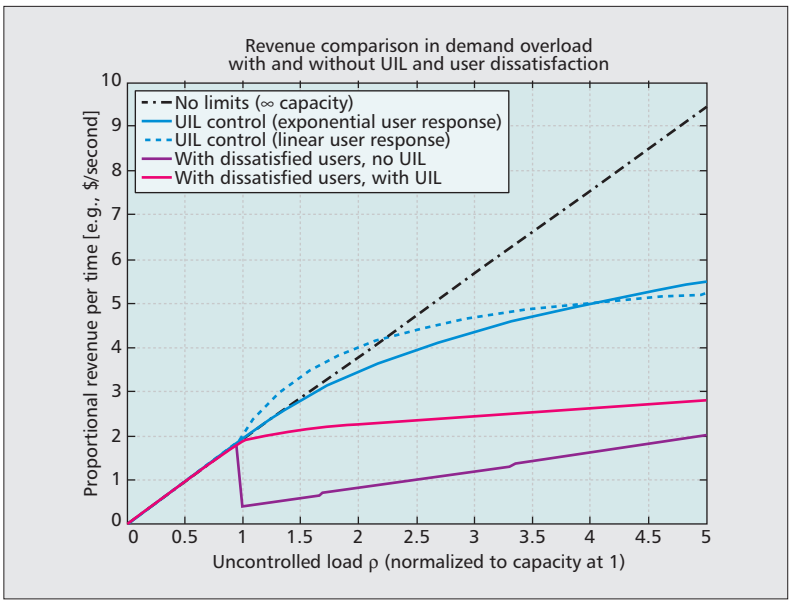

Figure 6. *Revenue comparison considering user satisfaction as a function of the unconstrained demand. The "no limits" case is hypothetical without capacity limitations. In all other cases, the excess load limit is* $\rho_{max}^{(u)} = 1$. *Increases in revenue beyond that point are possible with UIL due to dynamically increasing prices, keeping the real load* $\rho^{(c)} < 1$. *When taking "dissatisfaction" into account (all users dissatisfied who cannot carry the traffic due to costs or failures), UIL is perceived better, because the user has a choice, in contrast to failing applications in congestion, thus the better revenue. Without UIL the satisfaction shown here is > 0 only because voice traffic is still carried without disruptions.*

Criticism that this might be too invasive is not justified, because in underload there is no different price from normal (and not even an annoying message). In overload the alternative to compare UIL with is the situation of frequent "application failures" for all the participants, which is even much less desirable.

Apart from cellular networks, there is a number of applications of the closed-loop UIL approach in all areas where scarce resources need to be (fairly) distributed and capacity cannot be exceeded. The green training effect of UIL is psychologically very effective, because it includes immediate feedback (unlike a bill one month or even a year later) and the incentive is substantial (financial). The smart-grid research is already going to start changing people's habit towards off-peak usage [13]. The virtual power plant concept is equivalent to a virtual capacity in networks. In its full extent and in long term, this can help in limiting growth [1] of average power used per person and thus, for example, help in phasing out nuclear energy, and reduce fossil fuel combustion in order to stabilize the earth climate. In a world of limited resources [1], UIL allows to communicate the responsibility back to the point of origin, which is the individual consumption. This approach would help solving the tragedy of the commons [15].

Conclusion and Outlook

Although technology is improving to provide better capacity in wireless networks of the future, according to the current data it seems impossible to keep up with the increase in demand. The paradigm so far has been to over-provision capacity in order to avoid any blocked calls, failed applications or other effects of congestion. The picture can be changed completely.

This article investigates the novel UIL paradigm which suggests changes to the user session initiation in a closed control loop. Incentives (or penalties) are used in times when the network needs to reduce the demand or increase the total spectral efficiency in order to avoid congestion. This is only applied in critical overload conditions, whereas in regular (limited load) times there is no difference to the current engineering practice. For a fair comparison, the use of UIL and its implications must be related to the alternative scenario of permanent traffic overload, where users' QoE drops dramatically due to unsatisfied expectations. UIL also enables soft call admission control (CAC), in contrast to hard CAC or no CAC at all.

Recent simulation results based on survey data show that a stable operating point is easily obtained with UIL and the control is robust against model inaccuracies of the hard-to-predict user behavior. Operating in demand overload now becomes feasible, and economic reasoning shows that this even allows saving CAPEX and OPEX for new infrastructure. Results in Fig.2 indicate that UIL for voice might not be required at all, due to the low proportion of voice in the total estimated traffic in the future.

Abandoning the flat-rate tariff for non-voice traffic is an important step towards sustainable networks, but there are ways to motivate UIL by

equivalent incentives, as discussed in this article. Network neutrality is an important principle which is not compromised by UIL because it treats all users the same at a given time and location within the same QoS class. Important is also the full transparency of the reason for dynamic prices, as a customer right, in order to avoid any misuse by the stakeholders involved. How to treat machine-to-machine communication traffic is an interesting question for the future, but it only becomes relevant when its traffic will be in the same order as human-initiated traffic, and by that time business models can include usage-based pricing and an entity responsible for billing.

Research into this direction is just beginning [14] and it needs field experiences and trials with real users to convince skeptics, but the potential is enormous and applications are possible far beyond wireless communications, for instance, in transportation systems or the smart-grid [13].

Green aspect of wireless communications [12]
Wireless cellular networks consume 0.5% of the world total electricity which is approximately 20 PWh in 2010. The average monthly cellular wireless traffic was $240 \cdot 10^{15}$ bytes in 2010. Then the energy per byte can be found as $0.0347 \cdot 10^{-6}$ kWh and it is equal to 0.125 J. If the electricity is obtained from coal, then 975g CO_2 arises for 1 kWh of energy. This means that for one byte of wireless data 0.03383 mg of CO_2 arises, which is approximately equal to 34g of CO_2 for 1 MB. Provided this information, in a survey it was asked "If you are told that for each MB that you download, you produce around 30 grams of CO_2, on a scale between 0 and 10 how much do you care to adjust your data usage to be greener?" Results show that rather concerned people (score 5 and above) are in the majority by almost 3/4.

Box 2.

References

[1] J. Rockstroem *et al.*, "A Safe Operating Space for Humanity," *Nature*, vol. 461, 2009, pp. 472–75.
[2] "UMTS Forum report 44 — Mobile Traffic Forecasts 2010–2020," http://www.umts-forum.org/, UMTS Forum, Tech. Rep., Jan 2011.
[3] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010–2015," Cisco Systems Inc.," White Paper, Feb. 1, 2011, available: http://www.cisco.com/en/US/solutions/.
[4] "2010 Mobile Internet Phenomena Report," Sandvine Inc., Tech. Rep., 2010, available: http://www.sandvine.com/downloads/documents/.
[5] "Mobile Broadband Capacity Constraints and the Need for Optimization," Rysavy Inc., Tech. Rep., Feb. 2010, available: http://www.rysavy.com/Articles/.
[6] H. Ekstrom, "QoS Control in the 3GPP Evolved Packet System," *IEEE Commun. Mag.*, Feb 2009, pp. 76–83.
[7] R. Schoenen, H. Yanikomeroglu, and B. Walke, "User-in-the-Loop: Mobility Aware Users Substantially Boost Spectral Efficiency of Cellular OFDMA Systems," *IEEE Commun. Letters*, vol. 15, no. 5, May 2011, pp. 488–90.
[8] R. Schoenen *et al.*, "Quantified User Behavior in User-in-the-Loop Spatially and Demand Controlled Cellular Systems," *Proc. European Wireless*, Poznan, 2012.
[9] C. Courcoubetis *et al.*, "A Study of Simple Usage-based Charging Schemes for Broadband Networks," *Telecommun. Systems*, vol. 15, Dec. 2000, pp. 323–43.
[10] J. Altmann and K. Chu, "How to Charge for Network Services — Flat-Rate or Usage-based?," *Computer Networks*, vol. 36, no. 5–6, 2001, p. 519.
[11] L. A. DaSilva, "Pricing for QoS-Enabled Networks: A Survey," *IEEE Commun. Surveys Tutorials*, vol. 3, no. 2, 2000, pp. 2–8.
[12] R. Schoenen *et al.*, "Green Communications by Demand Shaping and User-in-the-Loop Tariff-based Control," *Proc. 2011 IEEE Online Green Commun. Conf. (IEEE GreenCom'11)*, Online, 2011.
[13] A. Vallejo *et al.*, "Next-Generation QoS Control Architectures for Distribution Smart Grid Communication Networks," *IEEE Commun. Mag.*, vol. 50, no. 5, May 2012, pp. 128–34.
[14] R. Schoenen, "User-in-the-Loop Project," http://userintheloop.org.
[15] G. Hardin, "The Tragedy of the Commons," *Science*, vol. 20, 1968, pp. 1243–47.

Biographies

RAINER SCHOENEN [SM'13] (rs@comnets.rwth-aachen.de) received his German Diplom-Ingenieur and Dr.-Ing. (Ph.D.) degrees from RWTH Aachen University, in Electrical Engineering in 1995 and 2000, respectively. His research interests include stochastic Petri nets and queuing systems, ATM, TCP/IP, switching, flow control, QoS, tariffs, User-in-the-loop (UIL), wireless resource and packet scheduling and the MAC layer of 4G and 5G systems. His Ph.D. thesis was *System Components for Broadband Universal Networks with QoS Guarantee* with the ISS group of Prof. Heinich Meyr at RWTH Aachen University, Germany, from 1995 to 2000. He started working self-employed in 2000. He was a senior researcher at the Communication Networks (ComNets) Research Group, RWTH Aachen with Professor Bernhard Walke from 2005 to 2009, working on computer networks, queuing theory, Petri nets, LTE-Advanced, FDD relaying, scheduling, OSI layer 2 (MAC) and IMT-Advanced Evaluation within WINNER+. He is currently a project manager at the Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada, working with Professor Halim Yanikomeroglu.

HALIM YANIKOMEROGLU [S'96, M'98, SM'12] was born in Giresun, Turkey, in 1968. He received the B.Sc. degree in Electrical and Electronics Engineering from the Middle East Technical University, Ankara, Turkey, in 1990, and the M.A.Sc. degree in Electrical Engineering (now ECE) and the Ph.D. degree in Electrical and Computer Engineering from the University of Toronto, Canada, in 1992 and 1998, respectively. He was with the R&D Group of Marconi Kominikasyon A.S., Ankara, Turkey, during 1993–1994. Since 1998 he has been with the Department of Systems and Computer Engineering at Carleton University, Ottawa, where he is now a Full Professor. His research interests cover many aspects of wireless technologies with special emphasis on cellular networks. He co-authored more than 50 IEEE journal papers, and has given a high number of tutorials in leading international conferences on wireless technologies. In recent years, his research has been funded by Huawei, BlackBerry, Samsung, Communications Research Center of Canada, Telus, and Nortel. This collaborative research resulted in about 15 patents (granted and applied). He has been involved in the organization of the IEEE Wireless Communications and Networking Conference (WCNC) from its inception, including serving as a Steering Committee member as well as the Technical Program Chair or Co-Chair of WCNC 2004, WCNC 2008, and WCNC 2014 to be held in Istanbul. He was the General Co-Chair of the IEEE Vehicular Technology Conference Fall 2010 held in Ottawa. He has served on the editorial boards of *IEEE Transactions on Communications*, *IEEE Transactions on Wireless Communications*, and *IEEE Communications Surveys & Tutorials*. He was the Chair of the IEEE's Technical Committee on Personal Communications (now called Wireless Technical Committee). He is a Distinguished Lecturer for the IEEE Vehicular Technology Society. He is a recipient of the Carleton University Faculty Graduate Mentoring Award in 2010, the Carleton University Graduate Students Association Excellence Award in Graduate Teaching in 2010, and the Carleton University Research Achievement Award in 2009. He spent the 2011–2012 academic year at TOBB University of Economics and Technology, Ankara, Turkey, as a Visiting Professor. He is a registered Professional Engineer in the province of Ontario, Canada.