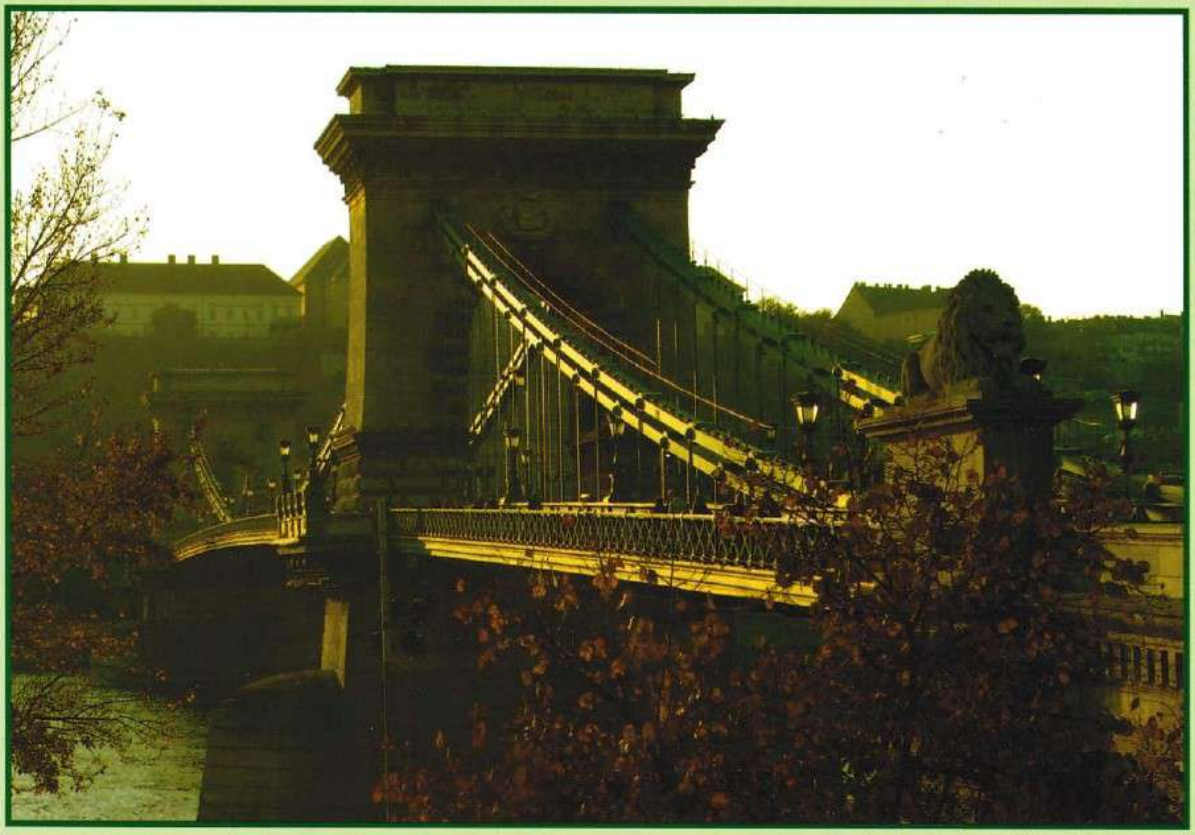
híradástechnika

1945 VOLUME LXI. 2006

info-communications-technology



Optical communications

Speech technology

World Telecommunications Congress 2006

Selected Papers

2006/7

Journal of the Scientific Association for Infocommunications with co-operation with the National Council of Hungary for Information and Communications Technology

Contents

FOREWORD 2

Orsolya E. Ferencz, Csaba Ferencz
New directions in the wave propagation theory 3

Péter Jeszenői, Jenő Szatmári
Physical limits of the applicability of 10 and 40 Gbps speed DWDM systems 8

András Kern, György Somogyi, Tibor Cinkler
Applying statistical multiplexing and traffic grooming in optical networks jointly 17

Klára Vicsi, Szabolcs Velkei, György Szaszák, Gábor Borostyán, Géza Gordos
Speech recognizer for preparing medical reports:
Development experiences of a Hungarian speaker independent continuous speech recognizer 22

Géza Kiss, Géza Németh
Machine learning algorithm for automatic labeling and its application in text-to-speech conversion 28

Tamás Bérczes, János Sztrik
Performance evaluation of Proxy Cache Servers 36

Ákos Horváth, Dániel Varró, Gergely Varró
Automatic generation of platform-specific transformation 40

Peter Janeck
NGN development at Magyar Telekom: The future of our fixed network 46

Anett Schülke, Daniele Abbadessa, Florian Winkler
Service delivery platform: Critical enabler to service provider's new revenue stream 50

Ágnes Füredi
WTC2006 – World Telecommunications Congress, April 30-May 3 Budapest, Hungary 56

ZTE IPTV: a great IPTV solution from China 58

Foreword

szabo@hit.bme.hu

We are continuing with the practice of publishing regularly English issues, at present twice a year, in July and in January. We hope we will be able to gradually increase their number from two per year to maybe four per year, yielding to shorter waiting times for those wishing to submit their results for these issues directly which we welcome.

This issue also reports on a recent important international conference, World Telecommunications Congress, WTC2006, organized jointly by the Scientific Association for Infocommunications, Hungary (HTE) and the Association for Electrical, Electronic & Information Technologies, Germany (VDE/ITG). The conference took place in Budapest, Hungary, on April 30 - May 3, 2006, under the title "Emerging Telecom Opportunities". We publish a short summary of the event and we have selected two papers for our English issue.

The paper by P. Janeck titled "NGN development at Magyar Telekom: The future of our fixed network". It reports on an ambitious development Magyar Telekom is running to deploy broadband access and to build an IMS based NGN network.

The paper by Schülke et al titled "Service Delivery Platform: Critical enabler to service providers's new revenue stream". It gives an overview of the Service Integration Environment as a potential part of a future SDP solution with an in-depth view of the respective market and its relation to the ongoing standardization activities.

Now let us briefly introduce the papers selected by the Editorial Board from those already published in the Hungarian issues for the past five months.

Orsolya Ferencz and Csaba Ferencz presents a new method for wave propagation calculations that resolves the contradictions in the existing methods. The method gives opportunity to find new, exact and right solutions, to avoid the former mistakes, and by the aid of which several measurements in space research can be successfully interpreted.

The paper of Jeszenői and Szatmári deals with an important issue of increasing the capacity of high speed optical backbone systems. Here we have to cope with the non-linear properties of the optical fibre which is the target of the investigation in this paper.

Kern et al approach to the capacity increasing of wavelength division multiplexing systems from another angle: they show that the combination of statistical multiplexing and traffic grooming can lead to the increase of the efficiency of the system.

Vicsi et al report on a new development in speech technology that can improve the efficiency of doctors' work in the hospitals by automatically generating the diagnosis report based on speech input.

Kiss and Németh present a novel machine learning approach usable for text labelling problems and illustrate the importance of the problem for text-to-speech systems and through that for telecommunication applications.

Bérczes and Sztrik targets the modeling of the Web traffic. The primary aim of their paper is to modify the performance model of Bose and Cheng to a more realistic case when external arrivals are also allowed to the remote Web servers with limited buffer capacity.

The paper of Horváth et al was specifically submitted for the English issue and presents a new approach using generic and meta-transformations for generating platform-specific transformer plug-ins.

Let us note that this time the selection was more difficult than usually due to a large number of quality research papers published during the preceeding five months, and we were not able to include in this issue two more papers, they will be published in the next English issue to appear in January.

László Zombory
Chairman of the Editorial Board

Csaba A. Szabó
Editor-in-Chief

New directions in the wave propagation theory

ORSOLYA E. FERENCZ, CSABA FERENCZ

Space Research Group, Institute of Geography and Earth Sciences, Eötvös University Budapest
spacerg@sas.elte.hu

Reviewed

Keywords: wave propagation, modes, inhomogeneity

The paper presents a contradiction, originated from a root fallacy, which is commonly accepted and applied in the wave propagation calculations up to now, but yields wrong results. Further, a solving method will be briefly overviewed, by application of which it became possible to deduce new and exact solutions, to avoid the former errors, and to interpret successfully several registrations in space research.

1. Introduction

In the case of many important wave propagation problems we cannot avoid to create more and more accurate models of the physical phenomena and the structure of the propagating signals. One of the most sensitive topics is the exact description of the signals rising in inhomogeneous media, apart from the extremely strong inhomogenities needing scattering calculations. The known and commonly applied models (e.g. W.K.B. description, Airy-functions, Stokes equation, eikonal-equation, generalized propagation vector, etc. [1]) involve fundamental misunderstanding regarding the structure of the signal. To enlighten this problem we demonstrate this contradiction in a simple example.

2. The structure of the signal

As a simple case, let a strictly monochromatic signal propagate in a linear, isotropic, time-invariant, lossless medium containing spatial inhomogeneity. In this case, a part of the signal reflects point by point during going through the medium, while the amplitude of the forward propagating signal-part will attenuate. From the simplicity of the model it is obvious that the permittivity can be defined as scalar $\varepsilon(\bar{r})$. Further simplification is assuming the permeability as μ_0.

Consequently, the form of the signal is:

$$\overline{G}(\bar{r},t) \triangleq \overline{G}_0(\bar{r})\, e^{j[\omega t - \varphi(\bar{r})]} \tag{1}$$

where $\bar{G}$ means $\bar{E},\bar{B},\bar{D},\bar{H}$, functions, $\bar{r}$ location vector, t is the time, ω is the angular frequency, φ is the phase.

In our case, the forms of Maxwell's equations are as follows:

$$\begin{aligned} \overline{\nabla}\times\overline{H} &= j\omega\varepsilon\overline{E}, \\ \overline{\nabla}\times\overline{E} &= -j\omega\mu_0\overline{H}, \\ \overline{\nabla}\cdot\overline{H} &= 0, \\ \overline{\nabla}\cdot(\varepsilon\overline{E}) &= 0 \end{aligned} \tag{2}$$

from which it can be obtained by the known way that the third and the fourth equations will be automatically fulfilled, if the first two equations are fulfilled, in so far as the medium is not characterized by distributions (the functions are derivable continuously).

So, the equations to be solved are the following:

$$\begin{aligned} (\overline{\nabla}\times\overline{H}_0) - j\overline{\nabla}\varphi\times\overline{H}_0 &= j\omega\varepsilon\overline{E}_0, \\ (\overline{\nabla}\times\overline{E}_0) - j\overline{\nabla}\varphi\times\overline{E}_0 &= -j\omega\mu_0\overline{H}_0. \end{aligned} \tag{3}$$

Introducing the $\bar{k} \triangleq \overline{\nabla}\varphi$ and $\bar{k}\times\bar{u} \triangleq \overline{\overline{k}}\cdot\bar{u}$ notations (where $\bar{u}$ is arbitrary vector), $\bar{G}_0$ and φ assuming that (as is usual in the simplest cases) and are real functions, the equation-system to be solved will be disintegrated into two groups. The real part is

$$\begin{aligned} \overline{\nabla}\times\overline{H}_0 &= 0, \\ \overline{\nabla}\times\overline{E}_0 &= 0; \end{aligned} \tag{4}$$

while the imaginary part is

$$\begin{aligned} \bar{k}\times\overline{H}_0 &= -\omega\varepsilon\overline{E}_0, \\ \bar{k}\times\overline{E}_0 &= \omega\mu_0\overline{H}_0. \end{aligned} \tag{5}$$

(This separation is explained in the literature by the argumentation that weakly inhomogeneous medium is considered, in which the variation of the medium-parameters is very slow. But it is obvious that this assumption itself means a strong restriction in the validity limits of these models.)

As this separation automatically results that the law of the conservation of the energy cannot be fulfilled for the two parts separately, the W.K.B. philosophy eliminate this contradiction by introducing an additional condition regarding the constancy of the energy of the propagating signal.

(4) and (5) are investigated one by one. On the one hand, solution of (5) leads to the well-known dispersion-equation,

$$\left|\overline{\overline{k}}\,\bar{k} + \omega^2\varepsilon\mu_0\overline{\overline{1}}\right| = 0, \tag{6}$$

from that

$$k^2 = \omega^2\varepsilon\,\mu_0, \text{ and } k = \pm\omega\sqrt{\varepsilon\mu_0}, \tag{7}$$

can be obtained for the propagation vector, forecasting a forward and a backward propagating solution as a result. On the other hand (4) delivers a solution,

completely independent from (5), the solution of which is always

$$\bar{H}_0 = \text{constant} \quad \text{and} \quad \bar{E}_0 = \text{constant} \tag{8}$$

for the amplitudes. However, (9) is theoretically impossible in an inhomogeneous medium, and leads back to an obvious contradiction in comparison with (8).

What can be the reason of this contradiction? This evidently has to be hidden in the structure of the signal. In the traditional conceptions dealing with inhomogeneity the forward propagating and the reflected signals are taken into consideration during the derivation, as if they were the solutions of Maxwell's equations singly. As it is a well-known fact in the mathematics, the sum of several independent solutions of a linear differential equation-system is also a solution of that. But decomposing a known solution into additive parts, it cannot be assumed to be generally true, that these parts could be solutions of the original equation-system. The physical picture is clearer. In order to handle the forward propagating and the reflected signal independently, we must consider them to exist alone, as the solution of Maxwell's equations (and some coupling or relation between them can be created during the computation, by defining additional assumptions). However, the presence of the inhomogeneity automatically causes the reflection of the signal, so the propagating and the reflected signal-parts can appear only and exclusively together in inhomogeneous media, and not independently of each other.

To see the problem in more detail, let the application of the Stokes equation and Airy functions be examined [1,5].

As it can be found e.g. in Budden's book ([1] – chapters 9. and 15.), it is a routine procedure to lead back Maxwell's equations to the so called Stokes equation for inhomogeneous cases:

$$\frac{d^2E_y}{dz^2} + k_0^2 q^2 E_y = 0 \tag{9}$$

where

$$q^2 = n^2 \text{ (for longitudinal propagation)} \tag{10}$$

n is the refraction coefficient.

As this is well seen in Budden's deduction, he supposes the starting form of the signal as the sum of the propagating and the reflected parts:

$$E_y = A \cdot \mathrm{e}^{-\mathrm{j}k_z z} + B \cdot \mathrm{e}^{\mathrm{j}k_z z} \tag{11}$$

where $k_z = k_0 \cdot n = \frac{\omega}{c} \cdot n$

In the further deduction Budden states, that this signal form shown in (12) is used during solving the Stokes equation, the known solutions of which are the Airy-functions.

But in the followings Budden substitutes back the forward and backward propagating parts into Maxwell's equations separately, obtaining formally identical equations. After this he solves Maxwell's equations also separately, for the forward propagating and the reflected signals, and not for the sum of them.

However, as we mentioned above, the solution of Maxwell's equations can be only and exclusively the resultant sum of the two signal-parts, because these signals cannot appear and fulfill the equations independently in the presence of spatial inhomogeneity. Let us control Budden's calculations for the resultant sum of the two signal-parts from (12), writing back their sum into the Stokes equation.

Considering Budden's assumption, whereas A and B are constants (although we must emphasize that this means strong restriction in the validity of the model) let (12) be rewritten into the Stokes equation. In this case the following will be yielded:

$$A = -B \cdot \mathrm{e}^{\mathrm{j}2k_z z} \tag{12}$$

This leads back to an obvious contradiction again, as from (13) A and B cannot be constants. Budden's solution therefore regards the forward and backward propagating signals independently, which cannot be assumed in inhomogeneous medium (and originally Budden has neither assumed.

If A and B are not constants and we write back (12) into the Stokes equation, the given forms will more widely differ from Budden's results, as no such differential equation will arise, the solution of which could be the Airy functions, but a more complicated relation can be written between A and B:

$$\left[-2\mathrm{j}\frac{\mathrm{d}A}{\mathrm{d}z}\left(\mp k_0 q\right) - \mathrm{j}A\left(\mp k_0 \frac{\mathrm{d}n}{\mathrm{d}z}\right) + \frac{\mathrm{d}^2 A}{\mathrm{d}z^2}\right] \cdot \mathrm{e}^{-\mathrm{j}k_z z} + \\ + \left[2\mathrm{j}\frac{\mathrm{d}B}{\mathrm{d}z}\left(\mp k_0 q\right) + \mathrm{j}B\left(\mp k_0 \frac{\mathrm{d}n}{\mathrm{d}z}\right) + \frac{\mathrm{d}^2 B}{\mathrm{d}z^2}\right] \cdot \mathrm{e}^{+\mathrm{j}k_z z} = 0 \tag{13}$$

This relation on the one hand is not identical with the one deduced by Budden, and on the other hand, this results in an unsolvable underdetermined mathematical description.

By our investigation it turned out obviously, that the inhomogeneous computing methods using the Stokes equation involve implicitly the wrong and contradictive assumption, according to which the propagating and the reflected signals can exist and can be deduced from Maxwell's equations independently. This conclusion is valid independently from the nature of the signal (monochromatic or UWB transient).

3. Method of Inhomogeneous Basic Modes (MIBM)

As it was enlightened in details above, a wrong approach referring to the structure of the signal can cause fundamental inherent inconsistency in the solution. How could it be possible to avoid this? We have to assume such signal structure, which contains the resultant sum of all the possibly existing signals in each spatial and temporal point along the propagation path. We have to start from the point that only and exclusively this resultant sum can satisfy Maxwell's equations, but its parts (modes) independently cannot. This approach is the Method of Inhomogeneous Modes (MIBM, [2]).

To show the method, let us consider a linear, time-invariant, bi-anisotropic medium, where for the field-strengths one can write the followings:

$$\begin{aligned} \overline{D} &= \overline{\overline{\varepsilon}}(\overline{r})\overline{E} + \overline{\overline{\kappa}}(\overline{r})\overline{H}, \\ \overline{B} &= \overline{\overline{\nu}}(\overline{r})\overline{E} + \overline{\overline{\mu}}(\overline{r})\overline{H}. \end{aligned} \tag{14}$$

Supposing monochromatic functions, the general form of the signal is

$$\overline{G} = \sum_{i=1}^{n} a_i(\overline{r}) \cdot \overline{G}_{0i}(\overline{r}) \cdot \exp\, \mathrm{j}(\omega t - \varphi_i(\overline{r})) . \tag{15}$$

where $a_i(\overline{r})$ is a general envelope function depending on space, n is the number of the possible modes.

Substituting (16) into Maxwell's equations and applying some mathematical simplifications, the following equations to be solved are yielded (16):

$$\sum_{i=1}^{n} \left[\overline{\nabla}(\ln a_i - \mathrm{j}\varphi_{ai}) \times \overline{H}_i + \overline{\overline{\nabla}}_{TH0i} \overline{H}_i - \mathrm{j}\overline{K}_i \times \overline{H}_i \right] = \sum_{i=1}^{n} \mathrm{j}\omega \left(\overline{\overline{\varepsilon}} \overline{E}_i + \overline{\overline{\kappa}} \overline{H}_i \right)$$

$$\sum_{i=1}^{n} \left[\overline{\nabla}(\ln a_i - \mathrm{j}\varphi_{ai}) \times \overline{E}_i + \overline{\overline{\nabla}}_{TE0i} \overline{E}_i - \mathrm{j}\overline{K}_i \times \overline{E}_i \right] = -\sum_{i=1}^{n} \mathrm{j}\omega \left(\overline{\overline{\nu}} \overline{E}_i + \overline{\overline{\mu}} \overline{H}_i \right)$$

where

$$\overline{\overline{\nabla}}_{TG0i} = \begin{bmatrix} 0 & -\dfrac{\partial \ln G_{20i}}{\partial x_3} & \dfrac{\partial \ln G_{30i}}{\partial x_2} \\ \dfrac{\partial \ln G_{10i}}{\partial x_3} & 0 & -\dfrac{\partial \ln G_{30i}}{\partial x_1} \\ -\dfrac{\partial \ln G_{10i}}{\partial x_2} & \dfrac{\partial \ln G_{20i}}{\partial x_1} & 0 \end{bmatrix};$$

$$\overline{K}_i = \overline{\nabla}\varphi_i;$$

$$; \quad \overline{\nabla}_{\overline{\overline{\alpha}}} = \overline{\nabla} \cdot \overline{\overline{\alpha}};$$

$$\left(\overline{\overline{\nabla}}_{\alpha i} \overline{G}_{0i} \right)_m = \alpha_{mn} \frac{\partial \ln G_{0in}}{\partial x_m};$$

(17)

Investigating (17) a very important feature can be recognized. This equation-system contains the whole solution arising in inhomogeneous medium, without any restriction. The final terms on the left side of the equations and the terms on the right side are completely identical with the ones valid for homogeneous case, while the first two terms on the left side are new; do not appear in homogeneous medium. As it seems to be reasonable to look for the solution in a form leads back to the known for homogeneous case, the further way of thinking is based upon this perception.

Let the inhomogeneous basic modes be defined in such a way that they deliver the solutions of the equation-parts remaining in homogeneous case, separately. But we must keep it in sight the fact, that these basic modes are not solutions of the full Maxwell's equation-system shown in (17), they fulfill just a part of it. But for homogeneous medium they trace back to the known solutions, a sin this case the first two terms disappear. Let the definition of the generalized propagation vector ($\overline{K}_i = \overline{\nabla}\varphi_i$) be the solution of the following dispersion relation, as follows

$$\left| \left(\overline{\overline{K}}_i + \omega\overline{\overline{\kappa}} \right) \overline{\overline{\mu}}^{-1} \left(\overline{\overline{K}}_i - \omega\overline{\overline{\nu}} \right) + \omega^2 \overline{\overline{\varepsilon}} \right| = 0 . \tag{18}$$

So, the inhomogeneous basic modes belonging to $\overline{K}_i$ are the solutions of the equations below:

$$\begin{aligned} \left(\overline{K}_i \times \overline{H}_i \right) &= -\omega \left(\overline{\overline{\varepsilon}} \overline{E}_i + \overline{\overline{\kappa}} \overline{H}_i \right) \\ \left(\overline{K}_i \times \overline{E}_i \right) &= \omega \left(\overline{\overline{\nu}} \overline{E}_i + \overline{\overline{\mu}} \overline{H}_i \right) \end{aligned} \tag{19}$$

Now, as an essentially new step differing significantly from the former methods, let the inhomogeneous modes given on the presented way be substituted into (17), into the full form of Maxwell's equations free from any eliminations. The envelope functions and the phase functions remain unknown variables. The parts remaining in homogeneous case now are cancelled out (as the inhomogeneous modes are solutions of these parts), the remnant equations are called as "coupling equations", as these will deliver the missing unknown parameters, they describe the relation among the modes and the excitation:

$$\begin{aligned} \sum_{i=1}^{n} \left[\overline{\nabla}(\ln a_i - \mathrm{j}\varphi_{ai}) \times \overline{H}_i + \overline{\overline{\nabla}}_{TH0i} \overline{H}_i \right] &= 0, \\ \sum_{i=1}^{n} \left[\overline{\nabla}(\ln a_i - \mathrm{j}\varphi_{ai}) \times \overline{E}_i + \overline{\overline{\nabla}}_{TE0i} \overline{E}_i \right] &= 0, \end{aligned} \tag{20}$$

By solving the coupling equations we can obtain the whole solution, all the simultaneously arising modes and the connection among them. This means in an inhomogeneous medium the resultant sum of the forward propagating and the reflected signal-parts, and their connection to the excitation as well.

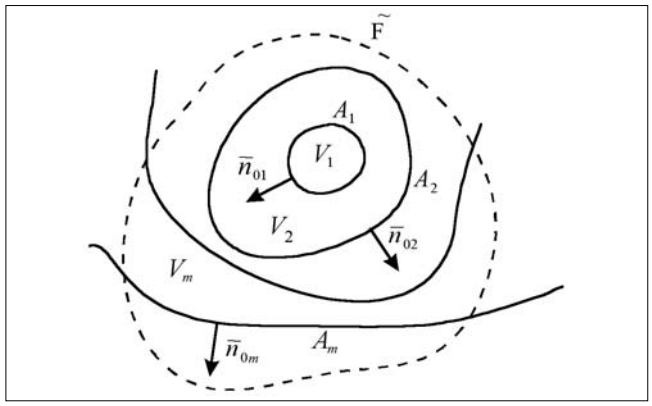
4. Solution of Maxwell's equations in the presence of distributions

Now, let the problem be examined in which the medium-parameters change suddenly at several opened or closed A_m surfaces not crossing each other *(Fig. 1)*. Let the variation of the medium-parameters within the V_m volumes between the surfaces be continuous functions, which connect to each other by steps at the surfaces. This case is the variation of medium-parameters describable by distributions (functionals) [3].

Considering further strictly monochromatic electromagnetic signals, and supposing exp j(ωt – φ) type solutions in volumes V_m, the forms valid for each volume are

$$\overline{G}_m = \left[\sum_i \overline{G}_i \right]_m = \left[\sum_i \left(a_i \cdot \mathrm{e}^{-\mathrm{j}\varphi_{ai}} \right) \overline{G}_{0i} \cdot \mathrm{e}^{\mathrm{j}(\omega t - \varphi_i)} \right]_m . \tag{21}$$

Figure 1. The structure of the medium



Furthermore, introducing the known $1(x)$ Heaviside (unit-step) and $\delta(x)$ Dirac-delta distributions, $\mathbb{1}[\bar{r}(p_m, q_m)]$ notes the distribution the value of which changes from 0 to 1 at the surface $\bar{r} = \bar{r}(p_m, q_m)$. The $\bar{r}(p_m, q_m)$ vector is the parameter of surface A_m. Let gate-functions be created from $\mathbb{1}[\bar{r}(p_m, q_m)]$ unit-steps belonging to surfaces A_m on the following way:

$$s_m(\bar{r}) = \{\mathbb{1}[\bar{r}(p_{m-1}, q_{m-1})] - \mathbb{1}[\bar{r}(p_m, q_m)]\}, \qquad (22)$$

the value of that is 1 between A_{m-1} and A_m and elsewhere 0.

By the application of the rules of derivation on these gate-functions, and keeping in mind that the generalized derivative of $1(x)$ is $\delta(x)$, one can get such a function the value of which differs from 0 only at the surfaces:

$$\bar{\nabla} \cdot s_m(\bar{r}) = \delta[\bar{r} - \bar{r}(p_{m-1}, q_{m-1})]\bar{n}_{0m-1} - \delta[\bar{r} - \bar{r}(p_m, q_m)]\bar{n}_{0m}, \qquad (23)$$

where

$\bar{n}_{0m}$ is the outward directed normal vector of A_m.

The whole solution is yielded again by the application of MIBM.

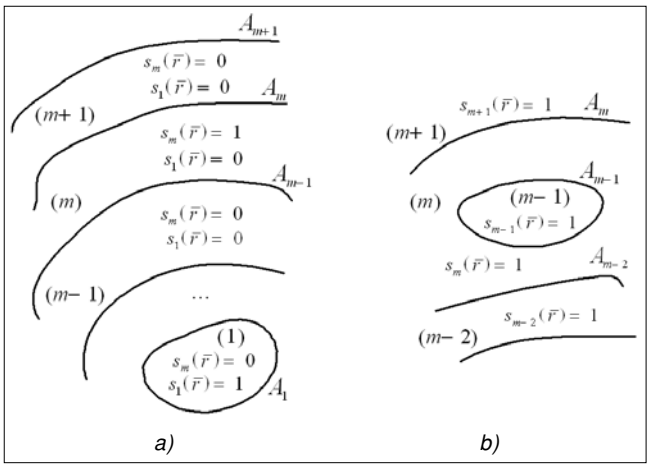


Figure 2. The distribution functions

Defining the gate-functions on a way shown on *Fig. 2.*, one can write in each volume $s_m(\bar{r})=1$ the whole sum of all the possibly existing basic modes, and sum these in the complete examined as it follows:

$$\bar{G} = \sum_{m=1}^{M} s_m(\bar{r}) \left[\sum_{i=1}^{n} \bar{G}_i \right]_m, \qquad (24)$$

where M is the number of the continuous V_m ranges.

The basic modes can be determined within each V_m on the way presented in Part 3., from the equations below:

$$\begin{aligned} \bar{K}_{im} \times \bar{H}_{im} &= -\omega\left(\bar{\bar{\varepsilon}}_m \bar{E}_{im} + \bar{\bar{\kappa}}_m \bar{H}_{im}\right), \\ \bar{K}_{im} \times \bar{E}_{im} &= \omega\left(\bar{\bar{\nu}}_m \bar{E}_{im} + \bar{\bar{\mu}}_m \bar{H}_{im}\right), \end{aligned} \qquad (25)$$

$$\left| \bar{\bar{K}}_{im} + \omega\bar{\bar{\kappa}}_m \right) \bar{\bar{\mu}}_m^{-1} \left(\bar{\bar{K}}_{im} - \omega\bar{\bar{\nu}}_m \right) + \omega^2 \bar{\bar{\varepsilon}}_m \Big| = 0 \qquad (26)$$

For the determination of the complete solution the obtained basic modes have to be substituted back into Maxwell's equations, and by solving the coupling equations the parameters still unknown can be delivered:

$$\begin{aligned} &\sum_{m=1}^{M} \bar{\nabla} \cdot s_m(\bar{r}) \times \left[\sum_{i=1}^{n} \bar{H}_i \right]_m = 0, \\ &\sum_{m=1}^{M} \bar{\nabla} \cdot s_m(\bar{r}) \times \left[\sum_{i=1}^{n} \bar{E}_i \right]_m = 0, \\ &\sum_{m=1}^{M} \bar{\nabla} \cdot s_m(\bar{r}) \left\{ \bar{\bar{\varepsilon}}_m \left[\sum_{i=1}^{n} \bar{E}_i \right]_m + \bar{\bar{\kappa}} \left[\sum_{i=1}^{n} \bar{H}_i \right]_m \right\} = 0, \\ &\sum_{m=1}^{M} \bar{\nabla} \cdot s_m(\bar{r}) \left\{ \bar{\bar{\nu}}_m \left[\sum_{i=1}^{n} \bar{E}_i \right]_m + \bar{\bar{\mu}} \left[\sum_{i=1}^{n} \bar{H}_i \right]_m \right\} = 0 \end{aligned} \qquad (27)$$

5. Results of the new model

Let us apply the presented method for monochromatic and transient (Ultra Wide Band, UWB) signals propagating in arbitrarily strongly inhomogeneous medium [4,5,6].

Let the medium be magnetized, anisotropic plasma (frequently occurring in the space research). Apart from the detailed overview of the deduction, here we show only several final solution formulas, illustrating that the new model modifies the structure of the signal essentially, in a large measure, in comparison with the former solutions.

In monochromatic cases (too), the solution given by MIBM is iterable by successive approximation. The zero-ordered solution of this is the well-known W.K.B. formula.

$$E_1(x) = C\sqrt{Z_0(x)} \qquad (28)$$

where C = constant

$$E_2 = \frac{E_{10}}{2}\sqrt{Z_0(x)} \int_x^{x_M} \frac{d(\ln Z_0)}{du} e^{-j2\int_0^u k(v)dv} du \qquad (29)$$

The following, first order approximation gives more accurate formulas, and the coupling of the energy between the signal-parts can be well seen in the structure of the formulas:

$$\begin{aligned} E_1 = E_{10}\sqrt{Z_0(x)} \Bigg\{ 1 - \frac{1}{4}\int_0^x \frac{d(\ln Z_0)}{du} e^{j2\int_0^u k\,dv} \\ \cdot \left[\int_u^{x_M} \frac{d(\ln Z_0)}{dw} e^{-j2\int_0^u k\,dv} dw \right] du \Bigg\} \end{aligned} \qquad (30)$$

Considering impulse propagation [7]

$$I_{x=0}(\omega) = \int_{-\infty}^{\infty} \left\{ \int_{-x_0}^{0} J_0\left(l, t + \frac{l}{c}\right) dl \right\} e^{-j\omega t} dt \qquad (31)$$

the solution for the reflected signal in the first step of the successive approximation is the following (32):

$$E_{z2}(x,t) = -\frac{j}{8\pi} \int_{-\infty}^{\infty} \left[\frac{C_0(\omega)}{\sqrt{k(x,\omega)}} \int_{\xi}^{x} \frac{1}{2k(u,\omega)} \frac{\partial k(u,\omega)}{\partial u} e^{-2j\int_0^u k(v,\omega)dv} du \right] e^{j\left(\omega t + \int_0^x k(h,\omega)dh\right)} d\omega$$

where

$$C_0(\omega) = I_{x=0}(\omega)\frac{k_0(\omega)\sqrt{k(x=0,\omega)}}{k_0(\omega)+k(x=0,\omega)} \tag{33}$$

$$k(x,\omega) = \frac{1}{c}\sqrt{\frac{\omega\omega_b(x)\omega_p^2(x)+\omega^2\left[\omega_p^2(x)+\omega_b^2(x)-\omega^2\right]}{\omega_b^2(x)-\omega^2}} \tag{34}$$

It can be well seen point by point in the structure of the solution, by the integrals nested into each other, that the propagating and the reflected parts of the energy are in closed connection with each other, varying point by point.

6. Summary

In this paper a fundamental theoretical misunderstanding of the known and commonly used inhomogeneous wave propagation models was presented. This error is originated from the wrong assumption regarding the structure of the signal.

We briefly outlined the Method of Inhomogeneous Basic Modes (MIBM), by the application of that this contradiction and error cannot arise, and really accurate and right wave propagation description can be obtained.

The importance of the presented problem and solving method is great, as the wave propagation results of the last 100 years have to be revised, and opens the way toward such new, exact descriptions, by the application of which the interpretation of our knowledge and ideas regarding our global surrounding environment may go through serious development. This exact determination of the reflection will influence the research in many fields (space research, radar-technique, telecommunication etc.).

Acknowledgements

The presented results were born by the support of Hungarian Space Research Office (Ministry of Informatics and Telecommunication) and the MTA-ELTE Research Group for Geoinformatics and Space Sciences, the Hungarian Academy of Sciences, further the OTKA T037611 and F037603 contracts (already closed).

References

[1] Budden K.G.:
Radio waves in the ionosphere;
Cambridge University Press, London 1966.
[2] Ferencz Cs.:
Elektromágneses hullámterjedés;
Akadémiai Kiadó, Budapest 1996.
[3] Idemen M.:
The Maxwell' equations in the Sense of Distributions;
IEEE Trans. on Ant. and Prop.; AP-21, 1973.
pp.736–738.
[4] Cs. Ferencz:
Real solution of monochromatic wave propagation
in inhomogeneous media;
Pramana Journal of Physics, Vol.62, No.4, 2004.
pp.943–955.
[5] Ferencz O.E.:
Full-wave solution of short impulses
in inhomogeneous plasma;
Pramana Journal of Physics, Vol.64, No.2, 2005.
pp.1–20.
[6] Erhardtné Ferencz O. és Ferencz Cs.:
Elektromágneses impulzusok terjedésének vizsgálata
különböző közegekben;
Híradástechnika, Vol.LIX, 2004/5, pp.18–24.
[7] Ferencz Cs., Ferencz O.E., Hamar D., Lichtenberger J.:
Whistler Phenomena, Short Impulse Propagation;
Kluwer Academic Publishers,
Astrophysics and Space Science Library,
Dordrecht, 2001.

Physical limits of the applicability of 10 and 40 Gbps speed DWDM systems

PÉTER JESZENŐI, JENŐ SZATMÁRI

Magyar Telekom PKI Telecommunications Development Institute
{jeszenoi.peter, szatmari.jeno}@t-com.hu

Keywords: DWDM, optical network, optical amplifier, optical fibre, dispersion, SPM, XPM, FWM, SBS, SRS, Q-factor

Due to the growing transmission demands and the technical evolution, the use of DWDM systems that have more and more channels for the transmission of bundles of higher and higher speeds is spreading. In case of the application of 10 Gbps, but especially 40 Gbps systems, the dispersion characteristics of the optical fibres come to the focus of attention. Due to the high optical levels that can be provided with optical amplifiers, non-linear phenomena in the optical fibres can be observed. The imperfection of the passive optical devices used for the multiplexing/de-multiplexing of the wavelengths causes channel cross-talks. The aforementioned phenomena are in close relation with the high-speed transmission and they have to be taken into account when designing, installing and operating such systems. In general, the problems of the physical layer appear much more in case of high speed multiplex wavelength transmissions than as it used to be in case of known lower speed systems.

1. Introduction

The current DWDM connections enable transmitting of 40-160 channels each of 10 Gbps, but the development of the 40 Gbps devices, too, has entered the phase, in which the first multi wavelength systems, transmitting traffic of industrial scale on 40 Gbps bundles, have appeared.

In the time of the early single channel, single mode optical systems, the optical fibre seemed to be almost an ideal transmitting medium. The distance that could be bridged over was only limited by the optical attenuation, while till the appearance of the 2.5 Gbps systems, in systems using single mode fibre the impact of the chromatic dispersion was also negligible. The applied laser sources could be quite simply directly modulated, preserving normal signal shape and good extinction ratio. At the applied optical outputs of maximum few milliwatt, the optical fibre behaved fully linearly, noise or cross-talk had not to be expected on the line. With the high speed (>10 Gbps) multi wavelength systems, we have got far from the aforesaid almost ideal status.

2. Physical limits of the optical transmission

In the optical fibre transmission systems – with the exception of the analogue cable television applications – we transmit digital signals. Often these signals are multiplex signals bundled with a certain time division multiplexing (TDM) method. In spite of the aforesaid, we can state that on optical level the transmission is performed in a fully analogue way. For the optical transmission of the digital information "pre-multiplexed" by TDM there is an intensity modulation applied on the optical carrier. We do not call this modulation method as amplitude modulation, because the optical carrier is not a single-frequency carrying wave, but usually a spectrum of several MHz bandwidth. The intensity modulation can be simply performed with the help of switching on/off the driving current of the semiconductor laser used as light source, or using an external modulator. On the reception side, a very simple direct detection is taking place and there is no need to create the carrier; the original digital signal is included in the change of the current of the receiver's photo detector. No special coding is required for the transmitting of the signal; scrambled NRZ or RZ coded signals can be transmitted.

Notwithstanding, the NRZ coding is not the best coding method from the point of view of the transmission, at least for two reasons: first, the radiated carrier does not carry information and it unnecessarily loads the optical amplifiers, secondly, it is quite sensitive to the Polarisation Modus Dispersion (PMD). The RZ coding is more favourable from the point of view of PMD, through the carrier is radiated in this case, too. There exist other modulation methods favourable from many aspects and some of them are close to being introduced in the practical use. Due to space limits this article does not allow us to describe these methods.

The characteristics of the transmission medium, the applied light sources, the passive and active devices installed on the transmission path and the features of the optical receiver fundamentally influence the high speed transmission. The maximum length of the optical section is limited first of all by the attenuation of the optical fibre and the passive elements inserted within the transmission path. Besides the attenuation, the signal distortion caused by the dispersions, the noise produced by the optical amplifiers, the interferences caused by the crosstalks, the signal shape distortions,

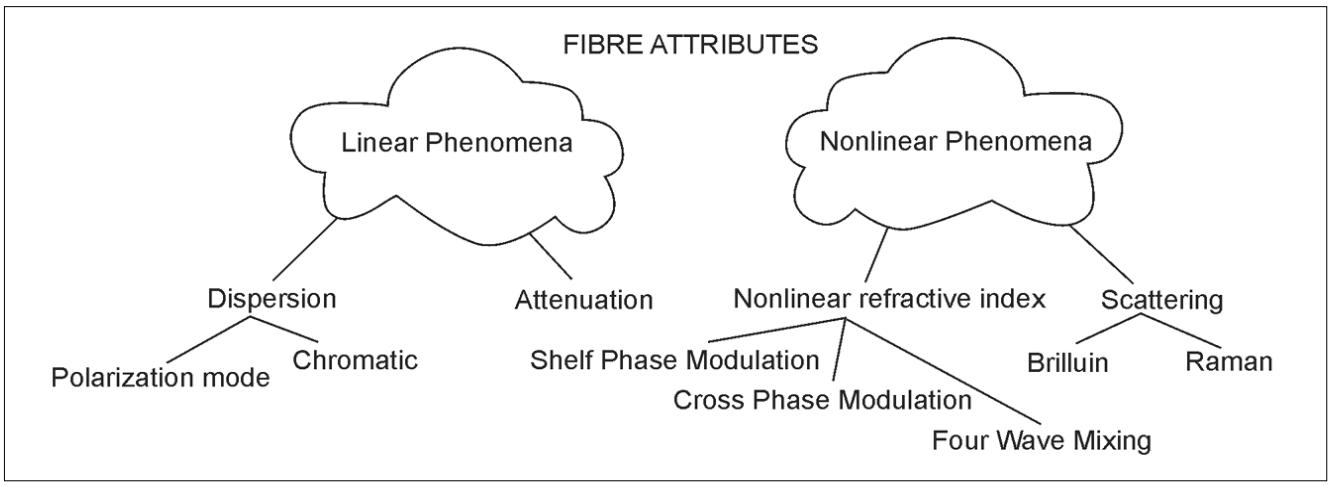


Figure 1. Characteristics of an optical fibre

noises and the jitter developing due to the non-linear characteristics occurring in the fibre altogether further limit the section that can be bridged over at an acceptable bit error rate. The application of a redundant error correction coding (Forward Error Correction, FEC) provides protection to a certain extent against the errors occurring on the line section. Activating FEC, a noise gain of 4...6 dB can be achieved.

In the followings we give an overview of factors that influence the high speed transmission on the physical level. The phenomena that on any way influence the propagation of the light pulse in the optical fibre are grouped into two categories and illustrated in *Figure1*.

2.1. The transmission medium

A basic element of DWDM systems is the optical fibre itself. ITU-T has standardized, in its recommendations, several types of single mode optical fibres that basically differ from each other in respect of their dispersion characteristics.

In the last one and the half decades most of the networks were built from single mode optical fibre cables described in ITU-T Recommendation G.652. This type of fibre is often called as “standard” single mode fibre (SSMF). The SSMF is optimised for 1310 nm wavelength, which means that the fibre has 0,3...0,5 dB/km attenuation and its chromatic dispersion value is low enough in this spectrum.

Shifted dispersion fibres (G.653) have appeared mostly for the application in long distance connections. The dispersion characteristics of these fibres are optimised for the lower attenuation 1550 nm wavelength. Thus, with the higher performance of the fibre the section distances could be further extended. At the same time, this type of fibres has explicit disadvantages from the point of view of high speed DWDM. Because of the smaller diameter of the mode field, the non-linear phenomena occur in an increased degree. In this context, it is an additional disadvantage that in the transmission range, the chromatic dispersion becomes zero and the dispersion coefficient reverses its sign.

Later, further fibre types have appeared having better parameters that match better to the needs of broadband and high speed DWDM transmission. Their common characteristic feature is that their dispersion parameters are optimised for the 1550 nm environmental conditions and that thanks to their relatively bigger and efficient diameter they tolerate higher output levels without increasing of the disadvantageous non-linear phenomena. The characteristics of these types of optical fibres are described in Recommendation G.655. The manufacturers differentiate the fibre types by different fancy names.

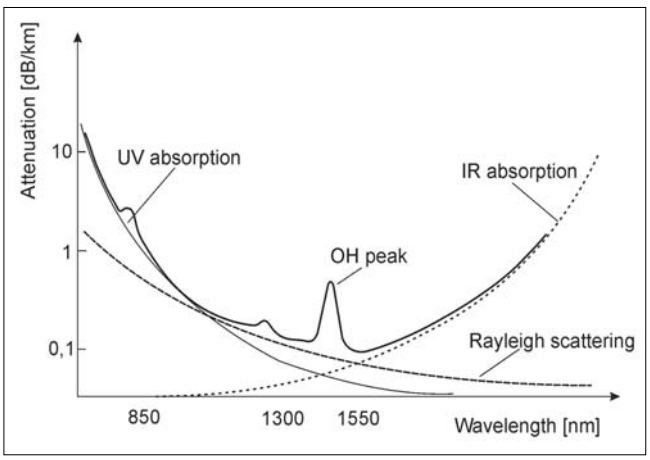
2.1.1. Linear characteristics

The most important transmission attributes of the optical fibres are the wavelength dependent attenuation, the chromatic and the polarization mode dispersion.

The attenuation of the silicon-based single-mode fibres arises from three factors: the absorption, the scattering and the wave guiding losses *(Figure 2)*.

- Absorption may be of intrinsic nature, which is caused by the electron transients falling into the UV range and the photons of the IR range; or may be generated by contaminations, caused by temporary metallic components, or the vibrations of H_2 and OH ions; and finally caused by problems due to inhomogeneity of the material.

Figure 2.
Components of the attenuation of an optical fibre



• A considerable part of dispersion losses is caused by Rayleigh-scattering, which is an inherent material characteristic of non-crystalline substances. Light dispersion may occur also on macroscopic defects of the material, such as blisters, cracks and other inhomogeneities, or the interfacial unevennesses of the core-shelf plane.

• Waveguiding losses may arise from macrobanding (losses originating from the curve of the waveguide) or from microbanding (losses caused by perturbation).

The size of the attenuation influences basically the signal transmission, but with the application of optical amplifiers the attenuation problem can easily be eliminated.

The light pulse components having various wavelengths propagate in the optical fibre at different velocities due to the wavelength dependency of the refractive index of the silicon oxide. This phenomenon is called chromatic dispersion (CD). The CD develops as the common result of several effects. From these factors, the waveguide dispersion can be influenced by shaping the profile of the optical fibre's refractive index *(Figure 3)*. This offers the possibility of producing optical fibres with different dispersion characteristics.

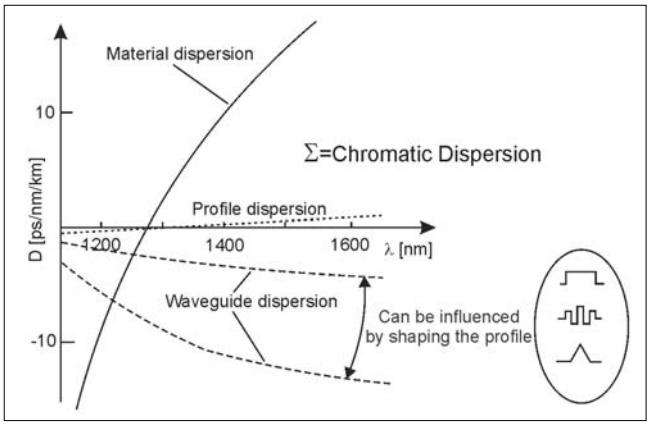


Figure 3. Chromatic dispersion

Due to chromatic dispersion the individual components of the light pulse coupled into the fibre arrive at different points of time at the place of reception and cause the broadening of the original pulse *(Figure 4)*.

If the broadening is so big that it leads to the overlapping of the subsequent pulses, it results in bit errors

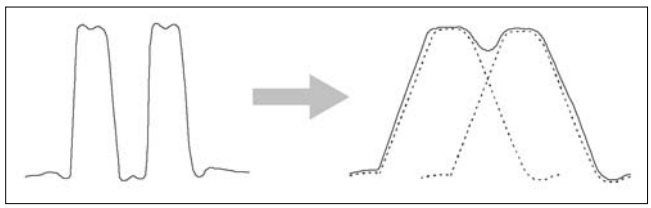


Figure 4.
The chromatic dispersion effect: pulse broadening and overlapping

in the transmission. The higher the transmission speed is, the more the chromatic dispersion affects the transmission quality, because the overlapping of the adjacent pulses takes place earlier due to the shortened bit time, and also the spectrum of the laser transmitter broadens much more due to the effect of the higher modulation frequency. As a result of these jointly occurring phenomena the dispersion-sensitivity increases nearly quadratically with the bit rate. A 40 Gbps system is more sensitive to dispersion by approximately 16 times than a 10 Gbps system and 256 times more sensitive than a 2.5 Gbps system.

The extent of the pulse broadening depends on the spectral characteristics of the transmitter. Application of narrow light sources, with a few MHz spectrum may be advantageous, but from other points of view (for instance Brillouin scattering) the application of them is definitely disadvantageous. The extent of pulse broadening can be calculated with the following formula:

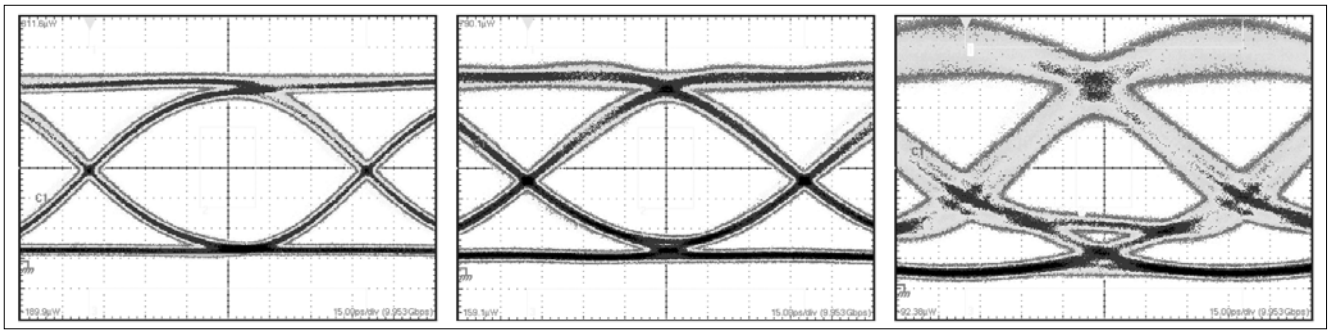
$$t_H = \delta\lambda * L * D ,$$

where $\delta\lambda$ is the spectrum width of the light source, L is the length of the link and D is the chromatic dispersion coefficient of the optical fibre.

Figure 5. shows a (10 Gbps) STM-64 signal shape spreading on a G.652 optical fibre, after concatenation of 5, 50 and 100 km length optical fibres. The receiver is a standard SDH reference receiver. The pulse broadening can be seen well and also that the noise has significantly increased due to the optical amplifier used at the measurement of the 100 km fibre length.

In spite of its isotropic material and circular cross section the optical fibre has slight birefringence. The birefringence introduced to the fibre is caused by the non-circularity, surface unevennesses developed during manufacturing, longitudinal and crosswise power

Figure 5.
STM-64 signal shape degradation due to chromatic dispersion on an inserted 5, 50 and 100 km length SSMF at 1550 nm



impacts during installation, longitudinal twisting or bending. The polarization mode dispersion attributable to the group delay difference of the two polarization components of the HE_{11} dominant modus of the light. The difference in the group delays belonging to the different polarization planes is called Differential Group Delay, DGD.

The polarization mode dispersion is the rms value of the differential group delay. Additional higher order PMD effects come to this first order DGD, such as: polarization-dependent chromatic dispersion, skewing of the dominant polarization planes, etc. Similarly to chromatic dispersion, the adverse impact of PMD to transmission appears as broadening of the transmitted pulse and pulse overlapping arising due to broadening.

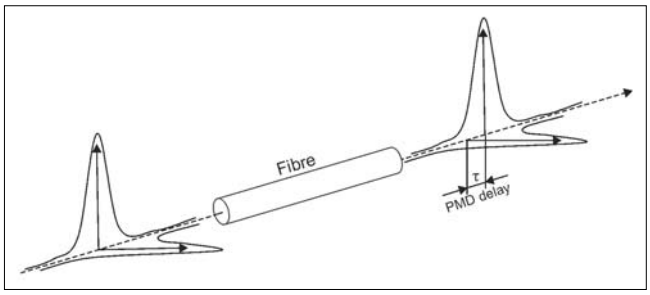


Figure 6. The PMD effect

The value of the PMD is proportionate with the square root of the cable length. The value of polarization mode dispersion permitted to a system is usually quoted in 1/10 of the periodic time typical for the given transmission system. For example, for a 10 Gbps system we allow maximum 10 ps. If the PMD coefficient of our cable is 0,5 ps/√km, then the maximum permitted section length (limited by the PMD effect) is $L = (10/0,5)^2 = 400$ km. The effect of PMD can be compensated by adequate techniques.

2.1.2. Non-linearities

The output optical power of the "traditional" optical systems exceeded the +3...5 dBm value only rarely. The application of optical amplifiers made it possible to achieve higher, even +20 dBm (100 mW) output levels. Thus, with the amplifiers applied periodically along the transmission link there can be high signal level sustained, and the system's sensitivity to the noise developing in the receiver decreases significantly. At the same time, due to the higher power level, the increased number of channels in the WDM systems we leave the range where the optical fibre shows linear behaviour with a good approximation. The non-linear characteristics of the fibre originate from the fact that the light – due to the enormous spectral density of order of magnitude of 100 MW/m^2 occurring in the core – enters into interaction with the glass fibre. Non-linearities can be grouped basically into two different categories:

Effects that arise because of the changing of the refractive index caused by high field strength belong to the first category. These are the followings:

– Self Phase Modulation, SPM,
– Cross Phase Modulation, XPM and
– Four Wave Mixing, FWM.

Scattering type phenomena belong to the second category. These are the followings:

– Stimulated Brillouin Scattering, SBS and
– Stimulated Raman Scattering, SRS.

Effects occurring due to the change of the refractive index

The refractive index of the optical fibre, even if to a small extent, depends on the intensity of the light. At the peaks of the pulses of the modulated light signal the refractive index changes (Kerr effect). The extent of the change is:

$$n = n_0 + n_2 |E|^2 ,$$

where n is the changed refractive index, n_0 is the original value of the refractive index, n_2 is the refractive index coefficient that depends on the non-linear field strength, E is the field strength. The approximate value of n_2 is $-2{,}2 \times 10^{-20}$ m^2/W, which practically does not depend on the type of the fibre. The increase of the refractive index can be expressed with a more practical formula as:

$$n = n_0 + \frac{n_2}{A_{eff}} P ,$$

where P is the power launched into the fibre, A_{eff} is the effective area of the optical fibre.

The change in the refractive index results in phase modulation and the latter changes the signal spectrum. The Self Phase Modulation (SPM), in case of negative chromatic dispersion, introduces the broadening of the light pulse, while in case of positive dispersion, it reduces the pulse. The spectral broadening caused by Self Phase Modulation may cause in multi-channel systems interference among the adjacent channels. Zero, or near to zero positive chromatic dispersion environment reduces the effect of this phenomenon. High bit rate, negative dispersion and several concatenated sections further increase the effect of SPM. In case of 10 or 40 Gbps systems, the effect of this phenomenon can be detected already at power levels above 10 mW. With the proper setting of the dispersion values of the fibre sections the degrading effect of SPM is more or less manageable in case of homogeneous optical links not longer than 1000 km.

Cross Phase Modulation (XPM) is caused by the signals of other systems working on other wavelengths of the WDM system that also cause changes in the refractive index and induce undesired phase couplings among the carrier waves. Cross Phase Modulation and Self Phase Modulation are always together at present. The effect of XPM is, of course, more intensive in case of DWDM systems with short channel distances. Higher optical powers lead to the broadening of the spectrum of the transmitter and cause timing jitter in the received signal. The chromatic dispersion appearing due to spectral broadening further worsens the situation on long-haul systems. Therefore, efforts shall be made for the optimal setting of chromatic dispersion of the links.

The recommended compensation settings can be calculated with the following empirical formula:

$$D_{PRE} = \frac{-D_{SMF}}{\alpha} \ln\left[\frac{2}{1+e^{-\alpha L}}\right],$$

where D_{PRE} is the recommended compensation value, α is the per km attenuation of the fibre, D_{SMF} is the dispersion and L is the length of the link. In practical systems the compensation value is set to -200 ps/nm, which means the over-compensation of the link. In general it can be stated that the XPM effect is not significant in case of 100 GHz or larger channel distances and at less than 5 mW launched power.

Four-Wave Mixing (FWM) is the most dangerous nonlinear phenomenon in WDM systems. Exceeding the critical power, due to the undesired phase couplings mixing products occur, the wavelengths of which fall on operating wavelengths in case of equal channel spacing.

In case of wavelengths ω_1 and ω_2 the created mixing products are: $2\omega_1-\omega_2$ and $2\omega_2-\omega_1$. In a system consisting of N channels the number (n_λ) of the developed "ghost" wavelengths is expressed as:

$$n_\lambda = N^2\left(\frac{N-1}{2}\right),$$

where N is the number of wavelengths applied in the given system. So thus, for instance, in a 32-channel DWDM system, there appear more than 15 thousand (!) mixing products. Four-Wave Mixing develops already on 10 km fibre lengths on wavelengths or in the vicinity of wavelengths, the chromatic dispersion value of which is zero. Thus, FWM is especially critical in case of low effective area, dispersion shifted G.653 fibres.

In this case the non-desired effect can only be reduced with carefully chosen, non-equal channel spacing. Mixing products – taking into account also the products generated by the noise of optical amplifiers – appear as noise in the given channel and cause the closure of the eye pattern and finally degrade the system's BER performance.

The effects of scattering type phenomena

Stimulated Brillouin Scattering (SBS) is attributable to the macroscopic interaction between the light and the density waves of the fibre's material (acoustic photons). Due to Brillouin scattering the power launched into the fibre at 1550 nm is partly reflected on a frequency shifted by appr. 11 GHz. Thus, the phenomenon is especially harmful if extremely low channel spacing is applied. The extent of backward scattering is independent of the number of channels applied in the system, but it very much limits the power that can be launched into the fibre, especially in case of transmitters with low spectral width. The power level that causes at least 1 dB degradation in the optical signal-to-noise ratio can be calculated using the following relation:

$$P_{th} = 21\frac{KA_{eff}}{gL_{eff}} \cdot \frac{\Delta\nu_p + \Delta\nu_B}{\Delta\nu_B},$$

where P_{th} denotes the threshold power, g denotes the Brillouin gain coefficient (constant) (~4×10^{-9} cm/W), A_{eff} is the fibre effective area, K is a constant determined by the degree of freedom of the fibre's polarization statuses (in case of G.652 fibres, K=2), $\Delta\nu_B$ and $\Delta\nu_p$ represent the Brillouin bandwidth and the spectral width of the pumping light. L_{eff} denotes the effective fibre length, which can be defined with the formula:

$$L_{eff} = \frac{1-e^{(-\alpha L)}}{\alpha},$$

where α is the fibre attenuation per length unit and L is the fibre length.

In case of sources having spectral width of $\frac{\Delta\nu_p}{\Delta\nu_B} << 1$, which is smaller than the Brillouin bandwidth, the critical (threshold) power can be calculated using the following relation:

$$P_{th} = 21\frac{KA_{eff}}{gL_{eff}}.$$

In the practice SBS phenomenon occurs already at power levels of approximately 80 mW (+19 dBm). Its effect can be reduced with some per cents low frequency (30...100 kHz) amplitude modulation applied on the carrier wave.

Stimulated Raman Scattering (SRS) occurs as the interaction of light and the SiO_2 molecules of the fibre which involves the high frequency vibration of the adjacent nucleuses (optical photon). The induced radiation has the same direction as the normal light propagation and its wavelength is shifted typically by 100 nm towards the lower wavelengths. The induced radiation has a spectral width of 50...60 nm.

In *Table 1 (on the next page)*, we provide a summary of the optical characteristics discussed before, the physical phenomena and their effects on digital transmission and the methods for the elimination or compensation of these effects.

3. Q-factor measurement methods

In intensity modulated digital optical transmission systems the information is represented by two possible signal levels. In real systems, different mean noise values are added to the two signal levels. It means that different electrical signal/noise ratios can be attributed to the two signal levels. When we would like to determine the occurrence probability of the bit errors of the transmission, we have to reckon with two various signal-to-noise ratios. The two signal-to-noise ratio values can be merged in one transmission quality parameter, and it is the Q-factor.

The Q-factor is interpretable as a signal-to-noise ratio at the input point of the decision circuit of the optical receiver. The Q-factor and the optical signal/noise ratio can be arranged together in the case only, if we take into account only the ASE noise generation of the optical amplifiers. In the reality – as we could see it before – there are several other effects influencing the quality

of the optical signal, thus the Q-factor and the optical signal/noise ratio can be calculated into each other with certain inaccuracy only.

To determine the relationship between BER and the eye opening it is necessary to determine statistically the amplitude noise. If there is no Inter Symbol Interference (ISI) present, the noise is statistically independent from the signal content and if the dominant amplitude noise is of Gaussian distribution, the Q-factor can be expressed by the following equation:

$$Q = \frac{(\mu_1 - \mu_0)}{(\sigma_1 + \sigma_0)},$$

where μ_1 and μ_0 represent the low and high average levels of the amplitude function, σ_1 and σ_0 represent the Gaussian distribution values of white noise *(see Fig. 7)*.

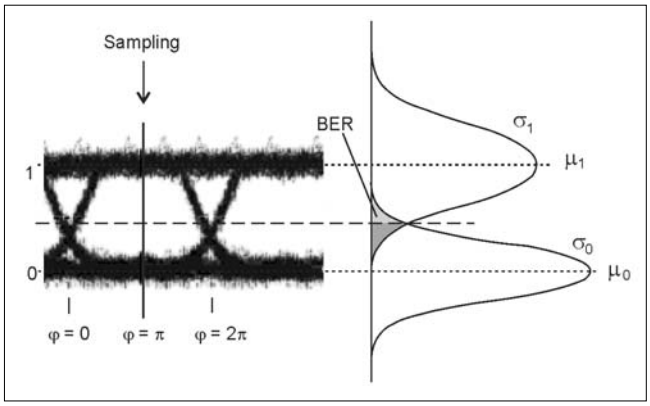


Figure 7.
Noise distribution, mean value and distribution of logical 1 and 0 levels

Analysing the probability curves, we can see that there are two possibilities for the occurrence of a wrong decision: to detect "0" instead of "1" and vice versa, to detect "1" instead of "0". The bit error rate is proportional to the sub-curve area belonging to the opposite logical level stretching over the decision threshold (grey area on Figure 7).

The decision threshold is in its optimal place (i.e. the probability of wrong decisions is the lowest), if the amount of the sub-curve areas belonging to the other logical level on the right and the left side is the minimum. This value is at the crossing point of the two bell-shaped curves, if they are fully identical. In real systems, however, the bell-shaped curves belonging to the two logical levels are always different from each other.

The optimal decision level is at:

$$\mu = \frac{\sigma_0\mu_1 + \sigma_1\mu_0}{\sigma_1 + \sigma_0}.$$

It can be seen from the eye pattern that the occurrence probability of the logical levels depends on the place of detection, too. Taking the width of the eye pattern as 2π, the optimal sampling phase is at the place where $\varphi = \pi$.

Relationship between BER value and Q-factor:

$$BER = \frac{1}{4}erfc\left(\frac{\mu - \mu_0}{\sqrt{2}\sigma_0}\right) + \frac{1}{4}erfc\left(\frac{\mu_1 - \mu}{\sqrt{2}\sigma_1}\right),$$

where *erfc* is the supplementary error function, integrated from x up to ∞, and μ is the decision threshold.

Table 1. Summary of non-linearities

Interference effect	Cause	Critical per channel power	Effect	Compensation
Attenuation/noise	Material absorption/circuit elements	Independent from power	Decreased power, BER	Shorter section, optical fibre with lower attenuation
CD	Wavelength-dependent group propagation velocity	Independent from power	Decreased power, BER, increased spectrum width	Insertion of dispersion with opposed sign
PMD	Random change of refractive index	Independent from power	Decreased power, BER, signal shape degeneration	Compensation of optical or electrical PMD
FWM	Signal interference	10 mW	Side-bands, BER	Accurate CD setting, irregular channel distribution
SPM/XPM	Intensity-dependent refractive index	10 mW	Spectral broadening, BER, channel crosstalk	Accurate CD setting
SRS	Interaction of photons and fibre molecules	1 mW	Power depletion, OSNR, crosstalk, BER	Conceptual power level planning
SBS	Interaction of photons and the fibre's density waves	5 mW	Power, OSNR depletion, signal instability, crosstalk, BER	Source with larger spectrum width

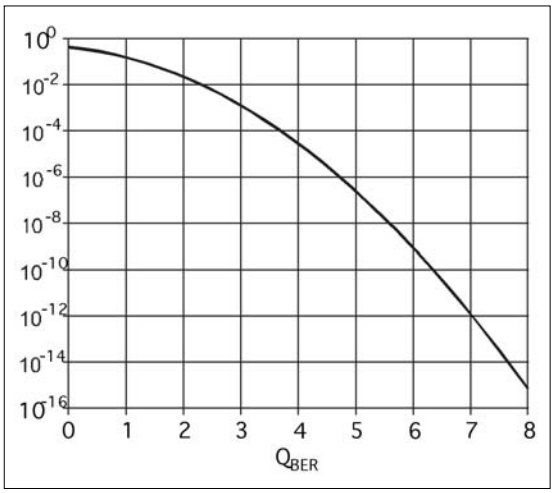


Figure 8.
Relationship between Q-factor and BER

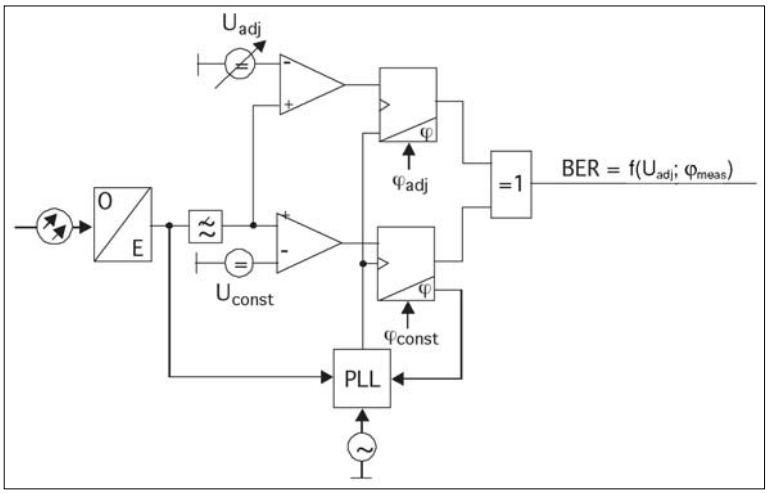


Figure 9.
Synchronous double threshold method for Q-factor measurement

ITU recommends the following formula for the approximation of the *erfc* function:

$$BER(Q) = \frac{1}{2} \times \left[2 \times e^{-\left(\frac{Q}{\sqrt{2}}\right)^2} \times \right.$$

$$\left. \times \frac{a0 + a1 \times \frac{Q}{\sqrt{2}} + a2 \times \left(\frac{Q}{\sqrt{2}}\right)^2}{\left[b0 + b1 \times \frac{Q}{\sqrt{2}} + b2 \times \left(\frac{Q}{\sqrt{2}}\right)^2 + b3 \times \left(\frac{Q}{\sqrt{2}}\right)^3 \right] \times \sqrt{\partial}} \right],$$

where

a0 = 1.69071595	b0 = 1.90764542
a1 = 1.45117156	b1 = 3.79485940
a2 = 0.50003230	b2 = 2.90845448
	b3 = 1.00000000

At values Q>1,5 the approximation provides a good accuracy and is simply programmable.

There are several known methods available for the determination of the Q-factor. The so called “synchronous double threshold” may be one of the most advantageous methods. In an arrangement illustrated in *Figure 9*, the signal is split into two parts and sent to two different decision circuits.

The decision level of one of the decision circuits is set fixed to the optimum value, while the threshold value of the other decision circuit is adjustable. The clock signal which corresponds to the ever current bit rate and with the help of which the optimal decision position of the decision circuits can be set, is provided by a PLL circuit.

Comparing the decision results of the two comparator branches (EXOR, EXCLUSIVE-OR) conclusions can be drawn for the BER value. The result is illustrated in a diagram as shown in *Figure 10*. The vertical axis shows the bit error ratio, while the horizontal axis shows the decision levels. The method allows measurement with acceptable accuracy of bit error ratios in the range between 10^{-4} and 10^{-8}. Applying a regression function, a good extrapolation to the smaller BER values can be done from the diagram.

By this method, in an indirect way, i.e. by calculating the Q-factor, it becomes possible to determine approximately the BER value of the transmission system, independently from the transmitted protocol and the bit contents. In case of extent Inter Symbol Interference (ISI) and non-Gaussian noise distribution, we have to reckon with measurement inaccuracy and be more care-

Figure 10.
Measurement results and extrapolation

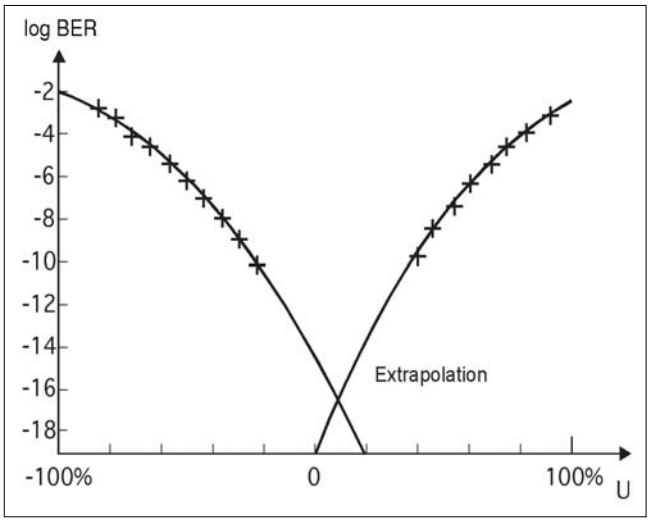
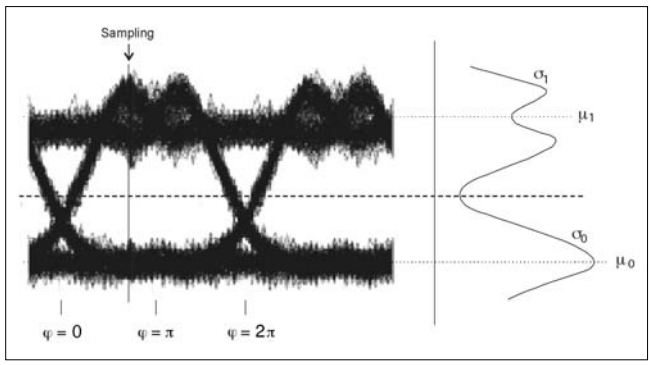


Figure 11. Non-Gaussian noise distribution



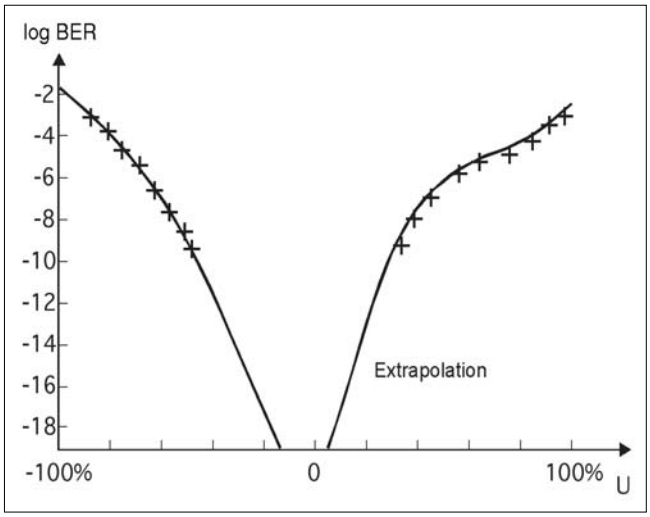


Figure 12. Degenerated Q curves in case of non-Gaussian noise distribution

ful at the applying of regression lines. It may be useful to control the eventual inter symbol interferences and the noise distribution on the eye pattern displayed on a digital oscilloscope.

For instance, in case of non-Gauss noise *(see Fig. 11 and 12)*, the adjustment of the regression line should only be done for the BER range below 10^{-8}.

3.1. Application possibilities for Q-factor measurement

The Q-factor method, of course, is not applicable for detecting errors occurring in the receiver. At the same time, it is an excellent tool for the indication of the various degradations of the optical transmitter, while with the help of it the proper adjustment of the chromatic dispersion compensation, the proper setting of which is very important in case of high speed systems, can be very well controlled, along with the eventual noise increase in the optical amplifiers, or non-linearity effects occurring at higher number of channels, or at higher optical levels.

The Q-factor measurement and assessment on the basis of it requires further considerations in case of non NRZ or RZ coding or in case of non intensity modulation (see [6]).

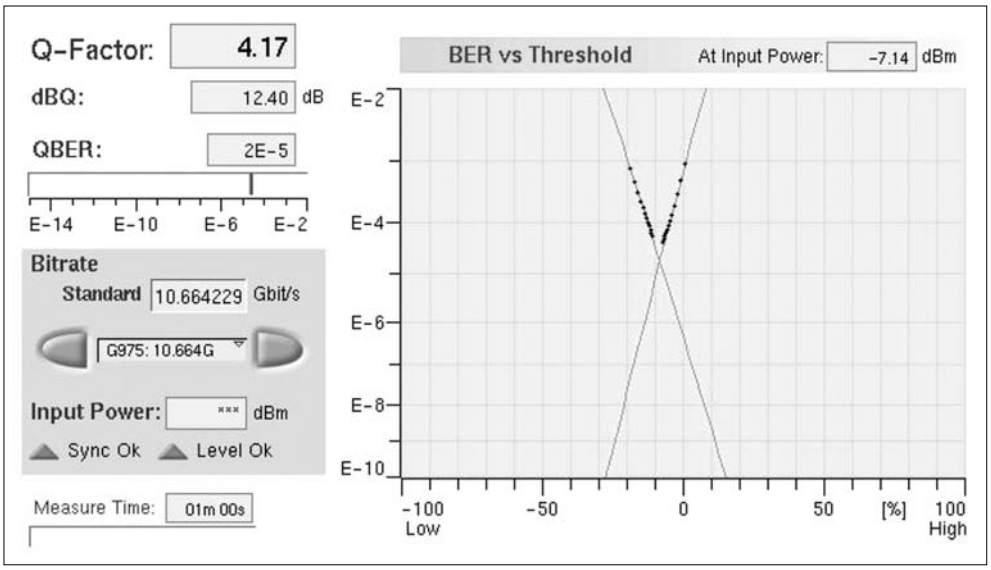
Following the installation of transmission systems one of the most important thing to do is to perform the appropriate control measurements regarding system performance. The BER (bit error ratio) test is one of the most important control tests. According to the specification requirements we expect the systems to be within the BER range of 10^{-12}...10^{-13}. It is quite time consuming to perform the BER test measurements. For instance, in case of a 10 Gbps system to measure statistically correctly a 10^{-13} bit error ratio takes at least 28 hours. One can imagine how long testing times would be required in case of putting into operation a DWDM system with several parallel channels. In such cases, the Q-factor measurement, performed in a few minutes, provides quite accurate approximation of the bit error rate characteristics of the optical signal. Thanks to the very quick measurement that can be performed with the Q-factor method, the malfunctioning network parts or components can be easily identified and separated in difficult cases.

The Q-factor measurements do not fully substitute the BER measurements performed with bit error ratio measuring instruments for the estimation of the system's performance capabilities. Nevertheless, it can help that long lasting tests take place only, when according to the Q-factor tests the system seems to be faultless or error free. Thus, one can save plenty of time and trouble. Last, but not least the number of 10...40 Gbps bit error rate measuring tools which are very expensive, can be reduced with the purchasing of some less costly instruments applicable for Q-factor measurement.

Figure 13 and Figure 14 show the measurement results of a Hungarian DWDM line section of 420 km length. As it can be drawn from Figure 13, the BER value is of $2x10^{-5}$. Measuring the Q-factor at several points of the line section, it could be detected that the laser transmitter installed at the beginning of the section was not working properly. Having it replaced, the Q-factor and BER values have improved significantly (see Figure 14).

Another exciting application field of Q-factor measurement is the optimized setting of the levels and dispersion compensation of the optical systems.

Figure 13. Q-factor measurement results of a defective STM-64 line section



The Q factor measurement offers the possibility of fast controlling after the changing of the compensation and levelling parameters, enabling the settings considered to be optimal from the point of view of the optical signals.

The great advantage of the Q-factor measurement technique is that it allows monitoring of the living system under operation, connecting the measuring device to the measuring points of it. This method can improve the efficiency of the fault detection and maintenance activities. Under operation testing and monitoring may be very useful also for the follow-up of particular maintenance activities or in SLA complaint investigations.

4. Summary

The physical layer of the network basically influences the quality of the high speed multi-wave optical systems. Practically there is analogue signal transmission taking place in the physical layer. In the design, planning and dimensioning of high speed DWDM networks, there are several parameters to be taken into consideration that were less important in the past. Typically such parameters are the non-linear characteristics of the optical fibres, the noise of the optical amplifiers, the exact synthesis of the system and the dispersion compensation. From the point of view of the performance of the systems, the perfect and coordinated functioning of the physical layer is essentially important. The 10...40 Gbps systems set new challenges from planning/designing and operations point of view.

The Q-factor measuring method supports well the operations and maintenance of the physical layer. Sending a feedback on experience gained by the tests for the planning and incorporation of them into the planning process may beneficially promote the accuracy of the planning work.

Figure 14.
Q-factor measurement results after fault correction on line section illustrated on Figure 13

Last but not least, the planning or the optimization of the systems has economic/financial aspects, too: the quality of the transmission is manageable, the network will not be unnecessarily oversized, or, for example the number of the optical amplifiers, and along with that the investment costs can be reduced, or investment savings can be achieved.

References

[1] Ines Brunn:
Dense Division Multiplexing, Pocket Guide;
Acterna Eningen GmBH.

[2] Vitus Zeller:
Q factors basics, Pocket Guide;
Acterna Eningen GmBH.

[3] Jan-Pierre Laude:
DWDM fundamentals, components and applications;
Artech House Inc. 2002.

[4] Hanik, N.:
Netze mit optischem Frequenzmultiplex;
Der Fernmelder Ingenieur, 1997/06.

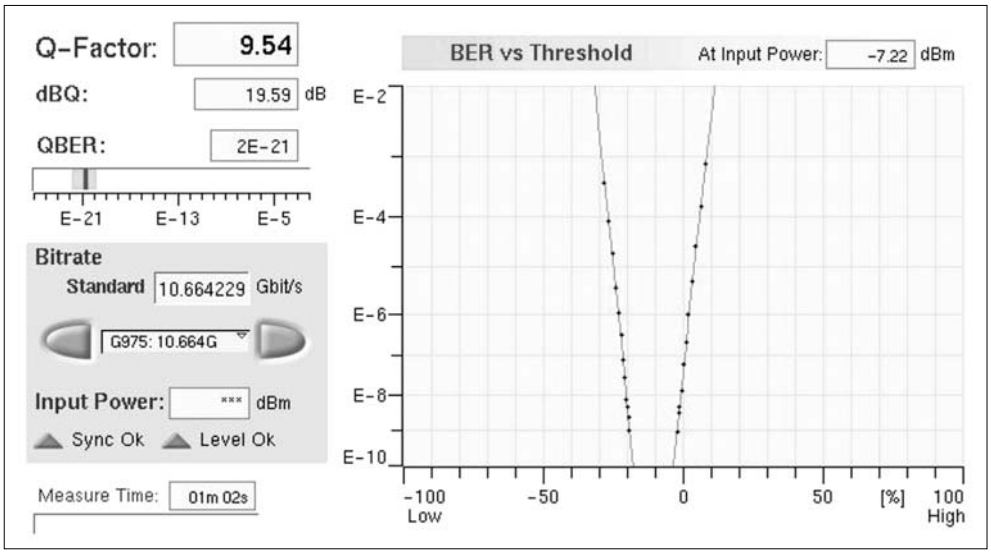
[5] Jeszenői P.:
DWDM rendszerek alkalmazhatósága meglévő optikai hálózaton; Előadások gyűjteménye – 13. Távk. és informatikai hálózatok kiáll. és szemin.

[6] G. Bosco, P. Poggiolini:
On the Accuracy of the Q-parameter to Assess BER in the Numerical Simulation of Optical DPSK Systems; ECOC 2003 Proceedings.

[7] Maxim I.C.:
Optical Receiver Performance Evaluation;
Application Note HFAN-03.0.2 (Rev. 0, 03/03).

[8] Marcuse, D., Chraplyvy, A.R., Tkach, R.W.:
"Dependence of cross-phase modulation on channel number in fiber WDM systems,"
IEEE Journal of Lightwave Technology,
Volume 12, Number 5, p.885, May 1994.

[9] ITU-T Rec. O.201:
Q-factor test equipment to estimate the transmission performance of optical channels.



Applying statistical multiplexing and traffic grooming in optical networks jointly

ANDRÁS KERN, GYÖRGY SOMOGYI, TIBOR CINKLER,

Budapest University of Technology and Economics, Dept. of Telecommunications and Media Informatics
{kern, somogyi, cinkler}@tmit.bme.hu

Reviewed

Keywords: dynamic optical network, GMPLS, traffic grooming, statistical multiplexing

Multilayer optical core networks are able to provide huge bandwidth. With traffic grooming we can utilize more efficiently the available resources. The principle of grooming: if the routes of two different traffic flows (or demands) have common links, their traffic can be joined in to the same wavelength channel. Another well-known solution for increased efficiency of resource usage is multiplexing the traffic. The statistical multiplexing does not allocate the maximal bandwidth for each traffic demand, but less than the maximal and more than the average. The aim of this article is to investigate the effects of applying both solutions.

1. Introduction

The optical core networks are based almost exclusively on optical transmission, because this technology provides huge bandwidth: A single optical channel typically carries data at a rate of 10 Gbps. In addition, the application of Wavelength Division Multiplexing (WDM) enables that a fibre can transmit more simultaneous signals using parallel channels. Depending on the number of parallel channels we differentiate Coarse Wavelength Division Multiplexing (CWDM) and Dense Wavelength Division Multiplexing (DWDM) systems. In DWDM systems more Tbps can be provided.

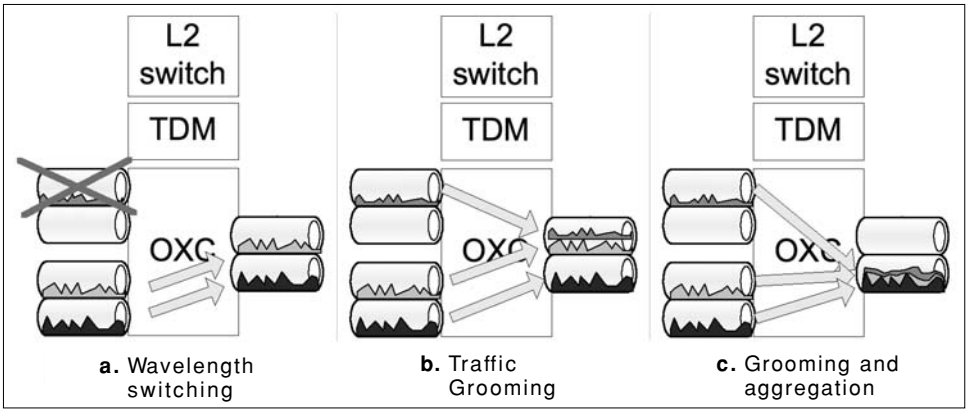
In such DWDM networks connections between distant nodes are realized using lightpaths that may be defined in advance or on-demand. Each lightpath is a sequence of wavelength channels and traffic enters and leaves it only through its endpoints, so these lightpaths can be imagined as pipes laid in the network. The clear purpose of the operator in such networks is to use the available resources efficiently via properly configuring the lightpaths. The network operator faces two problems related to resource allocation:

(1) the traffic demands have bandwidths by orders of magnitude lower than the size of a wavelength channel, and (2) the traffic rate fluctuates, so it does not use the whole allocated bandwidth in the significant part of time. The first problem is solved by the traffic grooming concept [2], while the second is by statistical multiplexing. The two areas have own considerable literature, but as far as we know the effects of joint application of these two solutions haven not been investigated in switched optical network, yet. We aimed this problem in our article.

We illustrate the above problem through the following example. *Figure 1.* depicts a node with three ports and each port has two wavelength channels. The traffic arrives from three different sources while their destinations (or the next node along the paths) are the same. The traffic from three sources arrives on different wavelength channels.

If wavelengths switching is allowed only (Figure 1.a), then only two of the three flows can be forwarded, thus, the third flow would be blocked at connection setup phase. If traffic grooming is supported, the traffic of all three sources could be combined to one channel. However, the sum of the maximal bandwidths given by traffic descriptors exceeds the capacity of the outgoing wavelength channel; therefore, only two of the three flows can be groomed into one channel and a further channel will be defined for the third flow (Figure 1.b). Finally, if we allow that less than the sum of the maximal bandwidth requirements is allocated – the exact method of how to calculate this value will be detailed later –, then all the three traffic flows can be carried in the same channel (Figure 1.c).

Figure 1.
Joint application of grooming and statistical multiplexing based aggregation for switching



Whenever, only whole wavelengths can be switched (there is no grooming), considering statistical multiplexing is meaningless in this example, since the three flows arrive in different channels and they cannot be groomed, so there would be no multiplexing gain at all.

2. Resource allocation and routing in switched optical network

Static resource allocation in switched optical networks is used, when the traffic demands are static, i.e., neither their bandwidth nor their endpoints change. The traffic matrix formed by these demands is constant: it can be defined in advance. In this case traffic routing problem can be formulated as an optimization task and good or even optimal solution can be found. On the contrary, in real networks the traffic demands arrive in different moments, and after different holding times they leave the network. It can be better described using a dynamic allocation model, where both the intensity and the space distribution of the traffic varies in time. In that case the arriving traffic demands are served in sequence one after the other.

2.1. Wavelength graph model

For the formal description of the dynamic routing we assumed the Wavelength Graph model. The base idea of this model is that the fibres between two nodes are modeled by many parallel edges as different wavelengths are used over that fiber. In addition all physical nodes are described with type-dependent subgraphs that make possible to describe different types of nodes in simple and expressive way. This property is one of the most important benefits of this model. In this paper we assumed two node types with different abilities:

The optical cross connects (OXC) realize switching whole wavelength channels between fibres. Additionally they have optical add-drop multiplexer (OADM) functionality in order to act as ingress and egress points for the traffic demands. On the contrary, *the grooming nodes* complete the property of OXC nodes; the grooming nodes can multiplex more traffic in one common channel, so they have grooming capability. The grooming is detailed in the next section.

2.2. Traffic grooming

The bandwidth requirements of traffic demands are typically less than the capacity of a wavelength channel. A whole channel assigned to one traffic demand wastes the resources in most cases, therefore, the ligthpaths should be shared among more traffic demands. When two or more demands have the same source and the same destination they can be *multiplexed* in the electronic layer of the ingress node and transferred along one ligthpath. At the egress node they can be de-multiplexed. However it cannot be applied in the case, when the traffic demands do not have the same source or destination.

In this latter case, the ligthpaths have to be torn down before and after the common part of their paths and at this point the traffic is transmitted to the electronic layer. The traffics are multiplexed there – e.g. by using time division – and in the common part transferred in one channel. This solution is referred as *traffic grooming*. In general traffic grooming is, when the traffic arriving from one or more channels is rearranged in higher – electronic – layer (e. g. according to their destination), and they are forwarded in a common channel combined together.

The clear benefit of grooming is the efficient usage of wavelength channels. However, at the same time the grooming requires expensive optoelectronic converters to route the traffic to the electric layer. Therefore in the dimensioning phase the cost of the grooming must be taken into account. Nevertheless, in this paper we do not deal with these design and dimensioning questions.

2.3. Statistical multiplexing

Significant part of the traffic in the core network is provided by data traffic, which has rate variance in time. This raises an issue of how much capacity should be allocated for the traffic demands? Traditional solution is to allocate resources equal to the sum of maximum bandwidths required by each demand. This is *deterministic multiplexing* that results in an over-dimensioned network. The amount of over-provisioned capacity can be decreased with the application of statistical multiplexing (or aggregation). In this case we exploit that more sources likely do not generate traffic at peak rates. So for the aggregated traffic a limit can be defined, for which the probability of aggregation extending the defined limit is a fixed low value. This latter parameter is the *overflow* (or the packet loss) *probability* and the name of the defined limit is *effective bandwidth*.

The theoretical principles can be found in F. Kelly's paper [4]. S. Floyd has proposed a simple method for defining the necessary capacity based on the Hoeffding bound:

$$BW = \sum_{i=1}^{n} m_i + \sqrt{\frac{\ln\left(\frac{1}{\varepsilon}\right) \cdot \sum_{i=1}^{n} p_i^2}{2}}, \qquad (1)$$

where m_i and p_i are the average and the maximal rate of the ith basic flow, and ε is the probability that the aggregated traffic exceeds the allocated capacity denoted by *BW*. It has two benefits: it is easily calculable and it is a conservative estimation (it guarantees that after the given border condition the bandwidth will not be larger than the calculated value). Serious drawback is that this model is rather inaccurate in core networks, since the traffic is already aggregated, so the fluctuation is smaller. Therefore, this model is not applicable.

To construct more accurate models we have to make assumption about the characteristics of the traffic. We assume that the arriving traffic flows are mutually independent and their sizes follow the Gaussian distribution. This assumption approximates well the reality, because

the traffic of each demand is also aggregated. In this case the Guerin model is applicable [6]. The allocated bandwidth can be calculated as follows:

$$BW = \sum_{i=1}^{n} m_i + \alpha \cdot \sigma , \qquad (2)$$

where m_i the average rate of the elemental sources, while σ is the deviation of the rate of the aggregated traffic. Since more basic and independent flows are assumed for each demand, the aggregated traffic will also follow the Gaussian distribution. This distribution remains when these flows are further aggregated. In this case the overflow probability (i.e., the probability that the aggregation exceeds the allocated capacity) is well characterized with parameter α. For instance, in order to keep the overflow probability at 0.01, the value of alpha must be 2.32, and in the case $\alpha = 5.61$, this probability will be 10^{-8}. Since an additional assumption is that all elementary streams are mutually independent, the variance of the aggregate is easy to calculate: it is equal to the sum of deviations of the elemental traffics.

In [7] an extension of this model is discussed, and it introduces several methods to calculate the α parameter. The benefit of this model is that the effective bandwidth can be easily calculated; furthermore, it describes well the real traffic. At the same time among the traffic descriptors the deviation of the traffic has to be given as well, or should be estimated from the other given parameters (e.g. from average or the maximal rates).

The principle of Lindberger's approximation is that it replaces the original cell rate distribution by a process composed of equivalent Poisson bursts [8]. The resulted formula defines the needed bandwidth for each elemental traffic flow. Summarising this we get the following formula, which is proportional to the average bandwidth requirement and to the variance, and inversely proportional to the capacity of the channel *(C)*:

$$BW = \sum_{i=1}^{n} a \cdot m_i + b \cdot \frac{\sigma^2}{C} , \qquad (3)$$

where *a* and *b* depend only on the packet loss probability:

$$a = 1 - \frac{\log P_{loss}}{50} , \qquad b = -6 \cdot \log P_{loss} .$$

In our examination we used the (1,18; 63) pair of parameter, with which the accessible overflow probability is $P_{loss} = 10^{-9}$.

With the following capacity estimation formula (PCRSCR) the allocated capacity is equal to the sum of the average bandwidth of each demand, and this amount has to be increased with the maximum among the difference of maximal and average bandwidths.

$$\sum_{i=1}^{n} m_i + \max_{i=1..n} \{p_i - m_i\}. \qquad (4)$$

Beside the models discussed here, several other models are known; however, their computational complexity is larger than these models. Nevertheless, our aim is to investigate the effects of introducing statistical multiplexing in optical networks having grooming capability and not to compare the different models.

However, the model proposed by S. Floyd works well only when the traffic flows has great variance: their peak rates are significantly higher than their mean rates. Here, in core networks the traffic flows are already aggregated, so their fluctuation is smaller. Because of this problem, the Floyd model has to be excluded. Hence, we focus on the three other models: the Guerin, the Lindberger and the PCRSCR ones.

3. Investigation of common usage of statistical multiplexing and grooming

The performance of the joint application of grooming and statistical multiplexing was investigated through simulation. For the simulation we have applied a simulation tool called Intra- and Inter-Domain Routing (IIDR). The IIDR is a discrete event simulator developed at our Department. It simulates among others the dynamic behaviour of a given network for different traffic loads.

The network provides connectivity between distant nodes and capacity is allocated for them. The parameters of the connections source and destination address, holding time, average and maximal bandwidth requirements define a *traffic demand*. During the simulations these demands arrive one after the other into the network and the routing algorithm serves them one by one. In the first step it looks for a path between the source and the destination nodes. Searching the route is performed over the logical graph, which is based on the previously introduced wavelength graph. Before the path searching for a demand, the edges having insufficient amount of free capacity are pruned temporarily from the graph. Along the paths found in this reduced graph there will be enough free capacity for the demand, therefore, a shortest path finding method (e.g., Dijkstra's) can be used. If a route exists between the source and destination pair the required amount of capacity can be allocated. Otherwise the demand will be blocked avoiding the latter congestion. In the case of *deleting* a demand the simulator frees the resources allocated to the demand in one step.

The effects of the different investigated aggregation models appear in the routing step. To check whether there is enough free capacity on the link is performed as follows. For each link the algorithm calculates how much capacity would be allocated if the demand to be routed was used the considered link. If this estimated bandwidth is less than the link of the capacity, free capacity will remain after routing the demand. Otherwise, there is no room for the demand on the considered link, thus, the link will be temporarily pruned form the graph.

3.1. Simulation environment

The performance evaluation is conducted in three steps. First, the traffic demands are generated in advance by an application developed for this purpose making possible to perform more independent simula-

tions on the same traffic sequence. Second, the simulations were performed using the generated traffic patterns. Finally, the collected results were evaluated.

3.2. Topologies

We conduct simulations on the reference network of COST 266 European Union project [8]. We used two versions: The first is the COST 266 core topology and the second is the COST 266 ring topology. A core topology consists of 16 nodes and 23 edges, the degree of the nodes less than three. The ring topology consists of 28 nodes and 35 edges, here the average degree of the nodes is 2.5. In the case of both topologies 4 wavelength channels are defined between each pair of nodes, and the capacity of a channel is 1000 Mbps.

3.3. Traffic demands

The traffic demands are described using six parameters: the source and destination nodes, the time when the demand is invoked, the holding time, and finally, the bandwidth requirement of the demand defined by its peak and its mean rates. The demands to be routed are generated randomly in advance to make possible the investigation of the different models over the same traffic sample. These samples are defined as follows. The arrivals are modeled as a Poisson process: the intensity is the inverse of the average interarrival time of consequent demands. The average of the holding time of the demands is also defined. The two descriptors of the bandwidths of demands are calculated as follows: The average peak rate is calculated from the link capacity and it is described with the peak-rate to channel capacity ratio (PCR/CH ratio). The mean rate is derived from the peak rate via multiplying the peak rate with the peak-to-mean ratio (PCR/SCR ratio).

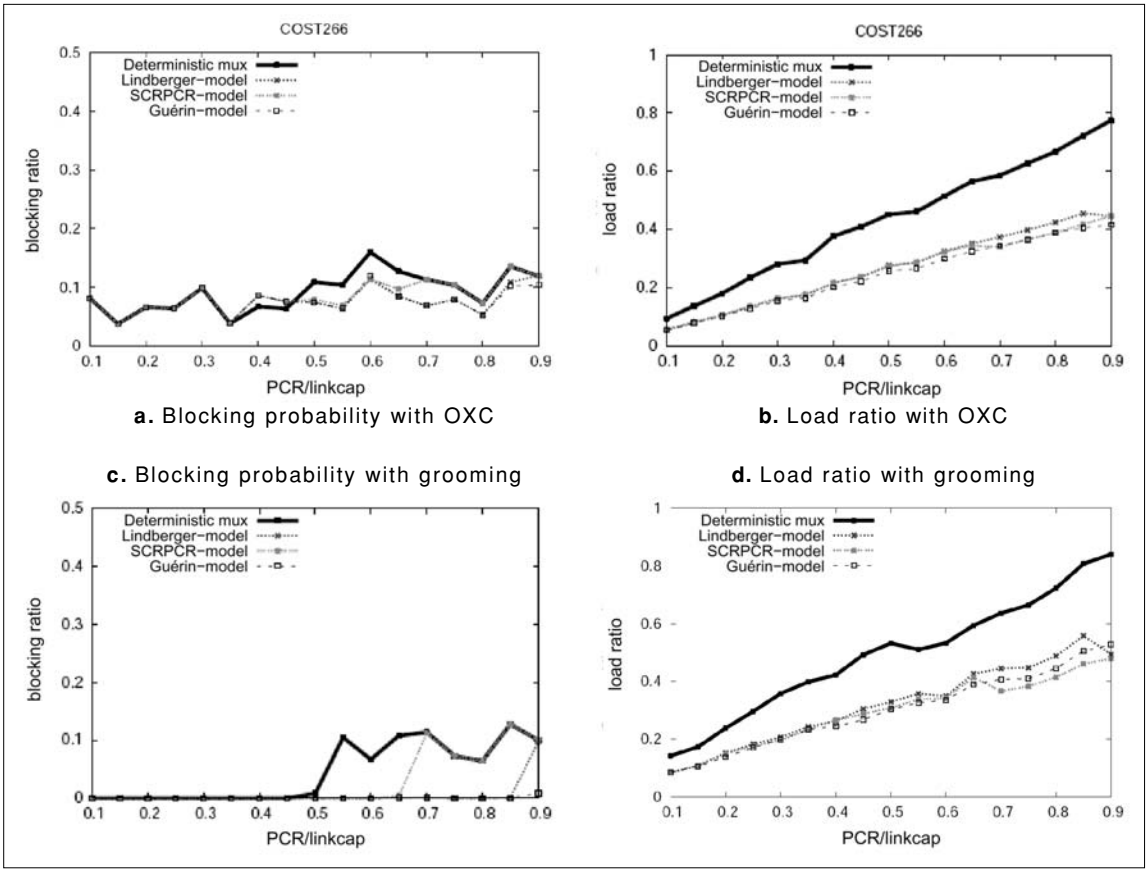
3.4. The investigated parameters

Blocking probability is maybe the most important property from the point of view of the network operation. It indicates how many traffic demands could be served by the network and how many remain *blocked*. If the blocking probability is less, then more demands can be served on the given network, which can result higher income.

Load ratio shows the average load of the network links. This parameter is a good *estimator* of the efficiency of the model and of the method applied. Since there are lightly loaded links in the network, they can serve more traffic, i.e., more demands. Therefore, it also decreases the blocking probability.

4. Results

The simulations are performed on both topologies. In the case of both topologies we define two basic *cases*. In the first case the nodes were OXCs, while in the second one the nodes had grooming capability. In these four base cases we measured the blocking probability and the load ratio. To describe the size and the dynamicity of the traffic we have introduced two metrics: the ratio of maximal bandwidth requirements of demands to the capacity of the wavelength (load), and the ratio



a. Blocking probability with OXC

b. Load ratio with OXC

c. Blocking probability with grooming

d. Load ratio with grooming

Figure 2.

of the average and the maximal sizes of the demands (variability). We investigate the blocking probability and the load ratio by changing these two parameters. On both topologies we have obtained similar results.

In the following simulations we varied the ratio of the size of each traffic demand to the channel capacity from 0.1 to 0.9. Additionally we assumed the ratio of the maximal to average bandwidth, so the variability was 2:1. The next figures show the results *(Fig.2.)*.

Let us assume that all nodes in the network have only wavelength switching capabilities, i.e., all nodes are OXCs. We have measured that difference between the network capacities allocated using deterministic and statistical multiplexing models steadily increases (see Figure 2.b). However, this capacity save is in vain: the blocking probabilities are roughly the same in all multiplexing models (Figure 2.a).

If all nodes have grooming capability a bit higher network load is measured (see Figure 2.d.) compared to the OXC case. However, in this case the capacity save greatly affects the blocking probabilities (see Figure 2.c): the deterministic multiplexing starts to block when the peak rate of a traffic flow reaches the 50% of the link capacity, while the blocking probabilities of the various statistical multiplexing models is roughly 0. These models start to block at higher peak-to-link-capacity ratios. First the SCRPCR blocks at 65%, and then the Lindberger and the Guerin at 85%. When the demands start to be blocked, then the number of traffic demands routed in the network also decreases. This results in the load drops observed on Figure 2.d.

The above measurement we executed by different variability of the demands. Our experiences show that, increasing the variability of the traffic the statistical multiplexing was more efficient against the deterministic multiplexing when grooming is allowed.

We can also see that, when OXCs are used there is no significant aggregation gain, so we cannot serve more traffic in the network. On the contrary, with traffic grooming the blocking probability decreases significantly in the investigated cases, so we can serve more traffic.

5. Conclusion

In this paper we investigated the joint application of traffic grooming and statistical multiplexing in multi-layer optical core networks. In the first step we presented the topic of optical core network focusing on traffic grooming and statistical multiplexing. The purpose of this paper is not the presentation of the entire *storehouse* of the models, hence we selected four approaches to investigate. The dynamic behaviour of the network is evaluated by simulations performed by a simulator tool developed at the Department. The simulations showed that, if the traffic is multiplexed (without grooming), then the gain of statistical multiplexing appears only as a decrease of the allocated resources, but the blocking ratio remains the same. Therefore more traffic cannot be served by the network.

We had expected that applying statistical multiplexing with traffic would decrease the blocking probability. In the paper we showed that, this difference can be large: e. g. with Guerin's model the network starts blocking at 0.9 PCR/linkcap load value (the ratio of maximal bandwidth requirement of traffic to capacity of one channel), while with deterministic multiplexing already at 50% of link capacity.

Acknowledgement

The authors of the paper would like to thank to János Szigeti and Péter Hegyi for the development of the basic version of the simulator, and Géza Geleji for the control scripts written in perl.

References

[1] B. Rajagopalan, J. Luciani, D. Awduche: "IP over Optical Networks: A Framework" IETF RFC 3717, http://www.rfc-archive.org/getrfc.php?rfc=3717.

[2] E. Modiano, P.J. Lin: "Traffic Grooming in WDM Networks", IEEE Communications Magazine, Vol.39., No.7, 2001., pp.124–129.

[3] Cs. Gáspár, G. Makács, T. Cinkler: "WR-DWDM hálózatok konfigurálása", Híradástechnika, Vol. LVIII., 2003/7, pp.2–9.

[4] F. Kelly: "Notes on effective bandwiths," in Stochastic Networks: Theory and Applications, F. Kelly, S. Zachary, and I. Ziedins, Eds. Oxford University Press, 1996., pp.141–168.

[5] S. Floyd: "Comments on Measurement-Based Admissions Control for Controlled-Load Services," beadva: CCR 1996. július http://www.icir.org/floyd/papers/admit.ps

[6] R. Guerin, H. Ahmadi: "Equivalent Capacity and its Applications to Bandwidth Allocation in High-Speed Networks", IEEE Journal on Selected Areas Communications, Vol.9, 1991. szeptember, pp.968–981.

[7] L. Noirie: "Mixed TDM and Packet Technologies as a Best Compromise Solution to Ensure a Cost-Effective Bandwidth Use with the Current Traffic Evolution" in Next Generation Internet Networks Conference (EuroNGI), Róma, 2005.

[8] K. Lindberger: "Dimensioning and Design Methods for Integrated ATM Networks," in Proc. 14th Int. Teletraffic Congress, 1994, pp.897–906.

[9] Robert Inkret, Anton Kuchar, Branko Mikac: "Advanced Infrastructure for Photonic Networks", Extended Final Reporf of Cost Action 266, p.20. http://www.ure.cas.cz/dpt240/cost266/docs/COST266_Extended_Final_Report.pdf

Speech recognizer for preparing medical reports:

Development experiences of a Hungarian speaker independent continuous speech recognizer

KLÁRA VICSI, SZABOLCS VELKEI, GYÖRGY SZASZÁK, GÁBOR BOROSTYÁN, GÉZA GORDOS

BME, Dept. for Telecommunication and Mediainformatics, Laboratory of Speech Acoustics
{vicsi,szaszak,gordos}@tmit.bme.hu

Reviewed

Keywords: automatic speech recognition, HMM models, n-gram models, bi-gram models, perplexity

A development tool (MKBF 1.0) for constructing continuous speech recognizers has been created under Windows XP. The system is based on a statistical approach (HMM phoneme models, and bi-gram language models with non linear smoothing) and works in real time. The tool is able to construct a middle sized speech recognizer with a vocabulary of 1000-20000 words. New solutions have been developed for the acoustical pre-processing, for the statistical model building of phonemes, and in syntactic level. Through our examination, different training sets were used with different vocabularies. Hungarian is a strongly agglutinative language, in which the number of the word forms is very high. This is the reason why two forms of bi-gram linguistic model were constructed: one is the traditional word forms based and the other is the morpheme based model, in which the vocabulary is much smaller. In this article, test results and the experiences drawn from them are presented. Recognition accuracy has been considerably increased using perplexity based linguistic adaptation.

Introduction

Hungarian belongs to the Finno-Ugric Language family, and – like the other members of this family – is a strongly agglutinative language. The number of different word forms is about hundreds of millions. Word forms are composed by oblique stem and suffixes. In addition, suffixes influence the form of stem in many cases.

The phonetic transcription of written form to spoken one can be easily generated using some rules, but the pronunciation of most words starting or ending with a consonant depends on the adjacent words, because difficult consonant combinations are replaced by simpler ones by a hierarchy of the phonological rules. The situation is more complicated in case of linking morphemes.

In the Laboratory of Speech Acoustics of the Budapest University of Technology and Economics a Hungarian continuous speech recognizer (ASR) has been developed according to the standard knowledge components in a state-of-the art ASR system. These components, the acoustic pre-processing, the acoustic-phonetic model [4] and the syntactic, morpho-syntactic models have been optimalized.

The acoustic pre-processing is the following: the sampling rate is set to 16 kHz, data is coded on 16 bits. The frequency analysis was done in Bark scale (Bark filterbank using 17 bands). The observation sequence vectors, including first order time derivatives are calculated every 10 ms: 17 delta frequency Bark coefficients +17 delta time frequency coefficients +1 energy coefficient are used alltogether.

Phoneme based *acoustic-phonetic models* were used for modeling the Hungarian phonemes. These models are Quasi Continuous Hidden Markov Models (QCMM) with 24 steps, and 5 states. These models were trained and tested by using the Hungarian Reference Speech Database [9]. The test results of the acoustic pre-processing and the phoneme based acoustic-phonetic models were presented in our earlier work [8].

In this article the development of *language models* in syntactic/morpho-syntactic level is presented. Bi-gram models were constructed in two different ways: in the first experimental setup, the basic linguistic units are the word forms; in the second setup, morpheme based bi-gram models were constructed. The training corpus consisted of medical reports collected from the Semmelweis University of Budapest (4000 records) and from the Medical University of Szeged (6365 records) in the field of endoscopy. In the first setup, the vocabulary of the word forms (with 14 331 words) and in the second one, the vocabulary of morphemes (with 6 824 morphemes) together with their pronunciation were prepared based on this corpus. The HUMOR morpheme analyser [5] was used to split the words into morphemes. Medical reports were composed automatically by the computer by recognition of the utterances pronounced by physicians during the examination of patients. These examination reports had been recorded from 5 speakers (4 records from each of these 5 physicians were used). These reports were recorded at the Semmelweis University of Budapest.

1. The bi-gram language model

1.1. Description of the language model

We used a probabilistic language model (LM) based on the assumption that the probability of a word occurence depends on the words preceeding it. If the language model computes the probability of a word

occurrence using the previous n-1 words, it is called an n-gram LM. In practice, language models are usually bi-grams or tri-grams.

The probability of an n-gram is computed from its frequency within a training text, or corpus. In most cases, corpus must be very large.

If the probability of a sequence of words is referred to as $\hat{P}(w_1, w_2, ..., w_m)$ then:

$$P(w_1, w_2, ...w_m) = P(w_1)\prod^{m} P(w_i | w_{i-1} ... w_1) \quad (1)$$

By limiting the context this can be replaced by the following approximation:

$$P(w_1, w_2, ...w_m) \cong P(w_1)\prod_{i=2}^{m} P(w_i | w_{i-1} ... w_{i-n+1}) \quad (2)$$

where $n>0$ is an arbitrary selected integer number.

The probability of a word occurrence using the previous n-1 words:

$$P(w_i | w_{i-1} ... w_{i-n+1}) = \frac{N(w_i ... w_{i-n+1})}{N(w_{i-1} ... w_{i-n+1})} \quad (3)$$

where *N(.)* is the number of occurence for a given word in the training set. We have used bi-gram model in our experiments.

1.2. Smoothing of an N-gram model

The correct estimation of the probability of rare word events is a primary concern in building language models. Generally, the training corpus must be very large in order to ensure that rare words appear at least some times. Instead of increasing the size of the text corpus, different smoothing technics can be used to compensate for data sparsity and to generalize the LM to better model unseen events [6].

We used a non-linear smoothing technique. This smoothing method, based on *absolute discounting,* generally outperforms the others proposed in the literature [7]. Formula applied to the evaluation of the conditional bi-gram probabilities becomes (4):

$$\hat{P}(w_j | w_i) = \max\left\{\frac{N(w_j, w_i) - D_i}{N(w_i)}, 0\right\} + D_i \frac{|V| - n_0(w_i)}{N(w_i)} P(w_i)$$

where $|V|$ is the size of the vocabulary, $n_0(w_i)$ is the number of bi-grams that have the predecessor word w_i and that never occur during the training, $P(w_j)$ is the probability of the unigram w_j, $0 \le D_i \le 1$ is a constant value.

The non-linear interpolated model has some noteworthy properties, interesting by modeling the conditional probabilities. Indeed, if a certain predecessor word is followed by a single word or by a few different words, the effect of the smoothing will be less than in case if the word is followed by many different words. If $D=1$, the events seen only once are handled in the same way as the unseen events.

$$D_i = \frac{|V| \cdot b}{n_0(w_i)}, \qquad \text{where } b = \frac{n_1}{n_1 + 2n_2} \quad (5)$$

Here n_1 and n_2 are the number of bi-grams detected exactly one and two times in the training set. We note that D has an index depending on the predecessor word, i.e. it is constant for all the bi-grams that have the same predecessor.

If the factor $\frac{|V|}{n_0(w_i)}$ in (5) is neglected then D becomes independent of the predecessor word [6].

2. Testing the language model, examination of perplexity

For the training of bi-gram LM models we used the training corpus mentioned in the Intruduction. Training texts were composed of corrected, annotated and phonetically transcribed medical reports collected from two hospitals. This training corpus was divided into 4 groups:

- **G.1** – Gastroscopy reports from Semmelweis University of Budapest, Faculty of Medicine, II. Department of Medicine (Budapest gastroscopy)
- **G.2** – Gastroscopy reports from University of Szeged, Faculty of Medicine, (Szeged gastroscopy)
- **G.3** – Colonoscopy reports from Semmelweis University of Budapest, Faculty of Medicine, II. Department of Medicine (Budapest colonoscopy)
- **G.4** – Colonoscopy reports from University of Szeged, Faculty of Medicine, (Szeged colonoscopy)

These four groups and their combinations were used for the training. For training of the acoustic-phonetic models, the Hungarian Reference Speech Database [9] was used.

2.1. Training conditions

Before training the LM, the vocabulary of the training texts of colonoscopy and gastroscopy were examined. It was found, that only a small part of the vocabularies of Budapest and Szeged reports was common, as it can be seen in the *Table 1* (on next page). The reason for this relatively poor coverage between Budapest and Szeged corpus groups is the use of different expressions for preparing medical reports in the two institutions. Finally, all reports included in the four groups were used together for LM training.

Our later analysis also showed that the vocabularies of the materials given for testing the recognizer contained some new words which were not included in the training corpus.

2.2. Perplexity based WER estimation

The recognition accuracy of speech recognition systems is usually characterized by the *Word Error Rate* (WER) in scientific literature.

The calculation of WER is an expensive process, moreover, it would be very practical to introduce such an indicator that can estimate the recognition accuracy irrespectively of the acoustic-phonetic level. Perplexity is such an estimation method which can help to examine

the language model separated from the acoustic one. The calculation of the perplexity is given by the following equation:

$$PP = \left(\prod_{i=1}^{N} P(w_i \mid w_{i-1}) \right)^{-\frac{1}{N}} \quad (6)$$

where N is the number of words in the test corpus, w_i and w_{i-1} are the i^{th} and $(i-1)^{th}$ words of the test corpus.

In case of a morpheme based LM, words are replaced by morphemes in the above formula. The value of perplexity is a real number greater than 1. The closer this value to 1, the better the coverage of the language model on the given corpus. Too high values refer to a language model that does not really cover the selected test corpus.

3. Test results

To evaluate recognition results, we use some metrics and abbreviations which are:

- ***Ref**: number of units (words or morphemes) to be recognized*
- ***Rec:** number of units recognized*
- ***Corr**: number of units correctly recognized*
- ***Ins:** number of units inserted*
- ***Del:** number of units deleted*
- ***Subs:** number of units substituted*
- ***Accuracy:*** $Acc = \frac{Corr - Ins}{ref}$ (7)
- ***Word Error Rate:*** $WER = 1 - \frac{Corr}{ref}$ (8)

Standard recognition results for word-based LM trained on the whole medical corpus (G.*) are shown in *Tables 2 and 3*. The reason for the poor recognition performance (high word error rates) was found to be that although the LM training corpus covers well the test corpus in terms of vocabulary, but not in terms of bi-gram entries. Typically conjunction words were mostly confused, which is explained by the different reporting standards in the 2 hospitals. By training the LM on the mixed corpus, we also incorporated a high "noise" into the bi-gram field.

If we have a look at Table 1, the differences between the vocabularies of Budapest and Szeged hospitals are obvious. Since speech data used for testing was recorded in Budapest, we trained the LM by using only the G.1 Budapest gastroscopy corpus. Results for this setup are shown in *Table 4*.

Table 2.
Test results for gastroscopy with word unit based bi-gram LM trained on G.1–G.2–G.3 and G.4 mixed; tested on recorded medical reports spoken by physicians.

Ref [# of units]	Rec [# of units]	Corr [# of units]	Ins [# of units]	Del [# of units]	Subs [# of units]	Acc [%]	WER [%]
1173	1580	750	451	22	401	25.4	36.1

Table 3.
Test results for colonoscopy with word unit based bi-gram LM trained on G.1- G.2 - G.3 and G.4 mixed; tested on recorded medical reports spoken by physicians

Ref [# of units]	Rec [# of units]	Corr [# of units]	Ins [# of units]	Del [# of units]	Subs [# of units]	Acc [%]	WER [%]
890	1326	504	822	8	370	-35.7	43.4

Ref [# of units]	Rec [# of units]	Corr [# of units]	Ins [# of units]	Del [# of units]	Subs [# of units]	Acc [%]	WER [%]	PP
1150	1417	799	283	8	343	44.8	30.5	73.59

Table 4.
Test results for gastroscopy with word unit based bi-gram LM trained on G.1, tested on recorded medical reports spoken by physicians

Results presented in Tables 2-4, were obtained by using the utterances of physicians. These records were however relatively noisy, articulation was also poor. Hence, we re-recorded the same 20 reports in a low-noise environment with accurate, standard articulation in order to examine the effect of acoustic quality on recognition performance. By using the same LM trained on G.1 corpus, results are shown in *Table 5.* As it can be seen, WER was reduced considerably. (The reason for the little increase in the number of reference units in *Table 5.* compared to Table 4. is explained by the accurate articulation.)

Ref [# of units]	Rec [# of units]	Corr [# of units]	Ins [# of units]	Del [# of units]	Subs [# of units]	Acc [%]	WER [%]	PP
1173	1451	922	280	1	250	54.7	21.3	73.59

Table 5.
Test results for gastroscopy with word unit based bi-gram LM trained on G.1, tested on medical reports recorded with accurate articulation in low-noise environment

For a special case, we evaluated the recognition performance in case of utterances included in LM training corpus. For this purpose we have chosen 10 reports included in the Budapest gastroscopy (G.1) training corpus, we recorded them in low-noise environment with accurate articulation. Results are shown in *Table 6.*, as expected, WER is much lower, perplexity is also very low. These results can be regarded as theoretical optimum, in this case it was excluded that a missing word from vocabulary, or a forbidden bigram entry influence recognition performance.

Ref [# of units]	Rec [# of units]	Corr [# of units]	Ins [# of units]	Del [# of units]	Subs [# of units]	Acc [%]	WER [%]	PP
416	444	380	28	0	36	84.6	8.6	9.36

Table 6.
Test results for gastroscopy with word unit based bi-gram LM trained on G.1, tested on medical reports included in G.1, recorded with accurate articulation in low-noise environment

3.1. Conclusion for word based tests

Summarizing the results for word based LM testing, it is obvious that Budapest gastroscopy and Budapest colonoscopy reports are very different to the ones of Szeged in terms of expression syntax. A bi-gram LM based on a mixed Budapest-Szeged corpus is more robust and includes more words in the vocabulary, but in practice, recognition performance is worse by 5.54% than in case of a bi-gram trained only on Budapest gastroscopic reports.

Another crucial point is a relatively correct, accurate articulation and a lowered noise level. We should remark that 1 physician of the 5 asked to test the recognizer spoke very low. If we eliminate his 4 reports from the test set WER decreases to 24.27% from 30.52%.

The bi-gram LM trained on G.1 does not cover sufficiently the area recommended by a physician user. This

is verified in the last experiment and seen in Table 6. Moreover, a detailed analysis of errors in the 20 test utterances recorded from physicians and in the 20 utterances recorded with accurate articulation, a very high correlation was found between the errors in the corresponding poor articulation – good articulation speech utterances, which refers to data sparsity problem concerning the LM training data.

Finally, a low rate for *Acc* might refer to an improper coverage of the LM, since in this case the number of insertions increases radically. Inserted units are usually frequent words consisting of one or two syllables, whose vowels correspond to the vowels of the original word misrecognized.

3.2. Perplexity analysis

Relying on perplexity defined in section 1.2., we have seen in section 1.3. that perplexity is occasionally a good predictor of word error rates during recognition. In this section we would like to examine if this prediction is feasible or not. Perplexity is usually used to characterize the language model, but does not deal with the acoustic level: hence, in theory, it is possible that a LM with relatively low perplexity on a corpus yields worse recognition results than a LM of higher perplexity because of eventual acoustic similarity of vocabulary elements.

On *Figure 1*, correlation between perplexity and word error rate is illustrated, measured on G.1 corpus for the original test set of 20 medical reports. As it can be seen, a linear dependence of WER on perplexity can be assumed [1]. The variation of values on Fig.1. is explained by the fact, that only a subset of the Hungarian language is investigated. The other reason for this has been already mentioned above: for some test utterances, acoustical distance between vocabulary elements is various, the closer they are in spectral properties, the more likely is the confusion. This causes the right "shift" of perplexity – WER value pairs seen on Fig.1 [4].

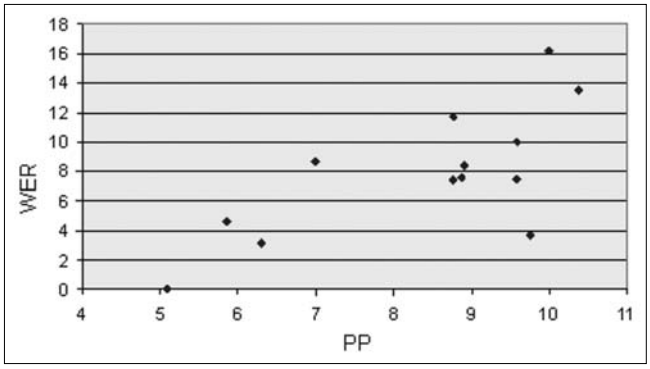


Figure 1. Correlation between perplexity and WER (LM trained on G.1, low noise utterances)

3.3. Word- or morpheme-based LM?

Beyond the "traditional" word-based bi-gram LM, a morpheme-based bi-gram was also prepared. Recognition tests for the morpheme-based LM were carried out with the same conditions as for the word-based one, the obtained results were quite similar for these two cases *(see Table 7)*.

The advantage of the morpheme-based language modeling is a considerable reduction in the size of vocabulary, hence the size of the bi-gram field to store and to use by recognition is also reduced, which is a critical issue, since bi-gram fields are usually stored in the memory during recognition to ensure a real-time operation. Based on the whole LM training corpus, the number of distinct words was found to be 14 331, but the number of distinct morphemes that covered fully the same corpus was only 6 706, less than the half of the number of words. Since the size of the bi-gram field is proportional to $|V|^2$ (the square of vocabulary size), this means that the bi-gram LM for words needs ~4.5 times more storage capacity in our case, moreover, the operation will also be slower. The values of bi-gram probabilities are higher in case of a minor vocabulary, which is also advantageous.

	Ref [# of units]	Rec [# of units]	Corr [# of units]	Ins [# of units]	Del [# of units]	Subs [# of units]	Acc [%]	WER [%]	PP
Word-based LM	1631	2045	1241	778	9	355	28.3	23.9	27.31
Morpheme-based LM	1173	1451	922	280	1	250	54.7	21.3	73.59

Table 7. Recognition results with word-based and morpheme-based LM trained on G.1, using test reports with accurate articulation

The disadvantage of morpheme-based language modeling is the difficult handling of assimilation phenomena across morpheme boundaries. The correct description of these events is not automatically feasible currently. Another problem might be the existence of some very short morphemes difficult to model (e.g. suffix *-t* in Hungarian to express accusative).

4. Expanding the language model based on perplexity measures

Perplexity can be used to predict recognition accuracy with the restrictions presented in section 2.2. We have also shown the difference between test data included directly or not included in the LM (see Tables 6 and 5). The results in Table 6 are obviously better, since the coverage of the LM was ensured for test corpus. In this section we would like to investigate, whether it is possible to increase recognition accuracy by including into the LM not the test corpus, but some typical word sequences from it. We would like to know also, how many times a sequence should be included to get a LM with corresponding weights for these bi-gram entries.

By training of missing word connections we were looking for word sequences that had not been included in the original training corpus G.1. This can be described by the following formalism (here '+' refers to one or more repetitions as in regular expressions):

```
<word included in G.1>
     <word not included in G.1>+
            <word included in G.1>          (9)
```

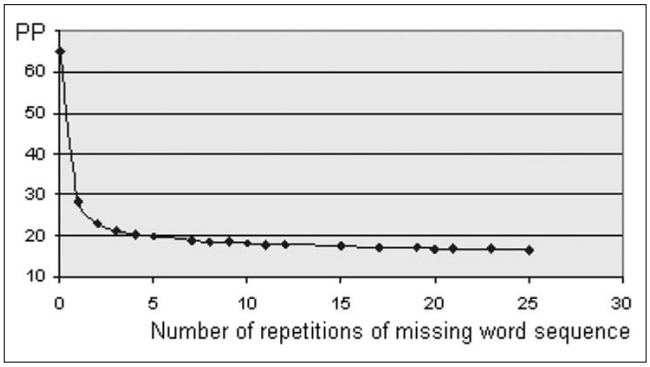


Figure 2.1.
Perplexity for test set depending on the number of repetition of missing word sequences

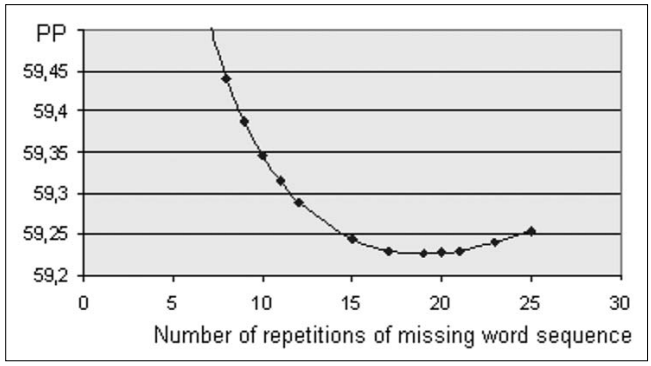


Figure 2.2.
Perplexity for control set depending on the number of repetition of missing word sequences

Our aim is to incorporate the missing bi-gram entries into the training corpus without distorting the actual LM. This explains why each word sequence selected for incorporation begins and ends with words already included in the corpus. Hereby, the context of new items will also be added.

4.1. Bi-gram weight optimization for new items

To determinate the number of times a selected word sequence should be added to the training corpus, we assume that relying on perplexity, we are able to predict recognition performance with the given LM training corpus.

Please note, that perplexity values are not comparable normally, but in this special case, it can be a good predictor because we carry out only minor modifications on the LM training corpus, preserving all other characteristics of data. The expanding process can be formalized like:

```
<original training corpus (G.1)>
    <selected word sequences>*
```
(10)

Table 8.1.
Recognition results for the selected 4 test reports with original (non-expanded) LM

Recognition of reports with LM from Original G.1 training corpus								
Report ID	Ref	Rec	Corr	Ins	Del	Subs	Acc	WER
3	95	116	67	21	0	28	48.4	**29.4**
33	63	80	48	17	0	15	49.2	**23.8**
53	55	68	50	13	0	5	67.2	**9.1**
92	55	62	46	7	0	9	70.9	**16.3**
Average WER:	**19.6%**		Average Acc:	**58.9%**				

Table 8.2.
Recognition results for the selected 4 test reports with expanded LM (20 repetitions of missing word sequences)

Recognition of reports with LM from expanded G.1 training corpus								
Report ID	Ref	Rec	Corr	Ins	Del	Subs	Acc	WER
3	95	110	78	17	1	16	64.2	**17.8**
33	63	69	61	6	0	2	87.3	**3.1**
53	55	58	53	3	0	2	90.9	**3.6**
92	55	61	49	6	0	6	78.1	**10.9**
Average WER:	**8.9%**		Average Acc:	**80.1%**				

Table 9.1.
Recognition results for 3 control test reports with original (non-expanded) LM

Recognition of reports with LM from Original G.1 training corpus								
Report ID	Ref	Rec	Corr	Ins	Del	Subs	Acc	WER
3	51	64	41	13	0	10	54.9	**19.6**
33	34	41	11	11	2	21	0.0	**67.6**
53	118	167	70	49	0	48	17.7	**40.6**
Average WER:	**42.6%**		Average Acc:	**24.2%**				

Table 9.2.
Recognition results for 3 control test reports with expanded LM (20 repetitions of missing word sequences)

Recognition of reports with LM from expanded G.1 training corpus								
Report ID	Ref	Rec	Corr	Ins	Del	Subs	Acc	WER
3	51	64	41	13	0	10	54.9	**19.6**
33	34	41	11	11	2	21	0.0	**67.6**
53	118	166	68	48	0	50	16.9	**42.3**
Average WER:	**43.2%**		Average Acc:	**23.9%**				

where '*' refers to the number of times a word sequence was added to the training corpus.

For testing, 4 medical reports were chosen from the test set. The other 16 test reports were also kept to control whether the LM becomes distorted. After the determination of missing word sequences included in these 4 reports, but missing form the LM training corpus, these word sequences were added, progressively increasing the number of times they were repeated, and controlling whether perplexity measures for the rest of the test reports (16) do not become worse. Our aim is to increase recognition accuracy (predicted by perplexity) for the 4 reports selected and keep or even improve perplexity of the rest 16 test reports.

In *Figure 2.1.* it can be seen that perplexity for the 4 test reports improves by adding the missing word sequences again and again. In parallel, in *Figure 2.2.* perplexity for the 16 control reports improves until 18-20 repetitions, but decreases after. According to Figures 2.1. and 2.2., the final repetition number was set to 20. To control recognition performance, a full test process was carried out again, after the expansion of LM training set. Results can be seen in *Tables 8.1-2. and 9.1-2.*

As expected, recognition results for the 4 selected reports are radically increased. Recognition results without incorporation of missing word sequences are presented in *Table 8.1.*, while after incorporation with 20 repetitions they change according to *Table 8.2.*

In *Tables 9.1. and 9.2.*, recognition performance for some control test reports are presented. As it can be seen, they are only slightly influenced by LM expansion, but there is no evidence of any LM distortion.

5. Conclusion

According to our investigations reviewed in this section, language model expansion (and its implicit re-weighting) is feasible, and by this expansion, the number of times a missing item should be repeated can be determined using perplexity. This procedure does not decrease perplexity and recognition performance for LM. This method can be used in the future to expand LM fast and efficiently. An implementation of a self adaptation (LM tuning for user's profile) algorithm might be a result of our current investigations.

On the other hand, the method we evaluated is not a universal solution to the problem of LM updating. By highly agglutinative languages, like Hungarian, usage of new items by the ASR user is always possible, but for a restricted area, like medical solutions for diagnostics, the method can be suitable to ensure better performance.

References

[1] Máté Szarvas, Sadaoki Furui:
Evaluation of the Stochastic Morpho-syntactic Language Model on a One Million Word Hungarian Dictation Task. Eurospeech, Genova, 2003. pp.2297–2300.

[2] Stanley Chen, Dougles Beeferman, Ronald Rosenfeld:
Evaluation Metrics For Language Models,
In: DARPA'98, National Institue of Standards and Technology (NIST), accessible:
www.nist.gov/speech/publications/darpa98/html/lm30/lm30.htm

[3] Philip Clarkson, Tony Robinson:
Towards improved language model evaluation measures, accessible:
http://Citeseer.ist.psu.edu/clarkson99toward.html

[4] Yonggang Deng, Milind Mahajan, Alex Acero:
Estimating Speech Recognition Error Rate without Acoustic Test Data, accessible:
http://research.microsoft.com/srg/papers/2003-milindm-eurospeech.pdf

[5] HUMOR: Hungarian Morpheme Analyser, accessible:
http://www.morphologic.hu/en_humor.htm

[6] Claudio Becchetti, Lucio Prina Ricotti:
Speech Recognition, Theory and C++ implementation, Fondazione Ugo Bordoni, Rome, 1999.
ISBN 0-471-97730-6

[7] Ney, H., Essen, U., Kneser, R.:
On Structuring Probabilistic Dependencies in Stochastic Language Modeling,
Computer Speech and Language, 1994/8. pp.1–38.

[8] Velkei Szabolcs, Vicsi Klára:
Beszédfelismerő modellépítési kísérletek akusztikai, fonetikai szinten, kórházi leletező beszédfelismerő kifejlesztése céljából
(ASR Model Building Experiments on Acoustic-phonetic Level for a Medical ASR Application),
in Hungarian: II. Magyar Számítógépes Nyelvészet Konferencia 2004.
pp.307–315.

[9] Vicsi Klára, Kocsor András, Teleki Csaba, Tóth László:
Beszédadatbázis irodai számítógép-felhasználói környezetben,
(A Speech Database for Office Environment),
in Hungarian: II. Magyar Számítógépes Nyelvészet Konferencia 2004.
pp.315–319.

Machine learning algorithm
for automatic labeling and its application in text-to-speech conversion

GÉZA KISS, GÉZA NÉMETH

BME, Department of Telecommunications and Media Informatics
{kgeza, nemeth}@tmit.bme.hu

Reviewed

Keywords: machine learning, language identification, LID, TTS

In this paper we present a novel machine learning approach usable for text labeling problems. We illustrate the importance of the problem for Text-to-Speech systems and through that for telecommunication applications. We introduce the proposed method, and demonstrate its effectiveness on the problem of language identification, using three different training sets and large test corpora.

1. Introduction

The number of speech-based telecommunication applications, which accept incoming calls through an IVR (Interactive Voice Response) system, is continuously increasing, both in Hungary and abroad. In the most modern systems the user can express his wishes not just by typing, but also through an ASR (Automatic Speech Recognition) system, and in response s/he will receive good quality speech. In the most simple case the response will contain prerecorded messages (so called "prompts") or messages assembled from a few separate items (limited vocabulary speech synthesis).

If the content of the message to utter is unpredictable or shows great variability, then we produce the speech using a Text-to-Speech system (TTS). A few examples for the latter from the services available in Hungary are the e-mail system of T-Mobile Hungary [1], its reverse directory system [2], or the loud SMS service of T-Com.

We expect TTS systems to generate good quality, understandable and correctly intonated speech using only the written form. But we can claim that the writing systems in use (whether it be Hungarian or of another language) contain only a fraction of the information content of speech and only hints at prosody using some punctuation marks. The human reader completes the text in his mind with the missing pieces of information using his world knowledge and the context, which helps him to read it out also (if necessary). For just determining the proper pronunciation and stress a system needs to correctly find out information such as the language of the text and of intruding foreign words and the role of the individual words in the sentence (part-of-speech, grammatical structure). In *Figure 1* you can see a few examples for Hungarian where this is not trivial because a different decision is needed even in the case of identical word forms or because the sentence contains a foreign language part.

In this paper we review the methods used for language identification from text, then describe a machine learning algorithm that learns from labeled text and can be used for various automatic labeling tasks. We explain it on the exemplar of language labeling, referring to the possibility of part-of-speech tagging; this kind of information is important for TTS programs, which are increasingly used in telecommunication also. Besides these, the method can probably be used in diverse other areas.

2. Methods for language identification from text

The notion of "Language Identification" (LID) can refer to the methods used for identifying the language of speech before ASR or to those used for identifying the language of texts. Language identification can be viewed as a special case of classification (or labeling) problems, so the lessons learned here can be applied for other similar problems, e.g. for part-of-speech tagging.

Figure 1. Examples of non-trivial language tagging and part-of-speech tagging tasks

***"a test"*– Hungarian or English expression:**	A lélek és *a test*. [The soul and *the body*.]	This is *a test*.
foreign word in Hungarian text:	A "Sok hűhó semmiért" *Shakespeare* műve. ["Much ado about nothing" is *Shakespeare*'s comedy.]	
***"egy"* – determiner or indefinite article:**	*Egy* vagy két alma. [*One* or two apples.]	*Egy* alma esett le a fáról. [*An* apple fell from the tree.]

Although at first glance most people who are interested in the topic will have ideas for automatically determining the language of texts, several issues harden the problem and make it far from trivial. While we can assign the correct language to a longer text with high probability using fairly simple techniques, the correct identification of the words in a mixed-language text with word forms occurring in more than one language is a much harder task. In addition, the effective solution, besides being precise, must be fast and need relatively small storage place.

2.1. Morphological analysis

Some of the systems, endeavoring to have correct identification on the word level, or even morpheme-level, use detailed morphological analysis, e.g. using DCG's (Definite Clause Grammar) [3], or indirectly using a spell checker [4]. In an intermediate solution no real morphological analysis takes place; instead they infer the language of words by matching items of a dictionary (made up of words and sub-word units) against the text, augmenting this with statistical methods [5].

The Humor [6] and Hunmorph [7] morphological analyzers can be used for Hungarian. However, if the decision to be made is not simply "Is it Hungarian or not?", but you have to decide on one of several potential languages, then you need a morphological analyzer for each one. This is hard to accomplish, moreover the necessary processing power may be problematic in certain applications.

2.2. Word-based method

Word-based methods [8] rely on the observation that in every language there is a fairly small set of words that are used very frequently. Therefore the presence of such words from a language indicates with high reliability that the text was written in that language. The most frequent 1000 words can make up 50 to 70% of all occurring word forms [9].

A disadvantage of the method is that the shorter the text, the more likely it is that no words from the list of its language will occur therein. Additionally you need considerable effort to assemble the wordlist, partly because some of the word forms will occur in several of the most frequent word lists of languages, as you can see from the examples in *Figure 2.*

Figure 2.
Examples of frequent word forms occurring in several of five European languages

word form	English	German	Spanish	Polish	Hungarian
de			X		X
a	X		X		X
na				X	X
to	X			X	
in	X	X			
do	X			X	
el			X		X
is	X				X
es		X	X		
mit		X			X
was	X	X			
ha			X		X
so	X	X			
on	X			X	
mi			X	X	X
most	X				X
ja		X		X	
be	X				X
ma				X	X

2.3. Vectorspace methods

The basic idea of vectorspace methods is to assign feature vectors to the document being examined and to the possible classification categories, which contain the numeric value of some properties, weighted by the importance of the feature. The feature vectors for the classification categories are created from belonging sample documents. The similarity of the document to a category is characterized by the scalar product of the feature vectors, where 0 means ortogonality and 1 means identity. In the approach described in [10], the features are the number of n-grams in the texts with N=2...5 and the number of short words or words with unlimited length.

2.4. Neural networks

In the approach used in [11] a Multi-Layer Perceptron (MLP) is trained. The input of the network are the characters within a window placed at each character position of the words, the output is a language probability at the position. A language decision is made for each word by combining the probabilities for each letter.

2.5. Decision-tree-based methods

A decision tree (different from the one in our proposal in section 3. is used in [12]: they train a separate decision tree for each different character, where the branches have questions about the identity of neighboring characters, and the leaves contain the most probable language. They make a word level decision, choosing the language candidate that was voted most often.

2.6. N-gram-based methods

N-gram based methods use sub-word items as the basic units for identification. These can be letter sequences of two, three or more characters or sequences of different lengths simultaneously. N-gram frequency statistics can be created from training corpora prepared for the languages.

N-grams are good at handling several problems that word-based solutions handle poorly. One of these is the frequent presence of spelling errors in electronic texts (coming from mistyping or OCR errors), as these have such great variability that they cannot be handled by storing all

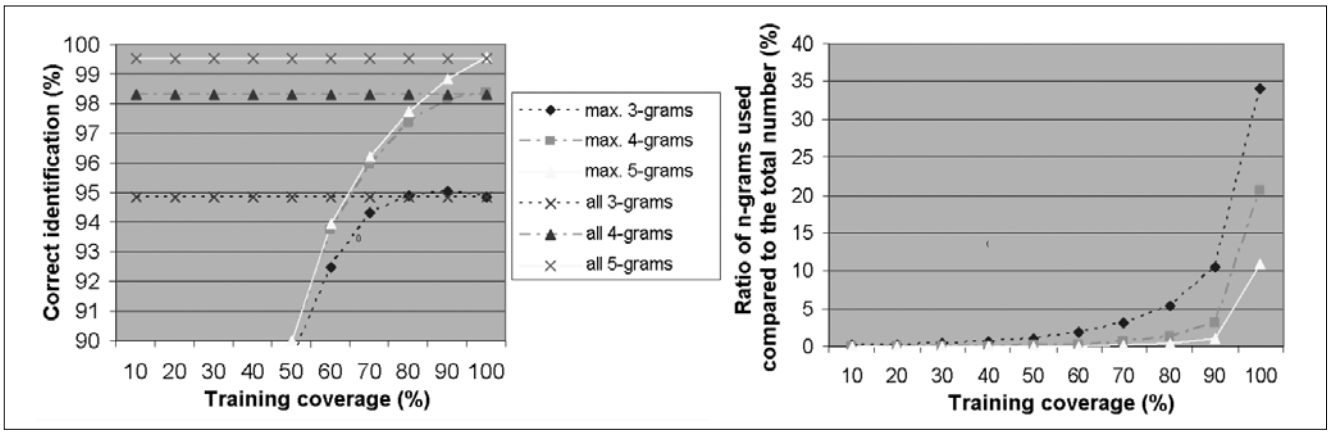


Figure 3.
Comparison of the performance and database size of the methods on the training set using ML estimation and either fixed or variable length character context

the incorrect word forms, but do not spoil the n-gram statistics very much. Another problem is data sparseness (i.e. you practically cannot collect sufficient data to have distribution information about all items that can occur; you will always encounter some that was not seen in the training corpus). This problem will be present to a much lesser degree than when working with words as the number of n-grams in a word is proportional to the square of the word length. A characteristic example of this approach is the method of Canvar and Trenkle [13].

2.7. Markov model

We know that we can get the exact word probability for a word of *l* characters using the chain rule of probability:

$$\mathrm{P}(\mathrm{word}\,|\,\mathrm{language}) = \prod_{i=1}^{l+1} \mathrm{P}(c_i \,|\, c_0 \ldots c_{i-1}, \mathrm{language}) \quad (1)$$

where $c_1,\ldots,c_l$ are the characters of the word, $c_i, i \leq 0$ and c_{l+1} are special word-starting and word-ending characters.

This probability can be approximated using a Markov model, i.e. assuming this is a random process where the probability distribution of the next character depends on the current state only. Traditionally the state is defined to consist of the previous *n-1* characters for LID (2):

$$\mathrm{P}(\mathrm{word}\,|\,\mathrm{language}) \approx \prod_{i=1}^{l+1} \mathrm{P}(c_i \,|\, c_{i-n+1} \ldots c_{i-1}, \mathrm{language})$$

We can approximate the conditional probability of characters after a given context using the ML (Maximum Likelihood) estimation:

$$\mathrm{P}(c_i \,|\, c_{i-n+1} \ldots c_{i-1}) \approx \frac{\#\, c_{i-n+1} \ldots c_{i-1} c_i}{\#\, c_{i-n+1} \ldots c_{i-1}} \quad (3)$$

Since we can expect that previously unseen n-grams will occur no matter what the size of the training set is, the use of some kind of smoothing technique is inevitable. The choice of this considerably determines the quality of the approximation.

The probability of the word in the language can in turn be used to estimate probability of language:

$$\mathrm{P}(\mathrm{language}\,|\,\mathrm{word}) = \frac{\mathrm{P}(\mathrm{word}\,|\,\mathrm{language}) \cdot \mathrm{P}(\mathrm{language})}{\mathrm{P}(\mathrm{word})} \quad (4)$$

When determining the most probable language, we can ignore the denominator (as this is characteristic of the input text and not of the language). We can regard the probability of the language as a constant value or one changing dynamically based on the context (5):

$$\mathrm{language} = \arg\max_{\mathrm{language}} \mathrm{P}(\mathrm{word}\,|\,\mathrm{language}) \cdot \mathrm{P}(\mathrm{language})$$

In theory the Markov model estimation in combination with a good smoothing technique can give arbitrarily good estimation if *n* is large enough. (If *n* is the maximum word length, we get the chain rule probabilities, i.e. the exact word probabilities for the training set.) But in practice generally *n*=2 (bigrams) or *n*=3 (trigrams) are used for two main reasons: because of the data sparseness problem and because larger *n*'s would need significant storage capacity.

Unfortunately such lengths do not allow for correct classification when deciding on multiple languages, as shown in *Figure 3* for the training set described in the first line of *Table 1*; these values are the theoretical limit for correct identification. As we can see, when training and testing on the same set, the correct identification rate does not reach 100% even for 5-grams but this would need a rather large database, furthermore it does not yet contain the smoothing inevitable for unknown texts. Using the method proposed in section 3., the size of the database is 10% to 35% of the previous with similar identification rates, plus it gives slightly better results when training for 100% of the training corpus.

Training (words)	Languages	Database	Training, word	Test, word	Test, sentence
2-9 million words	3	54 Kbytes	99,6%	94,2%	98,5-99,5%
600-700 words	3	7,4 Kbytes	95,5-97,8%	79,6-87,4%	91,7-97,2%
500-1700 words	77	5,4 Mbytes	70,0-99,8%	30,1-59,6%	71,0-84,0%

Table 1.
Results for different training sets with an independent test set for three languages, for word and sentence level identification

2.8. Summary

In this section we overviewed some of the methods used for language identification from text. The two main groups thereof are detailed analysis and statistical methods.

We can conclude that the purely statistical methods at present do not yield precise language identification on short texts in general; therefore they are not trustworthy enough for word level classification. Moreover, those that compare the document to so-called "language profiles" previously generated from training text corpora for the languages often need significant computing power in the identification phase also. In contrast, detailed morphological analysis is hard to accomplish, especially for a large number of languages, and the necessary processing power may be too high for some applications.

3. The proposed method

The method described below was first devised for the task of language identification from text; therefore we shall discuss it from this perspective. But by substituting "class" for "language" and "text unit" for "word", you can read it as the description of a general text labeling method.

3.1. Basic principle

Our aim was to create a method that has the ability to identify correctly even short sentences (down to one word), and can be held in hand in the sense that the system can be trained to correctly identify any required input, while it retains its ability to generalize, i.e. to identify unseen words correctly based on similarity to known training words. We also wanted to restrain the size of the resulting database.

A suitable way for this is to approximate P(word | language) with a precision set by a predefined criterion - e.g. with enough precision for the calculated probabilities to be the greatest for the language whose probability is greatest based on the training set. Another element of our method is to calculate language probability for every word based on its context and to decide on the language according to equation (5).

This approach makes it theoretically possible that we get correct language identification on word level, even in the case of homomorphs (in our case word forms belonging to several languages at the same time), furthermore that inserted foreign words can be labeled with their real language, rather than with the naive approach that deterministically decides based on the context. If one decision is needed for the whole piece of text (e.g. for a sentence), we can decide using the language labels determined for words, e.g. using the principle of majority voting.

This approach retains the advantage of the word-based method, i.e. it can be fully controlled, and extends it with generalizing abilities, allowing for correct word level identification. Using an appropriate probability estimation technique we may be able to estimate this probability for previously unseen words based on the spelling of known words. The key for the success of this method is the estimation of the conditional probabilities and language probabilities with enough precision.

3.2. Estimating conditional probabilities

The method we worked out for estimating P(word | language) relies on n-grams of variable length. While traditional LID solutions with Markov-model use a fixed length character context for estimating the conditional probability of the next character, in the proposed method we use variable length character context. (6)

$$\mathrm{P}(\mathrm{word}\,|\,\mathrm{language}) \approx \prod_{i=1}^{l+1} \mathrm{P}(c_i \,|\, c_{i-n_i+1}\ldots c_{i-1}, \mathrm{language}), n_i \geq 0$$

We determine the n_i lengths in the course of a training process. Training starts out with a 0 length character context for every character (this is the occurrence probability of the character), then increases this length in certain contexts for attaining the predefined probability estimation criterion, which can be e.g. the correct identification of a certain percentage of the most frequent words of the training set. Continuing the process without a limit yields the chain rule, and through this the word-probability (assuming an appropriate smoothing technique), therefore the training process converges to the correct identification for any training set. Using a larger database with longer n-grams will give more accurate identification results. Therefore the method is scalable, as one of the database size-desired identification rate-pair can be chosen freely.

Storing the conditional probabilities belonging to the n-gram contexts in a tree, we can view the method as training a kind of decision tree for the purpose of word probability estimation. The direction for expanding the tree is decided based on the "usefulness" of the expansion in regard to the estimation criterion. We worked with several such usefulness functions, examining them with regard to the correct identification rate on the training set, the results on an independent test set (i.e. generalizing ability), and how concise the resulting database is. We did not characterize conciseness simply with size – as it is not indifferent what identification rate a smaller database produces – but with the identification rate/size quotient, which we called the performance of the LID-database. We can see the best usefulness functions and the graphs that describe them in *Figure 4*. As we can see, different usefulness functions are needed for the best generalization ability and for the most concise database.

Another novelty is that instead of observing languages separately, we endeavor to estimate the conditional probabilities difference between languages with enough precision. From this we expect smaller database size, since this way we make the algorithm "concentrate" on the features that distinguish the languages.

In so doing the algorithm is not only scalable, but automatically scales itself to the complexity of the problem. E.g. if two languages have different character sets, then even if we aim at 100% correct identification, training stops at using unigrams (1 character long n-grams).

3.3. Using language probabilities

3.3.1 The notion of "language probabilities"

By "language probabilities" we mean the probability that a word with a certain language will occur in a given context; we use this in (5) in the place of P(language). The features that can be used in the estimation of this probability in our solution are the language of the surrounding words and the punctuation between them. We have chosen to include these among possible questions to ask because these seem to play a role in forming the human interpretation, but a different, smaller or broader set of questions is also possible.

We can view the modeling of language probabilities as training a decision tree for storing conditional probabilities similarly to the type given in 3.2., but it is different from that (and the generally used methods [15,16]) in that we do not only use the previous words, but the following ones also.

The difficulty in this approach is partly that we presume to know the language of the words around the word in question, which is not fulfilled in practice, and partly that, because of the large number of possible questions that guide the expansion of the decision tree during training, the number of possible alternatives is huge. We shall discuss the solution of these in the next section.

3.3.2 Finding the most probable label sequence

The first difficulty is a mathematically solvable problem, although the efficient realization of the calculation is not trivial. The task is to find the label sequence that maximizes the probability on the sentence made up of N words.

$$\{\mathrm{lang}_{l_i} \mid i \in [1..N]\} = \arg\max \prod_{i=1}^{N} \mathrm{P}(\mathrm{lang}_{l_i} \mid \mathrm{word}_i) = \tag{7}$$

$$= \arg\max \prod_{i=1}^{N} \frac{\mathrm{P}(\mathrm{word}_i \mid \mathrm{lang}_{l_i}) \cdot \mathrm{P}(\mathrm{lang}_{l_i})}{\mathrm{P}(\mathrm{word}_i)}$$

We can disregard the $\mathrm{P}(\mathrm{word}_i)$ coefficient here also, since it does not depend on the language, therefore it does not affect the result.

Finding the most probable label sequence with exhaustive search with L possible labels would mean L^N calculation steps, which defines a search space that grows exponentially with the length of the sentence; there are not many practical applications that can allow this. Therefore we need to find a method that approximates the optimal solution in some way.

The approximation used currently in our system is the following: first we calculate a label sequence with uniform language probability, and then using the language context received this way we recalculate the labels for the whole sequence from left to right. If a label is modified, then we recalculate the labels to its left that could be affected by the change, but (in order to avoid iteration) modify them only if the probability calculated for the new language label is higher than for the previous one. The algorithm can be improved, e.g. by using simulated annealing (allowing label modifications that do not immediately result in probability growth to a lesser and lesser degree).

3.3.3 Using rule templates

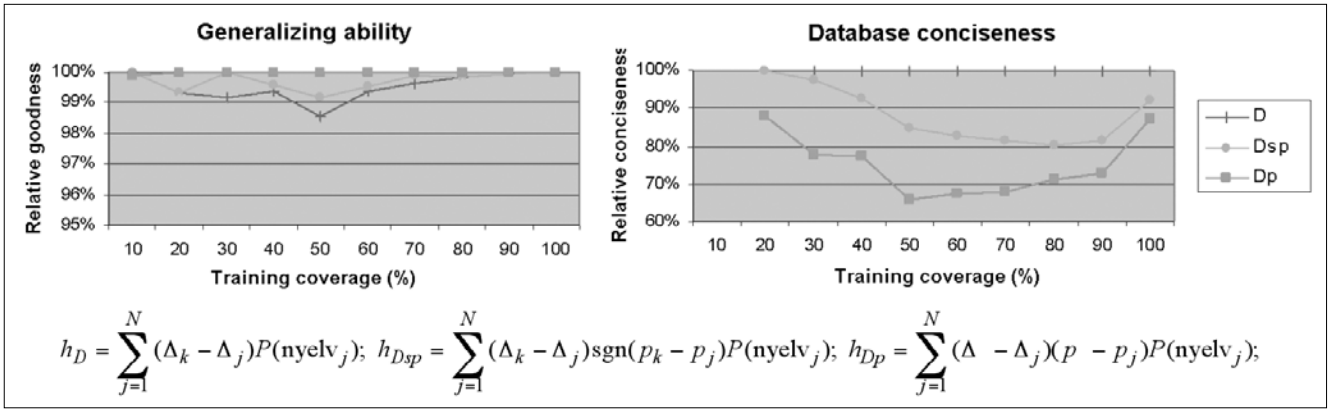
To avoid the second difficulty, in our sample system we use so-called rule templates, which gives a way for using linguistic knowledge and heuristics, and largely diminishes the number of alternatives to examine.

The rule templates can have the following form (in BNF notation):

<template> ::= { <label-description> [<separator-description>] }
<label-description> ::= L{[<label ID >][?|*] | "<label name>"}
<separator-description> ::= S{[<separator ID>][*] | "<separator>"}

The label is language in this case. Labels with the same label ID must be equal; the same is true for separators. The descriptions marked with an asterisk indicate items whose different values do not create independent rules, while those that are not marked create a separate rule for every different value of the label or separator.

Figure 4. Characterizing usefulness functions
(values are compared to the best result, training for diverse percentages of most frequent words);
N is the number of languages, k is the index of the real language of the word, p is the occurrence probability of the n-gram to add, Δ is the change in the probability value arising because of adding the n-gram.



$$h_D = \sum_{j=1}^{N} (\Delta_k - \Delta_j) P(\mathrm{nyelv}_j);\ h_{Dsp} = \sum_{j=1}^{N} (\Delta_k - \Delta_j) \mathrm{sgn}(p_k - p_j) P(\mathrm{nyelv}_j);\ h_{Dp} = \sum_{j=1}^{N} (\Delta_k - \Delta_j)(p_k - p_j) P(\mathrm{nyelv}_j);$$

The label-description marked with a question mark ("?") is the one for which we are observing the occurrence probability of the different labels. When creating the rules to use, we match the rule templates onto each word of the training set at this point, and where the template can be matched against it, we create a rule that is filled with the specific values found in the text. These rules will contain the occurrence probability of the labels at the end of the process (calculated from the number of places the rule could be used, and number of occurrences of the label types).

If several rules can be applied at a position with different number of conditions, then we use the one that has more conditions (following the principle of the decision tree). You can see examples for rule templates and the rules created from them later in *Figure 5*, for the topic of language identification.

4. Application to language identification from text

4.1. Assembling a training corpus

We can say that at present no large text corpus is available that is labeled with the correct language label on word level, moreover it is hard to assemble corpora that are actually monolingual (because of the foreign names, expressions and loanwords that are present in every language, plus foreign language parts or complete foreign texts find their way even into well-known monolingual corpora such as Project Gutenberg). Therefore assembling training sets for LID is also difficult despite the fact that vast amounts of text can be downloaded from the internet for practically any language, but with the aforementioned mixed nature.

A way to solve the problem can be to train the LID system for several languages assuming the collected texts to be monolingual (or at least to have the nominal language in majority), and then to automatically label the texts using the resulting LID system. After omitting complete texts or sentences that are identified as clearly differing from the language of the training corpus, we can repeat the training process, this time with a cleaner text that better approaches monolinguality, until no improvement can be realized.

In case of a corpus that contains relatively short texts (e.g. the archive of a newspaper), the headers and footers, which are usually present in texts and are basically the same in each one, can notably distort word and n-gram statistics, as these multiply the number of occurrences of some (perhaps otherwise rare) words or expressions. For the theoretically correct operation it is worth removing these from longer pieces of text also. This also helps in cleaning the corpus from foreign language parts, as the header and footer is normally written in the language of the corpus even for incidentally retained foreign language documents.

We need to pay attention to the fact that the texts collected for a language may be coded with diverse character sets. In order to handle these correctly, we need to know the encoding of these texts. If we can expect to know the character set of the input to be identified, then we just need to convert the texts to a common coding (e.g. Unicode). If it is not known, we can train the system with texts in different encodings at the same time, or we can label the texts of the training sets with the same language but different encodings with a name that includes language and character set, so that we can identify these two features in one step.

Another problem is that in certain application areas (e.g. when working with SMS or e-mail messages) the diacritics, which are used in Hungarian and many other languages, may be missing from the letters, which can hinder language identification if we do not cope with this. Possible approaches are to use a training set that contains both kinds of text (i.e. with and without diacritics); or to convert the training set and the text to be identified to a smaller common character set (the one without diacritics) and during identification to modify the language probability based on the original character set of the word [11].

4.2. The test corpus used

We performed several test with training and test corpora of different sizes. First we trained the system for three languages (English, German, Hungarian) on large corpora (British National Corpus, Project Gutenberg DE, Hungarian Electronic Library), without cleaning them of the foreign language parts, to correctly identify 90% of the most frequent words. We did the testing on independent test sets (Project Gutenberg, online Hungarian Newspapers).

We also did the training on small texts (5 kilobyte) belonging to 77 languages that are used in an implementation (http://odur.let.rug.nl/~vannoord/TextCat/Demo/) of the method introduced in [13].

4.3. Results without using language probabilities

The results for correct identification can be seen in *Table 1*. A manual overview of the classified texts showed that in the case given in the first line the texts classified to a different language often really did not belong to the language of their groups, or had mixed language; additionally we found that for precise word level operation we need to recognize certain expressions beforehand that can be viewed as language-independent (regarding their format, not their pronunciation), for example Roman numbers, internet and e-mail addresses, dates, international words (e.g."tel.","fax."), abbreviations, expressions containing measurements (e.g."2 cal").

4.4. Results using language probabilities

We also examined the effect of taking into account language probability calculated from the environment of words. For this we first labeled the training set using the LID database created for the third line of Table 1. We used this one because the small training set and the resulting relatively poorer identification rate means

rule templates
L1* S* L? S* L1*
L1* S* L? S* L2*

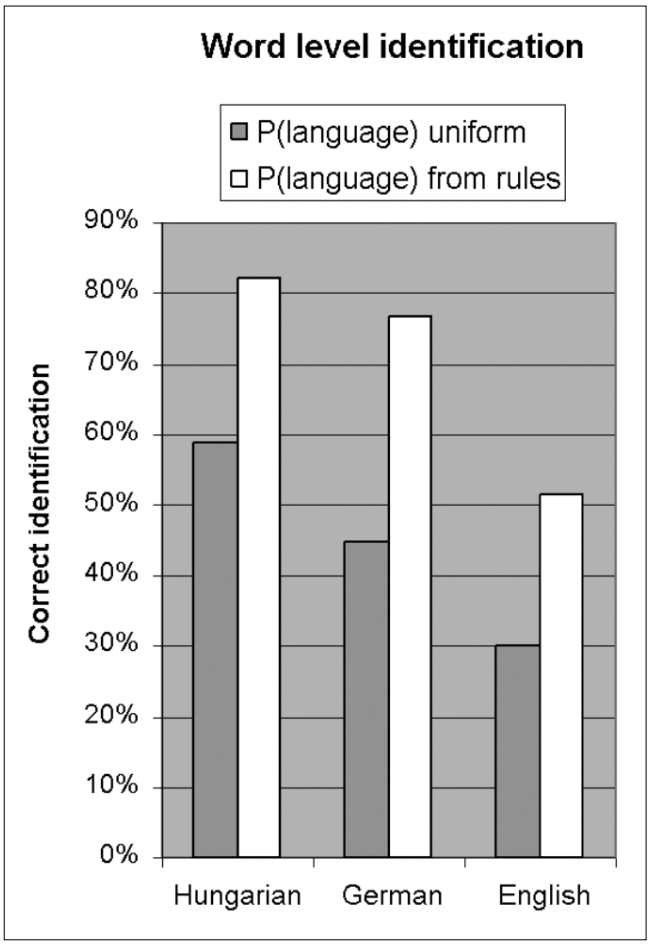
rules generated for the training corpus using the LID database for Table 1 Line 3
("": number of times the template could be applied;
"L": number of other labels – neither L1, nor L2)
L1* S* L? S* L1*; { "": 13300305, "L1":13230575, "L":69730 }
L1* S* L? S* L2*; { "": 20188476, "L1":16625250, "L2":16625298, "L":168503 }

Figure 5. The rule templates used and the rules generated by applying them

a greater challenge for our method. Then we applied the rule templates shown in *Figure 5* onto the labeled text and we arrived at the rules shown below the templates. Using these rules to estimate P(language), we relabeled the texts with the technique described in 3.3.2. With this, the correct identification rate on the German test corpus increased from the former 45% to 65% (assuming the text to be purely German), then iterating the rule generation – labeling steps it increased to 70%, and further on to 72%.

The improvement is shown in *Figure 6* for three languages. The bulk of errors for English (10% and 14%) resulted from erroneously deciding on the very similar Scottish language. It is noteworthy that such improvement was attained despite that we did not use the actual probability of each language but a rough approximation of the probability of neighboring words having the same language. Based on the figures in Table 1, such word level identification rate allows for 80% to 90% correct sentence-level identification using majority voting. Because of the availability of a great amount of text in various languages, e.g. through the internet, we are not obliged to use such small training sets. Therefore in practical applications we can expect an even better rate of correct identification that the one shown in the first line of Table 1, very close to 100%, by using language probabilities in the described way.

Figure 6. Examples for the improvement attained using language probabilities



5. Conclusions

In this article we sketched the significance of TTS systems in telecommunications and pointed out the importance of automatic labeling methods, like e.g. language identification and part-of-speech tagging.

We overviewed some techniques used for language identification. To address some of their weaknesses we introduced a new method that work with two kinds of conditional probabilities, estimating their values using decision trees. We demonstrated its effectiveness on the task of language identification from text. The results justify the viability of the approach. We expect it to be applicable for other topics also, e.g. for an approximation of part-of-speech tagging without using a morphological analyzer.

References

[1] Németh, G., Zainkó, Cs., Fekete, L., Olaszy, G., Endrédi, G., Olaszi, P., Kiss, G., Kis, P.: "The Design, Implementation and Operation of a Hungarian E-mail Reader", International Journal of Speech Technology, Volume 3, Numbers 3/4, December 2000, pp.217–236.

[2] G. Németh, Cs. Zainkó, G. Kiss, M. Fék, G. Olaszy and G. Gordos: "Language Processing for Name and Address Reading in Hungarian", Proc. of IEEE Natural Language Processing and Knowledge Engineering Workshop 2003, Oct. 26-29, Beijing, China. pp.238–243.

[3] Pfister, B., Romsdorfer, H.:
"Mixed-lingual text analysis for polyglot TTS synthesis",
Proc. of Eurospeech 2003,
pp.2037–2040.
[4] Halácsy, P., Kornai, A., Németh, L., Rung, L.,
Szakadát, I., Trón, V.:
"Creating open language resources for Hungarian",
Proc. of LREC 2004,
pp.203–210.
[5] Marcadet, J. C., Fischer, V., Waast-Richard, C.:
"A Transformation-based learning approach to
language identification for mixed-lingual
text-to-speech synthesis", Proc. of Eurospeech 2005,
pp.2249–2252.
[6] Prószéky, G.:
"Humor: a Morphological System for Corpus Analysis.
Language Resources for Language Technology",
Proc. of the First European TELRI Seminar,
Tihany, Hungary, 1995.
pp.149–158.
[7] Németh, L., Halácsy, P., Kornai, A., Trón, V.:
"Nyílt forráskódú morfológiai elemző"
(Open-source Morphological Analyser)
In: Csendes D, Alexin Z. (eds.):
II. Magyar Számítógépes Nyelvészeti Konferencia
(Hungarian Conference on Computer Linguistics),
Szeged, 2004.
pp.163–171.
[8] Ted Dunning:
"Statistical Identification of Languages",
Computing Research Laboratory,
New Mexico State University, 1994.
[9] G. Németh, Cs. Zainko:
"Multilingual statistical text analysis,
Zipf's law and Hungarian Speech Generation",
Acta Linguistica Hungarica, Vol. 49 (3-4), 2002.
pp.385–405.
[10] Prager, J. M. Linguini:
"Language Identification for Multilingual Documents",
Proc. of the 32nd Annual Hawaii International
Conference on System Sciences, Vol. 1., 1999.
pp.2035.
[11] Tian, J., Suontausta, J.:
"Scalable neural network based language
identification from written text",
Proc. of IEEE International Conference on Acoustics,
Speech, and Signal Processing, Vol. 1, 2003.
pp.48–51.
[12] Häkkinen, J., Tian, J.:
"N-gram and Decision Tree-based Language
Identification for Written Words",
Proc. of IEEE Workshop on Automatic Speech
Recognition and Understanding,
Madonna di Campiglio Trento, Italy, 2001.
[13] W. B. Canvar, J. M. Trenkle:
"N-gram based Text Categorization",
Symposium on Document Analysis and Information
Retrieval, University of Nevada, Las Vegas, 1994.
pp.161–176.
[14] Sproat, R., Riley, M.:
"Compilation of weighted finite state transducers
from decision trees"
In: Association for Computational Linguistics,
34th Annual Meeting, Santa Cruz, Canada, 1996.
pp.215–222.
[15] Suendermann, D., Ney, H.:
"Synther – a New M-Gram POS Tagger",
Proc. of the NLP-KE 2003,
International Conference on Natural Language
Processing and Knowledge Engineering,
Beijing, China, 2003.
[16] Halácsy P, Kornai A., Varga D.:
"Morfológiai egyértelműsítés
maximum entrópia módszerrel"
(Morphological Disambiguation with
Maximum Entropy Method),
Magyar Számítógépes Nyelvészeti Konferencia
(Hungarian Conference on Computer Linguistics),
2005.

Performance evaluation of Proxy Cache Servers

TAMÁS BÉRCZES
IFSZ Kft., Debrecen; berczes.tamas@ifsz.hu
JÁNOS SZTRIK
University of Debrecen, Dept. of Informatics Systems and Networks
jsztrik@inf.unideb.hu

Reviewed

Keywords: queueing network, Proxy Cache Server (PCS), performance models

Due to the rapid growth of internet users, the Web traffic also grows very fast. The primary aim of the present paper is to modify the performance model of Bose and Cheng to a more realistic case when external arrivals are also allowed to the remote Web servers and the Web servers have limited buffer capacity. We analyze how many parameters affect the performance of a Proxy Cache Server (PCS). Numerical results are obtained for the overall response time with and without a PCS. The numerics show that the benefit of a PCS depends on various factors. It is noticed that by increasing the cache hit rate or the external arrival rates the overall response time is smaller in case of installing a PCS.

1. Introduction

The World Wide Web (WWW) can give a quick and easy access to a large number of web servers where users can find all kind of information, documents and multimedia files. From the user's point of view it does not matter whether the requested files are on the firm's computer or on the other side of the world. The usage of the web has been growing very fast. The number of internet users increased from 474 million in 2001 to 590 million in 2002, and the forecast for 2006 is 948 million users. According to the facts, that in 1996 the number of users was only 627.000, the growth is rapid and we can justify and exponential grows in traffic, too. The users want to get a high quality service and modest response time.

The answer from the remote web server to the client often takes a long time. One of the problems is that the same copy of the file can be claimed by other users at the same time. Because of this situation, identical copies of many files pass through the same network links, resulting in an increased response time. A natural solution to avoid this situation is to store this information. In general, caching can be implemented at browser software; the originating Web sites; and the boundary between the local area network and the Internet. Browser cache are inefficient since they cache for only one user.

The caching at the Web sites can improve performance, although the requested files are still subject to delivery through the Internet. It has been suggested that the greatest improvement in response time for corporations will come from installing a Proxy Cache Server (PCS) at the boundary between the local area network and the Internet. Requested documents can be delivered directly from the web server or through a proxy cache server. A PCS has the same functionality as a web server when looked at from the client and the same functionality as a client when looked at from a web server.

The primary function of a proxy cache server is to store documents close to the users to avoid retrieving the same document several times over the same connection. In this paper, a modification of the performance model of Bose and Cheng [1] is given to deal with a more realistic case when external arrivals are also allowed to the remote Web servers and the Web servers have a limited buffer. For the easier understanding of the basic model and comparisons we follow the structure of the cited work.

In Section 2, we construct a queuing network model to study the dynamics of installing a PCS. Overall response-time formulas are developed for both the case with and without a PCS.

In Section 3, numerical experiments are conducted to examine the response time behavior of the PCS with respect to various parameters of the model.

Concluding remarks can be found in Section 4.

Table 1.

λ	arrival rate from the PCS
Λ	external arrival rate
F	average file size (in byte)
p	cache hit rate probability
B_{xc}	PCS output buffer (in byte)
B_s	Web server output buffer (in byte)
I_{xc}	lookup time of the PCS (in second)
Y_{xc}	static server time of the PCS (in second)
R_{xc}	dynamic server time of the PCS (in byte/sec)
I_s	lookup time of the Web server (in second)
Y_s	static server time of the Web server (in second)
R_s	dynamic server time of the Web server (in byte/sec)
N_c	client network bandwidth (in bit/sec)
N_s	server network bandwidth (in bit/sec)
K	The buffer size of the Web server (in requests)

2. An analytical model of PCS traffic

In this section, we briefly describe the mathematical model with the suggested modifications. Using proxy cache server, if any information or file is requested to be downloaded, first it is checked whether the document exists on the proxy cache server. (We denote the probability of this existence by *p)*. If the document can be found on the PCS then its copy is immediately transferred to the user. In the opposite case the request will be sent to the remote Web server. After the requested document arrived to the PCS then the copy of it is delivered to the user. The advantage of a PCS depends on several factors: The probability of the "cache hit rate" of the PCS, the speed of the PCS, the bandwidth of the firm's and the remote network and the speed of the remote web server [1].

Fig.1. illustrates the path of a request in the modified model starting from the user and finishing with the return of the answer to the user. The notations used in this model are collected in *Table 1*.

We assume that the requests of the PCS users arrive according to a Poisson process with rate λ and the external arrivals at the remote web server form a Poisson process with rate Λ. Let F be the average of the requested file size. Now we define the variables in the figure:

$$\lambda_1 = p^*\lambda; \quad (1)$$

$$\lambda_2 = (1-p)^*\lambda; \quad (2)$$

$$\lambda_3 = \lambda_2 + \Lambda; \quad (3)$$

The solid line in Fig 1. (λ_1) represents the traffic, when the requested file is available on the PCS and can be delivered directly to the user. The λ_2 traffic depicted by dotted line, represents those requests which could not be served by the PCS, therefore these requests must be delivered from the remote web server. Naturally the web server serves not only the requests of the studied PCS but it also serves requests of other external users. Let λ_3 denote the intensity of the overall requests arriving to the remote Web server. The traffic undergoes the process of initial handshaking to establish a onetime TCP connection [7,1]. We denote by I_s this initial setup.

According to [1], "The remote Web server performance is characterized by the capacity of its output buffer B_s, the static server time Y_s, and the dynamic server rate R_s." In our model, we assume that the Web server has a buffer of capacity K. Let P_b be the probability that a request will be denied by the Web server. As it is well-known from basic queueing theory, the blocking probability P_b for the M/M/1/K queueing system:

$$P_b = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}} \quad (4)$$

where

$$\rho = \frac{\lambda_3 F\left(Y_s R_s + B_s\right)}{R_s B_s}. \quad (5)$$

Now we can see that the requests arrive to the buffer of the Web server according to a Poisson process with rate

$$\lambda_4 = (1-P_b)^*\lambda_3 \quad (6)$$

The performance of the firm's PCS is characterized by the parameters B_{xc}, Y_{xc} and R_{xc}.

If the size of the requested file is greater than the Web server's output buffer, it will start a looping process until the delivery of all requested file's is completed.

Let

$$q = \min\left(1, \frac{B_s}{F}\right) \quad (7)$$

be the probability that the desired file can be delivered at the first attempt. Let λ_4' be the rate of the re-

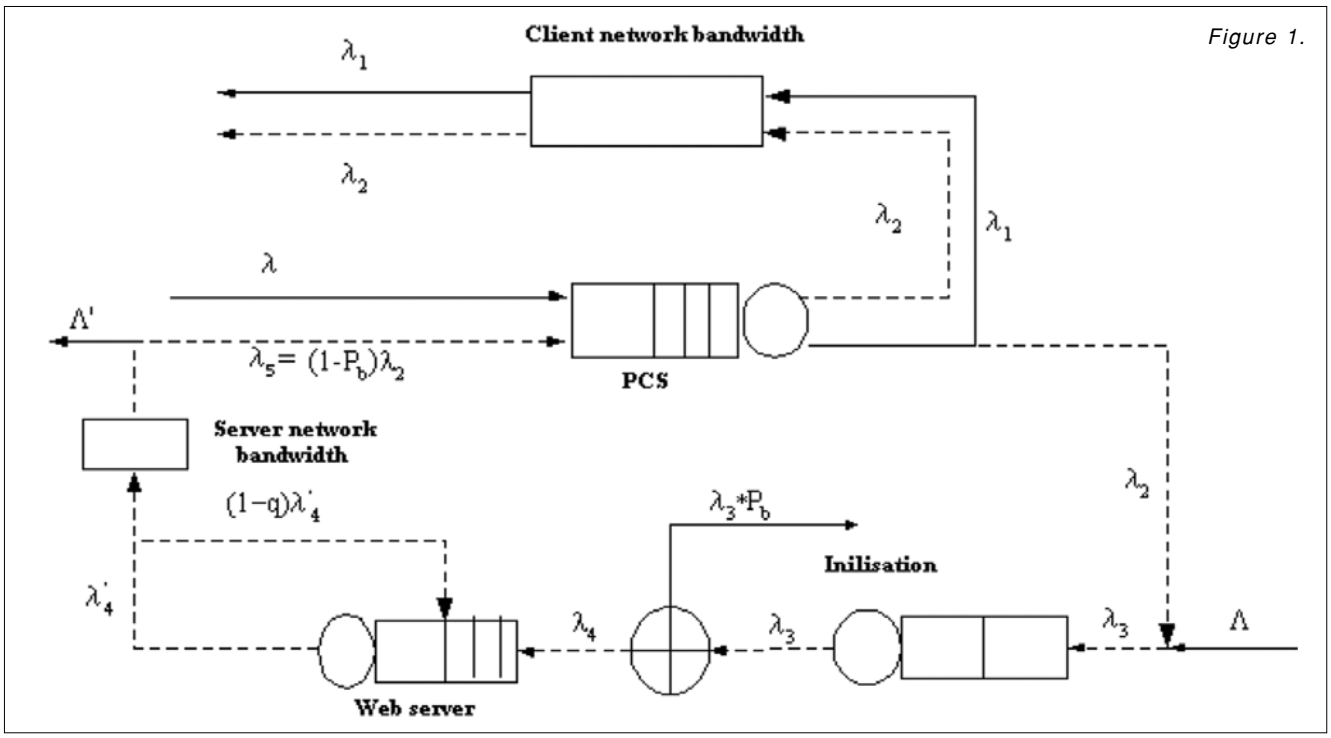


Figure 1.

quests arriving at the Web service considering the looping process. According to the conditions of equilibrium and the flow balance theory of queueing networks

$$\lambda_4 = q\,\lambda_4' \tag{8}$$

Then, we get the overall response time:

$$T_{xc} = \frac{1}{\frac{1}{I_{xc}} - \lambda} + p\left(\frac{1}{\frac{B_{xc}}{F\left(Y_{xc} + \frac{B_{xc}}{R_{xc}}\right)} - \lambda_1} + \frac{F}{N_c}\right) + \tag{9}$$

$$+(1-p)\left(\frac{1}{\frac{1}{I_s} - \lambda_3} + \frac{1}{\frac{B_s}{F\left(Y_s + \frac{B_s}{R_s}\right)} - \frac{\lambda_4}{q}} + \frac{F}{N_s} + \frac{1}{\frac{B_{xc}}{F\left(Y_{xc} + \frac{B_{xc}}{R_{xc}}\right)} - \lambda_5} + \frac{F}{N_c}\right)$$

The response time T_{xc} consists of three terms. The first term is the time to check whether the requested file is on the PCS or not. This is derived from the waiting time in an M/M/1 queueing system where the arrivals form a Poisson process with rate λ, the service rate is $1/I_{xc}$.

The second term is the response time in the case if the requested document exists on the PCS, the probability of which is p. The first item in this term is the waiting time on the PCS, where

the numerator $\frac{B_{xc}}{F*\left(Y_{xc} + \frac{B_{xc}}{R_{xc}}\right)}$ is the "service demand".

The second item in the second term corresponds to the required time for content to travel through the client network bandwidth. The third term is the response time when the requested file does not exist on the PCS. The probability of that event is (1-p). This term consists of three terms too. The first item is the expected one-time initialization time of the TCP connection between the PCS and the remote web server. The second item is the waiting time of the queueing system on the remote Web server, where $\lambda_4' = \lambda_4/q$ and F / N_S is the expected time of transferring the requested documents on the network of the bandwidth. The third term is the waiting time of the PCS when the copy of the requested document is transfered to the user.

When there is no PCS, the overall response time T, is given by the same arguments:

$$T = \frac{1}{\frac{1}{I_s} - (\lambda + \Lambda)} + \frac{1}{\frac{B_s}{F\left(Y_s + \frac{B_s}{R_s}\right)} - \frac{(1-P_b)(\lambda+\Lambda)}{q}} + \frac{F}{N_s} + \frac{F}{N_c} \tag{10}$$

3. Numerical results

For the numerical explorations the corresponding parameters of Cheng and Bose [1] are used. The value of the other parameters for numerical calculations are:

$I_s = I_{xc} = 0.004$ sec, $B_s = B_{xc} = 2000$ bytes,
$Y_s = Y_{xc} = 0.000016$ sec, $R_s = R_{xc} = 1250$ Mbyte/s,
$N_s = 1544$ Kbit/s and $N_c = 128$ Kbit/s.

In figures the dotted line plot the case with a PCS and the normal line depicts the case without a PCS.

3.1. Effect of arrival rate

In *Fig. 2.* the response time is depicted as a function of the arrival rate. In this figure the external arrival rate is 100 requests/s and the cache hit rate is 0.1. We see that in this case the response time will be greater when we install a PCS. In *Fig. 3.* we use the same parameters, but the cache hit rate is 0.25. In this case the response time is the same with and without a PCS. In *Fig. 4.* we use a higher external arrival rate ($\Lambda = 150$) with a smaller cache hit rate ($p = 0.1$). When λ is smaller than 70 requests/s the response time is larger with a PCS than without a PCS.

When we use a higher cache hit rate with a higher external arrival rate (*Fig. 5.*, $p = 0.25$, $\Lambda = 150$) the efficiency of PCS is clear. In this case the response time with a PCS will be smaller than the response time without a PCS for any value of the arrival rate. So, we can see that the performance of a PCS depends on a high scale of the firms behaviour, but when the intensity of the requests from the firm is greater than 70, and the external arrival rate is 150 requests/s then it is enough a small cache hit rate to access a smaller response time.

3.2. Effect of external arrival rate

Now we investigate the effect of the external arrival rate. In *Fig. 6.* the arrival rate from the PCS is 20 requests/s, the requested file size is 5000 byte and the cache hit rate is 0.1. We can see that with these parameters installing a PCS we get a higher response time. In *Fig. 7.* we modified only the cache hit rate probability to 0.25. In this situation when we have more than 140 external requests/s then the response time with PCS is smaller than without a PCS. When the cache hit rate probability is smaller ($p = 0.1$) and $\lambda = 70$ requests/s *(Fig. 8.)* then the response time with a PCS is smaller than without, when we use a greater external arrival rate ($\Lambda > 150$). When we use a higher cache hit rate probability (*Fig. 9.*, $p = 0.4$) then the response time with a PCS is smaller, independently of the external arrivals.

Observing Fig. 6-9., we can find that in general the response time with and without a PCS increases when the external arrival rate increases. When the arrival rate of the studied firm is modest (20 req./s) then the benefit of the PCS is visible when the external arrival rates are bigger or when the cache hit rate probability is higher.

4. Conclusions

We modified the queueing network model of Bose and Cheng [1] to a more realistic case when external arrivals are allowed to the remote web server and the web server has limited buffer. To examine this model we conducted numerical experiments adapted to realistic parameters. We noticed that, when the arrival rate of requests increases, then the response times increase as well regardless of the existence of a PCS. But in contrast with [1] when external arrivals are allowed to the remote web server, the PCS was beneficial with a low

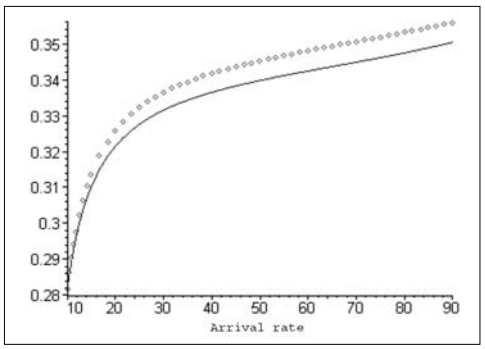


Figure 2.
p=0.1, F=5000 bytes, Λ=100, K=100

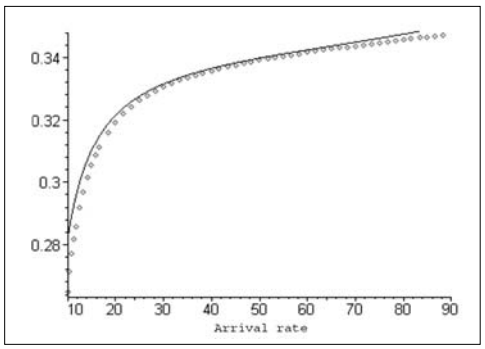


Figure 3.
p=0.25, F=5000 bytes, Λ=100, K=100

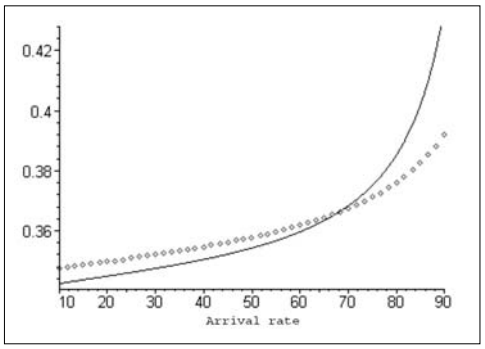


Figure 4.
p=0.1, F=5000 bytes, Λ=150, K=100

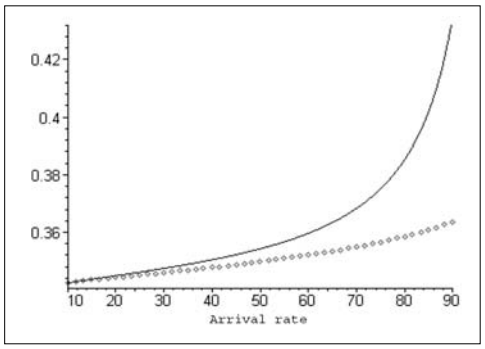


Figure 5.
p=0.25, F=5000 bytes, Λ=150, K=100

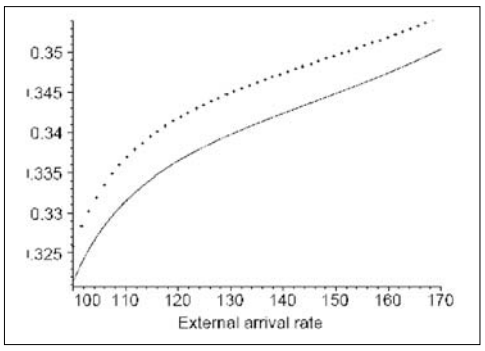


Figure 6.
λ=20, p=0.1, F=5000 bytes, K=100

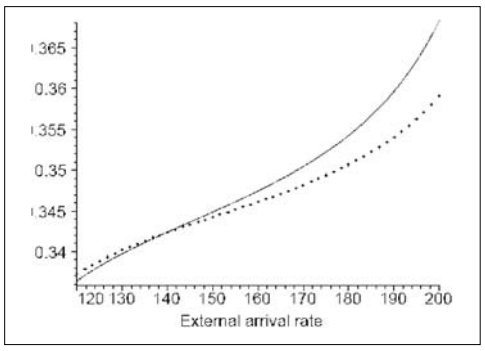


Figure 7.
λ=20, p=0.1, F=5000 bytes, K=100

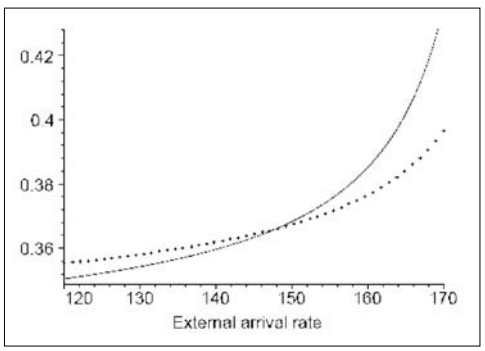


Figure 8.
λ=70, p=0.1, F=5000 bytes, K=100

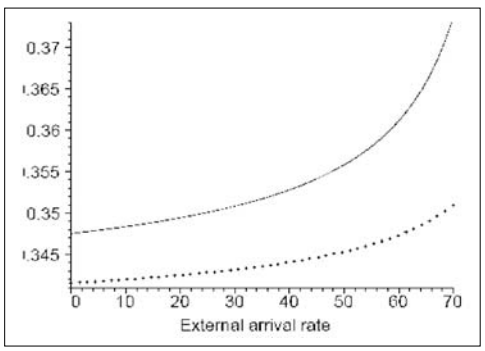


Figure 9.
λ=20, p=0.4, F=5000 bytes, K=100

traffic and a low cache hit rate. When we used a high arrival rate with a high cache hit rate probability, then the response time gap was more significant between the cases with and without a PCS.

To compare the two models we examined the effect of the external arrival rate. With low external arrival rate installing a PCS resulted higher response times. Increasing the external arrival rate, the difference between response time with and without a PCS was smaller and smaller until this difference vanished and the existence of a PCS resulted lower response times.

Examining numerical results it was clear that allowing external arrivals and limited buffer a more realistic model was obtained.

References

[1] Bose, I., Cheng, H.K.: Performance models of firms proxy cache server. Decision Support Systems and Electronic Commerce, 29 (2000), pp.45–57.

[2] CacheFlow Inc. (1999), CacheFlow White Papers Avaiable from http://cacheflow.com/technology/

[3] Menasce, D.A., Almeida, V.A.F.: Capacity Planning for Web Performance: Metric, Models and Methods. Prentice Hall (1998).

[4] L.P. Slothouber: A model of Web server performance. 5th International World Wide Web Conference, Paris, France (1996).

[5] C. Aggarwal, J.L. Wolf, P.S. Yu: Caching on the World Wide Web, IEEE Transactions on Knowledge and Data Engineering 11 (1999).

Automatic generation of platform-specific transformation

ÁKOS HORVÁTH, DÁNIEL VARRÓ, GERGELY VARRÓ*

Budapest University of Technology and Economics
Dept. of Measurement and Information Systems; ha442@hszk.bme.hu, varro@mit.bme.hu
** Dept. of Computer Science and Information Theory; gervarro@szit.bme.hu*

Reviewed

Keywords: meta-transformation, generic transformation, code generation

The current paper presents a new approach using generic and meta-transformations for generating platform-specific transformer plugins from model transformation specifications defined by a combination of graph transformation and abstract state machine rules (as used within the VIATRA2 framework). The essence of the approach is to store transformation rules as ordinary models in the model space, which can be processed later by the meta-transformations, which generates the Java transformer plugin. These meta rules highly rely on generic patterns (i.e. patterns with type parameters), which provide high-level reuse of basic transformation elements. Graph algorithms used for search plan generation are integrated as abstract state machines, while the final code generation step is carried out by code templates. As a result, the porting of a transformer plugin to a new underlying platform can be accelerated significantly.

1. Introduction

Nowadays, model-driven system development [3] (MDSD) is an emerging paradigm in software development. A main challenge for MDSD is accommodate to the accelerating changes of business and technology. Based on high-level model standards (such as the Unified Modeling Language, UML [12]), MDSD separates business and application logic from underlying platform technology.

Platform-independent models (PIM) capture the core business functionality independently from the underlying implementation technology, which are incorporated later on in platform-specific models (PSM). The source code of the system under design is generated afterwards from such platform-specific models. The success of the MDSD highly depends on automated model transformations (MT), which generate PSMs from PIMs, and executable source code from PSMs.

In MDSD, models are frequently captured by a graph structure, and the transformations are specified as graph transformations. Informally, a graph transformation (GT [11,7]) rule performs local manipulation on graph models by finding a matching of the pattern prescribed by its left-hand side (LHS) graph in the model, and changing it according to the right-hand side (RHS) graph.

The main objective of the VIATRA2 (VIsual Automated model TRAnsformations) framework developed at the Department of Measurement and Information Systems at Budapest University of Technology and Economics is to provide a general-purpose support for the entire life-cycle of engineering model transformations including the specification, design, execution, validation and maintenance of transformations within and between various modeling languages and domains. Since September 2005, VIATRA2 is part of the Eclipse Generative Modeling Tools subproject.

Advanced model transformation tools frequently aim at separating the design of a transformation from its execution by using high-level model transformation rules in design time and deriving executable platform-specific transformer plugins from these high level models. The role of design-time transformation frameworks (also called as platform independent transformers PIT) is to ease the development of model transformations, while the role of compiled standalone versions of a model transformation (Platform (lang.) specific transformers (PST)) in an underlying platform (e.g. Java) are more efficient from runtime performance aspects.

Code generators deriving the standalone transformers, are typically implemented in a standard programming language for specific model transformations, thus, it is difficult to reuse existing code generators to different platforms with conceptual similarities (e.g. from Java to Enterprise Java Beans) or to integrate them into other MT tools.

The current paper presents a new approach using generic and meta-transformations [14] for generating platform-specific transformer plugins from model transformation specifications defined by a combination of graph transformation and abstract state machine rules (as used within the VIATRA2 framework).

The essence of the approach is to store transformation rules as ordinary models in the model space, which can be processed later by the meta-transformations, which generates the standalone Java transformer plugin. These meta rules highly rely on generic patterns (i.e. patterns with type parameters), which provide high-level reuse of basic transformation elements. Graph algorithms used for search plan generation are integrated as abstract state machines, while the final code generation step is carried out by code templates.

As a result, the porting of a transformer plugin to a new underlying platform can be accelerated significantly.

2. Overview of the approach

The proposed workflow of the meta-transformation for PST generation is summarized in *Fig. 1.*

In VIATRA2, transformations can be defined by the combination of graph transformation (GT [7]) and abstract state machines (ASM [4]). The Transformation (XForm) metamodel (to be discussed in details in Sec. 3.2) consists of an ASM part for control structures and a graph transformation part for elementary model manipulation.

The steps of the plugin generator transformation are the following:

• As ASM and GT rules are processed differently, we separate them in the first step. Since ASM rules are (semantically) very close to traditional high-level programming language constructs, their handling is not discussed.

• GT rules are processed in two substeps. The LHS of the rule should be handled as a GT pattern, while the action part described by the difference of RHS and LHS (and potentially additional ASM rules).

• For each pattern call initiating a graph pattern matching process, different search graphs are generated. (See [15] for a detailed discussion of search graph generation.)

• An optimized search plan (i.e. the traversal order of pattern nodes) is generated for every search graph in order to sequence the matching of the GT pattern.

• Finally, Java output is generated by code templates. For every different implementation platform only these code templates have to be replaced.

Note that the presented transformer plugin generation approach is implemented in the VIATRA2 framework, which improves extensibility and portability. In the rest of the paper, we first provide a brief overview of the models and transformations used in VIATRA2 (in Sec. 3). Then, the main part of the paper discusses (in Sec. 4) the meta-transformation developed for the PST generation and focuses on the graph pattern matching phase, as it is the most critical step for the performance of graph transformation. Finally, Sec. 5 concludes the paper.

3. Models and Transformations in VIATRA2

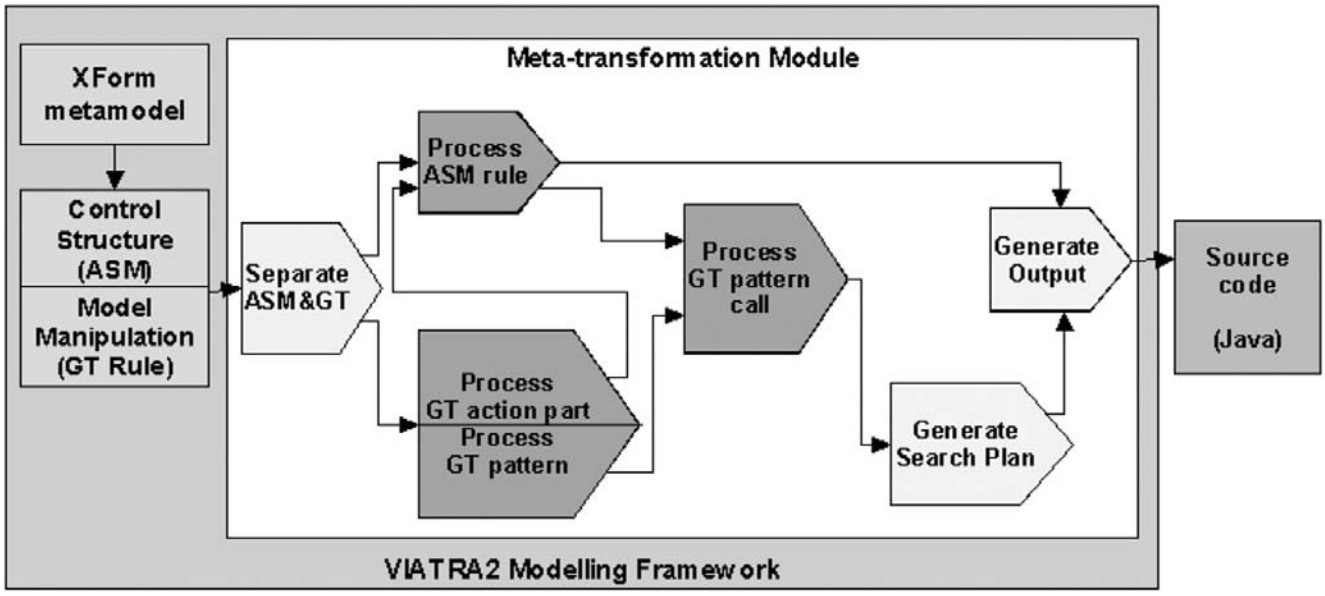
3.1. The VPM Metamodeling Language

Metamodeling is a fundamental part of model transformation design as it allows the structural definition (i.e. abstract syntax) of modeling languages. Metamodels are represented in a metamodeling language, which is another modeling language for capturing metamodels.

The VPM (Visual Precise Metamodeling) [13], which is the metamodel language of VIATRA2, consists of two basic elements: the entity (a generalization of MOF package, class, or object) and the relation (a generalization of MOF association end, attribute, link end, slot). Entities represent basic concepts of a (modeling) domain, while relations represent the relationships between other model elements. Model elements are arranged into a strict containment hierarchy, which constitute the VPM model space. Within a container entity, each model element has a unique local name, but each model element also has a globally unique identifier, which is called a fully qualified name (FQN).

There are two special relationships between model elements: the *supertypeOf* (inheritance, generalization) relation represents binary superclass-subclass relationships (like the UML generalization concept), while the *instanceOf* relation represents type-instance relationships (between meta-levels). By using explicit instanceOf relationship, metamodels and models can be stored in the same model space in a compact way.

Figure 1. Overview of the meta-transformation based generation



3.2. Transformation Language

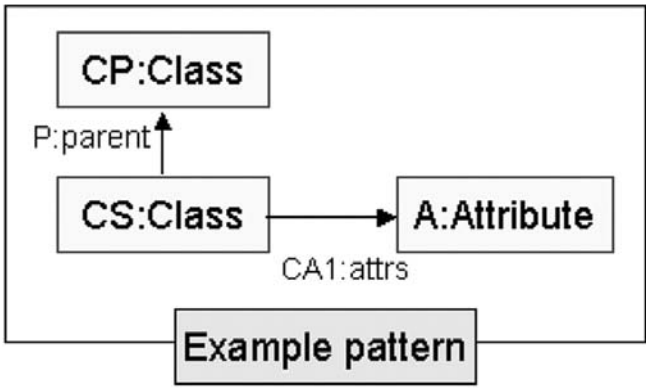
Transformation descriptions in VIATRA2 consist of the combination of three paradigms: (i) graph patterns, (ii) graph transformation (GT [7]) rules and (iii) abstract state machine (ASM [4]).

Graph patterns

Graph patterns (referred as GT patterns) are the atomic units of model transformations. They represent conditions (or constraints) that have to be fulfilled by a part of the model space in order to execute some manipulation steps on the model. A model (i.e. part of the model space) can satisfy a graph pattern, if the pattern can be matched to a subgraph of the model (by graph pattern matching).

An example GT pattern is depicted in *Fig. 2.* The GT pattern of Fig. 2. is fulfilled if there exists a class *CS* that has an attribute A and a parent class *CP*.

Figure 2. Example GT pattern



Graph transformation rules

While graph patterns define logical conditions (formulas) on models, the manipulation of models is defined by graph transformation rules, which heavily rely on graph patterns for defining the application criteria of transformation steps. The application of a GT rule on a given model replaces an image of its left-hand side (LHS) pattern with an image of its right-hand side (RHS) pattern (following the single pushout approach [8]).

The meta-model used for the graph transformation rules in VIATRA2 framework extends the core formalism by: (i) negative conditions can be embedded into each other in an arbitrary depth, (ii) supports the use of ASM rules in the action part of a GT rule, and (iii) supports the notation of standalone GT patterns.

The sample graph transformation rule in *Fig. 3.* defines a refactoring step, which moves an attribute from the child to the parent class. This means that if the child class has an attribute, it will be moved to its parent.

The rule contains a simple pattern (marked with keyword *condition*), that jointly defines the left hand side (LHS) of the graph transformation rule, and the actions to be carried out. Pattern elements marked with keyword new are created after a matching for the LHS is found (and therefore, they do not participate in the pattern matching), and elements marked with keyword del are deleted after pattern matching.

Control Structure

To control the execution order and the mode of graph transformation, abstract state machines [4] are used. ASMs provide complex model transformations with all the necessary control structures including the sequencing operator *(seq)*, ASM rule invocation *(call)*, variable declarations and updates (*let* and *update* constructs), *if-then-else* structures, non-deterministically selected *(random)* and executed rules *(choose)*, iterative execution (applying a rule as long as possible *iterate),* and the deterministic parallel rule application at all possible matchings (locations) satisfying a condition *(forall)*.

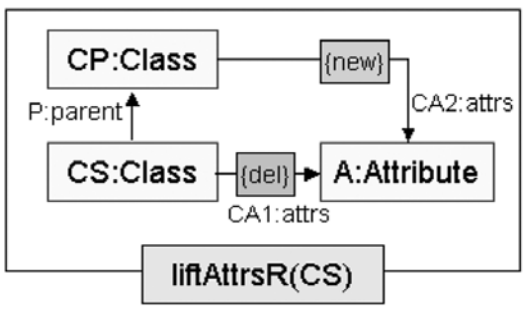
4. Generation of PST with Meta-transformations

To give an overview how the automatic PST generation process can be implemented over model transformations, three conceptually critical fragments are discussed in this section. The first example (in Sec. 4.1) shows how the type of the elements in a GT pattern is

Figure 3. The GT rule liftAttrs

```
gtrule liftAttrs(in CS) =
   {
        condition pattern cond (CS)
=
        {
      Class(CP);
      Class.attrs(P, CP, CS);
      Class(CS);
      Attribute(A);
      new class.attrs(CA1, CP, A);
      del class.attrs(CA2, CS, A);
       }
}
```

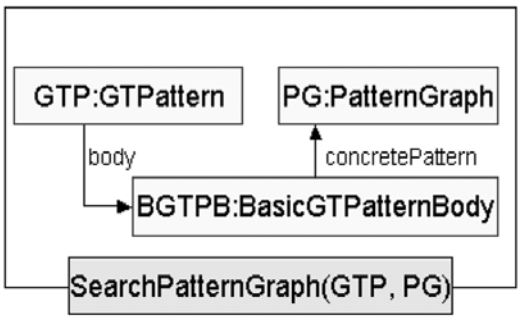
(a) code view



(b) graph view

```
//PG is the pattern graph
of GTP(GTPattern)
pattern aPatternGraph(GTP,PG) =
{
'GTPattern'(GTP);
'GTPattern'.'BasicGTPatternBody'
(BGTPB);
'GTPattern'.'BasicGTPatternBody'
.'PatternGraph'(PG);
'GTPattern'.body(Bo,GTP,BGTPB);
//relations
'GTPattern'.'BasicGTPatternBody'
.concretePattern(Con,BGTPB,PG)
}
```

(a) code view



(b) graph view

Figure 4. The SearchPatternGraph GT pattern

determined by a combination of GT patterns and ASM rules (using explicit instanceOf relations). The second example (in Sec. 4.2) gives an overview how the algorithms of the search plan generation are implemented in the framework. While the third (in Sec. 4.3) shows how the Java representation of relation (association) traversal is generated by a code template.

4.1. Processing the pattern elements of the graph transformation

Our approach is using generic model transformations on the graph pattern rules presented as models in the VIATRA2 modelspace. Generic patterns in VIATRA2 use explicit *instanceOf* relations, which denote type variables. This approach of the PST generation consists of two GT patterns *SearchPatternGraph, directType* and they are called from an ASM rule *processGTPattern*.

The meta-pattern SearchPatternGraph

The pattern *SearchPatternGraph* of *Fig. 4.* denotes that the PG is the pattern graph of the GT pattern GTP. In the transformation model, the *PatternGraph* is connected to the *GTPattern* through the *BasicGTPattern Body* (BGTPB) entity, along a body and a *concrete Pattern* relation.

The generic pattern directType

The pattern *directType* (depicted in *Fig. 5.*) is used to return the direct type of the input parameter X. The outer (positive) pattern matches the metamodel entity, which represents the type of X by the explicit instanceOf relation.

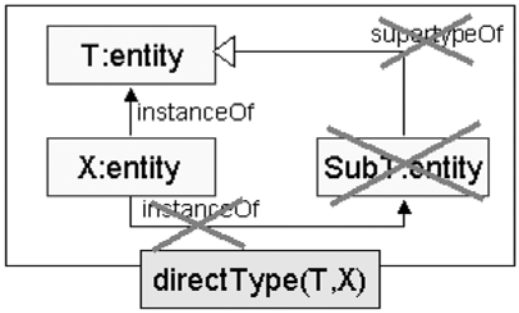
The inner (negative) pattern can be satisfied if the input entity T has a subType, which is connected to X by an instanceOf relation. In this case the execution of the whole rule is violated.

This generic pattern can handle several situations where essentially the same rule pattern should be applied on objects of different types. The type variables used in the pattern are instantiated by the instanceOf relation as concrete entities/relations from the metamodel (similarly to ordinary pattern variables).

Figure 5. The directType GT pattern

```
pattern directType(X, T) =
{//X is an entity of Type T
      entity(X);
      entity(T);
      instanceOf(X,T);
      //T is not a type of X's
      //supertypes
      neg pattern hasSubType(T,X) =
      {
         entity(T);
         entity(SubT);
         supertypeOf(T, SubT);
         instanceOf(X,SubT);
      }
```

(a) code view



(b) graph view

The ASM rule processGTPattern

The ASM rule *processGTPattern* determines the direct type of the elements in the graph pattern PG. Type entities must be under the input parameter *Metamodel* in the containment hierarchy, while PG is the pattern graph of the input parameter GT pattern *InGTPattern*. The steps of the rule are the following:

(i) The *choose* selects the pattern graph of the GT pattern *InGTPattern* with the GT pattern *SearchPatternGraph* and puts it into the variable PG.

(ii) The *forall* enumerates all the combinations of the elements given in the scope one by one and tries to match the *directType* GT pattern. If a part of the model satisfies the pattern then its values are stored in variables *X* and *T*.

(iii) The ASM rule *processEntityBuildSG* is called with parameters PG, *X* and *T* in order to add this new element to the search graph of the GT pattern *PG*.

The VIATRA2 transformation rule is as follows:

```
//GTPatternHolder holds the pattern, and MetaModel is the
//metamodel of the
//entities used in the GTPattern(s)
rule processGTPattern(in InGTPattern, in MetaModel) = seq
{
//selects the GTPatternGraph below the input GTPattern
choose PG with find aPatternGraph(InGTPattern,PG) do
//selects the the type(T) in the Metamodel of the entity X
     forall T below MetaModel, X in PG with
     find directType(X, T) do

     //processes the entity further and adds to the search graph
          call processEntityBuildSG(PG,X,T);
}
```

(iii) Finally, Java code is generated based on the search plan (discussed later in Sec. 4.3).

As abstract state machines are widely used to formalize algorithms [10], is straightforward to implement them in VIATRA2. The following example demonstrates this on the well known greedy algorithm used in the search plan evaluation to select a low cost search plan from a search tree.

Simple greedy algorithm

Initially, the list *P* consists only the starting node. The algorithm simply selects the smallest edge that goes out from the search graph nodes that are already in *P*, and adds the target of the selected edge as the last element of *P*.

The *iterate choose* construct selects the smallest edge that leading out of *P*, by using the ASM function *nodes* and *values* to store the edge with the smallest weight. Then the second *choose* selects the target node of the edge and adds it to *P* by setting the value of the node to *P*.

The recursion terminates when the counter of nodes in *P* reaches the number of the nodes in the search graph (stored in the ASM function values):

```
//SG is the spanning tree of the search graph
rule sPlan(in SG) = seq {
  update values("Min") = "infinite"; //init values
  update nodes("MinEdge") = "-1";
  //selecting the lowest weighted edge
iterate choose No below SG, NextNode below SG, Edge below SG, Owe
    below SG withfind(searchPlan(No,NextNode,Edge,Owe)) do
      //checking the edge values, smaller then Min and not in P
      if((value(Owe) < values("Min") && value(NextNode)!="P") seq
      {update values("Min") = value(Owe);
      update nodes("MinEdge") = Edge; //weights are updated
      };
  // update the value of the element by P
  choose No below SG, NextNode below SG, Owe below SG
   with find(searchPlan(No,NextNode,nodes("MinEdge"))) do seq
     {setValue(NextNode, "P"); //it is now part of the 'list' P
     update values("searchPlan") = values("searchPlan")+1;
     }
  if(values("searchPlan") != values("nodeMaxNumber"))
  call sPlan(SG); //recurvise call
}
```

4.2. Search plan evaluation

As the most critical step for the performance of a graph rewriting framework is the graph pattern matching phase, our approach uses local search algorithms for evaluating the traversal order of the pattern matching. A weighted search graph is a directed graph with numeric weights on its edges, having a starting node connected to each other node with an edge. A search tree is a spanning tree of the weighted search graph. As the starting node has no incoming edges, all other nodes should be reachable on a directed path from the starting node. A search plan is one possible traversal of a search tree. A traversal defines a sequence in which edges are traversed.

The Java code representation of the optimized traversal order is also generated by model transformation, which consist of three phases:

(i) In the first phase, a weighted search graph is generated from the input GT pattern also taking into account all constraints on VPM entities of the pattern.

(ii) By using Chu-Liu and Edmonds algorithm [5,6] combined with a simple greedy algorithm, a low cost search plan is calculated.

4.3. Source code generation

In this section, we propose a source code generation technique for model transformer plugins in Java based on VIATRA2 code templates. The template concept is similar to the one introduced in the Apache Velocity [1] language, but uses the formal ASM and GT paradigms as its control language whose constructs can be referred by the #() notation.

As an example, we use the template rule *printTraversalArb*, which generates the Java equivalent of a simple traversal of a relation with arbitrary multiplicity. In case of arbitrary multiplicity in the traversed direction (one-to-many or many-to-many), an *iterator* is generated to investigate all possible continuations.

The input of the template is the source *(Source)* and target *(Target)* entities of the relation, the type *(Type)* of the target element, the name of the relation *(Relation)* and the next *(Next)* element in the traversal order. The ASM function *name* returns the name of the model element. The steps of the traversal order are processed recursively by calling the ASM rule *processNextStep* in order to generate the Java equivalents of internal code blocks:

```
//code generation the traversal of a relation with arbitrary
multiplicity
template printTraversalArb(in Target, in Source, in Relation,in
Type, in Next)= {
Iterator iter_#(name(Target))=
#(name(Source)).get#(name(Relation))().iterator();
      while(iter_#(name(Target)).hasNext()){
      try{
            I#(name(Type)) #(name(Target)) =
                  (I#(name(Type))) iter_#(name(Target)).next();

            //call recursively the next step in the order of
            //traversal
            #(call processNextStep(Next);)
      } catch (ClassCastException e) {} }
}
```

5. Conclusion

In this paper, we proposed to use generic and meta-transformations for generating platform-specific transformer plugins from transformation specifications given by the combination of graph transformation rules and abstract state machines in the VIATRA2 framework.

The main advantage of our approach is reusability: only final code generation templates need to be altered when porting plugins to other object oriented languages. Up to now, we have a complete implementation for Java, but we plan to port the plugin transformers to other underlying platforms (e.g, Eclipse Model Framework, EMF) and to perform numeric measurements on the transformers.

Experimental evaluation of the generated transformer plugins was carried out in [2] using Enterprise Java Beans 3.0 [9] as the underlying plugin technology.

The generated transformer plugins were able to handle persistent models stored in relational databases with several million graph objects. A next challenge for the future is to integrate transformer plugins to the VIATRA2 framework itself. After successful integration, an optimized compiled version of native Java transformations can be executed instead of the interpreted version.

References

[1] Apache, Velocity homepage, http://jakarta.apache.org/velocity/index.html

[2] Balogh, A., G. Varró, D. Varró, A. Pataricza: Compiling model transformations to EJB3-specific transformer plugins, In: ACM Symposium on Applied Computing – Model Transformation Track (SAC 2006), pp.1288–1295.

[3] Bettin, J.: Ensuring structural constraints in graph-based models with type inheritance, In: M. Cerioli (editor), Proc. 8th Int. Conference on Fundamental Approaches to Software Engineering (FASE 2005), LNCS 3442. pp.64–79.

[4] Börger, E., R. Stark: "Abstract State Machines. A method for High-Level System Design and Analysis," Springer-Verlag, 2003.

[5] Chu, Y. J., T. H. Liu: On the shortest arborescence of a directed graph, Science Sinica 14 (1965), pp.1396–1400.

[6] Edmonds, J.: Optimum branchings, Journal Res. of the National Bureau of Standards (1967), pp.233–240.

[7] Ehrig, H., G. Engels, H.-J. Kreowski, G. Rozenberg, editors: "Handbook of Graph Grammars and Computing by Graph Transformation, Vol. 2: Applications, Languages and Tools," World Scientific, 1999.

[8] Ehrig, H., R. Heckel, M. Korff, M. Löwe, L. Ribeiro, A. Wagner, A. Corradini: In: [11], World Scientific, 1997. pp.247–312.

[9] Enterprise Java Beans 3.0, Sun Microsystems, http://java.sun.com/products/ejb/docs.html

[10] Gurevich, Y.: The sequential ASM thesis, Bulletin of the European Association for Theoretical Computer Science 67 (1999), pp.93–124.

[11] Rozenberg, G., editor: "Handbook of Graph Grammars and Computing by Graph Transformation, Vol. 1: Foundations," World Scientific, 1997.

[12] Rumbaugh, J., I. Jacobson, G. Booch: "The Unified Modeling Language Reference Manual", Addison-Wesley, 1999.

[13] Varró, D., A. Pataricza: VPM: A visual, precise and multilevel metamodeling framework for describing mathem. domains and UML, Journal of Software and Systems Modeling 2 (2003), pp.187–210.

[14] Varró, D., A. Pataricza: Generic and meta-transformations for model transformation engineering, In: T. Baar, A. Strohmeier, A. Moreira, S. Mellor, editors, Proc. 7th International Conference on the Unified Modeling Language (UML 2004), LNCS 3273. pp.290–304.

[15] Varró, G., D. Varró, K. Friedl: Adaptive graph pattern matching for model transformations using model-sensitive search plans, In: G. Karsai and G. Taentzer, editors, International Workshop on Graph and Model Transformations (GraMot 2005), ENTCS, Vol.42, pp.191–205.

NGN development at Magyar Telekom: The future of our fixed network

PETER JANECK

Head of Magyar Telekom Network Division
peter.janeck@t-com.hu

Reviewed

Keywords: next generation network, multimedia services, IMS

Fixed-line telecommunication is at crossroads. Mobile is dominating the voice market and is developing toward offering multimedia. Internet penetration is growing and applications on Internet are blossoming. Broadband access is spreading and the developing bandwidth capability more and more appeals for attractive content. Usage of legacy telephone service is declining. Next Generation Network (NGN) – having the Release 1 standard package published – gives new opportunities to renew fixed telecommunications offering feature-rich multimedia services and applications. Magyar Telekom is running an ambitious development to deploy broadband access and to build up an IMS based NGN network that will be also the technical basis of the ongoing integration of Magyar Telekom and T-Mobile Hungary.

1. Introduction

Magyar Telekom as the main incumbent telecom provider of Hungary provides a wide range of services. Magyar Telekom Group comprises lines of businesses and a family of subsidiaries. It holds a majority stake in MakTel of Macedonia and Telekom Montenegro and it is also present in Ukraine, Romania and Bulgaria. Magyar Telekom is the market leader in fixed telephony (78% by 2 753 thousand phone lines and ISDN channels), in mobile services (45% by 4 194 thousand subscribers), in broadband access (79,8% by 329,3 thousand DSL lines) and in Internet services (36,2% by 328,5 thousand subscribers)*.

Competition opened four years ago is extending as service demands, technical capabilities and business models are developing. Mobile is dominating the voice market. It is more characteristic in CEE countries, where mobile has overcome fixed telephony at a low/moderate penetration of the latter. Mobile has advantages based on intelligent terminals and services that make it attractive beyond the convenience of mobility.

Internet penetration is growing and applications are developing. Voice over Internet offers mean a threat to legacy voice services that is increasing with the deployment of broadband access. As broadband access becomes independent from PSTN subscription the substitution effect of VoIP will be significant. Demands for bandwidth, for broadband access is growing and the increasing bit-stream capability more and more appeals for video services. That leads to the objectives of triple-play: the combination of voice, data and video (entertainment) services.

User terminals are rapidly developing: growing intelligence, IP/Ethernet/SIP interfaces, portability and other convenient capabilities. With emerging broadband wireless technologies new possibilities are opening for upgrading fixed network offering certain kind of mobility, and for fixed-mobile convergence.

Telecom players operating on one or two markets are preparing to move and enter lines of telecommunications, e.g. PSTN resellers and alternative providers are expectedly going to enter the mobile market as virtual network operator (MVNO), CATV providers are offering also Internet and VoIP.

The main technology stream that utilizes the trends of broadband access deployment, the feature rich applications and flexibility of Internet (IP), the mobility, the portable intelligent devices, multiple access solutions including wireless ones, and aims a convergent telecommunication integrating voice, data, video (entertainment) is the development of Next Generation Network (NGN).

Magyar Telekom's strategy is to build NGN in harmony with broadband deployment as the technical enabler of becoming a next generation integrated telecom provider.

2. Broadband evolution

For NGN – as being an IP-based packet switching network – broadband access is a prerequisite. Having the objective of providing a wide scale of multimedia services from voice to high quality real-time video broadband access expansion is one of the most important strategic assets.

Broadband deployment was started for offering fast Internet access. From now it will be gradually re-positioned into multi-service access, the access of NGN. TV based entertainment services are the next step towards serving the communication and entertainment

** Figures refer to the end of 2005*

needs of a family. In short term IPTV implementation is starting with current access technology (4-7 Mbps) to get ready for the future and to compete with 3play offers on cable. In long term start segmented access network developments to enhance service delivery capabilities for the next generation connected home (HD IPTV, video telephony, multiple PCs, high bandwidth internet, interactive gaming).

Development objectives can be seen in *Fig. 1.* The targeted total number of broadband customers at the end of 2008 is 1 000 000 including residential and corporate users. Beside Internet usage VoIP, IPTV and broadband gaming will take off; in year 2008 dual and triple players will represent 15% and 22% of total access.

NGN will interoperate with every type of access technologies. The main short term goal is to offer a wide spectrum of fixed and wireless connection possibilities, which allows MT to provide broadband interfaces for different local situations. The mass market broadband demand will be served on the basis of ADSL and ADSL2+. Our goal is to increase ADSL penetration from 15% up to 45% on copper lines. WIMAX will be used as a Hot Zone infrastructure for high speed data communication, to cover DSL gaps, replacing Fix-GSM and for LTO expansion.

ADSL coverage will reach 90% of subscribers and 51% of settlements for 2008.

The long term strategy is moving forward to an optical infrastructure and replacing step by step the traditional copper network elements to enable broadband access up to 25 Mbps. In 2006 MT is testing the strength of a passive optical network (PON) and VDSL2 in pilot implementations for green field areas.

3. Evolution of NGN

The long term, perspective solution to the challenges of telecom is the converged common network: Next Generation Network. The main objectives of NGN may be summarised as: multi-service integrated network, technology-independent services/applications, and ubiquitous and personalised services. Building a common network for all services promises benefits of reduced investment and maintenance costs. Technology independence of services needs special next generation service concepts, like OSA/Parlay, and promises flexibility and the separation of operator and service provider roles. Ubiquity and personalisation of services means that users may compile service features for their demands, and the same set of services may be accessed via different access means and from different locations: the concept of mobility is extended, nomadic usage appear even at the edge of fixed networks.

NGN means a significant change in architecture: instead of monolithic exchanges, there are separated functional entities, and control functions earlier being hidden inside a switch, appear between devices – it promises flexibility and economy, but on the other hand this is challenging for standardisation.

Standardisation has been in progress very intensively and resulted in the Release 1 package of ETSI/TISPAN. The important fact is that standardisation shows convergence: IP Multimedia Subsystem (IMS) – being specified by 3GPP for 3G mobile has been accepted as the basis of the common architecture.

After an early period while NGN was considered as a new way of telephony promising cost reduction, we arrived to a milestone when the concept of a real multiservice integrated network is specified. The overall architecture highlighting the main characteristics can be seen on *Fig. 2.* (on next page).

Motivations for establishing and deploying NGN are: cost reduction, defending against competitors attacks (competitiveness), business opportunities via new services and the ending of old technology lifespan. Cost reduction could be realized only by replacing significant segments or the whole of old platform that can not be justified in short/medium terms at stagnating demands. PSTN/ISDN platform shall be sustained, that is assured by life cycle management agreements. The technical basis of defending our positions and to find new business opportunities is establishing an overlay NGN platform.

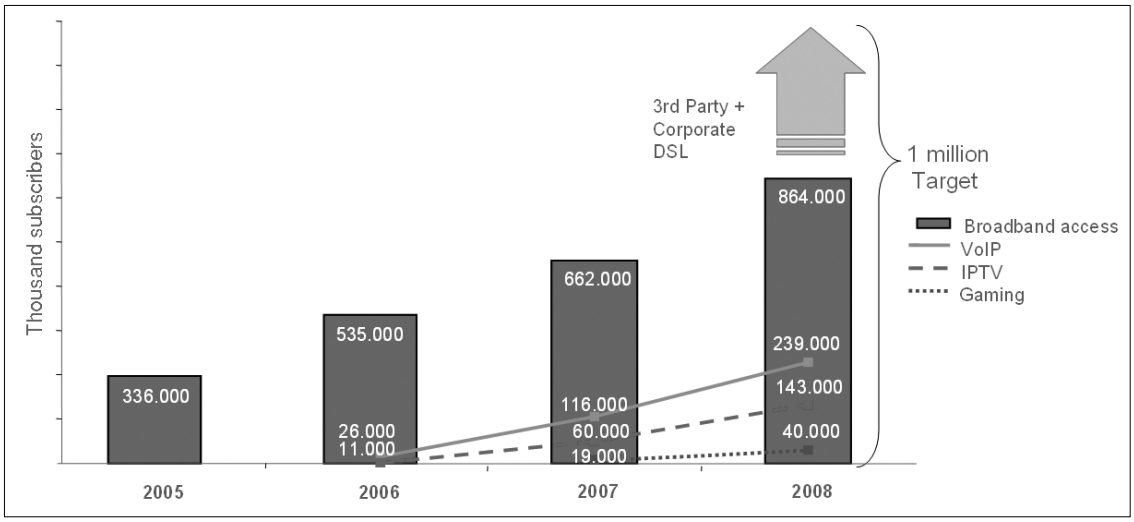


Figure 1. Broadband access development objectives

MT implemented a pre-NGN platform based on softswitches to get technical experience with an NGN platform and its interconnection to PSTN, and to give economic and perspective solution for actual VoIP demands: Voice over CATV, Voice over Internet, Unified Governmental Backbone, Romanian Presence. Having the standards of the IMS based NGN, the multimedia service objectives and the merge of MT and T-Mobile Hungary the overlay NGN platform is being built according to TISPAN NGN specifications.

The scope of our NGN project is:

- Exploiting potential synergies of integrated development, operation and maintenance of MT's telecommunication networks
- Common selection and implementation of a carrier grade NGN platform for MT Group
- Integration of pre NGN platform
- Common implementation of (new) services for existing market demands

Roles of overlay NGN platform:

- Data and video services accomplishing 3play service offers from one platform
- Carrier hosted business communication, integrated voice-data services to business market
- Migration of VoIP services on pre-NGN shall be judged case-by-case
- It will be the basis of platform consolidation: substituting legacy voice and data platform.

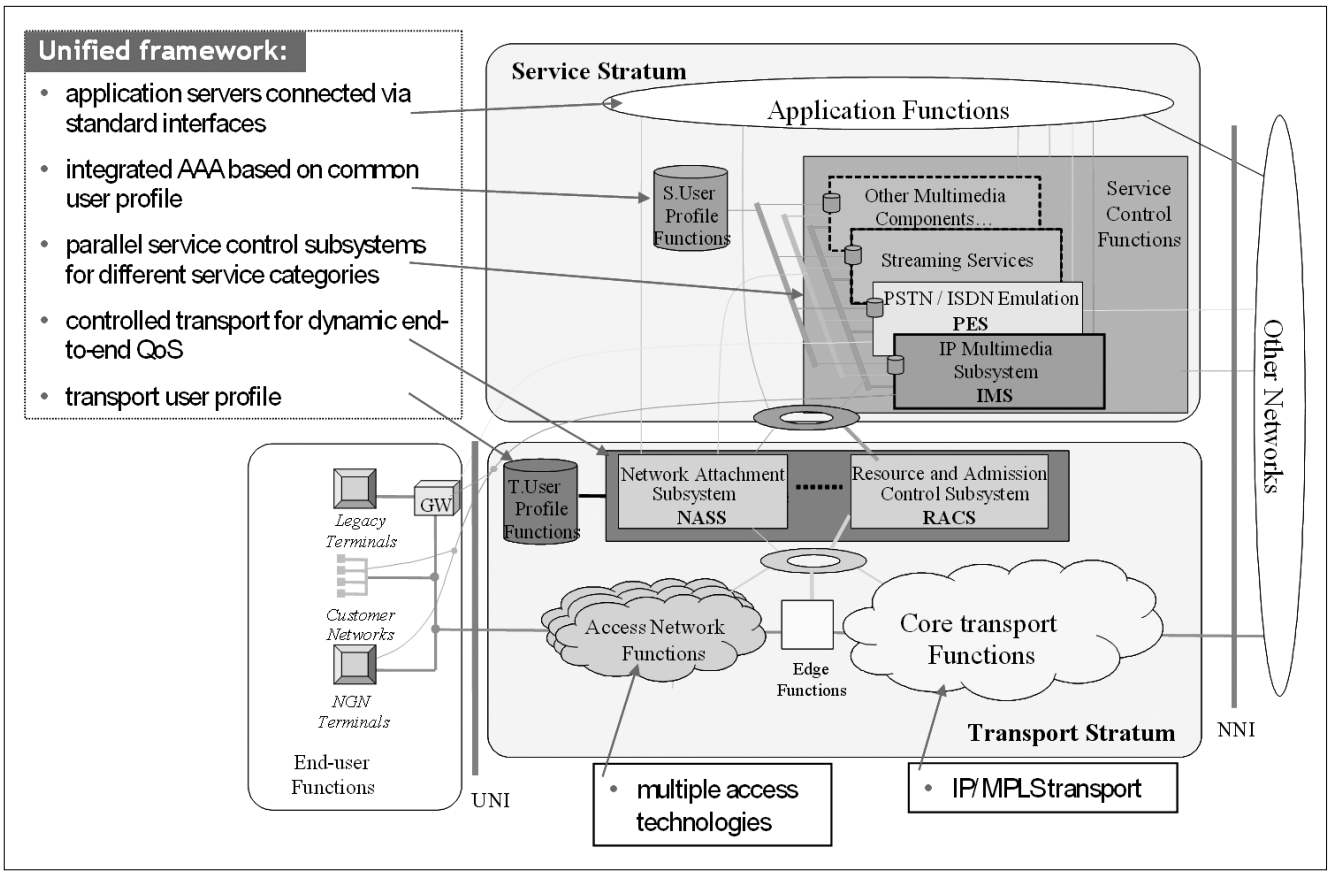
4. Network consolidation

The main objective of the convergent NGN to operate an integrated multi-service network instead of several networks of different technologies for different services will be reached if the legacy networks are substituted by it.

Although the substitution of PSTN/ISDN network in short term is not justified, offering VoIP services on the overlay NGN platform will stimulate subscribers to leave PSTN and use the new services. This service-driven migration will decrease the utilized capacity of the old platform. A gradual technical migration shall be started substituting exchanges of low utilisation. In the next five years more than half of voice lines will be provided by NGN *(see Fig. 3)*. For the objective of stopping decline of total voice lines a significant part of VoIP shall be offered in new area and as second line.

Consolidation of networks and network elements is generally planned. The actual steps are depending on the technology, the services produced and the status of the given network. The motivation is to decrease operational costs by ceasing technologies, platforms selecting those for further operation that matches to NGN architecture and requirements. In case of some technologies like PDH transmission it means the substitution by SDH or IP. In case of the legacy IN platform it may be replaced by SIP server that matches into the NGN architecture and is capable to control also the PSTN for providing IN services.

Figure 2. Overall NGN architecture



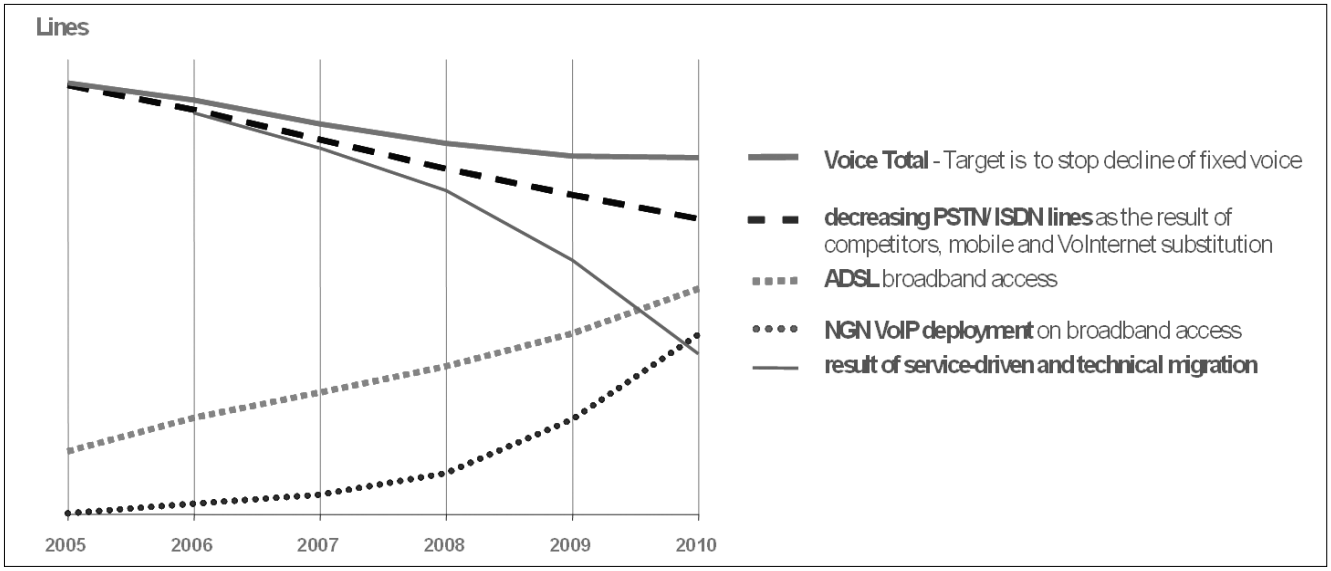


Figure 3.
Consolidation impacted by VoIP deployment on NGN platform

5. TMH integration

As a relevant part of merging T-Mobile Hungary (TMH) to Magyar Telekom an Integrated Network project is in progress. It focuses on synergies of integrated development, operation and maintenance of MT's telecommunication networks. It elaborates cost efficient integrated network concept based on market requirements as defined by residential and corporate segments, and analyzes the development opportunities of the current network infrastructure towards NGN *(see Fig. 4)*.

T-Com transport network provides transport service for all demands of MT group, including backbone for T-Mobile's UMTS. The core transport technology of NGN is IP and will be developed into a high-availability multiservice transport.

Fixed and mobile environment gives different requirements for IMS, that justifies implementing two parallel subsystems. The key is the compatibility that is important for interoperability with the application servers and between different service control subsystems.

On the top of the parallel service control subsystems (IMS-es) common application servers will provide fixed-mobile converged services:

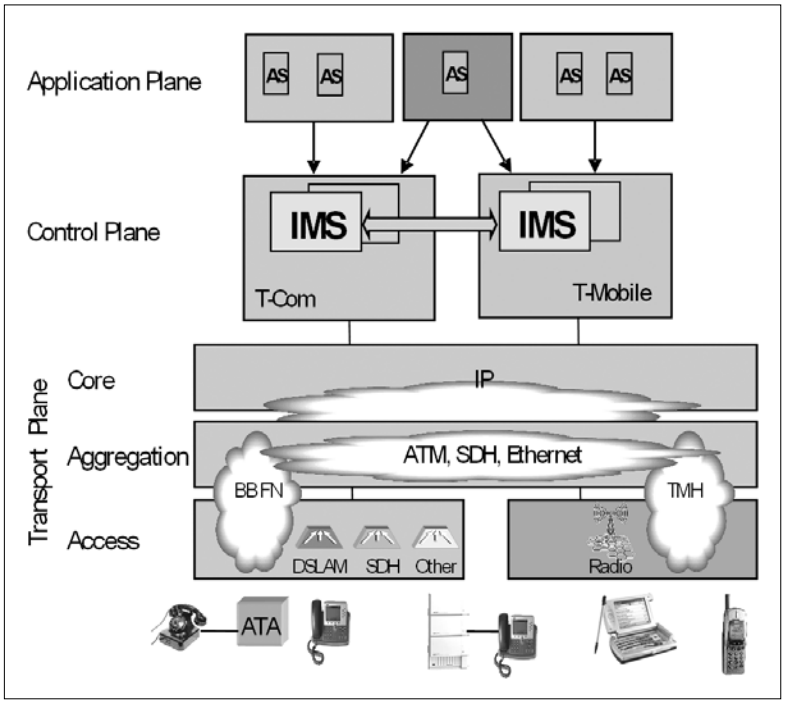
– provide flexible product development (faster to the market and simplified for the network operation);
– enable inter-working of IP services across networks (e.g. roaming) or across services (service inter-working);
– e.g.: TV, video sharing, multiplayer gaming.

6. Conclusions

Establishing IMS based overlay NGN is the main stream of Magyar Telekom's infrastructure development. NGN capabilities will give answers to challenges, providing fixed-mobile convergent and feature-rich communications, services with guaranteed quality and security, attractive content for broadband and a variety of applications on new CPEs.

NGN is the target where existing technologies shall converge – consolidation of legacy platforms will be made gradually: PCM, PDH, ATM, MLLN, PSTN.

Figure 4.
NGN infrastructure for fixed and mobile services



Service delivery platform:
Critical enabler to service provider's new revenue stream

ANETT SCHÜLKE, DANIELE ABBADESSA, FLORIAN WINKLER

NEC Network Laboratories, NEC Europe Ltd., Germany
{anett.schuelke, daniele.abbadessa, florian.winkler}@netlab.nec.de

Reviewed

Keywords: Next Generation Networks, Service Creation, Service Delivery Platforms, IP Multimedia Subsystem (IMS), Policy Control

Service creation and delivery platforms as key network components support objectives for next-generation services, such as the ability to tailor services quickly and flexible for individual customers and to provide an open platform for third-party service development. The Service Integration Environment (SIE) described in this paper aims at providing an enabling technology for developing advanced applications for fixed and mobile networks. The main focus of this realization is the integration of advanced value added services over the 3GPP IP Multimedia Subsystem. This paper will give an overview about the Service Integration Environment as a potential part of a future SDP solution. An in-depth view of the respective market and its relation to the ongoing standardization activities will be outlined. A detailed description of the functional structure of the architecture layers, the advantages of the involved integration techniques (including the dynamic policy management) are provided. The provided sample application will motivate the easiness for service creation by exploring the SIE-offered IMS and non-IMS service enablers.

1. Introduction

Today's mobile telecommunications service providers face strong competition to deliver new revenue-generating services to the market while decreasing related operational costs. Mobile users are getting more demanding in their requirements for useful, personalized application offered at a reasonable price. Future service creation and delivery platforms as key network components are targeting to deliver more creative services and more quickly to a service provider's target market. They support objectives for next-generation services, such as the ability to tailor services quickly and flexibly for individual customers and to provide an open platform for third-party service development. Systems integrators are bringing together multiple SDP products of different vendors, combining the strengths of these different products, and ensuring to be in a standards-based and open service-oriented environment.

Within 3rd Generation Mobile Networks, Internet-related concepts are being introduced more and more in the telecommunication environment. At the same time the industry is standardizing services and service enablers within the Open Mobile Alliance (OMA). OMA has defined a reference architecture called OSE that defines how basic services can be re-used and combined to integrate into new and advanced services. OMA's role is to create application level specifications for various services, agnostic to the underlying network technology. The missing part on the way to even more attractive new services is the glue that ties network and service enablers together and makes service creation easy and efficient.

The Service Integration Environment (SIE) described in this paper aims to provide an enabling technology for developing advanced applications for fixed and mobile networks.

This paper will give an overview of the Service Integration Environment as a potential part of a future SDP solution. An in-depth view of the respective market and its relation to the ongoing standardization activities will be outlined. A detailed description of the functional structure of the architectural layers, the advantages of the involved integration techniques (including the dynamic policy and resource management) are provided. The provided sample application will motivate the easiness of service life cycle exploring the SIE-offered IMS and non-IMS service enablers.

2. Market Review

The migration towards "All-IP" is transforming the telecommunication landscape and, together with voice price erosion, it is putting additional pressure on telecommunications operators to launch innovative services and create differentiation. In the past few years fixed line and wireless operators primarily focused on OPEX and CAPEX reduction. The business focus is now shifting towards developing new services and bringing them to market quickly and efficiently.

Service bundling has been and still is one of the approaches followed by telecommunications operators to gain or retain market shares. However, service bundling is only a short term solution since it offers little differentiation. Moreover, telecommunications operators which will make service bundling their key strategy for fighting competition, in the long term, will become simply flat-rate pipe providers and be constantly involved in price wars.

Despite the fact that wireline and wireless operators are rather different, a common set of requirements can be identified:

– Creating service differentiation
– Faster time-to-market
– Achieving cost efficiency
– Ensuring compelling user experience

The current approach to service creation has been characterized by a "stovepipe" approach which has lead to the proliferation of separate platforms, one for each specific service. This approach presents several limitations and it is suitable only in the case of a limited service portfolio with no integration and orchestration between services. Moreover, it is not optimal and unsustainable in the next-generation service environment which requires the support of convergence of data communications (IT) and voice communications (telco).

For this reason, OMA is now focusing on a new service enabling approach, which will enable operators to quickly implement service orchestration and therefore will enable them to create greater service differentiation.

Wireless operators have been traditionally more innovative and agile than wireline operators. Service Delivery Platforms (SDP) and IP Multimedia Subsystem (IMS) were both initiatives originating from the wireless industry. The fixed line industry is now catching up and there is a growing interest amongst fixed line operators in introducing SDP and IMS as key elements of their next-generation service layer.

Despite many initiatives around the world led mainly by the largest telecommunications operators (e.g. AT&T, BT, Bell Canada, Sprint), SDPs are not yet deployed ubiquitous across the industry. Costs, complexity, uncertain Return on Investment (ROI) and legacy companies' organizations are seen often as the major obstacles to the wide adoption of these platforms. Moreover, it is important to stress the fact that there is no comprehensive definition for SDP, because SDP functionalities do not reside on a single platform, but are rather comprised of an integrated set of software modules.

The introduction of SDP solutions in the operators' environments is also linked to the migration to IP-based platforms and services. SIP application servers and IMS are key enablers of this migration at the service level. Operators will adopt different strategies for the introduction of SIP Application Servers, IMS and SDP. Their strategies will be dictated by the local environment, e.g. competition, regulatory, etc., and it will result in a slower or faster deployment of IMS, SDP and SIP Application Servers. As illustrated in *Fig. 1.* [1], the initial IMS and SDP deployments will start in the next 12 to 24 months, whilst full deployment is expected to follow at later dates.

A study conducted by the Moriana Group [2] shows that SDP spending is expected to reach about $19 billion over the period 2003-2007 *(see Fig. 2)*.

NW Operators	Large	Medium	Small
Mobile	500	2,200	1,900
Fixed/Mobile	1,000	1,600	500
Fixed/IP	1,000	800	200
Total			**9,700 M€**

Service Providers	Large	Medium	Small
MVNO	420	630	336
VNO	630	525	336
ASP	1,050	630	252
ISP	2,100	2,250	880
Total			**10,039 M€**

Figure 2. SDP Forecast Spending (2003-2007)

The importance of SDP is also confirmed by the increasing number of companies growing from smaller solution vendors to large IT system providers, which populate the SDP marketplace. Interestingly, most of the SDP providers are not traditional telecom suppliers. This is a key sign that highlights how SDP is mainly an IT solution. In the future, the role of the Telco networks will be of pure (IP) transport bearers whilst the delivery of services recognizable by consumers and enterprises will be through an IT rather than a network infrastructure. The shift towards IT and the growing importance that SDP and IT technologies will play in the future for the telecommunications industry is also recognized by the corporate management.

According to a study conducted by the Yankee Group [5], the majority of corporate managers interviewed be-

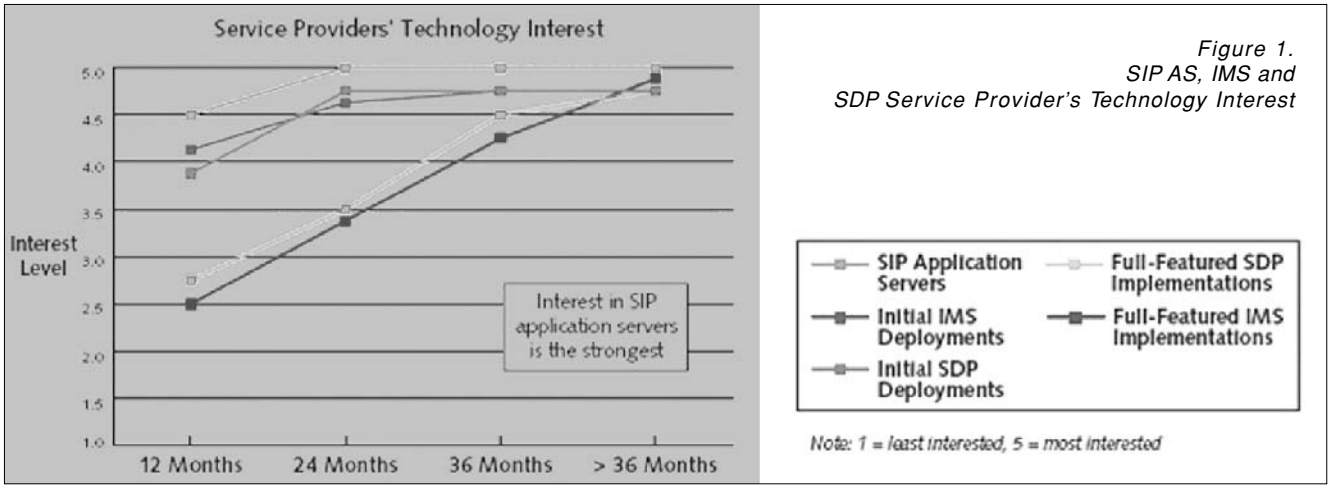


Figure 1.
SIP AS, IMS and
SDP Service Provider's Technology Interest

lieve that new products and service technologies will remain one of the key factors to success for the next 5 years. Therefore, investment in new products and service technologies represents a “strategic” investment for telecommunications companies.

3. Standardization

The Open Mobile Alliance (OMA) [4] is the leading standardization organization for mobile service related technologies. Founded in 2002, it has incorporated several existing fora like WAP Forum, Location Interoperability Forum (LIF), Wireless Village Initiative, Mobile Games Interoperability Forum (MGIF), Mobile Wireless Internet Forum (MWIF) and many others.

The OMA Service Environment (OSE) is a logical architecture that provides a conceptual environment for service enablers, interfaces to applications that make use of these enablers, interfaces to a Service Providers’ Execution Environment (e.g. software life cycle management) and the interfaces to invoke and use underlying capabilities and resources for enabler implementations. The IP Multimedia Subsystem (IMS) (as defined by 3GPP) is a Session Initiation Protocol (SIP) based IP multimedia infrastructure that provides a complete platform for globally interoperable IP multimedia services – especially for the mobile environment. The ISC (IMS Service Control) interface allows applications, i.e. commercial services, to access IMS capabilities. IMS provides service-enabling functions and IP transport, which are relevant for the OSE as defined by OMA. Applications and service enabler implementations may make use of IMS capabilities, e.g. charging, authentication, service management, etc.

The main requirements defined for an OSE can be summarized as

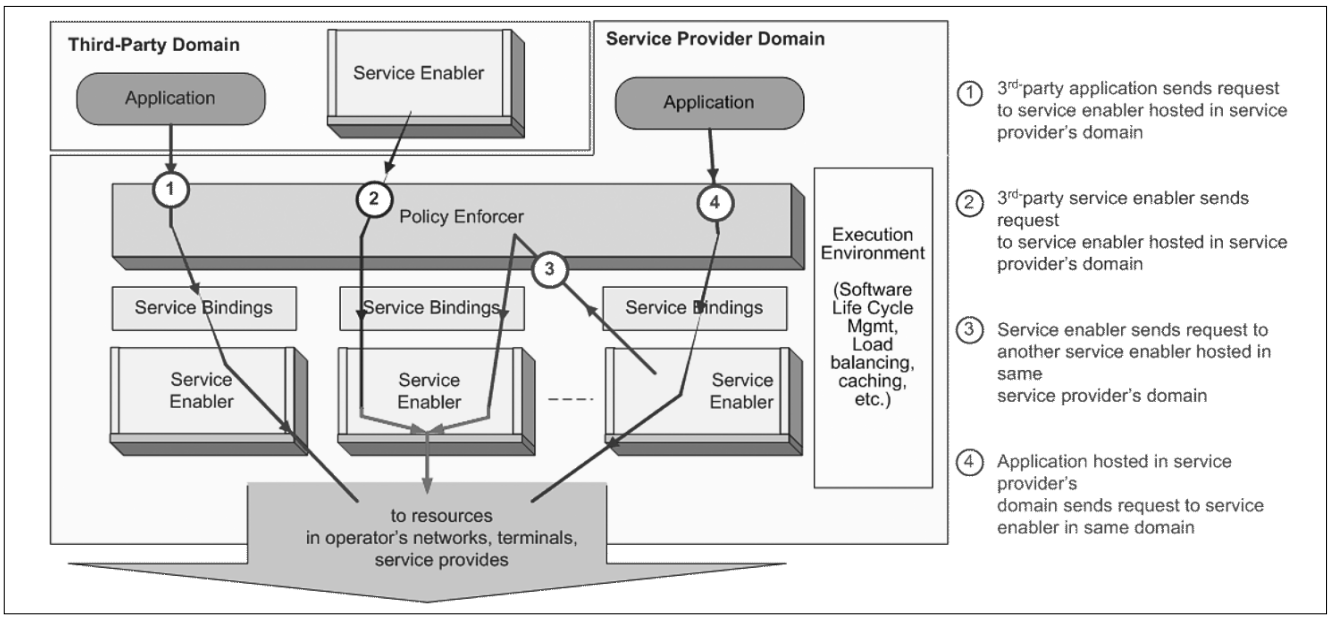
– Mechanisms for authentication, authorization, federated identity, subscription management,
– Single sign-on/log-out,
– Accounting and Charging Handling,
– Provisioning of services, service enablers, and user parameters,
– Service Registration, Discovery Mechanisms, Policy Management,
– O&M support (including service monitoring).

The conceptual architecture of the OSE is shown in *Fig.3* [4]. It also presents a simplified view of the OSE request flow from various entities towards the network. In OMA, a service performs useful work for users or service providers while applications are defined as a software or hardware implementation of a related set of functions. Therefore, applications are the way to access one or more services using service enablers. Applications may be deployed within or outside a service provider’s domain or in a terminal. Service enablers provide standard access to network and terminal resources, and to other service enablers using service bindings.

The OSE deals with service requests from applications (located inside or outside the service provider’s domain) as well as requests from other service enablers. Enablers shall be re-usable in the same or in different service provider domains. Enablers can be exposed by various binding mechanisms (e.g. Web Services, standardized APIs, CORBA, etc.).

To secure existing networks and their operations, the OSE may protect network resources by a policy management allowing fine grain control of access and system behaviour. The policy management is also used when application access service enablers or when service enablers interact with each other. Policy mechanisms are used to enforce access control by dynamic policy evaluation, to manage the use of network resources e.g. through appropriate charging, logging and enforcement of user privacy or preferences, and to allow extensibility by offering service-platform-controlled delegation between enablers.

Figure 3. Description of OSE and request flow



4. Service Integration Environment

At NEC's European Network Laboratories we develop a solution for a service architecture tailored for the creation of advanced services, specially tailored for IMS services [3]. The Service Integration Environment is the realization of our concept for the core engine of a framework for service creation and delivery. The main benefits in the approach of this unique Service Integration Environment are to develop advanced integrated applications by composing new service logic with re-usable service functions as basic building blocks. Its target for OMA-compliance is expressed in the architecture with emphasis on the centralized intelligent policy management and the flexibility for future extensibility in service creation. As service creation function the SIE provides service API bindings for a flexible service enabler portfolio offered for the application development.

4.1. Service Creation Concept

The Service Creation Process can be described as a 2-phase-process divided into knowledge domains – the *Business and Service Knowledge* and the *Network Capability and Operation Knowledge*. This separation is illustrated in *Fig. 4.*

The Network Capability and Operation Knowledge covers the network and operation control, network service management and control as well as the reliability for the entire service handling from invocation of a service up to the charging and revenue control. The Business and Service Knowledge builds upon this trusted knowledge from the operator's side. It allows for well-defined service interface access, service provider's policy setup under operator's management control, permission-controlled service invocation and revenue control.

A service created on the service provider's domain contains the business logic parts which are specific and unique to the new service. For its operation, this new service is re-using the well-operated services of the operator's domain without installing and managing its own lifecycle for those network-provided service building blocks.

This view supports the arguments for the strongly horizontal approach of the SIE. The separation of service creation from the service enabler's execution calls for a reliable concept for "test once, use forever" with an open graduation for easy and fast new service creation.

4.2. Architecture and Functional Description

The Service Integration Environment (SIE) architecture provides the technical realization for the exposure of network capabilities towards application developers and 3rd party service providers in a secure and manageable way. The SIE provides a combination of various service integration technologies in the different layers *(see Fig.5).*

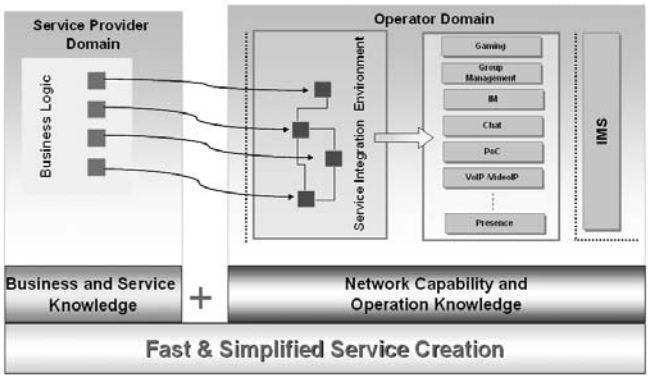


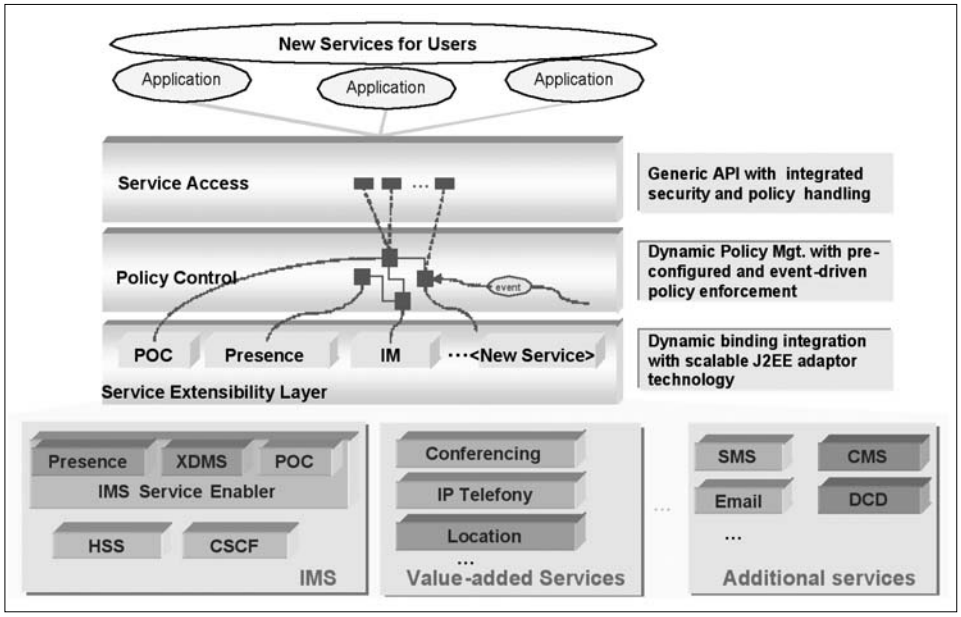
Figure 4. Visualization of the Service Creation

The top layer, called *Service Access*, allows the discovery of and the access to existing service enablers using e.g. Web Services, enabler discovery, dynamic proxies, and AAA. Service Access technology offers a generic API integration with intrinsic security and policy handling. This framework's gate opens services the access and control to the service enablers via Java (J2EE) APIs and Web Services. *Fig. 6.* (on next page) shows a simplified flow example.

Functionalities included in this layer are

– User administration for SIE platform (e.g. access control),
– permission handling towards the SIE platform and its managed underlying service platforms (e.g. Mutual authentication where user/application has mutual login to different underlying platforms),
– convergence and combination control setup with dynamic PEEM proxy delegation mechanism.

Figure 5.
Architecture of the Service Integration Environment (SIE)



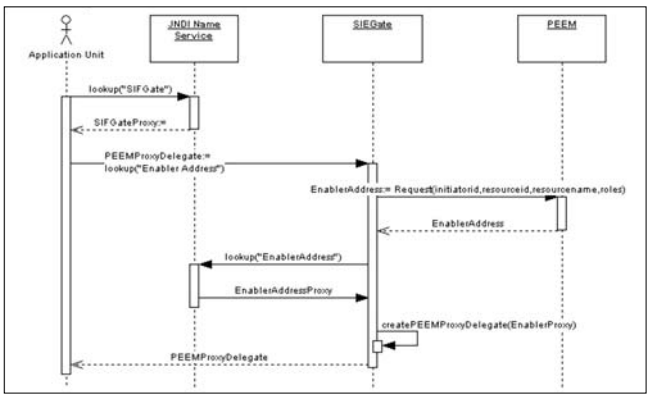


Figure 6.
Example Flow for service access via SIE Access gate

This service environment contains – as crucial point of this service integration platform – an embedded intelligent *Policy Control* layer that allows the creation and enforcement of rules to ensure safe and controllable access to any kind of resource.

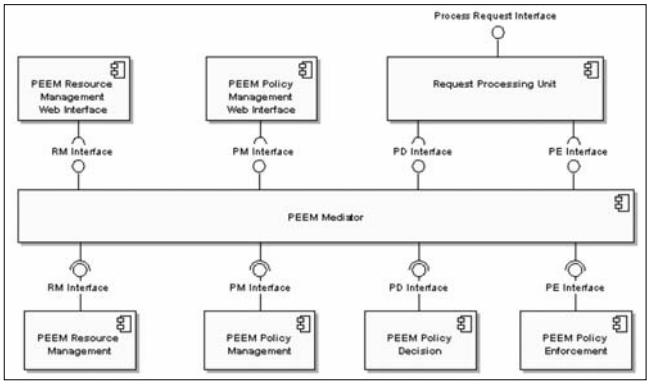
The middle SIE layer – containing this intelligent policy control – permits network operator to control requests between advanced application and service enablers. Within this concept, the design of a dynamically enabled service composition over policy control is a major part of SIE platform. Policy Control technology is designed for static pre-configured as well as dynamic event-based policy enforcement. This permits e.g. for dynamically enabled service invocation respecting current network conditions. The Policy Control's component overview is shown in *Fig.7.*

The access layer towards the underlying network capabilities and services – the *Service Environment Extensibility* allows dynamically expanding the set of service enablers using adaptor technologies for the dynamic installation of new APIs. The service adaptors shown in this layer in Fig. 5 and 8. are represented as Java or Web Service interface, depending on the underlying service platform.

4.3. Performance

Estimating the performance of the SIE prototype implementation, the performance of the centralized policy control layer for a request applied in the direct policy enforcement mode is measured. The time consumed by the policy decision and enforcement process has been measured by issuing 10000 local requests from within a single thread to the application server. By issuing the requests locally we avoid noise effects from network access. For our performance tests one policy had to be found, analyzed and enforced. Since policy decision and enforcement times depend on the number of applicable policies, processing times should increase in more complex cases.

Figure 7. Policy Control Component overview



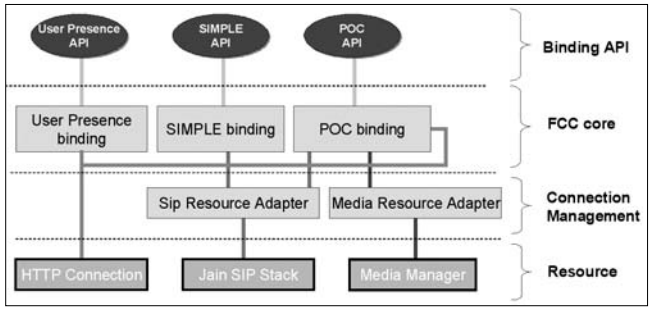


Figure 8. SIE's Extensibility Layer Component overview

In *Fig. 9,* the processing times for requests are varying between 16 to 30 ms normally, and 25 ms in average. The peaks are due to the application server's EJB pooling and allocation behavior that should reduce in a constantly running system.

Running conditions:

10000 requests issued locally from within a thread.

Measurement conditions:

JBoss 4.0.0 application server, single processor 2.0 GHz Intel Pentium 4, machine with 512 MB of memory.

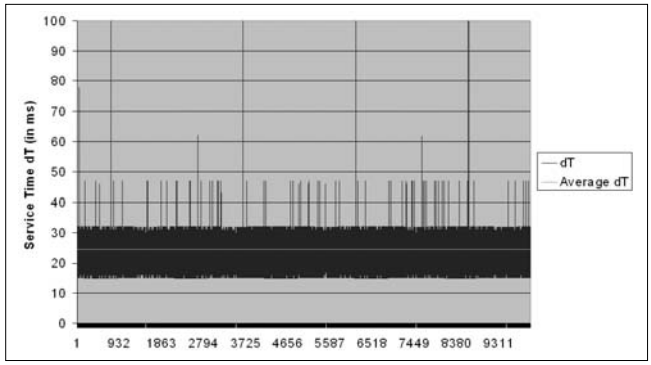


Figure 9. Performance measurement for the policy layer

In this system no code optimization was performed on the policy framework, so that further performance gains in the future can be expected. The response times show that policy enforcement can be done in an acceptable timeframe. Since we are using clustering and load balancing capabilities of the application servers, extending the cluster can insure availability if the need arises.

5. Advanced Application: Mobile Auction

This section will provide an example description of an advanced application using the service creation potential of the SIE platform.

The example is outlining an application for a *Mobile Auction*. This example application is based on the consumer market. The basic idea of auctions is a well known application which is deployed in various scenarios (e.g. internet (eBay), live auctions). Those existing scenarios are normally fixed scenarios in place and/or time and/or procedure. Internet auctions are providing the convenience of attending the final phase of bidding from any private or most suitable location with access to the internet, however it actually is lacking the real life-feeling to be connected to a group of people in order to share information/dispute about an item. The proposed Mobile Auction scenario is bridging this experience gap by expanding the fixed internet-based scenario into the mobile communication world. The focus is not just basically to provide mobile internet access, but rather to use context parameters to integrate an interactive, mobile communication phase adaptable by preferences given by the auction participants (seller and purchasers). The specific idea is the integration of IMS-based communication services with respect to context information (e.g. presence, location). Fig. 8 illustrates the basic request flow for the Mobile Auction application.

The Mobile Auction starts with "online" bidding for an auction item on the internet. The internet-based auction switches to the "Live" auction on "context condition", e.g. if enough bidders are online (can be distinguished by using their presence status) and when there are enough biddings issued to the portal.

The "Live" auction phase can utilize different IMS communication services for "live" bidding and price update e.g. Push-to-Talk ad hoc group call or a dedicated text chat room based on Instant Messages. The portal also has a functionality, so that the text chat service connects between mobile and fixed (internet) users.

Fig. 10. is illustrating this scenario. There might be certain conditions which prohibit the setup of the mobile phase by policy control to receive e.g. Presence Update information (due to charging restrictions), as shown in Fig 9. The SIE platform will record those activities as visualized above in the policy execution logging capability.

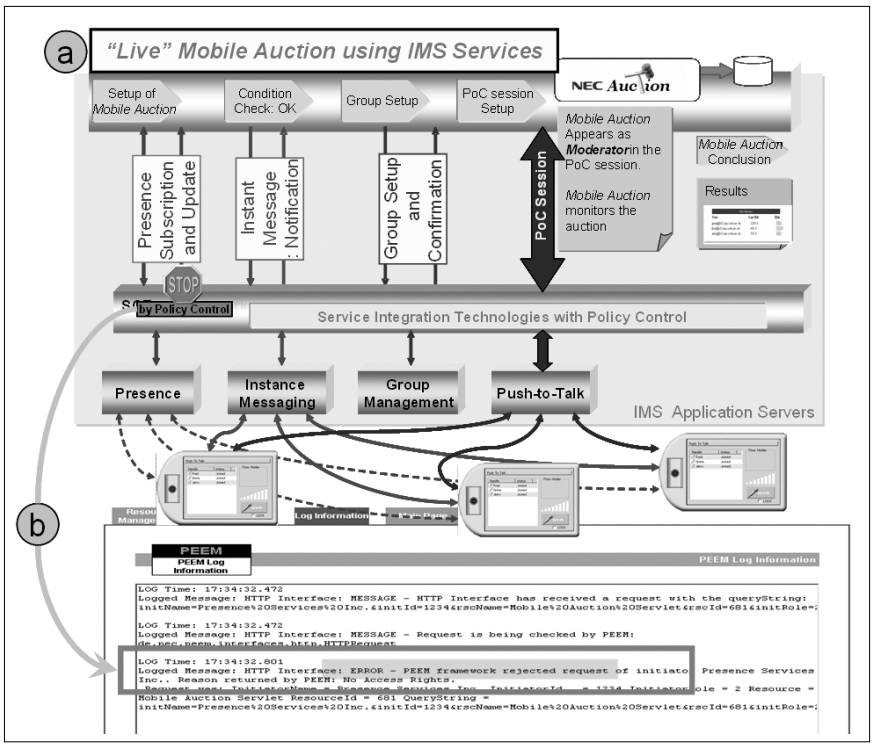
6. Conclusion

The world of Mobile Network operators and service providers in Europe is different from the situation in Japan in several aspects. Many large and small providers are competing with each other. Their main challenge is to be ahead of the competition regarding the roll-out of new and attractive services. In order to improve service development speed and to contain development costs, mobile operators want to build as much as possible on standardized architectures and solutions. OMA is the leading mobile services standards body.

This paper gave an overview of the Service Integration Environment (SIE) as a potential part of a future SDP solution accompanied with a respective market and OMA standards discussion. A detailed description of the functional structure of the SIE architecture layers, the involved integration techniques are provided. The centralized dynamic policy control has been explained as a major part of our platform. The provided sample application illustrates the easiness of creating advanced applications utilizing e.g. IMS service enablers.

The SIE developed at NEC's Network Laboratories in Heidelberg aims at complying with the OMA service model paradigms. The resulting SIE architectural model opens a big market opportunity for future service revenue for network operators.

Figure 10.
Illustration of Mobile Auction service request flows



References

[1] The Yankee Group, August 2005. "IP Multimedia Subsystems and Service Delivery Platform Will Drive SIP Application Server Adoption",

[2] The Moriana Group, June 2004. "Service Delivery Platforms and Telecom Web Services",

[3] IP Multimedia Subsystem (IMS), TS 22.228. http://www.3gpp.org/

[4] Open Mobile AllianceTM, http://www.openmobilealliance.org

[5] OSS Challenges and the Role of Integrators in IMS/SDP Deployments, The Yankee Group, September 2005.

WTC2006
World Telecommunications Congress, April 30-May 3 2006, Budapest, Hungary

FÜREDI ÁGNES

Scientific Association for Infocommunications, Hungary (HTE)

Two scientific societies, the Scientific Association for Infocommunications, Hungary (HTE) and the Association for Electrical, Electronic & Information Technologies, Germany (VDE/ITG) have succeeded to organize the World Telecommunications Congress (WTC2006) in Budapest, Hungary, between April 30 and May 3 2006, under the subtitle "Emerging Telecom Opportunities".

The event was based on traditions and novelties in many senses. The International Technical Committee (ITC) – associating highly distinguished professionals and scholars from different fields and sectors of communications from almost all over the world – responsible for the professional content brought together the three-decade heritage of providing platform for the telecommunications community by merging two series of events (ISS – the International Switching Symposium, and ISSLS – the International Symposium on Services and Local AccesS). The countries of the organizing bodies that joined forces have long relationship in history, culture, science, industry and business.

• The conference program was built around four tracks that might show the next transforming steps telecommunications could take in the 21st century. Five plenary sessions with invited keynote presentations from very high level international technical and business experts gave insights about possible trends, while 86 technical presentations in 21 technical (oral and poster) sessions, three with invited session leading talks covered all conference topics.

• 148 extended abstracts had been submitted and the ten reviewing committees of countries represented in the ITC selected about 90 topics based on their relevance, novelty and quality. Authors developed their final papers from extended abstracts in an interactive process moderated by their session chairpersons. Around 10 papers were peer reviewed by at least three opponents to provide scientific credit points to PhD student authors.

• Over 200 participants from 25 countries visited WTC2006.

• WTC2006 was technically co-sponsored by the Communications Society of the Institute of Electronics, Information and Communication Engineers (IEICE), and sponsored by Magyar Telekom, NEC, Siemens, the Ministry of Communications and Informatics, Hungary and the National Office for Research and Technology, Hungary.

• The organizers hope that this first joint event lived up to its motto, and would really mean for all who took part: Emerging Telecom Opportunities.

Tracks and Sessions of WTC2006

Track 1. Transforming the business of telecommunications

1. Regulatory & Policy Issues
2. Service Management
3. Network Management
4. Convergent Networks & Services
5. New Business Models & Business Issues

Track 2. Transforming the Quality of Services

1. QoS Engineering
2. QoS Topics
3. Quality of NGN Services
4. QoS Measurement & Monitoring (4)
5. Network Design and Analysis, Poster Session 2

Track 3. Transforming the networks' technology

1. IP Access & Networking
2. Access Networks: Technology & Modeling
3. IP Infrastructure Developments
4. Dynamic Wireless Networks
5. Broadband Access
6. Optical Core Networks
7. Routing

Track 4. Transforming the customer's experience

1. Multi-Services Networks & Applications
2. Broadband Services
3. NGN Infrastructure & Services
4. Broadband & Wireless, Poster Session 1

Keynote and Session Leading Speakers of WTC2006

Miklós Boda, President, National Office for Research and Technology, *Hungary*

Tadanobu Okada, Associate Senior Vice President, NTT, *Japan*

Bruno Orth, Deutsche Telekom AG, *Germany*

Pradeep Sindhu, Founder and Chief Technology Officer, Juniper, *USA*

Prof. Lajos Hanzo, Chair, University of Southampton, *UK*

Juan-Carlos Valverde Rey, Manager for Network Architecture Evolution, Telefónica Espana, *Spain*

Bob Cowie, Chief Engineer, Openreach, BT, *UK*

Peter Janeck, Chief Technical Officer, Magyar Telekom, *T-Com Hungary*

Tim Stone, Senior Marketing Manager, Cisco Systems, *UK*

Prof. John Cioffi, Professor, Stanford University, *USA*

Michael Chamberlain, Director of Solutions, Microsoft, *USA*

Nora Maene, Marketing Director Consumer Applications, Alcatel, *Belgium*

Riccardo Fiandra, Head of ICT Quality of Service Department, FASTWEB, *Italy*

Herbert Mueller, Chief Operation Officer, T-Com, *Slovak Telecom*

Wolfgang Schmitz, Senior Executive Vice President, *Deutsche Telekom AG*

Camille Mendler, Vice President, Yankee Group, *USA*

Oscar Gestblom, Marketing Manager, Business Unit Systems, Ericsson AB, *Sweden*

Dave B. Payne, Manager, Broadband Architecture and Optical Networks, BT One IT, *UK*

(Presentations can be downloaded from: www.wtc2006.com)

ZTE IPTV: a great IPTV solution from China

LIANG BING

An inspiring news comes in June 2006: ZTE wins the IPTV project of Shanghai Branch of China Telecom. Shanghai is the economic center of China, and China telecom is the largest fixed line operators in China. According to the contract, the network will serve 49,000 customers in Shanghai Pudong by provisioning video services including Live TV, video on demand (VoD), time-shifted TV (TSTV), network private video recorder (NPVR) and near video on demand (NVoD) as well as value-added services such as interactive games, information services, instant communications, media sharing etc.

The Shanghai IPTV project follows ZTE's three previous IPTV contracts winning from China Telecom which includes Jiangsu Telecom, Guangdong Telecom and Shannxi Telecom. Besides the repeated success in China Telecom, ZTE is also a winner of IPTV projects of Beijing Netcom which is the first formal IPTV bidding of China Netcom. In the overseas market ZTE was awarded the triple play contract by Columbia operator Emcali in October 2005, which covers 13,000 IPTV subscribers and 140,000 VoIP subscribers.

With the remarkable H.264 ZTE IPTV solution, ZTE gets the first IPTV market share in China and consolidated the position as one of the world's leading telecommunications equipment providers in IPTV market.

ZTE IPTV total solution is an end-to-end solution specifically designed for employing multimedia services over an IP network. ZTE's IPTV is a mature, advanced system with excellent operability, manageability and scalability, helping to expand the range of services of many fixed network operators. ZXBIV provides fixed network operators with a platform for providing abundant broadband interactive value-added services to a broad range of clients.

System Architecture & Components

The ZXBIV consists of head end system, middleware, the content distribution network (composed of VoD servers and corresponding management system), CAS & DRM system, HG, STB and IP network(s).

Head End System

The head end system receives, demodulates and decrypts video and audio content from a variety of sources including satellites, terrestrial, studio and storage, while encoders convert it into an IP multicast stream in the desired coding format. In the ZXBIV solution, the head end's MPEG4 AVC compression sets a new benchmark for low bit rate, high quality encoding of both video and audio.

Middleware

ZXBIV middleware provides concise subscriber, SP, content and report management in conjunction with EPG and STB version management functions, all while remaining open for future services integration. The ZXBIV IPTV system is a genuinely easy to operate and administrate, fully reliable, commercial system.

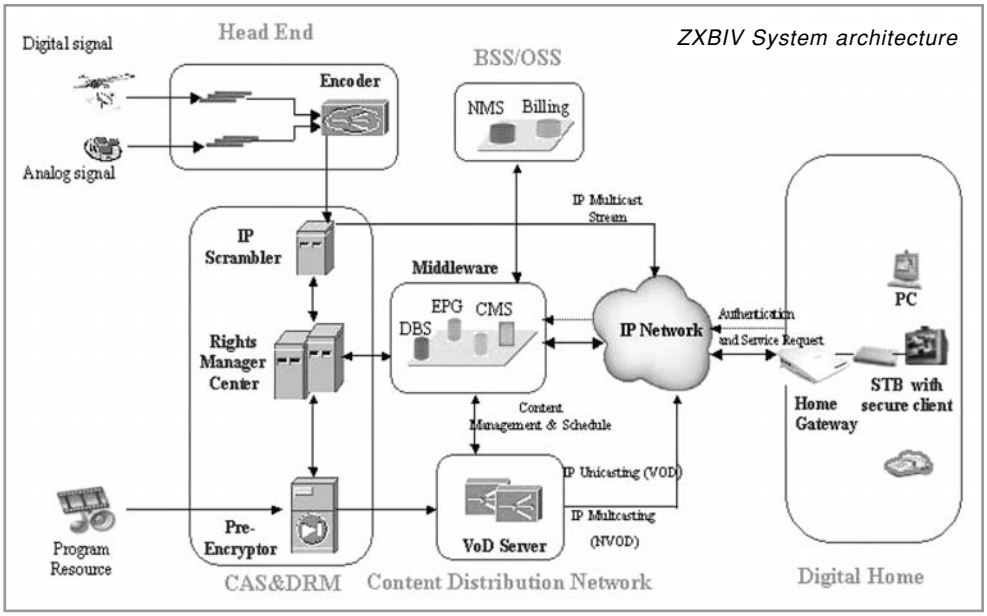
Content Distribution Network

The ZXBIV content distribution network (CDN) is made up of video server clusters, with contents distributed between CDN nodes according to flexible distribution policies. The distributed CDN structure allows operators to economically scale the system, proportionate the subscriber load and service demand.

CAS/DRM System

As we all know, the average IP network security is just that... average. How do telecom operators protect content from being captured while riding on an IP network? To get the content, operators must prove they are adequately securing the content they purchase.

ZXBIV System architecture



ZTE's IPTV solution provides end-to-end security policies to help solve these problems. A ZXBIV solution integrates the major vendor's CAS & DRM systems to protect content distribution including live TV descrambling, VOD content pre-encryption, and secure client integrated in STB.

Digital Home

HG

Home Gateway (HG) is a very important part of the ZTE IPTV Total Solution. It is the integrated access gateway for home users. As the core of the home network, HG controls and coordinates all equipment in the home; it also provides users with a uniform and convenient user interface. The ZTE home gateway HG has a diversified interface. The HG acts as a connecting link between upstream and downstream resources, connecting different home application equipment, and connecting this equipment to various external networks. It improves the scope, depth and entertainment level of all home equipment and applications.

STB

The ZXV10 series STB supports advanced MPEG4 AVC decoding. In addition, ZXV10 series STB can also provide HDTV, picture-in-picture (PIP), connectivity with external storage devices and so on, all of which help to make IPTV services more competitive.

Highlights of ZXBIV Solution

Diversified IPTV services, customized for operators

- Live Broadcast Television (IPTV) – TV viewing via IP network
- Video on Demand (VOD, NVOD) – Movie or TV, music, MTV etc.
- Network Personal Video Recorder (NPVR) – Allows users to record live TV content at a certain period and play it later at a time of their choice
- Time-shifted Television (TSTV) – When watching live-TV programs, users can pause and continue viewing later, according to their schedule and interests
- TV on demand (TVOD) – The live channels could be viewed any time
- Video Communication – Extend communications to television, a revolution of home communication. A Bluetooth handset is designed for conveniently calling and dialing.
- Short message service – Send short message to mobile network
- Instant messages
- Walled garden (for customized web site)
- Advertising – Insert advertisements or system messages as required
- Gaming
- Media sharing – Share the media, photo album with your friends

The perfect end-to-end solution

ZTE offers an end-to-end solution, which includes the head end system, middleware, VOD server, encoder, network platform, HG and STB as one complete package. This seamless integrity not only ensures the quality and interoperability of all products, it simplifies system implementation and maintenance.

Scalable CDN architecture, flexible for expansion

ZXBIV solution supports a flexible video distribution network architecture, which allows for centralized, distributed and hybrid CDN deployment. Networks can smoothly evolve at different development stages, using current network resources and improving the performance-cost ratio of IPTV as fast or slow as conditions allow.

Open system architecture

ZXBIV is initially designed in accordance with open system platform. ZXBIV has special interfaces to support third party application. The system may interoperate with third party IP TV components effectively and efficiently. The openness also guarantees new service generating capability.

A unified network management platform

Because the middleware, VOD server and STB are all from ZTE, the unified network management comes true. The middleware, VOD server and STB could be managed by ZXNM01 management platform.

At the same time, ZTE data network products, soft switch, DSLAM, home gateway products are all based on ZXNM01 network management platform. It eases the telco's network management.

Able to response the market customization requirement promptly

Because IPTV system is so complicated and made up of several parts, integration is necessary in an IPTV project. So any new function or service extension maybe effect all parts of the system.

How to control the project process? That's really a great challenge for integrators.

There's no worrying for this with ZXBIV IPTV solution. Because the main body of ZXBIV IPTV solution are from ZTE, we can response to customizing requirements promptly.

Advanced system architecture, commit to IPTV future

Grasping the IPTV core technology and taking advantages the experience accumulated in voice, broadband, data network, multi media, value added services for several years, ZTE will support more and more new IPTV services with the technical development. Concerning the IPTV services extending to the mobile field, ZTE has strength to support related services with the deep understanding and achievement of 3G.

(x)

Selected Papers

híradástechnika 2006/7

OPTICAL COMMUNICATIONS

New directions in the wave propagation theory

Keywords: wave propagation, modes, inhomogeneity

The paper presents a contradiction, originated from a root fallacy, which is commonly accepted and applied in the wave propagation calculations up to now, but yields wrong results. Further, a solving method will be briefly overviewed, by application of which it became possible to deduce new and exact solutions, to avoid the former errors, and to interpret successfully several registrations in space research.

(In: 2006/4, pp.2–6.)

Physical limits of the applicability of 10 and 40 Gbps speed DWDM systems

Keywords: DWDM, optical network, optical amplifier, optical fibre, dispersion, SPM, XPM, FWM, SBS, SRS

Due to the growing transmission demands and the technical evolution, the use of DWDM systems that have more and more channels for the transmission of bundles of higher and higher speeds is spreading. In case of the application of 10 Gbps, but especially 40 Gbps systems, the dispersion characteristics of the optical fibres come to the focus of attention. Due to the high optical levels that can be provided with optical amplifiers, non-linear phenomena in the optical fibres can be observed. The imperfection of the passive optical devices used for the multiplexing/de-multiplexing of the wavelengths causes channel cross-talks.

The aforementioned phenomena are in close relation with the high-speed transmission and they have to be taken into account when designing, installing and operating such systems. In general, the problems of the physical layer appear much more in case of high speed multiplex wavelength transmissions than as it used to be in case of known lower speed systems.

(In: 2006/2, pp.2–10.)

Applying statistical multiplexing and traffic grooming in optical networks jointly

Keywords: dynamic optical network, GMPLS, traffic grooming, statistical multiplexing

Multilayer optical core networks are able to provide huge bandwidth. With traffic grooming we can utilize more efficiently the available resources. The principle of grooming: if the routes of two different traffic flows (or demands) have common links, their traffic can be joined in to the same wavelength channel. Another well-known solution for increased efficiency of resource usage is multiplexing the traffic.

The statistical multiplexing does not allocate the maximal bandwidth for each traffic demand, but less than the maximal and more than the average. The aim of this article is to investigate the effects of applying both solutions.

(In: 2006/2, pp.35–39.)

SPEECH TECHNOLOGY

Speech recognizer for preparing medical reports: Development experiences of a hungarian speaker independent continuous speech recognizer

Keywords: automatic speech recognition, HMM models, n-gram models, bi-gram models, perplexity

A development tool (MKBF 1.0) for constructing continuous speech recognizers has been created under Windows XP. The system is based on a statistical approach (HMM phoneme models, and bi-gram language models with non linear smoothing) and works in real time. The tool is able to construct a middle sized speech recognizer with a vocabulary of 1000-20000 words. New solutions have been developed for the acoustical preprocessing, for the statistical model building of phonemes, and in syntactic level. Through our examination, different training sets were used with different vocabularies. Hungarian is a strongly agglutinative language, in which the number of the word forms is very high. This is the reason why two forms of bi-gram linguistic model were constructed: one is the traditional word forms based and the other is the morpheme based model, in which the vocabulary is much smaller.

In this article, test results and the experiences drawn from them are presented. Recognition accuracy has been considerably increased using perplexity based linguistic adaptation.

(In: 2006/3, pp.14–20.)

Machine learning algorithm for automatic labeling and its application in text-to-speech conversion

Keywords: machine learning, language identification, LID, TTS

In this paper we present a novel machine learning approach usable for text labeling problems. We illustrate the importance of the problem for Text-to-Speech systems and through that for telecommunication applications. We introduce the proposed method, and demonstrate its effectiveness on the problem of language identification, using three different training sets and large test corpora.

(In: 2006/3, pp.51–58.)

Summaries • of the papers published in this issue