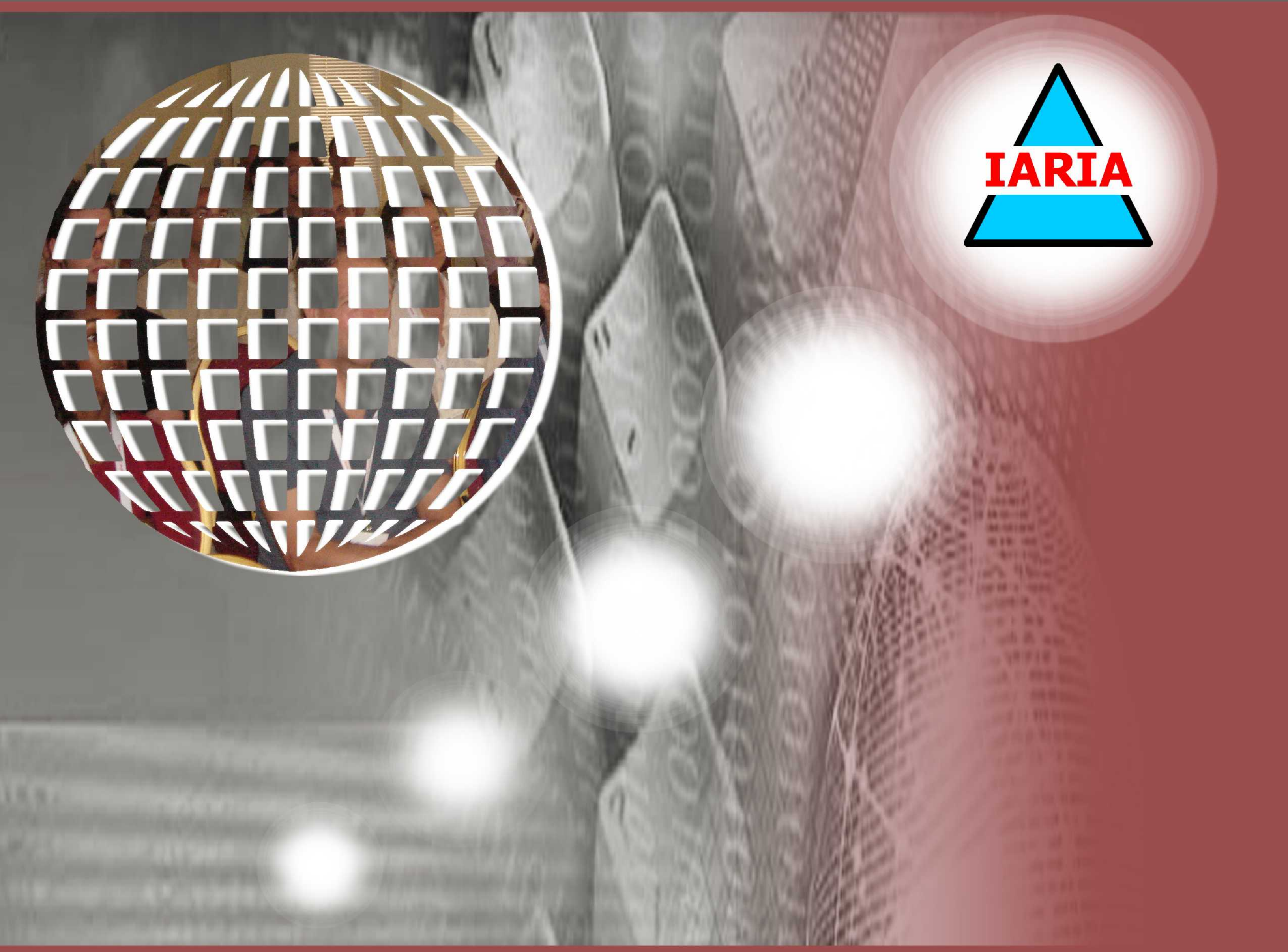
International Journal on

Advances in Telecommunications



2013 vol. 6 nr. 1&2

The *International Journal on Advances in Telecommunications* is published by IARIA.
ISSN: 1942-2601
journals site: http://www.iariajournals.org
contact: petre@iaria.org

International Journal on Advances in Telecommunications
Volume 6, Number 1 & 2, 2013

Ilija Basicevic, University of Novi Sad, Serbia
Carlos Becker Westphall, Federal University of Santa Catarina, Brazil
Mark Bentum, University of Twente, The Netherlands
David Bernstein, Huawei Technologies, Ltd., USA
Eugen Borgoci, University "Politehnica"of Bucharest (UPB), Romania
Fernando Boronat Seguí, Universidad Politecnica de Valencia, Spain
Christos Bouras, University of Patras, Greece
David Boyle, Tyndall National Institute, University College Cork, Ireland
Martin Brandl, Danube University Krems, Austria
Julien Broisin, IRIT, France
Dumitru Burdescu, University of Craiova, Romania
Andi Buzo, University "Politehnica" of Bucharest (UPB), Romania
Shkelzen Cakaj, Telecom of Kosovo / Prishtina University, Kosovo
Enzo Alberto Candreva, DEIS-University of Bologna, Italy
Rodrigo Capobianco Guido, University of Sao Paulo, Brazil
Hakima Chaouchi, Telecom SudParis, France
Emmanuel Chaput, IRIT-CNRS, France
Silviu Ciochina, Universitatea Politehnica din Bucuresti, Romania
José Coimbra, Universidade do Algarve, Portugal
Hugo Coll Ferri, Polytechnic University of Valencia, Spain
Denis Collange, Orange Labs, France
Noel Crespi, Institut TELECOM SudParis-Evry, France
Leonardo Dagui de Oliveira, Escola Politécnica da Universidade de São Paulo, Brazil
Gerard Damm, Alcatel-Lucent, USA
Danco Davcev, University for Information Science & Technology "St. Paul the Apostle", Ohrid, Macedonia
Francescantonio Della Rosa, Tampere University of Technology, Finland
Chérif Diallo, Consultant Sécurité des Systèmes d'Information, France
Klaus Drechsler, Fraunhofer Institute for Computer Graphics Research IGD, Germany
Jawad Drissi, Cameron University , USA
António Manuel Duarte Nogueira, University of Aveiro / Institute of Telecommunications, Portugal
Alban Duverdier, CNES (French Space Agency) Paris, France
Nicholas Evans, EURECOM, France
Fabrizio Falchi, ISTI - CNR, Italy
Mário F. S. Ferreira, University of Aveiro, Portugal
Bruno Filipe Marques, Polytechnic Institute of Viseu, Portugal
Robert Forster, Edgemount Solutions, USA
John-Austen Francisco, Rutgers, the State University of New Jersey, USA
Kaori Fujinami, Tokyo University of Agriculture and Technology, Japan
Shauneen Furlong , University of Ottawa, Canada / Liverpool John Moores University, UK
Ana-Belén García-Hernando, Universidad Politécnica de Madrid, Spain
Bezalel Gavish, Southern Methodist University, USA
Christos K. Georgiadis, University of Macedonia, Greece
Mariusz Glabowski, Poznan University of Technology, Poland
Katie Goeman, Hogeschool-Universiteit Brussel, Belgium
Hock Guan Goh, Universiti Tunku Abdul Rahman, Malaysia
Pedro Gonçalves, ESTGA - Universidade de Aveiro, Portugal

International Journal on Advances in Telecommunications
Volume 6, Numbers 1 & 2, 2013

CONTENTS

Analytical and Simulation Modeling of Limited-availability Systems with Multi-service Sources and Bandwidth Reservation

Mariusz Glabowski, Maciej Sobieraj, and Maciej Stasiak
Poznan University of Technology, Chair of Communication and Computer Networks, Poznan, Poland
Email: mariusz.glabowski@put.poznan.pl, maciej.sobieraj@put.poznan.pl, stasiak@et.put.poznan.pl

***Abstract*—The aim of this article is to present a new analytical calculation method for the occupancy distribution and the blocking probability in the so-called limited-availability group with multi-service sources and reservation mechanisms. The limited-availability group consists of links with different capacities. The article considers multi-service limited-availability systems with multi-service sources, in which each single traffic source can generate calls of different traffic classes. The results of analytical modeling of the limited-availability systems with multi-service sources and reservation mechanisms are compared with simulation data, which confirm a high accuracy of the method. Any possible application of the proposed model can be considered in the context of wireless networks with multi-service sources and reservation mechanisms. The proposed model can be also considered in the context of switching networks.**

Keywords-limited-availability systems; multi-service networks; multi-service sources; bandwidth reservation.

I. Introduction

Cellular networks are one of the most rapidly growing areas of telecommunications and one of the most popular systems of mobile communication. They can be used for voice transmission, but are also very efficient for sending data streams from different applications [1] [2] [3] [4]. Data transmission is a type of service that originally was not handled by cellular networks. Over time, data transmission has become more and more popular in expanding mobile networks. Data transmission services offered by operators include video conferencing services, streaming audio services, electronic mail and large file transmission [5] [6] [7] [8] [9].

The increase in intensity of traffic generated by data transmission services is accompanied by an increase in the requirements with respect to the volume of resources offered by networks. This also causes a growing necessity of working out mechanisms that introduce differentiation in Quality of Service (QoS) for particular data classes. The introduction of QoS differentiation mechanisms was conducive in turn to a development of new analytical models for dimensioning of multi-service mobile networks. The initial models of multi-service cellular networks assumed that a single traffic source of a given class could generate only one, strictly defined, type of the traffic stream (traffic sources class unequivocally determined the nature of a traffic stream). In the analytical models of such multi-service systems with single-service sources the class of traffic stream generated is defined by the class of its traffic source. Using the models of the multi-service systems with single-service source both cell groups without QoS mechanisms introduced [10] and cell groups with QoS differentiation mechanisms were considered, including resource reservation mechanisms [11] [12] [13].

With multi-service terminals becoming more and more universal in modern cellular networks, it has become necessary to develop new traffic models. In [14], a model of the multi-service network was presented for the first time, which assumed that a given and defined set of services was related to a single traffic source. The considered system was described as a multi-service system with multi-service sources. The considerations presented in [14] were limited, however, to a model of a full-availability group (single resource) without any QoS differentiation mechanisms introduced. In [1], the model of limited-availability group (a group of separated resources, e.g., a group of cells) with multi-service sources and bandwidth reservation has been presented.

This article proposes a generalized model of a limited-availability group of resources, in which — to facilitate resource usage — a reservation mechanism has been implemented. The model can be used for modeling groups of cells with different capacities in multi-service networks, thus expanding the results presented in [1] in which resources with the same capacities are discussed. In the proposed model, it is taken into consideration that a single terminal can generate various traffic streams corresponding to particular services implemented in the terminal. Additionally, it is assumed that the services cannot be used simultaneously by the terminal. This means that when the terminal is involved in the generation of traffic stream related to one service, it cannot at the same time generate traffic streams related to other services (A case when a single terminal can simultaneously generate a few traffic streams can be also taken into account in the analytical modeling – such a terminal can be treated as a few traffic sources).

The remaining part of the article is organized as follows. In Section II, the generalized model of the limited-availability systems with multi-service sources and resource reservation mechanism is proposed. In Section III, the simulator of multi-service systems with limited-availability is presented. In this section, the results of the blocking probability obtained for limited-availability systems with multi-service sources and reservation mechanisms are compared with the simulation data. Section IV concludes the article and presents further work.

II. Analytical Model of System with Multi-service Sources and Reservation Mechanisms

Let us consider a generalized model of the limited-availability group (GMLAG) with multi-service sources and the capacity V_L, presented in Figure 1. The generalized limited-availability group model is a model of a system that is composed of links with different capacities. The group is composed of z types of links. Each type h is unequivocally defined by the number υ_h of links of a given type and by the capacity f_h of each of the links of a given type (Figure 1). The total capacity V_L of the limited-availability group with different capacities of links is then $\sum_{h=1}^{z} \upsilon_h f_h$. The group can admit a call of a given class for service only when it can be entirely carried by the resources of one of the links.

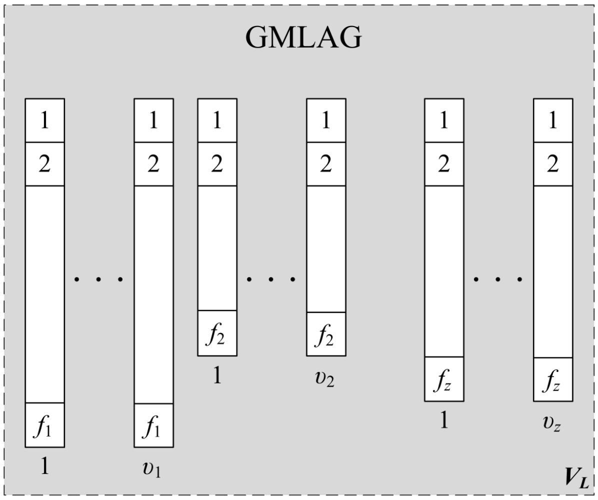


Figure 1. Generalized model of the limited-availability group

In the considered model, m traffic classes that belong to the set $\mathbb{M} = \{1, 2, ..., m\}$ are defined. A given class c requires t_c Basic Bandwidth Units (BBUs) to set up a new connection. The service time for class c calls has an exponential distribution with the parameter μ_c (service rate). In the group, the reservation mechanism has been applied. In accordance with the adopted reservation mechanism for a given class c, the reservation limit Q_c is introduced (Q_c is a certain occupancy state of the system, expressed in the number of BBUs being busy). The reservation mechanism can be applied to selected traffic classes from the set $\mathbb{M}$. The classes, in which the reservation limit has been introduced, are grouped into a new set of classes $\mathbb{R}$, which is a sub-set of the set $\mathbb{M}$. The parameter R_c that determines the reservation area (a certain number of occupancy states of the system) has also been defined. This parameter can be expressed by the following formula:

$$R_c = V_L - Q_c. \tag{1}$$

The system admits a call of class c that belongs to the set $\mathbb{R}$ for service only when this call can be entirely carried by the resources of an arbitrary single link and when the number of free BBUs in the group is higher or equal to the value of the reservation area R_c. A call of class c that does not belong to the set $\mathbb{R}$ can be serviced when this call can be entirely carried by the resources of an arbitrary single link. Thus, this is an example of a system with a state-dependent service process, in which the state dependence is the result of the structure of the group and the introduced reservation mechanism. An example of the limited-availability group with reservation mechanism applied for only class 1 is presented in Figure 2.

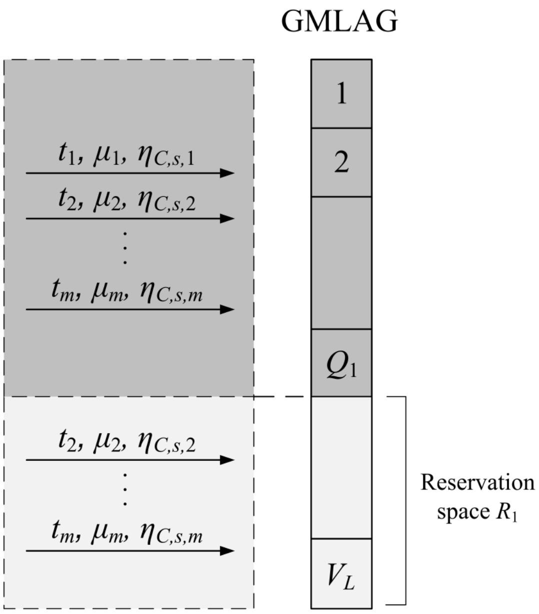


Figure 2. Generalized model of the limited-availability group with reservation mechanisms

The group is offered three types of Erlang (Poisson call streams), Engset (binomial call streams) and Pascal (negative binomial call streams) traffic streams [15]. The selected types of traffic cover three different types of the dependence between the mean arrival rates of calls and the occupancy state of the system: (1) the mean arrival rate of new calls does not depend on the occupancy state of the system (Erlang traffic), (2) the mean arrival rate of new calls decreases with the increase in the occupancy state of the system (Engset traffic), (3) the mean arrival rate of new calls increases with the increase in the occupancy state of the system (Pascal traffic).

Each traffic stream is generated by sources that belong to the corresponding set of traffic sources $\mathbb{Z}_{C,s}$. In set $\mathbb{Z}_{C,s}$, index C denotes the type of traffic stream generated by sources, which belong to this set and takes the value I for Erlang traffic stream, J for Engset traffic stream and K for Pascal traffic stream, respectively, while index s denotes the number of the set, the sources of which generate a given type of traffic stream. In the system, the s_I sets of traffic sources that generate Erlang traffic streams are defined, as well as s_J sets of traffic sources that generate Engset traffic streams and s_K sets of traffic sources that generate Pascal traffic streams. The total number of the sets of traffic sources is $S = s_I + s_J + s_K$. The sources that belong to the set $\mathbb{Z}_{C,s}$ can generate calls from the set $\mathbb{C}_{C,s} = \{1, 2, ..., c_{C,s}\}$ of traffic classes according to the available set of services.

The participation of class c (from the set $\mathbb{M}$) in the traffic structure of traffic generated by sources from the set $\mathbb{Z}_{C,s}$ is determined by the parameter $\eta_{C,s,c}$, which, for particular sets of Erlang, Engset and Pascal traffic sources, satisfies the following dependencies:

$$\sum_{c=1}^{c_{I,i}} \eta_{I,i,c} = 1, \qquad \sum_{c=1}^{c_{J,j}} \eta_{J,j,c} = 1, \qquad \sum_{c=1}^{c_{K,k}} \eta_{K,k,c} = 1. \tag{2}$$

To determine the value of traffic $A_{I,i,c}$ offered by Erlang sources that belong to the set $\mathbb{Z}_{I,i}$ as well as the traffic value $A_{J,j,c}(n)$ offered by Engset sources from the set $\mathbb{Z}_{J,j}$ and traffic $A_{K,k,c}(n)$ offered by Pascal sources from the set $\mathbb{Z}_{K,k}$ that generate calls of class c in the state of n busy BBUs, we use the following formulas [14]:

$$A_{I,i,c} = \eta_{I,i,c} \lambda_{I,i} / \mu_c, \tag{3}$$

$$A_{J,j,c}(n) = [\eta_{J,j,c} N_{J,j} - y_{J,j,c}(n)] \alpha_{J,j}, \tag{4}$$

$$A_{K,k,c}(n) = [\eta_{K,k,c} S_{K,k} + y_{K,k,c}(n)] \beta_{K,k}, \tag{5}$$

where:

- $\lambda_{I,i}$ – the mean arrival rate of new calls generated by a single Poisson source that belongs to the set $\mathbb{Z}_{I,i}$
- $\eta_{J,j,c}$ – the parameter that determines the participation of calls of class c in traffic generated by sources that belong to the set $\mathbb{Z}_{J,j}$,
- $\eta_{K,k,c}$ – the parameter that determines the participation of calls of class c in traffic generated by sources that belong to the set $\mathbb{Z}_{K,k}$,
- $N_{J,j}$ – the number of Engset traffic sources that belong to the set $\mathbb{Z}_{J,j}$,
- $S_{K,k}$ – the number of Pascal traffic sources that belong to the set $\mathbb{Z}_{K,k}$,
- $y_{J,j,c}(n)$ – the average number of calls of class c generated by Engset sources that belong to the set $\mathbb{Z}_{J,j}$ currently serviced in the system in the occupancy state n,
- $y_{K,k,c}(n)$ – the average number of calls of class c generated by Pascal sources that belong to the set $\mathbb{Z}_{K,k}$ currently serviced in the system in the occupancy state n,
- $\alpha_{J,j}$ – the average traffic intensity of traffic generated by a single Engset source that belongs to the set $\mathbb{Z}_{J,j}$, determined by the following formula:

  $$\alpha_{J,j} = \sum_{c=1}^{c_{J,j}} \eta_{J,j,c} \frac{\gamma_{J,j}}{\mu_c}, \tag{6}$$

  where $\gamma_{J,j}$ – the mean arrival rate of new calls generated by a single Engset source that belongs to the set $\mathbb{Z}_{J,j}$,
- $\beta_{K,k}$ – the average traffic intensity of traffic generated by a single Pascal source that belongs to the set $\mathbb{Z}_{K,k}$, defined by the following formula:

  $$\beta_{K,k} = \sum_{c=1}^{c_{K,k}} \eta_{K,k,c} \frac{\gamma_{K,k}}{\mu_c}, \tag{7}$$

  where $\gamma_{K,k}$ – the mean arrival rate of new calls generated by a single Pascal source that belongs to the set $\mathbb{Z}_{K,k}$.

We can notice that – according to Formulas (4) and (5) – with the case of Engset sources, the mean traffic intensity generated by new calls of individual traffic classes decreases with the increase in the occupancy state of the system, whereas in the case of Pascal sources the mean traffic intensity generated by new calls of individual traffic classes increases with the increase in the occupancy state of the system.

The first articles on modeling systems with limited-availability were limited to include multi-service systems with single-service sources only [16] [17] [18]. In [16], an approximate method for a determination of the blocking probability in the generalized limited-availability group model is proposed, while [17] discusses a group model with Engset traffic streams. Limited-availability groups servicing BPP traffic are studied in [18] [19].

The first studies on modeling multi-service systems with limited-availability and multi-service sources are presented in [20], while a method for modeling systems with multi-service sources and the reservation mechanism is described in [1]. The present article, taking advantage of the results presented in [16], proposes an expansion of the method proposed in [1] to provide a possibility of a determination of the occupancy distribution in the generalized limited-availability group model with the reservation mechanism. For this purpose, it is necessary to first determine conditional transition coefficients $\sigma_{c,S}(n)$ for calls of class c. These coefficients define the influence of a particular structure of a group on the process of the determination of the blocking probability. They can be determined on the basis of one of the two following methods: the one presented further on in the text of this section (the combinatorial method), and the convolutional method proposed in [21].

In the combinatorial method, the conditional transition probability $\sigma_{c,S}(n)$ for a traffic stream of class c in the generalized limited-availability group with the parameters: z, υ_h, f_h, V_L, is determined with the assumption, similarly as in the limited-availability basic model [22], that there are n busy BBUs in the considered group and that each distribution of free BBUs is treated as a division of busy BBUs of one class ($t_1 = 1$) between the links.

Let us consider first a limited-availability group composed of links of two types ($z = 2$). The capacity V_L of this group can be represented as the sum of the capacities of the links of the first and the second type, i.e.,

$$V_L = V_1 + V_2, \quad \text{where: } V_1 = \upsilon_1 f_1, V_2 = \upsilon_2 f_2. \quad (8)$$

A determination of the number of all possible distributions x of free BBUs in the group can in this case be successfully performed in two stages:

1) we determine the number of all possible combinations of distribution (division) of x free BBUs in the links of the two types, i.e., all combinations of the division of x BBUs into x_1 BBUs in the links of the first type and $x - x_1$ BBUs in the links of the second type;
2) we determine the number of possible distributions of a given number of BBUs in the links of the same type, i.e., the number of distributions x_1 BBUs in the links of the first type (with the limit of the capacity of a link to f_1 BBU taken into account), and $x-x_1$ BBUs in the links of the second type (with the limit of the capacity of a link to f_2 BBU taken into account) (Figure 3).

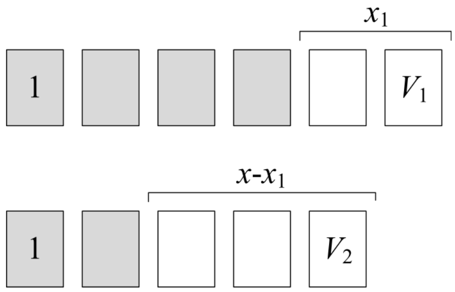


Figure 3. Possible distributions of x free BBUs in links of the two types

The number of distributions determined in the second stage is determined on the basis of the following formula which determines the number of arrangements of x free BBUs in υ links and the capacity of each link is equal to f BBUs and each link has at least t free BBUs:

$$F(x, \upsilon, f, t) = \sum_{r=0}^{\left\lfloor \frac{x-\upsilon t}{f-t+1} \right\rfloor} (-1)^r \binom{\upsilon}{r} \binom{x - \upsilon(t-1) - 1 - r(f-t+1)}{\upsilon - 1}. \quad (9)$$

Eventually, the number of possible distributions of x free BBUs in the links of the two types – with the assumption that in each link of the first type there are at least t_1 free BBUs and in each link of the second type there are at least t_2 free BBUs – can be expressed by the following formula:

$$F\{x, (\upsilon_1, \upsilon_2), (f_1, f_2), (t_1, t_2)\} = \sum_{x_1=0}^{x} F(x_1, \upsilon_1, f_1, t_1) F(x - x_1, \upsilon_2, f_2, t_2). \quad (10)$$

The number of possible distributions of x free BBUs in groups composed of links of three types can be determined in a similar way as for the group composed of links of two types:

$$F\{x, (\upsilon_1, \upsilon_2, \upsilon_3), (f_1, f_2, f_3), (t_1, t_2, t_3)\} = \sum_{x_1=0}^{x} \sum_{x_2=0}^{x-x_1} F(x_1, \upsilon_1, f_1, t_1) F(x_2, \upsilon_2, f_2, t_2) \cdot F(x - x_1 - x_2, \upsilon_3, f_3, t_3) \quad (11)$$

and then – generalizing the formula (11) – for groups composed of links of l types:

$$F\{(x, (\upsilon_1, \upsilon_2, \ldots, \upsilon_l), (f_1, f_2, \ldots, f_l), (t_1, t_2, \ldots, t_l)\} = \sum_{x_1=0}^{x} \sum_{x_2=0}^{x-x_1} \cdots \sum_{x_{l-1}=0}^{x-\sum_{r=1}^{l-2} x_r} F(x_1, \upsilon_1, f_1, t_1) \cdot F(x_2, \upsilon_2, f_2, t_2) \ldots F(x_{l-1}, \upsilon_{l-1}, f_{l-1}, t_{l-1}) \cdot F\left(x - \sum_{r=1}^{l-1} x_r, \upsilon_l, f_l, t_l\right). \quad (12)$$

Formula (12) can be eventually rewritten into the following form:

$$F\{(x, (\upsilon_1, \upsilon_2, \ldots, \upsilon_l), (f_1, f_2, \ldots, f_l), (t_1, t_2, \ldots, t_l)\} = \sum_{x_1=0}^{x} \cdots \sum_{x_{l-1}=0}^{x-\sum\limits_{r=1}^{l-2} x_r} \left\{ \prod_{z=1}^{l-1} F(x_z, \upsilon_z, f_z, t_z) \cdot F\left(x - \sum_{r=1}^{l-1} x_r, \upsilon_l, f_l, t_l\right) \right\}. \quad (13)$$

Formula (13) makes it possible to determine the transition coefficient for transitions between neighbouring states of the process occurring in the limited-availability group with z types of links in the state of n busy (i.e., $x = V - n$ free)

BBUs:

$$\sigma_{c,S}(n) = 1 - \frac{F\{V-n,(\upsilon_1,\ldots,\upsilon_l),(t_c-1,\ldots,t_c-1),(0,\ldots,0)\}}{F\{V-n,(\upsilon_1,\ldots,\upsilon_l),(f_1,\ldots,f_l),(0,\ldots,0)\}}. \tag{14}$$

The presented combinatorial method for a determination of conditional transition coefficients $\sigma_{c,S}(n)$ that determine a possibility of admittance of a call of class c for service in the occupancy state of n BBUs has been derived with the assumption that each distribution of free BBUs is treated as a result of the occupancy of resources by calls that demand 1 BBU [16]. To determine the parameter $\sigma_{c,S}(n)$, this method uses information on the capacity of individual links, but does not take into consideration, differences in the size of resources demanded by particular classes of calls.

Authors in [21] proposes a convolutional method for a determination of the conditional transition coefficient in the limited-availability group on the basis of independent unavailable distributions of BBUs in individual links. This method takes into consideration the size of required resources demanded by calls of individual traffic classes, but does not take into consideration link capacities. Comparative results provided by the studies for both of the methods, i.e., the combinatorial and the convolutional method, indicate their similar accuracy [21].

Observe that in the case of the considered model of the limited-availability group with multi-service traffic sources and reservation, the operation of the reservation mechanism introduces an additional dependence between the service stream in the system and the current state of the system. To determine this dependence, the parameter $\sigma_{c,R}(n)$ is introduced. The parameter $\sigma_{c,R}(n)$ can be calculated using the following formula:

$$\sigma_{c,R}(n) = \begin{cases} 1 & \text{for } n \leqslant Q_c \wedge c \in \mathbb{R}, \\ 0 & \text{for } n > Q_c \wedge c \in \mathbb{R}, \\ 1 & \text{for } c \notin \mathbb{R}. \end{cases} \tag{15}$$

The reservation mechanism is introduced to the group regardless of its structure, which allows us to carry on with product-form determination of the total coefficient of passing (transition coefficient) $\sigma_{c,\mathrm{Tot}}(n)$ in the generalized model of limited-availability group:

$$\sigma_{c,\mathrm{Tot}}(n) = \sigma_{c,S}(n) \cdot \sigma_{c,R}(n). \tag{16}$$

Having the values of offered traffic $A_{I,i,c}$, $A_{J,j,c}(n)$, $A_{K,k,c}(n)$ and the total coefficient of passing $\sigma_{c,\mathrm{Tot}}(n)$ at our disposal, we are in position to modify the original Kaufman-Roberts formula [23] [24] in order to determine the occupancy distribution in the limited-availability group with multi-service traffic sources and the reservation mechanism:

$$\begin{aligned} n[P_n]_{V_L} &= \sum_{i=1}^{s_I}\sum_{c=1}^{c_{I,i}} A_{I,i,c}\sigma_{c,\mathrm{Tot}}(n-t_c)t_c[P_{n-t_c}]_{V_L} + \\ &+ \sum_{j=1}^{s_J}\sum_{c=1}^{c_{J,j}} A_{J,j,c}(n-t_c)\sigma_{c,\mathrm{Tot}}(n-t_c)t_c[P_{n-t_c}]_{V_L} + \\ &+ \sum_{k=1}^{s_K}\sum_{c=1}^{c_{K,k}} A_{K,k,c}(n-t_c)\sigma_{c,\mathrm{Tot}}(n-t_c)t_c[P_{n-t_c}]_{V_L}, \end{aligned} \tag{17}$$

where $[P_n]_{V_L}$ is the occupancy distribution (the probability of n busy BBUs) in a system with the capacity V_L, and the parameter $\sigma_{c,\mathrm{Tot}}(n)$ determines the additional dependence between the service stream and the current state of the system resulting from the specific structure of the group and the applied reservation mechanism [25].

Having the values of individual state probabilities $[P_n]_{V_L}$, determined on the basis of Formula (17), we are in position to determine the average number of serviced calls of class c, generated by sources that belong to the sets $\mathbb{Z}_{J,j}$ (Engset sources) and $\mathbb{Z}_{K,k}$ (Pascal sources). For this purpose, we use the following formulas:

$$y_{J,j,c}(n) = \begin{cases} A_{J,j,c}(n-t_c)\sigma_{c,\mathrm{Tot}}(n-t_c)[P_{n-t_c}]_{V_L}/[P_n]_{V_L} & \text{for } n \leqslant V_L, \\ 0, & \text{for } n > V_L. \end{cases} \tag{18}$$

$$y_{K,k,c}(n) = \begin{cases} A_{K,k,c}(n-t_c)\sigma_{c,\mathrm{Tot}}(n-t_c)[P_{n-t_c}]_{V_L}/[P_n]_{V_L} & \text{for } n \leqslant V_L, \\ 0, & \text{for } n > V_L. \end{cases} \tag{19}$$

The knowledge of the occupancy $[P_n]_{V_L}$ is required to determine the parameters $y_{J,j,c}(n)$ and $y_{K,k,c}(n)$. Whereas, to determine the occupancy $[P_n]_{V_L}$ it is necessary to know the values of the parameters $y_{J,j,c}(n)$ and $y_{K,k,c}(n)$. Equations (18), (19), and (17) form thus a set of confounding equations. To solve a given set of confounding equations it is necessary to employ iterative methods [26] [27].

Assuming that the distribution $[P_n^{(l)}]_{V_L}$ is the occupancy distribution, determined in the l-th iteration, while $y_{J,j,c}^{(l)}(n)$ and $y_{K,k,c}^{(l)}(n)$ define the average number of serviced calls of class c generated by traffic sources that belong respectively to the sets $\mathbb{Z}_{J,j}$ and $\mathbb{Z}_{K,k}$, we can write:

$$y_{J,j,c}^{(l+1)}(n) = \begin{cases} A_{J,j,c}^{(l)}(n-t_c)\sigma_{c,\mathrm{Tot}}(n-t_c)[P_{n-t_c}^{(l)}]_{V_L}/[P_n^{(l)}]_{V_L} & \text{for } n \leqslant V_L, \\ 0, & \text{for } n > V_L. \end{cases} \tag{20}$$

$$y_{K,k,c}^{(l+1)}(n) = \begin{cases} A_{K,k,c}^{(l)}(n-t_c)\sigma_{c,\mathrm{Tot}}(n-t_c)[P_{n-t_c}^{(l)}]_{V_L}/[P_n^{(l)}]_{V_L} & \text{for } n \leqslant V_L, \\ 0, & \text{for } n > V_L. \end{cases} \tag{21}$$

The iteration process, involving Formulas (17), (20), and (21), terminates when the assumed accuracy ϵ of the iteration process is reached:

$$\forall_{0\leq n\leq V}\ \left|\frac{y^{l-1}_{J,j,c}(n) - y^{(l)}_{J,j,c}(n)}{y^{(l)}_{J,j,c}(n)}\right| \leqslant \epsilon, \tag{22}$$

$$\forall_{0\leq n\leq V}\ \left|\frac{y^{l-1}_{K,k,c}(n) - y^{(l)}_{K,k,c}(n)}{y^{(l)}_{K,k,c}(n)}\right| \leqslant \epsilon. \tag{23}$$

In case of the proposed model the assumed accuracy ϵ of the iteration process is always reached.

Having determined all probabilities $\sigma_{c,\mathrm{Tot}}(n)$ as well as the occupancy distribution $[P_n]_{V_L}$ in a limited-availability group composed of different types of links, we are in position to proceed with a determination of the blocking probability for individual traffic classes from the set $\mathbb{M} = \{1, 2, ..., m\}$. The blocking state in the generalized limited-availability group model occurs when none of the links have a sufficient number of free BBUs to service a call of class c. This means that each occupancy state n in a link of the type s, such as: $(f_s - t_c + 1 \leq n \leq f_s)$, is a blocking state. Possible blocking states for a traffic stream of class c in a limited-availability group composed of links of l types can be thus determined by the following formula:

$$V - \sum\nolimits_{s=1}^{l} \upsilon_s(t_c - 1) \leq n \leq V. \tag{24}$$

In addition, in the case of a limited-availability group with the reservation mechanism, the blocking state also occurs – for calls of class c – in all occupancy states of the group that are higher than the reservation limit Q_c, i.e., for $n > Q_c$. On the basis of the determined values of the conditional transition coefficients $\sigma_{c,\mathrm{Tot}}(n)$ and the occupancy distribution $[Q_n]_V$, the blocking probability for calls of class c, that belong to the set $\mathbb{M} = \{1, 2, ..., m\}$, can be expressed by the formula:

$$E_c = \sum_{n=0}^{V_L} [P_n]_{V_L} \left[1 - \sigma_{c,\mathrm{Tot}}(n)\right]. \tag{25}$$

III. Model Verification and Case Studies

A. Description of the simulator

In order to verify the accuracy of the proposed methods for analytical determination of the blocking probability in multi-service systems with the bandwidth reservation mechanism, the authors built a simulator of their own. The simulator was written in C++ with the application of the object-oriented programming rules. The process interaction approach was adopted as method of choice for simulations [28]. The devised simulator makes it possible, among other things, to determine the value of the blocking probability for individual traffic classes in the generalized limited-availability group model with the reservation mechanism.

1) Input data and end task condition: Capacity and the group structure provide input data to the simulator. In addition, for each simulated traffic class we give the number of demanded BBUs, service time and the parameters related to the introduced reservation mechanism (reservation limits). We also define the sets of traffic sources and their type (Erlang, Engset, Pascal). Finally, we give the average value of traffic offered to a single BBU of the system.

Thus, to carry out simulation studies of a system with the capacity V_L, composed of z types of links in which a given type h consists of υ_h links, while the capacity of a single link is f_h, we are to provide values for the following parameters that describe traffic classes:

- the number of defined traffic classes m,
- the number of demanded t_c BBUs necessary to set up a connection of class c and the average service time μ_c^{-1} of a call of class c,
- the number of sets of traffic classes S,
- the set of traffic sources $\mathbb{Z}_{C,s}$ of type C, the number of classes $c_{C,s}$ that belong to a given set, the share of $\eta_{C,s,c}$ calls of class c in traffic generated by traffic sources of a given set, and the number $N_{C,s}$ or $S_{C,s}$ of traffic sources in a given set,
- the set $\mathbb{R}$ of traffic classes to which the reservation mechanism has been introduced,
- the reservation limit Q_c defined for calls of class c from the set $\mathbb{R}$.

In addition, we also give the average traffic a offered to a single BBU of the system. The mean value of offered traffic a can be calculated using following equation:

$$a = \left[\sum_{i=1}^{s_I} \lambda_{I,i} \sum_{c=1}^{c_{I,i}} t_c\eta_c \sum_{c=1}^{c_{I,i}} \eta_c/\mu_c + \right. + \sum_{j=1}^{s_J} \gamma_{J,j} N_{J,j} \sum_{c=1}^{c_{J,j}} t_c\eta_c \sum_{c=1}^{c_{J,j}} \eta_c/\mu_c + \left. + \sum_{k=1}^{s_K} \gamma_{K,k} S_{K,k} \sum_{c=1}^{c_{K,k}} t_c\eta_c \sum_{c=1}^{c_{K,k}} \eta_c/\mu_c \right] / V_L. \tag{26}$$

On the basis of the above parameters the intensity $\lambda_{I,i}$, $\gamma_{J,j}$ or $\gamma_{K,k}$ of the occurrence of new calls depending on the type of a traffic stream is determined in the simulator. In the case of Engset and Pascal traffic streams, the intensities $\gamma_{J,j}$ and $\gamma_{K,k}$ designate the intensity of the appearance of new calls generated by a single free source. Therefore, the parameters $\lambda_{I,i}$, $\gamma_{J,j}$ and $\gamma_{K,k}$ can be determined on the basis of the following formulas:

$$\lambda_{I,i} = \frac{aV_L}{S\left[\sum_{c=1}^{c_{I,i}} t_c\eta_{I,i,c}\right]\left[\sum_{c=1}^{c_{I,i}} \mu_c\eta_{I,i,c}\right]}, \tag{27}$$

$$\gamma_{J,j} = \frac{aV_L}{S\left[\sum_{c=1}^{c_{J,j}} t_c\eta_{J,j,c}\right]\left[\sum_{c=1}^{c_{J,j}} \mu_c\eta_{J,j,c}\right] N_{J,j}}, \tag{28}$$

$$\gamma_{K,k} = \frac{aV_L}{S\left[\sum_{c=1}^{c_{K,k}} t_c \eta_{K,k,c}\right]\left[\sum_{c=1}^{c_{K,k}} \mu_c \eta_{K,k,c}\right] S_{K,k}}. \quad (29)$$

The next parameters that are determined on the basis of Formulas (27), (28), and (29) are used then as input parameters to call generators that are used by particular traffic sources.

The condition for the simulator to end the simulation is the counted appropriate number of generated calls of the least active class (with the lowest intensity of call generation). This number is selected in such a way as to obtain 95% confidence intervals. The average result of 5 series is then calculated. In practice, to obtain confidence intervals at the level of 95% we need 1,000,000 generated calls of the least active class.

2) Simulation algorithm: The simulation algorithm can be determined using the following steps:

1) Initial configuration of the simulation model.
2) Checking the end task condition of the simulation. If the end task condition is fulfilled, the simulation is terminated and the results are returned.
3) Updating of system time until the first event from the list appears.
4) Execution of the first event from the list.
5) Removal of the first event from the list and return to Step 2.

Two events have been defined in the simulation model: *appearance of a new call* and *termination of the call service*. According to the process interaction approach, these events are serviced by one function. This function has a different form for each type of Erlang, Engset and Pascal traffic streams.

The described approach enables us to define many different traffic classes in the system and to assign them to different types of the sets of traffic sources. All sources that generate calls of different traffic classes are to be created in the initial configuration of the simulation model.

3) Simulation of the system with sets of Erlang sources: Consider a system in which a set of Erlang traffic sources $\mathbb{Z}_{I,i}$ is defined. The set of traffic classes whose calls can be generated by sources from the set $\mathbb{Z}_{I,i}$ looks as follows: $\mathbb{C}_{I,i} = \{1, 2, ..., c_{I,i}\}$. In the initial configuration of the system, it is necessary to plan ahead the appearance of a call of class c from the set $\mathbb{C}_{I,i}$. Thus, the function that executes events that are related to the set of Erlang traffic sources will have the following form:

1) Planning of the appearance of a new call of class c according to the exponential distribution with the intensity $\lambda_{I,i}$ as the parameter. A choice of class c from the set $\mathbb{C}_{I,i}$ on the basis of the parameter $\eta_{I,i,c}$ according to the uniform distribution. Inclusion of the event on the list.
2) Checking network resources for the purpose of admitting a call for service:
   a) Checking if any of the links of the group has at least t_c free BBUs. If not, the call is lost.
   b) In the case of classes that belong to the set of classes $\mathbb{R}$, for which the reservation mechanism has been introduced, checking the occupancy state of the system in relation to the reservation limit Q_c. When the occupancy state is higher than the reservation limit Q_c, a call of class c is lost.

   If any of the conditions is not satisfied, the next steps are omitted.
3) Occupation of resources demanded by a call of class c.
4) Planning (scheduling) of a termination of service according to the exponential distribution with the intensity μ_c as the parameter. Inclusion of the event on the list.
5) Termination of service and release of resources.

4) Simulation of a system with sets of Engset traffic sources: Consider now a system in which a set of Engset traffic sources $\mathbb{Z}_{J,j}$ is defined. The set of traffic classes in which calls can be generated by $N_{J,j}$ sources from the set $\mathbb{Z}_{J,j}$ looks as follows: $\mathbb{C}_{J,j} = \{1, 2, ..., c_{J,j}\}$. In the initial configuration of the system, it is necessary to plan ahead the appearance of calls of class c from the set $\mathbb{C}_{J,j}$, generated by each of $N_{J,j}$ sources. Thus, the function that executes events related to the set of Engset traffic sources will take on the following form:

1) Checking network resources for the purpose of the admittance of a call for service:
   a) Checking if any of the links of the group has at least t_c free BBUs. If not, the call is lost.
   b) In the case of classes that belong to the set $\mathbb{R}$, for which the reservation mechanism has been introduced, checking the occupancy state of the state in relation to the reservation limit Q_c. When the occupancy state is higher than the reservation limit Q_c, a call of class c is lost.

   If any of the conditions is not satisfied, the simulation proceeds to Step 5.
2) Occupation of resources demanded by a call of class c.
3) Planning (Scheduling) of the termination of service according to the exponential distribution where the parameter is the intensity μ_c. Inclusion of the event on the list.
4) Termination of service and network release.
5) Planning of the appearance of a new call of class c according to the exponential distribution where the intensity $\gamma_{J,j}$ is the parameter. Choice of the class c from the set $\mathbb{C}_{J,j}$ on the basis of the parameter $\eta_{J,j,c}$ according to the uniform distribution. Inclusion of the event on the list.

5) Simulation of a system with Pascal traffic sources: Consider also a system in which a set of Pascal traffic sources $\mathbb{Z}_{K,k}$ is defined. The set of traffic classes in which calls can be generated by $S_{K,k}$ sources from the set $\mathbb{Z}_{K,k}$ looks as follows: $\mathbb{C}_{K,k} = \{1, 2, ..., c_{K,k}\}$. In the initial configuration of the system, it is necessary to plan ahead the appearance of calls of class c from the set $\mathbb{C}_{K,k}$, generated by each of $S_{K,k}$ sources. Thus, the function that executes events related to the set of Pascal traffic sources will take on the following form:

1) Planning (scheduling) of the appearance of a new call of class c according to the exponential distribution where the parameter is the intensity $\gamma_{K,k}$. Choice of class c from the set $\mathbb{C}_{K,k}$ on the basis of the parameter $\eta_{K,k,c}$ according to the uniform distribution. Inclusion of the event on the list.
2) Checking network resources for the purpose of the admitting a call for service:
   a) Checking if any of the links of the group has at least t_c free BBUs. If not, the call is lost.
   b) In the case of classes that belong to the set $\mathbb{R}$ classes, for which the reservation mechanism has been introduced, checking the occupancy state of the system in relation to the reservation limit Q_c. When the occupancy state is higher than the reservation limit Q_c, a call of class c is lost.

   If any of the conditions is not satisfied, the next steps are omitted.
3) Occupation of the resources demanded by a call of class c.
4) Planning (scheduling) of the termination of service according to the exponential distribution where the parameter is the intensity μ_c. Inclusion of the event on the list.
5) Addition of a new source related to the admitted call. Planning of the appearance of a new call of class c generated by the new traffic source according to the exponential distribution for which the parameter is the intensity $\gamma_{K,k}$. Choice of class c from the set $\mathbb{C}_{K,k}$ on the basis of the parameter $\eta_{K,k,c}$ according to the uniform distribution. Inclusion of the event on the list.
6) Termination of service and resource release.
   a) Removal of the source related to the call that has just been terminated in service (this source is added at the moment a call is admitted for service).
   b) Removal of the event related to the removed source.

B. Simulation studies of systems with limited-availability and bandwidth reservation

The presented method for a determination of the blocking probability in systems with multi-service traffic sources and reservation mechanisms is an approximate method. In order to confirm the adopted assumptions, the results of the analytical calculations were compared with the simulation data. The research was carried for seven systems described below:

1) Limited-availability system No. 1
   - Capacity: $z = 1$, $\upsilon_1 = 2$, $f_1 = 20$ BBUs, $V_L = 40$ BBUs,
   - Number of traffic classes: 3
   - Structure of traffic: $t_1 = 1$ BBU, $\mu_1^{-1} = 1$, $t_2 = 2$ BBUs, $\mu_2^{-1} = 1$,$t_3 = 6$ BBUs, $\mu_3^{-1} = 1$, $R_1 = R_2 = 33$ BBUs
   - Sets of sources: $\mathbb{C}_{I,1} = \{1, 2\}$, $\eta_{I,1,1} = 0.6$, $\eta_{I,1,2} = 0.4$, $\mathbb{C}_{J,2} = \{2, 3\}$, $\eta_{J,2,2} = 0.7$, $\eta_{J,2,3} = 0.3$, $N_2 = 60$
2) Limited-availability system No. 2
   - Capacity: $z = 1$, $\upsilon_1 = 2$, $f_1 = 30$ BBUs, $V_L = 60$ BBUs,
   - Number of traffic classes: 3
   - Structure of traffic: $t_1 = 1$ BBU, $\mu_1^{-1} = 1$, $t_2 = 3$ BBUs, $\mu_2^{-1} = 1$,$t_3 = 7$ BBUs, $\mu_3^{-1} = 1$, $R_1 = R_2 = 51$ BBUs
   - Sets of sources: $\mathbb{C}_{I,1} = \{1\}$, $\eta_{I,1,1} = 1.0$, $\mathbb{C}_{J,2} = \{1, 2\}$, $\eta_{J,2,1} = 0.6$, $\eta_{J,2,2} = 0.4$, $N_2 = 50$, $\mathbb{C}_{K,3} = \{2, 3\}$, $\eta_{K,3,2} = 0.7$, $\eta_{K,3,3} = 0.3$, $S_3 = 50$
3) Limited-availability system No. 3
   - Capacity: $z = 1$, $\upsilon_1 = 4$, $f_1 = 20$ BBUs, $V_L = 80$ BBUs,
   - Number of traffic classes: 4
   - Structure of traffic: $t_1 = 1$ BBU, $\mu_1^{-1} = 1$, $t_2 = 2$ BBUs, $\mu_2^{-1} = 1$,$t_3 = 4$ BBUs, $\mu_3^{-1} = 1$, $t_4 = 9$ BBUs, $\mu_4^{-1} = 1$, $R_1 = R_2 = R_3 = 63$ BBUs
   - Sets of sources: $\mathbb{C}_{I,1} = \{1, 2\}$, $\eta_{I,1,1} = 0.6$, $\eta_{I,1,2} = 0.4$, $\mathbb{C}_{J,2} = \{2, 3\}$, $\eta_{J,2,2} = 0.7$, $\eta_{J,2,3} = 0.3$, $N_2 = 70$, $\mathbb{C}_{K,3} = \{2, 3, 4\}$, $\eta_{K,3,2} = 0.3$, $\eta_{K,3,3} = 0.2$, $\eta_{K,3,4} = 0.5$, $S_3 = 140$
4) Limited-availability system No. 4
   - Capacity: $z = 2$, $\upsilon_1 = 1$, $f_1 = 20$ BBUs, $\upsilon_2 = 1$, $f_2 = 30$ BBUs, $V_L = 50$ BBUs,
   - Number of traffic classes: 3
   - Structure of traffic: $t_1 = 1$ BBU, $\mu_1^{-1} = 1$, $t_2 = 3$ BBUs, $\mu_2^{-1} = 1$,$t_3 = 5$ BBUs, $\mu_3^{-1} = 1$, $R_1 = R_2 = 35$ BBUs
   - Sets of sources: $\mathbb{C}_{I,1} = \{1, 2\}$, $\eta_{I,1,1} = 0.6$, $\eta_{I,1,2} = 0.4$, $\mathbb{C}_{J,2} = \{2, 3\}$, $\eta_{J,2,2} = 0.7$, $\eta_{J,2,3} = 0.3$, $N_2 = 60$
5) Limited-availability system No. 5
   - Capacity: $z = 2$, $\upsilon_1 = 1$, $f_1 = 30$ BBUs, $\upsilon_2 = 1$, $f_2 = 40$ BBUs, $V_L = 70$ BBUs,
   - Number of traffic classes: 3
   - Structure of traffic: $t_1 = 1$ BBU, $\mu_1^{-1} = 1$, $t_2 = 4$ BBUs, $\mu_2^{-1} = 1$,$t_3 = 7$ BBUs, $\mu_3^{-1} = 1$, $R_1 = R_2 = 53$ BBUs

- Sets of sources: $\mathbb{C}_{I,1} = \{1\}$, $\eta_{I,1,1} = 1.0$, $\mathbb{C}_{J,2} = \{1,2\}$, $\eta_{J,2,1} = 0.6$, $\eta_{J,2,2} = 0.4$, $N_2 = 50$, $\mathbb{C}_{K,3} = \{2,3\}$, $\eta_{K,3,2} = 0.7$, $\eta_{K,3,3} = 0.3$, $S_3 = 50$

6) Limited-availability system No. 6
   - Capacity: $z = 2$, $\upsilon_1 = 2$, $f_1 = 20$ BBUs, $\upsilon_2 = 2$, $f_2 = 30$ BBUs, $V_L = 100$ BBUs,
   - Number of traffic classes: 4
   - Structure of traffic: $t_1 = 1$ BBU, $\mu_1^{-1} = 1$, $t_2 = 3$ BBUs, $\mu_2^{-1} = 1$, $t_3 = 5$ BBUs, $\mu_3^{-1} = 1$, $t_4 = 7$ BBUs, $\mu_4^{-1} = 1$, $R_1 = R_2 = R_3 = 67$ BBUs
   - Sets of sources: $\mathbb{C}_{I,1} = \{1,2\}$, $\eta_{I,1,1} = 0.6$, $\eta_{I,1,2} = 0.4$, $\mathbb{C}_{J,2} = \{2,3\}$, $\eta_{J,2,2} = 0.7$, $\eta_{J,2,3} = 0.3$, $N_2 = 70$, $\mathbb{C}_{K,3} = \{2,3,4\}$, $\eta_{K,3,2} = 0.3$, $\eta_{K,3,3} = 0.2$, $\eta_{K,3,4} = 0.5$, $S_3 = 140$

7) Limited-availability system No. 7
   - Capacity: $z = 3$, $\upsilon_1 = 1$, $f_1 = 20$ BBUs, $\upsilon_2 = 1$, $f_2 = 30$ BBUs, $\upsilon_3 = 1$, $f_3 = 40$ BBUs, $V_L = 90$ BBUs,
   - Number of traffic classes: 4
   - Structure of traffic: $t_1 = 1$ BBU, $\mu_1^{-1} = 1$, $t_2 = 4$ BBUs, $\mu_2^{-1} = 1$, $t_3 = 6$ BBUs, $\mu_3^{-1} = 1$, $t_4 = 8$ BBUs, $\mu_4^{-1} = 1$, $t_5 = 10$ BBUs, $\mu_5^{-1} = 1$, $R_1 = R_2 = R_3 = 74$ BBUs
   - Sets of sources: $\mathbb{C}_{I,1} = \{1,2\}$, $\eta_{I,1,1} = 0.6$, $\eta_{I,1,2} = 0.4$, $\mathbb{C}_{J,2} = \{2,3\}$, $\eta_{J,2,2} = 0.7$, $\eta_{J,2,3} = 0.3$, $N_2 = 180$, $\mathbb{C}_{K,3} = \{3,4,5\}$, $\eta_{K,3,3} = 0.3$, $\eta_{K,3,4} = 0.2$, $\eta_{K,3,5} = 0.5$, $S_3 = 90$

The results of the research study are presented in Figures 4–10, depending on the value of traffic a offered to a single BBU. The results of the simulation are shown in the charts in the form of marks with 95% confidence intervals that have been calculated according to the t-Student distribution for the five series with 1,000,000 calls of each class. For each of the points of the simulation, the value of the confidence interval is at least one order lower than the mean value of the results of the simulation. In many cases, the value of the simulation interval is lower than the height of the sign used to indicate the value of the simulation experiment. We can notice that the results of analytical calculations agree with the simulation data for both lower and higher traffic intensities.

IV. Conclusion and Further Work

This article proposes a new method for a calculation of the occupancy distribution and the blocking probability in limited-availability systems with multi-service traffic sources and reservation mechanisms. The method can be used in modeling connection handoff between cells in cellular systems [10], as well as in modeling outgoing directions of switching networks [29]. The proposed method is based on the iterative algorithm for a determination of the average value of traffic sources being serviced in particular states

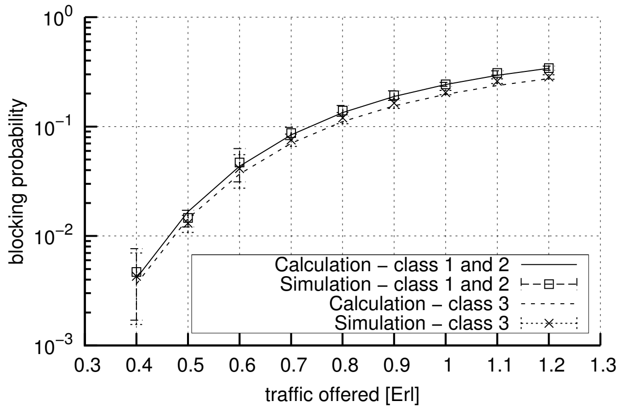


Figure 4. Blocking probability in the limited-availability group No. 1 with reservation mechanism; the reservation mechanism equalizes the blocking probability for calls of class 1 and 2

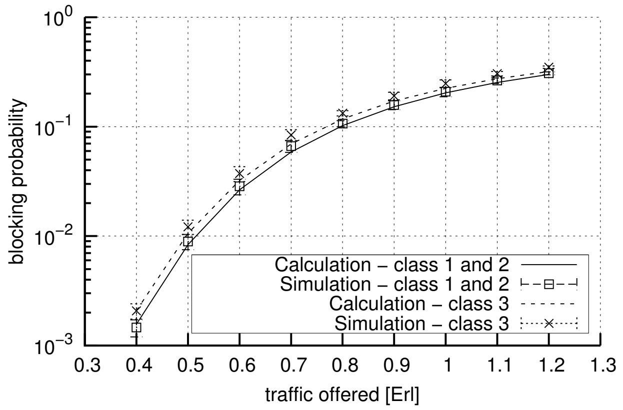


Figure 5. Blocking probability in the limited-availability system No. 2 with reservation mechanism; the reservation mechanism equalizes the blocking probability for calls of class 1 and 2

of the system. The results of analytical calculations were compared with the simulation data, which confirmed high accuracy of the proposed method. The proposed method is not complicated and can be easily implemented.

In the further work, we plan to develop analytical models of the multi-service systems with multi-service sources, in which different call admission control mechanisms will be applied, i.e., an analytical model of multi-service networks with threshold mechanisms and multi-service sources, and a model of multi-service systems with hysteresis and multi-service sources.

References

[1] M. Głąbowski, M. Sobieraj, and M. Stasiak, "Modelling limited-availability systems with multi-service sources and bandwidth reservation," in *Proceedings of the The Eighth Advanced International Conference on Telecommunications (AICT 2012)*, IARIA. Stuttgart, Germany: IARIA, May 2012, pp. 105–110. [Online]. Available: http://www.thinkmind.org/index.php?view=-

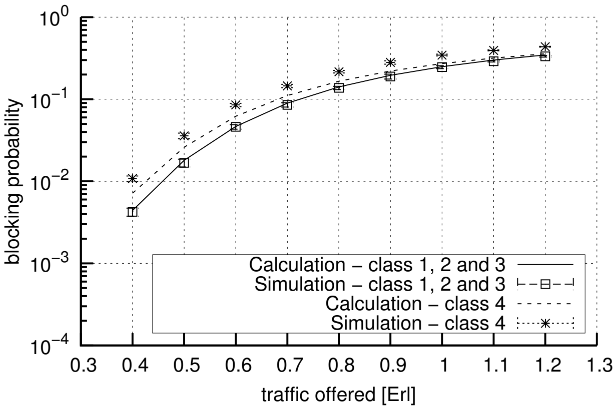


Figure 6. Blocking probability in the limited-availability system No. 3 with reservation mechanism; the reservation mechanism equalizes the blocking probability for calls of class 1, 2 and 3

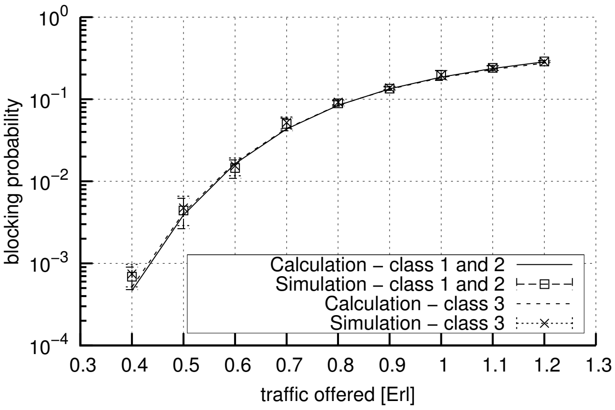


Figure 8. Blocking probability in the limited-availability system No. 5 with reservation mechanism; the reservation mechanism equalizes the blocking probability for calls of class 1 and 2

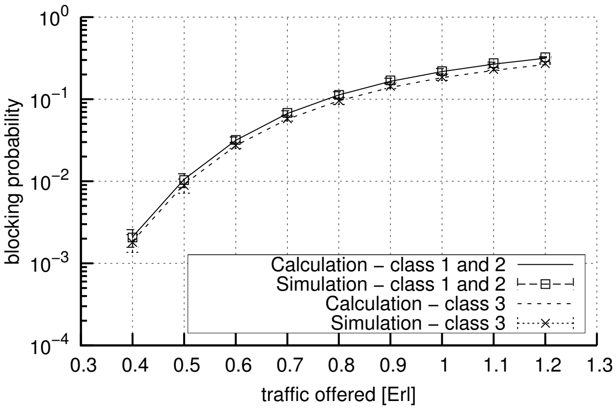


Figure 7. Blocking probability in the limited-availability system No. 4 with reservation mechanism; the reservation mechanism equalizes the blocking probability for calls of class 1 and 2

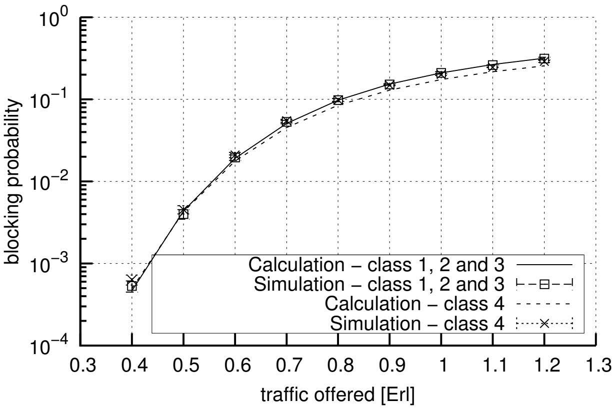


Figure 9. Blocking probability in the limited-availability system No. 6 with reservation mechanism; the reservation mechanism equalizes the blocking probability for calls of class 1, 2 and 3

article&articleid=aict_2012_5_30_10167 <retrieved: June, 2013>

[2] 3GPP, "High Speed Packet Access (HSPA) evolution; Frequency Division Duplex (FDD)," 3rd Generation Partnership Project (3GPP), TR 25.999, Mar. 2008. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/25999.htm <retrieved: June, 2013>

[3] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRAN); Overall description; Stage 2," 3rd Generation Partnership Project (3GPP), TS 36.300, Sep. 2008. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/36300.htm <retrieved: June, 2013>

[4] 3GPP, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access," 3rd Generation Partnership Project (3GPP), TS 23.401, Sep. 2008. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/23401.htm <retrieved: June, 2013>

[5] A. Hadden, "Mobile broadband - where the next generation leads us [industry perspectives]," *Wireless Communications, IEEE*, vol. 16, no. 6, pp. 6–9, Dec. 2009.

[6] S. Ortiz Jr., "4G wireless begins to take shape," *Computer*, vol. 40, no. 11, pp. 18–21, 2007.

[7] S. R. Subramanya, "Emerging mobile technologies and issues," in *Proceedings of the International Symposium on Collaborative Technologies and Systems*. Washington, DC, USA: IEEE Computer Society, 2006, p. 172.

[8] V. G. Vassilakis, I. D. Moscholios, and M. D. Logothetis, "Call-level performance modelling of elastic and adaptive service-classes with finite population," *IEICE Transactions on Communications*, vol. E91-B, no. 1, pp. 151–163, 2008.

[9] M. Głąbowski, M. Stasiak, A. Wiśniewski, and P. Zwierzykowski, "Uplink blocking probability calculation for cellular systems with WCDMA radio interface, finite source population and differently loaded neighbouring cells," in *Proceedings of Asia-Pacific Conference on Communications*, Perth, Australia, 2005, pp. 138–142, best paper award.

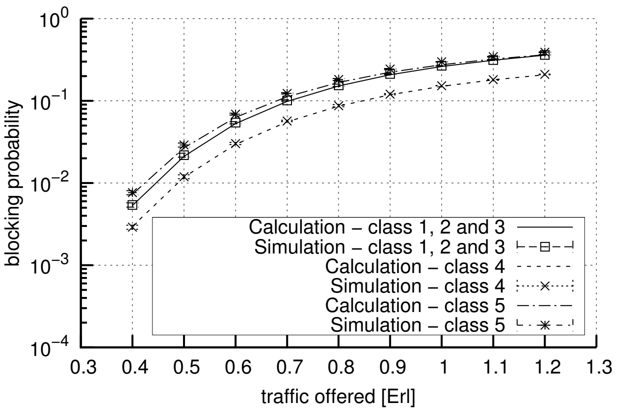


Figure 10. Blocking probability in the limited-availability system No. 7 with reservation mechanism; the reservation mechanism equalizes the blocking probability for calls of class 1, 2 and 3

[10] M. Głąbowski, M. Sobieraj, and M. Stasiak, "Analytical modeling of the influence of connection handoff on UMTS traffic characteristics," in *Proceedings of Third Advanced International Conference on Telecommunications*, Morne, may 2007.

[11] M. Sobieraj, M. Stasiak, J. Weissenberg, and P. Zwierzykowski, "Analytical model of the single threshold mechanism with hysteresis for multi-service networks," *IEICE Transactions on Communications*, vol. E95-B, no. 1, pp. 120–132, 2012.

[12] M. Głąbowski, M. Sobieraj, and M. Stasiak, "Evaluation of traffic characteristics of UMTS with bandwidth reservation and handoff mechanism," in *Proceedings of 14th IEEE International Conference On Telecommunications*, Penang, May 2007, pp. 736–740.

[13] M. Głąbowski, M. Sobieraj, and M. Stasiak, "Blocking probability calculation in UMTS networks with bandwidth reservation, handoff mechanism and finite source population," in *Proceedings of 7th International Symposium on Communications and Information Technologies*, Sydney, Oct. 2007, pp. 433–438.

[14] M. Głąbowski, M. Sobieraj, and M. Stasiak, "Analytical modeling of multi-service systems with multi-service sources," in *Proceedings of 16th Asia-Pacific Conference on Communications (APCC)*. Auckland, New Zealand: IEEE, Oct. 2010, pp. 285–290.

[15] H. Akimaru and K. Kawashima, *Teletraffic: Theory and Application*. Berlin-Heidelberg-New York: Springer, 1999.

[16] M. Głąbowski and M. Stasiak, "Multi-rate model of the group of separated transmission links of various capacities," in *Telecommunications and Networking - ICT 2004, 11th International Conference on Telecommunications, Fortaleza, Brazil, August 1-6, 2004, Proceedings*, ser. Lecture Notes in Computer Science, J. N. de Souza, P. Dini, and P. Lorenz, Eds., vol. 3124. Springer, 2004, pp. 1101–1106.

[17] M. Głąbowski and M. Stasiak, "Generalised model of the limited-availability group with finite source population," in *Proceedings of 2nd International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs)*, D. Kouvatsos, Ed. Ilkley, West Yorkshire, UK: Networks UK, Jun. 2004, pp. 40/1–40/10.

[18] M. Głąbowski, "Modelling of state-dependent multi-rate systems carrying BPP traffic," *Annals of Telecommunications*, vol. 63, no. 7-8, pp. 393–407, Aug. 2008.

[19] M. Głąbowski, S. Hanczewski, M. Stasiak, and J. Weissenberg, "Modeling erlang's ideal grading with multi-rate BPP traffic," *Mathematical Problems in Engineering*, vol. 2012, Article ID 456910, 2012.

[20] M. Głąbowski, M. Sobieraj, and M. Stasiak, "Analytical model of limited-availability group with multi-service sources," in *Information Systems Architecture and Technology*, A. Grzech, L. Borzemski, J. Świątek, and Z. Wilimowska, Eds. Oficyna Wydawnicza Politechniki Wrocławskiej, 2010, vol. Networks and Networks' Services, pp. 187–197.

[21] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Convolution algorithm for state-passage probabilities calculation in limited-availability group," in *Proceedings of Fourth Advanced International Conference on Telecommunications 2008*, Athens, Jun. 2008, pp. 215–220.

[22] M. Stasiak, "Blocking probability in a limited-availability group carrying mixture of different multichannel traffic streams," *Annales des Télécommunications*, vol. 48, no. 1-2, pp. 71–76, 1993.

[23] J. Kaufman, "Blocking in a shared resource environment," *IEEE Transactions on Communications*, vol. 29, no. 10, pp. 1474–1481, 1981.

[24] J. Roberts, "A service system with heterogeneous user requirements — application to multi-service telecommunications systems," in *Proceedings of Performance of Data Communications Systems and their Applications*, G. Pujolle, Ed. Amsterdam: North Holland, 1981, pp. 423–431.

[25] M. Głąbowski and M. Stasiak, "Point-to-point blocking probability in switching networks with reservation," *Annales des Télécommunications*, vol. 57, no. 7–8, pp. 798–831, 2002.

[26] M. Głąbowski, A. Kaliszan, and M. Stasiak, "Modeling product-form state-dependent systems with bpp traffic," *Journal of Performance Evaluation*, vol. 67, pp. 174–197, Mar. 2010.

[27] M. Głąbowski, M. Stasiak, and J. Weissenberg, "Properties of recurrent equations for the full-availability group with bpp traffic," *Mathematical Problems in Engineering*, vol. 2012, p. 17, 2012, article ID 547909.

[28] J. Tyszer, *Object-Oriented Computer Simulation Of Discrete-Event Systems*. Kluwer Academic Publishers Group, May 1999.

[29] M. Głąbowski and M. Sobieraj, "Point-to-group blocking probability in switching networks with threshold mechanisms," in *Proceedings of the Fifth Advanced International Conference on Telecommunications*. Venezia: IEEE Computer Society, 2009, pp. 95–100.

Shared Narratives as a New Interactive Medium: CrossTale as a prototype for Collaborative Storytelling

Joaquim Colás, Alan Tapscott, Ayman Moghnieh, and Josep Blat
Universitat Pompeu Fabra
C/Tanger 122-140, E-08018 Barcelona, Spain
{Joaquim.colas, alan.tapscott, ayman.moghnie, josep.blat} @upf.edu

***Abstract*—Through ages, storytelling has been used as one of the main ways for sharing knowledge. We envision the use of shared narrative spaces as a new kind of media that empowers the collaborative creation of vast narrative worlds. We identified existing information systems related to storytelling, and evaluated how they support multi-authored non-linear narratives. A pilot experiment was conducted to understand the user interaction model with shared narratives more profoundly, and we extracted the main interaction factors observed: the different modes of interaction performed towards the informative space, the exploration of a non-linear medium through linear storylines, and the preservation of literary consistency. This model was later transduced into a set of design implications for collaborative narrative systems, which were used as a premise for designing a prototypical tool called CrossTale. Finally, we conducted two experiments to evaluate CrossTale's interaction model and user experience. We discuss how the results of these experiments show that shared narratives have the potential of becoming a distinct type of interactive medium supporting a new genre of user experience.**

Keywords-shared narratives; information systems; user experience; multi-authored composition; storylines; consistency.

I. Introduction

The present article is an extended version of the work presented in CrossTale: Shared Narratives as a New Interactive Medium [1], which examined the use of shared narrative spaces as a collaborative medium through a prototype tool named CrossTale. In this article, we deepen on our understanding of the interaction aspects of shared narrative spaces and their implication on designing systems supporting multi-authored storytelling; and also present a complementary research on consistency preservation through the development of new experiments with CrossTale.

Traditionally, storytelling (from mythological parables through literature classics to modern literary fiction best-sellers) has been associated with the oral and written media, the first two channels of information transmission to appear in human history. Since then, different models of cultural expression had appeared, and those modalities had taken profit of technical advances giving birth to the main contemporary narrative vehicles such as novels, cinema, TV series or comic-books. All those kinds of narrative mediums share the trait of linearity, which suits the temporal causality of classic narratives. In spite of that, several experiments about experiencing narratives in a non-linear way were done (e.g., Moholy-Nagy "total theater" [2] and Borges' tales [3]). With the apparition of digital media, new opportunities arise for creating and experiencing narratives in new ways.

Many contemporary works focus on understanding and modeling storytelling as an interactive experience. Mehan's Talespin [4] is a pioneering approach for automatically generating stories from atomic parts, and is an instigator of a larger body of research focusing on computer-generated narratives. On the other hand, other works studied narratives from an HCI perspective, placing user interaction at the center: Brenda Laurel's work on interactive fictions, impacting HCI as a discipline by underscoring the properties human interaction with information [5]; and Chris Crawford's work on interactive storytelling [6], which addresses aspects of game design. A wide range of actual works focus on models for creating non-linear narratives [7] [8], but to the extent of our knowledge they do not address this task from the perspective of user experience and the study of the user's understanding of non-linearity and narrative consistency.

In terms of media evolution, interactivity represented a change from the mono-directionality of traditional storytelling to a new paradigm better suited for non-linear narratives: through hypertext fiction, conversational adventures or other videogames, users are an active part of the system, performing the exploration required by a non-linear medium. There were also changes concerning the human aspects of media authoring and experiencing. The "fanzine" movement (magazines created by fiction followers who create non-canonical stories) gave way to the apparition of internet communities of novel, movie and TV-series fans that gathered at forums, shared their stories, and catalogued their fiction worlds using tools as wikis. The professional production of fiction also benefited from technological advances, and software for the development of commercial fiction has made appearance, with some examples, as Celtx [9], operating in the cloud as a collaborative tool.

We use the term shared narrative space to address the informative spaces concerning vast narratives created, developed, and maintained through the collaboration of multiple authors. A vast narrative space can be defined as a set of narrative information units organically to form a non-linear story. Those units can have a wide range of granularity depending on the nature of the narrative (e.g., an issue of a comic-book in the informative space of the Marvel Universe, a book in Terry Pratchett's Discworld literary saga, or a chapter in Borges' The Garden of Forking Paths). It is a ludic and cultural medium of expression and communication. As a narrative, it is composed of a story and a discourse (storytelling). The story consists of a setting in time and space, characters, and events (or plots). It is usually thematically unified and logically coherent. Its elements are connected through cause and effect relations, thus temporal order is meaningful [10].

This non-linear medium is comparable to the real-life development of events: multiple stories are happening at the same time, and each can be told from different viewpoints. This points towards the suitability of non-linear narratives not only in developing fiction, but also as a way of sharing information like in online networks (e.g., forums, chats, and communities of creators). Theoretically, the content of social networks could be considered a narrative based on the sequential groupings of threads as scenes. Each forum thread could be regarded as one linear development inside a bigger story, and parts of the same thread could belong to different developments as a cause of this inter-relation. However, the relations between threads are usually vague or inexistent, and there is a need for a global connection between them to provide thematic unification and overall coherence.

Our purpose is to define the adequate system concepts and design to represent and interact with non-linear narratives. Therefore, we developed two empirical experiments with paper-based and digitally-implemented prototypes to extract and understand the user's mental model of interaction with a narrative space, as a basis for the development of modern interactive systems for narratives.

This paper is structured as follows. First, we give an overview of some relevant works related to shared narratives and point out their common concerns about accessing complex information structures and preserving its consistency. Next, we present six major types of information systems related to storytelling, and evaluate their support for shared narratives as a medium for content generation, collaboration and communication. Then, we illustrate a pilot experiment conducted to extract the user model of interacting with a shared narrative space. This model is explained distinguishing its major interaction aspects: First, we detail the different approaches taken by the users interacting with the narrative information depending on their reading or creative tasks; secondly, we describe the "time-space-development" mental model users follow when comprehending the informative space, and how they search for character and plot relations in order to establish linear reading paths or "storylines"; finally, we expose how readers request consistency to understand the narrative. In the succeeding section, the observed interaction factors are transduced into design implications for informing the design of multi-authored narrative systems, and we present CrossTale as a prototype based on these design suggestions. The next section describes CrossTale user evaluation showing the feasibility of supporting new elaborated user experiences with shared narratives. We then discuss how our results deepen our understanding of the characteristics of the interaction with shared narratives: we consolidate our space-time-storyline paradigm for exploring and contributing to narrative spaces, and point to how implemented rules can contribute to maintaining consistency, explaining how this affects the user experience. In our discussion, we argue in support of the potential of shared narrative spaces as new media for collective generation and development of content, communication, and human interaction. Finally, we conclude by summarizing our work and discussing its limitations, and then address their implications on future works.

II. Related Works on Shared Narrative Spaces

There seems to exist a common problematic when approaching the construction of non-linear, interactive narrative structures: the linearity of the narrative classical dramatic structure (that comes from a one-directional medium) seems to be in contradiction with the divergent, open structure that interaction needs as a bi-directional medium; so it is necessary to find a scheme that facilitates a balance between interaction and storytelling. This occurs at a large scale in mediums like videogames, where the interaction is ontological (the user interaction alters what occurs in the narrative world), but it also appears during explorative interaction in non-linear storytelling (where the user interaction consists on selecting which parts of a complex narrative space wants to visit and from which point-of-view), as different configurations of order and content create different narrative structures [3].

Several studies have approached the creation of platforms for the authoring of interactive stories. One notable example is StoryTec [11], a digital authoring tool for interactive multi-media storytelling based on the outcomes of the INSCAPE project [12]. This system employs an editor to model the story as a branching graph, establishing the conditions or triggers that let the user to jump from one state to another. The work of Tanembaum on cognitive hyperlinks for authoring non-linear narratives [13] presented an innovative approach on how to deal with the complex interconnections between different story parts taking place in different moments or places, linking them through suggestive concepts. This idea was an inspiration for the development of the our "Storylines" device, and we will discuss what kind of concepts can be used to establish a relationship between two story fragments as part of our work.

However, as far as we are concerned, many of those systems were conceived and designed as single authoring tools (although they can be used by multiple authors to create a single narrative). The end users do not have to understand the narrative structure underlying their interactive reading experience. Multi-authoring requires those structures to be easy to understand and manipulate (expanding, altering, etc.), and has also particularities which should by addressed by specifically designed mechanisms.

About those specificities on the side of multi-authored experiences, in their work about vast narratives [14] Harrigan et al. provide a large compilation of different scenarios of shared narrative spaces, as literary and television fiction franchises, games, or creator's communities. Harrigan's book exposes the problematic arising during the creation and maintenance of those vast narratives, and how different systems propose different approaches for those problems: the difficulty of making accessible, for both creators and audience, those large amounts of inter-dependent stories in an understandable way; and how to deal with the consistency problems that can appear when different authors participate in the same narrative space.

The work of Y. Cao et al. [7] proposes an interesting approach to a system for collaboratively generating non-linear multimedia stories. It employs the traditional approach of modeling non-linear storytelling as a node tree, but adding the use of story templates. One of those templates is based on Campbell's "heroes journey" stages [15], seeking to ensure the narrative quality of the output. This work also puts emphasis on describing the different kinds of roles that users take in an on-line narrative generation platform. The need of a template for constructing stories and the attention to the roles show how difficult is to generate a navigable narrative structure and maintain its coherence when the narrative space is non-linear and shared between different authors.

Del Fabro et al. [16] approached the theme from a different perspective in their work about real-life events summarization. This work proposes a system to automatically generate the summary of a public event as seen through the large quantities of participants that uploaded their videos, photos and comments to the Internet, and specifically the social networks. Although the output of this system is linear and automatically generated (it cannot be considered an "authoring" tool), the depiction of real-life events from simultaneous and interconnected points of view is an application of non-linear collaborative storytelling that we consider in our discussion section. Del Fabro's system actually has to face the same problems that most multi-authored (or in this case, multi-source) narratives seem to face: multiple points-of view are difficult to locate in a shared narrative space. It is difficult to spatially and temporally locate each single event in relation with the rest of the narrative, and also consistency problems arise.

In conclusion, from our point of view most shared narrative spaces face a similar problematic inherent to their essential nature. First, non-linear storytelling is an oxymoron: the non-linearity makes difficult for both authors and readers to organize and understand information that is subject to the narrative law of cause-and-effect. The "reading order" of the events is central to this understanding, and a non-linear space implies that this order is not defined. Second, the problem of accessing and understanding a complex space of interrelated information also contributes to the difficulty of maintaining narrative consistency when multiple authors expand the same informative space.

III. Contemporary interactive systems for storytelling and narratives

In [3], Ryan proposed a classification of interactive narrative types based on the nature of the user participation: users can either experience the narrative acting as an internal character of the story, or as an external agent; they can either alter the ontology of the narrative through interaction, ontologically alter the narrative world through interaction, or explore the narrative without inducing any change. This classification provides a framework to analyze and characterize contemporary systems for interacting with narrative by reflecting on how the user experience is contributing to the narrative, and how the narrative is influencing the user experience.

We have identified six major types of information systems directly related to interactive narratives: The first type are adventure books, which comprise a tale where the reader follows a character and makes choices that lead the story towards distinct developments; the second is tabletop role-playing games (or RPGs), in which the player creates a character and its story, and then devises the character's actions according to a set of rules; adventure videogames are the third type, and they put the player in the role of a character that resolves puzzles in order to advance in the story; the fourth type is role-playing videogames, where the player makes navigation decisions to reach one of several possible endings; the fifth type is Forum or chat-based RPGs, where players collaboratively create a story (usually with a few rules of engagement); the sixth and last type is web communities of fiction writers (fan-fiction), that create stories in the same fiction world, but not always collaboratively. A high number of fan-made wikis can be found on the web, compiling formation about events, characters, and places concerning those worlds. Harrigan gives a wide overview of the complications of maintaining these vast narrative spaces, and how the different systems or communities address them [14]. These systems are described in Table 1 according to Ryan's framework.

TABLE I. CONTEMPORARY SYSTEMS OF INTERACTIVE NARRATIVES

System	Example	Author /reader role	Main role	Author interaction	Reader interaction
Adventure Book	Choose your own adventure	Separated	Reader	-	External Ontological
Tabletop RPG	Dungeons & Dragons	Mixed	Author	Internal, Ontological	-
Adventure videogame	Monkey Island	Separated	Reader	-	Internal Exploratory
RPG-Videogame	Baldur's Gate	Separated	Reader	-	Internal Ontological
Forum / chat RPG	Aelyria.com	Mixed	Author	Internal, Ontological	External Exploratory
Fan-Fiction community	Fanfiction.net	Mixed	-	External, Ontological	External Exploratory

Seeing the particularities of the informative structure of narratives, we point at differences between existing systems and interactive storytelling. In particular, none of these types of systems entirely supports shared narratives as a medium of social interaction. Three of them (books, adventure videogames and RPG videogames) are unidirectional mediums, created by authors and consumed by other people as readers. They support a varied degree of interaction with the content, but they do not allow users to contribute. Tabletop RPGs and forum or chat RPGs, allows user-generated content to be added to ongoing discussions, which together do not constitute a coherent story that can later be consumed as part of the user experience. Only fan-fiction Internet communities fully support both the addition of user-generated content and its consumption as part of the user experience. But the lack of collaboration and cooperation between contributors tends to divide the narrative space into distinct and incoherent flows of events, which only share the original work as a point of reference, resulting in independent narratives.

The case of fan-fiction communities is the major exponent of a multi-authored narrative system in which usually no one acts as both reader and author to the same shared narrative, but each participant is only the author of his narrative sub-space and reader of others. A similar handicap exists in forum RPGs, where each contribution is by force situated directly after the previous one, and is the only possible type of contributions.

In conclusion, none of these systems fully empowers participants to contribute efficiently to the shared narrative space, nor to collaboratively organize and maintain its overall structure and narrative coherence. The aforementioned problems of finding some "order" to access and understand a complex space of interrelated information and maintaining its consistency are approached, mostly, by having different independent narrative spaces (so not sharing one single narrative space) or simply by restricting the contributions to that space. Therefore, there is still a need for supporting the users' ability to understand and navigate the space, allowing the narrative to grow in an organic way, and extending its contents from any desired point in the narrative flow.

Figure 1. The settings and development of the pilot experiment

IV. UNDERSTANDING INTERACTION FACTORS

The first experiment, "Story on a Wall" (Fig. 1), was an observation experience designed to allow users to freely create both a narrative and the rules that operate it. 20 subjects (university students) were provided with paper templates as a frame to create scenes and a set of elements (fairytale characters and objects) that they could use. Scenes were crafted creating an image (drawing and pasting elements) and a short literary text. The story was developed on a large glass wall posting the scenes and drawing arrows. Each arrow connected two temporally consequent scenes, but aside of this the narrative meaning of the relations (i.e. thematic or point-of-view connections) was left for the subjects to interpret. The subjects proceeded one by one to read the story on the glass, and then modify or expand it by creating new scenes, posting them in the wall, and drawing connections. Although subjects were encouraged to expand the story, there were no constraints on what the subject was allowed to do in the narrative space: they could rearrange, modify or eliminate previous scenes. Observations were made during this process, and the subjects were later asked to fill a questionnaire of 18 questions. The questionnaire evaluated the story comprehension and consistency as perceived by the subjects (asking them to rate story comprehensibility and coherence in a scale), and inquired about the reading or navigation paradigm that they used (how had they selected the relations between scenes, which narrative elements and concepts they followed throughout the story, and how the reading order was decided). It also asked about their contributions (number, content, location, etc.), and if they added scenes to the narratives or contended in modifying existing ones.

The kind of interaction performed is external, as the users do not assume the role of any particular character. It is also ontological during the creation, and exploratory during the reading. The analysis of the resulting story and the questionnaire answers revealed several aspects about the nature of the user comprehension and interaction with the shared narrative space. From these results we will derivate, in the next part of our work, a set of design implications to develop and study interactive narrative systems. We can resume the interaction factors observed as follows:

A. *Three interaction modes*

The subjects' interaction with the narrative space shows that at least three different views for three different purposes are needed for a multi-modal interaction with narratives. These views are illustrated in Fig. 2.

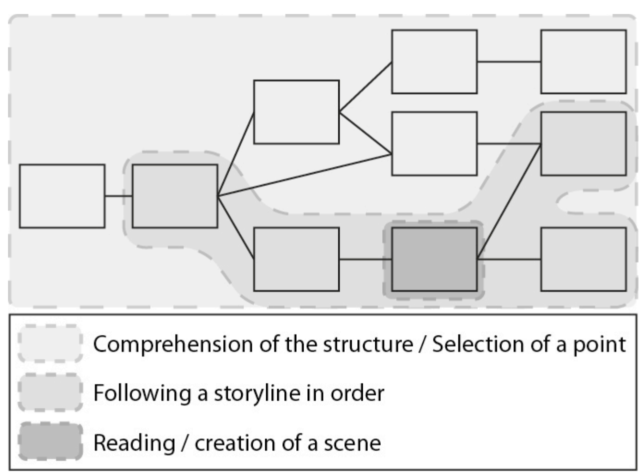


Figure 2. Viewpoints related to the interactions with the narrative space

A global view of the space is used to approach and comprehend the whole narrative space and its structure, as well as when selecting a point in time and place to add a new scene. Then, a "zoom-in" view is used for viewing a scene inside a storyline and understanding the other storylines related to it. Finally, the composition view allows users to create and edit scenes focusing on the crafting of a single scene.

B. *Navigation through storylines*

The results revealed that subjects project a "time-space-development" logic on the narrative. Although the story in this pilot experiment was mainly developed as a classic "choose your own adventure" narrative (this is, a set of branching paths), when reading and expanding the story, subjects considered higher level relations between those branches (e.g., the relations between events that subjects considered taking place simultaneously in different, unrelated spaces). This can be resumed as the story being mentally situated on a space with a temporal and causal logic, represented in two axes: the temporal relationship between the scenes (time), and the places where these scenes occur (space).

All subjects followed linear sequences (which we call storylines) for reading, being a linear sequence of connected scenes that track the development of a specific character or plot. The relations between scenes were only indicated through links, so the concepts that characterized the relation between the scenes were mainly determined by the readers' perception. When asking the subjects, all of them coincided on the scenes being related by following a character or a plot relationship, which also indicates that contributors, although maybe unconsciously, established the relations following that paradigm. 14 out of 20 followed those storylines throughout the narrative space from the first scene to a finishing one before backtracking (the others abandoning some storylines and jumping to new ones arbitrarily). In addition, 12 of them followed character developments, and 10 of them followed plot relationships.

Understanding how users navigate the narrative space leads us to consider a visualization that copes with this "time-space-development" logic and facilitates the creation and finding of storylines as the main way of explore and contribute to the space, consequently facilitating the user interaction.

C. *Preservation of literary consistency*

The results also show that the generated narrative space is unitary, coherent, and with a limited divergence. It is unitary in the sense that all the scenes are interrelated and are part of the same story. In fact, the divergence of the narrative space away from the central topic is limited: subjects found it easier and socially proactive to expand existing storylines instead of creating new ones. This notion of unity is directly derived from the fact that the entire story is predefined and all the storylines are happening simultaneously in the same time stream. This raises consistency issues in the literary fabrics of the narrative, which users thrive to treat by re-ordering scenes or inserting new ones.

The literary consistency of the narrative, defined as its elements and plots being in agreement/non-contradictory, is considered fundamental for understanding the story. Most subjects stated a dislike towards the notion of conflicting storylines, being consistency one of the main concerns when modifying/ adding scenes to the narrative space: maintaining consistency in the growing narrative was one of the main motivations for 8 out of 15 contributions, and 5 subjects used their contribution only for correcting consistency issues. Also, the totality of the changes made to previous contributions was for the sake of consistency. In the end of the experiment, only 5% of the scenes were considered inconsistent with the rest of the narrative.

Consistency therefore seems to have a great degree of influence on the resulting user experience as exposed by the result analysis. The subject-perceived consistency level (Figure 3) is relatively stable if slowly decaying, even though it tends to stay on the high portion of the scale. This may be happening due to the accumulative complexity of the story and the growing cost for achieving consistent scenes. For this reason, the system should implement mechanisms for helping to preserve literary consistency without restricting the non-linearity of the narrative.

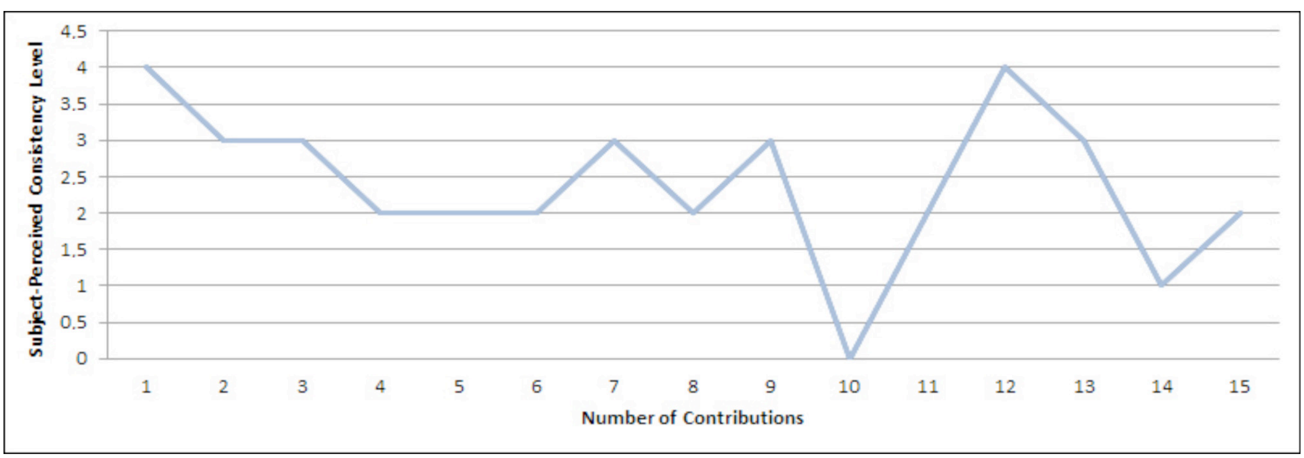


Figure 3. Evolution of the user-perceived consistency during the pilot experiment

V. PROTOTYPING SHARED NARRATIVE SPACES

The resulting interaction factors extracted from the first experiment were transduced as a set of implications (Table II) for the design of information systems that support interacting with shared narrative spaces. We developed a prototype named CrossTale based on those design implications to reproduce the user experience according to them.

TABLE II. DESIGN IMPLICATIONS EXTRACTED FROM THE INTERACTION MODEL

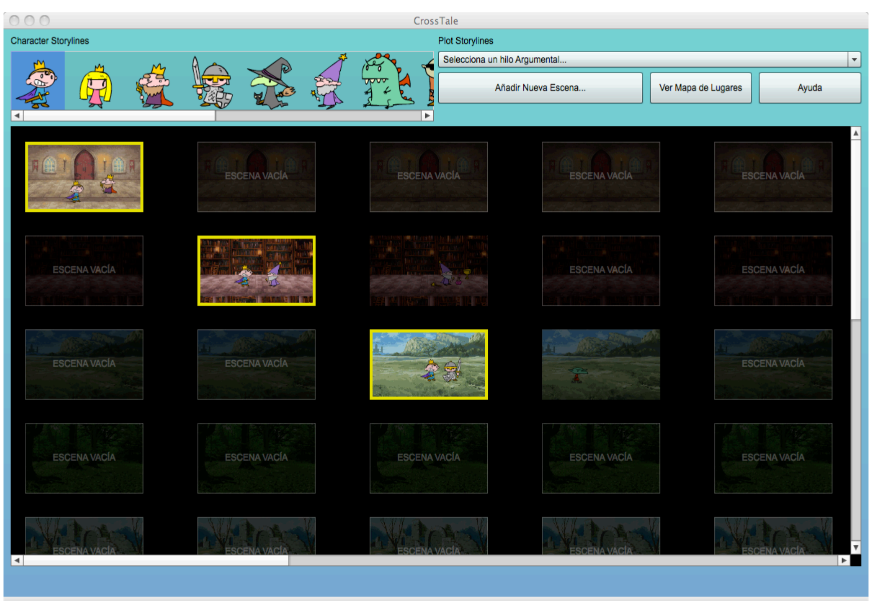
Interaction Factors	Design Implications
Projection of a logic based on time, space and developments.	Organization of the informative space based on time and space axis.
Reading by following linear sequences about a character or a plot.	Navigation through suggested plot and character storylines.
Unitary and coherent narratives.	Mechanisms for preserving congruence.
Global viewpoint for comprehending the whole story.	One interface mode for a global view of the informative space.
Reading a storyline through a zoom-in viewpoint.	One interface mode for following storylines.
Focusing on a single scene for creating and editing.	Independent interface mode for scene edition.

A. Three interface modes

CrossTale implements three interface modes corresponding to the three interaction modes defined previously. The global view (Fig. 4 a) lets users explore the whole narrative space, and has two differentiated contexts: the main context shows the narrative space visualized as a grid, with the axes representing time (from left to right, the scenes are ordered in temporal order) and space (each row representing a different place where the story takes place, as the kings castle or the enchanted forest). The upper context in this view allows the exploration through storylines (explained in the next section) by selecting characters and plots. When a character or a plot is selected, all the scenes belonging to that character/plot storyline appear highlighted forming a reading path. Selecting a scene changes the interface into the reading view (Fig. 4 b) in which the scene is maximized for reading. In this view, the user can also navigate back and forward by the current storyline. Finally, by selecting an empty frame in the global view, the user accesses the creation view (Fig. 4 c). In this view, s/he can create a scene selecting characters and objects from the right-side menu and by arranging them through drag-and-drop. The user also can introduce a title and the literary text describing the scene, and indicate to which plot storylines lines is related the current scene (allowing the definition of new ones). This context also implements the "consistency preservation" rules, which are further described in this section.

B. The storylines device

In traditional linear storytelling, a single author creates story (the content) and discourse (how is it presented). But in non-linear storytelling, the reader decides this partially (alters part of the discourse) when he chooses to follow some part or another of the narrative space in a desired order.



(a) The global view

(b) The reading view

(c) The creation view

Figure 4. Three interface modes of the CrossTale prototype

As we stated in our findings of the "Story on a Wall" experiment, subjects observing the "branching-structured" narrative established higher-level conceptual relations between its parts to find their own reading order, projecting temporal and spatial logic between all the elements. The concepts relating one scene with another were perceived as maintaining a character or plot relationship.

Following those pilot experiment discoveries we developed the "storylines" paradigm to allow the ordered exploration and development of the narrative space. The aforementioned grid of scenes, arranged in spatial and temporal axis, deconstructs the "scene tree" and presents all the information in a way where the user can easily recognize the spatial and temporal relation between all scenes. The paths of reading or "storylines" are provided by the interface when the user selects one character or plot to follow, highlighting a temporally ordered sequence of logically related scenes. This way, each scene can belong to an unlimited number of storylines based on what characters take part in it and to what plots the scene events are related.

Concerning the nature of those relations, character reactions are the easier to understand and work with. A character storyline is simply composed by all the scenes where a character appears, ordered temporally. Usually those scenes contain temporal and cause-to-effect relations between them, describing the story of the character. This type of plotlines evades easily the problem of non-linearity, because following a character implies following a linear development: the character, being at one place at a time, experiences the story partially and linearly, similar to the users' linear perception of time development.

The notion of plot is more abstract than the preceding, and also more difficult to implement. As authoring tool, in CrossTale plots are decided by subjects, so the authors decide to assign different created scenes to the same plot, considering that they are describing a thematically-unified "sub-story" inside all the story world (e.g., in our fairy-tale, where lots of different actions are occurring, the plotline about the kidnap of the princess and the different factions who are searching for her). Not all scenes are linear in time, because the same plot can follow different characters, jump in time, and/or develop in two or more places at the same time. The plot storyline is presented following the temporal sequence, and scenes taking place simultaneously are shown intercalated.

C. *Consistency preservation rules*

CrossTale implements a set of constraints that can be activated in order to maintain consistency during the scene composition process. These constraints were designed to prevent subjects from adding scenes that somehow disrupted the sense of space or time. For instance one character could not be present at different places simultaneously or no one could travel further than one location away between two timeframes (locations were connected in an arbitrary way and this information was transmitted to subjects by providing a map).

These rules were designed after the most common modifications performed by subjects during the previous experiment, attempting to anticipate potential displeasing outcomes and preventing them by blocking certain actions. A text dialog will be also displayed to inform the subject the reason why he cannot commit the change he is trying to (e.g., "The Princess cannot appear here because she is in another place at this moment"). If these mechanisms perform optimally, the whole story might experiment a global consistency increment and users might chose to concentrate on providing another kind of content.

VI. EVALUATING CROSSTALE

Our CrossTale system is a tool designed and implemented following the design implications extracted from the mental model of interaction observed during the pilot experiment. Therefore, we tested the tool through new experiments in order to assess the adequacy of this model in the development of systems to create and interact with non-linear, multi-authored narratives.

The objective of these experiences was to evaluate the adequacy of use of the different interaction modes, the paradigm of narrative space representation and exploration through the storylines device, and the usefulness and impact of the consistency preservation rules. With those aspects interrelated and being part of a complex system, we developed a couple of experiences to better understand the repercussions of those interaction factors. The first experience served us to evaluate the interaction mechanisms in terms of adequacy of design and user experience, with users performing creative tasks with the tool. Observations of the results of this creative task were used to analyze the nature of the created narratives and how users developed their storylines using the provided tool devices. Finally, a second experiment was conducted in order to test the impact of the consistency preservation rules on the produced narrative and the user experience. The description of the conditions and the result analysis of the evaluation experiments are presented next.

A. *Evaluating the interaction mechanisms*

The main experiment with the prototype, in order to test the adequacy of its design and the user experience, consisted of creating a narrative in a similar way to the pilot experiment. A total of 15 subjects (undergraduate students in media studies) were enlisted, and asked to freely use the interface to read and create a shared narrative with their own contributions. Each subject was briefly introduced to the interface controls, and then given an unlimited time to interact and the freedom to add as many scenes as wanted. Then, the subject executed eight interaction tasks provided by the evaluation team, and observations were made. Afterwards, each subject answered a questionnaire to rate the experience on a Likert scale, and evaluate the suitability of the design for reading and contributing and the overall user experience.

The results of the task-driven evaluation are summarized in Table III. It describes how many subjects employed each interface view for each task. The results show that 11 out of 15 subjects performed all tasks easily, and the remaining 4 subjects successfully performed 6 out of 8 tasks. The subjects used the global view and/or the reading view to identify and comprehend the narrative elements. Similarly to the first experiment, some subjects concentrated on characters while others on plots, but everyone used one of these two paradigms for finding storylines and navigating the narrative space. During the contribution task all the contributors also used the creation view to compose new scenes, but this view was never accessed for performing the identification tasks. These results indicate that the design supports the modes of interaction identified in the first experiment, and that these modes dispose of adequate functionalities. However, most subjects prefer having more information about the context of scenes while reading them. This means that the dissociation between the global and reading views could be revisited.

TABLE III. RESULTS OF THE DESIGN ADEQUACY EVALUATION

Task NB	Task	Correctly executed	Global View	Reading View	Both Views	Navigating with Storylines
1	**Identify the beginning scenes**	15	13	13	11	11
2	**Identify story end scenes**	15	13	9	7	7
3	**Identify main characters**	15	15	15	15	15
4	**Identify important places**	15	14	14	13	13
5	**Identify simultaneous scenes**	11	13	4	4	4
6	**Identify scenes in the same location**	12	14	2	2	2
7	**Approximate the duration**	15	13	13	11	10
8	**Find any inconsistency**	15	8	12	7	8
9	**Contribute (optional)**	13	13	6	6	5

Table IV shows the evaluation results of the user experience. All subjects appreciated the experience of interacting with narratives through CrossTale. In particular, they found that CrossTale supports reading a non-linear narrative (4.33/5), contributing to it (4.77/5), and finding and correcting inconsistencies (3.92/5).

TABLE IV. RESULTS OF THE USER EXPERIENCE EVALUATION

Question	Average Score
Overall experience	3.93 / 5
Found the system entertaining	4.33 / 5
Design makes reading easy	3.93 / 5
Design helps to maintaining consistency	3.92 / 5
Design facilitates contributing	4.77 / 5

Using Ryan's framework for the classification of interactivity with narrative systems, we can say that the users of CrossTale performed an external interaction during the whole experience, as they took on the role of agents external to the story, and read and contributed in it from outside the fiction world. This interaction is exploratory while reading, in the sense that the readers choose between storylines to follow but the reading itself does not change nor affect the structure of the narrative space. Finally an ontological participation is performed when the user takes the role of author and expands or alters the narrative world.

B. Use of storylines in the composed narrative

We present a general analysis of the resulting narrative compared with the narrative created in the pilot experiment. Figure 5 shows the structure of the resulting narrative. Users do not link scenes directly in CrossTale, but relations are established by the share of common characters/plotlines. According to this, we present the basic scheme with the time-consecutive scenes sharing characters connected (so they can be followed through character storylines), and each version of the figure showing what of those scenes belong to each of the four plot storylines the subjects used.

From the initial set of 8 starting scenes given to the first subject the resulting narrative ended up having 35 scenes, so 27 scenes were added by the subjects, with an average of 1.8 scenes per subject. This is slightly superior to the 1.4 scenes per subject added in the pilot experiment, which could indicate that the interface makes the addition of the scene relatively easy. Following the analysis of the character plotlines, the story has 6 bifurcating scenes (scenes from which different character plotlines going to different new scenes emerge) compared to the 9 bifurcations of the pilot experiment narrative, and 3 ending/unconcluded scenes compared to the 4 in the pilot. As a general conclusion, the properties of the generated narrative being unitary and with limited divergence are preserved, if not emphasized, through the use of CrossTale.

Analyzing the use of storylines, plot storylines generally do not create completely new reading paths but can be trailed following the transitions of the character storylines, so they don't create a new level of divergence. We can observe that the first storyline (already given in the starting narrative) starts temporally at two places at the same time, but these two parts can be read following the storylines of two main characters (the princess and the prince). The other three storylines used in the story are practically linear and follow the scenes where some concrete character appears. The initial setting of the experiment provided three storylines. Subjects only added two storylines, one of them being the linear continuation of a previous one (thus being presented as one single storyline in the figure). To sum up, most of the users included their created scenes in one existing plot storyline, but the creation of new plot storylines was very low.

As a conclusion according with this observations and the analysis done in the previous section, plot storylines are

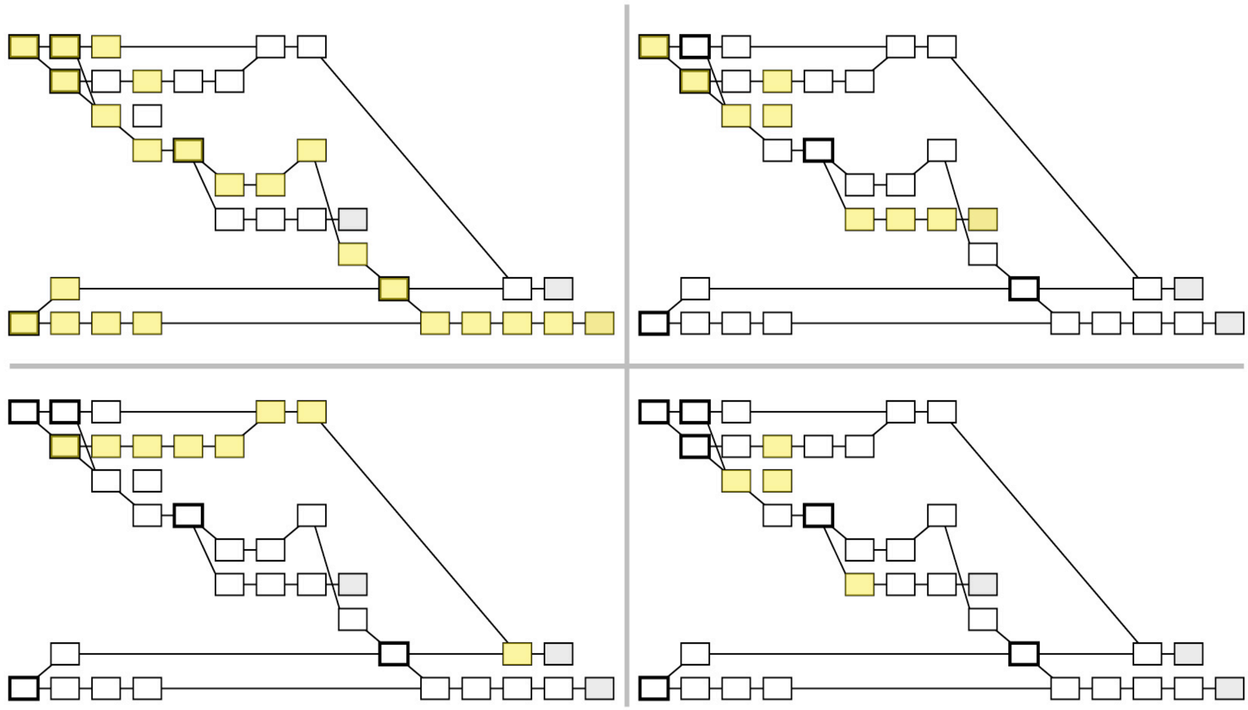


Figure 5. Plot storylines in the narrative produced with CrossTale

regarded as much useful as character storylines when reading, but they are largely more abstract, thus difficult to use when creating. While using a character implies the unconscious continuation of his/her storyline in the narrative space, the use of plot storylines needs to be deliberately planned by the user. Future investigations should focus on understanding deeply which concepts can relate scenes forming storylines, how to make easier to the creators find this relations (which can suggest new storylines), and consider if those relations should be automatically created (as in the case of the character storylines).

C. Evaluating the consistency preservation rules

In order to evaluate the consistency separately, we conducted an isolated experiment using two groups of ten people. The control group (group A) used CrossTale with the consistency constraints enabled, and the second group (group B) had these constraints disabled. Apart from this, both groups were exposed to the same experimental conditions: they were introduced to the use of the tool, and they contributed to the narrative one subject at a time, starting with the same set of initial scenes. The focus of the experiment was to evaluate the impact of the implemented system of rules in the congruence preservation and the user experience.

The tests performed in the end of the experiments revealed that users of CrossTale with consistency preservation rules perceived a slightly higher consistency level than users without constraints (3.8/5 vs. 3.4/5). The most remarkable observation was that the perceived consistency level seemed to decay more quickly over each contribution without the usage of consistency constraints (Figure 6). An optimistic interpretation of this phenomenon could be that enforcing a certain notion of time and space logic through the scenes (which was the purpose of the constraints) tends to produce more consistent results. Adding the constraints also seems to have affected the user experience of users who felt limited all the time (as seen during the video codification, where they keep complaining almost every time a constraint block pops up).

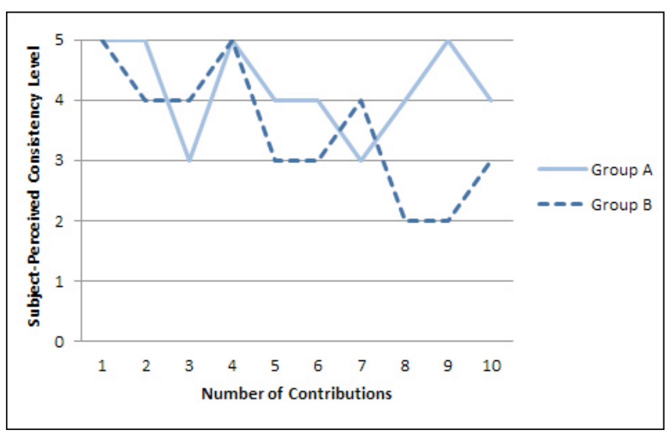


Figure 6. Perceived consistency level in the second experiment

Therefore, we expose an interesting situation; constraining scene composition to a specific outcome range way may lead to more consistent results while hampering the authoring process and negatively affecting its user experience. One possible answer to balancing the exposed tradeoff might be switching the nature of the introduced trigger from an enforcing one to a less intrusive recommending system. This way we can inform contributing subjects about the potential incoherence of their creation, maybe preventing them from introducing conflicting scenes that later users would want to correct.

D. Final comments on the CrossTale evaluation

The results of this experiments show that the concepts and design of CrossTale, as a prototype for interacting with narratives, are highly appreciated by the subjects. However, they also point out several issues that need to be addressed in future versions. In particular, the level of context awareness when users interact with one single scene in the informative space, the use of storylines when authoring narratives, and the degree of creative intrusion produced by the consistency constraints. We cannot neglect the repercussions of each one of those areas into the other: the ability to access the information and to mentally locate it in the whole narrative space, or the difficulties integrating a scene in the storyline structure, can condition the consistency level perceived or produced by the user, disregarding the actual consistency level of the complete informative space.

In addition, social interaction between different authors remains indirect: users cannot communicate directly and the authors' profiles and their contributions are not discernible in the current design. Future versions can include more support for this aspect and study its effects on the user experience and collaboration.

VII. Discussion

The nature of shared narratives presents several challenges over how the inherent information is constructed, presented, and accessed. In a sense, non-linear interactive storytelling has always faced challenges for having to reconcile the sequential nature of narratives with the reader ability to explore between different threads of reading (the paradox of coping "storytelling" with the "non-linear"). In this work we provided a first grounding basis for addressing these challenges and developing shared narratives as a new kind of media. Our research is a first step for consolidating a standardized system for sharing and collaboratively constructing narratives, given we extracted, understood, and evaluated the user mental model associated with this interaction.

Several use cases can be provided exemplifying our vision of shared narratives as a new kind of media. The focus of our work can be easily illustrated using the case in which our current study is based: the collaborative generation of fiction worlds. The first chapters of this work described how, since the apparition of the World Wide Web, internet communities of literature amateur practitioners have dedicated their conjoined efforts to develop and catalogue fiction sharing common narrative spaces, using tools as forums or wikis to organize the large amounts of interrelated information produced. The main issues those communities have to face are the difficulty to navigate and comprehend grand quantities of interrelated information, and maintaining coherency when lots of authors expand and/or modify those interdependent sets of data (e.g., maintaining the coherency of the biography of a character that appears in several stories created by different authors). With tools implementing the interaction and data-management principles proposed in CrossTale, we would be able to develop systems that largely facilitate the creative process, ensuring that the collaborative efforts work together and empowering creativity rather than endangering it.

Our studies have been developed working within the fiction genre (concretely fantasy tales), with amateur and/or non-professional authors. This setting for the study and the experiments reflected our focus on the emergent online communities of literature and fiction aficionados. But we can discuss how this paradigm of authoring and exploring vast narratives could be applied to other genres or other contexts of use. Of course, this would require following new lines of research in order to study their particularities

As we previously introduced, with the development of the information technologies and the raise of the mass-oriented cultural products, the professional sector of literary creators dedicated to fiction has started to use computerized tools, as Celtx, to develop their creative tasks. Some of these tools try to deal with our studied difficulties: organization and exploration of complex informative spaces, and congruence preservation. Certain among them (as Celtx) have developed the status of on-line, social tools, aiming for a multi-authored model. Assuming that research is conducted about the mental models and the interaction needs of a different standard of users as professional creators are, this scenario could largely benefit from our developments. Tools following the CrossTale paradigm for interacting shared narrative spaces could prove very useful in the field of commercial vast narratives [14], as long-running TV-series or especially complex trans-media fiction franchises, which involve lots of interrelated, multi-authored information.

VIII. Conclusions

In modern literature and fiction worlds, it is common to have multiple stories set in a complex chronology inside a common setting, such as in fiction franchises where narratives are constructed through the contributions of multiple professional authors. Tools based on the CrossTale interaction model would be capable of organizing all this encyclopedic knowledge in a structured narrative space that suits better the temporal, causal, and multi-lineal nature of a narrative, empowering the authors to contribute easily to expand the vast fiction worlds and empowering the readers to explore them naturally. With such tools, narrative spaces grow organically and collaboratively; the proactive role of participants consequently diffuses the mono-directionality of the author/audience relation. In that sense, non-linear interactive narratives can become a new kind of media of its own, suitable for creation, collaboration, information sharing, and learning.

By experimentation, we learnt how users perceive and procreate the narrative space in a unitary and consistent way, how they mentally structure the informative space in terms of time and place, and navigate it following structured sequences of character and plot-related scenes. This model was used as the basis for designing a functional prototype, CrossTale, which was subsequently evaluated with users. These evaluations show the success of the adopted approach in supporting complex interactions with narrative spaces, which assimilate its non-linearity. It provides a validation for further investigations on the potential of shared narratives as new media.

We can summarize the approach used as working in two fronts: the user interaction with the informative space, and the coherency computation of the contained information. Concerning the interaction when creating and exploring a shared narrative space, the critical aspect revealed to be how to provide a mean to obtain meaningful information when exploring a complex net of causal-related scenes or story parts. The use of the "storylines" mechanism is a first successful step towards this objective of providing meaningful "reading paths/orders", but further developments should be made to approach some relevant issues (e.g., the simplicity of suggesting character storylines versus the abstract concept of "plot storylines", which has to be consciously appointed by authors).

About the preservation of consistency during collaboration, our experiments pointed that consistency between all the story elements and scenes is the main conception that readers use to understand the narrative space, and one of the main concerns when expanding the space by adding new scenes. The experiments seem to indicate that implementing computational rules help to raise consistency but can also undermine the user experience. Thus, we have to choose carefully what rules determine the consistency of the informative space, which of them should be only suggested, and which of them should be strictly followed to ensure that consistency.

IX. FUTURE WORK

This work has several limitations inherent to the nature of the experimental settings and the prototype. The pilot experiment was not designed as a strict experience for testing concrete aspects, but was used as an observation to extract information about how participants reacted and interacted with a collaborative narrative space. For this reason, direct comparisons between results of the pilot experiment and the prototype experiments should be taken carefully. While the pilot experiment was conducted in a large wall where all the informative space was present, access to the information using CrossTale is done through a common computer screen and mediated by the designed visualizations. This implies that, although mechanisms as storylines have been proven useful, its effectiveness is conditioned to the existent limitations of the interface design.

Regarding the prototype nature, the visualizations used have functional limitations (e.g., visualizing all related scenes to a selected one). One important limitation was the aforementioned difficulty, when reading, to locate the actual scene in relation with the global set of scenes. Other improvable design issues, mostly usability-related, were identified during the evaluation (e.g., the composition view is not user friendly). Finally, the development of the experiments in a controlled environment does not reflect intimately interactions with shared narratives, nor the collaboration phenomenon (performed in an indirect way through the experiment), as ought to take place online during a greater amount of time.

Some aspects related to narrative composition remained outside this study. In our experiments, the literary traits of the narrative space were somewhat pre-defined, especially the main characters (prince, princess, witch, etc.) and places (tower, castle, woods, etc.). This discouraged users to think about expanding the literary reach of the narrative space with few exceptions. In a collaborative creation processes, as online social role-playing games, some people perform the role of content generators, adding story elements (characters, objects, places, etc.) to the informative space. Such behavior should be further studied in the future.

With this model and prototype as a starting point, our future step consists of addressing some issues detected during the experiments and conducting more focused experiments about the impact of the storylines device in the construction of the story and its resulting consistency, and how a recommendation system could better empower the consistency preservation without disrupting the creative process. Conducting those experiments with a larger number of users interacting and evaluating the resulting narrative during a larger period of time will better cope with a community co-creative process, and would allow us to study the nature of the resulting interaction and narrative structure under those conditions, as well as the potential of shared narrative spaces to empower long-term collaboration. It will also provide large quantities of data relative to the outcome of the continuous interaction of users in a multi-user deployment.

With those experiences we will refine our knowledge of the interaction factors and try to discover new factors that can alter the user experience. We will determine which features are more useful or which ones need improvement, distinguishing between the tasks of creating and reading (e.g., the use of plot storylines, which have been proven useful to read but not very used while creating). We also pretend to discover factors external to the interface that can alter the experience and the behavior of the creators (e.g., starting the experience with a pre-established set of scenes and storylines could alter which features the users will use or how the narrative will grow). Those long-time experiments will also determine how the level of consistency evolves at long term, when large amounts of contributions are made, and compare the outputs of using or not consistency rules. Finally, we will also be able to find relations between those mentioned interaction factors and how they affect the consistency found (e.g., starting from a consistent set of storylines could help to reach a more consistent narrative space after several contributions).

Next steps of this study are in the direction of evaluating which theoretical models of narrative (as those proposed by

Propp [17] and Campbell [15]) could help to structure the informative space and refine our understanding of the needs of creators and readers. Those models would lead us to discover which elements can be used to determine storyline relationships between the story fragments, and which elements can be used to compute the consistency of a creation. With this knowledge we will be also able to develop features for adding user-generated story elements (as characters and places) leading to a complete system for generating and maintaining shared narrative spaces.

REFERENCES

[1] J. Colàs, A. Tappscott, A. Moghnieh, and J. Blat, "CrossTale: Shared Narratives as a New Interactive Medium", MMEDIA, 2012.

[2] L. Moholy-Nagy, "Theater, Circus, Variety" (1924), in Multimedia, from Wagner to Virtual Reality, R. Packer and K. Jordan, Ed. New York: W. W. Norton & Company, 2001, pp. 16-26.

[3] ML. Ryan, "Beyond myth and metaphor: The case of narrative in digital media", Game Studies, 2001.

[4] J. R. Meehan, "TALE-SPIN, An Interactive Program that Writes Stories", in Proceedings of the Fifth International Joint Conference on Artificial Intelligence, 1977.

[5] B. Laurel, "Computers as Theatre", Addison-Wesley, 1991.

[6] C. Crawford, "Chris Crawford on Interactive Storytelling", Berkeley, Calif.: New Riders, 2005.

[7] Y. Cao, R. Klamma, and M. Jarke, "The Hero's Journey – Template-Based Storytelling for Ubiquitous Multimedia Management", Journal of Multimedia, vol. 6, no. 2, Apr. 2011, pp. 156-169, doi: 10.4304/jmm.6.2.156-169.

[8] Y. Cao, R. Klamma, and A. Martini, "Collaborative Storytelling in the Web 2.0", in Proceedings of the First International Workshop on Story-Telling and Educational Games, 2008.

[9] www.celtx.com (June 30, 2013).

[10] S. Chatman, "Story and Discourse: Narrative Structure in Fiction and Film", Cornwell University Press, 1990.

[11] S. Gobel, L. Salvatore, R. Konrad and F. Mehm, "StoryTec: A Digital Storytelling Platform for the Authoring and Experiencing of Interactive and Non-linear Stories", in Interactive Storytelling, 2008, pp. 325-326, doi: 10.1007/978-3-540-89454-4_40.

[12] O. Balet, "INSCAPE: An Authoring Platform for Interactive Storytelling" in Virtual Storytelling. Using Virtual Reality Technologies for Storytelling (ICVS 2007 Proceedings), Marc Cavazza, Stéphane Donikian (Eds.) Berlin: Springer-Verlag, 2007, pp. 176–177.

[13] Tanenbaum, K. Tanenbaum, M. Seif El-Nasr, and M. Hatala, "Authoring Tangible Interactive Narratives using Cognitive Hyperlinks", Proceedings of the Intelligent Narrative Technologies III Workshop, 2010.

[14] P. Harrigan and N. Wardrip-Fruin, "Third Person: Authoring and Exploring Vast Narratives", MIT Press, 2009.

[15] J. Campbell, "The Hero with a Thousand Faces", Princeton University Press, 1968.

[16] M. del Fabro, A. Sobe, and L. Böszörmenyi, "Summarization of Real-Life Events Based on Community-Contributed Content", MMEDIA, 2012.

[17] V. Propp, "Morphology of the Folktale", University of Texas Press, 1968.

Low Complexity Cross-Layer Scheduling and Resource Allocation for VoIP in 3G LTE

Richard Musabe and Hadi Larijani
Dept. Computer, Communication, and Interactive Systems (CCIS)
School of Engineering and Built Environment
Glasgow Caledonian University, Glasgow, Scotland, UK.
E-mails:richard.musabe@gcu.ac.uk, h.larijani@gcu.ac.uk

***Abstract*—3G Long Term Evolution (LTE) is an emerging and promising technology that aims at providing broadband ubiquitous Internet access and improving multimedia services. This is achieved through streamlining the system for packet services, since LTE is an all Internet Protocol(IP) based network. The fact that 3G LTE is a packet based network brings about some improvements in the form of higher bit rates, lower latencies, and a variety of service offerings. However, some technical challenges are expected to arise when voice traffic is transmitted over an LTE network. This has become an interesting area of research and different types of resource management schemes have been developed, which are quite challenging and complex. In this paper, we have projected the voice packet scheduling and resource allocation problem as a constrained optimization problem. Our optimization objective is formulated using channel state information such as, transmission rate at the physical layer as well as the queuing state information like queue length at the MAC layer. We provide the algorithmic implementation of the obtained solution and also investigate the performance, complexity, and fairness of our proposed cross-layer scheduling algorithm under different conditions such as, VoIP delay, packet loss, etc. We compared it with other algorithms in literature such as, proportional fair (PF) and exponential proportional fair (EXP-PF). Based on the numerical and simulation analysis, we found that our proposed algorithm performed better than PF and EXP-PF in most cases.**

Keywords*—LTE, Scheduling Schemes, VoIP, Complexity, Utility Function.

I. INTRODUCTION

In this paper, we investigate the performance, complexity, and fairness of our proposed cross-layer scheduling algorithm for Voice over Internet Protocol (VoIP) in 3G LTE. This work is an extension of the analysis done in [1]. 3G LTE was identified by the third generation partnership project (3GPP) as the preliminary version of next generation wireless communication systems because of its high data rates [2]. 3G LTE technology provides a maximum 100Mbps downlink and 50Mbps uplink while using 20 MHz bandwidth [3]. In the downlink physical layer, LTE uses Orthogonal Frequency-Division Multiple Access (OFDMA) radio technology to meet the LTE requirements for spectrum flexibility and enables cost-efficient solutions for wide carriers with high peak rates. In the uplink, LTE uses a pre-coded version of OFDMA known as Single-Carrier Frequency-Division Multiple Access (SCFDMA), in order to compensate for a drawback with normal OFDMA of a high Peak-to-Average-Power Ratio (PAPR) [4].

Wireless technology has expanded from *voice only*: to high-speed data, multimedia applications, and wireless Internet [5]. LTE requirements for high data rates are achieved by the fact that this technology is only designed for packet switched networks (PSN): hence, there is no need for the circuit switched mode. However, this design brings with it more technical challenges especially for voice services. VoIP services are both delay and packet loss sensitive. The biggest challenge of VoIP over LTE is to deliver Quality of Service (QoS). Normally users would expect voice with the same quality as that provided by circuit switched networks. However, traffic delivered over PSNs are subject to delay and packet loss [6]. A major issue with VoIP over LTE is that 3G LTE adopts a different method of resource transmission from other cellular systems like Code Division Multiple Access (CDMA). 3G LTE uses Physical Resource Blocks (PRB) as its transmission unit. PRBs can be defined as the basic unit with both frequency and time aspects [7]. Basically, the base station of 3G LTE, known as eNodeB has a fixed number of available PRBs according to their allocated bandwidth and it is supposed to assign PRBs repeatedly at every Transmission Time Interval (TTI) [3].

Our contributions in this paper are:

- Formulating the problem of scheduling and resource allocation using utility function optimization by extending the proposed approaches in [8][9] to include VoIP metrics
- Present a mathematical model for our extended version of problem formulation
- Use this technique to theoretically analyze the performance, complexity, and fairness of our proposed algorithm in [10] based on transmission rate, queue delay, and queue length parameters
- Through numerical and simulations analysis, we studied the performance, complexity, and fairness of the proposed algorithms
- Based on the numerical and simulation analysis, our proposed algorithm performed better than other algorithms proposed in literature in most cases

The rest of the paper is organized as follows: Section II discusses the related work, Section III analyzes VoIP QoS in 3G LTE, Section IV discusses the the system model, where we discuss the general problem formulation and the extended version of the problem formulation. Section V gives the summary of our proposed algorithm, PF, and EXP-PF. Section VI presents the simulation details where we discuss the PRB characteristics and scenario setup, Section VII presents the results analysis, which include numerical, performance, complexity, and fairness analysis of our proposed algorithm. Section VIII reviews the main conclusions and future work.

II. Related Work

Different techniques have been introduced in the literature to overcome the challenges faced when real time traffic is transmitted over an LTE network.

In [8], Jianwei Huang *et al.* addressed the gradient-based scheduling and resource allocation problem for the downlink OFDM system. They considered various practical features such as, integer tone allocation, different sub-channelization schemes, maximum SNR constraint per tone, self noise due to channel estimation errors, and phase noise. During each time slot, a sub-set of users are scheduled and the available tone and transmission power is allocated to them. Using the gradient based approach, they reduced this problem into an optimization problem, which can be solved in each time slot. Using the dual formulation, they were also able to give an optimal algorithm for this problem when multiple users can time share each tone. Their approach motivated us to further address the problem of gradient-based scheduling and resource allocation. We specifically focused on VoIP packets when transmitted over an LTE network. In our approach, we considered various parameters provided in channel state information such as, transmission rate. We also used the parameters provided at the Medium Access Control (MAC) layer (i.e., queue length) and the VoIP QoS requirements (such as, delay parameters).

In [11], Yaacoub *et al.* proposed two low complexity heuristic algorithms. The complexity of both algorithms was analyzed. The first algorithm had a linear complexity in the number of users and quadratic complexity in the number of resource blocks. The second algorithm had a linear complexity in both the number of users and resource blocks. It was shown that good results could be achieved by the proposed linear complexity algorithm (second algorithm). It was also shown through simulations that the maximization of total throughput leads to a higher cell throughput, although considering the logarithm of throughput as a utility function ensures proportional fairness, and thus constitutes a tradeoff between throughput and fairness.

In [12], Zhao *et al.* investigated two fairness criteria with regards to adaptive resource allocation for uplink OFDMA systems. These two criteria were Nash Bargaining Solution (NBS) fairness and proportional fairness (PF). These two criteria can provide attractive tradeoffs between total throughput and each user's capacity. Using Karush-Kuhn-Tucker (KKT) condition and iterative methods, two effective algorithms were designed to achieve NBS fairness and proportional fairness respectively. Through simulation results, NBS fairness criteria showed better performance in total capacity but the BS could not control the rate ratio because it only depends on the channel state of the users. PF criteria can provide a controllable rate ratio regardless of the channel condition for each user. However, to achieve the hard fairness, the system capacity degrades sharply.

With all these techniques introduced in the literature, there are still some challenges when real-time traffic such as, voice is transmitted over an LTE network. This is mostly due to the fading channels of wireless links and the delay and packet loss sensitive voice characteristics. So in this work, we extended the work in [8], where the problem of scheduling and resource allocation was formulated using the utility function optimization approach. We introduce the VoIP metrics to this approach and determine the resource allocation for VoIP users instead of power allocation, which was considered in [8]. Voice packet scheduling has some particular requirements such as, minimum end-to-end delay requirements, subchannel or subcarrier allocation constraints, etc.

We have projected the voice packet scheduling and resource allocation problem as a constrained optimization problem.This optimization objective aims at maximizing the expected total utility under different constraints. We implemented an algorithm for the proposed solution and analyzed its performance, complexity, and fairness. We then compared it with other algorithms in [13]. The simulation results were generated using the open source LTE system simulator called LTE-SIM [13]. It models different uplink and downlink scheduling strategies in multicell/multiuser environments: taking into account user mobility, radio resource optimization, frequency reuse techniques, the adaptive modulation, and coding (AMC). It is important to analyze the QoS requirements for voice when transmitted over an LTE network. This will give us an idea of what voice quality the end user can expect during the VoIP call.

III. VoIP QoS Analysis in LTE

A. VoIP in LTE Traffic and protocols

Conversation VoIP traffic in LTE can be assumed as the two state Markov model with a suitable voice activity factor (VAF). Different open source Codecs can be used in LTE but the most popular codec according to [2] is Adaptive Multirate (AMR). This codec provides 32-bytes voice payload in every 20 milliseconds while talking and 7-bytes payload every 160 millisecond while silent. The VoIP protocol stack, which utilizes the real-time transport protocol (RTP) is encapsulated to the user datagram protocol (UDP), which is in turn carried by IP. The use of all these protocols brings the total header size to 40 bytes for IPv4 header or a 60-bytes for IPv6 header. The overhead brought about by these headers causes serious degrading in the spectral efficiency supporting VoIP traffic in LTE. So to solve this problem, an efficient and robust header compression (ROHC) technique is used. This technique solves the overhead problem by minimizing the size of the IP/UDP/RTP headers to as little as 2 or 4 bytes using IETF RFC 3059 [7][14].

B. VoIP End-to-End Delay and Capacity

The main characteristic of voice traffic is strict delay requirements [15], according to [16] the allowed maximum mouth-to-ear delay for voice is 250ms. This delay requirements includes the assumption that the delay for the core network is approximately 100ms. The tolerable delay for radio link control (RLC), medium access control (MAC) buffering, scheduling, and detection should be strictly less than 150ms, this has been depicted in Figure 1 [2]. If we take into account that both end users are LTE users, then we can assume that the tolerable delay for buffering and scheduling is less than 80ms. To better account for unpredictability in network end-to-end delays, 3GPP performance evaluation has also chosen the delay of 50ms from eNodeB to UE [15]. Packets will be dropped when packet error and packet delay exceeds the target latency while VoIP traffic is transmitted over an LTE network. This may not affect the voice quality if the packet loss is less than outage threshold [15]. The outage limit means that packet error rate (PER) of VoIP users is kept below 2%. This gives us the actual limit that the maximum VoIP capacity for LTE is limited by the outage limit, which is described in TR 25.814 [7] and was later updated in R1-070674 [17]. We can finally describe VoIP capacity in LTE as the maximum number of VoIP users that can be supported without exceeding a given threshold and at least 95% of total VoIP users should meet the above described outage limits [2].

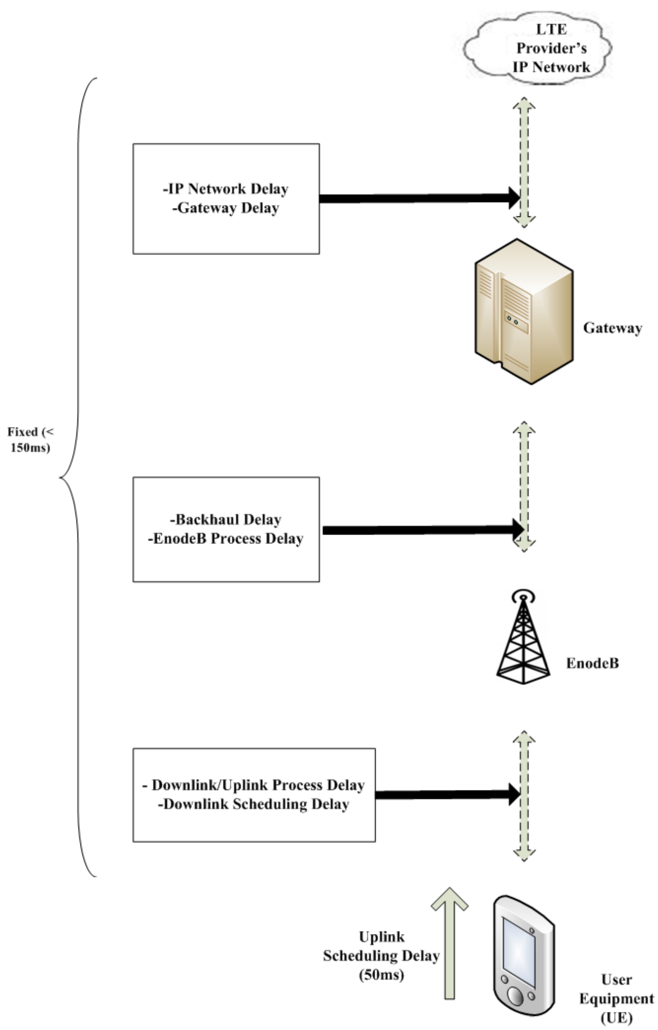


Figure 1 – VoIP End-to-End Delay Components in LTE

IV. SYSTEM MODEL

A. General Problem Formulation

In [8], the authors considered the downlink transmissions in an OFDM cell with base station and number of users. The authors considered K to be the maximum number of available users such that the number of users range from 1 to K, i.e., $k = \{1, ..., K\}$. So in every time slot, scheduling and resource allocation decision was done by choosing the rate vector $r_t = (r_{1,t}, ..., r_{K,t})$ from $R_{e_t} \subseteq R_+^K$, where e_t is the time varying channel state information available at time t. In short, the general problem is to find $r_t \epsilon R(e_t)$ that can maximize the system utility function $U(W_t) := \sum_{i=1}^{K} U_i(W_{i,t})$, where $U_i(W_{i,t})$ is the increasing concave utility function of user i's average throughput $W_{i,t}$, up to time t.

B. Our Extended Version of the Problem Formulation

1) utility function: Before extending the problem formulation in [8] to include VoIP metrics and other parameters, let us first describe the utility function. Utility functions can be useful in cross layer optimization as they can map network resources utilized by users into real numbers. The utility function can also indicate the level of satisfaction of the user, which in turn helps in the balancing the efficiency and fairness between the users. In 3G LTE, such as most wireless communication technologies, consistent transmission rate is the main factor that can determine the level of satisfaction of the user. So if we take m_j to be the transmission rate vector, then its utility $U(m_t)$ should be a nondecreasing function of the transmission rate m_j.

We adopted the utility function calculation from [9] and used it for the transmission rate m_t as:

$$U(m_j) = X_j\{\frac{1}{1 + e^{-p_j(m_j - R_j)}} - Y_j\} \quad (1)$$

With

$$X_j = \frac{1 + e^{p_j R_j}}{e^{p_j R_j}} \quad (2)$$

and

$$Y_j = \frac{1}{1 + e^{p_j R_j}} \quad (3)$$

where $U(m_j)$ is the utility function of user j with respect to their transmission rate. p_j is the priority tag assigned to VoIP users. R_j is the available resource blocks. X_j and Y_j are constants used to normalize the utility function.

2) Optimization Problem Formulation: The main aim of this problem formulation is to map the network resources of each user to their corresponding utility values. After that, the established utility function is optimized. Let K indexed by j, be the maximum number of available users such that $j \in \{1, ..., K\}$. If we consider the utility function of user j to be $U_j(.)$, then if user j has the transmission rate as m_j, we can say that the utility of user j is $U_j(m_j)$. Again if we let Q_l to be the length of user j's queue and Q to be the total number of queues for user j. Q is index by i, so $i \in \{1, ..., Q\}$. Then, the total utility function of user j is calculated from the utility function of its queue.

$$total_{utility} = Q_l * U_j(p_j R_j m_j) \quad (4)$$

where $U_j(p_j R_j m_j)$ can be equal to $U(m_j)$ in equation (1). If we take all the user j's queues in the network, then

$$Total_{utility} = \sum_{i=1}^{Q} Q_l * U_j(p_j R_j m_j) = \sum_{i=1}^{Q} Q_l * U_j(m_j) \quad (5)$$

So, our problem is to find a VoIP user that can maximize the total utility with respect to transmission rate and user queue values:

$$Max \sum_{i=1}^{Q} total_{utility} = Max \sum_{i=1}^{Q} Q_l * U_j(m_j) \quad (6)$$

However, the fact that we are dealing with VoIP application means that we need some constraints to control its QoS requirements. So, the above optimization objective equations should be subject to: $m_j \leq NC$ and $Q_d \leq D_{max}$. Where NC is the total available network capacity, Q_d is the queuing delay, and D_{max} is the maximum allowable mouth to ear delay.

Having seen that the main metrics in our problem formulation procedure are transmission rate (m_j), and queue length (Q_l), it is very important to know how these two metrics are obtained. This is described below.

3) Finding the Transmission Rate (m_j): During every transmission process, the user sends its instantaneous achievable signal to noise ratio (SNR) to its eNodeB. This value keeps on changing depending on different factors like mobility, selective fading channels, etc. So according to [9], user j's transmission rate at time t can be calculated as:

$$m_j(t) = \frac{n_{bits}}{symbols} * \frac{n_{symbols}}{slot} * \frac{n_{slots}}{TTI} * \frac{n_{subcarriers}}{RB} \quad (7)$$

where n_{bits}, $n_{symbols}$, n_{slots}, and $n_{subcarriers}$ are respectively the number of bits, number of symbols, number of slots, and number of subcarriers according to the PRB characteristic described earlier. TTI is the time transmission interval and RB is the resource blocks. These PRB characteristics are affected by path loss and fading channels but its values are kept constant for the entire PRB transmission time. According to [18], the channel gain of user j on a PRB at time t as a function of loss is calculated as:

$$CN_{gain_j}(t) = 10^{\frac{pathloss}{10}} * 10^{\frac{fading}{10}} \quad (8)$$

It should be noted that both $pathloss$ and $fading$ are measured in dB scale. Using this channel gain (CN_{gain}), the user knows the instantaneous SNR to send to eNodeB. Again according to [19], this SNR can be calculated as a function of CN_{gain}:

$$SNR_j(t) = \frac{P_{total} * CN_{gain}(t)}{R(N_o + I)} \quad (9)$$

Where P_{total} is the power with, which the eNodeB transmits, R is the total number available PRBs, I is neighboring interference, and N_o is the thermal noise measure.

4) Finding the Queue Length (Q_j): In order to obtain queue length metric, we adopted the queuing method in the LTE-SIM simulator. In this method, different traffic generators were developed, these generated packets that are transported by a dedicated radio bearer at the application layer. Using the application class, we were able to generate the packets and deliver them to the network. Once the packets reach the network, they are forwarded to the user-plane protocol stack to add protocol headers. Then, the packets are placed in the queue by the MAC queue class at the MAC layer before being sent to the destination. The MAC queue object has a counter, which increases or decreases when the packet is inserted or removed from the queue respectively.

Let $Q_j[l]$ be the amount of packets in user j's queue at time T_s. So, if the base station serves user j at rate $r_j[n]$ in time slot n, then user j's queue length at time $(n+1)T_s$, is expressed as:

$$Q_j[n+1] = Q_j[n] - r_j[n]T_s + a_j[n] \quad (10)$$

where $a_j[n]$ is the amount of arrival bits during time slot n.

5) Solution Approach: In order to solve our optimization problem in equation (6) that will maximize the network utility, we used the dual decomposition approach with lagrange multipliers. Solving this equation determines the user to be scheduled and assigned resource blocks according to the transmission rate (m_j) and queue values (Q_l, Q) parameters subject to $m_j \leq NC$ and $Q_d \leq D_max$ constraints. Writing up the optimization problem as a lagrange dual function, it becomes:

$$L(m_j, Q_d, \lambda, \mu) = \begin{array}{l} \sum_j U(m_j) + \lambda(NC - \sum_j m_j) + \\ \mu(D_{max} - \sum_j Q_d) \end{array} \quad (11)$$

The corresponding dual function can be written as:

$$L(\lambda, \mu) = MAX_{m_j, Q_d} L(m_j, Q_d, \lambda, \mu) \quad (12)$$

The inequality constraints in the optimization problem are put under consideration by augmenting the objective function with a weighted sum of the constraint function. Therefore, λ is called the lagrangian multiplier associated with $m_j \leq NC$ constraint and μ is the lagrangian multiplier associated with $Q_d < D_{max}$ constraint.

If we divide the objective function above into $|\lambda|$ and $|\mu|$ separate subproblems, then each subproblem can be solved separately if the values of λ and μ are known. the objective function of the dual problem then becomes:

$$D(\lambda) = MAX_{m_j \varepsilon NC} L(NC, \lambda) \quad (13)$$

and

$$D(\mu) = MAX_{Q_d \varepsilon Q_{max}} L(Q_{max}, \mu) \quad (14)$$

V. SUMMARY OF THE SCHEDULING ALGORITHMS

In this section, we are going to investigate how three different scheduling algorithms assign resources to their users in order to maximize their utility function. These algorithms are: our proposed algorithm, Proportional Fair algorithm (PF), and Exponential/Proportional Fair algorithm (EXP/PF).

A. Our Proposed Algorithm

Detailed explanations of this algorithm can be found in [10]. In our proposed algorithm, the scheduler assigns resources once every TTI and based on the user's current transmission rate (m_j), queuing delay (Q_d), and queue length (Q_l).

1) Summary of our Proposed Algorithm: The proposed algorithm is based on the modifications to the problem formulation in [8] and utility calculation in [9]. We introduced new parameters such as, transmission rate (m_j), queuing delay (Q_d), and queue length (Q_l). The first part of our algorithm is to determine the scheduling order for VoIP users. This can be done by ordering the users according to their decreasing sequence of their queuing delay (Q_d), and queue length (Q_l). Once the scheduling order is determined, the resource allocation is done by taking each user and determining the parameters that can maximize the utility of transmission rate (m_t). In order not to starve other applications in the network, we used the adaptive method proposed in [3].

This method provides limits to our proposed scheduling algorithm, which is adaptively changed between a pre-specific minimum and maximum according to the ratio of dropped packets. Higher drop ratio means that there are many ongoing VoIP calls, and hence it is necessary to increase the limits to allow more consecutive TTIs to be dedicated to VoIP calls. On the other hand, low drop ratio implies that QoS of VoIP calls are satisfied at decent levels, and thus it is safe to reduce the duration of the algorithm and serve other applications in the network.

2) Steps Involved in our Proposed Algorithm:

- **Step 1**: Determine the procedure of inserting users/packets into their queues
- **Step 2**: Scheduling starts at every TTI
- **Step 3:** Find out if there are any VoIP users/packets in the queues
- **Step 4:** If there are VoIP users/packets in the queue, apply our proposed algorithm and go to the next step otherwise apply the normal scheduling algorithm and exit
- **Step 5:** Arrange the VoIP users according to their decreasing values of their queuing delay (Q_d), and queue length (Q_l).
  Then, initialize $j = 1$, $m_{ext} = m_{max}$ and $Q_{ext} = Q_{max}$. Where m_{ext} and Q_{ext} are the extra/remaining transmission rate and queue length values at each stage
- **Step 6:** Determine if the successive counts of our proposed algorithm are not greater than the provided adaptive limit
- **Step 7:** If it is not greater than the limits then go to the next step, otherwise apply the normal algorithm, i.e., default algorithm such as, FIFO and exit
- **Step 8:** Find the parameter that maximizes the utility function for VoIP user j with $m_j \leq m_{ext}$ and $Q_j \leq Q_{ext}$
- **Step 9**: Schedule this VoIP user
- **Step 10:** Reduce m$_{ext}$ by m_j and Q$_{ext}$ by Q_j respectively
- **Step11:** If more resources blocks(RBs) and VoIP users exist, as well as $m_{ext} > 0$, $Q_{ext} > 0$ then set $j = j+1$ and repeat from step 8. If any of the three checks fails, then exit

The algorithm flow chart is presented in the Figure 2 below.

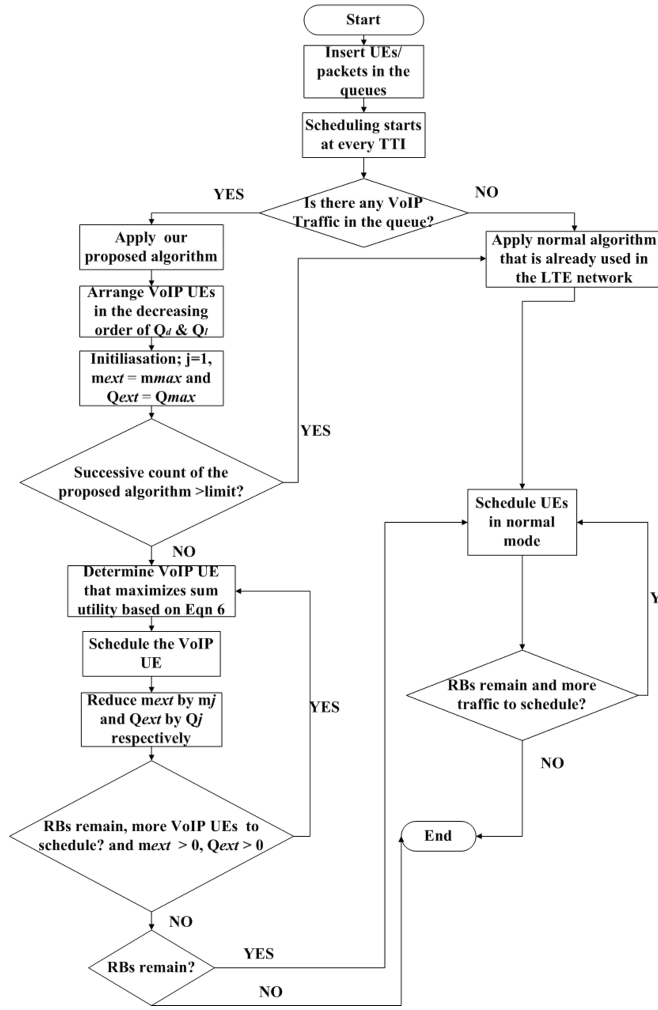


Figure 2 – Algorithm Flow Chart

B. Summary of Proportional Fair Algorithm (PF)

This scheduler was developed in [13][20]. Its main aim is to maximize the total network utility so that it can improve the network throughput and to guarantee fairness among flows. It assigns radio resources taking into account both the experienced channel quality and the past user throughput [21]. This scheduler uses a metric, which is defined as the ratio between the instantaneous available data rate and the average past rate with reference to the j-th flow in the i-th flow subchannel. This can be depicted in equation (15) below obtained from [13].

$$m_{i,j} = (\frac{r_{i,j}}{R_{i,j}}) \tag{15}$$

where $m_{i,j}$is the transmission rate, $R_{i,j}$ is the estimated average data rate, and $r_{i,j}$ is the instantaneous available data rate. $r_{i,j}$ is computed by the AMC module in LTE-Sim while considering the channel quality indicator (CQI) feedback that the UE hosting the j-th flow has sent for the i-th subchannel. It should also be noted that i and j are sub channel flows.

C. Summary of Exponential/Proportional Fair Algorithm (EXP-PF)

This scheduler was also developed in [13][20][22]. Its main aim is to increase the priority of real-time flows with respect to non-real-time flows, where their head-of-line packet (first packet in the queue) delay is very close to the delay threshold [21]. Its metrics were computed in [13] using the following equations.

$$m_{i,j} = exp\left(\frac{\alpha_i D_{HOL,i} - X}{1+\sqrt{X}}\right) \tag{16}$$

The variable X in equation (16) can be obtained from equation (17) below

$$X = \frac{1}{N_{r,t}} \sum_{i=1}^{N_{r,t}} \alpha_i D_{HOL,i} \tag{17}$$

where $N_{r,t}$ is the number of active downlink real-time flows. α_i in equation (17) can be described as the maximum probability that delay $D_{HOL,i}$ of the head-of-line packet exceeds the delay threshold. If we consider the packet threshold to be T_i, then α_i in equation (17) can be calculated as follows:

$$\alpha_i = -\frac{log_2 \alpha_i}{T_i} \tag{18}$$

Equations (17) and (18) proposed in [13], calculates the average total of the entire down link real time flows based on the probability that the first packet to be transmitted in the queue exceeds the delay threshold. This helps to prioritize down link real time flows.

VI. Simulation Details

A. PRB Characteristics

Before we go into the details of our simulation setup, lets first introduce the characteristics of PRBs, described as transmission resources in LTE. LTE systems can be analyzed both in time and frequency planes. The time plane is divided into 1 ms TTI, which consists of two slots of 0.5 ms to form 1 ms sub frames, where each sub frame contains 7 OFDMA symbols. In each TTI there are 14 OFDMA symbols, where 2 symbols out of 14 are reserved for uplink pilot transmission, while the other 12 symbols are used for data and control information transmission. TTIs can be defined as the minimum allocation unit in the time domain [23]. If we consider the frequency plane, the minimum allocation unit is the PRB, where each PRB contains 12 subcarrier of 15 KHz bandwidth each. The number of OFDMA symbols in a resource block depends on a cyclic prefix being used. All these can be depicted in Figure 3. It must be noted that VoIP packets must be transmitted per TTI and they can occupy one or more PRBs [6]. The amount of data bits that can be transmitted by one PRB depends on the link between the eNodeB and the user mobile terminal. This is due to the fact that 3G LTE uses adaptive modulation and coding (AMC), in order to change the modulation and coding schemes depending on the wireless link conditions.

B. LTE-Sim

In order to evaluate our proposed algorithm we used LTE-Sim simulation software. LTE-Sim is an open source

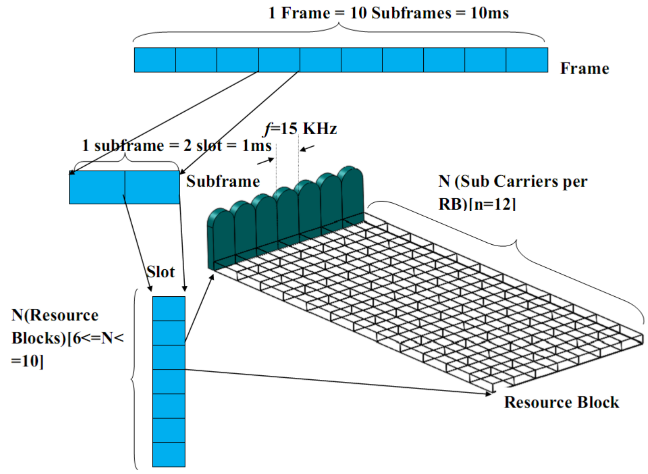


Figure 3 – Resource Grid

software, which is used to simulate LTE networks. It was developed in [13] and it is freely available under the GPLv3 licence. LTE-Sim is written in C++ using the object-oriented concept as an event driven simulator. This simulation software has all the important aspects of LTE networks notably the Evolved Universal Terrestrial Radio Access (E-UTRAN) and the Evolved Packet System (EPS). It maintains both single cell/multiuser and multiple cell/multiuser network topologies. This simulation software also supports different features, i.e., QoS management, user mobility, handover procedures, frequency reuse techniques, etc. Four different traffic generators are implemented and the management of data radio bearer is supported in this simulation software. The network nodes developed in this software are: User Equipment (UE), evolved Node B (eNB) and Mobility Management Entity/Gateway (MME/GW). Other features developed in LTE-Sim include: AMC scheme, channel quality indicator feedback, and some well known scheduling algorithms such as, PF and EXP-PF.

C. Scenario Setup

Our network topology is made up of a set of cells and different network nodes, which include: the eNodeB, mobility management/gateway (MME/GW), and user equipments (UEs). All the simulations were run in a three tier diamond-pattern macro scenario with 19-3-sector sites, which had a total of about 57 cells. The channel model used is the propagation loss channel model with channel realization. Most of the simulation parameters are presented in the Table 1. VoIP flows are generated by the traffic generator in LTE-SIM called VoIP application, which generates G.729 voice flows. The voice flow has been modelled with an ON/OFF Markov chain. The ON period is exponentially distributed with a mean value of 3s and the OFF period has a truncated exponential probability density function with an upper limit of 6.9s as well as an average value of 3s [24]. During the ON period, the source sends 20 bytes sized packets every 20 ms, which implies that the source data rate is 8 kb/s, on the other hand during the OFF period the rate is zero because the presence of voice activity detector is assumed. Three different scheduling algorithms were used in all simulation scenarios, these were: our proposed algorithm as well as EXP-PF and PF developed in [13]. Simulations were run for a number of iterations and in every iteration the seed number was updated. This was done in order to analyze the accuracy and the confidence interval of our simulation results.

TABLE I – Simulation Parameters

Simulation Parameters	Values
Bandwidth	5MHz
PRB structure	12subcarrier and 2subframes
TTI	1msec
Number of available PRBs	25
Number of sectors	3
Simulation time	1000 TTIs
Cyclic prefix	Normal
Number of Iterations	5
Scheduling Algorithms	Our Proposed Algorithm, EXP-PF, and PF
Cell radius	1 Km

VII. Results Analysis

A. Numerical Analysis

In this subsection, we present the numerical analysis to compare the analysis of our proposed scheduling algorithm with that of EXP-PF and PF proposed in the literature. The main reason for comparing our proposed scheduling algorithm with these two scheduling algorithms is that, they used the same PRBs allocation as ours and have similar simulation parameters except that they apply different metrics. They were also used as the benchmark scheduling algorithms in the LTE-Sim simulator. This made our comparison more feasible. Regarding the numerical analysis, we compared the level of VoIP user satisfaction (utility level) with the packet loss ratio, VoIP delay, and transmission rate. First let us analyze system utilization of all the three schedulers.

From the utilization point of view, as the transmission rate increases the bandwidth utilization also increases. This can be seen in Figure 4. This is mainly due to the fact that as the transmission rate increases, more users in the network are scheduled hence utilizing more bandwidth.

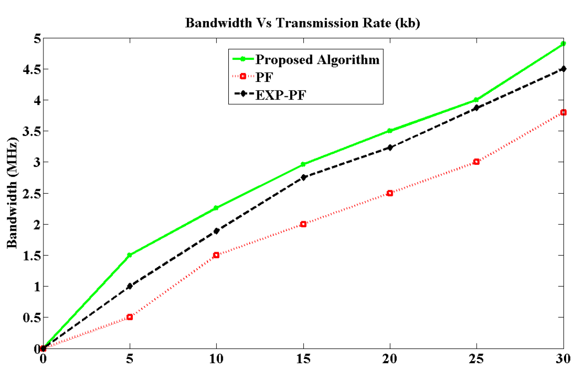


Figure 4 – Bandwidth (MHz) Vs Transmission Rate(kb)

From Figure 5, which shows the comparison between the utility levels and PLR, we can see that the VoIP user satisfaction levels were dropping as the PLR increased. PLR is the rate at, which VoIP packets are dropped during voice traffic transmission. So if more VoIP packets get dropped, the utility levels also start to fall. However, we note that there are some differences in the three schedulers. When PF and EXP-PF are used, the utility levels are lower than those of our proposed algorithm. This due to the fact that, with PF and EXP-PF schedulers, when there are high concurrent real-time flows, the probability of discarding packets for deadline expiration increases [11]. However, with our proposed algorithm, we do not calculate the deadline expiration factor for VoIP packets, it employs a simple method of scheduling users based on simple metrics, as well as the availability of the resource blocks.

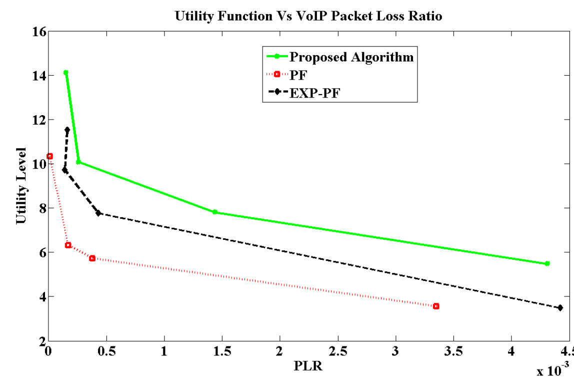


Figure 5 – Utility Function Vs PLR

Figure 6 compares the utility levels with VoIP delay. Again the utility levels decreases as the VoIP delay increases. When PF and EXP-PF are used, there is more reduction in the satisfaction levels than when our proposed algorithm is used. Our proposed algorithm employs a simple method of allocating resource blocks and scheduling VoIP user, which is less affected by high load factor as compared to the other two scheduling algorithm that employ packet deadline expiration procedure, which is highly affected by high load factor.

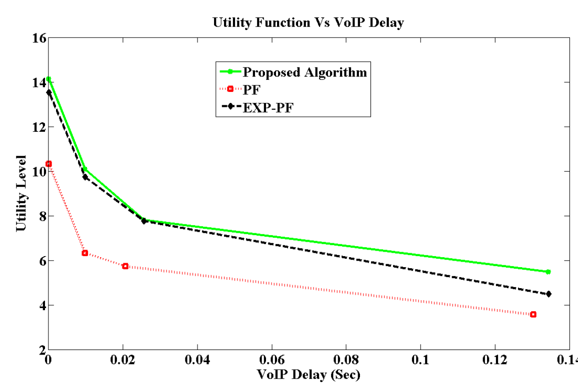


Figure 6 – Utility Function Vs VoIP Delay

Figure 7 compares the utility levels with transmission rates (m_j). The transmission rate was measured using the report provided by the user to eNodeB and was calculated using equations (7), (8), and (9) The better the transmission rate, the higher the levels of user satisfaction. As it can be seen in Figure 7, the utility levels are increasing as the m_j increases for all the three schedulers. Again, our proposed algorithm managed higher utility levels than EXP-PF and PF.

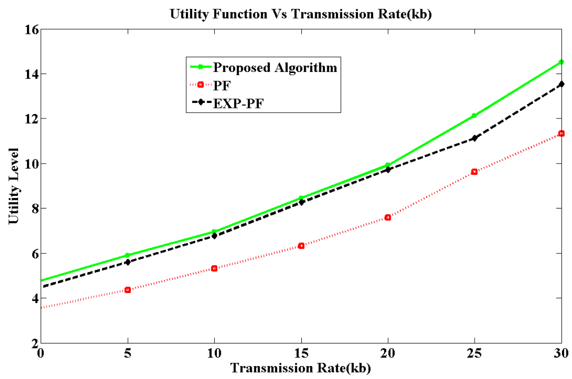


Figure 7 – Utility Function Vs Transmission Rate(kb)

B. Performance Analysis Based on VoIP Users

Regarding the performance analysis, we analyzed the same three VoIP metrics, which were throughput, VoIP delay, and packet loss ratio against the number of VoIP users. These metrics were compared with those of EXP-PF and PF scheduling algorithms. We analyzed user throughput for all three schedulers, while gradually increasing the number of VoIP users. As it can be seen in Figure 8, throughput decreased as the number of VoIP users increased in all algorithms. As the number of VoIP users increases, they overutilized the link and hence reduced the channel quality. This results in the VoIP packets being dropped as the number of VoIP users are increased, which led to less utilization of PRBs hence reducing the the total throughput achieved by VoIP users. As VoIP packets are small packets, many packets are needed to fully utilize the available PRBs. However, as congestion increased in the network, this led to VoIP packets being dropped, which led to less utilization of the available PRBs. But once again, our proposed algorithm obtained better throughput than the other two scheduling algorithms.

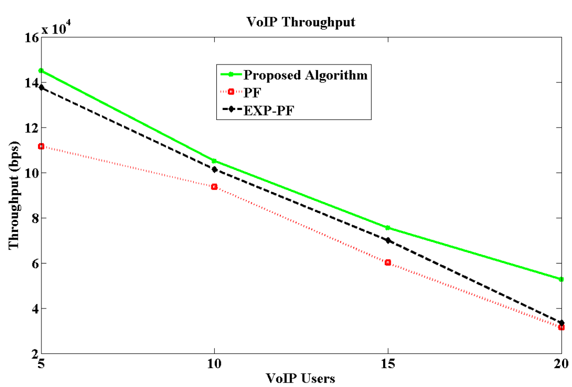


Figure 8 – Throughput Vs VoIP Users

We also analyzed VoIP delay while gradually increasing the number of VoIP users. This is shown in Figure 9. The VoIP delay is plotted on the Y axis in seconds as we increased the number of users steadily to twenty. As it can be seen, VoIP delay increased as the number of VoIP users increased. This is mainly due to the fact that as the number of users increase, they overutilize the link and cause congestion in the network. This will affect the transmission rate (m_j) and queue length (Q_l) metrics. This results in delay for VoIP packets. Even though there was an increase in VoIP delay for all the three schedulers, there are some differences in the three schedulers. When PF and EXP-PF are employed, the VoIP delay increases higher than of our proposed algorithm.

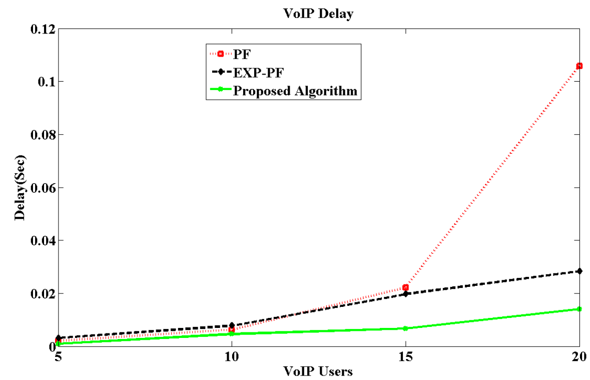


Figure 9 – Delay Vs VoIP Users

The packet drop ratio is analyzed and plotted on the Y axis as we increased the number of VoIP users steadily to the maximum of twenty users. Figure 10 shows the packet loss ratio for VoIP flows. As it can be seen, VoIP PLR increased as the number of VoIP users increased. This is mainly due to similar factors that affect VoIP delay. Even though there was an increase in VoIP PLR for all the three schedulers, we note some differences in the three schedulers. When PF and EXP-PF was used, the VoIP PLR increases more than that of our proposed algorithm. This is due to the fact that, with PF and EXP-PF schedulers, when there are high concurrent real-time flows, the probability of discarding packets for deadline expiration increases [8].

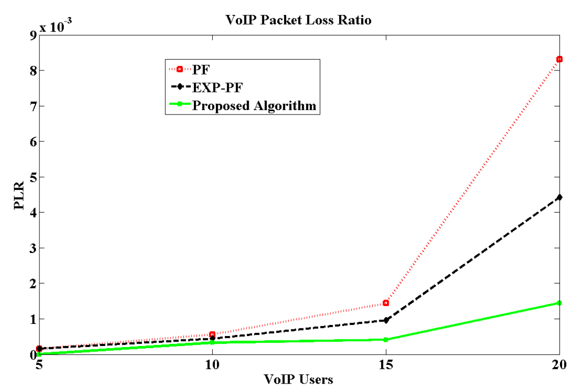


Figure 10 – PLR Vs VoIP Users

C. Complexity Analysis

Our proposed algorithm performs the scheduling operation after searching the user that can maximize utility function based on transmission rate (m_j) and queue length (Q_l) metrics. Therefore, the complexity to schedule the first user is

$O(KN)$, this will be the complexity for the first iteration. The complexity to schedule the second user is $O((K-1)N)$ and so on. In our algorithm, the number of iterations depends on the number of users K. As there are K iterations, the overall algorithm complexity can approximately be expressed as:

$$O(KN+(K-1)N+..+2N+N) = O(\frac{K(K+1)}{2}N) \approx O(K^2N) \tag{19}$$

Where $O(K^2N)$ implies that there is a second order complexity in the number of users based on m_j, Q_l metrics and there is also linear complexity in resource blocks N. This is due to the fact that there is no search done on the resource blocks, any available resource block is assigned to the user with highest metric. So for real complexity implementation of this algorithm, you only need to apply the possible values of N and K in the equation (19) to determine where it is efficient.

If we compare our algorithm to algorithm 1 in [11] that has a linear complexity in relation to the number of user and quadratic complexity in relation to the number of resource blocks, i.e., $O(N^2K)$, it is clear that our algorithm will only outperform it when the number of users are low since it will perform less iterations however when the number of users increases, algorithm 1 in [11] performs better.

D. Fairness Analysis

The fairness aspect is introduced mainly to solve the resource starvation problem, where users close to the base station are allocated more resources and edge users generally suffer from resource starvation [11]. Fairness can be described as a loose concept, which implies that all users are allocated equal amount of resources in order to meet QoS requirements. From the fairness point of view, we computed *jain* fairness index, which can be found in [25]. We compared our algorithm with PF and EXP-PF developed in [13]. Their fairness and complexity context constitutes an extension to algorithms described in [11][12] and they are also the bench mark schedulers in the simulator that we used. We analyzed the fairness index of all the scheduling schemes. As seen in Figure 11, fairness index decreased as the number of users increases. The fairness index of our proposed algorithm is higher than that of PF but lower than EXP-PF. It should be noted that the main advantage of our proposed algorithm is to improve the QoS of voice traffic when transmitted over an LTE network. At the same time, it reduces the negative impact, which may be caused by the introduction of the new algorithm on the entire systems performance. However, when we consider fairness analysis, EXP-PF out performs our proposed algorithm due to the following reason: EXP-PF employs the fairness concept in [11], which uses the algorithmic utility function that is associated with proportional fairness of the utility based optimization. This helps it in achieving a better fairness factor.

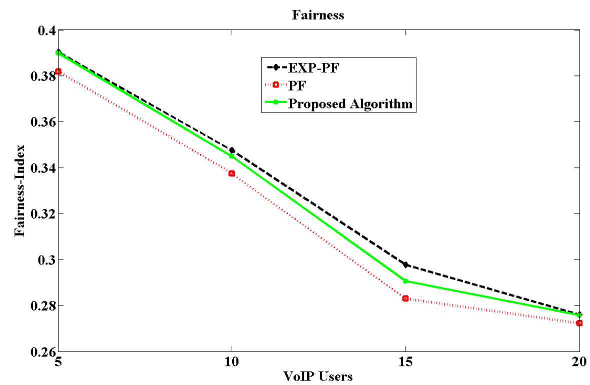


Figure 11 – Fairness Vs VoIP Users

VIII. Conclusion and Future Work

In this paper, we analyzed the problem of scheduling and resource allocation for VoIP in 3G LTE. We also investigated the performance, complexity, and fairness of our proposed cross-layer scheduling algorithm for VoIP in 3G LTE. We projected the voice packet scheduling and resource allocation problem as a constrained optimization problem. We solved this problem using a dual optimization approach with the goal of maximizing the expected total utility function under different constraints. Finally, we provided the algorithmic implementation of the obtained solution and also studied the performance of the proposed algorithm under different conditions and compared it with other algorithms in the literature, i.e., PF and EXP-PF.

Unlike other algorithms, which are time consuming and very complex, our proposed algorithm uses a metric maximization procedure to assign resource blocks to VoIP users. The main metrics used being queue length and transmission rate, this procedure makes our proposed algorithm less complex and it is executed in a short time.

Regarding complexity, our proposed algorithm performs better when the number of users is small since it schedules users after searching the user with highest utility metrics based on (m_j), (Q_l) and the search goes on for all available users. So for a small number of users, the search iterations done are less and hence the better performance. Our proposed algorithm performed better than PF but slightly poorer than EXP-PF.

Regarding numerical analysis, we compared the level of VoIP user satisfaction (utility level) with the packet loss ratio, VoIP delay, and transmission rate. These metrics were compared for all three scheduling algorithms, i.e., our proposed algorithm, EXP-PF, and PF. Our proposed scheduling algorithm performed better than the other two scheduling algorithms.

Regarding the performance analysis, we analyzed the same three VoIP metrics, which are throughput, VoIP delay, and packet loss ratio against the number of VoIP users. These metrics were compared with those of EXP-PF and PF scheduling algorithms. Again, our proposed algorithm out performed the other two algorithm.

Regarding fairness, the fairness index of our proposed algorithm is higher than that of PF but lower than EXP-PF. This is mainly due to the fact that EXP-PF uses the algorithmic utility function that is associated with proportional fairness of the utility based optimization. This helps in achieving a better fairness factor.

In future work, we will try to employ different tests such as, real life scenarios (existing networks) in order to analyze the difficulties and additional cost that would be required. It would also help us to analyze the practicability of our results and to make them more reliable. Our proposed algorithm can also be extended to other real-time applications, i.e., video, as well as extending it to latest LTE Advanced technology standards.

References

[1] R. Musabe and H. Larijani, "Complexity and Fairness Analysis of a new Scheduling Scheme for VoIP in 3G LTE," Eighth Advanced International Conference on Telecommunications (AICT), pp. 34-39, May 2012.

[2] S. Y. Kim, "An Efficient Scheduling Scheme to Enhance the Capacity of VoIP Services in Evolved UTRA Uplink," EURASIP journal of Wireless Communications and Networking, pp. 1-9, March 2008, doi:10.1155/2008/732418.

[3] S. Choi, K. Jun, Y. Shin, S. Kang, and B. Choi, "Mac Scheduling Scheme for VoIP Traffic Services in 3G LTE," IEEE 66th Vehicular Technology Conference, pp. 1441-1445, October 2007, doi: 10.1109/VETECF.2007.307, .

[4] Rohde and Schwarz, "UMTS long term evolution (LTE) Technology introduction," A report by Rohde and Schwarz, pp. 1-30, March 2007.

[5] M. C. Chuah and Q. Zhang, "Introduction to Wireless Communications," US: Springer, 2006.

[6] S. Saha and R. Quazi, "Priority-Coupling- A Semi-Persistent MAC Scheduling Scheme for VoIP Traffic on 3G LTE," ConTEL, 10th International Conference on Telecommunications, pp. 325-329, August 2009.

[7] 3GPP, "Physical layer aspects for Evolved UTRA," 3GPP Technical report 25.814, version 7.1.0, pp. 1-135, September 2006.

[8] J. Huang, V. Subramanian, R.Agrawal, and R. Berry, "Downlink scheduling and resource allocation for OFDM systems," IEEE Transactions on Wireless Communications, Vol.8, pp. 288-296, January 2009.

[9] S. Ali and M. Zeeshan, "A Utility Based Resource Allocation Scheme with Delay Scheduler for LTE Service-Class Support," IEEE Wireless Communications and Networking Conference: MAC and Cross-Layer Design, pp. 1460-1465, 2012.

[10] R. Musabe, H. Larijani, B. Stewart, and T. Boutaleb, "A New Scheduling Scheme for Voice Awareness in 3G LTE," IEEE Computer Society, Sixth International Conference on Broadband and Wireless Computing, Communication and Applications, pp. 1-8, December 2011.

[11] E.Yaacoub, H. Al-Asadi, and Z. Dawy, "Low Complexity Scheduling Algorithms for LTE Uplink," IEEE Symposium on Computers and Communications (ISCC), pp. 266-270, July 2009, doi: 10.1109/ISCC.2009.5202296.

[12] Y. Zhao, L. K. Zeng, G.Xie, Y.A. Liu, and F. Xiong, " Fairness based resource allocation for uplink OFDMA systems," Journal of China universities of post and telecommunications, pp. 50-55, Vol. 15, No. 2, June 2008.

[13] G. Piro, L. A. Grieco, G. Boggia, F. Capozzi, and P. Camarda, "Simulating LTE Cellular Systems: An Open-Source Framework," IEEE Transactions on Vehicular Technology, pp. 1-16, Vol. 60, No. 2, February 2011.

[14] IETF, "Attribute list extension for the service location protocol," pp. 1-7, IETF RFC 3059, February 2001.

[15] T. Henttonen, K. Aschan, J. Puttonen, and et al, "Performance of VoIP with Mobility in UTRA Long Term Evolution," Vehicular Technology Conference, pp. 2492-2496, May 2008, doi: 10.1109/VETECS.2008.549.

[16] ITU-T, "One way transmission time," ITU-T recommendation G.114, pp. 1-20, February 2003.

[17] TSG-RAN1, "LTE physical layer framework for performance verification," Technical Report, 3GPP TSG-RAN1 R1-070674, March 2007.

[18] X. Qiu and K. Chawla, "performance of adaptive modulation in cellular systems," IEEE Transactions on Communications, pp. 884-895, Vol.47, June 1999, doi: 10.1109/26.771345.

[19] H.Holma and A.Toskala, "WCDMA for UMTS: HSPA Evolution and LTE," John Wiley, fourth edition, 2007.

[20] H. Ramli, R. Basukala, K. Sandrasegaran, and R. Patachaianand, "Performance of well known packet scheduling algorithms in the downlink 3GPP LTE system," 9th IEEE Malaysia International Conference on Communications (MICC), pp. 815-820, December 2009.

[21] G. J. Choi and S. Bahk, "Cell-throughput Analysis of the proportional fair scheduler in the single-cell environment," IEEE Transaction Vehicular Technology, pp. 766-778, Vol.56, April 2007.

[22] R. Basukala, H. Mohd Ramli, and K. Sandrasegaran, "Performance analysis of EXP/PF and M-LWDF in downlink 3GPP LTE system," First Asian Himalayas International Conference on Internet, AH-ICI, pp. 1-5, Nov 2009.

[23] Roke, "LTE eNodeB MAC Schedular," Technical Report, Roke Manor Research Ltd, Siemens, January 2009.

[24] C. Chuah and R. H. Katz, "Characterizing packet audio streams from internet multimedia applications," Proc. IEEE, ICC, pp. 1199-1203, August 2002.

[25] R. Jain, "The Art of Computer Systems Performance Analysis," New York, Wiley, 1991.

TMDA: A Broadcast-Based Message Delivery Algorithm for VANETs

Yueyue Li, Stylianos Papanastasiou, Javad Akhlaghinia, and Evtim Peytchev
School of Science and Technology
Nottingham Trent University
Clifton Lane, Nottingham, NG11 8NS, UK
{yueyue.li, stylianos.papanastasiou, mohammad.akhlaghinia, evtim.peytchev}@ntu.ac.uk

***Abstract*—The challenges Intelligent Transportation Systems (ITS) address have only increased in time as, along with fast and cost-effective route planning, environmental sensibilities are nowadays also prominent. Currently, many ITS related studies focus on studying Car-to-Car (C2C) and Car-to-Infrastructure (C2I) communications in Vehicular Ad-hoc Network (VANET) environments. The utilization of such infrastructure-less and flexible topology networks poses several challenges for the routing mechanism aiming to achieve effective and efficient message delivery, specifically through broadcasts. This paper presents a new broadcast-based message delivery approach termed the Traffic Message Delivery Algorithm (TMDA) optimized for a city-based VANET setting. TMDA considers urban traffic travelling patterns and has been designed to exploit the route properties of different vehicle types such as cars and buses. In the case of the latter, TMDA adjusts its message propagation strategy so that bus routes, which intuitively involve elements (buses) with predictable routes, help propagate broadcast messages to a whole region which may otherwise have been disconnected. We investigate and compare the communications performance of vehicle groups under TMDA against other broadcast protocols through a set of experiments using NS-2 to simulate communications and SUMO to create representative mobility patterns. The simulation outcomes show that TMDA outperforms its competitors in efficiency and reliability while avoiding the deteriorating effects of a broadcast storm.**

Keywords- ITS; Car-to-Car (C2C) communication; VANET; broadcasting protocol; NS-2; SUMO

I. INTRODUCTION

Traffic demands in urban environments have been growing in recent years, following the proliferation of vehicles in the developing world and increasing worker mobility in modern economies. Utilising technology to help optimise existing traffic patterns and anticipate future demands has been a perpetual goal of Intelligent Transportation Systems (ITS). The availability of in-car wireless capabilities as well as the definition of interoperable vehicular communication standards worldwide has allowed for practically deployable decentralised ITS solutions alongside the more traditional centralised approach. As future trends indicate that in the near future urban centres will keep growing in importance and size, traffic coordination fuels and is fuelled by this growth, thus failing to manage it could stunt city development.

The importance of ITS in urban environments has been highlighted several times in the literature, with many developments helping define and expand its role [2-4]. A major theme in such works is decentralisation in the sense that traffic participants can be used as a practical tool to achieve efficient and reliable message propagation. Intuitively, vehicles are mobile entities with both predictable and unpredictable paths and so can be transports of messages to different sections of a city even if these are not covered by fixed ITS infrastructure. Further, fixed infrastructure facilities that can help propagate messages may be used in a complementary fashion helping the overall network to achieve an acceptable degree of connectivity. As a result of their great potential Car-to-Car (C2C) and Car-to-Infrastructure (C2I) solutions have received much attention [2], [5-8].

Several ITS related completed and on-going projects can be found in the seventh EU Framework Programme for Research and Technological Development [2], such as the Cooperative Vehicle-Infrastructure Systems (CVIS) project [5], which investigates interactions between vehicles and transport infrastructures for road safety. Another noteworthy and influential project is SAFESPOT [6], which considers intelligent information exchanges between vehicles and roadside units to realize safe and efficient transportation. Other works such as [7] recommend the deployment of roadside stations as it considers that infrastructure is a prerequisite for particular transportation monitors such as speed advisories and route navigation. These projects, among others, attempt to integrate C2C and C2I applications to achieve a combined benefit whilst highlighting that C2I can help improve ITS outcomes by utilizing roadside units (RSU), access points (AP) and cellular base-station information.

Even though C2I applications are broadly deployed compared to C2C applications, they exhibit well-studied weaknesses such as higher cost of infrastructure deployment [8], lower speed of connections [9] and smaller volume of handled data [10]. Importantly, C2C communications in Vehicular Ad-hoc Networks (VANETs), which are by definition infrastructure-free, self-governing and self-organized, can help effectively address such issues.

This study contributes to the C2C communication potential in a VANET architecture deployed in an urban (city) environment. It proposes a novel and efficient broadcast-based message dissemination algorithm termed the Traffic Message Delivery Algorithm (TMDA) which makes use of known travel route information for public (buses) or private vehicles (considering route intentions declared via Sat-Nav devices) to efficiently spread messages of interest through the network. The proposed algorithm is evaluated via extensive simulation, as opposed to testbed assessment [11-13], using the Network Simulator 2 (NS-2) [14] to model communications and the Simulation of Urban Mobility (SUMO) [15] to model the mobility pattern based on the topography of the city centre of Nottingham in the UK.

The rest of the paper is structured as follows. The next section presents a literature review of existing broadcasting techniques in ad-hoc networks outlining their respective strengths and weaknesses. Section 3 describes a VANET-based C2C communication architecture proposed in previous work, which works in tandem with existing fixed infrastructure features to more effectively accommodate ITS functions. Section 4 contains an in-depth description of our novel message delivery algorithm termed TMDA, intended for use in the VANET C2C setting described in the previous section. Section 5 describes the simulation parameters used in the subsequent evaluation of TMDA. Then, Section 6 presents the simulation results and an in-depth discussion of TMDA performance. Finally, Section 7 concludes the paper and offers suggestions for future work.

II. RELATED WORK

Flooding is a classic broadcasting method used in ad-hoc networks [3], [16], [17]. In this method, each of the nodes broadcasts or rebroadcasts a packet to all their neighbouring nodes the first time they receive it; if it has been received before they discard the packet to avoid redundant retransmissions. Ho et al. [18] state that simple flooding, also known as blind flooding [19] provides minimal state and high reliability, which are suitable for high mobility networks such as MANETs and VANETs.

Broadcast-based protocols suffer from high redundancy and transmission congestions, known as broadcast storms. Presently, several different broadcasting mechanisms are implemented in routing protocols to address this problem. Based on in-depth overviews of different protocols in [3] and [20], we only concentrate on those studies which are directly related to our area of focus, specifically those that propose delay-based, position-based and probability-based methods to mitigate broadcast storms.

A. Delay-based Protocols

In order to alleviate the negative effects of simultaneous broadcasting collisions, different methods have been adopted to generate forwarding delays.

In [21], the authors suggest a Vehicle-Density-based Emergency Broadcasting (VDEB) protocol which takes into account the sender-receiver distance. It obtains the distance from the interaction of neighbouring nodes, i.e., from acknowledgement (ACK) packets, and assigns a waiting time-slot for the rebroadcast action so that simultaneous forwarding can be avoided. The Beaconless routing algorithm (BLR) [22] applies a similar idea. Here, the delay is defined as deferring time which represents a relationship between transmission range and one-hop distance between the last sender and the current receiver. In Urban Multi-hop Broadcast (UMB) [23], which functions over IEEE 802.11, the farthest vehicles forward the received data first and then inform other nodes in between. Furthermore, Zhang et al. suggested a Neighbor Coverage-Based Probabilistic (NCPR) protocol [4], which defines a rebroadcast delay. The difference between NCPR and other protocols is that the delay is calculated based on a uncovered neighbour set $U(n_i)$ and covered neighbour sets of node s and $n_i - N(s)$ and $N(n_i)$.

B. Position-based Protocols

There exists a set of protocols which work with the position of nodes. In such cases, the nodes within or towards a particular area are the only ones that can rebroadcast messages. For example, the BLR [22] protocol specifies a Forwarding Area (FA). The nodes positioned inside the FA are the only ones that can rebroadcast the message. The shape of FA, which is defined as a sector in BLR [22] is considered as an issue. Position-based Opportunistic Routing (POR) [24] is another example of Position-based protocols and it defines the shape of FA as an area with positive progress towards a terminus.

C. Probability-based Protocols

Probability-based forwarding is another widely adopted approach to address the problem of broadcast storms. For example, Ni et al. [25] assume that the nodes should rebroadcast the received packet by following predetermined probabilities. So, if the probability is 100%, the scheme will be identical to blind flooding. The aforementioned NCPR [4] protocol uses a probability scheme to control the forwarding candidates. The authors in [25] introduce additional coverage ratio and connectivity factors to calculate the rebroadcast probability for each node. Higher coverage ratio means that there are more nodes that should be covered in the rebroadcast and therefore, the rebroadcast probability needs to be set to a high value (e.g., 0.75), allowing more nodes to dispose of Route Request (RREQ) packets. Meanwhile, although the connectivity factor increases the rebroadcast probability for the node in a sparse network with poor connectivity, it decreases the rebroadcast probability for nodes located in a dense network.

Current broadcasting protocols have improved the communication performance in terms of transmission efficiency and reliability. However, as Ros et al. [26] present, there are other factors such as uneven distributions of vehicles and their travelling speeds which are specific to

VANET networks. In sparse network scenarios such as VANETs, appropriate protocols should be applied to deal with the high number of disconnections and their impact on message exchanges. Many existing approaches [21], [24], [27] utilize acknowledgement packets for this purpose. For example, Lee et al. [28] introduce a method for periodically broadcasting to neighbouring nodes. In this approach, neighbouring nodes of one-hop distance will be able to disperse the message to higher distance hops when they are in one-hop distance with them because of their mobility freedom. As another method, Kitani et al. [29] suggest the concept of 'message ferrying' in inter-vehicle communications, introducing bus as the ferry rather than a vehicle. Message ferrying is proposed to improve the efficiency of information sharing in sparse areas by relying on buses' regular routes and their ability to help with collecting more traffic information.

In VANETs, there are diverse and changeable communication demands and traffic problems can occur at any time in different areas. These characteristics make it very difficult for a communication protocol to perform adequately in typical communication scenarios and, it would be fair to state, there has not yet been any comprehensive and popular message delivery algorithm to do so. Researchers have proposed algorithms which include particular traffic information, such as the inclusion of the acknowledgments into the periodic beacons for high reliability in [28] and the inclusion of vehicles' status and surrounding information in [30]. So far, on the basis of studies in the literature, explicitly utilising scheduled routes information has not been considered as a means to improving network coverage or message dissemination; the proposed TMDA algorithm in this paper addresses this void.

III. A VANET Framework

In previous work, we introduced a VANET architecture [31] which accommodates spontaneous wireless communications occurring within a group of wireless mobile nodes, as shown in Figure 1. The architecture integrates features of traditional ad-hoc networking and VANET technologies allowing for both intra-network and inter-network connections with gateway functions providing access to the Internet [32].

In this setting, we focus on the aspect of the communication system which utilizes vehicles (roaming agents) for routing purposes via the inclusion of traffic route information. To better utilise the potential of the vehicle/node as a message forwarding device when communications occur, the system distinguishes them in three types: mobile, semi-mobile and static nodes.

Mobile nodes, such as private vehicles, are defined as common ad-hoc nodes without pre-conceived routes that are able to consume and forward messages. As these are the most prolific in public road networks (privately owned vehicles outnumber publically or state owned ones in most settings), they form the majority group in acting as message forwarders and consumers. As vehicles can be generally equipped with highly capable electronic devices for message storage, they can also act as long-storage and dissemination agents for relevant messages through the network. The routes of such nodes cannot easily be predicted as it depends on the will of the owner at the time - i.e., the routes are largely unpredictable and cannot be known in advance.

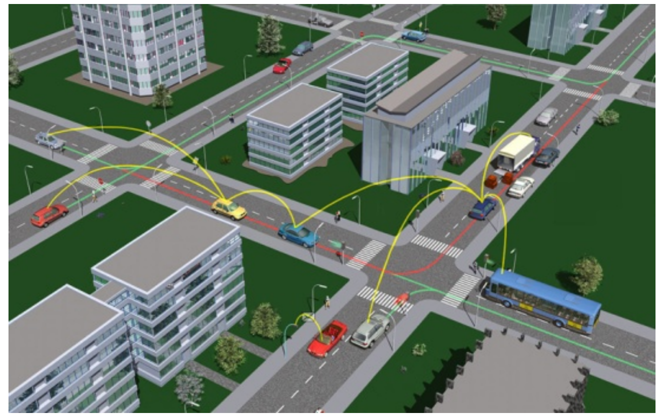


Figure 1. A VANET Scenario – city traffic communications [33]

Semi-mobile nodes, on the other hand, have predetermined routes which they follow at particular times; the most characteristic example of such vehicles is a bus following a particular bus route. In the particular case of buses, which is the semi-mobile node type considered in this study there is opportunity of including more powerful devices with greater storage than in cars due to the larger physical size of the bus vehicle. As such messages can be cached for longer before they are propagated (due to increased storage space), the communications range may be greater with the use of directional antennas (the cost of additional antennas is small compared to the cost of the bus) and more messages may be processed at the time (due to the availability of greater processing power). The presence of semi mobile nodes allows the predictable re-connection of possibly disconnected network segments and allows message propagation to reach, with predictable delay, areas in the road network where mobile node traffic is sparse. Further, in many cities there are dedicated bus-lanes or bus-priority lanes which help guarantee travel times for commuters even in times of peak traffic; as such the scheduled travelling times of semi-mobile bus nodes are more reliable than the predictable travelling time of other semi-mobile vehicles (such as police patrol cars).

Static nodes are fixed infrastructure units with the ability to relay and consume messages. These cannot ensure road network coverage on their own due to their high expense of deployment but instead cooperate with the other two node types to aid communication in the network. Broadly, these are termed roadside units and, in general, have few processing power and storage limitations.

The message delivery algorithm described in the next section is specifically designed to meet the requirements of C2C communications in this VANET framework by taking advantage of the three types of traffic participant mentioned above. We generally assume a sparse presence of static nodes (due to their cost) and mostly rely on mobile and semi-mobile vehicles to achieve message dissemination.

IV. TMDA – Novel Message Delivery Algorithm

Traffic Message Delivery Algorithm (TMDA) is a novel broadcasting-based algorithm designed for improving communication performance in the VANET network configuration described in the previous section. The TMDA version used here is identical to that used in previous work [1]; however, it is examined here more thoroughly. The main difference between TMDA and previously proposed broadcast methods is that it does not implement only a single broadcasting approach - such as simple flooding, delay-based, position-based or neighbourhood-based methods [34] - but also adopts intelligent broadcasting strategies by utilizing pre-existing travel information for semi-mobile vehicles. Specifically, every participating vehicle is aware of the route and schedule of semi-mobile nodes in the road network; this is used in the manner described below to disseminate a message to a single or several recipients. Initial route information of semi-mobile vehicles could be downloaded each day from the vehicle owner’s smartphone or through some other means.

Overall, TMDA aims to take advantage of the arbitrary, but presumably law abiding, route patterns of mobile nodes (cars) and exploits the benefits of controllable, scheduled, and predictable bus-nodes; it does not only use the simple broadcasting behaviours of cars, but also makes use of the higher processing power of bus-nodes to persistently store and forward messages. As well as handling local message delivery, TMDA allows for the possibility of Internet access through the static nodes infrastructure with road sign units acting as gateways.

A. Algorithm Description

TMDA is a receiver-oriented broadcasting protocol, meaning that receivers are solely responsible for determining re-broadcasting behaviour. The process can be distinguished in four sequences, namely (1) redundancy check; (2) position check; (3) distance check and (4) delay assignment. Algorithm I shows the pseudo-code of TMDA operations when a message reaches a receiver.

When a message is received by a receiver *R* but *R* is the originator *S* for this packet, then a broadcasting loop occurs (lines 2-3 in Algorithm I), so, in this case, *R* needs to discard the message. Each message has a unique message id *msg_id*. When a message reaches any receiver *R*, the *msg_id* is recorded into a broadcast table *brd_table* that every receiver *R maintains*. If the *msg_id* is found in *brd_table*, the redundant packet is directly discarded (lines 5-6); otherwise, *R* continues processing the message (lines 8 to 19).

ALGORITHM I. Pseudo-code of TMDA in Message Receiving

```
1 Event: the message received by R
2 if R = S then
3     discard the message;
4 else
5     if {msg_id ∩ brd_table} ≠ Ø then:
6         discard the message;
7     else
8         if P_r is on I-Routes or R is a static node then:
9             assign a WD_1;
10            forward the message at T_1;
11        else
12            if P_s is on I-Routes then:
13                discard the message;
14            else
15                if D_r = D_s then:
16                    assign a WD_2;
17                    forward the message at T_2;
18                else
19                    discard the message;
20                endif
21            endif
22        endif
23    endif
24 endif
25 END Event
```

In this VANET configuration, nodes are assumed to be equipped with GPS enabled devices. The location information derived from GPS allows the node to discover whether it is located in a special geographic location termed the *I-Route*. The term *I-Route* refers to predetermined routes on the map which are distinguished from regular road segments in that they are bus lanes where, normally, only buses travel, i.e., bus movement there is independent from normal traffic flow. If *R* is on an *I-Route* or it is a *static node* (e.g., bus stop), then *R* forwards the message to be propagated at time T_1 after a waiting delay WD_1. Another waiting delay WD_2 is defined with some value, which is inversely proportional to the distance between *S and R*. That is, the farthest *R* forwards the message first (because the calculated added delay will be smaller compared to that of other Rs). On the other hand, if *R* is not on the *I-Route* but the last-hop *S* is, then *R* discards the message so that redundant messages can be limited to a certain degree. With this method we attempt to guarantee that transmissions on an *I-Route* have higher priority to others. If the last-hop *S* is a non *I-Route* node, then *R* reads the message and knows that the position of last-hop sender (P_s) is not on the *I-Route*; it then compares the direction of the last sender (D_s) and itself (D_r). When they move towards the same direction, then the message is forwarded at T_2 after adding a delay of WD_2. WD_2 is always greater than WD_1 which in turn means that the resulting T_2 is greater than T_1. If the nodes move in opposite directions no propagating transmission takes place.

B. Main Mechanisms

1) I-Route: In TMDA, *I-Route* is a novel concept of assigning special importance to some linked road segments. The term implies a set of special routes considered in the message delivery algorithm that are used to determine the next operation of nodes. If a message reaches an *I-Route* receiver, the receiver forwards the message quickly using the pre-configured directions of the *I-Route*; otherwise, they follow the regular broadcasting strategy outlined in the algorithm. *I-Routes* are meant to represent locations where nodes will often be present and traffic will be dense. *I-Routes* also exist in sparse traffic places where there is a greater chance that a bus will be present compared to other parts of the network. In any case, *I-Routes* are predefined and they exist to attempt a best effort at finding an intermediate to propagate the message to its destination.

2) Farthest Node First Send: The final forwarding behaviour is defined by the Farthest Node First Send (FNFS) paradigm. This can be defined as follows; when a sender broadcast a message to all neighbours, the farthest one within the transmission range will retransmit the message before the others. This is represented by a probability value, as defined in (1):

$$P_{dis} = \frac{d_{<s,r>}}{d_{max}} \tag{1}$$

where P_{dis} is a probability of distance; $d_{<s,r>}$ represents a distance between last-hop sender and current receiver, as shown in (2); d_{max} is a maximum value of transmission range, e.g., 150 meters - this value depends on the signal propagation configuration.

$$d_{<s,r>} = \sqrt{\Delta x^2 + \Delta y^2} \tag{2}$$

where Δx and Δy is the difference of the x and y coordinates of the last sender and current receiver.

The above equations indicate that, if a receiver is positioned closer to the boundary of its last-hop sender, a larger P_{dis} is obtained, which contributes to a faster rebroadcast. In general, FNFS limits data collisions somewhat by reducing the number of transmissions along a transmission path.

3) Direction vector: The direction vector of the nodes is considered when both the sender and the receiver are not on any I-Routes. This is because nodes in such case are prone to broadcasting collision and congestion, and taking the direction of traffic into account can possibly reduce packet collisions. However, as described in Algorithm I the direction vector is not considered when the receiver is an I-Route node (a node located on an I-Route). The I-Route scheme includes a suppression mechanism to reduce the number of broadcast collisions - additional utilization of the direction vector may limit the quantity of forwarding candidates, resulting in lowering message reachability. Thus, the direction-based strategy is not essential for I-Route nodes. In (3), the direction vector is applied as:

$$P_{dct} = \frac{\pi - SR^\circ}{\pi} \tag{3}$$

where P_{dct} is termed a direction based probability; SR° represents the difference of direction vector of sender and receiver in degrees; π is equal to $180°$. In TMDA, when the angular distance between sender and receiver is less than $180°$, we consider these two nodes to move in the same direction, i.e., $D_r == D_s$.

4) Waiting Delay: In TMDA, the waiting delay WD is the most influential parameter used to control the time of forwarding. In generating a delay WD, the position and direction of senders and receivers as well as the distance between last hop and current receiver are the main factors. Forwarding with a WD strategy somewhat addresses the problem of broadcast storms. As mentioned before, VANET nodes are assumed to be equipped with GPS devices so they can detect their own positions and directions in a timely fashion and could place that information in the forwarded packets. So, when a receiver R receives the message, it is possible to deduce the distance between it and the sender (or last hop node). In TMDA, WD is calculated via (4) which adapts it for different node conditions:

$$T_{wd} = \begin{cases} (1 - P_{dis}) \times T_{hop} & RIR = 1 \\ (1 - P_{dis}) \times 2T_{hop} & RIR = 0, SIR = 0 \end{cases} \tag{4}$$

where T_{wd} is a waiting delay; T_{dis} is an minimum interval time of one-hop transmission.

According to the above equations, the waiting delay varies due to the position and direction of senders and receivers. The waiting delay is smaller when the receiver is an I-Route node than when both sender and receiver are non-I-Route nodes. In the latter condition, only communicating pairs facing the same directions have a chance to forward the message after some waiting delay. The predominant reason for this is to assign nodes different waiting delays before transmission, aiming to avoid or reduce the negative effects caused by simultaneous broadcasts.

V. Simulation Setup

To evaluate the proposed TMDA broadcasting protocols, we employ simulation and in particular use the NS-2 v.2.35 to create VANET network model for different scenarios. Further, we use SUMO to simulate real traffic flows in Nottingham (UK) city centre. Figure 2 depicts an overview of the simulation structure.

The NS-2 modules in the simulation structure consist of mobility information and broadcasting protocols. The raw data output from NS-2 is used to evaluate the network performance in Section 6.

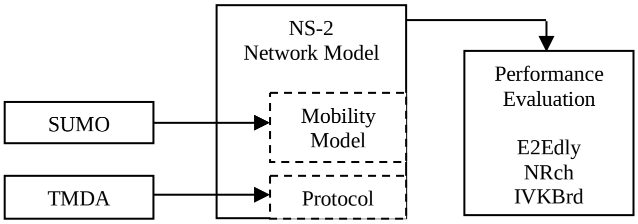


Figure 2. Simulation and evaluation modules

A. *Simulation Scenario*

In this study, the simulated area is chosen to be the city centre of Nottingham as shown on the map in Figure 3. Compared with T-shaped patterns [31] and #-shaped patterns [1] reported in our previous papers, this scenario contains more realistic elements such as road distributions, vehicle movements and the layout of bus lanes. Important elements in the simulation scenario are described as follows.

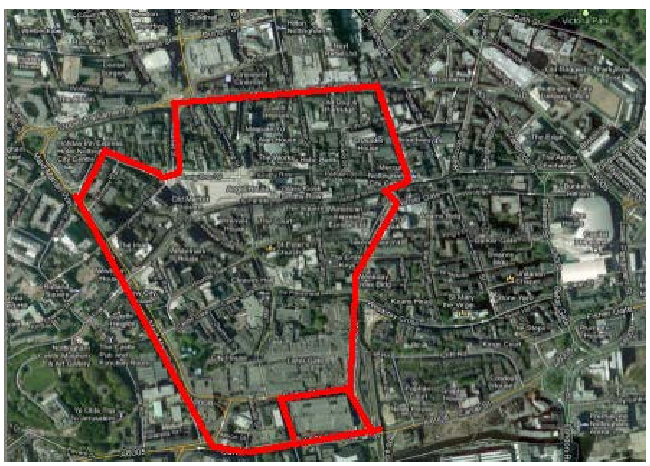


Figure 3. Nottingham (UK) City Centre with bus lanes

1) Nottingham City Centre: This is a medium scale traffic area consisting of intersections, roads in various shapes, buildings and so on. In this scenario, we need to find whether I-Routes provide efficient decisions for message delivery; whether different types of nodes can cooperate to provide high reachability under various network densities ;and whether broadcast storms can be alleviated effectively.

2) I-Route: This is a term used for a set of special routes in a target system. In I-Routes, the message transmission mechanism can process rebroadcasts based on their priorities and improve communication performance. In this simulation, we take advantage of buses in the city traffic by adopting bus lanes to shape the I-Routes; this is shown as red lines in Figure 3.

3) Network density: It represents the number of nodes in the network. In reality, the traffic density in an area can vary at different times and different days. Therefore, we consider simulation results for scenarios containing a number of mobile nodes, ranging from 50 to 200. We define terms 'very low density', 'low density', 'medium density', and 'high density' to represent networks with 50, 100, 150 and 200 nodes respectively. Density examples are shown in Figure 4.

Figure 4. Simulation models for very low and high density of networks

B. *Simulation Parameters*

The simulation parameters are shown in Table I. Further important simulation parameters are outlined below.

TABLE I. SIMULATION PARAMETERS

Network Simulator	NS-2 (version 2.35)
Traffic Simulator	SUMO
Channel	WirelessChannel
Bandwidth (Mbps)	6
Propagation	TwoRayGround
Network Interface	WirelessPhyExt
MAC Type	IEEE 802.11p
Protocol	TMDA, Flooding
Mobile vehicles	50, 100, 150, 200
Message size (byte)	128, 256, 512, 1024
Transmission range (meter)	150
Maximum Velocity (km/h)	30
Number of observation	50
Simulation time (second)	1500

1) Distance: In this work, the nodes are distributed following the topography of urban lanes. SUMO generates vehicular traffic using the car following model and the transmission is set to approximately 150 metres given that the signal propagation model (TwoRayGround) is deterministic.

2) Speed: The speed of vehicles can vary in different traffic conditions and in accordance with the traffic rules such as the road speed limit, etc. In this simulation, the speeds of nodes in city centre area are set to vary randomly with the limit of 30 km/h.

3) Time: The total simulation time is set to be 1500 seconds. NS-2 allows messages to be generated or sent by any nodes at any random time during the simulation time.

4) Message: Message contains three elements: message size, message id and other information such as the source node, current sender, the position, speed and direction of senders, the message expiry and current time. The size of message can vary for different communication purposes. In this paper, we observe message sizes from a minimum of 128 bytes to a maximum of 1024 bytes. We assume that only one message is transmitted per observation time.

VI. SIMULATION SETUP

In order to assess communication performance of TMDA by comparing it to conventional flooding protocol methods, we first introduce the performance metrics considered as follows, before presenting and discussing the simulation results.

A. Network Communication Performance Metrics

1) End-to-End delay (E2Ed): This metric refers to the time needed for a message from an originator to be received at the intended destination over the network, as in (5) and (6). E2Ed represents a capability of the network communication from a source to its possible destinations via one-hop or multi-hop transmissions.

$$\Delta T = T_e - T_0 \quad (5)$$

$$T_{e2ed} = T_{WD} + \Delta T \quad (6)$$

where T_e and T_0 stand for the time that the ultimate destination receives a packet and the first transmission time respectively.

2) Network Reachability (NRch): This parameter represents a ratio of network nodes receiving a message as in (7). In a broadcast storm, the NRch theoretically decreases due to more packet collisions or when link disconnections occur.

$$R_{NRch} = \left(\frac{N_{recv}}{N}\right) \times 100\% \quad (7)$$

where N_{recv} is the number of reached nodes and N is the total number of network nodes, i.e., vehicles and bus stops in this study.

3) Invoked Broadcast (IVKBrd): This metric measures the number of invoked broadcasts in the whole network. The particular goal of the metric is to observe how effectively the protocol can control the forwarding process.

B. Compared Protocols

1) TMDA: The proposed traffic message delivery algorithm in this study delivers messages using the concept of pre-configured routes (I-Routes) in the city scenarios. On the basis of general broadcasting methods, TMDA reduces broadcast storms via a selective forwarding mechanism, coupled with geographic information such as positions and directions.

2) Flooding: Flooding is the simplest broadcasting approach used in VANETs for data disseminations. This has been reviewed and further discussed in Section 2. Although there are obvious problems caused from the broadcast storm, flooding is a protocol with very high reachability in a particular range of networks. By comparing with this broadcasting protocol, we can clearly identify whether TMDA addresses broadcast storm and how effectively it does so.

Table II shows the anticipated features of TMDA, being given in advance, and these advantages and disadvantages are investigated through simulation results in the following sections.

TABLE II. ANTICIPATED CHARACTERISTICS OF TMDA

Advantages	Disadvantages
1) Simple broadcasting mechanism 2) No network topology maintenance 3) No complex route discovery algorithm 4) I-Routes are set up for controlling packet forwards 5) Broadcast storm is controlled and reduced by particular mechanism	1) Tolerate certain delays if nodes are not on pre-configured routes 2) Not suitable for emergency message exchanges in sparse networks based on C2C communications only

C. Results Evaluation and Analysis

This section compares the communication performance under the TMDA and flooding protocols with respect to E2Ed, NRch and IVKBrd. In this paper, the in-depth evaluation and analysis is presented in two parts: the effect of networks density and the effect of message size.

1) Effective of network density

Network density, as introduced earlier, represents the number of nodes in the network. In our target VANET scenario, both mobile vehicles and static bus stops are involved. In C2C communications, a dense network usually maintains good connectivity at the cost of serious broadcast storms. Therefore, our first investigation point on how to overcome heavy packet collisions and transmission congestions, which is closely related to the reachability performance metric. We compare the reachability of TDMA with the high reachability of flooding protocol to identify whether TMDA can exhibit an acceptable message delivery ratio.

A number of studies [2], [35], [36] have attempted to use infrastructures in C2I technology to reduce the problem of large disconnections and consequently lower reachability in sparse networks such as C2C. However, if we consider resource consumption and costs the C2C approach is more effective. At this stage, we are using TMDA to improve C2C communication performance without any connected infrastructure node. TMDA's capability to maintain C2C message delivery in both dense and sparsely connected

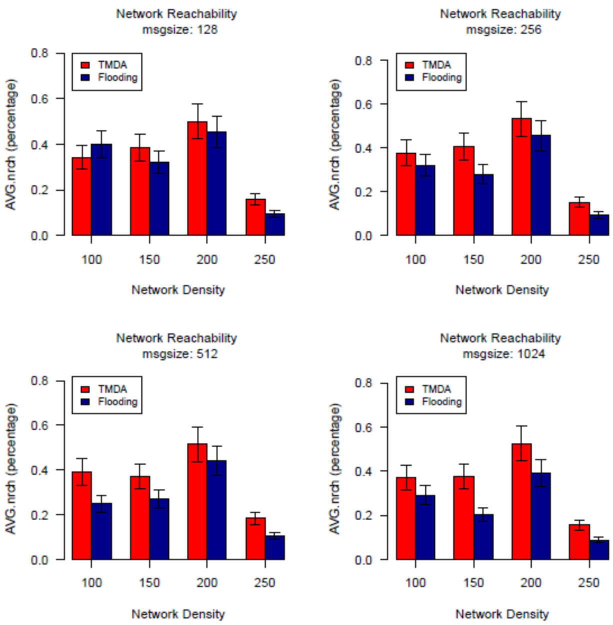


Figure 5. Reachability with the effect of density
(message size fixed)

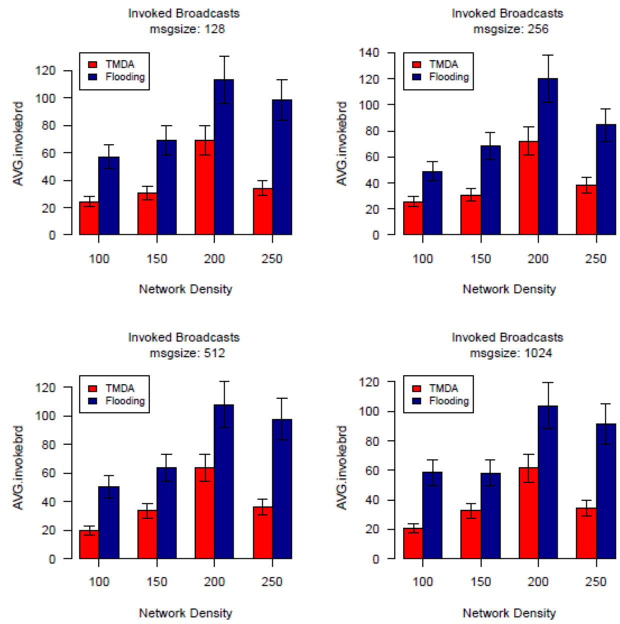


Figure 6. Invoked broadcasts with the effect of density
(message size fixed)

communication environments is investigated. This can become the basis for C2C and C2I communications in a more advanced TMDA in future work.

In Figure 5, TMDA's network reachability shows an upward trend of 0.35 to 0.5 when the number of nodes increases and then shows a significant decrease to approximately 0.15 when the network density exceeds 200 nodes. On the other hand, flooding protocol shows a slight network reachability fluctuation (0.4 to 0.32 to 0.48) when the number of nodes is less than 200. Flooding shows a significant decrease for network reachability similar to TMDA for a denser network (over 200 nodes). For fixed size of the messages, TMDA leads to around 5% to 20% higher reachability than flooding as the network becomes denser, except in the case where 128-byte messages are delivered in the 100-node network.

The reduction of rebroadcasts is a credible reason for TMDA's higher reachability. In Figure 6, TMDA shows a lower number of invoked broadcasts compared to the flooding protocol as the number of nodes in the network increases. For a fixed message size of 128 bytes, the IVKBrd is approximately 22, 30, 70 and 35 for TMDA and 58, 70, 115, and 98 for flooding, as network density increases. The range of difference in performance between the two protocols varies from about 30% to 60%. This outcome suggests that the concept of I-Route priority forwarding has reduced the number of redundant packets and thus alleviated broadcast storm.

The reachability results exhibit an unexpected trend when the size of the network increases from 200 nodes to 250 nodes. In most cases, TMDA controls the forwarding process adequately so we expected that the NRch in 250-node network should be higher than that in smaller networks. This is in contrast with the results for the 250 node case. Similarly, the flooding protocol results suffer from this descent. Here, we notice that the mobility model of 250-node network includes a large number of vehicles with short trips. In other words, many vehicles can leave the simulation

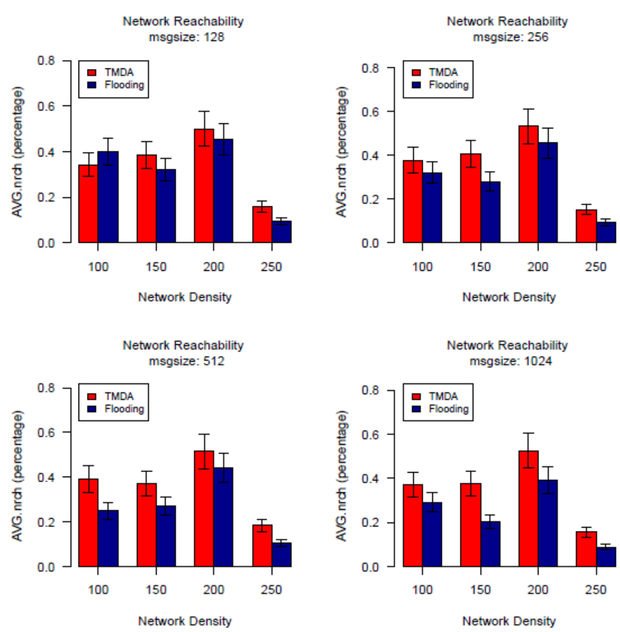


Figure 7. Delay with the effect of density
(message size fixed)

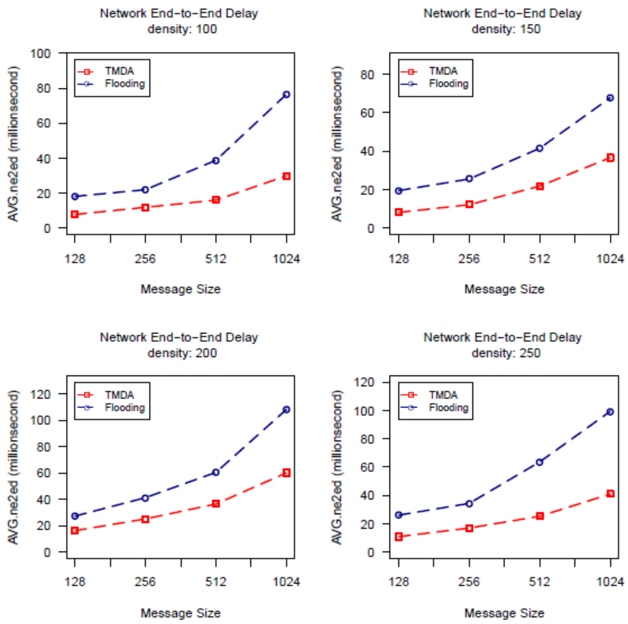


Figure 8. Delay with the effect of message size (network density fixed)

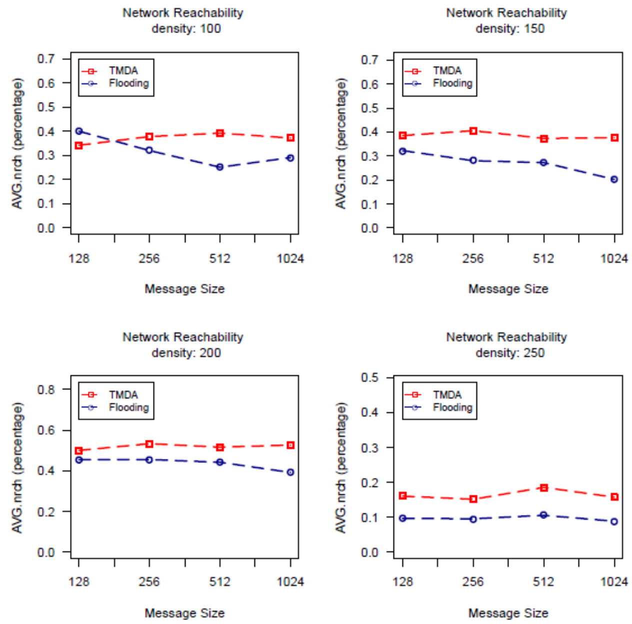


Figure 9. Reachability with the effect of message size (network density fixed)

region during simulation time. In this case, according to equation (7), the N_{recv} presents significant small value against the *N*. However, IVKBrd and EDT are not affected by this. This observations concludes our discussion of network reliability for TMDA compared to the flooding protocol. Now, we focus on the efficiency of message delivery.

In Figure 7, for both protocols the E2Ed metric generally increases as the number of nodes increases. Intuitively, if two protocols lead to similar or even the same level of message reachability, then the one with smaller E2Ed achieves more efficient transmissions. Results presented in Figures 5 and 7 show that TMDA can outperform flooding protocol with better NRch and E2Ed.

The best case for communication performance takes place in the 200-node network case where NRch is around 0.5 for TMDA and 0.42 for the flooding protocol when the message size is fixed to 128 bytes; in this case, E2Ed is 17ms and 27ms respectively. The differences between the two protocols vary in other cases. We can summarize that the schemes such as prior forwarding on I-Routes and waiting delay in TMDA work as anticipated in Table II. These schemes provide new optimisations addressing packet collisions caused by simultaneous forwarding.

Despite the fact that TMDA outperforms flooding for E2Ed, NRch and IVKBrd metrics, there is need and scope for further improvements addressing disconnections. In a sparse network (e.g., 100 nodes), if we were to observe high reachability, that would indicate that the network has an effective mechanism for addressing intermittent disconnections. In this case, TMDA shows rather better outcomes than the flooding protocol. However, it can be noticed that the value is still low over a C2C communication network – at around 0.35 regardless of the message size. We aim to address this in a future proposed improvement for TMDA.

2) Effect of message size

A critical factor effecting communication is not just the number of transmissions, but also the volume of data contained per message. To investigate the effect of the latter, in this paper, we set observed message sizes as 128

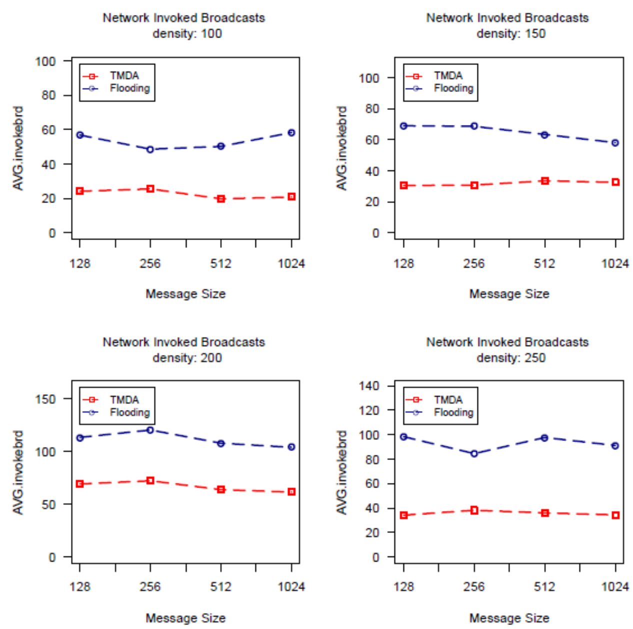


Figure 10. Invoked Broadcast with the effect of message size (network density fixed)

bytes, 256 bytes, 512 bytes and (at maximum) 1024 bytes. The effect of message size is represented in Figure 8 for E2Ed , Figure 9 for NRch and Figure 10 for IVKBrd.

There is an obvious impact on E2Ed as shown in Figure 8. Generally, for both protocols, E2Ed increases as the message size increases. This can be justified by the fact that the larger packets require longer time to be received by neighbouring nodes for a limited channel capacity. However, we consider more notable the fact that the E2Ed of TMDA is considerably less than that of flooding and the speed of growth under TMDA is much smaller (no more than 30ms) than that of flooding (up to approx. 75ms). The differences can be approximately 4ms to 60ms for very low density network, 5ms to 40ms for low density network, 11ms to 60ms for medium density network, and 15ms to 60ms for high density network. As shown above, TMDA outperforms flooding in terms of transmission delays and, in fact, the observed delays recorded are well acceptable for several VANET traffic applications. It then becomes of interest to examine whether packet size increase has an unacceptable deteriorating effect on reachability or observed end-to-end delay. In Figure 9, it can be seen that NRch for TMDA is not significantly affected by the increase in packet size at a given network density. Further, measurements of NRch in flooding show some notable decrease. This trend becomes less prominent when the network density increases to 250 nodes. As evinced in Figure 9, the best NRch performance noted in flooding is when the message size is 128 bytes and the number of nodes over the network is 100, i.e., the network is sparse. With regard to the number of IVKBrd shown in Figure 10, the impact of message size is not too significant as is the case with TMDA.

According to the above presented results, C2C communications in our chosen city scenario can only be reasonably achieved using the TMDA method when considering realistic packet sizes, such as 512-1024 bytes; flooding only exhibits reasonable performance at small packet sizes (128bytes) and in sparse networks. Overall, TMDA is a competent performer able to broadcast messages within acceptable delay and reachability margins, suitable for typical ITS applications.

VII. Conclusion and Future Work

This paper presented a new broadcast based Traffic Message Delivery Algorithm (TMDA) for VANETs and compared its communication performance to popular broadcast-based protocols in scenarios modelling road traffic of the centre of the city of Nottingham in the UK. TMDA adopts several principles of existing broadcast algorithms, such as delay-based and position-based forwarding techniques and, further, incorporates urban traffic route information. With respect to the latter, TMDA considers the concept of different node types in the VANET, such as cars and buses, and exploits the fact that some nodes' routes, termed the I-Routes, are predetermined and predictable. In TMDA nodes in such I-Routes are given higher transmission priorities so that propagation of messages, ultimately, occurs in areas where traffic is likely and predictably present. The broadcast techniques used in TMDA aim to alleviate the impact of the broadcast storms by controlling dissemination as opposed to indiscriminately re-transmitting broadcast messages.

We measured the performance of TMDA against flooding broadcast in sparse, medium and heavy traffic densities, measuring end-to-end delay, reachability and the number of broadcasts in the network with different packet sizes. Our results indicate that the end-to-end delay observed is always 25%-68% less when using TMDA in all traffic density and packet size configurations. The difference is in the higher range as the packet size used increases (from 128b to 1Kb). The same holds true with the number of invoked broadcasts; TMDA typically exhibits 30-60% less broadcasts than flooding for similar or higher network reachability. The latter is only observed to be less than flooding by 10% in a sparse network and at a very small packet size; in any other case reachability is observed to be higher than flooding by 12-45%. Overall, we observe that in all interesting use cases, i.e., medium packet/message size and across all network topology configurations, TMDA outperforms flooding in all metrics.

In the near future, we aim to examine TMDA performance against more modern flooding techniques, in particular, probabilistic and counter-based message dissemination methods because the advanced protocols show high capability to control the broadcast storm during message propagations. We also aim to consider other road topologies from different cities in the UK to extend our case study and observe if our results hold in different settings. Finally, we will study the offered broadcast load more thoroughly by examining TMDA behaviour when the number of simultaneous broadcasts increases.

Acknowledgment

The authors wish to thank all members of the Wireless Communications Simulation and Modelling group at Nottingham Trent University.

References

[1] Yueyue Li and Evtim Peytchev. "New Traffic Message Delivery Algorithm for a Novel VANET Architecture." In ICWMC 2012, The Eighth International Conference on Wireless and Mobile Communications, pp. 395-401. 2012.

[2] “European Commission Community Research-CORDIS.” [Online]. Available: http://cordis.europa.eu/fp7/home_en.html. [Accessed: May-2013].

[3] T. Camp and B. Williams. Comparison of broadcasting techniques for mobile ad hoc networks. In *Proceedings of The Third ACM International Symposium on Mobile Ad Hoc Networking and Computing (MOBIHOC 2002)*, Lausanne, Switzerland, Jun 2002.

[4] Xin Ming Zhang, En Bo Wang, Jing Jing Xia, and Dan Keun Sung, "A Neighbor Coverage-Based Probabilistic Rebroadcast for Reducing Routing Overhead in Mobile Ad Hoc Networks," *Mobile Computing, IEEE Transactions on* , vol.12, no.3, pp.424,433, March 2013.

[5] "CVIS Co-operative vehicle-infrastructure systems," 22-Jun-2010. [Online]. Available: http://www.cvisproject.org/. [Accessed: May-2013].

[6] L. Wischhof, A. Ebner, H. Rohling, M. Lott, and R. Halfmann, "Adaptive broadcast for travel and traffic information distribution based on inter-vehicle communication," presented at the IEEE Intelligent Vehicles Symp., Columbus, OH, June 2003.

[7] A. Paier, R. Tresch, A. Alonso, D. Smely, P. Meckel, Y. Zhou, and N. Czink, "Average Downstream Performance of Measured IEEE 802.11p Infrastructure-to-Vehicle Links," in *in Proc. IEEE Int. Conf. Commun*, Cape Town, South Africa, May 23–27, 2010, DOI: 10.1109/ICCW.2010.5503934.

[8] H. Hartenstein, B. Bochow, A. Ebner, M. Lott, M. Radimirsch, and D. Vollmer "Position-Aware Ad Hoc Wireless Networks for Inter-Vehicle Communications : the Fleetnet Project," in *In Proceedings of the second ACM international symposium on Mobile and ad hoc networking and computing (MobiHoc '01)*, 2001, pp. 259–262.

[9] L. Wischhof, A. Ebner, H. Rohling, M. Lott, and R. Halfmann, "Adaptive broadcast for travel and traffic information distribution based on inter-vehicle communication," presented at the IEEE Intelligent Vehicles Symp., Columbus, OH, June 2003.

[10] A. Paier, R. Tresch, A. Alonso, D. Smely, P. Meckel, Y. Zhou, and N. Czink, "Average Downstream Performance of Measured IEEE 802.11p Infrastructure-to-Vehicle Links," in *in Proc. IEEE Int. Conf. Commun*, Cape Town, South Africa, May 23–27, 2010, DOI: 10.1109/ICCW.2010.5503934.

[11] Y. Li and E. Peytchev, *Novel ad-hoc wireless mobile communication network routing model for location based sensor networks*, in proceedings of International Symposium on LBS & TeleCartography, pp. 383-396, September 2010.

[12] APE: *Ad hoc Protocol Evaluation testbed*, Department of Computer Systems at Uppsala, Sweden. [Online]. Available: http://apetestbed.sourceforge.net [Accessed: May-2013].

[13] J. Broch, D.A. Maltz and D.B. Johnson, Quantitative lessons from a full-scale multi-hop wireless ad hoc network testbed, in: Proceedings of the IEEE Wireless Communications and Network Conference 2000 (WCNC 2000), Chicago, IL, pp. 992-997.

[14] Fall, Kevin, and Kannan Varadhan. "The network simulator (ns-2)." [Online]. Available: http://www.isi.edu/nsnam/ns/ [Accessed: May-2013].

[15] "SUMO - Simulation of Urban MObility." Institute of Transportation Systems at the German Aerospace Centre. [Online]. Available: http://sumo.sourceforge.net / [Accessed: May-2013].

[16] A.M. Hanashi, A. Siddique, I. Awan, and M. Woodward, Performance evaluation of dynamic probabilistic broadcasting for flooding in mobile ad hoc networks, Simulation Modelling Practice and Theory, Volume 17, Issue 2, February 2009, Pages 364-375, ISSN 1569-190X, 10.1016/j.simpat.2008.09.012.

[17] J. Jetcheva, Y. Hu, D. Maltz, and D. Johnson. A simple protocol for multicast and broadcast in mobile ad hoc networks. Internet Draft: draft-ietf-manetsimple-mbcast-01.txt, July 2001.

[18] C. Ho, K. Obraczka, G. Tsudik, and K. Viswanath. Flooding for reliable multicast in multi-hop ad hoc networks. In *Proceedings of the International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communication (DIALM)*, pages 64–71, 1999.

[19] I. Stojmenovic and J. Wu, "Broadcasting and Activity Scheduling in Ad Hoc Networks," Mobile Ad Hoc Networking, S. Basagni, M. Conti, S. Giordano, and I. Stojmenovic, eds., pp. 205-229, IEEE Press, 2004.

[20] Panichpapiboon, S. and Pattara-Atikom, W., "A Review of Information Dissemination Protocols for Vehicular Ad Hoc Networks," *Communications Surveys & Tutorials, IEEE* , vol.14, no.3, pp.784,798, Third Quarter 2012.

[21] Yu-Tian Tseng, Rong-Hong Jan, Chien Chen, Chu-Fu Wang, and Hsia-Hsin Li, "A vehicle-density-based forwarding scheme for emergency message broadcasts in VANETs," *Mobile Adhoc and Sensor Systems (MASS), 2010 IEEE 7th International Conference on* , vol., no., pp.703,708, 8-12 Nov. 2010.

[22] K. Endo, Y. Inoue, and Y. Takahashi. 2012. Performance modeling of beaconless forwarding strategies in multi-hop wireless networks. *Comput. Commun.* 35, 1 (January 2012), 120-128.

[23] L.J. Li, H.F. Liu, Z.Y. Yang, L.J. Ge, and X.Y. Huang, Broadcasting Methods in Vehicular Ad Hoc Networks, *Journal of Software*, Vol.21, No.7, July 2010, pp.1620−1634 (in Chinese with English abstract).

[24] Shengbo Yang, Chai-Kiat Yeo, and Bu-Sung Lee, "Toward Reliable Data Delivery for Highly Dynamic Mobile Ad Hoc Networks," *Mobile Computing, IEEE Transactions on* , vol.11, no.1, pp.111,124, Jan. 2012.

[25] S. Ni, Y. Tseng, Y. Chen, and J. Sheu. The broadcast storm problem in a mobile ad hoc network. In *Proceedings of the ACM/IEEE International Conference on Mobile Computing and Networking (MOBICOM)*, pages 151–162, 1999.

[26] F.J. Ros, P.M. Ruiz, and I. Stojmenovic, "Acknowledgment-Based Broadcast Protocol for Reliable and Efficient Data Dissemination in Vehicular Ad Hoc Networks," *Mobile Computing, IEEE Transactions on* , vol.11, no.1, pp.33-46, Jan. 2012.

[27] C. E. Perkins, M. Park, and E. M. Royer, "Ad-Hoc On Demand Distance Vector Routing," in Proceedings of the Second IEEE Workshop on Mobile Computing Systems and Applications, pages 90–100. IEEE Computer Society, February 1999.

[28] U. Lee, J. Lee, J. Park, E. Amir, and M. Gerla, "FleaNet: A Virtual Market Place on Vehicular Networks," Proc. Third Ann. Int'l Conf. Mobile and Ubiquitous Systems: Networking and Services, pp. 1-8, July 2006.

[29] T. Kitani, T. Shinkawa, N. Shibata, K. Yasumoto, M. Ito, and T. Higashino, "Efficient vanet-based traffic information sharing using buses on regular routes," May 2008, pp. 3031–3036.

[30] J. Fukumoto, Sirokane, Y. Ishikawa, Wada, N. Tomotaka, K. Ohtsuki, and H. Okada, Analytic method for real-time traffic problems in VANET by using Contents Oriented Communications, 7th International Conference on ITS, 6-8 June 2007.

[31] Y. Li and E. Peytchev, *Novel ad-hoc wireless mobile communication network routing model for location based sensor networks*, in proceedings of International Symposium on LBS & TeleCartography, pp. 383-396, September 2010.

[32] I. Chlamtac, M. Conti, and J.J.-N. Liu, Mobile ad hoc networking: imperatives and challenges, Ad Hoc Networks, Volume 1, Issue 1, July 2003, Pages 13-64, ISSN 1570-8705, 10.1016/S1570-8705(03)00013-1.

[33] A VANET Scenario – city traffic communications, [Online]. Available: http://www.car-to-car.org/index.php. [Accessed: May-2013].

[34] N. Karthikeyan, V. Palanisamy, and K. Duraiswamy, *Performance Comparison of Broadcasting methods in Mobile Ad Hoc Network*, International Journal of Future Generation Communication and Networking , Vol. 2, No. 2, June, 2009.

[35] Mershad, K., Artail, H., and Gerla, M., "We Can Deliver Messages to Far Vehicles," *Intelligent Transportation Systems, IEEE Transactions on* , vol.13, no.3, pp.1099,1115, Sept. 2012.

[36] Abdrabou, A. and Weihua Zhuang, "Probabilistic Delay Control and Road Side Unit Placement for Vehicular Ad Hoc Networks with Disrupted Connectivity," *Selected Areas in Communications, IEEE Journal on* , vol.29, no.1, pp.129,139, January 2011.

Elckerlyc Goes Mobile
Enabling Natural Interaction in Mobile User Interfaces

Randy Klaassen, Jordi Hendrix, Dennis Reidsma,
Rieks op den Akker, Betsy van Dijk
Human Media Interaction
University of Twente
Enschede, the Netherlands
{*r.klaassen, j.hendrix, d.reidsma, h.j.a.opdenakker, e.m.a.g.vandijk*}*@utwente.nl*

Harm op den Akker
Roessingh Research and Development, Telemedicine group
University of Twente, Telemedicine group
Enschede, the Netherlands
h.opdenakker@rrd.nl

***Abstract*—The fast growth of computational resources and speech technology available on mobile devices makes it possible to entertain users of these devices in having a natural dialogue with service systems. These systems are sometimes perceived as social agents and this can be supported by presenting them on the interface by means of an animated embodied conversational agent. To take the full advantage of the power of embodied conversational agents in service systems it is important to support real-time, online and responsive interaction with the system through the embodied conversational agent. The design of responsive animated conversational agents is a daunting task. Elckerlyc is a model-based platform for the specification and animation of synchronised multi-modal responsive animated agents. This paper presents a new light-weight PictureEngine that allows to run this platform in mobile applications. We describe the integration of the PictureEngine in the user interface of two different coaching applications and discuss the findings from user evaluations. We also conducted a study to evaluate an editing tool for the specification of the agent's communicative behaviour. Twenty one participants had to specify the behaviour of an embodied conversational agent using the PictureEngine. We may conclude that this new light-weight back-end engine for the Elckerlyc platform makes it easier to build embodied conversational interfaces for mobile devices.**

***Keywords- Mobile ECA*;*User Interfaces*; *user evaluations*; *(mobile) coaching applications*.**

I. Introduction

Advances in user interface technology — speech recognition, speech synthesis and screen capacities — more and more allow people to engage in spoken interaction with services on their mobile phones. The use of a talking head or an embodied conversational agent (ECA) can support spoken interaction in different kind of user interfaces. In [1], a new light-weight PictureEngine was presented that allows to use ECAs in the user interfaces of mobile applications. The presentation of a service agent by means of a persona supports the idea of the computer as a social actor. Research has shown that animation of human-like social behaviours and expressions by means of a virtual human or embodied conversational agent strengthens the impression that the agent is present and engaged in the interaction [2]. They have a positive effect on user experience [3].

In human-human conversations, the one who has the speaker role is monitoring his addressees while speaking. Listeners give back-channels, short comments, and may also interrupt the speaker. By his gaze behaviour the speaker shows his interest with the addressee. By adjusting or stopping his speech, he shows being responsive to the listeners comments and that he is really engaged in the conversation. Gaze behaviour in conversations is important for interaction management, in particular for signalling that one wants to have the floor, that the speaker wants to keep the floor or is willing to yield the floor. Emotion expressions are prime indicators of engagement in what is going on in the conversation [4]. In designing virtual humans that are able to show these social signals and responsiveness one needs well designed model-based specification languages and tools. The SAIBA framework [5] provides a good starting point for designing the behaviours of interactive virtual humans. Its Behaviour Markup Language (BML) defines a specification of the form and relative timing of the behaviour (such as speech, facial expression, gesture) that a BML realizer should display on the embodiment of a virtual human. Elckerlyc is a state-of-the-art BML Realizer. In [6], its mixed dynamics capabilities are described and its focus on continuous interaction, which makes it very suitable for virtual human applications requiring high responsiveness to the behaviour of the user.

The Elckerlyc platform can act as a back-end realizer for different embodiments, like physical robots or realistic 3D full kinematic virtual humans. Using a full 3D virtual human on a mobile phone is however too heavy in terms of processing power and battery usage. To be able to use the Elckerlyc platform on a mobile phone, a light-weight animation embodiment is needed. The work presented here contributes to satisfy this need.

One of the many application areas for natural interaction with embodied agents is in healthcare services and coaching systems that users interact with through mobile devices. The field of Telemedicine — healthcare delivered remotely to the patient or user — is receiving a large amount of attention as a promising paradigm to reduce the burden on traditional healthcare services. As the population in the Western world is ageing, the prevalence of chronic disease is rising and the cost of healthcare increasing. Technology aided coaching on healthy behaviour (such as daily physical activity) can help prevent chronic diseases and influence the process of healthy ageing in general. Activity coaching is also potentially useful for everyone. The American College of Sports Medicine recommends in its 2011 position paper that every healthy adult engages in at least 30 minutes of moderate physical activity for five days per week [7]. Smart phones offer a unique opportunity to deliver coaching on physical activity to the user, as they become increasingly ubiquitous and are capable of running increasingly complex applications. They also contain built-in sensors, enabling context aware intelligent coaching through the use of location- and web based services. Research by, e.g., Bickmore [3] showed that personification of the user interface of coaching systems can have positive effects on the effectiveness of the coaching. Some examples of personification of the user interface of mobile coaching applications can be found in [8][9][10][11]. Some of these systems are distributed systems, while other system do not display (real-time) animations.

This paper presents the PictureEngine, a light-weight animation embodiment that enables our SAIBA-based BML realizer to be implemented and run stand alone on mobile Android devices. Compared to static pictures or pre-recorded movies, real-time animations are able to react immediately to the user, and this responsiveness increases the experience of engagement of the agent. Section II describes the Elckerlyc platform in more detail. The PictureEngine will be discussed in Section III, and the implementation of the platform on Android in Section IV. Section V presents the results of an evaluation of a design tool that helps designers in specifying their own ECA behaviour using the Behaviour Markup Language. This tool uses the PictureEngine to implement the multi-modal interactive behaviour specification. In Section VI two applications are discussed in which the Picture-Engine was integrated as part of the user interface. These applications are context-aware physical activity coaching applications. The ECA developed for these applications presents feedback of the digital coach by animated spoken interaction. We present some small user evaluation studies in which these coaching systems are evaluated. We conclude with describing future work on the development of the mobile embodied coach.

II. The Elckerlyc platform

In behaviour generation, at least two main aspects can be distinguished. The first aspect is the planning of the actions and movements as means to a certain goal that the agent intents to achieve. The second one is the actual detailed realization of the verbal and non-verbal behaviours in terms of "embodiments" of the (graphical) virtual human — including the generation of the speech by a text-to-speech synthesizer. This distinction between intent planning, behaviour planning and behaviour realization is the basis of the SAIBA framework (see [12] for more information about the SAIBA framework) [13]. According to this framework the detailed behaviours are specified in the Behaviour Markup Language [14].

The Elckerlyc platform is a BML realizer for real-time generation of behaviours of virtual humans (VHs). The Elckerlyc platform has been described and compared with other BML behaviour realizers (for example EMBR [15] and Greta [16]) in various papers [6][17][18].

Depending on the application and task that the intelligent system has, the virtual human presents for example the character of a tutor, an information assistant, or a conductor [19]. The goal is to make these embodied conversational agents look like believable and convincing communicative partners while interacting with humans. This requires the generation and coordination of "natural" behaviours and expressions.

Reidsma and Welbergen [18] discuss several features of the modular architecture of Elckerlyc and relates each of them to a number of user requirements. A general overview of the Elckerlyc system can be found in Figure 1. The input of the Elckerlyc platform is a BML specification. BML provides abstract behaviour elements to steer the behaviour of a virtual human. A BML realizer is free to make its own choices concerning how these abstract behaviours will be displayed on the embodiment. For example, in Elckerlyc, an abstract 'beat gesture' is by default mapped to a procedural animation from the Greta repertoire [20]. Greta is an expressive ECA that is able to show complex emotions. The developer may want to map the same abstract behaviour to a different form, i.e., to a high quality motion captured gesture [18]. The Elckerlyc system is easy to extend with other engines. Different Engines will handle their own parts of the behaviour specification and generate synchronized instructions for realizing, i.e., speech output, body gestures, postures and facial expressions. The output of all the engines are displayed on one embodiment, like a realistic 3D full kinematic virtual human, the Nabaztag or a graphical 2D cartoon like picture animation. Figure 2 shows three examples of embodiments supported by the Elckerlyc platform.

Not every embodiment is able to render all the behaviours that can be specified in BML. This depends on what the embodiment offers, i.e., a robot that is not able to smile or a picture animation that lacks a picture showing the smiling face cannot render the requested smiling behaviour. The interface between the output of Elckerlyc and the embodiment occurs in a Binding. A Binding is an XML

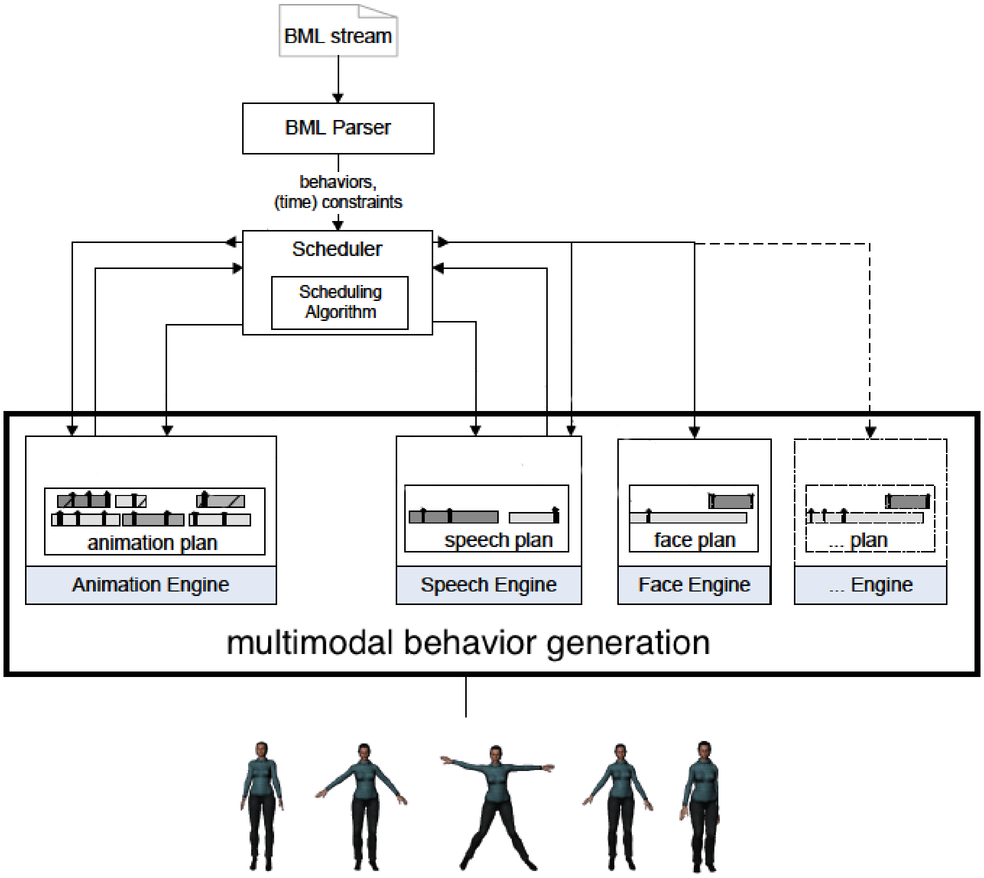


Figure 1. Overview of the Elckerlyc architecture. BML input is processed by the Elckerlyc system by different engines. The result is combined into one embodiment. New Engines, or Engines developed by others are indicated by dash lines.

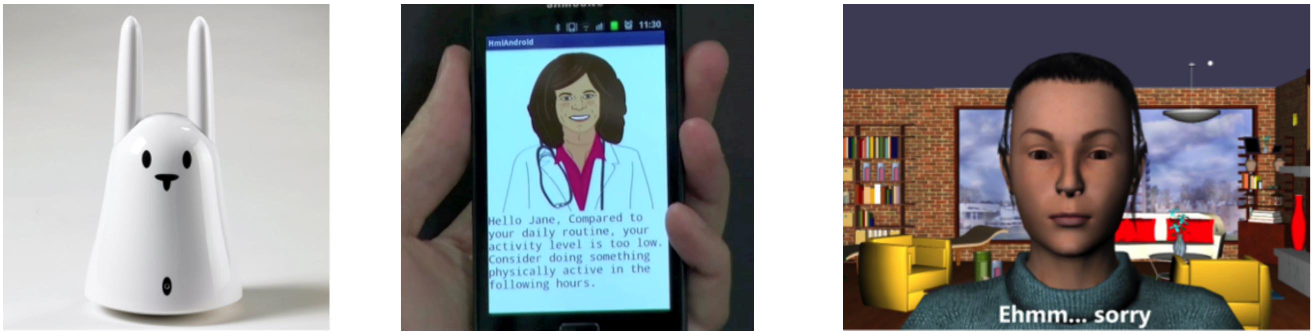
description to achieve a mapping from abstract BML behaviours to PlanUnits that determine how the behaviour will be displayed in the embodiment. Bindings can be customized by the application developer. Other Engines provide similar bindings.

This paper discusses how this feature was exploited. A light-weight PictureEngine was developed that makes it possible to run Elckerlyc on mobile Android platforms. Elckerlyc allows for a transparent and adjustable mapping from BML to output behaviours (rather than the mostly hard coded mappings in other realizers), and allows for easy integration of new modalities and embodiments, for example to control robotic embodiments, or full 3D embodiments. The PictureEngine that was developed allows rendering of behaviours and expressions using layers of pictures. The next section will discuss the PictureEngine in more detail.

III. THE PICTUREENGINE

Using a realistic 3D full kinematic virtual human embodiment is not suitable for use on mobile devices for multiple reasons. Not only do such devices lack the processing power to render this kind of environment, but displaying a full scene including a full body ECA on the relatively small screen of a mobile device is quite impractical. The displayed size of the ECA would make it so small that its expressions would hardly be visible. The high processing demands would also drain the devices' battery quickly. In order to avoid all these problems, Elckerlyc uses a different graphical embodiment on the Android platform, the PictureEngine.

The PictureEngine is a lightweight graphical embodiment that uses a collection of 2D images in order to display the ECA. While having a 2D image embodiment offers some limitations, it also has its advantages. First of all, it has low demands in terms of processing and memory. It also allows for great variation in the design of ECA's. One could for example design a cartoon figure ECA, an ECA based on more lifelike illustrations, or even an ECA based of photographic images of a real person, or pre-rendered images of a 3D lifelike ECA. Creating your own 2D ECA starts with designing a set of images for the appearance of



(a) The Nabaztag rabit (b) a 2D cartoon like picture animation (c) a realistic 3D full kinematic virtual human

Figure 2. Three types of embodiment used as back-end for the Elckerlyc platform

the ECA in different layers. More detailed information about layers can be found in Section III-A. Blinking, lip-sync and other nonverbal behaviour can be designed as a set of animations. Animations are designed by a subset of pictures that together can execute the animation and are defined in a small XML file. Section III-B will explain the animations in more detail. When these different images and animations are designed they have to be specified in a PictureBinding which will be explained in Section III-C.

A. Layers

In order to generate a dynamic ECA from a collection of images, the PictureEngine uses a layer-based approach. Different parts of the ECA are displayed on different layers of the final image, and can thus be in different states. For example, one layer may contain the eyes, while another contains the mouth. Figure 3 shows an example of how a couple of layers will result in a face of the ECA. The base layer normally contains the ECA in a base state, meaning that when the ECA is in a neutral or passive state, the user sees only this base layer. While each (facial) feature of the ECA does have its own layer, they are also present in the base layer. The base layer contains for example a full face with a neutral expression, even though the eyes and mouth may have their own layers. There can also be layers containing features that are not visible in the base state, such as hands that only move into the frame when executing a gesture. By using this layered approach, different parts of the ECA can be manipulated independently and combined in order to generate different expressions. This also allows the ECA to do several (connected or unconnected) things at once, such as blink while also speaking and pointing at something.

The layered approach does present some limitations. Since the features of the ECA are in separate layers, the base onto which these features are displayed (usually a face, and possibly part of the body) is generally static, so any movement of the entire ECA poses a problem. When an ECA contains facial features on different layers, the layered structure prevents it from moving around. This also applies

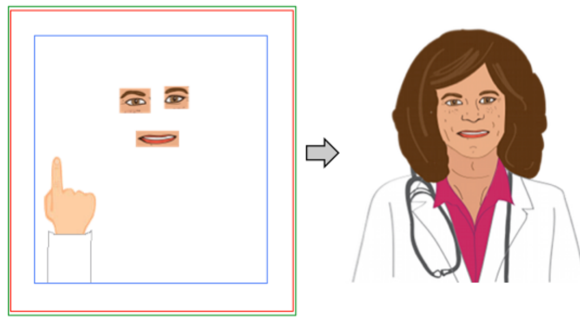


Figure 3. An example of different layers that finally results in the face of a ECA.

to smaller movements such as nodding, shaking and tilting of the head. However, because the PictureEngine is designed to be used on smaller screens, the ECA will generally be displayed as a talking head, a close-up of a face covering most of the available screen space. Subsequently, having the ECA perform locomotion is already impractical and, since there is hardly any room for the ECA's environment to contain anything but itself, arguably unnecessary.

B. Animations

While single images may suffice for portraying expressions in many cases, there are other cases where an ECA simply has to display some motion in order to come across as believable. To make this possible, the PictureEngine also allows the use of animations instead of single images. These animations are defined by using a simple XML format that allows a number of images to be listed, together with the duration for which they are to be displayed. While these durations are specified in seconds, the nature of the BML scheduler allows the duration of animations to be adjusted according to the BML code that is being realized, causing the animation to play faster or slower depending on the timespan determined by the scheduler.

An additional feature of these animation XML files that provides an advantage over using an already established

```xml
<PictureUnitSpec type="face">
    <constraints>
        <constraint name="type" value="LEXICALIZED"/>
        <constraint name="lexeme" value="smile"/>
    </constraints>
    <parametermap>
    </parametermap>
    <parameterdefaults>
        <parameterdefault name="filePath" value="animations/"/>
        <parameterdefault name="fileName" value="smile.xml"/>
        <parameterdefault name="layer" value="8"/>
    </parameterdefaults>
    <PictureUnit type="AddAnimationXMLPU"/>
</PictureUnitSpec>
```

Figure 4. PictureBinding entry for a smile.

format for image animations is the possibility to include synchronization information in the animation specification. Between any two frames of an animation, a synchronization point can be included in the specification. These synchronization points are available for use in the main BML code. In this way, it is possible to, e.g., synchronize the stroke of a beat gesture animation with a certain word within a speech element.

C. PictureBinding

In the same way that the other engines uses bindings, the PictureEngine uses a PictureBinding. This PictureBinding allows a combination of a BML behaviour class and (optionally) several constraints to be mapped to a certain image or animation. It is possible to include anywhere from zero constraints to all the constraints defined by the corresponding BML behaviour type. This allows the designer of a PictureEngine ECA to refine those behaviours that are most relevant to the ECA, and implement any others in a more general fashion.

The actual PictureBinding itself is defined in an XML file containing the behaviour classes and constraints and the PictureUnits and parameters they are to be mapped onto (see Figure 4 for an example). The accessibility of this format allows an ECA to be designed or modified by someone who does not have knowledge of the inner workings of Elckerlyc. Only knowledge of BML and the available PictureUnits and their parameters is required to be able to build a complete PictureBinding.

D. Lip-sync

In order to visually display the fact that the ECA is speaking, the PictureEngine provides a rudimentary lip-sync facility. This lip-sync feature is implemented in the same way as the lip-sync provided by the AnimationEngine. However, where the AnimationEngine provides a full mapping from visemes to animation units, the PictureEngine lip-sync currently does not make use of such a mapping (although it could be added in the future). In its current state the lip-sync allows a single animation to be specified which is played whenever the ECA is speaking. This animation is repeated for the number of times it fits into the duration of the speech unit (and slightly adjusted so that the amount of repetitions becomes a round number).

IV. Android implementation

Since the Elckerlyc platform is implemented almost entirely in Java, all of its core elements run on Android without any modification. However, since Android has its own environment for visual and audio output, some additions are required. This does not mean that the Android application uses a modified version of the core Elckerlyc platform. The fact that Elckerlyc uses an XML format to define the loading requirements for a specific ECA allows the Android application to simply load its own versions of a few key components. This allows the core Elckerlyc system to be used in the Android application as-is, so any changes to the Elckerlyc core can be directly used in the Android application without having to modify or port it first. The Android application requires Android Gingerbread (2.3) or higher. The subsystems for which the Android application contains its own versions are discussed in this section.

A. Graphical Output

The Android platform has its own graphical environment. Therefore the engines that provide graphical output use a modified component for printing this output in the Android application. This goes for both the PictureEngine, which handles the graphical display of the ECA, and the TextSpeechEngine, which outputs speech elements to a text area. Since PNG images can be handled without problems by the Android graphical environment, the additional code needed to replace the PictureEngines' default output subsystem with a version that works on Android is minimal. Displaying plain text is a basic function in Android.

B. SpeechEngine

In the case of the SpeechEngine (for the rendering of spoken text using text-to-speech(TTS)) the differences with Android are unfortunately more severe. The TTS engines used in the PC version of the SpeechEngine contain several dependencies on native PC systems and cannot be used on Android without significant changes. However Android does offer an internal TTS system. Using this internal system avoids the costly process of porting a TTS engine and any possible efficiency issues this may bring. In order to make use of the internal Android TTS system, an Android adaptation of the Elckerlyc SpeechEngine is needed. This includes the module that loads and initializes the engine, as well as the parts of the system dealing with the actual TTS operations.

The main problem with the Android TTS system is that it is not possible to obtain timing information for utterances, meaning there is no way to find out exactly at what time a word is spoken. This causes the BML scheduler to be unable to use synchronization points within utterances. This makes it hard to precisely synchronize other behaviours with specific words being spoken. A partial solution is that utterances are pre synthesized to a file in order to find the total duration of the utterance. This provides the crucial information for the Elckerlyc scheduler. This "preloading" of utterances causes a delay at start up before the ECA starts playback of the requested BML code. Furthermore, the TTS also does not offer any viseme information, making it impossible to use real lip-sync on Android. This is the main reason the PictureEngine on Android does not currently support true lip-sync.

C. Subtitles

Because the PictureEngine can run on a mobile device, the chances of the user having trouble hearing the text spoken by the TTS on the Android system are quite high. This could be caused by factors such as environment noise, low volume or bad speakers. In order for the user to still be able to interact with the ECA in these situations, the Android application also offers an on-screen representation of any spoken text, comparable to subtitling. The TextSpeechEngine (on-screen text display) receive the text handled by the SpeechEngine and displays spoken text to in a text area, synchronized (per utterance) with the TTS.

V. User evaluation with the PictureEngine

This section presents a user evaluation of a BML editor that uses the PictureEngine. The goal of this user evaluation was to investigate if it was possible for non expert users to specify the verbal and non-verbal behaviour of an ECA in BML. Users had to create their own BML script following a step-by-step task description. This results in verbal and non-verbal behaviour of an ECA which is displayed by means of the PictureEngine. With this evaluation we tried to find out what problems users meet when creating a BML script. We also wanted to know how users assessed the usability of the BML editing tool.

A. Procedure

Participants were students and employees of the University of Twente as well as partners from the Smarcos project. Participants had to simply download and install the BML editor including the PictureEngine. An editor screen is used to edit the BML script and a feedback screen shows feedback and error messages (see Figure 5). On a separate screen to the ECA is displayed (see Figure 6). Participants started with an initial BML script and added modifications step by step following the assignments. After finishing the assignments participants could check their specification by comparing the resulting behaviour with that shown in a movie available on YouTube. The BML script had to be sent back to the evaluators and the participant had to fill in an online questionnaire.

Figure 6. The ECA screen.

B. Materials

An initial BML script was prepared and given to the users. The participants had to add modifications to this script as described in the assignments. The assignments were selected in such a way that the participants had to discover all the functionalities of the PictureEngine available in the current version. The completed task would show the ECA introducing herself and let her do some physical exercises. Verbal and non-verbal behaviour should be synchronized together with different kind of gestures and gaze behaviour.

The online questionnaire that was filled in after the test is divided in several sections:

1) A section devoted to collecting some general background information about the users (gender, age, education, organisation in which the user works and his/her role in it, knowledge about ECAs and experience with designing/developing ECAs.
2) A section focussing on the process of making the assignments, in particular, how much time it took to complete all the assignments and how easy it was to complete the different parts of the assignment.
3) A section devoted to the usability of the user interface of the PictureEngine. In this section the users had to rate the clearness of different parts of the UI of the PictureEngine and were asked to provide suggestions for possible improvements.
4) A section focussing on the feedback provided by the PictureEngine, in which the users had to rate how clear they found the feedback messages provided by the PictureEngine.
5) A section where users were asked to give any further comments and suggestions.

C. Results

This section presents the results of the questionnaires and discusses the remarks returned by the participants.

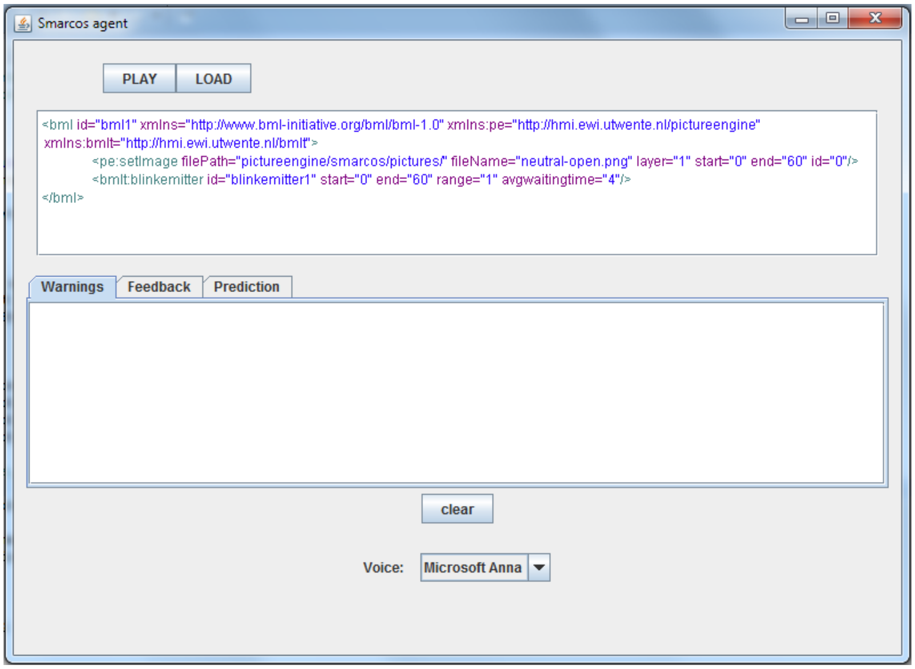


Figure 5. The BML editor screen

1) Participants: Twenty one persons participated in this user evaluation (5 females and 16 males), aged between 21 and 60, with an average age of 29. All participants were students or otherwise involved in research or education: from master students in Human Media Interaction, software engineers, to research or education in interface design and interactive systems, psychology and telemedicine. Twelve participants (5 female, 7 male) did not have any previous experience working with ECAs or specifying behaviour of embodied conversational agents and nine did have some experience with ECAs.

2) Questionnaires: Participants spend between 10 and 45 minutes to perform the whole user evaluation $\mu = 25$, $\sigma = 9.4$) The participants were asked to fill in answers to the following questions on a 10-point Likert scale. The questions, mean values and standard deviations are presented below.

Participant had to rate statements about the easiness of adding gestures and gaze behaviour, the synchronisation of gestures and speech, and the assignment in general. Results of these questions can be found in table V-C2. Questions were answered on a 10 point Likert scale where 1 is "very hard" and 10 "very easy".

From the figures in table V-C2 we conclude that overall the participants did not have many problems in completing the tasks.

3) Remarks: The participants were asked about what they found useful extensions of the capabilities to add sentences, gestures or gaze behaviours. Participants want to have an auto-complete function and like to get some hints and tips while writing their BML script. To add gestures or gaze behaviour, participants like to select predefined gestures or gazes from a menu or list. The user interface of the editor tool should have bigger text field to enter and edit the BML script. Participants want to have a stop and pause button to stop or pause the execution of the BML script. The error messages from the Elckerlyc system should be less complex and should not show too much details and the participants want to see line number in the error message. Errors should be highlighted in the text field.

Table I
SCORES OF COMPLEXITY QUESTIONS ON A 10 POINT LIKERT SCALE (1 IS VERY HARD, 10 IS VERY EASY)

Question	Mean	SD
How easy was it to add the gestures?	7.9	1.9
How easy was it to add the gaze behaviour?	8.1	2.0
How easy was it to synchronise gesture, gaze and speech?	7.5	2.5
How easy was it to finish the assignment?	8.0	1.5

D. Conclusion

This small first user evaluation shows that BML together with the PictureEngine and the editor tool offers good possibilities to specify and test verbal and non-verbal behaviour of an ECA for mobile devices and can help developing more natural interactions with mobile user interfaces.

In general the participants were able to write and run their own BML script, even the participants that indicated not to have any experience with ECAs or did not have a technical background or programming skills.

From the comments and results of the questionnaires it became clear that the editor of the PictureEngine needs some improvements to make it more easy to create and run a BML script with respect to the design of the user interfaces, and feedback and error messages. The user evaluation with the PictureEngine made use of the general desktop user interface of the Elckerlyc platform. Changing the editor to a Eclipse plugin for better support and feedback while creating a BML script can be interesting improvement to the PictureEngine. Having a plug-in for Eclipse will also makes it possible to upload and run the BML script directly to a mobile Android device.

VI. Applications

With the growing availability of online services and ubiquitous computing capabilities it becomes easier to develop systems that can present people information about their daily behaviours. This may help them to manage their lifestyle [21]. Sensor data and context information is available everywhere at any time. Some of these systems support people in their daily life by means of a human or digital coach. These systems can support users in coping with chronic diseases like COPD [22] and diabetics, but also to be more physical active [11][23]. Persuasive systems [24], and especially behaviour change support systems are information systems designed to form, alter or reinforce attitudes, behaviours or an act of complying without using deception, coercion or inducements [25].

The next sections present two different behaviour change support systems in which the PictureEngine was integrated in the user interface. Personalisation of the user interface by means of an ECA may affect the effectiveness of the behaviour change program and the user experience. Results from other studies indicate that the use of an ECA in a persuasive system has a positive effect on how the feedback is received by the user and on the results of the coaching program [26][27][28].

In the Smarcos application (see Section VI-A) the PictureEngine is integrated in the mobile Android devices that are part of the system. The C3PO system [22] described in Section VI-B integrated the PictureEngine on mobile Android devices and desktop PCs. Users were able to interact with the system on different systems and switch between the devices during the interaction.

A. Smarcos

In the EU Artemis project Smarcos we developed a personal digital health coach that supports users in attaining a healthy lifestyle by giving timely, context-aware feedback about daily activities through a range of interconnected devices [29]. The two targeted user groups of the coaching system are office workers and diabetic type II patients. Office workers will receive feedback about their physical activity level, while diabetic type II patients also receive feedback about their medication intake. Physical activity is measured by a 3D accelerometer and medication intake is tracked by a smart pill dispenser. The pill dispenser is using the mobile network to connect to the internet. The system is context-aware and multi-device which means that the (digital) coach can support the users in various contexts and on different devices. GPS information is provided by the mobile phone of the user. The system sends feedback to the mobile phone of the user, the laptop or PC and their television.

Figure 7 shows the overall architecture of the coaching system. All information from the monitoring devices and manual input from the users are uploaded to the cloud and stored in a central knowledge base. The coaching engine contains coaching rules and continuously keeps track of all user data. When the coach receives a trigger it starts to evaluate the coaching rules. When one of the rules is evaluated with a positive result, it will select a suitable message from the coaching content database and send the message to the user through one of the available output devices. Output devices that are able to run a BML realiser, like the PictureEngine, are called BML enabled devices. BML enabled devices in the Smarcos coaching system are delineated in purple in figure 7. These devices are able to present feedback by an animated spoken interaction with an ECA. Feedback from the system can be presented as a text message, by means of a graph or by an ECA.

A first user evaluation with a basic version of the Smarcos personal digital health coach compared two alternatives for providing digital coaching to users of a physical activity promotion service. Participants in the study (N=15) received personalized feedback on their physical activity levels for a period of six weeks. Feedback was provided weekly either by e-mail or through an embodied conversational agent. The messages by the ECA were prerecorded video message. User's perception of the digital coaching was assessed by means of validated questionnaires after three weeks and at the end of the study. Results show significantly higher attractiveness, intelligence and perceived quality of coaching for the ECA coach.

B. C3PO

The Telemedicine group of Roessingh Research and Development (RRD) has over the past few years been working on a technology platform for supporting physical activity behaviour change in patients suffering from various chronic diseases, as well as for healthy individuals. This platform, called the *Continuous Care & Coaching PlatfOrm*, or C3PO, consists of a 3D-accelerometer based sensor, a Smart phone, back-end Server and connected Web portals (see Figure 8). The platform has been used successfully in trials with

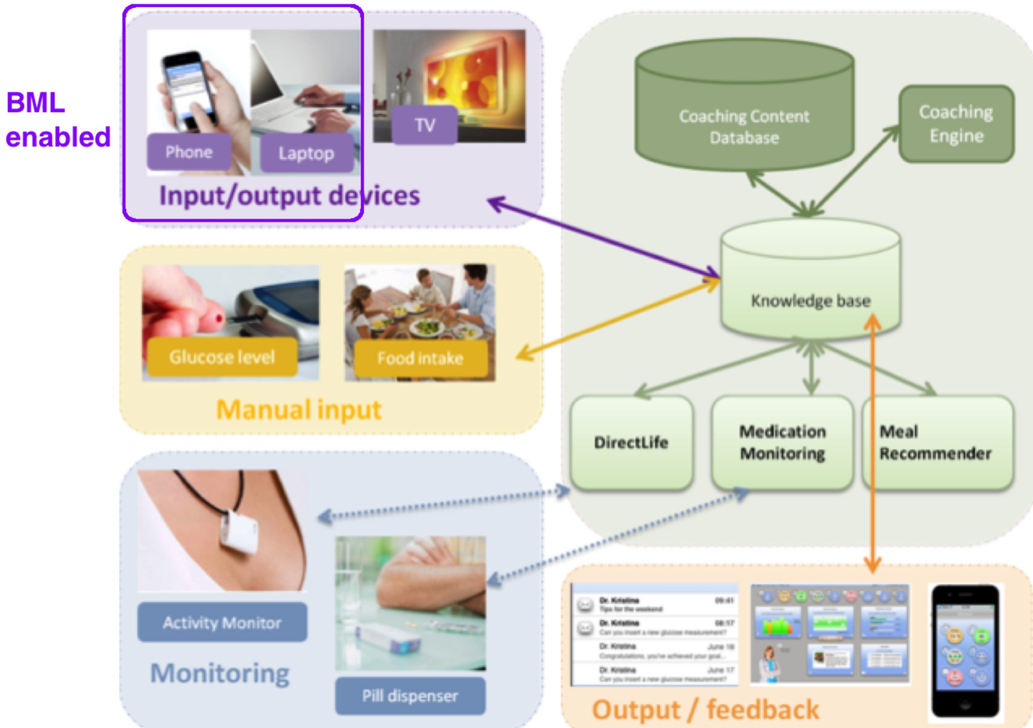


Figure 7. Overview of the Elckerlyc architecture. BML enabled output devices are marked in purple.

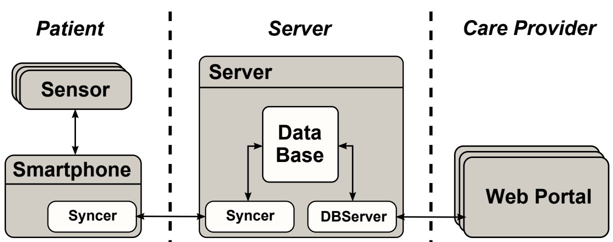


Figure 8. High level overview of the C3PO coaching platform [30].

patients suffering from Chronic Obstructive Pulmonary Disease (COPD), Chronic Low Back Pain (CLBP), Chronic Fatigue Syndrome (CFS), and Obesity [30], and is constantly under development to increase its effectiveness in real-time, tailored coaching. While earlier research focused on tailoring the delivery of motivational messages to the user in terms of timing and content (see, e.g., [22]), the visual representation of the feedback has been largely ignored. Due to the modular architecture of the Smart phone application, we were able to quickly integrate the PictureEngine in order to enable a more natural communication to the patient through the use of the ECA.

Two separate experiments where carried out to evaluate the PictureEngine integrated with the C3PO platform. The first is a controlled experiment to evaluate user perception and experience with receiving feedback from the ECA compared to the regular text-based user interface. For the second experiment, a more complex system was developed, similar to the Smarcos system described in the previous section, where feedback was presented to the user on various interconnected devices, including two BML enabled devices (PC and Smart phone). For both experiments, the target user groups where healthy office workers.

The first user evaluation comparing the use of an ECA to the standard text feedback message interface included 14 participants, aged between 22 and 61 ($\mu = 37$, $\sigma = 13.3$) and consisting of 8 males and 6 females. Participants where randomly assigned to either the text-first condition, in which they received the standard text based interface in the first week, and the ECA in the second week, or the ECA-first condition. All participants finished the evaluation, with 1 participant not being able to complete full measurement days due to a faulty sensor. In both conditions the system generated a *motivational cue* message every hour, based on the user's current activity progress compared to a predefined reference activity pattern. When asked about the user's preference for either of the two conditions, only 3 users preferred the ECA, 10 users preferred the text-only condition, and 1 user had no preference. The single most important reason given by the users for preferring the text-only condition is *glanceability*. As the ECA pronounces the feedback message with real-time subtitling (letter for letter), it takes a much longer time to convey the entire message compared to the text-based interface, where users can read it immediately. When asked directly about the ECA, users responses where varied. Four participants found the ECA fun and enjoyable, and three participants said the ECA added personality to the system. On the negative side it was commented that the ECA does not add anything significant to the system, and that the ECA was not believable because it was not a real person. Two participants commented that the ECA did not show enough enthusiasm, while another thought the ECA was too enthusiastic to the point of it not being believable any more. With only 3 out of 14 users preferring the ECA over the simple text version it can be concluded that the result of the evaluation is not in favour of using ECA's in this application. However, besides the fact that obvious improvements can still be made (such as better graphics and voice output), an interesting observation is the dichotomy between user's opinion about the ECA as personification of the application. This dichotomy was also mentioned as a key outcome of a study on Embodied Agents in 1996 by Koda & Maes [31]. It seems that the perception of an ECA is highly personal. Comments regarding future improvements to the system from the users are in line with this, which include suggestions about making the ECA more relevant to the target audience, enabling users to choose between different visual/auditory styles of the ECA, and enabling the user to choose whether or not to use the ECA at all.

One of the advantages of ECAs as user interface is the personification aspect. At least for some users, the ECA gives a certain personality to an application. This personality aspect can be exploited in the area of multi-device user interfaces to overcome some of the challenges related to continuity. Multi-device user interfaces exist in many different forms and levels of complexity. Using Paternó & Santoro's [32] framework and terminology, our multi-device platform can be described as a system that supports *UI Migration*, automatic *Trigger Activation* and *Multi-modal* devices using a client/server architecture. For the second evaluation we have implemented a multi-device component to the C3PO platform. For the target population of office workers, the idea is that throughout a regular work day, the user communicates with various devices that each offer unique capabilities in terms of physical activity coaching. While performing desk work, the PC would be the most suitable device for delivering communication from the system to the user, if the user is getting a coffee, a public screen mounted next to the coffee machine can offer coaching through social influencing, and while the user is travelling, the Smart phone can take over the coaching role for ubiquitous availability. In order to accommodate the virtual coach migrating across devices with the user, we have added a server component that manages so-called *user-requests*. Whenever a particular output device (PC, Smart phone, Public Screen) notices that the user is near it and available for communication, the device would send a user-request to the server. The server then decides which is the most suitable device for coaching, and notifies the devices of their ability to engage with the user. Also implemented for this evaluation is the ability to engage in short dialogues with the ECA on the PC or Smart phone. Users were presented with spoken questionnaires and were able to use speech input to answer the (multiple choice) questions. In order to accommodate user preferences regarding UI, the Smart phone supports either the ECA and a regular text-based interface for the questionnaires, between which the user can switch at any time. Six participants took part in the evaluation of this multi-device version of the coaching platform, 4 females and 2 males. The evaluation was small and focused on usability testing and a "thinking aloud" procedure. Users were observed while performing a set of tasks involving desktop computer work, going to the coffee machine, and walking around the office. Most of the findings from this early stage evaluation were related to the technical working of the system or missing functionalities, however from a usability perspective it again became clear that there are large differences between user preferences for the device selection, and user-interface selection.

From our experiences with both these applications, it became clear that the use of the PictureEngine can be a useful tool for tailoring the user interface. However, due to the large differences in users, the option of switching to a more classic interface (either automatically, or through user selection) should always be supported. As with other forms of tailoring, such as automatically adapting the coaching style (formal vs. informal), it remains an open problem how to automatically match the right user interface representation to the right user.

VII. Conclusion and future work

To take the full advantage of the known benefits of personification of the user interface of service systems a mobile platform that is able to present embodied conversational agents in mobile applications is presented. The platform makes use of the Elckerlyc system. Because it is too heavy to render realistic 3D virtual humans on mobile devices a lightweight PictureEngine was developed. The PictureEngine makes it possible to use the Elckerlyc system on the Android platform and to generate real-time animations of embodied conversational agents.

The PictureEngine is applied as user interfaces in three coaching applications. Initial results of short term user evaluations showed a large variety in the response of users to the ECA, in line with earlier research in this area by Koda & Maes [31]. As for some users the ECA was perceived very positively (adding personality to the system and increasing the feeling of consistency in the multi-device application), we believe that the use of ECA's should be offered as optional component in such coaching applications. Also, additional work is needed in allowing the user to make personal choices regarding the appearance of the ECA, both visually and in terms of voice and character.

When such modifications are in place, long term user evaluations with these coaching platforms, including the mobile embodied coach, are planned to investigate the effects of personalised coaching feedback on user experience, quality of coaching and effectiveness of the coaching program.

The results of user evaluation showed that BML together with the PictureEngine and the editor tool offers good possibilities to specify and test verbal and non-verbal behaviour of an ECA for all participants. All the participants, including the non expert users without any programming skills were able to complete the assignment.

Although it is shown that the PictureEngine can run on mobile Android devices it would be worth exploring options for using a different TTS system in the future. This would allow to regain the speech-related functionality that is currently unavailable on Android, such as synchronization within utterances and viseme-based lip-sync. A next step in the development of the PictureEngine is looking for techniques to allow small movements by the ECA, such as nodding and shaking of the head. These movements will make it possible for the ECA to show those communicative behaviours that have shown to be effective for turn-taking and attention signalling in real-time interactions.

ACKNOWLEDGMENT

This work was funded by the European Commission, within the framework of the ARTEMIS JU SP8 SMARCOS project 100249 - (www.smarcos-project.eu, last access date 28-06-2013).

The PictureEngine application for Android is released under GPL3 license. The PictureEngine is an Android implementation of the Asap BML realizer which is already available through http://asap-project.ewi.utwente.nl (last access date 28-06-2013).

REFERENCES

[1] R. Klaassen, J. Hendrix, D. Reidsma, and H. J. A. op den Akker, "Elckerlyc goes mobile: enabling technology for ECAs in mobile applications," in *The Sixth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, Barcelona, Spain*. XPS (Xpert Publishing Services), September 2012, pp. 41–47.

[2] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, B. Achorn, T. Becket, B. Douville, S. Prevost, and M. Stone, "Animated conversation: rule-based generation of facial expression, gesture & spoken intonation for multiple conversational agents," in *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '94. New York, NY, USA: ACM, 1994, pp. 413–420.

[3] T. Bickmore, D. Mauer, F. Crespo, and T. Brown, "Persuasion, task interruption and health regimen adherence," in *Proceedings of the 2nd international conference on Persuasive technology*, ser. PERSUASIVE'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 1–11.

[4] M. Argyle, *Bodily Communication*, MethuenEditors, Ed. Methuen, 1988, vol. 2nd.

[5] D. Traum, S. Marsella, J. Gratch, J. Lee, and A. Hartholt, "Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents," in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, H. Prendinger, J. Lester, and M. Ishizuka, Eds. Springer Berlin / Heidelberg, 2008, vol. 5208, pp. 117–130.

[6] H. van Welbergen, D. Reidsma, Z. M. Ruttkay, and J. Zwiers, "Elckerlyc - a BML realizer for continuous, multimodal interaction with a virtual human," *Journal on Multimodal User Interfaces*, vol. 3, no. 4, pp. 271–284, August 2010.

[7] C. E. Garber, B. Blissmer, M. R. Deschenes, B. A. Franklin, M. J. Lamonte, I.-M. Lee, D. C. Nieman, and D. P. Swain, "American College of Sports Medicine position stand. Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: guidance for prescribing exercise." *Medicine & Science in Sports & Exercise*, vol. 43, no. 7, pp. 1334–1359, 2011.

[8] F. Buttussi and L. Chittaro, "Mopet: A context-aware and user-adaptive wearable system for fitness training," *Artificial Intelligence in Medicine*, vol. 42, no. 2, pp. 153 – 163, 2008.

[9] M. Turunen, J. Hakulinen, O. Stahl, B. Gamback, P. Hansen, M. C. R. Gancedo, R. S. de la Cmara, C. Smith, D. Charlton, and M. Cavazza, "Multimodal and mobile conversational health and fitness companions," *Computer Speech & Language*, vol. 25, no. 2, pp. 192 – 209, 2011.

[10] M. W. Kadous and C. Sammut, "Inca: A mobile conversational agent," in *PRICAI 2004: Trends in Artificial Intelligence*, ser. Lecture Notes in Computer Science, C. Zhang, H. W. Guesgen, and W.-K. Yeap, Eds. Springer Berlin / Heidelberg, 2004, vol. 3157, pp. 644–653.

[11] O. Ståhl, B. Gambäck, M. Turunen, and J. Hakulinen, "A mobile health and fitness companion demonstrator," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, ser. EACL '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 65–68.

[12] (2013) Mindmakers.org. [Online]. Available: http://www.mindmakers.org (last access date 28-06-2013)

[13] E. Bevacqua, K. Prepin, E. de Sevin, R. Niewiadomski, and C. Pelachaud, "Reactive behaviors in SAIBA architecture," in *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, S. Decker, Sichman and Castelfranchi, Eds., May 2009.

[14] H. Vilhjalmsson, N. Cantelmo, J. Cassell, N. E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. N. Marshall, C. Pelachaud, Z. Ruttkay, K. R. Thórisson, H. V. Welbergen, and R. J. V. D. Werf, "The behaviour markup language: recent developments and challenges," in *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, ser. Electronic Notes in Artificial Intelligence, vol. 4722. Springer, 2007, pp. 99–11.

[15] A. Heloir and M. Kipp, "Real-time animation of interactive agents: specifciation and realization," *Applied Artificial Intelligence*, vol. 24(6), pp. 510–529, 2010.

[16] E. Bevacqua, M. Mancini, R. Niewiadomski, and C. Pelachaud, "An expressive ECA showing complex emotions," in *AISB'07 Annual convention, workshop "Language, Speech and Gesture for Expressive Characters"*, April 2007, pp. 208–216.

[17] D. Reidsma, I. de Kok, D. Neiberg, S. Pammi, B. van Straalen, K. Truong, and H. van Welbergen, "Continuous interaction with a virtual human," *Journal on Multimodal User Interfaces*, vol. 4, no. 2, pp. 97–118, 2011.

[18] D. Reidsma and H. v. Welbergen, "Elckerlyc in practice on the integration of a BML realizer in real applications," in *Proc. of Intetain 2011*, 2011.

[19] D. Heylen, M. Theune, H. J. A. op den Akker, and A. Nijholt, "Social agents: the first generations," in *ACII 2009: Affective Computing & Intelligent Interaction*, M. Pantic and A. Vinciarello, Eds. Los Alamitos: IEEE Computer Society Press, September 2009, pp. 114–120.

[20] E. Bevacqua, M. Mancini, R. Niewiadomski, and C. Pelachaud, "An expressive eca showing complex emotions." in *AISB'07, Artificial and Ambient Intelligence*, Newcastle, UK., 2007.

[21] M. Kasza, V. Szücs, A. Végh, and T. Török, "Passive vs. active measurement: The role of smart sensors," in *UBICOMM 2011, The Fifth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, 2011, pp. 333 – 337.

[22] H. op den Akker, V. Jones, and H. Hermens, "Predicting feedback compliance in a teletreatment application," in *Proceedings of ISABEL 2010: the 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies*, 2010.

[23] G. Geleijnse, A. van Halteren, and J. Diekhoff, "Towards a mobile application to create sedentary awareness," in *In Proceedings of the 2nd Int. Workshop on Persuasion, Influence, Nudge Coercion Through Mobile Devices (PINC2011)*, vol. 722, 2011, pp. 90–111.

[24] B. Fogg, *Persuasive Technology. Using computers to change what we think and do*. Morgan Kaufmann, 2003.

[25] H. Oinas-Kukkonen, "Behavior change support systems: A research model and agenda," in *Persuasive Technology*, ser. Lecture Notes in Computer Science, T. Ploug, P. Hasle, and H. Oinas-Kukkonen, Eds. Springer Berlin / Heidelberg, 2010, vol. 6137, pp. 4–14.

[26] O. A. B. Henkemans, P. van der Boog, J. Lindenberg, C. van der Mast, M. Neerincx, and B. J. H. M. Zwetsloot-Schonk, "An online lifestyle diary with a persuasive computer assistant providing feedback on self-management," *Technology & Health Care*, vol. 17, p. 253–257, 2009.

[27] D. Schulman and T. W. Bickmore, "Persuading users through counseling dialogue with a conversational agent," in *Persuasive '09: Proceedings of the 4th International Conference on Persuasive Technology*. New York, NY, USA: ACM, 2009, pp. 1–8.

[28] D. C. Berry, L. T. Butler, and F. de Rosis, "Evaluating a realistic agent in an advice-giving task," *Int. J. Hum.-Comput. Stud.*, vol. 63, no. 3, pp. 304–327, 2005.

[29] R. op den Akker, R. Klaassen, T. Lavrysen, G. Geleijnse, A. van Halteren, H. Schwietert, and M. van der Hout, "A personal context-aware multi-device coaching service that supports a healthy lifestyle," in *Proceedings of the 25th BCS Conference on Human-Computer Interaction*, ser. BCS-HCI '11. Swinton, UK, UK: British Computer Society, 2011, pp. 443–448.

[30] H. op den Akker, M. Tabak, M. Marin-Perianu, R. M. Huis In't Veld, V. M. Jones, D. Hofs, T. M. Tonis, B. W. van Schooten, M. M. Vollenbroek-Hutten, and H. J. Hermens, "Development and Evaluation of a Sensor-Based System for Remote Monitoring and Treatment of Chronic Diseases - The Continuous Care & Coaching Platform," in *Proceedings of the Sixth International Symposium on e-Health Services and Technologies (EHST 2012)*, Geneva, Switzerland, 2012, pp. 19–27.

[31] T. Koda and P. Maes, "Agents with Faces : The Effect of Personification," in *Proceedings of the 5th IEEE International Workshop on Robot and Human Comminication*, 1996, pp. 189–194.

[32] F. Paternò and C. Santoro, "A logical framework for multi-device user interfaces," *Proceedings of the 4th ACM SIGCHI symposium on Engineering interactive computing systems - EICS '12*, p. 45, 2012.

Smart Navigation in Intelligent Transportation Systems: Service Performance and Impact on Wireless Networks

Alessandro Bazzi, Barbara Mavì Masini, Gianni Pasolini, and Oreste Andrisano
CNR-IEIIT and DEI, University of Bologna at Wilab, Bologna, Italy
Email: alessandro.bazzi@cnr.it, barbara.masini@ieiit.cnr.it, gianni.pasolini@unibo.it, oreste.andrisano@unibo.it

***Abstract*—Wireless communications are nowadays considered as enablers of innovation in the field of smart mobility in smart cities. In this work, we focus on the smart navigation service, which is aimed at providing drivers with the best route to destination taking into account real time traffic conditions. Smart navigation is increasingly used today and expected to reduce traffic congestions, but the real impact on travel time and the cost in terms of wireless network resources are still open issues. These aspects are here discussed starting from the objectives and the outputs of the Italian project PEGASUS. More specifically, to what extent this application can reduce the travel duration and how frequently traffic information must be updated will be firstly discussed; then, the impact on wireless networks of both the uplink collection of traffic information and the downlink transmission to vehicles is shown, focusing on the UMTS cellular technology; finally, the use of short range IEEE 802.11p wireless communications technology is investigated to offload cellular networks. Through simulations performed in a dense urban scenario, it is shown that 30% to 50% travel time can be saved, that the needed information exchange might reduce the cellular network capacity available for other services of 20% or more, and that the deployment of few road side units and multi-hop transmissions can be effectively used to offload cellular networks.**

Keywords*—Smart navigation, Intelligent transportation systems (ITS), Vehicular networks (VANETs), Simulations, UMTS, IEEE 802.11p.

I. Introduction

Keeping traffic moving is a challenge that governments, industries and researchers are facing worldwide. Effective solutions can only be obtained with a capillary and continuously updated knowledge of traffic conditions: the creation of an infrastructure for communication between vehicles, service centers and sensors, is thus one of the main needs identified by international institutions, service providers and car manufacturers to address with satisfactory results the problems generated by traffic, justifying the big efforts that are being pushed both in Europe and in the rest of the world.

Many projects have been carried out in the last decade on these topics. Among the others, an interesting example is the Italian project PEGASUS [1], [2], which relies on over one million vehicles already equipped with devices periodically transmitting their position and speed to a control center. Safety, enhanced mobility, and smart navigation systems were the services targeted by the project.

With the expression *smart navigation* we refer to the best route discovery service, which is based on the collection of measurements from vehicles equipped with sensor devices, hereafter on board units (OBUs), and the provision of updated traffic information to those vehicles equipped with on board navigators, hereafter smart navigators (SNAVs). Each OBU periodically collects and sends data to a remote control center and each SNAV receives from the control center information related to the actual traveling speed on the interested road segments, exploiting this information to update the path toward its final destination. Indifferently, SNAVs can be either on-board navigation systems or personal smartphones with specific apps.

Starting from the outputs of PEGASUS, here we focus on the benefits provided by the smart navigation and its impact on the communications network. The various topics addressed by PEGASUS, discussed in [3]–[8] and summarized for the first time in [1], are here extended and discussed with an integrated approach to provide an overview of smart navigation even from the point of view of wireless communications.

More specifically, in this paper we will discuss: i) the impact of updated traffic information on travel time and the amount of data that must be transferred through the wireless networks to make the smart navigation effective, ii) the feasibility and the impact of uplink transmissions of data from a large number of OBUs through cellular networks, iii) the feasibility and the impact of downlink transmissions of updated traffic information to a large number of SNAVs through cellular networks, and iv) the feasibility and the performance of short range vehicle-to-vehicle (V2V) and vehicle-to-roadside (V2R) communications used to offload the cellular networks.

These issues are hereafter addressed through simulations carried out adopting a simulation platform that integrates a vehicular traffic simulator, which reproduces the urban mobility, and a wireless network simulator, which models the details of the communication protocols and the signal propagation.

The paper is organized as follows: In Section II, the envisioned application is detailed and the addressed issues are introduced. In Section III, the simulation platform is described. In Section IV, we focus on the smart navigation service and we evaluate the saved travel time. In Section V, the cellular system and the related performance in the considered scenario are investigated. In Section VI, short range communications

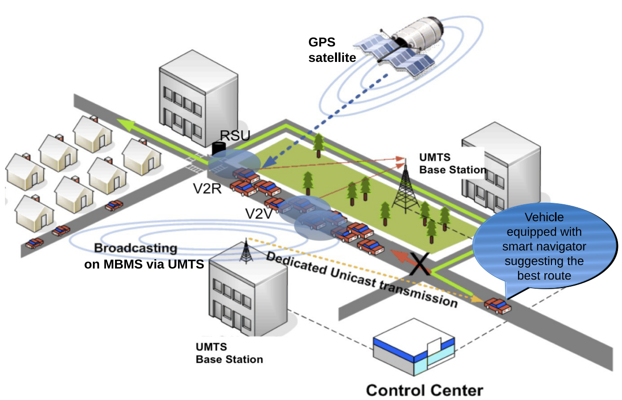


Fig. 1. Scenario: real time traffic information exchange and wireless technologies.

to offload the cellular network are presented and their performance investigated. Finally, in Section VII the conclusions are drawn and the future research activities are outlined.

II. Open Issues on Wireless Communications for the Smart Navigation Service

With the aim to study the benefits of smart navigation in terms of saved travel time and its feasibility from the communication network point of view, we consider the scenario shown in Fig. 1, where vehicles are equipped with OBUs acting as traffic sensors and periodically transmitting their position and speed to a remote control center, either through the cellular network or through multi-hop short range communications toward a road side unit (RSU). The data collected at the control center are processed in real time to estimate the actual traffic conditions (i.e., the travel time) on each monitored road segment; when the control center detects traffic conditions different from those foreseen by the static roadmap data base, it updates the roads travel times on a dynamic data base. Updated information is then transmitted from the control center to the vehicles equipped with the SNAVs through the cellular network in unicast or multicast mode. Then, the SNAVs calculate the optimal route; on a real time basis, they update their data and, in case, modify the route in order to avoid slowdowns.

From the communication technologies point of view, we need to acquire updated traffic information from OBUs (uplink) and to re-transmit processed data to the SNAVs (downlink). What amount of information needs to be transferred and which technologies to use are thus the obvious questions that arise. The first issue is strictly related to the impact of various parameters, such as the rate of data acquisition by OBUs or the rate of traffic updates provided to SNAVs, have on the effectiveness of the smart navigation service. The second issue finds cellular networks and short range wireless communication systems as possible candidates. In the short term, cellular networks appear as the only feasible solution, already guaranteeing high penetration and worldwide coverage. However, the resources needed for smart navigation are subtracted to other services, and their amount might be not negligible. A possibility to offload cellular networks could be to exploit short-range wireless communications based on wireless access in vehicular environment (WAVE) [9], which adopts IEEE 802.11p [10] at the MAC and physical layers and represents the future for V2V communications. This technology, well suited for safety applications, entertainment, gateway access, and road charging, can also be exploited in the envisioned scenario through the deployment of some RSUs and multi hop V2Vcommunications.

The above considerations motivated us to investigate several specific issues. First of all, in order to evaluate the potential benefits of updated traffic information provided by wireless communications systems, as detailed in Section IV we:

- investigate the travel time of vehicles equipped with SNAVs, periodically receiving updated traffic information, that enable to discover the best path. In particular we discuss the impact of the percentage of vehicles equipped with the OBU, the data collection rate, and traffic update timeliness.

The outcomes of these investigations are then used to discuss the role and the performance of wireless communications, both in the uplink and in the downlink, firstly focusing on the universal mobile telecommunications system (UMTS) cellular technology (in Section V) and then on the short range IEEE 802.11p technology (in Section VI). Focusing on UMTS we:

- investigate the feasibility of the acquisition of small but frequent amount of data from many OBUs (*uplink performance*);
- verify the feasibility of real-time transmissions to SNAVs of traffic information, both in unicast and multicast mode (*downlink performance*);
- investigate the impact of this service on the others already supported by the network (both in uplink and downlink).

Finally, considering IEEE802.11p-based V2V and V2R communications, we:

- investigate the impact of V2V and V2R short range communications to offload the cellular networks, with focus on the uplink delivery of measurement data generated by the OBUs.

III. Investigation Tools and Simulated Scenario

The investigation of the considered scenario requires a complete simulation of the enabling wireless networks, as well as a realistic simulation of vehicle movements. In fact, vehicular mobility significantly impacts on the performance of the telecommunication network and on the traffic redistribution itself. A realistic mobility model is thus needed, and it has to take into account all roads, with their speed limits, vehicles acceleration and deceleration, queues at traffic lights, etc. To this purpose we developed a simulation platform that integrates the VISSIM [11] vehicular traffic simulator and the SHINE [12] network simulator, that provided a realistic modeling of both the vehicular traffic (with queues, number of lanes, one

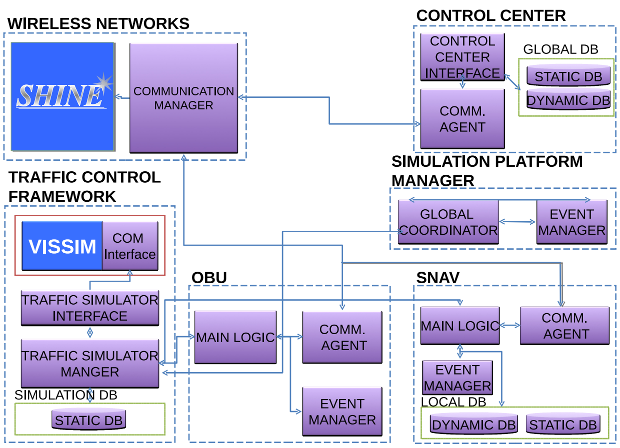


Fig. 2. Integrated platform for the simulation of smart navigation.

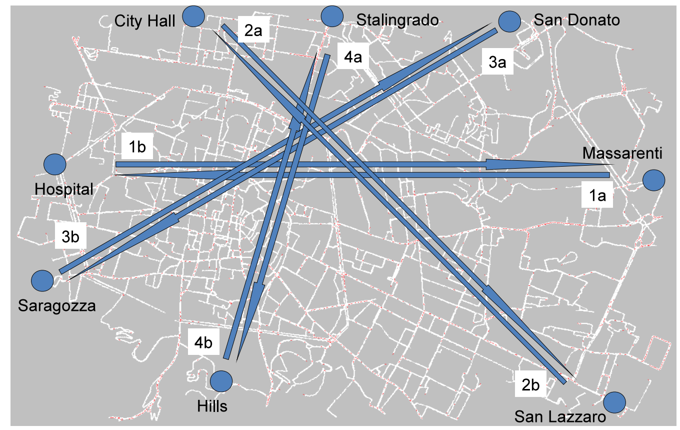


Fig. 3. Origins and destinations of the eight paths considered for the evaluation of the smart navigation impact on travel time.

way roads, etc.) and the network behavior [6]. The reasons behind this specific choice are detailed in [6], where pros and cons of several alternative vehicular traffic and wireless network simulators are discussed.

Vehicular simulator: VISSIM [11]. It is a commercial microscopic simulation tool that reproduces traffic flows in urban areas as well as interurban motorways, and allows to reproduce car-following and lane changing as in real scenarios. It uses a psycho-physical car following model for longitudinal vehicles movement and a rule-based algorithm for lateral movements. VISSIM can be controlled by external applications with the use of a component object model (COM): by the adoption of dynamic link libraries (DLL), it is possible for an application to control the movement of vehicles and to manage the whole simulation.

Wireless network simulator: SHINE [12]. It is an event driven dynamic simulator developed in our laboratories and written in C++, which allows to jointly take into account all aspects of the wireless networks related to the various layers of the network protocol pillar. SHINE has been designed to simulate heterogeneous wireless networks, according to a client-server structure; it is constituted by one server-core simulator (upper layers simulator, ULS) and one or more client simulators (lower layers simulators, LLSs), specific for the considered access technologies; among others, it includes an UMTS LLS and an IEEE 802.11p LLS.

Integrated platform. We realized a flexible architecture integrating VISSIM with SHINE, enabling the realistic simulation of vehicular traffic as well as of wireless networks [6]. This integrated platform allows the simulation of the whole smart navigation scenario: the vehicular mobility and the OBU transmissions, the data processing at the control center, the data base update with dynamic data, and the retransmission of the elaborated information to the SNAVs.

The architecture of the overall simulation platform is depicted in Fig. 2, and includes PostgreSQL databases and components written in C#; the interaction between components is provided through sockets and remote procedure calls (RCPs). More specifically, the following main blocks can be identified: *OBU*, which simulates the on board device collecting and transmitting position and speed; *SNAV*, which represents the fleet of vehicles equipped with the smart navigator, able to receive updated traffic information from the control center; *Control Center*, which is responsible of the gathering of data transmitted by the OBUs to update the dynamic data base; *Wireless Network*, which simulates the wireless network; *Traffic Control Framework*, which is based on VISSIM simulator and is in charge of controlling vehicles and their interactions; *Simulation Globals*, which manages the overall architecture. More details are available in [6].

Simulation scenario. The road-network layout of the reference scenario used for simulations consists of the medium sized Italian city of Bologna. In particular, we considered 13.636 road segments, corresponding to a length of about 600 Km. The digital-maps have been provided by TeleAtlas and given as input to VISSIM.

As detailed in the following, results concerning the cellular network performance will be obtained considering a portion of the entire scenario and assuming pedestrians that perform voice calls as background traffic added to the traffic generated by OBUs and SNAVs.

IV. Smart Navigation Impact on Travel Time

To evaluate the benefits, in terms of saved travel time, provided by timely updates of traffic information used by smart navigators, we assume that SNAVs receive updated information on the traffic conditions for each road segment from the current position to the destination and calculate the optimal route; on a timely basis, they update their data and, in case, modify the route in order to avoid slowdowns.

A. Simulation Settings and Output Figures

We consider non static traffic conditions, with a vehicle density that during each simulation varies in the range 1-10

TABLE I. SIMULATION PARAMETERS FOR THE SMART NAVIGATION SERVICE: DEFINITIONS AND NUMERICAL VALUES.

Parameter	Values
Starting instant of the journey of the monitored vehicle	600 s
Percentage of vehicles equipped with OBU ($\delta_{\mathrm{OBU}}x100$)	0%, 1%, 3%, 5%, 10%, 20%, 50%, 100%
Interval time for the transmission of the measured data τ	10 s, 30 s, 60 s
Interval time over which measured data are averaged T_{int}	10 s
Number of last consecutive intervals to obtain the current average speed	5
Interval time after which the average speed is updated in the smart navigator T_{update}	20 s, 60 s

vehicles/Km (please note that these are average values, also including secondary roads rarely used).

A parametric percentage of vehicles is assumed equipped with OBUs ($\delta_{\mathrm{OBU}} \times 100$, with $\delta_{\mathrm{OBU}} \in [0,1]$); every τ seconds, each OBU transmits several parameters including the actual vehicle position and speed to the control center. Measured data are processed by the control center to estimate the actual average speed of each road segment. Specifically, measured data are averaged on a parametric T_{int} interval time.

Then, every T_{update} seconds, the control center retransmits the processed data back to vehicles equipped with smart navigators. To avoid rough estimates in those roads where no vehicles or a too low percentage of them passed, we set up an average speed equal to that given by the static roadmap provided by TeleAtlas lowered by the 30%. In addition, when the measured speed is lower than 15 Km/h, we force the measurements exactly to 15 Km/h: this avoids to overestimate the travel time in the involved road segments.

We considered four paths, with different origins and destinations, in two directions (i.e., eight routes in total), as represented in Fig. 3. Each path is denoted in the following as path Nx, with N from 1 to 4, indicating the source-destination couple, and x either equal to a or b, indicating the direction. Since a smaller number of vehicles complete their route than those that are newly generated, the overall traffic increases in time, generating queues in an increasing number of junctions. In particular, about 1000 vehicles are present after 300 s, 2000 after 800 s, and 3000 after 1800 s.

The main parameters settings and their numerical values are summarized in Table I.

B. Numerical Results

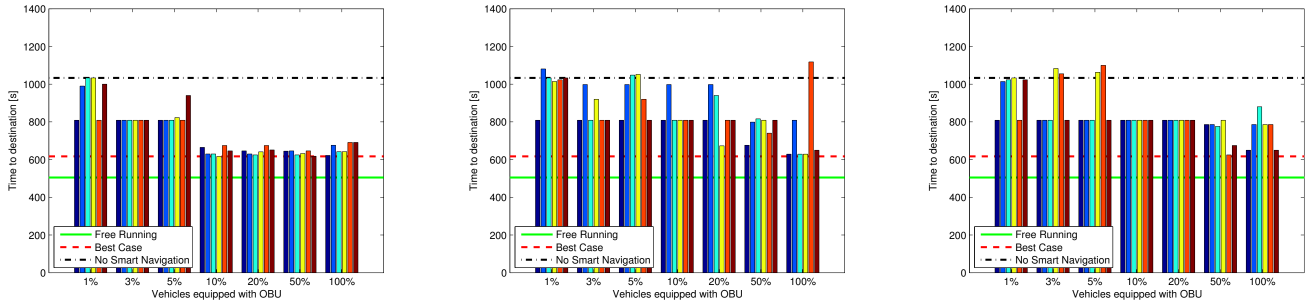
Numerical results are obtained assuming one vehicle equipped with the SNAV: this vehicle starts after 600 s and moves along one of the eight paths depicted in Fig. 3. For each path the time spent by the vehicle to reach the destination is investigated for different combinations of significant parameters (see Table I). Per each considered combination of parameters, six simulations were performed with different random variable initializations, that affect, for instance, the choice of the vehicles equipped with the OBU.

Results are presented for $T_{\mathrm{int}} = 10$ s, $T_{\mathrm{update}} = 20$ s and 60 s, and τ equal to 10, 30, or 60 s. Figs. 4 and 5 show the travel time, from origin to destination, of a monitored vehicle equipped with the SNAV for different percentage of vehicles equipped with the OBU and different scenarios. In each figure, results are compared with the following three cases adopted as benchmarks: i) *Free running*, referred to the case of a single vehicle moving alone on the entire scenario; ii) *Best case with smart navigation*, referred to a vehicle equipped with a smart navigator continuously updated with the best route, iii) *No smart navigation*, referred to the the same route as in Free Running in the presence of traffic (a navigator may be present, but without knowledge of real time traffic).

In Figs. 4 and 5, the travel time to destination is shown for path 1a for $T_{\mathrm{update}} = 20$ s and 60 s, respectively. Figs. 4(a), 4(b), and 4(c) (so as Figs. 5(a), 5(b), and 5(c)) refer to three different (uplink) transmission intervals τ: 10 s, 30 s, and 60 s, respectively. For each percentage of OBU equipped vehicles, six results are presented, corresponding to six different simulations providing time and space randomness (i.e., different vehicles are equipped with OBU, and the sampling process starts at different instants). By observing Figs. 4(a), 4(b), and 4(c), it can be noted that the time to destination increases with τ, showing a not negligible impact of a timely update of road segments traffic conditions. Focusing on Fig. 4(a), the impact of the percentage of OBU equipped vehicles can be appreciated: with a frequent traffic update in the uplink ($\tau = 10$ s), the 10% of vehicles equipped with OBUs is sufficient to have a time to destination very close the best case (i.e., about 600 s). As the uplink transmission interval τ increases, the 10% of vehicles is no more sufficient to obtain optimal results, that can be instead achieved (and not all times) only when all vehicles are equipped with OBUs.

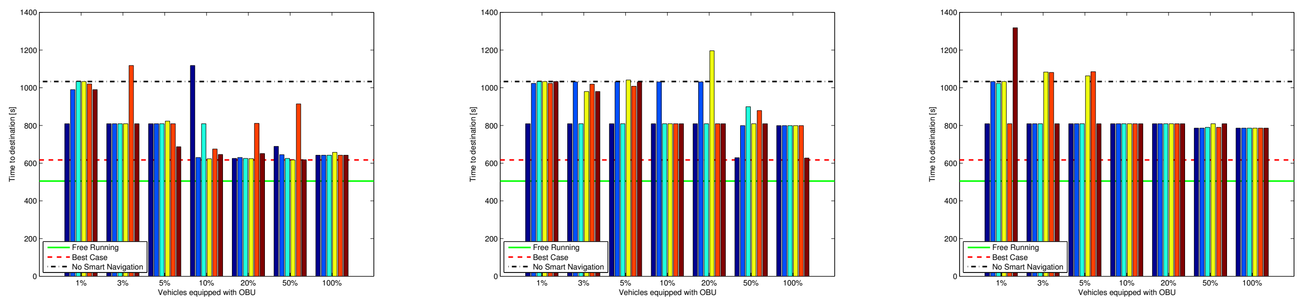
From a comparison between Fig. 4 and Fig. 5, referring to $T_{\mathrm{update}} = 20$ s and 60 s, respectively, it can be observed that the impact of the downlink (from the control center to SNAVs) transmission interval is lower than the impact of the uplink transmission interval τ (from OBUs to the control center),

Fig. 6 summarizes the results provided in Figs. 4 and 5, showing the average travel time of the six simulations and the confidence interval, corresponding to the standard deviation of the same values. The vehicle equipped with SNAV running in path 1a encounters various slowdowns and the free run time of 505 s increases up to 1033 s under the assumed traffic conditions. Smart navigation allows a reduction of the travel time to 629 s in the best case, which corresponds to about 40% of time saving. Results show that this saving can be achieved only with a sufficient amount of collected data: specifically, 10% of vehicles equipped with OBUs and $\tau = 10$ s are required. As can be observed, a similar advantage can be achieved in some cases also for a higher τ, but it is not guaranteed and depends on which vehicles are equipped and when they collect and send their data; for example, focusing on Fig. 4 and assuming $\tau = 30$ s and $\delta_{OBU} = 1$, in the forth case (yellow bar) the SNAV suggests the best possible route, whereas in the fifth case (red bar) it suggests a route, which is worse than the one followed with no smart navigator. Furthermore, comparing Figs. 6(a) and 6(b), referring to $T_{\mathrm{update}} = 20$ s and 60 s, respectively, a slight performance degradation can be noted when a lower update rate is assumed ($T_{\mathrm{update}} = 60$ s), but this effect is not



(a) $\tau = 10$ s. (b) $\tau = 30$ s. (c) $\tau = 60$ s.

Fig. 4. Travel time to destination for path 1a with different percentage of vehicles equipped with OBU and $T_{\text{update}} = 20$ s.



(a) $\tau = 10$ s. (b) $\tau = 30$ s. (c) $\tau = 60$ s.

Fig. 5. Travel time to destination for path 1a with different percentage of vehicles equipped with OBU and $T_{\text{update}} = 60$ s.

TABLE II. SUMMARY OF TRAVEL TIME FOR THE SMART NAVIGATION SERVICE.

Path	Free Run [s]	No Smart Nav. [s]	TI	Best Route [s]	SNI	Parameters to Allow the Best Performance of Smart Nav.
1a	505	1033	51.1%	617	40.3%	$\tau \leq 10$ s and at least 10% of OBU
1b	577	1016	43.2%	564	44.5%	$\tau \leq 10$ s or 100% of OBU
2a	477	651	26.7%	567	12.9%	Small advantage and easily achieved.
2b	482	800	39.7%	566	29.2%	$\tau \leq 10$ s and at least 10% of OBU
3a	494	575	14.1%	552	4.0%	Smart navigation is useless.
3b	439	533	17.6%	533	0.0%	Smart navigation is useless.
4a	374	449	16.7%	390	13.1%	$\tau \leq 10$ s and more than 10% of OBU
4b	386	864	55.3%	441	49.0%	$\tau \leq 10$ s and more than 10% of OBU or $\tau \leq 20$ s and minimum 50% of OBU

as remarkable as that observed varying τ.

Similar results can be obtained for all the other paths and are summarized in Table II. In particular, the columns show:

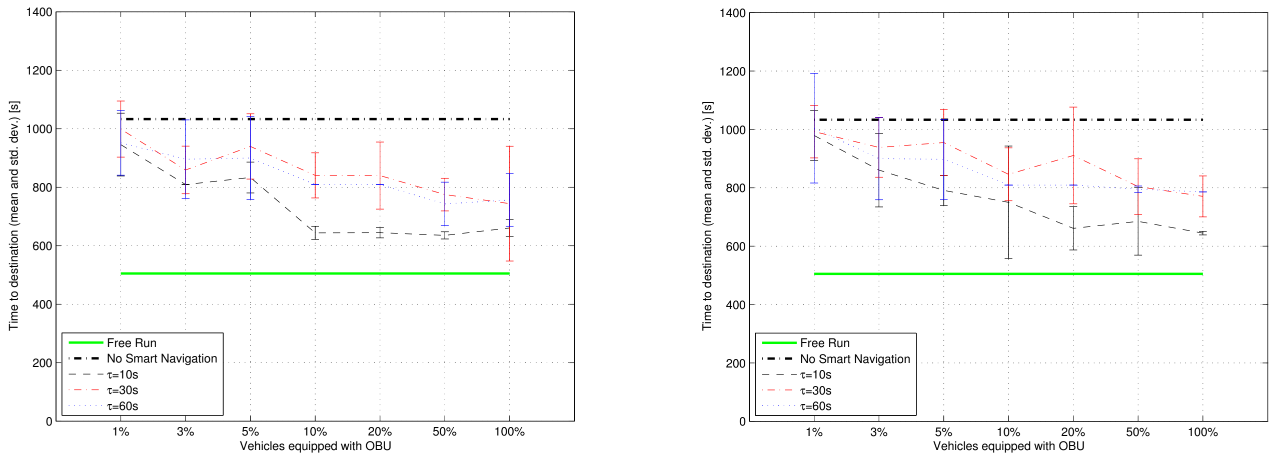
- the path name;
- the Free Run time in seconds, hereafter denoted with t_{FR};
- the No Smart Navigation time, hereafter denoted with t_{NoSN};
- the traffic impact TI evaluated through the following equation:

$$TI = \frac{t_{\text{NoSN}} - t_{\text{FR}}}{t_{\text{NoSN}}} \times 100 \tag{1}$$

  where (1) corresponds to the increase of the travel time along the best route withe respect to free run conditions caused by the presence of traffic in the absence of smart navigation, normalized and expressed in percentage;
- the Best Route time, hereafter denoted with t_{BR};
- the maximum achievable benefit of smart navigation SNI evaluated through the following equation:

$$SNI = \frac{t_{\text{NoSN}} - t_{\text{BR}}}{t_{\text{NoSN}}} \times 100 \tag{2}$$

  where (2) corresponds to the highest reduction of time enabled by smart navigation compared to the time needed in the same traffic conditions without smart navigation, normalized and expressed in percentage;
- a brief summary of the parameter settings that allow to achieve the maximum benefit from smart navigation,



(a) $T_{\text{update}} = 20$ s.

(b) $T_{\text{update}} = 60$ s.

Fig. 6. Mean and standard deviation of the time to destination for path 1a.

i.e., that allow the minimum travel duration with high probability.

The eight considered paths allow to compare the effect of smart navigation in various traffic conditions: when TI is high, i.e., when the traffic severely affects the travel duration (paths 1a, 1b, 2b, 4b), the smart navigation service allows to significantly reduce the travel time, with SNI between 30% and 50%; on the contrary, when TI is low (paths 2a, 3a, 3b, and 4a) the observed SNI is lower than 15%.

Another interesting conclusion arises comparing the impact of τ and T_{update}: the transmission rate from OBUs is much more important than the update rate at SNAVs. τ has, in fact, a direct impact on the traffic estimation, beacuse higher transmission rates from OBUs entail more available traffic measurements and thus better estimates, whereas T_{update} only impacts on the decision made by the SNAVs, taken at the few main junctions encountered.

Numerical results show thus that smart navigation allows the equipped vehicle to significantly reduce the travel time whenever the traffic has a significant impact on the travel duration and they highlight that a sufficient amount of data must be collected by OBUs. In particular, they show that a simple average of the collected speeds allows a good route selection when a transmission interval τ lower than 10 s is assumed and at least 10% of vehicles are equipped with OBUs. It may be possible that more sophisticated algorithms with more complex processing at either the control center or the OBUs would allow similar results with a lower τ, but the design of such algorithms is out of the scope of the present paper.

V. Cellular Communications Enabling Smart Navigation

The smart navigation service relies on wireless communications. Among all possibilities, cellular systems represent the obvious short term solution, both to collect data from OBUs (uplink) and to retransmit them to SNAVs (downlink). Due to the widespread diffusion of always-on navigators and smartphones with positioning applications, cellular systems allow the service implementation avoiding new set-ups or expensive installations [13]–[15]. Various activities are ongoing [4], [5], [16], [17], and products based on cellular technologies are already on the market.

Some studies on the performance of cellular systems in vehicular applications are coming out (see, e.g., [18]), but still few investigations focused on the impact that these new services have on other cellular services (such as voice calls) in terms of resource availability and quality of service (QoS).

Focusing on the uplink, the use of vehicles as communication-enabled moving sensors opens to a number of potential applications, as pointed out in [19]–[22]. Example applications are also presented in many related works, where vehicular ad hoc networks (VANETs) are envisioned, for instance, to alert upcoming vehicles when an accident occurs [23], to guarantee urban environment surveillance [24], to provide widespread pollution measurements [25], to enable traffic monitoring and smart navigation [4], and to perform civil infrastructure monitoring and automotive diagnostics [26]. For all these applications, cellular systems are today the unique solution. A study on the feasibility of data acquisition from vehicles through cellular systems has been carried out in the German project Aktiv CoCar [27], that defined a new protocol, called “traffic probe data protocol” (TPDP), to upload traffic data through UMTS common channels. However, results are given only in terms of cumulative distribution function (CDF) of the end-to-end delay, and no evaluation of the impact of this service on the QoS experienced by other users is given.

Focusing on the downlink, also broadcast technologies are available. A first example is represented by the traffic message channel (TMC) [28], [29], aimed at delivering traffic and travel information using the radio data system (RDS) on conventional frequency modulation (FM) radio broadcasts. However, these technologies and applications have some severe limitations. In fact, the data rate reserved for these services is limited, thus allowing the traffic conditions update for a very reduced number of roads. In addition, only the main events are described, with few approximate descriptions.

For this reason, in several Countries on-board navigation devices receive updated traffic information by means of cellular technologies [30]. Reduced data rate is sufficient until traffic information is not frequently updated nor capillary, but as the service coverage and timeliness increase the cellular network capacity and the expected QoS becomes a primary issue to be investigated.

Here, we focus on the most widely adopted cellular technology, UMTS, to collect data from OBUs and to provide updated traffic information back to SNAVs. Different strategies are considered in both directions and their impact is shown through simulations: in the uplink we compare the use of dedicated channels or random access channels, while in the downlink we compare the use of unicast or multicast transmissions. Since the quality perceived by users of other services is also of great interest, also pedestrians performing voice calls are considered.

A. Simulation Scenario and Non Vehicular Users

The portion of Bologna considered for UMTS simulations is shown in Fig. 7 and consists of a rectangular area of the city center sized 1.8 km (longitude) x 1.6 km (latitude) with 35 UMTS cells covered by 15 Nodes-B (1, 2 or 3 cells per Node-B are assumed), with a single frequency planning. This portion of the entire scenario has been considered for the assessment of the cellular network performance to avoid edge effects. Fig. 7 also shows the approximated coverage of each cell with random colors. Black segments represent roads where vehicles movements are constrained.

Concerning non vehicular users, a variable number of pedestrians randomly walking in the scenario is generated, with an uniform distribution and the same birth probability in each cell (this means that a higher density of voice users is assumed where smaller cells are considered). Hereafter, $\Lambda^{(v)}$ indicates the average offered voice load in Erlang per km^2.

As previously detailed, vehicles are constrained on roads and generated by VISSIM; a heavy traffic condition with many traffic queues is investigated, with 220 average vehicles per Km^2. A portion $\delta_{\mathrm{OBU}} \in [0,1]$ of vehicles are equipped with the OBU, while a portion $\delta_{SNAV} \in [0,1]$ of vehicles are equipped with the SNAV.

Particular attention has been paid to all physical level aspects, including propagation, interference, and power control mechanisms. For further details on both the scenario and the UMTS simulator the reader may refer to [3].

1) Background Voice Traffic: The duration of each voice call is randomly defined with Poisson distribution and 90 s as average. In the UMTS access network, voice traffic uses a 12.2 kbit/s bearer, corresponding to a logical dedicated traffic channel (DTCH), a transport dedicated channel (DCH), and

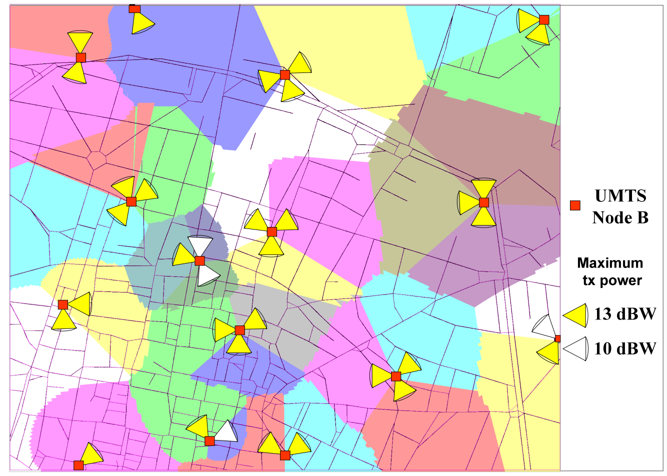


Fig. 7. UMTS planning in the considered scenario. Filled black squares correspond to Node-B locations. An approximated area of coverage is depicted for each cell with random colors.

a physical dedicated data channel (DPDCH). The DPDCH has a spreading factor (SF) equal to 64 in uplink and 128 in downlink.

Output figures. The evaluation of the QoS perceived by voice users is based on the following definitions: per each frame lasting 10 ms, a user (i.e., a voice call) is defined in outage if the BER after channel decoding of that frame is greater than 2% (uplink and downlink are evaluated independently to each other); an ended *voice call* is then considered *in outage* when either in downlink or in uplink, the outage intervals exceed a threshold of 5%. Hence, we have an outage voice call when one user is able to talk to the other party, but with poor audio quality. A voice call may also incur in the following situations: it may be blocked by the call admission control algorithm due to insufficient resources, or it may drop due to an excessive reduction of the received signal power.

Results will be presented in the following in terms of *satisfaction rate (SatR)*, that is the ratio between the number of users, which are not blocked, neither dropped, nor in outage, and the total number of call requests.

B. Uplink Acquisition from Vehicles

Among the cellular technologies, general packet radio service (GPRS) is nowadays the most adopted in vehicular scenarios for uplink measurements transmission. However, to transmit data over the GPRS network, the mobile device must first send a message on a common channel asking for a dedicated resource, with procedures requiring a not negligible access time, in the order of seconds [31]; for this reason in practical systems the OBU collects tens of measurements before transmitting them in a single packet. This approach obviously increases the data acquisition delay at the control center, and limits the effectiveness of the smart navigation service, as shown in Section IV. Differently to this, UMTS also allows the transmission of small amount of data over the shared random access channel (RACH), avoiding the set-up of dedicated resources [32]. This way, any measurement can be transmitted by the OBU as soon as it is taken, with minimum delay and reduced signaling overhead. This solution appears promising especially considering the expected increase of the number of equipped vehicles, but it clearly requires investigations on feasibility and resources occupation.

Here, we discuss the impact of real time data acquisition on the performance of cellular networks, foreseeing the realistic perspective of an explosion of the number of equipped vehicles. We consider DTCH using either dedicated or random access logical and physical channels. In the former case, a logical DCH and a DPDCH are used in uplink with a SF equal to 16, and the same channels are used in downlink with a SF equal to 32 (a dedicated channel is required also in the downlink for the TCP acknowledgment transmissions), both providing 64 kbit/s. In the latter case, a logical RACH mapped on the physical random access channel (PRACH) is used. Differently from the RACH, the dedicated transport channel is characterized by features such as fast power control, fast data rate change on a frame-by-frame basis, and soft handover. However, it has to be established each time a connection is required.

Vehicles transmit 80 byte packets every 10 seconds. When the random access is exploited in uplink, one RACH (out of the available ones) is exclusively used in each cell by the smart navigation service.

Output figures. The impact of the smart navigation service is primarily shown in terms of the effects it causes on the QoS of voice users. Besides this, the QoS of this service can be addressed investigating the probability that each measurement stored in vehicles is correctly received by the control center, independently on the specific source.

When the DCH is used, data cannot be lost due to the use of TCP at transport level and the delay introduced by the transmission over the cellular link is not very critical (the average delay remains lower than few seconds unless the network is heavily congested). Thus, observing the effect on the voice traffic is sufficient in this case.

When the RACH is used, the delay is even lower, but not all packets are delivered to the control center. In fact, a scheduled transmission fails in two cases: 1) when the RACH ramping procedure is unsuccessful, meaning that the propagation conditions and the perceived interference level are so disadvantageous that the maximum transmission power is not sufficient, and 2) when an error is detected at the receiver. In any case, the MAC layer may attempt a number of retransmissions before discarding the packet. Focusing on the smart navigation service with RACH, results are thus also expressed in terms of *packet discard rate* (R_D), that is the ratio between the number of discarded packets and the total number of packets generated by all OBUs.

UMTS Uplink: Numerical Results. The *SatR* for voice users and the R_D for the smart navigation service with RACH are plotted in Fig. 8 and Fig. 9, respectively, as a function of $\Lambda^{(v)}$.

In Fig. 8, the *SatR* of voice users is depicted with $\delta_{OBU} = 1$, and the case with no smart navigation service ($\delta_{OBU} = 0$) is shown for comparison. Observing Fig. 8, the presence of the smart navigation service seems not to impact on voice users, since their satisfaction remains almost unchanged. This is no more true when the dedicated channel DCH is adopted for the smart navigation service instead of RACH, that results in a *SatR* reduction. To grant a $SatR = 0.95$, about 740 voice users could be served both in the absence of the smart navigation service or using RACH, whereas the same *SatR* is satisfied with only 600 when the DCH is adopted (that entails a not negligible reduction of about the 20%).

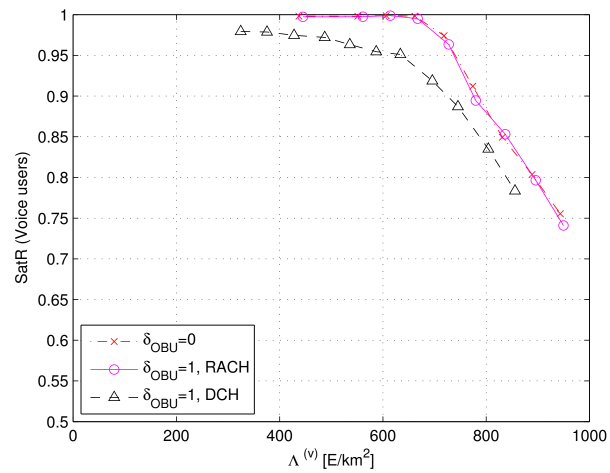


Fig. 8. Voice traffic performance: voice *SatR* vs. the offered voice traffic $\Lambda^{(v)}$. Comparison between no smart navigation ($\delta_{OBU} = 0$), smart navigation adopting RACH, and smart navigation adopting DCH.

From this preliminary result, related to voice users only, it seems that the use of RACH does not impact the network performance: for a given *SatR* the maximum $\Lambda^{(v)}$allowable is not affected by the underlying smart navigation service adopting RACH. However, looking at Fig. 9, where the R_D is also plotted as a function of $\Lambda^{(v)}$ for $\delta_{OBU} = 1$, this conclusion must be revised. As can be observed, the higher is the network load, the higher is the value of R_D, and the QoS of the smart navigation service results deteriorated. If $\Lambda^{(v)}$=740 (corresponding to $SatR = 0.95$) is taken as reference value, a packet loss higher than 5% can be observed, meaning that guaranteeing a $SatR = 0.95$ to voice users, does not imply that the smart navigation users are also served. To improve the QoS of the smart navigation service, a lower number of voice calls must be accepted. For instance, if R_D lower than 10^{-2} is targeted, with respect to a maximum of $\Lambda^{(v)}$=740 in the absence of the smart navigation service, a reduction of about 100 (13.4%) average voice users per Km^2 must be considered, drastically reducing the capacity left for voice users. Such reduction is, however, lower than that caused by the smart navigation service with the use of DCH.

C. Downlink Transmission to Vehicles

As far the adoption of UMTS to deliver traffic information to SNAVs is concerned, two options are available: either transmitting personalized data through unicast DCHs, or to distribute

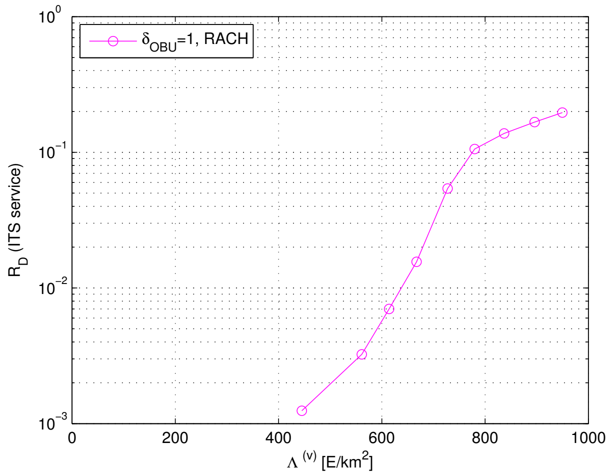


Fig. 9. Smart navigation performance: R_{D} varying the offered voice traffic $\Lambda^{(\mathrm{v})}$. $\delta_{\mathrm{OBU}} = 1$.

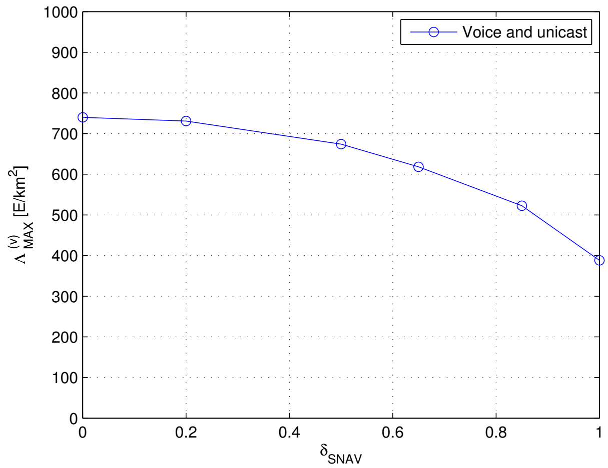


Fig. 10. Voice capacity as a function of the fraction δ_{SNAV} of vehicles receiving dedicated traffic information in unicast mode.

the same content to all users through multimedia broadcast multicast service (MBMS) channels. Hereafter, we compare these two cases, evaluating if the network can support the additional new load and the impact it has on the performance perceived by other UMTS users.

Due to the adoption of code division multiple access (CDMA) [3], the number of active channels in UMTS is a consequence of the trade-off between coverage and capacity, and the amount of resources occupied by each transmission is given in terms of used power: on the one hand, a higher data rate as well as a higher distance from the base requires a higher power for a sufficient QoS; on the other hand, a higher power reduces the cell capacity. The power is, in fact, a limited resource at the base station (in downlink) and each transmission turns into an interference to all other active communications (in both directions).

As for MBMS, it allows to share resources among many user. Hence, power is allocated to MBMS channels only once for any number of users in the cell receiving the service. When the multicast transmission is exploited in downlink, we assume that in each cell one MBMS channel is exclusively used by the smart navigation service, and that all SNAVs are enabled to join the multicast group, where traffic-related messages are distributed. It has to be remarked that MBMS uses part of the power available at the base station, thus limiting the number of DCHs that can be established. Moreover, the broadcast/multicast nature of the channel does not allow to exploit the fast power control feature that is of main importance for an interference limited system like UMTS; the base station pre-assigns a certain amount of power to MBMS services depending on the coverage planning and the desired bit rate.

The following strategies, thoroughly described in [5], are assumed for traffic updates:

- For the unicast mode, the update involves *road segments encompassed by an ellipse* whose focuses are the actual vehicle position and either an intermediate point or the final destination. This strategy avoids the transmission of information related to road segments too far from the actual vehicle position, which would be out of date when the vehicle needs it. Moreover, since only the transmission of the coordinates of two points is needed from the SNAV to the control center, the amount of data transmitted in the uplink is very small, thus limiting costs and resource occupation. Following [5], 1000 road segments are updated every 5 minutes.
- For the multicast mode, a *progressive coverage* strategy is considered, consisting in the transmission to the SNAV of the information related to the most important roads at national level and regional level, and to the minor roads at local level only. Following [5], 12000 road segments are updated in average.

Independently on the unicast or multicast communication technology, we assume the transmission of transport protocol experts group (TPEG) [33] messages at the highest layers of the protocol pillar with 60 bytes packet per each road segment [17].

1) Unicast Mode: Following the given assumptions, SNAVs receive updated traffic information through a 60000 bytes download (i.e., 1000 road segments $\times$ 60 bytes) every 5 minutes. Data are transmitted adopting the TCP protocol at the transport level, that assures data reception. A 64 kbit/s bearer is considered, corresponding to a logical DTCH, a transport DCH, and a physical DPDCH (the low amount of bytes and the relaxed delay requirements do not justify the use of more consuming bearers). The DPDCH is transmitted adopting a SF equal to 32 in downlink and 16 in uplink (note, in fact, that a dedicated unicast link is required also in the uplink direction for the TCP acknowledgment transmissions).

Output figures. A smart navigation user is satisfied if the update is received with a delay lower than 15 s (please consider that less than 10 s would be required if data were transmitted

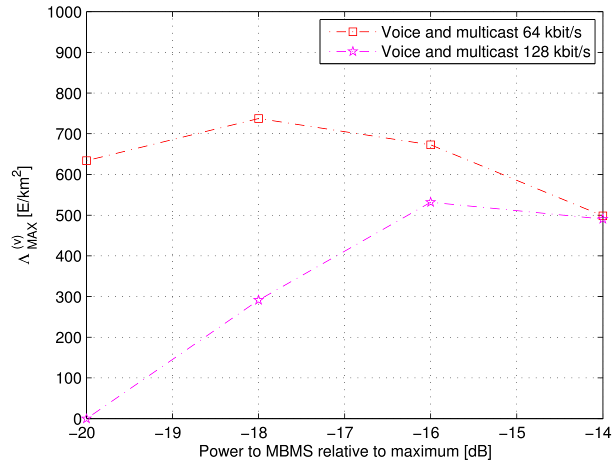


Fig. 11. Voice capacity as a function of the Node-B power dedicated to the multicast channel providing the envisioned service.

at 64 kbit/s with no errors and no TCP redundancy).

2) Multicast Mode: Data are transmitted adopting user datagram protocol (UDP) at transport level, which introduces limited redundancy but do not grant reliable communications; in this case, in fact, the absence of the uplink connection does not allow the transmission of acknowledgments. Two bearers at 64 and 128 kbit/s are considered, each corresponding to a logical MBMS transport channel (MTCH), a transport forward access channel (FACH), and a physical secondary common control physical channel (S-CCPCH). The S-CCPCH is transferred (obviously, in downlink) adopting a SF equal to 32 for 64 kbit/s and 16 for 128 kbit/s.

With the assumptions given above, all roads are updated every 90 or 45 seconds with the 64 and 128 kbit/s bearers, respectively.

Output figures. An ended smart navigation session is assumed in outage if less than the 95% of packets are correctly received. For smart navigation users the *satisfaction rate* is the ratio between the number of users that do not experience an outage and the total number of users.

UMTS Downlink: Numerical Results. In Fig. 10, the maximum $\Lambda^{(v)}$is plotted as a function of the number of equipped vehicles receiving updated traffic information via DCH. In particular, the x-axis represents the ratio δ_{SNAV} of vehicles that are equipped with the SNAV. The y-axis represents the maximum amount of voice calls that allow the system to serve both traffic classes (voice and smart navigation) with a satisfaction rate (i.e., ratio of satisfied users over the number of users of that class) greater than 95%. When the number of equipped vehicles is zero, we obtain results referred to the presence of voice only, considered as a benchmark (740 average voice calls). We can observe that, as the number of equipped vehicles receiving updated information increases, the maximum number of voice calls (i.e., the number of voice users) decreases, due to larger resources dedicated to the smart navigation service. It is however interesting to note that the decrease of $\Lambda^{(v)}_{MAX}$ is smooth for small values of δ_{SNAV}, and it becomes rapid as δ_{SNAV} approaches 1; more specifically, a reduction of only 10% voice users is observed if the 50% of vehicles are equipped with SNAVs ($\delta_{SNAV} = 0.5$), but a reduction of about 50% is observed when all vehicles are equipped ($\delta_{SNAV} = 1$).

Multicast results, shown in Fig. 11, are obtained varying the power used to transmit the S-CCPCH, which is a constant fraction of the maximum available power at the base station. Since an MBMS channel per cell is assumed independently on the number of SNAVs, the value of δ_{SNAV} is not relevant. Fig. 11 shows the maximum amount of voice calls (per km^2) that allow the system to serve both classes of traffic with at least 95% satisfaction rate as function of the fraction of Node-B power dedicated to MBMS. Multicast at 64 kbit/s and 128 kbit/s are compared. As can be observed, independently on the adopted bearer, the number of voice calls increases with the power assigned to MBMS until a maximum, then it starts decreasing: low power levels assigned to the MBMS service, in fact, require low interference (generated by voice users) in order to guarantee a full coverage to the smart navigation service, while high power levels generate strong interference that limits the number of voice calls that can be served. We can thus note that a trade off between voice and smart navigation services can be obtained for both 64 kbit/s and 128 kbit/s: -18 dB to MBMS with about 1% less average voice calls compared to the case of no smart navigation service in the former case; -16 dB to MBMS with about 30% less average voice calls in the latter case. These numbers also highlight that the adoption of a 128 kbit/s bearer greatly reduces the number of voice calls with respect to 64 kbit/s.

Comparing the unicast versus multicast approach, results show that the latter is preferable only for a high penetration of the service and only if an update every 90 s is acceptable (corresponding to a multicast channel at 64 kb/s).

VI. Short Range Communications to Offload Cellular Networks

As shown in previous sections, the increasing number of vehicles equipped with OBUs or SNAVs might overload the cellular networks, and this is particularly true in the uplink direction (acquisition of small but frequent packets from a very large number of OBUs). The use of alternative solutions, and in particular of short range V2V and V2R communications, could thus be conveniently exploited; in this case, RSUs could be deployed at proper positions and V2V and V2R could be used to collect data without the use of cellular communications.

Dealing with short range communications in vehicular scenarios, the WAVE [9]/IEEE 802.11p [10] is presently the reference standard, and several field trials and research activities are currently carried out focusing on it (see, e.g., [34]–[40]). WAVE defines the communication system architecture and the complementary set of services and interfaces for vehicular scenarios, whereas IEEE 802.11p, which is an amendment to the IEEE 802.11 standard conceived for vehicular communications at 5.9 GHz, describes the MAC and physical layer protocols. At the physical layer, IEEE 802.11p is based on the orthogonal frequency division multiplexing (OFDM) modulation, with seven non overlapping channels of 10 MHz each; one of these channels is reserved for control purposes and the other six are provided as service channels. In the control channel all OBUs are expected to periodically broadcast their identity and position (e.g., GPS coordinates) in packets denoted as beacons, so that each OBU has a real time knowledge of all its neighbors. A key amendment introduced by WAVE/IEEE 802.11p is the WAVE mode, which allows the transmission and reception of data frames with the wildcard basic service set (BSS) identity and without the need of belonging to a particular BSS. This feature enables very efficient communication-group setup without much of the overhead typically needed in nomadic IEEE 802.11a/g networks; it simplifies the BSS operations in a truly ad hoc manner for vehicular usage [10], [41], and can be used by devices for a fast exchange of data.

Here, we assume that OBUs are equipped with the WAVE/IEEE 802.11p [9], [10] communication interface with the objective to offload the cellular networks. Although it is not mandatory according to the specification, here we assume the use of two channels in parallel, with the control channel only used for beacon broadcasting and one service channel used for data exchange; this configuration guarantees that no reduction of performance is caused by the smart navigation service to safety applications. In our simulations, each OBU broadcasts its position in the control channel, with a beacon frequency of 10 Hz (in accordance with the considerations given in [42] and [43]), and transmits data in the service channel when needed.

To reduce the cellular network load, if an OBU is under the coverage of any RSU, it directly delivers its data through V2R communication. Otherwise, a routing algorithm is required to find the best route towards an RSU through V2V multiple hops. In particular, the scope of the routing algorithm is to search for a suitable next relay among the neighbor nodes. Among the many routing algorithms proposed in literature for VANETs (see, e.g., [44], [45]), here we adopt greedy forwarding (GF), which inspired most of those suited for not fully connected networks. In GF, the OBU knows the position of the nearest RSU (thanks to a location service, out of the scope of this work) and selects it as the destination; the OBU also knows the position of all its neighbors (thanks to the beaconing mechanism), and considers as possible relays those that are nearer to the destination; the OBU forwards data to the relay which is closest to the destination, if any, and stores the data otherwise.

Since an RSU or a suitable next relay are not always available and since the OBU buffer is limited and data cannot be stored for an unlimited amount of time, all packets are sent through the cellular network whenever one of the following conditions is met: (i) the number of packets inside the transmission buffer reaches a threshold N_{MAX}, or (ii) at least one of the queued packets was generated more than T_{MAX} seconds before the actual instant. N_{MAX} is related to the hardware implementation of the OBUs, while the choice of T_{MAX} is strictly related to the maximum acceptable delivery delay. Since, as shown in Section IV, the timely update of information impacts on the smart navigation effectiveness, the value of T_{MAX} is of main interest for the envisioned service.

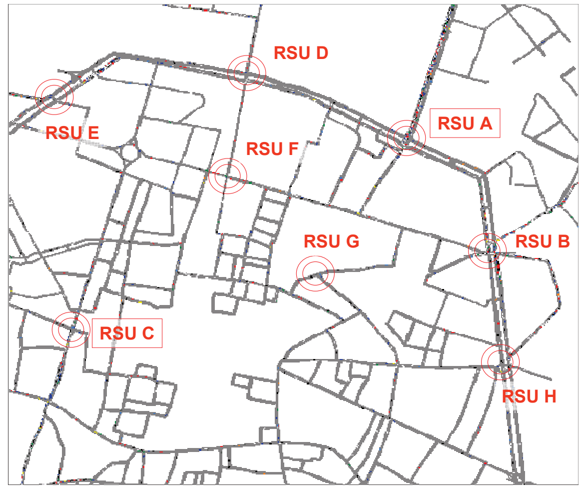


Fig. 12. RSU positions in the considered scenario.

In the following, we show the performance in terms of cellular resources that can be saved by exploiting V2V and V2R communications to deliver the sensed data to the control center through the RSUs.

A. Simulation Settings and Output Figures

As in previous sections, a parametric portion of vehicles δ_{OBU} ($\delta_{\mathrm{OBU}} \in [0,1]$) is assumed equipped with an OBU, which acquires some measurements every τ seconds. Like in the UMTS case, simulations are performed in a portion of Bologna with an high density of vehicles, corresponding to 220 vehicles per Km^2. In this case, one or more RSUs are deployed in the scenario, in the positions shown in Fig. 12. The selected sites correspond to the mostly crowded junctions, noting that major junctions are suitable sites also owing to the likely presence of lighting, traffic lights, and therefore of power supply. RSU A is positioned, in particular, in the busiest crossroad of the whole scenario, while RSU C is in the less loaded crossroad among the eight shown in Fig. 12.

As already discussed, the maximum number of packets N_{MAX} that can be queued is related to the hardware implementation of the OBUs, and is here set to a high value, 1000. On the contrary, different values are assumed for T_{MAX}.

In short range simulations, we refer to the following propagation model:

$$\mathrm{PL}(d) = \mathrm{PL}_0(1) + 10\beta \log_{10}(d) \tag{3}$$

where $\mathrm{PL}_0(1)$ is the free space path loss at 1 meter distance, β is the path loss exponent, and d is the distance in meters. A threshold model is then assumed for the packet error rate, with the shadowing effect due to buildings: a transmission between two devices is possible only if 1) the virtual line connecting them do not cross any building, 2) the received power is higher that the receiver sensitivity and 3) the signal-to-noise-plus-interference ratio is higher than a threshold. With the

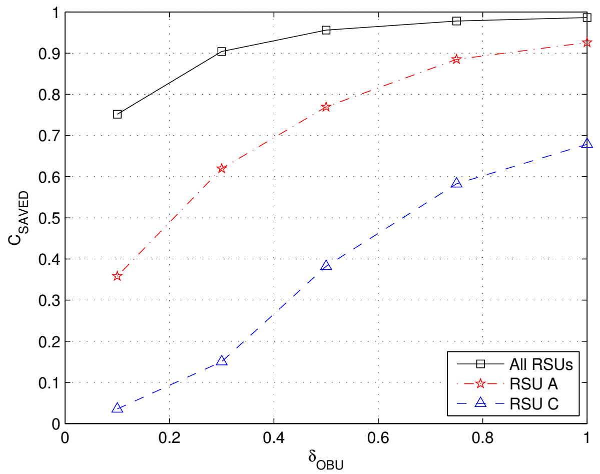


Fig. 13. Short range communications: impact of the OBU penetration δ_{OBU} and of RSU number and positions. Specifically, C_{SAVED} as a function of δ_{OBU}, with all eight RSUs, with RSU A only, and with RSU C only. $T_{MAX} = 10$ s.

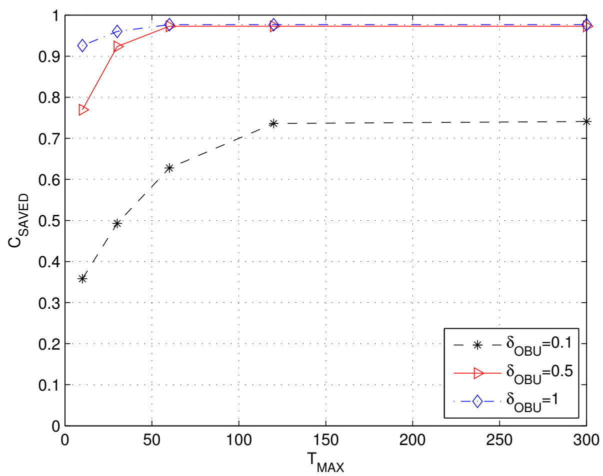


Fig. 14. Short range communications: impact of the maximum delay before transmission through the cellular link T_{MAX}. Specifically, C_{SAVED} as a function of T_{MAX}, with various values of δ_{OBU}. RSU A.

assumed parameters, the maximum communication distance in the absence of interferers is 200 m.

Numerical results are provided in terms of ratio of saved cellular resources C_{SAVED}, i.e., the ratio between the number of packets delivered through the RSUs and the number of packets generated by all vehicles.

B. Numerical results

The impact of the OBU penetration and the impact of the number and positions of RSUs are investigated in Fig. 13, where C_{SAVED} is plotted as a function of δ_{OBU}, with various RSU deployments. To guarantee an updated knowledge of traffic conditions at the control center, T_{MAX} is set to 10 s. Results show the large amount of cellular resources that can be saved through the joint adoption of V2V and V2R communications. As expected, when all the RSUs are deployed, C_{SAVED} outperforms the cases of a single RSU. However, even in case of single RSU A, C_{SAVED} is always larger than 0.3 when $\delta_{\mathrm{OBU}} > 0.1$, even reaching 0.9 when $\delta_{\mathrm{OBU}} = 1$; this result highlights that even a single RSU, properly positioned, can be exploited to save a large amount of cellular resources. Observing the curve that refers to RSU C, which is located in a less busy junction, it can be observed that it is less effective, owing to a reduced number of vehicles passing in its proximity. However, also in this case C_{SAVED} exceeds 0.6 when $\delta_{\mathrm{OBU}} = 1$, with a not negligible percentage of saved cellular resources.

Fig. 14 shows C_{SAVED} as a function of T_{MAX} for various values of δ_{OBU}, when only RSU A is deployed. Allowing a greater T_{MAX} increases the probability that vehicle mobility generates new paths toward the RSUs. As can be observed, if a higher T_{MAX} is acceptable, then C_{SAVED} noticeably increases. If a T_{MAX} higher than 30 s and a δ_{OBU} higher than 0.5 are assumed, then C_{SAVED} with RSU A becomes grater than 0.9, almost fully offloading data from the cellular network.

Short range communication is thus proved to represent an interesting solution to offload the cellular network with limited costs.

VII. Conclusion and Future Work

In this work, we have discussed the smart navigation service from the wireless communication networks point of view. In particular, we have firstly evaluated the performance of the smart navigation in terms of travel time, showing that a reduction of even the 50% is possible; results have also revealed that at least 10% of vehicles providing information every 10 s are needed to enable a correct estimation in most conditions, while the update at the smart navigator is less critical. Then, the feasibility from the communication systems point of view has been discussed. To this scope, we have considered UMTS as the enabling technology for the real time acquisition and transmission of traffic information, evaluating the impact of such a communication both on other services already provided by UMTS and on the QoS of vehicular users. Furthermore, the use of dedicated channels or shared channels was compared. Our studies have highlighted that the service appears feasible and that the number of equipped vehicles does not seem a critical issue. We have also pointed out, however, a not negligible loss in capacity; in particular, it has been shown that, to guarantee a satisfactory quality of service, 15% or more resources are subtracted to the other services. Finally, the adoption of short range V2V and V2R communications have been explored to offload the cellular network in the uplink, showing that 30% to 90% data traffic can be offloaded even with few RSUs, depending on the number of vehicles equipped with such technology and the accepted delivery delay.

Besides the several aspects discussed in this paper, a number of issues are still open for future work. Concerning the benefit of smart navigation for the driver, we will extend our study to the effect of smart navigation when all vehicles are routed at the same time. Furthermore, we will investigate if the same traffic estimation can be obtained with a lower amount of information, adaptively changing the rate of both uplink collection and downlink update following traffic conditions. Regarding cellular resources, we will investigate the adoption of long term evolution (LTE) and LTE advance instead of UMTS. Focusing on short range communications, we will investigate other routing algorithms and we will also address the downlink diffusion of information from the RSUs to the SNAVs.

VIII. Acknowledgments

The authors wish to thank the colleagues at WiLab for their contribution in the PEGASUS project and the reviewers for their valuable comments.

This work has been funded by the national project PEGASUS, financed by the Italian Ministry of the Economic Development in the framework of Industria 2015. The work has been developed at WiLab (Bologna, Italy) with the collaboration of researchers from CNR-IEIIT and DEI, University of Bologna.

References

[1] A. Bazzi, B. Masini, G. Pasolini, and O. Andrisano, "Wireless communications enabling smart mobility: Results from the project PEGASUS," in *Second International Conference on Mobile Services, Resources, and Users (MOBILITY) 2012*, Oct. 2012.

[2] PEGASUS project. Accessed on June 2013. [Online]. Available: http://pegasus.octotelematics.com

[3] A. Bazzi, B. Masini, and O. Andrisano, "On the frequent acquisition of small data through RACH in UMTS for ITS applications," *Vehicular Technology, IEEE Transactions on*, vol. 60, no. 7, pp. 2914–2926, 2011.

[4] A. Conti, A. Bazzi, B. Masini, and O. Andrisano, *"Vehicular Networks: Techniques, Standards, and Applications"*. Auerbach Pub., CRC Press, Taylor & Francis Group, 2009, ch. Heterogeneous Wireless Communications for Vehicular Networks, pp. 63–107, ISBN: 9781420085716.

[5] A. Bazzi, B. Masini, G. Pasolini, and P. Torreggiani, "Telecommunication systems enabling real time navigation," in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, 2010, pp. 1057–1064.

[6] A. Toppan, A. Bazzi, P. Toppan, B. Masini, and O. Andrisano, "Architecture of a simulation platform for the smart navigation service investigation," in *Wireless and Mobile Computing, Networking and Communications (WiMob), 2010 IEEE 6th International Conference on*, 2010, pp. 548–554.

[7] A. Bazzi and B. Masini, "Real time traffic updates via UMTS: Unicast versus multicast transmissions," in *Vehicular Technology Conference (VTC Fall), 2011 IEEE*, 2011, pp. 1–6.

[8] A. Bazzi, B. Masini, and G. Pasolini, "V2V and V2R for cellular resources saving in vehicular applications," in *Wireless Communications and Networking Conference (WCNC), 2012 IEEE*, 2012, pp. 3199–3203.

[9] *IEEE Trial-Use Standard for Wireless Access in Vehicular Environments (WAVE) - Multi-Channel Operation*, Std., 2006.

[10] *Standard for Information Technology- Telecommunications and Information Exchange between Systems- Local and Metropolitan Area Networks-Specific Requirements Part 11 - Amendment 6: Wireless Access in Vehicular Environment*, IEEE Std. 802.11p, 2010.

[11] Vissim. Accessed on June 2013. [Online]. Available: http://vision-traffic.ptvgroup.com/en-us/products/ptv-vissim

[12] A. Bazzi, G. Pasolini, and C. Gambetti, "SHINE: Simulation platform for heterogeneous interworking networks," in *Communications, 2006. ICC '06. IEEE International Conference on*, vol. 12, 2006, pp. 5534–5539.

[13] C. Sommer, A. Schmidt, Y. Chen, R. German, W. Koch, and F. Dressler, "On the feasibility of UMTS-based traffic information systems," *Ad Hoc Networks*, vol. 8, no. 5, pp. 506 – 517, 2010.

[14] I. Lequerica, P. Ruiz, and V. Cabrera, "Improvement of vehicular communications by using 3G capabilities to disseminate control information," *Network, IEEE*, vol. 24, no. 1, pp. 32–38, 2010.

[15] D. Valerio, A. D'Alconzo, F. Ricciato, and W. Wiedermann, "Exploiting cellular networks for road traffic estimation: A survey and a research roadmap," in *Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th*, 2009, pp. 1–5.

[16] The aktiv CoCar project. Accessed on June 2013. [Online]. Available: http://www.aktiv-online.org

[17] S. Cho, K. Geon, Y. Jeong, C.-H. Ahn, S.-I. Lee, and H. Lee, "Real time traffic information service using terrestrial digital multimedia broadcasting system," *Broadcasting, IEEE Transactions on*, vol. 52, no. 4, pp. 550–556, 2006.

[18] P. Belanovic, D. Valerio, A. Paier, T. Zemen, F. Ricciato, and C. Mecklenbrauker, "On wireless links for vehicle-to-infrastructure communications," *Vehicular Technology, IEEE Transactions on*, vol. 59, no. 1, pp. 269–282, 2010.

[19] A. Bazzi, B. M. Masini, A. Toppan, and O. Andrisano, "The benefits of a cellular enabled smart navigation in urban environments," in *Energy Conference and Exhibition (ENERGYCON), 2012 IEEE International*, 2012, pp. 1044–1048.

[20] U. Lee, E. Magistretti, B. Zhou, M. Gerla, P. Bellavista, and A. Corradi, "Efficient data harvesting in mobile sensor platforms," in *Pervasive Computing and Communications Workshops, 2006. PerCom Workshops 2006. Fourth Annual IEEE International Conference on*, 2006, pp. 5 pp.–356.

[21] U. Lee and M. Gerla, "A survey of urban vehicular sensing platforms," *Computer Networks*, vol. 54, no. 4, pp. 527 – 544, 2010.

[22] F. Losilla, A.-J. Garcia-Sanchez, F. Garcia-Sanchez, J. Garcia-Haro, and Z. J. Haas, "A comprehensive approach to WSN-based ITS applications: A survey," *Sensors*, vol. 11, no. 11, pp. 10 220–10 265, 2011.

[23] B. Masini, L. Zuliani, and O. Andrisano, "On the effectiveness of a GPRS based intelligent transportation system in a realistic scenario," in *Vehicular Technology Conference, 2006. VTC 2006-Spring. IEEE 63rd*, vol. 6, 2006, pp. 2997–3001.

[24] X. Yu, H. Zhao, L. Zhang, S. Wu, B. Krishnamachari, and V. Li, "Cooperative sensing and compression in vehicular sensor networks for urban monitoring," in *Communications (ICC), 2010 IEEE International Conference on*, 2010, pp. 1–5.

[25] A. R. Al-Ali, I. Zualkernan, and F. Aloul, "A mobile GPRS-sensors array for air pollution monitoring," *Sensors Journal, IEEE*, vol. 10, no. 10, pp. 1666–1671, 2010.

[26] B. Hull, V. Bychkovsky, Y. Zhang, K. Chen, M. Goraczko, A. Miu, E. Shih, H. Balakrishnan, and S. Madden, "CarTel: a distributed mobile sensor computing system," in *Proceedings of the 4th international conference on Embedded networked sensor systems*, ser. SenSys '06. New York, NY, USA: ACM, 2006, pp. 125–138.

[27] U. Dietz, "CoCar Feasibility Study: Technology, Business and Dissemination," CoCar Consortium, Report, May 2009.

[28] TMC. Accessed on June 2013. [Online]. Available: http://www.tisa.org/technologies/tmc/

[29] *Traffic and Traveller Information (TTI) TTI messages via traffic message coding Part 1: Coding protocol for Radio Data System Traffic Message Channel (RDS-TMC) using ALERT-C*, ISO Std. 14 819-1.

[30] Tom Tom. Accessed on June 2013. [Online]. Available: http://www.tomtom.com/

[31] P. Benko, G. Malicsko, and A. Veres, "A large-scale, passive analysis of end-to-end TCP performance over GPRS," in *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, vol. 3, 2004, pp. 1882–1892 vol.3.

[32] I. Vukovic and T. Brown, "Performance analysis of the random access channel (RACH) in WCDMA," in *Vehicular Technology Conference,*

2001. VTC 2001 Spring. IEEE VTS 53rd, vol. 1, 2001, pp. 532–536 vol.1.

[33] *Traffic and Travel Information (TTI) via Transport Protocol Expert Group (TPEG) data-streams: Parts 1 to 6*, ETSI Std. ISO 18 234.

[34] B. Bai, W. Chen, K. Letaief, and Z. Cao, "Low complexity outage optimal distributed channel allocation for vehicle-to-vehicle communications," *Selected Areas in Communications, IEEE Journal on*, vol. 29, no. 1, pp. 161–172, 2011.

[35] J. Karedal, N. Czink, A. Paier, F. Tufvesson, and A. Molisch, "Path loss modeling for vehicle-to-vehicle communications," *Vehicular Technology, IEEE Transactions on*, vol. 60, no. 1, pp. 323–328, 2011.

[36] C. Campolo and A. Molinaro, "Vehicle-to-roadside multihop data delivery in 802.11p/WAVE vehicular ad hoc networks," in *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*, 2010, pp. 1–5.

[37] A. Paier, D. Faetani, and C. Mecklenbrauker, "Performance evaluation of IEEE 802.11p physical layer infrastructure-to-vehicle real-world measurements," in *Applied Sciences in Biomedical and Communication Technologies (ISABEL), 2010 3rd International Symposium on*, 2010, pp. 1–5.

[38] P. Alexander, D. Haley, and A. Grant, "Cooperative intelligent transport systems: 5.9-GHz field trials," *Proceedings of the IEEE*, vol. 99, no. 7, pp. 1213–1235, 2011.

[39] V. Shivaldova, T. Paulin, A. Paier, and C. Mecklenbrauker, "Performance measurements of multi-hop communications in vehicular ad hoc networks," in *Communications (ICC), 2012 IEEE International Conference on*, 2012, pp. 7162–7166.

[40] A. Bazzi, B. Masini, A. Zanella, and G. Pasolini, "Vehicle-to-vehicle and vehicle-to-roadside multi-hop communications for vehicular sensor networks: Simulations and field trial," in *Communications, 2013. ICC '13. IEEE International Conference on*, June 2013.

[41] D. Jiang and L. Delgrossi, "IEEE 802.11p: Towards an international standard for wireless access in vehicular environments," in *Vehicular Technology Conference (VTC Spring), 2008 IEEE*, 2008, pp. 2036–2040.

[42] M. van Eenennaam, W. Wolterink, G. Karagiannis, and G. Heijenk, "Exploring the solution space of beaconing in VANETs," in *Vehicular Networking Conference (VNC), 2009 IEEE*, 2009, pp. 1–8.

[43] R. Schmidt, T. Leinmuller, E. Schoch, F. Kargl, and G. Schafer, "Exploration of adaptive beaconing for efficient intervehicle safety communication," *Network, IEEE*, vol. 24, no. 1, pp. 14–19, 2010.

[44] F. Li and Y. Wang, "Routing in vehicular ad hoc networks: A survey," *Vehicular Technology Magazine, IEEE*, vol. 2, no. 2, pp. 12–22, 2007.

[45] J. Bernsen and D. Manivannan, "Unicast routing protocols for vehicular ad hoc networks: A critical comparison and classification," *Pervasive and Mobile Computing*, vol. 5, no. 1, pp. 1 – 18, 2009.

www.iariajournals.org

International Journal On Advances in Intelligent Systems

ICAS, ACHI, ICCGI, UBICOMM, ADVCOMP, CENTRIC, GEOProcessing, SEMAPRO, BIOSYSCOM, BIOINFO, BIOTECHNO, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS, ENERGY, COLLA, IMMM, INTELLI, SMART, DATA ANALYTICS

issn: 1942-2679

International Journal On Advances in Internet Technology

ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING, MOBILITY, WEB

issn: 1942-2652

International Journal On Advances in Life Sciences

eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO, SOTICS, GLOBAL HEALTH

issn: 1942-2660

International Journal On Advances in Networks and Services

ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION, VEHICULAR, INNOV

issn: 1942-2644

International Journal On Advances in Security

ICQNM, SECURWARE, MESH, DEPEND, INTERNET, CYBERLAWS

issn: 1942-2636

International Journal On Advances in Software

ICSEA, ICCGI, ADVCOMP, GEOProcessing, DBKDA, INTENSIVE, VALID, SIMUL, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS, IMMM, MOBILITY, VEHICULAR, DATA ANALYTICS

issn: 1942-2628

International Journal On Advances in Systems and Measurements

ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL, INFOCOMP

issn: 1942-261x

International Journal On Advances in Telecommunications

AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA, COCORA, PESARO, INNOV

issn: 1942-2601