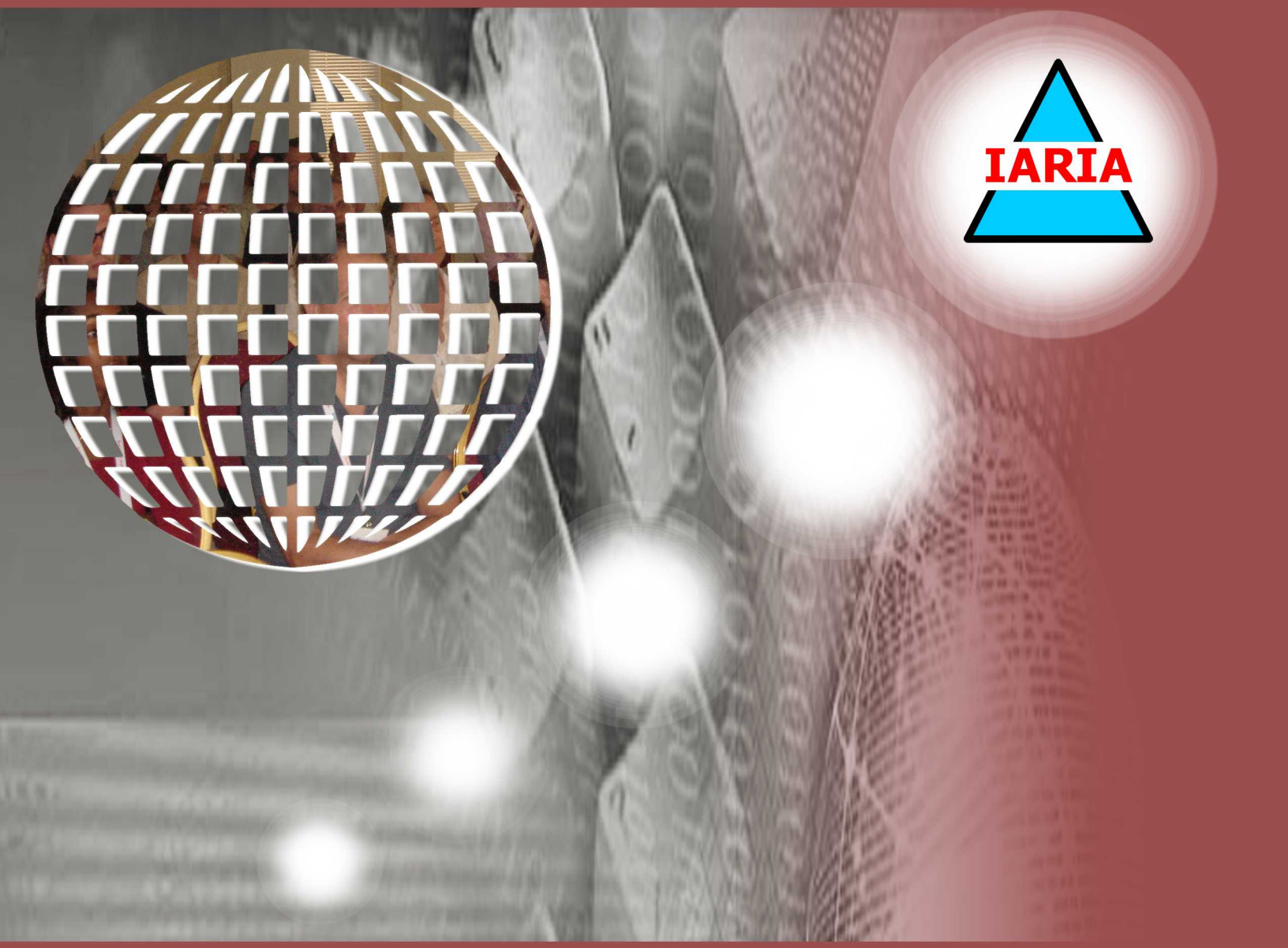
International Journal on Advances in Telecommunications



2015 vol. 8 nr. 3&4

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

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Telecommunications
Volume 8, Number 3 & 4, 2015

Eugen Borgoci, University "Politehnica"of Bucharest (UPB), Romania
Fernando Boronat Seguí, Universidad Politecnica de Valencia, Spain
Christos Bouras, University of Patras, Greece
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, São Paulo State University, Brazil
Hakima Chaouchi, Telecom SudParis, France
Silviu Ciochina, Universitatea Politehnica din Bucuresti, Romania
José Coimbra, Universidade do Algarve, Portugal
Hugo Coll Ferri, Polytechnic University of Valencia, Spain
Noel Crespi, Institut TELECOM SudParis-Evry, France
Leonardo Dagui de Oliveira, Escola Politécnica da Universidade de São Paulo, Brazil
Kevin Daimi, University of Detroit Mercy, USA
Gerard Damm, Alcatel-Lucent, USA
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
Valerie Gouet-Brunet, Conservatoire National des Arts et Métiers (CNAM), Paris
Christos Grecos, University of West of Scotland, UK
Stefanos Gritzalis, University of the Aegean, Greece
William I. Grosky, University of Michigan-Dearborn, USA
Vic Grout, Glyndwr University, UK
Xiang Gui, Massey University, New Zealand
Huaqun Guo, Institute for Infocomm Research, A*STAR, Singapore

Song Guo, University of Aizu, Japan
Ching-Hsien (Robert) Hsu, Chung Hua University, Taiwan
Javier Ibanez-Guzman, Renault S.A., France
Lamiaa Fattouh Ibrahim, King Abdul Aziz University, Saudi Arabia
Theodoros Iliou, University of the Aegean, Greece
Mohsen Jahanshahi, Islamic Azad University, Iran
Antonio Jara, University of Murcia, Spain
Carlos Juiz, Universitat de les Illes Balears, Spain
Adrian Kacso, Universität Siegen, Germany
György Kálmán, ABB AS, Norway
Eleni Kaplani, Technological Educational Institute of Patras, Greece
Behrouz Khoshnevis, University of Toronto, Canada
Ki Hong Kim, ETRI: Electronics and Telecommunications Research Institute, Korea
Atsushi Koike, Seikei University, Japan
Ousmane Kone, UPPA - University of Bordeaux, France
Dragana Krstic, University of Nis, Serbia
Archana Kumar, Delhi Institute of Technology & Management, Haryana, India
Romain Laborde, University Paul Sabatier (Toulouse III), France
Massimiliano Laddomada, Texas A&M University-Texarkana, USA
Wen-Hsing Lai, National Kaohsiung First University of Science and Technology, Taiwan
Zhihua Lai, Ranplan Wireless Network Design Ltd., UK
Jong-Hyouk Lee, INRIA, France
Wolfgang Leister, Norsk Regnesentral, Norway
Elizabeth I. Leonard, Naval Research Laboratory - Washington DC, USA
Jia-Chin Lin, National Central University, Taiwan
Chi (Harold) Liu, IBM Research - China, China
Diogo Lobato Acatauassu Nunes, Federal University of Pará, Brazil
Andreas Loeffler, Friedrich-Alexander-University of Erlangen-Nuremberg, Germany
Michael D. Logothetis, University of Patras, Greece
Renata Lopes Rosa, University of São Paulo, Brazil
Hongli Luo, Indiana University Purdue University Fort Wayne, USA
Christian Maciocco, Intel Corporation, USA
Dario Maggiorini, University of Milano, Italy
Maryam Tayefeh Mahmoudi, Research Institute for ICT, Iran
Krešimir Malarić, University of Zagreb, Croatia
Zoubir Mammeri, IRIT - Paul Sabatier University - Toulouse, France
Herwig Mannaert, University of Antwerp, Belgium
Adrian Matei, Orange Romania S.A, part of France Telecom Group, Romania
Natarajan Meghanathan, Jackson State University, USA
Emmanouel T. Michailidis, University of Piraeus, Greece
Ioannis D. Moscholios, University of Peloponnese, Greece
Djafar Mynbaev, City University of New York, USA
Pubudu N. Pathirana, Deakin University, Australia
Christopher Nguyen, Intel Corp., USA
Lim Nguyen, University of Nebraska-Lincoln, USA
Brian Niehöfer, TU Dortmund University, Germany

Serban Georgica Obreja, University Politehnica Bucharest, Romania
Peter Orosz, University of Debrecen, Hungary
Patrik Österberg, Mid Sweden University, Sweden
Harald Øverby, ITEM/NTNU, Norway
Tudor Palade, Technical University of Cluj-Napoca, Romania
Constantin Paleologu, University Politehnica of Bucharest, Romania
Stelios Papaharalabos, National Observatory of Athens, Greece
Gerard Parr, University of Ulster Coleraine, UK
Ling Pei, Finnish Geodetic Institute, Finland
Jun Peng, University of Texas - Pan American, USA
Cathryn Peoples, University of Ulster, UK
Dionysia Petraki, National Technical University of Athens, Greece
Dennis Pfisterer, University of Luebeck, Germany
Timothy Pham, Jet Propulsion Laboratory, California Institute of Technology, USA
Roger Pierre Fabris Hoefel, Federal University of Rio Grande do Sul (UFRGS), Brazil
Przemyslaw Pochec, University of New Brunswick, Canada
Anastasios Politis, Technological & Educational Institute of Serres, Greece
Adrian Popescu, Blekinge Institute of Technology, Sweden
Neeli R. Prasad, Aalborg University, Denmark
Dušan Radović, TES Electronic Solutions, Stuttgart, Germany
Victor Ramos, UAM Iztapalapa, Mexico
Gianluca Reali, Università degli Studi di Perugia, Italy
Eric Renault, Telecom SudParis, France
Leon Reznik, Rochester Institute of Technology, USA
Joel Rodrigues, Instituto de Telecomunicações / University of Beira Interior, Portugal
David Sánchez Rodríguez, University of Las Palmas de Gran Canaria (ULPGC), Spain
Panagiotis Sarigiannidis, University of Western Macedonia, Greece
Michael Sauer, Corning Incorporated, USA
Marialisa Scatà, University of Catania, Italy
Zary Segall, Chair Professor, Royal Institute of Technology, Sweden
Sergei Semenov, Broadcom, Finland
Sandra Sendra Compte, Polytechnic University of Valencia, Spain
Dimitrios Serpanos, University of Patras and ISI/RC Athena, Greece
Adão Silva, University of Aveiro / Institute of Telecommunications, Portugal
Pushpendra Bahadur Singh, MindTree Ltd, India
Mariusz Skrocki, Orange Labs Poland / Telekomunikacja Polska S.A., Poland
Leonel Sousa, INESC-ID/IST, TU-Lisbon, Portugal
Liana Stanescu, University of Craiova, Romania
Cosmin Stoica Spahiu, University of Craiova, Romania
Young-Joo Suh, POSTECH (Pohang University of Science and Technology), Korea
Hailong Sun, Beihang University, China
Jani Suomalainen, VTT Technical Research Centre of Finland, Finland
Fatma Tansu, Eastern Mediterranean University, Cyprus
Ioan Toma, STI Innsbruck/University Innsbruck, Austria
Božo Tomas, HT Mostar, Bosnia and Herzegovina
Piotr Tyczka, Poznan University of Technology, Poland

International Journal on Advances in Telecommunications
Volume 8, Numbers 3 & 4, 2015

CONTENTS

Bernd E. Wolfinger, Department of Computer Science, Telecommunications and Computer Networks University of Hamburg, Germany
Nico R. Wilzek, Department of Computer Science, Telecommunications and Computer Networks University of Hamburg, Germany
Edgar E. Báez, Superior School of Computing National Polytechnic Institute ESCOM-IPN, Mexico

Optimized Wireless Transmission of Stereo Images and 3-D Reconstruction on Hardware

Apurva Naik, Gourang Mulay, Arti Khaparde
Department of Electronics and Telecommunication
Maharashtra Institute of Technology
Pune, Maharashtra, India
Emails :{apurva.naik, gourang.mulay, arti.khaparde} @mitpune.edu.in

***Abstract*—Stereo images are captured using cameras connected to PC. These images are segmented and stored in the compressed form. The compressed segmented images are transmitted to the ARM-9 processor based system by using ZIGBEE wireless module. These images, when received at the receiver end, are recovered, and a 3-D image is generated on the TFT display. Depth levels are also estimated from segmented stereo images and transmitted through ZIGBEE module to ARM-9 processor based hardware. Depth levels received at the hardware are used to control a robot. This proposal is a prototype that can be implemented for vision based industrial applications. The present paper deals with the transmitter-receiver link for stereo images and movement of robot proportional to estimated depth levels.**

Keywords-ZIGBEE; Particle Swarm Optimization; Darwinian Particle Swarm Optimization; Fractional Order Darwinian Particle Swarm Optimization; robot; SAM-9M-10-G-45-EK.

I. INTRODUCTION

As humans, we perceive the three-dimensional structure of the world around us with apparent ease and also estimate depth of view very easily. There is a need for developing such perception of depth with ease using computer vision and embedded technology similar to humans. The methods designed for 3-D generation, especially using embedded system which has limited resources, should use lesser computation and lesser memory with greater accuracy. Focus of the present paper is binocular stereo vision, which uses two cameras placed at baseline distance and captures two views of image commonly known as left and right view of image. After basic processing steps like camera calibration and rectification, clustering based segmentation techniques are applied on left and right views.

Initially, well-known clustering algorithms like K-means and Mean shift are used for segmentation of stereo images. It has been shown that biologically inspired algorithms like Particle Swarm Optimization (PSO), Darwinian Particle Swarm Optimization (DPSO), Fractional Order Darwinian Particle Swarm Optimization (FO-DPSO) can be successfully used to segment the stereo images. There are significant advantages of using above algorithms for segmentation and also these methods give compression of stereo images. Segmentation based techniques are preferred over edge based technique for stereo matching because dense disparity map is obtained due to segmentation based stereo matching. Stereo matching algorithms are applied on segmented images to generate disparity maps and compressed 3D images are generated without disturbing depth levels in the image. The comparison between the disparity of the original stereo image pair and that of the segmented image pair is carried out. The reconstructed 3D images are analyzed based on compression ratio (CR) and Peak-Signal to Noise Ratio (PSNR). Segmented images are given to the disparity estimation algorithm. Depth levels are estimated with the help of disparity values obtained from the disparity estimation algorithm.

A single camera image does not give information about depth levels. Information about depth is required in several applications such as satellite imaging, robotic vision, target tracking and automatic map making. Hence, stereo matching uses minimum two views for processing. Basic aim of stereo matching is to extract depth information from image. Most of stereo imaging algorithms have been left largely unexplored and not implemented on hardware in last decade probably due to memory and hardware constraints and lack of resources on hardware. The primary aim of this paper is to analyze the existing stereo matching algorithms on segmented stereo images and wireless transmission of these images and estimated depth levels to a portable hardware. The hardware is able to display these images in 3-D form on TFT display. The hardware is also able to drive a robot (Simple DC motor driven linear assembly, which moves exactly the calculated distance) depending on depth levels which are received by receiver.

The movement of a robot can be used in robotic vision applications. Until recently, stereography was used either for entertainment purpose or DEM (Digital Elevation Model) for depth analysis of sea bed [1]. This novel approach will help us to control the unmanned vehicle to perform the numerous tasks in medical, mining applications and in the volumetric analysis of water reservoirs, etc., which requires the knowledge of depth. For example, one of the applications that can be developed is for computer-aided surgery. Images can be captured with the help of a stereoscopic endoscope. These images can be transferred to the control room. By performing an analysis and using depth information, the surgeon can instruct a robot to perform certain tasks. This paper is the extension of our previous work [1]. In the previous paper, wireless link between PC to PC was used. In the present paper, the second PC is replaced by ARM-9 evaluation kit SAM-9M-

10-G-45-EK. The second objective of this paper is also a development of low cost prototype for vision application. For capture of actual stereo images, the distance between two webcams should be at least 6cm, as shown in Fig. 1.

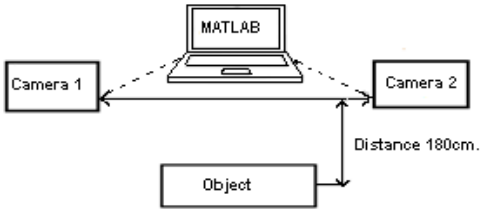


Figure 1. Block diagram of camera setup.

The camera should be placed at a distance of 1.8m from the object to be captured [1]. For estimation of the focal length, which is required for depth calculation camera calibration, the toolbox of MATLAB [2] has been used. For estimating the focal length of the camera, 20 images of a chess board, each image having a different orientation from the other were taken. A procedure [2] was followed and the focal length obtained was approximately 670 pixels. Focal length is verified with another method, i.e., simple 'lenses law' in optics and the focal length obtained was 17.47 cm. These values are further utilized in calculation of depth from disparity.

The paper is organized as follows. Section II describes the experimental setup of the entire system. Section III describes segmentation methodology and compression achieved for storage of 3-D images before transmission. Section IV describes stereo matching algorithms used and depth estimation. Section V describes the ZIGBEE module and protocol. Section VI gives details of hardware on which 3-D image is generated and also used for wireless transmission of depth and control of robot. Section VII gives results and discussion. Conclusion and future work that can be done in the present project are described in Section VIII.

II. EXPERIMENTAL SETUP

Experimental setup consists of two Logitech C-310 webcams, one general purpose PC, one ARM-9 based microprocessor SAM-9M-10-G-45 evaluation kit, one robot assembly, 4 ZIGBEE modules (two coordinator node and two router node). ZIGBEE modules have been used for transmission of real-time images and depth values. In the present setup, ZIGBEE module of DIGI Company (XBee RF Modules) is used. ZIGBEE standard operates on the IEEE 802.15.4 physical radio specification and operates in unlicensed bands including 2.4GHz, 900MHz and 868MHz. Each ZIGBEE module is connected to a PC via a Serial to USB Convertor for communication with MATLAB program. MATLAB has been used to segment the image data, and then transmit data to the router nodes. This segmented image data is used to generate a 3-D image on the hardware connected to router node. For segmentation of captured real-time stereo images various segmentation algorithms like Mean shift segmentation, K- means clustering, Particle swarm optimization, Darwinian Particle Swarm Optimization, Fractional order Darwinian particle swarm optimization [3-12] are used. The segmentation not only gives compression in stereo image size but also retains depth levels that are present in the image. Compression achieved due to segmentation saves wireless transmission bandwidth as well as reduces memory requirement when images are to be stored on the memory of hardware board. The Segmented images are applied as input to the disparity algorithm to estimate the depth values. The coordinator node of ZIGBEE module sends this depth data directly to the hardware. Another router node receives the depth data, which is then decoded, and these decoded values are used to control the robot. Block diagram of hardware implementation of stereo matching for wireless transmission is shown in Fig. 2. There is also a provision of transmission of the same data through USB port of PC if wireless link fails.

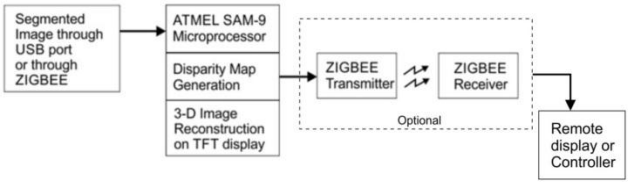


Figure 2. Hardware implementation of stereo matching .

III. SEGMENTATION METHODOLOGY

For some applications like stereo vision and matching, whole images cannot be processed, as it not only increases the computational complexity, but it also requires more memory [4]. Purely pixel-based methods are insufficient to express information of the image. The human identifies the objects by analyzing features of the objects such as color, texture and shape. The basic algorithm for stereo matching is not very complicated but is computationally exhaustive and limits its usage for real-time applications. Purely pixel-based methods used for stereo matching are insufficient to express information of the image. The quality of matching can be improved if a label is assigned to each pixel of the left and right image such that pixel with same label shares same intensity value. This forms different regions in the image that are more meaningful than individual pixels. This process of partitioning an image, commonly called as segmentation, is used prior to stereo matching. Thus, segmentation-based stereo matching is new methodology introduced.

Fig. 3 explains the complete methodology of proposed segmentation based stereo matching. Initially, the performance of algorithms is tested using the Middlebury

data set [13]. The algorithms are tested for real time images also. Dataset images do not require camera calibration and rectification steps. Real-time image inputs require camera calibration. These image data are applied to five different segmentation algorithms shown in Fig. 3. One more reason behind using segment based methods is that these techniques perform well in reducing the ambiguity associated with texture-less regions and enhancing noise tolerance.

These segmented images are applied to disparity estimation algorithms to create 3-D image and disparity map. There are two different disparity estimation algorithms to which segmented stereo images are applied as input. The result of stereo matching is disparity map and 3-D image. The performance of stereo matching is verified for various segmentation algorithms. To keep the computational complexity down, an algorithm relying on local Winner Take All (WTA) optimization was compared against Line Growing (LG) algorithm. The same sequence of steps is applied to real-time images also, and approximate depth values for real-time views are estimated.

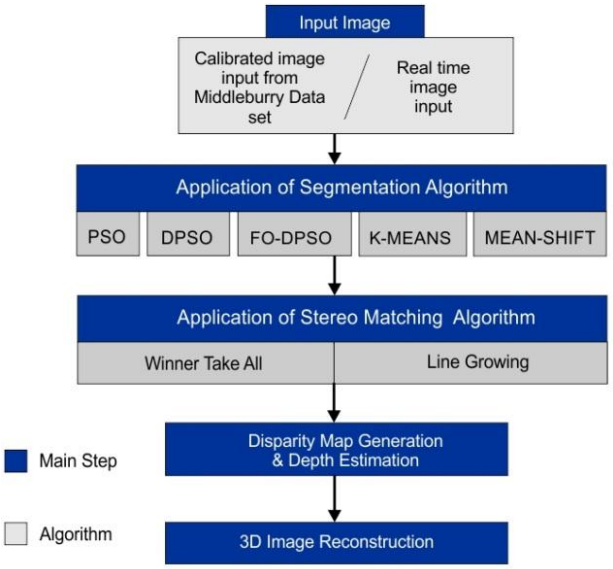


Figure 3. Complete Methodology for 3-D reconstruction and depth estimation.

Proposed segment-based stereo matching performs four consecutive steps. First, it segments the reference images using robust segmentation method; second, it gets initial disparity map using local match method; third, a plane fitting technique is employed to obtain disparity planes. Finally, optimal disparity plane assignment is approximated by using optimization methods. Proposed segmentation-based stereo matching system is shown in Fig. 4. After segmentation, Winner Take All or Line Growing algorithm is applied for stereo matching. For depth estimation Line Growing algorithm is used.

A. Clustering based segmentation Technique

Segmentation, the process of partitioning a digital image into multiple objects is widely used method in image classification and recognition. It is a low-level image processing task aiming at partitioning an image into homogeneous regions. The result of image segmentation is a set of regions. Image segmentation techniques can be grouped into several categories such as edge-based segmentation, region-oriented segmentation, histogram thresholding and clustering algorithms.

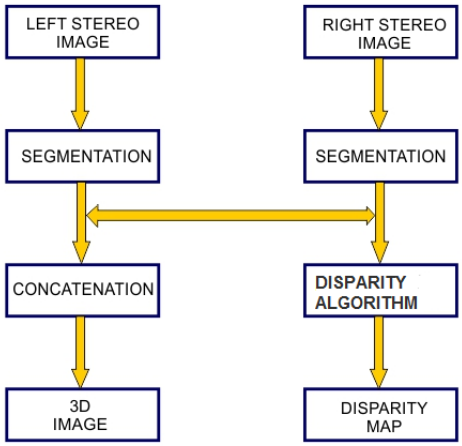


Figure 4. Segmentation based stereo matching system.

For present work, clustering based segmentation techniques were used to partition image into segments. The advantage of using segmentation-based matching over edge-based matching is that it reduces the mismatch in low texture region and occluded areas.

In the literature, various methods are available to cluster data sets. Broadly, they can be classified as parametric (a kind of density is known) and non-parametric (a form of density is not known) methods. In a parametric method like K-means clustering, prior assumptions of the number of clusters are made. This is a function minimization technique, where the objective function is the squared error distance measure. In non-parametric methods such as Mean shift clustering and Particle Swarm Optimization, no prior assumptions are made on the number of clusters. Mean shift is a procedure for locating the maxima of a mapped function given a set of discrete data points sampled from that function. The computational time and fitness value are most important indicators for clustering algorithms. All algorithms of this kind come into a class of statistical based algorithms. Statistical measures reduce dimensions of data and retain information. These kinds of methods are explored in the present work. Segmentation technique based on calculating mode is called Mean shift clustering [4]. The state of the art is to employ

Swarm-Based collective intelligence, also called biologically or nature inspired algorithms to image segmentation [6-12]. Key issues in the design of any clustering based segmentation are the choice of number and type of features used, the distance metric chosen to measure similarity, data reduction techniques used, and pre and post processing routines applied. Moreover, in real time applications, using high-speed algorithm is the main objective. Particle Swarm Optimization is a recently proposed population based stochastic optimization algorithm that is inspired by social behaviors of animals like fish schooling and bird flocking. PSO has superior search performance for many hard optimization problems with faster and more stable convergence rates [14]. PSO converges in the early stages of the searching process but saturates or terminates in the later stages. It is hard to obtain any significant improvements by examining neighboring solutions in the later stages of the search. Sometimes PSO algorithms may get trapped in local maxima or minima, and there is need to apply algorithms like Darwinian Particle Swarm Optimization (DPSO). Starting with basic segmentation algorithms such as Mean shift clustering and K-means, the proposed work implements bio-inspired methods for segmentation of stereo images like Particle Swarm Optimization (PSO), Darwinian Particle Swarm Optimization (DPSO) and Fractional Order Darwinian Particle Swarm Optimization (FO-DPSO). Comparison of traditional PSO with DPSO and FO-DPSO for stereo image segmentation is discussed in the sections given below. The following segmentation algorithms were implemented prior to stereo matching.

- Mean shift
- K- means
- Particle Swarm Optimization (PSO).
- Darwinian Particle Swarm Optimization (DPSO).
- Fractional order Darwinian Particle Swarm Optimization (FO-DPSO).

1) Mean Shift Segmentation

Mean shift is a nonparametric iterative algorithm or a nonparametric density gradient estimation using a generalized kernel approach. Mean shift is one of the most powerful clustering techniques. Mean shift algorithm was introduced by Fukunaga and Hostetler [4]. It considers feature space as an empirical probability density function. Probability distribution function for discrete image data values is given as the set of discrete pixels. Probability values cannot be larger than 1 (100%). Therefore, the first constraint is that the area under the entire probability distribution function should be 1:

$$\int_{-\infty}^{\infty} \mathrm{PDF}(x)dx = \sum_{\mathrm{pixel}=1}^{N} \Delta x \Delta y = N \Delta x \Delta y = 1 \quad (1)$$

where N is the number of the pixels in PDF image, Δx and Δy is the width and height of a pixel respectively. If the input is a set of points, then Mean shift considers them as sampled from the underlying probability density function. If dense regions (or clusters) are present in the feature space, then they correspond to the mode (or local maxima) of the probability density function. The groups associated with the given mode using Mean shift can also be identified. For each data point, Mean shift associates it with the nearby peak of the dataset's probability density function. For each data point, Mean shift defines a window around it and computes the average of the data point. Then it shifts the centre of the window to the mean value and repeats the algorithm till it converges. After each iteration, the window moves to a denser region of the dataset.

At the high level, the Mean shift algorithm can be stated as follows:

- Fix a window around each data point.
- Compute the mean of data within the window.
- Shift the window to the mean and repeat until convergence.

The Mean shift technique comprises of two basic steps: a Mean shift filtering of the original image data and a subsequent clustering of the filtered data points.

a) Mean Shift Filtering

Let x_1, x_2, x_n where n is the number of data points in d-dimensional space. In the Mean shift clustering, each data point is shifted to the average of the other data points in its neighborhood. This is done by using a Gaussian kernel, based on Euclidean distance between two data points (r), which is given by

$$K(r) = e^{-||r||^2} \quad (2)$$

The dense regions in the feature space correspond to the local maxima of the underlying distribution. The filtering step of the Mean shift segmentation algorithm consists of analyzing the probability density function underlying the image data in feature space. Consider the feature space composed of the original image data represented as the (x, y) location of each pixel, plus its color in L*u*v* (derived from lab color space with all components guaranteed to be positive) space. The modes of the probability density function underlying the data in this area will correspond to the locations with highest data density. For segmentation, the data points close to these high-density points (modes) should be clustered together. Filtering step in the Mean shift consists of finding the modes of the underlying probability density function (pdf) and associating with them any points in their basin of attraction. For a data point x in feature space, the density

gradient is estimated as being proportional to the Mean shift vector:

$$\hat{\nabla} f(x) \propto \frac{\sum_{i=1}^{n} x_i g\left(\left\|\frac{x-x_i}{h}\right\|\right)}{\sum_{i=1}^{n} g(\|\frac{x-x_i}{h}\|)} - x \quad (3)$$

where x_i are the data points, x is a point in the feature space, n is the number of data points (pixels in the image), and g is the profile of the symmetric kernel G.

Here, the simple case where G is the uniform kernel with radius vector h is used. Thus, the above equation simplifies to

$$\hat{\nabla} f(x) \propto \left[\frac{1}{|S_X h_s h_r} \sum_{x_i \in S_{x,h_s,h_r}} x_i\right] - x \quad (4)$$

where $S_{x,}h_s, h_r$ represents the sphere in feature space centred at x and having spatial radius h_s(spatial range to consider while computing mode) radius h_r (RGB range), and the x_i represent the data points within that sphere. For every data point (pixel in the original image) x, the gradient estimate (Eqn. (4)) is iteratively computed and x is moved in that direction, until the gradient is below a threshold T_h (threshold for the convergence). Thus, the points where $\hat{\nabla} f(x') = 0$, i.e., the modes of the density estimate were calculated. Afterwards, the point x was replaced with x', the mode with which it is associated. Finding the mode associated with each data point helps to smooth the image while preserving discontinuities. If two points x_i and x_j are far from each other in feature space, then x_j does not contribute to the Mean shift vector gradient estimate, and the trajectory of x_i will move it away from x_j. Hence, pixels on either side of a strong discontinuity will not attract each other. However, filtering alone does not provide segmentation as the modes found are noisy. This "noise" stems from two sources. First, the mode estimation is an iterative process; hence, when it converges within the threshold provided with some numerical error and secondly when an area in feature space is larger than S_{x_j}, h_s, h_r and where the colour features is uniform or has a gradient of 1. Since the pixel coordinates are identical by design, the Mean shift vector will be 0 in this region, and the data points will not move and hence may not converge to a single mode.

b) Mean Shift Clustering

After Mean shift filtering, each data point in the feature space has been replaced by its corresponding mode. Some points may have the same mode, but many may not have despite the fact that they may be less than one kernel radius apart. In the original Mean shift segmentation paper, clustering is described as a simple post-processing step in which any modes that are less than one kernel radius apart are grouped together and their basins of attraction (regions for which all trajectories lead to the same mode) are merged. This suggests using single linkage clustering, which actually converts the filtered points into segmentation. Typically, the Mean shift is run for each point, or sometimes points are selected uniformly from the feature space.

c) Effect of Mean Shift Parameter Variation

The Mean shift filtering stage has two parameters corresponding to the bandwidths (radii of the kernel) for the spatial (h_s) and color (h_r) features. Slight variations in h_r can cause large changes in the granularity of the segmentation. Fig. 5 shows left and right views of original Tsukuba image. By adjusting the color bandwidth, the different segmented views of Tsukuba are illustrated in Fig. 6 and Fig. 7. The optimum values obtained for RGB image are spatial range h_s of 40, RGB range h_r of 3, and the threshold for convergence as 3. This is significant problem with respect to using Mean shift segmentation as a reliable pre-processing step for other algorithms, such as stereo matching.

Mean shift clustering uses single point for locating modes (local maxima). Recently, researchers have become interested in finding multiple local optima of a given multi-modal function in a d–dimensional search space. For this purpose, nature-inspired techniques are used.

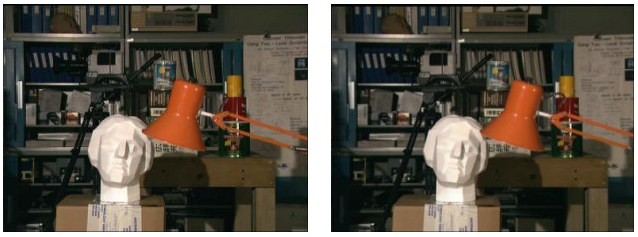


Figure 5. Left and right views of original Tsukuba Image.

Figure 6. Segmented left and right views of Tsukuba using Mean shift segmentation technique (hr=3).

Figure 7. Segmented Views of Tsukuba using Mean shift segmentation Technique (hr=2, hr=4).

2) K-means Clustering

The K-means algorithm does not have the above mentioned problems. The K-means algorithm typically requires only ***O*** (kN) operations, so that K-means algorithm can be applied to the relatively large dataset. To reduce computations, segmentation was carried out using K-means algorithm. K-means is one of most popular clustering algorithms. It is simple, fast and efficient. It can be compared with the Mean shift on the some parameters. One of the most significant differences is that K-means makes two assumptions – the number of clusters is given as input, and the clusters are shaped spherically (or elliptically).

K-means is one of the simplest unsupervised learning algorithms for solving clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (in present application k=3) fixed a priori. The main idea is to define k centroids, one for each cluster. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is remaining, the first step is completed, and an early grouping is done. At this point re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After these k new centroids are calculated, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop, the K centroids change their location step by step until no more changes are done, and centroids do not move anymore. Finally, this algorithm aims at minimizing an *objective function;* in this case, a squared error function. The objective function is:

$$J = \sum_{i-1}^{m} \|x_i^{(j)} - cj\|^2 \qquad (5)$$

where $\|x_i^{(j)} - cj\|^2$ is a Euclidean distance measure between a data point $x_i^{(j)}$ and the cluster center c_j an indicator of the distance of the n data points from their respective cluster centres. The algorithm is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

K-means is very sensitive to initializations. A wrong initialization can delay convergence or sometimes even result in false clusters. Similarly, K-means is sensitive to outliers but the Mean shift is not very sensitive. Results of

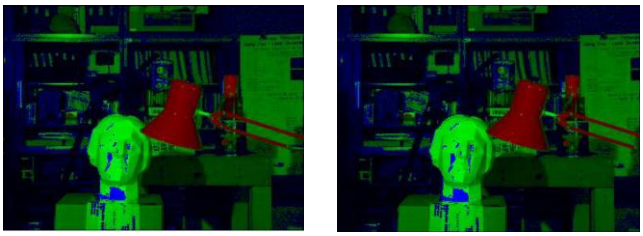


Figure 8. Segmented left and right views of Tsukuba using K-means segmentation technique.

K-means clustering are shown in Fig. 8 with the value of K=3. The performance of the above algorithm may be affected by the chosen value of K. Therefore, instead of using a single predefined K, a set of values might be adopted. It is important for the number of values considered to be reasonably large, to reflect the particular characteristics of the data sets. At the same time, the selected values have to be significantly smaller than the number of objects in the data sets, which is the primary motivation for performing clustering. To find a satisfactory clustering result, numbers of iterations are carried out with different values of K. The validity of the clustering result is assessed only visually without applying any formal performance measure. With this approach, it was difficult to evaluate the clustering result for multi-dimensional data set like images. Since K-means clustering is used as a pre-processing tool, the focus was on the effect of the clustering results on the performance of the stereo matching algorithm. In an attempt to improve performance for multidimensional data set, three different algorithms that are based on Swarm Intelligence were considered.

3) Swarm Intelligence based Clustering Algorithms

In image segmentation, the decision to assign a pixel to a particular class is simultaneously based on the feature vector of the pixel and some additional information derived from the segmentation step. To make this approach practical, an accurate segmentation of the image is needed [11]. *Thresholding* is one of the most commonly used methods for the segmentation of images into two or more clusters [7]. Thresholding techniques can be divided into two different types: optimal thresholding methods and property-based thresholding methods [19]. Algorithms in the former group search for the optimal thresholds that make the threshold classes on the histogram reach the desired characteristics. Usually, thresholds are selected by optimizing an objective function. The later group detects the thresholds by measuring some selected property of the histogram. Property-based thresholding methods are fast, which make them suitable for multilevel thresholding. The task of determining $n - 1$ optimal thresholds for n-level image thresholding could be formulated as a multidimensional optimization problem. To solve such a task, several biologically inspired algorithms have been

explored in image segmentation [6-19]. Bio-inspired algorithms have been used in situations where conventional optimization techniques cannot find a satisfactory solution, or they take too much time to find it, e.g., when the function to be optimized is discontinuous and cannot be differentiated and having too many nonlinearly related parameters [17]. One of the best-known bio-inspired algorithms is particle swarm optimization (PSO) [18]. The PSO consists of a number of particles that collectively move in the search space (e.g., pixels of the image) in search of the global optimum (e.g., maximizing the between-class variance of the distribution of intensity levels in the given image). A general problem with the PSO and similar optimization algorithms is that they may get trapped in local optimum points, and the algorithm may work in some problems but may fail in others. To overcome such a problem, the Darwinian PSO (DPSO) was presented [16]. In the DPSO, multiple swarms of test solutions performing just like an ordinary PSO may exist at any time, with rules governing the collection of swarms that are designed to simulate natural selection. More recently, an extension to the DPSO using fractional order calculus (FO-DPSO) to control the convergence rate of the algorithm is proposed. [17] The clustering algorithms mentioned above are applied to the segmentation of stereo images in the Middlebury dataset, and real-time images. Tuning of PSO parameter values for segmentation that will be useful in stereo applications is carried out. Experimental results show that the PSO based clustering algorithm performs better than well-known clustering algorithms (K-means and Mean shift that are already explained above) in all measured criteria. The introduction to these algorithms is presented in following sections.

a) Image segmentation using Particle Swarm Optimization (PSO)

Particle swarm optimization (PSO) is an optimization technique developed by Dr. Eberhart and Dr. Kennedy in 1995, inspired by social behavior of bird flocking or fish schooling. The particle swarm concept originated as a simulation of the simplified social system. The original intent was to simulate the choreography of birds of a bird flock or fish school graphically. However, it was found that particle swarm model can be used as an optimizer. Consider the following scenario: a group of birds are randomly searching for food in an area. There is only one piece of food in the area being searched. All the birds do not know where the food is. But they know how far the food is. So the effective strategy is to follow the bird that is nearest to the food. PSO learned from the scenario and used it to solve the optimization problems. In PSO, each single solution is a "bird" in the search space called as "particle." All of the particles have fitness values that are evaluated by the fitness function to be optimized and have velocities that direct the flying of the particles. The particles fly through the problem space by following the current optimum particles. Suppose a global optimum of an n –dimensional function is to be located. The function may be mathematically represented as

$$f(x_1, x_2, x_3, \dots\dots, x_n) = f(\overrightarrow{X}) \quad (6)$$

Where $\vec{x}$ is the search variable vector, which represents the set of independent variables of the given function. The task is to find out such a $\overrightarrow{x}$, that the function value $f(\overrightarrow{x})$ is either minimum or maximum denoted by f^* in the search range. If the components of $\vec{x}$ assume real values, then the task is to locate a particular point in the n-dimensional hyperspace that is a continuum of such points. There are two key steps when applying PSO to optimization problems viz. the representation of the solution and the fitness function. One of the advantages of PSO is that PSO takes real numbers as particles. For example, to find the solution for $f(x) = x1^2 + x2^2 + x3^2$, the particle can be set as (x1, x2, x3), and fitness function is f(x). Then the standard procedure can be used to find the optimum. The searching is a repetitive process, and the stop criterion is that either maximum iteration number is reached, or the minimum error condition is satisfied. It is not easy to find optima for some functions. To locate global optima quickly on such functions require parallel search techniques. Here, many agents start from different initial locations and go on exploring the search space until some of the agents reach the optimal global position. The agents may communicate among themselves and share the fitness function values found by them. PSO is multi-agent parallel search technique. Particles are conceptual entities, which fly through the multidimensional search space. At any particular instant, each particle has position and velocity. The position vector of the particle with respect to the origin of search space represents the trial solution of the search problem. At the beginning, a population of particles is initialized with random positions marked by vectors $(\overrightarrow{x_i})$ and random velocity $(\overrightarrow{v_i})$. The population of such particles is called a "swarm" S. A neighborhood relation N is defined in the swarm that determines whether any two particles P_i and P_j are neighbours or not. Thus, for any particle P, a neighborhood can be assigned as N (P), containing all the neighbors of that particle. A traditional strategy is N=S for each particle, i.e., any particle has all the remaining particles in the swarm in its neighborhood. Each particle P has two state variables viz., its current position x_t^n and its current velocity v_t^n . It is also equipped with small memory comprising its previous best position and velocity.

The PSO has following algorithmic parameters

- Maximum and minimum velocity (V_{max}), (V_{min}): it determines the maximum change one particle can take during each iteration.
- An inertial weight factor.
- Three uniformly distributed random numbers r_1, r_2 and r_3 that respectively determine an influence of global best and local best on the velocity update equation.

- Two constant multiplier terms ρ_1 and ρ_2 known as "swarm confidence" and "self-confidence", respectively along with multiplier ρ_3. The value selected for ρ_1, ρ_2, ρ_3 are such that equal weight is assigned to each term in PSO velocity equation.
- The number of particles N: the typical range is 20 - 40. Actually, for most of the problems, ten particles is large enough to get good results. For some severe or unusual problems, one can try 100 or 200 particles as well.

To model the swarm, each particle 'n' moves in a multidimensional space according to position (x_t^n) and velocity (v_t^n) values. The position and velocity values are highly dependent on

[i]. local best ($\check{x}_t^n$):

[ii]. personal best (pbest), which is the best solution (fitness) it has achieved so far

[iii]. neighborhood best ($\check{n}_t^n$) , i.e., best position of its neighbour and global best ($\check{g}_t^n$), i.e., the best value, obtained so far by any particle in the population of the swarm.

After finding the three best values, the particle updates its velocity and positions with the basic PSO equations

$$v_{t+1}^n = wv_t^n + \rho_1 r_1(\check{g}_t^n - x_t^n) + \rho_2 r_2(\check{x}_t^n - x_t^n) + \rho_3 r_3(\check{n}_t^n - x_t^n) \quad (7)$$

$$x_{t+1}^n = x_t^n + v_{t+1}^n \quad (8)$$

The coefficients w , ρ_1, ρ_2 and ρ_3 assign weights to the inertial influence, the global best, local best and the neighbourhood best when determining the new velocity respectively. Typically, the inertial weight is set to a value slightly less than 1. $\rho_1, \rho_2, and\ \rho_3$ are constant integer values that represent "cognitive" and "social" components. Different results can be obtained by assigning different influences for each component. For present work, which uses PSO for image segmentation, neighborhood best is not considered and hence, ρ_3 is set to zero. The parameters r_1, r_2, r_3 are random vectors with each component is a uniform random number between 0 and 1. The intent is to multiply a new random component per velocity dimension, rather than multiplying same component with each particle's velocity dimension.

The particles in the PSO are evaluated for the fitness function, which is defined as the between-class variance σ_B^2 of the image intensity distributions. Equations (7) and (8) are modified to (9) and (10) given below to suit the basic equation for image segmentation operation of red component in RGB image.

$$v_R = w * v_R + rand_1 * (\rho_1 * (X_{best} - X_R)) + rand_2(\rho_2 * (g_{auxR} * g_{best} - X_R)) \quad (9)$$

$$X_R = X_R + v_R \quad (10)$$

$\boldsymbol{v_R}$ is the particle velocity; $\boldsymbol{X_R}$ is the current particle (solution) and for the present application it is pixel intensity in the red component of the image and randomly generated using the maximum and minimum intensity values using the histogram. w is an inertial factor. The parameters 'pbest' and 'gbest' are defined as stated before. rand () is a random number between (0,1). ρ_1, ρ_2 are specified as above. g_{auxR} is a unity matrix of size (N,1) matching the matrix dimensions of gbest and X_R.

Each candidate solution can be thought of as a particle "flying" through the fitness landscape finding the maximum or minimum of the objective function. Similar equations can be written for green and blue components in the image.

Table I shows the steps in PSO algorithm used.

Table I. PSO ALGORITHM

Main Program Loop

Initialize swarm Position ($\boldsymbol{x_n}$) , Velocity ($\boldsymbol{v_n}$) **,** Local Best $\check{\boldsymbol{x}}_{\boldsymbol{n}}$, Neighbourhood Best ($\check{\boldsymbol{n}}_{\boldsymbol{n}}$), Global Best **($\check{\boldsymbol{g}}_{\boldsymbol{n}}$)**
Loop:
for all particles evaluate the fitness φ^C of each particle using Equation (4.17)
update ($\check{\boldsymbol{v}}_{\boldsymbol{n}}$) , ($\check{n}_n$) and ($\check{g}_n$)
update v_n *and* x_n
end
until stopping criteria (convergence)

In the beginning, the particles' position is randomly set within boundaries of the search space. The search area will depend on the number of intensity levels L. For the present application, images are 8-bit images and particles are deployed between 0 and 255.

Fig. 9 shows segmented views of Tsukuba using PSO algorithm. The segmentation time is not the same in each run for the same parameter values. It is found by averaging over ten runs of the algorithm.

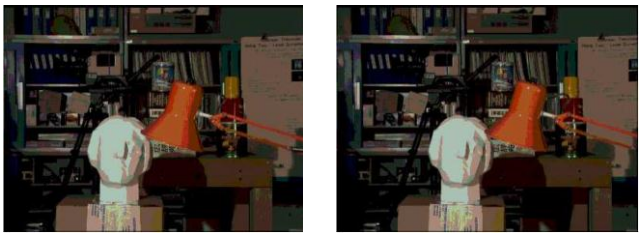


Figure 9. Segmented left and right views of Tsukuba using PSO segmentation technique.

b) Image Segmentation using DPSO

In search of a better model of natural selection using the PSO algorithm, the Darwinian Particle Swarm Optimization (DPSO) was formulated [16]. Here, many swarms of test solutions may exist at any time. DPSO is an extension of PSO algorithm. The concept of natural selection (Darwinian principle of survival of the fittest) is used to enhance the ability of PSO algorithm to escape from local optima. Many simultaneous PSO algorithms are run on groups of swarms in the same image. While running multiple swarms on the same image, a simple selection mechanism is applied. Each swarm individually performs just like an ordinary PSO algorithm with some rules governing the collection of swarms that are designed to simulate natural selection.

The traditional PSO-based segmentation is compared with the DPSO-based segmentation method to determine the *n*-1 optimal *n*-level thresholds on a given image. In DPSO, when a search tends to a local optimum, the search in that area is simply discarded and another area is explored. Here, at each step, swarms that get better are rewarded (extend particle life or spawn a new descendent) and swarms that stagnate are punished (reduce swarm life or delete particles). To analyze the general state of each swarm, the fitness of all particles is evaluated and the neighborhood and individual best positions of each of the particles are updated. If a new global solution is found, a new particle is spawned. A particle is deleted if the swarm fails to find a fitter state in a defined number of steps. Remove particles, and spawn a new swarm and new particle:

[1] When the swarm population falls below minimum bound, and

[2] The maximum threshold number of steps (search counter SC_C^{max}), without improving the fitness function, is reached.

After the deletion of the particle, instead of being set to zero, the counter is reset to a value approaching the threshold number, according to:

$$SC_C(N_{kill}) = SC_C^{max}\left[1 - \frac{1}{N_{kill}+1}\right] \qquad (11)$$

where N_{kill} is the number of particles deleted from the swarm over a period in which there was no improvement in fitness. To spawn a new swarm, a swarm must not have any particle ever deleted, and the maximum number of swarms must not be exceeded. Still, the new swarm is only created with a probability of $p = \frac{f}{NS}$ with f a random number in [0, 1] and NS the number of swarms. This factor avoids the creation of newer swarm S when there are large numbers of swarms in existence. The parent swarm is unaffected, and half of the parent's particles are selected at random for the child swarm and half of the particles of a random member of the swarm collection are also selected. If the swarm initial population number is not obtained, the rest of the particles are randomly initialized and added to the new swarm. A particle is spawned whenever a swarm achieves a new global best, and the maximum defined population of a swarm has not been reached. Like the PSO, a few parameters also need to be adjusted to run the algorithm efficiently:

- Initial swarm population.
- Maximum and minimum swarm population.
- Initial number of swarms
- Maximum and minimum number of swarms
- Stagnancy threshold

The basic assumptions made to implement Darwinian PSO are:

- The longer a swarm lives, the more chance it has of possessing offspring. This is achieved by giving each swarm a constant, small chance of spawning a new swarm.
- A swarm will have its lifetime extended (be rewarded) by finding a more healthy state.
- A swarm will have its life reduced for failing to find a more fit state.

DPSO algorithm is indicated in Table II. The results obtained after DPSO segmentation of Tsukuba image are shown in Fig.10.

Table II. DPSO ALGORITHM

Main Program Loop	**Evolve Swarm Algorithm**
For each swarm in the collection	1. For each particle 'n' in the swarm 'S'.
1. Evolve Swarm algorithm	2. Update Particle's objective function
2. For each swarm in the collection Allow the swarm to spawn a new swarm	3. Update Particle Bests
	4. Move Particle.
3. Delete "failed" swarms.	5. If swarm S gets better Reward swarm
	6. If swarm S has not improved Punish swarm

c) Image Segmentation using FO-DPSO

After application of PSO and DPSO algorithms to number of images, it has been observed that PSO is fast but not efficient (for finding the global optimum) and DPSO is efficient (for finding the global optimum) but speed of algorithm is less. It has been recently proved for benchmarking optimization problems that, the FO-DPSO is faster than the PSO (the most well-known optimization algorithm in terms of speed) and more efficient than the

DPSO (in order to find the global optimum while avoiding local optima) [17]. Therefore, FO-DPSO algorithm was selected as the next algorithm for segmentation of images achieving both important goals at once. More specifically, due to its convergence speed, this optimization method is a primary solution to a segmentation of high-resolution images.

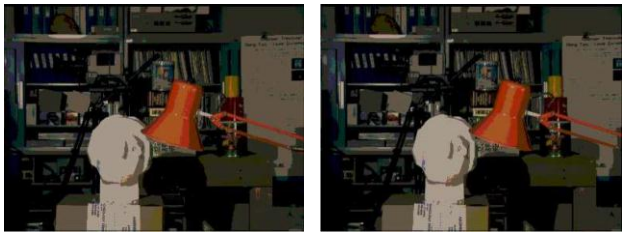


Figure 10. Segmented left and right views of Tsukuba using DPSO segmentation technique.

In Fractional Order Darwinian's Particle Swarm Optimization, several swarms compete using Darwin's survival-of-the-fittest principles and use fractional calculus to control the convergence rate of the algorithm. Using those principles, the FO-DPSO enhances the ability of the PSO algorithm to escape from local optima by running several simultaneous parallel PSO algorithms, each being a different swarm on the same test image, and applies a simple selection mechanism. When a search tends to a local optimum, the search in that area is just discarded, and another area is examined instead. As in PSO and DPSO discussed above, at each step, swarms that show improvement are rewarded (extend particle life or spawn a new descendent), and swarms that stagnate are punished (reduce swarm life or delete particles). The approximate *Grünwald–Letnikov FC* [17] definition allows using the concept of the fractional differential with (alpha) α, $0 \leq \alpha \leq 1$, to control the convergence rate of particles. Each particle a within each different swarm S moves in a multidimensional space according to position. $(x_a[t]), 0 \leq x_a[t] \leq L-1,$ and velocity $(v_a[t])$. The position and velocity values are highly dependent on the local best $\tilde{x}_a[t])$ and global best $(\tilde{g}_a[t])$ information. The coefficients w, ρ_1, and ρ_2 are assigned weights, which control the inertial influence, i.e., according to "the globally best" and "the locally best," respectively, when the new velocity is determined. Typically, the inertial influence is set to a value slightly less than 1. ρ_1 and ρ_2 are constant integer values, which represent "cognitive" and "social" components. Tuning these parameters properly will lead to better results. The parameters r_1 and r_2 are random vectors, with each component a uniform random number between 0 and 1. The intent is to multiply a new random component per velocity dimension, rather than to multiply the same component with the velocity dimension of each particle. The value greatly affects the inertial particles. With a small α, particles ignore their previous activities, thus ignoring the system dynamics and being susceptible to get stuck in local solutions (i.e., exploitation behavior). With a large alpha, particles will present a more diversified behavior, which allows exploration of new solutions and improves the long-term performance (i.e., exploration behavior). If the exploration level is too high, then the algorithm may take too much time to find the global solution. Based on the experimental results from [17], a fractional coefficient of $\alpha = 0.6$ is used, thus resulting in a balance between exploitation and exploration. Segmented Tsukuba image using FO-DPSO technique is shown in Fig. 11.

Memory complexity of the FO-DPSO is larger than the PSO and DPSO since it intrinsically has memory properties related to the fractional extension. Due to the truncation order of the approximate fractional derivative, it needs to track the last four steps of each particle's velocity that depends on the number of components C (i.e., R, G, and B) of the image. The computational complexity of the algorithms was considered, excluding the first computation of (7) and (8). This may be accomplished because the three algorithms require the same initial computation that depends on the size of the image. After that initial setup, the three algorithms may be adjusted in such a way to ensure a similar computational complexity. Likewise, the computational complexity of the three algorithms will increase with the number of desired thresholds n. The PSO computational complexity depends on the number of particles $\mathrm{N^P}$ within the population, the DPSO and FO-DPSO computational complexity depends on the accumulated number of particles within each swarm, i.e., $\forall$s $\mathrm{N^S}$. The Computational complexity of both DPSO and FO-DPSO will be inferior to the PSO by defining the maximum number of particles within each swarm as $\mathrm{N_{max}} = \frac{\mathrm{N^P}}{\mathrm{N_{S_{max}}}}$, wherein $\mathrm{N_{S_{max}}}$ represents the maximum number of allowed swarms.

Figure 11. Segmented left and right views of Tsukuba using FO- DPSO segmentation technique.

It has been observed after application of swarm based segmentation algorithms that these algorithms are robust algorithms. Once the initial fine tuning of the parameters is carried out for the particular application, the results are consistent. The intention of carrying out segmentation was to reduce the size of the stereo image that further reduces storage requirement for the 3-D generation.

The first level of optimization is achieved here, which stores stereo images in the compressed form. The compression will be useful for storing images in embedded prototype for the 3-D generation. Fig. 12 shows compression achieved due to various segmentation techniques for three images in the Middlebury dataset. The segmentation algorithms are tested for 50 different stereo images including actual camera images and it has been observed that 95% confidence interval for compression, for Mean shift, K-means, PSO, DPSO, FO-DPSO algorithms lies between (78.58, 86.01), (51.48, 61.54), (79.59, 86.22), (79.70, 86.22), (79.49, 86), respectively.

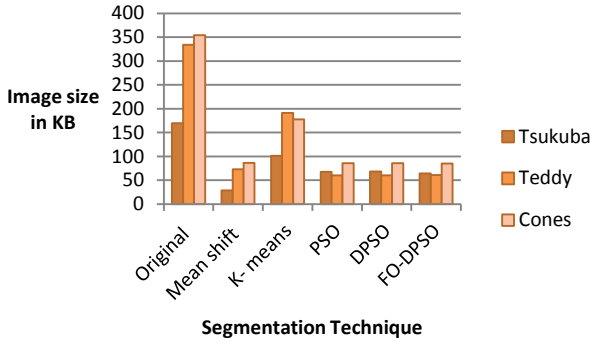


Figure 12. Reduction in size of three images from Middlebury dataset.

IV. STEREO MATCHING

The disparity refers to the difference in image location of an object seen by the left and right eyes, resulting from the eyes' horizontal separation (parallax). The brain uses the disparity to extract depth information from retinal images in stereopsis. In computer vision disparity refers to the difference in horizontal coordinates of similar features within two stereo images. Considering a single pixel in left image, to compute its correspondence in the right image a variety of search techniques can be used to match pixels based on their local appearance as well as the motions of neighboring pixels. In the case of stereo matching, some additional information is available, namely the positions and calibration data for the cameras that took the pictures of the same static scene. This information can be utilized to reduce the number of potential correspondences, and hence speed up the matching and increase the reliability of matching.

Stereo matching algorithms perform the following four steps:

1) Matching cost computation by applying global cost function
2) Cost (support) aggregation viz. Instead of comparing single pixels, compare small window areas
3) Disparity computation and optimization
4) Disparity refinement

In this paper, the disparity is computed using Winner Take All (WTA) algorithm, which uses step 1, 2, and 3. Also, the disparity is computed using a second algorithm, i.e., Line Growing algorithm that uses all the four steps mentioned above and it treats disparity as energy minimization function. The first algorithm comes in the class of local algorithms because it calculates disparities using a window centered on each pixel. The second algorithm comes in the class of global algorithms involving global optimization. The first component of any dense stereo matching algorithm is a similarity measure that compares pixel values in left and right views to determine how likely they are to be in correspondence. The most common pixel-based matching costs include sums of squared intensity differences (SSD) and absolute intensity differences (SAD). We have used SAD as measure for Winner Take All algorithm, and SSD for line growing algorithm. Sum of Absolute Differences (SAD) is one of the simplest of the similarity measures, which is calculated by subtracting pixels within a square neighborhood between the left or reference image I_L and the right or target image I_R followed by the aggregation of absolute differences within the square window, and optimization with the Winner Take All (WTA) strategy [21]. If the left and right images exactly match, the resultant will be zero. Disparity for each point is computed by finding the cost of matching point $I_L(x, y)$ in the left image to point $I_R(x + d, y)$ in the right image using Sum of Absolute Differences (SAD). It is described by the following equation:

$$\mathrm{SAD} = \sum_{(x,y)\in w} |\mathrm{I_L}(x, y) - \mathrm{I_R}((x + d), y)| \qquad (12)$$

A. Segmentation Based Stereo Matching

The disparity is computed with two techniques. First is the Winner Take All algorithm that is a local algorithm finding out disparity with SAD. The second algorithm the line growing algorithm is a global algorithm that employs Sum of Squared Differences (SSD) and carries out filtering of the disparity map. The second algorithm results in higher computation time due to SSD cost function and filtering steps.

1) Simple Winner Take All Algorithm

The disparity is computed using SAD cost function as shown in Fig. 13. As a result, we get three sets of disparity cost. Optimization of these three sets is done by using Winner Take All method. This method inspects the cost associated with each disparity set via window centered on each pixel. The disparity with smallest aggregated cost is selected and given as estimated disparity map.

Disparity estimation was done for different sets of image pairs as follows:

1. Left and right original images
2. Left and right segmented images using Mean shift algorithm, K-means, PSO, DPSO, FO-DPSO.

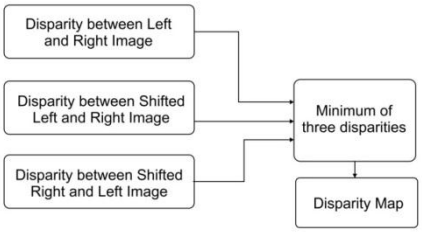


Figure 13. Process of disparity estimation.

The disparity maps obtained using various segmentation techniques are shown in Fig. 14. Fig. 14 (a) shows disparity map obtained using stereo matching of original Tsukuba image. Fig. 14(b), 14(c), 14(d), 14(e), and 14(f) show disparity map obtained using Mean shift, K-means, PSO, DPSO, FO-DPSO algorithms applied to original images and after application of WTA, respectively. The above algorithms were tested using a large number of epipolar rectified test image pairs. From the results obtained for each of them, it was observed that the algorithm gives good results for each type of image pairs and every kind of segmentation. Tsukuba image pairs contain texture-less areas such as the table and the lampshade. It also contains thin structures such as the rods of the lamp. Tsukuba image contains many objects of different size at different depths. Disparity map output is obtained in less than 1 second for five segmentation techniques.

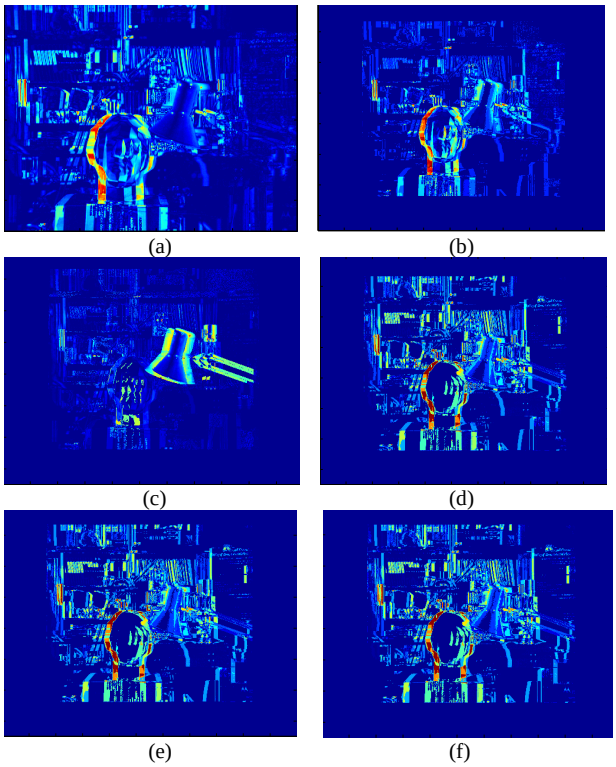


Figure 14. Disparity map using WTA and various segmentation techniques (a) Original (b) Mean shift (c) K-means (d) PSO (e) DPSO (f) FO-DPSO.

a) Advantages of Winner Take All Algorithm

- Fast Implementation and can most easily be optimized.
- Algorithm being simple can be easily implemented in a microcontroller.

b) Disadvantages of Winner Take All Algorithm

- Heavily dependent on stereo constraints.
- Identical pixels of cost matrix are assigned to reference pixel more than once.

2) Line Growing Algorithm

The area-based approach presented above falls into the category of "local methods" since the disparity computation is done for every single pixel. Another class of methods, which improve potential correspondences, are the global and semi-global methods. In these approaches, the task of computing disparities is treated as an energy minimization problem. Typically, energy function is formulated such as [21]:

$$E(d) = E_d(d) + \lambda E_S(d) \quad (13)$$

$\boldsymbol{E_d}$ **(Data Term)**: measures the pixel similarity, i.e., how well the disparity function d agrees with the input image pair.

$\boldsymbol{E_S}$ **(Smoothness):** penalizes disparity variations, i.e., how well does disparity match that of neighbors – regularization. The goal, in this case, is to minimize an objective function that includes some terms that model the costs associated with matching pixels at various disparities and others that seek to reward overall 'smoothness'.

Global stereo matching methods perform some optimization or iteration steps after the disparity computation phase and often skip the aggregation step altogether because the global smoothness constraints perform a similar function. Many global methods are formulated in an energy minimization framework, where the objective is to find a solution d that minimizes a global energy; this energy can be defined as

$$E_d(d) = \sum_{(x,y)} C(x, y, d(x, y)) \quad (14)$$

where C is the (initial or aggregated) matching cost of disparity. The smoothness term $E_S(d)$ encodes the smoothness assumptions made by the algorithm. To make the optimization computationally tractable, the smoothness term is often restricted to measuring only the differences between neighboring pixels' disparities. In the line growing algorithm, the idea is to minimize an objective function that includes some terms that model the costs associated with matching pixels at various disparities, and some terms may be added called 'smoothness 'term.

The resulting approach has some useful features. Firstly, it allows us to handle problems, such as stereo,

where the variable values are continuous without requiring any intermediate quantization. Secondly, penalty terms can be incorporated involving more complex functions of the disparity values.

a) Minimizing the Error Energy

Fig. 15 shows the global energy minimization technique. In this method, the block-matching technique is used to construct an Error Energy matrix for every disparity. L(i, j, c) denotes segmented left image in RGB format and R(i, j, c) denotes segmented right image in RGB format and e(i, j, d) denotes error energy .

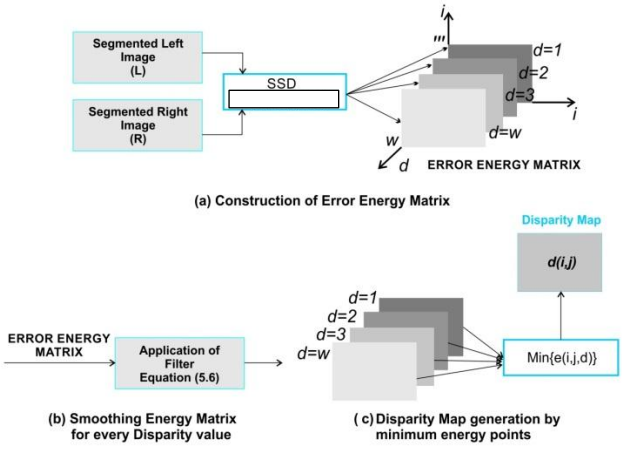


Figure 15. Method using global error energy minimization by smoothing functions.

For $n \times m$ window size of block matching, error energy e(i, j, d) can be expressed by,

$$e(i,j,d) = \frac{1}{3 \cdot n \cdot m} \sum_{x=i}^{i+n} \sum_{y=j}^{j+m} \sum_{k=1}^{3} (L(x, y+d, k) - R(x,y,k))^2 \quad (15)$$

where C represents RGB components of images and takes a value of {1, 2, 3} corresponding to red, blue and green, and 'd' is the disparity. For a predetermined disparity search range (w), every e(i, j, d) matrix related to the disparity is smoothed by applying averaging filter many times. Averaging filter (linear filter) removes the very sharp change in energy that belongs to incorrect matching. Another important property of repeating application of the averaging filter is that it makes apparent global trends in error energy. Considering the global trend in error energy makes this algorithm a region -based algorithm. For $n \times m$ window size, average filtering of e (i, j, d) can be expressed by the following equation,

$$\tilde{e}(i,j,d) = \frac{1}{n \cdot m} \cdot \sum_{x=i}^{i+n} \sum_{y=j}^{j+m} e(x,y,d) \quad (16)$$

After iterative application of averaging filter to error energy for each disparity, the disparity 'd' is selected, which has minimum error energy $\tilde{e}(i, j, d)$ as the most reliable disparity estimation for pixel (i, j) of disparity map. The necessary steps in the algorithm shown in Fig. 15 are

Step 1: For every disparity 'd' in disparity search range, calculate error energy matrix. Refer Fig.15 (a).
Step 2: Apply averaging filter iteratively to every error matrix calculated for a disparity value in the range of disparity search range. Refer Fig.15 (b).
Step 3: For every (i, j) pixel, find the minimum error energy $\tilde{e}(i, j, d)$, assign its disparity index 'd' to d (i, j) which is called disparity map. Refer Fig.15 (c).

b) Region Growing

The region growing is carried out in the direction of rows in the image since the disparity of stereo image is only in row directions. So, only one neighbor, which is the point after searched point, is inspected for region growing and hence algorithm is named as line growing as shown in Fig. 16.

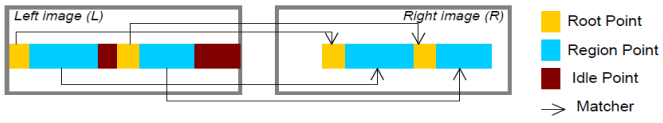


Figure 16. Method using Line Growing.

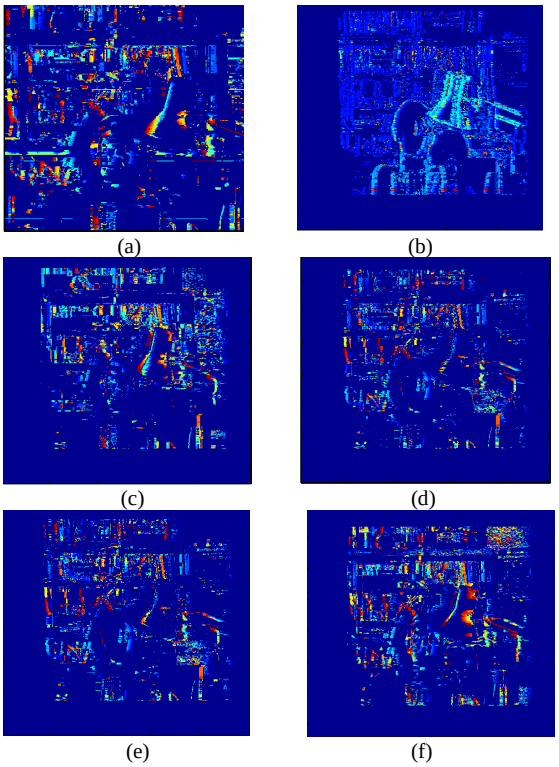


Figure 17. Disparity space images using Line Growing (a) Original (b) Mean shift (c) K-means (d) PSO (e) DPSO (f) FO-DPSO.

The steps in line growing are:

1. Root Selection process**:** Select a point on the row and find its disparity using energy function equation (16). If error energy of selected point is not less than or equal to line growing threshold, mark this point as idle. If error energy is less than a threshold, then mark the point as root point and go to step 2.

2. Line growing process**:** Calculate error energy of neighboring point using root point disparity, which was called region disparity. If it is lower than the predetermined error energy threshold, associate this point to the region.

3. Proceed for the steps 1 and 2 row by row until the end of the image. When all points in the image are processed, an algorithm is stopped. Grown disparity regions compose the disparity map d(i, j).

Disparity Space Images obtained using line growing algorithm employing various segmentations, i.e., Mean shift, K-Means, PSO, DPSO, FO-DPSO are shown in Fig. 17(b), 17(c), 17(d), 17(e), 17(f), respectively. Fig. 17(a) shows disparity map obtained using original Tsukuba image using Line growing algorithm.

B. Depth Estimation

Once the disparity values are calculated, the next step in stereo algorithms is finding out depth from disparity. Depth estimation is an important tool in several applications such as machine vision, robotics, and satellite terrain mapping. With recent advances in 3D consumer video communications technology, use of depth estimation is likely to grow significantly in near future. One of the objectives of this work is the depth estimation. Depth estimation using laser or infrared ranging techniques are precise and familiar. However, their applications are limited to certain tasks. For example, it is not advisable to laser scan a live human. Stereoscopic methods are purely passive and use a pair of cameras (left and right) to map a scene. The disparity is used to estimate the depth of different parts of a scène. Disparity estimation gives good results for finding short distances. The difference of each pixel position is calculated through one of the stereo matching algorithms explained above using segmented image as input. Using stereo camera parameters from the calibration and the disparity between corresponding stereo points, depths in the stereo images can be retrieved. The maximum range at which the stereo vision can be used for detecting obstacles depends on the image and depth resolution. Absolute differences of pixel intensities are used in the algorithm to compute stereo similarities between points. Using eqn. (17) below, depth for each pixel position is calculated.

$$Z = b\frac{f}{d} \tag{17}$$

There are only three parameters required to find depth or distance from disparity. The location of photoreceptors of the camera is called image plane. The focal length is the distance between photoreceptor and lens that is specified in the camera data sheet as ‘f’. Baseline width ‘b’ is the separation between stereo cameras and ‘d’ is the disparity of each pixel. Measurement of X and Y locations in the real view are carried out with the help of yard stick or measuring tape, and it is compared with the (x, y) pixel position on the camera. It is a mapping of a physical quantity in cm or meter to pixel scale. Due to this mapping all cm values are converted into pixel values, and the 3-D world coordinates of points corresponding to each pixel can be constructed from the disparity map. A disparity map or “depth map” image is an efficient method for storing the depth of each pixel in an image. Each pixel in the map corresponds to the same pixel in an image, but the grey level corresponds to the depth at that point rather than the grey shade or color. Disparity map construction can be summarized as follows:

- Find every corresponding point between the images.
- Assign a value 0 to 255 to each point based on the “disparity” calculation.
- Calculate depth.

To evaluate the method, four standard stereo image pairs were used: Cones, Teddy, Tsukuba and Venus. These RGB stereo image pairs are provided by the Middlebury database, processed with various algorithms discussed above. The data set images have different sizes and different values of maximum (d_{max}) and minimum (d_{min}) disparity. Tsukuba (384×288 pixels) is the smallest image pair and Teddy and Cones are the largest, both with pixels of size 450×375. Venus has a size of 434×383 pixels. This causes variations in the processing time since each pixel must be processed during the disparity estimation. This processing time remains the same due to the application of segmentation algorithms to these images but maintains the number of depth levels obtained.

The maximum disparity between the left and right image also affects the processing time. Tsukuba has the least disparity variation, only sixteen values from 0 to 15. The Venus disparities range from 0 to 19 while the Disparity ranges for Cones and Teddy are from 0 to 59. Actual depth calculations are explained in Section VII.

Fig. 18 shows depth maps obtained using various segmentation techniques and Winner Take All algorithm. A depth map of original Tsukuba stereo image pair is presented in Fig. 18(a). The depth map of original image pair of Tsukuba image gives eight depth levels after application of Winner Take All algorithm. Depth maps obtained after segmentation and Winner Take All algorithm

are shown in Fig. 18(b), 18(c), 18(d), 18(e), 18(f), respectively.

Depth levels obtained using Mean shift, K-means, PSO, DPSO, FO-DPSO algorithms are comparable with depth map of original image pair.

Depth map obtained after application of Line Growing algorithm on original images of Tsukuba is shown in Fig. 19(a). Depth maps obtained after application of Line Growing algorithm on segmented images of Tsukuba are shown in Fig. 19(b), 19(c), 19(d), 19(e), 19(f), respectively.

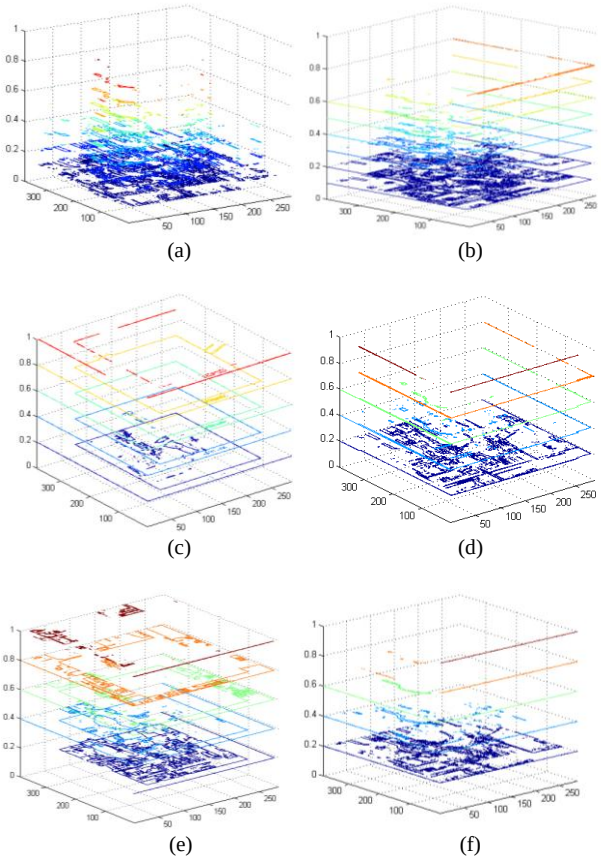


Figure 18. Depth Map using various segmentation techniques and WTA for Tsukuba (a) Original (b) Mean Shift (c) K-means (d) PSO (e) DPSO (f) FO-DPSO.

C. Reconstructed 3-D View after Segmentation

One of the significant results obtained from present work were 3-D views. By concatenation of original left and right stereo images 3-D view obtained is as shown in Fig. 20(a). There is tiny degradation in quality of the 3-D image obtained by segmentation and concatenation of segmented stereo pair.

Compressed 3-D images were generated using all the segmentation techniques described in Section III are shown in Fig. 20(b), 20(c), 20(d), 20(e), 20(f). A reconstructed 3-D using PSO variant is at par with the original 3-D image. The 3-D image in Fig. 20(f) is in much compressed form as compared to Fig. 20(a) but visual quality is not degraded. Compression achieved makes it suitable for storing it on systems with memory size constraints. The better option for portable application development was the implementation of above-mentioned stereo algorithms on embedded processor. Section VI describes the implementation of Winner Take All algorithm on a portable hardware, i.e., microprocessor of ARM 9 architecture. The reason for selecting this microprocessor was the popularity of this design when this project work was started and was available off –the-shelf.

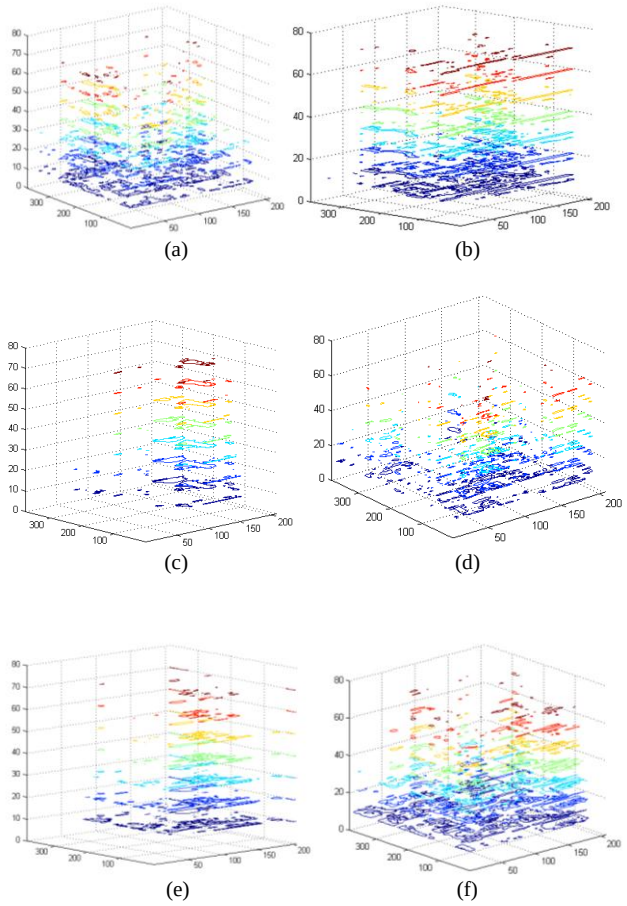


Figure 19. Depth Map using various segmentation techniques and LG for Tsukuba (a) Original (b) Mean Shift (c) K-means (d) PSO (e) DPSO (f) FO-DPSO.

V. ZIGBEE

The ZIGBEE Alliance [22] is a consortium of over 90 companies that is developing a wireless network standard for commercial and residential control and automation applications. Transmission of images by using Bluetooth network had been tried, but Bluetooth-based networks can cover the distance up to 10m while ZIGBEE based networks can be used up to 100m. Bluetooth takes three seconds to join a network while ZIGBEE joins a network in 30 milliseconds [22]. The main reason behind

selecting IEEE 802.15.4 over IEEE 802.11 is the low power consumption since the prototype developed is embedded product with limited batteries.

The Alliance has recently released its specifications for a low data rate on the wireless network. The design goals for the network have been driven by the need for a Machine-to-Machine (M2M) communication of small simple control packet and sensor data, and a desire to keep the cost of wireless transceivers to a minimum. ZIGBEE is a wireless technology developed as an open global standard to address the unique needs of low-cost, low-power wireless M2M networks, and it currently uses IEEE 802.15.4 MAC and PHY layers, as shown in Fig. 21 [23].

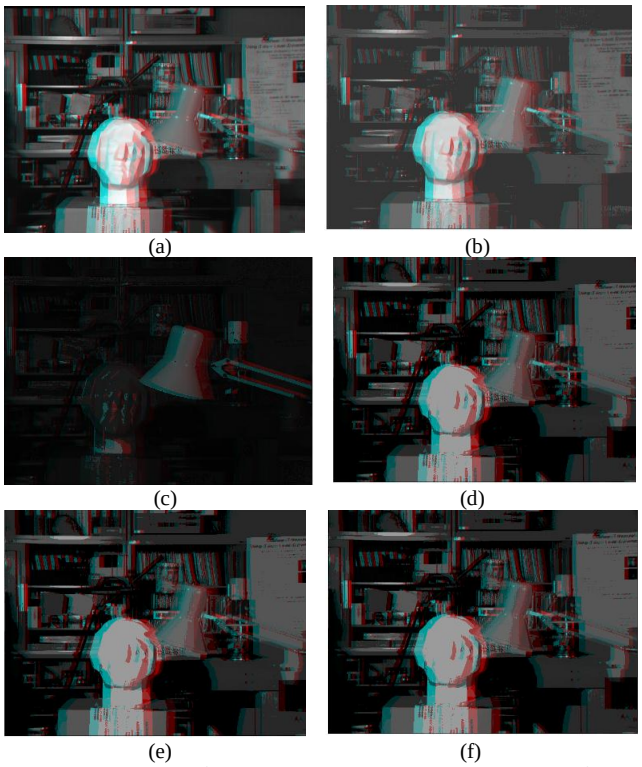


Figure 20. 3-D views after segmentation (a) Original (b) Mean shift (c) K-means (d) PSO (e) DPSO (f) FO-DPSO.

ZIGBEE uses a single channel for data transmission. A ZIGBEE module has three nodes, namely, coordinator node, a router node, and an end device node. End-device nodes communicate with each other through a coordinator node. A coordinator node handles starting the network and for choosing certain critical network parameters. The end-device nodes not only communicate with the coordinator node but also communicate with every router node. However, the router nodes processing a routing function cannot directly communicate with each other; they can communicate only with coordinator [23]. ZIGBEE network has three modes of transmission, namely, AT (by default), API and API with an escape character. In the AT (Transparent Mode), data coming into the Data IN (DIN) pin is directly transmitted over-the-air to the intended receiving radios without any modification. API (Application Programming Interface) mode is a frame-based method for sending and receiving data to and from a serial UART (Universal asynchronous receiver/transmitter). API with escape character is an extended version of API, which is used to prevent data loss in noisy environments. Both API and API with escape character are used to ensure secure communication. In this setup, AT (Transparent Mode) mode of transmission has been used as it is easy to configure ZIGBEE in this way and currently secure communication is not considered in the present prototype.

TABLE III. HARDWARE SPECIFICATIONS

ZIGBEE module
- Operating frequency: 2.4GHz.
- Low cost wireless module.
- Data rate: 250Kbps.
- Operating range: 100ft (30m).

Wireless camera
- Connection Type – Corded USB.
- USB Type –High Speed USB 2.0.

A. ZIGBEE Protocol

ZIGBEE is best described by referring to the 7-layers of the OSI model for layered communication systems. The Alliance specifies the bottom three layers (Physical, Data Link, and Network), as well as Application Programming Interface (API) that allows end developers the ability to design custom applications that uses the services provided by the lower layers. Fig. 21 shows the architecture adopted by the ZIGBEE alliance [23].

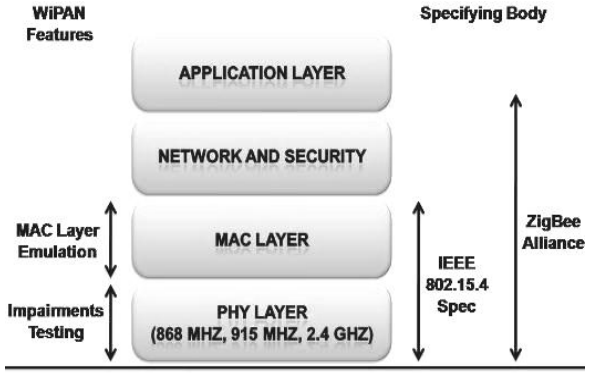


Figure 21. ZIGBEE stack [22].

B. Limitations of ZIGBEE protocol

The 2.4GHz band provides the highest bit rate of 50 Kbps in IEEE 802.15.4 PHY specification. The physical layer supports the transfer of only small sized packets,

which is limited to 127 bytes. Due to overhead at the network, each packet may contain at most 89 bytes for application data. This leads to loss of data during transmission. Therefore, there is a need for fragmentation of bit streams larger than 89 bytes. A flow-control mechanism is also needed to acknowledge and request retransmission of missing fragments above the network layer [22].

C. Transmission of image through ZIGBEE

If a large number of pixel values of an image are transmitted by using ZIGBEE, then there is a loss of data in an abrupt manner at the receiving end. For this, the data needs to be fragmented. In this case, an image of size 115 X 132 was transmitted using ZIGBEE. An image of size 115 X 132 has 15180-pixel values. The image is fragmented into small packets, and each packet contains approximately 2000 pixel values. For a complete transmission of the image, eight packages are required. Since each packet is transmitted separately, there is an increase in time taken for transmission of the entire image.

D. Control of robot by using depth information

The depth levels estimated from disparity data are transmitted through ZIGBEE module. The depth levels received by the receiver connected to hardware are used to control the robot.

VI. Hardware Implementation of Stereo Matching

Stereo vision algorithms require a very large number of computations and therefore, currently they are not widely used in portable systems. There is still a requirement of adequate hardware and support for the development of software for such systems. Realizing the importance of equipment that generates the 3-D image and gives object depth, prototype development was carried out. This prototype may work as a basic foundation for modern computer vision applications.

Winner Take All algorithm described in Section IV was implemented on ARM 9 microprocessor from ATMEL. The algorithm was optimized to suit lower processing power, using lower resolution images for better output performance. Also, 3-D image was generated on TFT display using concatenation of two images received at the receiver. More general programming platform like embedded C was used so as to satisfy any soft real-time system. There are no catastrophic consequences of missing deadlines in soft real-time system. Using a pair of stereo images, acquired through the camera or sent through USB port of PC hardware, system is able to provide a 3D image in real time, keeping the details of produced image acceptable to the human eye. Hardware also provides a disparity map that is a spatial representation of depths of various objects in the image on TFT display of hardware.

This hardware can be converted into prototype if it is to be used as an industrial product for the application like depth estimation. For verification of stereo matching algorithm on hardware microprocessor, SAM 9 from ATMEL was selected. Where SAM stands for “Smart Atmel Microprocessor” with ARM-9 architecture. The complete evaluation kit based on this microprocessor SAM9M10-G45-EK was available from ATMEL [26].

The segmented images generated using techniques described in Section III were stored in compressed form in the memory of SAM9M10-G45 evaluation kit. The depth levels and 3-D images were generated by applying a stereo algorithm on the segmented images. The 3-D images were displayed on TFT display of hardware board. Obtaining 3-D views on hardware enables robust and practical solutions to problems that are difficult or impossible to solve with conventional 2-D vision. 3-D allows easier discrimination between background and objects. It can also enable more reliable and more precise gesture interfaces, and it helps systems understand where objects are located with other objects. The specifications and other details of the SAM9M10-G45 evaluation kit are given in manual from ATMEL [27].

A. Solution Methodology

Because of availability of camera interface, high memory and high speed this kit ideal for image processing applications. The programming of this kit can be done through Keil µVision IDE and requires code written using Embedded C. Fig. 22 shows photograph of SAM-9-M-10M-EK.

Figure 22. Photograph of the SAM9M10-G45 evaluation kit.

Atmel SAM-BA® software provides an open set of tools for programming the SAM9M10-G45 evaluation kit for ARM® core-based microcontroller. The SAM Boot Assistant (SAM-BA) has been used as the programmer for the kit. This software is available from Atmel to download programs in SAM9M IC on SAM9M10-G45 evaluation kit. SAM-BA software provides means of programming different Atmel devices. They are based on a standard dynamic linked library (DLL), the sam-ba.dll. SAM-BA uses the DLL to communicate with the SAM9M10-G45

evaluation kit. Different stereo images were stored in DDRAM at address locations 0x70100000 and 0x70200000 in .raw format. Winner Take All algorithm was implemented (explained in Section IV) to get disparity map. Also, the 3-D view was generated on TFT display of the evaluation kit.

The SAM9M10-G45-EK features LCD controller. Portrait Mode LCD of dimensions 4×3" with resolution 480 x 272 provides the SAM9M10- G-45 evaluation kit with a low power LCD, a backlight unit, and a touch panel, similar to that used on commercial PDAs. Graphics and text can be displayed on the dot matrix panel with up to 16 million colors by supplying 24-bit data signals (8bit × RGB by default). It is possible for the user to develop graphical user interfaces for a broad range of end applications.

B. Displaying Image on LCD of the Kit

Two images were stored in the DDRAM of the kit, and 3-D view and disparity map were displayed on the LCD. Steps in the processing are

1. Create a project for SAM9M10-G45 evaluation kit using Keil µ-Vision 4
2. Build the project to obtain the .bin file.
3. Use the SAM-BA interface as shown in Fig. 23 to connect the SAM9M10-G45 evaluation kit to the computer.
4. Send the .bin file to the DDRAM of SAM9M10-G45 evaluation kit.
5. Send images to be displayed in .raw format to locations specified in the program.
6. Execute the .bin file using command window of SAM-BA.

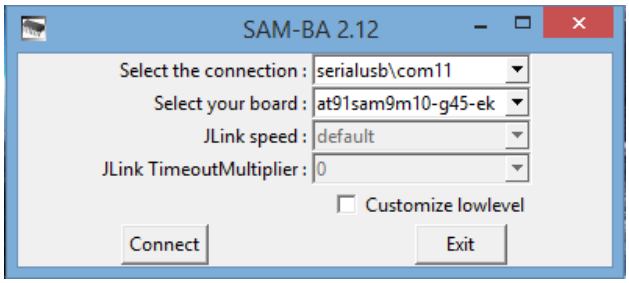


Figure 23. SAM-BA Interface.

C. 3-D reconstruction on SAM9M10-G45 Evaluation Kit

3D images obtained using MATLAB are shown in Fig. 20. The same function is implemented in an optimized way using Embedded C to obtain similar results on the SAM9M10-G45 evaluation kit. 3D images on the SAM9M10-G45 evaluation kit are shown in Fig. 24.

The segmented images were given to the disparity estimation algorithm to estimate the depth values and were transmitted through coordinator node of ZIGBEE module.

Segmented stereo images and depth values were received by router node. The image data and depth values received by the router node can be used for the further industrial application. Basic steps are shown in Fig. 25.

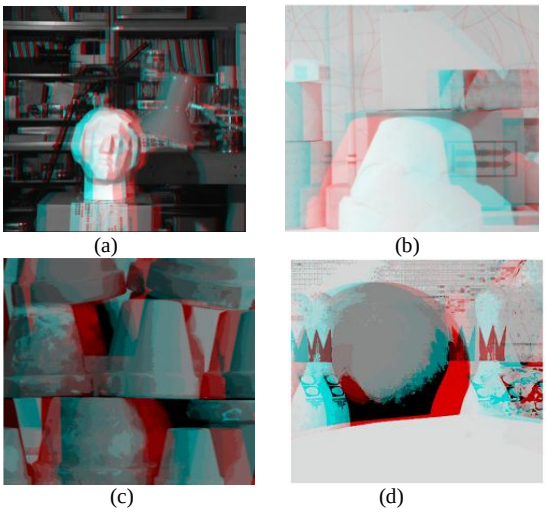


Figure 24. 3D views of various images on SAM -9 evaluation board.

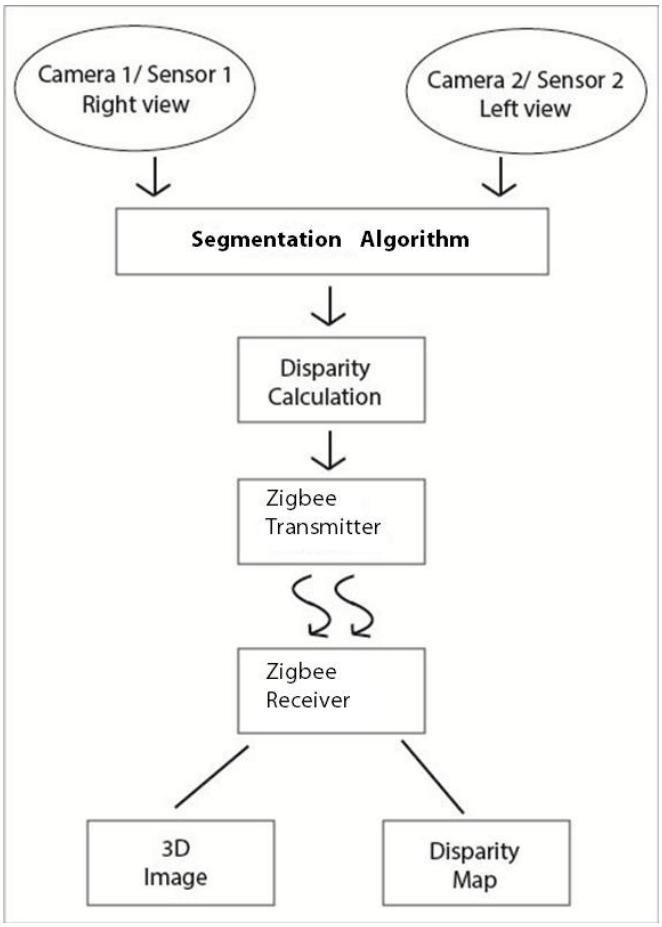


Figure 25. Steps in wireless transmission of stereo images and its disparity levels.

VII. RESULTS AND DISCUSSION

The results obtained after the implementation of different algorithms are presented in this section. The segmentation algorithms were tested on Middlebury database and were compared for the performance parameters like PSNR, compression ratio. 3-D images were generated using original left and right views as well as segmented left and right views and analyzed on the basis of subjective quality criterion. Comparison of stereo algorithms was carried out on the basis of a number of depth levels extracted. The number of depth levels extracted depends on the number of objects present in the image and also the stereo algorithm used. Fifteen different images (13 from Middlebury database) were used for the analysis purpose.

There are total five data sets provided on Middlebury website. These data set images provide rectified left and right view and ground truth images. This site also allows verifying the results. Revision of data set has been done in the years 2001, 2003, 2005, 2006, 2007, 2014. These images have different numbers of clusters with varying complexities; they consist of well-separated clusters, overlapping clusters or a combination of both. These images also contain different objects at different depths which make it easier to analyze the code written.

TABLE IV. PSNR VALUES IN DB WITH OPTIMUM PARAMETER VALUES SELECTED FOR EACH ALGORITHM

Image Name	Mean shift	K-means	PSO	DPSO	FO-DPSO
Tsukuba	20.47	10	14.42	15	17.08
Art	14.54	8	13	16	16
Books	16.81	6.31	13.99	14.39	13.04
Computer	15.37	7.15	15.99	15.13	17.12
Cones	16.26	8	16.26	14.5	15.47
Dolls	14	8.34	16	13	16
Drumstick	19.26	6.21	15.63	13	18
Dwarves	18	5	17.88	15.95	17.285
Laundry	14.89	5.746	16.83	16.72	14.77
Moebius	18.44	6.44	16.75	15.61	15.6
Reindeer	18.44	10.2	18.046	15.12	14.45
Teddy	18.22	7.41	18.6	17.51	11
Venus	17.11	10	17	16.5	17.11

It was observed that segmented images have very large MSE values about the original and hence low value of PSNR was obtained. These images show a negligible loss of perceived image quality. The PSNR values obtained are not so high, so that the visual quality of images after segmentation is still good for PSO variants. There is a loss of visual quality after K-means clustering algorithm that is also reflected in values of PSNR. The most reliable method for assessing the quality of images is through subjective testing since human observers are the ultimate users in most of the multimedia applications. According to human observers, the visual quality of PSO based techniques is good. Table IV shows PSNR values obtained for segmented images.

Fig. 26 shows the graph of time required for segmentation on intel i5 processor having 1.8GHz clock frequency. Segmented 3-D images were tested on 100 subjects for subjective analysis. Results of the individual analysis show that 3-D images constructed with the FO-DPSO technique were having better quality as compared to other techniques. Since the subjective quality of the 3D images obtained using PSO variant techniques is better, in future, it can be one of the best techniques of 3-D generation.

In subjective testing, a group of people were asked to give their opinion about the visual quality of 3-D each image. Subjective analysis of segmented 3-D images was carried out with 50 observers and results show that the FO-DPSO based segmentation technique gives good visual quality similar to original.

Hence, FO-DPSO can be considered as best segmentation technique because it not only gives good quality 3-D but takes less time for segmentation.

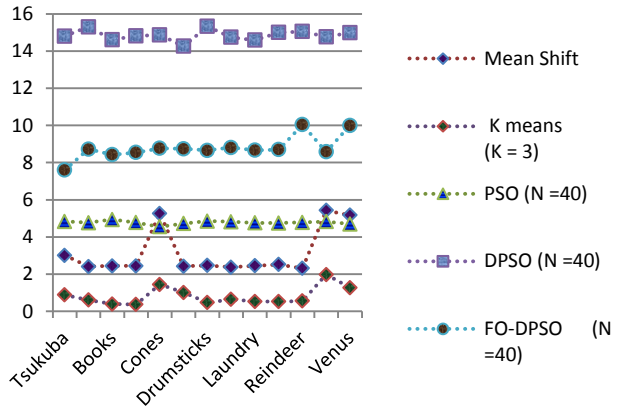


Figure 26. CPU time required for segmentation (in seconds) using various segmentation techniques.

Hence, it can be concluded that PSO algorithm retains the maximum original information of the image even after segmentation. PSO based segmented images provide better disparity estimation, with a good number of estimated depth levels.

A. Depth Levels Obtained

Table V shows the comparison of a number of estimated depth levels for different segmentation algorithms using Winner Take All stereo matching technique. A

number of depth levels determined using PSO segmentation are almost same to the number of depth levels estimated from an original image. This is the reason the subjective study for 3-D image reconstructed using PSO, DPSO, FO-DPSO segmentation provides better results compared to K-means and Mean shift segmentation techniques.

TABLE V. NUMBER OF DEPTH LEVELS OBTAINED USING WINNER TAKE ALL ALGORITHM

Image Name	Original	Mean shift	K-means	PSO	DPSO	FO-DPSO
Tsukuba	**7**	**8**	**5**	**5**	**5**	**5**
Art	**6**	**8**	**5**	**6**	**5**	**6**
Books	**7**	**6**	**5**	**7**	**5**	**7**
Cones	**6**	**6**	**4**	**6**	**5**	**5**
Dwarves	**5**	**4**	**6**	**5**	**5**	**5**
Laundry	**6**	**4**	**4**	**6**	**6**	**6**
Moebius	**5**	**3**	**6**	**5**	**5**	**5**
Reindeer	**6**	**5**	**7**	**5**	**5**	**5**
Teddy	**7**	**7**	**4**	**6**	**5**	**5**
Venus	**5**	**8**	**6**	**3**	**5**	**3**

Table VI shows the comparison of the number of estimated depth levels for different segmentation algorithms using Line growing stereo matching method. A number of depth levels determined using PSO segmentation are almost same to the number of depth levels estimated from the original image.

TABLE VI. NUMBER OF DEPTH LEVELS OBTAINED USING LINE GROWING ALGORITHM

Image Name	Original	Mean shift	K-means	PSO	DPSO	FO-DPSO
Tsukuba	7	7	7	7	7	4
Art	6	7	7	5	7	4
Books	7	7	3	5	7	5
Cones	6	7	4	7	7	5
Dwarves	5	7	7	5	7	5
Laundry	6	7	7	7	7	3
Moebius	5	5	7	6	6	3
Reindeer	6	4	7	7	7	5
Teddy	7	4	4	7	7	4
Venus	5	5	4	7	7	5

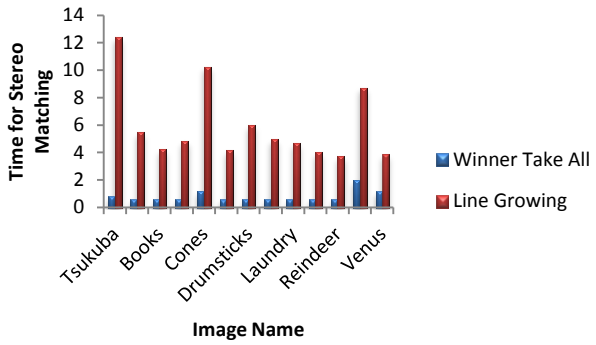


Figure 27. Time required for stereo matching in seconds using LG and WTA.

B. Stereo Matching Time

The time needed for stereo matching after application of stereo matching algorithms is shown in Fig. 27. It can be seen that the time required for stereo matching using line growing algorithm is high. This increase in stereo matching time is because of the additional filtering step that is carried out before finding disparity map in Line Growing algorithm.

C. Real view depth estimation

The arrangement shown in Fig. 1 was used for depth estimation of real time view. Initially, disparity of plane board was calculated after application of stereo algorithm as illustrated in Fig. 28. This figure also shows a color bar of disparity values indicating different colors for different disparities and disparity values varying between -1 to +1. Individual object disparities can be found using pseudo colors in the color bar if multiple objects are present in the view. Since plane board is not having any depth it is indicated by zero in the color bar.

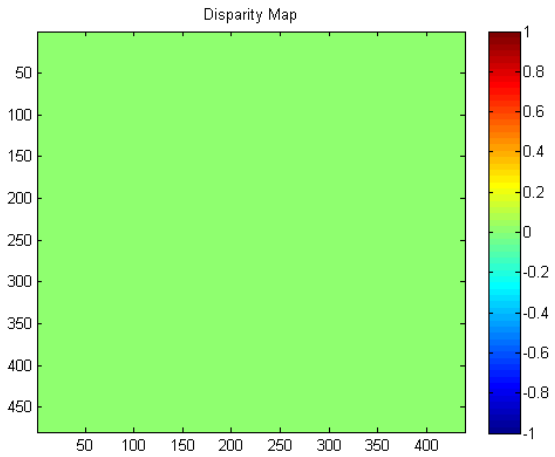
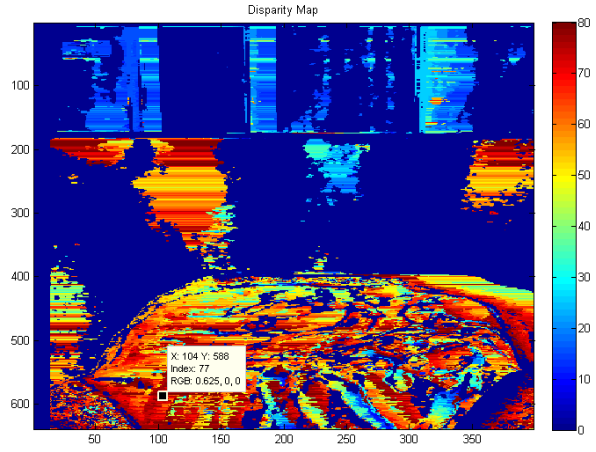


Figure 28. Disparity map of image having zero disparity.

(a)



(b)

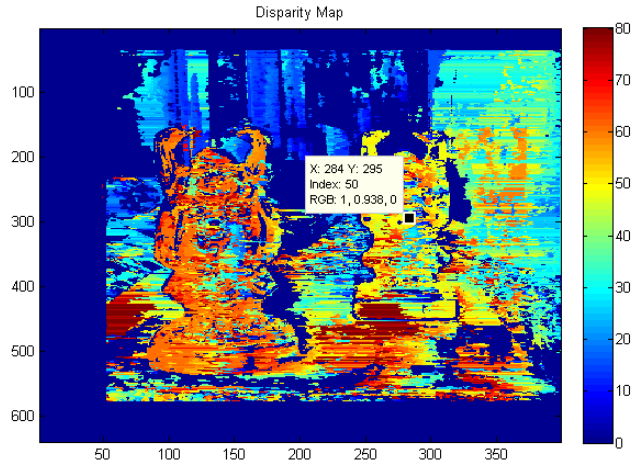
Figure 29. Absolute depth estimation (a) Left and Right Views of Actual Camera Image (b) Disparity map of the same image showing closer objects brighter. and distance obtained is approximately $43.02cm.$ ($Real\ distance\ 44 \pm 1cm$) (Baseline 5cm and focal length 17.47 cm).

There are two measures of depth, relative measure and absolute measure. Relative measure finds out if an object is farther or closer than another one. An absolute measure of depth finds out the distance between image pixels and camera. An absolute measure of depth, as well as relative measure of depth, is calculated in this work.

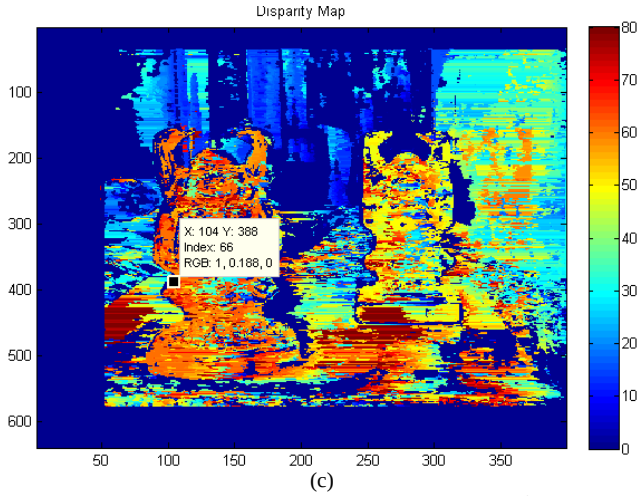
For measurement of depth, Fig. 29 (a) shows image pairs acquired through the camera with an object placed in front of the plane board. Fig. 29 (b) shows disparity map obtained for the same images, and it can be observed that the objects placed in front of the board are having higher disparity value than the background behind the board (shown in color shades of red and disparities in the range of 40 to 80). Far objects have the disparity in the range of 0 to 40.

Fig. 30 (a) shows left and right views of actual camera images that are used for finding the relative distance between two statues of Happy man placed 16 cm apart. Figs. 30 (b) and 30 (c) show disparity maps obtained for these images, and it can be observed that the disparity value obtained for Happy man in the front is 66, and that for the Happy man that is behind the first one is 50. Hence, the relative distance obtained between two statues is 15.97 cm.

(a)



(b)



(c)

Figure 30. Relative distance obtained between two statues of Happy man (a)Left and right views of actual camera images.(b) Disparity map showing disparity value of 50 for Happy man 2.(c) Disparity map showing disparity value of 50 for Happy man 1.

Since stereo algorithms discussed has a number of constraints there is variation in the accuracy achieved.

For the proposed work using the focal length of 17.47 cm and baseline of 5 cm and using Equation (17), different absolute and relative measures of range values approximately matching with the real distance were obtained.

D. Image Transmission

The testing of the present setup was done on several images from Middlebury data set [13]. One of the image pair, which was transmitted using ZIGBEE and received at the receiver ZIGBEE module, is shown in Fig. 31 and Fig. 32. Reconstructed 3-D image is shown in Fig. 33.

Figure 31. Left and Right view of images transmitted.

Figure 32. Left and Right view of images received.

Figure 33. Reconstructed 3-D image at the receiver.

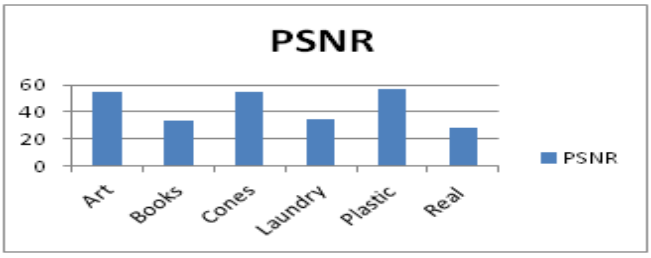


Figure 34. PSNR values obtained for images received at ZIGBEE receiver.

Six different images from Middlebury data set were transmitted and received at the receiver. The Peak Signal-to-Noise ratio (PSNR) values of received images in dB were plotted and are shown in Fig. 34.

VIII. Conclusion And Future work

A 3-D image was generated at the receiver end. It was observed that there is always a compromise between PSNR and time taken to transmit the image. The time taken for transmitting an image can be reduced by implementing a mesh or star topologies using a set of ZIGBEE modules, which may give rise to loss of data. Before implementing on real time, the above algorithms were tested for various data types such as .jpg, .png and results were found satisfactory for all types of images. In the future, the above segmentation algorithms like PSO, DPSO, FO-DPSO can be implemented on advanced DSP processor such as, Blackfin processor from Analog Devices. Also, CMOS cameras like OV 2640 can be interfaced with processor giving real time depth maps and also controlling robot movement from depth estimated.

References

[1] A. Naik, A. Khaparde, K. Velhal, and K. Shah, “Wireless Transmission of Stereo Images and Its Disparity levels,” in Proc. IMMM, 2014, The Fourth International Conference on Advances in Information Mining and Management, Paris, France, July 20- 24, pp. 41 – 44, ISBN**:** 978-1-61208-364-3

[2] http://www.vision.caltech.edu/bouguetj/calib_doc/ [retrieved: October 14, 2015]

[3] A. Khaparde, A. Naik, M. Deshpande, S. Khar, K. Pandhari, and M. Shewale, "Performance Analysis of Stereo Matching Using Segmentation Based Disparity Map," in Proc. ICDT, 2013, The Eighth International Conference on Digital Telecommunications, Venice, Italy, April 21-26, pp. 38-43.

[4] D. Comaniciu and P. Meer, “Mean shift:A Robust Approach Toward Feature Space Analysis,” Pattern Analysis and Machine Intelligence, IEEE Trans., pp. 603-619, 2002.

[5] K. Javed, “The Behaviour of K-means: An Empirical Study,” in Proc. ICEE 2008, Second International Conference on Electrical Engineering, Lahore, Pakistan, March 25-26, pp. 1-6.

[6] P. Ghamisi, “An efficient method for segmentation of images based on fractional calculus and natural selection,” Expert Syst. Appl., vol. 39, no. 16, pp. 12 407–12 417, Nov. 2012.

[7] P. Ghamisi , ”Multilevel Image Segmentation Based on Fractional-Order Darwinian Particle Swarm Optimization,” in IEEE Transactions On Geoscience And Remote Sensing, vol. 52, No. 5, May 2014

[8] Y.Kao and E.Zahara, "A hybridized approach to data clustering," Expert Systems with applications, vol. 34, no. 3, pp. 1754-1762, 2008.

[9] R.Kulkarni and G.Venayagmoorthy, “Bio-inspired algorithms for Autonomous Deployment and localization of sensor nodes,” SMC-C, vol. 40, no. 6, pp. 663-675, 2010.

[10] J. Tillet, T. Rao, and M. Sahin, "Darwinian Particle Swarm Optimization," in Proc. of 2nd Indian International Conference on Artificial Intelligence, Pune, India, 2005, pp. 1474-1487.

[11] P. Ghamisi, M. Couceiro, J. Benediktsson, and N. Ferreira, "An Efficient Method for segmentation of Images based on Fractional Calculus and Natural Selection," Expert Systems with Applications: An International Journal, vol. 39, iss. 16, pp. 1207-1217, November 2012.

[12] D. Fogel, Evolutionary Computation: Toward a new philosophy of machine intelligence. Piscataway, NJ, IEEE Press, 2000.

[13] http://vision.middlebury.edu/stereo/data/ [retrieved: August 15, 2014]

[14] J. Kennedy and R. Eberhart, "Swarm Intelligence," San Francisco, USA Academic Press, 2001.

[15] J. Kennedy and R. Eberhart, "A new optimizer using particle swarm theory," in Proc. IEEE 6th Int. Symp. Micro Mach. Human Sci., 1995, pp. 39–43.

[16] R. Szeliski, Computer Vision: Algorithms and Applications. Springer, 2010.

[17] M. Couceiro, P.Ghamisi, M.Martin, and J. Benediktsson "Multilevel Image Segmentation Based on Fractional-order Darwinian Particle Swarm Optimization," IEEE Trans. on Geoscience and Remote Sensing, vol. 52, pp. 2382-2394, June 2013.

[18] P.Ghamisi, M. Couceiro, M. Ferreria, L. Kumar, "Use of Darwinian Particle Swarm Optimization Technique for the Segmentation of Remote Sensing Images," in Proc. IGARSS 2012, The IEEE International Geoscience and Remote Sensing Symposium – Remote Sensing for a Dynamic Earth, Munich, Germany, July 22-27.

[19] D. Floreano and C. Mattiussi, Bio-Inspired Artificial Intelligence: Theories, Methods, Technologies. Cambridge, MA, USA: MIT Press, 2008.

[20] https://en.m.wikipedia.org/wiki/Binocular_disparity [retrieved: November 18, 2015]

[21] B. Alagoz, "Obtaining Depth Maps From Colour Images By Region Based Stereo Matching Algorithms,", OncuBlim Algorithm and Systems Labs, vol. 08, Art. No: 04, 2008.

[22] W. Chantharat and C. Pirak, "Image Transmission over ZigBee Network with Transmit Diversity," in Proc. IPCSIT 2011, International Conference on Circuits, System and Simulation, Singapore, vol. 7, pp. 139-143.

[23] www.zigbee.org. [retrieved: 2014]

[24] http://www.dashwood3d.com/blog/beginners-guide-to-shooting-stereoscopic-3d/[retrieved: September 10, 2011]

[25] IEEE Std 802.15.14: Wireless Medium and Physical Layer(PHY) Specification For Low-Rate Wireless Personal Area Networks (LR-WPANs), 2003.

[26] www.arm.com [retrieved :2014]

[27] http://www.atmel.com/Images/Atmel-6438-32-bit-ARM926-Embedded-Microprocessor-SAM9G45_Datasheet.pdf

Extending the Usable Ka-Band Spectrum for FSS Satellite Systems by using a FS Database

Wuchen Tang, Paul Thompson, Argyrios Kyrgiazos and Barry Evans
Institute for Communication Systems (ICS),
University of Surrey, Guildford
GU2 7XH, Surrey, United Kingdom
Email: {w.tang, p.thompson, a.kyrgiazos, b.evans}@surrey.ac.uk

Abstract— **Broadband access by satellite in Ka-band will become constrained by spectrum availability. In this context, the European Union (EU) FP7 project CoRaSat is examining the possible spectrum extension opportunities that could be exploited by a database approach in Ka-band via the use of cognitive mechanisms. The database approach utilizing spectrum scenarios between Fixed Satellite Services (FSS), Fixed Services (FS) and Broadcast Satellite Service (BSS) feeder links are considered. Database statistics for several EU countries are also provided for database analysis. Interference in the downlink scenarios are evaluated by the database approach using real databases and propagation models. The importance of using correct terrain profiles and accurate propagation models are shown. For the case of BSS interference to the FSS downlink (17.3-17.7GHz) it is demonstrated that in the UK an area of less than 2% is adversely affected. FS interference into the FSS downlink 17.7-19.7GHz is shown for the UK to only affect a small percentage of the band at any location. Some initial preliminary findings when considering earth stations on moving platforms are also presented. It is concluded that by using a database approach to allocate frequencies it is possible to use most of the band across different locations for satellites services in the shared Ka-band.**

Keywords – Database approach; frequency sharing; propagation models; area analysis; spectrum analysis.

I. INTRODUCTION

In this paper we address the extension of spectrum for satellite systems in the Ka-band using a database approach [1].

The demand for higher rate and reliable broadband communications is accelerating all over the world. Within Europe the Digital Agenda sets a target for universal broadband coverage of at least 30 Mbps across the whole of Europe by 2020 and 100 Mbps to at least 50% of the households [2]. Fixed connections and cellular cannot alone meet this target, particularly in the rural and remote areas but also in some black spots across the coverage. In these latter regions satellite broadband delivery is the only practical answer as satellite will cover the whole territory. Some recent studies of the roll out of broadband have shown that up to 50% of households in some regions will only have satellite available as a means of accessing broadband and thus 5-10 million households are potential satellite customers [3]. Current Ku band satellites do not have the capacity to deliver such services at a cost per bit that makes a business case and thus, the satellite community has turned to High Throughput Satellites (HTS) operating at Ka-band and above. Examples of early HTS Ka-band satellites dedicated to such services are Eutelsat's KaSat [4] and VIASAT-1 [5]. These satellites employ multiple (around 100) beams using fourfold frequency reuse over the coverage area to achieve capacity of the order of 100 Gbps per satellite. The latter is limited by the exclusive spectrum available to satellite (FSS) of 500 MHz in both the up and downlinks and this limits the feasible user rates to 10-20 Mbps. Thus, looking ahead to the increased user demands we have to look to larger satellites (maybe up to a Terabit/s [6][7]) and to more spectrum. Moving up to Q/V bands has already been suggested for feeder links but for user terminals the additional expense is not considered desirable so we return to the problem of getting more usable spectrum at Ka band.

The Ka-band exclusive bands for satellite are 19.7 to 20.2 GHz in the downlink and 29.5 to 30 GHz on the uplink. In these bands FSS terminals can operate in an uncoordinated manner, which means that they do not have to apply for and be granted a license by the national regulators, provided they meet set performance characteristics. The issue in other parts of the Ka-band is that the spectrum is allocated, not just to FSS but also to fixed links (FS) and to BSS (uplinks for broadcast satellites) as well as mobile services (MS).

The International Telecommunications Union (ITU) has allocated this spectrum in three regions of the world as shown in Table I for Ka-band (Europe is Region 1). In these so called 'shared bands' the different services need to co-exist and this is usually done by the process of coordination. For example, a larger gateway or feeder link may use this band but is coordinated and then licensed to operate and receives protection from interference from other service users.

Within Europe, the European Conference of Postal and Telecommunications Administrations (CEPT) [8] have adopted decisions that expand those of the ITU and produce tighter regulation as follows;

- 17.3-17.7 GHz: the BSS feeder links are determined as the incumbent links but uncoordinated FSS links are also permitted in this band.
- 17.7-19.7 GHz: FS links are considered incumbent but FSS terminals may be deployed anywhere but without right of protection.
- 27.5-29.5 GHz: CEPT provide a segmentation of the band between FSS and FS portions as shown in Figure 1. Within each segment there is a specified incumbent but for instance FSS terminals can operate in FS portions provided they do not interfere with the incumbent FS.

The work reported in this paper has been conducted within the EU FP7 project CoRaSat [9][10][11][12] that examines ways in which FSS satellite terminals in the Ka-band can co-exist with FS and BSS links given the regulatory regime discussed above. Specifically, a database approach for such coexistence schemes is investigated and demonstrated to exploit the frequency sharing opportunities for uncoordinated FSS terminals and verify its applicability. The aim is to show that future satellite systems can access additional spectrum beyond the exclusive bands that is needed to deliver cost effective broadband services.

Section II presents scenarios addressed and an outline of the database approach. Section III presents the database analysis and Section IV the database analysis for the specific scenarios under consideration.

Section V presents the impact on regulations and standards whilst Section VI draws the major conclusions on the work.

II. SCENARIOS AND DATABASE APPROACH

We report for the first time in the literature that a full interference analysis has been performed for frequency sharing within the frequency bands presented in Section I. We aim to show how a database approach will allow sharing between the satellite and terrestrial services. Within the CoRaSat project three scenarios have been investigated that reflect the three spectrum components detailed in the previous section. In Figure 2 we illustrate the interference paths in these scenarios. Two of the scenarios are downlink for the FSS; scenario A, 17.3-17.7 GHz where the potential interference is from BSS uplinks and scenario B, 17.7-19.7 GHz where the potential interference is from incumbent FS transmitters. In both of these cases the FSS is permitted to operate but is not protected by the regulatory regime and thus it is important to ascertain the level of the interference and its effect on the FSS received signal. The third scenario C, is in the transmit band of the FSS from 27.5-29.5 GHz and the interference is from the FSS transmitting earth station into the FS receivers, which are protected. The latter is more critical in that we need to demonstrate that the FSS does not contravene interference limits imposed by the regulatory regime.

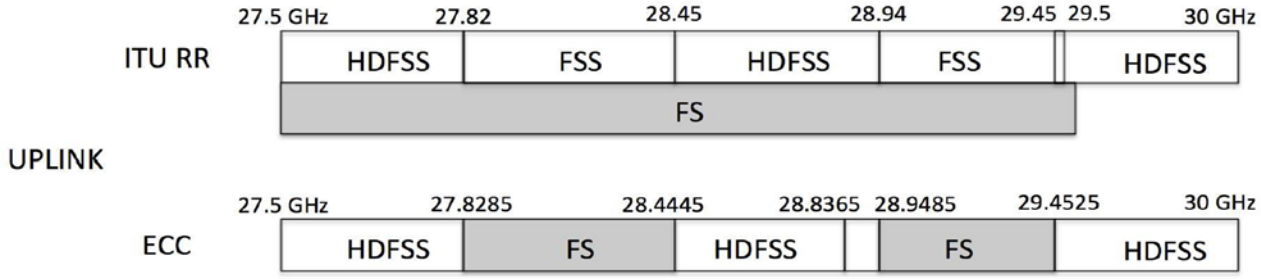


Figure 1. CEPT 27.5-29.5 GHz Segmentation

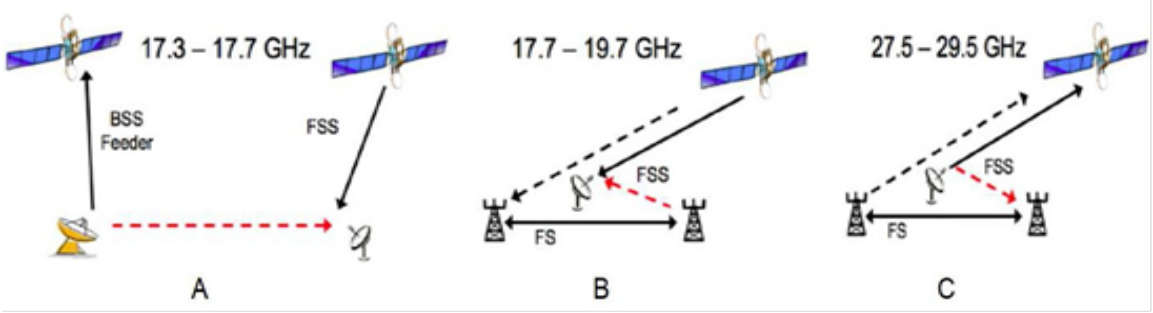


Figure 2. Scenarios in CoRaSat

The forward link, e.g., the downlink can be considered more important in that the ratio of downlink broadband to uplink broadband as operated via satellite is currently at around 6:1 and thus the acquisition of more spectrum here is key. In addition to this, operation in the downlink bands does not require regulatory changes but merely reassurance to the FSS users that the services need not be impaired.

The calculation of interference can be performed if one has obtained the corresponding accurate FS database, which includes the characteristics and locations of the potential interferers, then using this with equipment models, propagation models and the path details.

Similar ideas have been employed in Television White Space (TVWS) systems [13] to allow UHF frequencies to be used in the gaps between TV transmission regions. For scenario A the number of BSS uplinks in Europe is small and thus a database system is similar in magnitude to that of TVWS. However, for scenarios B and C the number of FS links runs into the tens of thousands and the database is much more complex. The data on the positions and the characteristics of the BSS and FS are generally held by national regulators and these need to be available for a database system to work.

The information from a real interferer database is interfaced to an interference modelling engine, which uses ITU-R Recommendation P.452-15 [14] procedures plus terrain and other databases. This is the latest version of this ITU Recommendation that contains a prediction method for the evaluation of path loss between stations. ITU-R P.452-15 includes all the propagation effects on the surface of the Earth at frequencies from 0.1 GHz to 50 GHz. In addition, other factors, which affect interference calculation, such as terrain height and bandwidth overlapping are also considered in the proposed database approach, which is illustrated in Figure 3. The typical interference threshold we determine is based on the long term interference, which can be expected to be present for at least 20% of the average year and it is set at 10 dB below the noise floor.

The interference thresholds for FSS reception and for FS reception are therefore -154 dBW/MHz and - 146 dBW/MHz, respectively as given in [15] and [16].

Having determined the interference level at the FSS (in scenarios A or B) it can be compared with the regulatory threshold. The action is then taken in the resource allocation at the gateway where a new carrier can be assigned either in another part of the 'shared band' where interference is acceptable or in the exclusive band. For scenario C the situation is different as the interference is caused by the FSS into the FS. Here the database is used to calculate the maximum permissible power that can be transmitted from the FSS in the vicinity in order to retain the threshold condition at the FS receivers. More details of the database approach are given in the following sections.

TABLE I. EXTRACT OF ITU TABLE OF FREQUENCY ALLOCATIONS

Frequency bands	ITU Region 1	ITU Region 2	ITU Region 3
17.3-17.7 GHz (Scenario A)	FSS (space-Earth) BSS (feeder links) Radiolocation	FSS (space-Earth) BSS (feeder links) Radiolocation	FSS (space-Earth) BSS (feeder links) Radiolocation
17.7-19.7 GHz (Scenario B)	FSS (space-Earth) BSS (feeder links 18.1 GHz) FS	FSS (space-Earth) FS	FSS (space-Earth) BSS (feeder links 18.1 GHz) FS
27.5-29.5 GHz (Scenario C)	FSS (Earth-space) FS MS (Mobile Services)	FSS (Earth-space) FS MS	FSS (Earth-space) FS MS

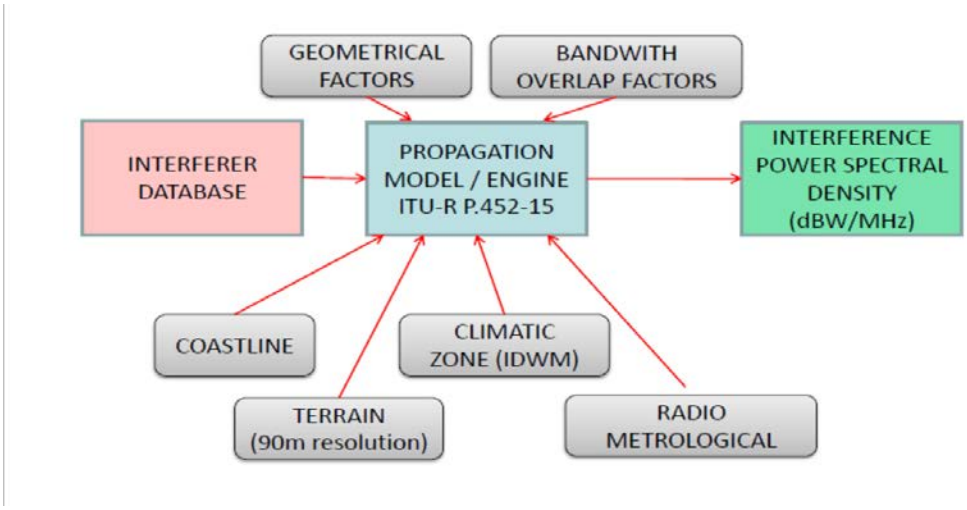
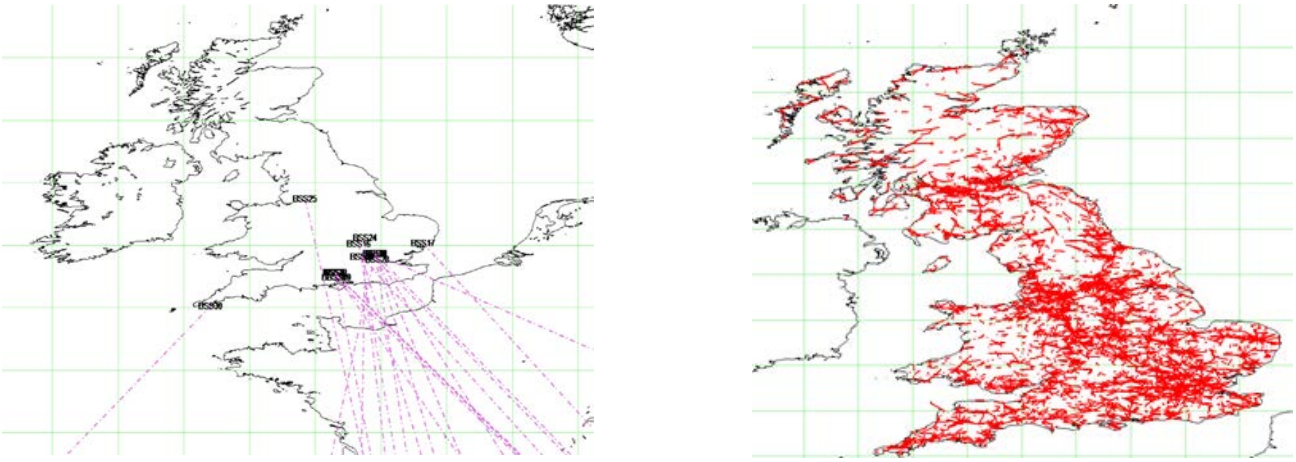


Figure 3. Interference modelling by ITU-R P.452-15



(a) Registered BSS feeder link stations in the UK. (b) Registered FS links of the whole band in the UK

Figure 4. Registered BSS and FS links in the UK.

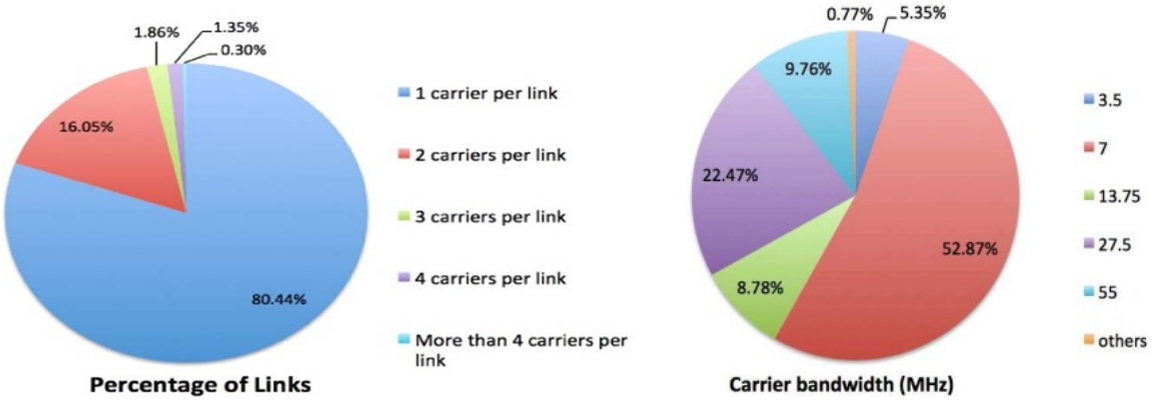


Figure 5. Pie chart of (a) UK FS carrier number of each link and (b) carrier bandwidths (MHz)

III. Database Analysis

The information in a database is normally listed on a carrier by carrier basis for a frequency band of interest. All carriers are usually detailed with their frequencies and channel bandwidths. When the database relates to satellite terminals the database should also contain details on the associated satellite in terms of satellite longitude and the earth stations azimuth and elevation angles. Polarization and antenna gain are also required along with the antenna radiation patterns as defined in ITU Recommendations for use in regulatory work or ETSI standards. In addition, transmission power and equivalent isotropic radiated power (EIRP) may also be included.

A UK BSS database made available for this study is used for scenario A and contains 442 carriers from a total of 31 BSS uplink earth stations at 8 physical sites, to 12 different satellites, which is shown as Figure 4(a). The locations of all these 31 BSS earth stations are marked with an indication of the direction of the beam to the satellite. The number of carriers of each BSS earth station ranges from 1 to 42. The carriers span the range 17.3 GHz to 18.35 GHz. The bandwidths of the carriers that belong to the same BSS earth station are the same while those that belong to different earth stations might be different and are typically 26 MHz, 33 MHz, 36 MHz or 66 MHz. The EIRP of these earth station antennas ranges from 69 dBW-84 dBW and all antenna radiation patterns are as defined in ITU-Recommendation S.465 [17] or S.580 [18].

FS databases at 18 GHz are required to evaluate scenario B. Again, an FS database was made available to this project (under the UK Freedom of Information Act). The database for the UK FS in the band 17.7 to 19.7 GHz is much larger than that for the UK BSS one and contains 12,712 links with 15,970 carriers recorded in the UK. A French database has also been examined at 18 GHz and is based on the latest ITU-R terrestrial services BR IFIC database [19], which contains 11,548 links with 17,384 carriers recorded. Figure 4(b) illustrates the FS links in the band 17.7 to 19.7 GHz in the UK and it can be seen that the FS links are much denser than for the BSS.

Figure 5 provides pie charts of numbers of carriers per link and carrier bandwidths based on the UK FS database in (a) and (b), respectively. It is indicated that more than 80% of links have only one carrier and more than 96% of links have up to 2 carriers. The majority of carriers have a bandwidth from 3.5 to 55 MHz. As a consequence, it can be deduced that at a particular location in the UK, little spectrum resource from the available 2 GHz band is used by the FS at a specific location. Thus, we are optimistic that spectrum available for FSS on a micro scale geographical basis is significant and can be exploited if the information of spectrum occupancy is known from the analysis of the database or is detectable by other mechanisms. A similar situation also exists for France.

We have implemented the ITU-R.P452-15 propagation and interference modelling to provide cognitive zones around incumbent terminals based on the available database. A cognitive zone here is defined as the geographical area around an incumbent user station where cognitive radio techniques such as spectrum sensing and beamforming should be employed to mitigate the interference to an acceptable level. In other words, the interference outside of this area is below the acceptable interference threshold thus, cognitive radio techniques are not necessary.

Figure 6 (a) and Figure 6 (b) show plots of cognitive zones around a BSS Station under scenario A case based on a free space loss model and the full ITU-R P.452-15 model, respectively. Similarly, Figure 6(c) and Figure 6(d) show plots of cognitive zones around a FS Station under scenario B case based on these two models. For the BSS cognitive zone the FSS terminal evaluated points to a satellite at 53 degrees E longitude and the BSS transmitting terminal points to a satellite at 28.2 degree while for the FS cognitive zone the FSS terminal is pointing to a satellite at 20 degrees E longitude and the FS transmitting terminal is pointing at a receive terminal on a bearing of 110 degrees East of True North (ETN). Clearly, for both cases the cognitive zones from the full model are much smaller and differently shaped than the ones under the free space model. On the average the areas are 9 times smaller at the -155 dBW/MHz and 3.5 times smaller at the -145 dBW/MHz thresholds. This is mainly because the diffraction effect based on the terrain data is considered in the full model while the free space loss model only includes line of sight propagation loss, which reflects the fact that the terrain databased diffraction effect is extremely significant in cognitive zone determination.

IV. Database Applications for the Scenarios

In this section we analyze for scenario A the areas that are affected by interference from BSS uplinks and for scenario B the availability of spectrum at FSS locations as a result of FS interference. Typical examples are provided to demonstrate the additional spectrum that could be available.

A. Scenario A: Area Analysis

Using the BSS database, area analysis for scenario A in the UK is provided to investigate how much area would be affected by interference from the BSS feeder links. The band of interest is split into 10 x 40 MHz sub-bands (SB1−SB10) and the analysis is then conducted in each sub-band to determine the area of the contours at different cognitive zone thresholds. These mirror the usual 40 MHz channel spacing adopted for BSS satellites. Area analysis is based on BSS database with the full ITU-R P.452-15 model employing the terrain and climatic zones and the FSS terminal evaluated points to a satellite at 53 degrees E longitude. The results are for long term interference (normally 20%), it being assumed that Adaptive Coding and Modulation (ACM) will mitigate short duration interference events including rain fades.

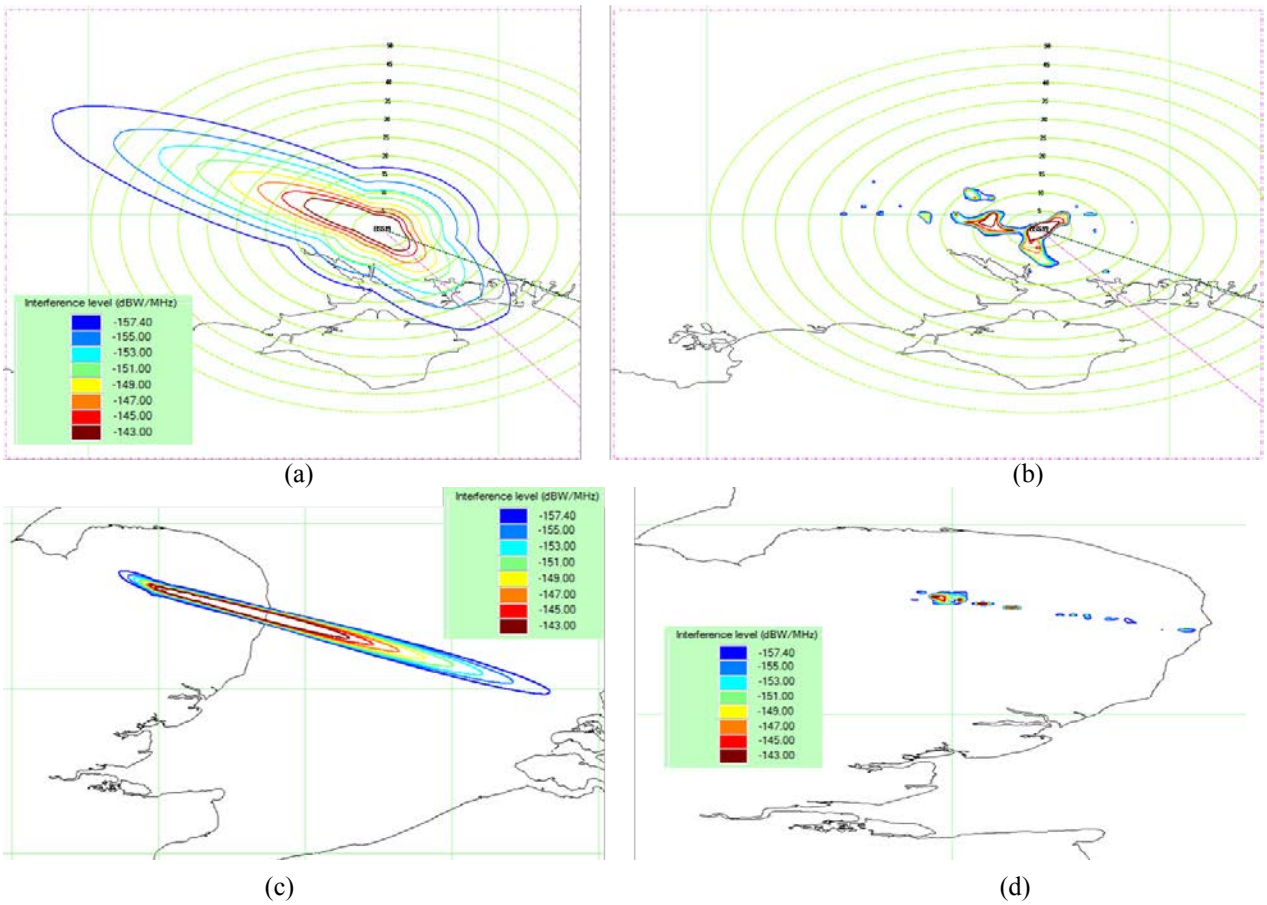


Figure 6. Example of cognitive zone for (a) BSS with free space loss model (b) BSS with full ITU model (c) FS with free space loss model (d) FS with full ITU model.

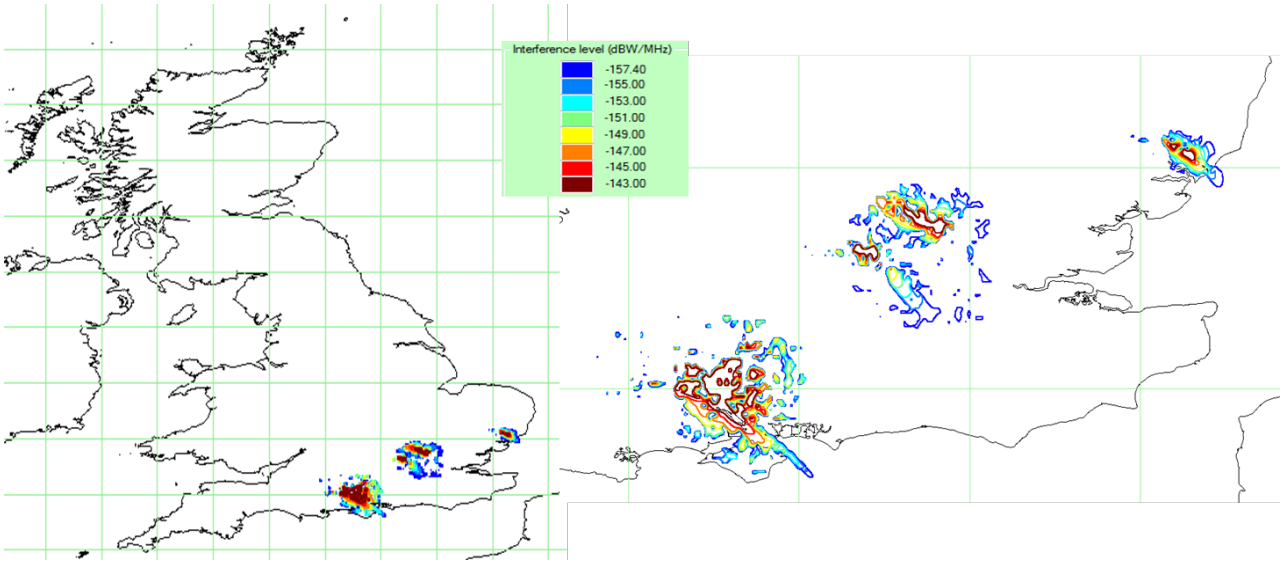


Figure 7. Example of cognitive zones for the sub-band 1 (17.3-17.34 GHz) based on full ITU model

One example of an affected area at different cognitive zone thresholds is shown in Figure 7, which represents SB1. Full data on the areas are given in the Table II. It can be seen that in general across the sub-bands at a -155 dBW/MHz threshold less than for 2% of the area of the UK is affected by BSS feeder links and thus more than 98% of the area of the UK can be used by an FSS terminal without the need for any further action. Some mitigation of excess interference may be required in these affected areas. Such mitigation could be achieved by suitable site shielding, beam-forming or reallocation to another frequency that is clear at the specific location. If such mitigation measures result in 10 dB suppression (a very conservative figure) then the remaining affected area would be of the order of 0.4% of the area of the UK. Re-farming the spectrum of such a small amount of traffic should not represent much of a challenge. This is very promising for future FSS deployment as the additional 400 MHz identified in scenario A (17.3-17.7 GHz) represents an 80% increase over the current exclusive band allocation (19.7-20.2 GHz).

Although we have presented results herein for an FSS terminal pointed at a specific orbit location we have examined a range of orbit locations from the UK and the results are very similar.

B. Scenario B: Spectrum Analysis

Unlike the situation in scenario A, the UK 18 GHz FS database comprises many more carrier records (15,036 records) over the 2 GHz band from 17.7 to 19.7 GHz. For scenario B we perform spectrum analysis for a particular location in the UK instead of geographical area analysis across the whole of the UK to determine what carrier(s) can be used by an FSS at a specific location. This information could then be integrated with a resource allocation algorithm in the satellite network to assign the carriers.

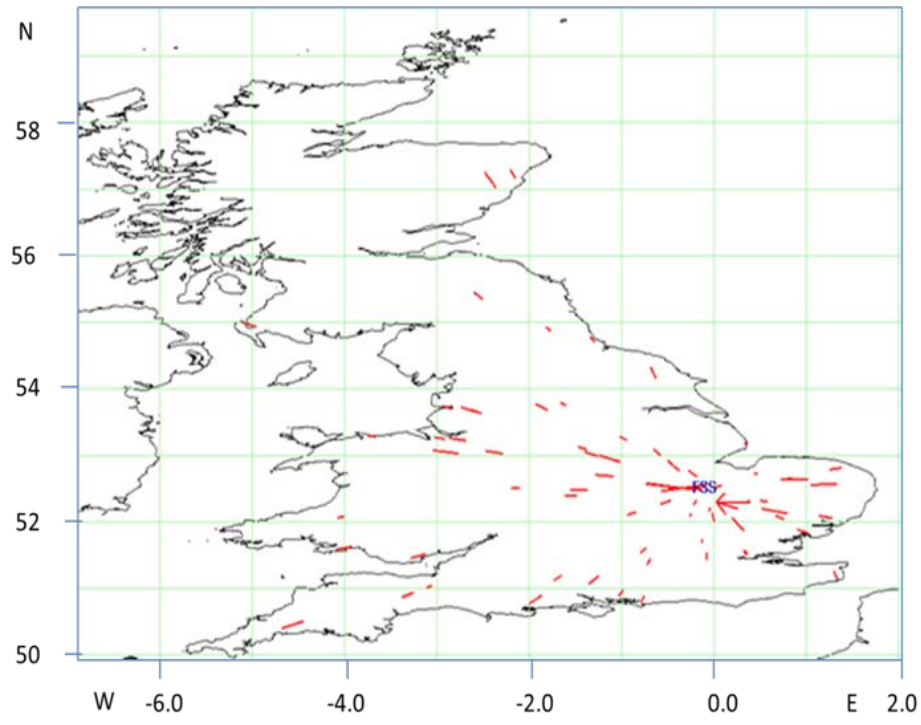
Spectrum analysis results for the UK FS links at 18GHz at a specified location (with latitude of 52.5 degrees and longitude of -0.1 degree) is shown here as an example. The analysis results of the location with both Line of Sight (LOS) and full model (ITU-R P.452-15) are shown as Figure 8 and Figure 9, respectively. The FSS terminal evaluated, points to the same satellite as the previous examples, which is located at 53 degrees E longitude. In each figure, a map of the links that exceed an interference level of -160 dBW/MHz is presented along with spectrum analysis as a plot of the interference power spectral density (PSD). Interference PSD is shown per MHz from 17.7 to 19.7 GHz. At this location, it can be seen that with the LOS model the interference from FS links can be much farther from the location of interest and these links are ones pointing directly at the location. There are only a few points with some angular offset and these are located very close so that interference is from their side lobes. From the interference PSD in Figure 8, it can be seen that there are significant white spaces in the plot and thus more than half of spectrum resource from 17.7 GHz to 19.7 GHz is available (with interference below the threshold) at this location for the LOS model. However, if the full terrain model is considered as in Figure 9, the number of interfering FS links dramatically decreases to less than ten, which means less than 0.1% of total FS links would cause problems at the location. Therefore, the majority of the 2 GHz band can be used by an uncoordinated FSS VSAT terminal site.

Complete maps of interference for locations in various European countries have been produced and these can be used as input to a resource allocation scheme that would then optimize the carrier allocation on the basis of the extra spectrum available. Examples for the UK and France are shown in Figure 10 and Figure 11 in terms of interfering spectrum occupied by the FS. It was noted that although the number of FS links in the database was large those that actually caused interference at a specific location and in a particular frequency band were quite small. It should also be noted that the available spectrum is not the same at each location and thus the database analysis can be used to optimize the carrier allocation as a function of FSS location.

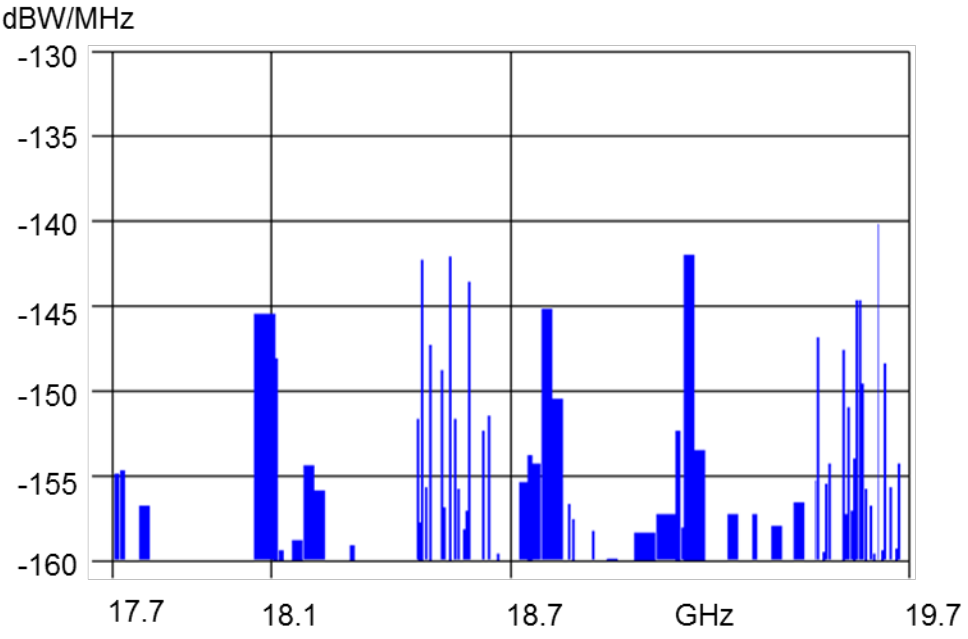
Calculations were also performed in order to assess the impact on a wider scale. In this case a range of FSS terminal locations from 8 degrees West to 2 degrees East at a longitude of 51.5 degrees North were used to assess the impact. In this case we present the number of FS links that would interfere (at -154 dBW/MHz interference threshold) for each location. The results are presented in Figure 12. It can be seen that the influence of the diffraction model is significant and for the more serious levels of interference the ratio of interfering links is of the order of 4.5:1.

TABLE II. AREA ANALYSIS (SQ. KM.) OF THE BAND 17.3-17.7 GHZ (BSS)

17.3−17.7GHz	**SB1**	**SB2**	**SB3**	**SB4**	**SB5**
−155 dBW/MHz	2, 420.9 (1.06%)	1, 692.4 (0.74%)	1, 692.4 (0.74%)	1, 683.3 (0.73%)	3, 570.9 (1.56%)
−145 dBW/MHz	683.0 (0.30%)	544.8 (0.24%)	544.8 (0.24%)	541.8 (0.24%)	926.0 (0.40%)
17.3−17.7GHz	**SB6**	**SB7**	**SB8**	**SB9**	**SB10**
−155 dBW/MHz	1, 683.3 (0.73%)	2, 411.0 (1.05%)	2, 535.6 (1.11%)	2, 367.6 (1.03%)	2, 936.4 (1.28%)
−145 dBW/MHz	541.8 (0.24%)	741.3 (0.32%)	774.2 (0.34%)	697.5 (0.30%)	928.6 (0.40%)

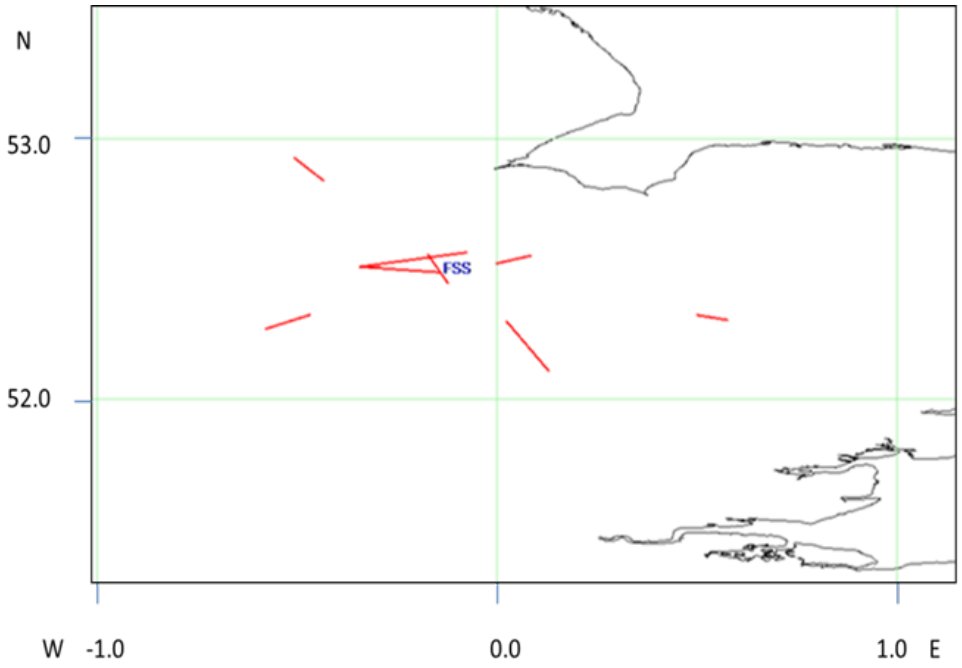


(a)

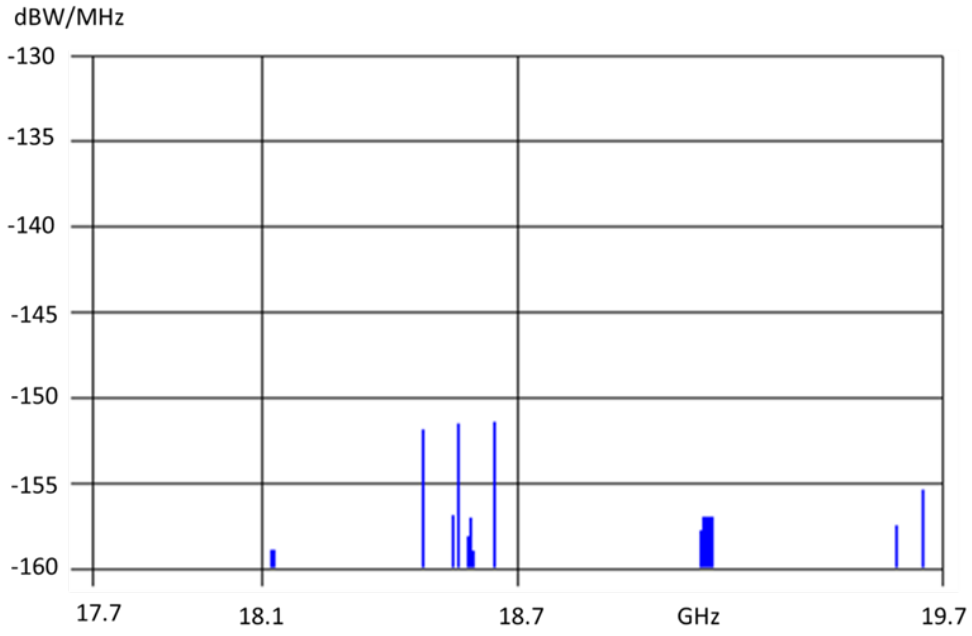


(b)

Figure 8. LOS result of all UK FS links, interfering to FSS terminal at latitude of 52.5 degs and longitude of -0.1 degs. (a) contributing links, (b) interference spectrum



(a)



(b)

Figure 9. Full ITU-R P452-15 result of all UK FS links, interfering to FSS terminal at latitude of 52.5 degs and longitude of -0.1 degs. (a) contributing links, (b) interference spectrum

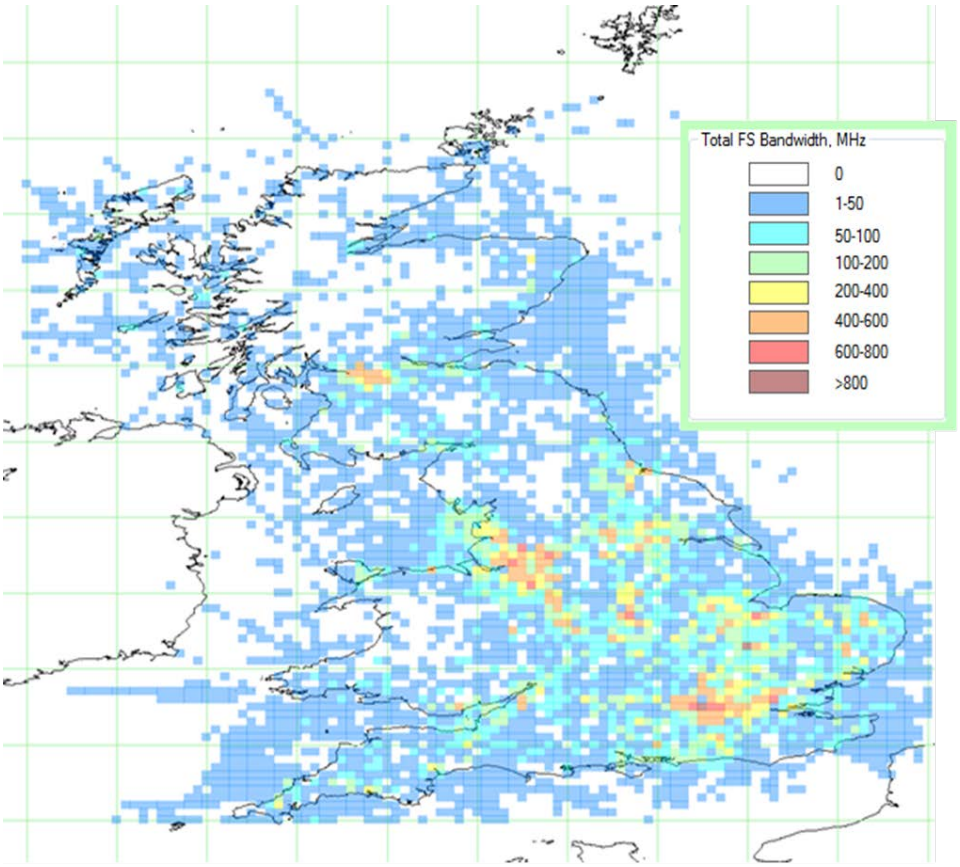


Figure 10. Interfering spectrum at -154 dBW/MHz for the UK

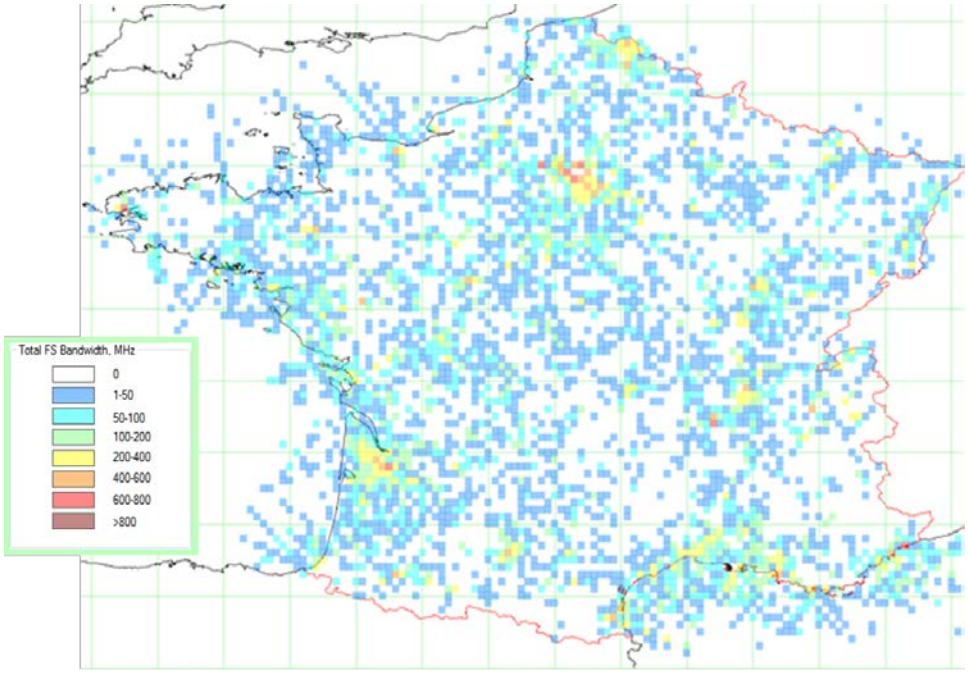


Figure 11. Interfering spectrum at -154 dBW/MHz for France

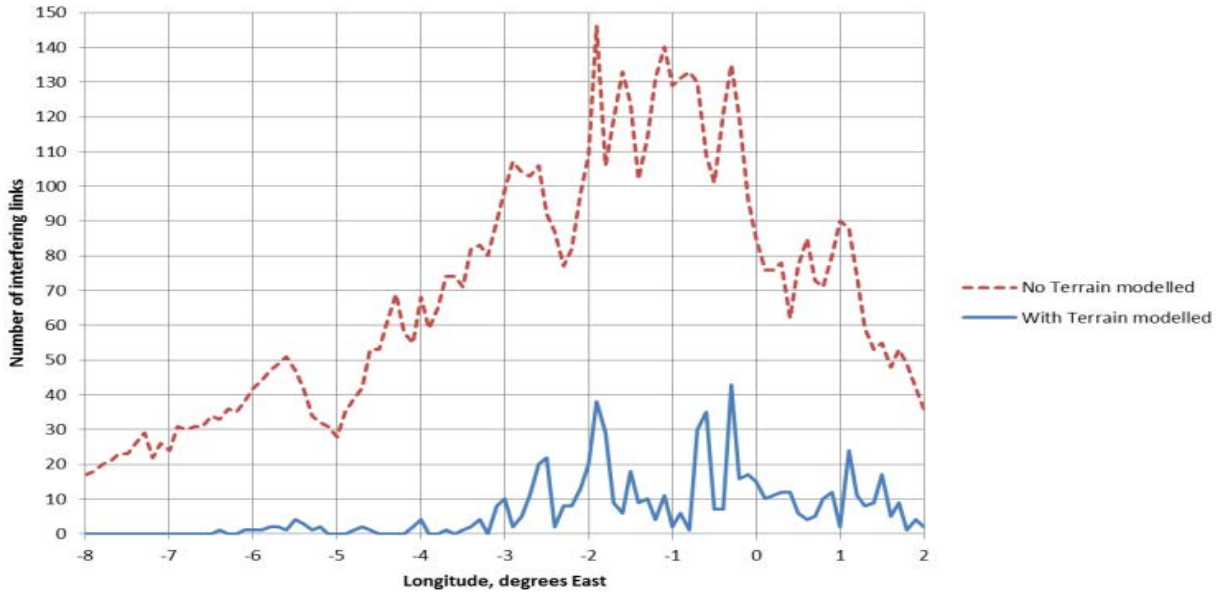


Figure 12. Number of links interfering (with and without terrain effects)

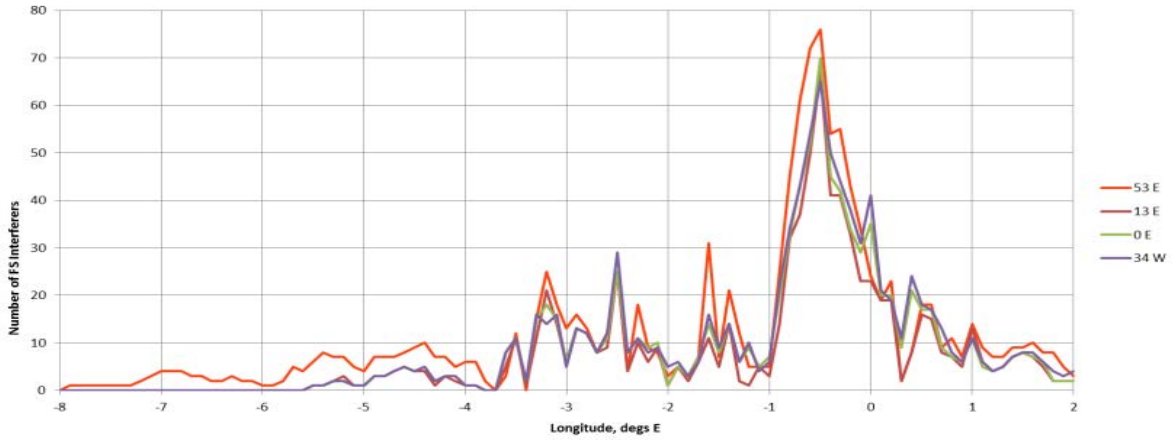


Figure 13. Number of FS interferers for different satellite locations (longitude)

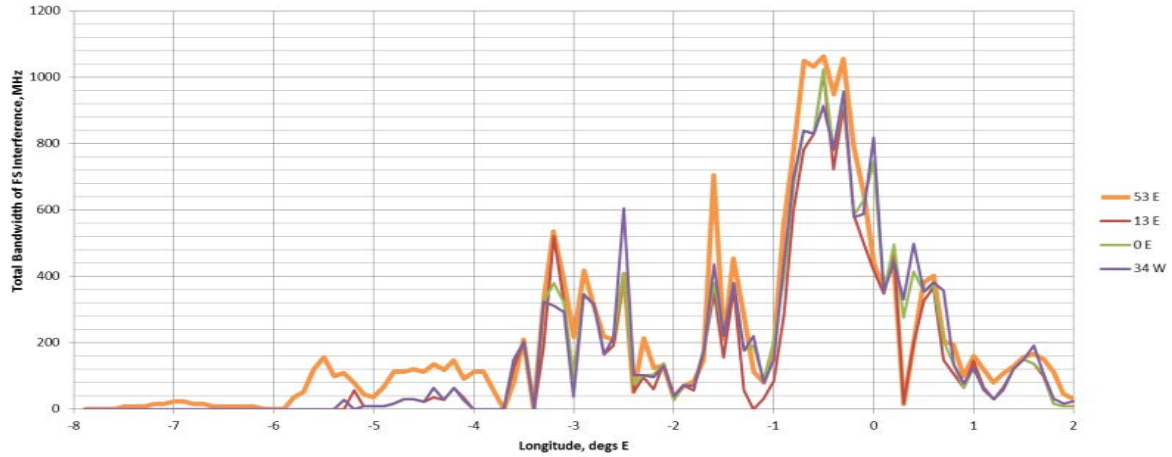


Figure 14. Total Bandwidth occupied by FS interferers for different satellite locations (longitude)

Impact of FSS Satellite Longitude

Calculations have been performed to assess the impact of the FSS antenna pointing to the satellite when the latter varies from 53E, 13E, 0E and 34W degrees longitude. Again, a range of FSS terminal locations from 8 degrees West to 2 degrees East at a longitude of 51.5 degrees North were used to assess the impact. Figure 13 shows the number of FS interferers for four different satellite locations (longitude). Figure 14 depicts the total bandwidth occupied by FS interferers for the same cases.

From these figures it can be seen that the assumption that the 53E longitude satellite represents the worst case scenario is validated.

Area Analysis of the FS interference

The ITU terrestrial services Radiocommunications Bureau (BR) International Frequency Information Circular (BRIFIC) provides us with databases to analyse the potential interference in the band of interest and from the results it is possible to get an increased insight into the situation. The analysis was conducted in UK, France, Poland, Hungary and Slovenia with the full diffraction model and statistics were derived from the results. To permit a fair comparison between the countries only the results for test point over land were included. The first set of statistics presented is the cummulative distribution function (CDF) of the number of interefering signals that exceed the -154 dBW/MHz threshold at each point. The resulting CDFs are presented in Figure 15.

A CDF was also produced for the total occupied bandwidth of the FS interferers at a point over the regions of interest. The resulting CDF is given in Figure 16.

NOTE: Also, in Figure 16, a second horizontal axis at the top indicates percentage of the total spectrum occupied by the FS.

The graphical results are also presented herein as tables.

Table III presents the CDF of number of FS links that interfer with a site in terms of the percentage of sites affected. Table IV presents the CDF of the total bandwidth of the interfering FS links that interfer with a site in MHz. Table V presents the same CDF but in terms of the percentage of the total bandwidth (17.7 – 19.7 GHz).

Maps of the total occupied bandwidth at locations within The UK and France were shown earlier in Figure 10 and Figure 11. These have also been produced at higher resolution for use as inputs to resource allocation software at the network gateway.

The results of the analysis for scenario B across various European countries is that there is over 90% of the 2GHz band between 17.7 and 19.7GHz available for most positions in the countries examined. The latter are considered to be typical and indeed represent the most dense distribution of FS in Europe. Thus, using a database or an interference map produced from the database to control the carrier allocations at the network gateway it should be possible to use this shared band spectrum for FSS down links.

Within the CoRaSat project there have been capacity gain calculations using a model multi beam satellite over Europe. These calculations are very dependent on the satellite characteristics and in particular on the satellite antenna C/I distribution across the beams. However, it was shown that in such a system using both the shared and exclusive bands a 400% increase in the forward capacity could be achieved [20].

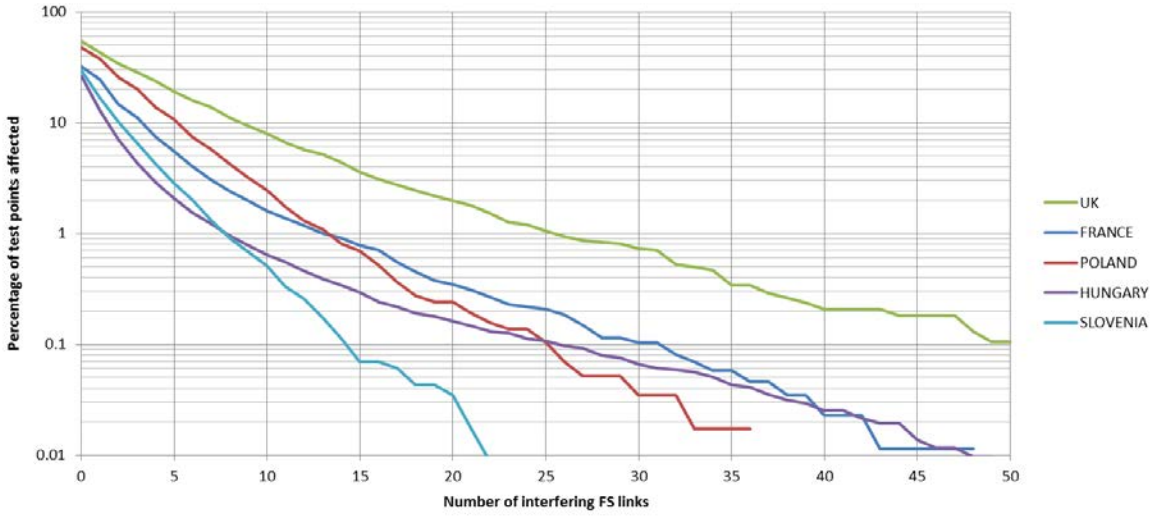


Figure 15. CDF of number of interferers at a test point (for -154 dBW/MHz threshold)

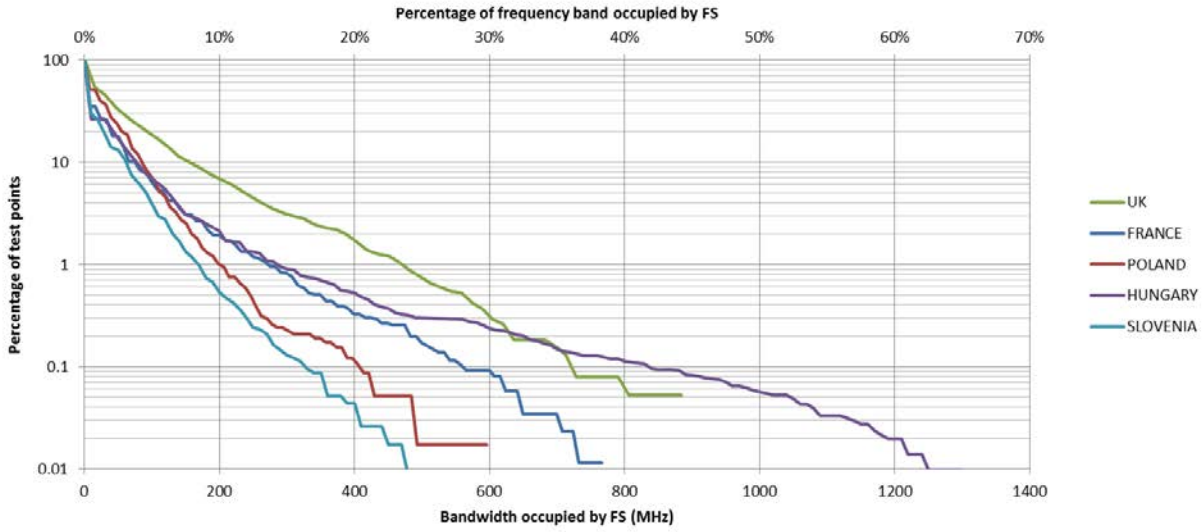


Figure 16. CDF of FS bandwidth occupied by interferers over the five regions (for -154 dBW/MHz threshold)

TABLE III. CDF OF NUMBER OF INTERFERES PER FSS SITE

% of sites	UK	FRANCE	POLAND	HUNGARY	SLOVENIA
>0	54%	31%	47%	26%	29%
10%	>9	>4	>5	>2	>2
1%	>25	>13	>13	>8	>8
0.1%	>50	>31	>25	>26	>13

TABLE IV. CDF OF TOTAL BANDWIDTH OF FS LINK INTERFERENCE PER FSS

% of	UK	FRANCE	POLAND	HUNGARY	SLOVENIA
sites	MHz	MHz	MHz	MHz	MHz
10%	139	58	80	45	50
1%	450	258	190	270	160
0.1%	700	550	400	820	405

TABLE V. CDF OF TOTAL BANDWIDTH OF FS LINK INTERFERENCE PER FSS SITE (% OF 17.7 – 19.7 GHZ)

% of sites	UK	FRANCE	POLAND	HUNGARY	SLOVENIA
10%	7%	3%	4%	2%	3%
1%	23%	13%	10%	14%	8%
0.1%	35%	28%	20%	41%	20%

C. Scenario C

Scenario C addresses the uplink from 27.5 to 29.5 GHz as shown in Figure 1. The ITU allows sharing across this band between FS and FSS but in Europe the CEPT have segmented the band between FS and HDFSS. For applications of HTS satellite systems designed for broadband internet access the uplink requirements are less than the down link. Current Ka-band systems in Europe are indicating an asymmetry of around 6:1. Thus, in Europe the availability of the two HDFSS bands may be adequate for early systems. However, in other regions of the world there could be a need to coexist in the uplink as well as the downlink.

In this case cognitive zones for scenario C are around FS stations and interference is from FSS terminals to FS links. This is a much more difficult case to address if we plan to use the whole of the shared band because we do not have access to adequate 28 GHz databases on which to operate. The techniques and software developed for scenario A and B can be used in a similar manner for scenario C if we have access to such a database. The results would be presented in a slightly different format as they would give the maximum allowable EIRP for the FSS at a given point. However, the HDFSS uplink band has been agreed for uncoordinated earth stations in all except 5 of the EU countries, therefore, perhaps the uplink increase in spectrum is not so urgent at this time. Some preliminary evaluation of scenario C was performed in [21] and the results of this indicated that only very close FS links (around 10km) would be affected and the density of FSS terminals would not cause a problem in the multi interference case. Some preliminary evaluations have also been done in the CoRaSat project using databases in Slovenia and Finland. These demonstrate that the reductions in FSS EIRP's are small and interference is not that much of a problem for the long term availability criteria. Thus, sharing of the uplink looks feasible but more evaluations are needed to be done for a wider selection of databases to confirm these results.

D. Earth Stations on Moving Platforms (ESOMPs)

More recently work has been focused upon Ka-band operation of ESOMPs in the shared bands.

The ESOMPs cases considered are as follows:

- Case 1: Aircraft- mounted ESOMP with downlink in the band 17.7-19.7 GHz and uplink in the band 27.5 to 30 GHz.
- Case 2: Ship-mounted ESOMP with downlink in the band 17.7-19.7 GHz and uplink in the band 27.5 to 30 GHz.

For Case 1 the situation is quite complex. The very directive nature of the FS antennas and the airborne antenna contribute very significantly to the level of interference and the number of significant interfering links, which vary quite rapidly as the aircraft travels along its flight path. Scenarios A, B and C can be considered for such cases where the FSS terminal is the ESOMPs terminal.

Analysis of Scenarios A and B for this case indicates that the above mentioned antenna effects will help reduce the cumulative interference levels to values that can be managed by appropriate mitigation techniques. An example of such analysis is given below.

For both cases, 1 and 2, for Scenario C we have limited our considerations to the case where in Europe the availability of the two HDFSS bands may be adequate for such systems and interference mitigation is not required. However, in other regions of the world this may not necessarily be the case. Work elsewhere has been addressing this matter in considerable detail [22], [23].

Case 1 Aeronautical ESOMP

The methodology adopted earlier using ITU-R P.452-15 can be extended to an aeronautical case by applying a number of critical modifications. These are:

1. Increase the height above mean sea level of the victim receiver so that it corresponds to the altitude of the aircraft;
2. Find the range from the FS transmitter to the aircraft for use in the calculations;
3. Find the azimuth and elevation angles of the aircraft as viewed from the FS transmitter;
4. Find the azimuth and elevation angles of the FS transmitter as viewed from the aircraft;
5. Using the above, determine the off-axis angle and thus gain of the FS transmitter antenna;
6. Using the above, determine the off-axis angle and thus gain of the aircraft receiving antenna;
7. Adopt the more complex ITU-R P.676-10 annex 1 model which is applicable to low elevation angles for calculating the gaseous losses;
8. Include the effect of aircraft fuselage attenuation;
9. Adjust the parameters to take account of the fact that the aircraft receive antenna diameter is 0.6 metres.

These modifications have been undertaken and wherever possible validated to be correct.

By way of an example we have performed calculations for an airborne ESOMP flying along the example path indicated in Figure 17 (a) and Figure 17 (b) at two different altitudes. The first altitude is 3.81 km (12,500 ft) and the second is for 11.88 km (39,000 ft), which is the average of the maximum altitude capability of known commercial airliners.

Figure 17 (b) has a background that is indicative of the general interference level in terms of bandwidth used by the FS but is indicative rather than being specific to ESOMPs.

The flight path is therefore over the more dense interference regions of the UK and typical of paths near Heathrow Airport. Detailed airborne ESOMP results are presented in Figure 17 (c), Figure 17 (d) and Figure 17 (e).

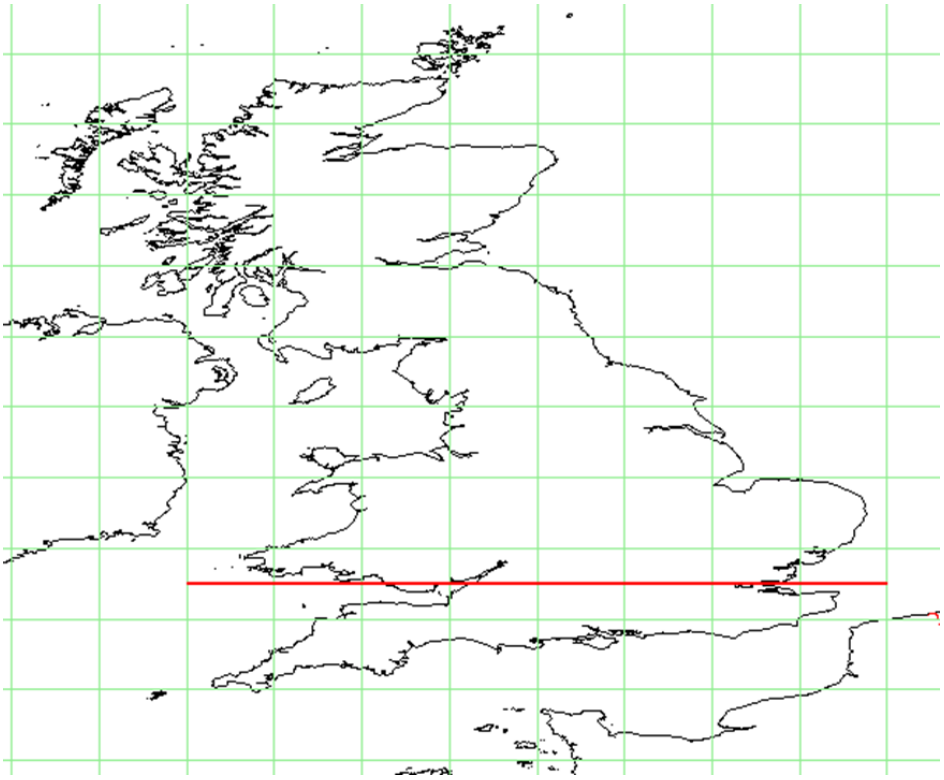


Figure 17. (a) ESOMP Flight Path

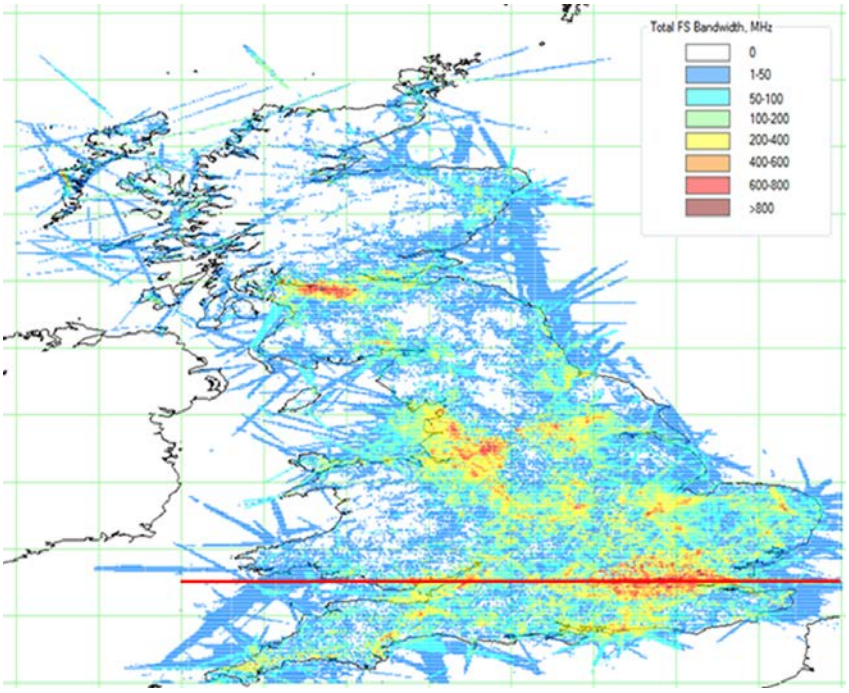


Figure 17. (b) Indicative Interference (FS BW used) around the flight path

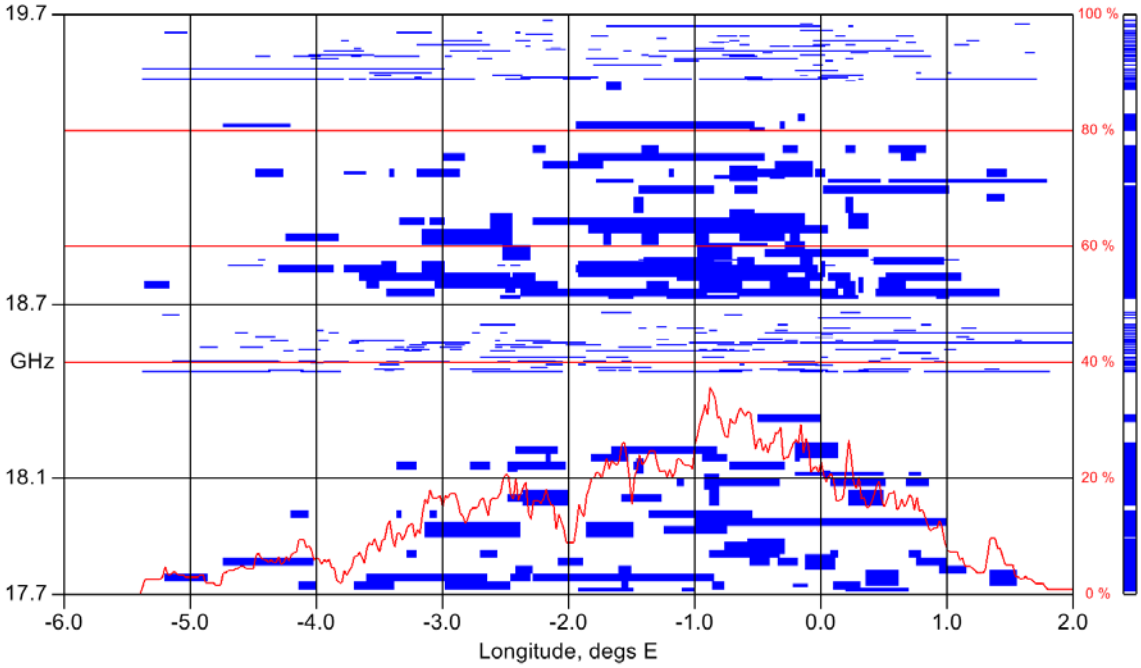


Figure 17. (c) ESOMP altitude 3.81 km (12,500 ft), threshold = -154.5 dBW/MHz

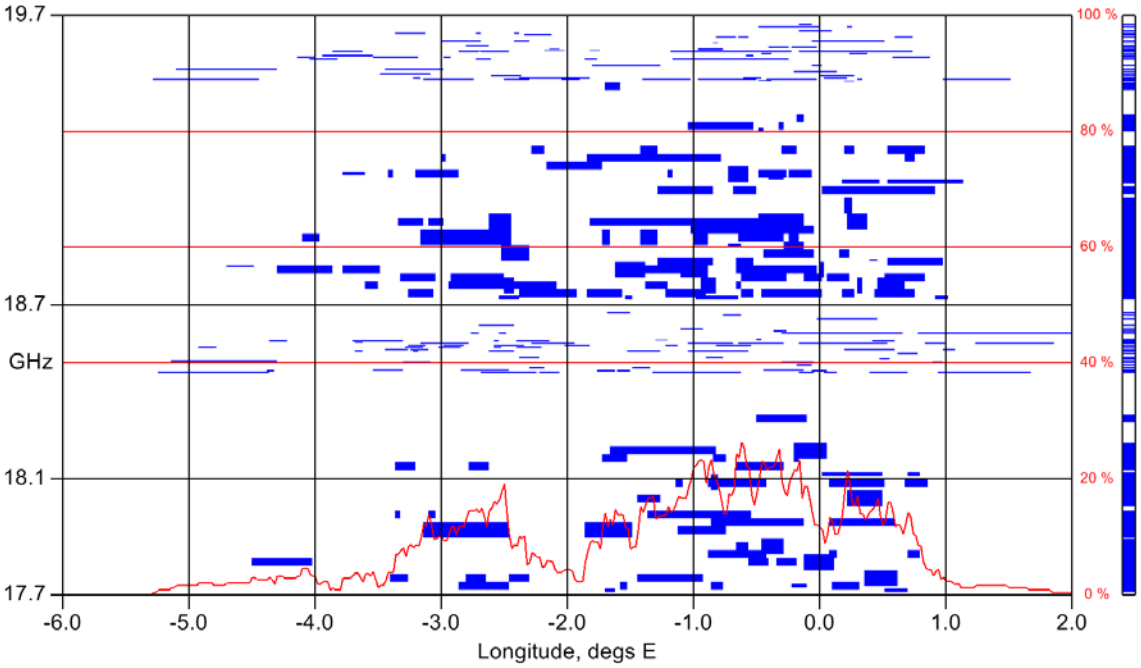


Figure 17. (d) ESOMP altitude 3.81 km (12,500 ft), threshold = -144.5 dBW/MHz

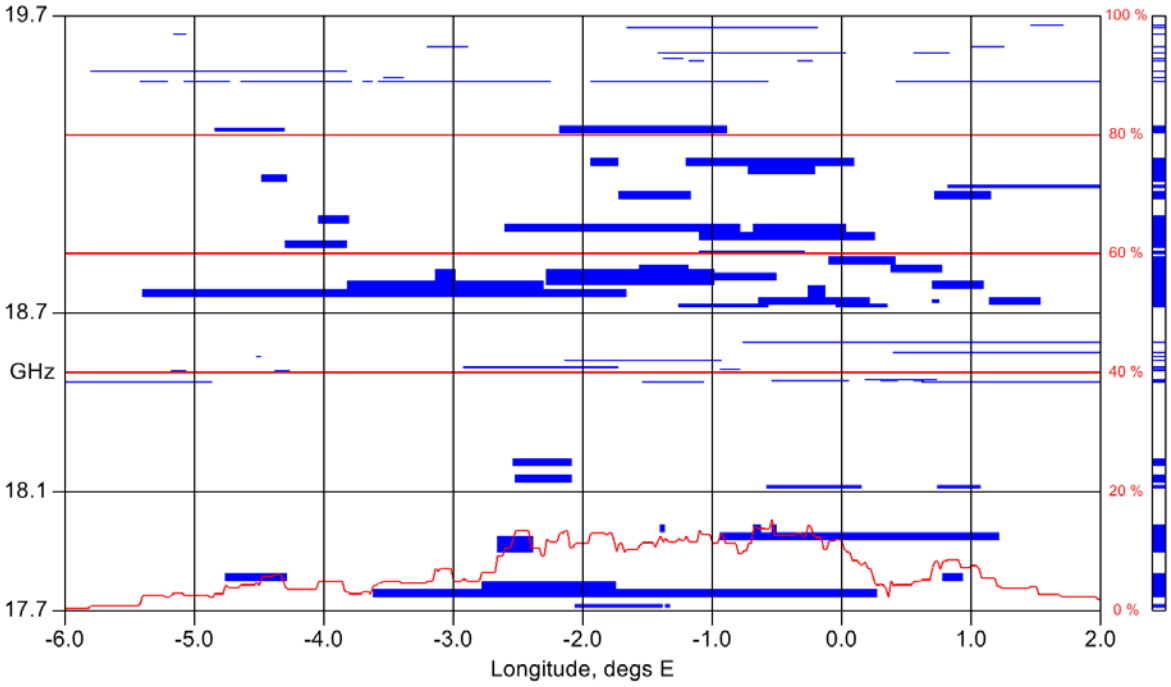


Figure 17. (e) ESOMP altitude 11.88 km (39,000 ft), threshold = -154.5 dBW/MHz

Figure 17 (c), Figure 17 (d) and Figure 17 (e) have several special features to note. Each interferer that exceeds the chosen threshold is shown as a blue band at the carrier frequency with its vertical extent indicating the carrier bandwidth. The length of the blue band indicates the longitude range (related to time) that the carrier remains an interferer. The narrow box on the right hand side of the figure indicates the composite amalgamation of all the interference entries. The red curve is associated with the scale on the right hand side of the figure and represents the percentage of the 2 GHz of available bandwidth that interference from the FS exceeds the given interference threshold.

Figure 17 (c) presents results for a low flying ESOMP at an altitude of 3.81 km (12,500 ft) with a threshold of -144.5 dBW/MHz. Figure 17 (d) is for the same case with a threshold of -144.5 dBW/MHz. Figure 17 (e) presents the results for an ESOMP flying at an altitude of altitude 11.88 km (39,000 ft) with a threshold of -154.5 dBW/MHz. It can be seen that the ESOMP experiences less FS interference at the higher altitudes and that some mitigation of the remaining interference should be possible with appropriate use of cognitive counter measures (especially interference aware radio resource management). There appears to be adequate bandwidth available for mitigating the FS interference, which in very encouraging.

Due to the movement of the terminal of interest causing a time variant element to the interference conditions the interference driven resource management mechanisms need to be much more dynamic than that required for the previously reported fixed location FSS cases.

Case 2 Maritime ESOMP

For case 2 when the ESOMP is maritime in nature then Scenarios A and B are simply extended in areas in the sea. In the case of Scenario B, with the ESOMPS ship operating in the 17.7 to 19.7 GHz band, example results are given for a ship sailing along the English Channel as depicted in Figure 18. Figure 19 presents the indicative interference field for total FS interfering bandwidth for a threshold of -154.5 dBW/MHz at any given point. The path of interest is shown in red and represents a journey of length 284 km.

In this example, it assumes that the ship is sailing in the interested path on the sea along the English Channel and that there are many UK based FS microwave stations on the land that may cause interference to the ship-borne ESOMP. Each FS link has its own frequency, bandwidth and the value of interference levels at each particular test point. The shipborne terminal is assumed to be pointing to a satellite located at 13 degrees East longitude with the ship receiving signals from the satellite. Antenna patterns, full terrain based propagation models and path losses were all taken into account for this calculation.

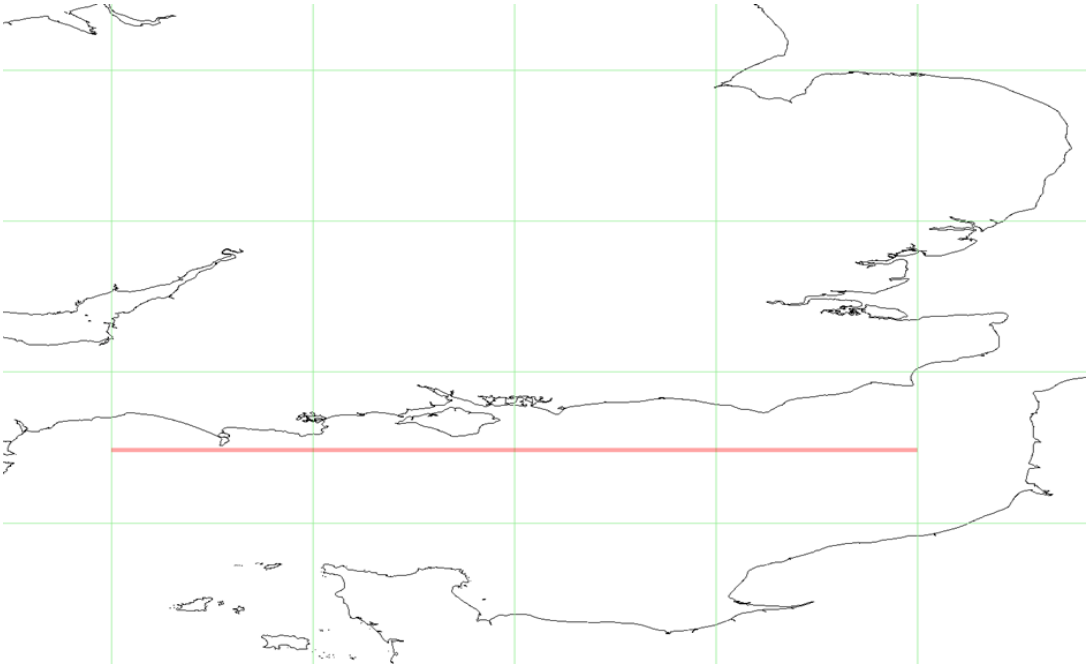


Figure 18. Example ESOMP vessel movement

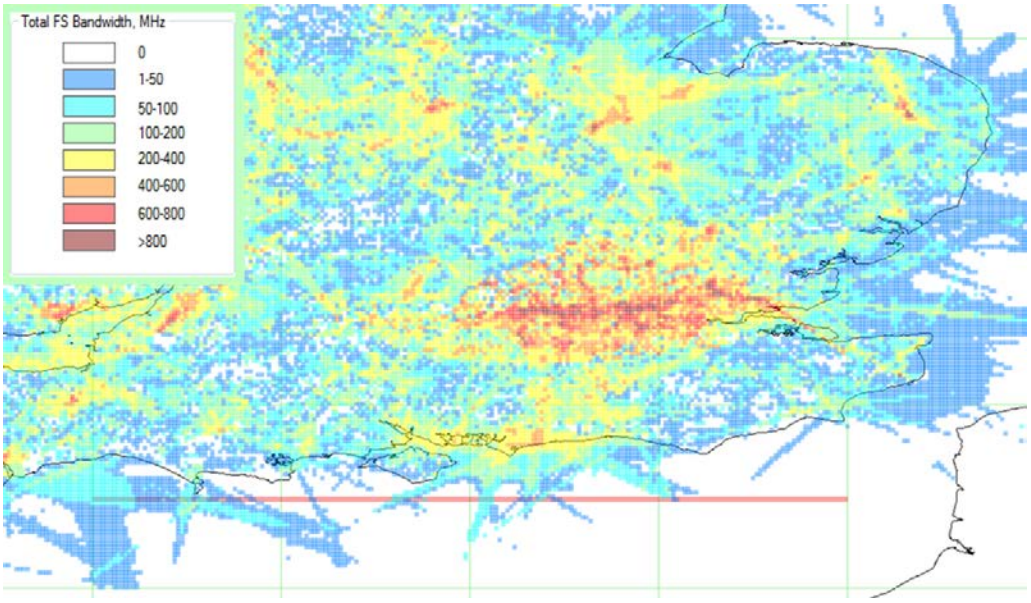


Figure 19. Indicative interference field for the maritime example

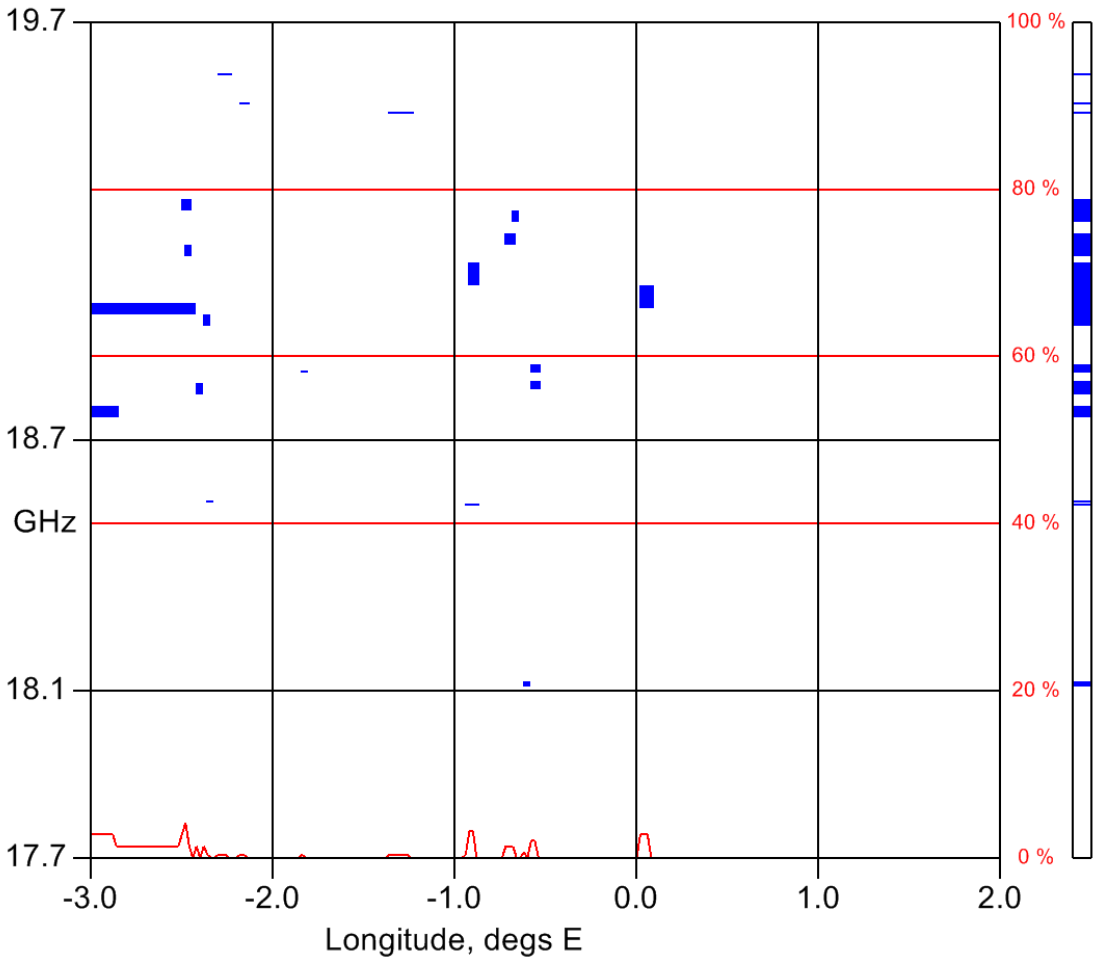


Figure 20. FS interference spectral plot along maritime path of interest (-155 dBW/MHz threshold)

The red line, associated with the right hand scale, in Figure 20 represents the percentage of the 2 GHz of available bandwidth that interference from the FS exceeds the given interference threshold along the example maritime path. It should be noted that such total bandwidth may be similar at different locations but be comprised of several carriers at quite different frequencies. Figure 20 indicates the spectral nature of the interfering carriers along the ships path. The presentation format is the same as that outlined for the Aeronautical ESOMP in Figure 17 (c), Figure 17 (d) and Figure 17 (e).

It can be seen that significant parts of the 17.7 to 19.7 GHz band are available for use with ESOMPs when the only mitigation approach required is spectrum management within the satellite resource allocation algorithms. The dynamic nature of the required interference driven resource management mechanism is not as fast or critical in the maritime case compared to the aeronautical one due to the lower speed of movement of the ESOMP.

ESOMP Summary

Example results for both shipborne and airborne operations have been presented indicating that with appropriate use of cognitive counter measures (especially interference aware radio resource management) there is adequate bandwidth available for mitigating the FS interference, which in very encouraging.

Due to the movement of the terminal of interest causing a time variant element to the interference conditions the interference driven resource management mechanisms need to be much more dynamic than that required for the previously reported FSS cases.

V. Regulations and Standards

As has already been discussed in the paper, acceptance by the Regulatory regime is crucial in order to access the additional spectrum in the shared band. In parallel to the

work being conducted in the CoRaSat project, work has been on going in the CEPT groups SE-40 and FM 44 on sharing in the band 17.7 to 19.7 GHz. Consultation documents have been issued in midyear 2015 by these committees. Some regulators have started to put their FS databases on the World Wide Web and it is hoped that more may follow. Others have been more reluctant to release information. SE-40 has been investigating software that could be made available to the national regulators so that they can interface with their databases and produce interference maps. The latter could then be made available to satellite operators and ground segment equipment providers to interface with resource allocation software at the gateways. Mechanisms are being sorted out amongst the regulators to allow the database systems with resource allocation to go ahead.

Within the standards arena a technical report based on the work of CoRaSat "ETSI .TR.103.263.v1.2- Electromagnetic compatibility and Radio spectrum Matters (ERM); System Reference document (SRdoc);Cognitive radio techniques for Satellite Communications operating in Ka-band" has already been published and is going through the updating process in 2015. Thus, manufacturers are engaged and buying into the use of systems as described in this paper.

With regard to ESOMPs, European Regulations are being put in place to permit the harmonized use, free circulation and exemption from individual licensing of ESOMPs within the frequency bands of interest [23].

VI. Conclusions

To meet future broadband access targets, in this paper we have described how the increased spectrum opportunities can be exploited by the proposed database approach together with interference mitigation techniques. We have demonstrated that in 17.3-17.7 GHz spectrum band 400 MHz of additional bandwidth is available across 98% of the UK, which houses the most dense BSS network in Europe, and similar results were obtained for Luxembourg. The evaluation needs to be repeated in other EU countries, but a similar if not better performance would be expected due to the lower density of BSS. If the FSS is required to be closer to a BSS, then cognitive means can be used to mitigate the interference.

We have also explored the availability of the 2 GHz of spectrum between 17.7 and 19.7 GHz (downlink) and the results have shown that the number of actual interfering FS links are limited due to terrain diffraction effects so that at a particular location substantial parts of the 17.7 to 19.7 GHz are available, but not the same frequencies at all locations. This indicates that a database interfaced with a resource allocation scheme could give access to the increased spectrum. This was demonstrated for the UK but needs to be validated in other European countries. In the case of the uplink 27.5-29.5 GHz The situation also looks very promising. Regulators and standards bodies are engaged with the sharing techniques and are now taking them forward to realization.

Our studies also indicate that the interference is also not too limiting for airborne and maritime ESOMPs operations in terms of interference from FS links into ESOMPs but a more dynamic interference aware resource management system will be required.

We would like to note that the work presented in this paper represents just part of the work conducted in the EU Project CoRaSat. In particular we note that other colleagues in the project have evaluated spectrum sensing using a novel SINR scheme, which can be used instead of a database or to augment its performance. In addition other colleagues have evaluated carrier resource allocation schemes to be used with databases. Finally, there has been a laboratory demonstration of all of these techniques, database, spectrum sensing and resource allocation with real satellite terminal equipment.

Acknowledgment

The authors would like to acknowledge the EU FP7 project CoRaSat, which has supported the work herein.

References

[1] W.Tang, P.Thompson and B.Evans, "A database approach to extending the usable Ka band spectrum for FSS satellite systems": SPACOMM -2015

[2] European Commission, "A Digital Agenda for Europe, FCC 02-155," European Commission COM 245, Brussels, Tech. Rep., 2010.

[3] "EU FP7 Project BATS," Available: http://www.batsproject.eu/ as at 1/11/2015.

[4] H. Fenech, E. Lance, and M. Kalama, "KA-SAT and the way forward," Ka-band Conference, Palermo, Italy, 2011.

[5] "Highest-capacity communications satellite," http://www.guinnessworldrecords.com/records-1/highest-capacity- communications-satellite/ as at 1/11/2015.

[6] P. Thompson, B. Evans, L. Castenet, M. Bousquet, and T. Mathiopoulos, "Concepts and technologies for a terabit/s satellite," in Proceedings of SPACOMM-2011, April 2011, Budapest, Hungary.

[7] A. Kyrgiazos, B. Evans, P. Thompson, P. T. Mathiopoulos, and S. Papaharalabos, "A terabit/second satellite system for european broadband access: a feasibility study," International Journal of Satellite Communications and Networking, vol. 32, no. 2, 2014, pp. 63–92.

[8] "The European conference of postal and telecommunications adminis trations," avaliable: http://www.cept.org/cept. As at 1/11/2015.

[9] "EU FP7 Project CoRaSat," available: http://www.ict-corasat.eu. As at 1/11/2015.

[10] K. Liolis, G. Schlueter, J. Krause, F. Zimmer, L. Combelles, J. Grotz, S. Chatzinotas, B. Evans, A. Guidotti, D. Tarchi, and A. Vanelli-Coralli, "Cognitive radio scenarios for satellite communications: The CoRaSat approach," in Future Network and Mobile Summit (FutureNetworkSummit), 2013, July 2013, pp. 1–10.

[11] S. Maleki, S. Chatzinotas, B. Evans, K. Liolis, J. Grotz, A. Vanelli- Coralli, and N. Chuberre, "Cognitive spectrum utilization in ka band multi-beam satellite communications," IEEE Communication Magazine, Vol 53, Issue 3, March 2015, pp 24-29..

[12] "Cognitive radio techniques for satellite communications operating in Ka band", ETSI System Reference document, available: http://webapp.etsi.org. As at 1/11/2015.

[13] "Standardization of TV white space systems," available: http://www.ict- crsi.eu/index.php/standardization-streams/tv-white-spaces. As at 1/11/2015.

[14] "Recommendation P.452-15: Prediction procedure for the evaluation of interference between stations on the surface of the earth at frequencies above about 0.1 GHz," International Telecommunication Union, 2013.

[15] "Methods for the determination of the coordination area around an earth station in frequency bands between 100 MHz and 105 GHz," ITU Radio Regulation Appendix 7, International Telecommunication Union, 2012.

[16] "Recommendation F.758-5: System parameters and considerations in the development of criteria for sharing or compatibility between digital fixed wireless systems in the fixed service and systems in other services and other sources of interference," International Telecommunication Union, 2012.

[17] "Recommendation ITU-R S.465: Reference radiation pattern for earth station antennas in the fixed- satellite service for use in coordination and interference assessment in the frequency range from 2 to 31 GHz," International Telecommunication Union, 2010.

[18] "Recommendation ITU-R S.580: Radiation diagrams for use as design objectives for antennas of earth stations operating with geostationary satellites" International Telecommunication Union, 2004.

[19] ITU-R Terrestrial BRIFIC, available: http://www.itu.int/ITU-R/index.asp?category=terrestrial&rlink=terrestrial-%brific&lang=en. As at 1/11/2015.

[20] S. Sharma, E. Lagunas, S.Maleki, S. Chatzinotas, J. Goetz, J. Krause, and B. Ottersten "Resource allocation for cognitive satellite communications in Ka band (17.7-19.7GHz)," IEEE - ICC 2015 Workshop CogRaN-Sat.

[21] A. Mohamed, M. Lopez-Benitez, and B. Evans, "Ka band satellite terrestrial co-existence: A statistical modelling approach," in Proceedings of 20th Ka and Broadband Communications, Navigation and Earth Observation Conference, October 2014.

[22] ECC Report 184 " The Use of Earth Stations on Mobile Platforms Operating with GSO Satellite Networks in the Frequency Range 17.3-20.2 GHz and 27.5-30.0GHz", approved February 2013, http://www.erodocdb.dk/docs/doc98/official/pdf/ECCRep184.pdf as at 1/11/2015.

[23] ECC Decision(13)01 "The harmonised use, free circulation and exemption from individual licensing of Earth Station On Mobile Platforms (ESOMPs) within the frequency bands 17.3-20.2 GHz and 27.5-30.0 GHz" Approved 8 March 2013, http://www.erodocdb.dk/does/doc98/official/pdf/ECCDec1301.pdf as at 1/11/2015.

Using SC-FDMA Waveform in a Shared Spectrum Context with High Efficiency

Benjamin Ros, Sonia Cazalens, Christelle Boustie
Satellite telecommunications systems department
CNES (French Space Agency)
18 avenue E. Belin
31400 Toulouse, France
e-mail: {benjamin.ros, sonia.cazalens, christelle.boustie}@cnes.fr

Xavier Fouchet
SILICOM
12 rue Caulet
31300 Toulouse, France
e-mail: xfouchet@silicom.fr

***Abstract*—The trend for future communication systems is the efficient sharing of frequency bands on one hand (Cognitive Radio) and the emergence of satellite/terrestrial integrated systems, which is a topic of interest in 5G research process, on the other hand. All this is in order to optimize the management of the spectrum. Single-Carrier Frequency Division Multiple Access (SC-FDMA) is well suited waveform for satellite link thanks to its low Peak-to-Average Power Ratio (PAPR) level. Its Orthogonal Frequency-Division Multiplexing (OFDM) based transmitter architecture, allowing a granular frequency access, reinforces its ability to perform well in shared spectrum context. More generally, SC-FDMA waveform seems to be a good candidate for both fixed-mobile convergence and satellite-terrestrial hybridization. This article aims to demonstrate two main points. Firstly, it will be proved that SC-FDMA waveform, even with holes enabling a dynamic use of the spectrum, is well suited to cope with satellite payload impairments, such as nonlinear amplifier and Input Multiplexer. Secondly, it will be assessed that doing frequency holes does not degrade so much air interface performance by using surboost of remaining carriers. As a whole, in this paper, the relevance of SC-FDMA in a shared spectrum context is demonstrated.**

Keywords-shared spectrum; integrated satellite terrestrial system; intentional jammer; (Extended and Weighted) Single-Carrier Frequency Division Multiple Access; amplifier non linearity; Input Multiplexer.

I. INTRODUCTION

Preliminary study results on insertion of frequency holes in Single-Carrier Frequency Division Multiple Access (SC-FDMA) waveform for a satellite system have been published in [1]. Complementary results are presented here.

Multi-carrier waveform is not historical waveform used on satellites. But, as European Union is promoting through incoming 5G integration of satellite and terrestrial components, paradigm is currently being modified [2]. Indeed, to make easier spectrum scalability, satellite should be able to choose sub-bands it uses depending on existing terrestrial systems. Moreover, in the case of integrated systems specifically, to encourage mass market terminal deployment, terminal should be able to receive the satellite signal or terrestrial one with the same chipset. This emphasizes the need for satellite to use a multi-carrier granular access, as it is already the case in terrestrial systems. The high crest factor of OFDM does not allow optimizing the efficiency of satellite amplifiers. An intermediate solution is the use of SC-FDMA. On forward link, its interest has already been demonstrated in the Digital Video Broadcasting (DVB) - Next Generation broadcasting system to Handheld (NGH) standardization process in S-band mobile system [3]. This waveform is also recommended in an International Telecommunication Union (ITU) working group for satellite International Mobile Telecommunications-Advanced (IMT-Advanced) systems [4]. More recently, interest of this waveform raised in European Telecommunications Standards Institute - Satellite Communication and Navigation (ETSI-SCN) and DVB - Second Generation Satellite Extensions (S2x) standardization process applicable to high frequency bands (Ku, Ka), especially for broadband systems.

In Section II, a brief state of the art is presented. In Section III, the considered scenarios are described. In Section IV, description of SC-FDMA waveform enabling frequency holes is done. Methodology used to analyze obtained results is introduced, with first simulation results. In Section V, performances with satellite payload impairments are discussed. In Section VI, impact of surboost on the air interface is studied before giving further work horizon in Section VII.

II. BRIEF STATE OF THE ART

In terrestrial context, a variant of OFDM known as non-contiguous OFDM (NC-OFDM) has been proposed for Cognitive Radio networks [5]. It allows the transmission of information in presence of primary users, by deactivating the subcarriers already occupied to avoid interferences. Efficient implementation of NC-OFDM transceiver has been proposed. It is based on an FFT pruning algorithm, which allows reducing the execution time [6]. Nevertheless, one major drawback of OFDM is its high Peak-to-Average Power Ratio (PAPR). SC-FDMA waveform allows reducing this PAPR, while still having the properties of frequency agility. Previous studies have demonstrated that Non-

Contiguous (NC) SC-FDMA can generally achieve better performance than NC-OFDM [1][7]. Some issues appears in the literature relative to the employment of non-contiguous multicarrier-modulation-based data transmission systems, like out of band interference, power amplification, synchronization, implementation complexity [8][9], but always in a terrestrial context.

In this paper, satellite scenarios and channels are considered: effects of satellite payload on NC-SC-FDMA are studied, as well as a mean to compensate loss of bandwidth by using power of unused subcarriers.

III. CONSIDERED SCENARIOS

Two main application cases have an interest. On the one hand, an integrated system example (the spectrum of the terrestrial and satellite components is managed by the same company) is given in Figure 1. It provides a service to nomadic or mobile devices. To optimize the spectrum usage according to the traffic, the scheduler decides if a part of the band is allocated to either satellite component or terrestrial one. On the other hand, a satellite system use frequency bands allocated to terrestrial systems provided it does not cause harmful interference. To enable good reception of satellite signal by satellite terminals, system operator may decide to null signal in the band where a certain amount of terminals are jammed, as it is depicted on Figure 2.

IV. SC-FDMA WAVEFORM MODELISATION WITH FREQUENCY HOLES

Firstly, the SC-FDMA modeling is described and waveform power fluctuations properties are analyzed. Then we explain how to insert frequency holes. The model of interference used is described. We present a method to compare the waveforms and finally, reference simulations are showed.

A. Basic SC-FDMA modeling

Because there have been several studies on SC-FDMA waveform, model will not be strictly detailed [10][11]. However, differences with OFDMA transceiver architecture will be emphasized. General SC-FDMA transceiver is summarized in Figure 3, and is described hereafter.

Firstly, interleaved and coded bits b_n are mapped into symbols c_i, $i \in [0, \dots, M-1]$. Spreading operation, specific to SC-FDMA transmission, is then done by applying an M–Discrete Fourier Transform (DFT) to the c_i symbols to get C_k symbols, $k \in [-\frac{M}{2}, \dots, \frac{M}{2} - 1]$:

$$C_k = \frac{1}{\sqrt{M}} \sum_{i=0}^{i=M-1} c_i \, e^{-j\frac{2\pi ki}{M}} \qquad (1)$$

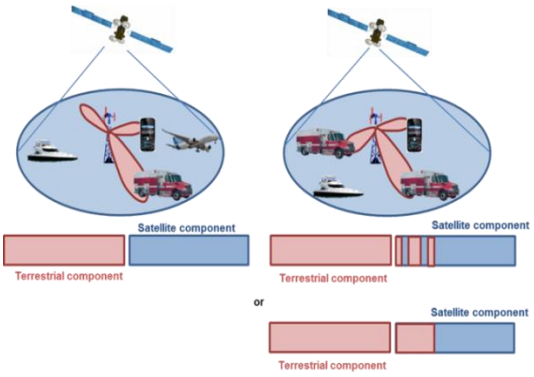


Figure 1. Example of integrated system managing efficiently its spectrum.

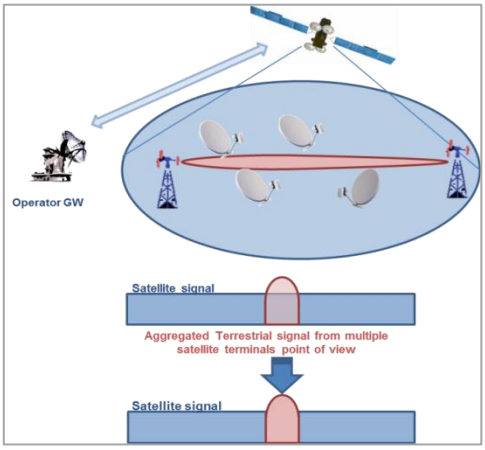


Figure 2. Example of satellite system using cognitive radio to transmit efficiently its signal.

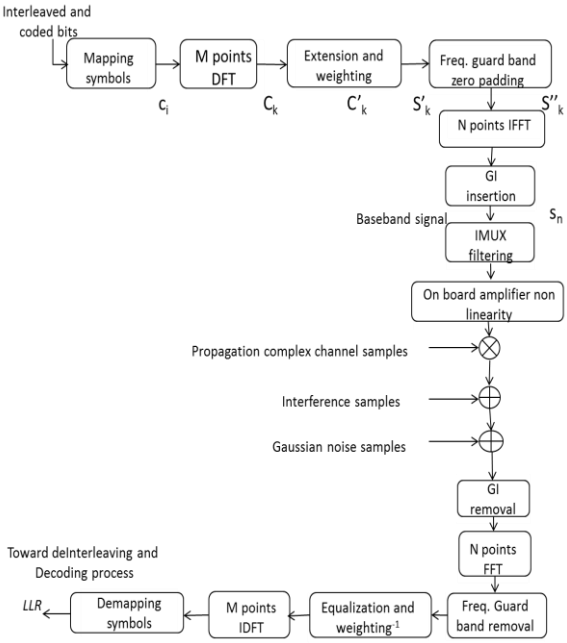


Figure 3. SC-FDMA transceiver architecture.

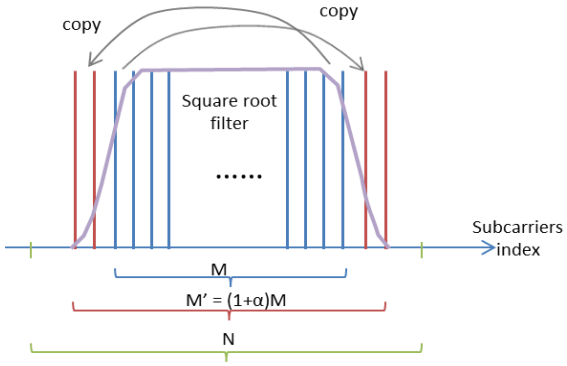


Figure 4. Extension and weighting process.

In the case of EW-SC-FDMA, as it can be seen in Figure 4, some of the edge C_k symbols are duplicated and put at the opposite side edge in the guard band. Main sizing parameter here is α, called the roll off factor. Extension process outputs M' (equals to $(1+\alpha)M$) complex symbols C_k'. Weighting process is finally applied over these subcarriers in order to have a square root raised cosine shaping [10]. It multiplies term by term $H_{SRC}(k')$ with C_k' to get S_k'.

Note that $H_{SRC}(k')$, $k' \in \left[-\frac{M'}{2}; +\frac{M'}{2}-1\right]$, is the frequency response of the square root raised cosine filter, given by:

$$H_{SRC}(k') = \begin{cases} 1, & 0 \le |k'| < \frac{(1-\alpha)}{2}M \\ \cos\left[\frac{\pi}{2\alpha M}\left(|k'| - \frac{(1-\alpha)}{2}M\right)\right], & \frac{(1-\alpha)}{2}M \le |k'| < \frac{(1+\alpha)}{2}M \end{cases} \quad (2)$$

It can be pointed out that when α equals to 0, process is strictly equivalent to SC-FDMA process. Next, unused subcarriers in the guard band are filled by zero, exactly $N - (1+\alpha)M$, to get a vector owning N complex symbols, $S_{k''}$. Note that to be compliant with OFDMA process, N should be a power of 2, which is not the case for M. Lastly, OFDMA modulation process can be done by applying an N-Inverse Fast Fourier Transform (IFFT) and inserting guard interval, getting baseband signal s_n.

For receiver considerations, the equalization should use a Minimum Mean Square Error (MMSE) frequency domain algorithm, working subcarrier by subcarrier [10][11]. Applied to SC-FDMA waveform (left hand side formula) and EW-SC-FDMA waveform (right hand side formula), estimated symbol $\widehat{x_k}$ is expressed hereafter:

$$\hat{x}_k = \frac{\hat{h}_k^*}{\left|\hat{h}_k\right|^2 + \hat{\sigma}^2} y_k \qquad \hat{x}_k = \frac{\hat{h}_k^* y_k + \hat{h}_{k+M}^* y_{k+M}}{\left|\hat{h}_k\right|^2 + \left|\hat{h}_{k+M}\right|^2 + \hat{\sigma}^2} \quad (3)$$

$\hat{h}_{k(+M)}$ are channel estimates in the useful bandwidth (M equals to 0) or extended bandwidth (M greater than 0). $y_{k(+M)}$ are received complex symbols after OFDM FFT matched filter. $\hat{\sigma}^2$ is the noise power estimate in a subcarrier bandwidth, that is to say Rs divided by the size of OFDM FFT, where Rs is I/Q sampling frequency.

Furthermore, a slightly difference that can be observed with SC-FDMA receiver compared to OFDMA one, is the way to compute the Log Likelihood Ratio (*LLR*) metrics at the demapping symbols step. Classic *LLR* formulation is reminded here:

$$LLR(b_i) = ln \frac{\sum_{x \in c_i^1} exp\left(-\frac{|I - \rho_I I_x|^2 + |Q - \rho_Q Q_x|^2}{2\sigma^2}\right)}{\sum_{x \in c_i^0} exp\left(-\frac{|I - \rho_I I_x|^2 + |Q - \rho_Q Q_x|^2}{2\sigma^2}\right)}, \quad (4)$$

where $x = I_x + jQ_x$ is a symbol of the Quadrature amplitude modulation (QAM) constellation, c_i^j represents the symbols of the constellation carrying the bit b_i when b_i equals to j, I and Q are the in phase and quadrature components of the received signal after OFDM FFT process, $\rho_{I/Q}$ is the fading on the I or Q component, $2\sigma^2$ is the Additive White Gaussian Noise (AWGN) variance, I_x and Q_x denote the reference symbols of the QAM constellation. In SC-FDMA receiver, because of the despreading process, i.e., there is a M points Inverse-DFT between equalization and demapping process, it is assumed that $\rho_{I/Q}$ can be approximated to the root mean square of the frequency channel response of the corresponding OFDMA symbol over the active subcarriers $H_c(k')$:

$$\rho_{I/Q} \approx \sqrt{\frac{\sum_{k'=-M'/2}^{\frac{M'}{2}-1} |H_C(k').H_{SRC}(k')|^2}{M}}. \quad (5)$$

B. *Waveform power fluctuations properties*

Because dealing with non-linearity effects, the study of envelope fluctuations for each waveform may help to understand further results. This is why complementary cumulated density function for instantaneous power is given for each studied waveform in Figure 5. As it can be found in literature, SC-FDMA waveform performs better than OFDMA one. Using extension can help to better decrease fluctuations. However, one shall note that this comparison is considering waveforms without frequency holes insertion.

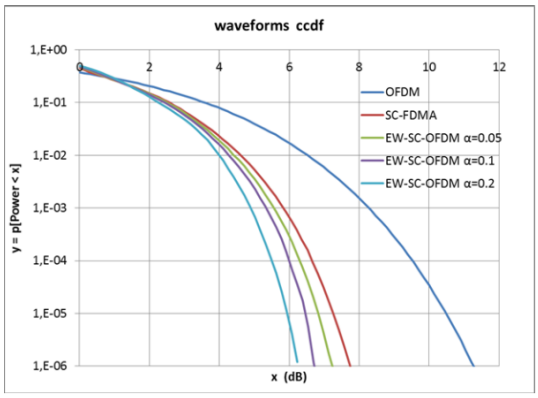


Figure 5. Complementary Cumulated Density Function of waveforms intantaneous power. 512 sub-carriers are used for simulations.

C. Frequency holes insertion

Frequency holes are inserted inside useful signal at the spreading process level. In fact, spreading applies a $L<M$ DFT over the incoming c_i symbols before filling M-L other symbols with zero, as it is drawn in Figure 6. Some vocabulary is necessary to define where frequency hole is located. β is defined as the relative bandwidth occupied by frequency holes over maximum achievable useful bandwidth:

$$\beta = \frac{M-L}{M}. \quad (6)$$

Δ is the relative shift of the frequency hole center relatively to the bandwidth center. To clarify these notations, an example is given in Figure 7.

D. Interference model

The interferer location is defined as it is done for the frequency holes, with β and Δ parameters. Generative model is quite simple: complex symbols are generated according to a normal law $\mathcal{N}(0;\beta)$, supposing power of the useful signal as unitary power when occupying all possible subcarriers. Interference symbols are applied to subcarriers as if it was an OFDMA signal. When frequency holes are defined, interferences are added exactly at the holes location.

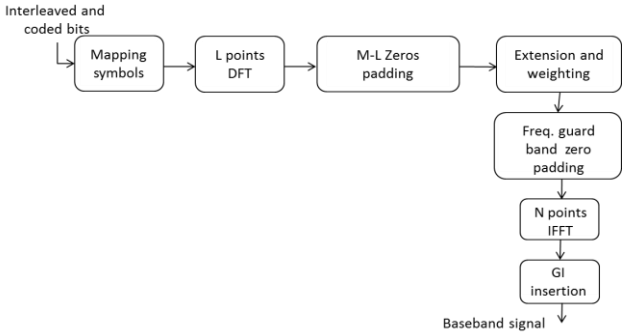


Figure 6. SC-FDMA transceiver architecture enabling frequency holes.

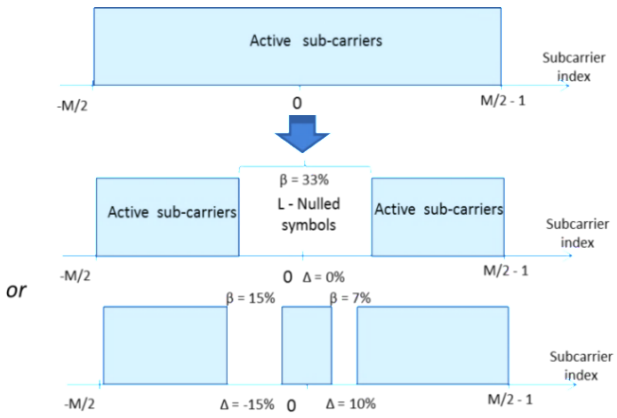


Figure 7. Position and size example of a frequency hole.

E. Comparative method analysis

We are about to compare different waveforms enabling creation of a frequency hole inside their spectrum. Here is proposed a method to compare waveforms, considering power loss, signal quality loss, and Carrier over Intermodulation ratio. *Power loss* corresponds to *OBO* value, and *Signal quality loss* equals to the performance gap considering non-linearity sub-block or not. *Total loss* metric is then defined as (all the quantities are in dB):

$$Total\ loss = Signal\ quality\ loss + Power\ loss$$

$$\begin{cases} power\ loss = |OBO| \\ signal\ quality\ loss = \left(\frac{C'+I_m}{N}\right)_{BER=10^{-5}} - \left(\frac{C}{N}\right)_{BER=10^{-5}} \end{cases} \quad (7)$$

where $\left(\frac{C}{N}\right)_{BER=10^{-5}}$ is the required signal to noise ratio in dB with ideal amplifier response at the bit error rate (*BER*) of 10^{-5}, and $\left(\frac{C'+I_m}{N}\right)_{BER=10^{-5}}$ is the required amplified signal power (pure signal plus intermodulated part) to noise ratio in dB at *BER* 10^{-5}. Both metrics are given at the receiver input location (see Figure 8). Assuming that I_m has a Gaussian behavior, pure Signal to Intermodulation power ratio $\left(\frac{C_I}{I_m}\right)$ can be derived (linear form):

$$\left(\frac{C'}{I_m}\right) = \frac{1 + \left(\frac{C'+I_m}{N}\right)_{BER=10^{-5}}}{\left(\frac{C}{N}\right)^{-1}_{BER=10^{-5}} * \left(\frac{C'+I_m}{N}\right)^{-1}_{BER=10^{-5}} - 1} \quad (8)$$

Signal to Intermodulation power ratio is an important criterion according to satellite operators, because it demonstrates the ability of the payload to work with any spectral efficiency. Indeed, a low $\left(\frac{C_I}{I_m}\right)$ ratio results in degradation of total $\left(\frac{C}{N+I}\right)$ ratio, and then limitation of spectral efficiency. As a result, performances of the waveforms will be compared at equal $\left(\frac{C'}{I_m}\right)$ ratio. It shall be pointed out that for a same spectral efficiency, $\left(\frac{C'}{I_m}\right)$ is directly linked to the signal quality loss considering (7) and (8). Thus, judicious representation may be, for each waveform, required $|OBO|$ to get *BER* = 10^{-5} vs $\left(\frac{C'}{I_m}\right)$ representation.

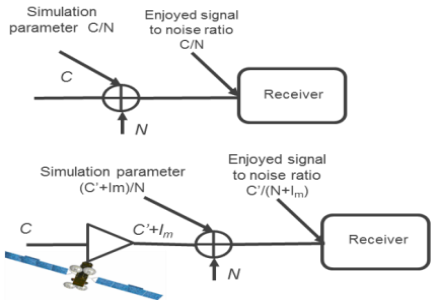


Figure 8. Receiver input signal to noise ratio, with/without non-linearity.

F. Reference simulations

The simulations are performed by using a DVB-NGH like transceiver chain [1]. DVB-NGH specifications enable SC-FDMA utilization without extension and weighting functionality. Besides, frequency hole, IMUX filtering, on-board nonlinear amplifier and interferences had to be considered as it is depicted in Figure 9. A significant importance should be given to the way how signal to noise ratios are computed when dealing with frequency holes. As power at output of the amplifier is unitary in simulations, it was chosen to decrease signal power with the same bandwidth reduction ratio. It enables comparing results at the same Es/N0. This decrease of signal power will be cancelled for simulations demonstrating of the surboosting effect benefits. Lastly, the simulation parameters are summarized in Table I.

V. STUDY OF SATELLITE PAYLOAD IMPAIRMENTS

Satellite payload impairments models usually include three main origins: nonlinear amplifier, selective input filters and phase noise.

For the current work, non-linearity effects and input filtering effects are considered separately. Phase noise is not considered because it is well known that effect is very weak over a Quadrature Phase Shift Keying (QPSK) modulation combined with current stability specifications over satellite local oscillators.

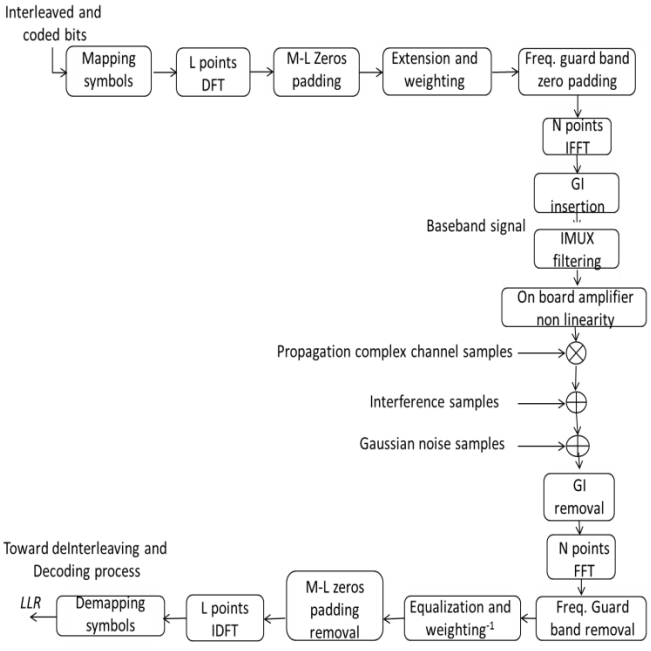


Figure 9. Simulation Chain, including SC-FDMA, frequency hole, interference and non linearity functionalities.

Input filters distortion may have an effect over the performances, but as the three studied waveforms (OFDMA, SC-FDMA, EW-SC-FDMA) have a subcarrier equalization process, it can be expected that the effect will be quite close in all cases, while not negligible. This is the reason why it was decided to focus on non-linearity effects, knowing there are envelope fluctuations differences between waveforms.

In the simulation tool, IMUX filtering is first applied to the baseband signal, before being processed by the nonlinear conversion. Signal passed through satellite amplifier, having a unitary power, is applied to propagation channel block, as it is shown in Figure 10.

TABLE I. MAIN PARAMETERS FOR PERFORMED SIMULATIONS

Parameter name	Value
Satellite signal Bandwidth	15 MHz
Sampling frequency Rs	120/7 MHz
I/Q sample duration	58.33 ns
Modulation and coding	QPSK 2/3 LDPC + BCH encoder 16200 bits codeword
Max active subcarriers (M)	426
OFDM FFT size (N)	512
OFDM Guard interval	1/16
Total OFDM symbol duration	31.73 µs
SC-FDMA	α = 0 (SC-FDMA) or α = 5% (EW-SC-FDMA)
Frequency hole insertion	Δ=0, β=33% when activated
Interference insertion	By default not activated. Δ=0, β=33% when activated
IMUX filtering	By default not activated
Satellite RF model	MTV by default.
Propagation channel	Ideal (AWGN) in this study

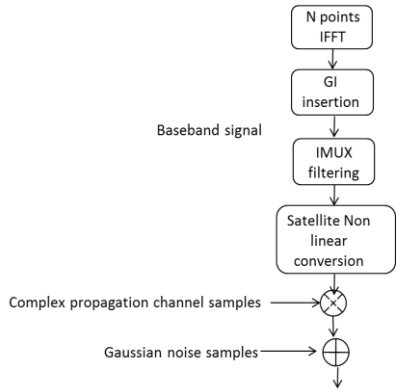


Figure 10. Insertion of non-linearity conversion and IMUX filtering in the transceiver.

A. *Nonlinear amplifier*

A typical model is chosen, called Mobile Television (MTV) model, giving Output Back Off (*OBO*) [dB] as a function of Input Back Off (*IBO*) [dB]. This model was chosen because it seemed to be quite realistic. Another model has been considered, NGH model, because it was used in DVB-NGH standardization process. As it seemed to be less realistic (especially phase characterization), it has not been used as a baseline. See on Figure 11 the non-linearity conversion curves of these amplifiers.

As a reference, *BER* simulations have been performed according to the method described previously. For a wide range of C/I_m, it appears in Figure 12 that SC-FDMA like waveforms outperform OFDMA in terms of *OBO* vs C/I_m, with a gap increasing when C/I_m is growing.

These results emphasize the fact that when no interferer is present, SC-FDMA is a good choice to be compatible with terrestrial multi-carrier legacy and to enjoy efficient use of satellite payload. For these reference simulations, the results with NGH amplifier have also been considered (see Figure 13). Conclusions about the interest of SC-FDMA are the same as with MTV amplifier, but with worse performances in required *OBO* at low C/I_m.

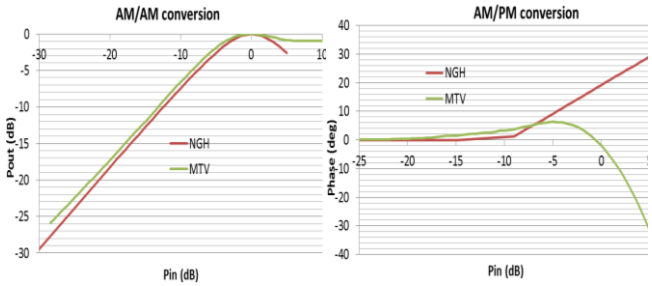


Figure 11. Non linearity conversion curves of MTV and NGH amplifiers.

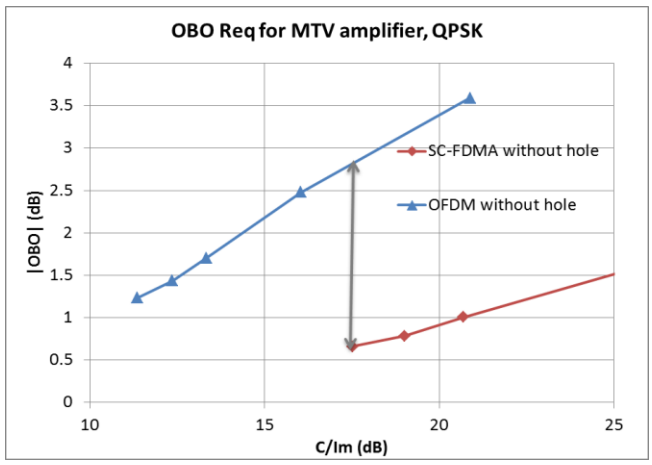


Figure 12. Comparison of waveforms in terms of OBO vs C/I_m when no frequency holes are inserted and without interference, for MTV amplifier.

Here, frequency hole was inserted with default parameters (Δ=0 and β=33%). Especially for single carrier waveforms, inserting such wide hole inside useful bandwidth may modify its fluctuations behavior. But, as it is depicted in Figure 14, single carrier waveforms performances are not so much degraded by the hole insertion, and remain quite competitive compared to OFDM, despite the large width of the hole. About OFDM results, it can be pointed out that results with and without frequency hole are quite similar. This can be explained by the multi-carrier effect of OFDM: multicarrier signal split in two parts remains a multi-carrier signal. This is the reason why in the following, only SC-FDMA waveform will be considered.

The influence of hole size was studied for SC-FDMA modulation (see Figure 15). It can be noted that for all hole sizes, between 2 and 50% width, a degradation of 0.5 dB on average has been observed.

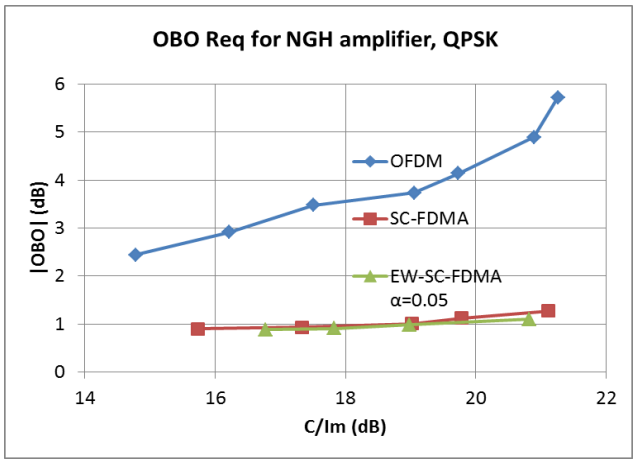


Figure 13. Comparison of waveforms in terms of OBO vs C/Im when no frequency holes are inserted and without interference, for NGH amplifier.

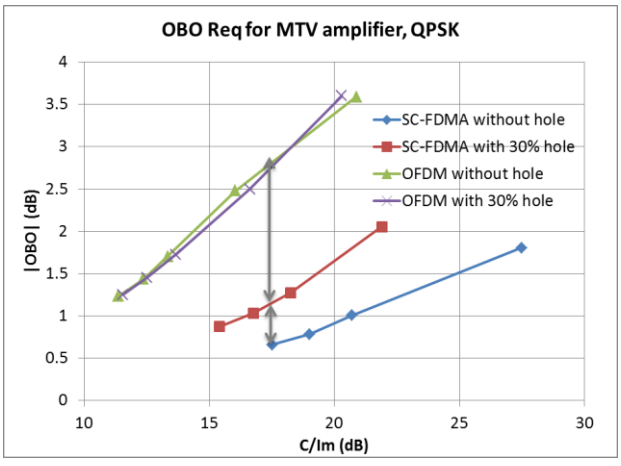


Figure 14. Comparison of waveforms in terms of OBO vs C/I_m when frequency holes are inserted.

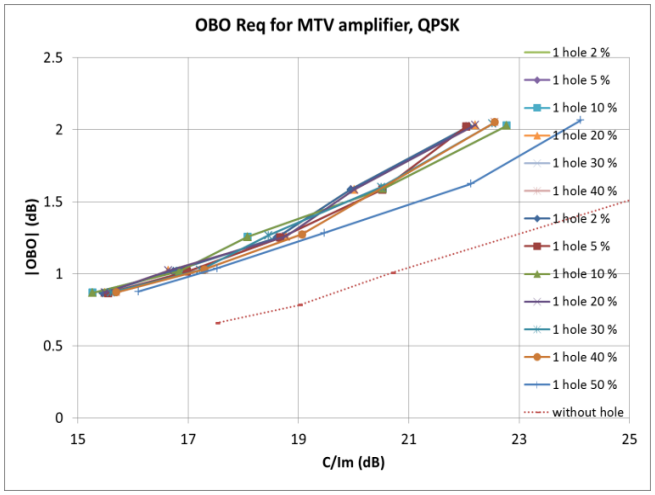


Figure 15. SC-FDMA performances in terms of OBO vs C/I_m with one frequency hole.

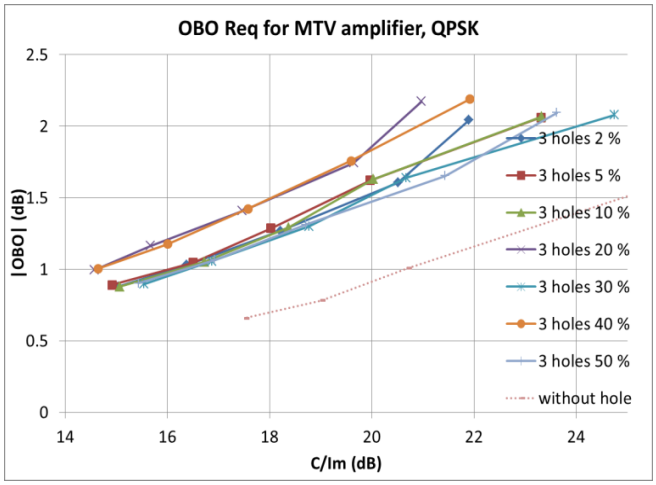


Figure 16. SC-FDMA performances in terms of OBO vs C/I_m with 3 frequency holes.

For integrated systems or satellite systems, which share frequency bands or have to deal with jammers, it could happen that several holes would be needed to avoid interference effects. Therefore, simulations have been performed with 3 holes, centered in $\Delta = [-27\%, 0\%, +27\%]$. Results can be seen in Figure 16, where percentile value means the total width of the frequency holes relatively to the useful bandwidth. Two sets of curves can be observed; first pack, closer than without hole performance curve, shows around 0.5 dB *OBO* degradation compared to without hole. This degradation is almost the same as single hole performance for SC-FDMA. For 3 holes, some combinations of width and holes positions make appear an additional degradation. Indeed, intermodulation power spectrum density is not similar to white Gaussian noise, which is frequency flat. Therefore, when useful signal is in front of intermodulation peak, it brings a bit more degradation.

B. *Input Multiplexer*

Input Multiplexer (IMUX) filter templates are given in Figure 17. Two filters are considered for the study, having a different cut-off frequency: first one has its 3 dB cutting-off frequency at 0.42 sampling frequency, where the second one is cutting-off at 0.5 sampling frequency. Both filters have a group delay greater than 800 ns, corresponding to more than thirteen times the *I/Q* symbol duration (see Table I).

In a satellite payload, IMUX filter is designed to split different incoming channels before amplifying them separately or by group. Because it is very important to filter noise and adjacent channels contribution, IMUX filters are usually designed with a margin on the channel bandwidth, in order to ensure that signal is not decreased at band edges, and to prevent group delay interferences effects at band edges also. This is the reason why two filters are considered for the simulations: aim is to demonstrate that SC-FDMA waveform enables extending signal bandwidth up to filters shoulders without losses, thanks to cyclic prefix adding up signal replicas.

As it is written in Table I, number of active OFDM subcarriers is 426, whereas OFDM FFT size is 512. It brings that occupied bandwidth relatively to *I/Q* symbol rate Rs is 0.84. That is to say signal is located between +/- 0.42 Rs. Thus, when using 0_5 filter as an IMUX, only group delay may have an impact on performances because filter has very weak fluctuations in the useful frequency band. Besides, using 0_42 filter, band edges of the signal are faded by the filter, so that it would be expected an additional degradation.

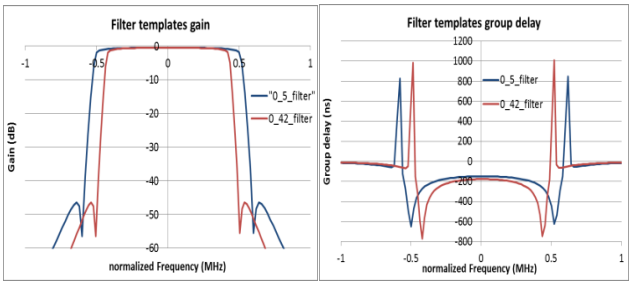


Figure 17. IMUX filter templates models : 0_42_filter and 0_5 filter.

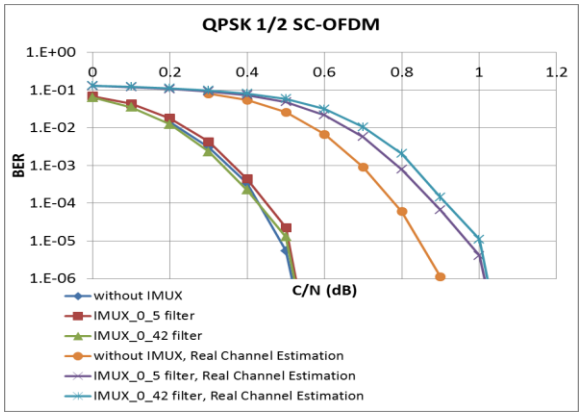


Figure 18. IMUX filter effect over SC-FDMA performances.

In this case study, simulations have been performed with and without real channel estimation, in order to assess real impact of IMUX filtering on receiver performances (see Figure 18). Results in QPSK ½ show that with perfect channel estimation, both filters have no impact over performances, whereas there is 0.1 dB degradation in real channel estimation, with respect to no IMUX insertion. This result demonstrates that multicarrier waveform helps filling at its best available bandwidth.

C. *Effects of interferences*

Interferences are inserted with Δ=0 and β=33% parameters. In addition, NGH amplifier was used for these simulations. Modulation and coding scheme is QPSK ½. I_0, the power spectrum density of interferer is the same as C_0, the power spectrum density of the signal.

That leads to relatively weak level of interferers but with quite large bandwidth (1/3 of achievable useful bandwidth). With no real surprise, it is first checked that inserting hole in any of the three frequency granular access waveforms results in a negligible degradation (see Figures 19, 20 and 21), because only secondary lobes of interferences are captured by active useful subcarriers when frequency hole is inserted.

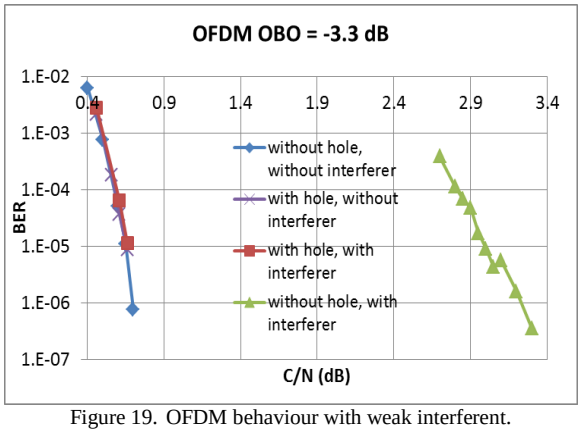


Figure 19. OFDM behaviour with weak interferent.

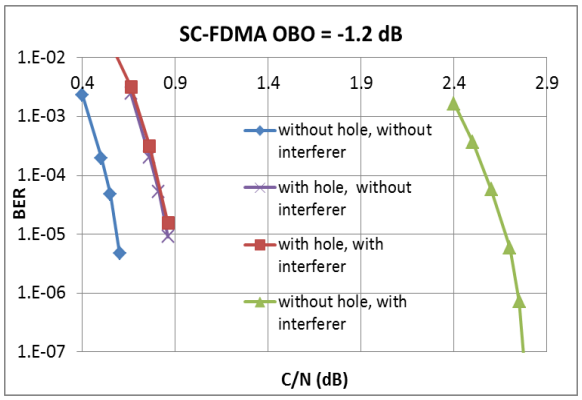


Figure 20. SC-FDMA behaviour with weak interferent.

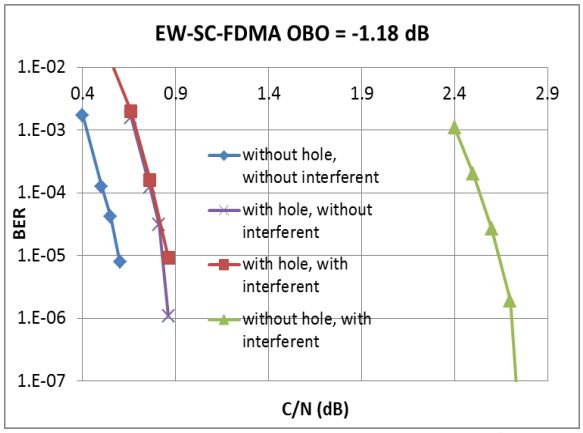


Figure 21. EW-SC-FDMA behaviour with weak interferent.

Besides, no creating hole when interferer is present is showing quite degraded results, even if the interferer has a weak level. Reader shall point out that weak interferer was chosen to visualize the degradation on the same curve. But, from a system point of view, power spectrum density of a terrestrial interferer would be greater than satellite useful signal one, and would emphasize the need for such frequency hole in the waveform.

VI. IMPACT OF SURBOOST ON PERFORMANCES

For this case study, satellite amplifier is providing constant power, whether or not frequency holes are present. Therefore, when a frequency band is unused, delivered power spreads in remaining useful sub-carriers. It creates a surboost of ratio η compared to baseline transmission:

$$\eta = \frac{BW_N}{BW_N - BW_H}, \quad (9)$$

BW_N is the useful nominal bandwidth, BW_H is the bandwidth corresponding to frequency hole.

The aim of this section is to demonstrate that surboost can compensate, from a bit rate point of view, useful bandwidth decrease by increasing spectral efficiency of the transmission. The steps of the process are given below. First, for different couples of modulation and coding schemes (MODCOD), compute frequency hole width in order to compensate, on a linear channel with a surboost η, required E_S/N_0 rising due to the changing of MODCOD (see Table II) [12]. Secondly, considering 3 frequency holes, searching *OBO* working point to reach *OBO* near from 17 dB, for both MODCOD of a couple. Lastly, compare for both MODCOD required Es/N_0 and associated bit rate.

Results are presented in Figure 22. The bit rate is computed considering MODCOD efficiency and overhead item such as pilot and frame building insertion, guard interval and a useful bandwidth of 14.2 MHz, according to Table I parameters [3].

TABLE II. MODULATION AND CODING SCHEMES COUPLES FOR SURBOOSTING EFFECT DEMONSTRATION

MODCOD 1, without hole and surboost	MODCOD 2, with hole and surboost	Frequency hole cumulated width (1-1/η)
QPSK ½ (4/9 real)	QPSK 3/5	32.4%
16QAM 4/5	16QAM 5/6	11.5%

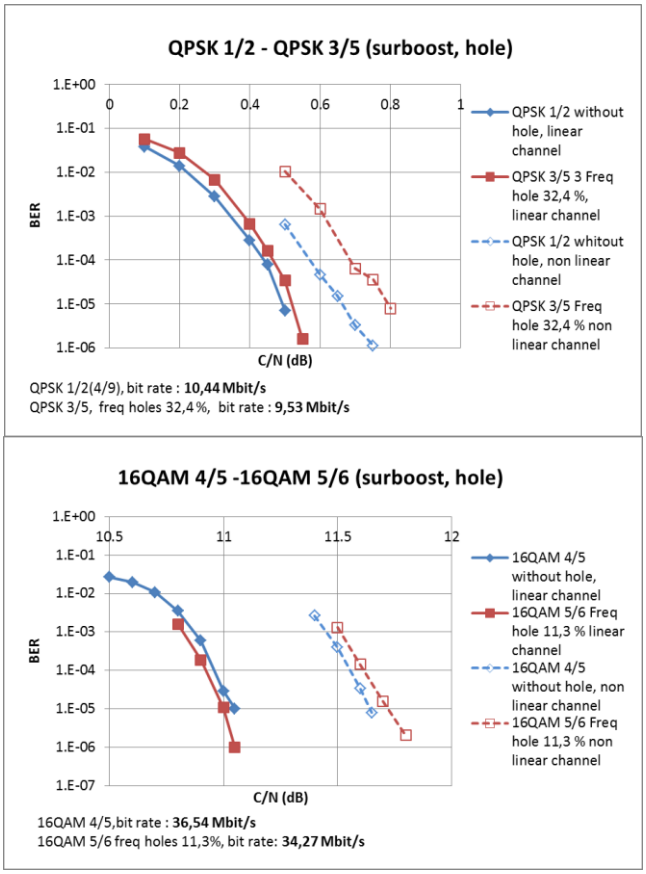


Figure 22. Using of the surboost to balance frequency hole insertion for two MODCOD couples.

Full line curves show that for linear channel, frequency hole width has been correctly chosen to balance E_S/N_0 requirement difference between two MODCOD. In dotted line, the same simulations over non-linear channel are performed. Lastly, physical layer input bit rate for two MODCOD show that the loss of bandwidth is worse than the loss of the bit rate thanks to the spectral efficiency increase.

VII. CONCLUSION AND FUTURE WORK

At a time where, on one hand, there is a need to make easier spectrum scalability and on the other hand, manufacturers are working to maximize the payload efficiency, a SC-FDMA based solution was introduced to address this issue. In this paper, it has been shown that SC-FDMA waveform fluctuations enable using satellite payload in an efficient way, and that creating frequency hole do not degrade so much air interface performances; SC-FDMA waveform remains in all cases relevant compared to OFDMA. Thus, using this waveform would take advantage of both frequency scalability and payload efficiency, while offering a solution at physical layer level for dynamic spectral resource sharing systems.

For further work, others topics should be processed. Standard modification impact of the introduction of frequency holes in the signal appears to be the first subject. Next, an accurate modeling of interferer would help to prove that frequency holes are enough to avoid interference effects. More generally, a cooperation with a satellite system manufacturer would be needed to assess the impact of using SC-FDMA at onboard level (amplifiers working point, signal routing) and ground level (terminal design, resource access methods).

ACKNOWLEDGMENT

The authors thank the Satellite Mobile Innovative Laboratory and Engineering project of French Space Agency allowing applicative research on such waveforms.

REFERENCES

[1] B. Ros, X. Fouchet, S. Cazalens, and C. Boustie, “SC-FDMA Waveform Enabling Frequency Holes in a Shared Spectrum Context,” IARIA SPACOMM 2015 : The Seventh International Conference on Advances in Satellite and Space Communications, Barcelona, Spain, April 2015, pp. 1-6, ISBN: 978-1-61208-397-1.

[2] 5G-Private Public Partnership project homepage. [Online]. Available from: <http://5g-ppp.eu/etp/> 2015.11.06

[3] ETSI, “Digital Video Broadcasting (DVB); Next Generation broadcasting system to Handheld, physical layer specification (DVB-NGH),” EN 303 105 V1.1.1, May 2013.

[4] ITU-R, Recommendation M.2047-0, “Detailed specifications of the satellite radio interfaces of International Mobile Telecommunications-Advanced (IMT-advanced),” December 2013. [Online]. Available from: <http://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2047-0-201312-I!!PDF-E.pdf> 2015.11.06

[5] H. Tang, “Some physical layer issues of wide-band cognitive radio systems,” 2005 First IEEE International Symposium on New Frontiers in Dynamic Spectrum Access Networks, Baltimore, USA, Nov. 2005, pp.151-159, ISBN: 1-4244-0013-9.

[6] R. Rajbanshi, A.M. Wyglinski, and G. J. Minden, “An efficient implementation of NC-OFDM transceivers for cognitive radios,” 2006 First International Conference on Cognitive Radio Oriented Wireless Networks and Communications, Mykonos Island, June 2006, pp. 1-5, ISBN: 1-4244-0381-2.

[7] H. Gao, "Comparison of SC-FDMA and NC-OFDM schemes for cognitive radio networks," 2010 Second International Conference on Computational Intelligence and Natural Computing, Wuhan, Sept. 2010, ISBN: 978-1-4244-7705-0.

[8] H. Bogucka, A.M. Wyglinski, S. Pagadarai, and A. Kliks, "Spectrally agile multicarrier waveforms for opportunistic wireless access," IEEE Communications Magazine Vol. 49, Issue 6, June 2011, pp. 108-115, DOI: 10.1109/MCOM.2011.5783994.

[9] H. Bogucka, P. Kryszkiewicz, and A. Kliks, "Dynamic Spectrum Aggregation for Future 5G Communications," IEEE Communications Magazine Vol. 53, Issue 5, May 2015, pp. 35-43, DOI: 10.1109/MCOM.2015.7105639.

[10] S. Okuyama, K. Takeda, and F. Adashi, "MMSE frequency-domain equalization using spectrum combining for Nyquist filtered broadband," 2010 IEEE 71st Vehicular Technology Conference, Taipei, Taiwan, May 2010, pp. 1-5, ISBN: 978-1-4244-2518-1.

[11] H. Kobayashi, T. Fukuharal, H. Yuant, and Y. Takeuchi, "Proposal of single carrier OFDM technique with adaptive modulation method," The 57th IEEE Semiannual Vehicular Technology Conference, April 2003, pp. 1915-1919, vol. 3, ISBN: 0-7803-7757-5.

[12] ETSI, "Digital Video Broadcasting (DVB); Implementation guidelines for a second generation digital terrestrial television broadcasting system (DVB-T2)," TS 102 831 V1.2.1, August 2012.

Ensuring Radio Frequency Compatibility (RFC) on-Board a Satellite by Early Analysis and Efficient Methods for Field Prediction

Jens Timmermann
Electrical Systems (TSPET32)
Airbus DS GmbH
Friedrichshafen, Germany
email: Jens.J.Timmermann@airbus.com

Christian Imhof
Electrical Systems (TSPET32)
Airbus DS GmbH
Friedrichshafen, Germany
email: Christian.Imhof@airbus.com

Dieter Lebherz
Electrical Systems (TSPET32)
Airbus DS GmbH
Friedrichshafen, Germany
email: Dieter.Lebherz@airbus.com

Jörg Lange
Electrical Systems (TSPET32)
Airbus DS GmbH
Friedrichshafen, Germany
email: Joerg.Lange@airbus.com

***Abstract*—Transmitters (=Tx) on-board a satellite generate an electromagnetic environment with potential impact on victim receivers (=Rx, e.g., instruments) placed nearby. Ensuring Radio Frequency Compatibility (RFC) on-board a satellite is hence an important point to be considered during satellite design and requires an optimized satellite configuration. This contribution concentrates on RFC issues in practical satellite design by considering the future MetOp-SG meteorological satellites: First, an overview is given summarizing the various transmitters and instrument receivers on-board the satellites. Then, the fundamentals of RFC analysis are presented showing the method how to compute the coupling factor between a Tx and a victim Rx. To improve the decoupling, MetOp-SG satellites are housing dedicated baffles between Tx and Rx antennas. Therefore, the contribution finally studies in detail the signal attenuation caused by a baffle by comparing two methods: field simulation and an extended knife-edge diffraction theory. By combining both methods, the overall engineering and computation effort to optimize the baffle design is minimized.**

Keywords- MetOp-SG; Radio Frequency Compatibility; coupling factor; knife-edge diffraction; baffle attenuation.

I. INTRODUCTION

Earth observation satellites typically house a variety of transmitters and very sensitive instruments receivers. Hereby, the signal is transmitted / received via dedicated antennas. Instruments may, e.g., sense the Earth atmosphere while the collected data is transmitted towards Ground by the on-board Tx antennas. It has to be ensured that the instrument receivers work properly in the electromagnetic environment generated by on-board Tx antennas. This means that the remaining signal at a victim receiver has to be below a specified value. As the dimension of a satellite is in the order of only a few meters, the distance between Tx and Rx antennas is quite small which makes it challenging to achieve Radio Frequency Compatibility (RFC) on-board a satellite.

Therefore, the configuration of a satellite has to be optimized w.r.t. RFC, which means that the positions and the orientations of Tx and Rx antennas play a significant role. Even in an early project phase, this aspect has to be considered to minimize the need for configuration changes in a later project phase. The approach is hence to define a preliminary configuration and to run an RFC analysis which investigates the coupling between critical Tx and Rx combinations. In an early project phase, the unintended signal at a victim receiver shall be well below (typically 20 dB) the maximum acceptable value, whereas the difference is called RFC margin. On the other side, a satellite configuration will not only be optimized w.r.t. RFC. Other aspects (such as center of mass, minimization of harness length etc.) have to be taken into account and will lead to some configuration changes. In the end, a compromise will be required ensuring positive margins in all considered disciplines.

As the optimization exercise is typically not finished in an early project phase, the approach is to run an RFC analysis based on a preliminary satellite configuration and to aim for high margins. After the global optimization exercise, the remaining RFC margins may be lower, typically above 6 dB and thus still fulfilling the needs.

This contribution is an extended version of [1] and considers RFC aspects for the future MetOp-SG satellites:

The European MetOp meteorological satellites currently in orbit will be replaced after 2020 by follow-on satellites with advanced instrumentation. MetOp-SG will ensure observations until approximately 2040 [2]. After successful finalization of ESA Phase A/B1 study by Airbus Defence and Space, the company has been nominated by EUMETSAT / ESA as prime contractor for the provision of the space segment of MetOp-SG. For this purpose, two satellites (Satellite A and Satellite B) with different scientific instruments are currently under development. Each satellite houses a variety of transmitters and instrument receivers being sensitive in the RF frequency range. Hence, ensuring RFC on-board the satellite is a major challenge.

Section II of this contribution gives an overview of the different transmitters and receivers on-board the MetOp-SG satellites. To improve the decoupling between critical Tx / Rx combinations, the satellite design encompasses baffles that shade the Line of Sight (LoS) between these critical combinations.

Section III deals with the fundamentals of RFC analysis by summarizing the computations to derive the coupling factor between a Tx and victim Rx antenna. This section also discusses possibilities to ensure sufficient decoupling.

The remaining sections deal in more detail with the influence of a baffle on the received field strength: Hereby, Section IV presents two general approaches (E-field simulation and a simplified method based on knife-edge diffraction) to determine the baffle attenuation. To improve the predicted field strength, Section V shows an expansion of knife-edge diffraction theory by inclusion of an angle-dependent antenna gain. Finally, this section compares the obtained results for the two approaches. It is shown that the simplified theory can be used during optimization of the baffle design while field simulations are used for final fine-tuning purposes. This helps to minimize the overall engineering and computation effort. Conclusions are given in Section VI.

II. OVERVIEW OF METOP-SG SATELLITES

MetOp-SG space segment will be composed of two Low Earth Orbit (LEO) satellites, called “Satellite A” and “Satellite B”. The satellites are housing different payload instruments sensing the Earth, see Figure 1.

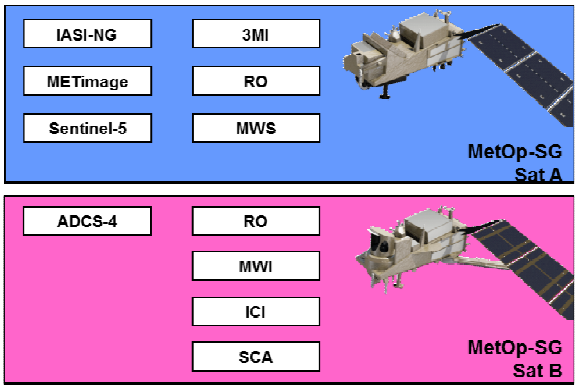


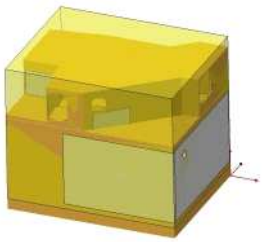
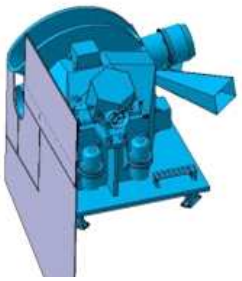
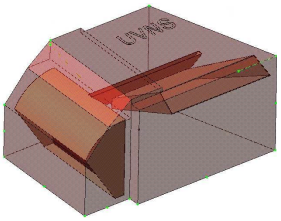
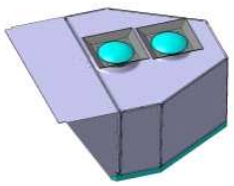
Figure 1. Allocation of payload instrument on-board the MetOp-SG satellites; left: Customer Furnished Instruments; right: Contractor Furnished Instruments

The full names of the instruments are:

- IASI-NG: Infrared Atmospheric Sounding Interferometer - New Generation
- 3MI: Multi-viewing, Multi-channel, Multi-polarisation Imager
- RO: Radio Occultation
- MWS: MicroWave Sounder
- METimage and Sentinel-5: no further name
- A-DCS 4: ARGOS Advanced Data Collection System 4
- MWI: MicroWave Imager
- ICI: Ice Cloud Imager
- SCA: SCAtterometer

Table I shows the geometry and the basic sensing functions of the instruments on-board Satellite A ([3], [4]) based on the status at System Requirements Review (SRR).

TABLE I. INSTRUMENTS ON-BOARD SATELLITE A

	**IASI-NG** Atmospheric temperature and humidity profiles; monitor various trace gases (for example ozone [O_3], carbon monoxide [CO], methane [CH_4], carbon dioxide [CO_2]) Frequency range: infrared sensing with wavenumber $k=2\pi/\lambda$ ranging from 645 cm^{-1} to 2760 cm^{-1} and a spectral resolution of 0.25 cm^{-1}.
	**METimage** High resolution information on clouds, cloud cover, land surface properties, sea, ice and land surface temperatures, etc. Frequency range: Optical imaging with 20 channels between 0.443 µm and 13.345 µm
	**Sentinel-5** Ozone and other atmospheric gases profile & column, aerosols optical depth; monitor various trace gases, monitor air quality and support climate monitoring Frequency range: From 0.27 µm (ultraviolet) to 2.385 µm (near infrared)
	**3MI** Aerosols (optical thickness, particle size, type, height, absorption) , volcanic ashes, surface albedo Frequency range: 12 channels from 0.41 µm to 2.13 µm

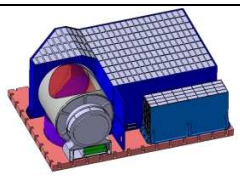
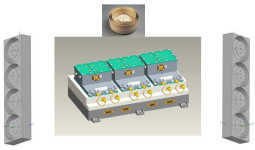
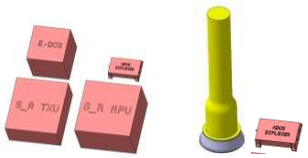
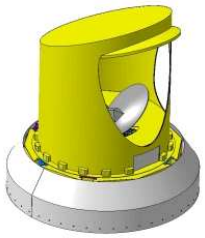
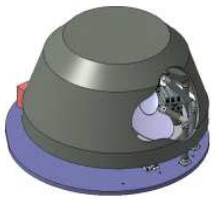
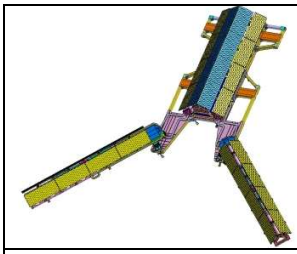
	**MWS** Atmospheric temperature and humidity profiles in clear and cloudy air, cloud liquid water total column Frequency range: RF channels at center frequencies between 23.8 GHz and 189 GHz
	**RO** Temperature, pressure and humidity profiles, electron contents in ionosphere Frequency range: Band L1: 1.57542 GHz +/- 10.23 MHz Band L5: 1.17645 GHz +/- 10.23 MHz

Table II shows the geometry and the basic sensing functions of the instruments on-board Satellite B ([2], [3]).

TABLE II. INSTRUMENTS ON-BOARD SATELLITE B

	**A-DCS 4** Collection of in-situ oceanographic and meteorological data Frequency range: ~ 400 MHz
	**MWI** Precipitation & cloud products, water vapour profiles & imagery, sea ice Frequency range: RF channels at center frequencies between 18.7 GHz and 191 GHz
	**ICI** Cloud products (ice clouds), snowfall detection and quantification Frequency range: Different RF channels between 180 GHz and 669 GHz

	**SCA (Tx + Rx)** Ocean surface wind vectors and soil moisture Frequency range: 5.355 GHz +/- 1 MHz
In addition: RO instrument (see Table 1)	

In addition, both satellites are housing a TT&C system in S-Band (transmitter and receiver) and transmitters in X-Band and Ka-Band for downlink of the sensed data towards Ground.

When the downlink transmitters are active (transmission via Tx antenna), it has to be ensured that the instrument receivers are not distorted by the emissions. Although the on-board Tx antennas are designed to radiate towards the Earth, the field strength around a Tx antenna is not negligible potentially leading to interference seen by the on-board receivers [5]. Limiting this effect is key to proper performance of the receivers. Reduction of unintended interference power can be achieved by, e.g., sufficiently large distances among Tx and Rx antennas, optimization of antenna patterns and inclusion of additional baffles to avoid Line-of-Sight links between Tx and Rx antennas.

For readability reasons, the remaining part of the contribution will use the wording “Transmitter (Tx)” and “Receiver (Rx)” in the sense of the dedicated antennas. Figure 2 shows a preliminary model (at System Requirements Review) of “Satellite A” together with the positions of an exemplary Tx radiating in the X-Band towards the Earth, the Microwave Sounder (MWS) instrument receiver, a baffle and the Nadir direction (towards the Earth during flight).

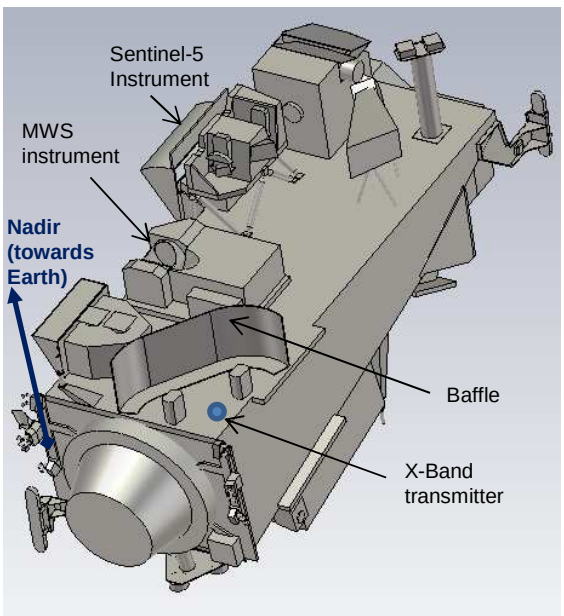


Figure 2. Model of “Satellite A” being part of MetOp Second Generation: Exemplary transmitter and instrument receiver positions

III. FUNDAMENTALS OF RFC ANALYSIS

When multiple transmitters and receivers operating in RF range are located on-board a satellite, potential interference is an issue, and a Radio Frequency Compatibility analysis has to be performed to ensure proper performance. For this purpose, coupling factors between involved Tx and Rx constellations are determined and the resulting interference level at the Rx position is compared to a specified limit.

In the following paragraph, the coupling factor is derived for free space propagation as a function of distance and the angle dependent antenna gain between Tx and Rx. In addition, the analysis takes into consideration

- Improvement of Tx-Rx decoupling for receivers integrating over a pulsed signal (MWI / SCA)
- Additional attenuation in case of No-Line of Sight between Tx and Rx (e.g., shading by structure or intended baffles); hereby, the attenuation value is based on 3D full-wave electro-magnetic simulations (CST Microwave Studio software).

A. Modeling of Interference Power

The approach presented below assumes free space propagation between a Tx and Rx, where the Tx radiates a power P_{Tx} at a frequency f and the victim Rx receives the signal in-band as an interference signal. Figure 3 shows a general constellation involving Tx, Rx, antenna patterns and the definition of elevation angles towards the LoS path.

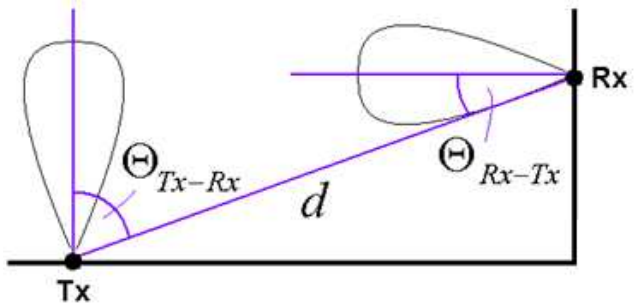


Figure 3. General definition of angles between transmitter and receiver

If the antenna patterns are also dependent on the azimuth angle, azimuth has to be considered as well.

In general, the power density S (in W/m²) at a victim Rx which has been generated by a Tx antenna can be determined as

$$S = \frac{P_{Tx} \cdot G_{Tx}\left(\Theta_{Tx-Rx}, \varphi_{Tx-Rx}\right)}{4\pi \cdot d^2} \quad (1)$$

where P_{Tx} is the total transmitted Power, $G_{Tx}(\Theta_{Tx\text{-}Rx}, \varphi_{Tx\text{-}Rx})$ is the Tx antenna gain at the considered frequency f in the direction of the Rx (direction described by the elevation angle $\Theta_{Tx\text{-}Rx}$ and the azimuth angle $\varphi_{Tx\text{-}Rx}$), and d is the distance between Tx and Rx.

The Rx antenna will suffer interference from the incident signal. The received interference power P_{Rx} at a distance of d is

$$P_{Rx} = S \cdot A_{eff} = \frac{P_{Tx} \cdot G_{Tx}\left(\Theta_{Tx-Rx}, \varphi_{Tx-Rx}\right)}{4\pi \cdot d^2} \cdot \frac{G_{Rx}\left(\Theta_{Rx-Tx}, \varphi_{Rx-Tx}\right) \cdot c_0^{\,2}}{4\pi \cdot f^2} \quad (2)$$

where $G_{Rx}(\Theta_{Rx\text{-}Tx,}\ \varphi_{Rx\text{-}Tx})$ is the antenna gain of the victim Rx towards the Tx, f is the Rx frequency, A_{eff} the effective area of the antenna, and c_0 the speed of light in vacuum. This equation describes free space propagation and is known as Friis equation. It assumes that the Rx is positioned in the far field of the transmitter: For antennas physically larger than $\lambda/2$ (where λ is radiated wavelength), the Rx is in the far field of the Tx for $d > d_f = 2D^2/\lambda$ (far field condition). The parameter D corresponds to the physical length of an antenna, or the diameter of a "dish" antenna. In addition, the following conditions have to be fulfilled: $d_f >> D$ and $d_f >> \lambda$.

The received power may be attenuated due to harness losses between the Rx antenna and the receiver input (e.g., 2 dB losses). This can be respected by a factor L_{har} ($\leqslant 1$) in the equation. The interference power at the receiver input is then

$$P_{Rx} = L_{har} \cdot \frac{P_{Tx} \cdot G_{Tx}\left(\Theta_{Tx-Rx}, \varphi_{Tx-Rx}\right)}{4\pi \cdot d^2} \cdot \frac{G_{Rx}\left(\Theta_{Rx-Tx}, \varphi_{Rx-Tx}\right) \cdot c_0^{\,2}}{4\pi \cdot f^2} \quad (3)$$

The coupling factor is defined as the ratio between the received and the transmitted power. Above equation leads to

$$C = \frac{P_{Rx}}{P_{Tx}} = L_{har} \cdot G_{Tx}\left(\Theta_{Tx-Rx}, \varphi_{Tx-Rx}\right) \cdot \frac{G_{Rx}\left(\Theta_{Rx-Tx}, \varphi_{Rx-Tx}\right) \cdot c_0^{\,2}}{\left(4\pi \cdot d \cdot f\right)^2}. \quad (4)$$

Hereby, the expression $c_0^2 / (4\ \pi\ d\ f)^2$ is also called free space loss. The coupling factor is hence the free space loss multiplied by the loss factor L_{har} at Rx side and the antenna gain of both Tx and Rx antenna in the LoS direction.

The coupling factor in dB is obtained by applying "10*log$_{10}$" of the linear value. In satellite design with distances in the range of meters and frequencies in RF range, a typical coupling factor is, e.g., -100 dB.

The sensitivity of the victim receiver describes the maximal allowed interference power (e.g., value in mW or dBm in logarithmic notation) at Rx side. The interference power P_{Rx} shall be smaller than the specified sensitivity, where the difference is called RFC margin. An RFC margin of 20 dB is typically recommended in an early project phase. Sometimes, the sensitivity of the receiver is not given in terms of power, but in terms of power spectral density (=PSD, e.g., in mW/Hz or dBm/Hz in logarithmic notation). In this case, power has to be replaced by PSD values in above equations. This leads to

$$PSD_{\mathrm{Rx}} = L_{\mathrm{har}} \cdot \frac{PSD_{\mathrm{Tx}} \cdot G_{\mathrm{Tx}}\left(\Theta_{\mathrm{Tx-Rx}}, \varphi_{\mathrm{Tx-Rx}}\right)}{4\pi \cdot d^2} \cdot \frac{G_{\mathrm{Rx}}\left(\Theta_{\mathrm{Rx-Tx}}, \varphi_{\mathrm{Rx-Tx}}\right) \cdot {c_0}^2}{4\pi \cdot f^2} \quad (5)$$

where PSD_{Tx} is the power spectral density of the Tx signal at frequency f and PSD_{Rx} is the power spectral density of the Rx signal at frequency f. In this case, the coupling factor is defined as

$$C = \frac{PSD_{\mathrm{Rx}}}{PSD_{\mathrm{Tx}}} = L_{\mathrm{har}} \cdot G_{\mathrm{Tx}}\left(\Theta_{\mathrm{Tx-Rx}}, \varphi_{\mathrm{Tx-Rx}}\right) \cdot \frac{G_{Rx}\left(\Theta_{Rx-Tx}, \phi_{\mathrm{Rx-Tx}}\right) \cdot {c_0}^2}{\left(4\pi \cdot d \cdot f\right)^2}. \quad (6)$$

The resulting coupling factor is hence the same when compared to the previous definition, which was based on power. The interfering power spectral density PSD_{Rx} shall be smaller than the specified sensitivity, where the difference is called RFC margin. If the LoS is shaded, e.g., by a dedicated baffle, an additional loss factor has to be considered in above equation. In this case, the coupling factor can be described by:

$$C = L_{\mathrm{har}} \cdot L_{\mathrm{baffle}} \cdot G_{\mathrm{Tx}}\left(\Theta_{\mathrm{Tx-Rx}}, \varphi_{\mathrm{Tx-Rx}}\right) \cdot \frac{G_{\mathrm{Rx}}\left(\Theta_{\mathrm{Rx-Tx}}, \varphi_{\mathrm{Rx-Tx}}\right) \cdot {c_0}^2}{\left(4\pi \cdot d \cdot f\right)^2}. \quad (7)$$

Above considerations are based on in-band interference. In general, this case can be avoided by proper selection of Tx and Rx frequencies. Nevertheless, interference may occur, e.g., since the Tx radiates, in addition to the desired signal.

- Out-of-band noise: This means that the Tx radiates noise power outside the desired Tx frequency range. The coupling factor is determined in the same way as for in-band considerations.
- Out-of-band spurious (harmonics): Hereby, the Tx radiates also an integer multiples of the carrier frequency f_c. The n-th harmonic is associated with a frequency of $n*f_c$. The coupling factor between the n-th harmonic and the Rx is calculated by the same equation as shown above, but f has to be replaced by $n*f_c$.

As a rule of thumb, if no sensitivity value is specified, the received interference power should be about 20 dB below the minimal input level of the receiver which may be, e.g., -120 dBm. In reality, a victim Rx may receive multiple interference signals simultaneously that originate, e.g., from different transmitters. In this case, the sum of all contributions at the considered frequency must still provide sufficient RFC margin (e.g., 20 dB).

B. Methods to Achieve Strong Decoupling

Equation (7) indicates that strong decoupling between a Tx and a Rx can be achieved by a high baffle attenuation, sufficiently low antenna gain at both Tx side and Rx side in LoS direction, large distance, low Tx power and high Tx frequency. In case of harmonics radiation, the radiated power of the harmonic signal can, e.g., be minimized by proper RF filtering.

C. Temporal Effects

On MetOp-SG "Satellite B", the Scatterometer (SCA) radiates pulsed signals. For combinations with SCA Tx and MWI Rx, the pulsed nature of SCA signals leads to an improvement of the decoupling between Tx and Rx, as described hereafter.

The SCA (only present on Satellite B) radiates a pulsed signal. This means that the signal (and hence the power spectral density in dBm/Hz) is only present during the pulse duration T_{P}. The MWI receiver hence observes a power spectral density of $S(t)=S_0$ during T_{P}, else zero. The MWI instrument integrates $S(t)$ over the integration time $T_{\mathrm{int}} > T_{\mathrm{P}}$ which leads to

$$\int_0^{T_{\mathrm{int}}} S(t)dt = S_0 \cdot T_{\mathrm{P}} = S_{0,\mathrm{red}} \cdot T_{\mathrm{int}} \quad (8)$$

where $S_{0,\mathrm{red}}$ is the effective reduced power spectral density as indicated in Figure 4.

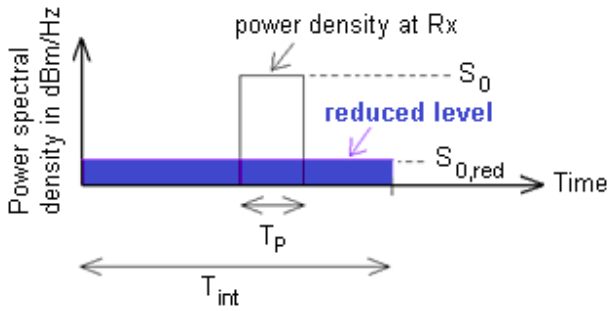


Figure 4. Temporal effects on power spectral density

This means:

$$S_{0,\mathrm{red}} = S_0 \cdot \frac{T_\mathrm{P}}{T_\mathrm{int}}. \qquad (9)$$

In logarithmic notation, the received power spectral density is hence $S_{0,red}$[dBm/Hz] = S_0[dBm/Hz] + 10*log (T_P/T_{int}) whereas the second term provides a negative value (hence $S_{0,red}$ is lower than S_0). |10*log (T_P/T_{int})| describes the improvement of the decoupling between Tx and Rx due to temporal effects, which translates into a respective improvement of the RFC margin.

IV. APPROACH TO DETERMINE BAFFLE INFLUENCE

This section assumes a metallic baffle (e.g., wall) between a Tx and a victim Rx to limit undesired signals at the Rx position. The physics of electromagnetic wave propagation at radio frequencies is the reason for an undesired signal still present at the Rx position, albeit strongly attenuated: Signal paths originating from diffraction at the baffle can travel towards the Rx as a result of Huygen's principle. In addition, further signal contributions may originate from reflections or scattering at objects in the vicinity of the Tx and Rx. The principle of this multipath propagation is visualized in Figure 5. Hereby, the shown diffracted path interacts with the baffle directly above the hypothetical LoS path. In general, further diffracted paths are possible with interaction points along the top of the baffle.

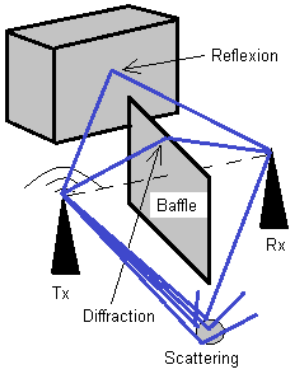


Figure 5. Multipath propagation

Since reflected and scattered paths can carry significant power levels, these contributions should be avoided by a proper design of the baffle (e.g., by an adequate height and an adequate length around the surrounding objects). In this case, the dominant contribution at Rx side only results from the diffraction at the baffle. Due to the physics of diffraction, the interfering signal decreases with steeper diffraction angle (e.g., increased baffle height) and frequency.

The influence of a baffle on the received signal can be determined either by:

- A simplified wave propagation model, e.g., theory of knife-edge diffraction.
- 3D field simulations: A simulation tool solves the corresponding electromagnetic field equations and determines the received field strength at the Rx. This method implicitly takes into account diffraction, reflection and scattering.

A. Analytical Approach by Knife-edge Diffraction

The scenario related to "knife-edge diffraction" is visualized in Figure 6. It assumes a "knife-edge" obstacle between Tx and Rx and shows the diffracted path between Tx and Rx. Hereby, the obstacle subdivides the distance between Tx and Rx into d_1 and d_2. Two cases are possible: In case 1, the upper edge of the obstacle appears at a height $h > 0$ w.r.t. the Line of Sight (LoS). This leads to a "No Line of Sight" (NLoS) scenario. In case 2, the upper edge of the obstacle appears at a height $h < 0$ w.r.t. LoS. This leads to a LoS scenario.

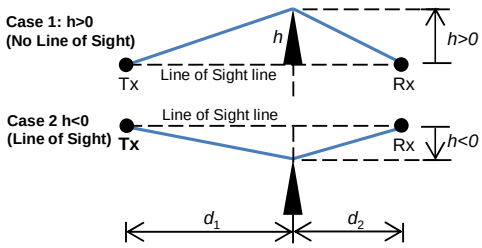


Figure 6. Diffraction at a "knife-edge" for two cases: "No Line of Sight" and "Line of Sight"

According to [6] and [9], the loss induced by the baffle (diffraction loss) is

$$L_{dB} = -20 \cdot \log_{10}\left|F(v)\right| \qquad (10)$$

with the complex Fresnel integral

$$F(v) = \frac{1+j}{2} \cdot \int_{v}^{\infty} e^{-j\pi t^2/2} dt \qquad (11)$$

and

$$v = h \cdot \sqrt{\frac{2}{\lambda} \cdot \left(\frac{1}{d_1} + \frac{1}{d_2}\right)} \qquad (12)$$

where v is the Fresnel-Kirchhoff diffraction parameter and $\lambda = c_0 / f$ is the wavelength of the considered signal. The

resulting diffraction loss ("baffle attenuation") as a function of v is plotted in Figure 7 for v = [-5 .. 5] as per [7].

Figure 7. Diffraction loss of a "knife-edge" versus parameter v [7]

The figure shows the level of the diffracted path in dB relative to freespace, which is negative for $v > -0.7$. Hereby, a level of "- x dB" corresponds to an attenuation of "x dB". According to (12), v and h are proportional, hence, $h > 0$ (NLoS) is associated with $v > 0$, yielding a baffle attenuation of at least 6 dB (see graph). The above graph can be approximated, e.g., by the following piecewise function [8]:

$$L_{dB} = \begin{cases} -(6 + 9 \cdot v - 1.27v^2) & \text{if } 0 \le v \le 2.4 \\ -(13 + 20 \cdot \log_{10}(v)) & \text{if } v > 2.4 \end{cases} \quad (13)$$

Note that above equation is the good one compared to a sign error related to $1.27v^2$ in [8]. To quickly determine the "baffle attenuation", the approach is to determine v by (12) and then to apply (13) for the obtained v.

Example (typical values on a satellite): For d_1 = 1.5 m, d_2 = 1.5 m and f = 8.2 GHz (X-Band as typical downlink case), Figure 8 visualizes the "baffle attenuation" as a function of the parameter h.

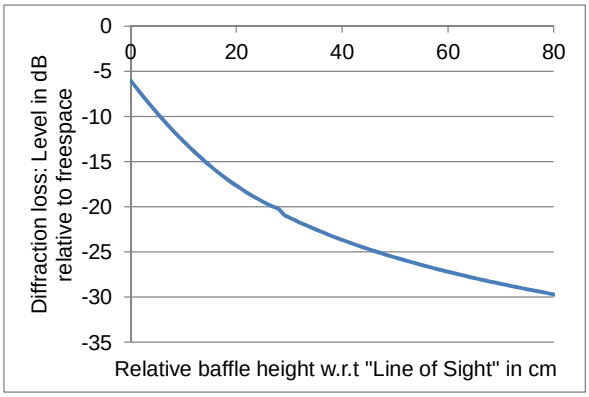


Figure 8. Diffraction loss of a "knife-edge" versus h assuming d_1 = 1.5 m, d_2 = 1.5 m and f = 8.2 GHz

The result reveals that the attenuation is very sensitive to the height. This behavior is due to the small wavelength which is only 3.7 cm in the considered case.

The other way around, the theory of knife-edge diffraction reveals that the baffle attenuation in X-Band frequency range can be improved significantly by only slightly increasing the baffle height. In practice, constraints on the height are given by the required field of views of the transmitters and instruments.

B. *Simulation based approach (CST field simulation)*

An approach based on solving electromagnetic field equations has the following advantages:

- Result available for any baffle geometry (not only for simple objects like a "knife-edge")
- All wave propagation phenomena implicitly taken into account (e.g., also reflection and scattering), not only diffraction as in the "knife-edge model"
- Environment (surrounding structure) can be taken into account

A well suited approach for satellite engineering is to use the simulation software "Microwave Studio" from the company CST. For example, this tool has also been used by Airbus Defence and Space to assess EMC/RFC for MTG satellites.

To determine the baffle attenuation, a dipole antenna is placed at the transmitter position and oriented in a way that the radiation towards the receiver position is maximized. The electric field strength in dB(mV/m) at a victim receiver is first simulated without baffle (reference, including Line of Sight path) and then with baffle. In both cases, the surrounding satellite structure is taken into account. The difference of the electric field strength in dB(mV/m) corresponds to the baffle attenuation in dB.

To obtain the simulation results reported in this paper, the integral equal solver based on Multi Level Fast Multipole Method (MLFMM) has been used. MLFMM is a technique based on the same principles as the traditional "Method of Moments" (MoM), but applicable to models of significantly larger electrical size. Given the geometrical dimensions of typical Earth observation satellites, simulations at frequencies as high as (roughly) 30 GHz can be performed applying this numerical technique. Higher frequencies (smaller wavelengths) require a mesh size that results in increased memory demand and simulation time. Should the need arise to overcome that constraint for practical limitations (e.g., memory size), the satellite structure can be restricted to a representative volume encompassing the Tx and Rx positions.

V. Comparison of Field Simulations w.r.t. Knife-Edge Theory

On Satellite A, the radiation of the X-Band transmitter towards the MWS instrument is reduced by a baffle (height:

65cm). Figure 9 visualizes a part of the satellite structure including the phase center of the transmitter (modeled as a dipole) radiating at 8.2 GHz, the baffle as well as the MWS victim receiver. Hereby, two Rx positions ("Position 1", "Position 2") are considered, where "Position 2" corresponds to the center of the MWS reflector plate. The figure also shows the position of the Sentinel-5 instrument.

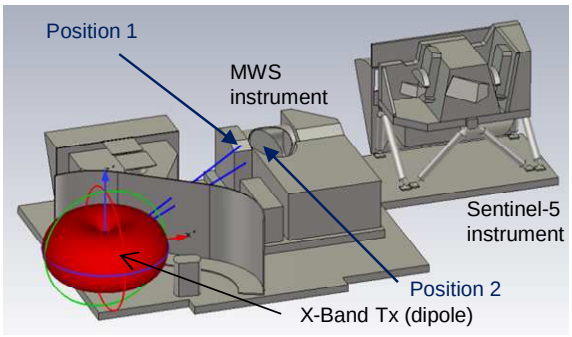


Figure 9. Part of the structure of Satellite A (dipole Tx)

The figure also indicates the LoS directions between Tx and the two Rx positions. The electric field strengths are simulated with the CST software for two scenarios:

- "without baffle"
- "with baffle".

Results are presented in Figure 10.

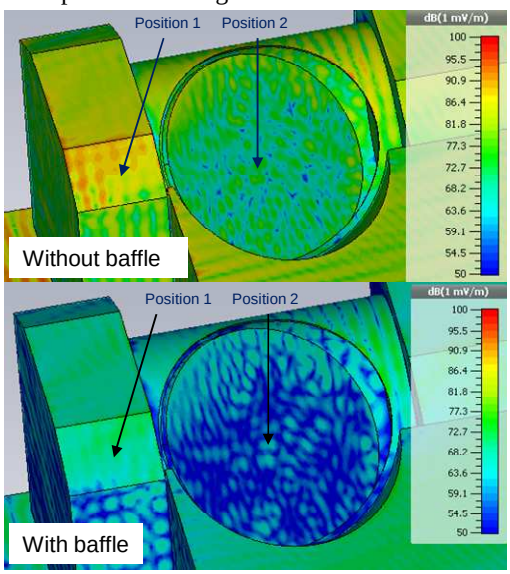


Figure 10. Simulated field strength at MWS assuming radiating dipole; f=8.2 GHz

Observation:
Position 1: The case "Without baffle" reveals a field strength of 90 ± 1 dBmV/m". The case "With baffle" reveals 72 ± 1 dBmV/m. Hence, the difference is 18 dB.

Position 2: The case "Without baffle" reveals a field strength of ≈ 77 dBmV/m". The case "With baffle" reveals ≈ 64 dBmV/m. Hence, the difference is 13 dB.

In a second step, the attenuation is estimated by applying the theory of knife-edge diffraction. As explained in the section on knife-edge theory, the baffle subdivides the theoretical LoS path into two distances (d_1, d_2) and a relative height h of the baffle. For "Position 1", the values are: d_1 = 1.07 m, d_2 = 1.08 m, h = 0.16 m. Assessment at f = 8.2 GHz yields an expected baffle attenuation of 17.2 dB while 18 dB has been simulated by CST software according to the previous figure. This shows a good agreement between simplified theory and CST simulations. Assessment for "Position 2" (d_1 = 1.05 m, d_2 = 1.43 m, h = 0.218 m) at f = 8.2 GHz yields an expected baffle attenuation of 18 dB while 13 dB has been simulated by CST software. This behavior can be explained as follows: In contrast to "Position 1", "Position 2" does not enable a path directly diffracted at the baffle towards the receiver position. The signal can arrive at "Position 2" only via multiple interactions, hence, the knife-edge diffraction theory based on a single baffle is not applicable.

Next, the radiation pattern of the Tx antenna is replaced by the measured characteristics of the physical X-Band helix antenna. Figure 11 visualizes the 3D pattern as well as the antenna gain as a function of elevation angle Θ.

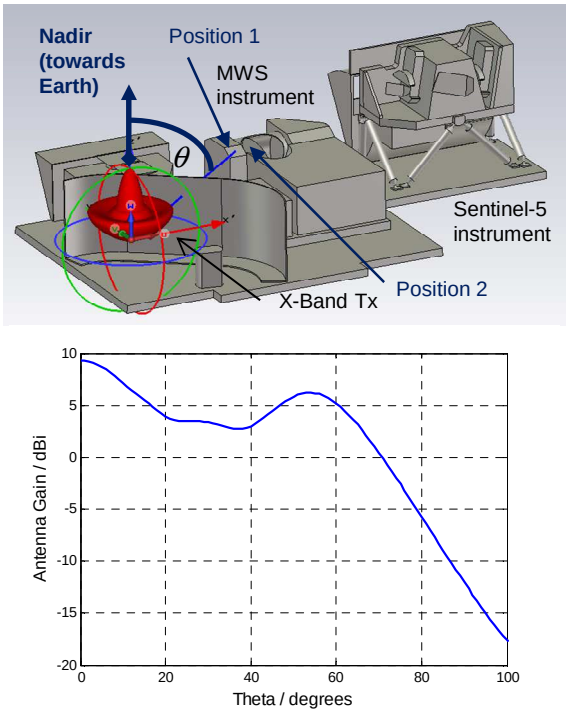


Figure 11. Scenario with real antenna pattern; antenna performance

For the analysis, "Position 1" is considered. The CST simulation as per Figure 12 reveals: The case "Without baffle" leads to a field strength of 80.8 ± 1

dBmV/m" while "With baffle" leads to 70.8 ± 1 dBmV/m. Hence, the difference caused by the baffle is 10 dB.

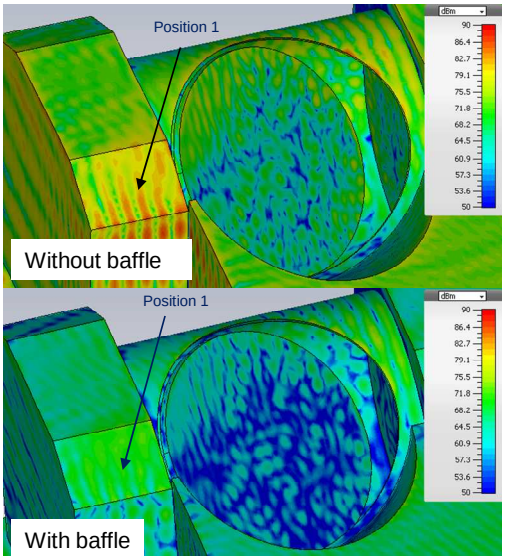


Figure 12. Simulated field strength at MWS assuming real antenna pattern; f=8.2 GHz

The question arises if this value of 10 dB attenuation can be predicted by the knife-edge diffraction theory. To do so, the angle-dependent antenna data has been incorporated into the knife-edge diffraction theory. The approach is described hereafter: First, the elevation angle is determined under which a propagation path leaves the transmitter. Figure 13 shows the principal scenario.

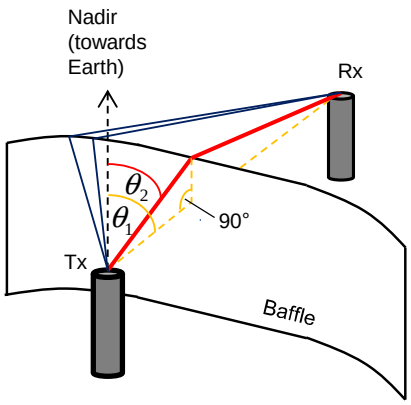


Figure 13. Principal scenario involving diffracted paths

- A dotted line indicates the propagation path in LoS direction which is present in absence of the baffle. The associated elevation angle is Θ_1.
- In presence of a baffle, a path originating from diffraction appears at an angle $\Theta_2 < \Theta_1$. Hereby, the interaction point with the baffle is inside the plane defined by the Nadir direction and the LoS direction.

For "Position 1", the elevation angles and the associated antenna gain according to Figure 11 are:

- Θ_1 = 89.9 deg, associated with a gain of -12.5 dBi.
- Θ_2 = 82.4 deg, associated with a gain of -7.3 dBi.

Hence, the diffracted path runs along a direction with higher gain when compared to the LoS direction. Therefore, it is expected that the influence of the baffle is lower compared to the dipole case. The expected attenuation by insertion of the baffle corresponds to the result of the dipole, corrected by the delta antenna gain, hence, the expected value is 17.2 dB – ((-7.3) - (-12.5)) dB = 12 dB.

For comparison, 10 dB attenuation has been determined using the CST simulation software. Limited differences in the result can be explained, e.g., by

- Multipath propagation:
  While above consideration assumes only one diffracted path, further diffracted paths are possible along the top of the baffle. These additional paths occur out of the plane which is defined by Nadir direction and LoS direction. Possible additional paths are already visualized in the left part of Figure 13. In principle, all paths have to be weighted by the angle-dependent antenna gain and then summed up. As the knife-edge theory does not predict multiple paths and the associated elevation angles, only weighting of the diffracted path "in-plane" is possible. A more complex channel model which predicts multiple paths and allows for insertion of an angle dependent antenna gain is Ray-tracing [10]. A disadvantage of this technique is however increased computational time.
- Baffle geometry:
  The baffle geometry differs from the ideal "knife-edge theory" as the baffle is bended and the distance between Tx and baffle differs along the baffle.
- Approximation of Fresnel integral :
  Equation (13) is only an approximation of (10).

To verify the effect of baffles on-board the MetOp-SG satellites prior to launch, measurements are envisaged in the frame of ground testing. These so-called mock-up tests will be performed in Q2/2016 and use transmitters and receivers with representative antenna pattern as well as a relevant part of the satellite structure.

A similar approach using an adapted knife-edge model is shown in [11], which considers the channel between a train and a satellite including a knife-edge obstacle that models structural elements on the roof of the train. In [11], classical knife-edge theory is expanded by only one antenna gain (the "train antenna gain"), whereas the present contribution takes into account both the characteristics of the transmitter and the receiver.

Finally, a general remark is given w.r.t. field predictions when involving antenna patterns: The radiation pattern of a transmit antenna differs between the near-field and the far field where far field conditions are achieved at distances of $d > d_{min} = 2\ D^2 / \lambda$ (D = antenna dimension). When using a far field antenna pattern in above approach, the distance between the transmit antenna and the baffle has to be at least d_{min} (fulfilled in above consideration). A near field approach considers possible pattern distortion by the baffle.

VI. CONCLUSIONS

Modern Earth observation satellites such as the MetOp-SG satellites accommodate manifold Radio Frequency transmitters and instrument receivers. The on-board transmitters generate an electromagnetic environment with potential impact on the performance of instrument receivers. Ensuring Radio Frequency Compatibility means that the level of the unintended signal at Rx side is kept below a certain threshold level so that the instrument performance is not degraded. The satellite configuration (for example, position and orientation of on-board transmitters and receivers) is vital to RFC. As a consequence, it deserves careful consideration throughout the satellite program, starting with a first optimization in a very early project phase.

The suitability of the configuration w.r.t. RFC is verified by an RFC analysis, which is based on the calculation of the coupling factor between critical combinations of Tx and Rx. An RFC analysis covers both in-band-radiation and out-of-band radiation (e.g., radiation of wideband noise or harmonics). A correction factor leading to improved decoupling should be applied when a pulsed signal is received by a receiver applying integration times in excess to the pulse width of the interfering signal. In case of insufficient decoupling between Tx and Rx, the situation can be improved by a dedicated baffle between Tx and Rx, optimization of antenna orientations, increased distance, lower Tx power or stronger filtering effort at Tx side.

Conclusions related to a proper design of a baffle are given hereafter: The height of the baffle shall be large enough to

- realize NLoS between Tx and Rx (and hence, a diffracted path towards the Rx)
- avoid reflections at, e.g., high objects in the vicinity of Tx and Rx

The length of the baffle shall be large enough to avoid reflections at objects next to the baffle which could carry significant power towards the Rx.

To determine the baffle attenuation for such a properly designed baffle, two methods have been studied: 3D field simulations and knife-edge diffraction theory (based on a single baffle), expanded by information on antenna gain. It has been shown that the results agree well in scenarios resembling the set-up illustrated in Figure 13, involving a single diffraction of the wave propagating from Tx to Rx. Hence, the simplified theory is an adequate method for assessing the effectiveness of the baffle prior to initiating extensive 3D full-wave simulations. This approach minimizes the overall engineering and computation effort. Verification of the derived results for MetOp-SG will be achieved by mock-up testing in Q2/2016.

REFERENCES

[1] J. Timmermann, C. Imhof, D. Lebherz, and J. Lange, "Application of Knife-Edge Diffraction Theory to Optimize Radio Frequency Compatibility On-board a Satellite," The Seventh International Conference on Advances in Satellite and Space Communications (SPACOMM), April 2015, ISBN: 978-1-61208-397-1, pp. 7-12.

[2] Statement Of Work for MetOp Second Generation (MetOp-SG) Phase B2/C/D/E, ESA UNCLASSIFIED – For Official Use, MOS-SOW-ESA-SYS-0494, issue 1, 09/09/2013.

[3] MetOp Second Generation (MetOp-SG) Space Segment Requirements Document (SSRD), MOS-RS-ESA-SYS-0001, 09/09/2013

[4] Monitoring Weather and Climate from Space – EPS-SG Overview for CSPP / IMAPP User's Group Meeting, 15 April 2015; http://www.ssec.wisc.edu/meetings/cspp/2015/Agenda%20PDF/Wednesday/Schluessel_EPS-SG_CSPP_IMAPP.pdf

[5] J.A. Miller and A.R. Horne, "Radio frequency compatibility design and testing on the polar platform spacecraft," Electromagnetic Compatibility, 10th International Conference on Electromagnetic Compatibility (Conf. Publ. No. 445), pp. 35-40, 1-3 Sept. 1997

[6] K. Du and M. Swamy, Wireless Communication Systems: From RF Subsystems to 4G Enabling Technologies. Cambridge University Press, 2010.

[7] http://www.mike-willis.com/Tutorial/PF7.htm [retrieved: Feb., 2015]

[8] www.wirelesscommunication.nl/reference/chaptr03/diffrac.htm [retrieved: Feb., 2015]

[9] C. Hasslet, Essentials of Radio Wave Propagation. Cambridge Wireless Essentials Series, Cambridge University Press, 2008.

[10] J. Timmermann, M. Porebska, C. Sturm, and W. Wiesbeck, "Investigating the Influence of the Antennas on UWB System Impulse Response in Indoor Environments," 37th European Microwave Week (EuMW), Oct. 2007, pp. 1562-1565.

[11] S. Scalise, H. Ernst, and G. Harles. "Measurements and modeling of the land mobile satellite channel at Ku-Band," IEEE Transactions on Vehicular Technology, 57 (2), pp. 693-703, March 2008.

Prefetching Schemes and Performance Analysis for TV on Demand Services

Manxing Du*†, Maria Kihl†, Åke Arvidsson‡§, Huimin Zhang¶, Christina Lagerstedt* and Anders Gavler*
*Acreo Swedish ICT, Sweden, Email: firstname.lastname@acreo.se
†Dept. of Electr. and Inform. Technology, Lund University, Sweden, Email: firstname.lastname@eit.lth.se
‡Business Unit Support Solutions, Ericsson, Sweden, Email: firstname.lastname@ericsson.com
§Department of Computer Science, Kristianstad University, Sweden, Email: firstname.lastname@hkr.se
¶Uppsala University, Sweden, Email: firstname.lastname.4997@student.uu.se

***Abstract*—TV-on-Demand services have become one of the most popular Internet applications that continuously attracts high user interest. With rapidly increasing user demands, the existing network conditions may not be able to ensure a low start-up delay of video playback. Prefetching has been broadly investigated to cope with the start-up latency problem, which is also known as user perceived latency. In this paper, two datasets from different IPTV providers are used to analyse the TV program request patterns. According to the results, we propose a prefetching scheme at the user end to preload videos before user requests. For both datasets, our prefetching scheme significantly improves the cache hit ratio compared to passive caching and we note that there is a potential to further improve prefetching performance by customizing prefetching schemes for different video categories. We further present a cost model to determine the optimal number of videos to prefetch. We also discuss if there is enough time for prefetching. Finally, more factors, which may have an impact on optimizing prefetching performance, are further discussed, such as the jump patterns over different time in a day and the the distribution of each video's viewing length.**

Keywords–TV-on-Demand services; user perceived latency; prefetching; jump patterns over time; viewing fractions;

I. INTRODUCTION

Internet has become a popular medium for distributing multimedia content like TV shows, movies, and user generated videos besides the tradional distribution channels such as air broadcasting, cable networks and physical media like Video Home System (VHS) or Digital Versatile Disc (DVD). The massive amount of multimedia traffic has imposed a significant burden on the Internet. Consequently, users sometimes have to endure long access delays for filling up the playout buffer before the content is displayed. Although web caching is widely used as a solution to lessen the web traffic congestion and improve network performance, the benefit of caches is limited. To further reduce user perceived latency, prefetching has become a popular technique. The objective of the prefetching system is to proactively preload certain content to the cache even before the user requests it. We have explored the subject previously in [1] and this paper extends the analysis and discussion presented previously.

Understanding the usage of IPTV services is very important when investigating prefetching schemes. In [2], the usage of a Peer-to-Peer IPTV service which includes both live and VoD content is presented. The daily VoD content receives more requests than live content. Videos belonging to different genres have different temporal distributions of popularity. In [3] [4] [5], the user behaviour for a VoD system and an IPTV service is investigated thoroughly in many aspects, such as user access rate, channel switching patterns, video popularity and so on. In [3], the impact of recommendations on the user viewing behaviour by recommending two sets of movies to the users is discussed. The result shows that recommending daily popular videos has a much more significant impact on the users choices than recommending popular videos over a longer period (15 days), which suggests that VoD content has short life time. We also know that the user's tolerance for downloading web pages is short. The results in [6] show that it is approximately 2 seconds and in [7], the result suggests that the more familiar the users are with a web site, the more sensitive they are to delays.

Thorough summaries of web caching and prefetching approaches and performance measures can be found in [8] [9] [10]. Domènech *et al.* in [11] compare different prefetching architectures and find that a maximum latency reduction of 67.7% can be obtained if the predictor is placed at a proxy while the collaborative prediction between proxy and server can reduce the latency by more than 97.2%. However, the results are obtained based on the ideal scenario that the prediction is always correct. Thus, the results can be seen as the upper limit of latency reductions.

Most of the existing prefetching approaches are access-history based which predict the future user requests depending on the observed content access patterns. Márquez *et al.* applies a double dependency graph prediction algorithm to a mobile web and observes that the performance of prefetching approaches rely on the underlying networking technologies [12]. Another history based model is the Markov model, which is an effective scheme to predict what users intend to request based on the sequence of the historical access [13]–[16]. The prefetching schemes proposed in [17]–[20] use data mining techniques to discover the users' access patterns.

Popularity-based prefetching approches are also widely used, especially for prefetching multimedia content. In [21], a trace driven simulation was performed to investigate prefetching schemes for YouTube videos. Their prefetching scheme is to prefetch the top 25 videos from each video's related video list to a proxy server. Combining both prefetching and conventional proxy caching, a hit ratio of 80.88% can be obtained and while requiring only a 2% increase in the network load.

However, this recommendation-based prefetching scheme requires an effective recommendation system, which can be a big challenge. Krishnappa *et al.* [22] apply a prefetching top-100 video scheme on the Hulu traffic trace, one of the most popular streaming media provider in America, from a campus network and compare the performance of prefetching with conventional proxy caching. The results prove that prefetching

is very effective for online TV services.

From these papers, we note that the research focus is mainly on proxy prefetching whereas the performance of terminal prefetching is less well known. Generally, the closer the content is to the users, the shorter the delay. To the best of our knowledge, our analysis is the first that focuses on using terminal storage to do prefetching for a TV-on-demand service.

In this study, we have used two datasets from different TV-on-demand services. Each TV-on-Demand program consists of a series of episodes which have a high consistency regarding content, thus, the index of each episode is a good indicator of the user's future access. We propose to use the intrinsic structural information of the episodes belonging to a specific TV series to make prefetching decisions. The criterion for prefetching is based on the index of episodes within each TV program. First, we analyse the potential of prefetching in our datasets, followed by an investigation into the optimal choice of prefetching. We show that a high terminal gain can be obtained by prefetching two adjacent episodes in a series for each viewed episode with minimum cost. This study also shows the potential of implementing terminal prefetching for TV-on-Demand service both effectively and economically. In addition, we investigate whether there is enough time to perform prefetching and detail possible improvements for making prefetching decisions.

The remainder of the paper is organized as follows. Section II describes the infrastructure of a prefetching system and the evaluation metrics used. Our datasets and prefetching methods are presented in Section III. Section IV shows the results and evaluations of our prefetching scheme. In Section V, we further discuss the available time for prefetching and in Section VI, factors which may have impact on improving the prefetching system are considered. At the end, we conclude in Section VII.

II. VIDEO PREFETCHING SYSTEM

By implementing a prefetching system, multimedia streams can be retrieved from a content provider and saved in a cache. In this way, users can quickly get access to their preferred content. In this section, we first introduce the components of a prefetching system, and then we present the evaluation metrics used for measuring the performance of prefetching schemes.

A. The infrastructure of a prefetching system

The prefetching system consists of two elements: the prediction engine and the prefetching engine. The prediction engine is responsible for forseeing what the users will watch next before they request the content. The prefetching engine proactively stores the video prefix to the local cache. In this case, both first-time views and repeats of the prefetched videos can be served from the local cache with short start-up delay. These two engines can be placed at any part of the web architecture, which is shown in Figure 1: at the terminals [21], at the proxies [18] [22], at the content servers [23] [24] or between these elements [11] [12].

In order to bring content closer to the end users, so as to the largest extent reduce their perceived latency, we assume that the prefetching engine is implemented at the user end, which is called terminal prefetching. This means that the prefetched content will be stored in a terminal cache at the user end. Anything from a short part of the video to the entire video

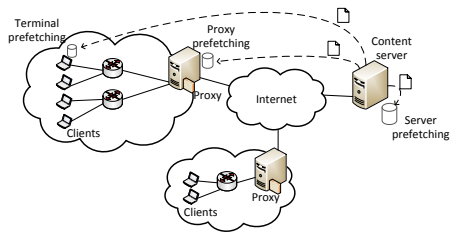


Figure 1. Network architecture with prefetching system

can be prefetched. Typically, to reduce the start-up latency, prefetching only part of the video stream is sufficient. When a user requests a video, it can be played directly from the cache, which gives more time for setting up the transport and filling up the playout buffer for the remaining parts of the video. Considering that the users may not always favour the beginning of each video [25] [26], the performance of caching different parts of the videos is discussed. Since the total length of each video is unknown and the starting and ending points of each request are also not available, in this paper, we do not address which portion of each video should be prefetched. Instead, each video is treated as a video unit, thus, the cache size is not considered. In this scenario, every video request from a user is first sent to a prefetching engine, which checks whether the video has already been retrieved earlier or not. If it has, the content will be served from the local cache instead of from the remote server. In Section VI, we made an assumption for video length and discussed the possible impact that viewing length proportion may have on improving the prefetching performance.

B. Evaluation metrics

In order to evaluate the performance of a prefetching scheme, this study uses three metrics. The first metric is precision (P), which is defined as the number of videos that are prefetched and requested by users (v) over the total number of prefetched videos (p).

$$P = \frac{v}{p} \tag{1}$$

A high precision value suggests that more prefetched content is requested by end users. Thus, the prediction engine works efficiently.

The second metric is recall (R), which is the share of prefetched and requested videos (v) over the total number of requested videos (r).

$$R = \frac{v}{r} \tag{2}$$

A higher recall value indicates that a larger share of the content requested by users can be correctly predicted by the prediction engine.

Precision and recall are constrained by each other. In principle, a higher recall value can be obtained by prefetching as many videos as possible in order to cover more content. Thus, it would be more likely that the prefetched content contains the videos requested by the users. However, in this case, p will also increase. Consequently, the prediction engine's precision will decrease. To prefetch a great amount of data, which will not be requested by the users will also deteriorate

the network congestion. These two metrics need to be balanced when designing a prefetching system. The F_1 score (balanced F-score) [27] is a weighted average of the precision and recall, which is used in this study to show the effectiveness of a prediction. The closer the F_1 score is to 1, the more effective the prediction is.

$$F_1 = 2\frac{PR}{P+R} \tag{3}$$

To evaluate whether the prefetched content is highly demanded by the users, we introduce the last metric, which is the cache hit ratio (H). It is defined as the number of requests for videos (h), which are retrieved from the prefetching cache over the total number of requests (t).

$$H = \frac{h}{t} \tag{4}$$

The cache hit ratio shows the share of repeated requests for videos, which can be served directly from the cache. A high hit ratio suggests that more requests can be served with reduced start-up delay and a high utilization of prefetched content.

III. Experiments

In this section, we will describe the experiments conducted using the prefetching method in this paper, which is to prefetch N adjacent episodes for each viewed episode. The objective of our analysis has been to investigate which episodes to prefetch for each viewed video to obtain the best performance of the prefetching system. The analysis is carried out by measuring the performance of prefetching and terminal caching in terms of effectiveness and hit ratio. To optimize the number of videos to prefetch, a cost model is proposed to find an appropriate trade-off between the cost of prefetching and the potential of response time improvement. Furthermore, we applied the prefetching scheme on another dataset to compare and validate the results.

A. Dataset

Our study is firstly conducted using the video requests from one of the most popular Swedish TV providers (*dataset 1*). The data was collected by Conviva who provides online video analytic solutions to media content providers around the world. The data is based on recorded TV requests for a subset of users in a city in Sweden. The users are so called subcribed users who have access to all the online TV content provided by the TV channel by paying a monthly fee. Thus, the users who do not have subscriptions are not included in this study. All the users are anonymized and no data can be traced back to any specific user.

There are nine video categories defined by the service provider as follows: *Children, Documentary, TV series, Home and leisure time, Entertainment, Default, Mixed, Sports and News.* There is too little data in the *Default* and *Mixed* categories for statistical analysis and usually the *Sports* and *News* content are distributed by live streaming, which cannot be prefetched like TV-on-demand content. Therefore, in the following sections, we include TV-on-demand content categorized as *Children, Documentary, TV series, Home and leisure time* and *Entertainment* and all the TV programs in each category consist of a series of episodes. To ensure unique identifiers for each episode, our dataset only includes requests for programs with only one season available.

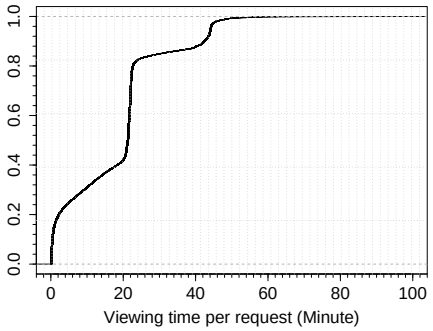


Figure 2. CDF of viewing time per request

TABLE I. Example of user request from Swedish TV provider

Title	The bridge episode 13
User Id	1044197
Start time	2012-12-31 09:00:08
End time	2012-12-31 09:21:48
Viewing minute	22
avgBitrateMbps	2.599
Program	The bridge
Category	TV series

Table I shows an example of the available information contained in each data entry. There are 7933 subscribed users who generated 104845 requests for 2427 videos belonging to 253 series over a period of 11 weeks from December 31, 2012 to March 17, 2013. Figure 2 depicts the Cumulative Distribution Function (CDF) of the viewing time per request. Around 20% of the total requests result in viewing times shorter than 2 minutes. We filter out these short viewing sessions to eliminate the impact of random clicks by the users, which are not suitable to serve as predictors. We also note that the users may request the prefetched videos after the end of time period in our dataset. Thus, the result of hit ratio is underestimated due to the finite time period of the dataset.

To test the proposed prefetching scheme on one more dataset and compare the results, a second dataset from a Portuguese catch-up TV service (*dataset 2*) is used. This dataset contains one month of request logs from its subscribers. The data was collected from June 1 to June 30 in 2014. Unlike traditional TV-on-demand services, this catch-up TV service has a 7 days access window for the content. The dataset contains over 17 million requests from about 570,000 users and more than 80 TV channels. Table II shows an example of a request log. Each video and each TV channel is identified by a unique ID, namely EpgPID and StationID. In contrast to *dateset 1*, the viewing time of each video is unknown. Since the category of TV programs provided by different channels varies, one of the most popular channels, which mainly contains TV series was chosen in the following study in order to be comparable with *dataset 1*. This channel has more than 3 million requests from about 275,000 users during the recording period.

B. User access patterns

In this section, traffic patterns for both datasets on different time scales are presented. To make it comparable with *dataset*

TABLE II. EXAMPLE OF USER REQUEST FROM CATCH-UP TV PROVIDER

Title	Belmonte - Ep. 192
User Id	A0573D6D4F9D7BC5018B3D17DC6DCB3
Start time	2014-06-07 09:31:09
Program release time	2014-06-03 21:51:00
Program end time	2014-06-03 22:47:00
EpgSeriesID	24952
EpgPID	6373423
StationID	327398

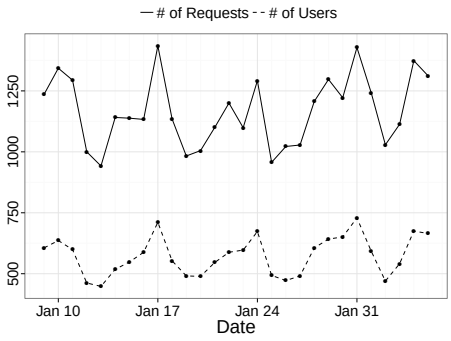


Figure 3. Daily pattern of dataset 1

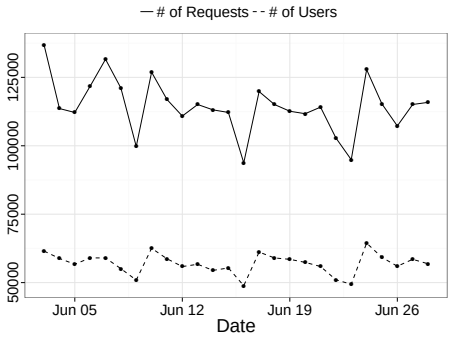


Figure 4. Daily pattern of dataset 2

2, one month of data from *dataset 1* was selected. As is shown in Figure 3 and Figure 4, on average, there are over 500 users online per day in *dataset 1* and over 50,000 users use the catch-up TV service in *dataset2* per day. Unlike the request pattern of VoD services such as YouTube in [28] in which the requests distributed evenly on weekdays, an interesting phenomenon here is that the number of users and the number of requests peaks on Thursdays in *dataset 1* and on Tuesdays in *dataset 2* and then drop rapidly. In *dataset 1*, each weekend has the least number of users and requests while in *dataset 2* users, the lowest usage is seen on Mondays.

Figure 5 and Figure 6 show the total number of requests and the total number of users for each hour of the day averaged over one month for both datasets. In *dataset 1*, peak hours are from 20:00 to 23:00 regardless of whether it is a weekday or a weekend. However, in *dataset 2*, multiple request peaks can be observed during the day. The peaks arrive earlier on weekends than on weekdays.

C. Video prefetching selection methods

In order to implement prefetching, the prediction engine needs to determine what content should be preloaded based on

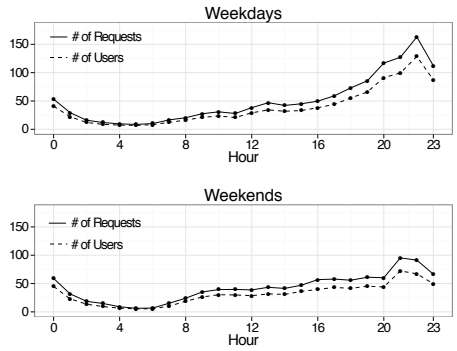


Figure 5. Average diurnal traffic patters in dataset 1

the user's viewing history of the same series. In the following, we propose and evaluate a scheme for how the prediction engine should make prefetching decisions.

In this study, we consider two extreme cases as baselines for evaluating the performance of the proposed prefetching scheme. One case is prefetching all of the available episodes in a series to the end user when the user has watched one episode in that series. In reality, the content provider has the information of the number of episodes in each TV series, both released and unreleased. In our dataset, the total number of episodes in each series is unknown, so we scan the users' viewing histories and collect unique video sets for each series. We assume that the videos in each set are all the available videos for each series. This coarse scenario may waste significant amounts of bandwidth since some of the content would never be accessed by users.

The other extreme scenario is using conventional terminal caching only, which means passively caching all of the user demand and preloading nothing prior to requests from the user. To distinguish this from prefetching, passive caching is used in the following. When passive caching is enabled, repeated requests for videos ,which have not been prefetched previously, can be served directly from the local cache. Passive caching can help to reduce initial delay only if the cached video is requested more than once.

Our approach is based on the intrinsic structure of TV content. Since each TV program contains a series of episodes that have high consistency and similarity in content, we propose to prefetch N adjacent episodes for each viewed episode. Unlike videos on traditional VoD websites such as YouTube, a TV series consists of a series of episodes, which will be released regularly. The user's request patterns of episodes in the series will be examined so that the prediction engine can decide which episodes to prefetch.

D. Cost model

Any prefetching scheme will make incorrect predictions and inevitably download more content than a system without prefetching. Consequently, the traffic overhead caused by prefetching may impose a burden on a bandwidth sensitive network. A congested network may lead to packet loss, longer transmission delay and poor quality of experience (QoE). Nervertheless, prefetching more content increases the probablity of meeting user demands and potentially reduces the user perceived latency. Sometimes spare network capacity is available during off-peak hours, e.g, during nights. It can be

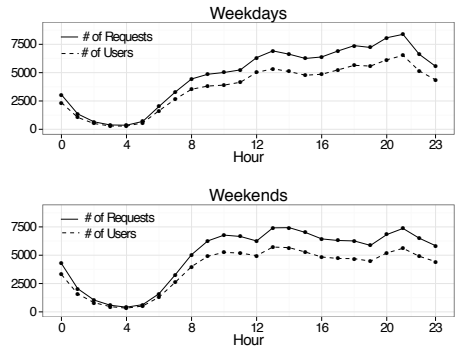


Figure 6. Average diurnal traffic patters in dataset 2

profitable to prefetch as much content as possible during that time to achieve high hit ratio. Hence, we propose a cost model to quantify the cost of prefetching in order to choose an optimal number of videos to prefetch.

We assume that the cost of real time video delivery equals 1 monetary unit per video. The cost of off-peak prefetching is x monetary units per video where x is smaller than 1 since prefetching can be done during off-peak hours costing less than real time downloading. In addition, the possible cost for poorer QoE can also be seen as a reason that real-time downloading is more expensive than prefetching.

We define the number of videos which are requested by each user but which is not prefetched as a video miss (M). The number of videos, which are prefetched by each user, is P. The subscript i indicates each user. The cost of prefetching for *k* users will then be:

$$C = 1 \cdot \sum_{i=1}^{k} M_i + x \cdot \sum_{i=1}^{k} P_i, \quad 0 < x < 1 \tag{5}$$

IV. Results

In this section, we present how to find the optimal number of episodes N to prefetch in order to achieve a high hit ratio that represents the potential of improving the response time.

We first use *dataset 1* to calculate the potential of the prefetching scheme. Then we compare the prefetching performances when different values of N are selected. The cost metric can be seen as a metric for measuring the performance of each prefetching scheme regarding transport cost, QoE cost and so forth. This aids in making optimal prefetching decisions when network condition change. Finally, we apply the same prefetching analysis to *dataset 2* for comparison.

A. The potential and benchmarks of prefetching

Before we go further into prefetching schemes, it is essential to evaluate the potential of prefetching.

In *dataset 1*, we assume a clairvoyant scheme. This means that once an episode in each series has been watched by a user, the following episodes in that series, which the user watches at a later time, can be 100% predicted and prefetched by the prefetching system. In this case, only the first requested video in each series cannot be predicted and preloaded. The precision of this prediction is 100%. The optimal recall in this scenario equals 73% calculated by equation (2), which means that in principle, 73% of the clicks to new videos are predictable. The result suggests that if all the episodes that are watched after the first one can be correctly predicted and stored in the local terminal cache, 73% of the requests to new videos will be served without delay. The corresponding F_1 score equals 0.84 that can be seen as the upper limit of the prediction effectiveness that our study can obtain based on *dataset 1*.

In order to evaluate the performance of a prefetching scheme, the two extreme cases described in Section III-C serve as benchmarks for comparison. First, we present the non-prefetching system, which only has passive caching enabled. The passive cache yields a hit ratio of 13.77% that means even with an infinite local cache, only 13.77% requests can be served locally. The result shows a great potential of prefetching since passive caching leaves over 85% of the requests unattended.

The other extreme case is to prefetch all the episodes in a series when an episode is watched. In this case, the cache hit ratio can reach up to 77%, which is the upper limit of the hit ratio that the prefetching system can achieve in this study. When all the episodes are prefetched, the maximum recall value of 73% is obtained. However, the precision is only 17% since the prefetching engine prefetched too much redundant data, which users are not interested in. As a result, the effectiveness of this prefetching scheme is only 0.28. Clearly, we need to intelligently select which content should be prefetched and stored.

B. Prefetch N episodes

In order to avoid prefetching too much data causing deteriorating network congestion, we propose to limit the number of videos to be prefetched by using a prefetching scheme, which only prefetches N videos in each series for each user.

Figure 7 shows the probability that a request for episode n will be followed by a request for episode $n+k$ as a function of k. We find that for a user who has watched episode n of a series, the probablity of that user watching episode $n+1$ next is over 50%. The number 0 on the x axis denotes that during the measurement period, there are approximately 26% of the users who will not watch any episode after watching episode n. According to Figure 7, prefetching videos with index of: $n+1$, $n+2$, $n+3$, $n-1$, $n+4$, $n+5$, and $n-2$ will account for 95% of the requests for a next video. Now, we need to decide the value of N and which videos to prefetch. In Figure 8, $N = 0$ represents passive caching when no videos are prefetched. We notice that when episode $n+1$ is prefetched, the hit ratio can reach 55%, which is a big improvement compared to the 13.7% hit ratio, which is reached using passive caching. To prefetch further videos after $n+1$ and $n+2$ gives very little increase in the hit ratio.

We also repeated the above analysis for different video categories. The request pattern for episodes in each video category were examined separately and we found that for the categories TV series and entertainment, the programs exhibit predictible consecutive request patterns while for the categories children's programs, documentaries and home and leisure time programs, a more random request pattern is exhibited. When we prefetch videos according to the request pattern for each video category, the hit ratio increase is about 1 percentage point compared to the results in Figure 8. Even if the improvement is not significant, which may due to the limited number of videos in each category, the impact of customizing prefetched videos

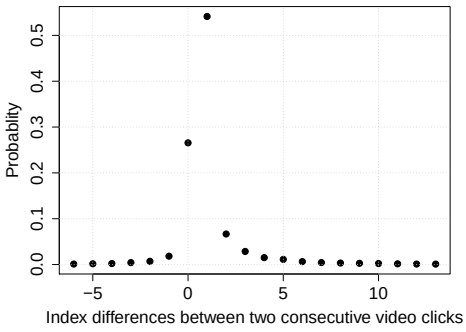


Figure 7. Transition probability of user requests within the same series in dataset 1

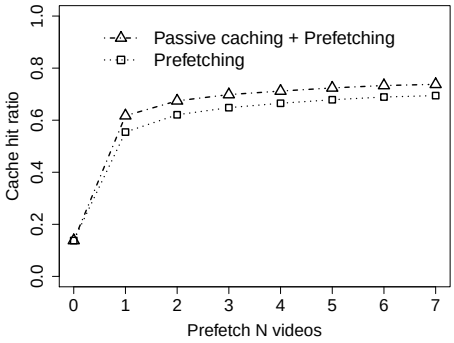


Figure 8. Hit ratios of terminal cache and prefetch in dataset 1

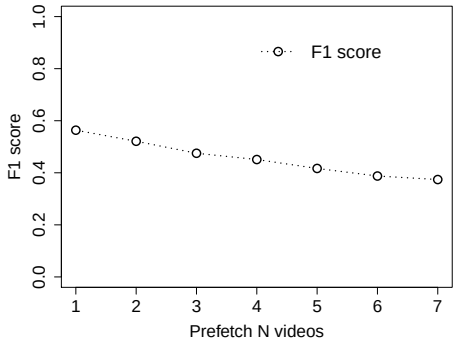


Figure 9. F1 score of prefetching system

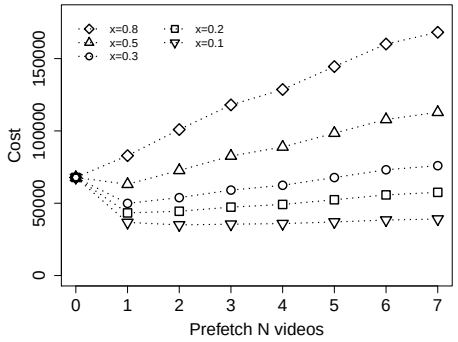


Figure 10. Total cost vs. Number of prefetched videos

according to video category should be further investigated by using larger datasets. In *dataset 2*, each video is tagged with multiple categories, thus, the categorization is too ambiguous to categorize videos for conducting a similar study.

In the lower curve in Figure 8, the videos, which are not prefetched, are not passively cached neither. They will be downloaded from the server during the playback and not stored locally. As is shown in the upper curve in Figure 8, the hit ratio increases about 6 percentage points which represents the gain for passively caching these videos. In our study, we treat all requests for the same video as cache hits. However, the users may watch part of a video and after some time continue to watch the rest. If we only treat the views after a user finishes watching the whole video as hits, the contribution from caching will be lower and the benefit brought by conventional passive caching will be more limited.

Now, we present the effectivness of prefetching schemes using varying values of N. Figure 9 shows the F_1 scores of the prefetching system when different N values are applied. In general, the more videos are prefetched, the less accurate and less effective the prefetching system will be. However, the hit ratio increases when more videos are prefetched, as shown in Figure 8. The decrease in effectiveness is caused by the amount of prefetched videos, which are not requested by users. If the network conditions are suitable to cope with this extra traffic, then a small hit ratio improvement with relatively large decrease of prefetching effectiveness is still profitable.

C. Cost

In this section, we calculate the cost of prefetching as a performance metric to find the optimal number of videos to prefetch.

Each point on the curves in Figure 10 shows the prefetching cost value (see Section III-D) versus the number of videos prefetched. The five curves represent the prefetching cost x in equation 5 equaling 0.8, 0.5, 0.3, 0.2, and 0.1, respectively. The total cost of passive caching when nothing is prefetched is shown when $N = 0$. Even though the top four curves show that the total cost of prefetching has a linear increase as more videos are prefetched, prefetching up to 7 videos is still cheaper than passive caching when off-peak downloading costs less than 30% of real time downloading. When off-peak downloading costs half of real time downloading, prefetching more than one video costs more than passive caching. However, in this case, prefetching only one video still outperforms passive caching. The cost for $x = 0.1$ is shown separately in Figure 11. An interesting phenomenon in this figure is that the total cost of prefetching declines when N increases from 1 to 2. This suggests that when prefetching two videos, the decrease in cost generated by a video miss (M) is larger than the cost increase generated by prefetching more videos. However, when N is larger than 2, the cost of prefetching starts to rise again. This indicates that there are more additional prefetched videos, which are not used by users. When the cost of off-peak downloading is 10% of the cost of real-time downloading, prefetching two videos yields the lowest cost.

D. Comparison

A similar study was conducted for *dataset 2*. First, we computed each user's requests and plotted the accumulated fraction of requests and the fraction of users. As can be seen in Figure 12, approximately 30% of the users generated 80% of requests. We define these 30% of the users in *dataset 2* as active users and in the following study, only the data from these active users is included.

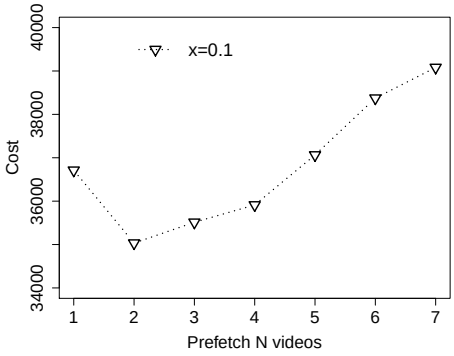


Figure 11. Total cost vs.Number of prefetched videos when cost factor = 0.1

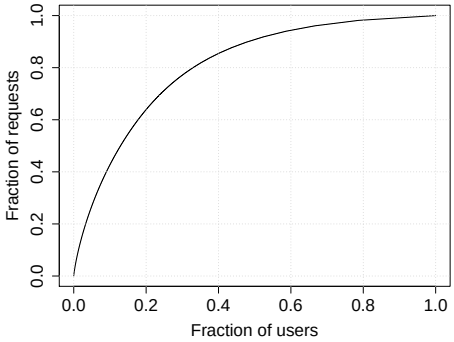


Figure 12. The distribution of requests between users

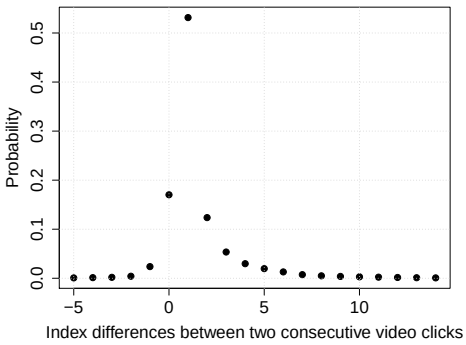


Figure 13. Transition probability of user requests within the same series in dataset 2

As was done for *dataset 1*, we examined the probability of what the user would watch next by using one month of data in *dataset 2*. As is shown in Figure 13, if episode n is watched by a user, then episodes with index of: $n+1$, $n+2$, $n+3$, $n-1$, $n+4$ have higher probabilities to be watched next. The probability that after watching episode n, no more episode from the same series will be requested by the user is nearly 20%.

Next, based on the request pattern, we applied our prefetching scheme on *dataset 2*. In Figure 14 is shown that if nothing is prefetched ($N = 0$), the hit ratio of passive caching is on average 27%. Prefetching the next episode, the hit ratio can reach up to 53%. With more than one episode prefetched, the hit ratio increase is more significant when two episodes with index $n+1$ and $n+2$ are prefetched. Prefetching 5 episodes for each request, a hit ratio of nearly 80% can be obtained, which is 3 times more than with passive caching. Even though the content and the group of users are different from those in

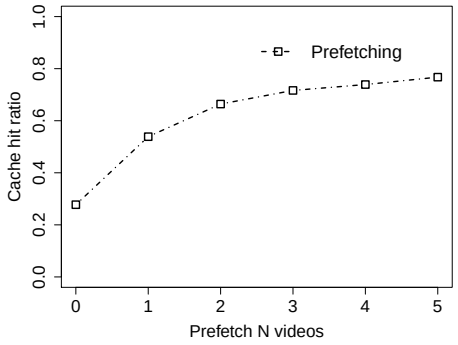


Figure 14. Hit ratios of terminal prefetching in *dataset 2*

dataset 1, our prefetching scheme can still be applied to *dataset 2*. From the results of both datasets, prefetching one episode can already greatly improve the cache hit ratio.

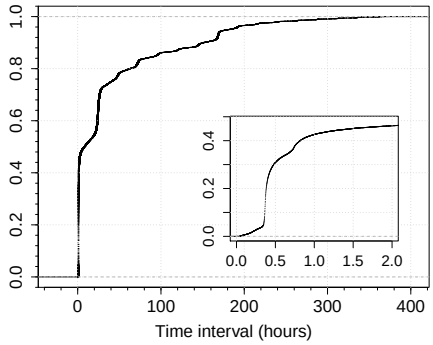
V. TIME TO PREFETCH

In this section, we estimate the available time for prefetching by measuring the time interval between two consecutive viewing sessions. Since in *dataset 2*, the length of each viewing session is unknown, we use only *dataset 1* in the following study.

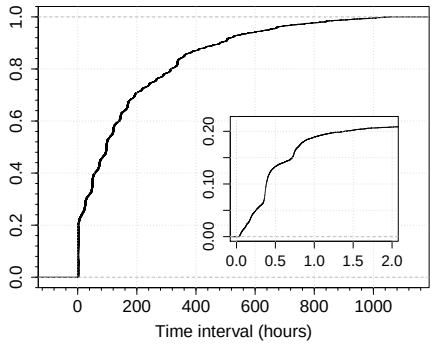
The time between the start of a viewing session and the start of the next one is defined as the upper bound for the time to make prefetching decisions. The time between the end of a viewing session and the start of the next session serves as the lower bound for the prefetching time. We differentiate between the behaviour of watching the $n+1$ video and watching any video not in sequential order and denote them as sequential views and non-sequential views correspondingly.

As is shown in Figure 15(a), two steep increases occur at 24h and 1 week respectively. Some programs are released on a daily or weekly basis and the increases suggest that a number of users follow these programs at the same pace. The inner graph shows the same CDF but zooms in the first two hours. This shows that 30% of the sequential views arrive within 20 minutes. Compared with the results in Figure 15(b) for non-sequential views, only approximately 15% of the requests are generated within 20 minutes. Figure 16 shows the CDF of the lower bound time. In general, it shows a similar trend as the one in Figure 15. The difference is seen in the inner figures, which suggests that about 40% of sequential views are generated within 1 minute after the end of a viewing session and only 15% of the non-sequential views are generated within the same time period. This indicates that if we choose to prefetch videos at the end of the current session, we have a limited time frame. As is shown in Figure 7, the risk of prefetching for sequential views is lower. Thus, it is more reasonable to prefetch the next video in order during the current session.

Prefetching for sequential views means prefetching one episode only beforehand, whereas prefetching for non-sequential views means prefetching more episodes after the next one in order. In this case, for example, when we decide to prefetch 2 episodes and episode n has already been requested, then episode $n+1$ and $n+2$ will be prefetched accordingly. If the user watches episode $n+1$ next, this requires episodes $n+2$ and $n+3$ to be prefetched, since episode $n+2$ is already

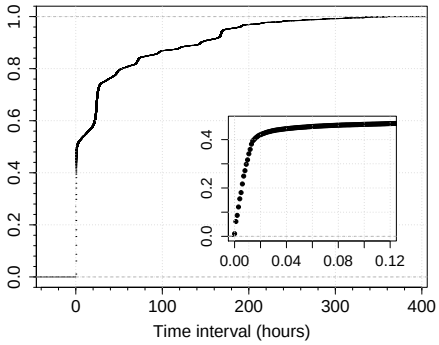


(a) Sequential views

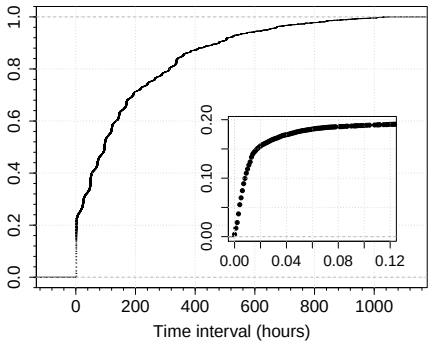


(b) Non-sequential views

Figure 15. CDF of time interval between two view events (Upper bound)



(a) Sequential views



(b) Non-sequential views

Figure 16. CDF of time interval between two view events (Lower bound)

stored in the cache. In this case there is a longer time period available to prefetch the second episode. For non-sequential views, the request pattern shows that people watch episodes from the same program on daily basis. This leaves us a longer time period for prefetching and we can delay until off-peak time.

Finally, we perform the same analysis for different video categories focusing particularly on the lower bound time. We observe that for TV series, 40% of the requests for the $n+1$ video come within 1 minute and 30% of the requests within 24 hours. This suggests that the next episode can be either pre-downloaded during the current episode's playtime or be downloaded during off-peak time. The entertainment programs show a similar pattern to the TV series. For children's programs, 60% to 70% of the requests are generated within 3 minutes no matter which episode is watched next. Similar patterns are observed for the videos of home and leisure time programs and documentaries. It suggests that in this case, prefetching during the video playback is more critical since the user has a high probablity to immediately request another video after watching the current one.

VI. Discussion

In this section, two factors that have potential to facilitate prefetching decisions are discussed. All the analysis in this section is based on *dataset 1*. We first investigate if the time of day can be used as a factor to adjust prefetching decisions. The other factor we discuss is the fraction of viewing length of a video. It implicitly reflects the users interests, which may be used as an indicator to predict the user's next move.

A. Temporal pattern of jumps

In this section, we examine whether a user's watching behaviour shows any preference depending on the time of day. Each day is divided into 5 time periods where period 1 is from 00:00 to 6:00, period 2 is from 6:00 to 12:00, period 3 is from 12:00 to 18:00, period 4 is from 18:00 to 21:00, and period 5 is from 21:00 to 00:00. The user's behaviour is categorized using *-1.0:* to denote watching the previous episode, *0:* to denote not watching any further episode within the same series, *1.0:* to denote watching the next episode and *IG:* to denote watching any other episodes within the same series. First, we plot the request distribution within each period in Figure 17. Not surprisingly, the most probable behaviour within each time period is to continue watching the next episode. The second highest possibility is not to request any further episode from the same series. Figure 18 shows the viewing behaviour distribution for the time periods as defined above. For instance, watching the previous episode is most likely to happen between 12:00 and 18:00 (time group 3) than during other time periods during the day. All the other behaviours are more preferable between 21:00 and 00:00 (time group 5), which is also the time of day when the users are most active. From Figure 17 and Figure 18, we can conclude that the jumps within each series do not show a clear temporal pattern in our dataset. Our observation confirms the results in [2], which suggest that if users watch an episode published in the late evening, they request further episodes from the same series in late afternoon of the following day. Thus, within time groups 3 to 5, the users may request episodes randomly from the same series or the next episode in order if it has been published.

Since watching next episode is the most probable be-

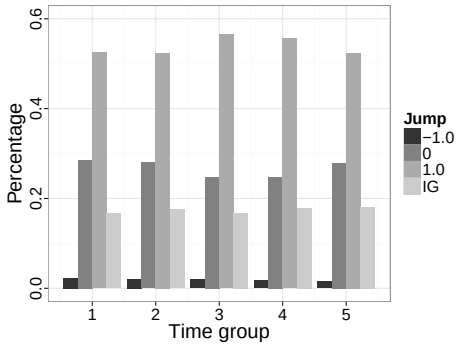


Figure 17. Jumps distribution within time periods

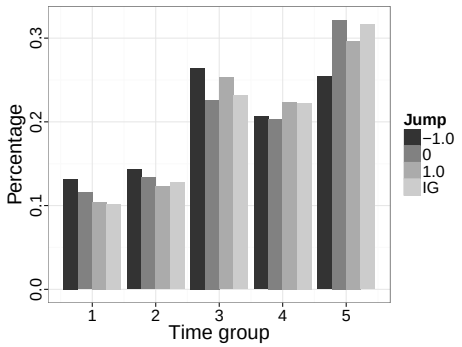


Figure 18. Jumps distribution over time periods

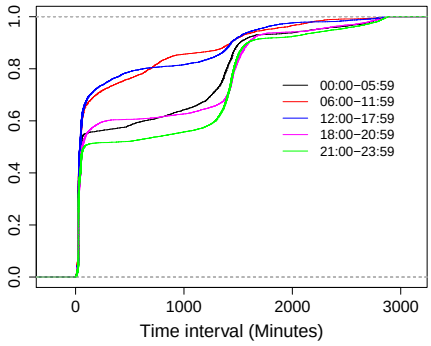


Figure 19. CDF of time interval watching n+1 video

haviour, we further investigate the time interval between the requests for episodes n and $n+1$. In Figure19, the longest time interval included in the graph was set to be two days and the CDF shows the cumulative distribution for the different time groups. It is clear that most requests for the $n+1$ episode are generated within the time period from 6:00 to 18:00 compared to the rest of the day. The three curves in the lower part of the figure show that nearly 40% of the requests for the next episode are generated within 24 hours. One explanation can be that people watch an episode regularly every day at a specific time and watch the next one the next day. The time of request also depends on the video availability. For instance, some series publish new episodes on a daily basis. Therefore, before new episodes are released, other older episodes from the same series have a larger chances of being requested. If the next episode has already been published, it accounts for 50% of the requests for the next episode being generated within a short time. However, in *dataset 1*, the video release time is unknown.

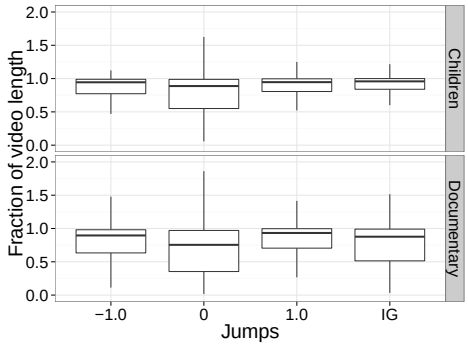


Figure 20. Boxplot of viewing length vs. jumps Part 1

B. *Viewing length distribution*

In the previous section, it was shown that the user viewing behaviour is almost evenly spread out across different time periods during the day. Having a model that shows whether the user behaviour is correlated with the viewing length of the current episode can help the prefetching system to make dynamic prefetching decisions as the video playback time accumulates. In this section, we investigate whether the viewing length of the current episode can be used as an indicator to predict what the user will watch next. Since the total video length is unknown in *dataset 1*, we assume that for each episode, at least one of the users has watched it until the end. Thus, the maximum viewing length of each episode over all the users who have requested it is estimated as the length of the episode.

We also investigate the viewing time fractions versus the watching behaviour for the different video categories. The viewing time of one episode is calculated by adding together the viewing length of consecutive requests for the same episode before the user requests the next episode. As can be seen in Figure 20, for childrens program, the users usually watch one episode until the end before requesting another one. However, if the user terminates the video before watching half of it, the user will not request any more episodes from the same series during the measurement period. Another interesting observation is, if a user watching the same episode more than once (the fraction of video length is larger than 1), then there is a higher probability that the user is only interested in this particular episode and will not watch anything more from the same series. Very similar behaviour can be observed for the documentary category. When users abort at the very beginning of the video playback, there is a higher probability that the user will not watch the next episode in order. Instead the user may browse other episodes from the same series. The short view time implies a low user interest, which is followed by a random viewing behaviour. The other three video categories have very similar patterns. The plot for TV series is shown in Figure 21 as an example. As the viewing time increases, it is hard to tell what the user's next move will be using only the viewing time as a factor.

VII. Conclusions

In this paper, we have proposed a prefetching scheme and performed analysis to evaluate its performance based on a dataset from a Swedish TV-on-demand service (*dataset 1*) in order to explore the potential of reducing start-up latency of streaming media services. The same method is applied to a

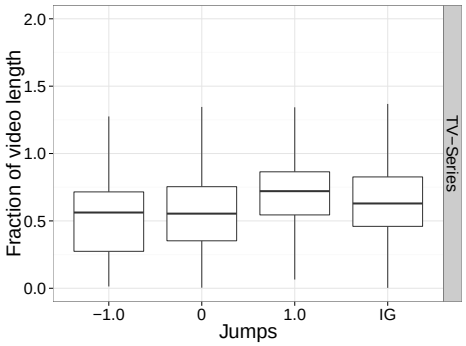


Figure 21. Boxplot of viewing length vs. jumps Part 2

second dataset from a catch-up TV service, (*dataset 2*), for comparison purposes and further analysis.

First, the paper has demonstrated that for *dataset 1,* 73% of the requests are predictable in an ideal scenario with 100% prediction accuracy. This suggests a great potential for prefetching. We propose to use the intrinsic structure of TV series in our dataset and prefetch N adjacent videos to terminal devices. A cost model is proposed to quantify the cost of prefetching and to provide an optimal solution for prefetching. The result shows that prefetching two adjacent videos yields 62% hit ratio, which is more than four times as much as terminal caching can obtain. We also demonstrate that with the simple prefetching scheme we propose, 69% of all the requested videos can be correctly predicted, a number which is very close to the ideal value of 73%. When prefetching costs 10% of the cost of real time downloading, prefetching two videos has the lowest cost and is the optimal choice for prefetching based on the dataset analyzed in this study. By comparing the prefetching results using two different datasets, we find that the more videos that are prefetched, the higher is the obtained hit ratio. This implicates that more requests can be served directly from the local cache with a short delay. Even though the two datasets we used are from different regions and different types of TV services, the prefetching scheme can easily be applied to both of them and the hit ratio improvement is significant.

We also found that the time of day suitable for prefetching depends on the user request patterns. For TV series, it is more reasonable to prefetch the next episode before the end of the current viewing session. For non-sequential requests, videos can be prefetched during off-peak hours. For programs, which have a more random request pattern, such as children's programs, it is better to make prefetching decisions during the current video playback time, even for non-sequential requests.

Two more factors are discussed as possible improvements for making prefetching decision. The first one is that user viewing behaviour does not show any strong preference regarding the time of a day. Approximately 10% more requests for the next episode is generated within a short time during the day from 6:00 to 18:00. From 18:00 to 0:00, the later during the day that the user watches the current episode, the higher probability that he will request the next episode in order within 24 hours. We also studied whether the viewing length of each video has any influence on what the user would watch next. For future work, more factors such as the number of episodes from the same series, which have already been watched by the user should also be considered to improve the prediction accuracy and hit ratio.

In this work, only viewing sessions longer than 2 minutes can trigger prefetch. However, users may be less tolerant to delays in short sessions than in long sessions. Thus, using prefetching to improve performance for these short sessions is worth being further investigated. Another prediction limit of our study is the number of first seen episodes in each series. In future work, we plan to use *dataset 2* to extend our research into cluster-based prefetching mechanisms to find user clusters and to make prefetching decisions based on similarity in user behaviour to predict even the first seen episode in each series.

ACKNOWLEDGMENT

This work has partly been financed by the Swedish Governmental Agency for Innovation Systems(VINNOVA) in the EFRAIM project and the NOTTS project. Maria Kihl is a member of the Lund Center for Control of Complex Engineering Systems(LCCC) and the Excellence Center Linkping - Lund in Information Technology(eLLIIT)

REFERENCES

[1] M. Du, M. Kihl, Å. Arvidsson, C. Lagerstedt, and A. Gavler, "Analysis of prefetching schemes for tv-on-demand service," in Proceedings of the Tenth International Conference on Digital Telecommunications, 2015, pp. 12–18.

[2] Y. Elkhatib, M. Mu, and N. Race, "Dataset on usage of a live amp; vod p2p iptv service," in Proceedings of Peer-to-Peer Computing (P2P), 14th IEEE International Conference, 2014, pp. 1–5.

[3] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng, "Understanding user behavior in large-scale video-on-demand systems," in Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems. ACM, 2006, pp. 333–344. [Online]. Available: http://doi.acm.org/10.1145/1217935.1217968

[4] M. Cha, P. Rodriguez, J. Crowcroft, S. Moon, and X. Amatriain, "Watching television over an ip network," in Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement. ACM, 2008, pp. 71–84. [Online]. Available: http://doi.acm.org/10.1145/1452520.1452529

[5] A. Ali-Eldin, M. Kihl, J. Tordsson, and E. Elmroth, "Analysis and characterization of a video-on-demand service workload," in Proceedings of the 6th ACM Multimedia Systems Conference. ACM, 2015, pp. 189–200. [Online]. Available: http://doi.acm.org/10.1145/2713168.2713183

[6] F. F.-H. Nah, "A study on tolerable waiting time: how long are web users willing to wait?" Behaviour & Information Technology, vol. 23, no. 3, 2004, pp. 153–163.

[7] D. F. Galletta, R. Henry, S. McCoy, and P. Polak, "Web site delays: How tolerant are users?" Journal of the Association for Information Systems, vol. 5, no. 1, 2004, pp. 1–28.

[8] W. Ali, S. M. Shamsuddin, and A. S. Ismail, "A survey of web caching and prefetching," Int. J. Advance. Soft Comput. Appl, vol. 3, no. 1, 2011, pp. 18–44.

[9] J. Xu, J. Liu, B. Li, and X. Jia, "Caching and prefetching for web content distribution," Computing in Science Engineering, vol. 6, no. 4, July 2004, pp. 54–59.

[10] Y. Jiang, M.-Y. Wu, and W. Shu, "Web prefetching: Costs, benefits and performance," in Proceedings of the 7th international workshop on web content caching and distribution, ser. WCW, 2002.

[11] J. Domenech, J. Sahuquillo, J. A. Gil, and A. Pont, "The impact of the web prefetching architecture on the limits of reducing user's perceived latency," in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2006, pp. 740–744. [Online]. Available: http://dx.doi.org/10.1109/WI.2006.166

[12] J. Marquez, J. Domenech, J. Gil, and A. Pont, "Exploring the benefits of caching and prefetching in the mobile web," in Second IFIP Symposium on Wireless Communications and Information Technology for Developing Countries, 2008.

[13] M. Deshpande and G. Karypis, "Selective markov models for predicting web page accesses," ACM Trans. Internet Technol., vol. 4, no. 2, 2004, pp. 163–184. [Online]. Available: http://doi.acm.org/10.1145/990301.990304

[14] X. Chen and X. Zhang, "Popularity-based ppm: an effective web prefetching technique for high accuracy and low storage," in Proceedings of International Conference on Parallel Processing, 2002, pp. 296–304.

[15] D. Joseph and D. Grunwald, "Prefetching using markov predictors," SIGARCH Comput. Archit. News, vol. 25, no. 2, 1997, pp. 252–263. [Online]. Available: http://doi.acm.org/10.1145/384286.264207

[16] Z. Ban, Z. Gu, and Y. Jin, "An online ppm prediction model for web prefetching," in Proceedings of the 9th Annual ACM International Workshop on Web Information and Data Management. ACM, 2007, pp. 89–96. [Online]. Available: http://doi.acm.org/10.1145/1316902.1316917

[17] G. Pallis, A. Vakali, and J. Pokorny, "A clustering-based prefetching scheme on a web cache environment," Computers & Electrical Engineering, vol. 34, no. 4, 2008, pp. 309–323.

[18] W.-G. Teng, C.-Y. Chang, and M.-S. Chen, "Integrating web caching and web prefetching in client-side proxies," Parallel and Distributed Systems, IEEE Transactions on, vol. 16, no. 5, 2005, pp. 444–455.

[19] Z. Su, Q. Yang, and H.-J. Zhang, "A prediction system for multimedia pre-fetching in internet," in Proceedings of the Eighth ACM International Conference on Multimedia. ACM, 2000, pp. 3–11. [Online]. Available: http://doi.acm.org/10.1145/354384.354394

[20] C. Bouras, A. Konidaris, and D. Kostoulas, "Predictive prefetching on the web and its potential impact in the wide area," World Wide Web, vol. 7, no. 2, 2004, pp. 143–179.

[21] S. Khemmarat, R. Zhou, D. K. Krishnappa, L. Gao, and M. Zink, "Watching user generated videos with prefetching," Image Commun., vol. 27, no. 4, 2012, pp. 343–359. [Online]. Available: http://dx.doi.org/10.1016/j.image.2011.10.008

[22] D. K. Krishnappa, S. Khemmarat, L. Gao, and M. Zink, "On the feasibility of prefetching and caching for online tv services: A measurement study on hulu," in Proceedings of the 12th International Conference on Passive and Active Measurement, 2011, pp. 72–80. [Online]. Available: http://dl.acm.org/citation.cfm?id=1987510.1987518

[23] Z. Zeng and B. Veeravalli, "Hk/t: A novel server-side web caching strategy for multimedia applications," in IEEE International Conference on Communications, 2008, pp. 1782–1786.

[24] Z. Zeng, B. Veeravalli, and K. Li, "A novel server-side proxy caching strategy for large-scale multimedia applications," J. Parallel Distrib. Comput., vol. 71, no. 4, 2011, pp. 525–536. [Online]. Available: http://dx.doi.org/10.1016/j.jpdc.2010.06.008

[25] S. Seny, J. Rexford, and D. Towsley, "Proxy prefix caching for multimedia streams," in Proceedings of Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 3, 1999, pp. 1310–1319 vol.3.

[26] W. Liu, C. T. Chou, Z. Yang, and X. Du, "Popularity-wise proxy caching for interactive streaming media," in Proceedings of 29th Annual IEEE International Conference on Local Computer Networks, 2004, pp. 250–257.

[27] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in Proceedings of the 27th European Conference on Information Retrieval, 2005, pp. 345–359.

[28] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "Youtube traffic characterization: A view from the edge," in Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement. ACM, 2007, pp. 15–28. [Online]. Available: http://doi.acm.org/10.1145/1298306.1298310

An Analytical Model and an Efficient Tool to Predict the Availability of IPTV Services in Vehicle-to-Infrastructure Networks

Bernd E. Wolfinger [1], Nico R. Wilzek [1], Edgar E. Báez [2]

[1] Department of Computer Science,
Telecommunications and Computer Networks
University of Hamburg
Hamburg, Germany
e-mail: wolfinger@informatik.uni-hamburg.de,
4wilzek@informatik.uni-hamburg.de

[2] Superior School of Computing
National Polytechnic Institute ESCOM-IPN,
Mexico City, Mexico
e-mail: ebaeze0700@alumno.ipn.mx

Abstract— **Entertainment services such as IP television (IPTV) are becoming increasingly important in vehicular ad-hoc networks (VANETs), which implies a strong need for quality of experience (QoE) studies. Therefore, in this paper we introduce analytical models to predict QoE of IPTV in VANET scenarios. Unlike earlier QoE research for IPTV, which has been mainly related to audio/video quality, our focus is on IPTV service availability. To evaluate our analytical models we offer a tool (ACTIVE) that can be used in a flexible and efficient manner. The paper also describes the architecture and the main components of the tool as well as its graphical user interface. Case studies based on our ACTIVE tool demonstrate how our models can be applied by a provider of IPTV services in order to satisfy QoE requirements regarding the service availability as given by the IPTV users. Moreover, some of the results are also of interest to the individual IPTV users.**

Keywords- Vehicular networks; IPTV; QoE; service availability; analytical model; model evaluation tool; validation.

I. INTRODUCTION

Current predictions for the car market claim that, in 2016, more than 80 % of all new cars sold will have access to the Internet (e.g., FOCUS Online [9]). Therefore, one can expect that the usage of Internet services by car passengers will become more and more wide-spread in the near future. Besides search-, information- and communication services also entertainment services (such as IPTV or Video-on-Demand) will probably play a significant role [5]. For that reason, quality assessment of Internet services with real-time requirements (as they are present, e.g., in IPTV services offered in vehicular ad-hoc networks – or VANETs for short) is getting increasingly important. Therefore, this topic is in the main focus of this paper.

Quality of service provisioning is relevant, in particular as it is experienced by the (human) end-users and thus it is denoted by Quality of Experience (QoE) [16]. In case of IPTV services, on the one hand, QoE refers to the quality of the received audio/video stream as perceived by the end-user. So, most of the existing studies concerned with QoE in the context of audio/video communications with real-time constraints have been related to audio-visual quality, which is judged by well-known QoE measures, such as PESQ/PEAQ/PEVQ (perceptual evaluation of speech/audio/video quality) [7] or by means of MOS (mean opinion score), a method that directly relies on the subjective judgement of the human end-users [25]. Belyaev et al. [6] propose an interesting approach for a dynamic adaptation of the video bit rate in order to maintain a certain level of video quality in a scenario of a vehicular video surveillance system (based on IEEE 802.p [12]). Here, QoE is evaluated in terms of visual quality and its impairment by packet losses.

Of course, the audio-visual quality is also relevant in the context of vehicular networks, more so, because the TV channels are offered to the corresponding car passengers via wireless access networks. And this may have a strongly negative impact on the quality of the stream delivered to the IPTV users.

Zhou et al. [28] measure user-satisfaction when users access media services via peer-to-peer (P2P)-based VANETs. In particular, they propose a scheme that solves content dissemination, cache update and fairness problems for P2P-based VANETs. However, unlike our studies presented in this paper, Zhou et al. do not consider IPTV services, nor do they assume multicast for the provisioning of the media services.

Besides the audio/video quality of the delivered TV channels, QoE is also influenced strongly by the delay it takes to switch from one channel to another, which usually is called *channel switching delay*. Several studies try to reduce these switching delays, cf. [2].

Last but not least, QoE in vehicular networks also comprises the degree of availability with which the user is able to access the IPTV service [15]. From the users' point of view the availability of TV channels may be even more important than the audio-visual quality of the IPTV service. As a measure of availability, we will take the probability that a desired TV channel can indeed be provided to the corresponding user though the bandwidth in the (access) network may be quite limited. Availability studies for IPTV services have been done in the past (by means of simulation models) in particular for DSL based access networks [14] as well as for WiMAX based access networks [2].

As currently no vehicular networks offering IPTV services are available to us for carrying out measurements, the only alternative for corresponding service availability studies is the use of models. To the best of our knowledge, up to now, only very few models exist, which allow one to predict the availability of IPTV services in VANETs. Detailed simulation models – and not analytical models as in the current paper – have been elaborated and applied in case studies by Momeni et al. [17]. Moreover, in recent past, first successful trials have been undertaken to predict IPTV availability in VANETs by means of analytical models, cf. Wolfinger et al. [27]. Other existing analytical models to predict QoE for IPTV, such as [11], [19], and [21], have a completely different emphasis: they are again concerned with audio-visual quality and not with channel availability. Moreover, in [11], users are watching TV via home networks and are not mobile at all and, in [19] the emphasis is on (low level) QoS and not on QoE, and finally, in [21], an architecture of an IPTV system is suggested for mobile devices, which is shown to satisfy some basic QoS/QoE requirements. Even most of the simulation tools that exist for studying vehicular networks [23] are quite useless as a basis for doing availability analyses, because they include very detailed submodels for the network services used but, on the other hand, user behavior is reflected in a rather superficial manner.

In a recently published paper [1], we significantly extended the results of [27] as we carried out an in-depth validation of the analytical model and, as a major new contribution, we presented a generalized procedure that allows us to predict the IPTV availability in a straight-forward manner for very different traffic scenarios and network technologies. We also applied our procedure in various comprehensive case studies. The current paper is a thoroughly revised and considerably extended version of our earlier publication [1]. We now not only present a class of analytical models but also various upper bounds for the unavailability of requested TV channels. Moreover, the newly developed ACTIVE tool is introduced, which is able to predict the availability of an IPTV service in vehicular networks in a very flexible and highly efficient manner. The ACTIVE tool not only provides the values for availability measures being of interest to providers of IPTV services but also those ones that are related to the channel availability as experienced by individual IPTV users. The availability measures cover both, blocking of TV channels during handover as well as during switching events.

The paper is structured as follows: Section II will give a short overview on IPTV services offered via VANETs including the availability measures that we will apply. The analytical model used will be introduced in Section III followed, in Section IV, by a thorough validation of this model. Our analytical models will be embedded in a model evaluation tool (ACTIVE), the architecture and user interface of which will be described in detail in Section V. A generalized procedure for a highly efficient usage of our models and the ACTIVE tool then is presented in Section VI. Application of the generalized procedure will be illustrated in the case studies of Section VII. These studies also show how our model can support a provider of an IPTV service (offered via a VANET) in dimensioning and configuring a network that satisfies the given QoE requirements of the IPTV subscribers.

II. IPTV Services in VANETs and Availability Measures for their Assessment

A. Provisioning of IPTV Services in VANETs

Two main classes of vehicular networks [20] are typically distinguished: networks supporting *vehicle-to-vehicle* (V2V) and those supporting *vehicle-to-infrastructure* (V2I) communication. V2V infrastructures are mainly used to improve vehicular safety [4]. For our studies, only V2I configurations are relevant because communication between vehicles is not of interest to us. V2I communication can be achieved in two variants that differ in the way how users in the vehicles can get access to the Internet: in the first variant (V1), the mobile station (e.g., a smart phone) could be communicating via a non-IP-based public mobile network and from there get access to the Internet. In the second variant (V2), the mobile station would access a dedicated road-side unit (RSU) via the base station (BS) / the access point (AP) of its local cell and from there get direct access to IP based routers (cf. proposal and prototype for so-called road-side backbone networks using RSUs to interconnect the Internet with the vehicles as described, e.g., by Gladisch et al. [10] and Krohn et al. [13]). In this paper, we assume that the IPTV services we analyze are provided in networks in which Internet access is established according to variant V2. Different network technologies (such as WLAN, LTE, WiMAX) can be used in principle to achieve communication between the mobile stations (in the vehicles) and the base station resp. access point in the corresponding cell. From point of view of IPTV service, provisioning different network technologies in the access network can have a strong impact on the service quality because they will typically support very different data rates and lead to very different cell sizes.

In the vehicular networks, we investigate the fact that ad-hoc networking is possible between vehicles is not really important for us. On the contrary, we are mainly interested in the delivery of IPTV services to the vehicles by means of vehicle to infrastructure (V2I) communications. Nevertheless, we argue that the IPTV service delivery studied in this paper does not only cover vehicular networks, but also VANETs and, accordingly, we use the formulation "IPTV Services in VANETs" throughout this paper.

If an IPTV service [26] is offered in a network with V2I communication, where Internet access is achieved by means of RSUs (as assumed in our studies), the basic network architecture will comprise the main components as depicted by Fig. 1:

- the IPTV Head-end, where all the TV channels are available that can be demanded by the IPTV users,
- that part of the Internet that is used to make communication between the Head-end and the set of

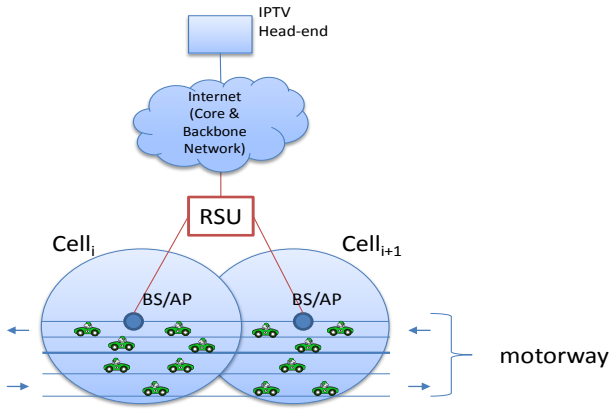


Figure 1. Basic architecture of an IPTV system for users of vehicular networks

RSUs possible (this subsystem could be the IP based network of an ISP providing the IPTV service),

- the access network representing the infrastructure for communication between an RSU and the mobile stations within the cells for which RSU is responsible.

Provisioning of IPTV services typically makes use of multicast (e.g., IP multicast [22]) leading to the advantage that a TV channel having been desired in an access network has only to be provided once by the corresponding RSU even in the case that the TV channel is currently watched by more than one user in this cell. A TV channel is no longer transmitted in a cell as soon as the last user watching this channel releases the channel (e.g., because he/she switches to another channel or is involved in a handover thus leaving the cell or the user may temporarily terminate usage of the IPTV service).

As a consequence of the limited data rate (bandwidth) of the cells representing the access networks, it is, of course, possible that a TV channel newly desired by a user cannot be provided at that moment when the request for the channel is issued. This happens exactly in the case that the desired channel currently is not yet delivered in the cell AND the total transmission capacity available for IPTV is completely exhausted currently because of having to transmit other channels. If a request for channel delivery has to be denied, we say that the channel is "blocked" for the user and call this event a *"blocking (event)"*.

So, we see that studies of IPTV service availability in VANETs that are based on detailed models will require that the corresponding models reflect

- the bandwidth utilized for IPTV at any instant,
- the list of TV channels currently being multicast in the corresponding cell,
- the behavior of the IPTV users in terms of the time instants at which TV channels are switched/changed and in terms of the id. (e.g., channel number) of the channel newly demanded.

Former investigations with respect to a realistic characterization of IPTV user behavior [2] [3] have shown that the popularity of TV channels can be approximated quite well by Zipf distributions [18].

In particular, the probability p_i that the i-th popular channel is requested is determined by the Zipf distribution as follows:

$$p_i = \frac{\frac{1}{i^\theta}}{\sum_{k=1}^{N} \left(\frac{1}{k^\theta}\right)} \quad (1)$$

where N denotes the total n^o of different channels offered, k is their rank and θ is the Zipf parameter reflecting the degree of popularity skew. A value of $\theta = 1.3$ is realistic according to measurements of IPTV user behavior [2].

B. Measures for IPTV Availability

The following two reasons exist that an IPTV user will demand a TV channel within a cell:

(1) A *channel-switching event*: Here, the user will demand a new channel to which he currently switches to (e.g., because he is "zapping" through a sequence of channels at time durations of just a few seconds or after he terminates a "viewing phase" with duration of several minutes or even hours during which he has received and watched just a single TV channel).

(2) A *handover event*: Here, the car will change the cell and, as a consequence, the channel currently received by a user in this car will no longer be needed by him in the "old" cell left but it will be needed in the "new" cell reached now.

In both cases, blocking of the desired channel may occur. Thus, we distinguish:

– *switching-induced* or *switching-related blocking*, and

– *handover-induced* or *handover-related blocking*.

Therefore, three channel blocking probabilities are of interest to us:

– *Channel Blocking Probability* (CBP) referring to all blocking events

– *Switching-induced Blocking Probability* (SBP) referring only to blockings being a consequence of channel switching

– *Handover-induced Blocking Probability* (HBP) referring only to blockings being a consequence of handover events.

As it is usual, we can approximate the three probabilities by the relative frequencies of the corresponding blockings choosing an observation interval that is sufficiently large.

Let T = [t1, t2] denote the observation interval and |T|=t2-t1 its length.

Let further denote:

- #r(T): n^o of all channel requests issued by all users in T
- #r$_h$(T): n^o of all handover-related requests in T
- #r$_s$(T): n^o of all switching-related requests in T

- #b(T): n^o of all blocked requests (blockings) in T
- #b_h(T): n^o of all handover-related blockings in T
- #b_s(T): n^o of all switching-related blockings in T.

Based on these variables, we can now define the following channel blocking frequencies for the interval T:

- CBF(T) $\triangleq \frac{\#b(T)}{\#r(T)}$ denoting the *overall channel blocking frequency*
- HBF(T) $\triangleq \frac{\#b_h(T)}{\#r(T)}$ denoting the *relative frequency of handover-related blockings*
- SBF(T) $\triangleq \frac{\#b_s(T)}{\#r(T)}$ denoting the *relative frequency of switching-related blockings.*

Evidently,

$$HBF(T) + SBF(T) = \frac{\#b_h(T)}{\#r(T)} + \frac{\#b_s(T)}{\#r(T)} = \frac{\#b_h(T) + \#b_s(T)}{\#r(T)} = \frac{\#b(T)}{\#r(T)} = CBF(T) \quad (2)$$

and – as the relative frequency converges to the probability for an interval length |T| tending to infinity:

$$CBP = \lim_{|T|\to\infty} CBF(T)$$
$$HBP = \lim_{|T|\to\infty} HBF(T)$$
$$SBP = \lim_{|T|\to\infty} SBF(T)$$

which implies that also CBP = HBP + SBP holds.

Instead of CBP we can alternatively use

$$CA \triangleq 1 - CBP$$

denoting the overall *channel availability*.

III. An Analytical Model to Predict TV Channel Availability

In [27], an analytical model was elaborated, which is the basis of this paper. This analytical model is used to determine CBP and it is able to take into account various traffic scenarios, access network technologies and IPTV service characteristics.

To present this model in this section and in the following sections, we use the variables and model parameters as introduced in Table I.

The basic ideas underlying the analytical model are the following ones:

(1) Calculate the probability that, for a given cell c, currently all bandwidth BW_c available for IPTV service is used to distribute required TV channels. In such a situation of lacking free bandwidth, the demand for a new channel (which is currently not yet transmitted in cell c) may have to be denied. This means that the transmission of the newly demanded TV channel may become blocked. Therefore, we call the cell as being in a *"potential blocking state"*.

(2) Calculate the probability that a currently unavailable channel is demanded when the cell is in a "potential blocking state".

TABLE I. LIST OF PARAMETERS AND VARIABLES USED

	Variable/ parameter	Meaning
Traffic-related variables	k	number of lanes per direction
	v_i	speed of vehicles on lane L_i assumed to be constant for this lane (in [km/h])
	d_i	distance between adjacent vehicles on L_i assumed to be constant for this lane (in [m])
	$\bar{d}$	mean distance between adjacent vehicles (averaged over all lanes)
Cell-related variables	C_r	radius of cell (in [km])
	BW_c	bandwidth available for IPTV service in cell c
	N_c	number of IPTV users in cell c
IPTV-related variables	N	number of TV channels offered in total
	α	percentage of vehicles using IPTV
	p_i	probability that channel i is required (according to Zipf distribution with parameter θ)

Calculation of CBP in our analytical model is based on the following four steps:

- STEP 1: Determine the probabilities P_i that, for given N and N_c, exactly i different channels are needed to satisfy the channel requests of N_c users, if N different channels are offered. P_i can be estimated by the relative frequency f_i that N_c users require exactly i different channels, where f_i can be determined in a straight-forward manner by means of Monte Carlo simulation [8], [24]. Throughout this paper, all of our Monte Carlo experiments are repeated one million times and, therefore, the size of the sample to calculate f_i is 10^6.
- STEP 2: Assume a certain cell bandwidth BW_c available for IPTV and determine P^* as probability that N_c users require more than BW_c different TV channels. So, P^* denotes the probability that the system is in a "potential blocking state".
- STEP 3: Assume that an IPTV user will require a new channel (channel number determined according to Zipf distribution) and determine the probability that the number of the channel demanded is larger than BW_c, which happens with probability

$$\sum_{i > BW_c}^{N} p_i \quad (3)$$

- STEP 4: We determine the probability (*CBP*) that a newly requested channel cannot be delivered, which happens with probability

$$CBP = P^* \bullet \sum_{i>BW_c}^{N} p_i , \quad (4)$$

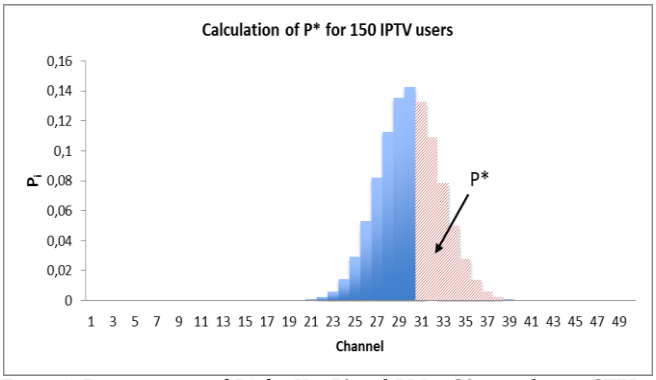


Figure 2. Determination of P^* for $N = 50$ and $BW_c = 30$ according to STEPs 1 and 2 of our calculation algorithm for the analytical model ($N_c = 150$).

if we make the favorable assumption that, in case of a "potential blocking situation (state)", exactly those channels are transmitted in the corresponding cell, which are the BW_c most popular ones.

Remark: It should be noted that, astonishingly, the (favorable) assumption that "if the system is in a potential blocking state then just the most popular channels are transmitted" is quite realistic indeed. This has been observed by us in simulation experiments based on detailed models of IPTV services in VANETs (cf. simulation models described in [17] and also used during our model validation in Section IV). □

Fig. 2 illustrates STEP 1 and STEP 2, by way of example, if we assume $N = 50$, $BW_c = 30$, $N_c = 150$. This figure depicts the histogram for P_i , $i \in \{1, 2, \ldots, 50\}$.

IV. Model Validation

What is left is the validation of our analytical model. We validate it by means of simulation and care mainly about the late (stationary) phase and situations where CBP ≤ 0.1, because we assume that if CBP > 0.1 this means that QoE is too low anyway and, therefore, model accuracy is not really important for those cases.

We validate the model by means of two series of experiments and observed rather good agreement between the analytical model and the simulation results. Therefore, we consider the analytical model as being sufficiently realistic.

Of course, our validation phase is limited by the fact that we do not have any access to measurements regarding IPTV service availability in vehicular networks because those systems currently do not yet exist. So, we find it acceptable to rely on IPTV service availability predictions based on a detailed and (hopefully) sufficiently realistic simulation model.

Comparisons with alternative analytical models for CBP predictions related to IPTV services in VANETs, developed by other researchers, have not been possible for us as, to the best of our knowledge, no such models do yet exist.

A. Series I of Validation Experiments

In Series I, we changed N_c (the number of users in the cell) and kept N (the number of channels available) and BW_c (the maximum number of channels that can be broadcasted at the same time) constant per set of experiments, with $N = 50$ and $BW_c = 30$ for *set 1* of Series I and $N = 100$ and $BW_c = 40$ for *set 2*. As can be seen in Table II, the analytical model and the simulation model are matching quite well with a few outliers at $N_c = 200$ in both sets. Also, the values of the analytical model in set 1 do not increase as fast as the values of the simulation model (with increasing N_c).

B. Series II of Validation Experiments

In Series II, we kept the number of users per cell constant ($N_c = 300$ for *set 1* of Series II and $N_c = 400$ for *set 2* of Series II). We changed N (the number of channels available) and BW_c (the maximum number of channels that can be broadcasted at the same time). Again, we observe a rather good agreement between the analytical model and the simulation results, with a few outliers at higher values for N, where the analytical model is a close upper bound; for details regarding the deviations, see Table III.

TABLE II. SERIES I OF VALIDATION EXPERIMENTS

Series I: Set 1 N=50 BW_c=30				
	CBP		**Deviation**	
N_c	**AM**	**SM**	**Relative [%]**	**Absolute**
100	0.0011	0.0024	-118.1818	-0.0013
200	0.0506	0.0339	33.0040	0.0167
300	0.0577	0.0549	4.8527	0.0028
400	0.0578	0.0615	-6.4014	-0.0037
500	0.0578	0.0649	-12.2837	-0.0071

Series I: Set 2 N=100 BW_c=40				
	CBP		**Deviation**	
N_c	**AM**	**SM**	**Relative [%]**	**Absolute**
100	0.00008	0.00009	-12.5000	-0.00001
200	0.0680	0.0327	51.9118	0.0353
300	0.0846	0.0642	24.1135	0.0204
400	0.0843	0.0832	1.3049	0.0011
500	0.0843	0.0869	-3.0842	-0.0026

TABLE III. SERIES II OF VALIDATION EXPERIMENTS

Series II: Set 1 N_C=300				
	CBP		Deviation	
N, BW_C	AM	SM	Relative [%]	Absolute
20, 10	0.1157	0.1364	-17.8911	-0.0207
20, 15	0.0456	0.0501	-9.8684	-0.0045
50, 20	0.1099	0.1222	-11.1920	-0.0123
50, 30	0.0577	0.0529	8.3189	0.0048
75, 30	0.0942	0.0956	-1.4862	-0.0014
75, 40	0.0607	0.0424	30.1483	0.0183
75, 50	0.0074	0.0057	22.9730	0.0017
100, 50	0.0473	0.0251	46.9345	0.0222
100, 60	0.0016	0.0017	-6.2500	-0.0001
150, 50	0.0889	0.0533	40.0450	0.0356
150, 60	0.0381	0.0182	52.2310	0.0199
150, 70	0.0011	0.0009	18.1818	0.0002

Series II: Set 2 N_C=400				
	CBP		Deviation	
N, BW_C	AM	SM	Relative [%]	Absolute
20, 10	0.1157	0.1381	-19.2770	-0.0223
20, 15	0.0456	0.0510	-11.8653	-0.0054
50, 20	0.1099	0.1218	-10.8434	-0.0119
50, 30	0.0578	0.0595	-3.0232	-0.0017
75, 30	0.0942	0.1033	-9.6317	-0.0091
75, 40	0.0619	0.0556	10.2112	0.0063
75, 50	0.0302	0.0199	34.2131	0.0103
100, 50	0.0616	0.0435	29.4651	0.0182
100, 60	0.0227	0.0128	43.5143	0.0099
150, 50	0.0900	0.0754	16.1957	0.0146
150, 60	0.0717	0.0382	46.7156	0.0335
150, 70	0.0293	0.0134	54.1613	0.0159

V. THE ACTIVE TOOL TO EVALUATE OUR ANALYTICAL AVAILABILITY MODELS

In this section we will introduce our recently developed tool, the „*A*vailability *C*alculator of *T*V Channels for *I*PTV Services in *Ve*hicular Networks“, or for short: ACTIVE.

ACTIVE is a means to simply and quickly calculate CBP and other IPTV related availability measures, with high precision and high efficiency.

A. Requirements to the Tool

The requirements we asked of this tool included:

(1) A simple and intuitive graphical user interface.
(2) The tool should be flexibly applicable, calculating the results for a great number of combinations of parameter values with low expenditure.
(3) In addition to CBP, these results should include the most important IPTV related availability measures for
   a. *the total of all IPTV users.* These QoE measures are especially interesting for the IPTV service provider.
   b. *an individual user, taking the speed of the vehicle into consideration.* These QoE measures are especially interesting for IPTV users and those (e.g., the IPTV service provider) who need the users' view.

There were a few ways we ensured that these requirements were met:

1. The tool has two different modes: one for high precision – the *simulation mode* – and one for high efficiency – the *approximation mode*.
2. The simulation mode takes - just as the name implies - the input and runs a new Monte Carlo simulation with 10^6 iterations for highest precision. To increase efficiency, if the exact values / value combinations for (N_c. N) and (BW_c , N) are saved, then the results are calculated directly without the need of a simulation.
3. The approximation mode uses a priori saved intermediate results of the simulations to calculate the results, which is done instantly, thus with maximum efficiency. To increase precision, every time the program runs the 10^6 Monte Carlo simulations in simulation mode, the new intermediate results will be saved, thus making the – already very precise – approximation mode more and more precise.
4. To keep the GUI simple and intuitive, the tool is a single window application, with all the settings on the left side, allowing a simple and flexible, parallel input of the parameters as well as the choice of mode. All results are shown on the right side of the same window at a glance. In approximation mode, the results of the next lower saved N_c and next higher saved N_c (corresponding to the N_c value that was asked for) are shown at the same time. If the difference of corresponding values (e.g., the "lower" CBP and the "higher" CBP) exceeds 20% the values will be shown in maroon.
5. To maximize flexibility, the tool accepts the input of the 8 parameters already introduced in Table I, with p_i and $\bar{d}$ not being input parameters but being calculated in the simulation and based on the d_i respectively. Also the input for both N_c as well as d_i are optional. N_c can easily be given directly or be calculated by the program, the value is shown directly below the parameter and updated "live" during parameter input for maximum clearness. N_c is then a function of k, C_r, α, d_i (or v_i). Same goes for d_i, if no d_i values are given, the tool will calculate values using the v_i.
6. The tool will show the results for 24 IPTV related availability measures in total.

The input data, which is used by the ACTIVE tool to produce the TV service availability results, is shown in

detail by Fig. 6 (cf. "Parameter input" section, left-side bottom part of the "Settings").

The main intermediate data computed by ACTIVE comprises in particular P*, cf. eq. (4), and the probabilities p_i, $i \in \{1, 2, \ldots, N\}$, cf. Table I.

In particular, the following results are provided by the ACTIVE tool:

a) For the total of all IPTV users, there is CBP (Channel Blocking Probability), SBP (Switching-induced Blocking Probability) and HBP (Handover-induced Blocking Probability) and

b) For an individual user, there is bph_s (blockings per hour, switching-induced), bph_h (blockings per hour, handover-induced) and bph (blockings per hour in total).

c) These results are all given
   1. for the next lower saved N_c and next higher saved N_c (corresponding to the N_c value that was asked for) and also
   2. for the "late" model, when the traffic evens out, which is the model discussed in this paper and the "early" model right after the lanes were empty and the traffic for this road started (cf. model presented in [27]). So, ACTIVE actually exceeds the model discussed in this paper and implements also a second one.

B. Tool Architecture and Internal Process Flows

With Fig. 3 we would like to present a brief overview on the architecture and the internal mechanics of the tool.

The user specifies the mode of operation and the parameters, which will be received by the *evaluation module*.

This module will give N_c, N and BW_c to either the database module - if approximation mode was selected - or the simulation module - if simulation mode was selected - and will get the corresponding intermediate results ("vector P", "zipf": the vector with the popularities for the N channels and "Z*", a value used exclusively by the early model) as return. It will then calculate the results and display them in the GUI.

If the *simulation module* is asked for the intermediate results, it will start a new Monte Carlo simulation with 10^6 iterations, by creating a new java thread from SwingWorker. This new thread calls the external program "runSimulation", which is written in python but compiled as an .exe. The new intermediate results will be given to the evaluation module for further processing and to the database module for storage.

If the *database module* is asked for the intermediate results, it will just return the corresponding results it either read from the database at program start-up or got from the simulation module. At program termination it will write all values back to the database. This database consists of three

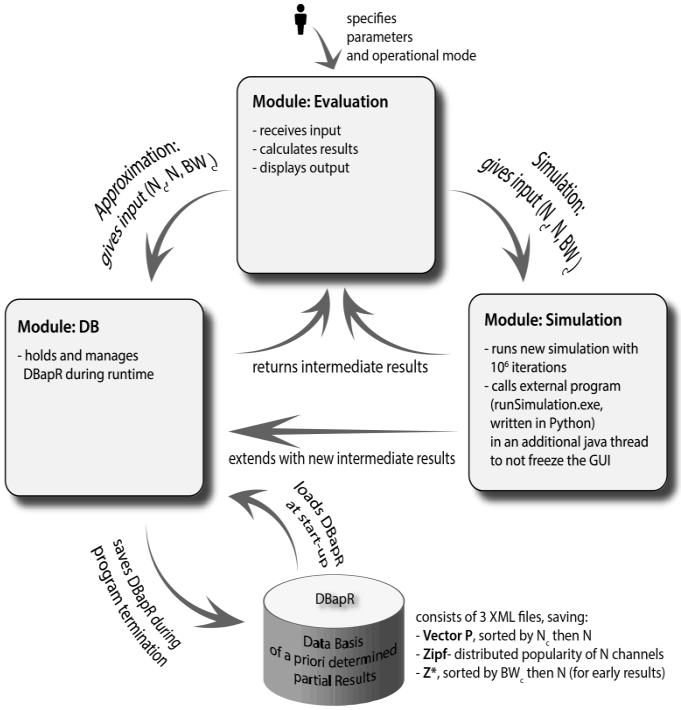


Figure 3. The internal mechanics of ACTIVE

XML files: "vectorP.xml", "zipf.xml" and "zStars.xml", which save their values, respectively.

Remark: The authors plan to provide their ACTIVE tool to interested researchers as an open source product as soon as it is completely finalized, i.e., before spring 2016 (cf. TKRN Web pages at the University of Hamburg).

C. Graphical User Interface

Fig. 4 shows the GUI in its entirety, as it looks on start-up. On the left side you see the settings panel. On top is the choice between approximation and simulation mode, as discussed. If you choose "Approximation" the results will be shown instantly, if you choose "Simulation" a new Monte Carlo simulation with 10^6 iterations will start and may take up to an hour, depending on the input of N_c and BW_c (on a 3 GHz processor core). The start button will change to a cancel button, the progress will be shown in the progress bar below the settings and the result panels, as shown in Fig. 5 and at the end of the simulation you will hear a "beep" sound for convenience. Below the choice of mode is the area for parameter input.

The variables N, BW_c, k, C_r, v_i are obligatory, d_i is always optional and below d_i you see a drop-down menu where you can choose between setting a value to N_c directly ("Set N_c manually") and setting a value to α, letting N_c be calculated ("Calculate N_c"). If you choose "Calculate N_c", the calculated value for N_c will be shown just below the value for α and will automatically update while you change the values for k, C_r, v_i, d_i or α.

On the right side is the result panel.

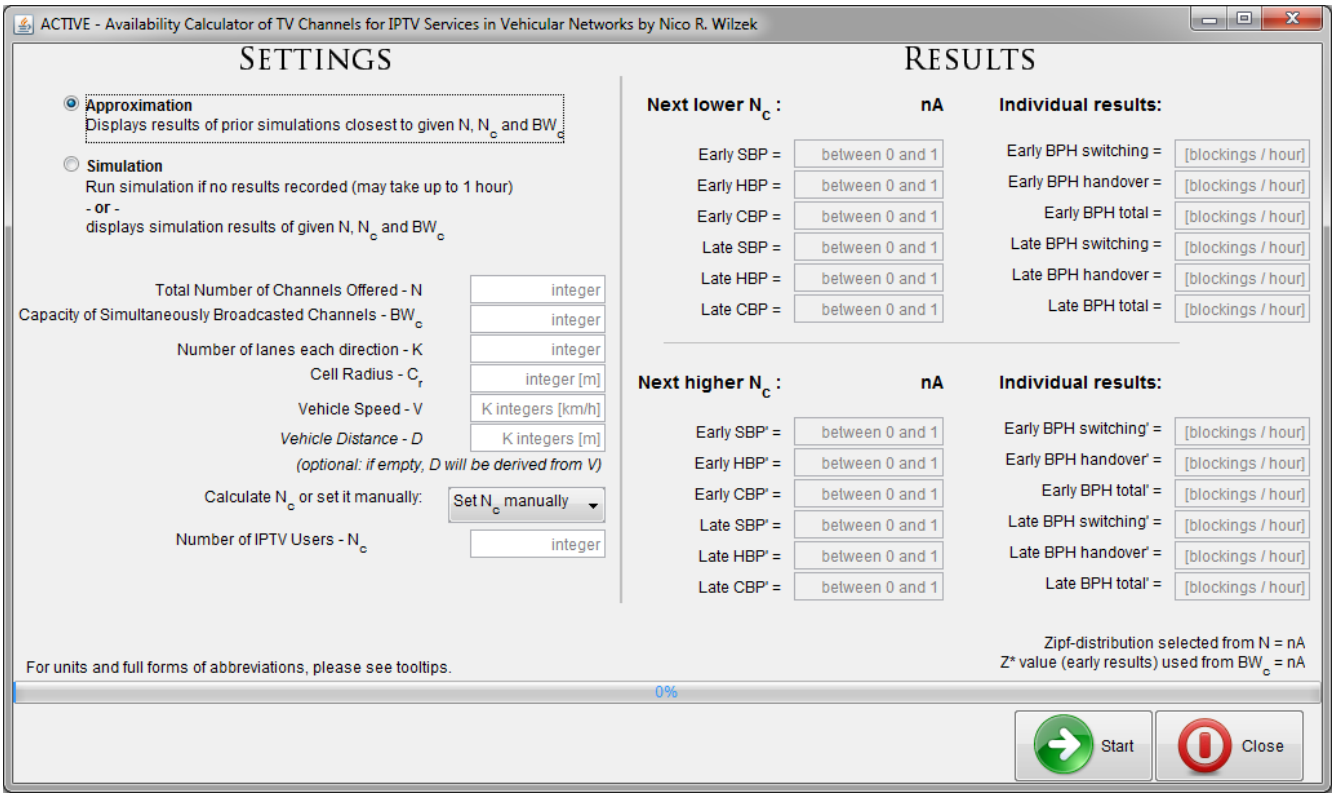


Figure 4. The GUI of ACTIVE at start-up

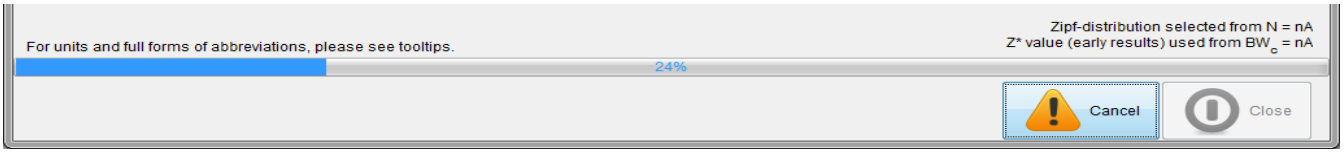


Figure 5. Change in GUI during simulation

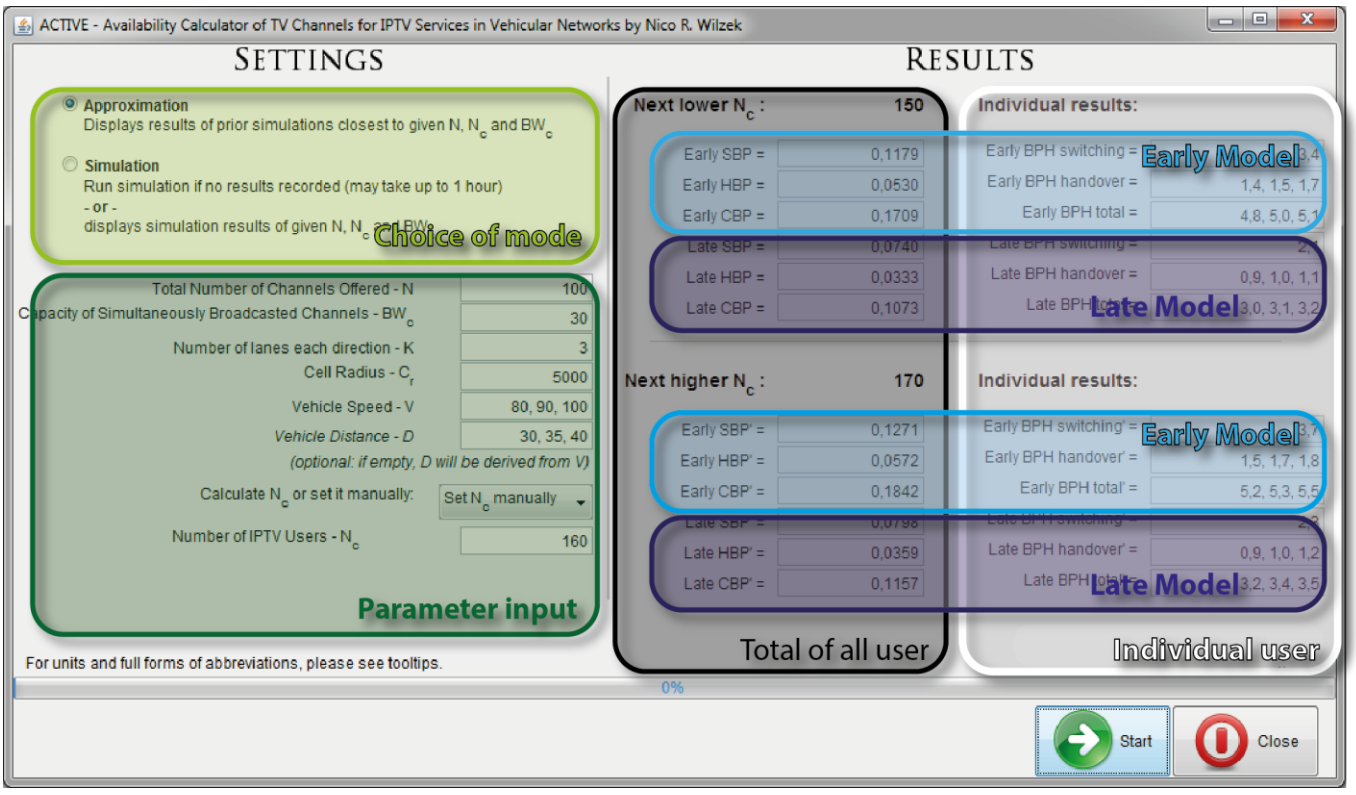


Figure 6. The (logical) partitioning of ACTIVE's GUI

If you split the result panel into left and right, you would find the results for the total of all users to the left and for an individual user to the right; if you split into top and bottom, you would find the results for the next lower saved N_c on top and the next higher saved N_c on the bottom.

Furthermore, you can split the halves for next higher/ lower saved N_c into top and bottom again, and you would find the results for the "early" model (additionally implemented to the model of this paper) on top and the results for the "late" model (discussed in this paper) on the bottom.

The complete (logical) partitioning is shown in Fig. 6.

D. A Sample Session

Let us use the tool for a specific scenario. We are interested in getting the results fast – if the given results are not close enough to the given scenario, we can still change our opinion. The scenario is a motorway with three lanes each direction. The cars drive faster on the inner lane, but not by a lot, let us say they drive 90 km/h on the outer lane, 100 km/h on the middle lane and 120 km/h on the inner lane. There is no easy way for us to measure the distances of the cars, traffic situation is dense (i.e., there is no unused space without cars), but they seem to have normal distances for their speed (that is the speculation of ACTIVE if no d_i is given).

We know that the IPTV service provider offers 100 channels in cells with 5 km radius, but can only stream 30 different ones at a time. Also, we assume that about every 20th vehicle will make use of the IPTV offer (but do not know the total number of users who make use of it).

To use the tool we just follow these 3 easy steps:

1. Keep/ change the mode setting to "Approximation".
2. Set input parameters by
   a. entering all values we have (except α, if α is hidden behind the drop-down menu), ignoring d_i and N_c, because we do not know the values for these parameters (the distances between adjacent vehicles and the total number of users),
   b. choosing "Calculate N_c" in the drop-down menu (since we do not know the total number of users, but the amount of IPTV users relative to all vehicles: every 20th),
   c. entering the value for α in % (every 20th = 5%), the number of users will be updated while we type, directly below our input.
3. Press the start button.

The results will be displayed immediately on the right side of the GUI.

Fig. 7 shows all steps and the corresponding results. Since all results are maroon in color it means the difference in results between the next lower saved N_c and the next higher saved N_c is more than 20%. We can now either choose to run the simulation by clicking on "Simulation" and then on "Start" to get more precise results, or say the N_c value that we asked for is close enough to one of the saved results. We would recommend the latter, since 78 is really close to 77.

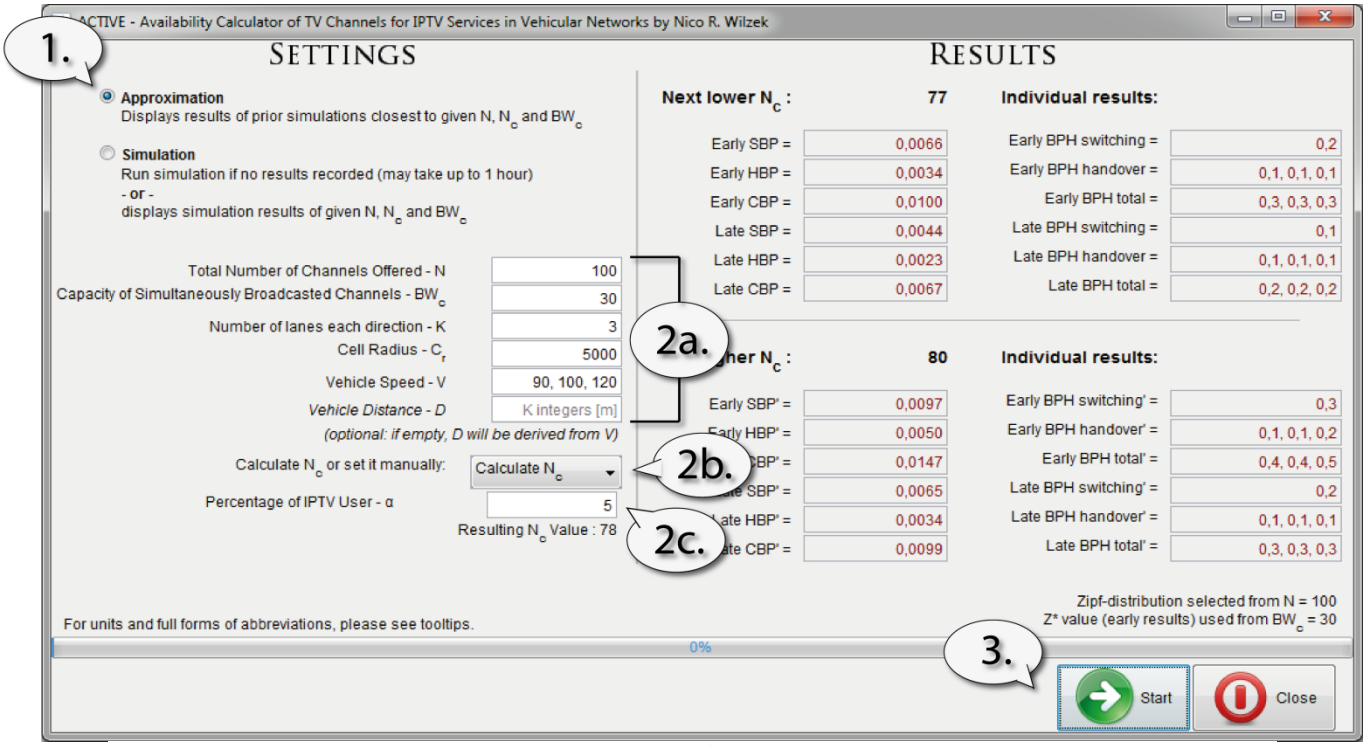


Figure 7. The 3 steps of the sample session

VI. CBP Bounds and Parameter Studies

In the following, our goal will be to use our analytical model, presented in Section III, to predict with only very little expenditure the availability of IPTV services in VANETs. In particular, our approach should cover a broad spectrum of traffic situations and of network technologies used to establish the access network for vehicle to RBU communication and, last not least, it should also cover numerous characteristics of the IPTV service offered.

Calculation of CBP based on our analytical model yields to the following formula:

$$CBP = P^* \bullet \sum_{i>BW_c}^{N} p_i , \qquad (5)$$

and this shows that CBP can be seen as a product of only two terms T_1 and T_2 with

$$T_1 \triangleq P^* \quad \text{and} \quad T_2 \triangleq \sum_{i>BW_c}^{N} p_i \qquad (6)$$

If we fix the value of the parameter θ in the Zipf distribution used to model IPTV user behavior, it becomes evident that

$T_1 = T_1(N, N_c, BW_c)$ *and* $T_2 = T_2(N, BW_c)$.

Therefore, it is possible to characterize T_1, as well as T_2 by means of elementary sets of curves.

Moreover, T_1 is a general upper bound for CBP because

$$T_1 = P^* > P^* \bullet \sum_{i>BW_c}^{N} p_i = CBP \qquad (7)$$

This is why the sets of curves related to term T_1 (resp. P*) are of particularly strong interest.

Similarly, T_2 is an upper bound of CBP, too, because

$$T_2 = \sum_{i>BW_c}^{N} p_i > P^* \bullet \sum_{i>BW_c}^{N} p_i = CBP \qquad (8)$$

Consequently, this also implies that

$$\min(T_1, T_2) > CBP$$

is a third and, in general, an even tighter upper bound of CBP.

A. Characterization of Upper Bound T_1, i.e., P*

Here, we want to investigate the influence of the available bandwidth BW_c on P* assuming that a certain number N of channels is offered and that the number N_c of IPTV users in the cell varies. In this study of P*, we assume $N \in \{20, 50, 100\}$ because $N = 20$ presents a small, $N = 50$ a medium and $N = 100$ a quite large number of channels offered.

Moreover, we suppose $N_c \in \{50, 100, 200, 300, 400\}$ because in realistic scenarios (e.g., for $\alpha = 0.05$) one nearly always will have no more than 400 IPTV users in a single cell (cf. below). Evidently, variation of BW_c only makes sense in the interval $[1, N]$.

For example, Fig. 8 directly shows that if $N = 100$ channels are offered, spending a bandwidth $BW_c = 70$ for IPTV will lead to a negligible value of P* and, therefore, also to a negligibly small CBP for all realistic cell populations considered by us ($N_c \leq 400$). And even a bandwidth $BW_c = 65$ reserved for IPTV will ensure that CBP < 10 % holds, if again $N_c \leq 400$ can be assumed.

B. Characterization of Upper Bound T_2

As T_2 is no longer dependent on N_c , investigations concerning this term become even more straight-forward than for T_1. In particular, the dependency of T_2 on the

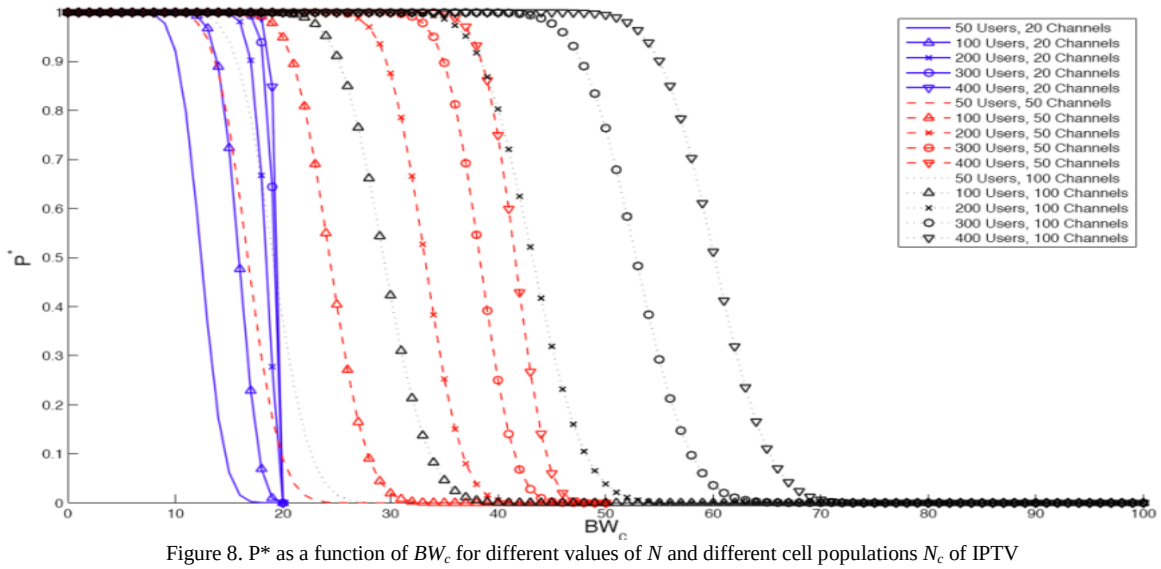


Figure 8. P* as a function of BW_c for different values of N and different cell populations N_c of IPTV

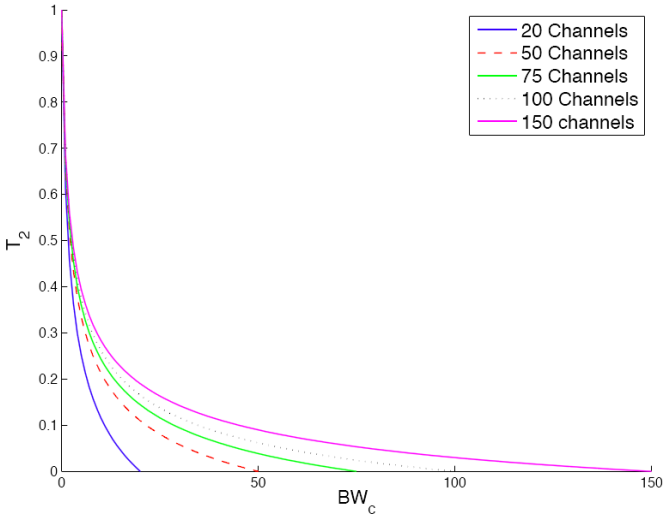


Figure 9. T_2 as a function of BW_c for different values of N

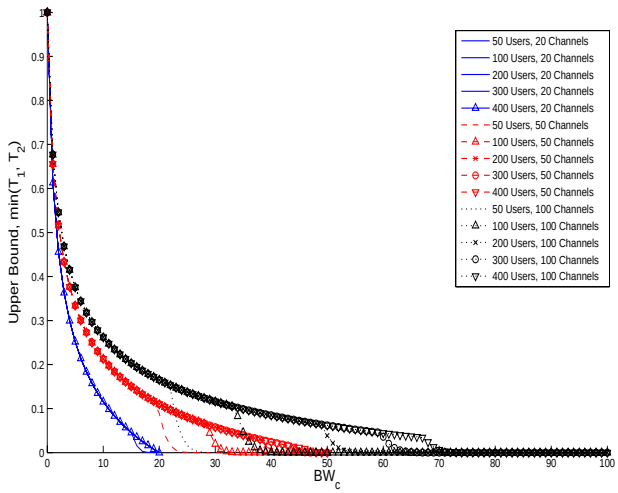


Figure 10. The upper CBP bound min (T_1, T_2) as a function of BW_c for different values of N and different cell populations N_c

bandwidth BW_c reserved for IPTV can be directly depicted for a given value of N.

Fig. 9 shows those dependencies for $N \in \{20, 50, 75, 100, 150\}$. This figure provides in-depth insight regarding the difficult decision of how much bandwidth should be spent for a given number N of offered channels. As examples, let us look at the case of $N = 20$ where it seems to be a good idea to choose $BW_c \geq 18$ (at least), for $N = 75$ a bandwidth of at least $BW_c = 50$ seems to be desirable and for $N = 150$ a chosen bandwidth of $BW_c \leq 80$ seems to be quite risky.

C. Characterization of Upper Bound min (T_1, T_2)

So far, we have discussed about how we can use the terms T_1 and T_2 as upper bounds of CBP separately. However, we are now interested in investigating the tighter upper bound min (T_1, T_2). So, we might ask the question: When does T_1 fit the best as upper bound of CBP and when does T_2? In order to avoid this uncertainty, we can define the function *min* (T_1, T_2) as a general and tighter upper bound as we stated before.

Fig. 10 depicts the behavior of the min (T_1, T_2) function. In this figure it becomes evident in which situations one term applies as upper bound and when the other one does, since we can see the inflection point at every curve.

From this figure, we can claim that when the resources are extremely scarce (w.r.t. the users in the system), T_2 fits the best as upper bound of CBP, however, when the resources are not that scarce, T_1 fits the best.

D. Expected Number of IPTV Users in a Cell

The number N_c of IPTV users to be expected in a cell will just depend on:

- average distance $\bar{d}$ between two adjacent vehicles (driving in the same lane), where the avg. is taken over all lanes
- the n^{o} of lanes per direction (k)
- the probability that in a vehicle IPTV is used (α)
- the cell radius (C_r).

In particular, N_c can be easily determined as follows:

$$N_c = \alpha \bullet 2k \bullet 2C_r / \bar{d} \qquad (9)$$

If we set $\alpha = 0.05$ and $k = 3$ to be constant and if we vary $\bar{d} \in \{5\text{m}, 10\text{m}, 20\text{ m}, 50\text{ m}, 100\text{ m}\}$ and assume cell radiuses of $C_r \in \{1\text{ km}, 3\text{ km}, 5\text{ km}, 10\text{ km}\}$, we get N_c values as depicted by Table IV. We see that with our assumptions, which we consider to be quite realistic, the value of N_c varies between 6 and 1200. We also can observe that rather different combinations of parameter values will lead to the same value of N_c, which facilitates the characterization of P* and thus also of CBP.

TABLE IV. N_c AS A FUNCTION OF d AND C_r

$\bar{d}$ \ C_r	1 km	3 km	5 km	10 km
5 m	120	360	600	1200
10 m	60	180	300	600
20 m	30	90	150	300
50 m	12	36	60	120
100 m	6	18	30	60

TABLE V. CHANNEL BLOCKING PROBABILITY (*CBP*) FOR DIFFERENT COMBINATIONS OF *N*, BW_c VALUES AND DIFFERENT N_c VALUES

CBP									
N_c / (N, BW_c)	50	75	100	125	150	200	300	400	500
(20, 15)	0.0028	0.0175	0.0330	0.0411	0.0442	0.0455	0.0456	0.0456	0.0456
(50, 20)	0.0098	0.0712	0.1043	0.1094	0.1099	0.1099	0.1099	0.1099	0.1099
(50, 30)	0	0.00002	0.0011	0.0086	0.0243	0.0506	0.0577	0.0578	0.0578
(75, 30)	0	0.0011	0.0193	0.0585	0.0843	0.0940	0.0942	0.0942	0.0942
(75, 50)	0	0	0	0	0	0.00003	0.0074	0.0302	0.0380
(100, 50)	0	0	0	0	0	0.0024	0.0473	0.0616	0.0619
(100, 60)	0	0	0	0	0	0	0.0016	0.0227	0.0414
(150, 60)	0	0	0	0	0	0.0002	0.0381	0.0717	0.0730
(150, 70)	0	0	0	0	0	0	0.0011	0.0293	0.0563
(150, 80)	0	0	0	0	0	0	0	0.0009	0.0162

E. Straight-forward Calculation of CBP for Numerous Scenarios of IPTV in VANETs

Combining the results achieved in this section up to now, we are able to propose a generalized proceeding that allows us to predict CBP for nearly any scenario of interest with nearly negligible expenditure (if we compare this with a CBP prediction based on simulation models for assessing IPTV availability in VANETs).

In particular, Table IV showed us what N_c values to assume to be realistic and the results of Figs. 8 and 9 can be directly combined (i.e., T_1 and T_2 can be multiplied) to determine CBP. Table V contains CBP predictions based on our analytical model for numerous scenarios of IPTV in VANETs. The results of Table V cover a broad spectrum of traffic situations (low, medium and high traffic load up to traffic jam), of access network technologies used having an impact on C_r and BW_c and of characteristics of the IPTV service (e.g., number N of channels offered).

To summarize, the results obtained in this section can allow one to significantly improve the understanding of the main factors and their mutual dependencies, which influence IPTV availability in VANETs.

VII. CASE STUDIES

In the previous section, we have shown how it is possible to determine CBP just as a function of *N*, N_c and BW_c, where, of course, N_c itself is a function of $\overline{d}$, C_r, *k* and α. We now want to indicate how the handover- and the switching-induced blocking probabilities HBP and SBP can be determined based on CBP.

A. Handover-induced Blockings: Calculation of #ho$_{ph}$

Let #ho$_{ph}$ denote the total number of handovers per hour of all vehicles using IPTV and leaving a given cell. We assume a mean speed of those vehicles of $\overline{v}$ and a mean distance between adjacent vehicles of $\overline{d}$, a cell radius C_r, *k* lanes per direction, as well as an IPTV watching probability of α. With these assumptions we can directly calculate N_c (cf. Section VI.D.).

#ho$_{ph}$ can be determined in a straight-forward manner as follows:

$$\#ho_{ph} = \alpha \bullet 2k \frac{\overline{v}[\frac{km}{h}]}{\overline{d} \bullet 10^{-3}\,[km]} = \alpha \bullet 2k \frac{\overline{v}}{\overline{d} \bullet 10^{-3}} [\frac{1}{h}] \quad (10)$$

For single IPTV users we thus obtain:

$$bph_h(v) = \#ho_{ph}(v) \bullet CBP, \quad (11)$$

if $bph_h(v)$ denotes the *number of handover-induced blockings per hour* experienced by a single user in a vehicle driving with speed v.

B. Switching-induced Blockings: Calculation of #sw$_{ph}$

Let #sw$_{ph}$ denote the total number of switching events per hour of all N_c vehicles using IPTV in a given cell. Let us assume a mean time Δt [min] between two successive channel switching events, where $\Delta t = 3$ [min].

Then, #sw$_{ph}$ can be determined as follows:

$$\#sw_{ph} = \frac{60}{\Delta t} \bullet N_c [\frac{1}{h}] \quad (12)$$

For single IPTV users we thus obtain:

$$bph_s = \#sw_{ph} \bullet CBP, \quad (13)$$

if bph_s denotes the *number of switching-induced blockings per hour* experienced by a single user.

And, evidently, the *total number of blockings per hour* experienced by a single user in a vehicle driving with speed v is expected to be

$$bph(v) = bph_h(v) + bph_s . \quad (14)$$

C. Calculation of HBP and SBP

HBP can be determined based on #ho$_{ph}$, #sw$_{ph}$ and CBP as follows:

$$HBP = (CBP \bullet \#ho_{ph})/(\#ho_{ph} + \#sw_{ph}) \quad (15)$$

Correspondingly:

$$SBP = (CBP \bullet \#sw_{ph})/(\#ho_{ph} + \#sw_{ph}) \quad (16)$$

which again confirms that: $HBP + SBP = CBP$. (17)

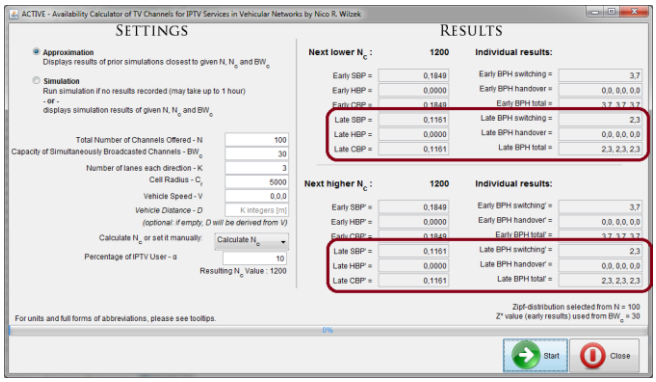


Figure 11. Case study results obtained based on the ACTIVE tool

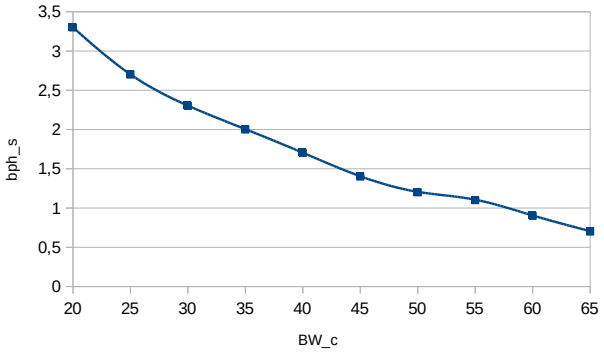


Figure 12. Corresponding bph_s to given BW_c

D. Case Studies

We will now use the ACTIVE tool to study different scenarios of interest in three case studies, switching our focus to the results for individual IPTV users for the first two of our studies. We will keep our focus on the model for the „late phase“, which are highlighted in Fig. 11.

The first two case studies will both be about a traffic jam. In the *first study* we will explore how much the IPTV service provider would need to increase the bandwidth for the TV channels, if the provider wishes to keep a certain level of QoE during a traffic jam (for example for places well known for their frequent traffic jams). The *second case study* will explore when the QoE will raise to an acceptable level while the traffic jam slowly dissolves and the speed of the users increases. We suggest a bph ≤ 2 as an „acceptable level“ (i.e., on average 1 blocking every 30 min). The *third case study* shows how to use the tool to gain information about highly fluctuating traffic scenarios.

We assume a traffic scenario with three lanes per direction on a motorway, which makes it likely enough to create traffic jams frequently: $N = 100$, $k = 3$, $C_r = 5$ km, $v = 0$ km/h (since traffic jam), $\alpha = 0.1$. These parameters imply that $N_c = 1200$ (which is an huge increase from the expected $N_c \leq 320$ we would have with $v = 50$ km/h). With $v = 0$ km/h the program assumes $d = 5$m, if no value for d is given. Let us assume this value is realistic in a traffic jam.

Case Study 1: During Traffic Jam

Since $v = 0$ km/h, bph_h will obviously stay 0, which results in bph_s = bph. So, for the rest of the case study we will only talk about the relationship between BW_c and bph_s. This relationship is shown in Fig. 12 as a graphical representation. We observe that for $BW_c \geq 30$ the decrease of bph_s is approximately linear.

Figures like that are very useful for IPTV service providers and users alike, because they show in a simple manner, what bandwidth you need for a certain QoE or what QoE to expect given a certain bandwidth, e.g., that you need about a bandwidth of 35 TV channels to drop under a bph_s of 2 blockings per hour. That is surprisingly low for such a huge increase in number of users.

Case Study 2: During Traffic Jam Termination

Assuming v slowly increases at the end of a traffic jam phase, we expect that bph_h will increase and bph_s will decrease in such a manner that bph will decrease.

Using the information learned from the first case study, we investigate those influences in detail for the three scenarios $BW_c \in \{40, 50, 60\}$. The results are shown in Tables VI - VIII and Fig. 13 summarizes the dynamic evolution of bph_s, bph_h and bph for all three scenarios in Case Study 2. Since v changes exactly the same in every lane, the values for all three lanes will be identical and, therefore, will be shown only once.

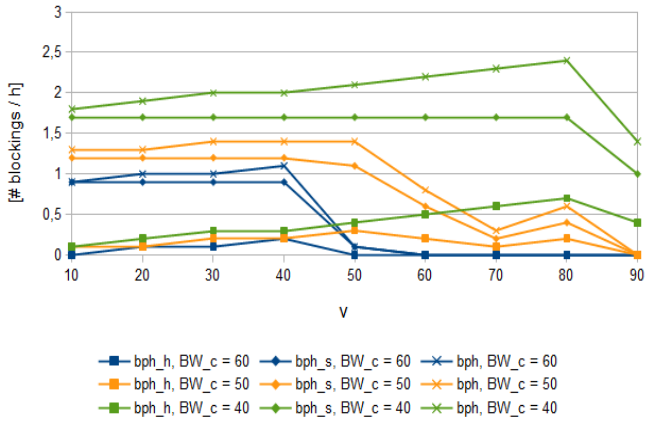


Figure 13. Corresponding bph to given v

TABLE VI. Results for scenario 1 of the traffic jam termination

Scenario 1: $BW_c = 40$			
v	bph_h	bph_s	bph
10	0.1	1.7	1.8
20	0.2	1.7	1.9
30	0.3	1.7	2.0
40	0.3	1.7	2.0
50	0.4	1.7	2.1
60	0.5	1.7	2.2
70	0.6	1.7	2.3
80	0.7	1.7	2.4
90	0.4	1.0	1.4

TABLE VII. Results for scenario 2 of the traffic jam termination

Scenario 2: $BW_c = 50$			
v	bph_h	bph_s	bph
10	0.1	1.2	1.3
20	0.1	1.2	1.3
30	0.2	1.2	1.4
40	0.2	1.2	1.4
50	0.3	1.1	1.4
60	0.2	0.6	0.8
70	0.1	0.2	0.3
80	0.2	0.4	0.6
90	0	0	0

TABLE VIII. Results for scenario 3 of the traffic jam termination

Scenario 3: $BW_c = 60$			
v	bph_h	bph_s	bph
10	0.0	0.9	0.9
20	0.1	0.9	1.0
30	0.1	0.9	1.0
40	0.2	0.9	1.1
50	0.0	0.1	0.1
60	0	0	0
70	0	0	0
80	0	0	0
90	0	0	0

Surprisingly enough, the results of Case Study 2 show that different traffic situations lead to very similar numbers of blockings, but there is always one speed where both bph_s and bph_h significantly drop (BW_c = 40: 80 km/h, BW_c = 50: 50 km/h, BW_c = 60: 40 km/h). These figures are very valuable to IPTV service providers, because you can easily answer a number of questions, including

- How long after a traffic jam will the QoE actually decrease, since the handover-induced blockings increase and the switching-induced blockings have not decreased yet?
- To which degree will the traffic jam need to dissolve for the QoE to become acceptable?
- What bandwidth will be needed to keep both, the QoE decrease after the traffic jam and the time for the QoE needed to increase to a certain level, acceptably small?

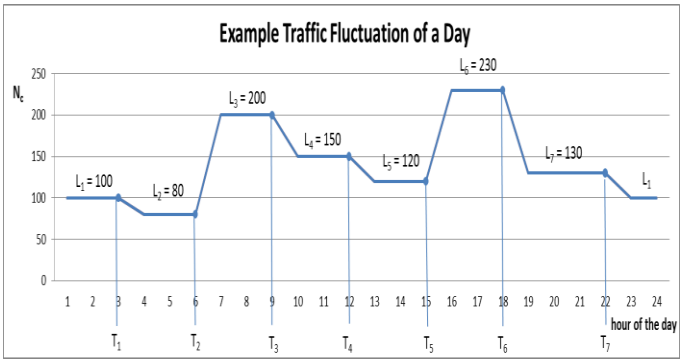


Figure 14. Example traffic fluctuation for case study 3

Case Study 3: Highly Fluctuating Traffic Scenario

This case study pays tribute to the fact, that the traffic situation can fluctuate strongly over time and will show how to use the ACTIVE tool accordingly.

Let us assume a traffic fluctuation over a day as shown in Fig. 14. As you can easily see, the traffic fluctuates over the course of the day, but is nearly constant for a few hours during several time intervals, so in our example scenario we get seven different levels of load.

Let $N_{c,i}$ be the mean number of IPTV users in cell c while the load is at level L_i.

We now use our analytical model to predict CBP for the dynamic fluctuating traffic scenario at the marked times, which represent the end of the level of load L_i.

We keep assuming N = 100 (and BW_c = 30) for this case study, too.

With ACTIVE we can easily and quickly get the CBP of all $N_{c,i}$ by entering successively the values for each L_i. The results of CBP as a function of N_c are shown in Table IX as well as in Fig. 15. And these results have been directly used to illustrate the daily CBP variation in Fig. 16.

Results as those shown by Fig. 16 are valuable for IPTV service providers, because they give answers quickly to a variety of questions, including but not limited to:

- What is the worst CBP/ QoE to be expected during the observed time interval (and when will it occur)?

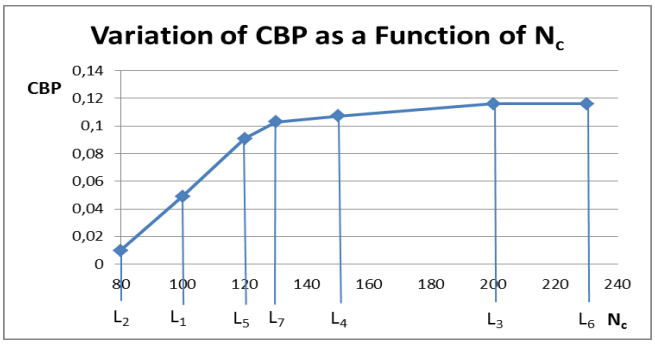


Figure 15. CBP values of the L_i from case study 3

TABLE IX. Resulting CBP for the seven $N_{c,i}$

N_c	CBP
80	0.0099
100	0.0490
120	0.0908
130	0.1030
150	0.1073
200	0.1161
230	0.1161

- Which changes in N_c have what kind of impact on CBP?
- How does the watched time interval partition into time slots with different QoE? In our example we see (cf. Fig. 16) that the day is basically partitioned into a night part with better QoE and a day part with worse QoE. The two rush hours do not seem to have a big impact on QoE.

VIII. Summary and Outlook

In this paper, we have tackled the challenging and difficult problem to predict the availability of IPTV services in vehicular networks. In particular, we have elaborated a class of analytical models that have been successfully validated by means of existing simulation models. The analytical models indicated that it is possible to obtain a quite realistic prediction of IPTV service availability by means of using just a few elementary parameters comprising, e.g., number of TV channels offered, number of active IPTV users in the corresponding vehicular network cell and total bandwidth available for IPTV in the cell. We also have presented a tool (ACTIVE) that allows us to determine IPTV service availability in a straight-forward and very efficient manner. The tool is based on our class of analytical models and it makes heavy use of a repository of partial results having already been calculated in advance to lay these results in stocks. Later, upon demand, the a priori determined partial results can be combined in a flexible manner to obtain an overall availability result for a new set of system parameters (regarding traffic situation, user behavior and IPTV service characteristics). Numerous case studies show how the ACTIVE tool can be used and they prove our claim that availability prediction by means of our analytical model and the new calculation method is highly efficient. What could take hours to be determined by executing simulation experiments can now be calculated in only a few seconds – and this can be done without loss of a significant amount of prediction accuracy as has been shown by our validation experiments. It is worth to be mentioned that our analytical modeling approach also directly covers the situation, that the watching probabilities p_i of the TV channels are not determined by means of a Zipf distribution but have been measured in an existing IPTV system as relative frequencies of channel accesses. This makes our model even more realistic.

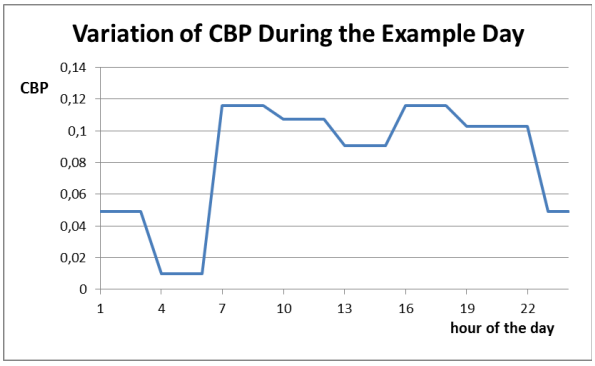


Figure 16. Variation of CBP during the example day.

In our future research related to IPTV in vehicular networks, we plan to investigate scenarios assuming IPTV in vehicles on rural roads (instead of motorways). An additional interesting research topic will be how IPTV user behavior will change if the users are passengers in cars instead of watching TV as "coach potatoes" being at home.

References

[1] B. E. Wolfinger, E. E. Báez, and N. R. Wilzek, "A Generalized Approach to Predict the Availability of IPTV Services in Vehicular Networks Using an Analytical Model," 14th Internat. Conf. on Networks, ICN 2015, Barcelona, Spain, 2015, pp. 163-170.

[2] A. Abdollahpouri, "QoS Aware Live IPTV Streaming Over Wireless Multi-hop Networks," Shaker Verlag 2012.

[3] A. Abdollahpouri, B. E. Wolfinger, J. Lai, and C. Vinti, "Modeling the Behavior of IPTV Users with Application to Call Blocking Probability Analysis," Praxis der Informationsverarbeitung und Kommunikation 2012, 35 (2), pp. 75-81.

[4] F. Ahmed-Zaid, H. Krishnan, M. Maile, L. Caminiti, S. Bai, and S. VanSickle, "Vehicle Safety Communications – Applications: System Design & Objective Testing Results," SAE Int. J. of Passeng. Cars – Mech. Syst., 2011, 4 (1), pp. 417-434.

[5] A. Baiocchi and F. Cuomo, "Infotainment services based on push-mode dissemination in an integrated VANET and 3G architecture," Journal of Communications and Networks, April 2013, 15 (2), pp. 179-190.

[6] E. Belyaev, A. Vinel, A. Surak, M. Gabbouj, M. Jonsson, and K. Egiazarian, "Robust Vehicle-to-Infrastructure Video Transmission for Road Surveillance Applications," IEEE Trans. Veh. Technol., 2014, 64 (7), pp. 2991-3003.

[7] T. Chen and R. R. Rao, "Audio-visual Integration in Multimodal Communication," IEEE Proc. 1998, 86, pp. 142-149.

[8] W. L. Dunn and J. K. Shultis, "Exploring Monte Carlo Methods," Elsevier Science 2011.

[9] FOCUS Online: "80 Prozent der Neuwagen mit Internet," www.focus.de/auto/news/revolution-im-auto-bis-2016-haben-80-prozent-der-neuwagen-internet_id_4216763.html (last access: Nov. 9, 2015).

[10] A. Gladisch, R. Daher, M. Krohn, and D. Tavangarian, OPAL-VCN: "Open-Air-Lab for Vehicular Communication Networks," WiMob'10, 2010, pp. 555-561.

[11] T. Guo, C. H. Foh, J. Cai, D. Niyato, and E. W. M. Wong, "Perfomance Evaluation of IPTV Over Wireless Home Networks," IEEE Trans. on Multimedia, 2011, 13 (5), pp. 1116-1126.

[12] D. Jiang and L. Delgrossi, "IEEE 802.11p: Towards an International Standard for Wirelesss Access in Vehicular Environments," Vehicular Technology Conf., 2011, pp. 2036-2040.

[13] M. Krohn, R. Daher, M. Arndt, and D. Tavangarian, "Aspects of roadside backbone networks," 1st Internat. Conf. Wireless Communication, Vehicular Technology, Information Theory and Aerospace & Electronic Systems Technology, Wireless VITAE 2009, pp.788-792.

[14] J. Lai, "Evaluation and Improvement of TV Channel Availability for IPTV Services," Shaker Verlag 2012.

[15] S. Malkos, E. Ucar, and R. Akdeniz, "Analysis of QoE key factors in IPTV systems: Channel switching," 5th Internat. Conf. on Application of Information and Communication Technologies (AICT2011), October 2011, Baku, Azerbaijan, pp. 1-5.

[16] S. Möller and A. Raake, (eds.), "Quality of Experience: Advanced Concepts, Applications and Methods," Springer 2014, pp. 11-35.

[17] S. Momeni and B. E. Wolfinger, "Availability Evaluation of IPTV in VANETs with different types of access networks," EURASIP Journal on Wireless Communications and Networking, Springer Open Journal, 2014: 117 (15 July 2014).

[18] M. E. J. Newman, "Power Laws, Pareto Distributions and Zipf's Law," Contemporary Physics 2005, 46 (5), pp.1-3.

[19] M. Oche, R. M. Noor, and J. I. Aghinya, "Network Centric QoS Performance Evaluation of IPTV Transmission Quality over VANETs," Computer Communications, 2015, 61, pp. 34-47.

[20] S. Olariu and M. C. Weigle, "Vehicular Networks: From Theory to Practice," Chapman & Hall / CRC Press 2009.

[21] S. Park, J. Jeong, and C. S. Hong, "QoS-guaranteed Mobile IPTV Service in Heterogeneous Access Networks," Computer Networks, 2014, 69, pp. 66-81.

[22] RFC 1112: "Host Extensions for IP Multicasting," August 1989.

[23] F. J. Ros, J. A. Martinez, and P. M. Ruiz, "A Survey on Modeling and Simulation of Vehicular Networks: Communications, Mobility, and Tools," Computer Communications, 2014, 43, pp. 1-15.

[24] R. Y. Rubinstein and D. P. Kroese, "Simulation and the Monte Carlo Method," 2nd ed., Wiley 2007.

[25] J. P. Urrea Duque and N. Gaviria Gomez, "Quality assessment for video streaming P2P application over wireless mesh network," XVII Symposium of Image, Signal Processing, and Artificial Vision (STSIVA2012), September 2012, Antioquia, Colombia, pp. 99-103.

[26] B. Veselinovska, M. Gusev, and T. Janevski, "State of the Art in IPTV," 37th Internat. Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2014), Opatija, Croatia, 2014, pp. 479-484.

[27] B. E. Wolfinger, A. Hübner, and S. Momeni, "A Validated Analytical Model for Availability Prediction of IPTV Services in VANETs," Electronics 2014, 3, pp. 689-711; doi: 10.3390/electronics3040689.

[28] L. Zhou, Y. Zhang, K. Song, W. Jing, and A. V. Vasilakos, "Distributed Media Services in P2P-Based Vehicular Networks," IEEE Trans. Veh. Technol., 2011, 60, pp. 692-703.

A Hardware and Software System for Information Interchange in Multinational Disaster Relief Operations

Peter Dorfinger, Ferdinand von Tüllenburg, Georg Panholzer, Thomas Pfeiffenberger
Advanced Networking Center
Salzburg Research Forschungsgesellschaft mbH
Salzburg, Austria
e-mail: {peter.dorfinger, ferdinand.tuellenburg, georg.panholzer, thomas.pfeiffenberger}@salzburgresearch.at

***Abstract*—Information and communication technology becomes more and more important for large scale disaster work as it allows for sharing information between stakeholders in short times. However, the interchange of disaster related information is often affected by missing communication coverage and outages of pre-existing infrastructure. Further problems arise, if multiple organizations from different countries shall share all their disaster related information. Until now, there is no ICT-System available for disaster relief work, which provides an integrated solution of communication technology and information sharing and visualization applications helping for creating a common understanding across national and organizational borders of what is happening in the disaster operation. In this paper a flexible infrastructure for information interchange of disaster related information is presented. The first main building block of this infrastructure is the disaster information system consisting of various IT systems and software applications used to produce, share and manage disaster data. The second building block is a mobile, and easy to handle 802.11 driven communication hardware and software system. This system is capable to bring communication coverage to almost every arbitrary location within a disaster area by setting up a meshed wireless network. The wireless network is used to connect end devices used by field personnel to the disaster information system. The communication system can be operated independently of any pre-existing infrastructure such as Internet access or power supply. Several training events showed the usability of the proposed solution and the advantages of a comprehensive ICT system in international disaster work.**

***Keywords*—Emergency network; 802.11 communication; interoperability; information interchange; disaster relief.**

I. INTRODUCTION

One central problem when coordinating relief units in large scale disaster relief operations is to provide the right information to the right people. Especially in the context of international relief operations new coordination problems arise from differing structures of the international associations. Different languages causing communication difficulties and making information distrustful if not properly formulated information is leaving room for interpretation. An example for the impact of this problem is the earthquake disaster in L'Aquila, Italy. Here, the local authorities declined offered international help as the effort to integrate foreign relief organizations in their own relief work was considered as too high, requiring too much manpower urgently needed in other places in the disaster area.

Information technology can prove extremely useful in information gathering, storing and sharing. However, it must be kept in mind, that information should be shared according to the principle "as much as necessary, as little as possible" in a way easily and clearly to understand in order to protect helpers from information overload. With the help of broadband communication technology the narrow information channels of radio messages or telephone calls could be prized by enriching transported information with videos, photos or sensor data. Experiences from the usage of such a system by end users in multinational disaster operations are presented in [1]. Within this paper we present in detail the technical framework behind such a broadband communication solution.

Following the principle of sharing precise and comprehensive information to the right person has multiple advantages to enhance efficiency of disaster relief work. First, it prevents helpers from becoming overwhelmed by the volume of information transmitted by e-mails, radio messages and telephone calls – a fact that is often claimed by relief workers nowadays [2]. Second, correct decisions can be made more quickly in order to provide urgent needed aid to affected people. For example, by bringing all meaningful information to a team leader working in the field, the team leader could take the right decision autonomously [2].

This is where the IDIRA (interoperability of data and procedures in large-scale multinational disaster response actions) project [3] comes into the picture. IDIRA's target is to enhance interoperability of organizations and their systems in order to streamline the cooperation in relief work.

IDIRA addresses this interoperability topic twofold. First, at an organizational level, IDIRA examines possibilities to reach administrative coordination of multi-national disaster relief organizations, each with their own workflows and procedures. Second, on the technical side, IDIRA provides a complete solution consisting of information systems, communication protocols, software applications, and standard data formats. The developed systems accompany the topics stated above: Bringing all meaningful information to exactly the people needing it to make correct decisions and present information in an unambiguous manner.

This solution allows exchanging disaster related information between administrative operators, executive personnel, and other disaster management systems connected to IDIRA. With IDIRA, information on incidents, resources, observations, and sensor data should be collected and shared to various other information systems like mobile devices and command and control systems (C&C). To reach the required level of interoperability and automatic information exchange, IDIRA has a strong focus on mobility of the developed systems. It is necessary to bring the systems directly to the scene of the disaster, because of severe infrastructural damage after the disaster.

A prerequisite for an information exchange is a working broadband communication infrastructure within the disaster area. One of the major problems we address with our proposed communication infrastructure is that after a large scale disaster, the existing public broadband network is often partially destroyed, overloaded, or hit by power outages. Consequently, first responders cannot rely on any pre-existing infrastructure which may fail as a consequence of the disaster.

As the communication network is essential for a more efficient collaboration between first responders, there is a critical need for the fast setup of alternative communication means. However, first responders are not experts in setting up communication equipment. Thus, easy setup and maintenance is heavily required for such systems.

This paper will present an easy to install adhoc communication infrastructure for first responders to be used for an enhanced collaboration in large scale disaster operations.

This paper is structured as follows: Section II presents technologies and standards used within IDIRA and gives a brief overview on alternative communication solutions for disaster operations. Section III describes the IDIRA information system, as an example for an enhanced information system useable in multinational disaster operations. Section IV presents the details of our proposed mobile communication solution. Section V concludes the paper and outlines further steps and ideas to improve our solution.

II. Related Work

Within the IDIRA applications, disaster information is represented using the Emergency Data Exchange Language (EDXL) [4]. This is an open XML-based messaging format and suite of standards aimed at the use of information exchange in emergency management systems. The EDXL-CAP (Common Altering Protocol) [5] data format is applied to data about occurred incidents registered, for example automatically by a sensor system or manually by a human user. Information about resources such as availability of relief units, emergency vehicles or electrical generators, are exchanged by EDXL-RM (Resource Messaging) [5] standard. The EDXL-SitRep (Situation Report) [7] messaging standard is used within the IDIRA context for exchanging information on observations and situation reports sent by commanders in the field via their mobile devices. These standards are not bandwidth optimized and potentially use large headers and large data payloads. Large scale disasters may result in thousands of such messages. Consequently, to transport this information a broadband communication infrastructure is needed.

Nowadays, a broad variety of communication technologies are used by first responder relief organizations and the most widespread technology used for many years was standard voice radio. However, with the advent of mobile communication standards such as 2G, 3G, and 4G these technologies are increasingly displaced. There also exists technology that is more tailor-made to disaster relief organizations regarding mobility and independence of pre-existing infrastructure such as working backbone networks and power supply. This makes sense as, for instance, mobile phone networks are often heavily overloaded or partly out of order after a disaster occurred. One approach to overcome these problems was TETRA [8]. TETRA allows both, range limited direct device to device communication without usage of a fixed infrastructure and range unlimited indirect communication via a fixed infrastructure. To enhance reliability of TETRA, the fixed infrastructure system was designed in a highly redundant way and is not made accessible to the public. The downside of TETRA is the limited bandwidth (28.8 kbit/s) available for data communication.

Other communication solutions are based on satellite communication systems like BGAN [9], VSAT [10], or Emergency.lu [11]. Satellite communication is used for both data and voice communication and is operable also in remote areas. The disadvantage of this technology is the usually high operational costs and, in case of BGAN, the very limited bandwidth. With the exception of BGAN satellite communication seems as a central Internet uplink technology in the disaster area and less as a communication technology used by field personnel to communicate with each other.

Especially the limited or expensive data communication capabilities are making these technologies only usable to a restricted degree. IDIRA heavily depends on data exchange with higher bandwidth demands and limited latency. One example is user interaction via IDIRAs web interface - the so called Common Operational Picture (COP). Here, data are exchanged between web clients of tactical personnel at the Command & Control Center and field commanders. To bootstrap a device using COP an initial data download of about 10 Mbyte of data is necessary and for a seamless operation a bandwidth of about 2 Mbit/s is recommended.

Another problem arises from special international operating permissions and licenses needed for some communication technologies such as WiMAX [12] equipment. Moreover, for a communication system specifically designed for public protection and disaster relief (PPDR) called Highly Mobile Network Node (HiMoNN) [13] operating licenses are only available for a few countries worldwide. HiMoNN is designed in compliance with ECC Recommendation (08)04 [14], and operates with transmission power of 8W in the 5GHz frequency band. It is able to transmit data over a distance of several kilometers with a bandwidth of 28Mbit/s.

To overcome the problems of international operation permissions also other approaches were considered. For example, in the work of Raffelsberger and Hellwagner it is proposed to build up a mobile ad-hoc network (MANET) using the end user devices of first responders as communication hops [15]. These devices are using their 802.11 wireless network interfaces to build up connections to other devices. This approach overcomes problems with missing operation permissions as 802.11 equipment can be operated all over the world without special license. During operation, first responders sending data via this MANET to a central host located in the Command and Control Center also employing special routing protocols. However, a quite dense concentration of devices is necessary making it difficult to use this technology to bridge distances of several kilometers to reach the central host. Another approach [16] overcomes the problem of low device density by placing mobile devices as stationary relays. It achieves this by using the mobile devices of disaster survivors to set up a disaster recovery network using 802.11 Wi-Fi. In contrast to other discussed approaches it focuses on connectivity for the survivors instead of the rescue teams, but can be used for both. Their approach and ours can be combined for greater range and flexibility.

The communication system proposed in this paper also uses 802.11 [17], as this technology is widespread, cheap and can be used all over the world without special licenses. One of the main achievements of this work was to provide a solution for the problem of limited range between two 802.11 end points.

As routing protocol we use the Optimized Link State Routing Protocol (OLSR) [18], which is optimized for constrained wireless LANs. OLSR is based on multipoint relays in order to reduce the routing overhead on the network.

III. IDIRA DISASTER INFORMATION SYSTEM

In IDIRA, an information system is developed that improves information sharing and information presentation in disaster relief work. On the technical side, this information system consists of various software applications and specific hardware solutions allowing, for example, optimal resource planning and decision finding across national and organizational borders. This section gives an overview of applications and hardware infrastructure developed within IDIRA.

Various software applications were developed to support information collection, data analysis, decision finding, and information sharing and presentation.

One main building block of IDIRA's information system is the disaster information data store, a central system used for storing all information related to the disaster. The data store comprises of information about incidents and observations or sensor data. Furthermore, information about available resources in the field - such as positions, tasks and utilization is stored as well as geographic information about infrastructural facilities like hospitals. Also, other important information such as weather data is contained within the data store.

To store and transmit this information, standardized protocols and data formats are used, for example, the Emergency Data Exchange Language (EDXL) standardized by OASIS [4].

An example for automatic information collection and using EDXL is the sensor data integration. IDIRA supports automatically inducing data generated by different sensors using the OGC Sensor Observation Service (SOS) interface. IDIRA uses a generic Senor Fusion Engine (SFE) for sensor data aggregation. In case of a pre-defined behavior is recognized, the SFE generates an alarm messages in EDXL-CAP format [19]. The EDXL-CAP is one specific message format defined by OASIS specifically for information interchange in disaster operations [5]. The standard CAP message contains information such as type of emergency, source of information, level of severity, location, and extent of disaster. A link to detailed information, such as state of damage and numbers of casualty for all settlements affected and, for example, coordinates of the nearest airports can be provided in the CAP message. Also, information related to availability and status of resources such as a fire fighting vehicle or some other technical equipment like water pumps is shared using EDXL. For this purpose, the EDXL-RM (Resource Messaging) [5] standard is used. The EDXL-SitRep (Situation Report) [7] messaging standard is used within the IDIRA context for exchanging information on observations and situation reports, e.g., generated by commanders in the field.

Several software applications were developed within IDIRA to insert, produce, share and process disaster data. To support optimal decision finding and risk management various simulation tools were integrated. For example, a fire simulation tool (FireSim) can be used to simulate the spread of forest fires [20] and a chemical accident simulation tool (ChemSim) is used to simulate the propagation of possibly toxic gases and chemicals. These simulators are mostly based on weather data such as wind strength and direction and several geographic parameters such as soil conditions and even more specific parameters such as dissolution rates of chemicals. An evacuation simulator can be used to calculate the safest way for evacuating people out of districts also considering geographic information and observations such as obstacles or dangerous areas. For improved resource allocation an optimal spatial partitioning algorithm was used [21].

Further decision support systems are integrated into IDIRA covering routing and load balancing capabilities. Using these applications feasible paths for vehicles or relief workers can be calculated. Also, questions can be answered such as which unit can reach a certain destination within a given amount of time, or which unit can be at the destination most quickly. Load balancing algorithms are supported, e.g., to distribute injured people optimally to medical facilities while preventing from overload.

One further application of IDIRA is the integrated reporting system. Using the reporting application reports can be generated containing all meaningful information depending on the person the report is generated for. This helps protecting relief workers from information overload.

Another useful application developed within IDIRA is missing person tracing. This application helps to match data of missed persons with data of rescued persons across different tracing systems.

Furthermore, a software interface exists in IDIRA, which allows creating connections to external applications like specialized management tools used by a certain unit or organization. Usually, for data exchange between IDIRA and the external application, a standardized message format such as EDXL is used. The IDIRA information system is designed to gather from and to provide information to local command and control infrastructures. For example, in case of earthquakes, the information of external agencies such as the German Geoforschungszentrum, the US Geological Survey or the European Mediterranean Seismological Center can be induced into the IDRIA information system.

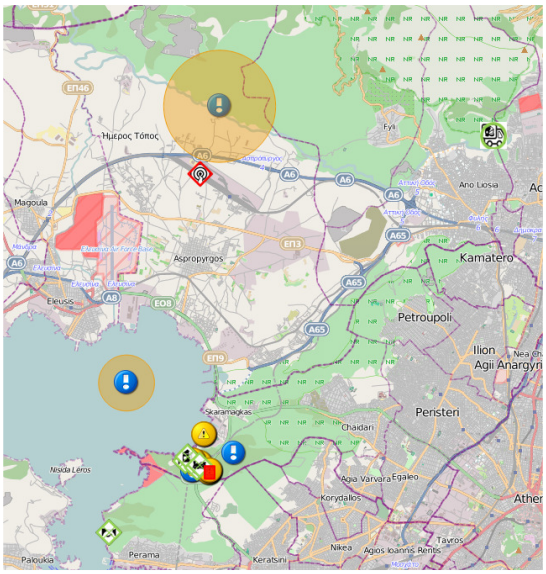


Figure 1: Disaster Information viewed by the COP.

The centerpiece of information presentation is the so called Common Operational Picture (COP). The COP is a web based application having its main goal to present all information in the system in an understandable and tailor-made way to the distinct users of the system, such as field commanders, tactical commanders, authoritative person or other stake holders within disaster relief work. Mainly, COP visualizes incidents, resources, tasks and other relevant information for disaster management on a map that can be seen in Figure 1. This map shows the epi center and affected regions of an earthquake, the position of resources or the location of incidents. Applying various filters and utilizing additional views on the available disaster information, COP helps to maintain an overview of the situation and protects from overlooking important details on a topic while simultaneously protecting the user from data overload. Furthermore, COP supports communication capabilities helping to get in touch with people involved in current relief work. Tasks can be assigned to field units, updates on tasks can be communicated by field units, text messages can be sent, or voice calls initiated.

The COP and most of the applications introduced above are running on hardware infrastructures specifically designed for IDIRA. This includes the so-called Fixed Infrastructure and a transportable compound called Mobile Information and Communication System (MICS).

The Fixed Infrastructure is a cloud computing infrastructure intended for high availability operation. It is located in a data center and reachable from the disaster area with permission as soon as there is access to the Internet. The Fixed Infrastructure together with the applications it is hosting acts as a central information hub where all disaster related information is stored and all persons and devices in disaster relief actions can access these data. The advantage of such a central data hub is that all users have the same view on the current state of a disaster.

While the Fixed Infrastructure is only accessible when Internet uplink is available in the disaster area and the Fixed Infrastructure itself is not affected by the disaster, a transportable version of this central information hub was designed with the MICS. The MICS can be shipped directly on-site and hosts the same services applications as the Fixed Infrastructure and runs them locally at the disaster area. This makes the need of an Internet uplink optional and the system is fully functional in absence of it. Nevertheless, if an Internet uplink is present the MICS establishes a VPN connection to the Fixed Infrastructure that is running an OpenVPN server. For the uplink, any broadband communication technology existing at the MICS location can be used. Likewise the Fixed Infrastructure, the MICS provides access to external expert systems locally (or via Internet – if accessible).

The IDIRA information system is the central information hub where disaster related information is accessible for relief workers and authoritative personnel. The second crucial part of the system is the communication system that grants access to the IDIRA information system and allows transmitting all meaningful information to mobile devices of field operators in action and will be presented in the next section.

IV. IDIRA Communication System

The IDIRA communication system is intended to connect devices at command and control centers as well as mobile devices to the IDIRA disaster information system. The IDIRA communication system fulfils a series of requirements, either brought in by first responder organizations or having its basis in the design principles of IDIRA itself.

The following requirements are the results of end user surveys and a detailed requirement analysis done during the project:

(1) The first requirement regards a largely unlimited and almost worldwide valid operation permission of the communication system. This is necessary as it would be a time-costly process to apply for operation permissions after a disaster has struck somewhere in the world. (2) Furthermore, the system should allow for integrating locally existing (and functional) broadband communication networks into the

IDIRA communication system. (3) The usage of open standards should ensure that a broad variety of end devices can easily be integrated into the communication network being able to exchange data with the IDIRA disaster information system. (4) Usually, relief organizations only have a few IT experts. This entails the requirement that especially the field components of the communication system can be easily installed and handled (even by non IT experts). (5) One further requirement concerning especially end devices is basic offline functionality. When the connection to the information hub is interrupted, users of end devices must be able to work with the system without great limitations. (6) The last requirement regards the provided bandwidth of the communication system. The system must provide enough bandwidth to guarantee a seamless interaction with the information hub (i.e., at least 2 Mbit/s for operating the COP on a mobile device). These requirements are described in more detail in [1].

With regard to the requirements for international operating permissions and the usage of open standards, the mobile communication system has been designed to use 802.11 based network technology. Additionally, 802.11 technologies are widely spread nowadays, thus, does not require high acquisition and operating expenditures. On the other hand, two major drawbacks were needed to overcome:

- Relief forces are not experienced in setting up a 802.11 communication network and,
- 802.11 provides communication coverage for small areas only.

This section presents a communication solution for a transportable 802.11 based communication network, which is intended to be installed right after a disaster strikes, bringing communication coverage to almost every place within a disaster area and, additionally, is fulfilling the requirements stated above.

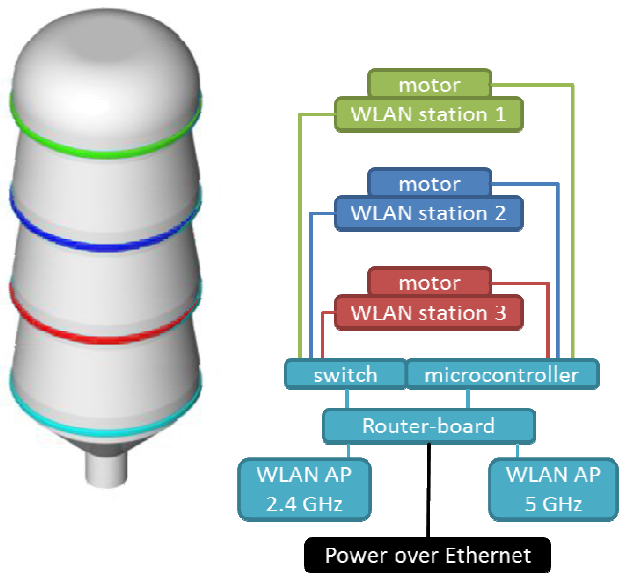


Figure 2: Schematic of the Wireless Gateway.

A. *Mobile Communication Equipment*

The central element of the proposed communication solution is a set of so called Wireless Gateway (WGW) devices.

The central ideas behind the WGW are that (1) multiple Wireless Gateways (WGWs) get automatically interconnected using directional antennas and 802.11 equipment. In comparison to omnidirectional antennas with directional antennas the signal quality between 2 WGWs can be improved. If two directional antennas are exactly aligned to each other the ratio between signal strength and noise level (SNR) increases and results in a higher possible throughput or larger distances between two WGWs. (2) After powering on, a WGW autonomously connects to other WGWs and starts building a meshed WLAN backbone network, which finally connects to the IDIRA disaster information system. (3) Additionally, each WGW provides communication coverage with a wireless hotspot and/or Ethernet LAN for end devices. (4) WGWs are designed as transportable devices and only need to be mounted on a pole and tripod before being powered on.

Following these 4 ideas, a system was designed that finally provides a solution for the two major drawbacks and the imposed requirements.

The main building blocks of the wireless gateway are presented as schematic in Figure 2. The system is mounted in a modular plastic housing consisting of four stacked layers. The top three layers (also referred to as modules) are equally built up. Each one is composed of a WLAN station, a directional antenna and a motor - all mounted on a turntable. The WLAN station supports wireless client mode and access point mode and is connected to a ~16 dBi directional antenna. The DC gear motor allows rotating the turntable by 360° on the horizontal plane. Keeping WLAN station, antenna and motor altogether on the turntable simplifies cabling of the devices and reduces the risk of entanglements when the turntable is rotating.

The bottom layer contains the control hardware and software of the WGW. A 5 port 100 Mbps Ethernet Switch connects the 3 top layer WLAN stations to a central router board containing the control logic, which is in charge of building the backbone connections to remote WGWs.

Figure 3 shows a schematic overview of the prototypes, which were developed based on this concept. The Router-board is an industrial grade embedded PC called Avila from manufacturer Gateworks. It is based on an Intel IXP425 CPU and features two 100 Mbit/s Ethernet ports and 4 MiniPCI slots. As operating system OpenWRT is running on the router board.

Furthermore, two out of the four MiniPCI slots of the router-board are equipped with CM9 MiniPCI WLAN cards from manufacturer Wistron NeWeb. The cards are based on Atheros AR5213A chips, one is operating in the 5 GHz band used for establishing the backbone connection between WGWs and the other operates in the 2.4 GHz band and is used for connecting end devices in the environment operated in the wireless cloud around the WGW.

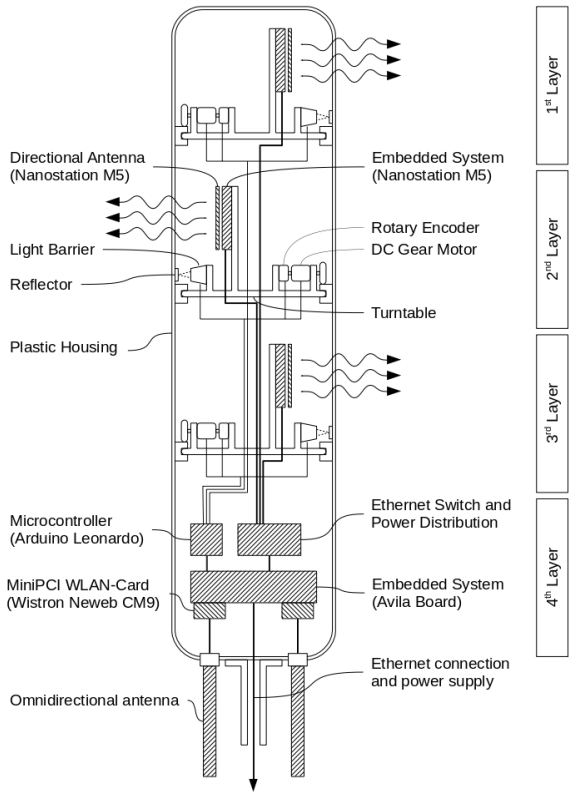


Figure 3: Prototype schematic overview.

The WLAN stations on the top three layers were implemented using a commercial product called Nanostation M5 from manufacturer Ubiquiti Networks [22]. For easier integration the plastic housing of the Nanostation M5 was removed and only the bare electronics were mounted on the turntables. To provide protection against physical influences the housing of the WGW covers all mechanical and electronic parts. The Nanostation M5 provides a built in 100 Mbit/s Ethernet interface used as data-link between the Nanostation M5 and the Router-board. The antenna inside the Nanostation M5 has a non-symmetrical radiation pattern of about 42° azimuth angle and 15° elevation angle. The devices are mounted 90° rotated so that the relevant radiation angle for the mechanical antenna alignment process is now the narrow 15° angle. The directional antenna is dual-polarized to support the MIMO feature of the 802.11n wireless interface. The installed MIMO antennas promise to enhance the signal quality and allow for higher bandwidth [23]. Figure 4 shows the radiation pattern of both polarization planes for the 15° beam width (green horizontal elevation, blue vertical elevation).

The Nanostations are configured to operate as router and run a modified Ubiquiti firmware supporting the OLSR routing protocol that is also used on the Avila router board. The OLSR configuration has been modified such that Ethernet links are generally preferred over wireless links.

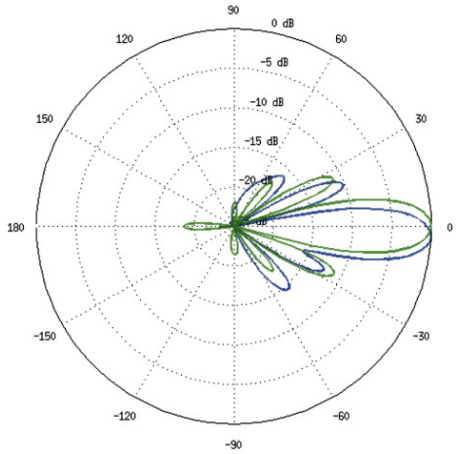


Figure 4: Nanostation M5 radioation pattern [22].

Additionally, an Arduino Leonardo microcontroller is connected to the router-board, which itself interfaces the motors of the top three layers to perform their rotation. The microcontroller tracks the position of each turntable by a light barrier. The light barrier is attached to the turntable itself together with a reflector attached to the outer housing of each layer. As soon as the light-barrier reaches the reflector the turntable is in home position.

The feedback signal of an incremental rotary controller (directly attached to the DC gear motor) is used to determine the exact position of the turntables. The rotary controller generates 2 signal patterns in order to determine both, the rotation direction of the turntable and to measure the turntable's alignment by counting the number of rotation steps.

The microcontroller is programmed to interpret text commands received over a serial interface. Furthermore, status information such as the current alignment can be requested from the microcontroller software. When the WGW is powered up, all turntables are aligned to their home position. During normal operation the microcontroller controls the DC motor and counts the pulses from the rotary encoders until the desired position is reached.

The microcontroller is not aware of cardinal directions. Instead it is only aware of the angular displacement of each turntable compared to its home position. The software accepts additional commands to store the actual turntable positions in a non-volatile memory. This stored position is recovered when the device powers up and the homing procedure is finished. The microcontroller is connected to the prototypes main logic board via serial interface.

The Switch on the base layer also provides power over Ethernet to the Nanostations on the 3 top layers. The microcontroller, the motors and the sensors are power supplied by a 5V DC/DC converter installed at the base layer.

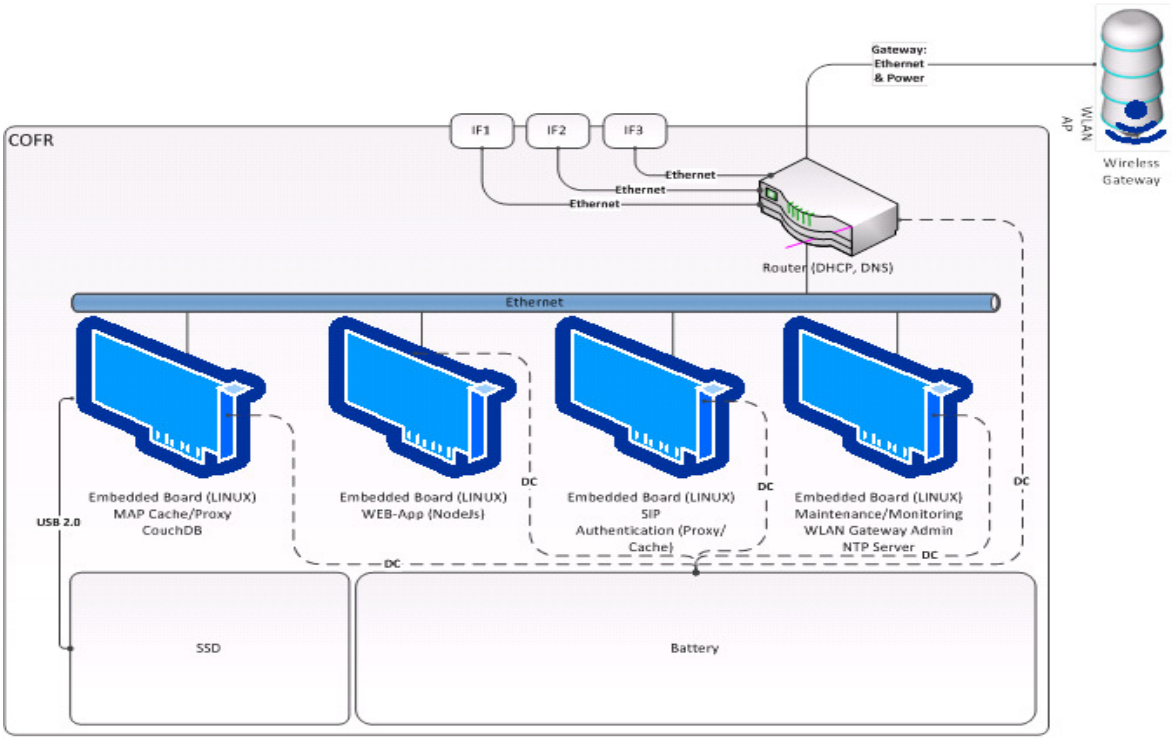


Figure 5: Building blocks of the Communication Field Relay.

The second part belonging to the mobile communication equipment is the so called Communication Field Relay or, for short, COFR. COFR and WGW are intended to be installed as a compound that provides connectivity to the MICS or Fixed Infrastructure for relief workers at any arbitrary location in the disaster area. The Communication Field Relay is positioned at the foot of the pole the WGW is mounted on, and it is connected to the WGW by Ethernet LAN and, thus, connected to the backbone network spanned by the WGWs. A schematic diagram of the COFR is shown in Figure 5.

The COFR is intended to provide several services for the IDIRA communication system. As a local communication hub for field commanders the COFR can provide a SIP service to allow for voice communication between field commanders without the need of a connection to the Fixed Infrastructure or the MICS – this grants more efficient bandwidth usage. Furthermore, the COFR can act as a communication proxy providing capabilities to cache data exchanged between end device applications and the IDIRA disaster information system. This guarantees a seamless operation even in cases the direct data exchange between an end device application and the central information system suffers from limited communication quality. The proxy service may also include a map server supplying geographic data to end devices. This has the advantage that potentially large map data are not needed to be transmitted multiple times for multiple end devices over a potentially constraint wireless backbone connection.

From a networking perspective the COFR offers a DHCP server and DNS server capabilities granting simplified end device configuration. End devices automatically get a valid IP configuration (DNS, IP address, gateway address, etc.) after the device is attached to the network. These services are supplied to devices using the wireless connection to the 802.11 hotspot (spanned by the WGW) or to devices directly wire-connected to the COFR via Ethernet. One Ethernet port of the COFR is especially considered for this case. A second Ethernet port is intended to be used as direct Internet uplink. Any arbitrary Internet uplink technology such as DSL, WiMAX, satellite-communication, UMTS, or LTE can be used. This Internet uplink can be shared by all clients connected to this COFR, to the local WGW or, to any remote COFR or WGW. The route is distributed by the OLSR dynamic gateway plugin. A third Ethernet port is used for the connection to the WGW and is providing - beneath communication capabilities - power supply to the WGW via Power over Ethernet (PoE). Therefore, the COFR can be connected to a power source either by a 230 V power socket or, if a power line or power generator is not available, to a battery via a 12 V cigarette lighter socket.

A COFR is assembled of several embedded Linux boards (Raspberry PI) providing the software services mentioned above. As data storage a fast and energy efficient solid-state disk is included. Furthermore, a router allowing for connecting different hosts via cable and ensuring the connection to the Wireless Gateway (WGW) is part of the system. The router is also responsible for the Internet uplink.

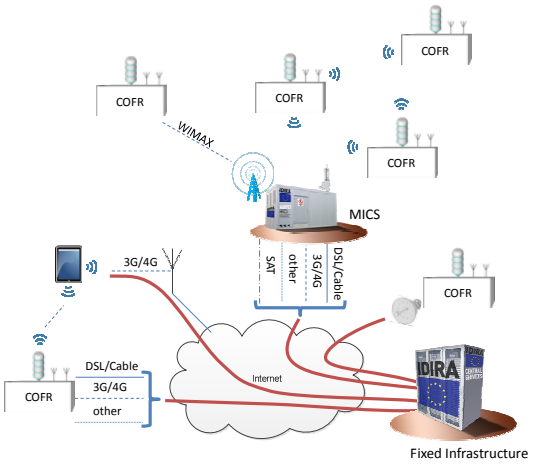


Figure 6: IDIRA communication network.

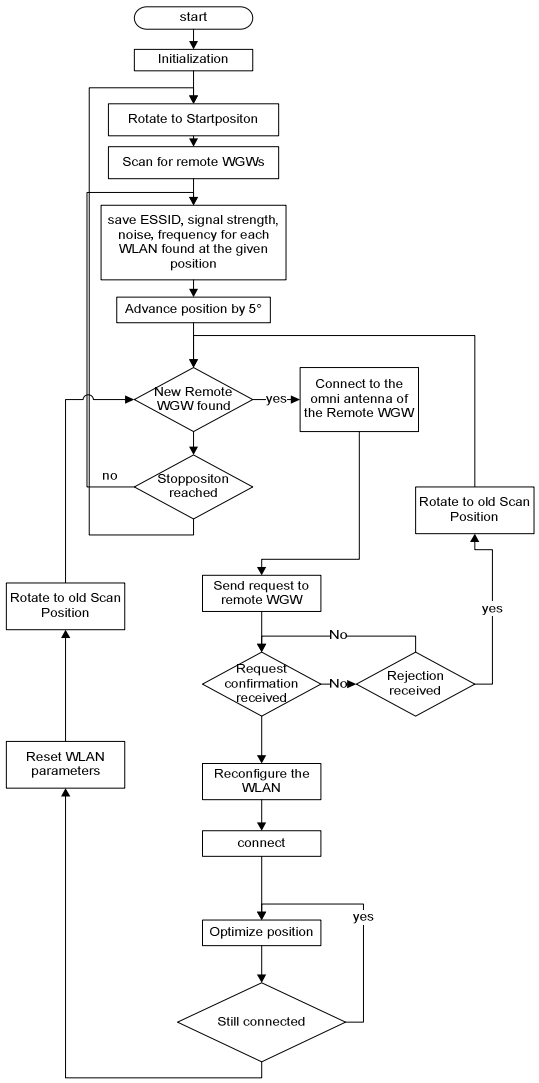


Figure 7: Local Alignment Algorithm

Figure 6 shows a full-featured disaster communication network based on the proposed components. The depicted network consists of the MICS (installed at the Command and Control Center on-site), several mobile communication sites (WGW/COFR) and also the Fixed Infrastructure which is connected over Internet. The figure also shows that WGW/COFR compounds are using a variety of connection methods either to the MICS (via Wireless LAN or WIMAX) or to the Fixed Infrastructure (via various Internet uplink technologies). A fully operational mobile communication node consists of a WGW/COFR compound, a battery pack with capacity for about 12 hours, together with a tripod and a telescopic 6 m pole.

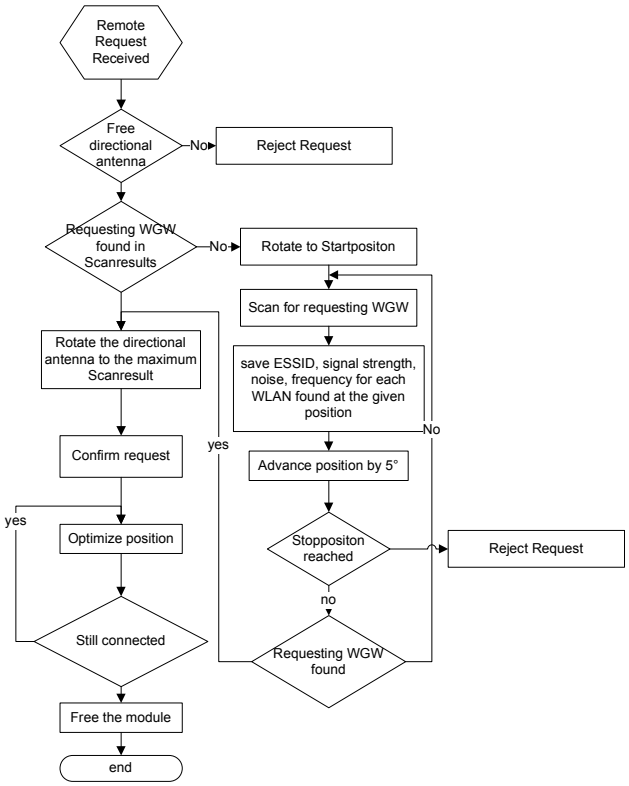


Figure 8: Remote Alignment Algorithm

B. *Antenna Alignment and Networking Configuration*

After powered on, the WGWs try to connect to remote WGWs following an alignment algorithm. Figure 7 and Figure 8 show a simplified flow-diagram of the alignment algorithm performed by a requesting (local) WGW and a responding (remote) WGW. Figure 9 shows the state diagram the alignment algorithm is based on. The state diagram shows the individual states of each of the three top layers of the WGW. The alignment algorithm is executed on the router-board on the lowest level of the WGW.

After supplying power to the WGW via PoE, the local adjustment sequence (see Figure 7) starts an initialization process. This initialization process identifies the home

positions of the modules and performs a self-check of the system. After the initialization, all three modules start to scan for remote WGW signals radiated by their 5GHz omni-directional antenna. For each module, individual start and stop positions are defined with an offset of 120° to each other. This allows for scanning the full 360° around the WGWs as fast as possible. During the scanning process, the antenna position is advanced by 5° in each step. One step lasts about 10 seconds, which is mainly determined by the time needed to execute the scan. The scan is stopped as soon as a module reaches its stop position or a remote WGW has been detected. When the stop position has been reached, the antenna is rotated to its starting position and the scanning process is restarted again.

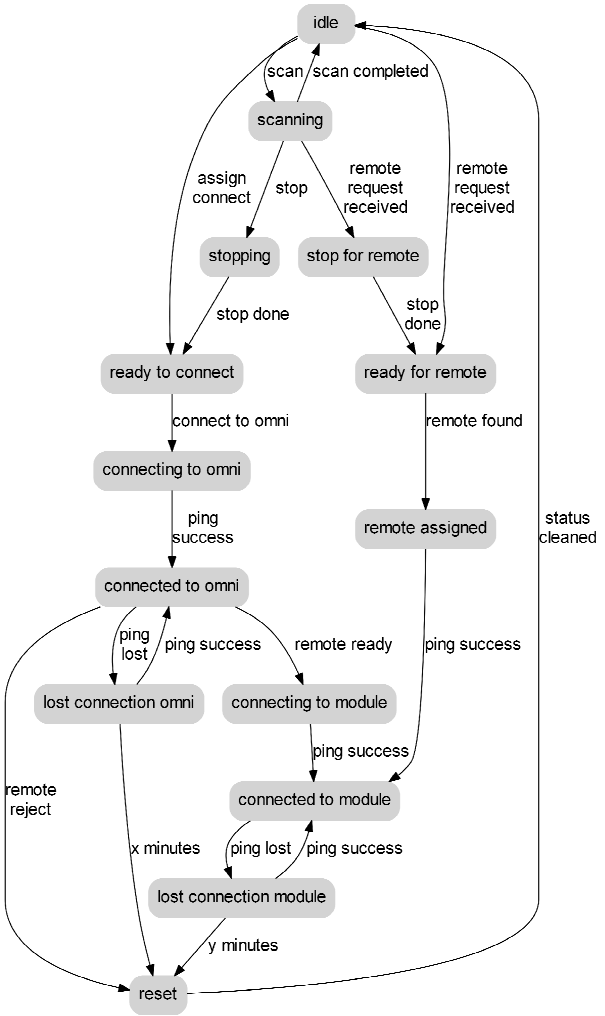


Figure 9: Alignment process state diagram.

In case of the 5GHz remote signal has been detected with signal strength greater than -87 dbm, one directional antenna of the detecting WGW rotates to the position with the best strength of this signal. This antenna may not be the same as the one detecting the signal.

During the scanning phase, the Nanostations M5 are configured to operate in access point (AP) mode as the firmware of the M5 only allows in AP mode to execute distinct scans for each scanning position. In station mode, in contrast, scan results are cached by the firmware over several scans. This, however, would make it impossible to map certain scan results to distinct positions.

After the module has reached the best known position for this signal, the module is configured in station mode to connect to the omnidirectional antenna of the remote WGW. As IP based communication is needed for the following steps, also the IP configuration of the local WGW is done such that the local module configures itself with an IP address of the remote WGW. As for performance constraints no DHCP server is running at the remote WGW and a distinct mapping of IP addresses to WLAN SSIDs is used. Each WLAN spanned by the omnidirectional antenna of the WGWs is sending a distinct SSID. Based on this SSID, the alignment and configuration algorithm knows the IP address a connecting module needs to build up an IP based connection. An example for a connection to an omnidirectional antenna is shown in Figure 10. After a connection is established and the corresponding IP configuration is done, the local WGW sends a request to connect to a directional antenna at the remote WGW.

At this time, the local WGW and the remote WGW are following distinct algorithm steps to establish the directional point-to-point connection. While the connecting WGW is following the remaining steps shown in Figure 7 (starting with checking if a request confirmation was received), the remote WGW will follow the sequence shown in Figure 8, as soon as a request has been received.

When a connection request is received, one module is determined for the directional connection. If no module is available (because all are used for other connections) a reject is sent to the requesting WGW. Otherwise, the scan results will be searched for results of the requesting WGW. If such a result exists the determined module is rotated to the corresponding position with the best signal strength and a confirmation is sent to the requesting WGW. If such a scan result cannot be found, all available modules starting a 360° full scan beginning from their start positions. This scan is executed until the requesting WGW is found or the stop position is reached. If the signal of the requesting WGW has been found, a confirmation is sent and a module is rotated to the position with the best signal strength known. Otherwise, the request is rejected.

When the requesting node receives the confirmation the connection to the remote WGW omnidirectional antenna is canceled, and the local WGW is configured to connect to the remote directional module. The local module is configured in station mode with the SSID of the remote module and the IP address is set according to the SSID of the remote module using a similar approach as when connecting to the omni-directional antenna. Furthermore, also the IP address is set appropriately for the remote module (see Figure 10). If the request is rejected the module will continue scanning for remote WGWs.

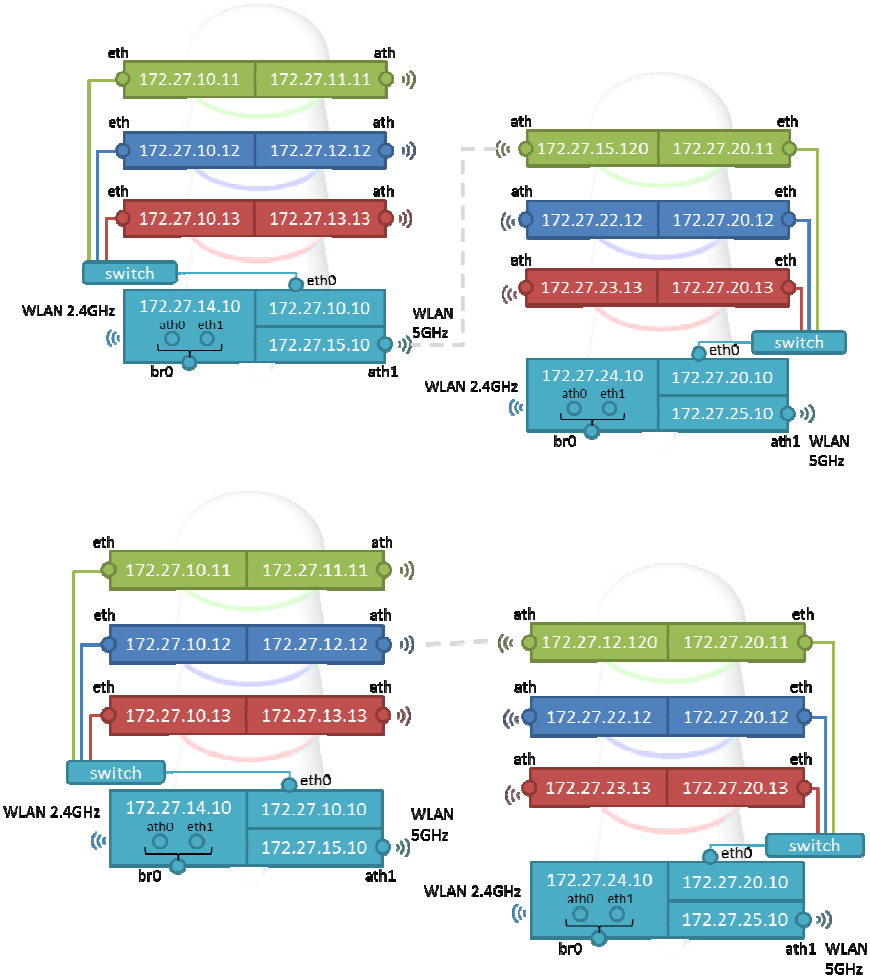


Figure 10: IP addressing scheme when connected to the omnidirectional antenna (above) and after successful connection between two WGW devices (below).

The alignment sequence typically lasts between five and 15 minutes. All established connections are monitored, and if one of them is lost the module will be reconfigured and starts to scan for remote WGWs again or connect to another previously located WGW. Not connected modules will continue to scan for remote WGWs. New scan results (coming from modules in scanning state) are used to adjust the position to ensure the best possible position to the remote WGWs based on the signal strength.

Another view on the alignment procedure can be given by the state diagram shown in Figure 9. This diagram shows the states of each module and all transitions between them. State transitions are executed by a central control instance separately for each module.

After powering up and during node initialization, the modules are in idle state. From the idle state, the control instance may trigger to perform a scan, trigger to connect to a remote WGW if a remote WGW is already known or being ready to be assigned to a remote WGW if a connection request is received from a remote WGW.

When a module completes a scan (after reaching the end position) it returns into the idle state and is realigned and triggered to start another scan. If a remote WGW is identified, the control instance choses one scanning or idle module that should connect to the remote node and switches it to state “ready to connect”.

If a remote request is received, the control instance also choses one scanning or idle module that is then used for establishing the directed point-to-point connection. The module is switched in state “ready for remote”. If the requesting WGW has been identified the module is rotated towards the remote node position and its state is changed to “remote assigned”.

Once the connection is established, the module’s state is changed to “connected to module” and this connection is regularly checked. If the connection is lost it will change the state to “lost connection module” and try to re-establish the connection for y minutes. If it fails to re-establish the connection, it will change the state to “reset” and finally to “idle”.

If a remote WGW has been identified in "scanning" state without receiving a request from it (i.e., the local WGW found the remote WGW first), one available local module will be reconfigured with an appropriate IP address to be able to connect to the 5GHz omni-directional antenna of the remote WGW. Also, the module is set to state "connected to omni" and the remote WGW is informed about the attempt to establish a point-to-point connection. If the connection to the remote WGW is lost, it is tried to re-establish the connection for x minutes, afterwards it will change the state to "reset" and finally to idle after all specific settings are reset. In case that the remote WGW node answers the request in a positive manner, the module is reconfigured and a connection to the remote module will be established.

C. Node Positioning Support

To install a field communication site is easy as it only requires mounting the WGW on a pole and connecting it to a COFR and a power source. But before a communication site can be established, one important question needs to be answered: Which location is particularly appropriate to setup a communication site where a WGW provide WLAN coverage for relief forces and is able to build up a backbone connection to other remote WGWs.

This section describes the Reachability Optimized Positioning (ROP) application of IDIRA. ROP provides a Web based interface, which is fully integrated into COP and helps to find the best possible locations for setting up communication sites. Commanding personnel can run WLAN coverage simulations at arbitrary locations on the COP map in order to evaluate the WLAN coverage at this place regarding to range and signal quality of the directional antennas. This information is then used by early responder teams to identify the optimal location for a communication site where a direct line of sight is available between multiple WGWs.

ROP calculates the radio signal propagation based on a digital earth surface model of the operational area. For this purpose, an extension of the open source tool SPLAT! [24] version 1.4 was developed, which uses a surface model with a resolution of $1/10^{th}$ of an arc second. SPLAT! provides radio signal propagation based on a terrain analysis for the electromagnetic spectrum between 20 MHz and 20 GHz. The calculations are based on the Longley-Rice Irregular Terrain [25] as well as the new Irregular Terrain with Obstructions (ITWOM v3.0) [26] model. In its base version SPLAT! uses the elevation data from the U.S. Geological Survey and Space Shuttle Radar Topography Mission [27]. These data have a resolution of 1 arc second for some areas of the Earth's surface and 3 arc seconds for the remaining areas.

To achieve precise results in a radio wave propagation simulation this resolution is too coarse grained. To solve this issue an Earth surface data basis with a high resolution of $1/10^{th}$ of the Earth surface was chosen, which is available from some satellite remote sensing programs such as TerraSAR-X [28] or from local authorities for some specific regions. This is where an extension of SPLAT! was necessary, as higher resolved Earth surface data are not supported by SPLAT!. To make highly resolved elevation data usable in SPLAT!, the application had to be extended in order to allow SPLAT! to read, use and visualize this kind of elevation and surface data and also the algorithm to compute radio wave propagation was slightly adapted to the new data basis. With the increased resolution to $1/10^{th}$ of an arc second the distance between points with available elevation data is approximately 3 m (for central Europe). This gives sufficient accurate propagation models to have guaranteed communication channels between WGWs. The coverage simulation also considers the operating height of the WGW of about 6 m and the results show if it is possible to establish line of sight communication between two WGWs absent of obstructions due to buildings, hills, or forests.

To find the appropriate places and areas, the commanding staff starts a signal propagation simulation with a pole at the location of the MICS. The result of the calculation is a picture of the signal propagation simulation shown in Figure 11 as an overlay of the COP map.

The white section in the circle indicates an area where it is possible to deploy the wireless gateways and to establish a communication channel to remote WGWs automatically. The red or dark grey area indicates that it is not possible to establish a communication infrastructure due to obstacles between the directional antennas. In the middle of the circle, the green or light grey area gives the commander the information that it is possible to support mobile equipment for communication in the incident area. These simulation results are presented within COP together with incident locations. Consequently, within one system tactical needs as well as communication needs can be taken into account when decisions for operational locations in the field of first responder field commanders have to be defined.

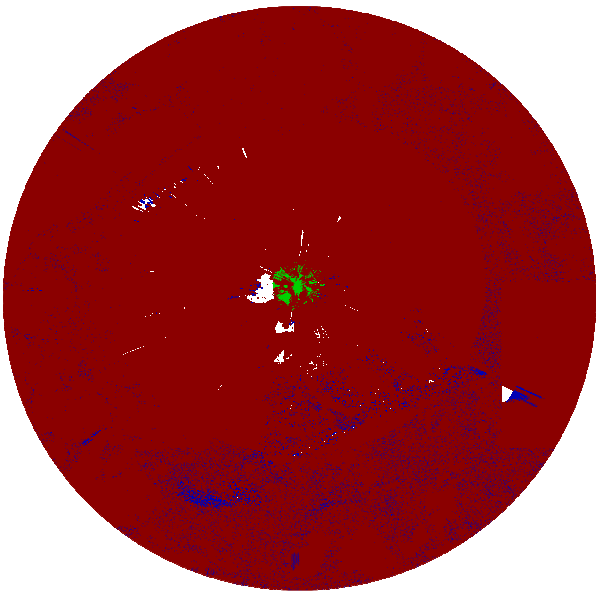


Figure 11: Result of a radio signal propagation simulation.

ROPs usability and correct way of working were proven in a validation test before it was used in multiple field trials

of the IDIRA communication system. The result of the validation test is shown in Figure 12.

Based on a simulation of the base area where the command and control center was located, ten places were defined to validate the possibility to establish a communication channel fully automatically (Point 1 - Point 10).

After the wireless gateways were deployed to the different places it was evaluated if a communication channel could be automatically established by the alignment algorithm:

- Wireless gateways placed on areas indicating a good signal to noise ratio (green and yellow, respectively light grey areas), could successfully setup a communication channel: Yellow pins (Points 2, 4, 5, 9, and 10).
- Wireless gateways placed in red or dark grey areas failed to setup a communication channel automatically: Red pins (Points 1, 3, 6, 7, and 8).

These results show that the accuracy of the radio signal propagation simulation was sufficient to give a reliable answer to the question where communication sites should be established in order to build a backbone network allowing for a connection to the IDIRA information system.

More detailed performance evaluation results of the IDIRA communication system can be found in [1][29] [30][31]. These papers contain results of several performance tests, end user training events, and large scale exercises held in context of the IDIRA project.

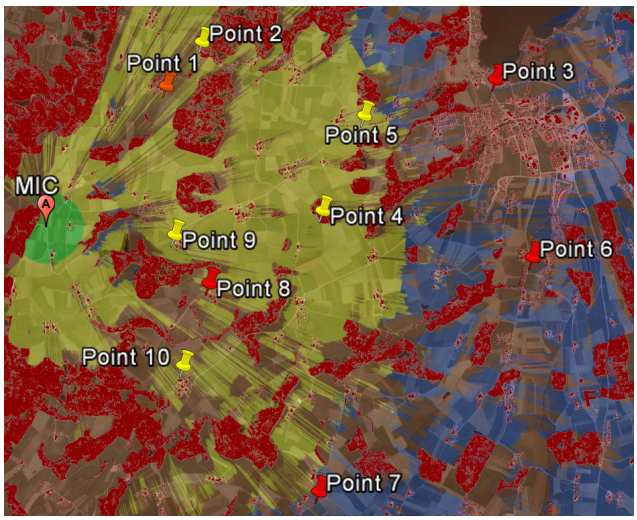


Figure 12: Simulation validation.

V. Conclustion and Future Work

This paper presents the information and communication systems developed within the EU funded project IDIRA. The main goal of IDIRA was to develop a solution to enhance interoperability and cooperation of relief units part of multinational disaster response organizations working together after a large scale disaster. Such a solution considers two aspects of interoperability. Organizational aspects dealing with the administrative coordination of various disaster relief organizations, and technical aspects to find technological solutions to enhance information interchange.

This paper focuses on the latter and presents various applications referred to as the IDIRA information system, which can help to find right decisions quickly and provide a common sight on what is happening within the disaster relief action. Furthermore, and with even more focus on details, a mobile communication system is presented providing wireless communication at almost any location within the disaster area.

This communication system complies with several requirements that have been introduced by action forces of relief organizations, such as easy installation and transportation, interoperability with existing communication systems and, international operation permission. The core of this system is the WGW/COFR compound to be installed out in the field of a disaster area granting wireless communication capabilities to field personnel. The field personnel is able to access the central information system of IDIRA. To allow this, in the background the WGW establishes a wireless connection to the central information system potentially using multiple other WGWs as wireless communication hops. The COFR provides power supply and Internet uplink to the compound. An additional application was developed helping first responders to setup the WGW/COFR compound at the right location, where it is possible to build up a wireless backbone network and supply an area near an operation site with a 802.11 wireless hotspot.

In several large scale exercises and user training events the usability of the IDIRA system has been proven. In these events, however, it was shown that several enhancements could improve the systems performance and should be considered in future. (1) Especially the WGWs mechanics should be built in a more robust way in order to make the system more capable for conditions in disaster operations. (2) The extension of the alignment algorithm with manual provided additional information could speed up the automatic alignment process. (3) Additional software interfaces to further existing disaster management tools and a broader variety of sensor sources will be provided.

Acknowledgment

This work was partially supported by the IDIRA European FP7 261726 research project.

References

[1] Peter Dorfinger, Ferdinand von Tüllenburg, Georg Panholzer, and Thomas Pfeiffenberger, “A Flexible Self-Aligning Communication Solution for Multinational Large Scale Disaster Operations,” Proc. of the ICN 2015: The Fourteenth International Conference on Networks (ICN2015), April 19 - 24, 2015 - Barcelona, Spain, pp. 230-236.

[2] Nalini Suparamaniam and Sidney Dekker, “Paradoxes of power: the separation of knowledge and authority in international disaster relief work,“ Disaster Prevention and Management, vol. 12, no. 4, pp. 312–318, 2003.

[3] IDIRA Project. *Interoperability of data and procedures in large-scale multinational disaster response actions, 2011-2015* [Online]. Available from: http://idira.eu/. 2015.11.30.

[4] OASIS Emergency Management TC. *Emergency Data Exchange Language (EDXL)* [Online]. Available from: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=emergency 2015.11.30.

[5] OASIS Emergency Management TC. *Common Alerting Protocol Version 1.2, 2010-07-01 - CAP-v1.2-os* [Online].Available from: https://www.oasis-open.org/standards#capv1.2 2015.11.30.

[6] OASIS Emergency Management TC. *Emergency Data Exchange Language Resource Messaging, EDXL-RM-v1.0-OS-errata-os, 22 Dec. 2009* [Online]. Available from: https://www.oasis-open.org/standards#edxlrm-v1.0 2015.11.30.

[7] OASIS Emergency Management TC. *Emergency Data Exchange Language Situation Reporting, edxl-sitrep-v1.0-wd19, Draft 02, 2012-08-07* [Online].Available from: http://docs.oasis-open.org/emergency/edxl-sitrep/v1.0/cs01/edxl-sitrep-v1.0-cs01.zip 2015.11.30.

[8] ETSI TR 103 269-1. *TETRA and Critical Communications Evolution (TCCE); Critical Communications Architecture; Part 1: Critical Communications Architecture Reference Model* [Online] Available from: http://www.etsi.org/deliver/etsi_tr/103200_103299/10326901/01.01.01_60/tr_10326901v010101p.pdf 2015.09.14.

[9] Immarsat BGAN. *Broadband Global Area Network* [Online]. Available from: http://www.inmarsat.com/service/bgan/ 2015.11.30.

[10] GVF VSAT. *Global Very Small Aperture Terminal Forum* [Online]. Available from: http://www.gvf.org 2015.09.14.

[11] Emergency.lu. [Online]. Available from: http://www.emergency.lu 2015.09.14.

[12] IEEE 802.16 WIMAX. *IEEE Standard for Local and metropolitan area networks* [Online]. Available from: http://standards.ieee.org/about/get/802/802.16.html 2015.11.30.

[13] IABG mbH. *HiMoNN Higly Mobile Network Node* [Online]. Available from: http://www.himonn.de 2015.08.31.

[14] Electronic Communications Committee (ECC). *The Identification of Frequency Bands for the Implementation of Broad Band Disaster Relief (BBDR) Radio Applications in the 5 GHz Frequency Range* [Online]. Available from: http://www.erodocdb.dk/docs/doc98/official/pdf/REC0804.pdf 2015.11.30.

[15] Christian Raffelsberger and Hermann Hellwagner, "Evaluation of MANET Routing Protocols in a Realistic Emergency Response Scenario," Proc. of the 10th Workshop of Intelligent Solutions in Embedded Systems (WISES'12). July 2012, pp. 88-92.

[16] Matthias Herlich and Shigeki Yamada. "Motivation for a Step-by-Step Guide to Set up Wireless Disaster Recovery Networks," Proc. of the International Conference on Information and Communication Technologies for Disaster Management (ICT-DM 2015).

[17] IEEE 802.11. *IEEE Standard for Information Technology--Telecommunications and Information Exchange Between Systems--Local and Metropolitan Area Networks--Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY)* [Online]. Available from: http://standards.ieee.org/about/get/802/802.11.html 2015.11.30.

[18] Thomas Clausen and Philippe Jacquet. (2003), "Optimized Link State Routing Protocol (OLSR)," Internet Engineering Task Force, IETF, RFC 3626.

[19] Harald Rieser, Peter Dorfinger, Vangelis Nomikos, and Vassilis Papataxiarhis, "Sensor Interoperability for Disaster Management," Proc. of the Sensor Applications Symposium (SAS2015) Zadar, Croatia; April 2015, pp. 389-395.

[20] Satways, OptiFire [Online] Available from: , http://www.satways.net/index.php?option=com_content&view=article&id=95:optifire-article&catid=53:simulation-a-modelling&Itemid=91&lang= 2015.11.30.

[21] Kostas Kolomvatsos, Kakia Panagidi, and Stathes Hadjiefthymiades, "Optimal Spatial Partitioning for Resource Allocation," Proc. of the 10th Int. ISCRAM Conference. May 2013.

[22] Ubiquiti networks. *NanostationM & NanostationlocoM Datasheets* [Online] Available from: http://dl.ubnt.com/datasheets/nanostationm/nsm_ds_web.pdf 2015.11.30.

[23] David Gesbert and Jabran Akhtar, "Breaking the barriers of Shannon's capacity: An overview of MIMO wireless systems", Telenor's Journal: Telektronikk, vol. 98, no 1, pp. 53-64, 2002.

[24] John A. Magliacane. (2002) Splat! [Online]. Available from: http://www.qsl.net/kd2bd/splat.html 2015.11.30.

[25] Anita Longley and Phill Rice, "Irregular terrain model," Intstitute for Telecommunication Sciences, 1968.

[26] Georgelpon Hufford, "The its irregular terrain model, ver 1.2.2," National Telecommunicatgions and Information Administration.

[27] U.s. geological survey. U.S. Geological Survey. [Online]. Available from: http://srtm.usgs.gov/ 2015.08.31..

[28] Airbus Defence & Space. TerraSAR-X Radar Satellite Imagery. [Online] Available from: http://www.geo-airbusds.com/terrasar-x/ 2015.11.30.

[29] Thomas Pfeiffenberger, Peter Dorfinger, and Ferdinand von Tüllenburg "Communication Coverage Awareness for Self-aligning Wireless Communication in Disaster Operations," Pervasive Networks for Emergency Management March 2015 St. Louis, Missouri, USA.

[30] Peter Dorfinger, Ferdinand von Tüllenburg, Georg Panholzer, Massimo Cristaldi, Giovanni Tusa, and Franz Böhm" An Offline Capable Communication Framework for Multinational Disaster Operations based on Self-aligning Wireless Gateways," Proceedings of International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC2015) Dubai, UAE January 2015.

[31] Peter Dorfinger, Georg Panholzer, Ferdinand von Tüllenburg, Massimo Christaldi, Giovanni Tusa, and Franz Böhm "Self-aligning Wireless Communication for First Responder Organizations in Interoperable Emergency Scenarios," In: Proc. of the 2014 International Conference on Wireless Networks, Las Vegas Nevada, USA.

ConEx Performance Evaluation and Application to Video Streaming

Ali Sanhaji, Philippe Niger, and Philippe Cadro
Orange Labs
2, avenue Pierre Marzin,
22300 Lannion, France
Email: {ali.sanhaji, philippe.niger, philippe.cadro}@orange.com

André-Luc Beylot
INP-ENSEEIHT, IRIT Laboratory,
2, rue Charles Camichel,
31071 Toulouse Cedex 7, France
Email: andre-luc.beylot@enseeiht.fr

***Abstract*—With Internet traffic ever increasing, network congestion should occur more and more frequently. During congestion periods, some users contribute more than others to the congestion in the network. It might be interesting for a network operator to differentiate between users proportionally to the congestion they induce, but the necessary information for this purpose is not available at the network layer, and is exchanged at the transport layer (e.g., Transmission Control Protocol (TCP) acks). This led the Internet Engineering Task Force (IETF) to design Congestion Exposure (ConEx), a new mechanism to expose to the network the amount of congestion a user is responsible for, allowing the network operator to improve the fairness between users. ConEx is designed to limit the added complexity, leveraging already existing mechanisms such as Random Early Detection (RED) and Explicit Congestion Notification (ECN), plus a number of modifications to the senders and receivers to be fully operational. Nonetheless, ConEx can also be deployed in a simplified mode of operation, relying only on loss information in DropTail queues to estimate congestion. The objective of this paper is to provide an in depth evaluation of ConEx mechanism. Firstly, we investigate how the setting of ConEx parameters (e.g., congestion policer) and the network configuration (e.g., router queuing, network delay, etc.,) impact the behavior of ConEx and influence its ability to improve fairness between users. Secondly, we compare the level of performance, in terms of fairness improvement, provided by different variants of ConEx of increasing complexity, i.e., from a simple implementation with modifications limited to the sender to a "full" ConEx approach implementing all proposed features. We show that, despite a reduced accuracy in congestion estimation, a simple variant of ConEx is already able to provide a good fairness improvement between users. This is particularly interesting in the context of an initial deployment scenario, allowing an incremental deployment of ConEx. Thirdly, we investigate and discuss the limitations and weaknesses presented by ConEx with regard to short-lived flows. Finally, based on a YouTube traffic model, we illustrate how ConEx can help to enhance the Quality of Experience (QoE) of video streaming users during congestion periods, significantly reducing the number and duration of stalling events.**

Keywords-ConEx; Performance; ECN; Congestion; Policing; YouTube; LEDBAT.

I. INTRODUCTION

This paper complements the investigation of ConEx presented in [1], adding new simulation results and more detailed discussions.

During the network's busy hours, an amount of traffic greater than what the network can handle leads to congestion, affecting the quality of experience of many users. Yet, this great amount of traffic is mainly caused by a small percentage of users, often referred as "heavy" users. For example, in Orange's Fiber To The Home (FTTH) access networks, 80% of downstream traffic is generated by 15% of the customers [2]. In order to improve the user's network experience, while restraining the network costs, the aim is to convince these heavy users to yield network resources during congestion periods for the benefit of everybody.

Some traffic management approaches are already implemented by network operators, like rate-limiting traffic or defining Data-Volume caps above which the users are slowed down or stopped. However, these solutions show limited efficiency because they do not consider the network state, i.e., if it is congested or not. A heavy user can be rate-limited even when he does not hamper the experience of the others, or when there are plenty network resources available, which would allow his traffic to be far much faster. Similarly, a heavy user might consume his allowed Data-Volume even when the network is not in a congestion phase, which can be perceived as largely unjustified. It would be fairer to limit the users according to how much congestion they induced. For this, we would need the information about the congestion encountered by the users. This valuable congestion information is generally available to the end-to-end flow control algorithms, for example, it can be exchanged between the users at the transport layer (e.g., through TCP acks), but it is transparent for the network layer. As the network elements operate at the network layer, they cannot have access to congestion information.

To counter this lack of information at the network layer, the IETF designed ConEx, which is a mechanism that allows the sender to inform the network about the congestion encountered [3]. The amount of lost and congestion marked packets exposed by a user defines a new metric called the Congestion-Volume, which is a more useful metric than Data-Volume because it reports directly the congestion in the network.

In order to minimize the implementation complexity, ConEx largely relies on existing mechanisms (e.g., RED, ECN capability on routers, TCP exchanges), and on new features added to both the sender and the receiver to be fully ConEx-capable. Considering the initial deployment of ConEx, we are interested in whether or not ConEx still presents good performance without the use of ECN in the network and relying only on minimal modifications to the user's end devices.

The additions to [1] are the following: firstly, the impact of the network configuration (e.g., router queueing, network delay) on ConEx mechanism is evaluated to determine how it may influence its ability to improve fairness between users. The sensitivity of ConEx to its environment is a key factor when considering its deployment in a real network. Secondly, the introduction of a new step of deployment in the analysis of the performance of ConEx variants with an increasing

implementation complexity, is also a valuable addition. It enhances the understanding of how all the ConEx components interact to achieve the goal of improving fairness between users.

We will first present in Section II the related work on ConEx. Section III will describe the ConEx principle and the mechanisms on which it relies. The performance evaluation of ConEx with and without ECN using long-lived flows is presented in Section IV while the short-lived flows issue will be discussed in Section V. Our interest will be focused, in Section VI, on how ConEx can be useful in the case of video streaming traffic to enhance the users' QoE, with scenarios using a YouTube traffic model, and how heavy users can take advantage in using a congestion control algorithm like Low Extra Delay Background Transport (LEDBAT). Section VII summarises the main outcomes of the study, finally, Section VIII discusses the future work, still waiting to be covered.

II. Related Work

The IETF has set up since June 2010 a working group to develop experimental specifications of ConEx in IPv6 networks [3]. A Request For Comments (RFC) [4] discussing the concepts and use cases has been published, and other documents concerning the ConEx mechanism have also been produced and are waiting for final adoption: the use of a destination option in the IPv6 Header to carry the ConEx markings [5], a mobile communications use case for congestion exposure [6] and the necessary modifications to TCP [7].

Re-ECN is a "pre-ConEx" implementation solution to allow congestion exposure for IPv4 networks. A thorough description and analysis of the Re-ECN mechanism has been done under the Trilogy project [8]. This work had a great influence for the emergence of the ConEx working group.

Some papers focused on the performance evaluation of the congestion exposure mechanism through the evaluation of Re-ECN in multiple scenarios. [9] developed a Linux implementation of Re-ECN and performed several simulations showing the great dependency of the Re-ECN information to the flow size, the Round Trip Time (RTT) and the Active Queue Management (AQM) parameters. [10] evaluates mobility issues with congestion exposure and shows that mobility is not a major concern for Re-ECN. [11] evaluates Re-ECN applicability in LTE networks and found that it can bring a significant improvement for these networks unless they experience a severe packet loss rate. All these papers rely on the use of ECN to signal congestion; to our knowledge, no performance evaluation of ConEx has been made solely based on loss exposure.

III. Congestion Exposure

In this section, we will describe ConEx, how it operates to expose congestion, along with the other mechanisms used to collect congestion information and control the user's traffic.

A. ConEx mechanism

Figure 1 shows the whole ConEx process and all the elements involved with it, in case of TCP traffic, which is the primarily target for ConEx. The ConEx mechanism works as follows: a transport sender starts by sending a data packet in the network, this packet might encounter one or several congested routers along its path. The packet will either be lost or ECN marked (by setting the Congestion Experienced (CE) codepoint in the IP header [12]) by the congested routers. This information about loss or marking will be detected by the transport receiver, and through the TCP acknowledgments, the receiver will feedback this information to the sender. With the use of ConEx, the sender will reinject this feedback to the network in the IP packet headers (e.g., use of the RE bit in Section III-D), which will hold the Re-Echo signals. Detecting a loss will generate a Re-Echo-Loss signal from the sender, while an ECN marked packet will generate a Re-Echo-ECN signal.

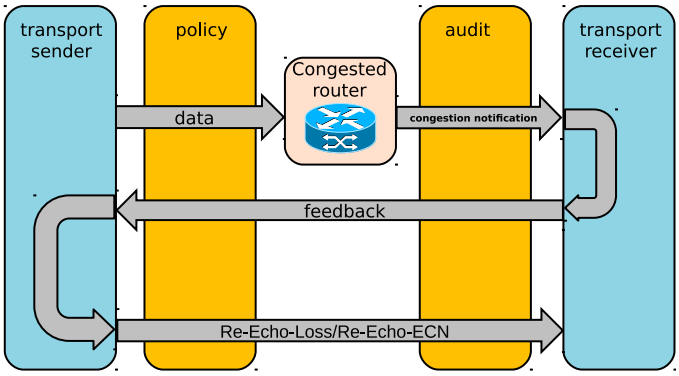


Figure 1. ConEx mechanism

The information provided by ConEx can then be used by the network operator for traffic management through a congestion policer for example. At the ingress of the network, a congestion policer counts the congested packets and takes traffic control policy decisions (e.g., discard, deprioritize packets using Differentiated Services (DiffServ)) if the user has consumed the congestion-volume he was allowed. At the egress of the network, an auditor might be used to ensure that the senders are exposing the right amount of congestion to the network. It helps as prevention from users understating the congestion their flows encounter, to preserve their congestion allowance and avoid policing. If the sources are trusted ones, for example, if the sources are controlled by the network operator or if there is an agreement between the sender and the network, the auditor is unnecessary and can be omitted. As reliable auditing is a complex task this greatly simplifies the deployment of ConEx.

B. Random Early Detection

Random Early Detection is an Active Queue Management technique, implemented on many routers, which was first introduced in [13]. It allows to randomly drop or ECN mark packets according to a probability that increases from 0 to the maximum probability p_{max} when the mean queue length increases from a minimum threshold to a maximum threshold (see Figure 2). Above the maximum threshold, all packets are either dropped or marked if ECN is used (the "gentle mode" was introduced later on to fix the problem caused by the discontinuity of the marking or drop probability when the queue length exceeds the max threshold).

C. Explicit Congestion Notification

Explicit Congestion Notification [12] is a way to indicate the occurrence of congestion in the network without having to drop packets. It uses two ECN bits [ECT,CE] of the IP header to signal congestion to the receiver.

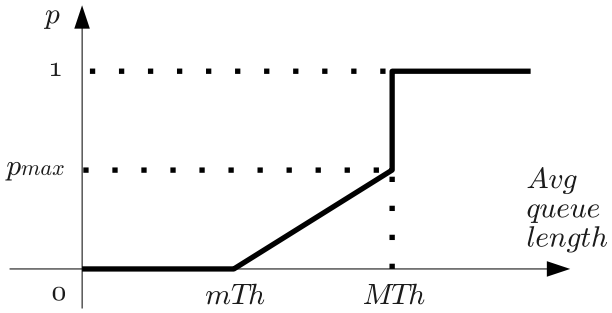


Figure 2. Random Early Detection dropping/ marking function

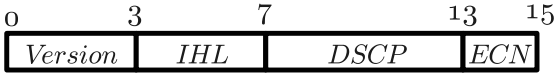


Figure 3. Bytes 1 and 2 of IPv4 header

D. Re-ECN

Re-ECN is a candidate implementation of ConEx for IPv4 [8]. It uses the bit 48 (RE bit) of the IPv4 header to extend the ECN field to a 3-bit field, allowing 8 codepoints. These codepoints identify the ConEx signals as described in Table I.

TABLE I. ConEx signals with Re-ECN encoding

ECN field	RE bit	ConEx signal
00	1	Credit (Used with the auditor)
01	1	ConEx-Not-Marked (ConEx-Capable)
01	0	Re-Echo-ECN or Re-Echo-Loss
11	1	ECN marked packet
11	0	Re-Echo packet and ECN-marked
10	0	ECN legacy (Not-ConEx)
00	0	Not-ECN (Not-ConEx)
10	1	Unused

E. TCP modifications

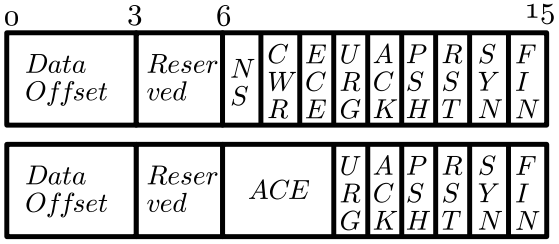


Figure 4. Bytes 13 and 14 of TCP header

The classic ECN mechanism as described in [12] allows the receiver to feedback only one CE mark per RTT. Indeed, even if several packets of the same flow get CE marked during one RTT, the receiver has only one bit (ECN-Echo (ECE) flag in the TCP header) to feedback all the marks. The information about how many packets have been marked is valuable for ConEx but also for other mechanisms like DCTCP [14], modifications to TCP are needed to provide more than one feedback per RTT. [15] and [16] propose a solution to achieve such a goal. They suggest overloading the three TCP flags ECE, Congestion Window Reduced (CWR) and Nonce Sum (NS) to form a 3-bit field, the ACE field as shown in Figure 4. This field would act as a counter for the number of CE marks seen by the receiver, which can feedback it to the sender. The sender is then able to accurately follow the evolution of ECN markings and report the right amount of Re-Echo-ECN signals. The use of accurate ECN feedback is negotiated during the TCP three-way handshake.

F. Congestion policer

The great advantage brought by ConEx is the possibility for the network operator to police the users proportionally to their contribution to congestion, thus to the impact they have on other users. Based on the ConEx signal the policing can be applied at the ingress of the network, which is far more efficient than a policing at the egress by the auditor as it prevents the heavy users from overloading the network.

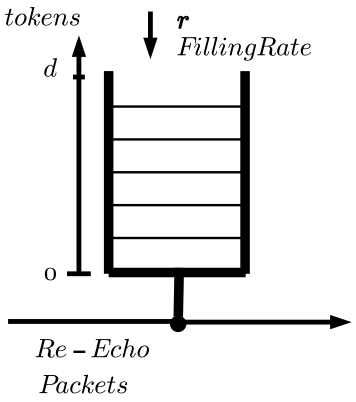


Figure 5. The congestion policer as a token bucket

The congestion policer can be implemented as a token bucket, as in Figure 5, with a filling rate r (the allowed Congestion-Rate) and a depth d (the allowed Congestion-Burst). In a byte-based mode of operation, the policer removes the same amount of tokens from the bucket as there are bytes in the Re-Echo-ECN/Re-Echo-Loss packets sent by a user. When the bucket empties, the policer proceeds to discard the packets of the user who exceeded his allowed Congestion-Volume. A packet-based policer, which does not consider the size of packets, can also be used, resulting in a simpler but potentially less accurate traffic control in case of heterogeneous packet sizes.

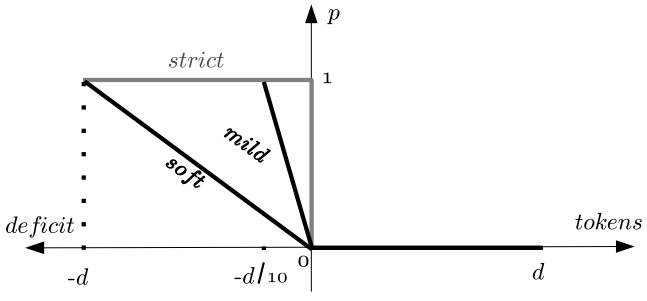


Figure 6. Drop function of the congestion policer

As shown in Figure 6, three levels of policing are used for performance evaluation characterized by their drop function: the strict policer discards all packets when the bucket is empty, the mild policer discards packets with a probability linearly

increasing from 0 to 1 when the bucket depth decreases from 0 to $-d/10$ and the soft policer discards packets with a probability linearly increasing from 0 to 1 when the bucket depth decreases from 0 to $-d$.

IV. LONG-LIVED FLOWS

A. Simulated Network

To perform the simulations, we used the Network Simulator 2 (NS2) [17] in which we implemented ConEx following the latest RFCs and drafts and we used the IPv4 proposal presented in Section III-D. The simulated network is depicted in Figure 7. There are 100 users on either side of the network, each single user on the right receiving traffic from a single user on the left. 90 of them are light users using only one File Transfer Protocol (FTP) flow each as a traffic source. The other 10 users are heavy users, they use 36 FTP flows each as a traffic source, they will thereby be responsible for 80% of the traffic on the bottleneck. The TCP senders use cubic as a congestion control algorithm with Selective Acknowledgments (SACK) and TimeStamps options. The TCP receivers can feedback ECN markings in an accurate count to the sender, which in turn will send a Re-Echo-ECN/Re-Echo-Loss signal for every ECN-marked/lost packet. The TCP maximum window value is equal to 64KB while the packet size is equal to 1500 bytes.

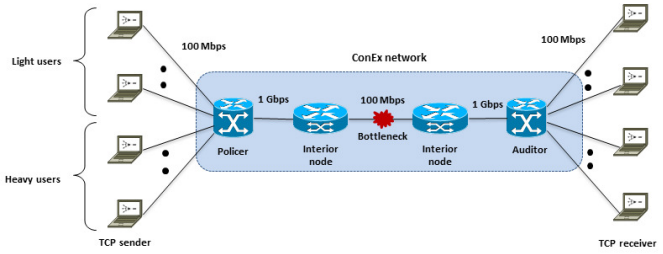


Figure 7. Simulated network topology

Unless specifically mentioned, the following configuration is used. All users have a minimum Round Trip Time of 100ms (due to the propagation time on the links) and share a 100Mbps bottleneck. At the ingress of the network, there is a per user congestion policer, which is implemented as described in Section III-F. The action taken by the policer is dropping the user's packets when the bucket, which has a depth of 64KB (or 45 packets), is emptied. On the bottleneck's router, there is a RED queue with a length equal to the Bandwidth Delay Product (BDP) in order to hold 100ms of the bottleneck's traffic. The probability of marking packets increases from 0 to $p_{max} = 1$ as the average queue length increases from 10% to 100% of the total queue length. At the egress of the network, there is an auditor that is deactivated because we use trusted sources, i.e., sources that are fully compliant to the behavior specified for ConEx.

Each simulation has a duration of 100s and is run 30 times to have proper 95% confidence intervals for each point. For greater visibility of the graphs, these intervals are not depicted when their value is around 1% of the metric's mean. The traffic sources are saturated (i.e., sending at their maximum possible rate) and each flow starts randomly and uniformly between 0 and 300ms.

TCP provides a flow-based fairness, meaning that a user can get more bandwidth share if he uses more flows. The per user congestion policer does not consider the user's flows individually but only the aggregate traffic of the user to monitor the amount of congestion induced in the network. The purpose of ConEx is to improve fairness between users, especially between the light user and the heavy user, which is useful for a network operator, as providing fairness between its customers in their use of the network is necessary. So, monitoring the impact of the mechanism on the fairness between the users is valuable. Therefore, we will be monitoring a metric defined in [18]:

$$unfairness = \frac{throughput\ of\ a\ heavy\ user}{throughput\ of\ a\ light\ user} \quad (1)$$

In the following sections, we will firstly evaluate the impact of the internal and external parameters of ConEx on its performance. The internal parameters are the ones that come with the implementation of ConEx i.e., the congestion policer configuration (filling rate, harshness of the policer and bucket depth). The external parameters are the ones that come from the environment in which ConEx operates: the main network parameters (i.e., queuing strategy and delay) and the TCP congestion control algorithm (i.e., cubic and compound). Afterwards, we will compare the performance of ConEx for a set of implementation variants of increasing complexity, corresponding to increasing steps of deployment, a very essential consideration for a network operator.

TABLE II. Parameter values summary

Parameter	Values
Policer Filling Rate r (packets/s)	$1-2-3-4-5$ $10-15-20-25-30$ $45-90-120-180$
Policer Depth d (packets)	$5-12-23-$ **45** -90
Policer Harshness	Soft−**Mild**−Strict
Fixed RTT (ms)	$20-50-$ **100** $-150-200$
TCP Congestion Control	**Cubic**−Compound
Queue Size q (ms)	$10-20-50-$ **100** -200
Queue MinThresh mTh (%q)	$2.5-5-$ **10** $-20-50$
Queue MaxThresh MTh (%q)	$20-40-60-80-$ **100**
Queue Maximum Mark/Drop Probability	$0.1-0.25-0.5-0.75-$ **1**
ConEx Complexity	DTConEx−REDConEx ECNConEx−**FullConEx**

Table II summarizes the evaluated parameters and their values in the simulations. The value in bold represents the default value of the parameter, used when no value is specifically mentioned in the text.

B. Policer harshness

Figure 8 represents the average unfairness versus the allowed filling rate of a user in the simulation. Each curve represents a level of harshness of the policer as explained in Section III-F. The straight red curve on top is the unfairness when no policing is applied (the policer is deactivated). Only TCP is performing congestion control and TCP induces fairness between flows; as a heavy user has 36 flows and a light user has only one, the unfairness is equal to 36 as expected. When the policer is activated (the three remaining curves), the heavy users are the ones that will be the most policed.

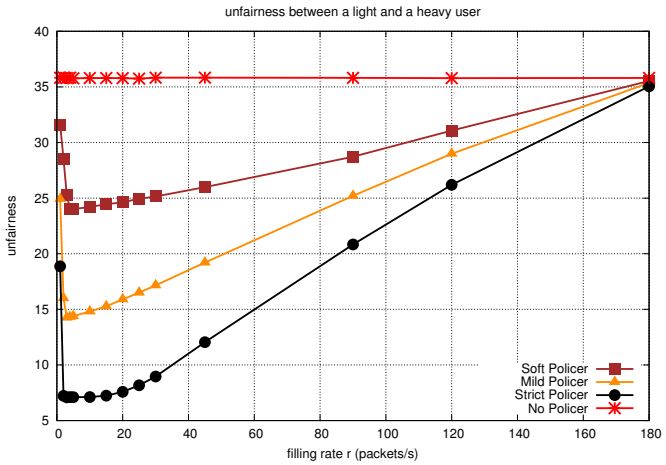


Figure 8. Unfairness between a heavy and a light user

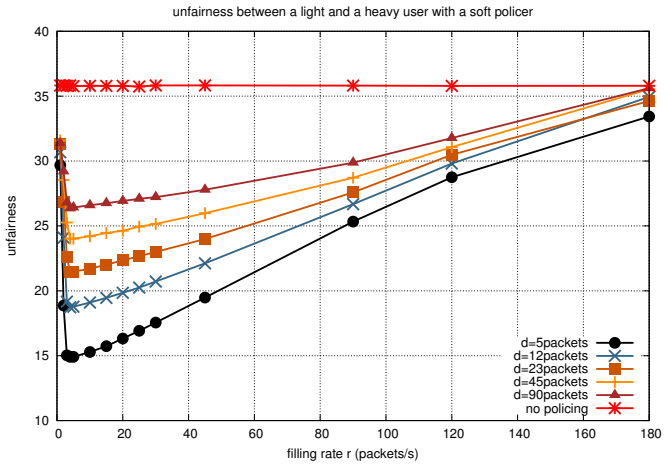


Figure 9. Unfairness with different depth values (in packets) with a soft policer

As the heavy users are forced to reduce their throughput, the light users occupy the freed bandwidth and the unfairness is reduced.

In Figure 8, the unfairness presents a minimum value suggesting an optimal filling rate. On the two sides of the optimum, the unfairness increases but for two different reasons. On the right side, as the filling rate increases, the heavy users undergo less policing. They get a higher throughput than with the optimal filling rate and the unfairness increases. When the filling rate is high enough, the heavy users avoid the policer's intervention, so the unfairness reaches the value obtained without policing ($unfairness = 36$). On the left side of the optimum, both the heavy users and the light users are policed because of the insufficient filling rate. The light users are forced to reduce their throughput and the unfairness increases compared to the unfairness obtained with the optimal filling rate. Policing the light users is counter-productive if the purpose is to reduce unfairness between light and heavy users; one has to attribute filling rates, which will avoid the light users from being policed while keeping the heavy users from overloading the network during busy hours.

To evaluate the impact of the harshness of the policer, a soft, a mild and a strict policer are used, which drop packets with increasing aggressiveness. Figure 8 shows that the three policers present the same optimal filling rate but are different in decreasing the unfairness. The harsher is the policing, the lower is the unfairness, because the heavy users will need to further reduce their throughput due to the policer's higher dropping probability. The difference between the policers is substantial because when the policer drops packets, the ConEx-enabled source will react by sending more Re-Echo-Loss packets, which will eventually lead to more policing. With a severe policer, the risk is to have a user continually decreasing his throughput because of the policer's actions, even when the network becomes uncongested. This potential artefact should be taken into account in the design of the policer's algorithm.

C. Token Bucket Depth

The depth of the token bucket corresponds to the burst of ConEx signals the network operator allows a user to send, it conditions the quickness of the policer to take action against a user inducing congestion. The token bucket depth is involved in the dropping function of the soft and mild policer (the smaller the bucket, the steeper the dropping function) as shown in Section III-F, thus the depth also represents the progressivity of the policing in relation to the congestion induced. Figure 9 and Figure 10 represent the unfairness between users versus the filling rate for different depth values in the case of a soft and a mild policer respectively. In both cases, as the token bucket depth decreases, policing becomes more reactive to congestion, and less permissive towards heavy users, so the unfairness decreases as the heavy users react to dropped packets by reducing their throughput.

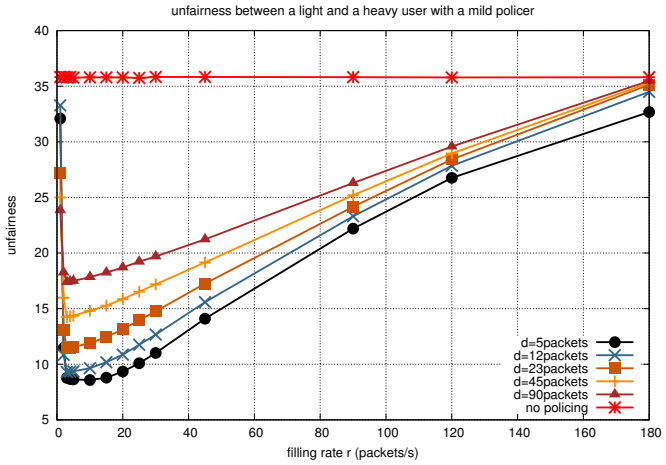


Figure 10. Unfairness with different depth values (in packets) with a mild policer

When the drop function is independent of the bucket length, like in the case of a strict policer (cf. Figure 11), only the reactivity of the policing is affected by the depth of the bucket, the harshness is not. As noticed in Figure 11, there is almost no difference in the decrease of the unfairness between the different depth values: long-lived flows can react to congestion by adapting their throughput, thus they induce a steady rate of congestion signals. If this congestion-rate is greater than the filling rate, it is only a matter of time for the bucket to be emptied completely, and for the policer to start taking action against the congestion inducing user, with the same harshness whatever the bucket depth. In the long run, the unfairness decrease would not be affected by the bucket depth

as much as by the filling rate and the dropping probability. So the reactivity of the policing, conditioned by the depth of the token bucket, is less significant than the harshness, conditioned by the filling rate and the dropping function steepness, in reducing the unfairness.

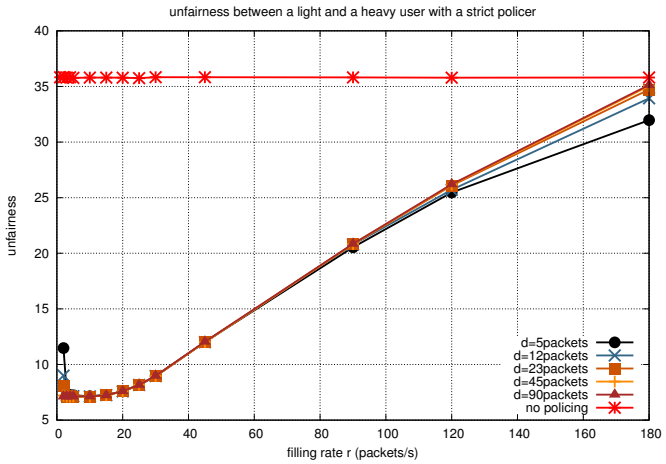


Figure 11. Unfairness with different depth values (in packets) with a strict policer

To proceed in the performance evaluation of ConEx, we use a mild policer and vary the values of the other parameters involved in the control of the mechanism or likely to influence its behavior. We will evaluate next the effect of the Round Trip Time.

D. Round Trip Time

We vary the minimum RTT (due to the propagation time in the links) of the users from 20ms to 200ms in the simulations. Having a short RTT allows a flow to quickly increase its congestion window, reaching a high throughput over a short time period. It also increases the probability for its packets to get marked or dropped: in a single second a flow have many "round trips" of traffic in the network, potentially resulting in many packets queued at the bottleneck and marked or dropped in case of congestion. This could drastically increase the number of Re-Echo-ECN and Re-Echo-Loss packets sent by a flow over a given time period. The user could then experience a high congestion-rate, rapidly consume the tokens in the token bucket when compared to the allocated filling rate, leading to a more severe policing. As shown in Figure 12, the unfairness significantly decreases with the RTT, and the difference observed between the curves is important, particularly for a RTT below 100ms. For the shortest round trip times (RTT below 100ms), even the highest filling rate ($r = 180 packets/s$) is not sufficient to allow the heavy users to deal with the congestion-rate they induce in the network, so the unfairness is still low even with a high filling rate. The results presented here clearly show that the round trip time is a very influential parameter for ConEx mechanism, especially in the design of the congestion policer algorithm.

E. TCP congestion control algorithm

In Figure 13, two popular TCP congestion control algorithms are compared, cubic and compound. Cubic is a more aggressive algorithm than compound that occupies more bandwidth, and can lead to more congestion on the bottleneck.

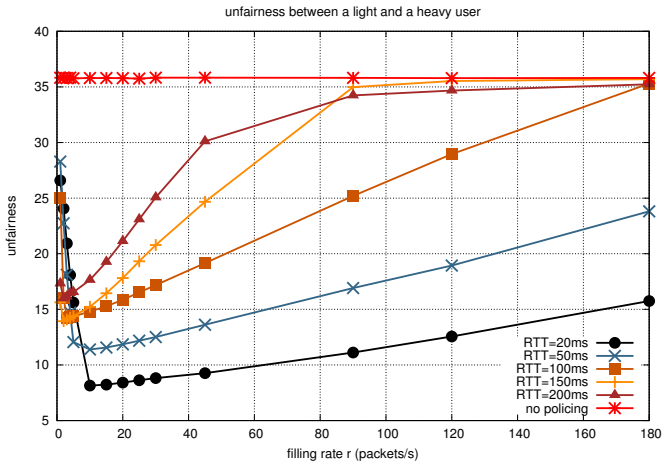


Figure 12. Unfairness with different RTTs

As a consequence, similarly to what was observed for a short RTT, it can be expected that cubic consumes more tokens, leads to more sanctions towards the heavy users hence results in a lower unfairness than compound TCP. On the contrary, the results show that compound TCP manages is more effective in reducing unfairness between users. In addition, the difference between cubic and compound remains almost constant when the filling rate varies, meaning that they do not have a significant impact on the behaviour of ConEx. The results show that the aggressiveness of cubic algorithm leads to more losses and markings during congestion, even for light users, forcing them to reduce their throughput more often and leading to a less effective unfairness reduction than with compound.

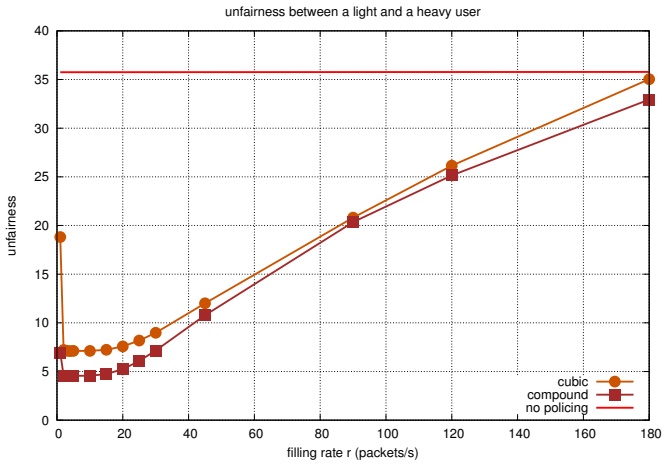


Figure 13. Unfairness with cubic and compound

Regarding more specifically ConEx, the obtained results show that both congestion control algorithms have a very similar behaviour in the presence of ConEx, resulting in roughly the same level of performance. This means that in the presence of ConEx, the users should be treated in the same way by the network, whatever their congestion control algorithm is, resulting in a relative independence regarding the type of device and/or the type of operating system.

F. Queue parameters

Many Active Queueing Management techniques and queueing algorithms can be used in a network depending on the objective the network operator aims at (reducing the queueing delay, the jitter, etc.), and these algorithms might have many parameters of their own, which will affect the behaviour of ConEx. In the case of the Random Early Detection algorithm, four parameters are involved: the queue size, the minimum threshold, the maximum threshold and the maximum marking probability. In this section, we will investigate the impact of each parameter on the behaviour of ConEx and try to quantify how it affects the effectiveness in reducing the unfairness.

In our evaluation the RED queue is used to mark packets, it only drops packets when it overflows. An ECN marked packets has two effects and provides two way to reduce the traffic load in the network. Firstly, it forces the traffic source to reduce its window size to lower its sending rate at a maximum rate of one time per RTT, similarly to what is done when detecting a packet loss. Secondly, it forces the traffic source to send a Re-Echo-ECN packet, consuming tokens at the policer and increasing the probability for the source to be policed.

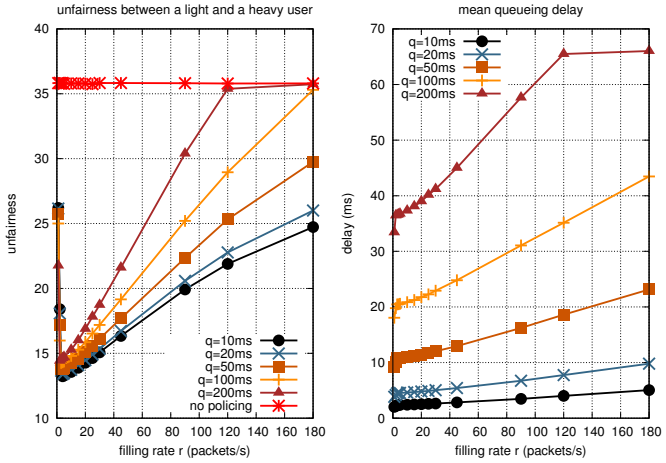


Figure 14. Unfairness with different queue sizes

1) Queue size: We vary the queue size to hold from 10ms to 200ms of the bottleneck traffic. A small queue quickly overloads, leading rapidly to an important number of packet drops. In the case of RED, as the marking probability increases with the mean queue length (cf. Section IV-A), the fraction of marked packets also increases rapidly for a small queue. Thus, the congestion induced by a user increases as the queue size decreases, with more and more Re-Echo-Loss and Re-Echo-ECN packets sent by the heavy users. These heavy users are policed and forced to reduce their throughput, and, as shown in Figure 14, the unfairness decreases significantly with the decrease of the queue length. It should be noticed that the decrease of the unfairness is also affected by the shorter RTT due to the shorter queue, as it is explained in Section IV-D, a short RTT leads to a reduced unfairness.

Policing the heavy users at the ingress of the network reduces the congestion in the network and allows the light users to have a greater share of the available bandwidth, but it also reduces the delay in the bottleneck queue. As the filling rate decreases, making the policing harsher, the mean queueing delay decreases, as shown in Figure 14. The value of the queueing delay without congestion policing is close to the value of the queueing delay when $r = 180 packets/s$, as with this filling rate, the heavy users are almost not policed.

For a given filling rate, the queueing delay exhibits important variations, which is natural as the queues varies in size, but compared to the impact on the unfairness, the impact of the queue size on the delay is much more significant. Thus, it can be considered that the impact of the queue size on the delay is much more a concern than its impact on ConEx.

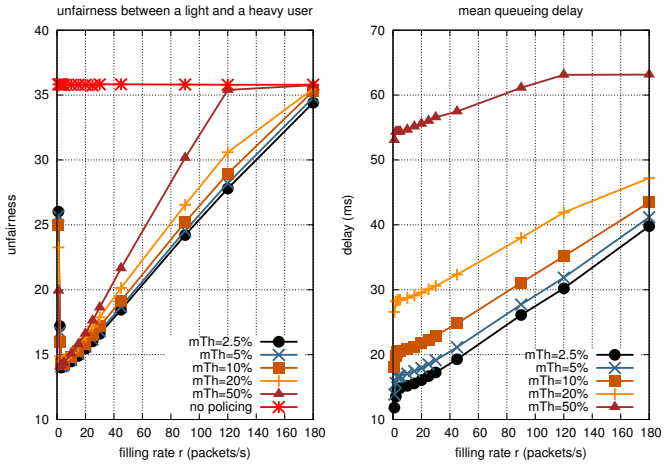


Figure 15. Unfairness with different minimum thresholds

2) Minimum Threshold: The minimum threshold determines when the queue will start marking packets. It can be expected that the sooner the marking will begin, the more Re-Echo-ECN packets will be sent by the users inducing congestion, consuming more tokens and being more severely policed. In Figure 15, the unfairness effectively decreases with the decrease of the minimum threshold, however, the unfairness is only slightly affected by the variation of the minimum threshold, particularly for the values of filling rate leading to the lower unfairness, making ConEx relatively insensitive to the minimum threshold setting. This can be explained by the fact that as the maximum threshold and the marking probability remained unchanged, a low minimum threshold results in a more reactive but less aggressive marking process.

On contrast the queueing delay is largely influenced by the variation of the minimum threshold, as already observed previously for the queue size variation case. This is quite natural as one of the main motivations for introducing RED was to control the queuing delay. Here again it can be concluded that setting the minimum threshold is primarily a question regarding the control of the queuing delay, with limited influence on the behavior of ConEx.

3) Maximum Threshold: When the mean queue length exceeds the maximum threshold, all packets are marked, so the lower the threshold the greater fraction of Re-Echo-ECN packets that will be sent by heavy users inducing congestion. Thus, the policer is more severe towards these heavy users and forces them to reduce their throughput. In Figure 16, the unfairness decreases with the decrease of the maximum threshold, because of the higher fraction of ECN marked packets in the queue, but as already observed for the minimum threshold, the unfairness is only slightly affected by the variation of the maximum threshold. The queueing delay also decreases with the maximum threshold, and here again, the queueing delay is

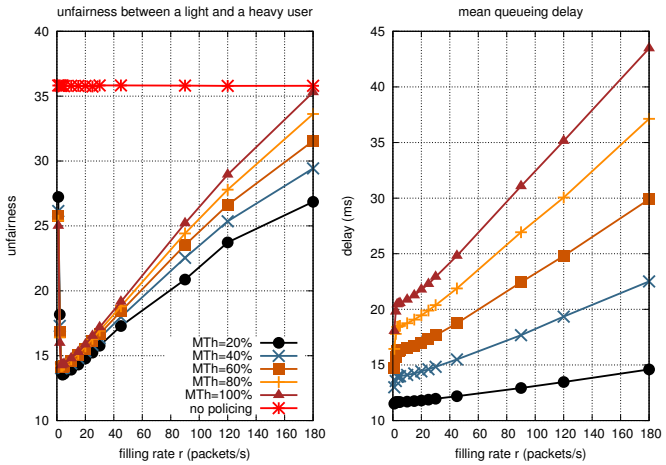


Figure 16. Unfairness with different maximum thresholds

widely impacted by this variation. From these results ConEx appears as also relatively insensitive to the maximum threshold setting. The drawback with a low maximum threshold is that the light users also experiment a high ratio of ECN marked packets in the queue. As they have to react to these marked packets, they are unable to increase their throughput when the heavy users are policed and the bottleneck freed.

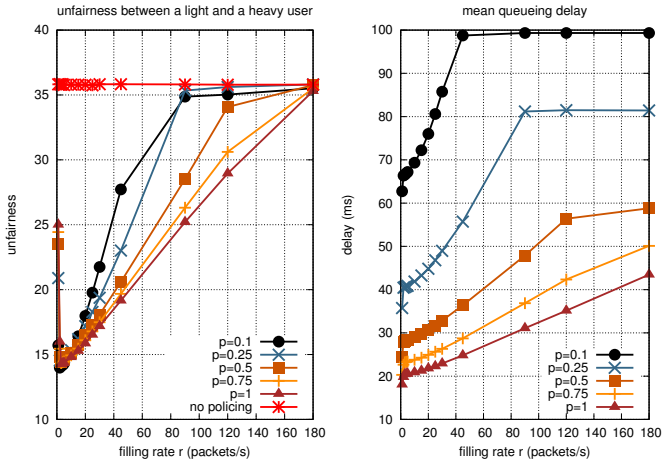


Figure 17. Unfairness with different maximum marking probabilities

4) Maximum marking probability: The maximum marking probability is reached when the mean queue length is equal to the maximum threshold, the higher the probability the more packets are marked so the more Re-Echo-ECN signals are sent by heavy users, making the policer harsher towards their traffic. As presented in Figure 17, a higher maximum marking probability decreases the unfairness, and it also decreases the queueing delay because of both the marking in the queue and the heavy users' policing at the ingress of the network.

The queueing delay is greatly affected by the variation of the marking probability. For a low marking probability the users experiment only small limitations and can largely (e.g., for a marking probability equal to 0.25), or completely (e.g., for a marking probability equal to 0.1), fill the queue, resulting in a high queuing delay. As the marking probability increases the user's limitation is more severe, resulting in a less loaded queue and a reduced delay. For low filling rate congestion policing is more severe and the queuing delay reduction is emphasized.

On contrast, unfairness exhibits a smaller sensitivity to the maximum marking probability. For a low filling rate (i.e., below 20 packets/s), the allowance of Re-Echo-ECN packets is so small that the marking probability has a limited impact on the unfairness reduction. For a high filling rate, on the opposite, the allowance is so important that here again the marking probability has a limited impact. It can even be observed that for a filling above 90 packets per second and a low maximum marking probability (i.e., below 0.25), the maximum rate of marked packets becomes close to the policer filling rate, resulting in heavy users being almost not policed.

From the investigation of the RED queue presented here, a trend can be observed: there is certainly an impact of the queue's parameters on the fairness improvement provided by ConEx, but it always goes with a significantly more important impact on the queuing delay. Thus, tuning of the RED parameters should focus much more on delay control, than on the influence of these parameters on ConEx. Additionally, whatever is the chosen tuning of the RED parameters, congestion policing can always provide a further reduction in the queueing delay, because of the traffic load reduction resulting from its action at the ingress of the network.

G. ConEx with increasing complexity

The deployability of ConEx is a major concern for both content providers and network operators. If a content provider can easily upgrade its servers, it does not control the traffic queuing mechanisms implemented on routers, nor the IP stack of receiving devices. On its part the network operator can modify the queuing strategy in its network but it does not control the IP stack on the senders and receivers. In that context the possibility of minimal modifications is a key factor for an introduction phase. ConEx allows incremental deployment by requiring only a few modifications to be operational. It can afterwards be upgraded, step by step, increasing the implementation complexity to provide a more and more accurate feedback of congestion information.

TABLE III. ConEx with increasing complexity

Case	queue	sender	receiver
DTConEx	DropTail	No ECN	No ECN
REDConEx	RED	No ECN	No ECN
ECNConEx	RED	Accurate ECN	Classic ECN
FullConEx	RED	Accurate ECN	Accurate ECN

The minimum modifications needed for ConEx are the modifications to the sender, which will react to a loss detection by sending a Re-Echo-Loss signal. In this case, ECN support is needed neither on the sender nor on the receiver and the RED queue can be replaced by a simple DropTail queue, which will drop packets when it overflows. In the next paragraphs, this case is referred to it as the $DTConEx$ case. The next step of modifications is when a RED queue is used on the router to improve reactivity to congestion appearance. ECN is not used and ConEx will react only to dropped packets by the RED queue. This is referred to it as the $REDConEx$ case. Another step of modifications is when ECN is activated on both the sender and the receiver, but the receiver does not provide an accurate account of the congestion signals it receives from the

network (cf. Section III-E), so only one congestion notification can be sent to the sender per RTT. This case is referred to it as the $ECNConEx$ case. The ultimate step of modifications is when ECN is used by both the sender and the receiver along with the modifications to the receiver to allow accurate ECN feedback (cf. Section III-E). This is referred to it as the $FullConEx$ case. The four cases are summurised in Table III.

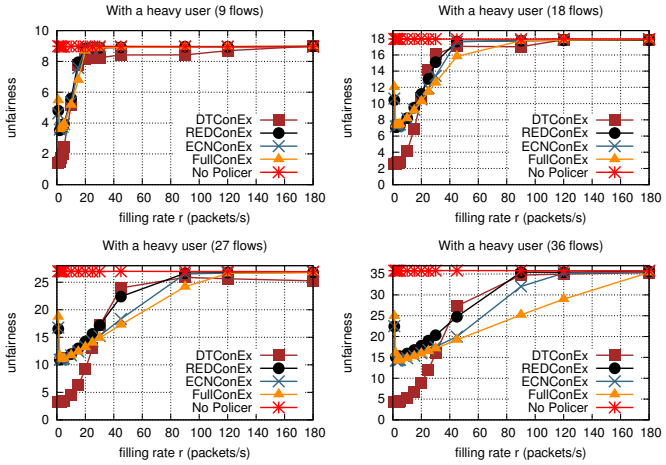


Figure 18. Unfairness with DTConEx, REDConEx, ECNConEx and FullConEx

Figure 18 depicts the average unfairness versus the filling rate in four scenarios where we vary the number of flows per heavy user (9, 18, 27, and 36), while the light user remains with a single flow. In each scenario, the red curve represents the unfairness without policing, while the four other curves represent the four cases explained above. As the number of flows of a heavy user increases, its contribution to congestion also increases. The user has to send more Re-Echo packets, he consumes more tokens and is more severely policed. As a consequence the range of filling rates allowing fairness improvement is widened.

In all scenarios, we see that $FullConEx$ and $ECNConEx$ have a similar behavior and decrease the unfairness more than $REDConEx$. The reason is that the two former cases provide the congestion information via both ECN and losses, which makes the policer more accurate than with $REDConEx$, which only provides the information on lost packets. In the same way, $FullConEx$ is slightly more effective than $ECNConEx$ in decreasing the unfairness, because it provides a more accurate congestion signal, particularly when the level of congestion increase (27 and 36 flows per user), allowing the congestion policer to more accurately restrain the heavy users.

The $DTConEx$ case provides even less congestion information than the other cases (i.e., only when the queue overflows), but manages to decrease more the unfairness in all scenarios in a range of filling rates around the optimum. $DTConEx$ is effective because it does not force the light users to reduce their throughput as early as for the other cases. Indeed, for $REDConEx$, $ECNConEx$ and $FullConEx$, the queue starts dropping or marking packets when its mean length exceeds a minimum threshold, forcing the heavy users, and in a smaller proportion the light user, to reduce their throughput. On opposite the DropTail queue only drops packets when the entire queue is filled, which gives the opportunity for the light users to increase their throughput when heavy users are restrained by the policer.

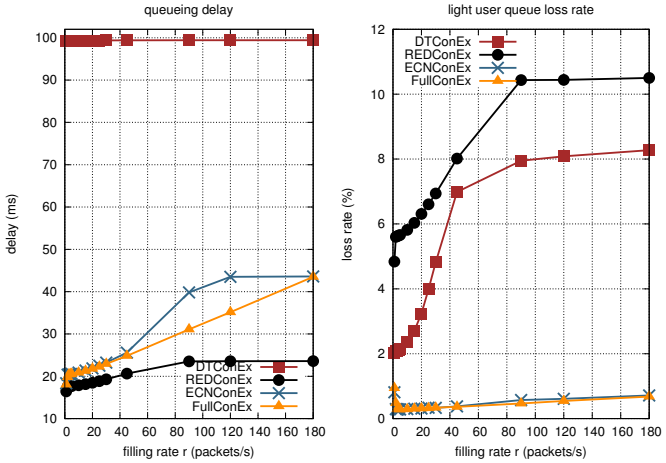


Figure 19. mean queueing delay and queue loss rate of a light user

Figure 19 represents the mean queueing delay and the loss rate that a light user encounters as a function of the filling rate (scenario with 36 flows per heavy user). Unlike RED, a DropTail queue does not allow reducing the queueing delay observed by the users, as we can see for the $DTConEx$ case. The DropTail queue is entirely filled when it starts dropping packets, so the users experience the highest delay equal to 100ms. For $REDConEx$, $ECNConEx$ and $FullConEx$, the queueing delay is reduced by the action of the RED queue. $REDConEx$ reduces the queueing delay more than $ECNConEx$ and $FullConEx$, because the RED queue drops packets while the two others only mark packets, leaving them in the queue. The congestion policer also contributes to the reduction of the queueing delay by limiting the amount of traffic entering the network. This effect is more visible as the filling rate decreases.

By reducing the traffic pressure on the bottleneck, the congestion policer also reduces the loss rate encountered by light users, especially in $DTConEx$ and $REDConEx$, which are based only on losses in order to notify congestion. In both cases, the light user's loss rate drastically decreases as the filling rate decreases. For all filling rates, $REDConEx$ results in a higher loss rate than $DTConEx$ because the RED queue begins dropping packets earlier than the DropTail queue. Finally, $ECNConEx$ and $FullConEx$, in which packets are ECN-marked rather than dropped, results in a similar and significantly lower loss rate for light users than the two other cases.

TABLE IV. Performance summary

Case	Fairness	Loss rate	Delay	Deployability
DTConEx	****	**	*	****
REDConEx	*	*	****	***
ECNConEx	**	****	**	**
FullConEx	***	****	***	*

Table IV summarises the advantages and drawbacks of each implementation variant in terms of fairness improvement, loss rate, queueing delay and deployability.

V. SHORT-LIVED FLOWS

Short-lived flows represent a great number of flows that cross the Internet (e.g., Domain Name System (DNS), Web objects). These flows are just a few packets long, they finish during the slow-start phase (in few RTTs) before reaching their fair-share rate [19]. This section aims to see how ConEx, which is a closed-loop mechanism requiring a number of RTTs to gather congestion information, behaves with short-lived flows and if it does bring an improvement to the completion time of these flows.

For performance evaluation, we use the same topology as in Section IV but modify the traffic sources from saturated long-lived flows to short-lived flows lasting only 10 packets. We use the aggregated traffic model described in [20], which uses a gamma distribution for the flow inter-arrival time, with a newly generated flow every 6ms on average. The 10 heavy users will generate 80% of the flows while the light users will generate the remaining 20%. In order to experience congestion in the network, a Not-ConEx cross traffic of 90Mbps over the 100Mbps bottleneck is generated. A strict policer is used as described in Section III-F. We monitor the flow completion time as a performance metric.

Each simulation lasts 600s, 30 simulations are performed to obtain a single point with a 95% confidence interval that is depicted on the graphs.

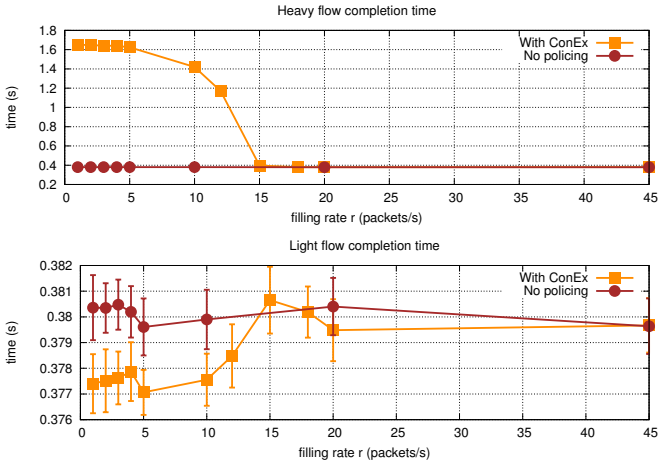


Figure 20. Completion time of a light and a heavy user's flow

Figure 20 represents the average completion time of a heavy user's flow and a light user's flow with and without the use of ConEx (FullConEx implementation in this simulation). The flows have an initial window of 2 segments and can be completed in 3 RTTs (300ms), which does not allow them to provide much congestion information for ConEx. Nevertheless, a heavy user can be policed when the filling rate is low enough ($r < 15 packets$), increasing greatly the completion time of his flows. The completion time of a 10-segment flow ranged from 380ms without policing up to 1.64s when policed. This is supposed to free the bottleneck for the light users' flows. Indeed, we see that when the heavy user is delayed, the light users benefit from a reduced completion time, but the decrease is only a few milliseconds, which is hardly a significant improvement.

Neither ConEx benefits from the use of short flows nor short flows benefit from ConEx. Short flows are not suited to retrieve congestion information for ConEx as they finish in few RTTs. In addition, these flows also finish before they can react to policing. When short flows lose packets, they can see their completion time increases dramatically from a few milliseconds to several seconds because they might need to wait for an RTO to perform retransmissions and complete. As expected, ConEx behaves poorly in presence of short flows, and it should be even less interesting if, as [19] suggests, the initial window is increased to 10 segments, which represents a less favorable scenario than the simulated one. However, the poor behaviour of ConEx observed with short flows does not lessen the interest of the mechanism considering that long flows are the main source of congestion. If a per user congestion policer is used, it should be more profitable to focus on long flows, which can retrieve congestion information and can efficiently react to policing.

VI. VIDEO STREAMING TRAFFIC: YOUTUBE USE CASE

We have observed over the last years an impressive growth of the video streaming traffic in both Orange's fixed and mobile networks (36% for FTTH, 26% for Asymmetric Digital Subscriber Line (ADSL) and 39% for mobile downstream [2]). This led us to analyse how ConEx can alleviate the pressure caused by video streaming traffic and we chose as a use case the very popular YouTube plateform.

A. YouTube server model

Many papers analysed the YouTube traffic generation. Among them, [21] [22] propose an algorithm to reproduce the behaviour of a YouTube server, which we implemented in NS2.

A server sends a video in two phases: the first phase is called the Initial Burst where 40s of video data is sent at maximum rate to provide sufficient buffering to the player. The second phase is called the Throttling phase, where the server sends the rest of the video data in chunks with a $sending\ rate = 1.25 \times encoding\ rate$ of the video. The chunk size is 64KB and the chunks are sent over a TCP socket with a 2MB sending buffer.

B. YouTube player model

We used the most precise monitoring approach proposed by [23] to implement a YouTube player in NS2. It is based on the status of the video buffer on the client player. The player starts playing the video when the buffered length exceeds a first threshold $\theta_0 = 2.2s$. If the buffer is depleted and the buffered length goes below a second threshold $\theta_1 = 0.4s$, the video stalls until the buffered length exceeds θ_0, then the video can start anew. We retrieve from the video player the number of stalling events N and their average length L to compute the QoE following a model suggested by [24] with the following equation:

$$QoE(L, N) = 3.50 \exp^{-(0.15L+0.19).N} + 1.50 \qquad (2)$$

C. YouTube results

The same topology as in Section IV is used to perform the simulations with 10 heavy users and 50 light users. The simulated scenario is the following: in the first 100s of the simulation, the heavy users have 20 FTP flows downloading at the maximum rate they can reach. No light user is present yet, the 10 heavy users can equally share the bottleneck. During

the next 100s, the light users begin requesting, randomly and uniformly over the 100s, a video from the servers. This video has a 300s duration and a bitrate of 1128kbps, which corresponds to the recommended bitrate for uploading 360p videos to YouTube (1000kbps for the video bitrate and 128kbps for the stereo audio bitrate [25]). The heavy users, which are responsible for 80% of the traffic, now have to share the network with the newcomers. At $t = 500s$, all light users should have finished watching their 300s video if no stalling events hampered the viewing, and the heavy users should be able to continue using the bottleneck until the end of the simulation 100s later. The mean QoE of the light users is computed at the end of each simulation.

A simple DropTail queue is used at the bottleneck. The policer is a strict policer as described in Section III-F and all users use cubic as a congestion control algorithm.

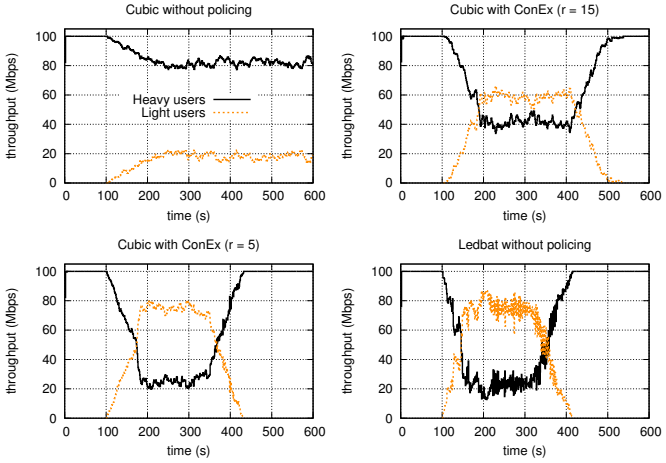


Figure 21. Throughput of heavy and light users versus time

Figure 21 shows the throughput of the heavy and the light users versus time. The three time periods of the simulated scenario are shown: before the arrival of the light users (0s-100s), during the light users' presence (100s-500s), and after the presumed departure of the light users if they watched the videos smoothly (500s-600s).

Figure 22 represents the computed QoE, the number of stalling events and the duration of a single stalling event for a light user in the following three cases: using cubic as a congestion control algorithm for heavy users without policing, using cubic for heavy users with ConEx policing and using LEDBAT as a congestion control algorithm for heavy users without policing.

1) Cubic without policing: When no policer is used, TCP with cubic will share the bottleneck equally between flows. The heavy users get 80% of the bottleneck and the light users will not be able to watch the video before the end of the second period. The light users will still be active during the third period, reducing the throughput of the heavy users when compared to the first period. The light users see their video stall many times and for a long duration, close to 10s, as shown in Figure 22, resulting in a $QoE = 1.5$, which is the lowest obtainable value with equation (2). It can be anticipated that in real life the users with such a low QoE would have stopped watching the video when the first stalling events occurred.

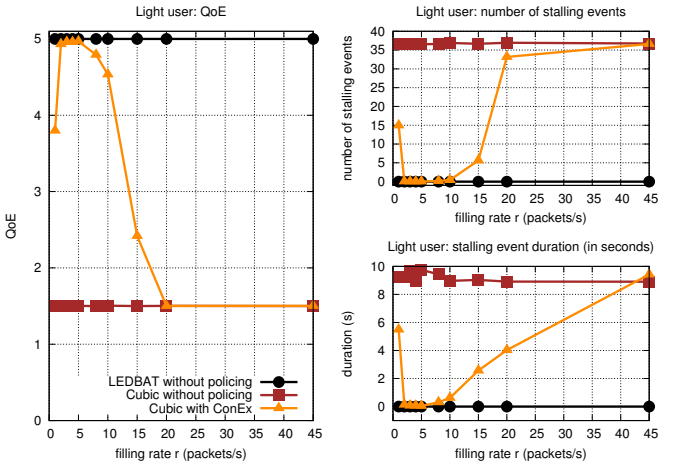


Figure 22. QoE of light users, the number of stalling events and the duration of a single stalling event

2) Cubic with ConEx: ConEx is activated in order to restrain the heavy users and improve the QoE of video users. Figure 22 shows that ConEx perfectly achieves this objective: as the filling rate decreases, the light users' QoE significantly increases due to a drastic reduction of the number of stalling events. For a filling rate below 15 packets/s video users benefit from a very good values ($QoE > 4$). The gain in QoE for the light users results from the heavy users reducing their throughput, in response to congestion policing, during the second period as represented in Figure 21. As represented for a filling rate of 15 and 5 packets/s, tuning the filling rate allows to finely control the bandwidth reparation between the heavy and light users. The light users are then able to finish viewing their video before the end of the second period. As the light users leave the bottleneck, the heavy users can increase their throughput during the third period.

3) LEDBAT without policing: The heavy users could avoid policing by being less aggressive towards video traffic. They could either postpone their activities until a less congested period, or they could use a less aggressive congestion control algorithm, which yields the network resources when encountering congestion. LEDBAT [26] is such a congestion control algorithm. It is designed to use the available bandwidth in a bottleneck and yields rapidly in presence of standard TCP. When LEDBAT is used (implementated in NS2 by [27]) instead of cubic for the heavy users, results in Figure 21 show that, without requiring any policing, the heavy users rapidly decrease their throughput when the video users become active. The light users are then able to watch their video with a very good QoE (Figure 22), similar to the results obtained by using cubic and ConEx policing ($r = 5$). In the same way when the light users' videos finish, LEDBAT is able to use the freed resources in the bottleneck.

The behavior observed with LEDBAT may raise some questions: what is the usefulness of ConEx? Why not jumping directly to LEDBAT? In fact, the use of a congestion control mechanism like LEDBAT for bulk data transfers could be seen as a target, but to favor its adoption it is necessary for the network to find a way to acknowledge the users who adopt a TCP-friendly behavior. As suggested in the ConEx charter [3], ConEx can be deployed in order to incentivize the heavy users to migrate a LEDBAT-like congestion control mecha-

nism. The use of LEDBAT prevented the heavy users from consuming tokens for applications like file transfer, preserving their congestion allowance for more critical applications, while allowing the light users to have a good quality of experience.

In this study, we have decided to use a progressive download mode of video delivery as it allows accurately quantifying the benefits of ConEx and LEDBAT. If the video delivery relies on HTTP-adaptive streaming, similar behavior can be expected: the light users would decrease the resolution of their video when encountering congestion, but after the heavy users have reduced their throughput using LEDBAT or in response to ConEx policing, the light users could increase the resolution of their video and benefit from a higher video quality.

VII. Summary and conclusion

ConEx is a new mechanism that allows a user to inform the network of the amount of congestion encountered. This allows the network operator to implement congestion-based policies proportionally to the amount of congestion a user has contributed to.

In Section IV, we have seen that ConEx allows us to differentiate between a light and a heavy user to improve the fairness between users. Many parameters (congestion policer parameters, RTT, TCP congestion control algorithm, RED queue) are involved in ConEx mechanism, and can influence its ability to improve fairness between users.

ConEx might be very sensitive with the harshness of the policing because of its interaction with the sent Re-Echo-Loss packets, as explained in Section IV-B. Tuning the policer harshness allows to precisely control the unfairness reduction for all level of congestion allowance (i.e., for all level of filling rate). Setting the token bucket depth allows to control the policer aggressivity without significantly impacting the level of performance. As for all mechanism operating in a close-loop mode between the sender and the receiver, the quickness of the congestion information retrieval, through a short RTT, also have a great impact on the behavior of ConEx and on its ability to improve the fairness between users. The results obtained with cubic and compound shows that ConEx is relatively insensitive to the TCP congestion control algorithms implemented by the end devices, meaning that users should be almost equally treated, whatever their congestion control is, resulting in a relative independence of ConEx regarding the type of end devices.

Finally, the investigation of the queuing strategy implemented by routers (i.e., RED queue) shows that the queue parameters have a limited impact on ConEx behavior, particularly on the fairness improvement through congestion policing, compared to their greater impact on other characteristics of the traffic like delay. The RED queue setting can then be optimized to control the queuing delay, with a limited impact on ConEx performance. We have also shown that ConEx can still improve fairness even with minimal modifications (the ability to react to lost packets by sending a Re-Echo-Loss signal) and the use of simple DropTail queues. So, an efficient initial deployment is possible, as suggests [4], before considering the deployment of a more accurate ConEx relying on ECN, which requires modifications to both the senders and the receivers, and the use of RED queues. The advantages and drawbacks of each step of modifications are summarized in Table IV.

In Section V, we illustrate the poor behavior of ConEx in presence of short-lived flows. We argued that neither ConEx benefits from the use of short flows nor short flows benefit from ConEx. Indeed, the short flows do not provide enough congestion information to ConEx, and policing them is not beneficial for their completion time. It is then more profitable to focus on policing long and responsive flows.

In Section VI, we have seen how video streaming like YouTube can benefit from ConEx. The results show the improvement that can be obtained by using ConEx alone, but also the benefits that can be expected from the combined used of ConEx and LEDBAT. ConEx can be used to restrain the heavy users who do not yield voluntarily under congestion, while leaving unpoliced those who do through a congestion control mechanism like LEDBAT. This should provide incentives for the heavy users to be more cooperative during congestion periods. The use of LEDBAT can protect the heavy users from being policed through ConEx while allowing the light users to have a great QoE.

To conclude, ConEx could be considered as a credible approach to improve user's fairness in case of high network loads, while being transparent otherwise. If ConEx operation is influenced by its environment (e.g., parameters setting, network topology, traffic demand, etc.), in all tested configurations its behavior remains robust, suggesting that reasonable margins exist for a network operator to deploy and provision ConEx in a real network. Considering its poor behavior when applied on short flows and the fact that long flows are the main cause of congestion, ConEx should be focused on long flows. In order to minimize the introduction cost, ConEx should be deployed first in a simple implementation mode, i.e., using only packet losses to estimate congestion. In that mode only the traffic sources (i.e., servers) have to be modified to be able to generate the ConEx signal. In a second phase, if a better level of performance is required, in particular regarding delay, upgrading ConEx could be envisaged, preferably for a full ConEx mode.

VIII. Future Work

Implementing a per user congestion policer requires the determination of the policer's parameters, the filling rate (the allowed Congestion-Rate) and the bucket depth (the allowed Congestion-Burst). Different kinds of flows with different behaviours need to be policed with the same allowance rate, which makes the determination of these parameters challenging. Further studies are required on this subject.

The congestion policing function is the key to improve fairness between users and to enforce some users to yield if they do not voluntarily. Designing a policer algorithm that achieves the goals we set is a crucial point in the deployment and is one of the main objectives of our future work.

Finally, the auditor can be necessary if there is a risk that the sources do not report the right Congestion-Volume they encounter. If auditing is relatively easy when ECN is used, ConEx on loss is more challenging as it requires detecting lost packets in the auditor. To address these issues, we can harness the substantial work concerning the auditor that has been done under the Trilogy project [8].

References

[1] A. Sanhaji, P. Niger, P. Cadro, and A.-L. Beylot, "DropTail Based ConEx Applied to Video Streaming," ICNS 2015, 2015, pp. 3–10.

[2] M. Feknous, T. Houdoin, B. Le Guyader, J. De Biasio, A. Gravey, and J. Torrijos Gijon, "Internet traffic analysis: A case study from two major european operators," in Computers and Communication (ISCC), 2014 IEEE Symposium on, June 2014, pp. 1–7.

[3] ConEx Working Group Charter. [Online]. Available: http://datatracker.ietf.org/wg/conex/charter/ [retrieved: November, 2015]

[4] B. Briscoe and al., "Congestion exposure (ConEx) concepts and use cases," December 2012. [Online]. Available: https://www.rfc-editor.org/rfc/rfc6789.txt [retrieved: November, 2015]

[5] S. Krishnan and M. Kuehlewind, "IPv6 destination option for congestion exposure," October 2015. [Online]. Available: https://www.ietf.org/id/draft-ietf-conex-destopt-10.txt [retrieved: November, 2015]

[6] D. Kutscher, F. Mir, S. Krishnan, Y. Zhang, and C. Bernardos, "Mobile communication congestion exposure scenario," October 2015. [Online]. Available: https://www.ietf.org/id/draft-ietf-conex-mobile-06.txt [retrieved: November, 2015]

[7] M. Kuehlewind and R. Scheffenegger, "TCP modifications for congestion exposure," October 2015. [Online]. Available: https://www.ietf.org/id/draft-ietf-conex-tcp-modifications-10.txt [retrieved: November, 2015]

[8] B. Briscoe and al., "Final report on resource control, including implementation report on prototype and evaluation of algorithms," December 2010. [Online]. Available: http://www.trilogy-project.org/fileadmin/publications/Deliverables/D13_-_Final_report_on_resource_control__including_implementation_report_on_prototype_and_evaluation_of_algorithms.pdf [retrieved: November, 2015]

[9] M. Kühlewind and M. Scharf, "Implementation and performance evaluation of the Re-ECN protocol," in Incentives, Overlays, and Economic Traffic Control, ser. Lecture Notes in Computer Science, B. Stiller, T. Hoßfeld, and G. Stamoulis, Eds. Springer Berlin Heidelberg, 2010, vol. 6236, pp. 39–50. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-15485-0_5

[10] F. Mir, D. Kutscher, and M. Brunner, "Congestion exposure in mobility scenarios," in Next Generation Internet (NGI), 2011 7th EURO-NGI Conference on, June 2011, pp. 1–8.

[11] Y. Zhang, I. Johansson, H. Green, and M. Tatipamula, "Metering re-ecn: Performance evaluation and its applicability in cellular networks," in Teletraffic Congress (ITC), 2011 23rd International, Sept. 2011, pp. 246–253.

[12] K. Ramakrishnan, S. Floyd, and D. Black, "The addition of explicit congestion notification (ECN) to IP," September 2001. [Online]. Available: http://www.rfc-editor.org/rfc/rfc3168.txt [retrieved: November, 2015]

[13] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance," IEEE/ACM Trans. Netw., vol. 1, no. 4, Aug. 1993, pp. 397–413. [Online]. Available: http://dx.doi.org/10.1109/90.251892

[14] M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan, "Data center TCP (DCTCP)," SIGCOMM Comput. Commun. Rev., vol. 41, no. 4, Aug. 2010, pp. –. [Online]. Available: http://dl.acm.org/citation.cfm?id=2043164.1851192

[15] M. Kuhlewind and R. Scheffenegger, "Design and evaluation of schemes for more accurate ECN feedback," in Communications (ICC), 2012 IEEE International Conference on, June 2012, pp. 6937–6941.

[16] B. Briscoe, R. Scheffenegger, and M. Kuehlewind, "More accurate ECN feedback in TCP," October 2015. [Online]. Available: https://www.ietf.org/id/draft-kuehlewind-tcpm-accurate-ecn-05.txt [retrieved: November, 2015]

[17] The Network Simulator - NS-2. [Online]. Available: http://www.isi.edu/nsnam/ns/ [retrieved: November, 2015]

[18] A. Martin and M. Menth, "ConEx-based congestion policing – first performance results," March 2012. [Online]. Available: http://www.ietf.org/proceedings/83/slides/slides-83-conex-5.pdf [retrieved: November, 2015]

[19] N. Dukkipati, T. Refice, Y. Cheng, J. Chu, T. Herbert, A. Agarwal, A. Jain, and N. Sutin, "An argument for increasing TCP's initial congestion window," SIGCOMM Comput. Commun. Rev., vol. 40, no. 3, June 2010, pp. 26–33. [Online]. Available: http://doi.acm.org/10.1145/1823844.1823848

[20] S. Gebert, R. Pries, D. Schlosser, and K. Heck, "Internet access traffic measurement and analysis," in Proceedings of the 4th International Conference on Traffic Monitoring and Analysis, ser. TMA'12. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 29–42. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-28534-9_3

[21] P. Ameigeiras, J. J. Ramos-Munoz, J. Navarro-Ortiz, and J. Lopez-Soler, "Analysis and modelling of youtube traffic," Transactions on Emerging Telecommunications Technologies, vol. 23, no. 4, 2012, pp. 360–377. [Online]. Available: http://dx.doi.org/10.1002/ett.2546

[22] J. Ramos-munoz, J. Prados-Garzon, P. Ameigeiras, J. Navarro-Ortiz, and J. Lopez-soler, "Characteristics of mobile youtube traffic," Wireless Communications, IEEE, vol. 21, no. 1, February 2014, pp. 18–25.

[23] R. Schatz, T. Hossfeld, and P. Casas, "Passive youtube qoe monitoring for isps," in Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2012 Sixth International Conference on, July 2012, pp. 358–364.

[24] T. Hoßfeld, R. Schatz, E. Biersack, and L. Plissonneau, "Internet video delivery in Youtube: From traffic measurements to quality of experience," in Data Traffic Monitoring and Analysis, ser. Lecture Notes in Computer Science, E. Biersack, C. Callegari, and M. Matijasevic, Eds. Springer Berlin Heidelberg, 2013, vol. 7754, pp. 264–301. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-36784-7_11

[25] Google, "Advanced encoding settings." [Online]. Available: https://support.google.com/youtube/answer/1722171 [retrieved: November, 2015]

[26] S. Shalunov and al., "Low extra delay background transport (LEDBAT)," December 2012. [Online]. Available: https://www.rfc-editor.org/rfc/rfc6817.txt [retrieved: November, 2015]

[27] D. Rossi, C. Testa, S. Valenti, and L. Muscariello, "LEDBAT: The new bittorrent congestion control protocol," in Computer Communications and Networks (ICCCN), 2010 Proceedings of 19th International Conference on, Aug. 2010, pp. 1–6.

Design of a Flexible Over the Top Content Streaming System with Dual Adaptation

Eugen Borcoci, Radu Iorga, Cristian Cernat, Marius Vochin, Serban Obreja
University POLITEHNICA of Bucharest
Bucharest, Romania
emails: eugen.borcoci@elcom.pub.ro
radu.iorga@elcom.pub.ro
cristian.cernat@elcom.pub.ro
marius.vochin@elcom.pub.ro
serban@radio.pub.ro

Jordi Mongay Batalla
National Institute of Telecommunications
Warsaw, Poland
email:jordim@interfree.it
Daniel Negru
LaBRI Lab, University of Bordeaux
Bordeaux, France
email:daniel.negru@labri.fr

***Abstract* — Real-time content delivery services have recently achieved high popularity in Internet, both for professional communities and also for entertainment. In contrast with some complex architectures like Content Delivery Networks, or Content Oriented Networks, this paper considers a light architecture, working on top of the current networking IP multi-domain infrastructures. It enhances the real-time (video) content delivery, by exploiting the cases where the content object exists on several servers. The system performs an initial content server selection (based on multi-criteria) considering the servers' load status and network conditions. Then it performs in-session media dynamic flow rate adaptation. Additionally, if necessary, a server handover is triggered. All above functionalities are assembled in a unified solution. This paper is a preliminary work, to identify the main requirements and then to develop the design for a family of implementations. Different design decisions variants are analyzed and explored, proving the solution flexibility. Currently the proposed system is under implementation in the framework of a European project.**

Keywords* — *content delivery; multi-criteria decision algorithms; server and path selection; media adaptation, monitoring, Future Internet; content-aware networking.

I. INTRODUCTION

The content-related real-time services are more and more present in the current and Future Internet, leading to recent significant developments [1][2]. Dedicated infrastructure like Content Delivery Networks (CDNs) improve the services quality [3], by distributing the content replica to caching servers, located close to groups of users; they are largely used in the real world. Content/Information Oriented/Centric Networking (CON/ICN/CCN) approaches [4][5], decouple names from location and introduce novel paradigms such as content-based routing, in-network (in routers) caching, etc.

However, all the above solutions involve highly complex architectures, high CAPEX and significant modifications in Service/Content Providers and Network Providers/Operators systems.

As an alternative, Service Providers (SP) might deliver services in *over-the-top* (OTT) style, over the current best effort Internet as a significantly cheaper solution. An OTT Service Provider (SP) could act as a separate entity from the traditional Internet Service Provider (ISP). Also, combined solutions could exist, with OTT-like SPs using the CDN Providers infrastructure to improve the quality of delivery. When transport problems appear in the network, the OTT approach to preserve or even improve the quality, are frequently based on adaptive solutions for media streams or servers. The overall goal is to maintain a good or at least an acceptable degree of the quality of experience (QoE) perceived at receiver side.

A light architecture (OTT-like), for content streaming systems is proposed by the European *DISEDAN* Chist-Era project [1][6], (*service and user-based **DI**stributed **SE**lection of content streaming source and **D**ual **A**daptatio**N,*** 2014-2015). The business actors involved are: *Service Provider (*an entity/actor which deliver content services and might own or not a transportation network); *End Users (EU)* which consumes the content; a *Content Provider (CP)* could exist, owning *Content Servers (CS)*. DISEDAN does not deal with CP/SP contractual relationships; we may assume that the content servers are owned by the SP.

This paper is an extension of a previous one [1] presented at CONNET 2015 Conference, dedicated to develop the design of a flexible system in the framework of DISEDAN project.

An initial assumption is that a given content object is present on at least one, or several content servers, geographically distributed over one or several network domains. In such conditions, a *novel concept* is introduced by DISEDAN, based on:

(1) *two-step server selection mechanism* (initial at SP and then at EU sides) using algorithms that consider context- and content-awareness and

(2) *dual in-session adaptation mechanism*, consisting in *media flow adaptation* (based on Dynamic Adapative Streaming over HTTP – DASH recent technology [7][8][9]) and/or content source adaptation (by *streaming server switching*) if quality degradation is observed at the EU Terminal (EUT) during the media session.

An effective solution is constructed in DISEDAN for the multi-criteria selection (hard problem) of the best *content source (server)*, while considering user context, server availability and requested content. The DISEDAN OTT-like architecture is attractive since it avoids the complexity of CON/ICN or CDN.

This work is mainly dedicated to identify the requirements, specify the architecture and then analyze several design decisions variants. Details on server/path selection, optimization algorithms and adaptation process combined with server switching are treated in other works [15][16][18].

Also it should be mentioned that this paper does not have as objective to detail the internal procedures of the functional blocks. These elements (i.e., algorithms, active or passive monitoring procedures, QoS/QoE evaluation, DASH details, design details and low level description of interfaces, SP or EU policies, etc.) are (or will be) the targets of other works, during the project development.

The DISEDAN system can be flexibly implemented in several variants, depending on the complexity/constraints envisaged and the EUs and SPs requirements. We explore different design decisions and trade-offs, versus the cost and implementation complexity. This work is preliminary; currently, the system is under its implementation.

Section II is a short overview of related work. Section III outlines the overall architecture, based on different sets of requirements. Section IV analyzes various design decisions and implementation-related implications. *Section V is a new contribution of the extended paper; it considers the previous design decisions and develops the functional architectures of the three main entities: Service Provider, End User Terminal and Content Server.* Section VI contains conclusions and future work outline.

II. Related Work

Adaptation techniques enhance the quality of streaming media at the consumer side when the transfer conditions deteriorate. They also support efficient network resource utilization, device-independent universal media access and optimized Quality of Experience (QoE). Many Service Providers apply adaptation, to solve the network variations [7]. Adaptation may act on media flow [7][8][9], and/or on Content server. The latter means in-session new server selection and switching (handover), depending on the consumer device capabilities, consumer location and/or network state [10][11].

Recent solutions for media adaptation use the HTTP protocol, while minimizing server processing power and being video codec agnostic [12]. Relevant examples are: Adobe Dynamic Streaming, Apple's HTTP Adaptive Live Streaming and Microsoft's IIS Smooth Streaming and open HTTP-based protocols like Dynamic Adaptive Streaming over HTTP (DASH) [9]. The DASH continuously selects, on-the-fly, the highest possible video representation quality that ensures smooth play-out in the current downloading conditions. The DISEDAN novelty [6] consists in "dual adaptation" by combining in a single solution the initial server selection (result of the initial cooperation between SP and EU) and in-session dual adaptation.

The initial server selection is based on optimization algorithms like *Multi-Criteria Decision Algorithms (MCDA)* [13][14], modified to be applied to DISEDAN context [15]16], or *Evolutionary Multi-objective Optimization algorithm (EMO)* [17]. The decision variables considered for selections are related to servers' load, network paths characteristics, SP and EUT policy related parameters.

In [15][16] several scenarios are proposed, analyzed and evaluated. The initial content selection problem is a multi-criteria one, given the different degree of availability of parameters of interest at SP, CS and respectively EUT levels. In particular, the availability of different static and/or dynamic input parameters for optimization algorithms is considered. Therefore, several designs are possible, different in terms of performance and complexity. It is the objective of this paper to analyze these variants, seen as design/ implementation decisions.

The main advantages of the DISEDAN approach versus others solutions are: simple architecture, working in OTT style and avoiding complex (as needed in ICN, CDN) management and control; multi-domain capabilities; embedding in a single solution the initial server selection, in-session dynamic media flow adaptation and/or server switching; backward compatibility versus current content streaming systems; low cost for implementation.

The principal limitation of the DISEDAN solution consists in its OTT-style of working; no strong QoS guarantees are offered to the end users.

III. DISEDAN System Architecture and Design Guidelines

While considering the above general concepts, assumptions and requirements should be identified, to provide inputs for the system design.

A. General framework and assumptions

The main business entities / actors have been mentioned above: SP, EU, CS. The connectivity between CSs and EU Terminals (EUT) are assured by traditional Internet Services Providers (ISP) / Network Providers (NP) - operators. Due to its OTT-style, DISEDAN does not consider, in its management architecture the connectivity – related relationships between SP and ISP/NPs. Note that some Service Level Agreements (SLAs) might exist, related to connectivity services, but they are not directly visible at our system level. The DISEDAN solution is also applicable to other business models, e.g., involving CPs, CDN providers, etc. The relationships between SP and such entities could exist, but their realization is out of scope of this study. The system works on top of the current TCP/IP mono and/or multi-domain network environment.

The EUTs might not have explicit knowledge about the managed/non-managed characteristics of the connectivity services. Network level resources reservation, or in-network connectivity services differentiation might exist, but they are not mandatory supposed. This approach shows the system flexibility: it can work in OTT low cost style, with no direct control of the connectivity services (from QoS point of view) or, in a more complex deployment, over a network having managed connectivity services. Therefore, in principle, the

SP envisaged in DISEDAN cannot offer strong QoS guarantees to EUs. Consequently, DISEDAN does not manage (but it does not exclude) possible EUs/SPs SLA contracts/relationships. However, it is assumed that a Media Description Server exists, managed by SP, to which EUT will directly interact.

The media streaming operations are independent from networking technology. The client-side streaming system, acts as a standalone application, (no mandatory modifications for SP are required); however, DISEDAN assumes that SP should provide some basic information to EUT, in order to help the initial server selection by the EUT. Then, the in-session decisions about dual adaptation are taken mainly locally at EUT, based on the real time delivery evaluation of the quality seen by the receiver. Based on the above approach, a complex EUT-SP signaling is avoided.

In a general case, several CSs exist (containing replicas of media objects), known by SP (geo-location, availability, access conditions for users), among which the SP and/or EUTs can operate server selection and/or switching. No restriction is imposed either on the geo-localization of EUTs or of CSs. Note that the proposed system does not consider how to solve network failures, except attempts to perform media flow DASH adaptation or CS switching. The terminal devices are supposed to have all the required subsystems and peripherals for video/ audio display and device control.

Note that several assumptions and requirements are general ones – needed in a content delivery system and they are not specific to only DISEDAN system.

B. End User Requirements

These requirements are expressed as End User needs, and are derived from *user scenarios* - when selecting and consuming media content - related services. The EUT (basically) but also the rest of the system should be designed as to fulfill the requirements coming from EU.

- The system must admit the usual user profiles. EU should be able to identify itself and login into the system through a controlled environment.
- The EU must be able to select among several SPs and among content items, servers and classes of quality – in the limits offered by the selected SP.
- The DISEDAN system must allow to EU: initial (optionally automatic or manual) server selection; in-session dual adaptation will be automatically enforced, to maximize the Quality of Experience (QoE).
- The EU should receive information from SP (on servers and possibly on network paths) to help him in selection. The EU should also have the possibility to finally decide on server selection/switching or amount of adaptation actions initiated and/or performed.
- The EUT must be still able to work by using only minimal information on server and network (e.g., server capacity or download bandwidth from the server) delivered by the SP. The final content server selection decision is basically locally taken, while avoiding complex signaling between user and SP.
- The EU should have the possibility to be informed about of QoE level delivered by the system.
- The client SW installed on the EUT should have maximum independence from the operating system running on the terminal.

C. Service Provider Requirements

These requirements are expressed as SP business and technical needs. The DISEDAN system:

- Should allow SP to develop multimedia content-based services, e.g., live - streamed IPTV services, Video on Demand (VoD) and its derivatives (e.g., streamed VoD, downloaded / pushed content).
- Must allow SP to filter the control information delivered to the EUs, but should not impose major architectural modification in the common SP Management and Control (M&C) architecture.
- May allow SP to apply different policies in its server selection (e.g., to maximize CS utilization and/or improve QoE)).
- Should be able to use in a flexible way the SP static/dynamic (monitored) information on servers and (possibly) on network paths status and availability, in mono or multi-domain contexts.
- Must not restrict the networking technologies (QoS capable or not) used by SP.
- Must support the SP-EU cooperation for dual adaptation purposes.
- Should offer to the SP the minimal capabilities to manage the Content Severs (if no distinct Content Provider business entity exists).

D. General System Requirements

These requirements are derived from the previous requirements for End User and Service Provider. They are related to the overall DISEDAN system, which:

- Must work in the traditional TCP/IP mono and multi-domain, in OTT style, on top of arbitrary network technology; the EUTs or CSs can be placed everywhere.
- Should provide a simple management with minimal architectural modifications at SP side or at EUT side.
- Must optimize multi-criteria content source selection, and then dynamic dual adaptation, considering user context, servers' availability, network conditions and content distribution mode. It will apply: a. *two-step server selection* (at SP and then at EUT) based on context/content - aware algorithms; b. *dual adaptation*, (media adaptation and/or server switching).
- At EU side, a standalone client application must exist. No mandatory modifications at SP M&C side are required; however, SP M&C should provide information to EUT, to help it in initial server selection.
- Should provide flexible possibilities to assign/balance the decision power between SP/EU, regarding sever selection/switching and dynamic adaptation.

Other, more specific EUT, SP and CS system requirements have been derived from the general ones but they are not detailed here.

E. General Architecture

Fig. 1 shows the high level – described, general architecture. The Service Provider entity includes the following functional modules:

- *Media Description generator* – dynamically generates Media Description (MD) XML file, containing media segments information (video resolution, bit rates etc.), ranked list of recommended CSs (for a given EU request) and possibly - CSs current state information and even information on network state (if applicable).
- *CS Selection (step 1) algorithm* - exploits MCDA or EMO, to rank the CSs and media representations, aiming to optimize servers' load and to maximize the system utilization.
- *Monitoring module* – collects information from Content Servers and estimates their current states.

The End User Terminal entity includes the modules:

- *DASH* – parses the MD file received from SP and handles the download of media segments from Content Servers.
- *Content Source Selection and Adaptation engine* – implements the dual adaptation mechanism.
- *Selection (*step 2) *algorithm*. - exploit MCDA, EMO, or other algorithms to select the best CS from the list recommended by SP.
- *Monitoring module* – monitors the local network conditions and – possibly - the server conditions.
- *Media Player* – playbacks the media segments.

The Content Server entity includes the modules:

- *Streaming module* – sends media segments requested by End Users.
- *Monitoring probe* – monitors CS performance (CPU utilization, network interfaces utilization, etc.). In a complex implementation of the CS, the monitoring probe could be replaced by a more capable monitoring module, to supervise both the active sessions and some connectivity characteristics from this CS to different groups of users.

The following (macro) procedural steps are:
1. The EUT issues a media file request to SP.
2. The SP analyzes the status of the CSs (involved in the request parameters) and runs the selection algorithm (optionally the SP could make first, a current probing of the CSs). Some SP policies could be enforced in this phase.
3. The SP returns a candidate CS list to EUT.
4. The EUT performs the final CS selection (by considering additional local information) and starts asking media segments from the selected CS.
5. During media session, the EUT measures the quality and evaluates the context. It applies DASH adaptation or if necessary, CS switching is decided.

When the user requests a Multimedia content, the SP sends an *xml* file containing Media Description (MD). This file is updated (from the static *xml* file) for each user request by considering the user profile, the SP policies for this user's class and other information at the SP side (e.g., state of the servers and possibly network-related information). The list of candidate CSs and other information are written inside the *xml* file. Also caching server *url* addresses can be added. The list may be ordered, following some desired metrics.

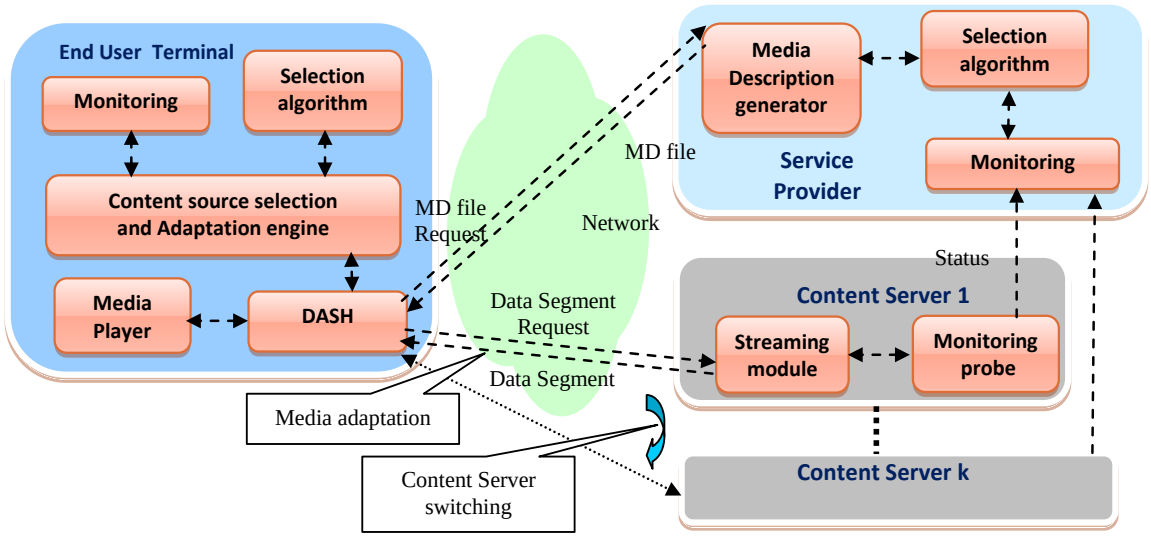


Figure 1. DISEDAN general architecture; DASH - Dynamic Adaptive Streaming over HTTP

When the user's application receives the MD file, it performs the final CS selection and possibly the network path choice (note that this last choice is not always possible in the current Internet with regular routing and forwarding). This decision can be based on user context, i.e., a MCDA can be run or, the EUT can simply select the first CS in the ordered list returned by the SP. Generally the CS selection achieves the final step of multi-objective optimization.

After final CS selection, the EUT starts to ask segments from the selected CS. During the receipt of consecutive chunks, the user's application can automatically change the rate of the content stream (DASH actions) or, if still problems exist, it can switch the CS.

The EUT performs measurements of the parameters of download process. If EUT detects deterioration of downloading rate, it can use SP information about alternate CSs and/or it can start probing CSs. After probing, the EUT decides on media or server adaptation.

IV. Design Decision Variants

The DISEDAN architecture is intended to be flexible. Several variants/versions of designs can be considered, i.e., basic ones or more complex, essentially depending on the roles of the business entities and their capabilities, interactions and also on SP and EU policies.

A. Monitoring procedures

The types and amount of static and dynamic monitoring data collected by various entities have a significant impact on the solutions. Consequently, the MCDA/EMO algorithms will have different sets of input parameters. Apart from static information available at SP, three types of monitoring contexts and possible policies can be identified.

An important component of the Control Plane is the Monitoring subsystem *(MON@DISEDAN*), whose components are distributed at SP, CS and EUT sides.

Apart from DASH defined metrics (in-session observed), the MON subsystem may collect information though its respective components, at each DISEDAN entity, as described below:

- *MON@EUT:* CS accessibility (probing); EUT local dynamic context; historical and prediction data on servers and paths utilization.
- *MON@SP:* CS status (collected from CS); active Users (i.e., those who are in-session) status; current load on some paths (here the network monitoring of the NP should cooperate); other dynamic, characteristics of some paths (e.g., loss, jitter); historical and prediction data on servers and paths utilization.
- *MON@CS:* CS status (load); CS environment data (network paths, dynamic characteristics of connectivity paths from CS to different groups of users - evaluated at overlay level; EUTs data, active user groups data.

The overall MON design is flexible, since it can combine different features of the above components.

Several types of monitoring activities can be performed.

- *Proactive monitoring:* executed in continuous mode; the monitoring information becomes input for the CS selection algorithm (Phase 1), when some new content requests arrive from a given EU to SP. At SP, this means supervision of different servers, maybe networks, and user communities, depending on its policies. SP/CS cooperation on this purpose is envisaged. Such data can be also used to construct a history and updated status of the environment envisaged by the SP. The CSs could be involved in proactive monitoring, provided they are capable to probe their connectivity characteristics towards different groups of users (indicated by the SP). At EU side, proactive monitoring might be performed, depending on capabilities of the EUT and its software. In some more complex scenarios the EU can construct history, dedicated to its usual content connections (if they are estimated to be repeated in the future). The terminal context can be evaluated by such measurements, including its access network status.

- *In-session monitoring:* monitoring is performed on a flow and data are collected in real time, to assess the level of QoS/QoE observed at EU side. These actions are basically performed by the EUT, in two ways:
  - collected by the DASH mechanisms, to serve internally as real time inputs to adaptation decision engine at EU,
  - collected by the MON@EUT, which can be consolidated with those produced by the DASH, thus offering a more complete view, not only about the reception of he media flow but also on general status and environment of the EUT.

In more complex DISEDAN variants, the SP and/or CS can be involved in such monitoring, at least in being aware of results (note that no SLA concerning mutual obligations of SP/EUs, related to QoE are established in DISEDAN system): for all active users or subsets; for all monitored data or summaries; full or summary monitored values.

- *Opportunity related monitoring:* measurements essentially performed by the EUT to test the opportunity of switching the CS that delivers the content to EU. An example of such category is the Probing of some CS candidates if a CS switching action is prepared.

B. Possible Roles of the Business Actors

The DISEDAN project outlines a set of optional Provider side modifications (w.r.t. useful information and metrics provided by SP to the client) that can further optimize server selection. The design can be backwards-compatible, ensuring that each modified client or SP can cooperate with the other

side, if the latter is using existing common content distribution solutions. Consequently, a range of solutions are proposed in this paper for SP, CS and EUT roles, i.e., several variants (named "use cases"), that are listed below.

Tables I, II and III illustrate different design choices, listed in increasing order of complexity and, consequently, in increasing order of performances and costs, for SP, CS and EUT. Note that, although the Monitoring subsystems could be included generally in the architectural Management Plane, the *Mon@SP*, *Mon@CS* or *Mon@EUT* are specified in the tables in a distinct way, in order to emphasize the dynamic character of the data collected. Depending on the specific requirements and constraints, different variants can be selected as design/implementation choices of the DISEDAN system.

Note also that the tables do not detail the monitoring capabilities embedded in the DASH adaptation subsystem.

TABLE I. SERVICE PROVIDER - DESIGN VERSIONS

	Information known (by SP) about:	**Obtained from**	**Type of information**	**Is Monitoring system involved?** (in collecting the Column 2 information)	**Remarks on SP role**
SP-V1	CS list and their locations	Mgmt@SP	Quasi-static	No	SP solves the user requests. SP is involved in initial server selection, or during media session (to help switching decision at EUT), based only on ordered list of servers and depending on their load. (*minimum complexity*)
	Content files (objects) mapping on different servers	Mgmt@SP	Quasi-Static or dynamic	No	
	CS status (current load)	CSs	Dynamic	Yes	
	User groups	Mgmt@SP	Quasi-static	No	
	Active (in-session) users (information is based on EUT request accounting only)	EUs	Dynamic	No	
SP-V2	**Idem as SP-V1, plus below items**				Idem as in SP-V1 but more qualified assistance in selection of the initial (server-path) pair. Problem: how can a given user invoke usage of a selected path if multiple paths are available? Usually the choice can address only the inter-domain paths.
	Potential user groups	Mgmt@SP	Quasi-static	No	
	Basic connectivity paths (from different CSs to different groups of users) static characteristics - obtained at overlay level	Mgmt@SP/ CSs	Quasi-static	No	
SP-V3	**Idem as SP-V2, plus below items**				Idem as in SP-V2 but more assistance in selection of the initial (server-path) pair, given the paths current load information. This is a powerful but expensive solution involving strong CS-SP interactions.
	Current loads of the paths (bandwidth availability)	CSs	Dynamic	Yes	
SP-V4	**Idem as SP-V3, plus below items**				Idem as in SP-V3 but more assistance is available in the selection process of the initial (server-path) pair.
	Other dynamic paths characteristics (delay, loss, jitter, etc.)	CSs	Dynamic	Yes	
SP-V5	**Idem as SP-V4, plus below items**				Idem as in SP-V4, but more flexibility from business point of view. The SP offered services can be better customized.
	SP Policy Information	Mgmt@SP	Static	No	
SP-V6	**Idem as SP-V5, plus below items**				Idem as in SP-V5, plus more powerful set of knowledge on system history. *(maximum complexity)*
	Historical and prediction data on servers and paths utilization	Mgmt@SP + Mon@SP	Dynamic	Yes	

TABLE II. CONTENT SERVER –DESIGN VERSIONS

	Information known (by CS) about:	**Obtained from**	**Type of information**	**Is Mon@CS involved?** (in collecting the Column 2 nformation)	**Remarks on CS role**
CS-V1	EU authorization data	Mgmt@SP	Quasi-static	No	The selected (by the EUT) CS solves the user content requests. CS status info is delivered to SP. CS info on active users can be also delivered to SP.
	EU requests	EUTs	Dynamic	Yes	
	CS status (current load)	Mgmt@CS	Dynamic	Yes	
	Active (in session) users	EUs	Dynamic	Yes	
CS-V2	**Idem as CS-V1, plus below items**				Idem as in CS-V1 but more assistance in offering (via SP) additional information for selection of the initial (server-path) pair.
	Potential user groups	SP	Quasi-static	No	
	Static characteristics of	Mon@CSs	Quasi-static	Yes	

	connectivity paths (evaluated at overlay level) from different CSs to different groups of users				
CS-V3	**Idem as CS-V2, plus below items**				These data can be be sent to SP to help for more efficient management of EU connections. If multiple paths are available, the CSs should have some source routing capabilities in order to force the stream to follow a given path.
	Active User groups	Mgmt@CS	Dynamic	Yes	
	Connectivity paths dynamic characteristics (evaluated at overlay level) from different CSs to different groups of users	Mon@CS	Dynamic	Yes	

TABLE III. END USER TERMINAL – DESIGN VERSIONS

	Information known about:	**Obtained from**	**Type of information**	**Is Mon@EUT involved?** (in collecting the Column 2 nformation)	**Remarks on EUT role**
EUT-V1	EUT local static context	Mgmt@EUT	Quasi-static	**No**	EUT issues content requests to SP. For server selection it uses the MD file sent by SP and its local static context information. For dual adaptation it uses the monitored data (including the DASH embedded one) and basic probing information.
	MD file	SP	Dynamic	No	
	QoE quality during session	Mon@EUT	Dynamic	Yes	
	CS accessibility (basic probing)	Mon@EUT	Dynamic	Yes	
EUT-V2	**Idem as EUT-V1, plus items below**				EUT issues content requests to SP. For server selection it uses the MD file sent by SP and its static context information. For dual adaptation it uses the monitored data and probing information.
	EUT local dynamic context	Mon@EUT	Dynamic	Yes	
	CS accessibility (advanced probing)	Mon@EUT	Dynamic	Yes	
EUT-V3	**Idem as EUT-V2, plus items below**				
	Local Policy information	Mgmt@EUT	Static	No	Possible local policy data are used in server selection and dual adaptation.
EUT-V4	**Idem as EUT-V3, plus items below**				
	Historical and prediction data on servers and paths utilization	Mgmt@EUT Mon@EUT	Dynamic	Yes	Possible history and prediction data are used in server selection and dual adaptation.

V. DESIGN DECISIONS DETAILS

This section will refine the functional blocks introduced in the previous one, in order to prepare the software functionalities specifications. We recall that objective of this section is limited to refine the architecture and proceed to design of the functional blocks, in such a way as to respond to the flexibility features proposed initially. The validation and performances of the system will be treated in other complementary works during the project.

The functional blocks inside each DISEDAN actor are presented in high level view, in the following subsections. Figures 2, 3 and 4 show the complete architecture (functional blocks and the relationship between them). However, in the basic version of the DISEDAN system, only the most important ones are actually implemented – i.e., those needed to prove the innovative concepts.

Figures 2, 3 and 4, use three marking types for functional blocks:

- the functions mandatory implemented in the basic DISEDAN version are depicted in *dark gray color* boxes;
- the *light gray colour* boxes can be implemented as static versions in the basic DISEDAN system and extended with dynamic capabilities for advanced SP versions;
- the *white color boxes* represent functionalities which could exist in advanced versions of SP (i.e., complete real life system implementation, with advanced functionality, in order to provide the best possible QoE).

A. Service Provider Functional Blocks

Figure 2 shows the SP functional blocks.

The (lightweight) SP functional blocks implemented for DISEDAN basic version are:

- *Comm Agent* –used for communication with external entities (in particular EUT and CS). It can also be used by the monitoring system. The *Comm Agent* will serve only the Management and Control

Plane communications between SP and EUTs and CSs.

- *MPD File generator* – dynamically generates *Media Presentation Description (MPD) XML* file, containing media segments information (video resolution, bit rates etc.), ranked list of recommended CSs and, optionally - current CSs state information and network state (if applicable).
- *MCDA* - performs the selection algorithm – i.e., runs Step 1 of the server selection process.
- *Monitoring module* – basically it collects monitoring information from CSs and performs the processing required to estimate the current state of each CS. If some EU-related information should go to SP, then this information are collected by the CS from EUT, and then aggregated and transited towards SP.
- *Data Base* - contains the static and dynamic information about CSs, EUT communities and profiles, etc. In the basic version these static information are filled offline into DB. In advanced SP implementations, the DB can be split in two modules Run-time DB and Quasi-static DB, containing respectively fast volatile data and respectively mid-long term data.
- *CS Discovery* - it has the role to discover the CSs locations, their main characteristics and content items available (mapping of the content objects - to - CS). In the basic version of SP these data can be statically introduced by the administrator.
- *EUT Discovery*- it has the role to discover the EUT groups locations and their main profiles. In the basic version of SP these data can be statically introduced by the administrator.

In advanced versions of SP implementations, other blocks can be added and also some of the existing ones are enhanced to have dynamic capabilities:

- *CS Dynamic Discovery* - it has the role to dynamically discover the CSs locations, their main characteristics and content items available (mapping of the content objects – to – CS). Periodical updates are necessary. This module should be existent in the SP-V2…SP-V6 versions of SP.
- *EUT Dynamic Discovery*- it has the role to dynamically discover the EUT groups locations and their main profiles. Periodical updates are necessary. This module should be existent in the SP-V2…SP-V6 versions of SP.

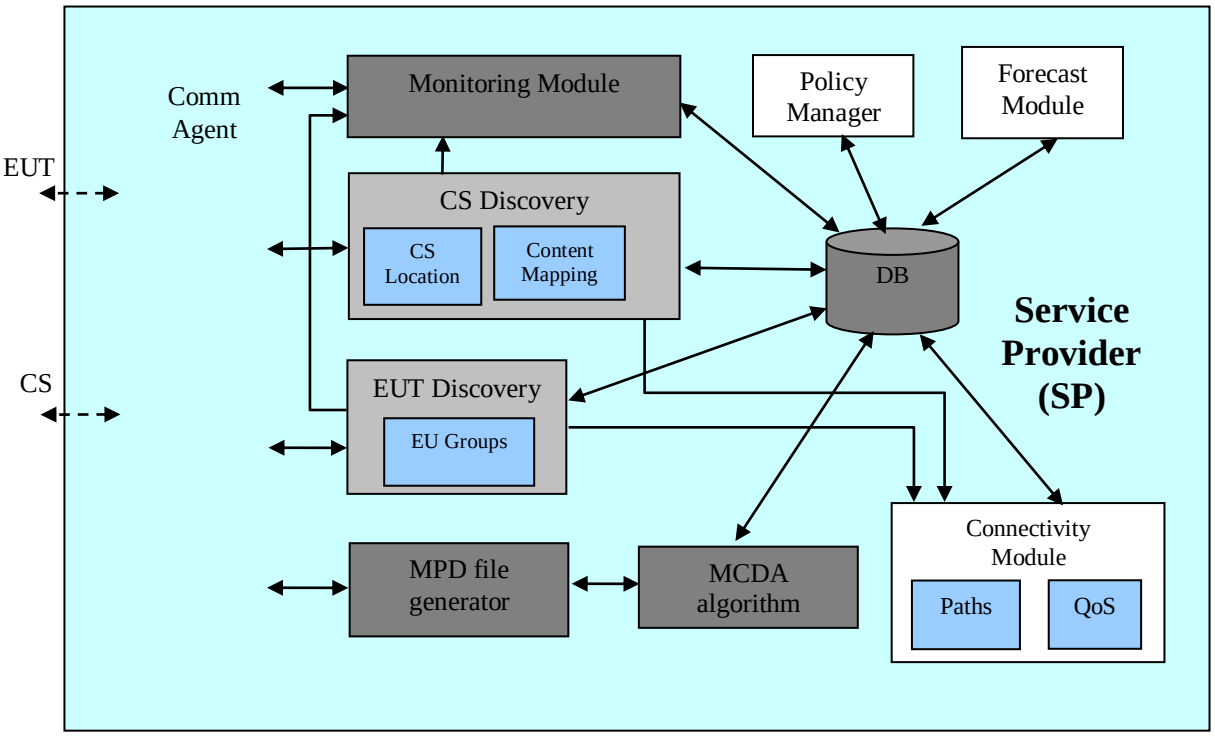


Figure 2. The Service Provider software architecture

- *Connectivity Module* – is an optional functional block, collecting and maintaining updates about connectivity characteristics (paths, QoS parameters) between the CSs of interest (list of them is delivered by the CS discovery module) and different groups of EUTs. If this module exists, then the SP level selection decision of the content server can be more powerful. The reason is that in such case not only CSs are selected but also some lists of "good" pairs {CS, path} could be selected and proposed to the EUT, in response of its request. A scalable solution (given the multi-domain characteristic of DISEDAN) for such a block implementation is to consider only overlay paths crossing one or several network domains. This module should be existent in the SP-V2…SP-V6 versions of SP.
- *Policy Manager* – is a module that can apply various policies at SP level (e.g., related to business and/or technical aspects). The effect will be the modification of the selection produced by MCDA, and, consequently, of the list returned to the EUT. This module should be existent in advanced SP-V5, SP-V6 versions of SP. Such policies can provide inputs to the MCDA process (see [15]), in two ways:
  - by assigning different weights to the existing decision variables – depending on policy considerations,
  - by defining new decision variables (derived from policies) to the MCDA matrix.
- *Forecast Module* – A module that can make educated predictions based on various data like history, communication preferences with other similar modules in places, like SP and CS. This module should be existent in the SP-V5…SP-V6 versions of SP.

B. End User Terminal Functional Blocks

The EUT high level architecture is presented in Fig. 3. The End User Terminal control logic contributes (by cooperating with SP) to the selection of the best available Content Server in order to provide the best possible experience to the user. On the request of the human End User, the EUT will send the request to the SP. The SP response contains an ordered list of preferred (after SP evaluation) Content Servers. The order of the CS in the list represents the preference of the servers from SP point of view.

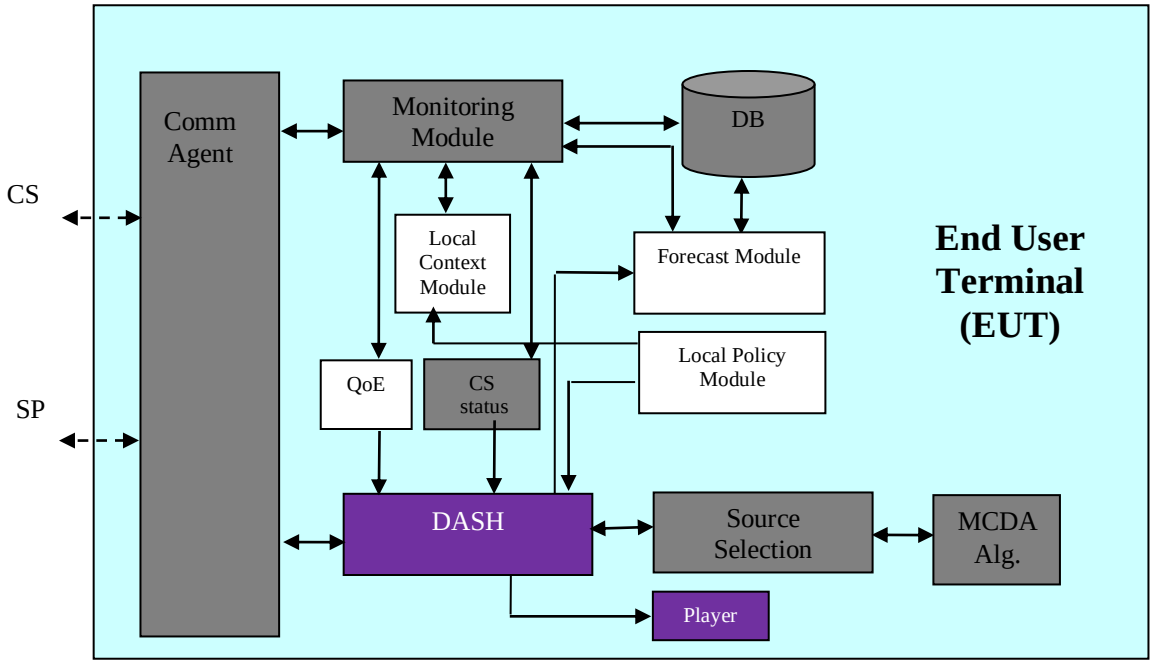


Figure 3. The End User Terminal software architecture

(DASH and Player boxes represent libraries of already existing software)

The functional blocks implemented in the DISEDAN basic version are:

- *Comm Agent* – used for communication with other external entities (in particular, SP and CS). It can also be used by the monitoring system.
- *CS status* – it can be seen as a monitoring system part, but here it is depicted as standalone, to emphasize its role in proof of concepts. This block finds out (considering the path from CSs to EUTs) the Round Trip Time (RTT) and Number_of_Hops towards each CS selected by the SP and reachable by the EUT. This information can be used later in the MCDA algorithm inside EUT (when making the final CS selection).
- *MCDA alg* – it runs the MCDA algorithm. A sub-block not represented in the figure does exist inside MCDA algorithm. This bloc creates the input data for the algorithm, ie., the matrix of values (see [15][16]), constructed from data like RTT, hop count and the priority as set by the SP. The result provided by the MCDA is sent to the *Source Selection* block.
- *Source Selection* – based on the decision of the MCDA algorithm this block creates the URL request for the elected CS and calls the DASH player.
- *Monitoring Module* – it can monitor various local aspects of the network or of the terminal itself. Note that it can be a simple block fed with static information – for the basic implementation; however, it could perform real monitoring in advanced EUT versions.

DB – A data base used to keep EUT staic and volatile data.Special blocks are coming form open source software that DISEDAN uses for the proof of concept:

- *DASH – The Dynamic Adaptive Streaming over HTTP* library. It is the DASH client that gets the movie from the CS where the DASH server is running.
- *Player* – the VLC media player is the ones that finally displays the movie content on the EUT screen.

The following blocks can be implemented in DISEDAN advanced versions, e.g., for some complete commercial implementation:

- *QoE* – Quality of Experience block; it is part of monitoring but is depicted separately in Fig.3, given its importance (recall that the final goal is to contribute to achieve the highest possible QoE). Additionally, the DASH subsystem itself has various internal mechanisms to adapt to environment conditions.
- *Local Module Policy* – it is a module that can apply various policies at the EUT level (e.g., related to business aspects and/or some special policies like for example parental-child control-related policies).
- *Local Context Module* – it is an agent that can gather information about the terminal (by aggregating static EUT information and dynamic monitoring information.

Forecast Module – makes based-on–learning predictions based on various data like history, preferences, communication with other similar modules in places like SP and CS.

C. Content Server Functional Blocks

The CS high level architecture is presented in Fig. 4. The functional blocks implemented for DISEDAN basic version are:

- *Comm Agent* – used for communication with other external entities (in particular SP and EUT). It can also be used by the monitoring system.
- *Data Plane - DASH Streaming module* – sends media segments requested by End Users.
- *Monitoring module* – monitors the basic CS performance metrics (CPU utilization, network interfaces utilization, etc.).
- *Current State module* – contains the main parameters describing this CS status (EU served, number of sessions, load, etc.).
- *Data Base* – contains information produced by the monitoring and also the data about currently EUs served by this CS and maybe some potential ones.

The following blocks can be implemented for advanced DISEDAN versions of the CS:

- *Advanced Monitoring module* - in a complex implementation of the CS, the monitoring can evolve from a simple probe to an advanced monitoring module, capable to supervise not only the active sessions but also some connectivity characteristics from this CS to different groups of users.
- *EU Module* – determines data about the EUs (status, groups) by using communication services offered by the communication agent.
- *Connectivity Module* – optional functional block, collecting and maintaining updates about connectivity characteristics (paths, QoS parameters) between the CS and different groups of EUTs. If existent, this module would provide additional information to the Connectivity Module of the SP.
- *AAA Module* – performs conventional Authentication, Authorization and Accounting functions. It is not essential for DISEDAN proof of concepts.
- *Forecast Module* – can make based-on–learning predictions based on various data like history,

preferences, communication with other similar modules in places like SP and CS.

VI. Conclusions and Future Work

This paper presented an analysis of design decisions for implementation variants of a novel and flexible light-architecture content delivery system, working on top of the current Internet networks. The system involves a Service Provider, End Users and Content Servers owned by the SP.

The novelty consists in including in a single solution of initial content server selection, (based on collaboration SP - EU, and multi-criteria optimization algorithms like MCDA, EMO, etc.) and session-time DASH adaptation and/or intelligent server switching (if the quality of the flow is degraded at the End User).

Several versions of designs are proposed, illustrating the architectural approach flexibility and comments are given on the associated complexity.

The main DISEDAN advantage consists in avoiding to develop complex M&C planes and signaling, while still offering sufficient QoE (due to adaptation capabilities) to the end users, in a cheap and fast implementable OTT-style solution. Note that the price paid for this lower cost solution is paid by the fact that DISEDAN cannot provide contracted (via SLA) hard QoS/QoE guarantees for its users.

However, the DISEDAN architecture is flexible, in the sense that it can benefit, if existent, of better managed connectivity services in the network; as well, it can benefit from information related to network static characteristics and dynamic monitoring data, possible to be collected by the Control Plane. In such cases the MCDA-based server selection algorithm can provide better solutions and consequently, higher QoE perceived by the End Users.

Preliminary results assessing the validity of the solution and performance of the algorithms are already reported in [15][16][18].Ongoing work is currently performed, to implement the described system (in the DISEDAN project).

In parallel with design and implementations, simulations have been performed (see extensive results in [18]), including large scale network environment, to prove the capabilities of the proposed architecture.

Details on the implementation of the functional blocks will be reported in future papers.

Complete results and performance evaluation of the implemented system, obtained for different use cases, will be also reported in some future papers. Comparisons with existing systems will be also provided.

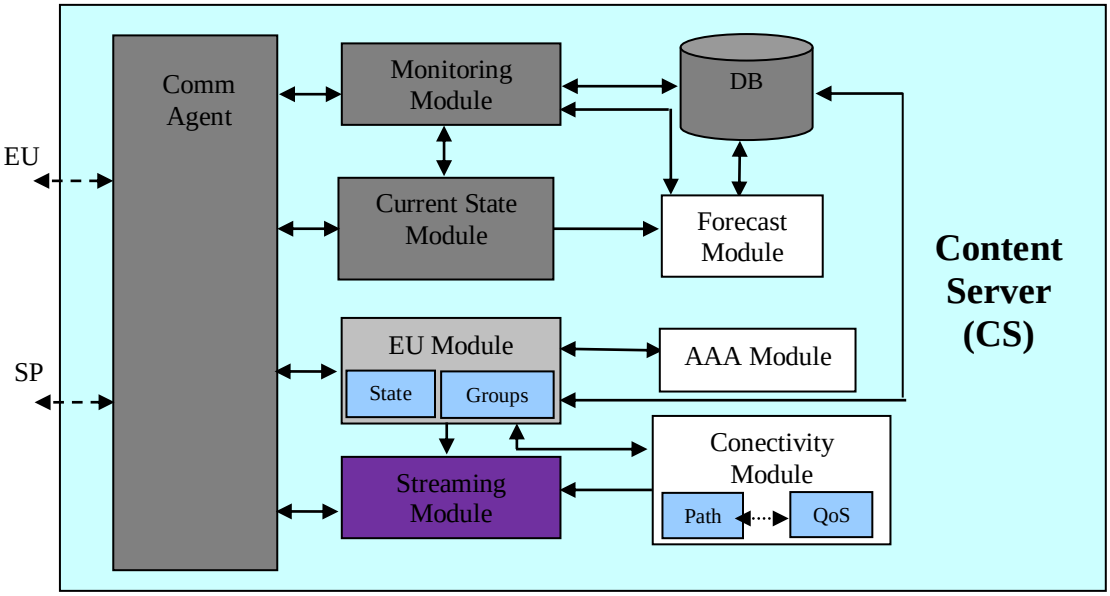


Figure 4. The Content Server software architecture

ACKNOWLEDGMENTS

This work has been partially supported by the Research Project DISEDAN, No.3-CHIST-ERA C3N, 2014- 2015.

REFERENCES

[1] E. Borcoci, C. Cernat, R. Iorga, J. M. Batalla, D. Negru, "Flexible Design of a Light-Architecture Content Streaming System with Dual Adaptation," The International Symposium on Advances in Content-oriented Networks and Systems, pp 114-119 http://www.iaria.org/conferences2015/CONNET.html, [Accessed October 2015].

[2] J. Pan, S. Paul, and R. Jain, "A survey of the research on future internet architectures," IEEE Communications Magazine, vol. 49, no. 7, pp. 26-36, July 2011.

[3] P. A. Khan and B. Rajkumar. "A Taxonomy and Survey of Content Delivery Networks". Department of Computer Science and Software Engineering, University of Melbourne. Australia : s.n., 2008, www.cloudbus.org/reports/CDN-Taxonomy.pdf, [Accessed October 2015].

[4] ***, "Information-Centric Networking-3", Dagstuhl Seminar, July 13-16 2014, Available from: http://www.dagstuhl.de/en/program/calendar/semhp/?seminar=14291, [Accessed October 2015].

[5] V. Jacobson et al., "Networking Named Content," CoNEXT '09, New York, NY, pp. 1–12, 2009,

[6] http://wp2.tele.pw.edu.pl/disedan/publications, [Accessed October 2015].

[7] T. Dreier, "Netflix sees cost savings in MPEG DASH adoption," 15 December 2011. [Online]. Available from: http://www.streamingmedia.com/Articles/ReadArticle.aspx?ArticleID=79409, [Accessed October 2014].

[8] S. Wenger, Y. Wang, T. Schierl and A. Eleftheriadis, "RTP Payload Format for Scalable Video Coding," IETF RFC 6190, 2011.

[9] ISO/IEC 23009-1, "Information technology -- Dynamic adaptive streaming over HTTP (DASH) -- Part 1: Media presentation description and segment formats," ISO/IEC, Geneva, 2014.

[10] M. Kawarasaki, K. Ooto, T. Nakanishi, and H. Suzuki, "Metadata driven seamless content handover in ubiquitous environment," in Proceedings of the 2004 International Symposium on Applications and the Internet SAINT'04, Tokyo, 2004.

[11] S. Park and S. Jeong, "Mobile IPTV: Approaches, Challenges, Standards and QoS Support," IEEE Internet Computing, vol. 13, no. 3, p. 23–31, 2009.

[12] L. De Cicco, S. Mascolo, and V. Palmisano, "Feedback control for adaptive live video streaming," in MMSys '11 Proceedings of the second annual ACM conference on Multimedia systems, San Jose, California, 2011.

[13] J. Figueira, S. Greco, and M. Ehrgott, "Multiple Criteria Decision Analysis: state of the art surveys," Kluwer Academic Publishers, 2005.

[14] A. Beben, J. M. Batalla, W. Chai, and J. Sliwinski, "Multi-criteria decision algorithms for efficient content delivery in content networks," Annals of Telecommunications, vol. 68, Issue 3, pp. 153-165, Springer, 2013,

[15] E. Borcoci, M. Vochin, M. Constantinescu, J. M. Batalla, D. Negru, "On Server and Path Selection Algorithms and Policies in a light Content-Aware Networking Architecture," ICSNC 2014 Conference, Available from: http://www.iaria.org/conferences2015/ICSNC15.html, [Accessed October 2015].

[16] O. Catrina, E. Borcoci, and P. Krawiec (WUT), "Two-Phase Multi-criteria Server Selection for Lightweight Video Distribution Systems," 27th IFIP TC7 Conference 2015 on System Modeling and Optimization, Integration of Optimization, Modeling and Data Analysis for Solving Real World Problems.

[17] E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach". IEEE Transactions on Evolutionary Computation, No. 3(4), pp. 257-271, November 1999.

[18] J. Mongay Batalla ed., A. Beben, P. Krawiec, P. Wisniewski, D. Negru, J. Bruneau-Queyreix,, E. Borcoci, R. Badea, DISEDAN D2.3 "Specification of the Dual adaptation mechanism," http://wp2.tele.pw.edu.pl/disedan/publications, [Accessed October 2015].

www.iariajournals.org

International Journal On Advances in Intelligent Systems
issn: 1942-2679

International Journal On Advances in Internet Technology
issn: 1942-2652

International Journal On Advances in Life Sciences
issn: 1942-2660

International Journal On Advances in Networks and Services
issn: 1942-2644

International Journal On Advances in Security
issn: 1942-2636

International Journal On Advances in Software
issn: 1942-2628

International Journal On Advances in Systems and Measurements
issn: 1942-261x

International Journal On Advances in Telecommunications
issn: 1942-2601