Infocommunications Journal

A PUBLICATION OF THE SCIENTIFIC ASSOCIATION FOR INFOCOMMUNICATIONS (HTE)

June 2014 Volume VI Number 2 ISSN 2061-2079

Technically Co-Sponsored by

IEEE COMMUNICATIONS SOCIETY

hte

IEEE HUNGARY SECTION

Indexing information

Infocommunications Journal is covered by Inspec, Compendex and Scopus.

Infocommunications Journal

Technically co-sponsored by IEEE Communications Society and IEEE Hungary Section

Supporters

GÁBOR BÓDI – president, National Council for Telecommunications and Informatics (NHIT)
GÁBOR MAGYAR – president, Scientific Association for Infocommunications (HTE)

Editorial Office (Subscription and Advertisements)**:**
Scientific Association for Infocommunications
H-1051 Budapest, Bajcsy-Zsilinszky str. 12, Room: 502
Phone: +36 1 353 1027, Fax: +36 1 353 0451
E-mail: info@hte.hu • Web: www.hte.hu

Articles can be sent also to the following address:
Budapest University of Technology and Economics
Department of Networked Systems and Services
Tel.: +36 1 463 3261, Fax: +36 1 463 3263
E-mail: szabo@hit.bme.hu

Subscription rates for foreign subscribers: 4 issues 50 USD, single copies 15 USD + postage

Publisher: PÉTER NAGY • Manager: ANDRÁS DANKÓ

HU ISSN 2061-2079 • Layout: MATT DTP Bt. • Printed by: FOM Media

Analysis of Assessment Alteration Phenomena of Subjective Quality of Experience Measurements in 2D and 3D Mobile Video Services

Péter András Kara†, *Member, IEEE*, László Bokor†, *Member, IEEE*, Sándor Imre†, *Senior Member, IEEE*

***Abstract*— The growing importance of Quality of Experience over Quality of Service demands precise results in the monitoring of experienced quality; empirical assessment of subjective QoE measurement on perceived quality is expected to deliver accurate reflection of reality. The goal of this paper is to highlight a specific potential error in existing subjective QoE measurement methodologies. Our approach focuses on a special topic of distortions caused by preconceptions based on prior technical knowledge of evaluation measurement test subjects. The paper presents two series of measurements where the test subjects were aware of the service parameters during the evaluation of the given services. The paper specifies the identified distortion phenomenon and shows how cognitive dissonance played a role in the formation of evaluation patterns and the distortion of the Mean Opinion Score.**

Keywords: Quality of Experience, Quality of Service, Mean Opinion Score, 3G HSDPA, 2D and 3D video services, cognitive dissonance

I. INTRODUCTION

One of the most important pillars of modern society is the provision and consumption of services. The list of properties of a service provides comparable information to the consumer. Although this does seem to be the universal method of comparison between services of the same kind, it must not be ignored that it is not the equivalent of actual user experience. This means that no matter how high such properties score if the service does not satisfy the consumer. For instance, in case of a video chat which uses mobile Internet connection, it is totally irrelevant how staggering the bandwidth is when the two participants of the conversation have a hard time understanding each other. This leads to the conclusion that the true value of a service rather lies in the "degree of delight or annoyance of the user" [1] (Quality of Experience – QoE) than the "totality of characteristics" [2] (Quality of Service – QoS).

Of course QoE and QoS are unquestionably connected, but their precise relationship is hard to define. However, there are some promising recent researches to flawlessly forge QoE values from a set of QoS parameters (e.g. [3]), yet a widely accepted method is still lacking. Service providers inevitably require user feedback on end-to-end performance to reach a cost effective level of QoE. Monitoring QoE primarily benefits for service providers, but on the other hand, it improves reception for subscribers.

Because of its importance, QoE monitoring is a well defined, standardized process [4]. However, the results of such measurements are affected by environmental information, for example the type of connection, location, device or even some available QoS parameters. In this study, we introduce that the usage of such information depends on the subject's prior technical knowledge and experience on the present technology (Level of Comprehension – LoC). Our term of Level of Comprehension [5] could be defined as "the amount of one's prior technical knowledge and experience which deduces and implies the possible usage of environmental data". In some cases, the awareness of parameters regarding the service cannot be avoided; therefore the results are preordained to be altered. Several examples can be mentioned from everyday life, where the preconceptions create distortions in user experience. The direction and power of these effects are quite far from triviality, yet it hasn't been circumspectly analyzed so far. Mobile video services demand accurate measurements and could benefit from the avoidance or at least the reduction of such distortions.

The complete definition of QoE also states that "it results from the fulfillment of his or her expectations" [1]. In this case, "expectation" refers to the desired level of quality which one has towards a specific service. However, a different interpretation of this word also plays a significant role. The word expectation also means the level of quality one anticipates to experience; a prior idea, a preconception of quality. One could easily presume perceived quality to utterly match these anticipations, but what happens when it doesn't? That would create a disharmonic state between the objective cognition of perception and the subjective cognition of preconception. The theorem of cognitive dissonance [6] explains the different methods of dissonance reduction that could occur in such a situation.

This article deals with the following topic: we study how the combination of aforementioned QoS parameters and different LoC levels alter the assessment results of two different QoE measurements. We also examine the unavoidable psychological reason which empowers preconceptions and manages to alter the refinement of perception. Measurement M1 was the evaluation of a video conference performed on a real-life 3G HSDPA network, while M2 was aiming at 3D multimedia streaming through a GPON transport and Wi-Fi access network. In both cases, the objective of the test participants was to grade the experienced quality, while possessing the parameters of the connection. The research goal was identify the alteration phenomena in both cases and to find correlation between the distinguished levels of LoC and the altered results; how prior knowledge and experience influence QoE.

The article begins with the introduction of assessment alteration approaches with some up-to-date examples and related work in Section II, followed by the configuration and the results of our experiments in Section III and IV. The last section concludes the paper, containing the possible future directions of this topic.

Manuscript received January 28, 2014, revised May 15, 2014
† Department of Networked Systems and Services, Budapest University of Technology and Economics, Budapest, H-1117, Magyar Tudósok körútja 2, Hungary, {kara, goodzi, imre}@mcl.hu

II. BACKGROUND AND RELATED WORK

The field of subjective determination of transmission quality has well defined standards, intensely detailed recommendations, and countless of exceptional papers sharing the experiences of researches and measurements. The ITU-T P series [7] provide a wide range of recommendations relating to the topic. A fine example for a subjective, context-aware, real-life QoE measurement was conducted by I. Ketykó et al. [8], dealing with the interference of the location and the number of surrounding people on perceived quality. Test subjects in both of our measurements were isolated in a fixed location, yet it is an exciting idea to investigate the explicit effects of the environment. We find it even more interesting to analyze the implicit effects of environmental information in case of varying location, which is a possible continuation of our topic. Explicit effects like the rise of environmental noise level due to the presence of vehicles, machinery or people affect perception and focus without a doubt, but we assume that awareness regarding the environmental information comes with its own distortions. Just for instance, a mobile location like urban public transportation implies mobile access to the service network. Mobility-awareness can create preconceptions and thus shape QoE measurement results, which is the topic of one of our current projects.

Of course the usage environment is not the only factor that affects quality. G. Exarchakos et al [9] highlights how the level of perceptual quality relies on the specific content and network impairments. Although our measurement M1 featured motion in the video conversation (e.g., the test moderator moves from one part of the camera field of view to another), still did not show high vulnerability towards packet loss due to the lack of numerous and recurrent interchanges between video frames. However, the content of measurement M2 was a high motion animated video stream, which made network impairments – especially packet loss – straightforwardly visible to the evaluator, usually in form of artifacts [10].

Terminals are also important variables of usage scenarios. The work of F. Agboma et al [11] details the correlations between the terminal of a given service and perceived quality. Indeed, while the conventional 2D PC monitor display of M1 posed no issues of technology acceptance, sometimes the active 3D display of M2 caused a real headache, literally [12].

Evaluation itself can be done in a qualitative or a quantitative matter. The paper of P. Brooks et al. [13] marks the importance of quantitative evaluation methodologies, since qualitative labels can question objectivity. A qualitative approach may indeed create distortions in evaluation due to the subjective meanings of different labels [14] and may result contrarily in different languages [15]. The paper of A. Watson et al. [16] also doubts the usability of such scales, indicates limitations and warns about the compression of measurement results to the lower half of the scale. V. Menkovski et al [17] argues with the qualitative absolute scale of ratings as well due to the uttermost subjectivity in their interpretations. Both of our measurements utilize the numerical evaluation of discrete scales.

Another work of V. Menkovski et al [18] also emphasize with the dense variety of factors responsible for the non-linear relationship between the physical and the perception domain. They present an active learning algorithm, an adaptive MLDS (Maximum Likelihood Difference Scaling) to increase the efficiency, scalability and the learning rate of the existing approaches [19]. The numerical results of both our subjective measurements contain psychometric functions, putting in relation physical and psychological scales. However, even with a greater number of participants compared to what we have had during measurement M2 (90 participants), variance and bias are expected to be included [20]. Objective solutions like MLDS are not only interesting due to the reduction of bias, but for the elimination of socio-psychological alterations as well.

In our study we deal with this specific type of assessment distortion. Evaluation during a measurement is nothing but a series of decisions. Due to this fact, cognitive dissonance [6] and especially post-decision dissonance [21] affect evaluator behavior. As mentioned earlier, cognitive dissonance is a disharmonic state between conflicting cognitions, which needs to be resolved in order to avoid discomfort, stress and other unwanted feelings. This is quite relevant in case of quality assessment since it encourages test participants to support prior ideas regarding the service instead of perception, resulting in the alteration of the actual experience and thus the scores as well. Post-decision dissonance protects the validity of prior decisions, which in case of assessment, forges a harmony between the results of evaluation tasks; evaluating a given test case in a measurement series is heavily affected by earlier evaluation decisions.

This interesting topic is investigated by others as well. The work of A. Sackl et al. [22] demonstrates the inevitable role of cognitive dissonance in QoE and underlines the correlation between experienced service quality and pricing. There is indeed a close linkage between quality perception and willingness-to-pay, and with the detailed phenomenon of post-decision dissonance, referred to as “post purchase cognitive dissonance”, they managed to clearly explain the background of their results. They emphasize the human action of justification, which is also a key element of the naissance of measurement result distortion in our study. While their first experiment of streaming video evaluation involved real-world currency and active user decisions, the second one lacked interaction. In order to justify the binary decisions of purchasing or not purchasing in the first experiment, the participants evaluated the given services with higher scores compared to the results of experiment without user decisions. Our works are rather related to the second experiment, since the only so-called interaction is the participation in a video conference in M1, as shall be seen later on. However, we still deal with justification in our series of measurements, due to the presence preconceptions; once an evaluator supports a specific idea with a finalized decision, it is likely to be repeated later on with the purpose of justifying the previous one.

The publications of M. O'Neill and A. Palmer [23][24] also gained our attention. Their research includes a time difference of one month, which enables post-decision dissonance to have a more significant, evolved impact. The intervals between evaluations in our measurements were merely a couple of minutes, resulting in short-term consequences of the phenomenon.

III. MEASUREMENT CONFIGURATION

A. Measurement methodology

As mentioned in the introduction, QoE monitoring plays an essential role in designing, initializing and maintaining services. The standard techniques for such measurements are defined by the recommendation [4] of the International Telecommunication Union. It contains all the important parameters that can be involved in the configuration of a QoE measurement. Subjective determination of transmission quality can be achieved by four different clusters of methods. The most popular ones are considered to be the conversation-opinion tests, since they are designed to replicate actual usage of two-way interactive services. Listening-opinion tests rather focus on ones perception, which makes them excellent to measure basic usability and acceptance. Interview and survey tests are efficient methods to extract information beyond a numerical judgment. A group labeled "other tests" is also defined. We decided to use conversation-opinion tests in measurement M1 and listening-opinion tests in M2, both with minor additions from interview tests methodology; test subjects were able to detail their decisions during recorded interviews. Additional verbal extension of evaluation supports understanding the motivations behind evaluator behavior.

Before the measurement itself, the Level of Comprehension of each subject was revealed by asking a set of questions related to the background of the concerned telecommunication technologies and solutions. For instance, in case of M2, questions on 3D display technologies and network security were necessarily included. These conversations, each taking approximately thirty minutes, were recorded for further analysis to precisely determine the LoC of the subjects. It needs to be noted that although this method of LoC determination required vast resources, we could not risk losing a desired level of accuracy regarding a proper selection. The determination process happened manually in an iterative manner; the ones with the greatest and the lowest technical competences were selected and the process was repeated until all participants were categorized. Although this method prevents the possible LoC overestimation, we shall use a more cost effective approach in the future.

Three different levels were distinguished; level –1 represented the group of those with the lowest, while level +1 represented the highest level of technical comprehension, and level 0 was in between. For more intense investigation of the correlation, more levels could be defined (e.g., M. R. Quintero et al. defined 6 [25]). To preserve the purity of LoC determination, the subjects were given no information about the nature of the measurement before it had begun. The variety of technical competence was not the only aspect during the selection of the test subjects, but it was also necessary to only select people who have never seen each other before in order to prevent information leak between measurements. The subjects haven't even met each other during the series of measurements, because of the different dates and times of the measurements. If any subject had received even the slightest information about the measurement before its date, it could have and probably would have resulted in LoC overestimation.

B. Configuration of Measurement M1

The basic set of measurements for our analysis of M1 was built on a video conference between the test moderator and the test participant, such emulating a typical mobile video service. The tests were performed on the laboratory network (see *Figure 1*) of the Mobile Innovation Centre [26]. Twenty test subjects participated in the series of measurements with different levels of prior technical knowledge, ranging from simple inexperienced user to IT engineer with PhD degree. Although test subject number may be considered to be low in the aspect of representative results, it is sufficient at this initial phase to expound the phenomenon and analyze evaluator behavior, not to mention our focus on user rating diversity [27].

The complete process of a measurement was divided into four sections, following each other without delay. The first part was the LoC level determination conversation, as mentioned before. This was extended by questions on general user behavior, involving the quality of previously experienced video conferences. After the basic instructions, began the third and most important part of the process, the mobile video conference and its evaluation. This was concluded by an oral evaluation of the experienced quality, which was also recorded like the first two conversations. The test moderator was the same in each and every part of the process and for all subjects.

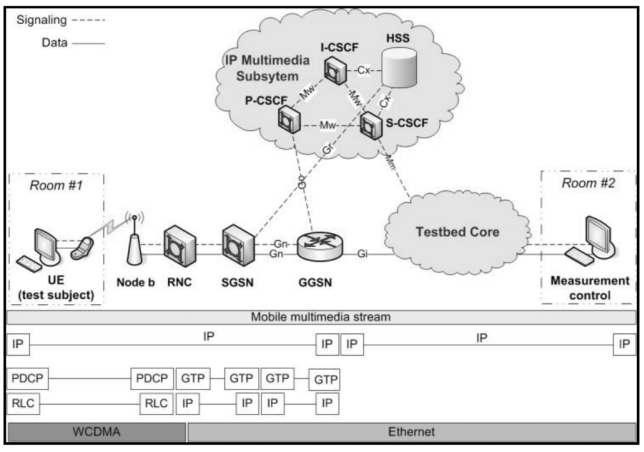


Figure 1. Network topology of measurement M1

During the video conference, the test moderator used a terminal in the laboratory of the Mobile Innovation Centre (Room #2 in *Figure 1*), while the subject was isolated in the

conference room of the laboratory. The audiovisual connection was established by a *Linphone 3.2.1* client [28] on an *Ubuntu 10.04* operation system. Both end terminals shared the same hardware and software, including multimedia equipment such as web camera and headset. Connection to the test network, however, was different. While the terminal at the laboratory connected via Ethernet, the computer at the conference room (Room #1 in *Figure 1*) used a Huawei 3G HSDPA wireless modem. IP Multimedia Subsystem (IMS) [29] was in control of the mobile multimedia traffic over the UMTS network.

The complete video conversation took approximately one hour, which of course may introduce the presence of the memory effect [30], but also enhances the usage of environmental information. Although it was divided into twenty subsections (referred to as test cases), the conversation itself was fluent and natural. It was enough to have test cases with 3 minutes of length, since longer test cases would not have led to significant differences in the perception of quality [31]. However, perception varies over time [32], so it was necessary to keep the complete length at a reasonable extent in order to comply with the attention span. Every subsection had a different artificial one-way QoS parameter load in terms of delay, jitter and packet loss, in addition to the real QoS values of the network. To achieve this, we used the command line based *netem* application [33] in order to change the output traffic of the laboratory terminal without the interruption or pause of the video conversation. The achieved impairment of QoS resulted different artifacts and stalling. The parameter values were given to the subject before commencing the conversation, in a form of a QoS parameter matrix (see *TABLE I*), together with the fix parameters of the measurement (see *TABLE II*), such as video resolution. The objective of the subject was to separately evaluate the audio and video quality of the twenty different test cases on a scale from one to ten, where ten represented the best score. Although five-point scales are indeed more popular in case of evaluation, we chose this size in order to support test subjects in distinguishing their experiences.

TABLE I. QOS PARAMETER MATRIX VARIABLE VALUES OF M1

Test case	Varying parameters		
	Additional delay	*Additional jitter*	*Additional packet loss*
1	0 ms	0 ms	0 %
2	50 ms	10 ms	0.5 %
3	200 ms	40 ms	2 %
4	800 ms	180 ms	8 %
5	0 ms	180 ms	8 %
6	0 ms	0 ms	8 %
7	0 ms	180 ms	0 %
8	800 ms	0 ms	0 %
9	800 ms	100 ms	1.2 %
10	400 ms	100 ms	1.2 %
11	200 ms	100 ms	1.2 %
12	100 ms	100 ms	1.2 %
13	100 ms	180 ms	0.5 %
14	100 ms	100 ms	0.5 %
15	100 ms	40 ms	0.5 %
16	100 ms	20 ms	0.5 %
17	200 ms	20 ms	0.5 %
18	200 ms	20 ms	2 %
19	200 ms	20 ms	4 %
20	200 ms	20 ms	8 %

TABLE II. DESCRIPTORS OF VIDEO CONFERENCE IN M1

Resolution: 640x480
Video codec: MPEG4
Audio codec: speex

C. Configuration of Measurement M2

QoE measurement series M2 was the assessment of 3D multimedia streams on a PC with Nvidia Vision active 3D technology [34]. The task of the participants was to rate five different aspects of quality of 20 test cases (see *TABLE III*). The chosen aspects were video continuity, image quality, 3D experience, audio/video synchronization and the overall experience. The variables of the test cases included jitter, packet loss, transmission power, and the binary presence of bandwidth limitation and network security.

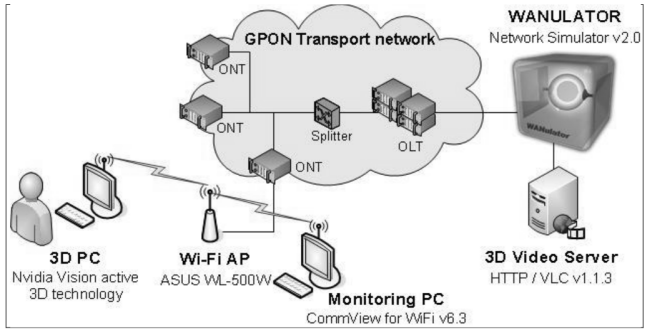


Figure 2. Network topology of measurement M2

The one-minute-long multimedia contents of measurement M2 (see *TABLE IV*) were streamed from a video server and delivered through a GPON network [35], which was accessed from client side via Wi-Fi (see *Figure 2*). While the simulation of varying network parameters was performed by WANulator on a separate computer, the intensity of transmission power was adjusted on the Wi-Fi

AP. The complete duration of a measurement required approximately 20-25 minutes, depending of the assessment speed of the participant.

TABLE III. QOS PARAMETER MATRIX VARIABLE VALUES OF M2

Test case	Varying parameters				
	Security	*TX Power*	*Jitter*	*Packet loss*	*Bandwidth limitation*
1	NO	71 mW	30 ms	0 %	NO
2	NO	71 mW	0 ms	1 %	NO
3	NO	71 mW	60 ms	1 %	NO
4	NO	71 mW	30 ms	1 %	NO
5	NO	71 mW	60 ms	0 %	NO
6	NO	71 mW	30 ms	0 %	YES
7	NO	71 mW	60 ms	2 %	NO
8	NO	71 mW	0 ms	2 %	NO
9	NO	71 mW	0 ms	0 %	NO
10	NO	71 mW	30 ms	2 %	NO
11	NO	71 mW	60 ms	0 %	YES
12	NO	71 mW	0 ms	0 %	YES
13	YES	71 mW	0 ms	0 %	NO
14	YES	71 mW	30 ms	1 %	NO
15	YES	71 mW	60 ms	2 %	NO
16	YES	71 mW	0 ms	0 %	YES
17	NO	35 mW	0 ms	0 %	NO
18	NO	35 mW	30 ms	1 %	NO
19	NO	251 mW	0 ms	0 %	NO
20	NO	251 mW	30 ms	1 %	NO

TABLE IV. DESCRIPTORS OF MULTIMEDIA CONTENT IN M2

Resolution: 3360x1050
Video codec: MPEG4
Audio codec: MP3

A total of 90 test subjects participated in M2. Similarly to M1, LoC was measured, but only in case of 34 participants. The rest performed so-called blind tests; they did not possess any direct information regarding the differentiation of test cases. The scale of evaluation was a 10-point quantitative discrete scale in this case as well, however, the highest score on the scale carried a slightly different interpretation. While in M1 score 10 was defined as the highest value that can be used for the evaluation of perceived quality, in case of M2 it represented the quality of the reference test case. Although the first test case of M1 can be deemed to be a reference of assessment, since participants were not informed explicitly about its nature, it cannot be considered to be a full-reference subjective QoE measurement, unlike M2. However, M2 also included the reference quality as a subject of evaluation, namely test case 9.

IV. MEASUREMENT RESULTS

A. Results of measurement M1

Before taking the different LoC levels into consideration, we took a look at the MOS results of M1 (see *Figure 3*). The first thing that grabbed our attention was that test case 8 with its additional 800 ms delay managed to achieve better video scores than the reference test case.

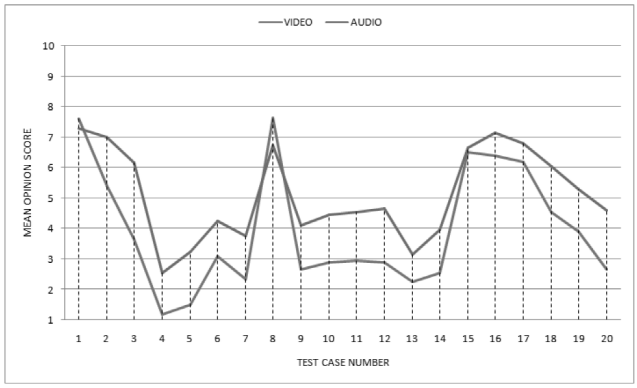


Figure 3. Mean Opinion Score of M1

Although this simulated network impairment rather had its effect on audio quality, it was still perceptible in video quality as well; however, the difference was barely noticeable. How was it possible that a test case with a minor degradation in quality received a higher score than the reference test case? By relying only on the MOS results, it would be quite exigent to give an accurate explanation to this phenomenon. After performing the LoC separation (7 participants in level +1 and –1, 6 participants in level 0) of the results and viewing the recorded video footages, the answer became clear (see *Figure 4*).

While the test participants of LoC level +1 and 0 were commonly controlled by the fact that delay is noxious to experienced quality and thus such measurement case cannot achieve a better score, members of level –1 were not aware of this. In fact, as heard on the recorded conversations, some of them were quite convinced that delay is beneficial and produces a higher level of quality. The other subjects were not affected by such misbelieves so not even a single participant gave test case 8 a better score. The devoted opinion of the evaluators in level –1 on the quality of these two cases was quite sufficient to create a distortion large enough to significantly alter the overall MOS results. It also needs to be noted that participants of level +1 indicated the difference in quality with more caution, even though their preconceptions were more reinforced by their technical knowledge and experience; many of them were more confident that they managed to detect the barely noticeable dissimilarities between the test cases, but they only distinguished them by a single unit on the measurement scale.

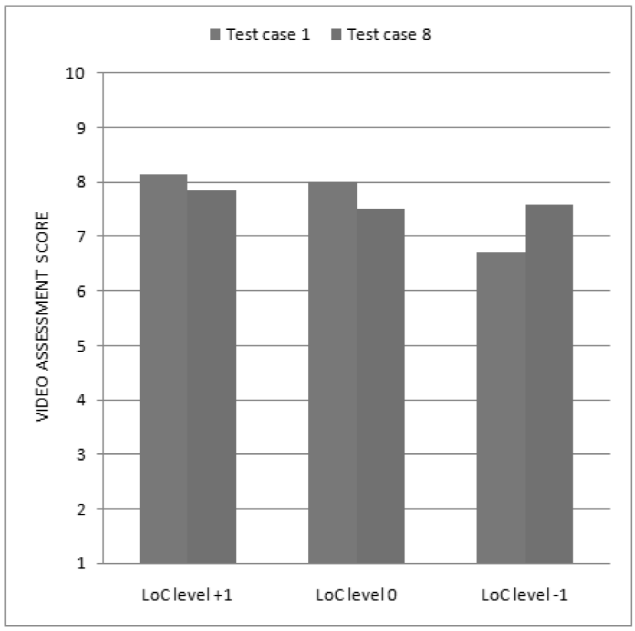


Figure 4. Video assessment scores of test case 1 and 8

Let us approach this issue from the angle of cognitive dissonance. On one hand, perception of many participants were not able to identify evident distinction between the two test cases, while on the other hand, preconception contained a clear direction of the difference. This dissonant state of cognitions was solved by either the alteration of perception ("I can clearly see the difference") or the reconsideration of its correctness ("I cannot clearly see the difference but I know it has to be there"). Those in LoC level –1 who supported the preconception of a beneficial delay in the aspect of video quality were fuelled by post-decision dissonance during the evaluation of test case 9 to 12. In these four test cases delay was reduced while the other parameters remained the same. Again, there were only the slightest differences in video quality, yet they made a decreasing score pattern, since preconception was also aided by a prior decision.

Those who utilized test case 1 as reference quality were strictly bounded by the rule that no other test case could ever exceed its score. However, only two participants from LoC level +1 granted it the maximal 10 points. This is a natural behavior when using an evaluation scale. Participants did not wish to limit the expressive ability of their evaluations; by using the top or the bottom end of the scale – especially during an early test case – participants forfeit the chance to express their thoughts should a test case with even greater or lesser quality appear. However, this implies the sacrifice of evaluation space; such participants were limited to use a 9-point or smaller scale. This idea also resulted these participants forcefully gave lower scores to each and every test case, even when no evident difference was found, as detailed earlier.

On the other hand, those who were not bounded by this test case were able to rate other test cases higher than the first one. The series from test case 13 to 16 was the reduction of the amount of additional jitter (see *Figure 5*). There was an immense difference between test case 14 and 15; while the video image of test case 14 was barely recognizable, test case 15 provided an acceptable video quality. This dire change of quality motivated some participants of LoC level –1 to give high scores, higher than test case 1.

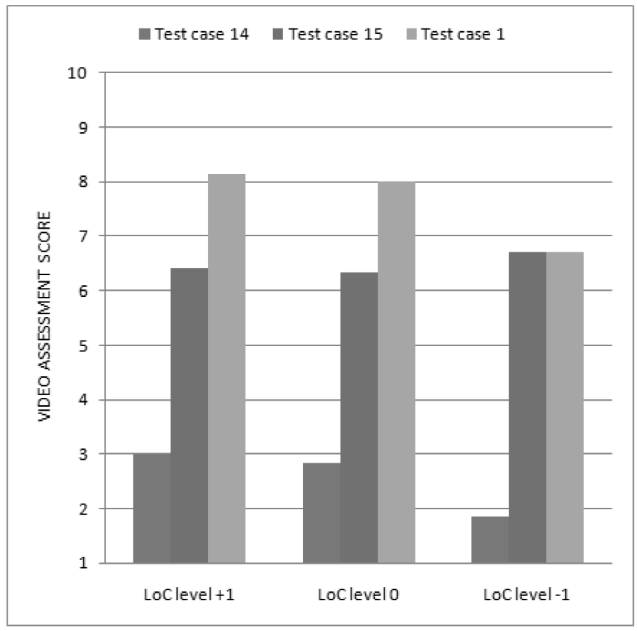


Figure 5. Video assessment scores of test case 14 and 15

The first four test cases represented a general decrement in QoS values; both delay, jitter and packet loss were increasing. In this case, it was quite interesting to see that the higher LoC level a participant had, the closer his/her evaluation was to uniformity (e.g., 10, 7, 4, 1) in both video and audio quality.

In audio quality evaluation scores, the progress from test case 9 to 12 was possibly the most interesting. These four test cases endured delay reduction while preserving a notable constant jitter. Presuming the experienced quality tendencies during these four test cases is not a trivial task. It was beneficial to have a smaller delay, however, the ratio of jitter and delay increased. The audio MOS shows a definite raise, even though none of the participants thought it that way. In level –1 and 0, there was no repeating behavior pattern. In fact, participants used a high variety of scoring patterns to assess, since there was no obvious difference in the overall experience of audio quality. On one hand, mutual speech interruptions were fewer, but on the other, audio quality was less enjoyable to some extent. The scores given by the participants were based on the personal decision whether the first or the second effect was more dominant. However, the audio assessments in LoC level +1 were shocking; 6 out of 7 participants used a constant evaluation pattern (see *Figure 6*). It means that preconceptions had such a high level impact on evaluation that these subjects ignored any lesser differences that they experienced between cases. They

considered the opposing effects nearly equal, which supposes an unvarying overall experience.

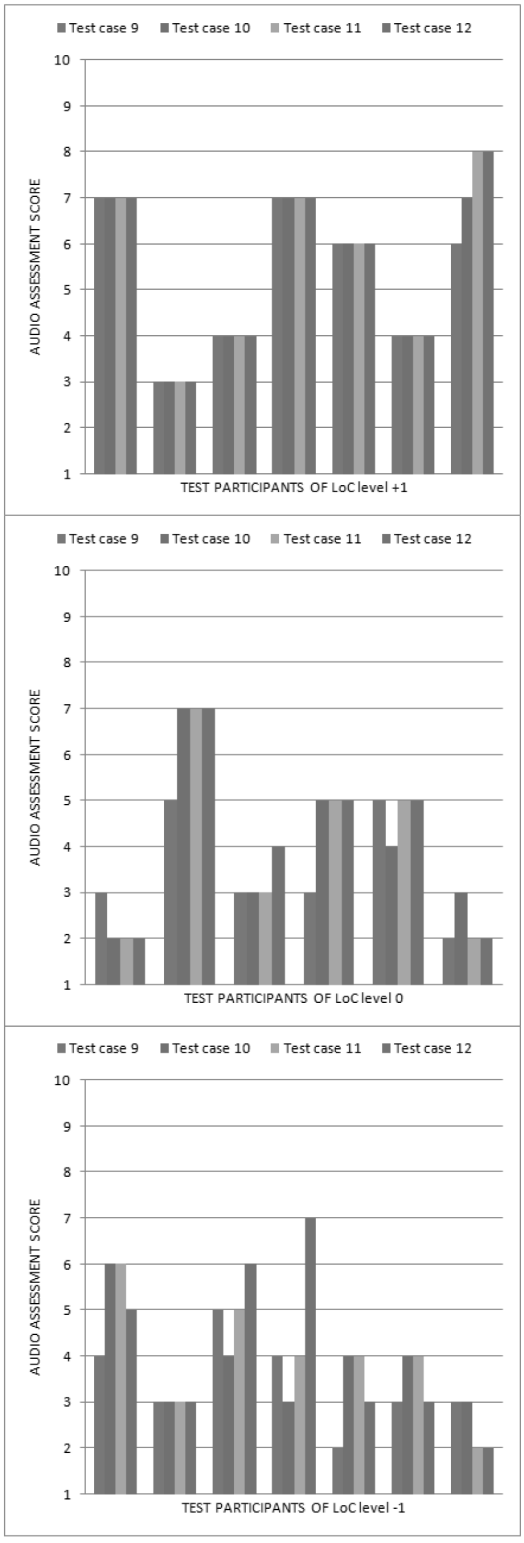


Figure 6. Audio assessment scores of test cases 9 to 12

A participant from LoC level –1 also used constant evaluation, but claimed that "jitter has no effect on audio quality".

It is also exciting to compare the audio results of test case 17 and 18. Due to the redundancy of the human voice, the given amount of packet loss caused no major difference between these two test cases. Many members of LoC level +1 and some of 0 indicated an apparent difference in scoring, since according to their preconceptions, audio quality should clearly lessen. However, there were participants in level –1 with the idea that packet loss is beneficial in the aspect of audio quality. This is also a great example for the disobedience of subjective prior cognition, since their scoring direction was inverted in the last two test cases. Even though preconception was supported by post-decision dissonance through a prior decision, the test participants had to abandon it when facing the obviously lessening sound quality of test case 19 and 20. Their assessment was so intense that it managed to make a clear impact on the audio MOS.

B. Results of measurement M2

This subsection mainly focuses on the assessment of the 34 participants with access to the alteration of the variables. The psychometric functions of the other results [36] are indeed also exciting, but this paper emphasizes more with the effects of direct environmental information. Moreover, this paper does not deal with the separated aspects of quality, but uses weighted averages, since participants were asked to weight these aspects based on personal importance with the sum of 10 (2 for each if all are equally important).

We approached the results of M2 from five directions: security presence, transmission power adjustment, jitter, packet loss and the limitation of bandwidth (see *TABLE V*). If we take a look at the MOS (see *Figure 7*), the tendencies of the evaluation results of the groups of participants with and without access to environmental information might seem to differ at some points.

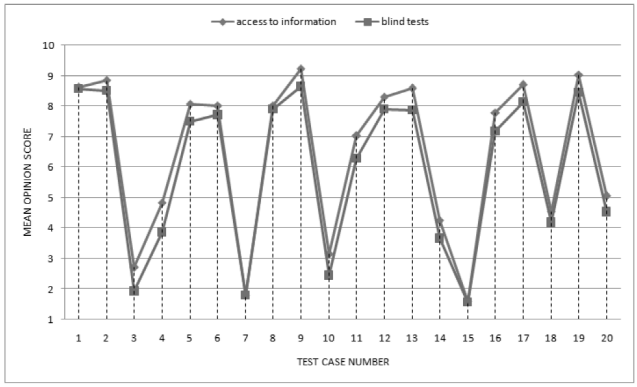


Figure 7. Mean Opinion Score of M2

If we just approach these data without any of the previously mentioned directions, the first thing we notice is that the mean assessment of those with QoS awareness is higher in scores. For instance, test case 9 – the hidden

TABLE V. QOE RESULTS OF THE INVESTIGATED ASPECTS IN M2

Test case	*Investigated aspect*	*MOS (blind tests)*	*MOS (with access)*
19	Transmission power	8.46	9.04
17		8.13	8.7
20	Transmission power	4.53	5.05
18		4.19	4.54
1	Bandwidth limitation	8.57	8.61
6		7.72	8.02
5	Bandwidth limitation	7.49	8.08
11		6.29	7.02
9	Bandwidth limitation	8.64	9.24
12		7.91	8.3
13	Bandwidth limitation	7.87	8.58
16		7.19	7.79
9	Security presence	8.64	9.24
13		7.87	8.58
4	Security presence	3.87	4.84
14		3.66	4.26
7	Security presence	1.79	1.8
15		1.57	1.6
12	Security presence	7.91	8.3
16		7.19	7.79
1	Jitter	8.57	8.61
5		7.49	8.08
6	Jitter	7.72	8.02
11		6.29	7.02
4	Jitter	3.87	4.84
3		1.93	2.71
10	Jitter	2.45	3.14
7		1.79	1.8
2	Packet loss	8.51	8.86
8		7.91	8.02
3	Packet loss	1.93	2.71
7		1.79	1.8
4	Packet loss	3.87	4.84
10		2.45	3.14

reference test case – achieved better evaluation results due to the fact that participants were aware that it was without any additional load.

The scores of test case 12 and 13 are also quite interesting; they both had a standard transmission power of 71 mW and no additional jitter or packet loss, but while test case 12 had limited bandwidth, test case 13 utilized secure transmission. The very similar situation can be witnessed in the relationship of test case 5 and 6. The parameters of bandwidth limitation on WANulator were chosen to imply a barely noticeable difference in quality. However, the parameter matrix only included this in a binary way, without any exact value. Preconceptions regarding bandwidth limitation were quite amplified, usually regardless of LoC level, since the majority depicted the word "limitation" as something harmful to quality, which it actually is.

In M2, the adjustment of additional jitter and packet loss had a rather evident effect on the experienced quality of the 3D stream transmission. Thus any prior idea of beneficial jitter or packet loss was nullified by perception.

Preconceptions regarding transmission power were a bit more diverse. There were quite some participants who approached the alteration of transmission power somewhat similar to sound volume, where too high is just as adverse as too low.

The information regarding these previous aspects was commonly used in the same way during assessment, apart from a few participants, whose evaluation scores did not affect the mean QoE results of their LoC levels. Similarly to M1, 3 different levels of LoC were distinguished (11 participants in level +1 and –1, 12 participants in level 0). The most interesting results appeared when we viewed the security presence aspect scores of LoC separation.

Although the mean results of those with access to the QoS parameters were similar to the others in scoring relations, it was revealed that there were many participants in LoC level –1 and some in level 0 with a steady preconception stating that secure transmission has to have better performance. It was so influential that it managed to become visible in the mean scores of level –1 (see *Figure 9, 10 and 11*).

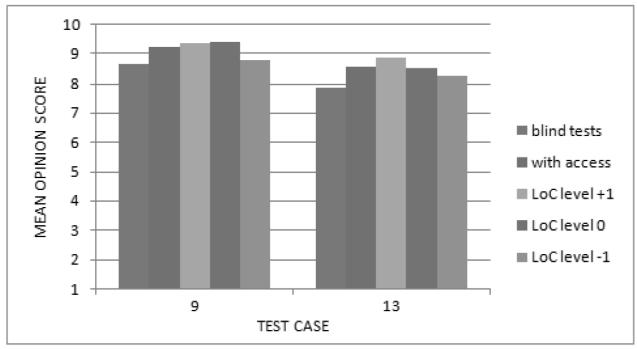


Figure 8. Mean Opinion Score of test case 9 and 13

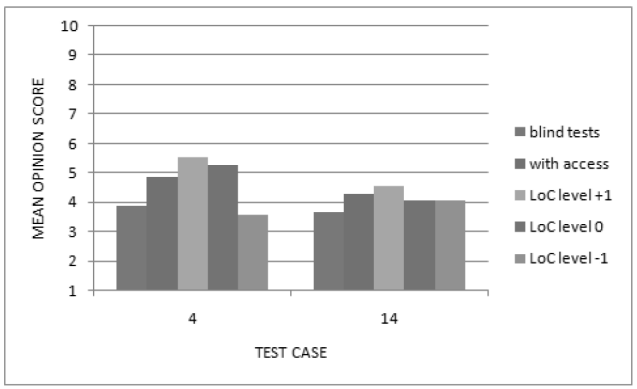


Figure 9. Mean Opinion Score of test case 4 and 14

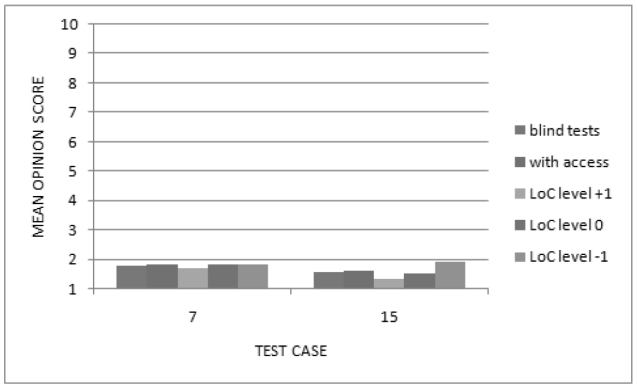


Figure 10. Mean Opinion Score of test case 7 and 15

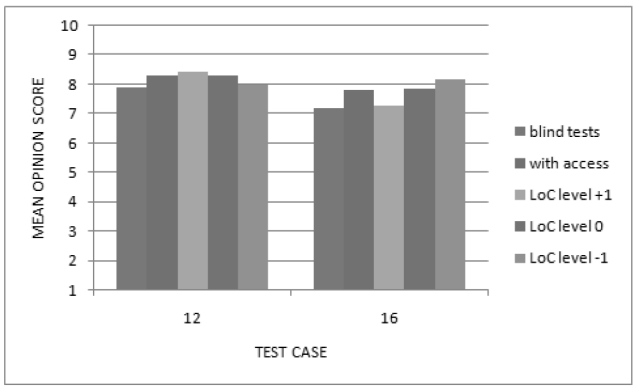


Figure 11. Mean Opinion Score of test case 12 and 16

Even though this phenomenon did not appear between the mean scores of test case 9 and 13 (see *Figure 8*), 3 out of 11 participants already supported that preconception; this number in case of test case 4 and 14 (see *Figure 9*) was 7 out of 11. Of course on the other hand, several participants belonging to LoC level +1 were confident that test cases with secure transmission had to be worse due to technical reasons. The difference of the experienced quality in practice was almost ignorable, which enabled preconception to dominate perception through cognitive dissonance.

V. CONCLUSION

The paper presented correlations between assessment alteration and the Level of Comprehension of test participants and detailed the socio-psychological background of the phenomenon. Environmental information regarding the given service can be considered the actual hotbed of preconceptions. Its relevancy is supported by the single fact that the majority of evaluation measurements cannot be considered to be so-called blind tests due to their configurations. The presented measurement utilized a radical amount and type of information, usually not public during service assessment and everyday service usage. However, in many cases basic information – like the type of connection – is very hard or impossible to hide.

Our results have demonstrated that any interpretation of environmental information can be influential and that misinterpretations usually occur on lower levels of LoC, given that perception can be suppressed by preconception. The outcome of the measurements also show that such influences can visibly affect the Mean Opinion Score, but of course distortion effects originating from preconceptions are best seen in separated mean values. Intense analysis of repeating assessment patterns and evaluator behavior shall be detailed in future papers of upcoming measurements.

Currently our researches deal with hard-to-hide environmental information, which are naturally present to evaluators. In the upcoming measurements, the methods for LoC determination will be simplified, however, at this initial phase of the research series we couldn't risk to lose any level of accuracy. Our future goals also contain the exhaustive analysis and comparison of automated and human assessment of quality, since objective solutions are invulnerable to the distortions presented in this paper, however, they have the flaws of their own.

ACKNOWLEDGEMENT

The research leading to these results has received funding from the European Union's Seventh Framework Programme ([FP7/2007-2013]) under grant agreement n° 288502 (CONCERTO project). This work was also supported by the Mobile Innovation Centre Hungary (MIK). We are grateful to the Department of Networked Systems and Services (HIT) and to the Department of Telecommunications and Media Informatics (TMIT) of the Budapest University of Technology and Economics (BME). We would also like to thank Ivett Kulik for her help and cooperation. Last but not least we would like to thank the many individuals whose work made this research possible.

REFERENCES

[1] Qualinet White Paper on Definitions of Quality of Experience. March 2013. http://www.qualinet.eu/images/stories/QoE_whitepaper_v1.2.pdf (retrieved January 2014)

[2] ITU-T Rec. E.800, Definitions of terms related to quality of service. Int. Telecomm. Union, Geneva, September 2008.

[3] M. Fiedler, T. Hossfeld, P. Tran-Gia. A generic quantitative relationship between quality of experience and quality of service. Network, IEEE , vol.24, no.2, pp.36–41, March-April, 2010.

[4] International Telecommunication Union. Methods for subjective determination of transmission quality. ITU Recommendation P.800 (08/96), August 1996.

[5] P.A. Kara, L. Bokor, S. Imre. Distortions in QoE measurements of ubiquitous mobile video services caused by the preconceptions of test subjects. IEEE/IPSJ International Symposium on Applications and the Internet SAINT2012. Izmir, Turkey, July 2012. pp. 409–413.

[6] L. Festinger. A theory of cognitive dissonance. Stanford, CA: Stanford University Press, 1957.

[7] International Telecommunication Union. P series: Terminals and subjective and objective assessment methods. ITU-T Recommendations, P series. http://www.itu.int/rec/T-REC-P/en (retrieved January 2014).

[8] I. Ketykó, K. De Moor, W. Joseph, L. Martens, L. De Marez. Performing QoE-measurements in an actual 3G network. In IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB 10), pp. 1-6, March 2010.

[9] G. Exarchakos, L. Druda, V. Menkovski, P. Bellavista, A. Liotta. Skype resilience to high motion videos. International Journal of Wavelets, Multiresolution and Information Processing, Vol.11(3), 2013, World Scientific Publishing.

[10] C.T.E.R. Hewage, M. G. Martini. Quality of experience for 3D video streaming. Communications Magazine, Volume 51, Issue 5, pp.101–107, May 2013.

[11] F. Agboma, A. Liotta. Quality of Experience Management in Mobile Content Delivery Systems, Journal of Telecommunication Systems, special issue on the Quality of Experience issues in Multimedia Provision. Vol. 49(1), pp. 85-98, Springer 2012.

[12] I. Kulik, P.A. Kara, T.A. Trinh, L. Bokor. Analysis of the Relationship between Quality of Experience and Service Attributes for 3D Future Internet Multimedia. IEEE 4th International Conference on Cognitive Infocommunications, Budapest, Hungary, 2-5 Dec. 2013, pp. 641–646.

[13] P. Brooks, B. Hestnes. User Measures of Quality of Experience: Why Being Objective and Quantitative Is Important. Network, IEEE, Vol. 24, No. 2. (March 2010), pp. 8–13.

[14] B.L. Jones, P.R. McManus. Graphic scaling of qualitative terms. SMPTE Journal, November 1986, pp. 1166–1171.

[15] N. Narita. Graphic scaling and validity of Japanese descriptive terms used in subjective-evaluation tests. SMPTE Journal, July 1993, pp. 616–622.

[16] A. Watson, M.A. Sasse. Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications. In MULTIMEDIA '98: Proceedings of the sixth ACM international conference on Multimedia (September 1998), pp. 55–60.

[17] V. Menkovski, G. Exarchakos, A. Liotta. The Value of Relative Quality in Video Delivery. Journal of Mobile Multimedia, Vol.7(3), pp. 151-162. Rinton Press, September 2011.

[18] V. Menkovski, A. Liotta. Adaptive Psychometric Scaling for Video Quality Assessment. Journal of Signal Processing: Image Communication. Vol.26(8), pp.788–799. Elsevier. 2012.

[19] C. Charrier, L.T. Maloney, H. Cherifi, K. Knoblauch. Maximum likelihood difference scaling of image quality in compression-degraded images. Journal of the Optical Society of America A24 (11), 2007, pp. 3418–3426.

[20] A.B. Watson. Proposal: measurement of a JND scale for video quality. IEEEG-2.1.6 Subcommittee on Video Compression Measurements, 2000.

[21] R.E. Knox, J.A. Inkster. Postdecision dissonance at post time. Journal of Personality and Social Psychology, 1968, pp. 319–323.

[22] A. Sackl, P. Zwickl, S. Egger, P. Reichl. The role of cognitive dissonance for QoE evaluation of multimedia services. 2012 IEEE Globecom Workshops (GC Wkshps), pp. 1352–1356.

[23] M. O'Neill, A. Palmer. Cognitive dissonance and the stability of service quality perceptions. Journal of Services Marketing, 2004, Volume 18, Issue 6, pp. 433-449.

[24] M. O'Neill, A. Palmer. Exploring the relationship between post-consumption dissonance and time-elapsed perceptions of service quality. Anzmac conference, New Zealand, 2001.

[25] M.R. Quintero, A. Raake. Is taking into account the subjects degree of knowledge and expertise enough when rating quality? QoMEX 2012: 194-199.

[26] BME-MIK. Budapest University of Technology and Economics - Mobile Innovation Centre, Official Website. https://www.mik.bme.hu/home/aboutus/, (retrieved January 2014).

[27] T. Hossfeld, R. Schatz, S. Egger. SOS: The MOS is not enough! Third International Workshop on Quality of Multimedia Experience (QoMEX), 2011, pp. 131–136.

[28] Linphone. Official Website. http://www.linphone.org/, (retrieved 2012 May).

[29] 3GPP TS 23.228. IP Multimedia Subsystem (IMS); Stage 2. Rel-8, 2008.

[30] T. Hossfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, M. Fiedler. The memory effect and its implications on Web QoE modeling. 23rd International Teletraffic Congress (ITC), 2011, pp. 103–110.

[31] P. Froehlich, S. Egger, M. Schatz, R. Muehlegger, K. Masuch, B. Gardlo. QoE in 10 Seconds: Are Short Video Clip Lengths Sufficient for Quality of Experience Assessment? Proceedings of the fourth International Workshop on Quality of Multimedia Experience QoMEX, 2012.

[32] E.B. Goldstein. Sensation and Perception, Eighth Edition. Cengage Learning, February 2009.

[33] netem. Linux Network Emulation Official Website. http://www.linuxfoundation.org/collaborate/workgroups/networking/netem, (retrieved January 2014).

[34] NvidiaVision Player website http://www.nvidia.com/object/3d-vision-video-player-1.7.5-driver.html (retrieved January 2014).

[35] I. Kulik, T.A. Trinh. Investigation of Quality of Experience for 3D Streams in GPON. Ralf Lehnert (Ed.) EUNICE 2011. LNCS, vol. 6955, pp.157 – 168 Springer, Heidelberg, 2011.

[36] I. Kulik, P.A. Kara, T.A. Trinh, L. Bokor. Attributes unmasked: Investigation of service aspects in subjective evaluation of wireless 3D multimedia. IEEE/ICIA Second International Conference of Informatics, Lodz, Poland, 23-25 Sept. 2013, IEEE, pp.270–275.

AUTHORS

László Bokor graduated in 2004 with M.Sc. degree in computer engineering from the Budapest University of Technology and Economics (BME) at the Department of Networked Systems and Services. In 2006 he got an M.Sc.+ degree in bank informatics from the same university's Faculty of Economic and Social Sciences. He is a Ph.D. candidate at BME, member of the IEEE, member of Multimedia Networks and Services Laboratory and Mobile Innovation Centre of BME where he participates in researches of wireless protocols and works on advanced mobility management related projects (e.g., FP6-IST PHOENIX and ANEMONE, EUREKA-Celtic BOSS, FP7-ICT OPTIMIX and CONCERTO, EURESCOM P1857, EUREKA-Celtic MEVICO and SIGMONA). His research interests include IPv6 mobility, mobile computing, next generation networks, mobile broadband networking architectures, network performance analyzing, and heterogeneous networks.

Péter András Kara received his B.Sc degree in 2011 and his M.Sc degree in 2013, both in Computer Engineering at the Budapest University of Technology and Economics (BME) at the Department of Networked Systems and Services (HIT). In 2013, he joined the Ph.D. doctorate course in Informatics Sciences of the same university. He is a member of the IEEE and participates in FP7-ICT CONCERTO. His research interest includes network performance analysis, quality of telecommunications services, quality assurance and management, and the assessment of subjective and objective quality.

Sándor Imre was born in Budapest in 1969. He received the M.Sc. degree in Electronic Engineering from the Budapest University of Technology (BME) in 1993. Next he started his Ph. D. studies at BME and obtained dr. univ. degree in 1996, Ph.D. degree in 1999 and DSc degree in 2007. Currently he is carrying his teaching activities as Head of the Dept. of Telecommunications of BME. He was invited to join the Mobile Innovation Centre of BME as R&D director in 2005. His research interest includes mobile and wireless systems, quantum computing and communications. Especially he has contributions on different wireless access technologies, mobility protocols and reconfigurable systems.

CALL FOR PAPERS
Special Issue on the Future Internet

Recent dramatic changes such as the rising number of Internet users, the penetration of portable and mobile devices, or the Internet of Things, have motivated a number of research initiatives, labeled "Future Internet" worldwide, supported by NSF in the USA and EU research framework programs in Europe. In Hungary, the ongoing "Future Internet Research, Services and Technology – FIRST" project, supported by the European Social Fund focuses on key theoretical, modeling, planning, application and experimental aspects of Future Internet.

Our journal is calling for original and unpublished contributions to this important area that will be peer-reviewed. Selected papers will appear in a two-part Special Issue to be published in September and December of 2014. Original and unpublished papers should be submitted by 30th of July, and by 30th of September in the form of pdf files in IEEE format according to the formatting instructions available at *http://www.ieee.org/publications_standards/publications/authors/authors_journals.html#sect2*

Contributions are expected from the following areas:

- The science of the Internet (basic research issues)
- Modeling, analysis and network design
- Network architectures for the Future Internet
- 3D Internet
- Internet of Things
- Cyber-physical systems and their applications
- Data and content technologies
- Cognitive infocommunications
- Community applications
- Standardization and regulatory issues
- Experimental systems
- Internet economics

Papers from the FIRST research community are particularly welcome, but the call is fully open to all authors wishing to share their research results in one of the fields listed above.

Guest Editors:

GYULA SALLAI received MSc degree from the Budapest University of Technology and Economics (BME), PhD and DSc degrees from the Hungarian Academy of Sciences (MTA), all in telecommunications. He was senior researcher in telecommunication network planning, then research director, strategic director, later deputy CEO with the Hungarian Telecom Company; then international vice president, after that executive vice president for the ICT regulation with the Communication Authority of Hungary. From 2002 to 2010 he was the head of the Department of Telecommunications and Media Informatics of the BME, and from 2004 to 2008 the vice-rector of the BME as well. From 2005 to 2011 he was also the chairman of the Telecommunication Committee of the MTA and the president of the Hungarian Scientific Association for Infocommunications (HTE). Recently he is full-professor at the BME, Scientific Director of Future Internet Research Coordination Centre, member of the FIRST Project Council and honorary president of the HTE. His main fields of interest are the ICT trends, strategic, management and regulatory issues, Future Internet engineering.

WOLFGANG SCHREINER is since 2004 associate professor of the Research Institute for Symbolic Computation (RISC) at the Johannes Kepler University Linz, Austria. His research areas are formal methods and parallel and distributed computing; he has directed in these areas various national and international research and development projects, participated in the program committees of 90 conferences, served as evaluator for various European projects, and is member of the editorial board of the Journal of Universal Computer Science. Prof. Schreiner has (co-)authored 13 book chapters, 9 journal publications, 46 other refereed publications, 12 non-refereed publications and 70 technical reports.

.

JÁNOS SZTRIK is a Full Professor at the Faculty of Informatics and Head of Department of Informatics Systems and Networks, University of Debrecen, Debrecen, Hungary. He received the M.Sc. degree in 1978, the Ph.D in 1980 both in probability theory and mathematical statistics from the University of Debrecen. He obtained the Candidate of Mathematical Sciences degree in probability theory and mathematical statistics in 1989 from the Kiev State University, Kiev, USSR, habilitation from University of Debrecen in 1999, Doctor of the Hungarian Academy of Sciences, Budapest, 2002. His research interests are in the field of production systems modeling and analysis, queueing theory, reliability theory, and computer science.

Ontology Evaluation with Protégé using OWLET

Thomas J. Lampoltshammer and Thomas Heistracher

***Abstract*—Amalgamation of formalised knowledge and real-world datasets is a pivotal challenge in the realm of information and communication technologies. Semi-automated classification of datasets can be performed by utilisation of ontologies. The detection process of image objects in Very High Resolution Satellite Imagery (VHRSI) gives a prominent example. The process of refinement of formalised expert knowledge within the related ontology still remains a challenging and time-consuming task. In this paper, the *JSON2OWL Converter (OWLET)* extension for Protégé is presented which supports experts during this refinement phase. The extension offers an integrated approach to transfer real-world dataset objects into the ontology modelling software for semi-automated classification. This transfer is achieved by combining open standard formats from both domains, the (Geo) Web domain (GeoJSON) and the Web ontology domain (OWL2). Thereby *OWLET* supports the process of accuracy analysis and accuracy fostering. By utilising the OWLET extension, experts can not only speed up their classification procedure considerably, but they can also refine their formalised knowledge by using the results of the classification process in conjunction with the outcomes of the accuracy analysis.**

***Index Terms*—OWL, Ontology, Protégé, Remote Sensing, GIS, Knowledge Formalisation**

I. INTRODUCTION

AS the formalisation of expert knowledge is required in almost all research domains, ontologies represent one solution towards this issue. The challenge associated to this procedure is twofold: on the hand the formalised knowledge has to be applied to the real-world datasets, on the other hand the real- dataset should also enhance and develop the formal knowledge. By serving both sides, semi-automated classification of data becomes possible.

For instance, Object Based Image Analysis (OBIA) presents an accepted and efficient method concerning the classification of high-resolution imagery datasets [1]. The main idea of this approach is to segment the original image into homogeneous units, based on pre-define criteria. In the next step, the image analysis process builds on these segmented objects [2]. Up to now, OBIA is expert knowledge-driven. In consequence, the accuracy of the classification and the time consumption for this process strongly depend on the experts‘ knowledge about the properties of objects represented in the image. Furthermore, he or she has to consider the relations between the objects as well as the context of the image. These circumstances exacerbate the process of embedding OBIA-based methods into operational frameworks, where speed and flexibility of information retrieval are important assets [3], [4]. To overcome this issue, a formalisation of *a priori* knowledge is necessary for the purpose of information extraction from satellite imagery by image analysis systems. One possible solution is presented in form of ontologies. Ontologies can be defined within the Artificial Intelligence (AI) domain as ‘an explicit specification of a shared conceptualisation’ [5, p. 1]. In other words: An ontology can be seen as the knowledge of a domain expert, formalised in a machine-understandable way. Various research endeavours have been conducted to employ ontologies for the automation of the image analysis and interpretation process [6], [7], [8], [9]. For a detailed review of ontology-based applications in remote sensing, please refer to [10]. In summary, it can be argued that the process of image analysis and image interpretation necessitates two kinds of knowledge: i) domain-specific knowledge, and ii) knowledge for image analysis [11]. The domain-specific knowledge represents the domain-specific terminology and incorporated semantics. The image interpretation knowledge can be separated into *qualitative information* and *quantitative information*. The first refers to spectral and spatial properties of objects inherent in e.g. satellite imagery. These properties can be characterised in natural language: for instance, rivers feature an elongated form, or buildings are represented by rectangular objects when seen from above. These qualitative descriptions need then to be mapped to information from the source image, in particular, the delineated objects. One of the most challenging steps presented by the associated engineering process is to form a knowledge base which features the necessary domain-specific semantics at a meaningful granularity. This issues is known in the literature as *‘the ontology grounding problem’* [12]. To contribute towards a possible solution to this problem, the JSON2OWL Converter (OWLET) presented in this paper aims at supporting domain experts during the development and refinement phase of their ontologies based on real-world data as fundament. This process bridges the ontology domain and the domain of real-world applications.

For the development of before-mentioned ontologies, the ontology creation tool Protégé [13] represents the *de facto* standard in the remote sensing community and beyond. This tool is not an expert system itself. Instead, it is intended to provide the necessary environment to develop custom-tailored tools for the process of knowledge-acquisition. The current version of Protégé is Java-based and therefore platform-independent. Furthermore, it is possible to extend the Protégé environment by plug-ins such as ontology visualisation [14] or fuzzy logics [15], [16]. Therefore, OWLET was realised as such a plugin[1].

Thomas J. Lampoltshammer is with the Department of Geoinformatics - Z_GIS; University of Salzburg; Hellbrunnerstrasse 34; 5020 Salzburg; Austria
Thomas Heistracher and Thomas J. Lampoltshammer are with the School of Information Technology and Systems Management; Salzburg University of Applied Sciences; Urstein Süd 1; 5412 Puch/Salzburg; Austria

[1] A demo version of the plugin, together with sample data and a sample ontology for testing purposes can be found at: http://lampoltshammer.com/owlet/demo.zip

The remainder of the paper is as follows: First, the overall architecture of OWLET is presented, together with the associated data flow and workflow. Second, the transformation process of geo data to the ontology modelling language is described. Subsequently, the presented plugin is demonstrated via an example workflow. A discussion about important aspects of the suggested solution and the conclusion end this paper.

II. Ontology evaluation

Throughout the literature, various approaches exist to evaluate given ontologies. According to [17], four main evaluation directions can be identified: i) 'gold standard' comparison, ii) application-based evaluation, iii) data source comparison, and iv) human-centric evaluation. The first category is dedicated to compare the given ontology to a 'gold standard', that is a predefined, well-formed dataset against other datasets are measured [18]. The second category describes the use of an ontology within an application and then to evaluate the outcome based on the employed ontology [19]. The third category utilises a repository of documents about a certain domain, which is then compared to the ontology that should cover this domain and the associated knowledge [20]. The last category describes a human-centric approach. Here, experts assess the quality of the given ontology by comparing it to a defined set of criteria [21].

The example workflow described in this paper is based on object features from literature, in particular building features. This qualitative and quantitative knowledge is then employed to manually build a 'gold standard', against which the ontology-based classification results are matched.

III. The integrated system architecture and process flow

Figure 1 depicts the integrated system architecture as well as the associated process flow. The rhomboid boxes represent data for input/output, while the rectangular-shaped boxes depict processing modules. The arrows within the figure visualise the process flow within the architecture. The architecture itself comprises several layers from bottom to top of Fig. 1. : i) the data layer, ii) the image processing layer, iii) the reasoning layer, and iv) the expert knowledge layer. The first layer serves as a repository for Very High Resolution Satellite Imagery (VHRSI). From this database, the image to be analysed is handled by the image processing layer. This layer is represented by a remote sensing application for image processing and analysis, e.g. eCognition[2]. Via this object-based image analysis tool, image segmentation algorithms are applied to the satellite image. The process of segmentation can be described as the partitioning of an image into distinct areas or regions. These regions are non-overlapping and are homogeneous considering certain predefined attributes [22]. The resulting delineated objects are then exported into the GeoJSON format [23]. The export functionality is either included in the remote sensing application or can be achieved via the use of external services

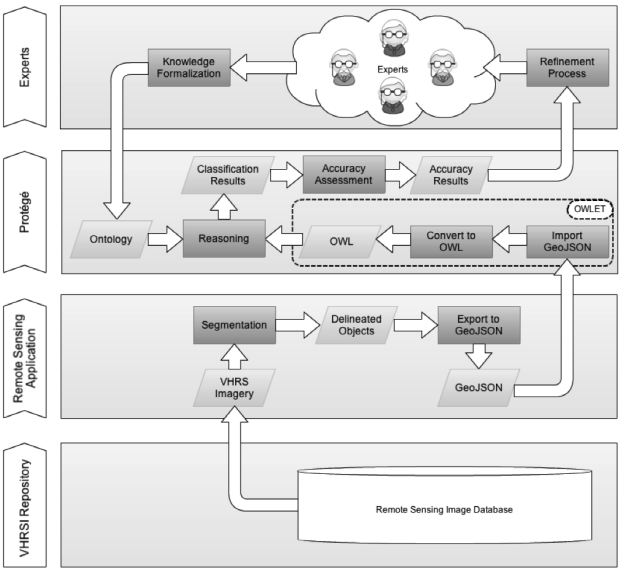


Fig. 1: Integrated architecture of the OWLET plugin

or libraries, such as *WorldMap*[3] or the *Geospatial Data Abstraction Library (GDAL)*[4]. The resulting GeoJSON file is then imported into Protégé via the *OWLET* plugin which is denoted by dashed lines within the processing modules in Fig. 1. In the next step, the imported objects (delineated objects from the prior segmentation) are translated into the Ontology Web Language Version 2 (OWL2) [24]. Subsequently, the ontology and the objects modelled in OWL (so-called Individuals) are now merged for the reasoning process. The ontology itself comprises the qualitative class description of the domain, as well as the quantitative descriptions based on the actual image. The classification results can then be used to perform an accuracy assessment based on *precision* and *recall* [25]. In the final step, experts can analyse characteristics of misclassified objects to refine the qualitative and quantitative knowledge modelled within the ontology, until satisfying accuracy results are achieved.

IV. Geo data transformation

In order to validate and evaluate the developed ontology, the segmented objects from the satellite image have to be imported into Protégé. As Protégé does not 'understand' shape files, the segmented objects have to be exported and converted in order to be compatible. In the first step, the given shape file has to be exported into the GeoJSON format [23]. GeoJSON provides encoding capabilities for various geographic data structures, such as geometry, a feature, or a collection of features. After the export is completed, the newly produced GeoJSON file can be imported into Protégé via *OWLET*.

For the process of reasoning (in this case the process of classification based on formalised knowledge), Protégé employs description logics in form of various reasoners. These reasoners make use of the specifically designed elements of

[2] eCognition - http://www.ecognition.com

[3] WorldMap - http://worldmap.harvard.edu/

[4] GDAL - http:www.gdal.org/

the Ontology Web Language Version 2 (OWL) by W3C [24]. For a comprehensive treatise of the origin of OWL2, the reader may refer to [26]. The OWL vocabulary contains three main artefacts as there are classes, individuals and, properties. Classes as such can be described as set of individuals, while properties describe the relationships between the classes and in consequence the associated individuals. These artefacts are formalised within OWL based on Description Logics (DL) – here called OWL DL. By these logic statements, automated testing by a reasoner becomes possible. A reasoner can be described as application the can infer logical relationships within the ontology and in consequence can perform consistency, equivalence and instantiation testing. There exist three profiles within the OWL2 standard - also called fragment or sublanguage - available. These sublanguages represent altered versions of OWL, namely: i) OWL 2 EL, ii) OWL 2 QL, and iii) OWL 2 RL (Motik et al., 2009). For the ontology in this research work, the author chose Protégé-OWL [27], an adapted and optimised version for Protégé.

This part of the paper introduces the main components utilised to build a basic ontology in OWL. First, a class hierarchy is built by classes and sub-classes. Listing 1 describes the class 'Area' to be a sub-class of the parent class 'Classifiers'.

```
<Declaration>
        <Class IRI="#Area"/>
</Declaration>
<Declaration>
        <Class IRI="#Classifiers"/>
</Declaration>
<SubClassOf>
        <Class IRI="#Area"/>
        <Class IRI="#Classifiers"/>
</SubClassOf>
```

Listing 1: Defining OWL classes and sub-classes

In order to link two classes with each other, object properties are employed. In addition, these relationships can be employed to describe entire classes - called 'EquivalentClasses'. Listing 2 demonstrates how such properties and 'EquivalentClasses' are described in OWL. In particular, it is defined that a class 'ResidentialAres' is equivalent to a class that is linked to the class 'LowArea' via the object property 'hasArea'.

```
<EquivalentClasses>
  <Class IRI="#ResidentialArea"/>
  <ObjectIntersectionOf>
    <ObjectSomeValuesFrom>
         <ObjectProperty IRI="#hasArea"/>
          <Class IRI="#LowArea"/>
    </ObjectSomeValuesFrom>
  </ObjectIntersectionOf>
</EquivalentClasses>
```

Listing 2: Defining OWL EquivalenClasses by ObjectProperties and Relationships

The next step consists of mapping these qualitative descriptions to quantitative values. Again, the principle of 'EquivalentClasses' is used. In addition, data properties are utilised to map concrete values to qualitative descriptions (see List. 3). Here, the definition of 'LowArea' is described as a double value of greater than 1,500. The property which holds this value is denoted as 'area_pxl'.

```
<EquivalentClasses>
      <Class IRI="#LowArea"/>
      <DataSomeValuesFrom>
          <DataProperty IRI="#area_pxl"/>
          <DatatypeRestriction>
              <Datatype abbreviatedIRI="xsd:double"/>
              <FacetRestriction facet="&xsd;maxExclusive">
                   Literal datatypeIRI="&xsd;double
                        ">1500.0</Literal>
              </FacetRestriction>
          </DatatypeRestriction>
      </DataSomeValuesFrom>
</EquivalentClasses>
```

Listing 3: Mapping of qualitative and quantitative knowledge

What the OWLET plugin does is to parse the objects denoted in the GeoJSON file and ports them into the OWL syntax to be included into the ontology. Listing 4 shows an example ontology entry for a parsed object as an individual.

```
<Declaration>
    <NamedIndividual IRI="#DataSet_building_set.2"/>
</Declaration>

<DataPropertyAssertion>
    <DataProperty IRI="#area_pxl"/>
    <NamedIndividual IRI="#DataSet_building_set.2"/>
    <Literal datatypeIRI="&xsd;double">1230.0</Literal>
</DataPropertyAssertion>
```

Listing 4: Defining OWL classes and sub-classes

This process is repeated for all properties of one object and for all objects included within the GeoJSON file.

V. Application Example

The following example describes a typical application scenario for OWLET. The example at hand is simplified for demonstration purposes. The employed quantitative and qualitative descriptions within the ontology are by no means representative. However, the described workflow can easily be extended to cover complex knowledge acquisition projects as well. For a more complex example dedicated to classification of buildings from Light detection and ranging (LiDAR)-based data, involving an early prototypical version of this plugin, the reader may refer to [28]. As a start, a remote sensing image is envisioned (for instance VHRSI) after the segmentation process in eCognition. The desired task is to identify different types of building classes within the segmented objects. In the next step, the segmented objects are exported to the GeoJSON format. The features utilised in this example are building features from literature. In particular, features mentioned in the work of [29] are employed such as i) the occupied area of the building, ii) the density of buildings in a specific area, iii) the two-dimensional shape of the building, and iv) the roof type of the building. An example set of segmented objects after the export to GeoJSON can be seen in Fig. 2.

The 'manual_classification' field holds the associated manual (visual) classification by the expert. Type '1' represents an 'ResidentialArea' and Type '2' represents an 'IndustrialArea'. The generated GeoJSON file can then be imported by the OWLET plugin. The user interface of the extension as such is clean and simple. Via a file explorer, the user can select the specific GeoJSON file which is then imported into the ontology.

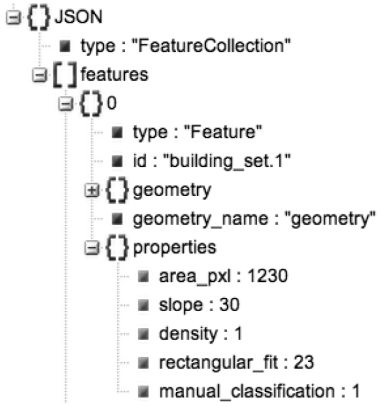


Fig. 2: GeoJSON example file after export

It is important to use the same object properties from the segmented objects within the ontology. A visualisation of the modelled *a priori* knowledge is depicted in Fig. 3. This figure represents a snapshot of the larger ontology, which was embedded in the *Corine* land-use/land-cover ontology[5]. After the import of the segmented objects within the GeoJSON file, the objects are contained as individuals within the ontology. In the next step, the reasoning process can be started to perform the ontology-based classification.

Based on the combined results per class (Tab. I and Tab. II), *precision* (1) and *recall* (2) can be calculated. In this case, the precision and recall values reach 100% and 70% respectively for the class 'ResidentialArea', while the associated precision and recall values for the "IndustrialArea' class reach 80% for precision and 50% for recall.

	Condition pos.	Condition neg.
Test pos.	392	0
Test neg.	168	240

TABLE I: Precision and Recall for 'ResidentialArea' class

	Condition pos.	Condition neg.
Test pos.	120	30
Test neg.	120	530

TABLE II: Precision and Recall for 'IndustrialArea' class

$$Precision = \frac{true\,positive}{true\,positive + false\,positive} \quad (1)$$

$$Recall = \frac{true\,positive}{true\,positive + false\,negative} \quad (2)$$

This situation occurred due to the 'unsharp' properties of some objects. An example object can be seen in Fig. 4. This object was manually classified as 'IndustrialArea', but was neither recognised as 'ResidentialArea' nor 'IndustrialArea'. Some properties are very close to both types of classes and therefore it is not possible to come up with a distinct result.

[5] Corine - http://harmonisa.uni-klu.ac.at/de

At this point, it is the expert's task to decide upon the quantitative description of the classes and their refinement. In addition, the expert might need to add additional qualitative descriptions as well. Furthermore, it could turn out that the existing object properties are not enough to describe the given classes in a proper way and additional properties have to be included or existing properties have to be omitted due to their generality.

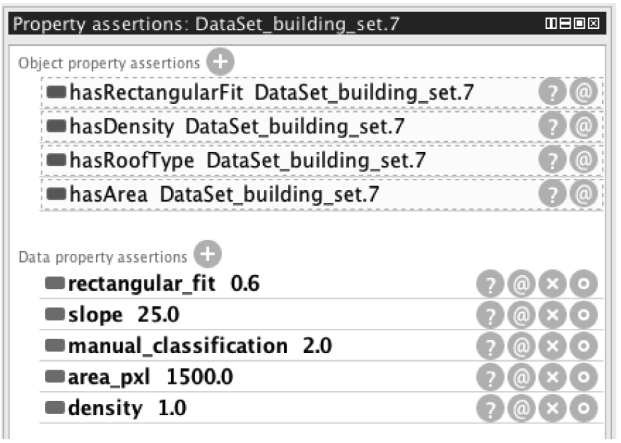


Fig. 4: Misclassified object

VI. Discussion and conclusion

As Protégé is the *de facto* software in regard to ontology development, the introduced extension does not only provide a higher degree of comfort, but also speeds up the entire evaluation process. As manually defining hundreds of segmented objects and their associated attributes would demand a tremendous amount of time, the import and translation capabilities of the OWLET plugin perform these tasks in a few minutes or even a few seconds - depending on the amount of objects.

However, several potential pitfalls remain, which should not be neglected. Three issues are related to the work with ontologies and quantitative modelling, while the fourth issue is an inherent problem of the remote sensing domain. The first issue related to the work with ontologies is represented by the so-called *semantic gap* [30]. This issue describes the fact that visual description of data is biased by the analysts' own perception and experiences. Hence, assuming *n* experts work on the interpretation task, *n* different interpretations may result. However, this issue of 'an objective reality' is not novel and was (and still is) discussed in philosophy under the paradigm of 'constructivism' [31].

Another issue can be identified as the problem of 'overfitting' [32]. This issues occurs if a model comprises more details than necessary (such as attributes or terms) to define a given concept. This can lead to worse decisions during the classification process, as the 'tweaked' ontology may cover some types of object classes better, while others are now misclassified. In addition, overfitting has a severe impact on the ontology's transferability.

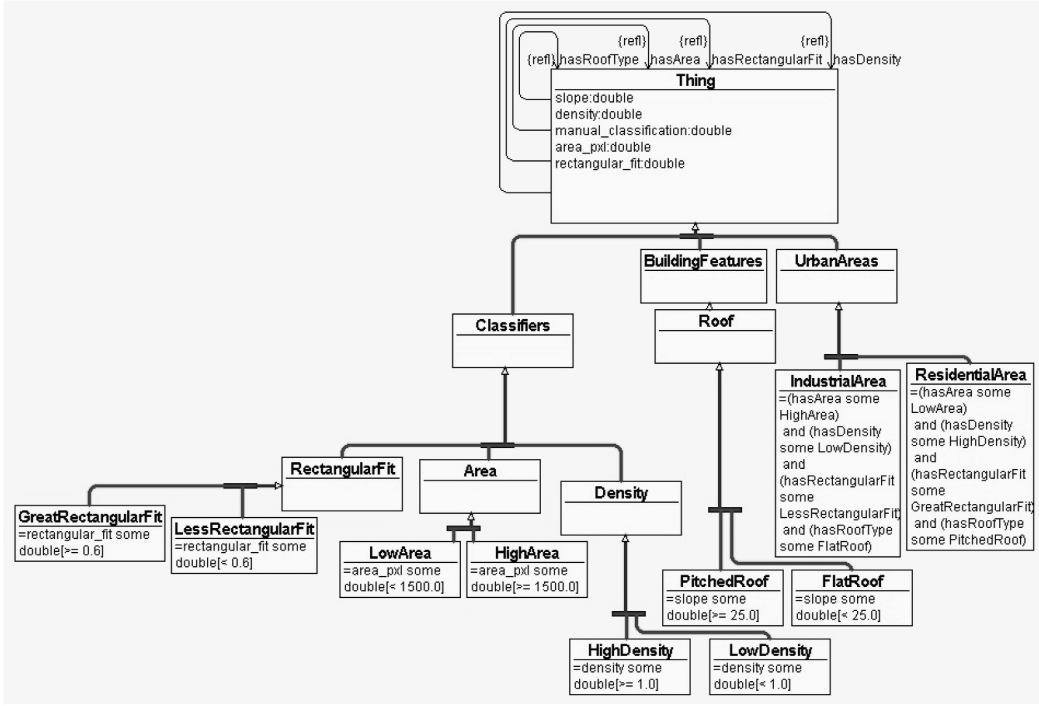


Fig. 3: Experts' a priori knowledge formalised within an OWL ontology in Protégé

For instance, an expert refines the ontology's quantitative and qualitative descriptors to match the given image close to 100% of the 'gold standard'. If the actual image has some special image properties, these are modelled within the ontology as well. If the expert now tries to transfer the newly gained ontology to another image from the same domain – e.g. another snapshot of a different area of the same city – these additional modelled properties have now the potential to negatively influence the classification results.

When working with ontology reasoning, the topic performance is an important aspect. Performance issues related to reasoning in terms of complexity or computational resources for various reasoners were studied by the authors of [33] and [34].

There come several reasoners included in Protégé 'from the shelf'. One of them is called *Pellet* and comprises state-of-the-art optimisation techniques such as Normalisation, Simplification, Absorption and Semantic Branching [35]. In addition, it features novel approaches to improve performance when handling nominal (enumerated classes) and individuals. As the approach in this paper strongly deals with individuals, Pellet poses as one solution for the reasoning process. Still, the user has to bear in mind that large numbers of individuals (several thousands) may heavily impact on the overall performance in terms of computation times and memory resources.

Figure 5 depicts the overall classification performance of the three common reasoners for Protégé. It can be seen that if the number of objects increases, the classification time is rapidly rising around 1,000 objects. This behaviour can become mission critical when dealing with near-real-time applications.

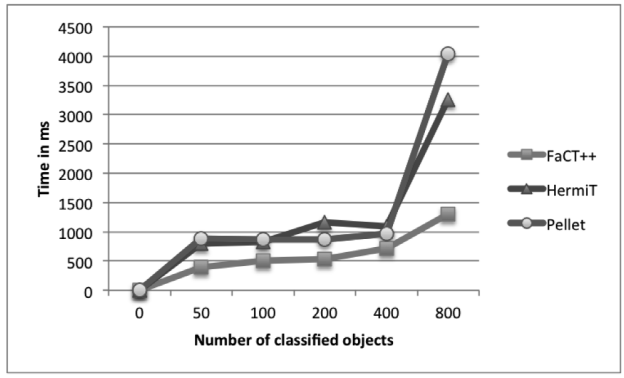


Fig. 5: Performance and run time comparison of the main reasoners of Protégé. The x-axis represents the number of objects classified, while the y-axis shows the associated classification time in ms.

The last issue to be discussed is an inherent problem of the remote sensing domain itself – the segmentation process [36]. Image segmentation refers to the process of aggregating adjacent pixels based on their similarities such as texture. As the related algorithms for the segmentation process tend to react heavily on small changes to the segmentation parameters, the outcomes of the segmentation may vary significantly. Users of the presented plugin may therefore not neglect that this important step in the process flow will impact the overall accuracy of the ontology – even if the ontology itself would be '100%' accurate.

This paper presented the OWLET plugin for Protégé to deliver an integrated solution for the process of ontology eval-

uation. This open-source extension supports domain experts from numerous field during the iterative process of formalising his or her knowledge, while performing collateral evaluation and refinement. The introduced integrated methodology for semi-automated classification of (image) objects, together with the implemented plugin serves as the missing link in the defined process. In addition, the suggested process, as well as its implementation, relies on multiple international standard technologies and tools such as JSON, OWL2 and Protégé. For the classification task, we map extracted (image) object information against a formalised a priori expert knowledge in form of an ontology. Furthermore, we demonstrate how to refine the modelled knowledge based on the classification outcomes. Our approach is built on state-of-the-art technologies; it is open and generic and can thus be adopted by people from various research fields.

ACKNOWLEDGMENT

This research is funded by the Austrian Science Fund (FWF) and the Salzburg University of Applied Sciences through the Doctoral College GIScience (DK W 1237-N23).

REFERENCES

[1] T. Blaschke, "Object based image analysis for remote sensing," *ISPRS journal of photogrammetry and remote sensing*, vol. 65, no. 1, pp. 2–16, 2010.

[2] P. Hofmann, J. Strobl, and A. Nazarkulova, "Mapping green spaces in bishkek—how reliable can spatial analysis be?" *Remote Sensing*, vol. 3, no. 6, pp. 1088–1103, 2011.

[3] N. S. Anders, A. C. Seijmonsbergen, and W. Bouten, "Segmentation optimization and stratified object-based analysis for semi-automated geomorphological mapping," *Remote Sensing of Environment*, vol. 115, no. 12, pp. 2976–2985, 2011.

[4] L. Moller-Jensen, "Classification of urban land cover based on expert systems, object models and texture," *Computers, environment and urban systems*, vol. 21, no. 3, pp. 291–302, 1997.

[5] T. R. Gruber *et al.*, "A translation approach to portable ontology specifications," *Knowledge acquisition*, vol. 5, no. 2, pp. 199–220, 1993.

[6] G. Forestier, A. Puissant, C. Wemmert, and P. Gançarski, "Knowledge-based region labeling for remote sensing image interpretation," *Computers, Environment and Urban Systems*, vol. 36, no. 5, pp. 470–480, 2012.

[7] F. de Bertrand de Beuvron, S. Marc-Zwecker, A. Puissant, and C. Zanni-Merk, "From expert knowledge to formal ontologies for semantic interpretation of the urban environment from satellite images," *International Journal of Knowledge-Based and Intelligent Engineering Systems*, vol. 17, no. 1, pp. 55–65, 2013.

[8] M. Thonnat, "Knowledge-based techniques for image processing and for image understanding," *Journal de Physique 4*, vol. 12, no. 1, pp. Pr1–189, 2002.

[9] P. Hofmann, P. Lettmayer, T. Blaschke, M. Belgiu, S. Wegenkittl, R. Graf, T. J. Lampoltshammer, and V. Andrejchenko, "Abia – a conceptional framework for agent based image analysis," *South-Eastern European Journal of Earth Observation and Geomatics*, vol. 3, no. 2s, pp. 125–130, 2014.

[10] D. Arvor, L. Durieux, S. Andrés, and M.-A. Laporte, "Advances in geographic object-based image analysis with ontologies: A review of main contributions and limitations from a remote sensing perspective," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 82, pp. 125–137, 2013.

[11] C. Hudelot and M. Thonnat, "A cognitive vision platform for automatic recognition of natural complex objects," in *Tools with Artificial Intelligence, 2003. Proceedings. 15th IEEE International Conference on*. IEEE, 2003, pp. 398–405.

[12] W. Kuhn, "Ontologies in support of activities in geographical space," *International Journal of Geographical Information Science*, vol. 15, no. 7, pp. 613–631, 2001.

[13] J. H. Gennari, M. A. Musen, R. W. Fergerson, W. E. Grosso, M. Crubézy, H. Eriksson, N. F. Noy, and S. W. Tu, "The evolution of protégé: an environment for knowledge-based systems development," *International Journal of Human-computer studies*, vol. 58, no. 1, pp. 89–123, 2003.

[14] A. Katifori, E. Torou, C. Halatsis, G. Lepouras, and C. Vassilakis, "A comparative study of four ontology visualization techniques in protege: Experiment setup and preliminary results," in *Information Visualization, 2006. IV 2006. Tenth International Conference on*. IEEE, 2006, pp. 417–423.

[15] S. Calegari and D. Ciucci, "Fuzzy ontology, fuzzy description logics and fuzzy-owl," in *Applications of Fuzzy Sets Theory*. Springer, 2007, pp. 118–126.

[16] M. Belgiu, T. Lampoltshammer, and B. Hofer, "An extension of an ontology-based land cover designation approach for fuzzy rules," in *GI_Forum 2013. Creating the GISociety*, A. Car, T. Jekel, and J. Strobl, Eds. Vienna: Austrian Academy of Sciences Press, 2013, pp. 59–70.

[17] J. Brank, M. Grobelnik, and D. Mladenić, "A survey of ontology evaluation techniques," in *In In Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005*. Citeseer, 2005.

[18] A. Maedche and S. Staab, "Measuring similarity between ontologies," in *Knowledge engineering and knowledge management: Ontologies and the semantic web*. Springer, 2002, pp. 251–263.

[19] R. Porzel and R. Malaka, "A task-based approach for ontology evaluation," in *ECAI Workshop on Ontology Learning and Population, Valencia, Spain*. Citeseer, 2004.

[20] C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks, "Data driven ontology evaluation," in *International Conference on Language Resources and Evaluation (LREC 2004), 24-30 May 2004, Lisbon, Portugal.*, 2004.

[21] A. Lozano-Tello and A. Gómez-Pérez, "Ontometric: A method to choose the appropriate ontology," *Journal of Database Management*, vol. 2, no. 15, pp. 1–18, 2004.

[22] D. L. Pham, C. Xu, and J. L. Prince, "Current methods in medical image segmentation 1," *Annual review of biomedical engineering*, vol. 2, no. 1, pp. 315–337, 2000.

[23] H. Butler, M. Daly, A. Doyle, S. Gillies, T. Schaub, and C. Schmidt, "The geojson format specification," 2008.

[24] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, "Owl 2 web ontology language primer," *W3C recommendation*, vol. 27, pp. 1–123, 2009.

[25] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[26] I. Horrocks, P. F. Patel-Schneider, and F. Van Harmelen, "From shiq and rdf to owl: The making of a web ontology language," *Web semantics: science, services and agents on the World Wide Web*, vol. 1, no. 1, pp. 7–26, 2003.

[27] H. Knublauch, R. W. Fergerson, N. F. Noy, and M. A. Musen, "The protégé owl plugin: An open development environment for semantic web applications," in *The Semantic Web–ISWC 2004*. Springer, 2004, pp. 229–243.

[28] M. Belgiu, I. Tomljenovic, T. J. Lampoltshammer, T. Blaschke, and B. Höfle, "Ontology-based classification of building types detected from airborne laser scanning data," *Remote Sensing*, vol. 6, no. 2, pp. 1347–1366, 2014. [Online]. Available: http://www.mdpi.com/2072-4292/6/2/1347

[29] G. Sohn and I. Dowman, "Extraction of buildings from high resolution satellite data," *Automated Extraction of Man-Made Objects from Aerial and Space Images (III). Balkema Publishers, Lisse*, pp. 345–355, 2001.

[30] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1349–1380, 2000.

[31] D. H. Jonassen, "Objectivism versus constructivism: Do we need a new philosophical paradigm?" *Educational technology research and development*, vol. 39, no. 3, pp. 5–14, 1991.

[32] D. M. Hawkins, "The problem of overfitting," *Journal of chemical information and computer sciences*, vol. 44, no. 1, pp. 1–12, 2004.

[33] J. Bock, P. Haase, Q. Ji, and R. Volz, "Benchmarking owl reasoners," in *Proc. of the ARea2008 Workshop, Tenerife, Spain (June 2008)*, 2008.

[34] Y. Li, Y. Yu, and J. Heflin, "Evaluating reasoners under realistic semantic web conditions," in *Proceedings of the 2012 OWL Reasoner Evaluation Workshop*, 2012.

[35] F. Baader, *The description logic handbook: theory, implementation, and applications*. Cambridge university press, 2003.

[36] G. Meinel and M. Neubert, "A comparison of segmentation programs for high resolution remote sensing data," *International Archives of Photogrammetry and Remote Sensing*, vol. 35, no. Part B, pp. 1097–1105, 2004.

Design Optimization of Tracking Area List for Reducing Total Signaling Overhead in LTE Systems

Lamia Osman Widaa, Sami Mohamed Sharif

Abstract—**In Long Term Evolution systems (LTE) the concept of Tracking Area List (TAL) is introduced. TAL consists of a group of Tracking Areas (TAs), and it is assigned to a User Equipment (UE), the UE does not need to register its location when it moves within TAs of the assigned TAL. If an optimum TAL design is implemented this may result in the reduction of total signaling overhead cost introduced by location update and paging procedures. One of the challenges of mobility management in cellular networks is the reduction of overall signaling overhead while maintaining acceptable performance. In this paper we propose a TAL design optimization for reducing overall signaling overhead algorithm. To achieve optimum TAL design a cell may change its current TAL, and this may cause a service interruption for active UEs in that cell. A budget constrain parameter is introduced to define the maximum number of cells change TALs while maintaining improved performance. For UE mobility modeling the data required can be obtain from the network management system which are cell load and handover statistical data. Markov model is used for UE mobility modeling. We present numerical results, from which we can say that the optimum TAL design returned from the proposed algorithm gave considerable reduction of overall signaling overhead cost compared to the traditional TA design.**

Index Terms*— *LTE, , Location Update, Paging, Signaling Overhead, TAL.

I. INTRODUCTION

In cellular communication networks, mobility management consists of location registration (location update) and paging, and it is one of the most essential techniques to provide communication services to UEs. Mobility management in Long Term Evolution (LTE) is different from that in the third generation mobile telecommunication networks. In LTE, the Mobility Management Entity (MME) is responsible for the mobility management function, and it's connected to a large number of evolved Node Bs (cells) that

are grouped into the Tracking Areas (TAs). The TA is defined as an area in which the user equipment (UE) may move freely without updating the MME. Tracking Area (TA) is a cluster of evolved Node Bs (eNBs) having the same TA code [9],[10]. In the standard TA scheme, having TA with small size (few number of cells) eliminates the paging signaling overhead; on the other hand having TA with a large size eliminates the location update overhead. When a UE receives a call, the network must page cells within the Location Area (polling) to find that user as quickly as possible. This process all induces system overhead in both system signal and wireless bandwidth consumption. If the wired network knows the exact location of a UE, the paging cost can be reduced to a minimum by polling only the cell in which the UE is situated. On the other extreme case, if the wired network does not have any information about the location of the UE, cells all over the wireless network have to be polled. This costs the maximum system overhead. So the problem is how to find an optimal TA design which gives reduction in signaling overhead and optimal balance between Tracking Area Update (TAU) and paging signaling overhead. Limitations of the tracking Area scheme can be summarized as follows:

- Ping Pong effect: Here if the UE moves back and forth between two or three neighboring TAs and this will cause excessive TAU, because when UE enters a new TA it will perform TAU.
- Massive Mobility Signaling Congestion: Here If a large number of users simultaneously move into a hotspot cell, this will cause excessive TAU from the UEs, in a short period of time.

In LTE networks Tracking Area List (TAL) scheme has been introduced to solve a problems existing in the standard TA scheme such as ping-pong, massive mobility problems discussed above and localized spikes in uplink traffic problem [1],[5]. The MME provides the UE with a list of TAs where the UE registration is valid. The network allocates a list with one or more TAs to the UE. The UE may move freely in all TAs of the list without updating the MME. When the MME pages a UE, a paging message is sent to all cells in the TAL.

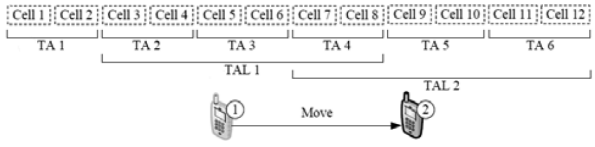


Fig. 1: An example of a TAL scheme

Manuscript received Febuary 6, 2014, revised May 1, 2014.
Lamia O.Widaa is with the Department of Telecommunication Engineering, University of Science & Technology, Khartoum Sudan. (e-mail: lolawidaa@gmail.com).
Sami M.Sharif is with the Department of Electrical & Electronic Engineering, University of Khartoum, Sudan. (e-mail: smsharif@uofk.edu).

Fig.1 above shows an example of a TAL scheme. In Long Term Evolution (LTE), the Mobility Management Entity is connected to a group of evolved Node Bs (eNBs; the LTE term for base stations). The radio coverage of an eNB (or a sector of the eNB) is called a cell (see the dashed squares. Every cell has a unique cell identity. The cells are grouped into the Tracking Areas (TAs; e.g., TA 1 contains Cell 1 and Cell 2. Every TA has a unique TA Identity (TAI). The TAs are further grouped into TA Lists (TALs). In Fig.1 TAL1 consists of TA 2, TA3, and TA4. A UE stores the TAL that includes the TA where the UE resides. In Fig.1, the UE is covered by Cell 5, and the TAL it stores is TAL1={TA 2, TA 3, TA 4}. If the LTE network attempts to connect to the UE, it asks the cells in the TAL (e.g., Cell 3-Cell 8) to page the UE. Every eNB periodically broadcasts its TAI. The UE listens to the broadcast TAI and checks if the received TAI is in its TAL. If so, it means that the UE does not move out of the current location. When the UE moves from Cell 5 to Cell 7, it receives the TA4 identity broadcast from eNB7. Since TA4 is included in TAL1, the UE still resides in the same location. When the UE moves to Cell 9 the received TA5 identity (broadcast from eNB9) but is not found in TAL 1, which means that the UE has moved out of the current location. In this case the UE executes the location update procedure to inform the MME that it has left TAL1. The MME then assigns a new TAL to the UE. Now the new TAL is TAL2= {TA4, TA5, TA6}. In LTE systems TALs for different UEs may have different sizes, and the newly assigned TAL may be overlapped with the previously assigned TAL (as shown in the previous example) [6].

If the information of each individual UE's movement and calls were available for the network, then designing an optimum TAL would become trivial and could essentially result to elimination of signaling overhead. In this situation the cell could give a specific, tailored list to each UE including all the cells the UE is intended to pass before it will be called. This information, if available at all, is costly to obtain. Moreover, the validity of the information expires fast, because the trace is the history of the UE's movement, and the UE's intention of where and when to move in future is unknown [4].

Finding an optimum design of TA and TAL (which gives minimum signaling overhead) is a challenging task in LTE network, and in general the goal of the location management is to find balance between the following:

- More frequent update (reduce polling cost).
- Rare location update, storing less information about users, reduces computational overhead (higher polling cost).
- Optimize the design of TAs or TALs, less handoff, quicker locating of user.

In this paper we proposed a TAL design optimization for reducing overall signaling overhead algorithm, this algorithm returns an optimum design of TALs (which gives minimum total signaling overhead) by using the available data obtained from the network management system which are cell load and handover statistical data during a given period of time. The optimum design was obtained by initially chosen random cells and changing the TAL of those cells till we reach the optimum design for all TALs in the network. In general the TAL of a cell can be modified at a time by either deleting or adding one of the cells in the TAL. For every cell in the network and if the cell changes its TAL this will causes service interruption, the cell load of the cell (number of active and idle UEs in the cell) is used to measure this service interruption in that cell. When applying the design to the suggested network it gave considerable reduction of total signaling overhead.

The rest of this paper is organized as follows: In Section II related works is summarized. Section III descries the proposed algorithm. Numerical results are presented in Section IV. Finally, Section V concludes the paper.

II. RELATED WORK

In Personal Communication Services Networks (PCS) several strategies were proposed to reduce the location update cost. In [7] the researchers studied a special case of a location tracking algorithm called the Alternative Location Algorithm (ALA) . This special case is referred to as the Two Location Algorithm (TLA). An analytical model is proposed to compare the performance of TLA and the IS-41 protocol. The study indicates that the performance of TLA is significantly affected by the user moving patterns and the call traffic. If the user mobility is higher than the call frequency or the user tends to move back to the previously visited registration areas, then TLA may significantly outperform IS-41.

Many researches in the literature aimed to contribute in reducing total signaling overhead in LTE networks. TAL concept is expecting to reduce the overall signaling overhead when compared to traditional TA concept. In [2] the authors proposed a method for allocating and assigning TALs for LTE networks. This method is called "Rule of thumb". The optimum conventional TA design was compared with the proposed TAL; it found that TAL works best if a dynamic frequent reconfiguration is applied for different time intervals. The Rule of thumb method is simple and cannot guarantee to give an optimum TAL design because each cell in the network is selfishly optimizing the signaling overhead according to their own data and does not considered the impact of the other cells on their modified TAL.

In [4] the authors introduced an approach for allocating and assigning TALs, here the impact of neighbor cells was considered, and the data required for TAL allocation is the same data used of TA design, which are the cell load and handover statistics. These data can be obtained from the mobility management system in the network. In the proposed scheme different users UEs in one cell are holding different TALs according to the original cell they are registered in. The authors aimed to show that even with a simple algorithm a TAL design is able to reach a lower overall overhead than the conventional TA design. The proposed algorithm takes into account the impact of neighbor cells in the allocation of TALs. The impact of the neighbor of neighbor cells is not considered, if it was considered this will gives more accurate design, but the algorithm will be complicated.

In [1] the local search algorithm is introduced for designing TALs. The input data required for implemented the algorithm was the same data required for designing standard TA (cell

load and handover statistics). The basic operation of the algorithm was to modify the TAL of a cell at a time by either deleting or adding one of the cells in the TAL, at each trial the overall signaling overhead is calculated, the algorithm was repeated until lower signaling overhead was reached. Here one of the disadvantages is that an optimum TA design was required before starting the algorithm. The algorithm showed lower overall signaling overhead when compared with the standard TA scheme.

All of the algorithms mentioned above were implemented using the available statistical data from the network mobility management system and were independent of UE traces. The design of TALs based on UEs traces is also possible. Here each cell is able to assign different TALs to different UEs. But if the UEs change their movement, which is quite probable the TALs would become inefficient.

III. TAL DESIGN OPTIMIZATION ALGORITHM

From the design of the proposed algorithm we tried to find an optimum TALs design, which gives minimum overall signaling overhead. Here a budget parameter is defined because the TAL of a cell can be modified at a time t by either deleting or adding one of the cells in the TAL. For every cell in the network and if the cell changes its TA (change its e nodeB) this will causes service interruption for active UEs on that cell, the cell load (Number of active and idle UEs in the cell at time t) of a cell is used to measure the service interruption in that cell. In [1] the budget constrains parameter B is defined by the following equation:

$$\sum_{i \in N} u_i \, d_i(t, t^o) \leq B \qquad (1)$$

Where:

N is the number of cells in the network.

u_i is the cell load of cell i.

t^0 is initial TA of cell i and t is the TA returned by the algorithm for cell i

$d_i(t,t^0)$ is a binary vector , $d_i(t,t^0)$ =1 if and only if t^0 of cell i $\neq t_i$.

To find an optimum design of TAL the service interruption is taken into account and the problem can be defined as finding an optimum design of TAL which gives minimum total signaling overhead and satisfies (1) above. The algorithm is examined for different suggested values of parameter B and for the network which descried in Section A below.

A. NETWORK DESCRIPTION

Network description is similar to [3]. The network has hexagonal cellular configuration with 61 cells, every cell has a unique cell identity or index (see fig.2 below), and every TA has a unique TA identity (TAI). When UE moves into cell i it resides in the cell for a random period of time and then moves out in the direction of one of six neighbor cells. In the proposed algorithm TAL overlapping was considered which means that one TA can be included in more than one TAL. The UE mobility model used is markov model. We assumed that the TA consists only of one cell, and no restriction on number of TAs within a TAL. A UE stores the TAL that includes the TA where the UE resides, when the UE moves out from its current TAL in this case it will executes the location update procedure to inform the MME that it has left current TAL. Then the MME assigns a new TAL to the UE. The newly assigned TAL maybe overlapped with the previously assigned TAL as discussed in section I. We assumed that there is no integration between the suggested LTE system and any Radio Access Technology (RAT).

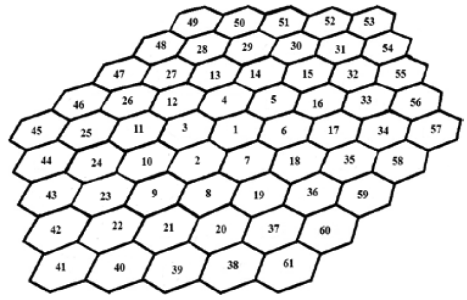


Fig. 2. Network cells

In the proposed network the calculation of the overall signaling is depends on the available statistical data obtained from the network management system. Here the data used for calculation are the cell load and handover statistical data. These data are generated using a UE mobility model. Also the design of TAs and TALs affects the calculation of overall signaling overhead.

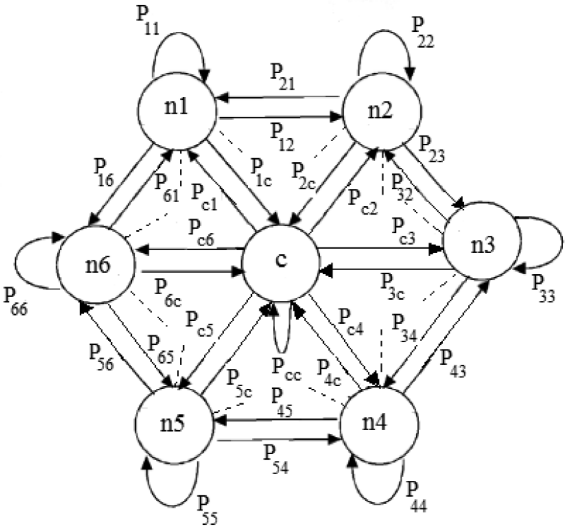
B. UE MOBILITY MODEL

UE movement and mobility behavior in a cellular network can be described by cell residence time and handover probability calculated for each cell in the network based on the time series of the visited cells of the UEs [4]. Cell residence time and handover statistics can be obtained from the management system of the cellular network. From the literature there are different UE mobility models that can be used to describe UE mobility behavior [11],[12],[13]. Markov model is widely used for this purpose. And in the proposed algorithm it's used for UE mobility modeling. It's a mathematical model that undergoes transitions from one state to another, between a finite or countable number of possible states. It is a random process usually characterized as memory less and the next state depends only on the current state and not on the sequence of events that preceded it. This specific kind of "memory lessens" is called the Markov property. Markov model have many applications as statistical models of real-world processes.

The Markov model, also known as the random-walk model and it can be applied to cellular systems to describe individual movement behavior of UEs. In this model, the UE at any given time slot will either remain within a cell or move to an

adjacent cell according to a transition probability distribution, this probability is often adjusted to practical observations of UE behavior in cells. In general the markov model can be described by the following:

- A markov chains with finite set S of m states.
  S ={s_1, s_2, ……,s_m}. And in the proposed algorithm we used 7-state markov model (m = 7) where each state represents a radio cell in the network. For each cell in the network there are 6 neighbors (or less if the cell is terminal cell such as cell with index 57 in Fig. 2 which has only 3 neighbors).
- The State Transition Probability distribution matrix P with size (m×m). Element p_{ij} in this matrix represents the probability of the movement of the UE from cell i to cell j in next time slot. Fig.3 below shows the state diagram and the transition matrix of 7-state markov model. Cell c represent the current cell of the UE, and cell n_z is one of the neighbor cells of cell c, where z=1,2,……,6.



(a) The State Diagram

P =

P_{cc}	P_{1c}	P_{2c}	P_{3c}	P_{4c}	P_{5c}	P_{6c}
P_{c1}	P_{11}	P_{21}	P_{31}	P_{41}	P_{51}	P_{61}
P_{c2}	P_{12}	P_{22}	P_{32}	P_{42}	P_{52}	P_{62}
P_{c3}	P_{13}	P_{23}	P_{33}	P_{43}	P_{53}	P_{63}
P_{c4}	P_{14}	P_{24}	P_{34}	P_{44}	P_{54}	P_{64}
P_{c5}	P_{15}	P_{25}	P_{35}	P_{45}	P_{55}	P_{65}
P_{c6}	P_{16}	P_{26}	P_{36}	P_{46}	P_{56}	P_{66}

(b) The Transition Probability Matrix

Fig.3. The state diagram and the transition probability matrix of a 7-state markov model

Probabilities introduced in the transition probability matrix shown in Fig.3 above are defined as the handover probabilities, and can be calculated from UEs handover statistical data obtained from the network management system [3]. The UE can be located in 7 different states during each time slot depending on handover probability of each neighbor cell.

- The steady state probability distribution vector in which π_i is the probability of a UE being in state i (from which cell residence time can be calculated). We assumed that cell residence times are Independent Identical Distributed random variables (IID) with average residence time 1/ƛ. The steady state distribution of the cell residence time satisfies the following equation:

$$\pi_j = \sum_{i=1}^{7} \pi_i P_{ij} \quad (2)$$

And

$$\sum_{j=1}^{7} \pi_j = 1 \quad (3)$$

C. UE TRACE MATRIX

In the proposed algorithm the UE trace matrix (V×T) is implemented similar to [1], in which V UEs are traced. Fig.4 below shows one raw in the trace matrix, the time interval in which UEs are traced is T and this interval was divided into n time slots (t), the serving cell of UE_v at time slot t_i is stored at element (v ,t_i) in the matrix , i=1,2,3……,n.

The starting cell (Cs) for each UE in the trace matrix can be obtained from (4) below:

$$P_{cs} = \frac{u_i}{\sum_{j=1}^{N} u_j} \quad (4)$$

Where:
P_{Cs} is the probability that cell i will be the starting cell for user UE_v,
u_i is the cell load (number of Active and Ideal users in cell i) of cell i.
N is the number of cells in the network.

For the design of the proposed algorithm we assumed the following:

- The initial cell load of each cell in the network was generated randomly, but in real world it may obtain from the network management system. The cell with high probability (and with high cell load) will be chosen to be the starting cell for UE_v. The first column in the trace matrix represents the starting cell for all UEs.
  According to the steady state vector obtained from markov model the cell residence time for each user is Independent and Identical Distributed (IID) random variable. The cell residence time is the time the UE spends in the cell before moving to one of the neighbor cells of current cell (handover process). In the algorithm the Residence Time (RT) for cell i will be calculated as follows:

$$RT = \pi_i * T\,min \quad (5)$$

Where
π_i represents the portion of time T that the UE spent in cell i.

- In the trace matrix the UE makes handover to the neighbor cell with high handover probability; this probability is calculated from the following formula:

$$P_{ij} = \frac{h_{ij}}{\sum_{p \in A_i} h_{ip}} \qquad (6)$$

Where
$P_{i,j}$ is the probability that UE will make handover from cell i to cell j,
A_i is a set contains indexes of neighbor cells of cell i.
h_{ij} is the number of users making handover from cell i to cell j,
h_{ip} is the total number of users making handover from cell i to neighbor cell p.

We assumed that at any given time slot in the trace matrix and if the UE decided to make a handover process, it will move to the neighbor cell with high handover probability. Handover values (number of users making handover) which are used to calculate the handover probability discussed above are generated randomly but in real world it can be obtained from network management system.

Fig.4. An example of one raw in the trace matrix

In the example shown in Fig.4 the time interval T is divided into 20 time slots. The cell load u_i is defined as the total number of UEs in cell i scaled by the time proportions that the UEs spend in cell i. Therefore, the load of each cell in the network is aggregated by the scaled values of UEs staying in the cell using all the elements of the trace matrix. The aggregated handover value is the number of moves from one cell to another [1].
Consider the example above, the aggregated cell load and handover for UE_v are:
$u_i = 0.35$, $u_j = 0.3$, $u_k = 0.35$, and $h_{ij} = 1$, $h_{jk} = 1$,$h_{kj}=2$, $h_{ji}=2$, $h_{ki}=2$, $h_{ik}=3$.Where $\frac{\Delta\tau}{T} = 0.05$.

D. CALCULATION OF OVERALL SIGNALING OVERHEAD

To calculate overall signaling overhead the cell load and handover data between each neighbor cells are required. These data are calculated directly from the trace matrix discussed in Section *C*, Here the load of each cell was aggregated by the scaled values of UE staying in the cell using all elements in the trace matrix. In the proposed algorithm we assumed that the TA consists only of one cell. For the traditional TA scheme handover data was aggregated from the trace matrix as follows, if a UE makes handover from cell i to cell j this means that h_{ij} will increase by one, So the h_{ij} value also calculated using all elements in the trace matrix.

Equation (7) below is used to calculate the overall signaling overhead for the TA scheme.

$$C_{SO}(t) = \sum_{i \in N} \sum_{j \in N: j \neq i} \left(h_{ij}\, c^u \left(1 - S_{ij}(t)\right) + \propto u_i c^p S_{ij}(t)\right) \qquad (7)$$

Where
h_{ij} is the number of users performed handover from cell i to cell j (if i and j are not in the same TA).
c^u is the update cost caused by one UE,
c^p is the amount of overhead of one paging.
$\propto$ is the probability that a UE has to be paged (also called call intensity factor).
u_i is the total number of UEs in cell i scaled by time proportion each UE spent in cell i. It can be obtained by collecting UE statistics over a given time (using the trace matrix).
In (7) the first term represents the overhead caused by the location update process for users moving from cell i to cell j (if the two cells are not in the same TA), and the second term represents the overhead caused by the paging process (if the two cells are in the same TA) .

To calculate the overall signaling overhead a TA design is required, the design of the TA is represented by the S(t) matrix, its N×N binary matrix where N is the number of cells in the networks. The value of element $S_{ij}(t)$ is determined as follows:

$$s_{ij}(\mathrm{t}) = \begin{cases} 1 & \text{if } \mathrm{t_i} = \mathrm{t_j} \\ 0 & \text{otherwise} \end{cases}$$

Where
t_i is the TA of cell i, and t_j is the TA of cell j.

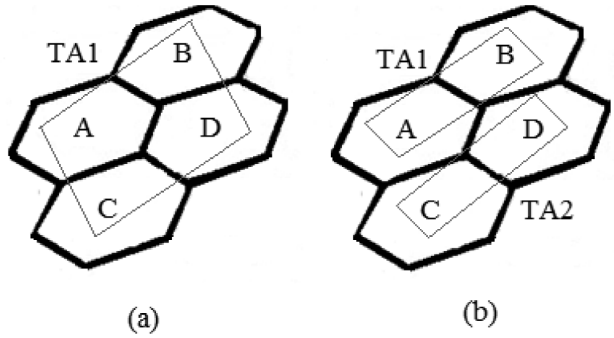


Fig.5. An example of the TA scheme

If all cells are included in one TA Fig(5.a) then the S(t) matrix is

$$S(t) = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

In this case the overall signaling overhead calculated is:

$$C_{so1}(\mathrm{t}) = 3 \propto \mathrm{c^p}(u_A + \mathrm{u_B} + \mathrm{u_C} + \mathrm{u_D}) \qquad (8)$$

Where
u_A, u_B, u_C and u_D are the cell load of cells A, B, C, and D respectively.
The signaling overhead calculated in (8) is introduced by paging process because all cells are in the same TA and there is no location update cost.
If cell A and B are in one TA and cell C and D are in another TA as shown in Fig (5.b), then the s(t) matrix is

$$S(t) = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

And the overall signaling overhead calculated is:

$$C_{so2}(\mathrm{t}) = \mathrm{c^u}(\mathrm{h_{AC}} + \mathrm{h_{AD}} + \mathrm{h_{BC}} + \mathrm{h_{BD}} + \mathrm{h_{CA}} + \mathrm{h_{CB}} + \mathrm{h_{DA}} + \mathrm{h_{DB}}) + \propto \mathrm{c^p}(u_A + \mathrm{u_B} + \mathrm{u_C} + u_D) \quad (9)$$

In the TAL scheme, different UEs in one cell are holding different TALs according to the original cell they are registered in. If we define *sij*(**t**) as the number of UEs in cell *i*, who have *j* in their list divided by the whole number of UEs in cell *i*, then **S**(**t**) cannot remain as a binary matrix, but it will rather contain some fractional values between 0 and 1. The authors in [4] gave an example and showed that if cell A percepts that cell B is in another TA but cell B assumes that they are in the same TA, this flexibility could lead to a lower overhead. As discussed in [4] *sij*(**t**) is given by the following formula.

$$S_{ij}(t) = \frac{u_i l_{ij} + \gamma \sum_{n \in Q_i} h_{ni} l_{nj}}{u_i + \gamma \sum_{n \in Q_i} h_{ni}} \quad (10)$$

Where:
$S_{ij}(t)$ presents the percentage of users in cell i having j in their TAL.
Qi is the set of neighbor cells of cell i having i in their TAL,
l_{ij} is a parameter and it's equal to 1 if i and j are in the same TAL and 0 otherwise.
γ is the probability of UEs entering cell i from a neighbor cell and have i in their TAL, in this case UEs will enter cell i without performing update process.

Equation (7) is also valid in the calculation of the overall signaling overhead for TAL scheme. For both traditional TA and TAL schemes, the calculation of overall signaling overhead requires:

- A well defined S(t) matrix which represents the design of TAs or TALs in the network,
- The cell load u_i for each cell in the network, and it can be obtained from the trace matrix. Cell load is aggregated by the scaled values of UEs staying in the cell using all the elements of the trace matrix as discussed in the example shown in fig3.
- Handover statistics which are the number of users performed handover from cell i to cell j. Also it can be obtained from the trace matrix discussed in Section C.

E. ALGORITHM BODY

The proposed algorithm returns the overall signaling overhead cost for an optimum TAL design. First the value of parameter B should be determined (It's the budget constrain parameter), this value is a percentage of the total cells load.
An important variable also defined which is bl variable. Initial value of this variable is 0, and it's used to check whether the movement of cell j from its current TAL to another TAL is within the budget constrain or not. When cell j changes its current and initial TAL bl value will increase by u_j (cell load value). And if cell i moves back from its current TAL to its initial TAL bl value will decrease by u_j.

Here the condition is that the value of bl after cell movement should be always less than the value of parameter B. In the design of this algorithm TALs overlapping is considered which means that one cell may be included in two or more TAL. Algorithm code can be summarized as follows:

- Define initial TAL design by the definition of S(t) matrix and determine the starting TAL for each cell in the network by using (4).
- Calculate the overall signaling overhead for the initial TAL design (Variable CS).
- Calculate B according to (11) below:

$$B = q \times U \quad (11)$$

Where $U = \sum_{i=1}^{N} u_i$ and $q \leq 1$

- Select a random cell i, and then for all cell j in the network check whether j is in the same TAL of i or not. If j and i are in the same TAL remove j from TAL of i.
- Update S_{ij} element defined by equation ( ) in the S(t) matrix. Update ADL which is a set containing all adjacent cells to TAL_i and are not included in the list.
- For all cell p which is a cell included in TAL_i and TAL_j update $s_{jp}(t)$ and $s_{pj}(t)$ elements.
  Calculate the overall signaling overhead cost for the new TAL design.
- If the new cost is less than the initial cost update bl value. And replace the initial TAL design with the new one, otherwise keep current TAL design.
- Repeat the steps mentioned above till we get an optimum TAL design which gives minimum overall signaling overhead cost. Or bl value reached B value.

Design Optimization of TAL Algorithm:

1- Definition:
2- ST is the initial S(t) matrix.
3- STT is the current S(t) matrix, which describes the current TAL design.
4- ADL is a set contains the indexes of the neighbor cells of TAL_i.
5- CS is the initial overall signaling overhead calculated based on ST matrix.

6- CSO is the minimum overall signaling overhead returned by the algorithm.

7- Construct the trace matrix.
8- Determine the starting TAL for each cell in the network to create the first column in the matrix.
9- Create the ST matrix according to equation ( ).
10- STT=ST; , CS=T_cost(ST);
11- calculate B.
12- bl=0; , l=1;
13- while bl<B
14- select a random cell i
15- for j=1:61
16- If j is in same TAL with i , remove j from the TAL of i
17- update s_{ij} element in the STT matrix,
18- For all p which is a cell included in TAL_i and TAL_j update s_{jp} and s_{pj} elements. and ADL set
19- end
20- CSO=T_cost(STT);
21- if CSO<CS
22- ST=STT; CS=CSO; ADL=TN;
23- bl=bl+U(j);
24- else
25- STT=ST; TN=ADL;
26- end
27- If j is not included in the list of I, added j to the list and then update s_{ij} element in the STT matrix.
28- For all p which is a cell included in TAL_i and TAl_j update s_{jp} and s_{pj} . and ADL set
29- end
30- CSO=T_cost(STT);
31- if CSO<CS
32- ST=STT; CS=CSO;, ADL=TN;
33- bl=bl+U(j);
34- else
35- STT=ST; , TN=ADL;
36- end
37- end
38- end

IV. NUMERICAL RESULTS

Given the UE traces matrix and a TAL design the exact S(t) matrix elements can be calculated and the aggregated cell load and handover data can be obtained as discussed in Section C. The state transition probability matrix shown in Fib.6 below is used for handover decision in the trace matrix. In this matrix probability values are calculated using (6).

0.30	0.12	0.05	0.02	0.30	0.28	0.04
0.10	0.36	0.12	0.10	0.02	0.05	0.03
0.25	0.13	0.36	0.18	0.12	0.20	0.10
0.02	0.22	0.22	0.12	0.16	0.02	0.20
0.11	0.02	0.10	0.18	0.21	0.12	0.03
0.04	0.10	0.02	0.30	0.09	0.03	0.40
0.18	0.05	0.13	0.10	0.10	0.30	0.20

Fig.5: State Transition Probability Matrix

As mentioned in Section B cell residence times are IID random variables. Fig.6 below shows the Steady State probability vector, which is used in the algorithm to calculate cell residence time.

$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{1}{7}$

Fig.6: Steady State Probability Vector

The tables below show the overall signaling overhead calculated from both standard TA optimum design algorithm implemented using the same methodology as in [1] and the proposed TAL design optimization algorithm. In each table the results calculated for ten different scenarios, each scenario represents a cell load and handover data set. Each table represents the results for a given value for both B and $\propto$ parameters. B value is a percentage of total number of UEs (U) in the network. We assumed that the number of UEs in all UE trace scenarios is 6000. UEs were traced for one hour period, this period was divided into 60 equal intervals and every time interval is equal to one minute. We assumed that Cu =1 , Cp =1 (common in the literature), Also we assumed that the average value of γ *is 0.05.*

Both standard TA optimum design algorithm and TAL optimum design algorithm were implemented using MATLAB, the algorithms run on a processor of type Intel core™ i3 with clock speed of 2.35 GHz. In the design of the network we assumed that there are 23 TAs. And each TA contains only one cell (to simplify calculations). Each TAL may contain two or three TAs (depends on the suggested initial TAL design). Tables below show the overall signaling overhead calculations for different scenarios and data sets. For each scenario the location update cost and paging cost were calculated. The dimension of the trace matrix in all scenarios is 6000×60. Numerical results presented in tables below show the location update, paging cost and total cost for ten different scenarios.

TABLE I
NUMERICAL RESULTS FOR B=15%U & $\propto$ =0.01.

Scen-ario	Optimum STA scheme			Optimum TAL scheme		
	TAU	Paging	Overall	TAU	Paging	Overall
1	25011	581	25592	19341	352	19693
2	25623	575	26198	19834	370	20204
3	25412	582	25994	20021	355	20376
4	25838	602	26440	20051	302	20353
5	24321	543	24864	19863	331	20194
6	25452	562	26014	19761	321	20082
7	24998	576	25574	20023	348	20371
8	25902	530	26432	20021	358	20379
9	25645	610	26255	19781	341	20122
10	24984	552	25536	20023	332	20355

TABLE II
NUMERICAL RESULTS FOR B=50%U & $\propto$ =0.02

Scen-ario	Optimum STA scheme			Optimum TAL scheme		
	TAU	Paging	Overall	TAU	Paging	Overall
1	27168	1018	28186	20068	996	21064
2	27532	989	28521	20112	872	20984
3	26972	1105	28077	19664	1012	20676
4	26884	1231	28115	18841	1002	19843
5	27153	1222	28375	18542	1132	19674
6	26341	996	27337	18731	865	19596
7	27132	1003	28135	18643	1423	20066
8	27634	979	28613	19997	1322	21319
9	27166	1017	28183	19821	1231	21052
10	26878	1033	27911	18732	1399	20131

TABLE III
NUMERICAL RESULTS FOR B=100%U & $\propto$ =0.05

Scen-ario	Optimum STA scheme			Optimum TAL scheme		
	TAU	Paging	Overall	TAU	Paging	Overall
1	27511	931	28442	4226	1798	6024
2	27504	971	28475	4321	1781	6102
3	27555	967	28522	4223	1756	5979
4	26998	943	27941	4651	1795	6446
5	27541	930	28471	4659	1752	6411
6	27510	895	28405	4213	1734	5947
7	27453	984	28437	4551	1722	6273
8	27423	932	28355	4223	1793	6016
9	26978	951	27929	4512	1721	6233
10	27132	940	28072	4224	1790	6014

From tables shown above we may conclude the following:

- The total signaling overhead calculated and recorded in table I shows that the optimum design of TAL is 19% to 23% better than standard TA design. The paging cost is less than the location update cost and it depends on the value of parameter $\propto$. Results were calculated for B=15% and $\propto$=0.01 which means that 1% of the UEs will be paged in every cell.
- The overall signaling overhead calculated and shown in table II shows that the optimum design of TAL is 25% to 31% better than standard TA design. The paging cost is less than the location update cost and greater than the paging cost shown in table I this is because here $\propto$ = 0.02.
- The overall signaling overhead shown in table III shows that the optimum design of TAL is 77% to 79% better than standard TA design. And this because B is 100%U which means that there is no restriction in the design of the TALs and we can freely move cells from TAL to another until we reach the optimum design. The table gives best results when compared to other tables, but with no budget constrain the service interruption when changing the TAL of a cell will affect the performance of the network in a great manner. In the proposed algorithm average value of γ was chosen to be 0.05, and it's required to give a good estimation of γ because it influences the TAL design and the resulting signaling overhead.

V. CONCLUSION

In this paper we proposed design optimization of TAL for reducing overall signaling overhead algorithm. The proposed algorithm returns the minimum overall signaling overhead calculated based on the optimum TAL design. And markov model is used as the UE mobility model. The numerical results obtained from the proposed algorithm show that the design of TAL scheme with TAL overlapping reduces the overall signaling overhead compared to the standard TA scheme. And with large value of B we got better performance of the TAL scheme but we should take into account the service interruption caused by cell movement from TAL to another to keep the acceptable performance of the system.

For future work we suggest to extend the idea of this paper to include a comparison between different mobility models to examine their effect on the reduction of overall signaling overhead when applied to the same scenario.

REFERENCES

[1] Modarres Razavi, Sara. "Tracking Area Planning in Cellular Networks: Optimization and Performance Evaluation." PhD diss., Linköping, 2011.

[2] Razavi, S. Modarres, Di Yuan, Fredrik Gunnarsson, and Johan Moe. "Dynamic tracking area list configuration and performance evaluation in LTE." In GLOBECOM Workshops (GC Wkshps), 2010 IEEE, pp. 49-53. IEEE, 2010.

[3] Szalka, Tamas, Sandor Szabo, and PÉTER FÜLÖP. "Markov model based location prediction in wireless cellular networks." Infocommunications Journal (2009): 40.

[4] Razavi, Sara Modarres, Di Yuan, Fredrik Gunnarsson, and Johan Moe. "Exploiting tracking area list for improving signalling overhead in LTE." In Vehicular Technology Conference (VTC 2010-Spring), 2010 IEEE 71st, pp. 1-5. IEEE, 2010.

[5] Wu, Chien-Hsing, Huang-Pao Lin, and Leu-Shing Lan. "A new analytic framework for dynamic mobility management of PCS networks." Mobile Computing, IEEE Transactions on 1, no. 3 (2002): 208-220.

[6] Liou, R., Y. Lin, and S. Tsai. " An investigation on LTE mobility management." IEEE Transaction on mobile computing(2011), vol 12, no. 1,pp 166-176,2011.

[7] Lin, Yi-Bing. "Reducing location update cost in a PCS network." IEEE/ACM Transactions on Networking (TON),vol 5, no. 1 ,pp 25-33, 1997.

[8] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access. Technical Specification 3G TS 23.401 version 10.0.0 (2010-06), 2010.

[9] 3GPP. 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2. Technical Specification 3G TS 36.300 version 10.1.0 (2010-09), 2010.

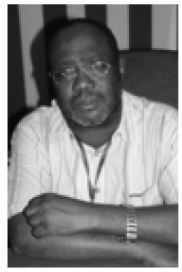
[10] Razavi, Sara Modarres, and Di Yuan. "Performance improvement of LTE tracking area design: a re-optimization approach." In Proceedings of the 6th ACM international symposium on Mobility management and wireless access, pp. 77-84. ACM, 2008.

[11] Jardosh, Amit, Elizabeth M. Belding-Royer, Kevin C. Almeroth, and Subhash Suri. "Towards realistic mobility models for mobile ad hoc networks." In Proceedings of the 9th annual international conference on Mobile computing and networking, pp. 217-229. ACM, 2003.

[12] Zonoozi, Mahmood M., and Prem Dassanayake. "User mobility modeling and characterization of mobility patterns." Selected Areas in Communications, IEEE Journal on 15, no. 7 (1997): 1239-1252.

[13] Liang, Ben, and Zygmunt J. Haas. "Predictive distance-based mobility management for PCS networks." In INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, vol. 3, pp. 1377-1384. IEEE, 1999.

Lamia Osman Widaa received her BSc in Electrical Engineering from University of Khartoum, Sudan in 1996. She did her MSC in Telecommunication & Information Systems from department of Electrical & Electronic Engineering, faulty of Engineering, university of Khartoum (UofK), Sudan in 2002. Currently she is doing her PhD research in telecommunication & information systems at the faculty of Electrical Engineering, University of Khartoum (UofK), Sudan, under the supervision of Prof Dr.Sami Mohamed Sharif. Her current research interests are in the area of optimum design of Tracking Area Lists for LTE systems.



Sami Mohamed Sharif received his BSc in Electrical Engineering from Faculty of Engineering, University of Khartoum (UofK), Sudan in 1980. PhD degree in electrical engineering from UofK, Faculty of Electrical Engineering in 1988, and PhD title "Effect of Dust storms on Microwave signal propagation". Currently, he is a Professor with the Faculty of Electrical Engineering, (UofK), and the secretary of academic affairs in UofK. His current research interests are in Networking, Switching and teletraffic engineering, Radio wave propagation and Electromagnetic, Information Technology and Communications, Electronic Systems, ICT Economics and regulation.

Multimedia Communications: Technologies, Services, Perspectives

Part I. Technologies and Delivery Systems

Leonardo Chiariglione and Csaba A. Szabó

***Abstract*—This survey/position paper gives an overview of the state-of-the art multimedia communications technologies and services, analyses their evolution over the last decade, points out to their present significance and expected future role, and attempts to identify development trends. The paper consists of two parts. Part I deals with the technologies and systems for multimedia delivery. It covers the dedicated networks such as digital broadcasting systems and IPTV as well as the technologies of Internet based multimedia delivery. Networking issues including delivery over future Internet architectures and enabling technologies such as streaming and content delivery networks are dealt with in this part. Part II, to be published in the next issue of this journal, will address applications, services, and future directions.**

***Index Terms* — Multimedia communication, IP networks, Internet, mobile communications.**

I. INTRODUCTION

A decade ago, Stephen Weinstein and Alexander Gelman, recognized professionals in communications and media technologies, published a paper in the "Topics in Emerging Technologies" section of IEEE Communications Magazine, titled "Networked Multimedia: Issues and Perspectives" [1]. This excellent survey paper discussed the state-of-the-art of network infrastructures for carrying multimedia content, enumerated several existing and promising multimedia applications and proposed approaches that were supposed to lift the that time existing barriers on the way of the penetration of these applications and services. The authors said: "…resolution of several business models, public policy, and technical issues would enable a new era of networked multimedia services and, along the way, could revitalize the communications industry. It may take some time to get there, but we believe that the future broadband Internet, with both wired and wireless access, will carry the dominant mass market media services."

It is quite interesting to see where we stand now and what trends can be observed, after ten years since the paper was published, and, in particular, to address three questions: (i) how the networking infrastructures and services have developed, (ii) have the forecasted applications gained wide acceptance and implementations and (iii) are there any new trends not foreseen that time by Weinstein and Gelman. This paper attempts to answer these questions.

As for networking and services infrastructures, the authors stated: "Access networking is the bottleneck preventing us from using the optical core network to its full potential." Furthermore, "…the infrastructure for commercial quality audio/video streaming and interactive media communication is not yet in place." This paper discusses the progress that has occurred since then and tries to draw a necessarily high-level picture of the multimedia distribution and delivery networks and services of today and of the near future.

Let us refer to two other visionaries regarding the trends in multimedia networking:

Charles Judice, the father of JPEG, in his keynote speech [2], forecasted that digital storytelling could be a source of generating huge volumes of content on the Internet. Michael L. Brodie, that time Chief Scientist of Verizon, emphasized the rapidly growing user generated content [3].

The figures in recent forecasts for the expected growth of networked multimedia are really impressive. As an example, Intel said that there will be 12 billion connected devices worldwide in 2015, delivering 500 billion hours of TV and other video content. Note that the world population is expected to be around 7 billion [4].

Coming back to the forecasts by Weinstein and Gelman, they enumerated several that time existing or promising multimedia applications, including peer-to-peer exchanges of media materials, exchange of personal digital photographs and movie clips, web-based retailing of physical products, furthermore educational, government and medical services. In our paper, we address these, grouped into key application areas of networked multimedia, starting from entertainment applications through e-health, visual collaboration to smart city applications and services.

The rest of this paper is organized as follows. In Section II, we give an overview of multimedia coding techniques and standards that are of fundamental importance for digital video and sound broadcasting as well as for Internet-based multimedia delivery systems. Section III, titled "Multimedia delivery over dedicated networks" covers digital TV and

Submitted on May 14, 2014, revised June 17, 2014.

L. Chiariglione is with CEDEO.net, Via Borgionera, 103, 10040 Villar Dora (TO) Italy (e-mail: leonardo@leonardo.net).

Cs. A. Szabo is with the Department of Networked Systems and Services, Budapest University of Technology and Economics, Magyar Tudósok krt. 2, Budapest, 1117, Hungary (e-mail: szabo@hit.bme.hu).

sound broadcasting (Sub-section A) and IP-based TV distribution over dedicated networks, commonly called IPTV (Sub-section B). The underlying technologies are briefly dealt with and benefits from the point of view of both service providers and customers are addressed. Sub-sections C and D discuss the issues around mobile multimedia and media delivery over heterogenous networks. Sub-section E completes Section III by an overview of IMS – IP Multimedia Subsystem – that supports service development, implementation and provisioning in IP-based multimedia networked systems.

In Section IV, we discuss some networking and access technology issues (in Sub-section A) and enabling technologies (in Sub-section B) that support the dramatic move of media distribution, delivery and consumption from dedicated systems to IP-based networks and to the public Internet. First, networking aspects will be dealt with, trying to answer the question whether we will have a totally new Future Internet network infrastructure or several incremental steps are being accomplished to satisfy the requirements posed by multimedia applications, including 3D and mobile. Challenges of providing ubiquitous Internet access are addressed next. Then an overview of some enabling technologies will be given, namely media streaming and CDNs – Content Delivery Networks.

This concludes Part I of this paper. In Part II, we shall discuss the service aspects of TV broadcasting, IPTV and Internet TV, the role and specific forms of the social element in multimedia applications, key application areas of multimedia communications, and, in the last section, which concludes this two-part paper, we shall point out to some future directions.

II. ENABLING MULTIMEDIA TECHNOLOGIES: MULTIMEDIA CODING

Studies of digitisation of multimedia information – essentially audio and video – started at the instigation of the global multi-decade plan hatched by telecommunication operators to convert their copper-based analogue networks to digital first and fiber optics-based networks later.

In the mid-1970s, European Action 211 of COST Area 2 Telecommunications became the focus of video coding activities that led to the development of a 1.5/2 Mbps videoconference codec that used DPCM and Conditional Replenishment and became the basis of the ITU-T Recommendation H.120. Later on, COST 211 became a major contributor to H.261, another video-related ITU-T recommendation for px64 kbit/s video coding (p-1, …, 30) that used a more sophisticated and efficient linear transformation with motion compensated prediction algorithm.

ITU-T was also involved in speech coding since the early 1960s. The first standard in this area – G.711 – has two non-linear quantisation characteristics that take into account the logarithmic sensitivity of the ear to the audio intensity. Since then, ITU-T and other telecommunication-related standards organisations have continued producing speech coding standards.

With the appearance of MPEG, multimedia coding has become a high-profile area of endeavour, standardisation and exploitation. In its 25 + years of activity MPEG has produced five major generations of video coding standards and has pushed forward the frontiers of video coding performance.

At the target bitrate of 1.5 Mbps, MPEG-1 Video yields a quality comparable to the VHS cassette (comparison is made with the analogue version of video used at that time). The quality of MPEG-2 Video, measured in 1995, showed that at 6 Mbps the quality was indistinguishable from the composite (PAL or NTSC) original and at 8 Mbps the quality was indistinguishable from the component (YUV) original. The first deployments used a bitrate of 4 Mbps but the current operational bitrate is at 2 Mbps with approximately the same quality. In 1998, 4 years after approval of MPEG-2, MPEG-4 Visual yielded a reduction in bitrate of about 25% and 5 years later MPEG-4 Advanced Video Coding (AVC) yielded a further reduction of 30%. Finally, the latest MPEG video compression standard approved in 2013 yielded an astonishing 60% reduction in bitrate compared to AVC. Note that the H.264 standard specified in ITU-T is identical with MPEG-4 AVC. The two specifications are maintained jointly by MPEG and the Video Coding Experts Group (VCEG) of ITU-T. MPEG-H HEVC, too, has been developed jointly withVCEG, and it has the name H.265 within the family of ITU-T standards.

Compression is an important dimension because the spatial – but partly also temporal – resolution of video continuously increases. MPEG-1 Video was designed to work particularly for ¼ of the spatial resolution of regular television, MPEG-2 for standard definition (even though in the USA it was deployed for Digital Terrestrial Television HDTV). MPEG-4 AVC is typically used also for HDTV and the latest HEVC standard is poised to take over the so-called 4k (i. e. about 4000 pixels per line) application field.

However, the video application fields are manifold. In some cases scalability – i.e. the ability to extract meaningfully decodable sub-bitstreams from a bistream, e.g. 1 Mbps from a 2 Mbps bistream – is required. MPEG has continued working on this aspect of the video coding field for many years with increasingly better results. The MPEG-2 Video and MPEG-4 Visual scalable video compression modes save 10% of the bitrate compared to "simulcast" (i. e. transmitting two individual non-scalable bitstreams). In other terms, if the application needs two bitstreams one at 1 Mbps and another at 2 Mbps, the scalable coding mode enables the transmission of a single scalable bitstream at 2.7 Mbps. This is probably not a sufficiently high gain to justify the use of a scalable mode, but the AVC and HEVC scalable modes offer a saving of 25%. In the example above, instead of 2 bitstreams at a total bitrate of 3 Mbps the scalable bitstream has just 2.25 Mbps.

In other application domains the transmission of two signals from two slightly separated cameras are used to provide a stereo image at the receiver. This has been done in several attempts at deploying "3D TV services" by simply transmitting two separately encoded bitstreams. Starting from MPEG-2 Visual, however, MPEG has provided a "stereo mode" that saves up to about 15% for MPEG-2 and MPEG-4 Visual and up to about 25% for AVC and HEVC. The comparison for the

last case can thus be between 2 bistreams at 2 Mbps each for a total of 4 Mbps against a "stereo bitstream" at 3 Mbps.

3D Video is a world in itself whose surface MPEG has barely started scratching. Another technology to represent a 3D Video is "Texture + Depth". In this case every pixel of an image have the usual RGB or YUV values and are supplemented by a value that represents the distance of the pixel on the camera from the object that creates the pixel. This technology has only been applied to the more recent AVC and HEVC standards and offers an additional 20% saving compared to the stereo mode. Still in this space another possibility offered by MPEG standards is the ability of a user at the receiving end to define an arbitrary viewpoint of the scene and to use the available information to synthesize the missing image. Obviously this functionality entails an increase of the bitrate – minimal, at the cost of 5-10% more bitrate.

It should be noted that there is no absolute value in the numbers reported above, just a rough statistical and usually subjective assessment of the performance of the algorithms on which the standards are based.

So many things are common but also so many things are different in the field of audio, a word that is in this paper is used to mean "music".

The first MPEG attempt in the stereo audio coding field was MPEG-1 Audio (a standard approved in 1992) with a choice of 3 versions ("layers") of the standard: Layer 1, used for the now defunct Digital Compact Cassette (DCC); Layer 2, used for terrestrial, satellite and cable set top boxes; and Layer 3 soon christened as MP3, an acronym that needs no introduction. Tests carried out in 1992 showed that the 3 layers offered a "quality subjectively transparent with the original" at 384, 256 and 192 kbps, respectively. The 192 kbps of MP3 is a reference bitrate: transparency can be achieved at a higher bitrate or at a lower bitrate, depending on how "smart" the encoder is in exploiting the characteristics of the human hearing system.

The second attempt began with the extension of MPEG-1 Audio to multichannel, a kind of "bottom-up" scalability because the new multichannel audio coding had to contain the already defined MPEG-1 Audio stream. This did not provide sufficiently attractive results, so a new MPEG-2 Audio standard – Advanced Audio Coding (AAC) – was designed focused on providing broadcast quality performance for 5-channel music signals at a total bit rate of 320 kbps. This standard was further developed as MPEG-4 AAC which provides subjective transparency at 128 kbps and excellent performance down to 48 kbps. The MPEG-4 High Efficiency AAC (HE AAC) uses Spectral Band Replication (SBR) which encodes the lower frequency part of the spectrum using a waveform coder and reconstructs the high frequency part by transposing the lower frequencies. HE AAC further improves performance at lower bitrates.

Another MPEG Audio coding standards developed more recently is MPEG Surround which encodes multi-channel audio by adding a low-rate side-information channel to a compressed stereo or mono audio program. A stereo/mono player receiving an MPEG Surround bitstream still produces a useful output while new-generation players can produce the full multi-channel experience. Another MPEG Audio coding standard is Spatial Audio Object Coding (SAOC) which allows access to individual audio objects (e.g. voices, instruments, ambience etc.) in an audio mix, so that listeners can adjust the mix to suit their personal tastes. Finally Unified Speech and Audio Coding (USAC) achieves consistently state-of-the-art (as of 2011) compression performance for any arbitrary content composed of speech, music or a mix of speech and music in the sense that it provides better performance than individual codecs designed for either speech or audio and significantly improves state-of-the-art performance at bit rates ranging from 8 kbps for mono signals to 32 kbps for stereo signals, and for bitrates to 64 kbps for stereo and beyond.

The latest standard still under development is 3D Audio, an MPEG Audio coding standard suitable for all scenarios – such as in home theater, automotive, headphones connected to a tablet/smartphone – where a multi-channel audio program (e.g. 22.2) needs to be compressed and rendered to a number of loudspeakers that is not necessarily the same as used at the source.

The objective of this section was to cover video and audio standards developed within the MPEG community. Let us finally mention other audio compression formats, first of all the Dolby Digital technology, a.k.a. AC-3, which is widespreadly used in DVD and Blu-ray players and in digital broadcasting.

III. MULTIMEDIA DELIVERY OVER DEDICATED NETWORKS

Media delivery and consumption is in the process of transition from using dedicated – vertically integrated – systems, namely the radio and TV broadcast networks, through dedicated and managed IP networks, to the public Internet. This section deals with digital TV and sound broadcasting systems, and IP-based TV distribution over dedicated networks, commonly called IPTV. In this section, we will also discuss the issues around mobile multimedia and media delivery over heterogenous networks. Finally, the IMS – IP Multimedia Subsystem – that supports multimedia service development, implementation and delivery will be introduced.

A. Digital broadcasting systems

1) Digital television systems

The advantages of digital TV broadcasting, in comparison with the old analogue broadcasting, are obvious for all stakeholders. Broadcasters can broadcast more TV channels without having to buy new frequency bands. Regulators and governments can sell the bandwidth freed up by the digital switchover, the so-called digital dividend. And, last but not least, consumers get improved video quality, also in wide screen (16:9) format, mono, stereo and surround sound, several audio tracks plus new features and services (subtitling, EPG – Electronic Program Guide, interactivity…). The price the customer pays for these new features and services is not really significant as most new TV sets are already digital ones

and set-top boxes for analogue sets are inexpensive, although this may be a problem for low-income population groups. To help them, governments usually implement various support programs.

The history of digital TV broadcasting started about a decade ago, when, in 1993, the satellite system, DVB-S [5], shortly thereafter, in 1994, the cable system, DVB-C were standardized [6]. In 1996, FCC adopted the ATSC (Advanced Television System Committee) standard for digital television broadcasting in the USA. About the same time, in 1997, the ISDB (Integrated Services Digital Broadcasting) standard was adopted in Japan. In 2000 DVB-T, the terrestrial system was born [7], followed by the mobile version, DVB-H in 2004. During the years from 2005 through 2010 the 2nd generation of DVB-X standards were established: DVB-S2 (2005), DVB-T2 (2008), and DVB-C2 (2010) [8].

Digital television systems are rather interesting from the technology point of view because of the sophisticated communication and coding technologies used to take into account the specific properties of the satellite, cable or terrestrial channels. The common elements of all three systems are as follows.

• Transport stream (MPEG-2 TS). The input of the systems is the audio/video transport stream, coded and packaged according to the MPEG-2 standard, see e.g. [9].

• An energy dispersal module. This unit, also called scrambler or randomizer, is used to generate a flat spectral density and to eliminate long sequences of "0"s and "1"s, by pseudo-randomising the MPEG-2 TS packet stream.

• FEC module, also called "outer FEC", since, in DVB-T system, a second FEC module, called "inner FEC" is used. It applies a Reed-Solomon code with error correcting capability of 8 symbols in a 204-symbol MPEG2-TS packet.

• Interleaver. The purpose of this unit is to rearrange the bytes in order to randomize the channel errors and improve the error-correcting capability of the Reed-Solomon code. It uses a convolutional interleaver of depth 12, that increases the error correcting to approx. 12x8=96 symbols (bytes).

In the three digital broadcasting systems, different transmission methods and additional error correcting modules are used to take into account the different nature of the transmission channels in the three cases. In the satellite channel, only attenuation and thermal noise (AWGN) plays role, there is no multipath propagation, and the bandwidth is not as limited as in the case of the other two systems. In cable systems, the bandwidth per channel is more limited. The terrestrial transmission channel is the most challenging one, with noises and interferences and multipath propagation.

Fig. 1 shows a conceptual block diagram of the three DVB systems.

The digital TV systems in North America (ATSC) and Japan (IMDB) are built along the same principles, for a comparison see the textbook [10] and the recent survey paper [9].

In the second generation digital TV standards, further improved transmission and coding techniques have been incorporated. For example, in satellite systems the main goal was to increase the data throughput in a given bandwidth (to increase the spectral efficiency). In the terrestrial system similar goals were set and modifications carried out. In cable systems, OFDM (Orthogonal Frequency Multiplexing) technique was incorporated instead of the single-carrier modulation schemes.

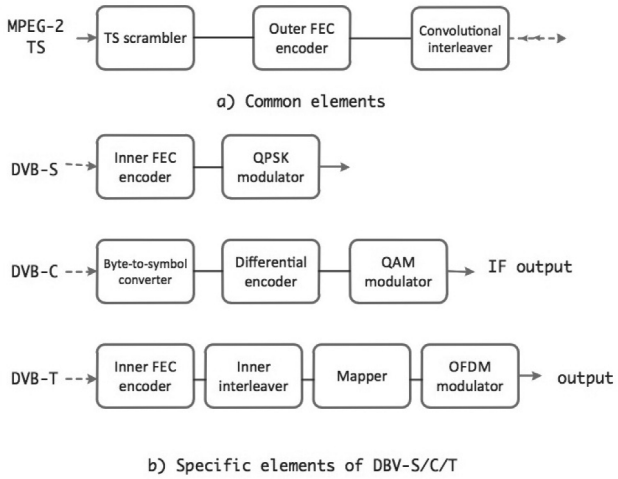


Figure 1 Conceptual block diagram of the three DVB systems

Currently the different countries around the globe have either already completed the switchover (true for most of the developed countries), or are in the process of completing it. Most of European countries completed the transition during the last years. In the United States, the switchover took place in 2009, in Australia and New Zealand in 2013. Mexico and Turkey will be among the last ones with planned switchover in 2015.

Finally let us mention the interesting member of the DVB-X family, the DVB-H (Digital Video Broadcast to Handsets), see e.g. [11]. While DVB-T was designed for use for living-room TV sets with rooftop antennas, DVB-H extends this terrestrial service to handheld devices. The technology is based on that of DVB-T which has been modified to take into account the specific properties of handheld devices, mainly the power consumption requirement but also the smaller screen and antenna, mobility and the like. The first commercial DVB-H service in Europe was introduced during the Football World Cup in 2006. After an initial fast growth of subscriber numbers (in particular in Italy where there were more than 1 million users in 2009) a decline followed and the DVB-H broadcasting was terminated in several European countries during 2000-2012.

2) Higher resolution or more dimensions in television? HDTV, 3D and beyond

During the last few years, the HD quality, meaning 1920x1080 pixels ("full-HD"), has become ubiquitous in entertainment industry, in digital cameras, TV sets and digital television broadcasting (HDTV). It is incorporated in coding standards such as MPEG-2 and MPEG-4, monitors and TV sets are now HD-capable and most TV programs are being

broadcasted in HD format. The next step seems to be a more recent 4k (and 8k) technology also called UHDTV (adopted by CEA – Customer Electronic Association, USA - in 2012) and Super Hi-Vision (introduced by NHK in Japan). EBU, the European Broadcasting Union calls this new technology UHD-1 and UHD-2. This technology, providing a resolution of 3840x2160 pixels (thus almost four thousand pixels in horizontal direction, hence the notation 4k) and 7680x4320 pixels (8k), was integrated first in monitors and projectors starting from 2011, then in TV sets starting from 2013. An ITU-R Recommendation was approved in 2012 [12].

3D broadcasting technology has been around for several years, most TV sets in the market are 3D-capable (to be viewed with polarized glasses) and several broadcasters started 3D trials. For instance, BBC began a two-year 3D trial in 2011, and broadcasted several shows and events in 3D, including the Olympic Games. Half of the estimated 1.5 million households in the UK with a 3D-enabled television watched the opening ceremony of the 2012 Olympic games in 3D [13]. However, BBC has recently postponed the trials, and will make no further 3D programmes for 3 years. In the USA, ESPN have decided to suspend the use of 3D technology for broadcasting. The Australian Pay-TV operator Foxtel has also terminated its dedicated 3D broadcast channel [14].

Why 3D TV (based on current technologies) is not breaking through? Reasons include the viewing inconvenience due to the need of wearing glasses, and the sometimes not adequate image quality. Also 3D has added value only for a few genres, and the content offering is far from satisfactory. Why, on the other hand, it seems that ultra-high resolution 2D TV could eventually break through? It clearly offers enhanced viewing experience without a discomfort caused by a supplementary device (the 3D glasses), provides larger field of view, it is 2D but nevertheless offers a better sense of realness, and causes less fatigue for the eye and brain. The picture may change in few years from now when glassless 3D TV technology becomes available for public.

3) Digital sound broadcasting

According to ETSI, "Digital Audio Broadcasting (DAB) was conceived as a means of digitizing audio programmes in order to offer distortion-free reception and CD quality sound" [15]. Digital sound broadcasting standards include DAB, its more recent variant DAB+ and DMB. For a comprehensive treatment of digital radio broadcasting, refer to [16], and for up-to-date information, visit the website [17].

DAB specifies sound broadcasting with MPEG Audio Layer III (MP3) coding and DAB+ sound broadcasting with MPEG-4 (AAC) coding. DMB is about adding video/multimedia capabilities to audio broadcasting thus allowing DAB to become a digital mobile television platform. All three have the same physical layer just the transport etc. protocols are different and they offer different services. The main operating frequency band is VHF III (174-230 MHz / 240 MHz in some countries). In this band, a large area can be covered with an external antenna and good penetration into buildings can be achieved. L-band (1452-1479.5 MHz) is used in some countries where Band III is not available yet or as the supplemental broadcasting band. In these frequency bands, no external antenna is needed which is an advantage particularly for mobile phones). This band is usable in urban areas where good reception can be achieved even in non-line-of-sight conditions. However, penetration into the buildings is limited and reception inside can be bad.

Advantages of digital sound broadcasting for consumers are CD quality, possibility of mobile reception, and enhanced receiver features. For operators and regulators, the advantages are spectral efficiency as compared with analogue broadcasting and lower transmitter power. Standardisation in Europe is well established.

In spite of these advantages, digital sound broadcasting is penetrating in a much slower pace than digital television has been. No country has done a complete switch-off of FM radio stations yet. Norway is the closest to that, it was announced that there will be 99.5% coverage in 2014, and that Norway was planning a switch-off of FM radio in 2017. There are signs of penetration in other countries as well. In the UK, 46% of households have DAB and the national coverage is 94%. 44% of new cars are equipped with digital receivers. Germany plans full national coverage by 2014 [18], [19], [20].

Let us finally mention DRM – Digital Radio Mondiale, which has been designed specifically as a high quality digital replacement for current analogue radio broadcasting in the AM and FM/VHF bands [21]. There is no significant penetration, many countries in Europe started then stopped their trials and did not launch commercial DRM broadcasting.

In spite of the standardization efforts in the aforementioned organizations and introduction plans in various countries, the future of digital sound broadcasting is at least unclear. Users can listen to a large amouont of radio stations on the Internet (we shall come back to this issue later), and as music is the primary genre in radio broadcasting, downloading MP3 songs from the Internet and enjoying them on mobile devices is just enough for most listeners.

B. Multimedia distribution over dedicated IP networks: IPTV

According to ITU-T Focus Group: „IPTV is defined as the service delivery of video/audio, text, graphics and interactivity over IP based networks managed to provide the required level of QoS/QoE, security and reliability".

IPTV is an opportunity for "classical" telecom operators to enter into the broadcasting business. Since they already play the role of an ISP by providing Internet access, typically over their xDSL networks, by adding TV they become a "triple play" provider of TV+Internet+Telephone services. IPTV offers services such as interactivity, time shifting (playback after the initial broadcasting of the content), VoD – Video-on-Demand - content consumption, program recording, and EPG – Eletronic Program Guide. The latter is an electronic program that allows intelligent selection and sorting of programmes as well as obtaining all kind of information about specific programs.

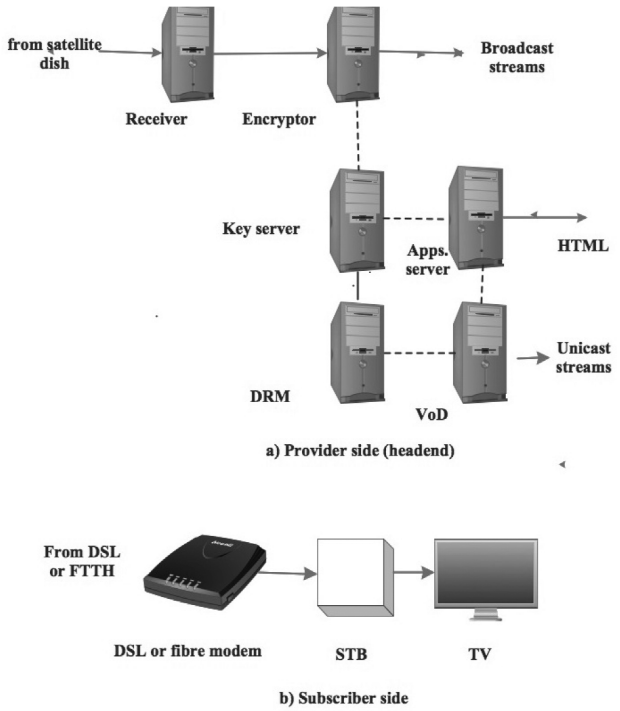


Figure 2 Functional diagram of an IPTV system

The technical aspects of the IPTV service are illustrated in Figs 2-3. On the functional diagram of Fig. 2, the headend (a term borrowed from cable TV systems) is where the content is collected and processed. Content can be live TV programs from a satellite or terrestrial distribution network, or can be a stored one from local media servers. Live or stored video then coded/transcoded, encrypted and transmitted to clients. Electronic Program Guide support is also part of the headend. The client side functional unit is the set-top-box (STB) which performs the media decoding, decryption, EPG client functions.

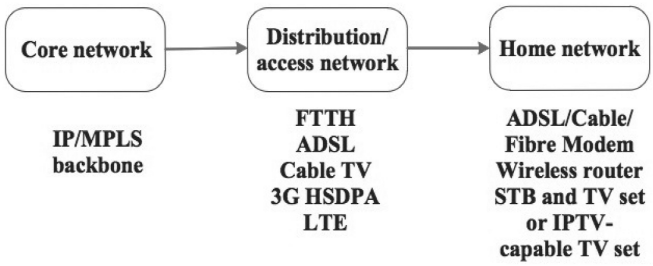


Figure 3 The IPTV delivery network

Fig. 3 shows the high level networking infrastructure used for IPTV delivery. The core part is the IP/MPLS backbone of the telecom service provider. The access network or first/last mile network is mostly xDSL or FTTH – Fiber-to-the-home, but it can be a Cable TV distribution network or a broadband wireless network (3G – HSDPA, LTE or WiMAX) as well. The home network usually consists of an ADSL modem, a wireless LAN router, IPTV-capable TV set with a set-top-box or with built-in IPTV capability and additional client devices.

The IPTV protocol architecture is shown in Fig. 4. The media stream coming from the application layer is coded into PES – Program Elementary Stream according to the MPEG standard, then it is packaged into MPEG Transport Stream packets (the same are used in digital television standards). Media transport is supported by RTP – Real Time Transmission Protocol that provides sequence numbering and time stamping services. RTP packets then carried in the payload of UDP – Universal Datagram Protocol packets. The protocol overhead added to the 188 bytes long MPEG TS packets is total 40 bytes plus the MAC/PHY overhead. For an extensive treatment of IPTV technology, see the textbook [22] and the paper [23].

To meet Quality-of-Service requirements and Quality-of-Experience expectations of the customers, a series of technical challenges have to be addressed. An IPTV system itself is a pretty complex one, so even if the input stream is ok, which is not always the case, sources of quality deterioration can be the failures in the core network (rarely), in the distribution and access networks (more frequently) and of course within the subscriber's home network. From the customer point of view, all this should be the service providers responsibility, however, the latter is not in the position of managing all the aforementioned components from a central place. (E.g. media streams are often sourced from third parties.)

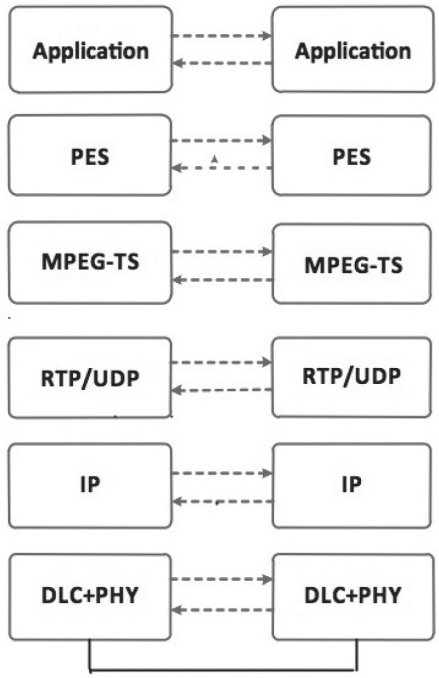


Figure 4 IPTV protocol architecture

Coming back to the customer side: What can IPTV offer (compared with digital TV broadcasting)?

- The same high quality picture and sound as in digital broadcasting.
- Time Shift - allowing playback of content after its initial transmission.
- EPG or Electronic Program Guide.

- Personalized interactive media consumption in many ways (recording programs, video on demand, alert messages for favorites programs etc.).
- Communication (video conferencing) and online training service.

In classical television broadcasting, there have been improvements in picture quality (HDTV), in the channel offering (multiplexes in digital broadcasting), however it remained basically a one-way distribution vehicle from the service provider to the end-user with a very limited interactivity. At the same time, a large share of TV users, in particular the younger generations, having already accustomed to the freedom when consuming media, including TV programs, on the Internet, are no longer satisfied with what the traditional TV broadcasting systems offer. For them, IPTV might be attractive.

C. *Mobile multimedia*

Providing multimedia services, such as the distribution of TV programs, for mobile users satisfies the growing demand for accessing these services anytime, anywhere and on any device. Mobile multimedia refers to transmission and delivery of multimedia information to mobile customers who access the Internet via cellular mobile services. Because of the specific properties of the wireless mobile channels - high error rate and packet loss rate, lower bandwidth and bandwidth depending on location and the heterogeneity of access networks and user devices -, serious technical issues have to be solved, including coding and presentation of multimedia content for mobile devices, end-to-end error control, multicast transmission, mobility management and other network-related issues. For example, the H.264 multimedia coding standard provides specific coding technique called Flexible Macroblock Ordering to cope with error propagation and error accumulation. Scalable Video Coding (SVC), an extension of H.264/MPEG-4 AVC video compression standard, provides adaptation of the coding rate to the estimated bandwidth of the wireless channel [24].

TV broadcasting to mobile devices requires coding formats suitable for mobile screens (QCIP, CIF, QVGA resolution), although there will be more and more devices with enhanced resolution (full HD), thus adaptation to the aforementioned formats might become unnecessary in the future. In addition, mobility management is needed even at high speeds (usage in cars on motorways), and multicast transmission is required. As for the latter, the 3GPP Release 6 standard includes a service called MBMS – Multimedia Broadcast Multicast Service, which is a general point-to-multipont service for IP packets offering data rates up to 256 kbps. Subsequent releases extend it for 3G/HSDPA and 4G/LTE mobile cellular services.

Lastly, ensuring mobility needs sophisticated methods and protocols starting from mobile IP at the network layer, through transport layer mobility protocols to solving mobility in application layer using SIP, the Session Initiation Protocol. A specific case is when the user moves accross wireless and mobile networks that are based on different technologies. The handover between cells in this case is called „vertical handover“, to distinguish the task from the usual handover when the user moves across cells of the same mobile cellular network. Let us briefly explain it by the example of a past project that was carried out by one of the authors and his team.

D. *Media delivery over heterogenous networks*

In a multi-platform access network environment, the user has several physical connections to access the Internet, hence the same resource could be accessed via different wireless networks, even simultaneously. This opportunity could be utilized to achieve higher quality service, i.e. faster download or higher quality media streaming solution by using all access networks simultaneously or selecting the best access network(s) dynamically. On the other hand, the available wireless access networks have quite different characteristics and properties such as average and peak bandwidth, availability, delay and jitter, packet loss rate and bit-error rate, optimal packet size, and pricing. Furthermore, these properties usually depend on the actual state of the network and on the user's location. In the media streaming architecture outlined in [25], a best-effort single-connection scheme is used i.e. the media streaming system uses the best connection (active connection) to transmit the media stream and avoid the other (idle) connections. In the single connection scheme, the moment of the handover (namely the change of the active connection) must be invisible to the user and he or she becomes aware of the handover only by observing a degradation or improvement of the media quality, depending on the characteristics of the earlier network connection and the new one. The decision on the switching of streams is based on the client's measurements. Based on the measured parameters (current packet loss rate and the access network type), the optimal bandwidth is estimated, the ranking of the access networks are made, and the best bandwidth/quality version of the content is determined and the switching is carried out in case of need. To accomplish this, the media server should provide the same media content in different resolutions continuously to allow the system to choose the appropriate resolution according to the quality of the active connection and the properties of the client device.

The media streaming architecture in [25] has the following key features:

- Vertical handover among different access networks, including 2G and 2.5G technologies (GSM, GPRS, EDGE), 3G cellular (UMTS), WLAN (Wi-Fi), WMAN (WiMAX) and even some wireline access such as xDSL.
- Horizontal handover, i.e. handover between the same kinds of wireless networks of different service providers.
- Content- and environment-adaptive charging, accounting, billing and payment schemes.
- Digital rights management schemes.

The generic system architecture is shown in Fig. 5. In the testbed, the UMTS and GPRS/EDGE access networks belonged to the same service provider, whereas the WLAN, WMAN and xDSL access networks were provided by a different operator. The xDSL wireline network was accessed via a Wi-Fi wireless access router.

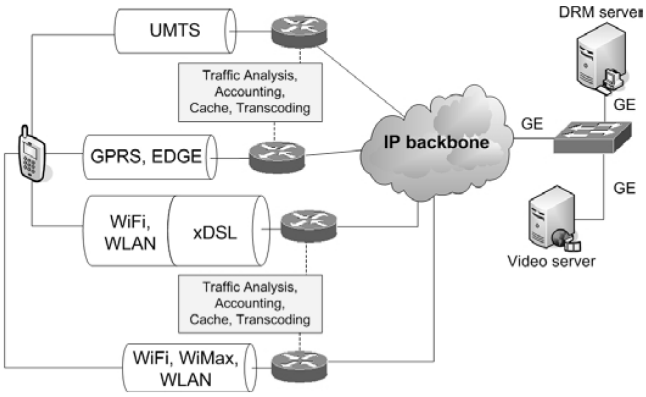


Figure 5 The generic scheme of the media streaming testbed [25]

E. Supporting service development, implementation and delivery: IMS

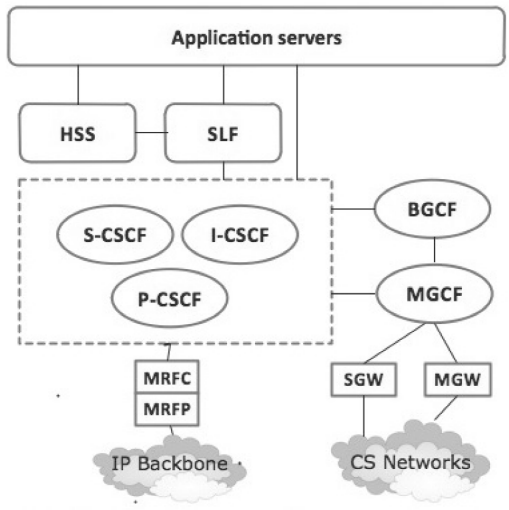
Starting from the introduction of VoIP – voice over the Internet Protocol – in mid-late 90s, telecom service providers – both incumbents and new ones – have been gradually moving from circuit-switched to packet switched voice services. While in the first VoIP systems the signalling/session control protocol was ITU's H.323 [26], SIP or Session Initiation Protocol, developed within IETF [27], emerged almost in parallel. While their functionalities are similar, SIP is a more flexible and better scalable protocol that can be easily integrated into web-based applications. At this point SIP seems to be the future. In the process of the development of newer versions of mobile communications systems, on the one hand, and moving towards a new concept of NGN or Next Generation Networks, on the other hand, it turned out that however important the session control can be, it is just one of the functionalities needed for supporting the development, implementation and provisioning of multimedia services over packet switched networks. Therefore, in the standardization body of the mobile world, 3GPP - Third Generation Partnership Project, a more complex new element of the network architecture, incorporating also SIP, called IMS - IP Multimedia Subsystem – was specified in their Release 6 [28].

The need for such a functionality has also emerged in the telecommunication world within the context of NGN standardization in ETSI TISPAN. (ETSI TISPAN – Telecommunications and Internet converged Services and Protocols for Advanced Networking - has been the key standardization body in creating the NGN - Next Generation Networks specifications.) NGN represents a paradigm shift from the classic telecom service model of independent, vertically integrated networks to a new architecture that comprises a variety of access networks and has a new horizontal layer or platform that supports service provisioning with important functionalities such as call control, quality of service provisioning, media gateways, authentication, authorization, and accounting (AAA) and the like. This new architecture allows telecom companies to successfully compete with Internet-based services, and in general, supports the convergence of the Internet, telecommunication and media industries. On the other hand, the horizontal separation of functionalities in telecom networks allow third parties to come and put their services on top of the network infrastructure of network service providers, see e.g. [29]. IMS was standardized both in the mobile world and telecom world, another big step in the process of their convergence.

IMS entities and key functionalities include:

- Session management and routing, based on SIP – the Session Initiation Protocol.
- Databases (like HLR – Home Location Register - in cellular mobile systems),
- Interworking elements (e.g. media gateways).
- Application servers and services, e.g. AAA – Authentication, Authorization and Accounting - based on Diameter protocol.

The IMS architecture is illustrated in Fig. 6. As it is shown on the top of the block diagram, the IMS system supports interfacing with legacy mobile call control systems as well as interworking with non-IP networks such as with the circuit-switched PSTN. To go a bit more into the IMS system, let us briefly mention the functionalities of its building blocks. The three CSCF – Call Session Control Function – nodes implement the SIP session control protocol. P-CSCF or Proxy-CSCF is the closest to the user agent and functions as SIP proxy server. I-CSCF or Interrogating CSCF determines the route of a call to the called UA, while S-CSCF or Serving CSCF serves the UA. These units communicate with HSS or Home Subscriber Server (identical to Location Server in SIP) and with SLF or Subscriber Location Function. BGCF or Breakout Gateway Control Function handles calls originated by the IMS and destined to PSTN. MGCF or Media Gateway Control Function takes care of the interworking process, while



HSS - Home Subscriber Server, SLF - Subscriber Location Function
S/I/P-CSCF - Serving/Interrogating/Proxy Call Session Control Functions
BGCF - Breakout Gateway Control Function
MGCF - Media Gateway Control Function
SGW - Signalling Gateway, MGW - Media Gateway
MRFC/MRFP - Multimedia Resource Function Controller/Processor

Figure 6 High-level architecture of IMS

MDW or Media Gateway carries out the necessary media conversions. Finally, the units MRF or Multimedia Resource Function and MRFC, MRF Controller, and MRFP, MRF Processor handle multiparty calls such as for multiparty multimedia conferencing.

Additional services, provided by IMS and could not be illustrated in Fig. 6, include the already mentioned AAA or Authentication, Authorization and Accounting, as an example. For an extensive treatment of IMS architecture and services, see [30], [31], [32].

Based on the services provided, IMS can also be considered as a multimedia SDP or Service Delivery Platform, offering the necessary support for multimedia services to be provided by telecom operators or third parties. SDP however is a term more often used in a broader sense, and denotes facilitation of service composition and integration, so that IMS can be considered as an additional layer on top of SDP.

In spite of the relative maturity of IMS and the potential advantages it can offer, its penetration has been so far slower than one might have expected. At the beginning the majority of significant telecom operators have purchased IMS systems from leading vendors such as Ericsson, Nokia or Huawei, primarily for testing purposes. There are several reasons why the commercial deployment has been not so fast. One of them might be that mobile operators are going pretty well without it and are reluctant to make a significant investment. It looks like that the advantage of IMS we mentioned above, i.e. that it is an unified platform for developing, deploying and providing multimedia services over IP networks, and that the operator can do it more efficiently using the "toolset" IMS provides, have not been transformed into specific business benefits so far. Also operators seem to be not too enthusiastic attracting third parties to bring their services and putting them on the operators network. The driving force will apparently be some new services that only IMS offers and that can immediately generate revenues. These services include push-to-talk, presence, multimedia sharing, emergency calls etc. Without them, the future of IMS will be unclear.

IV. Moving from Dedicated Networks to the Internet

A move from dedicated and managed IP-based networks to the public Internet seems to be simple, since the communication protocols of the TCP/IP stack are common, but it is a huge step at least in two aspects. One, delivering broadcasting content over the public Internet represents challenges in terms of ensuring access bandwidth, reliability, quality of service and like. Two, specific distribution and consumption models, including business models, arise. In this section, the networking aspects will be addressed.

A. Networking and access issues

1) On the architecture of the Future Internet

Will there be a radically new architecture? "Clean slate" or "evolutionary" design shall be followed? What shall be the design requirements and principles of the Future Internet, in particular of the Future Media Internet? How will this new architecture relate to the already existing and standardized in 3GPP and ETSI TISPAN NGN architecture?

These and similar questions have been posed and answers sought by several projects and working groups, labeled by the term "Future Internet" or FI, around the world, supported in particular by NSF in the USA and EU research framework programs in Europe. NSF launched its FIA – Future Internet Architecture program in 2010 and funded four projects [33], [34], then launched the second round of in 2013. In EU, the "Future Media Internet - Think Tank (FMIA-TT) supported by the nextMEDIA project aimed at working out a reference architecture model of the "Future Media Internet", "covering the delivery, in the network adaptation/enrichment and consumption of media over the Future Internet ecosystem" [35]. According to the leading professionals teamed together in this project, the existing Internet architecture should be replaced by a new three-layer one. In this hierarchical FMI architecture, the lowest layer is the Service/Network Provider Infrastructure Overlay. This is where the users who are both Content Producers and Consumers (therefore called "Prosumers") are located. They are connected through the infrastructure of the ISPs and network service providers. The nodes of this infrastructure have limited functionality and intelligence. The second layer is the Distributed Content/Services Aware Overlay, contains content-aware network nodes which are more intelligent as compared with the infrastructure nodes and are capable of identifying and qualifying content and services and reporting to the third layer of the architecture (Content/Services Information Overlay). It consists of intelligent nodes or servers that have a distributed knowledge of the locations and cashing of the content and of the conditions in the network. Based on this information, decisions can be made e.g. on the optimal delivery of content to the subscribers. We should note, however, that while introducing content aware network nodes and layers is certainly a good approach to the building of the "Future Media Internet", it somewhat contradicts to the network neutrality principle currently required from the ISPs and network service providers.

By now it has become clear that there will not be a radically new FI architecture. However, new approaches, design and improvements are needed in areas including:

- New networking protocols, in particular cross-layer solutions.
- Efficient methods to handle multimedia traffic which is already dominant and continues to grow.
- Ensuring throughput, Quality of Service, Quality of Experience.
- Providing access from anywhere, from any device, with the desired quality to users who are prosumers, that is consumers of media as well as creators of content.
- Ensuring seamless mobility, between arbitrary network technologies and systems.
- Adaptivity to the capabilities of user devices and network, ensuring the desired quality.
- Meeting the requirements of the Internet of Things towards the network, e.g. wireless (multimedia) sensor networks, with self-organizing capabilities.

Techniques and solutions outlined in the sub-sections to follow address some of these challenges.

2) Challenges of providing ubiquitous Internet access

When discussing multimedia services, it is often assumed that access to the public Internet is available everywhere with the desired speed („bandwidth") and quality of service. In a NSF study, we can read: "Historical infrastructures – the automobile/gasoline/roadway system, electrical grids, railways, telephony, and most recently the Internet – become ubiquitous, accessible, reliable and transparent as they mature." [36] While it is true for some historical infrastructures, ubiquitous access and reliability certainly cannot be taken for granted in the case of telecommunication networks and the Internet. And we are not talking about developing countries only and their under-developed regions, where providing just basic telecom access presents a huge problem. Ensuring broadband access to everyone and everywhere is also a challenge in developed countries because relying merely on market economy cannot solve this problem. Telecom and Internet companies operate according to their specific business models, which do not allow expanding their infrastructures to sparsely populated and/or geographically challenged areas, therefore, these areas remain underserved. This is one of the manifestations of the so-called "digital divide", a gap between those having proper Internet access and those who do not. Therefore, providing broadband access to citizens, communities, public institutions and developing businesses has become a strategic objective for state and local governments worldwide. A large number of initiatives, under the collecting name "community networks" or "municipal wireless" have been launched in North America as well as in Europe (see [37], [38], [39], [40]). By creating telecom infrastructure in underserved regions, local governments can prevent remote communities from digital divide, and are able to create a healthy climate for economic development, can help startups grow, and bring new businesses into the region. Often cited examples include the municipal network pioneer city of Corpus Christi, TX in the USA or the more recently deployed municipal network in Barcelona, Spain and the large scale network of the Province of Trento in Italy [41]. Solving the digital divide issue by building and operating city-wide or regional network infrastructures, local administrations create possibilities for advanced multimedia services such as telemedicine, e-learning applications, portals for tourists, regional TV channels, surveillance systems and the like, thus bringing additional benefits to the citizens and businesses as well as making these networks sustainable.

B. Enabling technologies for the implementation and provision of multimedia services

1) Streaming techniques

Audio/video streaming or multimedia streaming has been around for quite a long time and is today perhaps the most important technology component in networked multimedia applications and services. Its history started in 1995 when Real Networks launched RealAudio then RealVideo in 1997. In 1998 Apple announced QuickTime Streaming. A decade later, in 2007, Hulu launched its streaming service, offering ad-supported streaming video of TV shows and movies from many networks and studios. Today there are several thousands of TV stations available online on the Internet. In 2013 YouTube reached one billion monthly users with 4 billion views per day. Today, the most commonly used streaming technologies are Microsoft's Windows Media [42], RealNetwork's RealPlayer [43], and Apple's QuickTime [44].

Multimedia streaming is a technology that enables clients to download audio/video files from servers and to start viewing them immediately without waiting for complete download, and continue viewing without interruption. In addition, the user is provided with some DVD-like functions such as pause, resume, fast forward, rewind, etc. Key elements of the streaming system are playout buffer on the client side, protocols ensuring or supporting quality of service and specific protocols for streaming applications.

There are three classes of streaming applications: (i) stored media streaming, (ii) uni-directional real-time (live) streaming such as TV stations, and (iii) bi-directional real-time (live) streaming e.g. video conferencing. The technology and protocols used are essentially the same for all three classes. A playout buffer is used on the client side to compensate the fluctuations in the transmission delay and handle lost or out-of-order packets. The three classes significantly differ in the required quality of service parameters in particular delay, jitter and packet loss. For example, unidirectional live streaming requires less than 10 ms initial delay, less than 2 ms delay variance and < 2% packet loss. For interactive streaming applications, the end-to-end delay shall be around 150 ms, the delay variation < 1 ms and packet loss < 1%. In addition to the QoS parameters that are measurable in an objective way, QoE or Quality of Experience plays an important role, too. QoE is a subjective measure of the user experience which is influenced by many factors.

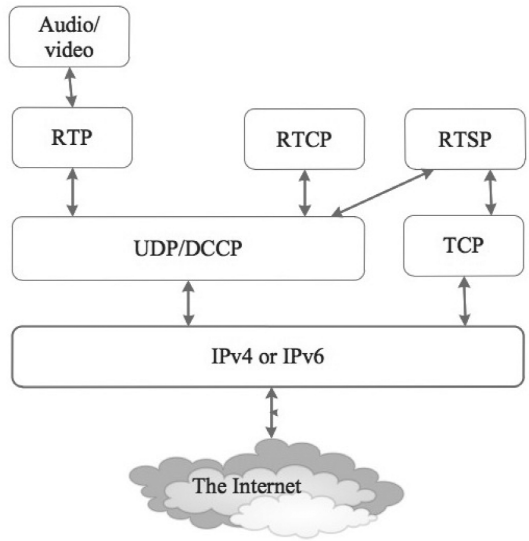


Figure 7 Streaming protocol architecture

Figure 7 shows the protocol architecture and the protocols commonly used in streaming applications. Going from bottom upwards along the architecture, the network protocol is obviously IP. At the transport layer, UDP is generally used for media transmission. Its limited functionality (only multiplexing/de-multiplexing, no error control via retransmissions, no congestion control) makes it robust and suitable for media transmission since it introduces almost no delay. TCP is used for control purposes. The transport protocols are not media-specific, therefore we need additional ones that support media transmission, such as RTP or Real-Time Transmission Protocol which uses sequence numbers and timestamps to help reconstruct the media stream on the receiver side. Its companion protocol, RTCP or Real-Time Transmission Control Protocol provides measurement information on the quality of transmission to the sender and receiver. Finally SIP or Session Initiation Protocol is used for session establishment and control, and RTSP - Real-Time Streaming Protocol is an application level protocol to provide the user with some DVD-like control functions during the streaming session.

More recent streaming technology is HTTP streaming. As the name suggests, it uses the HTTP protocol, and media is transmitted, using HTTP, in the form of successive short pieces (short files called chunks) and the client reconstructs the media stream from these independent chunks. HTTP streaming was first introduced by Apple for its QuickTime software. It is called HTTP Live Streaming or HLS. Its relatives are Microsoft's IIS Smooth Streaming, Adobe's Flash Dynamic Streaming and DASH, Dynamic Adaptive Streaming over HTTP.

HTTP Live Streaming is an adaptive protocol. At the sender side, multiple files are created for distribution to the player, which can switch between streams in an adaptive way to optimize the playback experience. The media stream at the source is encoded into multiple files at different data rates and is divided into short chunks of 5-10 seconds long. These are loaded onto an HTTP server along with a text-based manifest file that directs the player to additional manifest files for each of the encoded streams.

HTTP-based streaming has several advantages; no streaming server is required and the download of the media chunks should use HTTP caching servers located at different places of the networks of service providers, cellular providers, resulting in improved video quality for clients served from these caches. An important advantage is that content via the HTTP protocol can pass through most firewalls and proxy servers which is not the case with RTP over UDP.

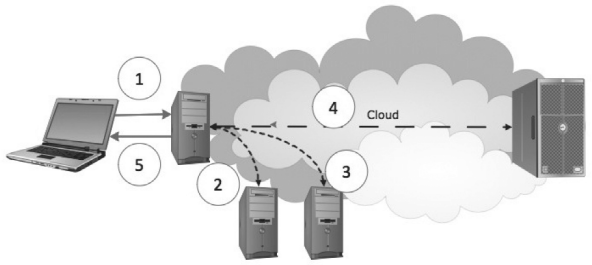
HLS is currently being standardized in IETF and at the time of writing (beginning of 2014) its specification is an Internet Draft [45].

2) Content Delivery Networks

As the sharing and consumption of multimedia content on the Internet has been growing rapidly, it has become obvious that many web servers hosting content and applications are unable to handle this demand, not to speak about bandwidth and quality of service requirements. The concept of CDN or Content Delivery Network has emerged to cope with the exponentially growing demand for exchange of multimedia information on the public Internet, to ensure scalability of multimedia networks and to enhance quality of experience of users.

To put it simply, CDN is a set of web servers, collaborating with each other, and hosting multiple copies of the same content to accomplish more efficient delivery of the desired content to the end users. The CDN concept is not entirely new as caching has been used to deliver general web content for many years, however, moving to delivery of on-demand or pre-recorded video and even of live video required new architectures and protocols. According to [46], CDNs have evolved from their first generation that delivered general static and dynamic content through 2nd generation that supported video-on-demand, streaming media and also mobile media applications during late 2000s to their 3rd generation, the community-based CDNs at the beginning of 2010s.

Main functional elements of a CDN architecture are: (i) origin servers where the content is put by the content owner and stored, (ii) edge servers or surrogate servers (caches) servers where copies of the multimedia content are distributed to and stored, (iii) distribution network which delivers content requests to the optimal location, (iv) redirector or request routing system that identifies the optimal (closest, not only in geographical sense) edge server for each user, and (v) some accounting mechanism for the origin server. Fig. 8 serves as an illustration. User request for the desired content is re-directed to the optimally closest edge server (1). The latter then searches for the content on its storage facility and if not available, checks other edge servers in its proximity (2, 3). If content is not found in the proximity of the end user, the request is sent to the origin server (4) which then delivers the content to the edge server and the latter delivers it to the end user (5).



1: User agent is directed to the closest edge cache server
2-3: Edge cache checks other edge caches since the content is not cashed there or it is old
4: Content is not cashed in edge servers or it is old so edge cache request it from the origin server
5:Content is delivered to user

Figure 8 The simplified scheme of obtaining content via a CDN

The largest CDN service providers include Akamai, the market leader [47] and Limelight Networks [48]. Akamai's market share is estimated to be over 80%, it operates 12000+ servers in 60+ countries.

A recent direction of CDN development is to support collaborative media streaming services using the Hierarchical Cooperative Control Protocol (HCOOP) [46].

The role of CDNs in multimedia communications is already significant and will continue to grow. According to [49], Content Delivery Networks (CDNs) will carry over half of Internet traffic in 2017, up from 34 percent in 2012, and the share of video traffic delivered over CDNs will be over two-thirds of total video traffic by 2017.

REFERENCES

[1] S. Weinstein and A. Gelman, "Networked Multimedia: Issues and Perspectives", IEEE Communications Magazine, Vol. …, No. 6, pp. 138-143, June 2003.

[2] Charles N. Judice, „Digital Storytelling: The Next Killer App ", Keynote speech at „Multimedia Services Access Networks" conference, Orlando, Florida, USA, June 13-15, 2005, http://msan.org.

[3] Michael L. Brodie, "The end of Computing Era", 2nd IEEE Conference on Digital Ecosystems Technologies (DEST), Phitsanulok, Thailand, February 2008.

[4] Sheau Ng, "A Brief History of Entertainment Technologies", Vol. 100, May 13, 2012.

[5] Digital Video Broadcasting (DVB); Framing structure, channel coding and modulation for 11/12 GHz satellite services. ETSI European Standard EN 300 421.

[6] Digital Video Broadcasting (DVB); Framing structure, channel coding and modulation for cable systems. ETSI EN 300 429 European Standard.

[7] Digital Video Broadcasting (DVB); Framing structure, channel coding and modulation for digital terrestrial television. ETSI EN 300 744 European Standard.

[8] ETSI Standards for 2nd generation DVB systems.

[9] El-Hajjar – L. Hanzo, "A Survey of Digital Television Broadcast Transmission Techniques", IEEE Comm. Survey&Tutorials, 2013.

[10] J.-N. Hwang: "Multimedia Networking – from Theory to Practice", Cambridge University Press, 2009.

[11] Amitabh Kumar, "Mobile TV: DVB-H, DMB, 3G Systems and Rich Media Applications", Elsevier, 2007.

[12] Parameter values for ultra-high definition television systems for production and international programme exchange. Recommendation ITU-R BT.2020 (08/2012).

[13] http://www.bbc.co.uk/news/entertainment-arts-23195479, 5 July 2013, BBC News, downloaded January 19, 2014.

[14] http://www.telegraph.co.uk/technology/news/10334043/Broadcasters-failing-to-keep-up-with-3D-TV-demand.html, The Telegraph, downloaded January 19, 2014.

[15] http://www.etsi.org/technologies-clusters/technologies/broadcast/dab

[16] W. Hoeg and Thomas Lauterbach. "Digital Audio Broadcasting: Principles and Applications of Digital Radio". John Wiley, 2003.

[17] www.worlddab.org

[18] OECD Communications Outlook 2013. Digital audio broadcasting in OECD countries. DOI:10.1787/comms_outlook-2013-table119-en

[19] WorldDBM seminar", Riva del Garda, Italy, 14 April 2013.

[20] P. Hannon, "Digital radio in Europe", World DAB Workhsop, Budapest, January 28, 2014, downloadable from http://www.worlddab.org/events/detail/335#informations

[21] http://www.drm.org

[22] G. O. Driscoll: Next generation IPTV services and technologies. Wiley Interscience, 2009.

[23] E. Mikoczy, D. Sivchenko, X. Bangnan, and J. I. Moreno, ''IPTV systems, standards and architectures: Part II. IPTV services over IMS: Architecture and standardization,'' IEEE Commun. Mag., vol. 46, no. 5, pp. 128–135, May 2008.

[24] Naeem Ramzan and Ebroul Izquierdo, "Scalable and Adaptable Media Coding Techniques for Future Internet", J. Domingue et al. (Eds.): Future Internet Assembly, LNCS 6656, pp. 381–389, 2011.

[25] László Bokor, László Lois, Csaba A. Szabó, Sándor Szabó, "Testbed of a Novel Media Streaming Architecture for Heterogeneous Wireless Environment," Proc. Tridentcom07, the 3rd Int'l IEEE Conference on Testbeds and Research Infrastructures, Orlando, USA, May 21-23, 2007, 10 pages. Available in IEEE XPlore, http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4444715.

[26] http://www.itu.int/rec/T-REC-H.323-200912-I/en

[27] http://www.ietf.org/rfc/rfc3261.txt

[28] Third Generation Partnership Project (3GPP), ''Technical specification group service and system aspects; telecommunication management; charging management; IP multimedia subsystem (IMS) charging (release 6),'' Tech. Specification 3G TS 32.260 version 6.8.0, 2007.

[29] K. Knightson, N. Morita, and T. Towle, ''NGN architecture: Generic principles, functional architecture, and implementation,'' IEEE Commun. Mag., vol. 43, no. 10, pp. 49–56, Oct. 2005.

[30] Gonzalo Camarillo, Miguel-Angel García-Martín: The 3G IP Multimedia Subsystem (IMS): Merging the Internet and the Cellular Worlds, 3rd Edition, John Wiley & Sons, 2008.

[31] Miikka Poikselkä, Georg Mayer: The IMS: IP Multimedia Concepts and Services, 3rd Edition, John Wiley & Sons, 2009.

[32] A. Sanches-Esguevillas, B. Caro, G. Camarillo, Y. B. Lin, M. A. Garcia-Martin, and L. Hanzo, "IMS: The New Generation of Internet-Protocol-Based Multimedia Services", Proc. IEEE, Vol. 101, No. 8, August 2013, pp. 1860-1881.

[33] http://www.nets-fia.net

[34] J. Pan, S. Paul, and R. Jain, "A Survey of the Research on Future Internet Architectures", IEEE Commun. Mag. July 2011, pp. 26-36.

[35] Future Media Internet Architecture Reference Model - White Paper: FMIA-TT_Reference Model_Ver1 0 _20110301

[36] Edwards et al, Understanding Infrastructure: Dynamics, Tensions and Design, An NSF Report, Jan. 2007. Available on: http://cohesion.rice.edu/Conferences/Hewlett/emplibrary/UI_Final_Report.pdf

[37] C. Szabo, "European broadband initiatives with public participation," in: I. Chlamtac, A. Gumaste and C. Szabo (Eds.): Broadband Services: Business Models and Technologies for Community Networks, John Wiley&Sons, 2005.

[38] Csaba A. Szabó and Károly Farkas,, „Planning wireless municipal networks based on Wi-Fi/WiMAX mesh networks – applications, technologies and business models", tutorial lecture, ICC2010, Cape Town, South Africa, May 2010.

[39] C. Szabo, "Services to Meet Society-related Needs", in: I. Chlamtac, A. Gumaste and C. Szabo (Eds.): Broadband Services: Business Models and Technologies for Community Networks, John Wiley&Sons, 2005.

[40] C. A. Szabo, "Planning Wireless Cities and Regions To Support Telemedicine Applications", in Malina Jordanova (ed.): "Global Telemedicine and eHealth Updates: Knowledge Resources", Vol. 1, pp. 169-173. Publ. Luxexpo, Luxembourg, 2008, ISSN 1998-5509.

[41] Károly Farkas, Csaba A. Szabó, Zoltán Horváth, „Motivations, Technologies and Sustainability Models of Wireless Municipal Networks", IEEE Communications Magazine, Vol. 47, No. 12, pp. 76-83, December 2009.

[42] Microsoft Windows Media Technologies, http://www.microsoft.com/windows/windowsmedia/default.mspx.

[43] RealNetworks RealPlayer, http://www.realplayer.com/.

[44] Apple QuickTime, http://www.apple.com/quicktime/.

[45] http://tools.ietf.org/html/draft-pantos-http-live-streaming-12

[46] R. Buyya, M. Pathan and A. Vakali (Eds.), Content Delivery Networks. Lecture Notes in Electrical Engineering 9. Springer-Verlag Berlin-Heidelberg, 2008.

[47] www.akamai.com

[48] www.limelight.com

[49] Cisco Visual Networking Index: Forecast and Methodology, 2012–2017. Cisco White Paper, May 29, 2013. Downloaded from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html, on February 24, 2014.

Leonardo Chiariglione obtained his Ph. D. degree from the University of Tokyo in 1973. During his career he launched or participated in several initiatives. Among these are MPEG and the Digital Media Project (DMP). He is currently CEO of CEDEO.net a company developing its own portfolio of technologies, products and services in the area of digital media, such as WimTV (http://wim.tv/).
Leonardo is the recipient of several awards: among these the Eduard Rhein Foundation Award, the IBC John Tucker Award, the IEEE Masaru Ibuka Consumer Electronics Award and the Kilby Foundation Award.

Csaba A. Szabó obtained his MSc from St. Petersburg State University of Telecommunications, Dr. Univ., Ph. D. and Dr. Habil. from the Budapest University of Technology and Economics (BME) and the title Doctor of Technical Science from Hungarian Academy of Sciences. He is a Professor and Head of Laboratory of Multimedia Networks and Services at the Dept. of Networked Systems and Services of BME.
Prof. Szabo has long-term experience in academia, R&D and telecommunication business. He co-founded a leading network system integrator and service provider company that was subsequently integrated into Hungarian Telecom. Prof. Szabo has been a member of editorial boards and EiC of several international journals. He is Senior Member of IEEE.

Measuring and Controlling IT Services
The Case of Telecom Enterprises

Péter Fehér, Péter Kristóf

Abstract — **Telecom enterprises are organizations with complex processes and structure the operation of which is highly dependent on information technology (IT) and the services provided by it. The landscape of the IT architecture reflects the complexity of the business processes. As this entanglement makes IT-based solutions and applications mission critical, securing the availability and operation of the services belongs to the top cost drivers at telecom enterprises, causing serious risks as well. Handling these costs and risks and aligning business strategy with IT-based solutions the use of Enterprise Architecture and IT controlling is suggested. The joint deployment of the two disciplines secure the measurement and controlling of IT services: the elaboration of the current state, the fact-based planning of the desired future state and the transition between them. Aligning business strategy and IT strategy helps optimizing infrastructure costs and mitigating risks. The paper gives a brief overview of the newest methodologies and emphasizes a holistic approach while bringing examples from the telecom industry.**

Index Terms — **Enterprise Architecture, IT controlling, business-IT alignment, TOGAF, cost accounting.**

I. INTRODUCTION

Telecommunication enterprises offer complex business services to their customers that are more and more dependent on information technology solutions. Both the enterprise architecture of the companies, as a whole, and the IT architecture, as part of it, mirror the complexity of the offered services. IT services are critical for providing business services, therefore business performance, and even business costs are heavily dependent on IT solutions. Not a surprise, telecommunication enterprises are among the top spenders to information technology: IT costs of a telecom company (comparing to its revenue) are double or even triple of an average manufacturing, utilities or food industry company [9].

The economic crisis of the last years, and the changes in the compliance environment (extra taxes, roaming limitations in EU) challenged the profitability of this sector. In the EU, Hungary and Greece introduced the highest speciality tax rates of respectively 40% and 36%. While the business environment became less predictable, sharing the business risk with customers by increasing customer prices is limited, because of the price sensitive nature of the market [8].

Manuscript received May 30, 2014.

Péter Fehér is Department Head of Computer Science at the Corvinus University of Budapest, 1093 Budapest, Fővám tér 8., Hungary (e-mail: pfeher@informatika.uni-corvinus.hu)

Péter Kristóf is PhD candidate at University of Pécs Faculty of Business and Economics, 7622 Pécs Rákóczi út 80., Hungary (e-mail: kristofp@ktk.pte.hu)

As telecom companies cannot increase their revenues significantly, they have to concentrate on their cost structure: postpone development and investments, and generally decrease operational costs, including IT investments and IT operational costs. Because of this business requirement, there is an increased pressure on IT departments in the telecom sector to reduce IT costs while keeping the value creation ability of IT services.

The paper analyses this challenge from an enterprise architecture point of view, while describing both business and IT solutions of the telecommunication industry. Enterprise architecture management is '[...] the practice that tries to describe and control an organisation's structure, processes, applications, systems, and technology in an integrated way' [14].

II. HANDLING COMPLEX TELECOM ENTERPRISE ARCHITECTURES

The complexity of the telecom enterprise architecture poses both an opportunity and a limitation of a more conscious and more controlled behaviour of the industry. The main reasons and challenges behind the complexity of enterprise architectures, especially IT architectures are the following:

- Complex telecom services require complex business processes and IT solutions.
- Mergers and acquisitions are common in the telecom industry, but these transactions generally result in an increased complexity of architectures.
- The change of the environment requires fast responses from telecom companies that leave less time for architecture optimisation.

A. Examples for complexity

To illustrate the impact of these reasons a few examples are presented.

Example 1

A telecommunication company that has roots in the cable TV industry, and now offers broadband internet and other telecommunication services, rapidly expanded in the last decade. It acquired companies from all of its service areas: cable TV providers, landline telecommunication companies, and internet providers. The expansion was strategically and financially justified, and the number of the company's customers increased quickly, and now it provides service to several million households, at several hundred thousand sites.

Although the customers benefited from the technological development and increased service quality, they suffered from the too bureaucratic processes, and poor customer care. The company received several customer complaints, and even consumer protection websites cited alarming events about its practice.

The company realised, that over the expansion, its enterprise architecture (both business processes and IT services) could not follow its changes: because of the rapid growth business processes were rarely integrated, nor consolidated, and the company has kept the very heterogeneous IT architecture of the acquired companies. In an extreme case, the company acquired a local cable TV and internet providing company that has just 10 months ago acquired a smaller cable TV company, and was in the middle of its architecture consolidation process. By the end of the process, the telecommunication company in our example had got three additional customer relationship management software, including the customer database.

As the result, the telecommunication company had to deal with at least duplicated, but mostly multiplicated databases, functional systems, and complex integration solutions. Employees in the customer office often applied excel tables to integrate the data from different sources. The company had to realize that its business and IT architecture raises a barrier to effective business management.

Example 2

An integrated multinational telecommunication company decided to integrate its previously semi-independent subsidiaries of mobile and landline telecommunication, internet, satellite- and cable TV services. These services supported the business and IT architecture of the subsidiaries, but because of the redundancies, the company had to face similar difficulties like in Example 1. Because of the different historic background of the subsidiaries, it was hard to judge which processes and systems are better, and how the systems should be consolidated.

Having complex architecture is not necessarily wrong, but needlessly complex architectures mean both development and operational risks for telecom companies, and requires extra resources to maintain the necessary level of services. Moreover, there is an increasing gap between business strategy, business requirements and the existing enterprise architecture, especially in the field of information technology [19]. As Example 1 and 2 illustrated, mergers and acquisitions result in redundant IT systems and databases. But even shutting down unnecessary applications could retain the existing complexity: previously dedicated background applications and services became later the background of other applications, and to avoid further risks, companies keep this complexity. As result applications and databases are connected to other applications and databases in a spaghetti-like architecture [16].

In order to support the cost cutting challenges of the companies, we have to understand and manage this complex architecture, and to provide a structured, transparent and simple approach to control and even to simplify complexity.

B. The need for enterprise architecture and IT controlling

Growing enterprises usually forget to perform a variety of critical and complex tasks which used to be in the focus of their daily activities after they started up: thinking about why the company was established, what the main activities are and how the client needs are fulfilled. They also forget about planning the future, building their strategy and defining the desired state of the business. Additionally they have no clear view about how their today's technology can perform and how it should be developed in order to fulfil future needs. As they are getting larger, the complexity of their business activities and the architecture will be also growing. These challenges – deeply impacting the design of the enterprises – are the result of changes in the environment (e.g. globalisation, service orientation), the competition and mergers or acquisitions.

This is especially true for telecom companies the operation of which is highly dependent on services provided and supported by information technology. Handling billions of transactions and providing services for millions of clients result in a complex IT architecture. For dealing with this complexity and designing and controlling IT architecture of large telecom companies the paper suggests using the approach and toolset of enterprise architecture and IT controlling.

III. LITERATURE OVERVIEW

Business performance highly depends on a balanced and integrated design of the enterprise. This includes staff (with all competencies), organizational structure, business processes, production and services, finance and business environment. Facing the challenges the management needs to make conscious decisions about the design of the enterprise [10].

At this point enterprise architecture (EA) is used as an instrument to steer an enterprise's future, and serves as a coordinating mechanism toward the actual transformation. In articulating an enterprise's future direction, the multi-perspective approach, which is typical of enterprise architecture, enables the achievement of organizational cohesion and integration. Furthermore enterprise architecture provides the clear overview required for translating strategy into execution, enabling the top-level management to take the ownership of decisions about the design of the future enterprise [10].

From this point of view the goal of enterprise architecture is to act as a guide (or pathfinder) and take the enterprise on a transformational journey – from an incoherent and complex world to a more rationally designed organization with multi-service, revenue-generating platforms and efficient operational structure. The aim should be clear: enterprise architecture is required to deliver tangible business benefits [5] and has to play a significant role in aligning business needs and technology solutions while offering

- an insight into the current utilization of IT in business operations,
- a vision for the future utilization of IT in business operations, and
- a roadmap for the evolution of the IT from the current state to the desired future state. [5]

Helping the fulfilment of this aim, different kinds of frameworks have been elaborated. Architecture frameworks help the companies to consciously go through the process of architecture development, and also offer an initial structure of activities. In the following sections an overview of the necessary background for applying the EA approach in the telecom industry is presented.

A. Frameworks

Enterprises can be of any size and can have different level of complexity – their architecture will tend to be similar. Over time different kinds of models, tools and frameworks have been developed to address this complexity and to support enterprise architecture [1]. Frameworks support structuring architecture description techniques by identifying and relating different architectural viewpoints and the modelling techniques associated with them. Some of them are quite specific about what kind of elements a structure should have [14]. Besides, the frameworks intend to help professionals dealing with architectural issues by providing also an ontology using different levels of abstraction for mapping all kind of required information and helping the communication between the professionals involved [15].

B. Why TOGAF?

On the field of enterprise architecture, The Open Group Architecture Framework (TOGAF) is a well-known framework that links four different types of architecture contents: business, data, application and technology [2]. Originally it was a generic framework and methodology for development of technical architectures, but evolved into an enterprise architecture framework and method. It has four main components [14]:

- Architecture Development Method (ADM): provides a 'way of working' for architects. The ADM is considered to be the most significant component of TOGAF, and consists of a stepwise cyclic approach for the development of the overall enterprise architecture and defines a full life-cycle process for planning, designing, realizing and governing enterprise architecture. Doing so it brings discipline to the architecture development process [5].
- Architecture Content Framework: considers an overall enterprise architecture as composed of four (above mentioned) closely interrelated architectures: business architecture, data architecture, application architecture, and technology (IT) architecture. It describes what the architecture should look like by the end of the transition process.
- Architecture Capability Framework: addresses the organisation, processes, skills, roles, and responsibilities required to establish and operate an architecture function within an enterprise. It also provides a set of concepts for a successful implementation of enterprise architecture governance.
- Enterprise Continuum: comprises various reference models. It illustrates how architectures are developed across a continuum ranging from foundational architectures, through common systems architectures and industry-specific architectures, to an enterprise's own individual architecture.

Architecture is the art and science of designing complex structures. In this context, enterprise architecture is defined as a coherent whole of principles, methods, and models that are used in the design and realisation of an enterprise's organisational structure, business processes, information systems, and infrastructure. [14]

C. IT controlling

To a great extent (and also at telecom companies) the complexity of IT is also the result of an uncontrolled proliferation of redundant systems and solutions. In a fragmented organizational structure each business unit has its own IT budget – therefore it will follow its own way in procuring, building and operating its technology. Furthermore IT departments are organized on a project basis and are free in making decisions. As the decision making happens locally, it prevents leveraging enterprise-wide synergies and leads to a nearly unmanageable jungle of redundant applications and data [5] and a widening communication gap between business and information technology. Architecture is an essential tool in controlling the complexity of the enterprise. Better alignment between business and IT leads to lower cost, higher quality, better time-to-market, and greater customer satisfaction. [14] For ensuring enterprise IT being stable, agile, adaptable and efficient, the methodology and toolbox of IT controlling is available. IT controlling provides support for managers when making decisions about IT-related resources. [17]

Furthermore IT controlling is about the control of IT-related operations in the organization. Its goal is to ensure efficiency and effectiveness of IT operations while providing quality, functionality and compliance to deadlines in information processing. IT controlling has a monitoring function as well as a coordination function for the management of information.

IT controlling focuses on a certain controlling application and covers common aspects such as data aggregation and processing, budget planning, and coordination of individual information and data sets. The formal aim of IT controlling is to ensure the effective and efficient usage of IT resources. Alongside formal aims, aims with regard to content are pursued, whereby business value, costs, quality, functionality, and in-time delivery are the goals with highest priority. IT controlling can be divided into three core processes [12]:

1. Planning: setting performance targets in alignment with the business and IT strategy.
2. Monitoring: to measure the status-quo and expected deviation from it or the deviation from performance targets.

3. Steering: continuous target-performance comparison. In the event that a deflection is identified, appropriate steering actions should be initiated.

IT controlling consists of making decisions about acquisition, change, and disposal of IT, as well as monitoring IT performance data in order to be able to control IT more effectively and efficiently.

D. Bridging the business-IT gap

At companies operating in sectors which are highly dependent on technology, the stakeholders from business and the stakeholders from IT often misunderstand each other and forget that they are approaching the same question from very different directions. One group is focusing on business demands while the other has to satisfy IT demands in a timely manner while managing the complexity and performance of the entire application landscape. Over time both of these groups have developed own theories, practices and rules which lead to an ever widening gap between them [3].

While IT controlling is proceeding from a business perspective and defines metrics and indicators to measure the status and contribution of information technology, enterprise architecture management is a procedure to create transparency, to clarify the role of IT and to spur the alignment of business and IT. As IT controlling and enterprise architecture are addressing the topic of business-IT alignment, they are also providing means for communication.

The architecture of an enterprise is commonly regarded as the cornerstone for its long-term evolution. In order to support this holistic approach it should be included in IT controlling. As a complementary to IT controlling, enterprise architecture management is a continuous and iterative methodology with a major goal of controlling and improving the existing and future IT support of an organization. Putting this methodology into practice the enterprise architecture process not only considers the information technology of an enterprise, but also takes business processes, business goals and strategies into consideration in order to build a holistic and integrated view of the whole enterprise. 'The goal of enterprise architecture management is a common vision regarding the status quo of business and IT as well as of opportunities and problems arising from these fields, used as a basis for a continually aligned steering of IT and business.' [3]

E. Business Process Framework (eTOM)

The Business Process Framework (formerly known as eTOM – enhanced Telecom Operations Map) is a TM Forum (a non-profit industry association for service providers and their suppliers in the telecommunications industry) initiative. Its purpose is to deliver a process framework for service providers within telecommunications industry. The framework maps and describes all the enterprise processes required by a telecom service provider and analyses them to different levels of detail according to their business significance and priority. Beside serving as the blueprint for process direction and providing a reference point for internal process reengineering needs, the framework outlines potential boundaries of software components to align with the customers' needs and allows an overview on the required functions, inputs and outputs that must be supported by products. [4]

The Business Process Framework can be used as a tool for analysing an organization's existing business processes and for developing new ones. In applying the methodology, different processes delivering the same business functionality can be identified, and so duplications eliminated, gaps revealed, new process design speeded up and risks reduced. It also helps measuring and assessing the value, cost and performance of these processes [4] and so supports controlling them.

Telekom Malaysia can serve as a good example. The country's largest integrated solutions provider started a project in 2009 for optimizing end-to-end business processes, enhancing operational efficiency and improving customer experience. The project was carried out on Business Process Framework basis and resulted in a consolidated and single platform for consumer, business and wholesale services in just 10 months. A significant result was also achieved by China Mobile. The world's largest operator launched a cloud mobility imitative for its 90 million subscribers, using Business Process Framework. With this project they achieved savings of $120 million and increased revenue of $40 million annually [18].

The Framework is defined as generically as possible, so that it is independent of organization, technology and service. It is basically intuitive, business driven and customer focused. To reflect the way businesses look at their processes, the Framework supports a horizontal (functionality-related processes) and a vertical (end-to-end processes) perspective on the grouping of the process elements. The overlay of the horizontal functional processes and the vertical end-to-end process groupings form a matrix structure, which is the core of innovations and fundamental benefits of the Framework. It offers a standard language and structure for the process elements that can be understood and used in specifying and operating end-to-end processes and creating the capability that enable these processes. [4]

Using the eTOM framework is a good start to structure the business architecture part of any telecom industry organisation, through providing a standard processes group and business architecture relationships.

After having established the foundations of enterprise architecture (EA) and IT controlling we will continue with practical questions.

IV. PRACTICAL QUESTIONS OF THE EA APPROACH

The first question of applying the enterprise architecture approach is how to create a relationship between business strategy and organizational activities. Without analysing the strategy development process, the enterprise architecture approach should apply the results and implications of the business strategy [16]. Of course the strategy development process can build on the existing and future opportunities of ICT solutions [13], but analysing this topic related to the EA approach is beyond the objectives of this paper.

Creating the relationship between business strategy and enterprise architecture is part of the company-wide governance, especially IT governance structures and processes [7]: The business strategy should be decomposed into business objectives, and tools to achieve them. To translate the abstraction of a business strategy the offered products and services should be identified.

Therefore, following the Enterprise Architecture based approach, telecom companies should understand and model their service portfolio. Defining the service portfolio enables to examine the possible existing strategic gap between business and IT services. In our research we explored and analysed the enterprise architecture structure of several organizations, mostly in the financial and telecommunication sector.

The findings in this section are generalized, but illustrated by telecom examples. As basis for developing the presented architecture model, the content metamodel of TOGAF [11] was selected with the following expansions: process modelling, governance, services and infrastructure consolidation. During research these elements were tailor made for the efforts of representing our service based enterprise architecture model.

A. Business architecture

In order to map the business architecture of telecom companies, strategic questions of target customer segment markets and offered services should be answered. In the last 15 years, telecommunication companies widely diversified their service portfolio, and beside the traditional telecommunication services (landline and mobile telecommunication), data communication (landline and mobile), IT services (e.g. cloud services), and content services (cable TV, mobile TV, internet-based contents) are also provided.

Each of these strategic services consists of several options. A very simple service breakdown is the following: a mobile strategic service is divided into pre-paid and subscription-based alternatives, and in each of them there is a wide variety of costumer packages that identifies the different fees. But each package contains additional services beside the basic call service: internet, sms/mms, voice mail, missed call service. These services are labelled as *customer-facing business services*, and customers are expected to pay for these services.

The business architecture, business services and the enabling processes are standardised in the (eTOM) Business Process Framework that defines the main business activity areas, that concentrates mostly on operational activities.

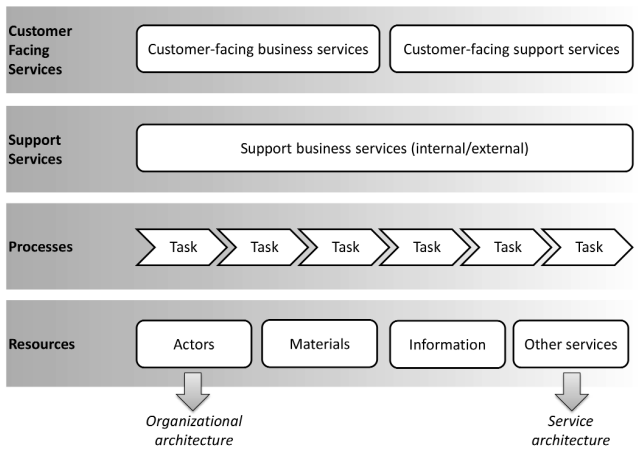
Beside customer-facing business services customer demand additional support services, like services desk, customer-care, sales points or device support. Generally these services are not considered as value added services by the customers, but a poor service desk can heavily impact the perception of the overall service quality. These services are labelled as *customer-facing support services*.

All of the customer-facing services are provided through the realisation of either internal or external business and IT processes that are not directly visible for the customers: market research, accounting, marketing, business operations, controlling, etc. These services are labelled as *support business services*, and enablers of customer-facing services.

Example 3

A customer wants to buy a new mobile phone with a new subscription. The customer-facing service is sales, and it creates income for the company, and value for the costumer. But in order to realise this service the sales clerk has to check the identity of the customer, record the transaction, invoice the fee, and provide the mobile phone itself. Providing mobile phones to customers requires procurement, external and internal logistics, that are not visible, and in fact customers do not care how the company solves this issue. Even most of the steps in case of a personal purchase are not perceived as value adding activities by the customer. Customers does not appreciate activities like identification, recording customer data, although these steps are necessary for the whole processes, and even serve customer's interest.

FIGURE I
MAIN ELEMENTS OF THE BUSINESS SERVICE ARCHITECTURE
(SOURCE: P. FEHÉR)



All business services are realised through business processes. This part would be considered as the domain of business process management, but because the business service hierarchy depends on the processes, it should be considered as part of the enterprise architecture [11]. Business processes are sequences of tasks that are often labelled as activities or process steps. Processes can run through different organizational units, but on the task level responsibilities belong to a well-defined organizational unit, even to a well-defined role. In the enterprise architecture approach, this is the domain of organizational architecture that defines the organizational structure, and determines the required roles and actors for each task. During performing a task, an actor uses resources, like materials, information from documents or electronic databases. Performing a task can also require the consumption of additional business or IT services.

B. *IT service architecture*

Business services and business processes require the availability of information technology services. As in the case of business services, there is only a limited number of IT service that is visible for the business departments, and consider them as value added services. These services are labelled as *business facing* (or direct) *IT services*.

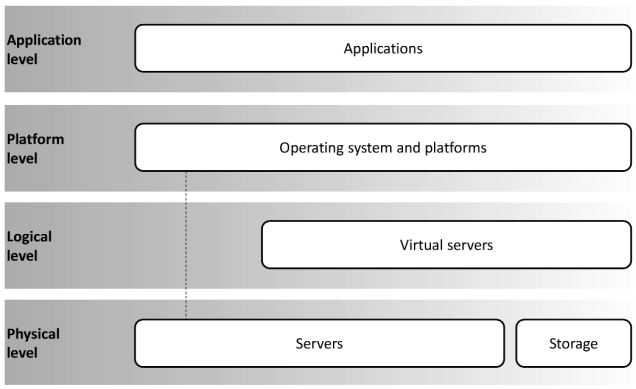
These services present themselves as frontend application functions, that is visible and usable for the business side. But in the background, to provide any service, other application functions are required, that are not visible for the business side. These services are labelled as *support* (or indirect) *IT services.* Practically using a business facing IT service means initiating an IT process, a workflow that is connected to the related applications, and uses their functionality.

C. *IT architecture*

Application functions are considered as IT-providing services, in case of complex organizations, such as telecom enterprises, applications use the functionality of other applications, and use data from various databases.

In order to provide the required environment for applications, a physical infrastructure (even external or internal) is required: servers and storage facilities provide the required processing and data storage resources for applications. In case of complex telecom architectures server virtualisation became popular, and provides more flexibility, and a more consoledated infrastructure in the physical level. On the virtual level, however, even several hundred virtual servers provide services. Virtual servers with the operating system and other platform software (e.g. a framework, a running environment, etc.) provide the basis to run the applications. Elements of this environment are labelled as *platform services*: 'A technical capability required to provide enabling infrastructure that supports the delivery of applications' [11].

FIGURE II
MAIN ELEMENTS OF THE IT ARCHITECTURE
(SOURCE: P. FEHÉR)



Additional architecture elements, such us network appliances or message brokers are also part of the IT architecture, that provide IT services (data transfer, messaging, communications).

V. COST ALLOCATION OF IT SERVICES

From business point of view it is important to see how information technology services support and enable business operations, and to understand the costs of these services. IT controlling, more specifically cost controlling and cost allocation helps to understand the main cost factors. In case of enterprise architecture approach the relationship between different levels of the architecture is defined, therefore cost allocation into higher level is possible.

The overview of an enterprise architecture presented the main elements of a complex telecom architecture. This understanding is required in order to calculate the costs of business, especially customer facing business services. During the budget planning process, and during the cost accounting main cost factors are planned and summarised. So even not knowing every detail of the cost factors of information technology services, the total sum is usually available for telecom companies.

In a wider view, even total cost of each department (such as IT) can be calculated, but without knowing the costs and contribution to business services neither architecture consolidation, nor cost reduction can be performed.

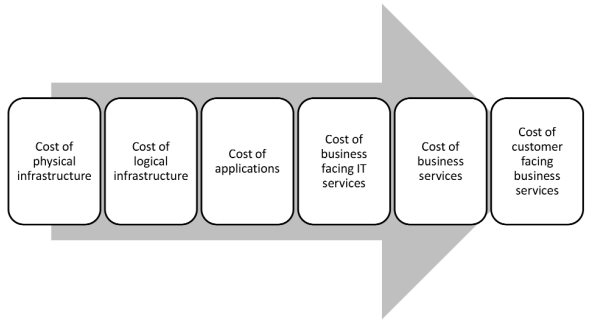
The biggest challenge in allocating IT costs is how to deal with this complexity, and how to provide a clear understanding of IT and business costs.

The main cost drivers are the following:

- Depreciation of infrastructure elements (based on previous years' CAPEX spending)
- Human costs
- Other contracted costs (OPEX)

These costs are allocated through the levels of the enterprise architecture.

FIGURE III
PROCESS OF COST ALLOCATION IN CASE OF TELECOM ENTERPRISES
(SOURCE: P. FEHÉR)



At this point we have to differentiate between the financial (cash-flow) and accountancy understanding. In case of budgetary planning and budget control the financial approach is applied, therefore incomes and expenditures are registered. But expenditures on investments like buying a new server, or implementing a new CRM application does not change the value of company assets, only cash is transformed into another kind of asset.

From an accountancy point of view the decrease of the value of these assets (depreciation) is considered as costs. In case of IT assets, the value of new investments is depreciated through 3-5 years. Companies should therefore consider, that changes in their IT budgets will have immediate impact on their cash flow, but a longer impact in their cost structure (Table 1), but over the long run, considering the life-cycle of equipment these expenditures and costs.

TABLE I
COMPARING THE CASH FLOW AND ACCOUNTANCY APPROACH

(Infrastructure development)	Year 0	Year 1	Year 2	Year 3	Total
Cash Flow	-€600 000	-€60 000	-€60 000	-€60 000	-€780 000
Investment	-€600 000				-€600 000
Support		-€60 000	-€60 000	-€60 000	-€180 000
Costs	€0	-€260 000	-€260 000	-€260 000	-€780 000
Depreciation		-€200 000	-€200 000	-€200 000	-€600 000
Support		-€60 000	-€60 000	-€60 000	-€180 000

Organizations should also consider that consolidation of IT architecture could mean an enhancement both in their cash-flow (e.g. price of sold servers, savings on support and maintenance) and in their cost structure (no more depreciation and support costs).

TABLE II
COMPARING THE CASH FLOW AND ACCOUNTANCY APPROACH IN CASE OF IT INFRASTRUCTURE CONSOLIDATION

(Infrastructure consolidation for Y2)	Year 0	Year 1	Year 2	Year 3	Total
Cash Flow	-€600 000	-€60 000	€400 000	€0	-€260 000
Investment	-€600 000				-€600 000
Support		-€60 000			-€60 000
Sold equipment			€400 000		€400 000
Costs	€0	-€260 000	€0	€0	-€260 000
Depreciation		-€200 000			-€200 000
Support		-€60 000			-€60 000

In order to understand how costs are allocated along the layers of the enterprise architecture, we have to differentiate between direct and indirect costs. Direct costs are directly related to a specific service, while indirect costs are shared among multiple services [6]. Although full accuracy can be achieved only by using direct services, in case of complex telecom architectures it is impossible. Even, because of the complexity, the measurement of each infrastructure element would result in almost impossible work efforts from the organization.

In case of the physical IT infrastructure, elements can be grouped by performance categories (e.g. mid-range of high-end storage facilities, or differentiating based on CPU performance of the servers), and costs of groups should be calculated. Of course, if a company has a very detailed configuration management database that is connected to the accounting system, the costs of each element are visible. In case of telecom enterprises this connection is rare, therefore simplification, wherever possible, is required.

The cost of IT infrastructure elements consists of the depreciation costs of infrastructure element, infrastructure related external services (e.g. support, maintenance), and the cost of human efforts. In order to measure human efforts, telecom organizations usually use dedicated people for infrastructure or applications environment, or use a time reporting system.

The cost of the virtual environment consists of the direct costs of virtualisation (virtualisation of software licenses, external service and HR costs), but also the indirect allocated costs of the physical infrastructure. The cost of each service can be calculated (beside the direct costs) by the used resources (CPU, memory, storage) of the physical infrastructure.

Cost of applications consists of their direct costs (software license, external services, HR costs), but the biggest cost factor is the allocation of the logical architecture costs, based on the used virtual servers and the used capacities.

Example 4

Several telecom companies use robust central servers, such as IBM AS/400 (formally renamed to iSeries) to provide the required capacity to their core operations, such as customer database, enterprise resource planning, or e-mail services.

In case of an AS 400 based service portfolio, the total cost of the AS 400 environment can be easily calculated, through its dedicated depreciation, support, maintenance and HR costs. The applications, running in this environment, use about 70% of its capacities. In this shared environment, costs are allocated by their relative usage, because allocated costs should cover the total cost of the environment. In another case the main application of a telecom company runs in a distributed environment and its components use several virtual servers. The total cost of this application builds up of the allocation of these virtual server costs.

In order to calculate the cost of business facing IT services, another method is required: IT services do not have necessarily direct costs (no depreciation, external services and HR costs and calculated on application levels). Therefore the costs of IT services are the allocations of IT application and databases. In order to calculate the cost of an IT service the identification of the used applications, and its capacity usage is required.

Example 5

In case of a webshop of a mobile telecom company when a customer wants to buy a new phone, this process requires the usage of other main background applications: customer database, identification module, payment module, logistic application. In order to calculate the cost of this IT service the allocation of other application costs is required.

As the process and the examples show, the use of the enterprise architecture logic helps to identify the main cost drivers of each IT service, but requires a very complex modelling of the enterprise. In order to avoid complex modelling efforts for simple services, it is suggested to concentrate on the most critical, most expensive IT services.

VI. CONCLUSION AND SUMMARY

Applying the enterprise architecture approach requires strong consciousness to explore and understand the existing structure of an organisation (as-is state). Understanding this structure and the interdependencies helps to calculate the costs of each customer facing business service, so the company can decide on shaping its service portfolio.

Based on the enterprise architecture approach, even business-facing IT services became transparent, and each business unit can decide on how to optimise its processes. Maintaining the IT service portfolio is not about only consolidating IT services, but also an opportunity to optimise business unit costs through automating business tasks. But in order to make the decision on these questions, information of IT costs can be a good basis for creating a reliable business case.

The paper showed that transparency and controllability of enterprises require a strategic approach. In order to provide profitable telecom services, the definition and the cost-based pricing of them is necessary. Complex architectures of telecom enterprises are difficult to overview, and the clarification of them can only start with modelling the organization and the business processes. Enterprise architecture is an efficient tool and method in this activity.

After having analysed large-size telecom companies we found that a unified model of enterprise architecture and IT controlling can bring significant results in business-IT alignment and cost optimization. In our paper this unification was introduced and confirmed with practical examples.

REFERENCES

[1] F. Ahlemann, E. Stettiner, M. Messerschmidt, and C. Legner, Eds., *Strategic Enterprise Architecture Management*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[2] P. Bernus (ed.), *Enterprise architecture, integration and interoperability IFIP TC 5 international conference, EAI2N 2010, held as part of WCC 2010, Brisbane, Australia, September 20-23, 2010 ; proceedings*. Berlin [u.a.]: Springer, 2010.

[3] S. Buckl, A. Ernst, J. Lankes, F. Matthes, and C. M. Schweda, *Application Landscape Metrics: Motivation, Expectations, and a Case Study*, p. 10.

[4] *Business Process Framework (eTOM) – Concepts and Principles*, version 9.2, TM Forum, 2011.

[5] S. Bente, U. Bombosch, S. Langade, *Collaborative Enterprise Architecture – Enriching EA with Lean, Agile, and Enterprise 2.0 practices*. Waltham, MA: Morgan Kaufmann, 2012.

[6] D. Cannon, *ITIL Service Strategy 2011 Edition*, 2nd edition. London: The Stationery Office, 2011.

[7] N. O. Fonstad and D. Robertson, "Transforming a Company, Project by Project: The IT engagement model," *MIS Quarterly Executive*, vol. 5, no. 1, 2006.

[8] F. Dickgreber and J. Mattson, "Taxing Telecom: The case for reform." A.T. Kaerney, 2013.

[9] M. Eul, A. Freyberg, and R. Jaeger, "IT in the Telecom Industry - Reaching the next level". A.T. Kaerney, 2010.

[10] D. Greefhors, E. Proper, *Architecture Principles – The Cornerstones of Enterprise Architecture*. Berlin-Heidelberg: Springer, 2011.

[11] T. O. Group, *TOGAF® Version 9.1*, 10th New edition edition. Van Haren Publishing, 2011.

[12] F. Hamel, T. Herz, F. Uebernickel, and W. Brenner, *Management of IT Costs and Performance in Business Groups: Analysis of Unaddressed Requirements*, in PACIS, 2011, p. 12.

[13] J. C. Henderson and N. Venkatraman, "Strategic Alignment: Leveraging Information Technology for Transforming Organizations," *IBM Syst. J.*, vol. 38, no. 2–3, pp. 472–484, Jun. 1999.

[14] M. Lankhorst, *Enterprise Architecture at Work – Modelling, Communication, and Analysis*. Berlin Heidelberg: Springer Verlag, 2005.

[15] M. Op't Land, E. Proper, M. Waage, J. Cloo, and C. Steghuis, *Enterprise Architecture*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009.

[16] J. W. Ross, P. Weill, D. Robertson, *Enterprise Architecture as Strategy: Creating a Foundation for Business Execution*. Boston, Mass: Harvard Business Review Press, 2006.

[17] A. Schwertsik, P. Wolf, and H. Krcmar, *IT-controlling in Federal Organizations*, 17th European Conference on Information Systems, p. 13, Jan. 2009.

[18] A. Turner (ed.): *TM Forum Case Study Handbook – Showcasing Innovation and Success*. New Jersey: TM Forum, 2014.

[19] P. Weill and J. W. Ross, *IT Governance: How Top Performers Manage IT Decision Rights for Superior Results*. Boston: Harvard Business School Press, 2004.

Dr. Péter Fehér was born in Szekszárd, Hungary, in 1977. He received his MSc degree in economics from the Budapest University of Economic Sciences and Public Administration, Budapest, Hungary 2000. He received his PhD degree in Management in 2005 at the Corvinus University of Budapest, Hungary.

He is an associate professor at the Corvinus University of Budapest, and serves as Department Head of Computer Science, and Associate Dean for Education Affairs, responsible for Master and Postgraduate programs at the Faculty of Business Administration.

His teaching activities cover Information management, Technology Innovation and Knowledge Management, IT Governance, IT Service Management and IT Controlling at different levels. Beside academic research he participated in several company financed research and development projects, and was leading consulting projects in the above mentioned field, focusing on the telecom and financial sectors.

He is the member the John von Neumann Computer Society, member of the IT Chapter Board of the Hungarian Economic Association, and founder of the Future Internet Livinglab Budapest Association. He is editor of the Electronic Journal of Knowledge Management and official reviewer of the Journal of Knowledge Management and Practice.

Péter Kristóf was born in Budapest, Hungary, in 1981. He received his MSc degree in business administration from the University of Pécs, his MSc degree in economics from the FernUniversität in Hagen, Germany and his BSc degree in computer science from the Denis Gabor College while completing his studies at the Jönköping International Business School, Sweden.

He is a PhD candidate at the University of Pécs focusing on IT controlling and enterprise architecture. He is the CEO of a software development enterprise and the founder of a consultancy firm focusing on innovative projects and business development. He co-founded Pannon Entrepreneur Meetup and is advisory board member at the National Council for Telecommunications and Informatics.

High speed compression algorithm for columnar data storage

György Balogh

Abstract—**Lossless compression algorithms with very high compression and decompression speed are widely used in data warehouses today. Even small improvement of these algorithms can have high impact on storage space but more importantly on data access speed which effects response time of data analysis systems. We present a generic column storage compression algorithm (DictComp) with very fast compression and decompression speed. At worst the performance of the algorithm falls back to the LZ4 [4] compression algorithm but on data dominated by few values (which is very frequent in unnormalized database tables) over 5 GB/sec/CPU core decompression speed can be achieved.**

Index Terms—**Big Data, Hadoop, Impala, log analysis, cloud computing, decompression, algorithm**

I. INTRODUCTION

Storage capacity and data access speed of storage devices evolve exponentially albeit with very different exponent: data access speed is getting exponentially slower relative to storage capacity. In 1991 a typical disk had 40 MB storage which could be scanned and processed in around a minute while today the typical capacity is 2 TB yet the full scan time is more than 4 hours! Yet in cloud based log analysis services huge amount of log data have to be stored and analysed. The presented method provides an efficient storage for log data achieving compression ratio of 10-30 and similar data access speedup which translates to query response time speedup.

Distributed storage and processing (e.g.: Hadoop) can mitigate this problem: more disks can work parallel so data access speed can be scaled up linearly with the number of disks. Another possibility to boost data access speed is high speed lossless data compression. CPU speed is also evolving much faster than data access speed so there is expanding opportunities for the implementation of more and more clever compression algorithms. A new generation of so called real-time compression algorithms has developed in the last couple of years. There is nothing new in these algorithms theoretically, but the very efficient implementations and ratio of CPU speed and data access speed makes the decompression time almost negligible today. Examples of such algorithms are LZO, Snappy and LZ4. LZ4 can achieve around 1 GB/sec/core decompression speed on log data with a compression ratio of 10 boosting the 100 MB/sec sequential disk I/O limit with an order of magnitude.

LZO, Snappy and LZ4 are generic compression algorithms, however in database storage engines other structural information such as field and record boundaries are also available, which can further boost the speed of compression and decompression.

In analytical databases data is typically stored in columnar way: values of one field for a larger set of records are stored in a compressed block. Columnar storage has many advantages for analytical workloads. Columns that are not participating in a query don't have to be read, thereby sparing significant disk I/O. A list of values for the same field typically can be compressed better. More and more data is stored in an unnormalized way with many repeating values. Log files are a typical example of this: the same value (e.g.: server address) is stored over and over again even if it is the same in each case. Unnormalized storage has the advantage of data locality: all information for a record is available locally; there is no need for indirections for potentially remote data.

In this publication we present a compression algorithm for columnar data. The algorithm falls back to an LZ4 compression in worst case, but can achieve decompression speed over 5 GB/sec/core for column data dominated by a few values.

The presented compression algorithm can be applied in analytical database storage engines. One of our goals is to integrate this result to the Parquet storage format to further speed up analytical queries over data stored in Parquet format. This would reduce query response time of queries over "big data" data sets. In practice this effect BI tool performance on all kind of data sets (financial, click stream, sensor data etc.).

II. THE ALGORITHM

In the case of columnar compression, the input for the compression algorithm is a sequence of string values and the output is a byte sequence. This compression schema can be applied to table columns with boolean, text and enum types. Numerical data types needs different class of algorithms.

In case of decompression, the input is the coded byte sequence and the output is the original string sequence. However, for many query operations the original string sequence does not have to be fully materialized. These operations can be performed on compressed or 'half compressed' data.

Compression is performed in blocks with the size of typically 1000-10000 items. Each block independently

Manuscript received June 06, revised June 6. György Balogh, CTO of Logdrill Kft, a subsidiary of Kürt Zrt. Contact: gyorgy.balogh@kurt.hu

compressed contains all information for decompression. The main idea of the compression algorithm is to split the data into three parts: literals, literal lengths and dictionary indexes as shown in Figure 1. Unique literals are concatenated without string terminator and compressed with LZ4. Length of the unique literals are collected separately and compressed with a very efficient integer compression algorithm described later. The same integer compression algorithm is used to compress the dictionary indexes.

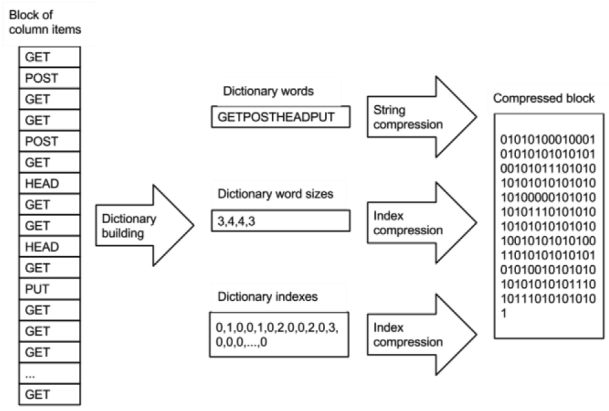


Fig. 1. The block compression process

Fig. 2. shows the compression algorithm. The dictionary index starts with 0 and increased for every new item. A fast hash function is used to recognize repeating items.

```
    virtual void addFromString(const char * str, size_t len)
    {
     Index h = murmurHash64A(str, len) % BUCKETS;
     Index ind = m_indexesByHash[h];
     if ( ind != m_unknownItemIndex && len ==
m_sizes[ind] &&
          strncmp(str, m_strings.data() + m_offsets[ind],
len) == 0)
      {
       // existing item
       m_indexes.push_back(ind);
      }
      else
      {
       // new item
       m_offsets.push_back(m_strings.size());
       m_strings.insert(m_strings.end(), str, str + len);
       m_sizes.push_back(len);
       m_indexes.push_back(0);
       m_indexesByHash[h] = ++m_nextId;
      }
    }
```

Fig. 2. Dictionary building algorithm.

III. INDEX COMPRESSION

Daniel Lemire and Leonid Boytsov have recently presented a comprehensive evaluation of different integer compression algorithms [8]. We evaluated the algorithms based on decompression speed including the disk I/O. The best candidate was the index compression algorithm (Simple8b) published by [7].

Key idea of the algorithm is to use 64 bit words to code input sequences. 4 bits are used to select the encoding schema the rest 60 bits holds the data. Example encoding schemes: 60 pieces of 1-bit numbers, 30 pieces of 2-bit numbers, 20 pieces of 3-bit numbers ... 1 pieces of 60 bit number. During decoding the 64 bit code hold a register and by repeatedly applying shift in conjunction with bitwise AND operations, this method can produce the output numbers at an extremely fast decoding speed.

```
    template<size_t LENGTH, uint8_t BITS, size_t K =
LENGTH>
    struct DecodeLiteral
    {
     static void decode(uint16_t *& out, uint64_t code)
     {
      *out++ = (code >> (60 - BITS * (LENGTH - K +
1))) & ((1U << BITS) - 1);
      DecodeLiteral<LENGTH, BITS, K -
1>::decode(out, code);
     }
    };
    template<size_t LENGTH, uint8_t BITS>
    struct DecodeLiteral<LENGTH, BITS, 1>
    {
     static void decode(uint16_t *& out, uint64_t code)
     {
      *out++ = (code >> (60 - BITS * LENGTH)) & ((1U
<< BITS) - 1);
     }
    };
```

Fig. 3. Fast integer decoding routines. To decode a number, shift and bitwise AND operations are needed. Compared to this the loop administration cost gets too high. With the help of template meta-programming, loop unrolling can be enforced leading to 100% speed up compared to a simple loop.

```
    template<size_t MAX_LENGTH, uint8_t BITS,
uint8_t SCHEMA_ID>
    static size_t encodeLiteral(const uint16_t * data, size_t
len, uint64_t * out)
    {
     const uint64_t max = (1U << BITS) - 1;
     const uint32_t n = std::min(len, MAX_LENGTH);
     *out = SCHEMA_ID;
     for (size_t i = 0; i < n; ++i)
     {
      if (data[i] > max) // termination condition
       return 0;
      *out = (*out << BITS) | data[i]; // coding
     }
     *out <<= 60 - BITS * n;
     return n;
    }
```

Fig. 4. Fast integer encoding routine. Doing the test and encoding in one loop is the key idea to speed up the coding.

Compared to the benchmark implementation of Daniel Lemire and Leonid Boytsov, we achieved significant speed up

both in compression and decompression speed. With loop unrolling (with template meta-programming) we achieved 100% decompression speedup. Decompression speed can get above 2 billion integers per second which translates to 1-2 CPU cycles per integer. In case of 32 bit numbers this translates to 8 GB/sec decompression speed.

In the baseline implementation, the test of coding schema and the actual coding are done in two steps. First it finds the best schema in a greedy manner then performs the coding in another pass. We modified the coding algorithm to do the schema test and coding in one step. The termination condition and the coding are independent and can be executed in parallel. Presumably as an effect of the superscalar execution, the added coding does not slow down the test which gives us an almost 100% coding speed up in compression speed. If a coding schema terminates due to a too large number, then the next coding will simply overwrite the invalid code generated by the previous schema.

The original index coding algorithm has only two schemas for encoding long 0 runs (run length with 120 and 240). We extended the algorithm with a new schema that can encode 6 runs in the 60 bit data part. 7 bits are used to encode the run length and 3 bits encode the data. This schema significantly increased the compression ratio (50-100%) in case of zero dominated distributions.

IV. Evaluation

First let's consider two special input distributions: all items are different and only a few items are different. In the first case dictionary index will always be zero (zero index means new item). In this case the all zero indexes will compress extremely well (with run length encoding) with negligible decompression time overhead so the compression will be an LZ4 compression basically. In the other case the dictionary indexes will be small and again compress very well. In this case the literal string will be short so the decompression time will be dominated by the index decompression speed. Depending on the distribution, the decompression will be a mixture of LZ4 decompression and index decompression with decompression speed ranging from around 1 GB/sec (LZ4 dominated) to over 5 GB/sec (Index decompression dominated).

We tested the compression algorithm on two realistic big data datasets:

- US domestic flight statistical database [6]. (10 million records, 29 fields, 1.6 GB in CSV).
- Web server logs of the 1998 World Cup [1]. (500 million records, 1 fields, 1.9 GB in CSV).

Measurements were performed on a $1000 class laptop with 8GB RAM, Intel Core I7 processor having 4 cores running at 2GHz. Operating system was Ubuntu 11.04. Measured sequential disk read speed is 70MB/sec.

We measured compression performance (bulk load) and some simple SQL queries. Query plans are hand coded over our column compression storage. We selected three columnar database engines for comparison: InfoBright [3], MonetDB [5] and Cloudera Impala [2]. We tested the InfoBright Community Edition (version 4.0.7), MonetDB v1.0 Jul2012-SP2 and Impala 1.1. All are 64-bit versions.

Table 1 shows the measured execution times.

	Flight29			*Web500M*		
	Bulk load	Cold query	Warm query	Bulk load	Cold query	Warm query
MonetDB	59.21	0.92	**0.04**	167.32	7.12	0.95
DictComp	**17.31**	**0.61**	0.05	**29.51**	**2.34**	**0.70**
InfoBright	78.13	1.67	0.92	130.34	70.96	70.94
Cloudera Impala	57.35	4.90	0.46	97.79	47.7	15.82

Table 1. Bulk load and query execution times (in seconds) of the engines on the two datasets. In case of bulk load the input file was always in the file cache. For queries we tested the query when the file and database caches were cleared (cold run) and after multiple runs of the same query (caches are warmed up).

In query performance DictComp and MonetDB are close. DictComp gets significantly better in I/O bound cold queries, for warm queries they almost exactly match. InfoBright simply fell short with the *Web500M* dataset, even a grep can perform better (6 seconds from file cache) on the original uncompressed dataset. The main reason why InfoBright cannot handle record numbers of this magnitude comes from its old architecture, which is inherited from the MySQL framework. The MySQL storage interface forces the InfoBright storage engine to copy every single item to the MySQL plan executor. So the whole dataset have to be decompressed and copied item by item. This makes the InfoBright query execution CPU bound on the *Web500M* dataset, which then results in the same execution time for cold and warm queries.

V. Acknowledgement

This publication/research has been supported by the European Union and Hungary and co-financed by the European Social Fund through the project TÁMOP-4.2.2.C-11/1/KONV-2012-0004 - National Research Center for Development and Market Introduction of Advanced Information and Communication Technologies.

VI. References

[1] 1998 world cup web site access logs. http://ita.ee.lbl.gov/html/contrib/WorldCup.html

[2] Cloudera Impala. http://www.cloudera.com/content/cloudera/en/products-and-services/cdh/impala.html

[3] Infobright. http://www.infobright.com

[4] The LZ4 lossless compression algorithm. http://code.google.com/p/lz4

[5] Monetdb. http://www.monetdb.org

[6] USA domestic flights info data. http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time

[7] Vo Ngoc Anh and Alistair Moffat. Index compression using 64-bit words. Softw., Pract. Exper., 40(2):131–147, 2010

[8] Daniel Lemire and Leonid Boytsov. Decoding billions of integers per second through vectorization. CoRR, abs/1209.2137, 2012.

György Balogh is the CTO of LogDrill Kft. György received his computer science degree from the University of Szeged and has 20 years of data mining and machine learning experience. György spent 6 years at Vanderbilt University in Tennessee as a researcher and developed the sensor fusion algorithms of the first distributed shooter localization system. Currently György Balogh is working on the LogDrill product family specialized for log and big data analytics.

New challenges and comprehensive risk management approaches in modern supply chains

Gyula Hegedűs, Kirilka Peeva-Müller

Abstract*—The highly specialized processes of modern supply chains, the diverse geographic distribution of suppliers and the complex electronic control systems are posing numerous new threats to companies in a global market environment more competitive than ever before. This article is about the risk-related challenges corporations face today as well as about the methods and practices they should consider applying in order to ensure the continuity of their daily operation and their supply chain in particular. In the introduction we present the current trends and fundamental challenges regarding modern supply chains. In the next chapter we introduce the method of ITBCP as a relatively new, but highly useful tool of managing IT-related operational risk factors and provide a high-level comparison with classic BCP methodologies. The third chapter is about the new challenges companies face in regard to various operational risks, while the fourth chapter briefly describes the most commonly used techniques of identifying and managing these threats. The last chapter serves as a conclusion, a summary why corporations need to take sophisticated preventive measures in order to minimize their risks threatening operational continuity.

***Index Terms*—business impact analysis (BIA), global challenges, IT business continuity planning (ITBCP), just-in-time (JIT), risk management, supply chains**

I. INTRODUCTION

CAN we precisely tell the amount of risk a large enterprise bears by not paying adequate attention to ensuring its operational continuity? Do we know about the risks threatening its business operations? Do we know about the nature and size of losses that may occur for each particular business area in case of any threat causing significant negative effects on the productivity of its resources?

The root of the problem is that companies are unable to answer these types of questions without executing proper in-depth systematic analyses on a regular basis. One possible consequence of this lack of knowledge is that organizations get exposed to an increasing number of risk factors making them vulnerable in a number of different, unknown ways potentially causing severe unexpected downtime.

As corporations are innovating and developing faster than ever in the course of history, the number of potential risk factors is increasing proportionately with the application of the more advanced, more sophisticated and more complex technologies infiltrating into all areas of the business. Always keeping one eye on the risks and having an up-to-date risk management plan for unexpected crisis situations can often mean life and death in the extreme competitive world we live in today. A leading-edge company can't afford regular unexpected outage of its resources, since it can lead to losing the competitive advantage over a competitor who pays greater attention to keeping its risks on a manageable level.

That is particularly true for modern supply chains, where cost and time became the two factors that are more important today than ever before. The increasingly popular just-in-time (JIT) methodology is just one of the several ways of cutting costs and optimizing the supply chain by reducing the amount of goods and materials a firm holds in stock in one moment in time, while other time management techniques of the lean methodology help companies in delivering to its clients faster and in a more reliable way.

One of the great examples of modern supply chain management is the automotive industry. Cars are an example of global products containing research and design features which involve the collaboration of experts of various nations. The components are usually produced in many separate locations and involve a range of assembly and sub-assembly activities [1]. They are likely to meet at a final assembly line, where parts from all across the world are waiting to get installed in the particular vehicle they were ordered to several months earlier. Such an extremely interdependent system can't tolerate unexpected downtime, thereby paying constant and profound management of operational risks in the supply chain is critical business success.

II. THE CRITICAL ROLE OF ITBCP IN SUPPLY CHAINS

As it is already obvious it's impermissible for a large corporation to fall out of its normal way of operation for a significant amount of time, which considering the increased competition as well as the high technological dependency becomes shorter every day. Business continuity is endangered by increasingly less transparent threats, due to the complex partnerships with suppliers, the increasingly complex

Manuscript received May 15, revised June 4.

Gyula Hegedűs is the Head of Business Analytics Department at KÜRT Information Management and Data Recovery Co. in Budaörs, Hungary (e-mail: gyula.hegedus@kurt.hu).

Kirilka Peeva-Müller is an Associate Professor of Management at KBTU Business School in Almaty, Kazakhstan and a Visiting Lecturer in Management Skills at the International Business School in Budapest, Hungary (e-mail: kira.mueller@gmail.com).

technologies and the increasing IT and telecommunications support necessary for the daily operation of different business processes.

A business continuity strategy should ideally address all possible areas of a company where a failure or resource outage could cause significant damage or operational disruption, and achieving the desired KPIs becomes simply impossible. As almost all major processes concerning supply chains today are based on different IT-related resources and services, within the general business continuity the IT-oriented business continuity (IT Business Continuity Planning - ITBCP) bears more weight day by day.

As a result of a series of modernization projects supply chain management systems are able to control the processes with minimal human intervention, if the necessary resources and functions are continuously ensured. However a failure of the system or its electrical or IT components can directly set back the supply chain processes, or worse, the loss of stored data which can cause even greater problems.

The methodology for executing an IT-focused business continuity planning project is highly similar to that of a regular BCP project, yet the pool of resources to be examined as well as the list of threats and the content of the action plans will be significantly different. Critical processes and resources should be identified with the help of a thorough business impact analysis, the result of which should serve as the basis for determining the list of scenarios for developing the action plans. Apart from the detailed and operative description of duties that need to be performed in case of a crisis event the plans should also include the preparatory tasks required to run the alternative work process, as well as the duties of the aftercare phase in order to make sure all the data are loaded back into the systems once it's again up and running.

III. EMERGING RISK FACTORS AFFECTING SUPPLY CHAIN STABILITY

The unstable global economy we have experienced in the last couple of years has multiplied supplier failures worldwide, thereby assessing and continuously monitoring suppliers' likelihood of potential financial failure has also become more important than ever. The process of regularly assessing and identifying critical suppliers that are most critical to the business has become an everyday routine for most successful corporations, and generating business continuity plans for supplier failure scenarios has also been the general practice in the last several years. Identifying key risk indicators and setting up early warning signals is essential to preventing or mitigating the impact of supply chain breakdowns. Since no matter what the root cause of such a disaster is, the public will always hold the company that owns the brand accountable for the negative event [2].

Adapting to shifts in the market – such as high currency fluctuations or quick changes in demand – has also become a major difficulty corporations need to tackle every day; the company most successful in foreseeing and reacting to rapid market events might be the one leading the market ahead of its competition sometimes for years to come.

Fig. 1. General categories of business risks affecting supply chains

Global sourcing has been the trend for the last two decades in a number of different sectors, however numerous threats have emerged in the recent years that have not been thought about when choosing this form of operation in the first place. These include the constantly increasing wages in developing markets, the increasing costs of logistics due to the increase of oil prices and the growing price of services such as translation, legal fees and licenses [3]. Usually the savings are much smaller today than they were in the time of setting up this form of operation. The types of risks most frequently associated with global supply chains can be related to distance, communicational difficulties and cultural differences.

Understanding cultural differences are one of the most critical factors corporations need to manage when considering operating a global supply chain. In the western world generating profit is usually the fundamental goal behind most business decisions. Companies always try to employ the optimal number of people to get the job done effectively. However, in some parts of Asia, the mentality is to keep as many people employed as possible no matter how little they are contributing to the overall goal. Companies setting up operations for example in China must precisely verify production capacities and workforce quality to ensure that everyone has a meaningful job. Another example is Central America, where workers may refuse to work until an exorcism has been performed in the premises they think may be haunted by ghosts [4]. When evaluating a target country for outsourcing particular parts of the operation it's not enough to analyze the region's political and legal environment, it's also crucial to get as much information as possible about the local habits, traditions and religion. Not taking these into consideration will lead to the emergence of various unknown risk factors that can cause serious headache for the top management and in some cases undermine the complete operation in a foreign country.

IV. IDENTIFYING AND MANAGING RISKS

Identification is the first, yet the most fundamental step in any risk management methodology. Once a certain risk has been identified it becomes a so-called "known" risk. Known

risks can be assigned to different resources or processes, a probability of occurrence and a potential impact can be estimated with the help of various techniques. Unidentified risks can be called "known unknown" or "unknown" risks. The former can be managed using contingencies, while the second is totally unknown and therefore isn't possible to prevent and consequently to manage. A challenge during a risk identification process is to reduce the presence of unknown risks, thereby enhancing the effectiveness and accuracy of the company's risk management plan.

There are four fundamental ways of managing known risks:

1) Risk Avoidance: Completely avoiding the activity that poses the potential risk. While it might sound attractive at first glance, it will likely forfeit most potential gains that would come with the particular resource or activity, thereby applying it requires careful analysis of all possible outcomes.

2) Risk Reduction: This is the classic idea of reducing the extent or possibility of a potential loss. This can be done by increasing precautions or limiting the amount of risky activity.

3) Risk Transfer: Risk is transferred to a third-party entity (usually to an insurance company). In this case the probability of occurrence or potential impact is not reduced in any way, only the financial responsibility is transferred to the external party.

4) Risk Retention: It simply means accepting the risk. It's usually effective in case of smaller risks that don't pose a significant financial threat to the company [5].

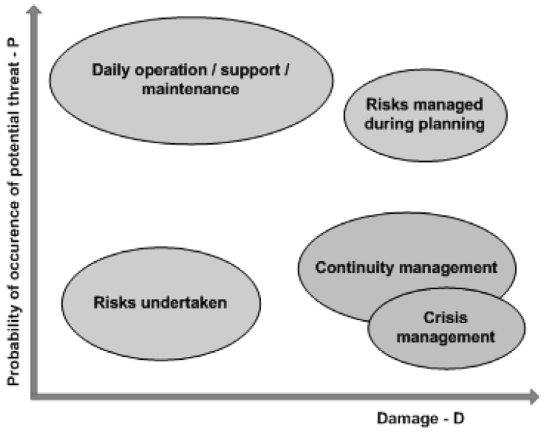


Fig. 2. Types of risk management measures required under different circumstances

Business continuity planning – a method touched upon earlier in this article – is a classic example of risk reduction. After risks, potentially affected resources and critical business processes have been identified with the help of a business impact analysis, alternative working methods are prepared to enable the continuity of processes crucial for normal operation. This way the financial loss resulted from the incident is significantly less than if the affected business processes stopped operating for the entire length of the breakdown. It's clear to see why taking preventive measures is such an important factor in risk management, while corrective measures (actions need to be performed once a risk has already occurred) should be kept at a minimum possible level in order to minimize the potential loss a threat poses to the everyday operation (and financials) of a company.

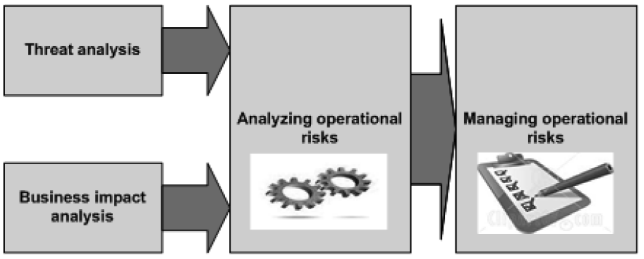


Fig. 3. Classic operational risk management workflow

V. SUMMARY

It is crucial for any company today to minimize the number of unplanned outages in their supply chains as well as the time spent with unscheduled troubleshooting and maintenance activities in order to meet the extremely high requirements set by the industry's other players in an environment more competitive than ever before.

In the same time supply chain processes, geographic distribution of suppliers and supporting electronic control systems are becoming more and more complex and interrelated, therefore business continuity is threatened by new, less transparent and more diverse risk factors every day. Business and supply chain continuity cannot be achieved by itself, today companies need highly sophisticated and regularly updated risk management strategies in order to meet the high industry standards required to stay competitive in the market.

ACKNOWLEDGMENT

This publication has been supported by the European Union and Hungary and co-financed by the European Social Fund through the project TÁMOP-4.2.2.C-11/1/KONV-2012-0004 - National Research Center for Development and Market Introduction of Advanced Information and Communication Technologies.

REFERENCES

[1] "Managing the supply chain for globally integrated products" – Available at: http://businesscasestudies.co.uk/exel/managing-the-supply-chain-for-globally-integrated-products/just-in-time-and-lean-production.html

[2] "Supply chain risk assessment", PwC's Knowledgeline, June 2009 – Available at: http://www.pwc.com/en_GX/gx/pharma-life-sciences/pdf/supply-chain-risk-assessment.pdf

[3] Barry Cross, Jason Bonin, "How to manage risk in a global supply chain", Ivey Business Journal Nov/Dec 2010 – Available at: http://iveybusinessjournal.com/topics/strategy/how-to-manage-risk-in-a-global-supply-chain

[4] Scott Hudson, "Cultural Effects on the Global Supply Chain", April 2005 – Available at: http://scm.ncsu.edu/scm-articles/article/cultural-affects-on-the-global-supply-chain

[5] "Ways of managing risks", Mays Financial – Available at: http://www.maysfinancial.com/insurance/ways-managing-risk/

Gyula Hegedűs received his B.Sc. degree in economics from the Corvinus University of Budapest and his M.Sc. in IT management and M.B.A. degrees from the Central European University Business School in Budapest, Hungary.

He has been working as a senior consultant for a number of years specialized in business continuity planning in the CEE region. By having extensive experience in the governmental, commercial and manufacturing sectors he has lead projects and successfully implemented complete business continuity management systems at a number of different corporations and participated in the development of a unique business continuity method designed specifically for manufacturing environments.

Kirilka Peeva-Müller is an Associate Professor of Management at KBTU Business School in Almaty, Kazakhstan and a Visiting Lecturer in Management Skills at the International Business School in Budapest, Hungary. She specializes in management and innovation development with particular focus on transition economies; her fields of work include researching the risks regarding the human factor in today's corporate environments.

WCNC 2015

IEEE WIRELESS COMMUNICATIONS AND NETWORKING CONFERENCE

NEW ORLEANS, LA, USA • 9-12 March, 2015

IEEE | IEEE COMMUNICATIONS SOCIETY

www.ieee-wcnc.org

IEEE WCNC is the premier event for wireless communications researchers, industry professionals, and academics interested in the latest development and design of wireless systems and networks. Sponsored by the IEEE Communications Society, IEEE WCNC has a long history of bringing together industry, academia, and regulatory bodies. In 2015, New Orleans will become the wireless capital by hosting IEEE WCNC 2015. The conference will include technical sessions, tutorials, workshops, and technology and business panels. You are invited to submit papers in all areas of wireless communications, networks, services, and applications. The instructions for authors will be posted on the conference website www.ieee-wcnc.org/2015. Potential topics include, but are not limited to:

CALL FOR PAPERS

Track 1: PHY and Fundamentals

- Interference characterization
- Cognitive radio, ultra-wideband
- Multihop and cooperative communications
- Modulation, coding, diversity
- Equalization, synchronization, channel estimation
- Space-time coding, MIMO, adaptive antennas
- OFDM, CDMA, spread spectrum
- Channel modeling and characterization
- Interference cancellation and multiuser detection
- Iterative techniques
- Information-theoretic aspects of wireless communications
- Signal processing for wireless communications
- Propagation models for high frequency channels

Track 2: MAC and Cross-Layer Design

- Multiple access techniques
- Cognitive and cooperative MAC
- Collaborative algorithms
- MAC for mesh, ad hoc, relay, and sensor networks
- Network information theory
- Radio resource management and allocation, scheduling
- Cross-layer design, cross-layer security
- Software defined radio, RFID
- Adaptability and reconfigurability
- Wireless MAC protocols: design and analysis
- B3G/4G Systems, WiMAX, WLAN, WPAN
- QoS provisioning in MAC

Track 3: Mobile and Wireless Networks

- Localization for wireless networks
- Network estimation and processing techniques
- Mesh, relay, sensor and ad hoc networks
- Mobility, location, and handoff management
- Mobile and wireless IP
- Wireless multicasting, routing
- Robust routing
- Multimedia QoS and traffic management
- Wireless broadcast, multicast and streaming
- Congestion and admission control
- Proxies and middleware for wireless networks
- Wireless network security and privacy
- Performance of E2E protocols over wireless networks
- Interworking heterogeneous wireless/wireline networks
- Capacity, throughput, outage, coverage

Track 4: Services, Applications, and Business

- Emerging wireless/mobile applications
- Context and location-aware wireless services & applications
- Wireless telemedicine and e-health services
- Intelligent transportation systems
- Cognitive radio and sensor-based applications
- Content distribution in wireless home environment
- Wireless emergency and security systems
- Service oriented architectures, service portability
- SIP based services, multimedia, QoS support, middleware
- Innovative user interfaces, P2P services for multimedia
- Dynamic services, autonomic services
- Regulations, standards, spectrum management
- Test-bed and prototype implementation of wireless services
- Personalization, service discovery, profiles and profiling

CALL FOR TUTORIALS AND WORKSHOPS

Proposals for tutorials and workshops are solicited on hot topics for future wireless communications systems and applications.

CALL FOR PANELS

Panel proposals are also solicited on technical, business and policy-related issues and opportunities for the wireless communications industry.

PLEASE NOTE: To be published in the IEEE WCNC 2015 Conference Proceedings and to be eligible for publication in IEEE Xplore®, an author of an accepted paper is required to register for the conference at the full or limited (member or non-member) rate and the paper must be presented by an author of that paper at the conference unless the TPC Chair grants permission for a substitute presenter arranged in advance of the event and who is qualified both to present and answer questions. Non-refundable registration fees must be paid prior to uploading the final IEEE formatted, publication-ready version of the paper. For authors with multiple accepted papers, one full or limited registration is valid for up to 3 papers. Accepted and presented papers will be published in the IEEE IEEE WCNC 2015 Conference Proceedings and submitted to IEEE Xplore®.

A portion of the accepted papers will be presented as posters. The IEEE WCNC 2015 Technical Program Committee will decide which papers will be presented in oral (lecture type) sessions and which papers as posters, and the decisions will be announced to the authors. Choice between oral and poster presentations will be totally independent of the review scores and of the paper quality.

IMPORTANT DATES

Paper submission deadline:	**1 August 2014**	**Tutorial proposals:**	**1 August 2014**
Notification of acceptance:	**15 November 2014**	**Workshop proposals:**	**1 June 2014**
Final Camera-ready papers due:	**15 December 2014**	**Panel proposals:**	**1 August 2014**

General Chair:
José Roberto B. de Marca,
Pontifical Catholic University of Rio de Janeiro, Brazil

Executive Chair:
Richard Miller,
AT&T (retired), USA

Technical Program Chair:
Nirwan Ansari,
New Jersey Institute of Technology, USA

Technical Program Vice-Chair:
Edit Kaminsky Bourgeois,
University of New Orleans, USA

WCNC Steering Committee Chair:
Khaled Letaief,
Hong Kong University of Science and Technology, Hong Kong

WCNC Advisor:
Sherman Shen,
University of Waterloo, Canada

SCIENTIFIC ASSOCIATION FOR INFOCOMMUNICATIONS

Who we are

Founded in 1949, the Scientific Association for Infocommunications (formerly known as Scientific Society for Telecommunications) is a voluntary and autonomous professional society of engineers and economists, researchers and businessmen, managers and educational, regulatory and other professionals working in the fields of telecommunications, broadcasting, electronics, information and media technologies in Hungary.

Besides its more than 1300 individual members, the Scientific Association for Infocommunications (in Hungarian: Híradástechnikai Tudományos Egyesület, HTE) has more than 60 corporate members as well. Among them there are large companies and small-and-medium enterprises with industrial, trade, service-providing, research and development activities, as well as educational institutions and research centers.

HTE is a Sister Society of the Institute of Electrical and Electronics Engineers, Inc. (IEEE) and the IEEE Communications Society. HTE is corporate member of International Telecommunications Society (ITS).

What we do

HTE has a broad range of activities that aim to promote the convergence of information and communication technologies and the deployment of synergic applications and services, to broaden the knowledge and skills of our members, to facilitate the exchange of ideas and experiences, as well as to integrate and harmonize the professional opinions and standpoints derived from various group interests and market dynamics.

To achieve these goals, we…

- contribute to the analysis of technical, economic, and social questions related to our field of competence, and forward the synthesized opinion of our experts to scientific, legislative, industrial and educational organizations and institutions;
- follow the national and international trends and results related to our field of competence, foster the professional and business relations between foreign and Hungarian companies and institutes;
- organize an extensive range of lectures, seminars, debates, conferences, exhibitions, company presentations, and club events in order to transfer and deploy scientific, technical and economic knowledge and skills;
- promote professional secondary and higher education and take active part in the development of professional education, teaching and training;
- establish and maintain relations with other domestic and foreign fellow associations, IEEE sister societies;
- award prizes for outstanding scientific, educational, managerial, commercial and/or societal activities and achievements in the fields of infocommunication.

Contact information

President: **DR. GÁBOR HUSZTY** • *ghuszty@entel.hu*

Secretary-General: **DR. ISTVÁN BARTOLITS** • *bartolits@nmhh.hu*

Managing Director, Deputy Secretary-General: **PÉTER NAGY** • *nagy.peter@hte.hu*

International Affairs: **ROLLAND VIDA, PhD** • *vida@tmit.bme.hu*

Addresses

Office: H-1051 Budapest, Bajcsy-Zsilinszky str. 12, Room: 502

Phone: +36 1 353 1027, Fax: +36 1 353 0451

E-mail: *info@hte.hu*, Web: *www.hte.hu*