Erlang Analysis of Cellular Networks using Stochastic Petri Nets and User-in-the-Loop Extension for Demand Control

Rainer Schoenen, Halim Yanikomeroglu
Department of Systems and Computer Engineering, Carleton University, Canada

***Abstract*—Cellular networks face severe challenges due to the expected growth of application data rate demand with an increase rate of 100% per year. Over-provisioning capacity has been the standard approach to reduce the risk of overload situations. Traditionally in telephony networks, call blocking and overload probability have been analyzed using the Erlang-B and Erlang-C formulas, which model limited capacity communication systems without or with session request buffers, respectively.**
While a closed-form expression exists for the blocking probability for constant load and service, a steady-state Markov chain (MC) analysis can always provide more detailed data, as long as the Markov property of the arrival and service processes hold. However, there is a significant modeling advantage by using the stochastic Petri net (SPN) paradigm to model the details of such a system. In addition, software tool support allows getting numeric analysis results quickly by solving the state probabilities in the background and without the need to run any simulation. Because of this efficiency, the equivalent SPN model of the Engset, Erlang-B and Erlang-C situation is introduced as novelty in this paper. Going beyond the original Erlang scenario, the user-in-the-loop (UIL) approach of demand shaping by closed-loop control is studied as an extension. In UIL, demand control is implemented by a dynamic usage-based tariff which motivates users to reduce or postpone the use of applications on their smart phone in times of light to severe congestion. In this paper, the effect of load on the price and demand reduction is modeled with an SPN based on the classical Erlang Markov chain structure. Numeric results are easily obtained and presented in this paper, including probability density functions (PDF) of the load situation, and a parameter analysis showing the effectiveness of UIL to reduce the overload probability.

Keywords*—*User-in-the-loop (UIL); demand shaping; demand control; congestion; Erlang; stochastic Petri-net (SPN).

I. Introduction

In cellular networks the trend towards increasing data rates continues, with predictions of up to 100% increase per year. Figure 1 shows what this would mean for a system where the capacity cannot be raised by the same factor. At some point in time the capacity is exceeded by the demand (not the traffic, which will be choked by packet losses). This will happen at a different times for different cellular locations and is subject to daily fluctuations as well. As we are particularly concerned about the busy hours, i.e., the times where congestion is likely to happen, every broadband wireless access point will face this problem at some time. There is a well-known theory for the blocking probability of such scenarios [1]. The novel UIL paradigm goes beyond that and allows soft-CAC compared to hard-CAC (CAC is call admission control).

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

Recently, the suitability of the Erlang approach for Internet traffic has been validated [2] and there are still active publications in this area [3]. In the wireless context the situation is similar, given the limited capacity, which would only allow a few simultaneous high-definition video transmissions at a time in the same cell. The scenario assumes stationary users and quasi stationary capacity. The classical Erlang-B and C formulas provide a closed-form result for the blocking probability P_b and the waiting probability P_w, respectively, with numeric complexity O(C), but it does not provide advanced statistics such as probability mass functions (PMF) of the channel usage and does not allow any modification in the Markov chain.

Stochastic Petri nets (SPN) are known to generate Markov chains (MC) [4]. They have rarely been applied to communications problems yet, but few examples include communication networks [5], protocols [6], WiMAX [7], wireless scheduling [8], ad-hoc networks [9], radio channels [10] and flow control [11]. For a quick introduction on SPN refer to [12].

In this paper we present SPN models which are able to reproduce the Erlang results precisely, because they correspond to the same Markov chain. In addition, the SPN allows calculation of PMF, so that we are able further analyze the channel utilization and waiting statistics by specifying reward measures of interest. Tool-support is available [13] and makes the generation of result graphs a job of 10 seconds. Having the SPN models is a significant achievement, because it allows SPN methodology to be used to extend the theory beyond the classical Erlang results. In this paper the SPN is extended to include the user-in-the-loop (UIL) [14], [15] paradigm, where demand shaping changes the arrival rate of new sessions depending on the congestion status, i.e., the number of already active sessions by an anticipated dynamic price increase and demand decrease. Analysis results in this paper are obtained by direct numeric MC solution from the SPN, without any need for simulation. Results show how the blocking and waiting probability can be reduced significantly by the UIL method, by reducing excess traffic demand, thus limiting congestion in the network.

The paper organization is as follows. The classical Erlang reasoning and the SPN models are introduced in section II. Then the UIL concept and its SPN model are introduced. Next analysis results are provided before the paper ends with a conclusion.

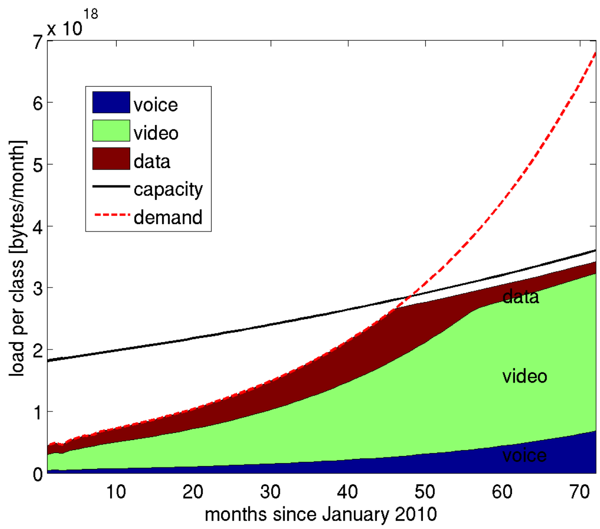

Fig. 1. This figure shows the exponential growth of demand for mobile traffic $r^{(u)}(t)$ (u=unconstrained), the limited capacity $\hat{R}(t) = r^{(t)}(t)$ and the effect of temporal UIL control to different traffic classes [16].

TABLE I. PARAMETERS FOR THE ERLANG SCENARIOS

Parameter	var	assumption,value
Connection/call arrival rate	λ	variable
Average connection holding time, e.g., video session duration	h	240 s
Offered traffic in Erlangs	$u = \lambda \cdot h$	50...100
Number of trunks/lines/circuits/resources, capacity	C	100
Max. number of sources	S	200
Blocking probability	P_b	to be calculated
Probability for waiting	P_W	to be calculated

II. ERLANG MODEL

This section provides SPN models which represent the known scenarios for Erlang-B, Erlang-C and Engset. As the reader will see, the SPN steady-state solution contains all the reward measures of interest, whereas the analytic solutions only provide mean values. Some require numeric iterations and summations, and are known as numerically instable.

The Erlang-B scenario models session arrivals as memoryless with exponential interarrival time $1/\lambda$, and a capacity of C times the requirements per session (C servers) [1]. Thus the $M/M/C$ queueing model and its Markov chain (MC) are applicable. Figure 2 shows the MC assuming $\lambda_{i,i+1} = \lambda$ and no states beyond C. Equation 1 [1] provides the blocking probability P_b:

$$P_b = B(u, C) = \frac{\frac{u^C}{C!}}{\sum_{i=0}^{C} \frac{u^i}{i!}} \tag{1}$$

$$u = \lambda \cdot h \qquad (in\ Erlangs) \tag{2}$$

The SPN model of the Erlang-B scenario is shown in Figure 3. $T1$ and $T0$ are the generating and serving transitions. $P0$ is the supply ($n = S$ can generally be assumed very large compared to all other numbers, $S \gg C$). It is not relevant for the queueing model, but only for bounding the MC state space of the SPN. $P1$ represents the state of active sessions (e.g., currently carried video streams). It is limited by capacity $m = C$, thus the disabling arc to $T2$, which is an immediate transition without timing. If $T2$ cannot fire due to full capacity, $T3$ takes over, which has lower priority than $T2$. Thus for

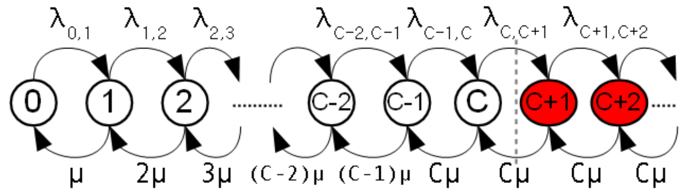

Fig. 2. Markov chain for the Erlang scenarios. All have in common that the service per established session (number given by the state index i) is constant, thus the departure rate is given by the aggregate of i servers, i.e., $i \cdot \mu$. The aggregate arrival rate $\lambda_{i,i+1}$ is usually constant λ for Erlang-B and Erlang-C and only differs for Engset. The states $C+1$ and following only exist for Erlang-C and the UIL extension discussed in this paper.

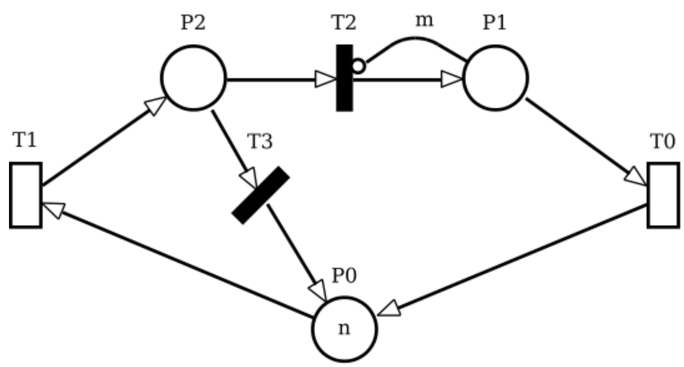

Fig. 3. SPN model for Erlang-B scenario, as explained in the text below.

Erlang-B, $P2$ would never contain tokens in a tangible state (e.g., does not constitute a state of the MC). The state index i is determined by $i = \#P1 + \#P0$, where $\#Px$ denotes the number of tokens in a place. The timing is defined as follows: $\lambda(T1) = u/h$ or $\tau(T1) = h/u$ specify the aggregate generation rate of sessions. The session have an average duration of h seconds, thus each active session is served by $T0$ with $\tau = h$, therefore $\tau(T0) = h/\#P1 = h/i$ or $r(T0) = i \cdot h$ is the state-dependent server timing. In Figure 2, μ equals h.

The Engset scenario is like the Erlang-B scenario, but with limited customers ($n = S$). Its MC is Figure 2 assuming $\lambda_{i,i+1} = (S-i) \cdot \lambda$ and no states beyond C. Its blocking probability exists in closed form, but requires recursion and many iterations to converge [1]:

$$P_b = \frac{\left[\frac{(S-1)!}{C! \cdot (S-1-C)!}\right] \cdot M^C}{\sum_{X=1}^{C} \left[\frac{(S-1)!}{X! \cdot (S-1-X)!}\right] \cdot M^X} \tag{3}$$

$$M = \frac{u}{S - u \cdot (1 - P_b)} \tag{4}$$

The Engset case is included in the SPN of Figure 3 by reducing the initial tokens in $\#P0(t = 0)$ to $n = S$ (supply). In all other cases the supply is chosen well beyond the numbers of relevance.

The Erlang-C scenario differs from Erlang-B by the existence of a waiting buffer for sessions which cannot be carried at the moment, but are taken into account as soon as capacity becomes available again. This is typical for an application scenario where the user clicks on a video link, and it takes a few up to many seconds until the video really starts. Its MC is Figure 2 assuming constant $\lambda_{i,i+1} = \lambda$ and infinitely many states beyond C. The service departure beyond state C is constant $C \cdot \mu$, as this is the maximum capacity to serve the active sessions. The Erlang-C waiting probability is known as

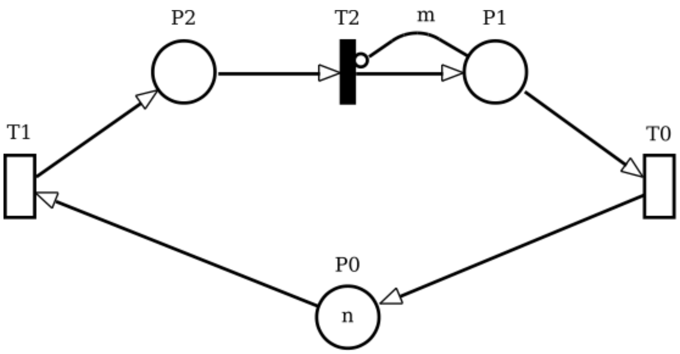

Fig. 4. SPN model for Erlang-C scenario. Compared to Fig. 3, there is no $T3$ and thus $P2$ serves as the waiting buffer for sessions.

in Equation 5:

$$P_w = \frac{\frac{u^C}{C!}\frac{C}{C-u}}{\sum_{i=0}^{C-1}\frac{A^i}{i!} + \frac{u^C}{C!}\frac{C}{C-u}} \quad (5)$$

$$u = \lambda \cdot h \quad (6)$$

Figure 4 shows the SPN for the Erlang-C scenario. It is similar to Erlang-B in Fig. 3, but this contains a buffer $P2$ which is not flushed at overload. Instead, if the capacity is exceeded, session requests wait in $P2$ to be served soon. This makes up for a bit more load than Erlang-B, thus P_w is generally higher than P_b.

Using SPN tools [13], the MC could be generated and solved very quickly (less than $10s$). Figures 5,6,7 show numeric results of the MC analysis of the Erlang-C scenario. Results for Erlang-B are omitted due to space limitations. As can be seen, the PMF of tokens in place $P1$ reveals the stochastic load distribution around the average load of u Erlangs, which has been studied for $u \in [50..100]$. For a load of $u = 75$, a significant probability of overload (congestion) is visible, just below 1%. Once we are in congestion, unserved sessions wait in $P2$, with a probability of exceeding any x given in Fig. 7. For $u = 100$ and beyond the system is not stable. In the next section a solution for leviating this congestion is introduced.

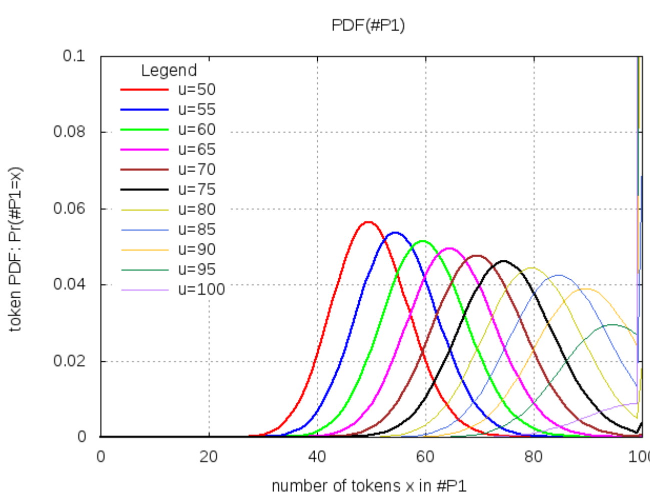

Fig. 5. Probability mass function (PMF, linear scale) of tokens in place $P1$ which represents the number of active sessions. The same PMF is shown in Fig. 6 in logarithmic scale.

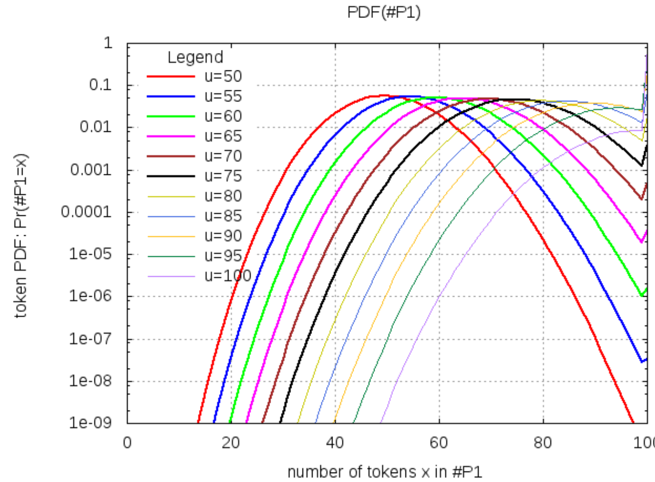

Fig. 6. Probability mass function of tokens in place $P1$ in logarithmic scale. $P1$ models the number of sessions currently active and carried by the system. The accuracy 10^{-9} cannot be achieved so clearly by means of simulation. Artefacts at the rightmost position ($\#P1 = 100$) are correct and contain the sum of all probabilities which would lead to traffic beyond 100% and are waiting in $P2$ instead.

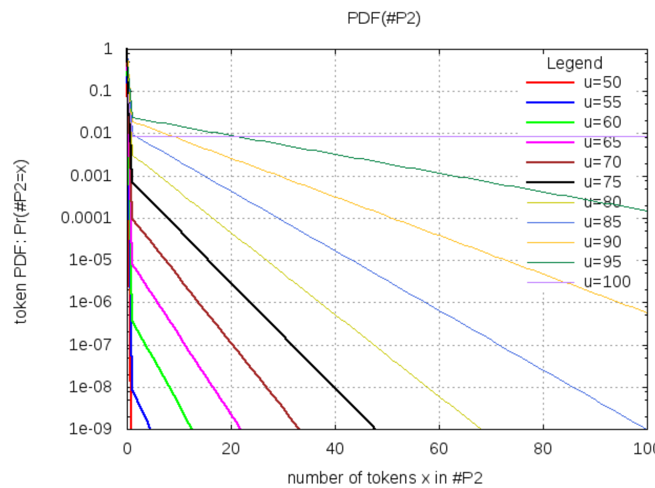

Fig. 7. Probability mass function of tokens in place $P2$ which represent the number of sessions waiting for available capacity. Obviously $u = 100\%$ is absolute overload.

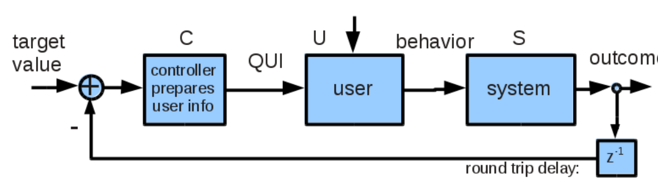

Fig. 8. User-in-the-loop (UIL): control of user and system [17]. Quantified User Information (QUI) in this paper is the indication of a dynamic price.

III. User in the Loop

The UIL paradigm is a shift from assuming user traffic as constant, given from outside of the system, towards assuming now this traffic (more precisely, the demand) can be influenced or shaped by the system itself. For wireless cellular communications there are two flavors of UIL, the spatial [17] by suggesting relocation to a point of better spectral efficiency and the temporal [14] by convincing users to

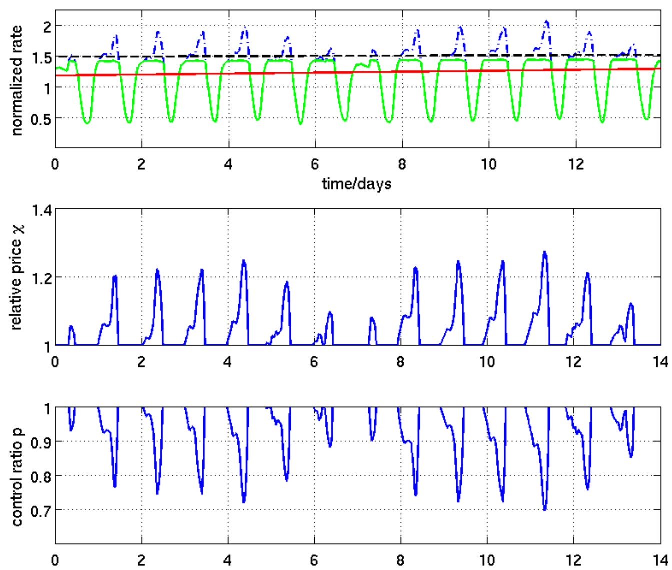

Fig. 9. UIL temporal control [16] in times of predicted congestion during the busy hours, assuming video traffic only. These 14 days represent typical traffic weeks days from a Sunday to a Saturday. The blue dash-dot line is the unconstrained traffic demand, with average shown by the red line. The black dashed line is the capacity of the system. The green line is the rate after using UIL temporal control. The controller calculates the normalized price increase χ. The users answer with a demand reduction given by the control ratio p.

postpone or rethink their demand. In both cases weak or strong incentives can be used, e.g., by financial boni or mali. Figure 8 shows this principle by a closed loop control. Incentives are adjusted dynamically so that the real traffic load stays below the capacity. The target value is set to, e.g., $\rho^{(t)} = 90\%$ here, in order to leave a margin of 10%. Recent survey results [18], [19] provide quantitative data on how users react to different levels of incentives, under different applications (QoS classes), and various scenarios. The individual user behavior cannot be known, but only the aggregate behavior of all users in a coverage region is required. This control ratio $p \in [0; 1]$ defines how much the original demand is reduced. The incentive i to control ratio p reaction data can be modeled by a linear or exponential fit. Table II shows the assumptions for this paper and Figure 10 shows the survey results. The price π of a transaction would differ from the nominal price $\pi^{(N)}$ by the factor $(1+\chi)$, where χ is the controller output (penalty, negative incentive):

$$\pi = \pi^{(N)} \cdot (1+\chi). \tag{7}$$

For the purpose of this paper, only one QoS class is assumed, which would be video because of its presumed dominance in the future. Figure 9 shows the traffic of an example week and the outcome of the UIL control.

The controller [16] has to determine a price increase χ as the incentive to the user, but internally installs a control ratio p.

Now the SPN model of section II is extended to incorporate the UIL control aspect, as can be seen in Figure 11. Most elements are equal to Figure 4, basically the loop $P0 - T1 - P2 - T2 - P1 - T0$. The main modification is $P3$ with $T3$, which has a (hidden) enabling funtion of $\#P1 <= u_T$,

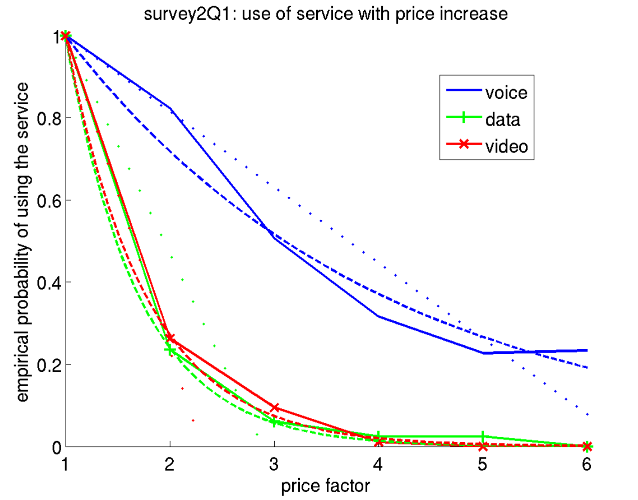

Fig. 10. User reaction [19] to a price increase of $p(\chi)$. The dotted and dashed lines show the linear and exponential fits. In this paper the focus is on video, a very elastic demand class, with $p(\chi) = e^{-1.3\cdot\chi}$

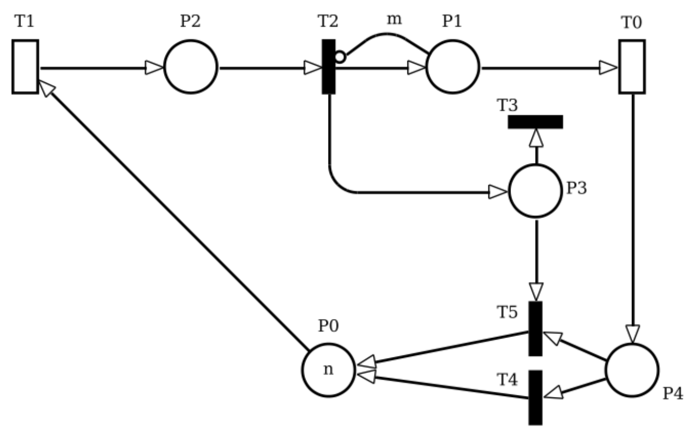

Fig. 11. SPN model for Erlang-C scenario with UIL demand control.

which flushes all tokens out of $P3$ as long as we are below the target threshold u_T. Above u_T, $P3$ holds tokens proportional to the severity of excess, $\#P3 = \Delta u = (i - u_T)$. The UIL controller calculates the dynamic price as $\chi = c_P \cdot \Delta u$, using the proportionality (P) factor c_P. There is no integral (I) or differential (D) component here. With this χ we know the user reaction according to Fig. 10. Therefore, the demand would reduce from u to

$$u \cdot e^{-\eta \cdot c_P \dot{\Delta} u} \tag{8}$$

and this is installed by adjusting the generator rate to be

$$\tau(T1) = h/u \cdot e^{\eta \cdot c_P \dot{\Delta} u} \tag{9}$$

The additional transitions $T4$ and $T5$ are there in order to empty place $P3$ in sync with $P1$, by setting priority levels $p(T3) > p(T2) > p(T5) > p(T4)$, so that

$$\#P3 = max(0; \#P1 - u_T) \tag{10}$$

Results for the UIL scenario are shown in Figure 12 and following. As written in Table II, the average demand load was set to $u = 90$ Erlangs. The parameter $cf = c_P$ is the proportional control factor. For $c_P = 0$ there is no UIL control and results reflect the Erlang-C results. Thus Figure 12 is the CCDF of Figure 5 for $c_P = 0$ and $u = 90$. With stronger

TABLE II. UIL ASSUMPTIONS FOR THE USER AND CONTROLLER BOX

Property	var	setting
Load threshold value (in SPN)	$\rho^{(t)}$ u_T	90% 90
price increase	χ	[0...2]
Exponential fit for user reaction model	$p^{(exp)}(\chi)$	function of χ $e^{-\eta\chi}$
Elasticity (log) [19]	η	1.3
Control factor (P as of PID)	c_P	[0...0.1]
Average load for Fig. 12 and following	u	90

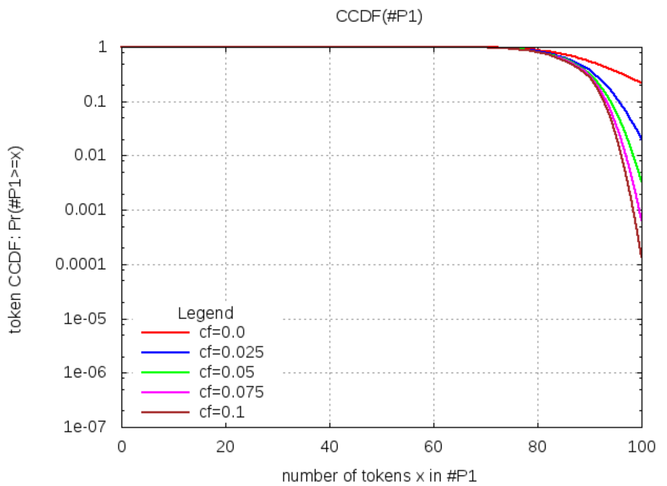

Fig. 12. Effect of UIL control on the number of sessions active, which is depicted by the $CCDF(\#P1) = Pr\{\#P1 \geq x\}$. We observe that only sessions beyond the threshold of 90% are controlled down.

contol factors up to $c_p = 0.1$, the probability of blocking or waiting is reduced from 0.2 to 10^{-4}. Figure 13 shows what happens in overload situations, as tokens in $P2$ represent (video) sessions waiting for capacity. Without UIL control, there is a significant number of sessions unserved. Even 100 unserved sessions (while 100 are served) are possible with probability in the order of 10^{-5}. The graph drops at $\#P2 = 100$ only because of the limited supply of 200 sessions, but a log-linear extrapolation is possible. With UIL control, $Pr\{\#P2 > x\}$ drops to very low numbers, as expected. Figure 14 is basically a zoom into Fig. 12 due to Eq. 10, but it can be observed how precisely the UIL control "bends" the demand above the threshold. The following figures show scalar results by varying the control factor. In Figure 15 the probability of exceeding the target threshold is studied. Naturally, as $u = 90$ and (independently) $u_T = 90$ was chosen, this probability if 50% without UIL control. Using Little's formula, the average waiting time ws determined and shown in Figure 16. There is basically no waiting for $c_p = 0.1$, as instead some users decided not to watch the video in the current overload situation. Figure 17 displays how likely the capacity is exceeded in the given scenario of average load 90%. 20% is a relatively high number of users who would be frustrated not being able to use the service. Instead, with UIL, a comparable number of users would not use the service, but for a different reason: Well informed that this is a congestion situation, and sorted by willingness to pay more, i.e., the more urgent use case is preferred compared to the less important application.

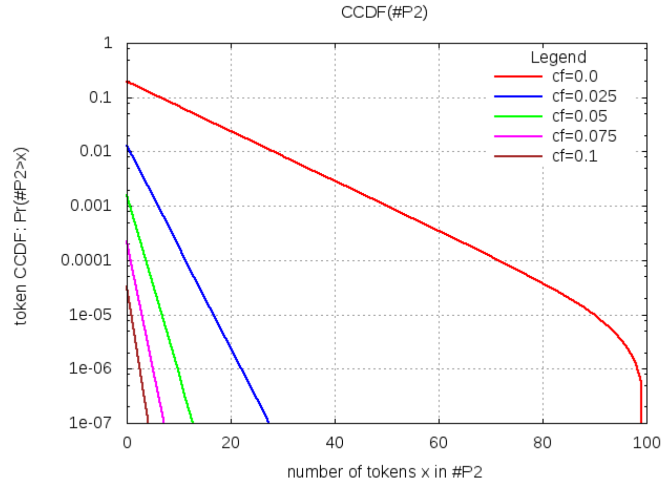

Fig. 13. Effect of UIL control on the number of sessions not served and waiting (in buffer $P2$), which is depicted by the $CCDF(\#P2) = Pr\{\#P2 > x\}$. This graph can be seen as a continuation of Figure 12 from left to right.

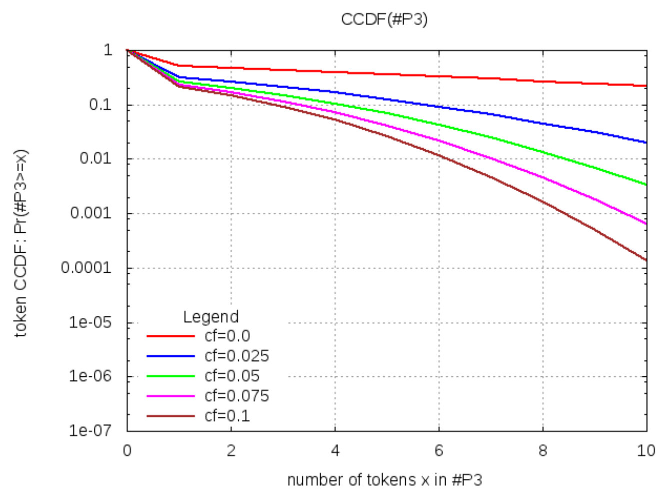

Fig. 14. $CCDF(\#P3) = Pr\{\#P3 \geq x\}$ quantifies the number of sessions affected by UIL control, namely those which exceed the target value of 90%, modeled by $P3$.

IV. CONCLUSION

In this paper stochastic Petri net models for the Erlang-B and Erlang-C traffic scenarios are presented, as well as an extension to incorporate UIL demand shaping. As can be observed, the modeling efficiency of SPN allows modifying the underlying Markov chain by simple means of (functional) parameters of the SPN model. In addition, tool-supported Markov chain analysis does not require simulations and naturally delivers accuracies in the order of 10^{-9} or better within a few seconds of run time. It is also easy to obtain higher order statistics, e.g., PMF, CDF and CCDF graphs without any extra effort, because the steady-state Markov chain contains all the information already. The case of UIL analyzed and discussed in this paper shows how demand control can be incorporated into networks and reduces the overload probabilities significantly, compared to the Erlang-C scenario. Especially in wireless networks the capacity is assumed to be in congestion more and more often in the future. As an outlook, fading channel

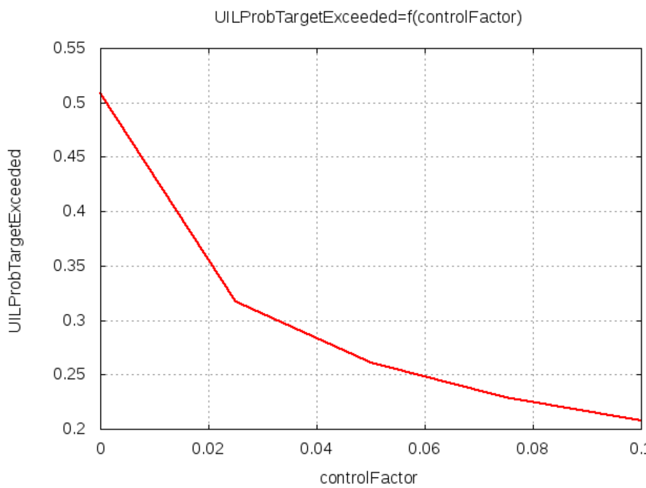

Fig. 15. The probability of traffic exceeding the threshold of 90% equals the reward measure $Pr\{\#P1 > 90\}$.

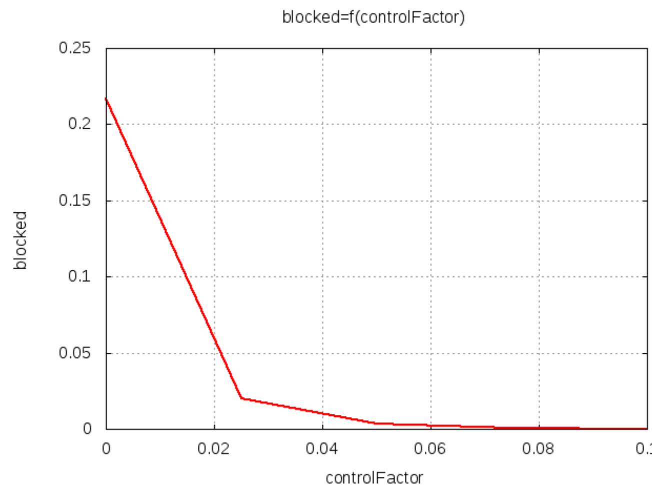

Fig. 17. The probability of blocked sessions (not enough capacity), depending on the UIL control factor (0 = no UIL control).

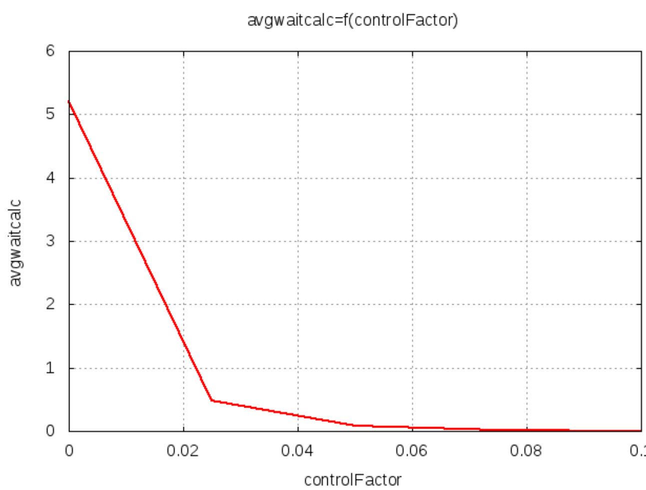

Fig. 16. Average waiting time (in seconds) for free capacity, depending on the UIL control factor (0 = no UIL control).

models and capacity fluctuations due to user mobility can be incorporated into transition $T0$. The UIL principle remains functional in this case.

References

[1] I. Angus, "Introduction to Erlang B and Erlang C," "Telemanagement #187", July 2001.

[2] T. Bonald and J. Roberts, "Internet and the Erlang formula," *ACS SIGCOMM Computer Communication Review*, vol. 42, no. 1, pp. 24–30, Jan 2012.

[3] V. Shakhov, "Simple approximation for the Erlang B formula," in *SIBIRCON-2010*, Irkutsk, Russia, Jul 2010.

[4] M. Marsan, *Modelling with generalized stochastic Petri nets*. Wiley, 1996, ISBN 0-471-93059-8.

[5] J. Billington *et al.*, *Application of Petri Nets to Communication Networks*. Springer, 1999, ISBN 3-540-65870-X.

[6] M. Bosch and G. Schmid, "Generic Petri net models of protocol mechanisms in communication systems," *Computer Communications*, vol. 14, no. 3, pp. 143 – 156, 1991.

[7] S. Geetha and R. Jayaparvathy, "Modeling and analysis of bandwidth allocation in IEEE 802.16 MAC: A stochastic reward net approach," *Int. J. Communications, Network ans System Sciences*, vol. 3, no. 7, pp. 631–637, July 2010.

[8] L. Lei, C. Lin, J. Cai, and X. Shen, "Performancs analysis of wireless opportunistic schedulers using stochastic Petri nets," *IEEE Transactions on Wireless Communications*, vol. 8, no. 4, April 2009.

[9] C. Zhang and M. Zhou, "A stochastic Petri net-approach to modeling and analysis of ad hoc network," in *Proceedings of the ITRE*, Aug 2003.

[10] H. Wang and N. Moayeri, "Finite-state Markov channel – a useful model for radio communication channels," *IEEE Transactions on Vehicular Technology*, vol. 44, no. 1, pp. 163–171, Feb 1995.

[11] R. Schoenen, G. Post, and A. Müller, "Analysis and dimensioning of credit-based flow control for the ABR service in ATM networks," in *Proceedings of the IEEE GLOBECOM*, 1998, vol.4 p.2399.

[12] R. Schoenen, "Credit-based flow control for multihop wireless networks and stochastic Petri nets analysis," in *Proceedings of the CNSR*, Ottawa, May 2011.

[13] R. German, "A toolkit for evaluating non-markovian stochastic Petri nets," *Performance Evaluation*, vol. 24, pp. 69–87, 1995.

[14] R. Schoenen, G. Bulu, A. Mirtaheri, and H. Yanikomeroglu, "Green communications by demand shaping and User-in-the-Loop tariff-based control," in *Proc. 2011 IEEE Online Green Communications Conference (IEEE GreenCom'11)*, Online, 2011.

[15] R. Schoenen and H. Yanikomeroglu, "User-in-the-Loop: Spatial and Temporal Demand Shaping for Sustainable Wireless Networks," *IEEE Communications Magazine*, accepted for publication 2013.

[16] ——, "Dynamic demand control with differentiated QoS in user-in-the-loop controlled cellular networks," in *Proceedings of the VTC'Spring 2013*, 2013.

[17] 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.

[18] R. Schoenen, G. Bulu, A. Mirtaheri, T. Beitelmal, and H. Yanikomeroglu, "First survey results of quantified user behavior in user-in-the-loop scenarios for sustainable wireless networks," in *Proceedings of the 2012 IEEE VTC Fall Conference*, Quebec City, September 2012.

[19] ——, "Quantified user behavior in user-in-the-loop spatially and demand controlled cellular systems," in *Proc. European Wireless*, Poznan, 2012.