Green Communications by Demand Shaping and User-in-the-Loop Tariff-based Control

Rainer Schoenen^{1,2}, Gurhan Bulu^{1,3}, Amir Mirtaheri¹, Halim Yanikomeroglu¹

¹Department of Systems and Computer Engineering, Carleton University, Canada ²Communication Networks (ComNets), Faculty 6, RWTH Aachen University, Germany ³Department of Electrical & Electronics Engineering, Hacettepe University, Turkey rs@sce.carleton.ca

Abstract—The new field of green communications can be divided into a) energy-efficient communications equipment or b) using information and communication technology to improve the world to become more energy aware. Mobile data traffic and utility energy consumption have a lot in common. There is a limited supply due to limited resources, and only growth (of quantities, technology) can increase this, at the cost of a higher carbon footprint. The green index is defined here for cellular wireless.

On the other hand there is a demand which is user-generated, variable over time and space, and ever increasing at a fast pace. Flat rates or almost flat utility tariffs have spoiled users and established high user expectations. Instead of engineering for the growth of supply, this paper investigates how to engineer for controlling the demand side. Dealing with congestion is a consequence of the supply=demand regime and the end of overprovisioning. New tariffs are required that are tailored to the major QoS classes and help to shape demand at the user and application level. This paper investigates QoS-aware demand shaping and control by user-in-the-loop and tariff-based control.

Index Terms—Demand shaping, user-in-the-loop (UIL), congestion, tariff, flow control, green index, sustainability, cross-layer

I. INTRODUCTION

G REEN Communications has recently got more and more attention. Power consumed by wireless infrastructure like base stations, switching centers currently already accounts for 0.5% of the global electric power consumption and therefore the carbon emissions. Putting contemporary data together results in a carbon footprint of 34 g CO₂ (or 17 dm³) for 1 MB of transmitted data. We can call this the current *green index* of wireless cellular communications. One bit corresponds to $5.8 \cdot 10^{16}$ molecules of CO₂ is the specific bit emission¹.

However, technical solutions which aim at making the wireless network more efficient in terms of specific bit energy are doomed to fail solving the big picture, because any

capacity gained by more efficiency will be used up soon, so that in effect the total energy consumption will increase and not decrease. We can learn here from experiences also made in the computer/processor sector. With the current increase rate of wireless application traffic (92% p.a.) it is absolutely impossible to come up with technical solutions to provide the required capacity in a few years, because boundary conditions like economic feasibility put a limit on any number and type of base station or access point device. Recent reports clearly state this dilemma of demand exceeding capacity [1]-[3]. Relay nodes [4], [5] or femtocells will only postpone the time shortly at which demand exceeds supply [6]. Clearly on average the customer is not willing to pay more than he does now for his tariff plan. On the operator side, companies spend \$50 billion per year into cellular infrastructure investments. How can this be balanced? Due to the complexity of this dilemma some analyst said "Networks have to adopt a more holistic solution". A very similar dilemma exists these days with the demand-supply balance of the electric power grid, when parts of the society want nuclear power supply to be phased out, while demand is expected to rise, especially due to emerging 'applications' like electric cars and more desalination plants.

It is a belief statement that the battle between supply and demand cannot be won, if supply in the end depends on limited and non-renewable resources (like coal, gas, oil, uranium). Without suitable measures, demand will approach capacity more and more often during the day until congestion (or temporary blackouts) become the rule and not the exception. Congestion is a serious condition for a network as packet delays become infinite or packet losses become intolerably high.

Instead we propose to solve the problem on the demand side. The approach is called demand shaping, a generalization of traffic shaping (or electric power usage shaping). The idea is to control the resource usage, i.e. data transfer, to a level which is sustainable. All devices and users are subject to control, so that in total the demand stays securely below the supply. One very old solution example is known as call admission control (CAC), which requires a protocol to request a connection (call) and has a chance of rejecting it. While this is useful (and recommended) for real time traffic like e.g. voice or video on a pay-per-view channel, data traffic is not controllable like this.

Gurhan Bulu is supported under postdoctoral research fund by The Scientific and Technological Research Council of Turkey (TUBITAK)

¹Wireless cellular networks consume 0.5 % of the world total electricity which is approximately 20 PWh in 2010. The average monthly cellular wireless traffic is $240 \cdot 10^{15}$ bytes which is totally 2880 PB in 2010. Then 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 975 g CO₂ arises for 1 kWh of energy. Then for one byte of wireless data 0.0338325 mg of CO₂ arises, which is approximately equal to 34 g of CO₂ for 1 MB.

In the future the by far largest traffic sector will be data traffic. Net-neutrality is not a issue because our approach encourages fair usage and emphasizes the larger share of responsibility of the heavy users. By the smarter utilization of network resources the total energy consumption is reduced not only by less transmit power, but mostly by avoiding or postponing the installation of new cell towers, which will keep the green index low.

Demand shaping for data traffic (DS) in our approach is achieved by the following components:

- A closed loop control between all network components, where hop-by-hop is most effective, but larger loops will be tolerable
- The user himself must be part of the User-in-the-loop closed loop control
- Each wireless bottleneck has a network controller (e.g., in the BS/eNB) to send information to the user's application or the user himself via the operating system on his user terminal (UT)
- Usage-based tariff plans give incentive to reduce data usage until a stable operating point is reached per user
- Long-term trends in traffic will be controlled by longterm tariff-plan announcements, but dynamically, not with a fixed contract
- Short-term traffic crunches (e.g., during busy hour or in crowded locations) are controlled by a dynamic price, so that each user can compare the urgency of his use case against the price he is willing to pay
- Remaining congestion situations are resolved by a wireless flow control [7] and fair scheduling [8], so that no data is lost and the data rate per user is fairly shared

The paper organisation is as follows. First the model of userin-the-loop is introduced. Then the control loops are identified in the ISO/OSI model. A tariff model that allows elastic demand response is proposed next. In the last section numeric simulation results are presented. The key contributions in this paper are summarized in the conclusion.

II. USER IN THE LOOP

The concept of user-in-the-loop was developed to control the user behavior in a wireless network so that a higher goal was achieved [9]. This goal can be the total spectral efficiency of a cellular network by convincing users to move from one location to a better one [10] or to postpone their use case out of the busy hour. There must be an incentive for the user, which is a better data rate in case of wireless best effort (BE) data traffic applications or a better price (discount, saving) on the cost per minute for voice and other real-time (RT) traffic applications.

This incentive plus the suggestions what to do to improve both (the network utility and the advantage for the user) is aggregated into the *quantified user information* (QUI). See Figure 1 for the complete picture. The user reacts with a generic behavior, which can be spatial (a movement as in [9]), temporal (postpone the phone call) or by any other change in usage of the network system **S**.



Fig. 1. Closed loop with the user as the system to control.

In this paper the user **U** receives tariff information on his display, which will influence his usage of the cellular network. So the expected behavior is a reduction in traffic volume if the bill is going to be too high and an increase as long as it is very affordable.

The controller C is a part of the management functionality in the network. It prepares the data to display to the user in a user-friendly and informative way. The input of the controller is the difference between the target value (objective) and the measured state (outcome) of the system. In practice this is a multidimensional problem and each individual user can get a different display or incentive, while the target value usually is a scalar and easy to understand for network operators. Section IV will show the application of this method.

For the purpose of demand shaping, there are two time scales. One is the long-term control to avoid traffic to exceed the capacity (Figure 3), and the other is a short-term control within hours, minutes or seconds to avoid short-term congestion situations due to busy hour peaks and bursty applications. The UIL concept is usable for both.

In the long-term case, the display shows a user-friendly and comprehensive explanation of the current tariff plan (of this month, week, day), separated by quality-of-service (QoS) class plus the current data volume accumulated so far this month and the predicted monthly cost on the next bill. This can be made mandatory by policy in order to make costs completely transparent to the user. No hidden costs or fees.

In the short-term case the user should get the cost information related to a transaction he is going to make. For example, before a phone call gets through the display (or a voice telling machine) already shows the cost per minute (CPM) for the number dialled (should be a customer advantage/right to see the costs beforehand, especially when roaming in other countries). Another example for data or video applications: The display shows the price tag² before clicking on a link (and committing the purchase, i.e. agreeing to the contract). E.g. the user moves the pointer on a link which leads to another web page, a file download, or a video stream and without clicking an infobox appears that shows the estimated price (on-the-fly) for clicking on that link. The network can simply calculate the price given the size of the file, the OoS class and the user's tariff plan. In case of expensive transactions (lets say \$0.50 and above) there will be a dialog box asking to confirm the transfer. All the thresholds for displaying the price on-the-fly

²Options for the price tag (Figure 2(b)) are 1) the application knows it and uses an API of the operating system (OS), 2) the OS uses a system popup or friendly voice, 3) an agent software (e.g. download manager app) knows acceptable limits by the user's preferences and autonomously accepts prices below the limit or asks otherwise.



Fig. 2. Moving from an uncontrolled to a controlled network with short loops including the user.

or as a confirm dialog or to have it locked at all should be settings in the user terminal (UT) to tradeoff between user convenience, price transparency and parental control issues.

There is huge potential in this approach, because the control is very fast (the round trip delay z^{-1}) can be within the order of seconds. That is why the user apparently gets immediate feedback and enough information to make an educated and responsible decision based on all required information. It is impossible that companies cheat on him and present surprise bills of \$6000 one month later. The fast feedback allows training and adaptation of the user to the typical situations, like avoiding busy hours or thinking twice before watching a video over the cellular network when it costs more than \$1 per ten minutes. The training effect does in the long term lead to a more responsible behavior of the aggregate set of users, i.e. a green behavior of using resources only when it is appropriate (affordable, sustainable) compared to the demand. The total solution is green because network expansion and the resulting energy costs can be postponed or avoided.

The author believes that this will also be a solution for other areas of demand vs. supply of scarce resources, e.g. when each electric appliance in a smart grid (connected to narrowband powerline network) would display the price tag (on a small LCD display) before or while using it.

III. GAPLESS CONTROL IN THE ISO/OSI MODEL

The closed loop control in the last section (Figure 1) did not specify where in the network the QUI data is generated and how the data is propagated. The idea of the display of QUI information here is similar to the UIL display in [9]. In the ISO/OSI reference model, each network protocol stack consists of seven layers from physical layer (1) to application layer (7). Between client (UT) and server there can be gateways and routers in the Internet, and a base station (BS) and optionally relay nodes (RN) in the wireless network (Figure 2). These days mobile applications on the UT are called 'apps', so it's easy to imagine that an app itself can prepare the tariff data, e.g. by showing a price tag next to each hyperlink (clicking of which would cost more than a certain threshold). Alternatively also the operating system (OS) on the UT can prepare and display the data, e.g. by showing a dialog box after the user clicks on the link or calls a number. The OS would communicate with an agent in the network to ask for the price of this requested URL. In Figure 2 the two alternatives are



Fig. 3. Aggregate mobile traffic prediction r(t) and capacity limitation $\hat{R}(t)$. According to [3] demand will exceed the supply around the year 2014, if there are no measures taken to protect from instability.



Fig. 4. PDF of the monthly data usage per user. It is a heavy-tailed distribution with a Pareto-like shape ($\kappa = 0.95$, $\sigma = 5$).

indicated by the arrows going up to the user "layer" (8). This agent would be in the wired Internet and can determine the size of the file before transfering it. That is what the application layer on the server side could also be responsible for.

In Figure 2 there are also flow control loops shown which can (optionally) ensure that data transmissions are not lost in the network even if there is congestion [7].

IV. DEMAND SHAPING BY USAGE-BASED TARIFFS

The dilemma of the unlike increase of demand and supply is shown in Figure 3. Any unbounded exponential increase of r(t)can inherently be called unstable or unsustainable, especially



Fig. 5. Tariff plan options are @ flat rate (w/o cap ③), ① linear pricing or ④ nonlinear (e.g. logarithmic) pricing.

if the Laplace transform of the per-user traffic r(t) is

$$r(t) = 86.1164MB/month \cdot e^{(0.053163 \cdot t/month)}$$
(1)

$$\mathcal{L}\{r(t)\} = R(s) = 86.1164/(s - 0.053163)$$
 (2)

and has a pole on the right side of th s-plane. The data based on [1] shows a serious problem around 2014 when nothing is changed. The crash date itself is speculation, but the different increase rates for wireless traffic r(t) and cellular network capacity $\hat{R}(t)$ is what makes this moment come sooner or later. Engineers work hard to provide more capacity, with new generations only every 10 years, and companies work hard to finance the \$50 billion per year for infrastructure (and engineers). On the other hand some of the data usage these days is questionable and only appears because people have flat rate plans and do not see (and care about) the cost per individual transaction. The massive success of smartphones and apps has led to a traffic profile that is very heavy-tailed, e.g. the top 5% of the users contribute for 50% of the traffic. The mobile wireless data data profile is shown in Figure 4 with values obtained from [2]. We easily fitted the data into a (heavy-tailed) Pareto distribution with $\kappa = 0.95$, $\sigma = 5$ and $\mu = 0$, which is valid for an average monthly usage of $\bar{R} = 100MB/month$ (February 2010):

$$pdf(r) = \frac{1}{\sigma} \cdot (1 + \frac{\kappa}{\sigma} \cdot (r - \mu))^{-\frac{1}{\kappa} - 1}$$
(3)

The current tariff plan options are shown in Figure 5. Linear pricing ① is simple, fair to a certain degree, but does not match the heavy-tailed demand (Fig. 4). Of course, a flat rate plan ② does not motivate for any change in behavior, but it is also a worry among network operators that the spirits that they called turn against them, because of the perspective of Figure 3. Introducing a cap ③ into a flat rate plan (rate limit) is only an emergency measure. Case ① and ② are special cases of ③ when c = 0 or b = 0. Logarithmic pricing ④ is a new idea and defined in Eq. 4 and Figure 7, where b is the tariff



Fig. 6. Elasticity η means quantifying the user's reaction (reducing demand) when the costs relate to (go up with) the demand. With a nominal bill \bar{K} there will be no change, i.e. p = 1.



Fig. 7. Logarithmic tariff for QoS class 1=wireless real-time(voice,video), 2=wireless data, 3=Internet data

per decade. E.g., for each 10 times more traffic, b = \$10 could be charged. Logarithmic pricing is defined by:

$$K(r) := \max(K_{min}; b \cdot \log_{10}(r/\bar{R} + 10^{-6}) + \bar{K}).$$
(4)

The literature on pricing packet data is in a niche but the majority of experts favor usage-based pricing [11]-[15]. The idea is here to have a usage-based pricing, a plan known and displayed openly to the user. The pricing scheme will change the user behavior in a certain way. The output of the U in Figure 1 is a new PDF, modeling the behavior of all users (aggregate). How it affects the user behavior is modeled by the elasticity between price and demand. Figure 6 shows that a user does not change $(r = \overline{R})$ if his bill K stays the same (at all-users average level \bar{K}). This corresponds to a user (or download manager agent on behalf of him) observing the regular price on a price tag. The user will use less traffic when the bill goes up, and use more when it gets cheaper. For simplicity we use the following linear assumption here for the control ratio p, defined as the quotient of the controlled output rate r_c divided by the uncontrolled output rate r_u :

$$p := \frac{r_c}{r_u} = \max(1 + \eta \cdot \frac{K - K}{\bar{K}}; 0).$$
(5)

The elasticity factor is assumed $\eta = -1/4$ in Eq. 5, i.e. -25% rate for a price twice as average. By definition this is already the final stationary value after all user considerations. The control ratio p can alternatively be interpreted as the proportion of users that stay with their original demand, while 1-p of the users change their mind and do not trigger the data



Fig. 8. PDF of the monthly data usage per user after tariff-based control.



Fig. 9. Control ratio p = f(r) for different tariff controls.

transmission. In this interpretation p = 1.25 means that 25% more users are attracted to use the service. The change of p will reshape the PDF of the monthly usage such that the heavy tail becomes less and therefore the mean value decreases. Quantitatively, Eq. 5 means that a user at $K = \bar{K}$ does not change, which corresponds to a monthly average given by the tariff functions in Figure 5.

V. NUMERIC RESULTS

For demonstrating the effect of pricing as a control feedback to steer user behavior, the analysis here uses the realistic predictions of Figure 4 (which represent the open-loop = uncontrolled behavior) and inserts the tariff-based UIL controller with a target value of $R_{target} = 0.8 \cdot \hat{R}$. $\bar{K} = \$25/month$. Figure 8 shows the PDF of the demand per user (February 2014) with different pricing schemes. For each scheme, the user reacts differently because of the control ratio p(r) (Eq. 5), shown in Figure 9. The logarithmic pricing ④ in Figure 5 charges the user gracefully over all ranges of possible monthly rates, without punishing him too extreme, which is what all linear or cap-based tariffs do.

A. Long-term traffic control

The trend in Figure 3 is assumed to be still valid for the intrinsic user demand and the resulting uncontrolled rate output. Proportional to the demand the PDF(r) is assumed to stay Pareto-shaped (Eq. 3 with $\kappa = 0.95$), but scaled towards higher averages with $\sigma(t) = \overline{R}(t) \cdot (1 - \kappa)$. Figure 8 shows the



Fig. 10. Aggregate mobile traffic prediction r(t), capacity limitation $\hat{R}(t)$ and controlled traffic $r_C(t)$ under Poisson traffic assumption.

PDF before and after control with different pricing schemes. Pricing 2 by flat-rate with $a = \overline{K}$ will not change the PDF and is considered the uncontrolled situation.

For each time step *t* (by month or day here) the controller knows the error $\epsilon(t) = R_{target}(t) - r(t)$. In order to reduce the uncontrolled traffic load r_u to R_{target} , the control ratio *p* of Eq. 5 must be chosen as $p = R_{target}/r$. From Eq. 5 the required price change can be obtained:

$$\chi := \frac{K - \bar{K}}{\bar{K}} = (p - 1)/\eta. \tag{6}$$

The next step is the adaptation of a pricing parameter, depending on the pricing model from Figure 5. First we assume proportional pricing with a = 0 and b as free parameter. Then b is adapted according to $b = \overline{b} \cdot (\chi + 1)$ where \overline{b} is the regular price factor (without control). This closes the control loop, because the tariff model and price information are the input to the user block. The user acts according to Eq. 5 and the target value is achieved.

Figure 10 shows the resulting traffic (day-by-day traffic fluctuations included) after introducing the closed-loop control with a target value of 80%. This is sufficient as a long-term control strategy. Of course, statistical variations of the load situation exist in space and time. In this case the QUI has to be dynamic depending on the short-term load situation. Relocating can be an option [10], but conditional prices are possible easily with the UIL approach.

B. Short-term traffic control

The next problem is to control of the short-term traffic pattern, at least at session level, so that exceeding the capacity becomes unlikely. Due to the immediate feedback, UIL provides instantaneous user persuasion (and training). The costs are now adapted and displayed per transaction. Figure 11 shows the typical daily traffic fluctuations (averaged per hour and normalized to 100% at maximum). Because of its periodicity a Fourier transform appears natural and gives the dominant harmonics to be

$$r(t) = 0.6254 - 0.3583 \cdot \sin(2\pi \frac{t/h - 1.428}{24}).$$
(7)

The average level is expected to be at $\overline{R}/\hat{R} = 80\%$ of the capacity due to the long-term control of the previous subsection. Therefore the uncontrolled system would exceed the capacity by 28% (0.8/0.6254 \approx 1.28). Short-term control will set a different target value to be at 95% of the capacity. A dynamic pricing control can reduce the load. According to the linear price model in Figure 5 with a = 0 and assuming the same elasticity from Eq. 5, $\eta = -1/4$ to reduce the highest traffic by 28% would roughly require a busy-hour price 85% higher than the average level. The highest price level in the busy-hour decreases as the traffic decreases and it reaches the average value if the traffic reaches the average value. Results are shown in Figure 12.

In the examples above the numbers for elasticity and pricing schemes are not known exactly, so reasonable assumptions had to be taken. Future work will quantify the user response.

VI. CONCLUSION

In this paper the green aspect of communications is improved by emphasizing the awareness of the carbon footprint (green index) in terms of CO₂ emissions per MB transmitted. With resource aware users, the user-in-the-loop approach enables controlling the traffic demand by incentives. A usagebased pricing is used as interface data between controller output and user display. This solves the problem of congestion and instability for the near future when traffic exceeds capacity. The green idea behind this approach is that user demand is elastic and users are not willing to pay certain prices for services of low importance but high data rates (e.g. streaming video). The approach is called demand shaping and is the opposite of overprovisioning capacity. The solution in this paper aims at sustainability (green) by reducing excessive use of limited resources while at the same time average users will stay on the same level of average monthly bills. Therefore infrastructure needs to be built less and at a later time, whereas users are trained to understand why certain high-rate applications are prohibitive in the busy hours. It is expected that users will not get used to satisfy high demand (e.g. for videos if it is not worth paying the price) when we change to usage-based pricing soon, before it is too late and leads to customer dissatisfaction. Future work will describe the detailed control system parameters for all tariff plan options and provide results from a user survey.

References

- "Cisco visual networking index: Global mobile data traffic forecast update, 2010–2015," Cisco Systems Inc., White Paper, February 1, 2011.
 [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.html
- [2] "2010 mobile internet phenomena report," Sandvine Inc., White Paper, October 2010. [Online]. Available: http://www.sandvine.com/downloads/documents/2010% 20Mobile%20Internet%20Phenomena%20Report.pdf
- [3] "Mobile broadband capacity constraints and the need for optimization," Rysavy Inc., White Paper, February 2010. [Online]. Available: http://www.rysavy.com/Articles/2010_02_Rysavy_ Mobile Broadband Capacity Constraints.pdf
- [4] R. Pabst, B. Walke, D. C. Schultz, H. Yanikomeroglu *et al.*, "Relaybased deployment concepts for wireless and mobile broadband radio," *IEEE Communications Magazine*, pp. 80–89, Sep 2004.



Fig. 11. Traffic variation in 0..100% during one day 0..24h and single sine approximation.



Fig. 12. Short-term controlled traffic by changing the user tariff.

- [5] M. Salem, H. Yanikomeroglu, D. Falconer *et al.*, "An overview of radio resource management in relay-enhanced OFDMA-based networks," *IEEE Communications Surveys and Tutorials*, vol. 12, no. 3, 2010.
- [6] M. Dohler, R. Heath, A. Lozano, C. Papadias, and R. Valenzuela, "Is the PHY layer dead?" *IEEE Communications Magazine*, vol. 49, no. 4, pp. 159–165, april 2011.
- [7] R. Schoenen, "Credit-based flow control for multihop wireless networks and stochastic Petri nets analysis," in *Proceedings of the CNSR*, Ottawa, May 2011.
- [8] R. Schoenen and A. Otyakmaz, "QoS and Flow Management for Future Multi-Hop Mobile Radio Networks," in *Proceedings of the IEEE VTC Fall*, Ottawa, Canada, Sep 2010.
- [9] R. Schoenen, H. Yanikomeroglu, and B. Walke, "User-in-the-loop: Mobility aware users substantially boost spectral efficiency of cellular OFDMA systems," *IEEE Communications Letters*, vol. 15, no. 5, pp. 488–490, may 2011.
- [10] R. Schoenen, "On increasing the spectral efficiency more than 100% by user-in-the-control-loop," in *Proceedings of the 16th Asia-Pacific Conference on Communications (APCC)*, Auckland, October 2010.
- [11] M. Honig and K. Steiglitz, "Usage-based pricing of packet data generated by a heterogeneous user population," *Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, p. 867, 1995.
- [12] J. MacKie-Mason and H. Varian, "Pricing congestible network resources," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1141–1149, Sep. 1995.
- [13] C. Courcoubetis, F. Kelly, V. Siris, and R. Weber, "A study of simple usagebased charging schemes for broadband networks," *Telecommunication Systems*, vol. 15, pp. 323–343, 2000.
- [14] J. Altmann and K. Chu, "How to charge for network services flat-rate or usage-based?" *Computer Networks*, vol. 36, no. 5-6, pp. 519 – 531, 2001.
- [15] L. A. DaSilva, "Pricing for QoS-enabled networks: A survey," *IEEE Communications Surveys Tutorials*, vol. 3, no. 2, pp. 2 –8, 2000.