

School of Information Technology
Information Technology papers

Bond University

Year 2007

Dimensioning and Optimization of
Push-To-Talk over Cellular Server

Muhammad T. Alam*

Zheng da Wu[†]

*Bond University, Muhammad_Alam@bond.edu.au

[†]Bond University, Zheng-Da.Wu@bond.edu.au

This paper is posted at ePublications@bond.

http://epublications.bond.edu.au/infotech_pubs/29

Dimensioning and Optimization of Push-To-Talk over Cellular Server

M. T. Alam, *student member, IEEE* and Z. D. Wu, *member, IEEE*

School of IT, Bond University

Gold Coast, Australia

Fax: +61 7 55953320

malam@bond.edu.au

Abstract- The PoC (Push-to-talk over Cellular) application allows point-to-point, or point-to-multipoint voice communication between mobile network users. The related work over PoC focuses on the performance analysis only and is completely ignorant about dimensioning PoC controller to optimize revenue for service providers. In this paper, we dimension a PoC controller with the assumption that the network grade of service is provided. The on-demand sessions should have access priority over pre-established sessions. A PoC controller should be able to terminate a PoC session based on an optimal timer. Moreover, the number of simultaneous session initiation by a PoC client is also a configurable parameter. We derived relations to provide access priority to special PoC sessions based on available Transmit/Receive Units (TRU) and threshold level. Load starting expressions are reported for a PoC controller using Lagrange multiplier technique. A simple relation to control the PoC session timer is proposed. Finally, the derivation of maximum number of allowable simultaneous session is depicted using two state Markov models. Values have been computed with the corresponding derivation to provide useful insight into the system behaviour. A PoC controller can benefit from these optimal values of our work during busy hour.

Index Terms- Timer Control, Load Balancing, Optimization, Markov Model

I. INTRODUCTION

Push-To-Talk can be viewed as an Instant Messaging service, enhanced with voice functionality. Ericsson, Motorola, Nokia and Siemens were the first vendors to team up to develop the open Push-To-Talk industry standard called PoC (Push-To-Talk over Cellular) [2]. This jointly defined specification was submitted to OMA (Open Mobile Alliance, [5]) to facilitate multi-vendor interoperability for Push-to-Talk products. The specification is based on 3GPP's (Third Generation Partnership Project) IMS (IP Multimedia Subsystem, [6]) architecture and PoC is to bring the first commercial implementations of the IMS architecture into mobile networks. A discussion on strategic actions related to standardization, system architecture and service diffusion of PoC has been discussed in [4]. An exploratory discussion of Voice over IP and CDMA usage in 2.5G/3G systems relating to Push-to-talk service has been furnished by DaSilva et al in [3].

The PoC application allows point-to-point, or point-to multipoint voice communication between mobile network users [1]. The communication is strictly unidirectional, where at any point of time only one of the participants may talk (talker), all other participants are listeners. In order to get the right to speak, listeners first have

to push a “talk” button on their mobile terminals. Floor control mechanisms ensure that the “right to speak” is arbitrated correctly between participants. The PoC application may become a highly popular service for the mobile telecommunications market if its responsiveness and voice quality meet end-user expectations.

“Push-to-Talk is a forerunner to peer-to-peer services over IP, for which IMS provides the capabilities and foundation. PoC is the first commercial application based on IMS” [7]. The driving forces behind the operators’ Push-to-talk initiatives are the search for new revenue opportunities and finding ways to increase subscriber acquisition and reduce churn. In this article, we depict some of the key areas based on the OMA release [5] to dimension a PoC network service.

Most of the related work available today focuses on the performance analysis over PoC. An architecture for enabling PoC services in 3GPP networks has been furnished by Raktale S. in [8]. Similar work is reported by Parthasarathy A. [9]. The design of a PoC service operated over a General Packet Radio service / Universal Mobile Telecommunications System (GPRS/UMTS) network is depicted by Kim et al [10]. The PoC performance is analysed over GPRS by Balazs [1]. However, these work are completely ignorant about dimensioning PoC controller to optimize revenue for service providers. The basic challenges that affect the end-to-end service performance for PoC are: a) Network configuration and dimensioning, b) Timer settings in terminals and networks, c) Traffic handling priorities used, d) Service option choices such as early media session establishment; and e) Client implementations on the terminals native operating systems. To the best of our knowledge, adding such controls to a PoC server during busy hour is a novel work. We dimension the PoC service based on the assumption that the network Grade of Service (GoS) is provided. This way a PoC server is able to control PoC functionalities to the optimal level. GoS is a measure of the blocking probability of an incoming call. Usually, a PoC Radio Access Network (RAN) infrastructure is dimensioned for 1%-2% GoS for PoC sessions. This means that the network should block less than 1%-2% of all incoming PoC sessions during busy hour. The contributions of this work are:

- i. Optimize traffic flows for the available Transmit/Receive Units (TRU) of a PoC Base Station (BS);
- ii. Controlling access of special sessions based on available TRUs;
- iii. Optimize the session timer for a PoC controller;
- iv. Optimize number of session initiation for a PoC client during busy hour.

The rest of the paper is organized as follows: Our system background and model assumptions are described in Section II and III. Section IV is dedicated to analyse the access control of on- demand session against pre-established session. We derive an optimal traffic flow expression and an optimal timer lifetime in section V and VI respectively. Section VII depicts the derivation of optimal number of simultaneous sessions for a PoC client and finally, Section VIII concludes the paper.

II. BACKGROUND

The PoC Server implements the application level network functionality for the PoC service. The PoC Server performs a Controlling PoC Function and/or Participating PoC Function. The Controlling PoC Function and Participating PoC Function are different roles of the PoC Server [5]. The determination of the PoC Server role (Controlling PoC Function and Participating PoC Function) takes place during the PoC Session setup and lasts for the duration of the whole PoC Session. The main difference

between participating PoC and controlling PoC function is that PoC server performing the PoC controlling function will have no direct communication to the PoC client but will interact with the PoC client via the server performing the participating function for the PoC client. However, local policy in the PoC server performing the participating PoC function may allow the PoC server performing the controlling PoC function to have a direct communication path for media and media related signalling to each PoC client.

Each session is controlled by one controlling function. PoC server performs the following when it fulfils the controlling PoC function: a) Provides centralized PoC Session handling, b) Provides the centralized media distribution, c) Provides the centralized Talk Burst Control functionality including Talker Identification, d) Provides Session Initiation Protocol (SIP) Session handling, such as SIP Session origination, release, etc. e) Provides policy enforcement for participation in Group Sessions, f) Provides the Participants information, g) Provides for privacy of the PoC Addresses of Participants, h) Collects and provides centralized media quality information, i) Provides centralized charging reports, j) Supports User Plane adaptation procedures and k) Support Talk Burst Control Protocol negotiation [5]. The work presented here is to dimension a PoC service based on resources available at the cell base station or Radio Access Network (RAN) infrastructure. The PoC controlling/participating function of the PoC servers will be able to perform according to the blocking requirement of the RAN.

As mentioned in the report of Northstream [7], PoC usage in GPRS has two main scenarios: 1. Short interactive sessions (Type 1) and 2. Long session with sporadic, interactive talk periods (type 2). Figure 1 illustrates these two types of PoC sessions. The distinction between the two talk is that one contains chat sessions after long intervals within a single session where as the other refers to the separate sessions for each talk. The key challenge is to reduce the delay involved in state transition from STANDBY state to READY state in the pre-established sessions. When the READY timer expires, a PoC terminal shall return to STANDBY state. The READY timer that controls the time a PoC terminal remains in READY state is set by the operator. The steps to be performed for state change are

- a) Paging with which the PoC server defines the location of the PoC terminal on cell level,
- b) Cell update with which the terminal tells the PoC server in which cell it is located
- c) Radio resource assignment procedures which are the part of session set up procedure and finally
- d) PoC signalling.

Obviously the long sessions will prefer a pre-established session than on demand session set up. We define the access control of these two kinds of session set up. Priority is provided to on demand session set up based on number of available and busy TRUs. The capacity of PoC framework is measured by TRUs to the base station which has the direct impact on cost analysis of the PoC service. A TRU can transmit on eight time-slots and receive on eight time-slots i.e., eight time-slot pairs. One time slot can share 5.48 sessions in GPRS and a TRU can support 43 simultaneous PoC sessions [7]. Usually, a cell will have 5 installed TRUs. The on-demand sessions can use any free TRU while pre-established sessions can use a TRU only when total number of busy TRUs is less than some fixed number (threshold/protection level). This way pre-established sessions will be forced to be initiated as on-demand sessions after the

protection level of TRUs is exceeded. Thus a number of message flows will be reduced for each session as there are few extra steps involved in the state change for pre-established sessions.

One of the basic challenges for a PoC service provider is the timer setting for a PoC session. The length of a PoC session timer should be carefully chosen in dynamic manner. A constant cut off time of a session will affect system performance and consequently reputation of service providers. A long timer setting will incur traffic overhead at the PoC server queue whereas a short timer setting will generate frequent requests from the PoC clients. We derive a simple relation to control the session lifetime based on GoS, time slot duration and number of TRUs installed. The PoC controller is able to terminate any session if it exceeds the timer setting during busy hour from the derivation provided in this paper.

The detail of all the PoC traffic flow scenarios can be found in the OMA release [5]. A Base Station (BS) can be thought of as a combination of TRUs and each TRU having a number of time slots. Each time slot can serve multiple number of PoC sessions. A PoC session flows can be shown as in figure 2. The initial INVITE messages of PoC clients go through the SIP/IP Cores (Session Initiation Protocol/Internet Protocol Cores). The SIP/IP Core is the reference point that supports/provides session signalling between PoC client and server, address resolution services, charging information, publication / subscription / notification of presence information, indication capabilities and relaying service settings including answering mode indication, incoming PoC session barring and incoming instant personal alert, etc. These types of huge traffic flows arise the niche of path optimization while passing through a TRU (Transmit/Receive Unit) of a base station. The lost traffic from a source PoC client to a destination PoC client must be minimized to avoid message re-generations. The solution to minimize lost traffic depends on the parameters of link offered traffic and the blocking probabilities at the TRUs. In this research work, we compute the optimized path for the available TRU of the base station to share the load and minimize the traffic overflow. The traffic flows are controlled by the controlling PoC function of the PoC server assigned to the originating PoC client.

A PoC client should not be allowed to initiate or take part in as many long sessions as it wants to initiate/join in busy hour as that may introduce congestion and performance degradation at the PoC server. The controlling PoC function must be able to limit the number of simultaneous sessions initiated by a PoC client. Simultaneous PoC session means a PoC client being the participant in more than one PoC session simultaneously. We introduce a simple two state Markov mechanism to optimize the number of simultaneous session for a PoC client during rush hour. Our derivation leads to an optimal number based on system resources. The time complexity of all the derived computation is negligible which strengthens the justification of our work. We believe a PoC service provider will be benefited from the adoption of the models presented in this paper.

III. MODEL OVERVIEW AND ASSUMPTIONS

The typical nature of a PoC session is depicted in figure 1. The PoC session consist of chats and pauses. The number of chats of long sessions is greater than that of small sessions. In fact, the statistical analysis shows that the voice activity factor has found to be 67%. That means that 33% of a conversation is actually pauses and silence [7]. The throttled arrivals with Poisson model has been extensively studied in many text books.

The inter-arrival time of session follows the negative exponential distribution (NED) and the probability density function (PDF) with arrival rate λ takes the form:

$$p(u) = \frac{\lambda}{u^2} e^{-\lambda/u}; \text{ and the corresponding cumulative distribution function is: } C(u) = e^{-\lambda/u}.$$

Appendix B provides the corresponding proof which may be found in many textbooks. The graphical representation of these functions has been studied by Liu in [14]. The interval rate takes a skew for growing value of the function. The chat arrivals of one session are also considered to be Poisson process. We control the access of these types of inter-arrival rates during rush hour which is discussed in section IV.

For a mobile network operator launching a Push-to-talk service, some investments in new RAN infrastructure are required. Our models are derived based on the resources available at the BS/RAN. The assumptions made for the presented models are

1. Capacity dimensioning of the RAN is directly linked to the expenditure of installed TRUs,
2. Each cell has installed TRUs,
3. One TRU contains 8 time slots,
4. The minimum unit of service rate is based on the number of PoC session-chats getting serviced by a time slot.

Since, we dimension the PoC service based on the given GoS and since each PoC session is handled by a controlling function of a PoC server, we assume the controlling policy will be set at the PoC server to function according to the models derived in this paper.

IV. CONTROLLING SESSION ACCESS

As mentioned earlier, Type 2 (pre established) sessions should not be allowed during the busy hour where as type 1 (on demand) sessions should be able to use any free TRU. Let, a Type 2 session can use a time slot only when the total number of busy TRU is less than some protection level of number b .

We let, λ , λ_1 , and λ_2 be total arrival rate of PoC session chats, arrival rate of type 1 session-chats and arrival rate of type 2 session-chats respectively. If we let both session chat arrivals as Poisson processes, then the system can be viewed as the birth and death a process of M/M/b TRUs. The corresponding Markov state change model with probabilities is presented in figure 3. μ represents the mean service rate of chats for both Type 1 and Type 2 sessions where number of Type 2 chats is greater than that of Type 1 chats. A state κ represents the number of chats present in the PoC BS. Under the above assumptions, we find:

$$\lambda_{\kappa} = \begin{cases} \lambda_1 + \lambda_2 & \text{if } \kappa < b \\ \lambda_1 & \text{otherwise} \end{cases} \quad (1)$$

$$\mu_{\kappa} = \kappa\mu \quad (2)$$

Where, μ_{κ} is the mean service rate in state κ

The steady state probabilities are:

$$p_{\kappa} = \begin{cases} p_0 \frac{1}{\kappa!} \left(\frac{\lambda_1 + \lambda_2}{\mu} \right)^{\kappa} & \text{if } \kappa < b \\ p_0 \frac{1}{\kappa!} \left(\frac{\lambda_1 + \lambda_2}{\mu} \right)^b \left(\frac{\lambda_1}{\mu} \right)^{\kappa-b} & \text{otherwise} \end{cases} \quad (3)$$

In this system, p_0 is the ordinary normalization condition. Given the state probabilities, it is possible to compute all the moments of the traffic, in particular of the variance.

From figure 1 we have, mean of total traffic offered, $a = \frac{\lambda}{\mu} = \frac{(\lambda_1 + \lambda_2)}{\mu}$ and the ratio of

Type 2 to total traffic, $r = \frac{\lambda_2}{a\mu}$. Let,

N = Total number of TRUs in the base station

B = The probability that all of the N TRUs are occupied

B_{b-1} = The probability that more than $b-1$ TRUs are busy

B , B_{b-1} determine how much of the two types of PoC sessions streams will be blocked.

We have,

$$B = p_0 \frac{a^N}{N!} (1-r)^{N-b} \quad (4)$$

$$1 - B_{b-1} = p_0 \sum_{j=0}^{b-1} \frac{a^j}{j!} \quad (5)$$

$$p_0 = \frac{1}{\sum_{\kappa=0}^b \frac{a^{\kappa}}{\kappa!} + \sum_{\kappa=b+1}^N \frac{a^{\kappa}}{\kappa!} (1-r)^{\kappa-b}} \quad (6)$$

This model can be used whenever a PoC controller wants to offer better service to one type of PoC sessions by restricting the availability of TRUs to the other type of sessions. The protection level can be adjusted based on given B and B_{b-1} in the above

equations. The time complexity is dominated by p_0 which is $O(a^b b)$ for $b > \frac{N}{2}$. Since, a cell will have only a few number of TRUs installed, the computation time becomes negligible to that context.

The behaviour of B and B_{b-1} are depicted in figure 4 and in figure 5 respectively. The experiment was performed for

$N=5$, $\lambda_1 = 5000/s$, $\mu = 5*40/20ms$ or $10000/s$. μ was taken in consideration that a TRU can have 8 time slot pairs and that a time slot can serve 5 simultaneous session-chats.

The arrival rate of Type 2 session-chats was varied from 5000/s to 10000/s and the blocking probabilities were computed for different protection level of TRUs. The total blocking probability goes up as the protection level goes high where as B_{b-1} goes down with raising protection level. The results in table 1, for

$N=10$, $\lambda_1 = \lambda_2 = 10000/s$, $\mu = 10*40/20ms$ or $20000/s$ exhibit similar behaviour. This

is obvious from the fact that higher values of b will allow more Type 2 (pre-established) sessions. If the value of B i.e, network GoS is provided then, p_0 can be defined and used in Eq. (5) to determine B_{b-1} . The analysis presented here can be used to fix a threshold level b , based on GoS and session-chat arrivals. Once there is a fixed protection level, any pre-established session will be blocked to initiate as on-demand session by the PoC controlling function after the total PoC session arrivals exceed the threshold number of TRUs.

V. LOAD SHARING AT POC BS

We assume that the traffic offered is given to compute the amount of overflow traffic offered to the TRUs for each PoC client. An optimization problem can be formulated based on the link offered traffic of the source PoC clients to the BS and of the BS to the destination PoC clients. This can lead to route optimization for the TRUs in a PoC BS. Our objective is to minimize the total traffic lost in the network i.e., from figure 2 we have:

$$\min_{A_k^{i,j}} z = \sum_{i,k} \hat{a}_{i,k} + \sum_{k,j} \hat{a}_{k,j} \quad (7)$$

with the constraints

$$\begin{aligned} \sum_k A_k^{i,j} &= A^{i,j} \\ A_k^{i,j} &\geq 0 \end{aligned} \quad (8)$$

Where $\hat{a}_{i,k}$ and $\hat{a}_{k,j}$ denote total blocked or overflow traffic at link (i,k) and (k,j) respectively;

And,

$A_k^{i,j}$ = The amount of traffic offered to TRU k from PoC client i to PoC client j

$a_{i,k}$ = The total traffic offered to link (i, k) .

$a_{k,j}$ = The total traffic offered to link (k, j) .

The total offered traffic to a link is computed as traffic arrival rate divided by traffic service rate. The Lagrange function to this load sharing optimal problem is

$$L(A,u,v) = \sum_{i,k} \hat{a}_{i,k} + \sum_{k,j} \hat{a}_{k,j} - \sum_{i,j} v^{i,j} \left(\sum_k A_k^{i,j} - A^{i,j} \right) - \sum_{i,j,k} u_k^{i,j} A_k^{i,j}. \quad (9)$$

Where,

u, v = Vectors of Lagrange multipliers

The first order conditions are given by $\frac{\partial L}{\partial A_k^{i,j}} = 0$, which can be expressed as,

$$u_k^{i,j} = -v^{i,j} + \sum_{l,n} \gamma_{l,n} \frac{\partial a_{l,n}}{\partial A_k^{i,j}} + \sum_{n,m} \gamma_{n,m} \frac{\partial a_{n,m}}{\partial A_k^{i,j}}, \quad (10)$$

Where,

$$\gamma_{i,k} = \frac{\partial \hat{a}_{i,k}}{\partial a_{i,k}} \quad (11)$$

Eq. (11) is called the marginal overflow of link (i,k) . In other words γ is the increment of overflow traffic corresponding to a small increase in the offered traffic. Indices, l , m and n represent the origin (PoC client l), destination (PoC client m) and TRU respectively. Eq. (10) can be reduced to

$$u_k^{i,j} = -v^{i,j} + \gamma_{i,k} + \gamma_{k,j} \quad (12)$$

Eq. (12) has the following interpretation. Consider a particular PoC session flow (i,j) . The sum $\gamma_{i,k} + \gamma_{k,j}$ is the total marginal overflow on the path through TRU k for this traffic stream. Because $v^{i,j}$ is independent of k , the optimal load sharing is as follows. If the right hand side of Eq. (12) is positive, we should not use the path through TRU k . Conversely, for all paths where there is some flow $\varepsilon_i^k > 0$, share the load to equalize the marginal overflow on all paths. This is obvious from the form of the optimality equation, which becomes $v^{i,j} = \gamma_{i,k} + \gamma_{k,j}$ i.e., the marginal blocking probabilities must be the same on all paths and must be equal to $v^{i,j}$.

However, Eq. (12) is not satisfied for overload condition since it does not take into account the path blocking and the lost traffic. In order to address this issue, we use the following case:

$$a_{i,k} = \sum_j A_k^{i,j} \quad (13)$$

$$a_{k,j} = \sum_i A_k^{i,j} [1 - B_{i,k}] \quad (14)$$

Where, $B_{i,k}$ is the blocking probability of the path (i,k) . This kind of path optimization in circuit-switched networks is studied extensively by Kelly in [12, 13]. Here, we assume that the traffic offered to the first link in a path is independent of the blocking on the second link, but that the converse is not true; i.e., the traffic offered to the second link has been thinned by an amount proportional to the blocking probability of the first link. This is in fact practical as the incoming traffic to a destination PoC client is dependent on the traffic from PoC BS. In this case the optimality equation becomes (see the appendix A for details):

$$u_k^{j,j} = -v^{j,j} + \gamma_{i,k} + (1 - B_{i,k}) \gamma_{k,j} - \frac{\partial B_{i,k}}{\partial a_{i,k}} \sum_m A_k^{i,m} \gamma_{k,m} \quad (15)$$

For the paths of positive flows $\varepsilon_i^k > 0$, we have, the complementary condition:

$$v^{i,j} = \gamma_{i,k} + (1 - B_{i,k}) \gamma_{k,j} - \frac{\partial B_{i,k}}{\partial a_{i,k}} \sum_m A_k^{i,m} \gamma_{k,m} \quad (16)$$

Where, the sum is taken over all destination clients m . In this load sharing model at PoC BS, we considered that the blocking at SIP/IP core is negligible and the coupling term is small which is practical. Thus the equal marginal overflow may lead to an optimal path via the PoC BS TRUs. A PoC controlling function will be able to use Eq. (15) and (16) to minimize the traffic overflow in busy time. The time complexity here is only $O(m)$ provided that offered traffic and blocking values are given.

VI. TIMER CONTROL

Our objective in this section is to control lifetime of the long PoC sessions for instance, session of figure 1(b) for a PoC controller. The following derivation can be used with the assumption that the service GoS, time slot duration and the service rate of time slots are provided. Note that the derivation below is based on one long PoC session only. Define,

$q(x)$ = The probability that x number of times a PoC session goes through a time slot of a TRU during time interval T .

t = Duration of a time slot.

p = The probability of all time slots being occupied at a point of time interval T , i.e., the probability that all time slots of a PoC BS are found to be occupied during a chat of a session passing through a time slot; by definition p is equal to the network GoS.

If a session is composed of only single chat, then the session can be serviced by a timeslot and does not need to be constrained. Thus, a session needs to be upper bounded if it contains more than one chat i.e., $x \geq 2$.

The relation between a session chats and the GoS is:

$$p = \sum_{x=2}^{\infty} q(x) \quad (17)$$

The assumption here is that the chat arrivals of a PoC session is Poisson process. As traffic is unequally distributed in reality at the TRUs of a BS, it is more correct to calculate the timer with regards to time slot duration [7]. Assuming a session will be active during the whole interval T we let, $q(x)$ to have the mean tT .

Thus, the Poisson distribution $q(x)$ is:

$$q(x) = \frac{(tT)^x}{x!} e^{-\left(t+\frac{1}{\mu}\right)T} \quad (18)$$

Here, μ represents the mean service rate of each TRU ($40/t$) considering that a TRU has 8 time slot pairs and that each slot can serve 5 chat sessions. Since chat arrivals of a session are continuous, $1/\mu$ is added in order to have the impact of mean service time of a chat in the session chat distribution. A session may go through any of the N TRUs in a PoC BS. Therefore,

$$q(x) \leq \frac{(tT)^x}{x!} e^{-\left(t+\frac{1}{\mu}\right)T} \left(\frac{1}{N}\right) \quad (19)$$

From Eq. (19) and (17) we get,

$$p \leq \sum_{x=2}^{\infty} \frac{(tT)^x}{x!} e^{-\left(t+\frac{1}{\mu}\right)T} \left(\frac{1}{N}\right) \quad (20)$$

Using the Taylor series

$$e^{(tT)} = 1 + tT + \frac{(tT)^2}{2!} + \dots \quad (21)$$

We find,

$$p \leq \frac{e^{tT} - tT - 1}{Ne^{\left(t+\frac{1}{\mu}\right)T}} \quad (22)$$

Solving Eq. (22) provides a bound for T . Here the computation complexity with 2nd degree approximation is dominated by $1 + \left(t + \frac{1}{\mu}\right)T + O\left\{\left(t + \frac{1}{\mu}\right)T\right\}^2$.

We use up to the 2nd degree approximation of Taylor expansion to solve Eq. (22):

$$\begin{aligned} p &\approx \frac{(tT)^2}{2N + 2N\left(t + \frac{1}{\mu}\right)T + N\left(t + \frac{1}{\mu}\right)^2 T^2} \\ \Rightarrow 2Np + 2Np\left(t + \frac{1}{\mu}\right)T + Np\left(t + \frac{1}{\mu}\right)^2 T^2 - (tT)^2 &= 0 \quad (23) \end{aligned}$$

Taking the 1st derivative with respect to T , we get a simple relation:

$$\begin{aligned} 2Np\left(t + \frac{1}{\mu}\right) + 2Np\left(t + \frac{1}{\mu}\right)^2 T - 2t^2 T &= 0 \\ \Rightarrow T &= \frac{Np\left(t + \frac{1}{\mu}\right)}{t^2 - Np\left(t + \frac{1}{\mu}\right)^2} \quad (24) \end{aligned}$$

The relationship between a session lifetime and a PoC server blocking probability is provided in figure 6 for $t = 0.02s$, $\frac{1}{\mu} = 0.0005s$. The result shows that a PoC client can have longer session with higher network GoS and higher number of installed TRUs in the network.

VII. OPTIMIZATION OF SIMULTANEOUS SESSIONS:

Our objective in this section is to control the number of simultaneous sessions for a PoC client during busy time. Since, the Northstream report suggests that cost analysis based on time slots of PoC servers produce equal outcomes as that of TRUs, we consider our next analysis based on number of time slots. The two-state Markov chain model has been extensively used for the voice traffic. Gilbert's model [15] and recent works in [16-20] have shown that a simple two-state Markov chain can measure packet loss over the internet efficiently. We use similar approach to compute the number of the optimal sessions for a PoC client. The analysis presented here is to limit number of simultaneous long sporadic/pre-established sessions (Type 2) for a PoC client during busy hour. Thus the notations λ , μ , a denote arrival rate, mean service rate and traffic intensity respectively of Type 2 sessions in this section. The other notations that will be used throughout this sections as follows:

N_T = The total number of time slots of a PoC access network,

N_c = The number of PoC clients being served by a PoC BS/the whole network.

The two state natures of figure 7 and figure 8 can capture the bursty nature of the number of simultaneous sessions in busy hour. The former represents the states of the BS where as the later represents the states of a PoC client. The model in figure 7 has two states: Blocking or busy and Not busy. H_1 and H_2 are the state transition probabilities. The PoC BS goes to Blocking state $\mathbf{0}$, when all channels/time slots are busy at a random point of time that can be computed from Erlang's loss formula. In this state, number of session arrival in the server is greater than $5N_T$, assuming that a time slot serves 5 PoC sessions at the same time on the average.

$$H_2 = \frac{\frac{a^{N_T}}{N_T!}}{\sum_{d=0}^{N_T} \frac{a^d}{d!}} \quad (25)$$

H_2 is the transition probability that causes the BS enter into Blocking state i.e., by definition H_2 is the given GoS. It goes to Not busy state $\mathbf{1}$, when there is at least one time slot available that can be computed from the Binomial distribution. Any new session will be blocked when the server is in state $\mathbf{0}$. A successful session set up only depends on the current state. Because of the throttled nature of the PoC sessions, a session changes between idle (inactive) and busy (active), the offered traffic per session is

$$\alpha = \frac{T_{busy}}{T_{idle} + T_{busy}} = \frac{1/\mu}{1/\lambda + 1/\mu} = \frac{a}{1+a} \quad (26)$$

Then, for non busy state,

$$H_1 = \sum_{d=0}^{N_T-1} \binom{N_T}{d} \alpha^d (1-\alpha)^{N_T-d} \quad (27)$$

The session Blocking is equal to the state probability $P(0)$. Similarly the probability of successful session set up is equal to the state probability $P(1)$. The transition between two states occurs at each session set up/termination. Thus in steady state:

$$P(0) + P(1) = 1 \quad (28)$$

The state transition matrix is given by

$$P_H = \begin{bmatrix} 1-H_1 & H_1 \\ H_2 & 1-H_2 \end{bmatrix} \quad (29)$$

Figure 8 represents the nature of a session initiation situation of a PoC client. State D represents a client initiating one session and state E represents multiple session initiation. I_1 and I_2 are the transition probabilities. The probability that a PoC client initiates a session is the mean arrival rate of all PoC clients i.e.,

$$I_2 = \frac{\lambda}{N_c} = \frac{\sum_{i=1}^{N_c} \lambda_i}{N_c} \quad (30)$$

The probability of simultaneous session initiation of a PoC client during a known period T can be determined by one less the probability of one session initiation of a PoC client. Since, we assume that the session initiations are Poisson streams we have,

$$\begin{aligned}
I_1 &= 1 - \Pr[\text{one session} \mid T = t_s] \\
&= 1 - I_2 t_s e^{-I_2 t_s} \\
&= 1 - \frac{\lambda}{N_c} t_s e^{-\left(\frac{\lambda}{N_c} t_s\right)}
\end{aligned} \tag{31}$$

Where, t_s is the session lifetime of a PoC client. The probability of a successful session set up of a particular client is equal to the state probability $P(D)$. Similarly, the probability that a client is successful in establishing more than or equal to two simultaneous sessions is equal to the state probability $P(E)$. In steady state,

$$P(D) + P(E) = 1 \tag{32}$$

The state transition matrix is:

$$P_I = \begin{bmatrix} 1 - I_1 & I_1 \\ I_2 & 1 - I_2 \end{bmatrix} \tag{33}$$

Since the PoC system going to busy state depends on total number of sessions, we concatenate two models as shown in figure 10.

In this model, state $(0D)$ and $(0E)$ represent session blocking whereas, state $(1D)$ and $(1E)$ represent successful session set ups. Again success or failure of session set up depends on the current state. At steady state:

$$P(0D) + P(0E) + P(1D) + P(1E) = 1 \tag{34}$$

The state transition probability matrix is given as:

$$P_{HI} = \begin{bmatrix} (1-H_1)(1-I_1) & H_1(1-I_1) & I_1(1-H_1) & H_1 I_1 \\ H_2(1-I_1) & (1-H_2)(1-I_1) & H_2 I_1 & I_1(1-H_2) \\ I_2(1-H_1) & I_2 H_1 & (1-I_2)(1-H_1) & H_1(1-I_2) \\ H_2 I_2 & I_2(1-H_2) & H_2(1-I_2) & (1-H_2)(1-I_2) \end{bmatrix} \tag{35}$$

The matrix multiplication takes $O(16)$ time only. Let, the total probability of simultaneous session being successful be π .

$$\begin{aligned}
\pi &= P(1E) \\
&= 1 - [P(0D) + P(1D) + P(0E)]
\end{aligned} \tag{36}$$

Where $P(0D)$ is the probability that a single session set up is blocked; $P(1D)$ is the probability that a single session can be established; and $P(0E)$ is the probability that simultaneous sessions is blocked.

Therefore, the mean random variable, \bar{n} , i.e., the number of simultaneous session for a PoC client can be obtained by:

$$\bar{n} = \left\lceil \pi \left(\frac{\lambda}{N_c} t_s \right) \right\rceil \tag{37}$$

Usually, a PoC client should not be allowed to initiate more than 3 simultaneous sessions during rush hour [7]. A list of values for \bar{n} has been furnished in the table 2, table 3 and table 4 for variable parameters:

VIII. CONCLUSIONS

In this paper, we derived and analysed several optimal characteristics to dimension a PoC service. The performance for PoC is highly dependent on tuning the service from an end-to-end perspective. Deployment requires optimized expertise in the entire service delivery chain from a cost point of view. A service provider can benefit from the analyses performed in this paper. We have shown the effects of providing controlled access to two different types of sessions, optimized load sharing expressions for a PoC controller as a decision criterion, a simple relation to control the session timer and finally an expression to compute the maximum number of allowable simultaneous session for each PoC client during busy hour. We are currently investigating the case of prioritizing and classifying PoC traffic in terms of session dropping probabilities.

APPENDIX A

The load sharing optimal Eq. (9) becomes (from [13]) :

$$\begin{aligned} U(A,u,v) = & \sum_{l,n} a_{l,n} B_{l,n} + \sum_{n,m} a_{n,m} B_{n,m} \\ & + \sum_{l,m} \lambda^{l,m} (\sum_n A_n^{l,m} - A^m) - \sum_{l,m,n} \mu_n^{l,m} A_n^m \end{aligned} \quad (\text{A.1})$$

Taking the derivative with respect to $A_k^{i,j}$, we get

$$\begin{aligned} u_k^{i,j} = & v^{i,j} + \sum_{l,n} \frac{\partial a_{l,n}}{\partial A_k^{i,j}} \left[B_{l,n} + a_{l,n} \frac{\partial B_{l,n}}{\partial a_{l,n}} \right] \\ & + \sum_{n,m} \frac{\partial a_{n,m}}{\partial A_k^{i,j}} \left[B_{n,m} + a_{n,m} \frac{\partial B_{n,m}}{\partial a_{n,m}} \right] \end{aligned} \quad (\text{A.2})$$

Using the fact that

$$\frac{\partial a_{l,m}}{\partial A_k^{i,j}} = \sum_m \frac{\partial A^{l,m}}{\partial A_k^{i,j}} = \delta_{i,l} \delta_{k,n} \quad (\text{A.3})$$

and

$$\begin{aligned} \frac{\partial a_{n,m}}{\partial A_k^{i,j}} = & \sum_l \frac{\partial A^{l,m} (1 - B_{l,n})}{\partial A_k^{i,j}} \\ = & \sum_l \left[(1 - B_{l,n}) \delta_{j,m} \delta_{n,k} \delta_{i,l} - A_n^{l,m} \frac{\partial B_{l,n}}{\partial a_{l,n}} \frac{\partial a_{l,n}}{\partial A_k^{i,j}} \right] \\ = & (1 - B_{i,n}) \delta_{j,m} \delta_{k,n} - A_n^{i,m} \frac{\partial B_{i,n}}{\partial a_{i,n}} \delta_{k,n}. \end{aligned} \quad (\text{A.4})$$

Where, $\delta_{i,j}$ = The Kronecker symbol = 1 if $i=j$, and 0 otherwise

Replacing these derivatives in Eq. (A.2) and doing the appropriate sums, we get the optimal expression in Eq. (15).

APPENDIX B

By the definition, in the Poisson model, the inter-arrival time of data units follows the negative exponential distribution (NED), i.e., the probability density function (PDF) is

$$f(t) = \lambda e^{-\lambda t} \quad (t > 0) \quad (\text{B.1})$$

And the cumulative distribution function (CDF) is

$$F(t) = 1 - e^{-\lambda t} \quad (t > 0) \quad (\text{B.2})$$

From Eq.(B.1) and Eq.(B.2), it can be derived that the inter-arrival time of session follows the negative exponential distribution (NED), then the probability density function (PDF) with arrival rate λ takes the form:

$p(w) = \frac{\lambda}{w^2} e^{-\lambda/w}$; and the corresponding cumulative distribution function is:

$$C(w) = e^{-\lambda/w}.$$

The mathematical detail of the derivation is furnished below.

Let X be a continuous random variable with the probability density function $f(x)$ such that $f(x) \neq 0$

In interval $[a,b]$. Let $y = g(x)$ be a real function differentiable everywhere. If $g'(x)$ does not change its sign in $[a,b]$, then $Y = g(X)$ is a continuous random variable with the probability density function

$$\psi(y) = \begin{cases} f(h(y)) |h'(y)| & \text{if } \alpha < y < \beta \\ 0 & \text{otherwise} \end{cases} \quad (\text{B.3})$$

Where $h(y)$ is the inverse function of $g(x)$, and

$$\begin{aligned} \alpha &= \min\{g(a), g(b)\} \\ \beta &= \max\{g(a), g(b)\} \end{aligned} \quad (\text{B.4})$$

If $y = g(x) = 1/x$, then the process connects both the inter-arrival time and inter-arrival rate. As a result, we have

$$\begin{aligned} x = h(y) &= 1/y \\ \text{and } h'(y) &= -1/y^2 \end{aligned} \quad (\text{B.5})$$

Thus for probability density function of Eq.(B.1), i.e., for the inter-arrival time following the negative exponential distribution, the PDF of the inter-arrival rate would be

$$\psi(y) = \frac{1}{y^2} f\left(\frac{1}{y}\right) = \frac{\lambda}{y^2} e^{-\lambda/y} \quad (y > 0) \quad (\text{B.6})$$

And the corresponding cumulative distribution function (CDF) is

$$C(y) = e^{-\lambda/y} \quad (y > 0) \quad (\text{B.7})$$

REFERENCES

- [1] Balazs A. QoS in wireless networks: Push-to-talk performance over GPRS. *Proceedings of the 7th ACM international symposium on Modeling, analysis and simulation of wireless and mobile systems*, 2004; 182-187.
- [2] URL: <http://www.ericsson.com/products/>
- [3] DaSilva L. A., Morgan G. E., Bostian C. W., Sweeney D. G., Midkiff S. F., Reed J. H. and Thompson C. The Resurgence of Push-to-Talk Technologies. *IEEE Communication Magazine*, 2006; 48-55.
- [4] Ali-Vehmas T., Luukkainen S. Service Diffusion Strategies for Push to Talk Over Cellular. *IEEE International Conference on Mobile Business, ICMB*, 2005; 427-433.
- [5] OMA, Open Mobile Alliance. Push to talk over Cellular (PoC)-Architecture. Push to Talk Over Cellular Working Group. 2005; URL: <http://www.openmobilealliance.org>
- [6] 3GPP. The 3rd generation Partnership Project, IP Multimedia Subsystem, Stage 2. 2005; URL: <http://www.3gpp.org>
- [7] Northstream AB. Overview and comparison of Push-to-talk. 2004; URL: <http://www.northstream.se>
- [8] Raktale S. 3PoC: An Architecture for enabling Push To Talk services in 3GPP Networks. *IEEE International Conference on Personal wireless communications (ICPWC)*, 2005; 202-206.
- [9] Parthasarathy A. Push to Talk over Cellular (PoC) Server. *IEEE International Conference on Networking, Sensing and Control*, 2005; 772-776.
- [10] Kim P., Balazs A., Broek E., Kieselmann G., Bohm W. IMS-based Push-to-Talk over GPRS/UMTS. *IEEE Wireless Communications and Networking Conference*, 2005; **4**: 2472-2477.
- [11] Kelly F. P. Blocking probabilities in large circuit switched networks. *Advances in Applied Probability*, 1986; **18**: 473-505.
- [12] Kelly F. P. Adaptive routing in circuit-switched networks, *Statistical Laboratory*, Cambridge University, Cambridge, England. 1986
- [13] Kelly F. P. Routing in circuit-switched networks: optimization, shadow prices and decentralization. *Advances in Applied Probability*, 1988; **20**: 112-144.
- [14] Liu X. Network optimization with stochastic traffic flows. *International Journal of Network Management*, 2002; **12**: 225-234.
- [15] Gilbert E. N. Capacity of a burst-noise channel, *Bell Syst. Tech. J.*, 1960; **39**: 1253-1265.
- [16] Swarts F. and Ferreira H. C. Markov characterization of digital fading mobile VHF channels. *IEEE Trans. Veh. Technol.*, 1994; **43**: 977-985.
- [17] Wang H. S. and Moayeri N. Finite-state Markov channel: A useful model for radio communications channels. *IEEE Trans. Veh. Technol.*, 1995; **44**: 163-171.
- [18] Altman E., Avrachenkov K. and Barakat C. TCP in presence of bursty losses. *in Proc. ACM SIGMETRICS*, 2000: 124-133.
- [19] Bolot J., Parisi S. and Towsley D. Adaptive FEC-based error control for Internet telephony. *in Proc. IEEE INFOCOM*, 1999; **3**: 1453-1460.

[20] Akan B., Akyildiz I. F. ARC: The Analytical Rate Control Scheme for Real-Time Traffic in Wireless Networks. *IEEE/ACM Transactions on Networking*, 2004; 12(4): 634-544.

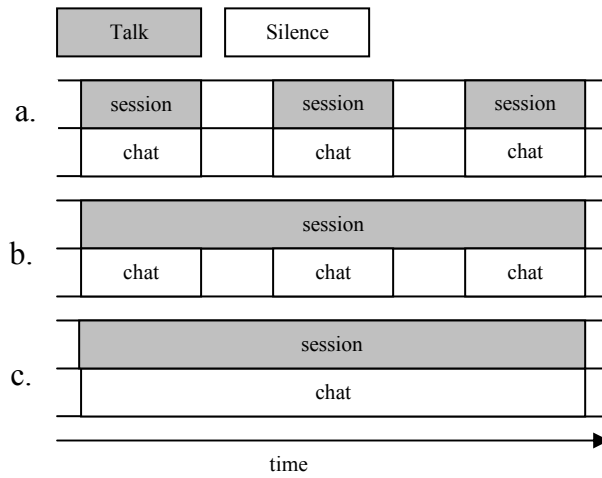


Figure 1. a) PoC short session b) PoC long session and c) Normal phone call [7]

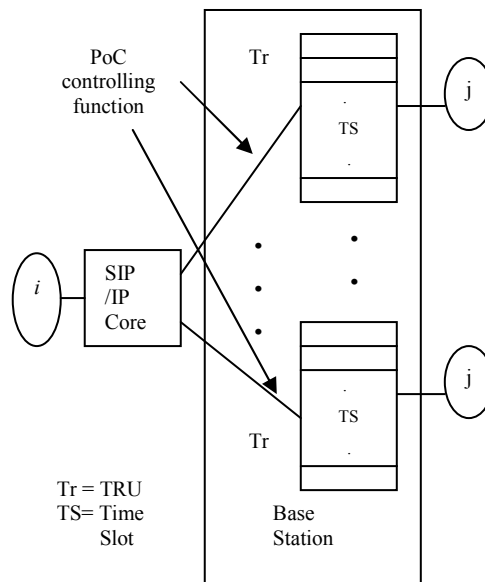


Figure 2-PoC route optimization between two PoC clients: i and j

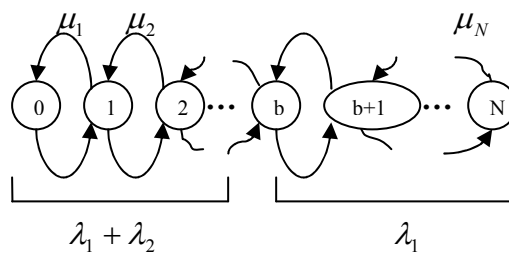


Figure 3-Markov model for accessing session

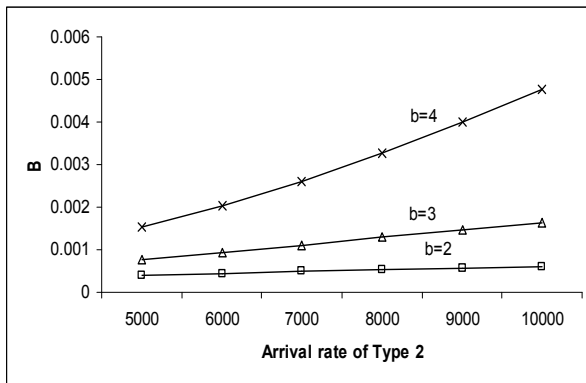


Figure 4-Total blocking probability for different protection level

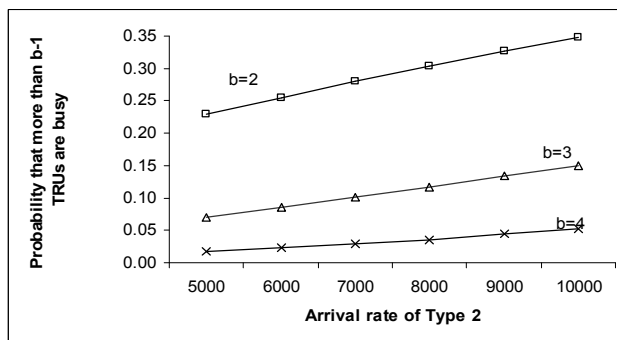


Figure 5-Blocking probability for protection level

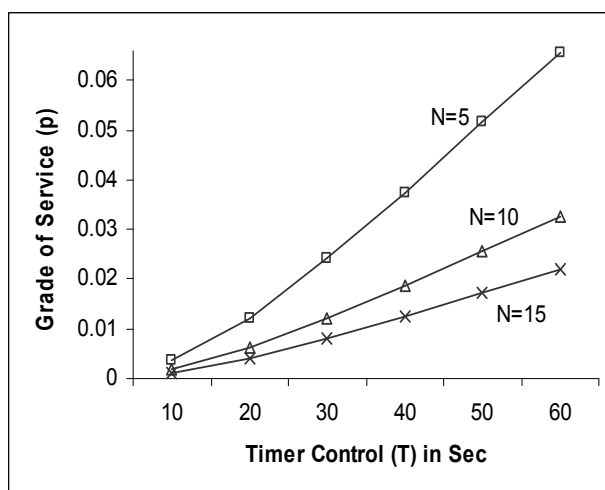


Figure 6-Effect of T for multiple installed TRUs

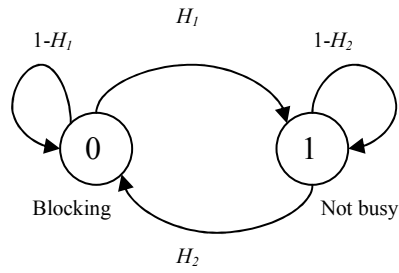


Figure 7-Markov model for the PoC BS states

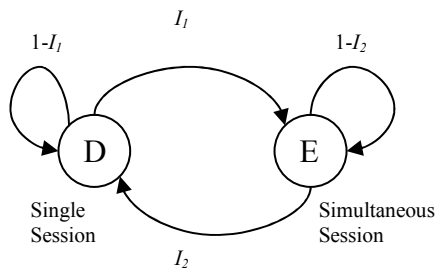


Figure 8-Session states of a PoC Client

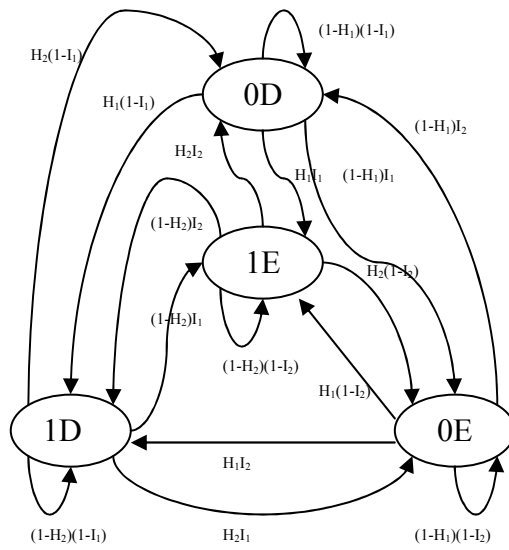


Figure 9-Four state Markov chain for session set up

Table 1-Blocking probabilities for N=10

b	B	B_{b-1}
2	0.0000000004	0.2292529587
3	0.0000000008	0.0705525580
4	0.0000000016	0.0170324915
5	0.0000000032	0.0033416557

6	0.0000000063	0.0005498781
7	0.0000000127	0.0000778205
8	0.0000000253	0.0000096562
9	0.0000000507	0.0000010645

Table 2-Number of allowable simultaneous sessions for a PoC client

$\alpha = \frac{0.99}{1+0.99}, N_T = \{40 \rightarrow 120\} N_c = 500, H_2 = 0.02, T = 40s$								
λ	H_1	I_1	I_2	$P(1E)$	$P(0E)$	$P(1D)$	$P(0D)$	\bar{n}
50	0.99999	0.92673	0.1	0.88351	0.01908	0.09548	0.00190	3.53406
75	0.99999	0.98512	0.15	0.84866	0.01918	0.12955	0.00258	5.09201
100	0.99999	0.99731	0.2	0.81374	0.01921	0.16377	0.00326	6.50994
125	0.99999	0.99954	0.25	0.78070	0.01922	0.19616	0.00390	7.80703
150	0.99999	0.99992	0.3	0.74999	0.01922	0.22628	0.00450	8.99991
175	0.99999	0.99998	0.35	0.72151	0.01922	0.25421	0.00505	10.10117
200	0.99999	0.99999	0.4	0.69505	0.01922	0.28015	0.00556	11.12094
225	0.99999	0.99999	0.45	0.67042	0.01922	0.30431	0.00603	12.06771
250	0.99999	0.99999	0.5	0.64743	0.01922	0.32685	0.00647	12.94879
275	0.99999	0.99999	0.55	0.62593	0.01922	0.34795	0.00688	13.77054
300	0.99999	0.99999	0.6	0.60577	0.01922	0.36773	0.00726	14.53853
325	0.99999	0.99999	0.65	0.58683	0.01922	0.38631	0.00762	15.25764
350	0.99999	0.99999	0.7	0.56900	0.01922	0.40379	0.00796	15.93219
375	0.99999	0.99999	0.75	0.55219	0.01922	0.42028	0.00828	16.56599

Table 3-Number of allowable simultaneous sessions for a PoC client

$\alpha = \frac{0.99}{1+0.99}, N_T = \{40 \rightarrow 120\} N_c = 500, H_2 = 0.02, \lambda = 50/s$								
$T(sec)$	H_1	I_1	I_2	$P(1E)$	$P(0E)$	$P(1D)$	$P(0D)$	\bar{n}
20	0.99999	0.72932	0.1	0.86082	0.01859	0.11821	0.00236	1.72164
30	0.99999	0.85063	0.1	0.87588	0.01892	0.10313	0.00205	2.62765
40	0.99999	0.92673	0.1	0.88351	0.01908	0.09548	0.00190	3.53406
50	0.99999	0.96631	0.1	0.88705	0.01916	0.09194	0.00183	4.43527
60	0.99999	0.98512	0.1	0.88864	0.01919	0.09035	0.00180	5.33188
70	0.99999	0.99361	0.1	0.88934	0.01921	0.08964	0.00179	6.22543

80	0.99999	0.99731	0.1	0.88964	0.01921	0.08934	0.00178	7.11719
90	0.99999	0.99888	0.1	0.88977	0.01922	0.08921	0.00178	8.00799

Table 4-Number of allowable simultaneous sessions for a PoC client

$\alpha = \frac{0.99}{1+0.99}, N_T = \{40 \rightarrow 48\}, N_c = 500, T = 40s, \lambda = 50/s$								
H_2	H_1	I_1	I_2	$P(1E)$	$P(0E)$	$P(1D)$	$P(0D)$	\bar{n}
0.01	0.99999	0.92673	0.1	0.89287	0.00973	0.09643	.00096	3.57148
0.02	0.99999	0.92673	0.1	0.88351	0.01908	0.09548	0.00190	3.53406
0.03	0.99999	0.92673	0.1	0.87452	0.02807	0.09456	0.00283	3.49810
0.04	0.99999	0.92673	0.1	0.86588	0.03672	0.09365	0.00373	3.46352
0.05	0.99999	0.92673	0.1	0.85757	0.04503	0.09276	0.00462	3.43028
0.06	0.99999	0.92673	0.1	0.84957	0.05302	0.09189	0.00550	3.39831
0.07	0.99999	0.92673	0.1	0.84189	0.06071	0.09103	0.00636	3.36756
0.08	0.99999	0.92673	0.1	0.83449	0.06810	0.09019	0.00720	3.33799
0.09	0.99999	0.92673	0.1	0.82738	0.07521	0.08936	0.00803	3.30955

Biography:



Muhammad T. Alam received the BS and the MS degrees, both from the Department of Computer Science, North South University, Dhaka, Bangladesh and Oklahoma State University, Oklahoma, USA in 1999 and 2002, respectively. He is currently a PhD student and a casual teaching fellow at Bond University, Australia. His research interests include quality of service support in mobile networks, mobility management, network optimization, TCP performance in mobile networks and MAC-layer resource allocation and scheduling.



Zheng Da Wu received the Master degree of computer technology in the Graduate School of University of Sciences and Technology of China, Beijing, from the Chinese Academy of Sciences in 1981, and the PhD degree in computer science from University of Kent at Canterbury, U.K., in 1987. He is currently an associate professor of Computer Science in the School of Information Technology, Bond University, Gold Coast, Australia. His current research is in the area of mobile networks and computing, multimedia communications and massive multiplayer networked games. His website is <http://www.bond.edu.au/it/staff/zhengda.htm>.