híradástechnika

VOLUME LVII.

2002/7

Selected papers

Mobil systems and networks

IP traffic

Informatics

Scientific Association for Infocommunications

Contents

The picture on the cover skeet demonstrates the realty of the fluid modell (Tamás Varga, Péter Benkő, Tamás Bőhm, Attila Eschwigh-Hajts)

Foreword to the Selected Papers

We always try to collect several articles of the previous half a year in order to show the latest development results, technical novelties, and our most distinguished authors to our readers not speaking Hungarian. A selection cannot ever be objective and even if we realised this fact, specifying the principle of selection may represent an additional problem. A possible principle is to show what our researchers achieved in the previous period. However, the thought that we make known our authors' name also beyond the borders of Hungary occurred, as well. And, of course, we should not forget our readers, but we have to show them topics and solutions that are interesting for them.

To harmonise these three viewpoints in the given volume is extremely difficult while also recognised foreign authors strive for publishing their new results. Having considered all of these issues, we thought to compile our current number of July out of the topics being in the forefront of interest of the world and to include articles in the English number that represent the results of the areas that develop at the highest speed.

Considering either the quantitative development or the technical novelties that reside in the service, it seems to be unambiguous that the mobile systems are ranked first among the telecommunications services of the recent years. However, the availability of radio frequencies is limited, therefore the questions associated with utilising these limited natural resources the most efficiently are the most vital. For this, the application of the Smart Antenna concept is one of the promising methods. Our first article deals with this principle and with its dimensioning method. This is followed by the survey of quality issues associated with the launch of the promising UMTS kept frequently mentioned. The quality of the WLAN systems is closely associated with this topic. One of the acknowledged experts of Siemens provides a comprehensive overview about the perspectives. Among the mobile novelties, the software radio is one of the most user friendly one by the use of which the handset containing the same hardware can automatically fit networks operating at various wavelengths and using various systems. The sender controls the mobile terminal and if it is authorised, it will be connected to the network. We hope that this picture is suitable both for showing the domestic development results and at the same time it depicts the application opportunities of the newest ideas to our readers.

The other highlighted topics are to transmit the traffic increased due to the use of the Internet and to utilise the available transmission capacities optimally. In this area the vital question is to launch the Call/Connection admission control (CAC) methods and to choose the best solution among them. After having demonstrated the opportunities of the liquid model simulation that can be applied for the purposes of dimensioning, as well as those of the multimedia transmission, we can be familiarised with important basic questions, that is, with settlement and tariffing.

After having surveyed these two big areas, we show a new solution out of the topics of televiewing and the Automatic Wizard Generation each.

Having scanned our selection, we are not convinced that we have solved the task the best. It could surely have been published more characteristic articles that would demonstrate the domestic researches deeper. To remove our doubts, we thought of that we have the opportunity after half a year to publish an English number again to our readers. Our authors' activity experienced so far grants the hope that we will publish important novelties also from other areas first in Hungarian language and later also in the English number. Therefore, we would like to emphasise that this is a selection and not an evaluation but rather a sampling.

Overview of Smart Antenna Concept

ZOLTÁN NÉMETH, SÁNDOR IMRE, FERENC BALÁZS

BUTE Department of Telecommunications

1. Introduction

Nowadays wireless communication is used in many fields of the life. Significant part of this communication is arranged though personal mobile communication systems which are used by hundred millions of people. Therefore the load in radio channel is quite heavy and the capacity is limited. Capacity in such a system can be defined as the total bit rate per bandwidth per unit area or bit/s/Hz/m^2. It seems to be evident that the main purpose is to increase the capacity, but there are many limiting factors of course. If two or more devices use the same frequency while being in the radio range of each other, then disturbance will occur. This disturbance appearing during parallel transmissions is called interference. The interference falls into two types: common channel and adjacent channel interference. The first ensues in the case of same frequency sub-band communications, and the reason for the second is that the adjacent bands cannot be separated perfectly. If we divide the given frequency band rationally, and we use this division in the whole area that should be covered, then number of users will be insufficiently small because of the common channel interference. We can say that the capacity is limited by the interference [4].

The problem can be solved if we divide the area into smaller areas called cells. In each cell not all of the available frequencies are usable. Naturally the cells have number of adjacent ones. In a cell the used frequencies should differ from other ones used in adjacent cells. But far from it these frequencies become available again, because the device will get out of the radio range of the interfering one and the signal is attenuated. This theory is applied in mobile communication systems and called frequency reusing. The smallest group of adjacent cells in which all of the available frequencies are utilized is called cluster [4]. In Figure 1 one can see a set of cells where the cluster size is 7. Every different colors mean different frequencies applied in each cell.

In presence of large number of subscribers it is conceivable that this solution is insufficient. Further enforcement is needed for providing adequate service (i.e. adequate SIR – Signal to Interference Ratio). The

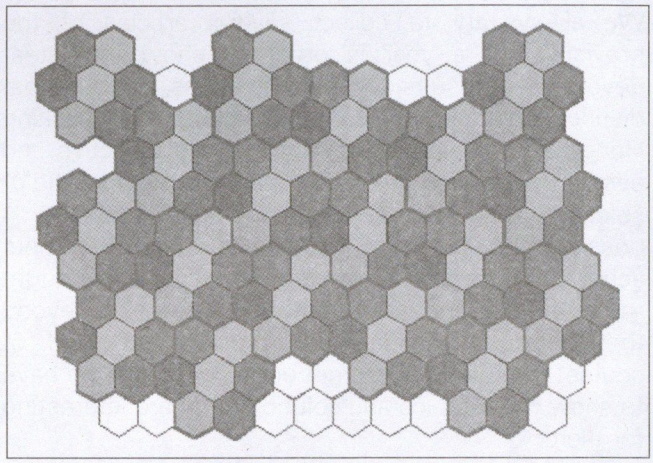


Figure 1. Cellular arrangement (cluster: 7)

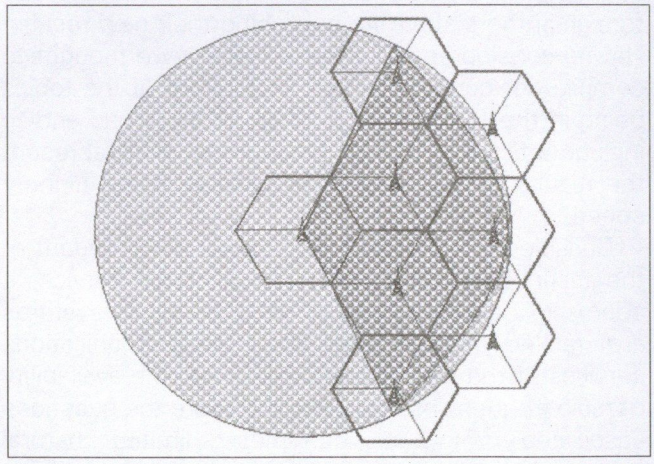


Figure 2. Sectorization

fundamental idea is that one transceiver (transmitter + receiver) in base station system should not transmit in the whole cell, but only in a given part of it. This part is called sector [4]. An example for sectorial transmission is showed in Figure 2. Supposing that we have sectorial transceiver in a GSM system with angle of 120 degree in spite of 360, the number of interfering cells decrease from 6 to 2, which means 4.77 dB growth in SIR. These numbers for sectorial transceiver of 60 degree are: decreasing in interference from 6 to 1, and 7.78 dB growth in SIR. All of the numbers are valid for cluster size of 7.

Even in the case of sectorial radiation we can note that rest of the energy is wasted uselessly and only a very small ratio is transmitted towards the user. It would be desired to concentrate the radiated energy into a narrow beam because it has two advances. On the first hand a fraction of transmitting power would be sufficient, on the second hand – since the radiated wanted signal of any cell means interfering signal for other ones – this method would significantly reduce the interference. The ideal solution would be that case when one beam is attached for each user [1]. This idea leads to the theory of smart antennas. Application of smart antennas is a very promising field in cellular radio communication, so in the following sections we are going to examine operation, construction and main types of smart antennas in detail.

2. Smart antennas

First we should define smart (or adaptive) antennas. They are kinds of antenna types in cellular radio systems that confine energy in a narrow beam instead of broadcasting it over the whole cell. This radiation mode has many of advantages. For instance the signal gain is increased depending on the antenna radiation pattern more exactly the directivity. Range of the signal path is greater, spectral efficiency is improved, multipath reflection is reduced. Similarly to the last one, the level of interfering signals from adjacent cells or beams is also reduced. Due to these properties the network capacity increases, and this is the main effort in smart antenna systems [1].

But apart from advantages they also have drawbacks of course. The most important of these that beams have to be redirected continuously in order to tracking the angular position of mobile terminals situated in the cell is needed. The angular position usually called Direction of Arrival (DoA) in smart antenna systems [1]. The measurement and mainlobe positioning require some additional devices in the transceiver system, which are described in the followings.

Normally, the term "antenna" comprises only the mechanical construction transforming free electromagnetic (EM) waves into conducted radio frequency (RF) signals traveling on a shielded cable and vice versa. In this system it only fills the part of radiating element(s). In the context of smart antennas, the term "antenna" has an extended meaning. It consists of a number of radiating elements, phase shifters, a combining/dividing network and a control unit [3]. Concept of a smart antenna is shown in Figure 3.

In the left hand side of the figure the triangles represents the radiating elements. They can be for instance monopole or microstrip antennas. $W_{n,m}$ denote the weighting parameters and belong to the combining network. T_0 boxes are responsible for time

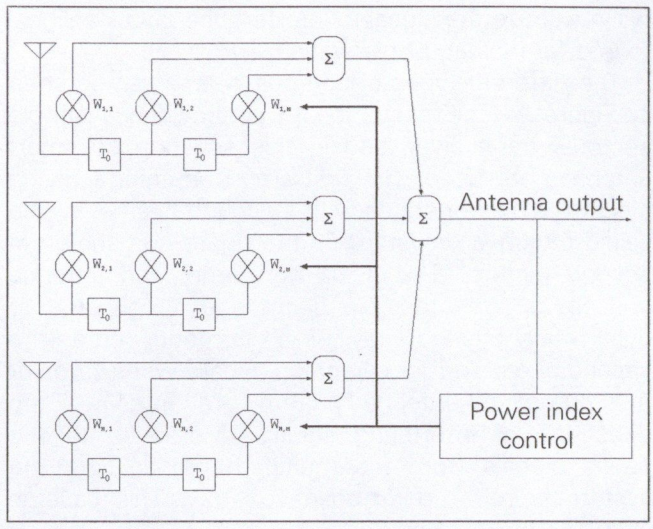


Figure 3. Smart antenna model

delay, which arises as phase shifting, so these units also called phase shifters. The control unit can be mentioned as intelligence of the smart antenna, and normally realized by using a digital signal processor (DSP). The processor controls feeder parameters of the antenna, based on several inputs, in order to optimize the communication link. Different optimization criteria can be used, as will be explained later.

The problem of positioning can be solved by different methods. Smart antennas can be grouped from this viewpoint. In the literature it is also mentioned as „level of intelligence" of smart antenna, and results in the following types [2]:

The switched lobe or also called switched beam is the simplest technique. It comprises only a basic switching function between separate directive antennas or predefined beams of an antenna array. The setting that gives the best performance, usually in terms of received power, is chosen. Because of the higher directivity compared to a conventional antenna, some gain is achieved. Such an antenna will be easier to implement in existing cell structures than the more sophisticated adaptive arrays, but it provides only limited improvement.

The next, improved and more difficult variation of smart antennas is the dynamically phased array. By including a direction of arrival algorithm for the signal received from the user, continuous tracking can be achieved and it can be viewed as a generalization of the switched lobe concept. The purpose is similar to the former one, the received power is maximized.

The third type is referred as adaptive arrays. In this case, a DoA algorithm to determine the direction toward interference sources (e.g., other users) is added. Then radiation pattern can be adjusted to null out the interferers. In addition, by using special algorithms and space diversity techniques, the radiation pattern can be adapted to receive multipath signals, which can be combined. These techniques

will maximize the signal to interference ratio (or Signal to Interference and Noise ratio (SINR)).

The different types of antenna beams can be seen in Figure 4. The sector beam is provided a simple sectorial transceiver, and in this situation no smart antennas are used. The first smart antenna applying system, which operates with switched beams, is also called multibeam system. In the figure it is shown in second picture. The next more improved antenna concept is the steerable beam and it concerns to dynamically phased array, which is mentioned as the second one in former paragraph. Finally when zeros or minimums of antenna pattern are fitted to the directions of interfering signals, moreover, to the useful signals from the multipath propagation, the system is referred as adaptive array. This case is shown in the last picture.

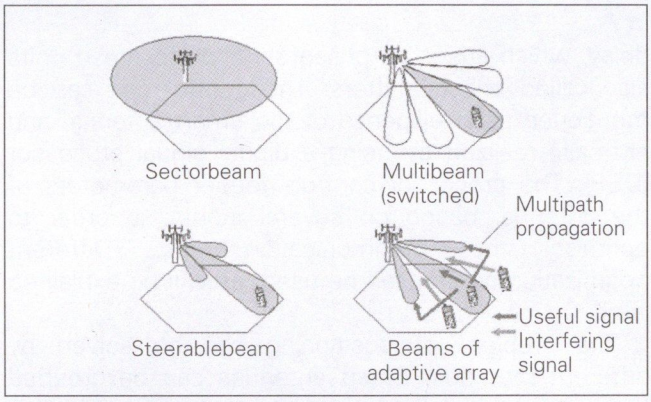


Figure 4. Possible beam types

All the levels of intelligence described beforehand are technologically realizable today. However, in the domain of personal and mobile communications, a kind of evolution can be foreseen in the utilization of smart antennas toward gradually more advanced solutions. The evolution can be divided into three phases [2].

Firstly smart antennas are used on uplink only. By definition we mean uplink that situation when the user is transmitting and the base station is receiving. Downlink means the opposite of it. By using a smart antenna to increase the gain at the base station, both the sensitivity and range are increased. This concept is called high sensitivity receiver (HSR) and does not differs in principle from the diversity techniques implemented in today's mobile communications systems.

In the second phase, directed antenna beams are also used in downlink direction in addition to HSR. In this way, the antenna gain is increased both in uplink and downlink, which implies a spatial filtering in both directions. Frequencies can be more closely reused, thus the capacity of the system is increased. The method is called spatial filtering for interference reduction (SFIR). It is possible to introduce this in second-generation systems.

The last stage in the development will be full space division multiple access (SDMA). This implies that more than one user can be allocated to the same physical communications channel simultaneously in the same cell, only separated by angle. Certainly, this solution generates some new problems that should be handled. For instance these can be problems of handover (handover occurs when the user is assigned to other transceiver), unmanageable number of users and beam-crossing. For steerable and adaptive array beams the measurement of DoA is required. For the second one the DoA stands for not only the desired signal but interfering signals, too. Thereinafter we are going to describe algorithms for the adaptive arrays, which are suitable for directivity measurement and noise suppression. Methods for steerable beams can be derived by simplifying the algorithms described here.

The main idea of determining the direction of the communicating transceiver is finding that angle to which belongs the maximal signal to noise ratio. It is achieved by scanning the whole space domain or the viewable part of it. Moreover, advanced algorithms can fit the zeros of the characteristic in directions of the interfering signals [5]. This is solvable if the number of antenna elements is greater or equal than the number of interferers. It should be noticed as a rule that direction of interference can be measured more exactly than direction of the signal because of having steeper descents around zeros. The signals from multipath propagation can also be taken into consideration. If they are useful then can be added to the main signal phase-correctly else zeros can be fitted to their direction or the harmful signals can be added up in anti-phase. The scanning operation is shown in first picture of Figure 5.

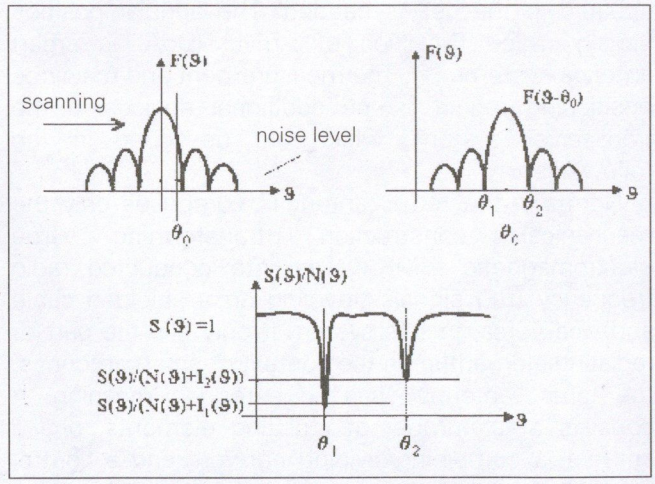


Figure 5. DoA estimating methods

$F(\vartheta)$ means the antenna pattern as a function of the angle ϑ. The procedure should be imagined as follows: the pattern scans in the direction shown by the arrow, meanwhile the θ_0 the DoA is fixed.

After the scanning there are several possibilities to calculate the adequate direction. Let's consider the methods in order of performance. Firstly we can mention the classical or Bartlett method, which can be seen in the second picture of Figure 5. It measures the two adjacent zeros of the characteristic which are closest to the mainlobe. Then it calculates the DoA as the mean of the two measured angles:

$$\theta_0 = \frac{\theta_1+\theta_2}{2}.$$

This can be explained by the fact that the characteristic changes rapidly near the zeros, so the estimation of direction is relatively exact. In spite of this fact, the function changes very slow near the top of the mainlobe, and the exact estimation based on measurement of this area, becomes impossible. Though this problem is eliminated, the Bartlett method has many disadvantages. It does not handle the interfering signals, thus in presence of them the measurement becomes unstable. In addition we can say this method is rather imprecise and the available relative angular resolution is poor. The measurable range is also small, the mainlobe-sidelobe ratio (or sidelobe suppression) is only 10-20 dB with the real assumption of existing 10 elements. Development of better methods was required.

One of these is the Capon-method [5] or also called MSINR (Maximizing Signal to Interference and Noise Ratio) -method. This utilizes the property of the antenna pattern that it has steep descent near the zeros, thus exact measurement is possible. Constant transmitter power density assumed to be (usually unity). Formation of SNR is shown in the third picture of Figure 5. Rapid changes in the function can be seen. $S(\vartheta)$ stands for the spectral density function of the transmitter, $N(\theta)$ for the noise and finally $I(\theta)$ for interfering signal. This method is better than the former one. The attainable measurability is about 30 dB in case of 20 dB SNR and the relative angular resolution is roughly 0.1, while 10 antenna elements supposed to be.

One another algorithm should be mentioned that is called MEM (Maximum Entropy Method). This is a more difficult method, but provides better performance. For the measurement and for determining the error function it uses a delay unit, prediction filter, and other devices as elements of its whitening filter. The whitening filter is required for whitening process [6]. The purpose of whitening is to transform the covariance matrix (created from the wanted signal, interference and noise) to a diagonal one. Thus the useful signal becomes distinguishable from the others. The available measurement range is about 60 dB and the relative angular resolution is 0.01. The conditions (SNR, elements) are the same as in case of methods discussed above.

Apart from these three algorithms a number of well-documented methods exist for estimating the DoA, for instance, MUSIC, ESPRIT, or SAGE [5]. Closing this section we give the estimating rules for the three mentioned algorithms, which are the follows:

Bartlett: $\hat{S}(\vartheta) = S^H(\vartheta)R_nS(\vartheta)$

Capon: $\hat{S}(\vartheta) = (S^H(\vartheta)R_n^{-1}\ S(\vartheta))^{-1}$

NEM: $\hat{S}(\vartheta) = |\ S^H(\vartheta)R_n^{-1}\delta\ |^{-2}$

Where H means the operator of transposition and conjugation, R_n is the covariance matrix of the noise and δ is the variance of the error function for the maximum entropy estimation.

3. Realization of smart antennas

After the general description of smart antennas, discussing pattern fitting algorithms and direction measuring methods we should deal with realization of the antennas since the lobe, we would like to work with, must be generated and formed somehow. The required equipment is an antenna array, each element with adequate feeders [2]. The array often has a relatively low number of elements in order to avoid unnecessarily high complexity in the signal processing. The geometrical parameters of the array can be various depending on the application.

The first two structures are used for beam-forming in the horizontal plane (azimuth) only. This is normally sufficient for outdoor environments, at least in large cells. The first is one-dimensional linear array with

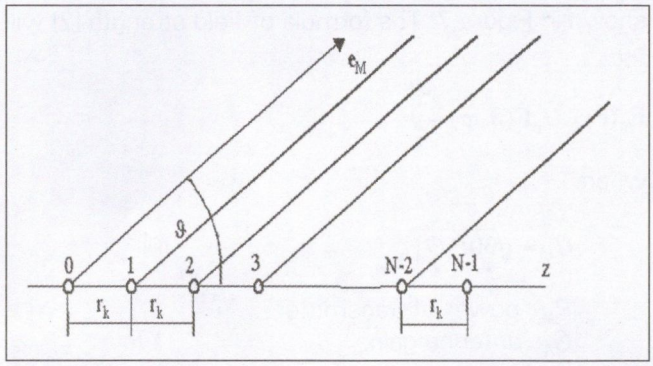


Figure 6. Antenna row

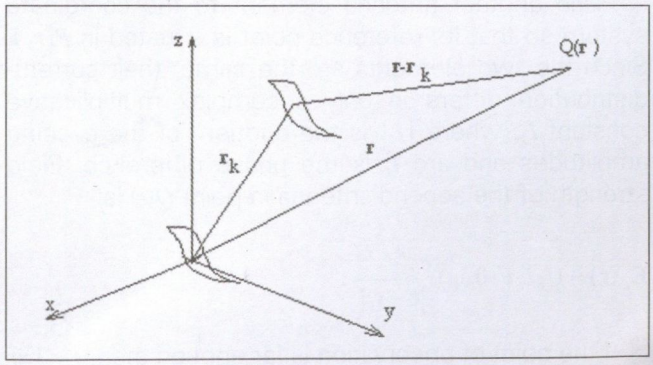


Figure 7. Model for calculations

uniform element spacing of rk. The arrangement is shown in Figure 6.

This structure can perform beam-forming in azimuth angle within an angular sector. This is the most common structure due to its low complexity (subsequently we will discuss them in detail). The second structure is the circular array with angular element spacing of

$$\Delta\Phi = \frac{2\pi}{N},$$

where N stands for the number of elements. This structure can perform beam-forming in all azimuth angles.

The other three structures are used for performing two-dimensional beam-forming, in both azimuth and elevation angles. This may be desirable for indoor or dense urban environments. The two-dimensional linear array structure can perform beam-forming within a solid angle. Beam-forming in the entire space, within all solid angles, requires some sort of cubic or spherical structure. In case of two- or tree-dimensional structures the element spacing can be different in each directions. As mentioned in this article the antenna rows will be discussed in detail because of the simplicity of them, and in addition the theory of higher dimension arrays can be traced back to theirs.

First of all we should define the model going to be applied. For the sake of simplicity in relations and formulas we assume that the distance of the observation point and the antenna is much greater than the linear dimension of the antenna system. Let's place the antenna into a (r,ϑ,φ) polar coordinate system and fit the first element into the origin. This situation is shown in Figure 7. The formula of field strength [2] will be:

$$\mathbf{E_0}(\mathbf{r}) = U_0 \mathbf{F}(\vartheta,\varphi)\frac{e^{-j\beta r}}{r},$$

where

$$U_0 = \sqrt{30 P_A G_A},$$

P_A power of transmitter,
G_A antenna gain,
F characteristic of each elements,
β phase coefficient.

Place another antenna element to the coordinate system so that its reference point is situated in $P(\mathbf{r}_k)$. Since the two elements are the same, their current-distribution differs in only a complex multiplicative constant I_k, where $|I_k|$ is the quotient of the exciting amplitudes and arc I_k is the phase difference. Field strength of the second antenna in point $Q(\mathbf{r})$ is:

$$\mathbf{E}_k(\mathbf{r}) = U_0 I_k \mathbf{F}(\vartheta,\varphi)\frac{e^{-j\beta|\mathbf{r}-\mathbf{r}_k|}}{|\mathbf{r}-\mathbf{r}_k|}.$$

If the point of observation is far enough then $|r\text{-}r_k| \cong r$ in the denominator, while in the exponent:

$$|\mathbf{r}-\mathbf{r}_k| \cong r - \mathbf{r}_k \cdot \mathbf{e}_r,$$

where er is unity vector in direction of r. Utilizing it the first formula can be rewritten:

$$\mathbf{E}_k(\mathbf{r}) = \mathbf{E}_0(\mathbf{r}) I_k e^{j\beta\,\mathbf{r}_k\cdot\mathbf{e}_r}.$$

If we deploy N antennas in the space radiating the same frequency, the resultant field strength will be:

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_0(\mathbf{r})\sum_{k=0}^{N-1} I_k e^{j\beta\,\mathbf{r}_k\cdot\mathbf{e}_r},$$

so

$$\mathbf{E}(\mathbf{r}) = \mathbf{E}_0(\mathbf{r})\cdot F_i(\vartheta,\varphi).$$

This means that function of field strength of an antenna system – elements of which radiates parallel and are equivalent – can be written as multiplication of field strength of the element situated in origin and another coefficient, which is only a function of spatial arrangement and excitement (and independent from element properties) [2]. The latter multiplicator is called directivity and its formula:

$$F_i(\vartheta,\varphi) = \sum_{k=0}^{N-1} I_k e^{j\beta\, r_k\cdot\cos\alpha_k},$$

where

$$\cos\alpha_k = \mathbf{e}_k\cdot\mathbf{e}_r = \sin\vartheta\cdot\sin\vartheta_k\cdot\cos(\varphi-\varphi_k)+\cos\vartheta\cdot\cos\vartheta_k,$$

$$\mathbf{e}_k = \frac{\mathbf{r}_k}{r_k},$$

$$\mathbf{r}_k = |r_k|.$$

The formula of directivity is still valid for three-dimensional arrangements. If we would like to prove maximal field strength in a direction $\mathbf{e}_M$ with given amplitude distribution, then it achievable by adequate phase distribution. One possible condition for it:

$$\delta_k = -\beta\, r_k \cos\alpha_{Mk},$$

where

$$\cos\alpha_{Mk} = \mathbf{e}_k\cdot\mathbf{e}_M = \sin\vartheta_M\cdot\sin\vartheta_k\cdot\cos(\varphi_M-\varphi_k)+\cos\vartheta_M\cdot\cos\vartheta_k.$$

Fulfilling this requirement we get the travelling wave phase distribution [2]. The main point is that in a favored direction the phase difference caused by the element distance is compensated by the phase difference of the feeder current. The directivity in the direction of $\mathbf{e}_M$

$$F_i(\vartheta_M,\varphi_M) = \sum_{k=0}^{N-1}|I_k|$$

and in any direction

$$F_i(\vartheta,\varphi) = \sum_{k=0}^{N-1}|I_k|e^{j\beta\, r_k\cdot(\cos\alpha_k-\cos\alpha_{Mk})}.$$

We should mention that travelling wave feeding proves maximal field strength in case of given sum of current, but in case of given sum of power the phase criteria for feeding of elements will change.

Suppose that we would like to use an adaptive antenna row. We can have a question: which requirements should be met in the arrangement? In this respect we should examine the possible distance of elements, does it have minimum or maximum? The answer is that the maximal distance is limited. To explain it, at first, the domain of visibility should be dealt with. Let's take out the δ phase shift from I_k coefficient of excitement and initiate the next new variable:

$$\Psi=\delta+\beta d\cos\vartheta$$

where d denotes the distance of elements. The physical meaning of variable Ψ: phase difference between remote field strengths generated by two adjacent elements of the row, which has components arising out of both geometry (βd) and feeding (δ). The modified formula is:

$$F_i(\Psi)=\sum_{k=0}^{N-1} I_k e^{jk\Psi}.$$

By initiation of this transformation the formula of directivity becomes independent from δ and d parameters, and can be given by a more general function $F_i(\Psi)$ having periodicity of 2π. Using the same $F_i(\Psi)$ but different choice of δ and d, many kind of $F_i(\Psi)$-s can be created. Therefore we should examine the relation between $F_i(\Psi)$ and $F_i(\vartheta)$ in each situation. The first question is that which domain of Ψ belongs to the physically understandable ϑ=0-180° domain (the directivity is symmetrical to the ϑ=0 axis). $\Psi(0)=\delta+\beta d$ belongs to ϑ=0, $\Psi(90)=\delta$ to ϑ=90° and $\Psi(180)=\delta-\beta d$ to ϑ=180°. The domain of variable Ψ which belongs to ϑ=0-180° is called domain of visibility. In accordance with it width of domain of visibility is $2\beta d$ and the centre of it is d. If the distance between elements is smaller than the half of the wavelength ($d<\lambda/2$) then Ψ has a domain out of visibility, while if it is larger ($d>\lambda/2$) then some sections of $F_i(\Psi)$ repeats in $F_i(\vartheta)$. In case of this last one the appearance of protruding sidelobes is quite detrimental, which is a consequence of that the side of visibility domain around Ψ=0 hangs into the mainlobe positioned at Ψ= -2π or 2π. Protruding sidelobes are avoidable in case of adequate choice for distance of elements which means that the domain of visibility should be extended only the to the zeros bordering the mainlobes at $\pm 2\pi$ [5]. The condition for it:

$$\frac{d}{\lambda}=\frac{N-1}{N(1+|\cos\vartheta_M|)}.$$

Finally we should mention another interesting fact, that an adaptive array can be imagined as a FIR (Finite Impulse Response) filter in space domain. In compliance with it some analogues can be declared. The next concepts can be paired (in time domain – space domain): time – distance, frequency – frequency in space domain, correlation – correlation in space domain, spectrum – spectrum in space domain, frequency filter – spatial filter. From these analogues some interesting formulas can be derived. For example in this context the Shannon's sampling law is also valid, thus we can state that the maximal distance of elements is the half of the wavelength ($d\leq\lambda/2$). After negotiation of adaptive antenna rows in general, the main types of them we are going to discuss in the next section.

4. Types of antenna rows

Adaptive antenna rows are distinguished by spatial arrangement of the elements and their feeding. In practice equidistant arrays (distance between any adjacent elements is the same) are used and arrays with different distance between elements are not. Because of it we only deal with the equidistant arrays, which are the simplest antenna systems. As it comes from the arrangement, if the maximum of function F_i is set by travelling wave current distribution, then it means constant phase difference between adjacent elements, in other words along the row in direction of maximal radiation we get monotonously increasing (progressive) phase shift [2]. It has many advantages in practice. The progressive phase shift can be formulated as follows:

$$\text{arc } I_k - \text{arc } I_{k-1} = \delta.$$

Let arc I_0 =0, thus

$$\text{arc } I_k = k\delta,$$

where

k=0,1,2...N-1

The travelling wave progressive phase shift for the main direction ϑ_M:

$$\delta=\beta d\cos\vartheta_M.$$

Two of main directions have stressed significance. If ϑ_M=0° or 180° then the row is called endfire. The condition for it in case of travelling wave feeding

$\delta=-\beta d$ ha ϑ_M=0°

and

$\delta=\beta d$ ha ϑ_M=180°.

The other row is called broadside. Main direction of it is ϑ_M=90°. The condition for if: δ=0.

Antenna rows can also be distinguished by feeding [2]. The identically distributed amplitudes are not only simple but having great importance in practice, since it can be easily realized. Utilizing this amplitude distribution in case of large number of elements the level of first sidelobe (or sidelobe suppression) is 2/3π (-13.5 dB) if the level of mainlobe is set to be 1. Decrease of sidelobes is in direct proportion to 1/Ψ. In case of identically distributed amplitudes the shape of directivity and width of mainlobe can be set by adequate parameter setting.

But the level of sidelobes cannot be decreased arbitrary by changing d or δ. The sidelobe suppression

can be (intuitively) increased by generating a row of which directivity is equals the square of directivity of identically distributed amplitude row. Thus we get a row with triangular distribution. In case of large number of elements the level of sidelobe converges -27 dB, and the mainlobe is much narrower. But twice as many elements are required! The function near its zeros is roughly parabolic, so the function is rather flat. This fact can be useful if we want to eliminate an interfering transmitter in a direction and wouldn't like the antenna to be sensitive for setting.

As the number of sidelobes are determined by zeros of $F_i(\Psi)$ situated in domain of visibility, so there is only mainlobe in $F_i(\vartheta)$ is all of zeros are placed out of visibility. In case of $d=\lambda/2$ the sidelobes can be avoided by feeding binomially. This means that the coefficients of the row are members of binomial (mathematical) row, which can be calculated utilising the Pascal-triangle.

In case of distributions having decreasing amplitudes toward the ends, like binomial and triangular, the sidelobes are smaller than in case of identical distribution. Price of decrease in sidelobes is widening in the mainlobe and decline in directivity. Improvement can be achieved by increasing the level of the far and smaller sidelobes [2]. As we reach equality in level of sidelobes, we get optimal solution since the width of the mainlobe will be the smallest one with given sidelobe-level. For the solution the width of mainlobe or the sidelobe suppression is prescribed. The mathematical drafting of the task is similar to the case of elliptical filters.

The optimal antenna row is firstly worked out by Dolph for broadside rows having element distance of $\lambda/2$ and he showed that the problem can be solved by using Chebishev polynomials (which can be classified as orthogonal polynomials), so these antenna rows are called Dolph-Chebishev rows. If $d<\lambda/2$ then the method does not give optimal solution. If $\delta= -\beta d$, i.e. we would like to create endfire row the protruding sidelobe can be avoided if only $d<\lambda/4$, which is in the domain of super gain, thus among other errors it gives a quite tolerance sensitive solution. Therefore the Dolph-Chebishev method is not applied for measuring out endfire rows, but it is widespread method for synthesis of broadside rows.

5. Summary

The main purpose of this article has been to provide an overview of smart antenna technologies, which are promising for third-generation land mobile systems, like UMTS. It is obvious that smart antenna technology is important for providing the necessary capacity and coverage. In addition to existing method, dividing the space into cells, it is now also possible to employ space division inside each cell. The most important smart antenna technologies, which involve switching between lobes as much as advanced algorithms maximizing the SNR, has been described here. The main types of smart antennas and methods for beam-forming with them are also overviewed.

References

1. Sören Andersson, Bengt Carlqvist, Bo Hagerman and Robert Lagerholm: Enhancing cellular network capacity with adaptive antennas, Ericsson Review No. 3, 1999
2. Per H. Lehne, Magne Pettersen: An overview of smart antenna technology for mobile communications systems, IEEE Communications Surveys, Fourth Quarter 1999, vol. 2 no. 4
3. Seshaiah Ponnekanti, An overview of smart antenna technology for heterogeneous networks, IEEE Communications Surveys, Fourth Quarter 1999, vol. 2 no. 4
4. Pap László, Imre Sándor: Az interferencia elnyomásának módszerei mikrocellás rádióhálózatokban, MTA Tudomány Napja 2001
5. Joseph C. Liberti Jr., Theodore S. Rappaport: Smart antennas for wireless communications, Prentice Hall, 1999
6. Bernard Widrow, Samuel D. Stearns: Adaptive signal processing, Prentice Hall, 1985

Performance Evaluation of UMTS Terrestrial Radio Access Networks

SZABOLCS MALOMSOKY, SZILVESZTER NÁDAS, BALÁZS SONKOLY

Traffic Analysis and Network Performance Laboratory, Ericsson Research
Laborc u. 1., Budapest, H-1037, Hungary

On transport links of UMTS terrestrial radio access networks (UTRAN) but especially on the Iub interface, which connects base stations and radio network controllers, resource allocation is complex, because packet delay and loss requirements are strict and the amount of transmission resources is relatively low. In this paper, link admission control (LAC) methods applicable in UTRAN nodes are proposed. Both FIFO and priority scheduling are considered. Connection admission control (CAC) decisions in the network are taken based on independent LAC decisions along routes assigned to connections. Therefore, we also investigate end-to-end QoS within the network. Multiplexing efficiency of various ATM/AAL2 switching (traffic concentration) methods between the base station and the radio network controller are analyzed. Our primary goal is to investigate ATM/AAL2 based UTRAN, but the proposed model and analysis can also be applied to IP based network infrastructures.

1. Introduction

Switching and multiplexing technologies used for the first releases of UTRAN are mainly based on Asynchronous Transfer Mode (ATM) in combination with the ATM Adaptation Layer type 2 (AAL2) [4][7]. Future releases will be deployed using also IP (Internet Protocol) technologies [2]. Resource management is a key problem in the radio access network, where a large amount of transport infrastructure is needed to deliver traffic to (or from) a possibly large number of base stations.

Some recent papers have dealt with resource management in UTRAN. The application of ATM/AAL2 in UTRAN is explained in [7]. Performance and QoS issues related to AAL2 switching networks are discussed, and different multiplexing scenarios are evaluated by simulation. Simple AAL2 link admission control (LAC) methods using the Chernoff-bound are evaluated in [8]. A method for dimensioning of voice carrying network links is described in [9]. In [10] and [11], bandwidth management in AAL2 networks and the performance of AAL2 cell assembly and disassembly are analyzed. Based on conclusions of these papers, it is proposed in [12] to use the UBR (Unspecified Bit Rate) VC (Virtual Channel) switching alternative in UMTS.

Motivations of our work are the following:

- to establish an adequate model for the UTRAN transport network (compared to [7] and [8], the Iub specific traffic behavior is modeled),
- to develop LAC algorithms for both FIFO and priority scheduling, and to validate them using simulations and mathematical analysis based on e.g., [14], [16] and [18],
- to evaluate whether a series of LAC decisions is able to fulfil end-to-end QoS requirements, and
- to investigate performance of switching alternatives, similarly to [12], but using a novel methodology, which enables to analyze large UTRAN networks with typical traffic mixes and realistic admission control.

The paper is organized as follows. Section 2 presents the system architecture. In Sections 3 and 4, the *single link* case is considered. Section 3 presents the related queuing model. In Section 4, LAC algorithms are proposed for FIFO and priority scheduling, and numerical examples are presented, which demonstrate their applicability. Section 5 considers the *multiple link* case. End-to-end QoS within the network is investigated, and the performance of various ATM/AAL2 switching (traffic concentration) methods are evaluated. The paper is concluded in Section 6.

2. System architecture

In this section the UTRAN system is described and ATM/AAL2 switching alternatives applicable in UTRAN are introduced.

2.1. UTRAN System

The UMTS network architecture [6] is depicted in Figure 1. An UMTS network consists of the user equipment (UE), the UMTS terrestrial radio access network (UTRAN) and the core network (CN). Main

elements of UTRAN are base stations (Node Bs) and radio network controllers (RNCs).

UTRAN handles all tasks related to radio access, such as radio resource management, handover control, etc. The core network is the backbone of UMTS connecting the access network to external networks (e.g. PSTN, Internet). The mobile (UE) is connected to base stations (Node Bs) over the WCDMA (wide-band CDMA) radio interface (Uu). One UE can communicate with several Node Bs at the same time during soft handover.

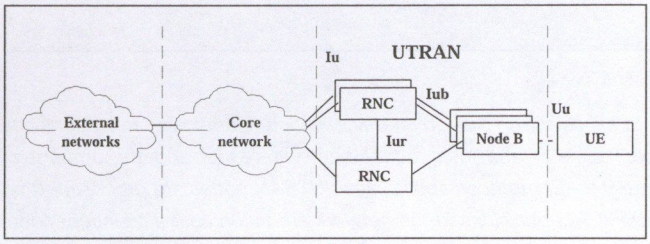


Figure 1. UMTS network architecture

The ATM/AAL2 based protocol stack of UTRAN is depicted in Figure 2. The retransmission mechanism of the radio link control (RLC) protocol ensures reliable transmission of loss-sensitive traffic over the radio interface. The medium access control (MAC) protocol forms radio frames and schedules these periodically according to the timing requirements of the radio interface. This period is called TTI (transmission time interval), and its length can be a multiple of 10 ms. Bitrates of radio connections (so called RABs – radio access bearers) take typical values between 8 kbps and 384 kbps. MAC frame sizes and TTI lengths are RAB-specific. Considering the simplest case, when a user uses a single service, one RLC and one MAC entity are created in the RNC (and the peer entities in the UE) for each actually connected UE.

In the transport network, MAC frames are encapsulated into Iub frames. Iub frames are segmented and packed into AAL2 CPS (common part sublayer) packets, which are multiplexed into ATM cells. AAL2 payload can be of variable length (up to 45 bytes), and the AAL2 header is 3 bytes long. ATM cells are 53 bytes long including a 5 bytes long header. By AAL2 multiplexing, several AAL2 packets from different connections can be carried within an ATM cell. In an ATM network, cells are transported along predefined paths using the VPI/VCI (virtual path and virtual channel identifier) fields in the ATM header. The CID (connection identifier) field in the AAL2 header identifies a specific AAL2 connection within an ATM VC, and the VCI identifies a VC in a VP. In the AAL2 multiplexer, the so called CU timer (T_{CU}) determines how long the multiplexer should wait for arriving AAL2 packets if an ATM cell is partly filled. Therefore, multiplexing efficiency also depends on the value of T_{CU} [11]. In this paper, we focus on situations when the transport links are highly loaded. In this region, the effect of T_{CU} on delay performance is expected to be negligible (see also [1]).

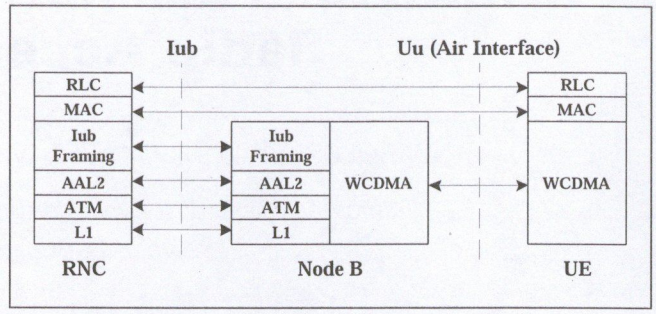


Figure 2. ATM/AAL2 based UTRAN protocol stack

In UTRAN, a new AAL2 connection is set up for each new RAB. AAL2 CAC allocates resources for the new AAL2 connections in the transport network. It makes its decisions based on traffic descriptors and QoS parameters.

On Iub, to decrease the probability of packet congestion in transmission network queues, start positions of frames intended for different UEs should not coincide in time. Therefore, phases of the periodic frame flows of different connections are randomly distributed over the TTI. The user/application level traffic process is reflected in the UTRAN transport network such that the carried traffic is not a continuous periodic packet flow, but can be modeled by a series of active and inactive intervals. We will refer to these intervals as ON and OFF periods, respectively. In an ON period MAC frames are sent in each TTI, while in an OFF period packets are not sent at all. For example, in case of voice traffic the characteristics of the ON and OFF periods are determined by the interaction of the speech process (the speaker behavior) and the voice activity detector in the voice coder. Admission decisions are made based on *traffic descriptors*, which are sent in signaling messages between transport nodes. The traffic descriptors associated with each connection are the following: (Iub) frame size (we will also call it packet size), TTI (the inter-arrival time of frames), and the so called "activity factor". The activity factor is a number between 0 and 1, and it is defined as the average length of ON periods divided by the sum of the average lengths of ON and OFF periods. (It means that reliable information on the distributions of the lengths of ON and OFF periods are not available for the LAC.) In UMTS, the following TTI values are possible: 10, 20, 40 and 80 ms. For example, typical parameters for voice service can be the following: 40 bytes long packets arriving in 20 ms periods (TTI) with activity factor 0.6. Note that the activity factor is an effective value, which is set by the operator in order to exploit statistical multiplexing gain.

Considering *QoS requirements*, for several reasons, the packet delay is the most important performance measure in the transport network. For example, for

voice traffic the phone-to-phone (or phone-to-gateway) delay budget determines the amount of tolerable queuing delay in the UTRAN transport network, which is around 5-7 ms [1]. The delay requirements for other services are not very different from that of voice. If a UE has simultaneous RABs to two or more Node Bs (during soft handover), the radio frames scheduled in downlink have to be sent out from every Node B to the UE at the same time (t_{out}). Therefore, nodes must be synchronized. For the same reason, it has to be ensured that each frame arrives to the Node Bs before t_{out}. This determines a delay requirement on the UTRAN transport network. It means that even for best-effort services there are delay requirements in UTRAN. Moreover, since the round-trip time of RLC packets should be minimized in order to maximize the throughput of best-effort traffic, these delay requirements are strict [13]. Therefore, short buffers are applied in the system, and thus mainly short time-scale traffic fluctuations can be absorbed by buffering. We assume that the end-to-end (Node-B-to-RNC) delay requirement associated with a RAB is smaller than (or equal to) the TTI characterizing the RAB.

If packets of all traffic classes (classification is done based on traffic descriptors) wait in the same queue and the packets are served in the order of arrival (first-in-first-out, FIFO), the most stringent delay requirement (typically the delay requirement of voice) has to be fulfilled. This can be avoided by QoS differentiation, where we have separate queues for traffic classes with different delay requirements. Capability Set 1 of the AAL2 signaling protocol does not support *QoS differentiation* (i.e., QoS based separation) of AAL2 connections, while Capability Set 2 enables the selection of the carrying VC according to the requested QoS [4]. This way, AAL2 connections with different QoS requirements can be transported in separate ATM VCs.

2.2. ATM/AAL2 switching alternatives

In this paper, the following ATM/AAL2 switching alternatives are discussed (see Figure 3.):

A *Traffic aggregation using ATM VC switches and CBR VCs:* In this alternative we have end-to-end CBR (Constant Bit Rate) VCs, LAC is only performed at the end-point, where the AAL2 multiplexing is done. VC capacities need to be defined such that the CID-limit (max 248 AAL2 connections can be multiplexed in a VC) is taken into account. By choosing too large VCs, the CID limit may bar full use of the VC capacity (having many narrow-band connections). By choosing too small VCs, packet-level statistical gains are decreased and also granularity problems arise.

B *Traffic aggregation using AAL2 switches and CBR VCs:* This alternative is similar to alternative A, but there are no end-to-end VCs. It means that in a concentration node AAL2 connections can be switched from one VC to another, and AAL2 LAC needs to be performed in each AAL2 switch.

C.1 *Traffic aggregation using ATM VC switches and UBR VCs (but CBR VPs):* This is similar to alternative **A**, but UBR VCs are used. This alternative is proposed in [12]. The CID limit is not a resource related limit here, because there is no bandwidth associated with the VCs. However, using this alternative, AAL2 signaling messages have to be processed by ATM switches, because AAL2 LAC needs to be done in each concentration node over the VP resource (see also [12]). The performance of this alternative is comparable to that of C.2. In fact our model is the same for C.1 and C.2.

C.2 *Traffic aggregation using AAL2 switches and UBR VCs (but CBR VPs):* This is similar to alternative **B**, but UBR VCs are used. The difference from C.1 is that in the switching points AAL2 LAC (over the VP resources) can easily be done, because there are AAL2 switches there. This alternative is expected to have the best performance from statistical multiplexing point of view.

We have the following comments on using rt-VBR (real time-Variable Bit Rate) VCs instead of CBR VCs: Resource allocation of AAL2 connections over rt-VBR VCs is more complex than over CBR VCs. It is also a problem that in AAL2 traffic descriptors the burstiness of the traffic sources is not characterized. It means that assumptions on burstiness have to be used, and it is difficult to find the most cost-effective rt-VBR parameters. In this paper we do not consider rt-VBR in more detail, for an analysis of multiplexing AAL2 connections over rt-VBR VCs see [21].

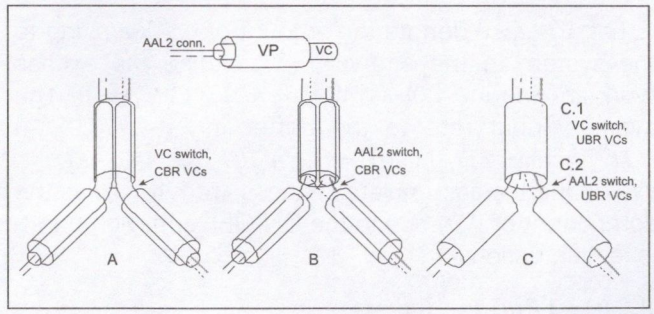


Figure 3. ATM/AAL2 switching alternatives in UTRAN

3. Queuing model

In this section, a queuing model considering a single link with FIFO or priority scheduling is proposed.

3.1. FIFO Scheduling

Figure 4 shows an example system with two connections from different traffic classes. We are interested in the probability of the packet delay

criterion violation: $\Pr\{D_i > \tilde{D}_i\}$, where D_i is a random variable representing the delay of a packet from class i, and $\tilde{D}_i$ is the delay criterion (or target maximum delay) of packets from traffic class i. The total delay of a packet from class i consists of two parts: $D_i = Q_i + S_i$, where S_i is the service time and Q_i is the waiting time in the queue. Packet losses are considered as infinite delays. In the proposed model, we consider delays

- due to the ON-OFF behavior, which results in temporary system overloads and
- due to the periodic packet emission during ON states, where the emission phases associated with the connections are random, which can result in packet congestion.

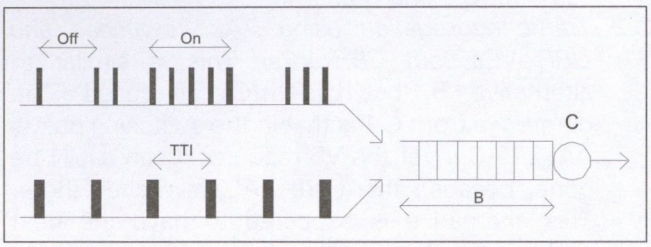


Figure 4. System fed by two periodic ON-OFF connections with different TTIs and packet sizes

Denote the number of traffic classes by K. The number of connections present in the system from class i is N_i. Connections within the same class are characterized by the activity factor α_i, the packet size b_i and the packet inter-arrival time TTI_i. The server capacity is denoted by C. The allowed delay criterion violation probability for class i is $\tilde{\varepsilon}_i$, meaning that the LAC should ensure that $\Pr\{D_i > \tilde{D}_i\} > \tilde{\varepsilon}_i$ for $i=1,\ldots,K$.

Decomposition of burst-scale and packet-scale effects

Let $A(t)$, $t \geq 0$ denote the amount of work arriving to the system in the interval $[-t,0)$. Define the excess work arriving in $[-t,0)$ as: $W(t) = A(t) - C\,t$, $t \geq 0$. The average input rate to the buffer in $[-t - TTI^{max}, -t)$ $(TTI^{max} = \max_i TTI_i)$ is: $R(t) = (A(t + TTI^{max}) - A(t)) / TTI^{max}$. We define the accumulated excess work $W_{acc}(t)$ as the component of $W(t)$ due to the ON-OFF behavior (burst-level fluctuations):

$$W_{acc}(t) = \int_0^t R(u)\,du - C\,t\,.$$

The evolution of $W_{acc}(t)$ depends on the distribution of the ON and OFF periods, the dependency among the sources, etc. It can be positive only if the average input rate $R(t)$ is larger than the server rate C.

In a FIFO queue the waiting time is approximated using the *workload* (or virtual waiting time) [16]. The system is stationary so that 0 represents an arbitrary time instant. The workload is calculated as: $V(0) = sup_{t\geq 0}\, W(t)$. The "burst component" of the workload, i.e., the component considering only $W_{acc}(t)$ is: $V^{burst}(0) = sup_{t\geq 0}\, W_{acc}(t)$. Then, the "packet component" can simply be defined as the difference of the workload and its burst component: $V^{packet}(0) = V(0) - V^{burst}(0)$.

We are interested in the complementary distribution function of $V(0)$: $Q(x) = \Pr\{V(0) > x\}$.

The queuing process can be decomposed as follows:

$$Q(x) = \Pr\{V^{packet}(0) + V^{burst}(0) > x \mid V^{burst}(0) > 0\} \cdot \Pr\{V^{burst}(0) > 0\} + \Pr\{V^{packet}(0) > x \mid V^{burst}(0) = 0\} \cdot \Pr\{V^{burst}(0) = 0\} \quad (1)$$

In [14] it is shown that if the burst component is positive, then, in general, the expectation and the variance of the packet component $\boldsymbol{E}\{V^{packet}(0) \mid V^{burst}(0) > 0\}$ and $\boldsymbol{Var}\{V^{packet}(0) \mid V^{burst}(0) > 0\}$ are rather small compared to the value of the burst component, $V^{burst}(0)$. Therefore, the following approximation can be used:

$$\Pr\{V^{packet}(0) + V^{burst}(0) > x \mid V^{burst}(0) > 0\} \cdot \Pr\{V^{burst}(0) > 0\} \approx \Pr\{V^{burst}(0) > x\} \quad (2)$$

Note however, that the traffic descriptors α_i, b_i and TTI_i do not include any characterization of the ON and OFF period lengths. It means, that the distribution of the burst component of the workload $\Pr\{V^{burst}(0) > x\}$ can not be evaluated.

Since our system has strict delay requirements $(\tilde{D}_i > TTI_i)$, the waiting time in an overload situation (when $R(t) > C$ during a time interval) reaches very fast the predefined delay criterion. In other words, unless the overload situations are rather short, the queue can not smooth out the temporary overload efficiently, even if considering infinite buffer. Therefore, we take the conservative assumption that the delay of each packet, arriving in an overload situation, is always larger than the delay criterion. Using this assumption, instead of $\Pr\{V^{burst}(0) > x\}$ we will evaluate the probability that a packet arrives at an overload situation, and by this we will approximate the packet loss probability.

When the burst component is zero, $V^{burst}(0) = 0$, the queue is emptied periodically (with period TTI^{max}), and the distribution of the workload is determined by the cell emissions in a short interval before considered time 0.

The proposed model

Applying the assumptions, we set up a combined model. Considering packets in the system we can observe two types of delay criterion violation events; some packets are lost due to buffer overflow and some packets are only exceeding the delay criterion. We define two measures as

$$\hat{a}_i^{lost} = \frac{\#\text{lost packets}}{\#\text{packets}} \quad \text{and} \quad \hat{a}_i^{delayed} = \frac{\#\text{delayed packets}}{\#\text{packets}}.$$

Denote the number of active connections (the connections in ON period) at time t of class i by $N_i^{act}(t)$

and let the vector of active connections at time t be $\underline{N}^{act}(t)=[N_1^{act}(t),N_2^{act}(t),...,N_K^{act}(t)]$. At a fixed time t_0, we say that the system is in state $\underline{n}$ if the random vector $\underline{N}^{act}(t_0)$ takes the value $\underline{n}$ (i.e., $N_i^{act}(t_0)=n_i$; $i = 1, 2,..., K$). The probability, $\Pi_i(n_i)$, that the number of active connections from class i is n_i (i.e., $N_i^{act}(t_0)=n_i$) can be obtained with a binomial distribution, while the probability that the system is in state $\underline{n}$, denoted by $\Pi\underline{n}$, is calculated using a multi-dimensional binomial distribution as follows:

$$\Pi(\underline{n}) = \prod_{i=1}^{K} \Pi_i(n_i) = \prod_{i=1}^{K} \binom{N_i}{n_i} \alpha_i^{n_i} (1-\alpha_i)^{N_i - n_i}$$

We define the load of an active connection as the packet size divided by the period length: $\rho_i = b_i/TTI_i$. Then the input rate in a state is:

$$R(\underline{n}) = \sum_{i=1}^{K} n_i \rho_i$$

The probability that a packet of class i arrives at an overload situation (when $R(\underline{n}) > C$) can be calculated as:

$$\varepsilon_i^{lost} = \Pr\{\text{packet arrives at overload situation}\} =$$

$$= \frac{\sum_{\underline{n}:R(\underline{n})>C} n_i \Pi(\underline{n})}{\sum_{\forall \underline{n}} n_i \Pi(\underline{n})} \approx \hat{\varepsilon}_i^{lost} \qquad (3)$$

In case of a normal situation $(R(\underline{n}) > C)$ the waiting time is dominated by the periodic packet emission. Thus the probability that a packet, arriving at time t0, is exceeding its delay criterion, but does not get lost can be calculated as:

$$\varepsilon_i^{delayed} = \frac{\sum_{\underline{n}:R(\underline{n})\le C} n_i \Pi(\underline{n}) \cdot \Pr\{D_i > \tilde{D}_i \mid \underline{N}^{act} = \underline{n}\}}{\sum_{\forall \underline{n}} n_i \Pi(\underline{n})} \approx \hat{\varepsilon}_i^{delayed}. \qquad (4)$$

Finally, similarly to the decomposition in Eq.(1), the probability of delay criterion violation is the sum of two probabilities: $\varepsilon_i = \varepsilon_i^{lost} + \varepsilon_i^{delayed}$.

To check the assumptions of the proposed model, we simulated different ON-OFF sources. Figure 5 compares delays of Markov modulated sources with average ON period lengths of 0.4 sec, 4 sec and 40 sec to the result with the proposed model. Up to $\tilde{D} \approx 10$ ms the delays hardly depend on the length of ON periods, and delay violations are dominated by the periodic packet emission. For larger $\tilde{D}$ values the proposed model is conservative and the delay violation is mainly caused by too many active sources filling up the buffer.

3.2. Priority Scheduling

Prioritization means that a packet from a lower priority queue can be served only if all the higher priority queues are empty. Segmentation is used to minimize the influence of large low priority packets, which are already in the server, on high priority traffic. The segment size s is an additional model parameter. Q_i and S_i apply to the last segment instead of the whole packet. An important difference from the FIFO case is that Q_i cannot be calculated directly using the workload of the system because higher priority packets can overtake lower priority packets. When calculating ε_i^{lost} and $\varepsilon_i^{delayed}$, the only difference from the FIFO case is that from the class i point of view the system is overloaded only if the input rate of traffic classes with higher (or equal) priority is higher than C.

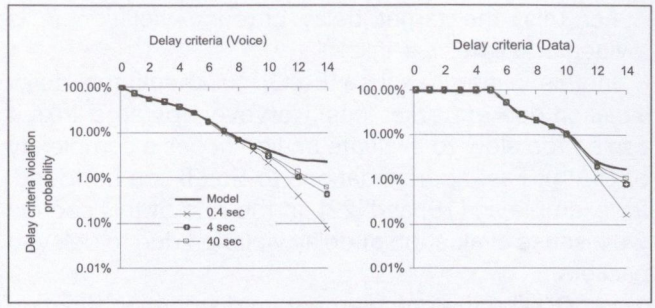


Figure 5. Simulation of delays on a link of capacity C=520 kbps with 15 ms long buffer fed by 30 voice (TTI_1 = 20 ms, b_1 = 40 bytes, α_1 = 0.6) and 2 data (TTI_2 = 40 ms, b_2 = 360 bytes, α_2 = 1) connections

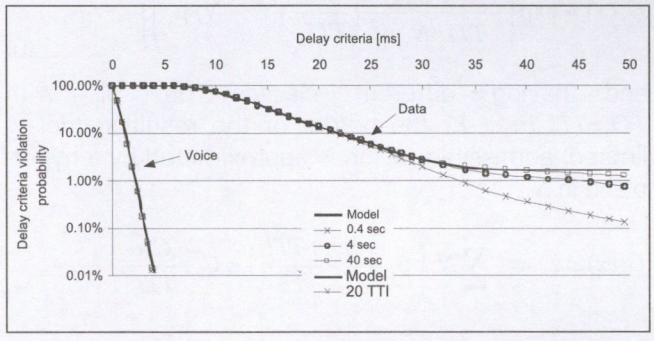


Figure 6. Simulation of delays. Model parameters are the same as in Figure 5, s = 47 bytes, voice has higher priority than data

Similarly to the FIFO case the delay violation events are: lost and delayed packets. Simulation results validating the model in case of prioritization are depicted in Figure 6. We can observe that the model is conservative and that the delay criterion violation probability for voice is much smaller than in the FIFO case.

4. LAC algorithm for FIFO and priority scheduling

The task of the LAC algorithm is to check on-line whether a certain traffic mix is within the admissible region. In other words, when a new connection arrives, the LAC needs to check:

- the delay violation due to packet loss,
- the delay violation due to delayed packets.

For this, the target delay criterion violation $\tilde{\varepsilon}_i$ is divided into $\tilde{\varepsilon}_i=\tilde{\varepsilon}_i^{lost}+\tilde{\varepsilon}_i^{delayed}$.

In this paper, we use Eq.(3) to check the delay violation due to packet loss. However, for large links it can be too slow to evaluate on-line (it has a complexity of $O(N^K)$). Fast approximations to Eq.(3) can be found, for example, in [19] and [22]. In the rest of this section we discuss evaluation of delay violation due to delayed packets.

Let $Q(x)$ be the complementary distribution function of the workload in a FIFO queue. The arrival process has stationary increments, therefore $Q(x)$ can be obtained using a general queuing theory result [16]:

$$Q(x)=\Pr\left\{V^{packet}(0)>Cx\right\}=\Pr\left\{\sup_{\tau\geq 0}\left(A(\tau)-C\tau\right)>Cx\right\}. \tag{5}$$

Since only traffic mixes with $R(\underline{n})\leq C$ are considered, which always results that the queue empties periodically; we only regard $\tau\in[0,TTI^{max}]$. When several independent periodic sources from class i, with $\alpha_i=1$ are superimposed, the resulting arrival process can be approximated by a Brownian bridge [18]. Using the Brownian bridge approximation enables us to obtain the solution of Eq.(5) in a closed form [17], [18]:

$$Q_i(x)=\exp\left\{-\frac{2Cx}{TTI_i N_i \rho_i^2}\left(\frac{Cx}{TTI_i}+C-N_i\rho_i\right)\right\}, \tag{6}$$

and having different classes but all with $TTI_i=TTI_i, i=1,\ldots,K$, the border of the resulting (delay-limited) admissible region is approximated by a hyper-plane in N_i:

$$Q_i(x)\leq\varepsilon \Leftrightarrow \sum_i N_i\left(\rho_i+\frac{\rho_i^2}{C}\frac{\gamma\, TTI}{2x}\right)\leq C+\frac{Cx}{TTI}, \tag{7}$$

where $\gamma=-\ln(\varepsilon)$, and all $\tilde{\varepsilon}_i^{delayed}=\varepsilon$. The accuracy of the Brownian bridge approximation is good [16]. Note that since Eq.(6) is exact, the hyper-plane approximation is justified by the inherent relation between the real arrival process and the Brownian bridge approximation.

We have simulated admissible regions extensively with UTRAN-specific traffic parameters (for classes inhomogeneous in TTI_i and $\alpha_i\leq 1$), and found that delay-limited borders of the admissible regions are approximately linear (this observation is also supported by the observations in [15] and by our analysis in [22]).

Since we found the linear approximation to be good, we need to calculate only the edges of the hyper-planes. Define TN_{ij} as the maximum number of class i connections assuming that a single packet from class j would fulfill the QoS requirement of class j ($\mathrm{PR}\{D_j>\tilde{D}_j\}\leq\tilde{\varepsilon}_i^{delayed}$). We approximate by TN_{ij} the maximum number of class i connections if one additional class j connection is present in the system. (More details on this approach and alternative formulas for computing TN_{ij} can be found in [22].) This way, considering only the delay criterion, the necessary condition of accepting the traffic mix $(N_1,N_2,\ldots,N_K)$ is:

$$\sum_{i=1}^{K}\frac{TN_{jj}}{TN_{ij}}\cdot N_i\leq TN_{jj}+1 \qquad j=1,2,\ldots,K;\ N_j>0. \tag{8}$$

When determining TN_{ij} values, we consider three cases depending on the priority levels of class i and class j.

If *class i and class j have the same priority* level (same as if FIFO scheduling is applied), then we calculate TN_{ij} using Eq.(6): $(\mathrm{PR}\{D_j>\tilde{D}_j\}=Q_i(\tilde{D}_j-b_j/C)$.

If *class i has higher priority* we follow a different method. The event that the last segment (of size s_{last}) of the class j packet could not be served before time $\tilde{D}_j$ is equivalent to the event that all segments before the last one could not be served before $D'=\tilde{D}_j-s_{last}/C$. Formally, denote $B(0,t)$ the server availability in $[0,t]$ seen by the class-j packet when it has arrived at time 0, and then:

$$\Pr\{D_j>\tilde{D}_j\}=\Pr\left\{B\left(0,\tilde{D}_j-\frac{s_{last}}{C}\right)<\frac{b_j-s_{last}}{C}\right\}. \tag{9}$$

We use a conservative approximation of the server availability process: $B(0,t)\approx t-A_i(0,t)/C$ and we use the Brownian bridge approximation of the arrival process, which results in

$$\Pr\{D_j>\tilde{D}_j\}\approx\Phi\{b_j-s_{last};\ (C-N_i\rho_i)D',\ N_i\rho_i D'(TTI_i-D')\}, \tag{10}$$

where $\Phi\{x;\mu,\sigma^2\}$ denotes the normal distribution. In Table 1, TN_{ij} values calculated using Eq.(10) are compared with exact values (C=920 kbps, b_j=320 bit, s_{last}=320 bit, TTI_i=20 ms, $\tilde{D}_j$=10 ms, $\tilde{\varepsilon}_i^{delayed}$= 0.1%).

b_j [bit]	320	640	960	1920	2880	3840
TN_{ij}^{exact}	41	38	36	30	25	20
TN_{ij}^{approx}	37	36	34	29	24	20

Table 1. Example demonstrating the accuracy of Eq.(10)

If *class i has lower priority* we neglect the effect of a segment possibly under service from class i on delays of class j packets. (A precise solution, which counts with this effect is presented in [20].) It means that TN_{ij} values are not calculated for this case (i.e., $TN_{ij}=\infty$ in Eq.(8)).

4.1. Numerical examples for the single link case

Consider the example defined by Table 2 and the following parameters: C=1504 kbps, s_{last}=384 bit, $\tilde{\varepsilon}_i^{lost}$= 0.05 %, $\tilde{\varepsilon}_i^{delayed}$= 0.05 % for all i.

Considering signaling traffic (PCH, FACH and DCCH), there are always 3 PCH, 3 FACH1 and 3 FACH2 connections in the system, while the number of DCCH connections equals the sum of the number of voice, 64k RAB and 384k RAB connections. (On roles of signaling channels see [3].)

The following FIFO and two-priority systems are examined: ***FIFO*** (strictest delay requirement: 5 ms),

PRIO-7.5 (delay requirements are: high: 5 ms, low: 7.5 ms), ***PRIO-10*** (high: 5 ms, low: 10 ms), ***PRIO-15*** (high: 5 ms, low: 15 ms). Admissible regions, obtained using the algorithms in the previous section, are presented in Table 3. Simulations showed that for all mixes the LAC algorithms are conservative, i.e., the QoS requirements are met.

RAB type	*TTI* [ms]	*b* [bit]	α	priority level
voice	20	336	0.55	high
64k RAB	20	1480	1	low
384k RAB	10	4360	1	low
DCCH	40	216	0.2	high
PCH	10	480	0.5	low
FACH1	10	432	0.5	low
FACH2	10	456	0.5	low

Table 2. Parameters of the numerical example

It is clear from the results that prioritization (QoS separation) increases the number of admissible low-priority connections as the low priority delay requirement gets looser. The number of admissible high-priority connections may decrease, if the low priority delay requirement is close to the high priority delay requirement. Regarding this numerical example, we can conclude that, already with the 10 ms low priority delay requirement, priority scheduling is more advantageous than FIFO.

	0 of 384k RAB	**1** of 384k RAB	**2** of 384k RAB
0 of 64k RAB	90, 73, 83, 90	43, 39, 53, 53	--, 8, 16, 16
1 of 64k RAB	84, 65, 76, 84	33, 33, 46, 46	--, 1, 10, 10
2 of 64k RAB	77, 60, 71, 77	24, 28, 40, 40	--, --, 4, 4
3 of 64k RAB	71, 56, 67, 71	14, 24, 34, 34	
4 of 64k RAB	64, 51, 63, 64	4, 20, 27, 27	
5 of 64k RAB	58, 47, 58, 58	--, 15, 21, 21	
6 of 64k RAB	52, 43, 52, 52	--, 11, 15, 15	
7 of 64k RAB	45, 38, 45, 45	--, 6, 9, 9	
8 of 64k RAB	39, 34, 39, 39	--, 2, 3, 3	
9 of 64k RAB	30, 29, 33, 33		
10 of 64k RAB	21, 25, 26, 26		
11 of 64k RAB	11, 20, 20, 20		
12 of 64k RAB	1, 14, 14, 14		
13 of 64k RAB	--, 8, 8, 8		
14 of 64k RAB	--, 2, 2, 2		

Table 3. Number of admitted voice connections in cases ***FIFO, PRIO-7.5, PRIO-10*** and ***PRIO-15***

5. The multiple link case

In this section, statistical multiplexing gains achieved with different switching alternatives, described in Section 2.2, are evaluated. Numerical examples are given to show that, when using AAL2 switching (i.e., we have more hops) the end-to-end QoS requirement can be fulfilled with sufficient accuracy.

5.1. Efficiency of switching alternatives

If a new AAL2 connection arrives, $\varepsilon_i = \mathrm{PR}\{D_i > \tilde{D}_i\}$ is evaluated by LAC algorithms in each node, where AAL2 multiplexing or switching happens along the route selected for the AAL2 connection. Although, $\tilde{D}_i$ is the end-to-end delay requirement, we set this value in each LAC. (In Section 5.3, we will show that this is a good approach, and it is not needed to divide $\tilde{D}_i$ among the links carrying a connection of class i.) Capacity C, used by the LAC, represents the capacity of a CBR VC, or a CBR VP if UBR VCs are multiplexed in CBR VPs. The AAL2 connection is admitted if $\varepsilon_i < \tilde{\varepsilon}_i$ for all classes at each LAC.

GoS requirements of class i are met if the blocking probability of class i connections is below a predefined threshold, i.e., $B_i < \tilde{B}_i$. Probabilities of the occurrences of traffic mixes (and therefore also blocking probabilities) depend on the connection arrival process. B_i values also depend on the CAC algorithm, i.e., on which traffic mixes are rejected by LAC decisions. If more VCs (or VPs) are available for connections between two nodes, also the strategy of connection distribution over these resources affects blocking probabilities.

To calculate blocking probabilities, simple Markov-chain based methods (such as extensions of the Erlang-B formula) are difficult to use because: (1) an admissible region provided by the LAC is typically not a single hyper-plane [22], (2) connection distribution strategies are not easy to include in Markov-chains, (3) the connection arrival process is in general not Poisson, (4) the blocking probabilities on links of different hierarchy levels in the network are not independent.

Since the LAC algorithm is designed to work on-line in UTRAN (i.e., it is fast), it is also applicable in a connection-level simulator. Therefore, we have developed such a simulator, in which packet level traffic descriptors are attributes of generated connections. This way, the LAC algorithm can be executed in traffic concentration nodes and the ratio of connections blocked by admission controllers (LACs) can be measured.

5.2. Numerical example

In this section, the scenario in Figure 7 is evaluated. We consider downlink traffic, meaning that connections are generated in the RNC and terminated in Node Bs. During the discussion we will sometimes refer to the hierarchy levels as Level 1, 2 and 3 (L1: between Node Bs and SN2, L2: between SN2 and SN1, L3: between SN1 and RNC). The traffic is uniformly distributed over the Node Bs.

We assume that on L1, 2 E1s are used. To achieve higher utilization there, we apply priority scheduling and the connection distribution mechanism depicted in Figure 8. On L2 and L3, large physical links are used (e.g., 155 Mbps) and FIFO scheduling is applied. To simplify the discussion, we assume that the granularity of capacity allocation (VP and VC capacities) corresponds to the capacity of an E1. This way, on L2 and L3, X and Y "E1s" are used, respectively. The gross capacity of an E1 is *1920* kbps. In this work, we generate only user-plane traffic and assume that for

each Node B *212* kbps is allocated for operation and maintenance traffic, and *234* kbps for common control channels (FACH and PCH). This way, if CBR VCs are used on L1, the capacities of these VCs without ATM overhead can be calculated as 47/53◊(2◊1920-(212+234))/2 = *1504.8* kbps. For the sake of simplicity, we will assume that this capacity (*1504.8* kbps) is available per E1 on each Level, and to obtain the gross capacities, we add the signaling and O&M overheads at the end of the calculations. For example if *Y*=10 and *J*=3 (see Figure 7), the gross capacity on L3 is 53/47·Y·1504.8+3·J·(212+234) = *20983* kbps.

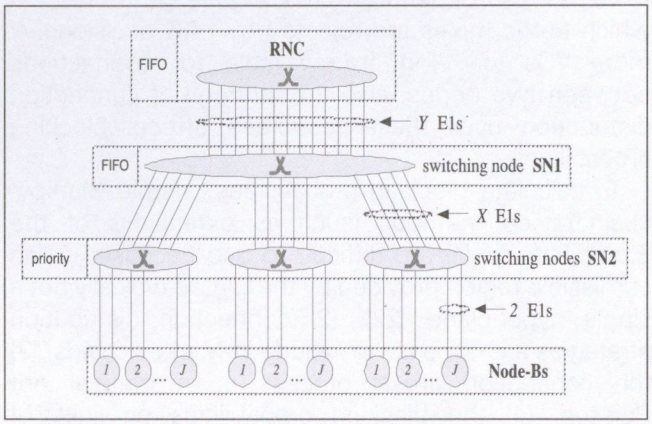


Figure 7. Example scenario

First, *X* and *Y* is set such that there is no blocking on L2 and L3, and the number of RAB users per Node B is increased until $B_{i,\ L1}$ is reached. Next, *X* is decreased until $B_{i,L2}$ is reached. Finally, *Y* is decreased until $B_{i,L3}$ is reached. Results with homogeneous traffic and switching alternative ***B*** are presented in Table 5 (*J=3*).

Traffic parameters, QoS and GoS requirements are presented in Table 4. In case of data traffic, 64k and 384k RABs are used. For data traffic, it is difficult to make assumptions on the RAB usage, because it depends on system as well as user and application-behavior. Therefore connection level traffic parameters are hypothetical ones. In this example, connections are generated according to a Poisson process. If a connection is blocked, no retries are assumed (e.g., if a 384k bearer is blocked, there is no retry on a 64k bearer.) The above simplifications are not necessary when using the simulator.

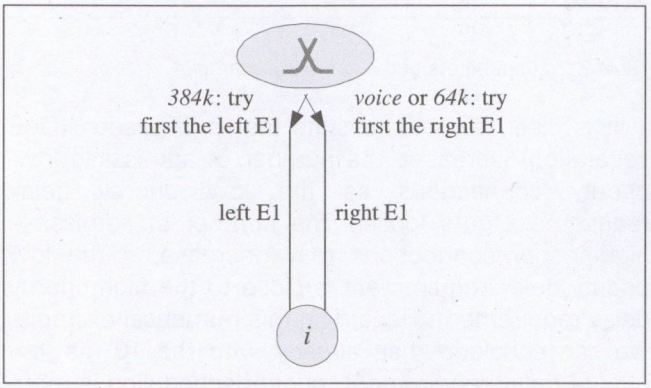


Figure 8. Connection distribution to decrease resource fragmentation on L1

The results in the table show that for homogeneous voice traffic little gains can be achieved (e.g. 15 instead of 18 VCs on L3 and no gain on L2), but for larger data bearers gains are significant (e.g. 9 instead of 18 VCs for 384k on L3).[%]

	Packet Level			Connection Level	QoS req.		GoS requirements		
	TTI_i [ms]	b_i [bit]	α_i	Offered tr. / RAB user [mErlang]	D_i [ms]	ε_i [%]	$B_{i,L1}$ [%]	$B_{i,L2}$ [%]	$B_{i,L3}$ [%]
voice	20	336	0.55	15	5	0.1	0.1	0.2	0.3
64k	20	1480	1	2.775	15	0.1	1	2	3
384k	10	4360	1	0.925	15	0.1	1	2	3

Table 4. Parameters of the numerical example

RAB type	# RAB users / Node B	X	Y
voice	14634	6	15
64k	10527	5	14
384k	2105	4	9

Table 5. Results in case of homogeneous traffic

Consider a voice oriented traffic mix, where the RAB users are: 80% voice, 18% 64k and 2% 384k. First, assume that if alternative **B** or **C.2** is used in SN1 or RNC, then always **B** is used in SN2. This way, the number of RAB users per Node B is 8700 (104.4

	Switching alternative	*X*	*Y*	Gross bitrate on L2	Gross bitrate on L3	Gain over *A* on L2	Gain over *A* on L3
3	**B**	5	12	29447 bit/s	24359 bit/s	15 %	29 %
	C.2	4	11	24359 bit/s	22663 bit/s	29 %	34 %
6	**B**	9	23	53805 bit/s	47021 bit/s	22 %	32 %
	C.2	8	19	48718 bit/s	40238 bit/s	29 %	42 %
9	**B**	13	32	78165 bit/s	66293 bit/s	25 %	36 %
	C.2	11	27	67988 bit/s	57813 bit/s	34 %	44 %

Table 6. Results in case of inhomogeneous traffic, statistical multiplexing gains of switching alternatives ***B*** and ***C.2*** over alternative ***A***

Erlang voice, 4.34 Erlang 64k and 0.16 Erlang 384k traffic). Results are shown in Table 6. The gains are computed as follows: *(gross bitrate with **A** – gross bitrate with **B** or **C.2**)/(gross bitrate with **A**)*.

Next, assume that if alternative ***C.2*** is used on SN1 or RNC, then also ***C.2*** is used on SN2. This way, the number of RAB users per Node B increases to *9600*. If *J=3*, then *X=5* and *Y=11*, and the statistical multiplexing gains on L2 and L3 are *15%* and *34%*, respectively.

Results presented in Table 6 indicate that using AAL2 switching, significant statistical multiplexing gains can be achieved, even with a voice oriented traffic mix. Having UBR VCs instead of CBR (***C.2*** instead of ***B***) is the most efficient solution. However, using ***C.2*** one must be careful, because it must be ensured that no other traffic than the considered UBR VCs share the VP resource. For example, if leased VCs are used, then only ***B*** can be realized. Also note that even if there is no further concentration on L3 (VC endpoints are the RNC and SN2 nodes, and Y=3·X), statistical multiplexing gains are still significant (see column 'Gain over A on L2' in Table 6).

5.3. End-to-end packet delay

In UTRAN it is not possible to execute a global CAC, which considers the state of the whole network. Instead a LAC algorithm is evaluated in each node with AAL2 multiplexing. Thus the end-to-end delay requirement must be transformed to link-by-link delay requirements. This transformation cannot be made by the switching nodes, because they do not have information about the network topology. On the other hand, because of the probabilistic delay requirement and the correlation of delays on different links, the end-to-end delay requirement cannot be simply divided among the hops. We assume that the end-to-end delay is dominated by the delay on a bottleneck link. Thus we configure the end-to-end delay requirement to each LAC. In this section we will show that it is a practically good approach.

The number of hierarchy levels, which is determined by the number of AAL2 switching points, is typically 2-3 in UTRAN. We used the simple two level scenario depicted on Figure 9 to verify end-to-end delays (L1: between Node Bs and SN, L2: between SN and RNC). FIFO scheduling is used on both levels. The capacity of all VCs is 1504 kbps. The traffic descriptors are the same as in Table 4. Switching alternative ***B*** is used in the SN.

Whenever a voice connection is rejected by the LAC on a selected L2 VC, the connection level simulator logs the state of the system (the number of connections and their route). This generated log is used as input for a packet level simulator. The end-to-end delay of packets on the connections going through the selected L2 VC is measured in this simulator. An example for these traffic flows is marked with dashed lines on Figure 9. The LAC on the selected L2 VC rejects a voice connection if this VC is full (i.e., the actual traffic mix is at the border of its admissible region). But in this case, the L1 VCs are not necessarily full. To evaluate extreme cases (from end-to-end delay point of view), in some examples we increased the number of connections on L1 VCs until the border of their admissible region is reached. These additional connections could not be transported on L2, because the number of VC there is less than on L1. However it does not cause problem in the simulator, because on L2 we simulate only the traffic of the selected L2 VC (in the simulator, the packets which would be going through other L2 VCs are terminated in the SN).

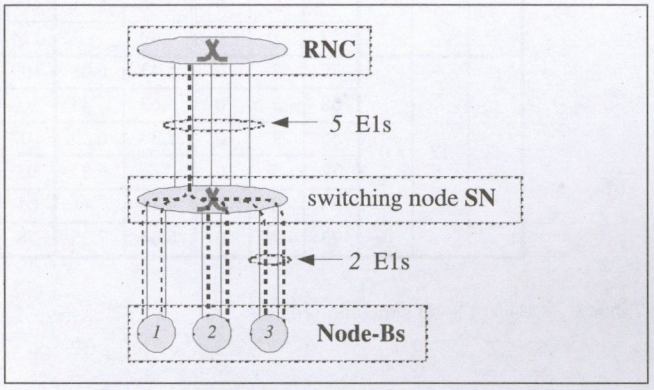


Figure 9. Simulation scenario

Table 7 summarizes three simulation examples. In case of each example we evaluated a 'Normal case' when the log generated by the connection level simulator is directly used as input for the packet level simulator. Then we evaluated a case when the number of voice connections on L1 VCs was increased until the border of the admissible region was reached. Finally we evaluated a case when first the system was stuffed with 384k then the rest with 64k and finally with voice connections. The 0.1% quantile of the end-to end delay is indicated in the table for each simulation.

Remember that the admissible region determined by the LAC is the intersection of two types of regions. One type is determined by the delay violation probabilities due to packet loss and the other is determined by the delay violation probabilities due to delayed packets (see Section 4). The three simulation examples of Table 7 were chosen such that the arriving voice connection was rejected on the selected L2 VC due to different reasons. In the first and second case the packet loss probability was too high. In the second case the voice delay violation probability due to delayed packets was also very close to its limit ($\tilde{\varepsilon}_i^{delayed}$). In the third case the voice delay violation probability due to delayed packets was too high.

The results indicate that even in these rare situations and extreme cases, the maximum end-to-end voice delay was below 7 ms, which is acceptable

Number of connections on the selected L2 VC			Normal case					Increased # of voice					Increased # of 384k-64k-voice				
			# of connections on L1 VCs			end-to-end delay [ms]		# of connections on L1 VCs			end-to-end delay [ms]		# of connections on L1 VCs			end-to-end delay [ms]	
voice	64k	384k	voice	64k	384k	voice	64k	voice	64k	384k	voice	64k	voice	64k	384k	voice	64k
128	0	0	50	1	0	3.21	-	121	1	0	4.21	-	50	1	1	5.22	-
			79	3	0	4	-	107	3	0	4.75	-	86	6	0	5.54	-
			67	1	0	3.35	-	121	1	0	4.33	-	67	8	0	5.72	-
			67	1	0	3.35	-	121	1	0	4.34	-	67	8	0	5.74	-
			71	1	0	3.36	-	121	1	0	4.29	-	78	7	0	5.87	-
			50	3	0	3.57	-	107	3	0	4.81	-	56	9	0	5.82	-
86	6	0	46	0	1	5.92	-	61	0	1	6.03	-	50	1	1	6.06	-
			59	8	0	6.36	6.9	67	8	0	6.55	7.05	67	8	0	6.51	7.07
			54	0	1	5.92	-	61	0	1	5.95	-	61	0	1	5.88	-
			63	7	0	6.31	6.74	78	7	0	6.92	7.2	67	8	0	6.64	7.09
			67	5	0	5.38	5.91	93	5	0	6.18	6.48	67	8	0	6.31	6.81
			57	2	0	4.49	-	114	2	0	5.72	-	67	8	0	6.21	-
23	12	0	77	3	0	5.43	6.08	107	3	0	6.02	6.56	78	7	0	6.57	7.16
			68	5	0	5.63	6.45	93	5	0	6.29	6.89	78	7	0	6.66	7.24
			55	3	0	5.36	6.09	107	3	0	6.25	6.72	56	9	0	6.71	7.36
			77	5	0	5.7	6.35	93	5	0	6.2	6.73	78	7	0	6.5	7.03
			57	0	1	6.11	-	61	0	1	6.15	-	61	0	1	6.13	-
			73	7	0	6.64	7.25	78	7	0	6.85	7.33	78	7	0	6.85	7.37

Table 7. Packet level simulation results

in practice. The delay of 64k were much below its requirement, because we used FIFO scheduling. Note that if priority scheduling were used in the system and the rate of L2 VCs were higher, voice packet delays would be lower. In practice, one could reduce end-to-end delays also by slightly rendering the LAC delay requirements more strict.

6. Conclusions

The UTRAN transport infrastructure has to be efficient, because transmission resources are expensive. On the other hand, strict QoS requirements have to be met. The key to manage this trade-off is a good and reliable connection admission control method.

In our work we introduced a queuing model of the UTRAN system, and used this model to develop connection admission control algorithms for both FIFO and priority scheduling. Assuming that the proposed admission control method is implemented in the system, we evaluated the efficiency of various switching (traffic concentration) alternatives, and found the application of AAL2 switches beneficial. Once AAL2 switching is used, a connection admission controller is implemented in each AAL2 switch, and a connection is admitted in UTRAN only if it is admitted in each AAL2 switching node. We performed packet level simulations, and argued that in such a system, the end-to-end delay can be controlled with sufficient accuracy.

References

1. 3GPP, Delay Budget within the Access Stratum, TR 25.853, 2001.
2. 3GPP, IP Transport in UTRAN, TR 25.933, 2001.
3. 3GPP, Multiplexing and channel coding (FDD), TS 25.212, 2001.
4. ITU-T, AAL Type 2 Signalling Protocol (Capability Set 1 and 2), New ITU-T Rec. Q.2630.1 and Q.2630.2
5. The ATM Forum Technical Committee, Traffic Management Specification V. 4.0, 1996.
6. T. Ojanpera and R. Prasad, editors, Wideband CDMA for Third Generation Mobile Communications, Artech House, 1998.
7. G. Eneroth, et al, Applying ATM/AAL2 as a Switching Technology in 3G Mobile Networks, IEEE Comm. Mag., 37(6):112-122, 1999.
8. G. Fodor, et al, Comparison of Call Admission Control Algorithms in ATM/AAL2 Based 3rd Generation Mobile Access Networks, Proc. IEEE WCNC, 1999.
9. K.M.F. Elsayed, N. Gerlich and P. Tran-Gia, Efficient Design of Voice Carrying Fixed-Network Links in CDMA Mobile Communication Systems, Telecom. Systems, 17(1,2): 9-29, 2001.
10. H. Saito, Bandwidth Management for AAL2 Traffic, IEEE Tr. on Vehicular Tech., 49(4): 1364-1377, 2000.
11. H. Saito, Performance Evaluation and Dimensioning for AAL2 CLAD, Proc. IEEE INFOCOM, pp. 153-160, 1999.

12. H. Saito, Effectiveness of UBR VC Approach in AAL2 Networks and Its Application to IMT-2000, IEICE Transactions on Communications, E83-B, 11, 2000.
13. J. Peisa and M. Meyer, Analytical Model for TCP File Transfers over UMTS, Proc. 3G Wireless 2001, 2001.
14. I. Norros, et al, The Superposition of Variable Bit Rate Sources in an ATM Multiplexer, IEEE JSAC, 9(3): 378-387, 1991.
15. L. He and A. Wong, Connection Admission Control Design for GlobeView-2000 ATM Core Switches, Bell Labs Technical Journal, pp. 94-110, Jan.-March, 1998.
16. Methods for the performance evaluation and design of broadband multiservice networks, Part III, Traffic models and queuing analysis, The COST 242 Final Report, 1996.
17. B. Hayek, A Queue with Periodic Arrivals and Constant Service Rate, in Kelly, F.P. (ed.) Probability, Statistics and Optimisation, a Tribute to Peter Whittle., Wiley, 147-157, 1994.
18. F. P. Kelly, Notes on Effective Bandwidths, Stochastic Networks: Theory and App., Vol. 4., Oxford University Press, 141-168, 1996.
19. A. M. Makowski, Bounding On-Off sources – Variability ordering and majorization to the rescue, ISR TR 2001-13, http://www.isr.umd.edu/TechReports/ISR/2001/TR_2001-13/TR_2001-13.pdf
20. K. Iida, et al, Delay Analysis for CBR Traffic under Static-Priority Scheduling, IEEE/ACM Transactions on Networking, 9(2), April, 2001.
21. A. Rácz, N. Fias, P. Rácz, Effective Bandwidth and Associated CAC Procedure for Traffic Streams Multiplexed over a VBR Link, Proc. SPECTS, pp. 72-79., Vancouver, 2000.
22. Sz. Malomsoky, S. Rácz and Sz. Nádas, Connection Admission Control in UMTS Radio Access Networks, submitted to Computer Communications, Elsevier Science, 2002.
23. Sz. Nádas, S. Rácz and Sz. Malomsoky, On Quality of Service Differentiation in UMTS Radio Access Networks, submitted to IEEE GLOBECOM, Taipei, Taiwan, 2002

News

Internet Telephone Numbering System (ENUM) offers promise of a single point of contact for all communication devices. The International Telecommunication Union (ITU) annouce approval of interim procedures for ENUM, the technology that builds a bridge between the public switched telephone network and the Internet.

It makes possible for consumers to use a single number to access many types of terminals and services, such as: phone, fax, e-mail, pager, mobile tlephones, websites of any other serices available through an internet addressing scheme.

EMUM will make possible for the first time the ability to call a PC from the public-switched telecommunication network (PSTN) and to determine what type of terminal is associated with the number.

A call to a telephone number can invoke Internet type services. For example, calling an ENUM-enabled telephone number from a 3 rd generation multimedia handset could allow access to a location-based mobile web service, thus avoiding entering Internet-type addresses on numeric keypads.

In keeping with the need to allow for voluntary implementation of the scheme andrecognizing that ENUM services are prymarily national issues, rapid progress at the international leves is necessary to create a stable environment in which investment can be made in the wordwide deployment of ENUM.

Analysis of WaveLAN Systems' Performance

Ákos Juhász, Ferenc Ulrich
Dr. Bertalan Eged, Ferenc Kubinszky

Budapest University of Technology and Economics
Department of Broadband Infocommunication Systems
Wireless Information Technology Laboratory

Introduction

As a consequence of the evolution of the Internet technology more and more computers got connected to networks. Almost every computer is connected into wired LAN/WAN systems. The main advantages of wired LANs are the high bandwidth and the low costs of the devices. It's safe and reliable. The main disadvantages are the fixed location and the circumstantial installation.

A network should give more mobility and freedom. Few years ago the idea of wireless LANs came up as a solution for building commercial/business LANs. The main features of WLAN (Wireless LAN) are the flexibility, mobility and the easy installation.

A stable and reliable WLAN system needs careful planning. The wireless radio channel can cause many problems according to the high frequency radio interface. Noisy channel and crowded built-in scenarios can cause communication problems.

This article shows some investigation of a WLAN systems real-world performance.

A WLAN system

According to the IEEE 802.11 standard WLAN devices communicate in the Industrial, Scientific and Medical (ISM) band (2.4 – 2.483GHz). The ISM band is freely usable in either Europe and in the US. The maximal transmittable power is region-dependent. ETSI (EU) defined 100 mW, while FCC (US) 1W as the maximum.

To avoid narrow-band jamming the standard defines spread spectrum modulation, Frequency-Hopping Spread Spectrum (FHSS) and Direct Sequence Spread Spectrum (DSSS). Nowadays almost all devices utilize DSSS, that scrambles the data using a 11 chip-long Barker code. The scrambled data is then modulated to decrease the signal's bandwidth. Different kinds of modulations can results in higher or lower throughput. The data modulation schemes are DBPSK, DQPSK, resulting 1, 2. 5.5 and 11Mbps throughputs are achieved via CCK coding mechanism.

The signal's bandwidth in the radio channel is 22MHz. According to the standard 13 channels can be established in the 83MHz-wide ISM band. Spectral overlapping between adjacent channels is significant, so maintaining two or more communications on the same area without any interference needs minimum3 channels' distance.

Because of the common channel, multiple access machanism is required. Wired networks employs CSMA/CD (Carrier Sense, Multiple Access with Collision Detection). This protocol is really hard to implement on wireless LANs, according to the following problems:

1. Collision Detection process needs a full duplex radio
2. Collision Detection protocol assumes that all nodes can receive each other`s signal, cannot be established all the time in a wireless environment.

To solve these problems IEEE 802.11 defines collision avoidance (CA) mechanism with positive acknowledgement:

1. The transmitter node checks the radio channel. If the channel is detected to be idle for a pre-defined time (DIFS, Distributed Inter Frame Space) the node starts its transmission, else it postpones the transmission.
2. The receiver station checks the frame's CRC (Cyclic Redundancy Code) and sends an Acknowledgement packet if it is correct. If the acknowledgement isn't received in a certain time by the sender, then the packet is resent.

A WLAN system consisting of Lucent WaveLANTM cards were the point of our measurements. These cards correspond to IEEE 802.11b standard. The cards support both adhoc and managed modes. An adhoc featured network offers great flexibility and can be built up dynamically. Every single node is a part of the whole wireless network, and can communicate with another one without a common controller. In managed mode an "access point" appears in the wireless infrastructure communicating with the wired backoff. Multiple Access points can create a cellular system. Roaming between cells is defined in the standard, and

the mobile station initiates it. In managed mode every single packet traverses the access point, which controls every single cell's communication.

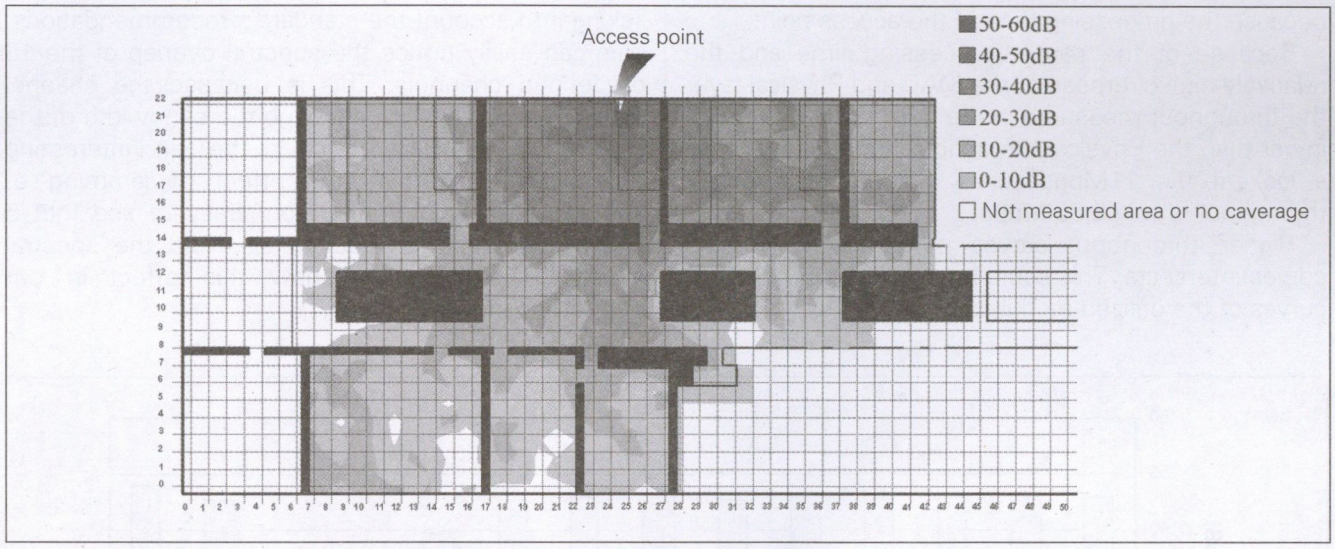
Test conditions

Investigations were made to determine the maximum usable indoor distance and to characterize the wireless links' performance. Different signal to noise ratios were applied to assure both the presence and the absence of wireless errors. Different test-schemes were used to measure the performance of the wireless network. Measurements were made both in adhoc and managed modes. To achieve the highest performance possible minimal interference was assured for the tests.

The tests were run under Linux operating system. Information on overall link quality and packet discardings are available under several system files. These system files are refreshed whenever a packet arrives or leaves. Thus measurements can be automated via reading these files periodically. Specific C programs performed the collecting and analization of the measured values. Netperf 2.2 was used to generate TCP traffic over the wireless channel for our data-link measurements.

Coverage measurement

To determine the maximum available indoor distance and to characterize the attendance of different obstacles a two node managed mode arrangement was used. One of these nodes was connected to the access point routing the wired Ethernet network and the other node was a mobile computer. The complete coverage layout of the whole department was made. The access point was placed inside of our laboratory as shown on the figure below. The signal to noise ratio was measured in 1m step size so we got the full coverage map.



Coverage map

According to the figure the average attendance is 4.3dB per meter, and the maximum usable indoor distance is 25m. Of course these values are highly depend on the built in density, the building materials. The figure truly shows the higher/lower attendance of different materials. For example a glass wall represents much lower attendance than a reinforced concrete wall. Doors are also almost transparent to radio waves. Water is the highest attenuator factor in the 2,4GHz ISM band. Since human body is almost water One could be a great attenuator. So that's why we had chosen our measurement time to non-crowded periods. This way we tried to maximize the usable indoor distance. Doing the measurement in crowded scenarios can result in radically lower measurement values.

A coverage map could be useful if you consider planning a cellular WLAN system. Defining a minimal Signal to Noise Ratio optimal places can be found for other access points to achieve full coverage.

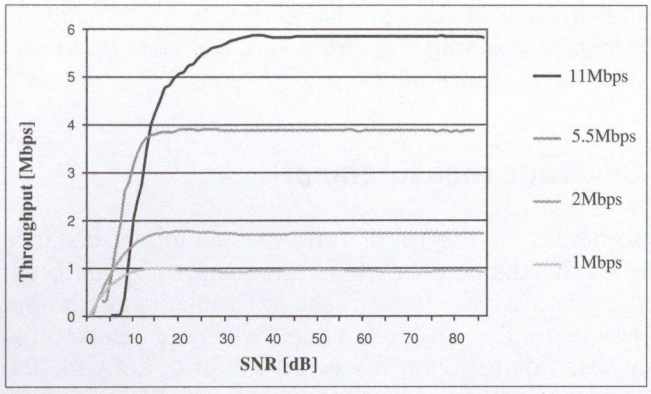
Throughput measurements

The overall Signal to Noise Ratio (SNR) may fall at higher distances or on crowded/noisy arrangements. Below a specific SNR threshold value Bit Error Rate (BER) rises, resulting in higher frame loss rate and lower throughput. At higher data rates certain parts of the whole radio frame are transmitted at lower speeds for more security.

The effect of Signal to Noise Ratio fluctuation to the throughput on different data rates was the point of these investigations.

We measured throughput on different data rates and both in managed and adhoc modes. To achieve

lower signal to noise ratio we increased the distance between the two communicating nodes.

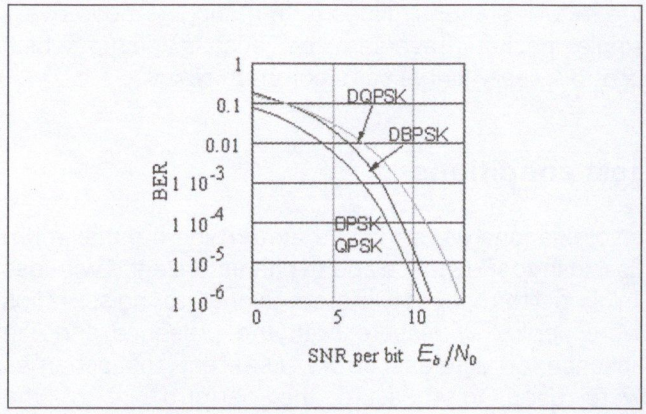


Throughput vs SNR

The figure shows the throughput on different data rates in the function of the SNR measured in adhoc mode. At higher data rates more signal to noise ratio is required to reach a constant throughput. In managed mode 10% lower speed was measured. This is because the processing time of the access point.

Because of the packet processing time and the relatively high overhead of the MAC and Physical layer the throughput measured in the third layer was much lower than the Physical throughput. For example take a look at the 11Mbps curve, where the effective throughput is only 5,85 Mbps.

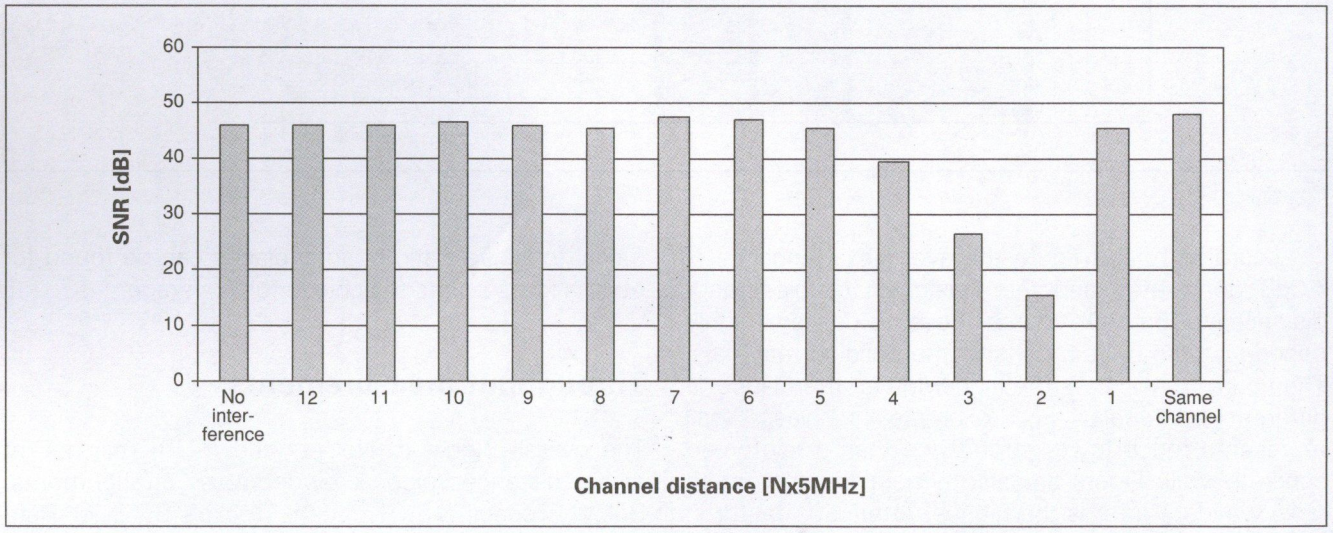
Higher throughput curves' rises faster, so the curves intersects. This could be predicted by the BER curves of the utilized modulations:



The minimum required signal to noise ratio (mentioned in the previous section) can be defined using this diagram depending on the required throughput.

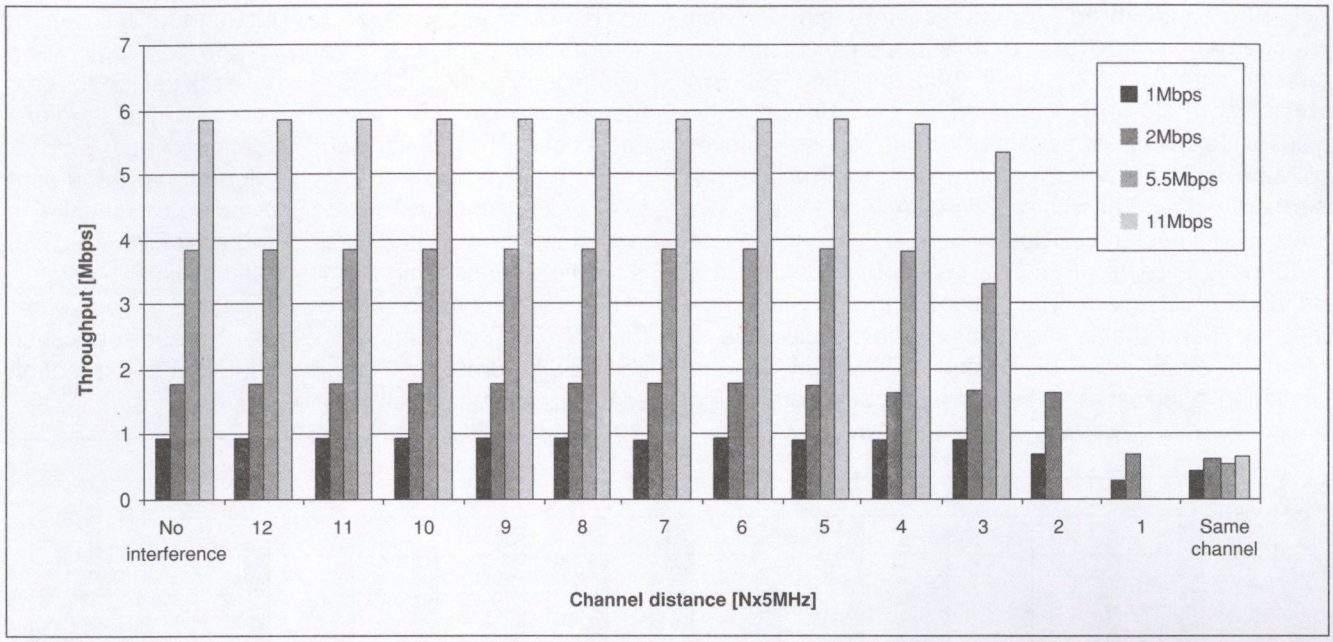
Interference measurements

Taking into account the standard's recommendations, One can easily notice the spectral overlap of the 13 predefined channels. This is because the channel distance is only 5MHz, although the bandwidth of the channel is 22MHz. It could be an interesting possibility to examine the effect of jamming of another WLAN communication. Easy to see that 5 channel distance is required to avoid the spectral overlap. We'd like to show this effect in our interference investigations.



SNR vs Interference

The vary of the signal to noise ratio versus the channel distance of the two communication is shown on figure no.3. The effect of jamming appears firstly in a distance of 4 channels'. Very interesting point we have come to. The value of the SNR is high again on the same or neighbour's channel. Due to the communicating frequency band and the spreading code are the same for both communication, the device recognises the other communication as a signal not noise and therefore the CSMA/CA algorithm starts to operate.



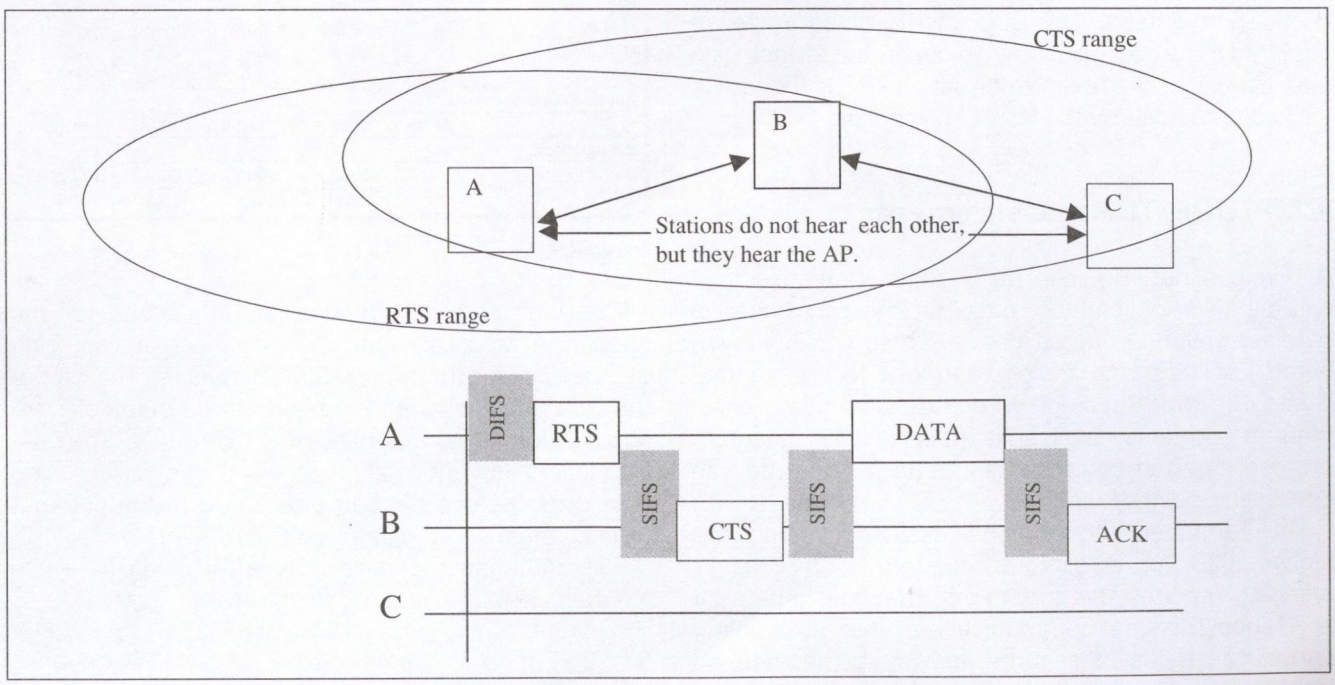
Throughput vs. interference

The figure shows the throughput versus the channel distance. The effect of the jamming communication turns up first only within 3 channel's distance. The insensibility to narrow band jamming of spread spectrum technique unveils in 4 channels` distance, when the two communicating channels overlapping is only a few MHz. As a result the SNR decreasing, and the throughput remain stable.

Now we can understand why the firm recommends 3 channel's distance for multiple communications in the ISM band. You may notice the fall of the throughput on the same channel according to the use of CSMA/CA protocol. At higher bitrates (11, 5.5Mbps) low speed cards (2Mbps) were used for jamming, that's why we measured significantly lower throughput than half speed. Effectively more time is spent on the communication at a lower speed.

RTS/CTS mechanism

CSMA/CA protocol can only help preventing packet collision if the two sender node can receive each other's signals. There can be different network arrangement where the two senders cannot receive each other's signals, so a sender can sense the channel to be idle, but a collision can occur at the receiver. One of these problematic arrangement is the hidden terminal.

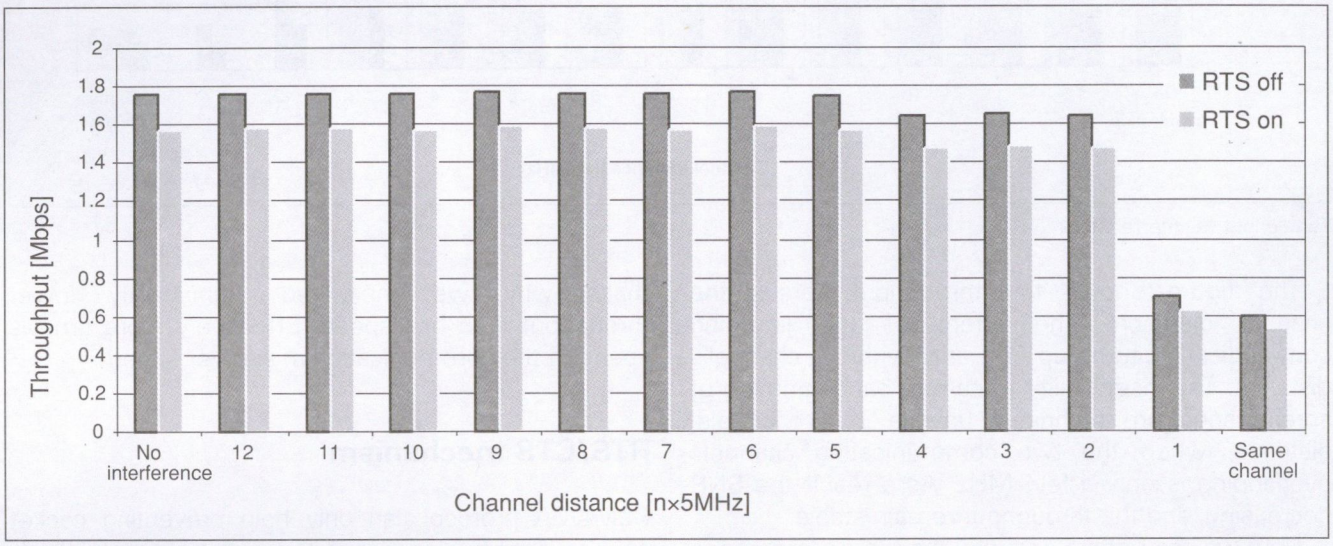


Hidden terminal problem and RTS-CTS mechanism

If there is three communicating node in an arrangement where two of this nodes (A,C) cannot receive each other's signals but the third (B) can receive both of them (A,C), then we can talk about the hidden terminal arrangement that cause packet collision at B. If A and C want to send their packets at the same time, then the CSMA/CA won't work and there'll be a packet collision at B.

To avoid this, the IEEE 802.11 standard defines the RTS/CTS mechanism. This means a packet sequence like the hand-shake mechanism which provides a communication free of packet collisions. At the beginning A indicates its sending intention via sending an RTS (Request To Send) packet to node B. C cannot receive this packet according to the topology. As a response B sends a CTS (Clear To Send) packet to A, so that every node in B's range is warned not to send any packet at all. Then A can send its data packet to B. After the data packet an ACK (Acknowledge) packet is sent to A which can be received by C and so the channel is signed to be free. DIFS (Distributed Inter-Frame Space) is the time for sensing the channel if it's idle.

If all of the 3 nodes can communicate to each other the RTS/CTS mechanism is not needed because CSMA/CA protocol assures sufficient protection against packet collisions.



Throughput vs. Interference at 2Mbps

The above figure shows the data speed lowering effect of the RTS/CTS protocol if every node can communicate to each other. Not all the operating mode suffer the same throughput decreasing due to the RTS/CTS overhead. The throughput values decrease occurred to be approximately 10 percent at 2Mbps data rate thanks to the extra information (RTS, CTS, ACK packets) provided by the RTS/CTS mechanism.

ICMP/UDP/TCP measurements

A wireless network can be a good alternative to a wired LAN. Applications must run the same way in wireless environment, so it have to be compatible with wired LAN assuring transparentment for higher OSI layers. Bit Error Ratios on wired LANs are much lower than in wireless LANs. A transfer and datagram protocol performance measurement can be an interesting point of view.

As mentioned above, higher data speeds requires higher SNR for error free transfers. The 802.11 standard requires the control informations to be sent at 1Mbps for safe communication. The table below shows the fields of a frame, indicating the required transmission time versus the applicated data mode.

Parameter	Byte	11 Mbps	5.5 Mbps	2 Mbps	1 Mbps
UDP/IP header	28	20,4µs	40,7µs	112µs	224µs
Mac overheader					
SNAP	8	5,8µs	11,6µs	32µs	64µs
MPDU/FCS	34	24,7µs	49,5µs	136µs	272µs
ACK (*)	14	56µs	56µs	56µs	112µs
RTS (*)	20	80µs	80µs	80µs	160µs
CTS (*)	14	56µs	56µs	56µs	112µs
Physical layer overhead					
PLCP	24	192µs	192µs	192µs	192µs
IFS overhead					
DIFS		50µs	50µs	50µs	50µs
SIFS		10µs	10µs	10µs	10µs

Frame elements

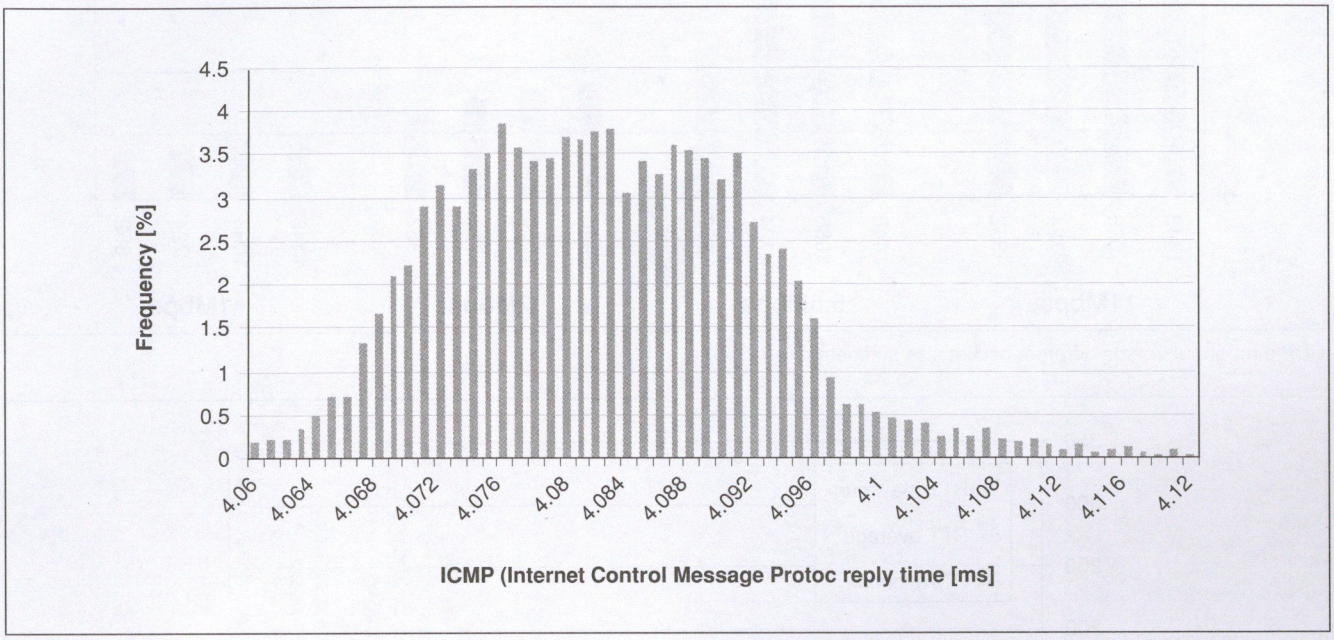
The indicated data modes' speeds stand for the speeds in physical layer. Due to the relatively high overhead the effective speed decreases. The higher the data rate is chosen the higher throughput loss will occur according to the mandatory overheads. The next throughput lowering factor is the packet size. Using smaller packets in the same data rate the throughput will fall, thanks to the fixed sized overhead.

The throughput can be estimated using a simple rule below:

$$T_{put} = \frac{1}{t_{data} + t_{ovehead} + t_{backoff}}$$

where:
t_{data} is the time for transmitting the payload
$t_{overhead}$ is the time for transmitting the overhead
$t_{backoff}$ is the random backoff time used by CSMA/CA

The size of the packet and the utilized data rate determinates t_{data}. The mandatory overhead's overall time gives $t_{overhead}$. The only problematic component is the backoff time ($t_{backoff}$). Measuring the backoff time requires special arrangements and equipments, but the density and the dispersion of it can be measured easily using ICMP (Internet Control Message Protocol) packets. The density and dispersion of ICMP reply time follows the backoff time's density and dispersion. The diagram below shows the dispersion profile of the ICMP reply times.



As a result the average ICMP Reply time is 4080us. This includes an ICMP echo-request and an ICMP echo-reply packet.

UDP is the protocol for real-time applications, where a few packets' loss don't take serious effect.

The following table shows the measured and the calculated throughput values, in function of data rate and size of the packet.

	packet size	calculated throughput	measured throughput	difference
	1448 byte	6,73 Mbps	6,13 Mbps	8,92%
	1086 byte	5,96 Mbps	5,34 Mbps	10,4%
	724 byte	4,85 Mbps	4,23 Mbps	12,78%
11Mbps	362 byte	3,11 Mbps	2,62 Mbps	15,76%
	1448 byte	4,1 Mbps	3,9 Mbps	4,88%
	1086 byte	3,78 Mbps	3,58 Mbps	5,29%
	724 byte	3,27 Mbps	3,01 Mbps	7,95%
5.5Mbps	362 byte	2,32 Mbps	2,08 Mbps	10,34%
	1448 byte	1,73 Mbps	1,77 Mbps	2,31%
	1086 byte	1,66 Mbps	1,68 Mbps	1,2%
	724 byte	1,53 Mbps	1,51 Mbps	1,31%
2Mbps	362 byte	1,23 Mbps	1,21 Mbps	1,63%
	1448 byte	0,9 Mbps	0,98 Mbps	8,89%
	1086 byte	0,88 Mbps	0,96 Mbps	9,09%
	724 byte	0,82 Mbps	0,86 Mbps	4,88%
1Mbps	362 byte	0,7 Mbps	0,74 Mbps	5,71%

The throughput can drastically decrease on noisy channels. The next diagram shows the throughput measured over noisy and ideal channel in function of different data modes and packet sizes. The average level of SNR was 15dB in noisy scenario.

The figure illustrates the modulations employed by higher data rates are much more sensitive to noise. The higher the packet size the higher the probability of an error. Even more increasing the level of noise can easily lead to a radical throughput fall.

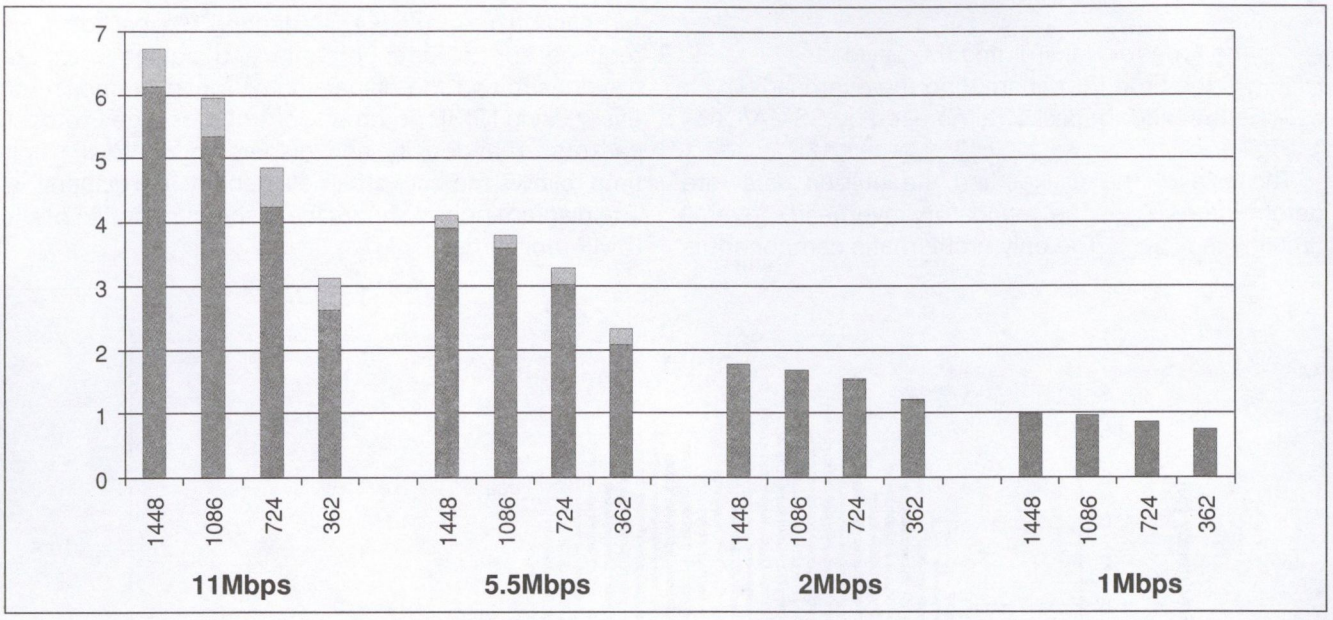
The TCP Protocol can establish reliable end to end connections and provides packet retransmission. An acknowledgement packet follows a successful data transmission. The RTT (Round Trip Time) is a time interval between passing a packet to lower layers and the arrival of the acknowledgement in the TCP layer. The higher the bit error rate, the higher the packet retransmissions, resulting in higher RTT.

Applicating the previous noisy scenario the following RTT profile draws out.

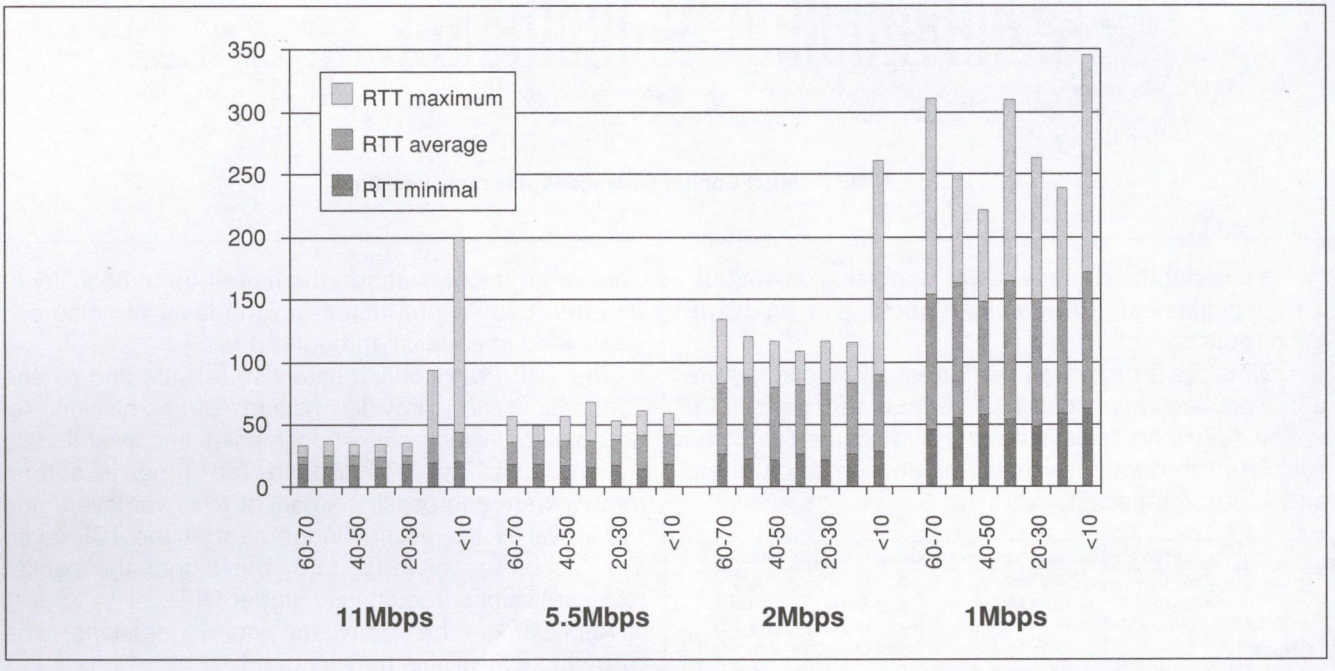
The round-trip time increases drastically in higher data rate modes due to the frequent retransmissions caused by the higher noise sensitivity. According to a random error occur the RTT min is almost constant, while RTT max can be extremely high, especially when a link drop occur. At higher data rates RTT is more sensitive to the noise because of the higher sensitivity of the utilized modulations.

Conclusions

According to our measurements a testbed is needed for developing/measuring a wireless system. We can encounter many serious problems if the adhoc testbed is built in open environment. Filtering noises



UDP throughput versus different packet size and data rate



Round Trip Time versus different SNR and data rate

and jammings are almost impossible. Special arrangements such as hidden terminal can only be established by using obstacles or extreme high distances. For eliminating these problems wired interconnections are needed, so any arrangements can be built up in a small area according to the demands.

References:

1. Robert C. Dixon: Spread Spectrum Systems with Commercial Applications
2. Spread Spectrum Scene at www.sss-mag.com.
3. Lucent Technologies WaveLAN IEEE 802.11 Pc Card User's Guide
4. D. Duchamp and N. F. Reynolds. Measured performance of a wireless LAN. In Proceedings of the 17th Conference on Local Computer Networks, pages 494–499. IEEE, Sep 1992.
5. B. Tuch. Development of WaveLAN, an ISM band wireless LAN. AT&T Technical Journal, pages 27–37, July/August 1993.

Efficient Resource Controller for Software Radios

LÁNYI ÁRPÁD, IMRE SÁNDOR, RÁBAI GYULA

Department of Telecommunications, Budapest University of Technology and Economics

The role of software radio in the world of telecommunications

The evolution of today's telecommunication networks is oriented towards mobility. This means that there is a need to use a unified technology to be able to transmit video, audio and data independent of location. The fact, that different countries and continents are using different radio standards, which require different handsets, is also becoming a major problem. These problems led the creators of the 3rd generation (3G) technology to designing a world- wide family of systems called the IMT-2000 (International Mobile Telecommunication). The implementation of this system however has to overcome many difficulties. The most important of these is that, well performing systems already exist and these systems should be developed to a level, where they can be called 3G [1].

This aim is satisfied in many ways in the 4th generation software radio concept [2], which targets the integration of telecommunication systems existing in different countries and continents. Imagining ourselves as a businessman traveling between different parts of the world, we need to carry different mobile handsets with different phone numbers and we need to have different telephone subscriptions from place-to-place. Wouldn't it be better to only carry one universal handset, which is capable of downloading the appropriate software for the current network? This universal mobile handset would be able to work in all parts of the world.

The downloaded software could be an application converted to JAVA byte code to operate on the JAVA virtual machine implemented in the universal hardware. This way a platform independent environment could be used in these handsets and all reconfigurable hardware implementations could be treated in the same way (Figure 1).

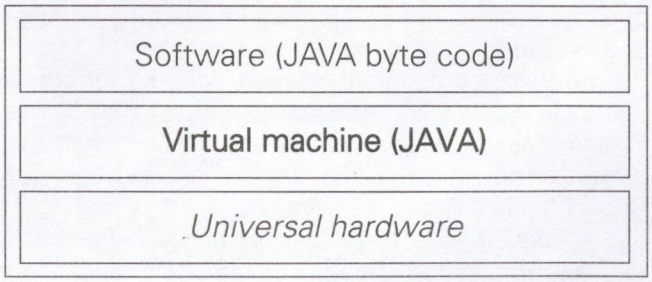


Figure 1. Software Radio Layers

In the following part we are going to present the architecture introduced in the CAST Software Radio project supported and sponsored by the European Union and we are going to describe the concept that helps the architecture used to use the resources of the system efficiently.

The CAST software radio system

The CAST (Configurable Radio with Advanced Software Technology) project is part of the European 5th Framework Scientific Research program, introduced in February 1999. Currently seven companies and universities are taking part in this research, with MTA SZTAKI being the only Hungarian participant. As part of the MTA SZTAKI team, we are taking part in this research.

The goal of the project is to develop an end-to-end software radio (SWR) system. To be more precise we need to construct a demonstration which implements a well-defined set of services needed in the software radio environment, to prove that the software radio concept can be implemented. In the CAST project we are working with two standards: GSM and UTRA-FDD. The user can place a call in one of these systems and can communicate with another user, who is not necessary using the same system. The terminals must be capable of switching between standards without interrupting the call. This is only possible if the software code needed to use the other system is available in the terminal. If it is not, this code needs to be downloaded from the network and should be installed on the hardware. The software for communication must be available in the base station as well in order to be able to use the service.

The system has two major parts from the point of view of network management. One part is the global management, which is responsible for the

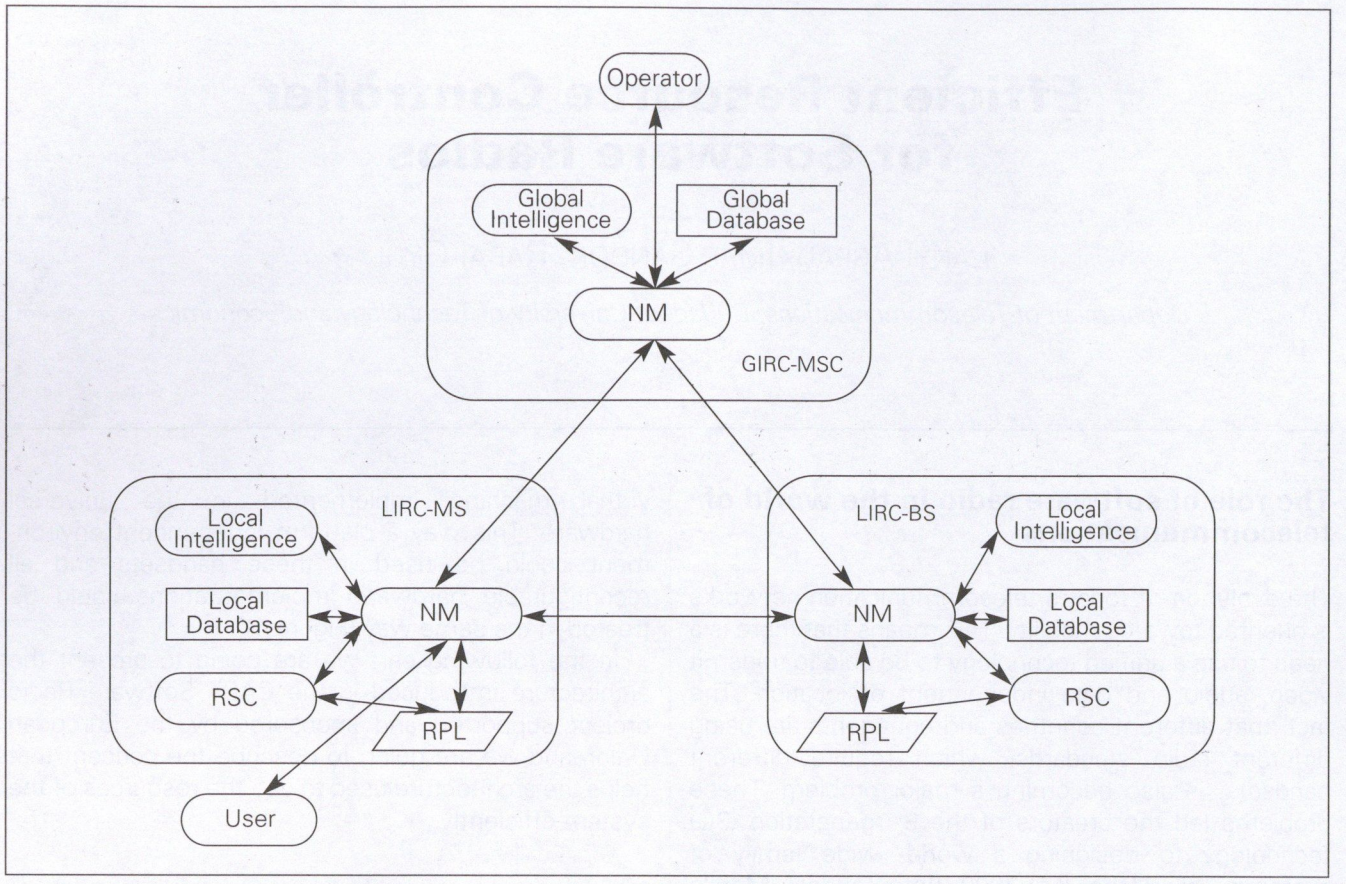


Figure 2. The CAST SWR system's detailed diagram

functionality of the whole network. The other part is the local management, which is providing the intelligence needed by the local entities (mobile handset, base station) in order to operate [3].

To have a clearer picture we should look into the details of the network architecture.

As seen In Figure 2. The local and global part of the intelligence and the management are well separated. The name of the global part is GIRC (Global Intelligent Reconfiguration Controller) and the name of the local part is LIRC (Local Intelligent Reconfiguration Controller). Both components are depending on the NM (Network Management). The role of the Network Management is to make decisions, to direct other components and to establish communication between the global and local part. The users of the network (operator, customer) are also in contact with the network management.

The work of NM is based on the intelligence. Its role is to monitor the network and if it detects some anomalities it notifies the NM and makes some suggestions. For example in the local part, the intelligence is monitoring the quality of service in the radio channel and if it detects that the signal/noise ratio reaches a level where the desired service cannot be provided, it can suggest the NM to use a more robust coding scheme for channel coding. Another example would be that the intelligence could make predictions based on historical data. These predictions can be used to configure the hardware to specific services even before the user requests the services. This way the system can provide a more convenient service for the user.

The database used by the system can be split into different parts based on where the data is stored. In the global part the data is stored in a central database. In this database, among others, user information and network configuration is kept. The user information includes the possible services a user can use based on the contract between him and the service provider. The network information is about the configuration and the state of the whole network. In the local parts data is stored at base stations and mobile handsets. At the base station, data describing the mobile handsets and users that are in the base station's coverage area is present. In addition to these detailed information about the base station's hardware and the nearby base stations are also stored here. In the mobile handsets, due to lack of resources information about the terminal and its user is present.

The third component needed for the effective operation of the system is the RSC. The implementation of this component is our task in the project. The short name is an abbreviation for Reconfigurable Resource Controller.

The role of RSC is to satisfy the reconfiguration requests originated from the LIRC-NM and feed information to the network management about the

configuration status. There are many types of reconfiguration requests, for example service installation, service startup or service suspension. A service can be voice transmission, video transition or data transmission with different parameters [5]. Such a service is implemented by the concatenation of basic functions in a way where the output of a function is fed as input to another function. These set of functions are called processing chains. The functions of the processing chains are operating on physical devices, which form the RPL (Reconfigurable Physical Layer).

The reconfiguration is performed in the following way: The LIRC-NM sends a request to the RSC, which acts. If the request was a service installation, RSC tries to find optimal locations for the functions on the hardware. If it fails to find a suitable configuration it notifies the Network Management, otherwise it downloads the appropriate software and installs it on the hardware. After the installation the RPL returns the final status of the configuration to the LIRC-NM and the RSC. The LIRC-NM needs this information for future decision-making.

All parts of the CAST system are developed using advanced, object oriented programming languages, such as JAVA and C++. One of the major advantages of these languages is their ability to help trouble shooting and that the final code created by them is small. Small code size provides two benefits. First of all it only uses a small portion of the terminals valuable resource: the memory, second it takes small time to download. Another advantage of the used technologies is that parallel development of well-defined parts of the software is easy with them and the cooperation between the different parts can be easily implemented. The language used to develop the RSC is JAVA.

In the next part the architecture and the operation of the RSC is demonstrated and the algorithms that make optimal resource management possible are overviewed.

The RSC architecture

RSC has a modular design (Figure 3.). The core module is the CCI (Core Controller Intelligence), which is the heart of RSC. Its role is to receive the requests coming form the above layers and place them into the Queue Handler. The Queue Handler is storing these requests, not like a tradition FIFO storage, but in a preemptive way. The Queue Handler's aim is to balance the incoming requests, so the RSC won't be overloaded, which would result in decreased performance of the system. The CCI is responsible for the creation and management of the so-called processing modules. The modules are doing the actual processing, in other words they implement the algorithms necessary to find the appropriate hardware for a certain function.

The processing modules are not part of the CCI. They are totally independent. This way parallel processing is possible in multiple threads, which is an advantage. This way the time spent on reconfiguration can be decreased. Another advantage is that this way redesign and reconstruction of a certain module is possible without touching the rest of the system.

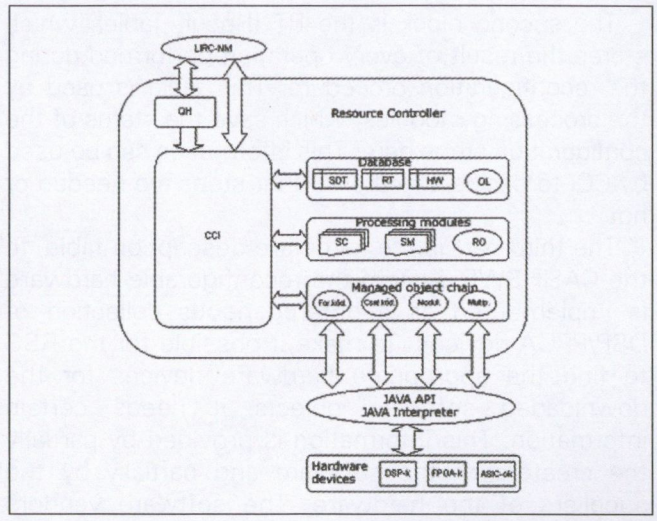


Figure 3. The CAST Resource Controller

The following three types of the processing modules are available:

- The SC (Service Configurator) is the service configuration module. Its role is to do the main computations needed by the installation and configuration. The CCI always instantiates this module when a new service needs to be installed or an existing service should be reconfigured.
- The SM (Service Modifier) is the module for modifying services, which means that only the parameters of the installed functions are changed when reconfiguration is performed. This is the case when the coding scheme or the compression technique needs to be modified, because circumstances have changed.
- The RO (Resource Optimizer) is the resource optimization module. It works on creating an optimal configuration automatically without any external requests. One of the optimization technique implemented by this module is load balancing. This automatic optimization leads to more efficient usage of the hardware. This module can only exist in one instance since it operates on the whole hardware.

The database block seen in Figure 3 is split into four major parts. The first part is the SDT (Service Description Table). This table stores the already installed services (the working, suspended and partitioned chains) and the requests that were taken from the Queue Handler. Each record in the table has a status indicator, which stores the current status of the installation procedure. SDT also stores the data representing the request and an object reference. The

value of the object reference depends on the state of the installation procedure. If the chain is in the installation phase the reference to the object performing the installation is kept here, if the chain is already installed, the reference to the installed chain object is stored.

The second block is the RT (Result Table), which stores the result of every operation performed during the reconfiguration procedure. This table is used by the processing modules, which save the status of the configuration steps here. This information can be used by CCI to decide whether further steps are needed or not.

The third part is the hardware description table. In the CAST SWR system the reconfigurable hardware is implemented as a heterogeneous collection of DSP/FPGA devices. To make it possible for the RSC to find the appropriate hardware devices for the downloaded software objects it needs certain information. This information is provided by partially the creators of the software and partially by the suppliers of the hardware. The software vendors must supply information about what kind of hardware is required for operating the software. For example in case of a DSP, the memory requirements and the desired CPU time should be specified. On the other the hand RSC needs detailed description of the hardware. The hardware provider must supply the initial description. This description is used and modified during the operation of the hardware based on the usage statistics. The hardware can be modeled as a graph. An example graph can be seen in Figure 4.

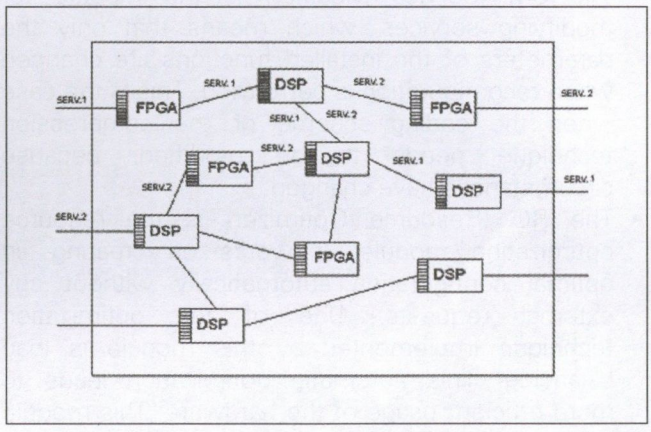


Figure 4. The hardware model – a graph

The nodes of the graph represent the different hardware devices while the edges describe the connection between them. The nodes of the graph can have certain properties, which describe the usage of the device represented by the node. For example a DSP would have a property describing the available memory resources and the available CPU time. These values are modified by RSC according to the device configuration.

The fourth part is the OL (Object Library), which stores the code of the functions implementing the different service building blocks. These objects are called BPC (Base band Processing Cell), which includes the device-dependent code implementing the given function on a specific device (on a DSP device this would be an assembly code, on an FPGA a configuration bit steam). Also the JAVA application needed to manage the hardware device is stored here as a byte code representation. These can be found in Figure 3 in the block named Managed Object Chain.

The RSC Framework

During operation RSC is communicating with two neighboring units (Figure 3.), the network management (LIRC-NM) and the hardware unit (RPL). This communication requires well-defined interfaces.

One of these interfaces is the channel that is used to send requests to the resource controller and to receive the responses to these requests. First of all the format of a reconfiguration request needs to be designed. We need to decide what kind of information is needed from the LIRC-NM. If we require only a small amount of information, the processing can be done faster and the communication time can be reduced, however we need to have a larger database in the RSC. For example if all services are stored in the local database at the RSC, the LIRC-NM only has to pass the identifier of the service. In this case the local database would be quite large, even if we consider only the two standards of the CAST project. The database would have hundreds of records with high redundancy. On the other hand if we require more information from the NM, we need to implement less intelligence in the resource controller. We can use a smaller database. In this scenario the tradeoff is that the LIRC-NM would need detailed information of the hardware to be able to provide reconfiguration requests.

Considering these aspects, in the implemented solution we use a smaller database, mainly because a mobile handset is usually lack of resources such, as memory. This means that we expect most of the information from the LIRC-NM. In this case the object library we only store the required BPCs (the java code needed for management and the hardware dependent code.) Each BPC has an identifier, which can be referenced by the NM. Thus the LIRC-NM should send

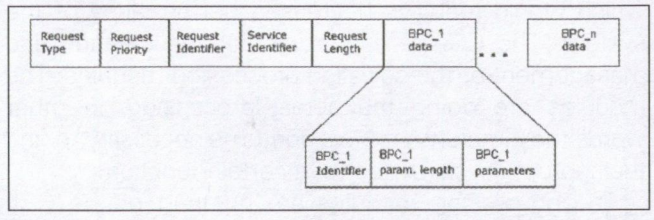


Figure 5. A reconfiguration request

the following information as part of the request: what type of BPC-s, with what kind of parameters and in what order should be installed on the hardware to implement a service.

Now let's take a look at Figure 5, to examine how a reconfiguration request is made up:

The first field determines the type of the request, which could be installation, activation of an already installed chain, deactivation and deletion of the chain.

The second field represents the priority of the request. This is needed by the QH for the preemptive handling. The third is the request the fourth is the service identifier. These are needed when the NM instructs the resource controller to terminate a service. They are used to identify the service. The fifth field shows the length of the request. These first five fields are required in each request while the following fields are only needed for installation. The additional fields contain the BPC identifiers implementing the service. The BPCs are JAVA classes, with appropriate constructors, that make it possible to feed different parameters to the objects while instantiation. The network management also provides these parameters. The order of the BPCs in the request determines in which order should they be concatenated. RSC should provide a positive or negative response to the request for the upper layer. The positive answer only consists of the request identifier and the „OK" signal. The negative response besides the identifier contains the cause of the failure. This cause is received from the PLC-RSC communication.

The other interface is located between the RSC and the RPL. This interface implements the following functionality:

- Downloading of the hardware dependent code implementing the functions.
- The concatenation of these functions to form a chain
- Startup and shutdown of a service
- The removal of the code from the hardware device
- Handling of error messages

The developers of the hardware layer provide us with a JAVA API [6]. With the help of this API we attach a JAVA object to each device in the system. This way we can handle the hardware as if it was a set of JAVA objects. To work with the hardware the RSC performs method calls on the objects. For example the PlaceFunction() method installs a function on the device and the Remove() method deletes it. Besides this the RPL must feed information to the RSC about the outcome of the installation procedures.

The RSC resource management

The primary goal of the RSC is to quickly and efficiently install the functions onto the hardware devices, which are needed to receive and process the data stream arriving through the radio channel.

Efficiency in our case means two things: The selection of the optimal hardware device based on parameters describing the capacity (e.g.: memory, processing time, IO ports, etc.). This hardware device needs to be capable of fulfilling the requirements of the desired service. And we need to reach even utilization of the hardware.

It is very important for the RSC to be able to perform this selection as fast as possible, since the response for a reconfiguration request should be returned in a certain amount of time. Also a quick response is important because making a resource management decision also requires resources and big delays can result in user dissatisfaction. Thus the RSC must comply with requirements opposed to each other: efficient resource selection and short response time is needed at the same time. To be able to satisfy these requirements RSC has two options:

- If the service is not already configured, it selects the components that will be fit for the job with the help of an optimization algorithm.
- If the NM requires the termination of a service, it might not bi wise for the RSC to delete it from the hardware, because in the next request the already installed chains might be used again. This is true not only for installed chains, but for configured chain portions, processing functions as well. The unused, but already installed BPCs can be valuable building blocks in the construction of the next processing chain. If we use these reconfigured building blocks the total configuration time can be significantly reduced. The resource controller maintains information about the unused, but installed BPCs in the SDT.

Top sum it up, if the upper layer (NM) sends a request, the RSC examines the service identifier, and scans the SDT list for the desired service to find an already installed chain that can fulfill it. (Figure 6.) If such a chain was found RSC determines whether it is in use or can be activated to start the service. If no

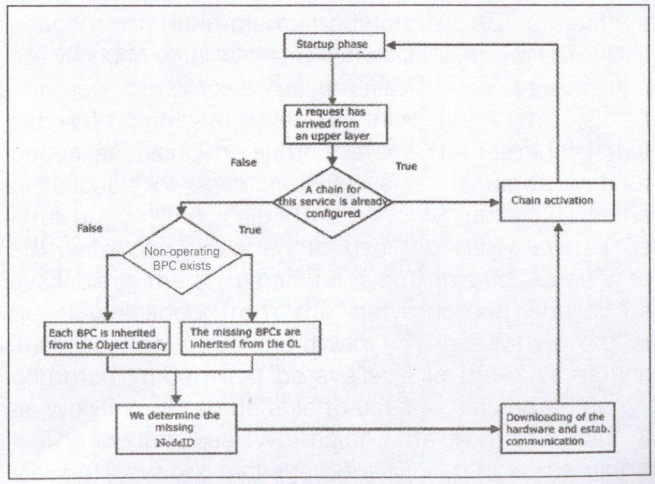


Figure 6. The procedure of hardware device selection

such chain is found, it checks the table listing of the unused BPCs and looks for BPCs, which can be used while constructing a new processing chain. If it finds some, it only installs the missing chain elements and includes the already configured BPCs in the chain. The installation is done by instantiating a chain class and finding a hardware device that will be able to run the function. The hardware devices are identified with the NodeID field. If no already configured portions are found the installation procedure should be performed for all parts of the chain. Finally the installed or pre-configured functions must be connected in order to form a processing chain.

Conclusions

As we mentioned in the introduction, the demand for a worldwide mobile communication system, that is capable of carrying voice, video and data, is growing. The aim of the 4th generation software radio systems is to provide similar service to the user on all mobile networks.

We have worked on creating a software radio system in the frames of the CAST project sponsored by the European Union:

The concept of Software Radio is, that if a user would like to use a service (e.g.: GSM voice, IS-95 voice or UMTS data communication), he needs to download software from the network that implements the communication standard and the specific service. To be able to satisfy this concept a general-purpose hardware device is required, that is capable of executing the downloaded code. Software implementing a specific service can be divided into smaller units called functions. These functions form a processing chain. In GSM for example the following functions are members of this chain: segmentation, voice coding, channel coding, interleaving, modulation. After the software functions are downloaded, they need to be installed on the hardware. This procedure requires an efficient resource controller framework, which is capable of creating a nearly optimal hardware configuration. In the CAST project our task was to create such a framework. We named it RSC. During the design we had two major goals: efficiency and speed.

The final version of the system is going to be ready by the fall of 2002.

References

1. Dr. Dárdai Árpád: Mobil Távközlés, 1996
2. W. H. Tuttlebee: The Softwer radio concept, IEEE Communications Magazine, 2000
3. J. Mitola: The Softwer radio architecture, IEEE Communications Magazine, 1995
4. Matthias Forster: Proposed Design for the network management for CAST, 2001
5. CAST Initial Description of the Baseband and its Modules, 2001
6. Recommendations for an Object Oriented Interface to the Reconfiguration Resource Controller, HW Communications Ltd, 2001
7. CAST Structured Object Representation-Initial Draft Specification, HW Communications Ltd, 2001
8. Joseph Mitola-Software Radio Architecture, Object-Oriented Approaches to Wireless Systems Engineerind, Wireless Architectutes for the 21st Century, Fairgax, Virginia, Usa, October 2000, ISBN 0-471-38492-5

Abbreviations

UMTS	Universal Mobile Telecommunication System
CAST	Configurable radio with Advanced Software Technology
GIRC	Global Intelligent Reconfiguration Controller
LIRC	Local Intelligent Reconfiguration Controller
RSC	Reconfiguration Resource Controller
RPL	Reconfigurable Physical Layer
ASIC	Application Specific Integrated Circuit
DSP	Digital Signal Processor
FPGA	Field Programmable Gate Array
UTRA-FDD	UMTS Terrestrial Radio Access
CCI	Core Controller Intelligence
BPC	Baseband Processing Cell
IMT	International Mobile Telecommunication
SWR	Software Radio
QH	Queue Handler
SC	Service Configurator
SM	Service Modifier
RO	Resource Optimiser
SDT	Service Description Table
RT	Result Table
OL	Object Library

Wireless Local Area Networks – Business Opportunity or Niche?

J. F. HUBER

Siemens AG

An increasing number of promoters in the industry, WLAN vendors and Internet service providers highlight the merits of public WLAN access. The paper gives an insight view on WLAN technology, which could be seen as a means to complement wireless wide area technology such as 3G/UMTS, but which could also attack in some service areas. The emerging wireless personal area technology called Bluetooth will have a further impact on these developments.

1. Introduction

Wireless access to the Internet allows a more convenient way to work and to play with lightweight computing devices, which are flooding into the market. Computer software allows connectivity to local areas and to wide area networks. Business users can be connected to their Intranets via corporate Wireless LANs (WLANs), access to the public Internet emerges in a number of countries. Bluetooth is on its way to the market place. In the light of development, the question arises: "Will WLANs succeed in the public market, will they beat or complement Wireless Wide Area Networks such as 3G/UMTS and what will happen, if new technologies like Bluetooth will emerge?" Additional questions relate to frequency resources, co-existence of such technologies in the same bands, whether they are unlicensed or for exclusive operator use.

A WLAN is an open data communication system for wireless access to the Internet/ intranets. It also allows LAN-to-LAN connectivity within a building or a campus. A WLAN can be integrated and used in conjunction with wired networks and can be integrated with wide area wireless networks under certain restrictions. The WLAN bitrate requires adequate transmission support from the backbone network.

Currently there are several standardisation organisations (Table 1.) dealing with WLANs. The dominating standard is presently IEEE 802.11b, emerging for higher bit rates is IEEE 802.11a. HiperLAN2 will converge with the emerging IEEE 802.11a standard and will operate in the 5 GHz range. The Worlds Radio Conference 2003 will decide on the detailed specifications of these bands. The 5 GHz bands are not subject to interference from other technologies, which operate in the crowded 2.4 GHz ISM band. WLAN developments also exist in Japan.

2. Standards

The IEEE 802.11 Wireless Local Area Network Standard

WLANs are used mainly in the corporate sector, but also in small office environments and in homes. The client/server network as shown in Fig. 1. uses an access point that controls the allocation of transmit time for all stations and allows mobile stations limited roaming from cell to cell within the same WLAN environment. Roaming between different WLANs may cause interoperability problems and is dependent on the backbone network. The access point is used to handle traffic from the mobile radio access system to the wired or wireless backbone of the client/server networks. WLAN cards give laptop and desktop users access up to theoretically 11 Mbps. The cards are also

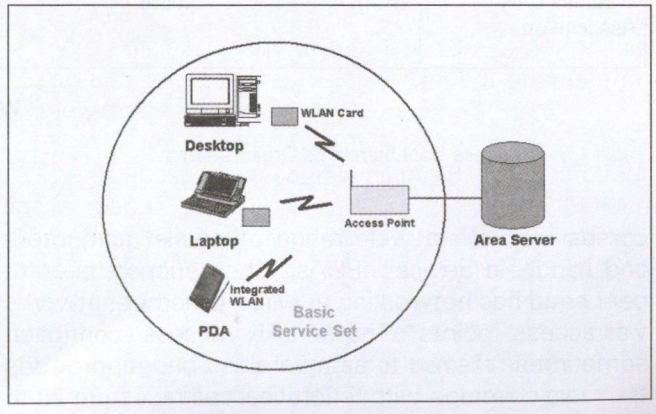


Figure 1. WLAN Connectivity

Organisation	Acronym	Country	Purpose
Institute of Electrical and Electronics Engineers	IEEE	global	Promote the engineering process of electrical and information technologies and sciences.
Federal Communications Commission	FCC	USA	Establish policies that govern interstate and international communications by TV, radio, wire, satellite, and cable.
Multimedia Mobile Access Communications Promotion Council	MMAP-PC	Japan	Develop high performance wireless systems.
Conference Europeenne des Administrations des Postes et Telecommunications	CEPT	Europe, Africa parts of Asia Arabian States	Establish frequency spectrum policies for Europe, GUS and African States.
International Telecommunications Union	ITU-R	global	Radio Standards Frequency Regulations
Wireless Ethernet Compatibility Alliance	WECA	global	To certify interoperability of IEEE 802.11b products. Products bearing the Wi-Fi [1] symbol have passed independent testing for interoperability with other Wi-Fi certified products.
Wireless LAN Association	WLANA	global	Education organisation.

[1] WiFi – Wireless Fidelity

Table 1. Wireless LAN Standards Organisations

compatible with a wide range of pocket computers and handheld devices (PDAs). They support peer-to-peer or ad hoc networking to wired Ethernet networks via access points. The WLAN access controller, sometimes referred to as a wireless bridge, provides the radio coverage with theoretical cell radii from 20 to 100 m indoors [2].

IEEE 802.11a

This standard specifies a different physical layer than 802.11 (and 802.11b) by using orthogonal frequency division multiple access (OFDMA). The target is to achieve 54-Mbps bit rates with a frequency bandwidth of 20 MHz. The IEEE 802.11a specification amendment from 1999 defines the physical and the MAC layer. 802.11a products are foreseen to operate and, according to the ITU WRC 2000 resolution 736, also in the 5.470 to 5.725 GHz and 5.15 to 5.35 GHz range. A new development is Wi-Fi 5: The Wireless Ethernet Compatibility Alliance (WECA) says Wi-Fi products based on 802.11a could be introduced in Europe by the end of the year 2002. WECA has completed the test plan for 802.11a products and plans to start certifying interoperability depending on the availability of hardware.

IEEE 802.11b

This standard is expected to be dominant in the market until 2003/04. Products based on this specification work in the ISM band. Their physical layer uses DSSS, whereby low power consumption is achieved in contrast to earlier IEEE 802.11 systems and in contrast to IEEE 802.11a.

802.11 b provides theoretically max. 11 Mbps and has a carrier bandwidth of 22 MHz, up to 3 systems can operate within the total bandwidth of the 2.4 GHz ISM band. IEEE 802.11b products are available in all parts of the world. Their technical data are as follows:

- Frequency range ISM band, 2.4-2.483 GHz;
- 5 MHz carrier channel raster in CEPT countries; ≤ 22 MHz carrier bandwidth;
- Direct sequence spread spectrum modulation technique (11 chips code);
- PCMCIA type II interface;
- Media access protocol CSMA/CD with ACK;
- Frame error rate < 8%;
- Peer-to-peer or point-to-multipoint communication;
- Power consumption:
  Transmit approximately 300mA;

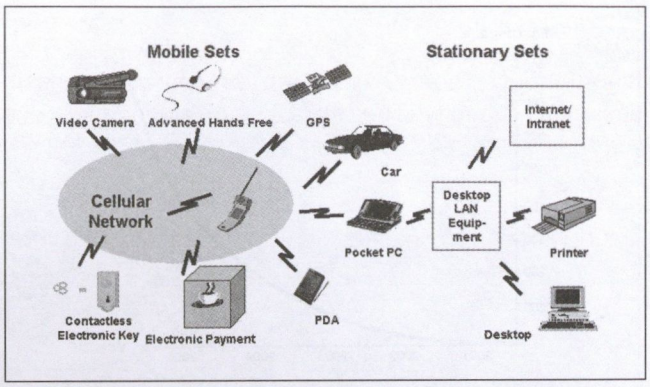


Figure 3. Bluetooth Connectivity, Source [2]

Each device is equipped with a microchip transceiver that transmits and receives data in the industrial-scientific-medical (ISM) frequency band of 2.4 to 2.4835 GHz that is available global (with some variation of bandwidth in the different countries). In addition to data, up to three voice channels are available. Each device will have a unique 48-bit address from the IEEE 802 standard. Connections are one-to-one. The maximum range is 10 m. The aggregate bit rate is 1 Mbps (up to 2 Mbps).

A frequency hop scheme (slotted TDMA) allows devices to communicate even in areas with a great deal of electromagnetic interference. Built-in encryption and verification is provided to make sure that the right information arrives secure at the right partners. Bluetooth's master-slave network configuration allows point-to-point communication, piconets (a cluster of eight devices), and scatternets (interconnected piconets, 10 clusters maximum). Manufacturers offer products, which provide voice and data connectivity.

There are five primary criteria for the deployment of Bluetooth:

1. Small implementation;
2. Open specification;
3. Low power;
4. Low cost;
5. Ad hoc connectivity.

From a networking point of view, there is great interest to integrate Bluetooth with wide area networks like UMTS and the Internet. One remaining item of Bluetooth is interoperability. The Special Interest Group (SIG) has resolved some of these problems, although there is still some work to do. For example, Bluetooth products need to be tested at a qualified test facility to ensure compliance with specifications. Testing of Bluetooth products is currently done against designated protocol test products called Blue Units. These can test a number of key functions, but their use is limited to partial testing of the baseband and link management functions. Co-existence with other radio interfaces integrated into the same device (e.g., GSM, GPRS, UMTS, IS-95) is of great importance [2].

Main Bluetooth Parameters

- Class 1 allows +20 dBm maximum output power with a maximum distance of up to 100 m (depending on the environment).
- Class 2 allows +4 dBm maximum output power with a maximum distance of up to 10 m.
- Class 3 allows 0 dBm maximum output power with a maximum distance of 1 to 2 m.

Some companies are already working on Bluetooth Version 2, which will have greater range (approximately 100 m), more bandwidth, and more participants. Some companies of the Bluetooth SIG want to increase the capabilities of Bluetooth so that it becomes more like a wireless LAN and maybe even someday replaces the wireless LAN in various scenarios.

WLAN Practical Bit Rates, Bandwidth Demand and Capacity

WLANs are preferably promoted with 11 Mbps however, this theoretical value does not take into account the radio transmission protocol and the impacts of the 'hidden terminal' problem.

The practical bit rates are therefore lower (55% and less) and they are dependent on the number of users and their environments. Some manufacturers also provide solutions that can be integrated into a public wide area network (e.g., by using the SIM card for GSM access, the wireless LAN terminal will get the SIM inserted and have access software to communicate with GSM). The WLAN is connected to the GSM network by a server, which takes over authentication and mobility management. For 3G networks like UMTS, such functionality is presently under discussion in the 3GPP standardisation groups.

Practical cell radii vary depending on the product. Benchmark tests have shown that IEEE 802.11 frequency hopping spread spectrum (FHSS) techniques reach their limits at 3 Mbps, while the direct sequence spread spectrum (DSSS) techniques may achieve higher bit rates up to 6 Mbps depending on transmit power, cell size, number of simultaneous users, application, and cellular environment. Products from different manufacturers were benchmarked with file transfer of 64-MB date files and TC/IP traffic with 4-KB block size. The throughput (0,5-5 Mbps) results varied depending on the terminal distance (max. 45 m) to the access point.

The total throughput of a WLAN depends further on the number of active users per cell Fig. 4.

WLANs according to the IEEE 802.11b standard require 22 MHz minimum bandwidth (Hotspot). Up to three parallel WLANs can operate within the ISM band of ~80 MHz. The capacity of a system is a non-linear

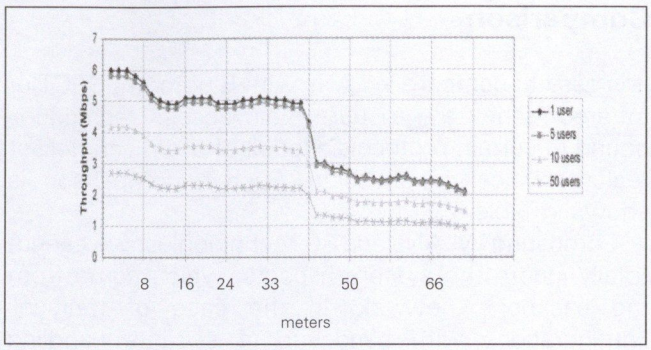


Figure 4. Performance of 802.11b in Hot Spot (Nomadic/Non-Contiguous Coverage)

function depending on environment distance and number of simultaneous users. As shown in Fig. 4. bit rates between 500 Kbps and 2 Mbps seem to be feasible, without considering the backbone network transmission capacity. Up to 2 Mbps bit rates will almost be available in hotels, airports, public buildings etc. A sample calculation for a given number of users is shown in Table 3.

Radio Coverage and Transmission Capacity Requirements

There is no doubt, that WLAN technology should not be used for universal coverage. Even in cities with extensions with 10 by 10 km, the number of access points would go up to the order of thousands. For example a typical 3G/UMTS cell area of 1 km^2 would require 1000 WLAN sites. All these sites would have to be supplied with expensive wireline transmission backbone capacity. The bit rate of these connections has to be around 4-6 Mbps in order to allow the WLAN access rates.

Wireless Wide Area Networks 3G/UMTS

Until the year 2002 more than 100 frequency licenses have exclusively been granted for 3G network deployments guaranteeing exclusive frequency spectrum in the order of approximately 2 x 15 MHz + 5 MHz spectrum per operator. In addition, CEPT has designated license exempt spectrum for the use of UMTS/TDD technology in the bands from 2010-2020 MHz. Dualmode 2G/3G terminal technology will further enable every user to access a mobile network with either UMTS or GSM/GPRS technology. This means for the user nearly global 2G/3G coverage is provided – at least for basic services, because there are 500 GSM networks in more than 170 countries in operation. UMTS offers bit rates up to 10 Mbps (HSDPA – High Speed Downlink Packet Access) depending on mobility parameters and up-/downlink directional traffic requirements. However, besides such peak rates, UMTS is designed mainly for high capacity with medium bit rates per users, say to several hundred KBPS. Practical user bit rates depend

Active users	Expected total throughput	Cells	Cell size (km^2)	Frequency bandwidth demand	Required backbone capacity
10	2 Mbps	1	0,05	~ 20 MHz	3-4 Mbps
20	4 Mbps	2	0,1	~ 44 MHz	5-6 Mbps
30	6 Mbps	3	0,13	~ 66 MHz	7-8 Mbps

Table 3. WLAN Capacity per Hotspot

User Mobility	High Mobility 500 km/h	Medium Mobility 120 km/h	Low Mobility
Cell Type	Macro	Macro/Micro	Pico
Maximum User Bit Rate	144 kbps	384 kbps	2 Mbps < 10 Mbps Downlink

Table 4. UMTS User Bit Rates

on the services needed. The continued development of compression techniques for audio/video will allow streaming for small screens from 40 kbps up to 384 kbps (high quality). As UMTS is already in operational use, the first 3G/UMTS terminals confirm such bit rates up to 384 kbps.

The bit rates in Table 4. are specified for both operational modes, UTRA-FDD and UTRA-TDD. TDD is a comparable radio access technique to WLANs, however, it allows delivery of all services (e.g. voice, video and data) in all environments from pico to macro cells.

Besides business related questions, standardisation impacts have to be classified, the 3GPP standardisation project discusses three issues:

- Complexity and degree of inter-working.
- Is 802.11 the right technology?
- Timing of specifications.

Regarding the technology question, a comparison of the terrestrial radio interfaces shows differences in features and in spectral efficiency figures in contrast to UTRA-TDD (Fig.5.).

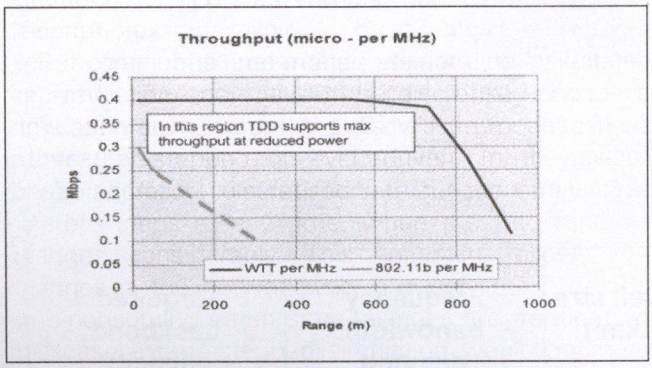


Figure 5. Throughput for Single Cell – Micro Deployment UMTS/TDD vs. 802.11b

Another issue is the network integration, which needs to be investigated. On the 3G side, the USIM card is associated with the mobile backbone network. WLAN access would only be possible with the USIM related network operator and its roaming partner operators. Thus a solution with a Radius Server and/or Multiplexer Switch to the Internet will be required Fig. 6.

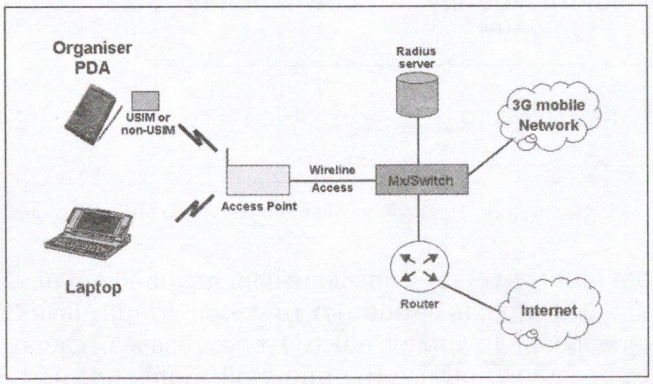


Figure 6. WLAN Connection to Backbone Networks

Comparison

It is easy to compare WLANs with Bluetooth, DECT or Infrared – they are all pure radio access techniques bound to certain protocols. Not all of them can deliver realtime voice transmission. Such a comparison is shown in Table 5.

Comparing WLANs and 3G technologies, we cannot totally ignore the system aspects, which come from the backbone network. In the case of multicell configurations with handover, in security handling transmission quality are depending on backbone capacity, interoperator roaming etc.

WLANs in many cases do not support functions which could be activated from the backbone side. Some functions will be required for public applications. A few examples are given below.

From a radio access' point of view, the main differences are given by the characteristics of the radio technologies regarding cell size, mobility (low mobility/pedestrian or vehicular), multi-user radio control (e.g. power control), roaming. From a system's point of view, the differences lie in the support of the users for registration, security, QoS management of features and charging, in the handling of handover and inter-operator roaming. So far, WLANs cannot directly be compared with wireless Wide Area Networks, because

- they are radio access technologies only
- they are dependent on their backbone network.

A direct comparison between the TDD radio access of 3G/UMTS and WLANs was given in Fig. 5. Spectral efficiency figures for UTRA/TDD are available and they are dependent on user mobility and service [2]. Spectral efficiency figures are dependent on many parameters. Also, the crowded ISM band may impact efficiency, depending on the environment.

The integration of WLAN into fixed or mobile networks is seen as one possibility for network operators to offer public wireless Internet access according to the IEEE 802.11 standard in order to enhance the 802.11 capability with 3G features, the infrastructure network and USIM/SIM security. Restrictions will come from the WLAN itself; they are related e. g. to the voice service, as long as the backbone does not support VoIP or to the maximum bit rate, which the backbone network can provide. Mobility support, user roaming, may be another area of restrictions.

Available Frequency Spectrum

The bandwidth allocations for wireless systems are partly harmonised world wide, however, they differ in size from country to country. This is valid for the ISM band (WLAN, Bluetooth) as well as for cellular 2G/3G bands.

	WLAN 802.11a	802.11b	Bluetooth	UTRA TDD	DECT+
Theoretical Bit Rate	54 Mbps	11 Mbps	1 Mbps	2(<10)Mbps(DL)	2 Mbps
Practical max. Bit Rate	n. a.	5.5 Mbps		1 – 2 Mbps	n. a.
Distance Max (pract.)	n. a.	60 m	10 m	<1 - 3	100 m
Interference	Low	High	Medium	Low	Low
Frequency Range	5 GHz	2.4 GHz	2.4 GHz	2 GHz	1.8 GHz
Carrier Bandwidth	20 MHz	22 MHz	1 MHz	1.6 < 5 MHz	1.8 MHz
Pros	Spectral Efficiency	Products available	Data + voice Products avail. Low price	Data + voice lic. spectrum UMTS component	Data + voice market exists, products available
Cons	Data focused costs	Data focused		Products not yet available	No global spectrum

Table 5. Comparison of Radio Access Techniques

Bluetooth and WLAN according to IEEE 802.11b share the same ISM band from 2400 to 2483.5 MHz. France, Japan and Spain have limited bandwidth of 22 to 44 MHz inside this band. The situation will be improved for 802.11a (and HiperLAN), because these systems will probably get more than 200 MHz as exclusive bands for these technologies. Cellular 2G/3G networks will have in total 560 to 580 MHz; the new extension bands in 2.5 GHz included. Fig. 7. gives an overview on the band allocations. For IEEE 802.11a and HiperLAN products, frequency spectrum is partly existing and foreseen in the 5 GHz range. The Worlds Radio Conference 2003 will decide upon these band allocations.

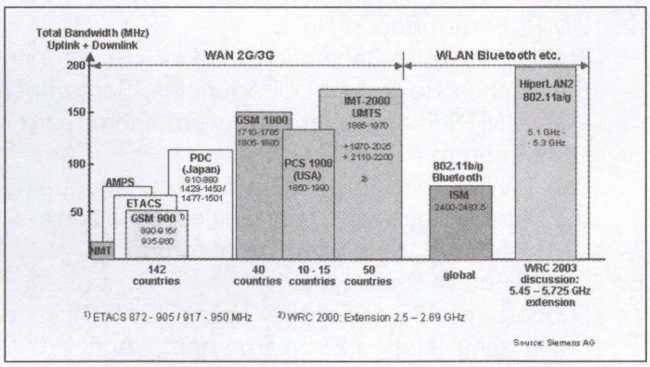


Figure 7. Available Spectrum for 2G/3G Services and WLANs

It should be known, that the 2G/3G cellular bands are mainly used as exclusive bands per operator. This is not the case for WLAN, Bluetooth – the systems work in unlicensed bands and have therefore to cope with interference from various technologies – whether they are standardised or not. Bluetooth seems to be more robust then IEEE 802.11b, because of its 1 MHz carrier combined with frequency hopping across the whole ISM band.

Impact on 3G Services Revenues Caused by WLANs?

The UMTS Forum identified six services categories, that represent the forecested user demand. These service categories are clearly defined from a user's perspective. As shown in Fig. 8. 3G goes away from a voice centric environment towards data, whereof a great portion will be independent from Internet (e. g. Multimedia Messaging with user generated content) or will be very much more mobile specific than Internet specific. For example, location based services will require additional support from the mobile network to content provisioning (e. g. position information, personal data). 'Mobile Internet/Intranet-Extranet Access': Mobile Internet Access offers

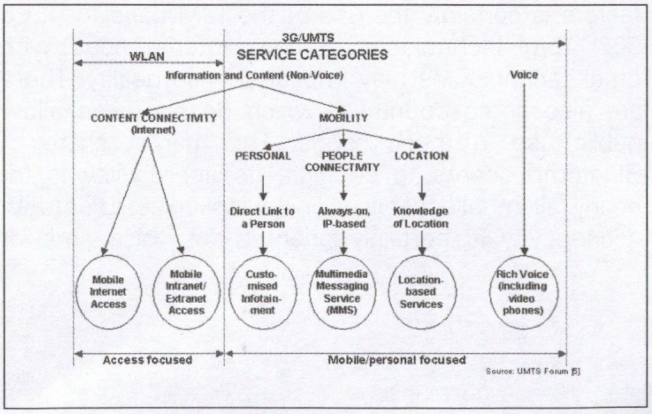


Figure 8. 3G Service Categories and Relevance to WLANs, Source: UMTS Forum [4, 5]

mobile access to fixed ISPs including full Web access to the Internet as well as file transfer, e-mail and streaming audio/video. Mobile Intranet/Extranet Access is a business 3G service that provides secure mobile access to Corporate Networks allowing Virtual Private Networks. In the latter case, the question about providing wireless access either via WLAN or UMTS radio arises. An existing 3G mobile network operator or a new entrant could operate this service category.

Conclusions

The growing number of WLANs in the corporate sector, in public buildings and campus raise the question to public operators, what business plans could speak for providing WLAN coverage in selected hotspot areas of the public domain (Fig. 9.). WLAN could successfully sell on the advantage over UMTS hotspots regarding the bandwidth offered and the service being available now. Company-internal WLAN usage could further encourage the usage of WLAN both in public and company internally. However, it must be understood that decisions on remote mobile access from a public domain to the corporate network is completely different from that of a decision mobile voice usage, only because many IT-systems have to be re-designed for the mobile access to the corporate networks. Questions of data security, compatibility of systems, and the unknown reliability of the service provider have to be considered by the corporation. Area wide coverage with WLAN technology is economically not feasible. The hotspot area will be mainly WLAN coverage – as it is focused on data – can only be seen optionally to 3G service coverage. Handhelds will remain to be served by the 3G networks only. A number of Telecom operators and start-up companies are already providing public WLAN coverage. It makes clear, that for start-ups providing WLAN only services, the profitability of the projects remains a big question mark. For public operators remains the question about frequency spectrum use, in public areas and in corporates. One of the risk factors is certainly the use of the ISM band for IEEE 802.11b/g technologies, where interferences with other technologies may impact service quality. There are also some countries, which do not even allow public use in these bands. The market share of Bluetooth seems to be high, its applicability is for nearly all mobile services including voice. Bluetooth connectivity is certainly oriented for voice and for

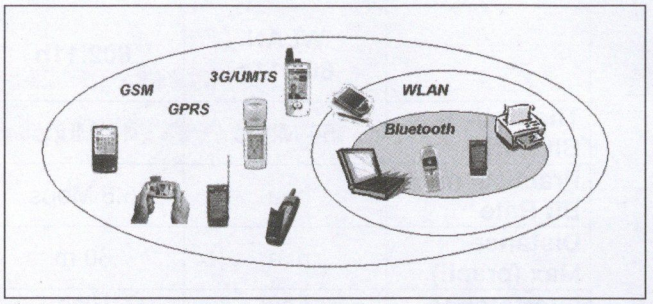


Figure 9. Integration of all Radio Standards into a 3G Network economical?

broad applications and may therefore be flexible to different systems in addition to 3G. Depending on the frequency use and interference, WLANs may come to the market sooner than 3G, but will always be limited to hotspot coverage in buildings. It can be estimated, that a typical 3G cell area of 1 km_ (r ~ 0,6 km) would require 1000 WLAN cells, assumed that the ISM band allows such a multicell approach. Even the wireline backbone provisioning would be a logistic and financial problem. However, the WLAN supply of useful business places, at conventions, for dedicated tasks could be a complement to 3G networks. The emerging Bluetooth could take away some of the WLAN business in the long-term.

Literature

1. Ahlens, E. and P. M. Ziegler 'Luftbrücken – USB-Adapter und Basisstationen für die Funkvernetzung', c't magazine Vol.18, 2001 pp 126-133
2. Book: UMTS and Mobile Computing Josef and Alexander Huber Artech House, April 2002, London, Web: www.artechhouse.com ISBN 1-58053-264-0 90000>
3. UMTS Forum Report No. 5 Minimum Spectrum Demand per Public Terrestrial UMTS Operator in the Initial Phase, 1998 UMTS Forum, Web: www.umts-forum.org
4. UMTS Forum Report No. 9 The UMTS Third Generation Market – Structuring the Service Revenues Opportunities, September 2000 UMTS Forum, Web: www.umts-forum.org
5. UMTS Forum Report No. 13 The UMTS Third Generation Market: Structuring the Service Revenue Opportunities, April 2001 UMTS Forum, Web: www.umts-forum.org
6. UMTS Forum Report No. 14 Support of 3G Services using UMTS in a Converging Network Environment, April 2002 UMTS Forum, Web: www.umts-forum.org

Fluid Simulation in Telecommunication Networks

TAMÁS VARGA[1], PÉTER BENKŐ[2], TAMÁS BŐHM[1], ATTILA ESCHWIGH-HAJTS[1]

[1]*Budapest University of Technology and Economics, Department of Telecommunications and Telematics*
[2]*ERICSSON Hungary, Network Performance and Traffic Analysis Laboratory*

Performance evaluation of packet-switched telecommunication networks faces to more difficult challenges due to the growing size and higher bandwidths in the network. Simulation tools often provide adequate means for this purpose, however, packet-based simulation is infeasible within reasonable run-time. Fluid simulation is deemed in the literature as an alternative technique, but the flattering speed-up potential cannot be always exploited due to the extensive state-space maintenance. An update aggregation algorithm is presented here to enhance the performance of the event-driven fluid simulation. Our findings showed moderate performance enhancement, which is still not adequate for large-scale models.

Introduction

Simulation tools play important role in performance evaluation of telecommunication networks, since it provides a cost-effective alternative to the expensive and time-consuming prototype implementation of the system. The widely used packet-based simulation tools suffer from performance problems in case of large-scale or high-speed networks. The reason is that the simulator spends too much time with state space maintenance caused by the large number of packets in the simulated network.

A number of proposals have been already made to speed-up network simulation, which can be grouped in three major categories: increasing the computational power, applying special simulation techniques and using higher-level abstraction models. Beside the application of faster processors or memories, multi-processor and distributed systems may provide better performance. However, adaptive traffic introduces such dependencies in the system, which makes difficult or even infeasible to partition the network model [6]. Secondly, special simulation techniques are focusing to optimise the state space maintenance [2] or to explore important but rarely occuring events [3]. More abstract simulation models provide simplified and more efficient evaluation, albeit the accuracy of performance measures is getting relaxed and their mapping to the physical world becomes more complex.

Continuous approximation models of the discrete (packet-based) network traffic are known for a long time [1], usually they are referred to fluid models. Analogously to handling natural liquids, they operate on continuous amount of data rather than individual packets. The difference is that telecommunication fluids do not mix when they are multiplexed in network queues, they remain isolated after being served. Evenly spaced packets arriving close to each other may be compacted in fluid-chunks, which can be described with average rate at first-order or along with variance at second-order approximation. We will consider first-order approximation throughout the article.

Fluid simulators are thus dealing with fluid-chunks, which describe the properties of the data flow during a certain period of time. At first look, it seems much more efficient than handling individual packets, since a simple fluid-chunk may substitute many packets. This will hold only for simple fluid networks when the data flows do not consume all the resources. Whenever the flows need to share the available bandwidth due to congestion, the performance of the fluid simulation will degrade, and it may even provide poorer performance than a corresponding packet-based simulator. The reason is intrinsic to the state-space maintenance of fluid simulation. When the bandwidth sharing of fluid flows changes at a buffer during congestion, the flow rates need to be propagated through downstream queues towards the receivers. Thus a single rate change may affect a large number of flows as in an avalanche; which is often referred to the ripple-effect. A formula can be established for calculating the speed-up factor of fluid simulation with respect to the packet sending rate and the number of flows in case of on-off sources [8]. As a conclusion, the authors suggest to aggregate all flows, which are out of scope of the investigation to one background flow to decrease the number of update events. Often this aggregation is infeasible due the disjoint routes in the network where these flows are found.

In this article, we are investigating the possible event reduction by aggregated rate updates rather than by flow aggregation. We will show that

unfortunately in most of the practical cases, event reduction does not improve significantly the performance of the simulator. Therefore, simulation of large-scale networks has still open issues.

Fundamentals of fluid simulation

In the fluid modelling technique, buffers in the network are responsible for sharing bandwidth among competing flows. While in a packet-based FIFO buffer only a single flow is served at a time, but fluid flows are served in parallel at fluid buffers (processor sharing). Continuous time fluid models are using cumulative arrival and departure functions, which cumulates the arrived or departed fluid (or bytes) until time t. In Figure 1, the first diagram shows packet arrivals and the second one shows the corresponding cumulative arrival function $A(t)$. Fluid models are inherited from $A(t)$ through a smoothing process to obtain a continuous cumulative arrival function $A^*(t)$. From simulation point of view, the derivative of the cumulative arrival function, the rate function $r(t)$, is much more useful for system description.

But continuous time models are not directly suitable for simulation, they have to be discretised first. In a simulation environment, the discretised model can be either time-stepped or event-driven. In the event-driven approach, discretised rate function may have discrete steps at any time, when the rate function changes considerably. This approach follows the changes in the network very precisely but it can be time-consuming when state changes are frequent. In this case a time-stepped simulation would be more efficient [4]. In this case the average rate is calculated in fixed size timeslots.

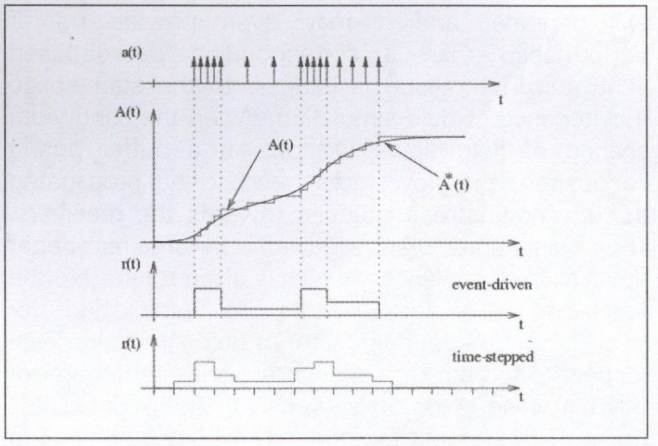


Figure 1. The origin and the description of fluid models

In event-driven fluid simulation, let $r_i(t)$ denote the arrival rate of flow i to the buffer at time t, C the service capacity and $V(t)$ the length of the queue. Then the output flow rate $s_i(t)$ can be expressed as follows:

$$s_i\left(t+V(t)/C\right)=\alpha r_i(t) \qquad \text{if } \sum r_i(t)\geq C$$

where

$$\alpha=\begin{cases}1 & \text{if } \sum r_i(t)\leq C\\ \dfrac{C}{\sum r_i(t)} & \text{otherwise}\end{cases}$$

is the overload factor. The change of the fluid queue length can be determined by:

$$\frac{dV(t)}{dt}=\begin{cases}\sum r_i(t)-C & 0<V(t)<B\\ 0 & V(t)=0 \text{ or } V(t)=B\end{cases}$$

In other words, flows are served with their arrival rate as far as the buffer is not congested. In case of congestion, when the total arrival rate of the flows exceeds the service capacity, the server will share the total service capacity among the flows proportional to their input rate. That is, flows will be scaled down with a factor of α. In this case, the fluid queue length will increase with a rate of $\Sigma r_i(t)-C$ until a maximal buffer length B. If it is exceeded, loss will occur, where the overall loss rate can be calculated as:

$$\frac{dl(t)}{dt}=\begin{cases}0 & \text{if } \sum r_i(t)\leq C \text{ and } V(t)\leq B\\ \sum r_i(t)-C & \text{otherwise}\end{cases}$$

If the queue is not empty, a rate change will only become effective on the output after a $V(t)/C$ amount of time, when all the preceding data in the queue will be served. In this case, different flow sharing patterns will be generated which will arrive to the output one by one, delayed with their corresponding delay.

When the congestion period is over, the queue will empty with rate of $C-\Sigma r_i(t)$ and a different fluid pattern appears if the queue was not empty or the total input rate is non-zero. If $V(t)>0$, the buffer enters into a temporary phase during which flow rates are scaled up to the capacity while the data belonging to the old pattern is sniffed out from the queue. Then the appropriate flow rate pattern becomes effective.

$$s_i\left(t+\frac{V(t)}{C-\sum r_i(t)}\right)=r_i(t) \quad \text{if } \sum r_i(t)<C$$

$$s_i\left(t+\frac{V(t)}{C}\right)=\frac{\sum r_i(t)}{C}r_i(t) \quad \text{if } \sum r_i(t)<C \text{ and } V(t)>0$$

This mechanism is demonstrated with an example in Figure 2. Let the service capacity be 3 units per second, and let us assume that the system was empty at start. At time $t=0$ the first traffic source starts to transmit data with a rate of 1 units per second. At time $t=1$ the second source turns on with rate 2 units/sec. Since the buffer can just serve both flows, no congestion will occur until time $t=2$, when the third source starts to transmit with 3 units/sec. Then the total input rate will become 6 units/sec, therefore 3/6 proportion of all flows will be served while the remaining part will be queued. When the first source

stops transmitting at time *t=3*, the buffer is still congested with an overload factor of 3/5. Since the queue is not empty at that time, a new sharing pattern will be generated, which will become effective at the output after the 3 units of buffered data have been served at time *t=4*. At the same time, the second source turns off and the buffer becomes uncongested. However, this change will only be apparent on the output after the 3 units of data (generated by the first sharing pattern) and the 2 units of data (generated by the second sharing pattern) have been served at time *t=5.67*. At last, also the third source turns off at *t=5*, having its effect apparent at the output at time *t=6.67*.

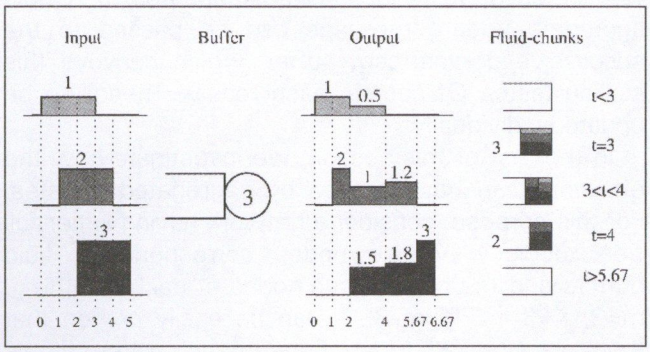


Figure 2. Visualisation of buffer dynamics

In a generic network, which consists of several nodes, the output of a buffer arrives to the input of the successor buffer. Therefore, the transmission rate change of a single traffic source can generate a number of rate changes in the network. When a buffer is in congestion, or it is getting to or leaving the congested state, the rates of all traversing flows have to be updated and propagated through the network (see Figure 3). This is the so-called ripple-effect, which, as we will see later on, sets a fundamental bound on the efficiency of the simulation.

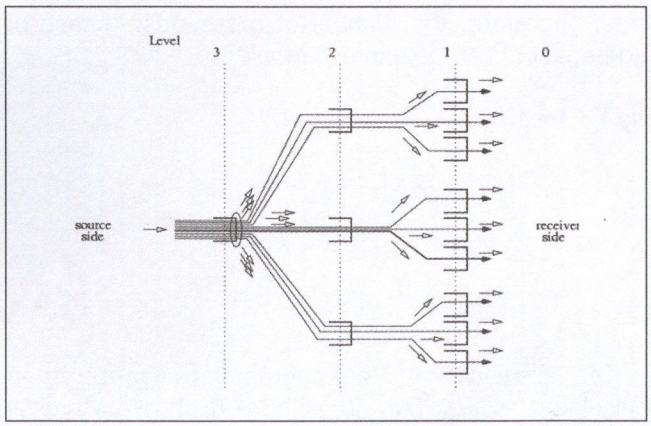


Figure 3. Rate updates in a tree topology network

Precision of the fluid simulation

An important measure of fluid simulation is the ability how it estimates the network performance parameters compared to traditional packet level models. The most important parameters, like end-to-end delay, loss, throughput and link load, have been already investigated in [5,6,7]. For instance, Nicol et al. reported 2% delay and 1% loss error using ON-OFF sources in a single buffer simulation. When calculating error values, one has to consider that an inherent property of the fluid simulation is that it can only cope with long-term congestion conditions accurately. Transient congestion resulting from packet arrival synchronisation is not modelled correctly and does not consider when calculating fluid queue lengths.

Since most performance parameters are closely coupled with buffer queue lengths, the investigation of queue length distributions provides an overall view on the accuracy of the estimation. For this study, we have used a simple Markov modulated ON-OFF source with an ON rate of constant 10 Kbps, emitting 100 byte packets. The state transition probabilities of the Markov process have been calculated from a fixed 0.6 activity factor and from variable average burst sizes (the number of packets a fluid chunk represents). The service capacity was set dynamically according to the target link utilisation value.

The generated traffic flew through a single buffer with a length of B=10000 bytes. The actual queue length was sampled periodically and the mean squared error of the packet and the fluid cumulative distribution function (CDF) of the queue length was calculated. If we denote the CDF resulting from the fluid and from the packet level simulation with $F_f(x)$ and $F_p(x)$ respectively, then the mean squared error on n points is defined as:

$$h(F_f(x),\ F_p(x)) = \frac{1}{n+1}\sum_{i=0}^{n}\left(F_f\left(\frac{i}{n}B\right) - F_p\left(\frac{i}{n}B\right)\right)^2$$

where B is the maximum buffer length. Since CDFs are normalised, the mean squared error will be normalised too. A small mean squared error stands for a small estimation error between the fluid and packet models. In contrast to the mean value, the mean squared error value is more sensitive to differences, however, it will not indicate higher error for larger queue length values.

The simulations were carried out using our FluIPsim platform, running on Sun Ultra-5 workstation. The error of the queue length estimation is plotted in Figure 4. It is worth to note that if the number of flows present in the system is low, then the estimation error is less than 10^{-4}. As the number of flows increases, the error first rises with an order of magnitude, then it falls back again. The reason behind this is twofold: on the one hand, in case of a small number of flows there are only a few rate changes in the buffer, therefore the queue length can be accurately estimated by the fluid model. On the other hand, when the number of flows is very high, the asymptotic error caused by frequent rate changes will be relatively small.

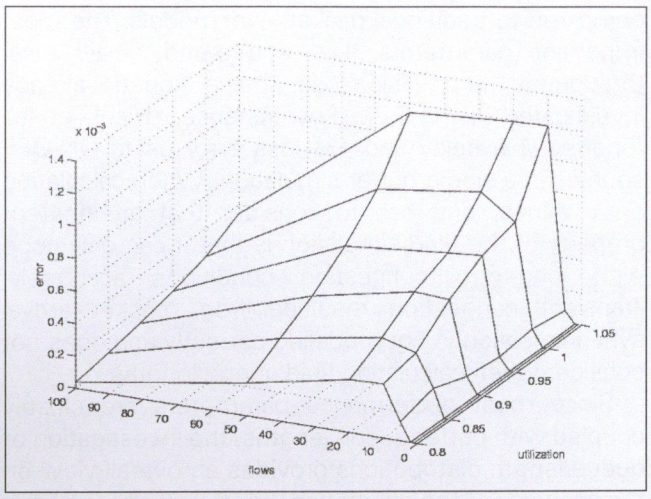


Figure 4. The estimation error of fluid simulation

Scalability of the fluid simulation

The easiest way to measure the efficiency of fluid simulation is to compare its performance to an equivalent packet-level simulation. By deriving the speed-up factor as the fraction of simulation events occurred in fluid and packet simulation, we get a simple scalability measure.

For the investigations, consider the tree topology shown in Figure 3. There are N flows in the network, each one is routed through the neck of the tree and distributed evenly among the downstream paths. Each flow is driven by a Markovian ON/OFF source with activity of 0.6. Link capacities are dimensioned to generate utilisation of 0.95 on the long-term. When congestion occurs at the neck of the tree, the whole tree needs to be updated, thus the appearing ripple-effect will degrade the performance.

Figure 5 depicts the contour plot of the speed-up factor in respect to the burst size and the number of flows. Each line corresponds to those burst size and flow number pairs where the speed-up ratio is the same. We can realise that in one of the two domains

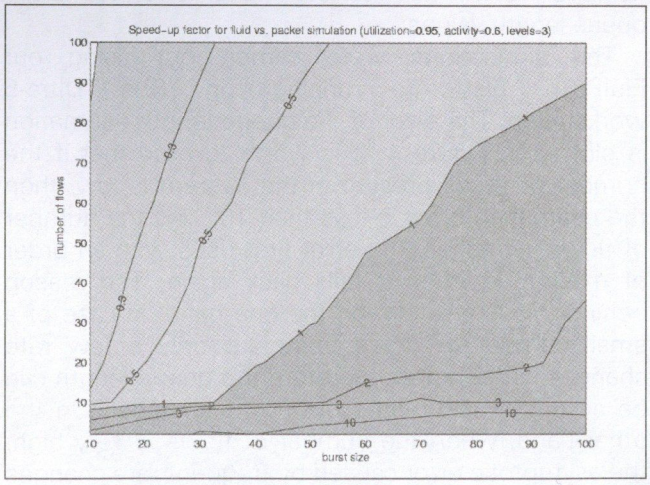


Figure 5. Contour plot of the speed-up factor

(bottom right), fluid simulation performs better, and in the other domain (top left), packet-level simulation is more efficient. It can be also seen that the fraction of the burst size and the number of flows yield the achievable speed-up factor in this scenario. We can conclude that the compression effect of the fluid-chunks cannot compensate the work caused by the network updates in case of large number of flows.

The efficiency of network updates can be improved by introducing aggregated update messages between fluid buffers. Apart from certain special cases, the output flow rates are changing with the same factor when the congested situation establishes, changes or terminates. Instead of individual updates, a simple aggregate update message can be passed to the successive downstream buffer, which conveys this scaling factor. Of course, each receiver needs to be updated individually.

In the rest of this section, we investigate how the efficiency can be improved by aggregated updates. For this purpose, consider a network having m-tier full tree topology, whose nodes correspond to fluid buffers and there are K sub-nodes at each level (e.g. m=3, K=3 in Figure 3). It can be easily proven that there are $l_K(m)=K^{m-1}$ leaf nodes in such a tree. Flows are routed from the neck of the tree towards the leaf nodes so that M flows traverse through each leaf node. Whenever a buffer at level i $(0<i\leq m)$ is congested, it generates

$$n_{K,M}(i) = Mil_K(i) = MiK^{i-1}$$

update messages in the native case, since it has to maintain the flow rates in the sub-tree beneath that level. In case of aggregation, a single message can be used for multiple flows between intermediate nodes and individual messages needs to be sent for the $Ml_K(i)$ receivers. In an i-tier sub-tree, the number of messages passed through level j $(0<j<i)$ is $l_K(i-j)$ since it can be seen as the number of leafs in an i-j-tier sub-tree. Therefore, the number of messages, in case of aggregation, can be expressed as:

$$n^*_{K,M}(i) = Ml_K(i) + \sum_{j=1}^{i-1} l_K(i-j+1) =$$

$$= MK^{i-1} + \sum_{j=1}^{i-1} K^{i-j} =$$

$$= \begin{cases} MK^{i-1} + i - 1 & K = 1 \\ MK^{i-1} + \dfrac{K^i - 1}{K - 1} & K > 1 \end{cases}$$

The alleviation of the aggregation can be then calculated as $n^*_{K,M}(i)/n_{K,M}(i)$. Figure 6 shows a typical alleviation surface (for the K=3). We can see that the number of update messages is inverse proportionally decreasing as the number of tree levels increases and the number of flows influences it only lightly, only in case of few flows. This yields a moderate alleviation in practical cases, e.g. around 33% for a 3-tier access network, as the thick line indicates in Figure 6. This

theoretical maximum can be achieved only if the network is always congested, otherwise, it is less (greater value) and depends on the utilisation as simulation results show, see Figure 7.

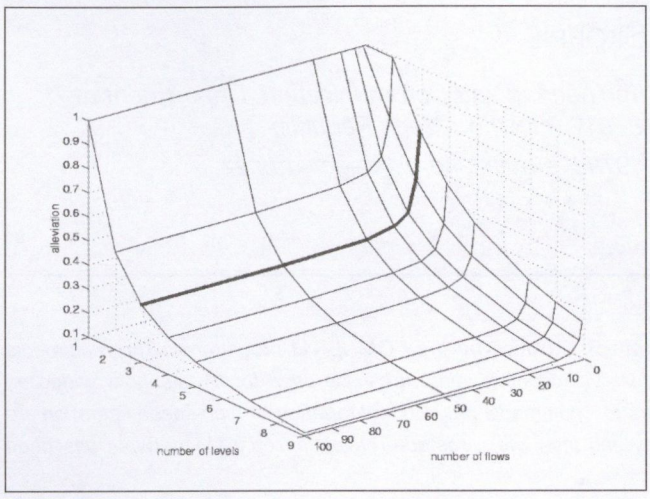


Figure 6. The alleviation of the aggregated updates

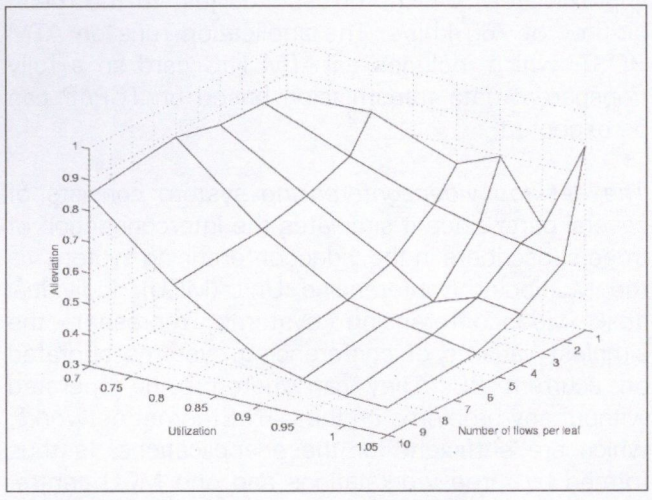


Figure 7. The alleviation versus utilisation

The efficiency of the event-driven fluid simulation cannot be increased more by exact rate update mechanisms. By introducing partial updates, the rate maintenance work may be further reduced at cost of loosing synchronicity in the fluid network. Partial updates can be either realised by time driven or by rate driven manner. In case of time driven aggregation, the update messages are collected first until a predefined period of time and a summary is sent over the network. In the other case, only those flows are updated whose output rate changes more than a predefined threshold ratio. However, these methods may further increase the simulator performance, but the accuracy is getting weaker on the counterpart.

Conclusions

Fluid simulation is deemed in the literature as a powerful technique for performance evaluation of high-speed networks. The more abstract traffic model suggests a more efficient simulation, however, its potentials cannot be always exploited due to state-space maintenance overhead. By applying aggregated update mechanism, the simulation performance can be proportionally increased with the system size. Unfortunately, as the number of flows in the system gets higher, the gain of the aggregation is less effective to the simulation performance. Albeit partial updates may provide further improvement by neglecting small changes, the accuracy of the simulation gets weaker. Time-stepped fluid simulation seems to be a better solution, because it is suitable for distributed computing. The accuracy of this latest method largely depends on the size of the timeslot it uses, therefore its application is reasonable for very high-speeds only.

References

1. L. Kleinrock, Queueing Systems, Volume II: Applications, Wiley & Sons Inc., ISBN 0-471-49111-X, pp. 56-97, 1976
2. R. Röngren, J. Riboe, R. Ayani, A comparative study of some priority queues suitable for implementation of the pending event set, Department of Teleinformatics, Computer Systems Division, Royal Institute of Technology, Sweden, 1993
3. M. Villén-Altamirano, J. Villén-Altamirano, RESTART: A Straightforward Method for Fast Simulation of Rare Events, Proceedings of the 1994 Winter Simulation Conference, pp. 282-289, 1994
4. A. Yan, W. Gong, Fluid Simulation for High Speed Networks, Department of Electrical & Computer Engineering, University of Massachusetts, Amherst, TR-96-CCS-1, 1996
5. D. Nicol, M. Goldsby, M. Johnson, Fluid-based Simulation of Communication Networks using SSF, European Simulation Symposium, Erlangen-Nuremberg, 1999
6. P. Benko, Accerelated Simulation of TCP/IP Traffic Using the Modified Fluid Model, Budapest University of Technology and Economics, M.Sc. Thesis, Budapest, 1999
7. D. Ros, R. Marie, Estimation of end-to-end delay in high-speed networks through simulation of fluid models, SPECTS'99, Chicago, 1999
8. B. Liu, D. R. Figueiredo, Y. Guo, J. Kurose, D. Towsley, A Study of Networks Simulation Efficiency: Fluid Simulation vs. Packet-level Simulation, INFOCOM2001, 2001

Multimedia Network Optimization

VLADISLAV SKORPIL

Brno University of Technology, Faculty of Electrical Engineering and Communication, Department of Telecommunications, Purkynova 118, 612 00 Brno, Czech Republic

tel.: +420 5 41149212, fax: +420 5 41149192, e-mail: skorpil@feec.vutbr.cz

This paper describes the main results received in the ATM network traffic simulation by the COMNET III programme. The multimedia network model contains important applications, namely the hall and desk videoconferencing, It can be used for IP telephony, video on demand, TV and radio distribution and data transmission. The viewpoints for multimedia network optimization of high-speed operation are scrutinized. Transmission speeds network reliability transmission delay and jitter were tested. The optimized ATM network has been designed according to simulation results.

Multimedia network model

At the Department of Telecommunications in Brno, a flexible ATM (Asynchronous Transfer Mode) network is simulated. This ATM network consists of two switches (Cisco) and it is connected to the Brno university high-speed network. A simple ring network is realized in this way. It is in operation, but it is also good for research.

The multimedia network considered in this paper represents a classical network on which some specific demands are made, the two most important tem are:

- providing sufficient band width: some applications require a high transmission speed, which the network must provide and which must be kept within a prescribed interval.
- maintaining the time relations: the packet (cell) delay while passing through the network from the source to the destination has rigorous limits, and, simitor, demands are also made on the deviations of individual transmission times from the mean value.

The proposed model must support selected applications described in the following chapters.

Video hall conferencing simulated by a professional videoconferencing system. The most important parameter that characterizes it is the required 768 kbit/s bit flow. Currently, this value represents the upper bound of standard realizations of videoconferencing systems. The system is designed as a videoconferencing facility for the Ethernet, Token Ring and ATM networks. The model of hall videoconferencing is conceived as a computer application located in the network in an ATM host. The model generates standard ATM cells approximately, which corresponds just to the mean bit-flow of 768 kbit/s. The application runs on ATM HOST, which includes an ATM NIC card so a fully transparent data transmission based on TCP/IP can be expected.

The desktop videoconferencing system consists of several parts since it simulates the interconnection of three subscribers in the videoconferencing system via the Multipoint Conferencing Unit (MCU). Individual desk videoconferencing systems represent the simplest variants of conferencing systems operated on acurrent PC. They can therefore be operated without any problems on the LAN Ethernet networks, which are sufficient for these applications. Is thus formed by three workstations and one MCU centre. Each workstation generates a flow of 64 kbit/s and receives the picture and sound from the other connected workstations (in our case two workstations) at a speed of 128 kbit/s.

E1 circuits are used for audio or for other services requiring real-time operation. The ATM Forum has specified two modes for circuit emulation in ATM networks (CES – Circuit Emulation Service), namely the structured and unstructured modes. The unstructured mode maps the full E1 interface into one virtual ATM channel. The structured mode maps the individual 64 kbit/s E1 circuits into separate virtual channels.

The model of private branch exchange may contain the session instruction, which represents the establishment CBR of type virtual connection. All 32 channel intervals of the E1 frame are then mapped into the a CBR virtual connection. The allocated transmission band remains reserved even if no information transmission is necessary.

IP telephony is modelled by two terminal, i.e. two computers. In the model the first computers signalling to the other the preporing the call through a this instruction. With the instruction executed, the connection is set up via the session instruction and, likewise, the other computer sets up the connection using the session instruction. The above procedure represents a simplification. Actually, the ealling terminal is signalling a reseruation to realize a call and simultaneously it contracts the network management for the required transmission capacity. It will get a message about call admission. The terminal device is made up of a receiver and a transmitter in order that the requested data flow can be realized. The C&C node cannot simultaneously receive and send packets if the flow is formed with the aid of packeting delay. The packeting delay must be introduced in order to realize the calling partyís ringing. The calling C&C node generates a bit flow of 7.6 kbit/s so in every 97 ms it creates a VoIP protocol packet. The receiving node will process this bit flow and provide an adequate response. The private branch exchange for IP telephony is in fact formed by a computer and it is again modelled by a C&C node. The model considers the connection of 100 calls and thus represents a bit flow of 760 kbit/s. The computer, acting as a private branch exchange for the packet transmission of audio signals, operates simultaneously as a LAN switch. It thus ensures a transparent transmission of VoIP via the ATM network and, at the same time, it modifies the non-uniform flow of VoIP coming from the LAN such that it will be as constant as possible for the sake of an economical transmission in the ATM part of the network. In this way it is then possible to simulate a transparent transmission of VoIP packets over ATM. The data flow in the model focuses on the transmission of calls and is formed by repeated execution of the transport instruction.

The Video on demand application consists of two parts. The first is a film server, which contains the titles of requested programmes by the multimedia workstations which form the other part of the model. The transmission of non-distorted MPEG-1 signal requires a band of ca 1.5 Mbit/s.

The model of film server is formed as an ATM host realized by a C&C node with a direct connection of 155 Mbit/s to the ATM switch. The applications that represent individual films are formed by the session instructions, which realize bit flows of 1.5 Mbit/s.

Similarly by to the computers, multimedia workstations are modelled also by the C&C node and they differ only by their location in the multimedia network. A workstation that requires to follow title A is connected to the classical Ethernet segment, where the capability of this segment to transmit multimedia data is strongly affected by the behaviour of the other nodes that share its transfer capacity. The model thus represents the simplest connection of the workstation that can, realize the transmission of compressed video of the MPEG-I format, ifcertain conditions are satisfied.

The workstation that requires to follow title B is connected in the network directly to an ATM switch with 2 Mbit/s. In this way the Primary Rate Access (PRA) line of ISDN is simulated, whose transmission capacity is quite sufficient for the transmission of compressed video data. Another model of the connection to the multimedia network represents the ADSL technology. Via the ADSL modem, workstation C is connected to a classical subscriber line, through which access to the ATM network is realized. In the model, the ADSL modem is simulated with basic 2 Mbit/s for down-stream and 16 kbit/s for up-stream.

TV and radio distribution can be discussed separetly The bandwidth required for one TV programme ranges between 25 and 34 Mbit/s. HDTV television requires as much as 100 Mbit/s. The model of distributed TV signal only simulates the transmission of data flow between the ATM host nodes, which represent decentralized TV workplaces.

The distribution of radio broadcast can be modelled by a bit flow of 1 Mbit/s, which represents, with some reserve, non-compressed transmission of two radio channels Q (50–15 000 Hz) sampled by a frequency of 32 kHz and encoded by 12 bits. Radio distribution can also be modelled by a flow of no more than 128 kbit/s, which give the maximum limit for the reception via the Basic Rate Access (BRA) line of ISDN. Reducing the bit flow to 128 bit/s or lower (depending on the quality required) is obtained by compression according to the ISO/MPEG Layer-3 standard, which can on powerful workstations in distribution centres be done in real time.

Receivers of distributed radio broadcast are modelled like C&C nodes with access to the network either via the ADSL modem (for the reception of non-compressed radio broadcast) or via the BRA line of ISDN 2x64 kbit/s (for the compressed version).

Data transmissions applications are quite general in classical networks based on the existing network protocols. For their cooperation with the high-speed part of a multimedia network built on ATM architecture the emulation of the LAN networks over ATM – LANE – is used. In this way the mutual cooperation of data sources with Ethernet interface over the ATM backbone network can be modelled.

It is necessary to distinguish between two groups of computers connected in the network working on standard protocols of LAN networks:

The first group is formed by the existing Ethernet LANs, which are connected to the ATM through a LAN switch, which is provided with an ATM NIC card on the ATM side.

The other group is represented by computers connected to the ports of ATM/Ethernet asymmetric

switches, in which it is possible to assign any port to a selected virtual network.

LANE then ensures that the environment of ATM backbone network is fully transparent for Ethernet users. In this mode, ATM emulates the MAC layer of LAN networks such that the applications can cooperate mutually while using the current LAN protocols.

Optimization of high-speed operation Optimizing the high-speed operation in networks based on ATM technology means looking for the optimum properties of the network. With optimum properties, the network can provide the required high-speed services. If multimedia applications additionally require real-time data transfer, further optimization criteria are necessary that define network behaviour and time relationships. Other (this time) general optimization criteria include network reliability, the capability of network management to respond dynamically to network changes (jamming of some nodes, fall-out of nodes or transmission lines, Ö), physical realization of network elements (possibility of extending the network, increasing the transmission capacity, introducing support for newly defined recommendations and standards, etc.). Selected optimization criteria are proposed below.

Transmission speed viewpoint can be optimized in an already realized network only on a limited scale. An example can be seen in the optimization via creating permanent virtual channels where it has some meaning. On shared Ethernet segments optimization is only possible by reducing the number of connected workstations. This always implies rearranging the available transmission bandwidth among its users. It is much more important to optimize the network in the course of design. It is necessary to have the proper knowledge of the requirements of the connected workstations and to consider the future requirements and the increasing number of connected workstations. After that the elements are proposed that can realize the desioned requirements. Engineering a network model, there were two points. A precondition was the existence of LAN networks, which were connected to the newly designed multimedia network. In the initial segments of LAN networks the number of applications had to be reduced in order that data of broadband applications should be transmitted. Another possibility was to create from the existing structured cables more LAN networks interconnected only on the port of the backbone network. By adding another LAN router the available bandwidth can be increased for individual applications. We are concerned here with real bandwidth since Ethernet represents a shared medium of a maximum theoretical throughput of 10 Mbit/s and by reducing the number of sharing workstations the real available bandwidth is increased for individual applications. In the backbone part 155 Mbit/s transmission speed is used in the model. This lower value (maximum throughput values are 622 Mbit/s and 2.5 Gbit/s) is quite sufficient for the proposed applications.

Network reliability Network reliability and its ability to operate successfully even in case of the outage of any section must be considered in the coure of engineering the network topology. The choice of architecture and network management can only offer a routing to make use of the possibilities offered by suitably located and interconnected network elements. Duplicated physical transmission routes must be establesked, which can be used in the case that when cannot be delivered by the original routes. The need for these back-up paths grows with the network hierarchy level on which the given problem is being solved. No back-up is provided in the model for the case of drop-out on the lowest levels, i.e. in the subscriber access parts. In the backbone part, the provision of back-up routes is of much importance.

The ATM backbone network is designed such that it can provide its services to all applications in the model simultaneously and still has a sufficient reserve. For this reason there is no sense in modelling a flooded backbone part since the designed applications cannot achieve such a state. The ATM backbone is thus optimized emphasising the transmission speeds.

Data transmission delay The operation of multimedia applications deprived of both the transmissions of files and internal operation in the LAN segments was modelled. Results of this modelling represent the reference values of delay times necessary for the transmissions of packets of individual network applications. These values must be as lowas possible and in the case of actual network operation (the network is also used to transmit files) they will be in most cases higher. The maximum times of packet delays in the network are in the case of time-sensitive applications below 1 ms which is excellent comparing with the roung-trip delay, which is limited by ITU in 400 ms. This condition is satisfactory. With the TV distribution application the maximum time of packet transmission is as much as 70 ms. This value is markedly lower than the transmission times of packets generated in a lower bit flow but it is still sufficient for single-direction distribution of multimedia data.

The next simulation step is the determining the effect of internal data operation in individual LAN parts on applications that shared the same medium. Along with this simulation, data applications for file transmission over the ATM network were also run. All this in an extent that the network had been designed for. The resulting delay values are such as to ensure a reliable operation of all applications.

The simulation proved the network stability. There were no extreme increases in the delay. The packet transmission time of some applications was close to

100 ms. Video-on-Demand applications were only tens of ms. Videoconferencing and other multimedia applications could not be affected by this array. The result of this simulation leads to the statement that as long as data segments are loaded only with such flows that will not cause absolute network flooding (accompanied even by loss of packets if buffers are overfilled), the simulated network can also provide multimedia services. The worst values of packet delay still guarantee reliable working of applications sensitive to real time.

In the next simulation step a 80 Mbit/s bit stream was integrated into the backbone network, which shifted the utilization of high-speed lines to the optimum limit that ensures reliable support and guaranty of offered services with sufficient reserve. This did not affect the transmission parameters of the other bit flows.

Delay jitter is affecting the quality of multimedia applications. Multimedia applications generate a continuous flow of bits, which passing through the packet network changes into a surge of bursts. This is due to the multiplexing of packets in switches or routers from various input connections to one output connection and also to packet queuing. The effect of jitter can be by storage for, which, increasing redvced delay time of the packets is the badedraw. The ATM part of the network can hold the jitter within defined limits by means of QoS and the limits.

Information on packet jitter can be calculated as the standard deviation of the delay of individual messages, where the mean and the maximum delay values are also available. In evaluating the transmission of individual packets only the mean and maximum packet delay times in the network are given. Even so, comparing these two values gives an idea of the time irregularity of package arrivals. The value obtained represents approximately the time for which it is necessary to store the packets in order to obtain a continuous bit flow on the output of the network.

Conclusion

The multimedia network model was prepared to test as many applications as possible. The number of applications realized in the model was, a compromise that includes:

- 3 routers
- 8 ATM switches
- 5 LAN switches
- 5 Ethernet-based LANs
- 4 PBXs
- 2 IP telephones
- 5 servers
- numerous workstations with application sources

The basic part of optimization was performed while preparing the individual models of applications, which were first tested and checked separately. Subsequently their simultaneous working was simulated. At this stage, maximum emphasis was laid on the model behaving in the COMNET simulation program environment as faithfully as possible. The effort was to approach the actual processes and sequences of events in multimedia networks. In conclusion, four complete simulations were conducted, each to test and check the overall model from one point of wiev. The network behaviour was tested during a sudden arrival of data in a certain part of the multimedia network. The network response was evaluated as the change in transmission parameters of individual bit flows of the running applications. The results of the simulations can be summed up and on their basis it can be said that the simulated network represented the model prepared in the COMNET environment, it has been optimized for the operation of multimedia applications. A response to data arrays representing approximately the maximum utilization of the lines increased the time of paekets passing through the network, inclusive of a slight increase in packet jitter. The results were within limits that guarantee a safe provision of multimedia services.

Acknowledgement
This research was supported by the grants:
No CZ 400011(CEZ 262200011) Research of communication systems and technologies (Research design)
No LP 0088 Internet Journal Elektrorevue (grant of the Czech Ministry of Education, Youth and Sports)

References

1. PERROS, H.G.: An introduction to ATM networks. Wiley, New York 2002
2. MIKALSEN, A.: Local Area Network Management, Design and Security. Wiley, New York 2001
3. Brazda, R.: Optimization of high-speed operation in an experimental ATM network (in Czech). Final-year Project BUT Brno, 1999
4. COMNET III, User's Manual, Release 1.1 Beta. CACI Product Company, LA Jolla California 1994
5. Jordan,R.-Shawwa,L.: ATM Technology and Applications. http://tularosa.eece.unm.edu/faculty/rjordan/595-025/sudheer/atm3.htm

Accounting and Pricing in DiffServ Networks

LADÁNYI ZSUZSANNA, SZÁSZ ANDRÁS

Budapest University of Technology and Economics
Budapest, Hungary

The main goal of this paper is to demonstrate the necessity of accounting and pricing and collecting their main functions. It surveys several pricing methods, which can be important in affecting people to use the Internet according to their needs which could help planning the expected traffic of the network. Finally, the architecture of a realized system is presented, together with the implemented pricing method.

Overview

The developments and spreading of newer applications made the Internet a widely used medium. Hence, the volume of the forwarded traffic through the Internet is increasing, which means that the problem of sharing resources fairly and efficiently is getting more and more difficult. Not only e-mails can be sent and files can be copied from distant computers, but Internet Telephony services can be reached and video can be watched through packet switched networks, too. There are newer ways of using the Internet, traffic types are differentiated which require different forwarding methods. However, the transmission based on the Internet Protocol (IP) only can apply the best-effort algorithm currently, which is not enough for forwarding high quality voice and video.

Therefore various Quality of Service (QoS) parameters have to be provided, different guaranties are required at different types of traffic. Two architectures have been developed to solve this problem. One of these is the Integrated Services (IntServ) [3] architecture where the routers handle the flows separately. This solution becomes difficult at larger networks because the network components have to manage too many flows, which raises scalability problems. The second solution is the Differentiated Services (DiffServ) [2] architecture where the flows are classified into service classes by their quality requirements (i.e., maximum packet loss, delay, jitter etc.). In this case, the traffic within the DiffServ domain is handled on per flow basis. This is a more scalable way of forwarding traffic hence it is more promising in the near future. We are focusing on DiffServ networks in this paper.

In this system, there are Internet Service Providers (ISP) and users. The user contracts with the ISP to be able to connect to the Internet. A common way of accessing the Internet is over regular telephone networks. The standard form of pricing is a monthly subscription charge to the ISP together with any telephony charges associated with the dial-up connection, which is normally charged at the local call rate. In the United States, the local calls are normally free, so users stay connected for long periods of a day. This causes enormous overload of the Internet. Furthermore at present if a user managed to access the Internet, he can generate as much traffic as he would and there is no limitation. This traffic is transferred with the best effort algorithm, so there is no way of serving quality. These problems motivate the ISPs to emphesis pricing.

If we look at the Internet, it is divided into hundreds of subnetworks, which are called primary domains and several hosts belong to each domain. These domains can be divided into further domains and so on. The idea of this paper is that domains act as network providers, and they demand bandwidth from one another for their aggregated traffic. In this manner the customer is a domain too. The customer domain contracts with the provider domain. In this contract the customer determines when, from which IP address to which domain, how much bandwidth and what kind of quality guarantee is required (determining one of the offered service classes). These contracts can be the bases of a traffic prediction, and the main goal of this prediction is that the network can be configured in advance by these. The customer can generate more amount of traffic than it is mentioned in the traffic prediction, but then some penalty have to be paid, which can force the customer to comply with the traffic prediction. The accurate traffic prediction is very important, so a proper pricing method is required, which can influence the customers.

Pricing cannot exist without accounting. The main function of accounting management is to collect information about the network usage. This is the basis of determining QoS parameters and providing the pricing with detailed information about the generated traffic.

The implemented system performs accounting and pricing as well. The accounting module collects data from the routers and stores them in a database. The pricing module compares these data with the traffic predictions and determines the fee to be charged. The system has a web interface so the customers are able to track the changes of the account.

Accounting management

The Internet Engineering Task Force (IETF) is the standardization body that coordinates research and development for the Internet. They have published the RFC 2975 [1], "Introduction to Accounting Management". This is concerned with the collection of resource consumption data for the purpose of capacity and trend analysis. This means that arising demands on bandwidth can be determined based on the measured data and conclusion can be drawn to determine the changes in the future. Furthermore, it is a purpose to make the possibility of auditing and charging.

The accounting management architecture involves interaction between network devices, accounting servers, and pricing servers. The network devices collect resource consumption data and make a session record. This record is a summary of the resource consumption of a user over the entire session and it is transferred to the accounting server via a standard accounting protocol or by an own consent. This server processes the received data and stores or forwards them. Intra-domain and inter-domain accounting events can be distinguished and they have to be transferred to the appropriate direction. At an inter-domain accounting event (the session record contains the Network Access Identifier (NAI), which determines the routing information) the packet is sent to the accounting server of another administrative domain. In the case of intra-domain accounting the information goes directly to the local charging server.

For charging it is very important that the loss of accounting information should be avoided because at usage-based pricing it is essential. To prevent this possibility, the usage of interim accounting is recommended, which provides checkpoint information and in case of any trouble (e.g., device reboot) all of the accounting data, session records can be restored.

Quality of Service

Nowadays several telecommunication services exist, which have different characteristics, and have to meet different requirements. The different types of traffic are divided into service classes. Regarding to [7], there are three measures of quality of service:

Transparency refers to the time and semantic integrity of transferred data. For real-time traffic, delay is not permitted while certain degree of packet loss is tolerable. On the contrary, for data transfer delay is not a critical problem, but semantic integrity is generally required.

Accessibility makes the possibility of refusing admission (data packets can not enter the network) and delay can be set up in case of blocking. Currently after authentication there is no admission control on the Internet. If a new request arrives it is accepted and there is no way to refuse it, so it reduces the amount of bandwidth allocated to the ongoing transfers.

Realized throughput is the most important measurable value of the Quality of Service, e.g., at data transfer it shows the rate of downloading. Its unit of measurement is kbit/s, for example.

These properties can determine the service classes and based on them the customer can choose the most suitable class for the traffic.

Differentiated Services

The Differentiated Services (DiffServ) [2] architecture determines the ability of end-to-end quality management. The DiffServ offers various service classes, which have different parameters.

The advantage of DiffServ is that only the border routers do the shaping, dropping, classifying, marking and traffic measurement functions, while the nodes inside of a domain (core routers) only forward the packets reducing the administration, providing a scalable way for data transmission.

Charging methods

Today, pricing of the Internet is based on flat rate tariffs. The Quality of Service guarantees can change this situation since the "guarantees" will represent the new value, this is what the customers pay for to the network operators. Hence, choosing a proper pricing method becomes important. In this section, we briefly introduce some important pricing methods[5].

The flat rate pricing means a fixed monthly charge, which is independent of the volume of the traffic, that the customer produce. The advantage of this method is its simplicity, but it is unfair because a "light" user has to pay as much as a "heavy" user.

The next method is based on the principle that a customer has to pay money amount proportional to the network usage. This method has two more types, time charge and volume charge. The time-charge pricing method counts the time, which was spent by customer connected to the service provider and the value of the bill is proportional to this. The volume-charge method measures the sent or/and the received data. Different tariffs can be determined considering the parts of the day, the peak period.

The Smart market pricing is based on an economic model. The users add value attributes to their packets, which specify the user's willingness to pay for the transportation of their packet. Then the packets "compete". Those packets are forwarded first, which lie within the capacity threshold and have the higher pay attribute, others are dropped. Each user is charged with the value of the highest valued packet dropped when their packet is carried by the resource, which can deter users to attach higher and higher willingness to pay values. The economic language says that it is an auction.

The Paris Metro pricing has only theoretical importance. The network has to be divided into two subnetworks with equal capacities, but two different prices are charged (first and second class). The principle is that users have to pay more for better quality. The goal of this method is that probably less customers pay the tariff of the first class so less customers use that part of the network and they can get higher bandwidth, better quality.

The finally introduced method is based on the effective bandwidth [6], which is accounted as a powerful method so this is the most interesting method for us. It is important to use a pricing method, which is usage-based and it is easy to understand by the customer [4]. Figure 1 shows how the expense of the service provider changes as the bandwidth is rising.

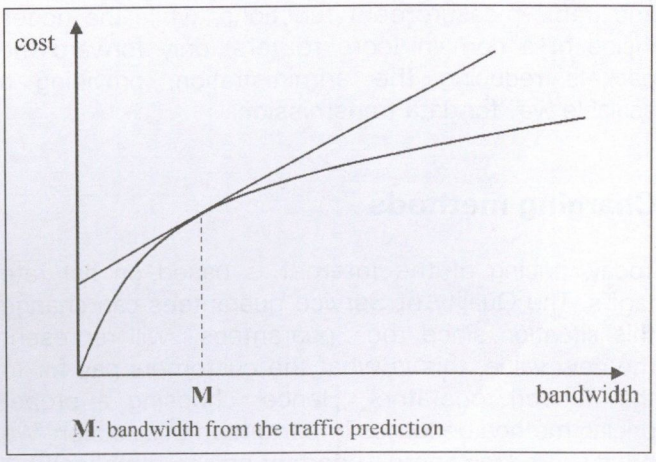


Figure 1. Pricing based on the effective bandwidth

The effective bandwidth is a scalar which summarizes the amount of resources required by a connection in order to preserve its own QoS requirements, and the requirements of the other connections it is multiplexed with. In the figure, the effective bandwidth is not a linear graph, while the graph of the charged money is linear, which results in a simple charging function. The customer makes a contract with the service provider, and in the prediction the M parameter is determined, which shows that how much bandwidth is required during a definite period. The service provider draws a tangent in the M point to the graph of the effective bandwidth, and if the customer deviates from the prediction, some penalty have to be paid which follows the tangent. As the figure shows the difference between the tangent and the effective bandwidth is growing, so this can be an incentive force for the customer to comply with the prediction. Some extra money also have to be paid, if the customer deviates to negative direction, because the service provider has to be compensated for the reserved but unused bandwidth.

The realized system

The system was developed within the cooperation of the Swedish Telia Research AB and the High Speed Networks Laboratory at the Budapest University of Technology and Economics. The activity started as a part of the TraFIPAX project, which sets the main goal on providing End-to-End Quality of Service.

Considering the business aspects, the network providers' own interest is to keep the utilization of their networks as high as possible – as long as it does not cause congestion. To realize this (together with providing the needed higher forwarding qualities) they need information in advance about the expected bandwidth needs. This need of advance information can be fulfilled using the previously introduced traffic predictions.

These traffic predictions are kept in a database. Each record from this database can be associated with a given route in the network. Being more specific, it describes a relationship between an interface of a border router (this acts as the entry point for a packet into to network) and a destination network (which is the exit point). The records contain the following fields:

- the identifier of the chosen quality (the DiffServ CodePoint – DSCP),
- the time of validity (the begin and the end),
- the required bandwidth,
- the identifier of the entry interface of the border router (IP address and network mask), which is used to identify the sender,
- the identifier of the target network (IP number and network mask), which can be used for determining the exit point of the flow.

Using this information, the network provider can calculate the expected traffic amount in advance. The resources for a specific sender in a given quality towards a specific target can be allocated in this way. There is a need for some tools, which can be used to check the traffic generated by the users, and also can be used to make the users comply with their traffic predictions.

The realized system can be used to solve this problem. It is capable of measuring the traffic in the network, detecting the traffic which does not conform to the prediction, possibly taking actions against them and charging the user for the used resources.

This system consists of two main parts: an accounting and a charging part, which are tightly interconnected. The accounting part collects and stores the network usage data, the charging part calculates and displays the bill.

The accounting system

The task of this system is to measure the traffic in the network differentiated and sorted by service classes and connecting domains acting as customers, and to serve the charging system with appropriate information. We consider the neighboring networks (or subnetworks) as customers, who send and receive data from third party networks through the monitored DiffServ network. In this manner, only the data that passes through the network has to be monitored.

To measure the traffic that passes through the network, it is enough to monitor the amount of traffic that enters the network. It can be done in this way, because we can suppose that our network acts as a transit network – at any rate the major amount of traffic passes through it. This is typically true in the case of backbones, which interconnect more networks. In this way, traffic is only measured on the entering side of the border routers.

In the realized system each border router monitors the entering traffic, and temporarily stores the results. After a certain amount of collected data, it sends it to a central server, which is capable of storing the measured data for a longer time. This central server is also in charge of supplying the charging system with the stored data.

The accounting system consists of three main modules (Figure 2):

- accounting client (on the border routers),
- accounting server (on a central server),
- accounting database manager (together with the server).

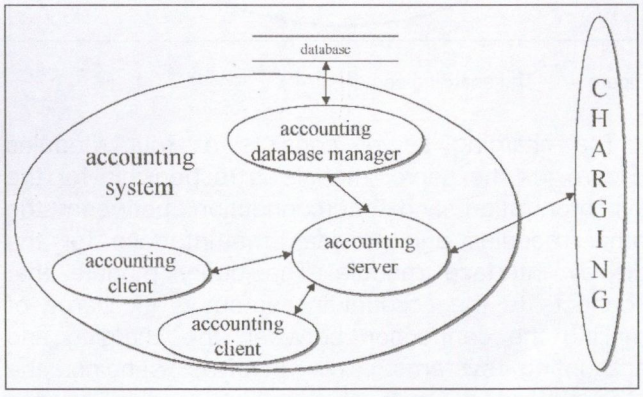


Figure 2. The accounting system

The accounting client, The task of this module is to measure the amount of traffic that enters the network, and to forward the collected data to a central server. The module is able to identify a packet which passes through the network, and another which has a destination inside the network, because we are interested only in the passing packets.

The client module can take some configuring commands even when the system is already running. This makes the system flexible, since it can be controlled during runtime even from the server. Another useful property is that it is able to provide live data to a display client, thus it can be used for debugging purposes. It is an important point how it can cut down the extra traffic caused by the accounting using buffering and a well-chosen protocol between the accounting client and the server.

The accounting client can work in two operating modes considering the way in which it sends the collected data to the server. In the event-driven mode, it sends the data if an event occurs (e.g., an alarm event from an expired timer). In the polling mode, data is sent if the server specifically asks for it.

There are two main options besides the one for the operating mode, which also have a high impact on the client. Modifying the time for a summation, an optimum working mode can be tuned in between an accurate charging and a lower network overhead. The number of the summations is used for determining the maximum amount of data stored temporarily before sending them to the server.

The accounting server. The main task of the server module is to operate as a connector between the different modules. It receives and stores the accounting data from the accounting client and serves the charging system with the requested data.

The accounting clients are connected to the server. The clients send the collected data to the server from time to time, which receives them simultaneously. If the clients are in polling mode, a request has to be sent out first.

A separate interface is used to connect the database module to the server. Since the database manager module is a part of the server the communication between them consists of object and function calls. The parallel multiprocessing is handled on the database management part.

The communication between the accounting server and the charging system (actually the charging server) is settled over TCP/IP. It forwards the received query to the database management module and returns the answer (bandwidth usage data) to the charging server.

In addition, it provides an interactive command line based user interface. Administration tasks can be carried out with the help of this interface, logging can be turned on, working parameters can be sent to the clients, and the server can be stopped.

The accounting database manager. The role of this module is to store the data received from the server and later to return the data that conforms to the request. It uses the miniSQL database manager system [8] to store the data in. Since this module uses standard SQL commands, the database system running under it is exchangeable.

The modules, which want to store data in the database or want to retrieve information from there, have to turn to the server, because the database module is integrated in it. The communication between the database manager module and the server is realized by using FIFO buffers. Each thread of the server has a database module, so locks have been used to handle the parallel access to the database.

The stored data in the database correspond to the general description (described in Chapter 3). There are only two differences: the identifier of the measuring router is also stored, and instead of the required bandwidth, the sum of the transmitted data is stored.

The charging system

The role of the charging system is to generate the users' bill using the collected usage information from the accounting system. This can be done upon a request from the operator or scheduled periodically. Then it has to query for the data from the accounting system to make the calculations, and it has to display the results to the operator.

It is necessary to use the information from the traffic prediction to generate the correct bill. The problem is that not everybody complies with the amount given in the traffic prediction. If less bandwidth is used as expected the difference could be the network provider's loss. In the case of overuse, there could be problems inside the network such as congestion, higher packet-loss ratio, and the assumed forwarding qualities can not be fulfilled even for other traffic flows.

The realized system offers the solution of the introduced problem. Using it the users can be forced to comply with their traffic predictions, in other words to give predictions that are more exact. It uses a similar pricing method to the one based on the effective bandwidth as described previously. If the users do not comply with their prediction penalty tariff is charged. Thus, they have to pay more, so they are persuaded to try to give exact information.

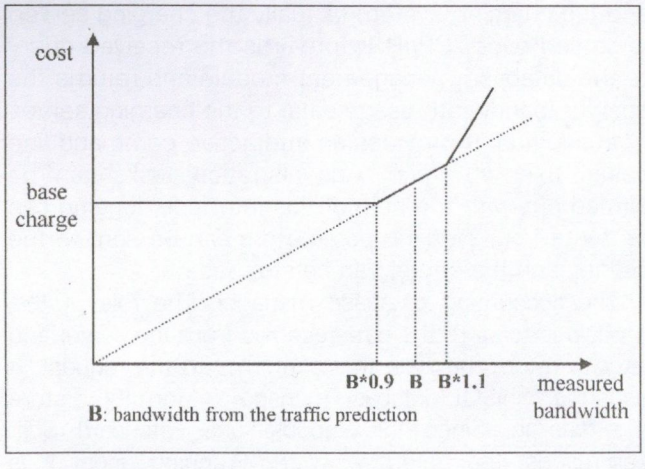


Figure 3. The used charging function

The function used for the charging can be seen on Figure 3. The effectively used bandwidth is compared to the predicted one and if it is within a specified limit (for example ±10% in the figure), the regular usage based prices are used. If it goes below the lower limit, a default base amount is charged. Finally, if it is above the upper limit the price is calculated using a penalty tariff that is higher than the regular rates.

Using the charging method above fair charging can be achieved: the users are encouraged to give precise predictions for their own interest, which is also favorable for the network provider.

The charging system consists of two main parts, the charging server and the display interface.

The charging server. The role of the charging server is to wait for a request from the operator, to generate a correct bill and to send it back. The server gets the traffic predictions from a database, and uses them to calculate the bills. The bandwidth usage information is stored within the accounting system, so the charging server has to query the accounting server for it. The charging function described in the last chapter is changeable thus a new charging method easily can be implemented. The server is multithreaded so more clients can connect at the same time, thus allowing multiple queries to be run simultaneously. The server is connected over TCP/IP to the display interface and to the accounting system.

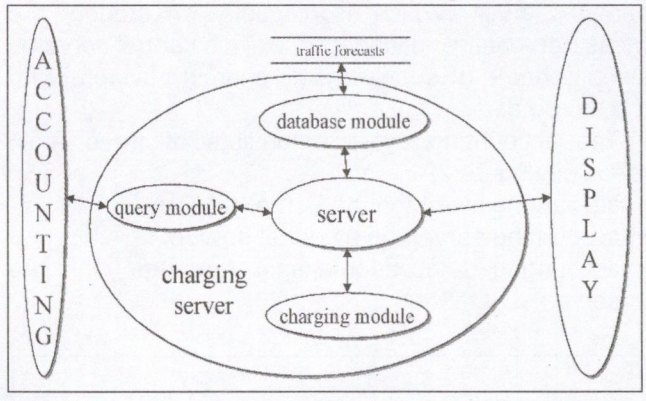


Figure 4. The charging server

The charging server consists of four modules (Figure 4). The server module is responsible for the synchronization and interconnection between the other modules and provides the interface for the display interface module. The query module that connects to the accounting system is in charge of making the connection between the charging and accounting systems. This involves sending the received query from the server module to the accounting system and returning the received response back to the server module. The database module is responsible for checking the existence of a traffic prediction and retrieving the expected bandwidth from it. Finally the task of the charging module is to calculate the payable amount from the

expected and the used bandwidths using the appropriate charging function.

The display interface. This module serves as a user interface to the whole accounting and charging system. It is realized as a CGI [9] module. Thus the query can be filled out in a form on a web page and the results are returned on a new web page. The results can be formed as a table or additionally as a diagram. The supplied information in the query is subjected to an exhaustive check before submitting it to the server, thus avoiding senseless queries. The occurring errors are presented to the operator among with the other errors in other modules of the system.

Summary

The basic idea of the introduced charging systems – which allow the traffic to be predicted in advance – is to make the users be interested in making an accurate traffic prediction. Using these predictions, a network operator can count on the expected traffic amount and the network can be set up in advance for this, so the utilization can be kept on a suitable level.

The charging method in the realized system is based on the same principles as the one based on the effective bandwidth. However, while the latter defines the penalty price as the difference between a logarithmic and a linear function, our method always uses the difference of two linear functions. This can be calculated easier, and in the case of a bigger difference from the predicted amount, it inspires better the user to make a more accurate prediction. To serve the charging system with up-to-date information, the implemented accounting system realizes both of the operational modes (polling and event-driven). It uses a special accounting protocol for transferring data, which is also capable of transporting commands. The accounting data is guarded against unpredictable deletion by storing it in a permanent database on the server side.

The users have an important role: they have to guess the required bandwidth in advance. This can be done by bringing the process of charging closer to the user, using methods which can be understood easily, and allowing the users to track their real bandwidth usage. The web-based display interface can be used to fulfill this latter need.

In these days flat rate charging is used typically which does not give the opportunity to predict the traffic of a network. With the further evolution of the Internet, the advance of the charging methods is expected. Traffic prediction based charging methods can help to solve the problems introduced by the previously used charging methods. They can complement the quality guarantees provided on a DiffServ network.

References

1. B. Aboba, J. Arkko, D. Harrington, Introduction to Accounting Management, IETF RFC 2975, October 2000
2. S. Blake et al., An Architecture for Differentiated Services, IETF RFC 2475, December 1998
3. R. Braden, D. Clark, S. Shenker, Integrated Services in the Internet Architecture: an Overview, IETF RFC 1633, June 1994
4. C. Courcoubetis, Vasilios A. Siris, An Evaluation of Pricing Schemes that are based on Effective Usage, University of Crete, February 1998
5. R. Gibbens, Control and pricing for communication networks, Statistical Laboratory, University of Cambridge, 1999
6. F. P. Kelly, Notes on effective bandwidths. In F. P. Kelly, S. Zachary, and I. Zeidins, editors, Stochastic Networks: Theory and Applications, pages 141–168. Oxford University Press, 1996.
7. J. W. Roberts, Engineering for Quality of Service, France Télécom – CNET, July 1998
8. MiniSQL: www.hughes.com.au/library/msql/
9. CGI: Common Gateway Interface: http://www.w3.org/CGI/

Automatic Wizard Generation

DÁNIEL SZEGŐ

Budapest University of Technology and Economics, Department of Measurement and Information Systems
Budapest, Hungary

Introduction

As software systems are getting more complex, the difficulties of an average user increase. Nowadays an average application offers hundreds or even thousands of different functionalities and services. To cope with this complexity requires significant experience and steady learning. For example WinZip, one of the simplest software utilities, has got more than fifty pages of documentation and dozens of functionalities, and we did not mention more complicated applications like a word processor yet.

Wizards means a significant solution for easing the users' work. Wizards collect the most common user activities and offer them as well defined sequences of functionalities. The main attitude of a wizard is that it decreases information overload by offering the user a limited set of choices, hopefully the good ones.

Most state of the art software systems contain preprogrammed wizards, hardcoded into the software. This article investigates theoretical and practical approaches of writing programs with which can dynamically generate wizards on demand. A dynamically generated wizard is not written during software development, but it evolves automatically during software use, and thus it can optimally reflect actual user's needs.

Users could profit from dynamically generated wizards in two ways. On one hand, a user could get a wizard which consists of her most common tasks and would function as a high quality shortcut for these tasks. On the other hand, wizards for the most common tasks could be generated by professional users by simply using the software, and later these wizards could be used by non-professionals. This significantly differs from the traditional style of wizard development where programmers try to predict the most common tasks during design time.

User modeling and its applications has been studied for years [8,9], and several success has been achieved, for example in the fields of adaptive hypermedia [2,3,4], information retrieval [6,7], or web search [5]. Adaptive user models play important roles in user agents as well [12]. On the other hand, to write a traditional wizard seems to be more a development than a research problem, although there has been several traditional wizard studies performed [10,11] and software environments supporting wizards [10] are available. However, little effort has been made for integrating the two approaches, and creating an architecture, which supports automatic evolution of wizards via an adaptive user model.

The next section introduces the basic architecture and working mechanism of automatic wizard generation. The two middle sections introduce theoretical concepts of model and validation engine, and the article is closed up with some implementation details.

Basic architecture

One of the key problems of automatic wizard generation relates to nature of generic software systems' architecture. To analyze a software system from the user's point of view, four main layers could be considered. Generation of *components* takes place at the lowermost layer. In a technological sense components may be JavaBeans, CORBA, COM, COM+, or even simple objects of an object oriented programming language. The *user interface* resides at the top level of the architecture. It usually manifests as a set of buttons, textboxes, links or text items.

Services can be regarded as an abstraction of calls to components' methods. The major difference is that a service has a semantic meaning to the user, like *'file copy'*, as opposed to a method call, like *'int cpf (File f)'*, which is just an internal call specification of the program. Actually, when the user has access to the program through the user interface, he manipulates the services. Sometimes, there is no strict distinction between services and method calls of a component, because component developers usually write methods which also have semantic meaning to the user [13]. Consequently, services can be regarded as a subset of components methods having semantic meaning to the user; and it is the software developers responsibility to identify this subset.

There is a fourth layer between service and user interface, called *glue code,* which contains the program

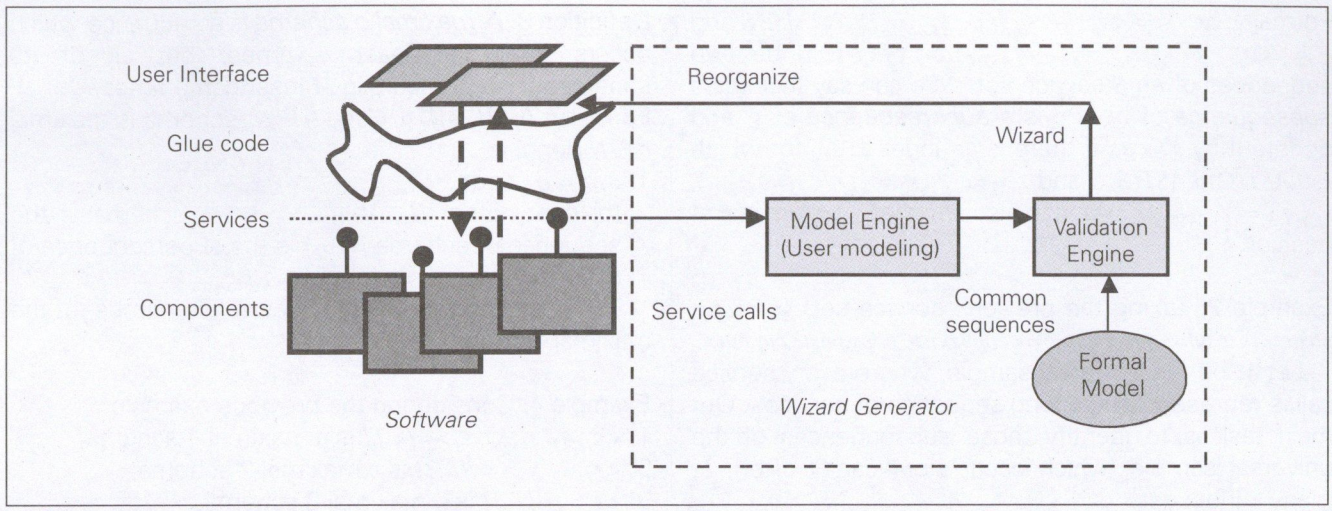


Figure 1. Basic Architecture of Automatic Wizard Generation

code delegating the user's actions from the interface to the service calls. It usually consists of event handlers and other useful methods.

Thus it seems natural to consider services to be the basic steps of a dynamically created wizard. Possible user interface elements and implementations show such a huge variety that it is even hard to take them into account. Similarly, the glue code level is too fuzzy to handle, while the component level is far away from the user interface.

Our wizard generator architecture can bee seen on the right side of figure 1. Its main task is to create a wizard from user's behavior and embed directly into the software so that it can be applied by the user. The wizard generator consists of two major parts, the *model engine,* and the *validation engine,* which will be introduced in the following.

Service calls are collected into a universal sample, while the user is using the software. In other words, all service calls are registered as a single sequence. Wizards should be manifested as common subsequences of all service calls. The model engine is responsible for computing common sequences from all service calls, usually called *schemas*. Its task can be realized by different data mining or dedicated algorithms.

The model engine generates common sequences of service calls from the user's behavior, however these sequences cannot be considered directly as wizards because validation is missing. The validation engine gets wizard candidates from the model engine, and checks the validity of these candidates, with the help of a domain specific formal model. The valid candidates become real wizards; the non-valid ones are dropped or transformed into valid.

The working mechanism of a dynamic wizard generation architecture can be outlined as follows:

1. The user applies the software manifested by service calls at the service level.
2. Service calls are registered as samples.
3. The model engine creates schemas (candidates for wizards) from the sample.
4. Validation engine checks transforms and drops the candidates.
5. Valid candidates become wizards.
6. The user interface is forced to reorganize itself so that the new wizard can appear.
7. The user can have access to the new wizard.

In the following, fundamental concepts of model and validation engine will be introduced.

Basics of model engine

As it has been mentioned above, service calls are registered in a sample. A wizard will somehow match those parts of the sample, which occur quite often. In this chapter, some of our results are outlined, which demonstrate what this *'occur quite often'* statement could mean.

In the following sections the following notations are used. Sets are denoted by boldface characters. E.g. N_0 represents the set of non negative integer numbers. The cardinality of set $\boldsymbol{A}$ is denoted by $|\boldsymbol{A}|$. Since here cardinality is only used for finite sets, it equals the number of elements of a set.

A sequence of services, shortly *sequence*. A sequence is a well defined ordering above a service set. Like $\underline{s} = <s_1, s_2, s_3, s_4 \ldots s_{LAST(S)}>$; $\forall i: si \in \boldsymbol{S}$, where $\boldsymbol{S}$ is the set of all services, and s_i is a service at the i-th position of the sequence. An element of the service set can appear more than once in the sequence, distinguished by its position number. The *LAST* integer represents the position of the last element in a sequence. The sequence itself is denoted by underlined letters.

Example 1. Supposing that the set of all services $\boldsymbol{S}$ = *{copy, paste, cut}*, a possible sequence of $\boldsymbol{S}$ can be *<copy, paste, copy, paste, cut, cut>*, where for example the last 'cut' occurs at the 6th position.

Definition 1. An sequence $\underline{s}$ is a subsequence of another sequence $\underline{c}$, if $\underline{s}$ can be found in c.

Formally: Let $\underline{s} = <s_1, s_2, s_3, s_4 \ldots s_{LAST(S)}> \forall i{:}si \in \boldsymbol{S}$ and $\underline{c} = <c_1, c_2, c_3, c_4 \ldots c_{LAST(C)}> \forall j{:}cj \in \boldsymbol{S}$ be two sequences of an $\boldsymbol{S}$ service set. We can say that $\underline{s}$ is a ***subsequence*** of $\underline{c}$, or $\underline{c}$ is a ***supersequence*** of $\underline{s}$, and denote it by $\underline{s} \subseteq \underline{c}$, if there is an index $k \in N_0$ for which $k<LAST(C)-LAST(S)$ and $s_1=c_k$, $s_2=c_{k+1}$, $s_3=c_{k+2}, \ldots, s_{LAST(S)}=c_{k+LAST(S)-1}$. We can also say that $\underline{s}$ ***occurs*** in $\underline{c}$ at position k.

Example 2. Taking the previous service set, $\underline{s}$=<*copy, paste*> *is a subsequence of* $\underline{c}$=<*copy,paste,copy,paste,cut,cut*>.

Let $\boldsymbol{M}$ be the universal sample, where every service call is represented in a long sequence of services. Our main task is to identify those subsequences of the universal sample, which occur significantly often, at many k indexes.

Definition 2. An ***[l,o] schema*** of $\boldsymbol{M}$ is a sequence of length l, which occurs in $\boldsymbol{M}$ exactly o times.

Formally: An $[l,o]$, $o \in N_0$, $l \in N_0$ schema is an $\underline{s}=<s_1, s_2, s_3, s_4 \ldots s_i>$ sequence of length l in M if it occurs in $\boldsymbol{M}$ at $\boldsymbol{K}=\{k_1, k_2, \ldots, k_o\}$ positions, $|\boldsymbol{K}|=o$, and there is no other position where $\underline{s}$. Sequence $\underline{s}$ itself is the ***sequence of the schema***, l denotes the length of $\underline{s}$, o denotes the occurrence number of $\boldsymbol{K}$ is the position set of the schema, and the length l sequences in $\boldsymbol{M}$ starting according to the elements of the position set are the occurrences of the schema.

Example 3. In $\boldsymbol{M}=<s_1, s_2, s_1, s_2, s_1>$ there are:

1. two length one schemas $<s1>=[1,3]$, and $<s2> = [1,2]$;
2. one length two schema $<s_1, s_2>=[2,2]$;
3. two length three schema $<s_1, s_2, s_1>=[3,2]$ and $<s_2, s_1, s_2>= [3,1]$;
4. two length four schema $<s_1, s_2, s_1, s_2>=[4,1]$ and $<s_2, s_1, s_2, s_1>= [4,1]$;
5. one length five schema $<s_1, s_2, s_1, s_2, s_1>=[5,1]$.

It is important to notice that the two occurrences of the length three schema in $\boldsymbol{M}$ overlap each others. On one hand this is correct according to definition 2. On the other hand, it is a desired property, because most common patterns of user behavior are being looked for, independently when and how they occur.

If a sequence of *length* occurs only one time in the universal sample, it becomes a *[length,1]* schema. Consequently every subsequence of the main sample are *[length,1]* schemas, or *[length,o]* schemas, where $o>1$. A sequence which is not a subsequence of the universal sample is a *[length,0]* schema, in other words a *zero schema.* Zero schemas is paid little attention this time.

Considering the previous definitions, a wizard should be established upon schemas having high occurrence numbers. Unfortunately, it would result a high number of wizards even for a relatively small universal sample, because if a schema occurs in a universal sample, then all schemas of its subsequences also occur. So schemas should be created which are in some sense maximal.

Definition 3. A ***maximal o schema*** is a sequence which occurs in $\boldsymbol{M}$ at least o times, but all of its supersequences occurs in $\boldsymbol{M}$ less than o times.

Formally: An $[l_1,o_1]$, $o_1 \in N_0$, $l_1 \in N_0$ schema is maximal $o \in N_0$ long, if

1. $o_1 >= o$, and
2. for every other $[l_2,o_2]$, $o_2 \in N_0$, $l_2 \in N_0$ schema: if the sequence of schema $[l_2,o_2]$ is the supersequence of schema $[l_1,o_1]$, $o_2<o$ follows.

The o integer denotes the occurrence number of the maximal schema.

Example 4. Considering the previous example:

1. $<s_1, s_2, s_1, s_2, s_1>=[5,1]$ is a maximal 1 schema;
2. $<s_1, s_2, s_1> = [3,2]$ is a maximal 2 schema;
3. $<s_1> = [1,3]$ is a maximal 3 schema.

In our approach, wizards are generated based upon the concept of maximal o schemas. Consequently the sequence of a wizard occurs o times in the sample, and it is maximal, so there is no other wizard which contains its subsequence. An input o is defined with the help of different heuristics. The aim of tuning o is that the number of wizards should not be either zero or very high. Based on running experiences, this can be reached by choosing $o = c*\lg(n)$.

Three algorithms were designed to solve maximal sequence problem: Mschema0, Mschema1, Mschema10. The first one is a brute force algorithm with $O(n^3)$ time and space complexity, where n is the length of the universal sample. Space complexity has been succeeded to reduce in Mschema1, from $O(n^3)$ to O(n), unfortunately the speed of the algorithm has remained the same. Mschema10 is the mixture of the two previous algorithms with $O(n^3)$ time and $O(n^2)$ space complexity, however it produces a user model beside maximal schemas. In real life applications, Mschema1 and Mschema10 algorithms seem to be efficient enough, whilst Mschema0 is not adequate because of its space complexity.

Theoretical framework of validation engine

As it has been mentioned previously, validation engine gets sequences of service calls from the model engine, and filters them with the help of a formal model. The principle questions are what the formal model should contain, and how the filtering should be realized.

Fortunately, the algorithm of the model engine has some nice properties. If it produces an $\underline{s}=<s_1, s_2, s_3, s_4 \ldots>$ sequence as a wizard candidate, it can be supposed that all services in the sequence are in valid order. In other words, if s_i precedes s_j in an $\underline{s}$ wizard candidate they must be in a valid order. This is a consequence of the user model, if one calls s_i and directly after s_j, it means that the two services can be called in this order

(supposing that the software itself is not faulty). The only problem which could remain is that there might not be a necessary service in the sequence. For example, an *open_connection* service call always has to precede a *remote_copy* call.

Consequently, formal model has to express two constraints, a preceding and a following constraint. Certainly, they could be expressed by different formalisms, like LTL (Linear Temporal Logic), CTL (Computational Tree Logic) or modal μ-calculus. Unfortunately, a fast possibly linear approach is needed for evaluating the constraints, and for transforming a non-valid wizard candidate to a valid one. That is the reason why a relational formalism have been chosen.

Definition 4. The must precede $MP \subseteq S \times S$ and the must follow $MF \subseteq S \times S$ relations are binary homogenous, irreflexive, antisymmetric relations above the S set of services.

Homogenous and binary means, that relations make connections between two elements of S.

The consequence of irreflexivity is that an $s_i \in S$ cannot precede or follow itself, and the antisymmetric property results that no $(s_i, s_j) \in MP$ and $(s_j, s_i) \in MP$ can be valid at the same time.

Example 5.

Supposing that the set of services include $S=\{open\_connection, remote\_copy, close\_connection\}$, a valid must precede relation could be $(open\_connection, remote\_copy) \in MP$, and $(remote\_copy, close\_connection) \in MF$. In other words, if we have a *remote_copy* service call in our wizard, there must be an *open_connection* previously, and a *close_connection* following.

Definition 5. The *transitive closure* of MP is the minimal $R \subseteq S \times S$ relation (minimal in the sense of consisting the minimal pair of services) for which $MP \subseteq R$, and if $(s_1, s_2) \in MF$ and $(s_2, s_3) \in MF$ implies that $(s_1, s_3) \in MF$. The R relation of MP is usually denoted by $tr(MP)$. The definition is the same for the transitive closure of MF.

Example 6. Defining $S=\{start\_program, open\_connection, remote\_copy\}$ service set, and $MP = \{(start\_program, open\_connection), (open\_connection, remote\_copy)\}$ must precede relation $tr(MP) = \{(start\_program, open\_connection), (open\_connection, remote\_copy) (start\_program, remote\_copy)\}$.

In this example *start_program* denotes the service which starts the software capable of making the remote copy.

Simple *must precede* and *must follow* relations represent the domain specific formal model given by the developer or programmer, as opposed to transitive closures, which is actually used during wizard validation. This ease the developers work, since only a small number of preceding or followings have to be specified, but every indirect preceding (like *(start_program, remote_copy)* in the previous example) is also considered.

Definition 6. Graph representation of MP, is a $G(MP)=<S,E>$ directed acyclic graph for which $E \subseteq S \times S$, and $(s_1,s_2) \in E$ if $(s_1,s_2) \in MP$. It implies that if $(s_1,s_2) \in tr(MP)$ there is a directed path in $G(MP)$ from node s_1 to s_2.

Example 7. Considering the previous example, the figure 2 shows the graph representation of MP; arcs are elements of MP, and paths are elements of tr(MP).

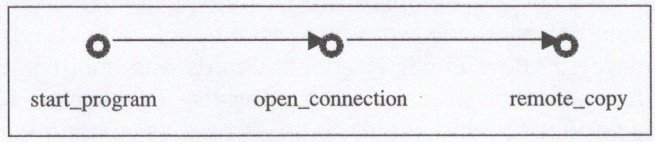


Figure 2. Graph representation of a must precede relation

Graph representation is a formalism which consists of both the relation and its transitive closure. Since it is a directed graph, algorithms for any purpose could be easily created.

Definition 7. An $sch=<s_1, s_2, s_3, s_4 ....>$ sequence of services satisfies an MP must precede relation denoted by $sch \models MP$, if $s_i \in S$ occurs in sch and $\exists\, c \in S$ for which $(c,s_i) \in tr(MP)$ than c also occurs in sch at position j, and $j<i$.

Must follow relation can be defined similarly, however it describes that c service call has to follow s_i.

Example 8. Let the service set be:

- $S = \{start\_program, open\_connection, remote\_copy, close\_connection, exit\_program\}$,

the two relations:

- $MP = \{(start\_program, open\_connection), (open\_connection, remote\_copy)\}$

the transitive closure is:

- $tr(MP) = \{(start\_program, open\_connection), (open\_connection, remote\_copy), (start\_program, remote\_copy)\}$,

The $sch_1 = <open\_connection, remote\_copy, remote\_copy>$ sequence doesn't satisfy MP relation, because there must be a service call before open_connection *(start_program)*.

The $sch_2=<start\_program, open\_connection, remote\_copy, remote\_copy, close\_connection, exit\_program>$ satisfies MP relation.

The previous definitions represent the theoretical framework of the validation engine. The developer specifies a formal model, which consists of the two relations, for each software. The model engine generates the candidate wizards which are checked by the validation engine. it is valid if it satisfies both the must precede and the must follow relation, so it

become a wizard, else it is non-valid, so it is dropped, or transformed (extended) into a valid sequence.

Implementation

Perhaps the most critical part of every theory is its application. Therefore, a shell system architecture for generating dynamic wizards has been developed and tested in Java, based on the previous results. The major goal is to create an object oriented piece of software to support automatic wizard generation.

The problem is not quite trivial, because the user interface has to be reorganized after generating a new wizard. In the optimal case everything would occur automatically and software developers should not have to know anything about wizards and about the reorganization process of the interface. To reach such a perfect solution seems to be as hard as Hunting for the Holy Grail. Instead, a design pattern has been introduced, which the user has to follow during the implementation. To apply this design pattern does not require much extra work, just the interface of the software has to be developed in a well defined way. Based on this design pattern, all other work is done automatically by the system shell.

Conclusion and further research

This paper has examined the possibility of writing a software in a way that it would be capable of automatically creating wizards. The basic architecture of automatic wizard generation, possible interpretation of wizards, and theoretical framework of validation engine, and some hints about implementation were presented.

Since the architecture consists of three major parts, model engine, validation engine and technological side, further research will be focused on developing these. Model and validation engine can be extended and developed in mathematical sense, whilst better implementations and design patterns can be given at technological side.

References

1. Thomas, H. Cormen, Charles E. Leiserson, Ronald L. Rivest. Introduction to Algorithms. The Massachusetts Institute of Technology Press 1990.
2. Giovanni Fulantelli, Riccardo Rizzo, Marco Arrigo, Rossella Corrao. An Adaptive Open Hypermedia System on the Web. Lecture Notes in Computer Science Volume 1892, pp 305-310. 2000.
3. Brusilovsky, P. Methods and Techniques of Adaptive Hypermedia. User Modeling and User Adapted Interactions 6, pp 87-129, 1996
4. Hongjing Wu, Paul De Bra, Ad Aerts, Geert-Jan Houben. Adaptation Control in Adaptive Hypermedia Systems. Lecture Notes in Computer Science pp. 250-259. 2000.
5. Maria Fasli, Udo Kruschwitz. Using Implicit Relevance Feedback in a Web Search Assistant. Lecture Notes in Artificial Intellingence. Volume 2198, pp 356-360. 2001.
6. Fan Lin, Liu Wenyin, Zheng Chen, Hongjiang Zhang, Tang Long. User Modeling for Efficient Use of Multimedia Files. Lecture Notes in Computer Science. Volume 2088, pp 182-189. 2001.
7. R. I. John, G. J. Mooney. Fuzzy User Modeling for Information Retrieval on the World Wide Web. Knowledge and Information Systems. Volume 3, Issue 1, pp 81-95. 2001.
8. C. Stary. User diversity and design representation: Towards increased effectiveness in Design for All. Universal Access in the Information Society. Volume 1, Issue 1, pp 16-30.
9. Readings in Inteligent User Interfaces. Morgan Kaufmann Publisher, Inc. 1998.
10. Doug Tidwell, Jeanette Fuccella. TaskGuides: instant wizards on the Web. Annual ACM Conference on Systems Documentation, Proceedings of the 15th annual international conference on Computer documentation 1997.
11. Lori Phelps. Active Documentation: Wizards as a Medium for Meeting User Needs. Proceedings of the 15th annual international conference on Computer documentation 1997.
12. Jeanne Murray, David Schell, Cari Willis. User centered design in action: Developing an Intelligent Agent Application. Proceedings of the 15th annual international conference on Computer documentation 1997
13. Guy Eddon, Henry Eddon, Inside COM+ Base Services PUBLISHED BY Microsoft Press A Division of Microsoft Corporation One Microsoft Way Redmond, Washington 98052-6399 1999.

Towards a Win-Win Outsourcing Relationship

DR. GYÖRGY BŐGEL

Strategic Advisor of the KFKI Computer Systems Corporation
Associate Professor of the IMC Graduate School of Business, CEO, Budapest

An outsourcing is successful from economic aspects, when new, added value is generated by that, which can be distributed between the customer and the service provider: the revenues increase, costs decrease, the invested capital brings substantial profit, generation of added value and sharing such values makes the key element of the outsourcing issue.

No doubt, there is some mystery about outsourcing. Outsourcing is when an organization (for example, a company or any other institution) decides to purchase some service – that was delivered earlier inside the company – from other, external organization. In this sense of word outsourcing means to „move the resources outside the company" i.e. the action of taking out. If a company decides not to have an own warden function, and wants to purchase the related activities (including maintenance, cleaning, painting, reception service) from another company, then we say that warden activities have been outsourced. At the same time we would not call it outsourcing if a company had never had in mind to do something in-house, (e.g. produce electricity with its own generators or operate water works), as these services were performed by other providers from the very beginning.

To keep it simple, lets insist on the interpretation that – as far as the activities relating to outsourcing are concerned, the basic idea is that to-date some sort of activity was carried out in-house, while from now on it is purchased from outside, from another organization. A process was performed to-date by the units of the organization, while in the future it will be performed within the scope of authority of another company which organization is in market relation with the outsourcing company. As a result of outsourcing, the hierarchic relationship (my bookkeeper is my employee) will be replaced by a market-type relationship (my bookkeeper is my supplier).

The question is, what will be improved with this change, and what are the conditions of making things better. We must not forget that many of the outsourcing campaigns end up with failure, and do not realize at all the hopes associated with outsourcing. From economic aspects, outsourcing is successful, if added value is generated by that, which can be shared by the customer and the service provider, revenues increase, the costs decrease, the invested capital pays back. Generation of added value and sharing of such added value are the key elements of the outsourcing issue.

The logic of choosing between internal and external resources is simple: do what you can do the best, and leave the rest to those who can do it the best. The ideological background was created by the book of Prahalad and Hamel (1) which emphasis the importance of basic competencies: everybody must clarify what abilities he has, and based on that decide what are the activities he wants to perform himself.

The major targets of outsourcing are normally the services performed inside the organization, so, for example, at a typical manufacturing company the maintenance of production lines, cleaning the offices, bookkeeping, copying, recruitment of employees, IT, designing advertising materials, operating the workplace canteen. The service provider organized as functional unit, or cost center is in monopoly position, especially when it does not have to compete for other customers. Instead, they wink upward, try to develop good terms with the management. They make efforts to influence planning, to get increasing budget and comfortable, easy-to-perform assignments. No saving is made on costs. The effect of occasional initiatives of the management to rationalize the costs vanish quickly: quality is not improved, costs do not decrease, even grow constantly. (2)

The solution seems simple: the service provider should be removed from its monopoly position, competitive environment should be created, the service provider should be made interested in the rationalized business.- that is, hierarchic relations should be replaced by market relations, where the

customers are free to select a service provider whereas the service providers must run after work. In the case of full-scale outsourcing the internal service provider disappears: the remaining units and service providers face each other as market players, as customer and supplier, whose rights and obligations are set down in market agreements. There is no hierarchic, pears' relationship whereas market interests, profit interests are prevailing, and parties are under permanent pressure of improvement of efficiency.

An activity or a process can be moved outside the company, whereas the party doing outsourcing maintains the control over the particular activity. Just to mention an example: Manufacturing of certain parts is transferred to a partner who must follow the technological guidelines, instructions and the schedule, etc. received from the customer, while the control is also maintained by the customer. In English this version is called „contracting". The category of „contracting„ also includes the cases, when a company employs its staff persons on contracting basis – primarily for reasons of taxation and flexibility. In legal terms an „outsourcing" is performed also in this case. that is, the customer purchases service from outside, while in fact they do not transfer the control over the particular activities, therefore it would be an error to call such activity an outsourcing.

As you will see later on, it is an important issue to make distinction between outsourcing (a process of acquiring services from external source with the transfer of control) and contracting (acquiring service from external source without transferring the control over the activities).

From the history of IT outsourcing

If you approach the issue of outsourcing from historical aspects you can see that the circle of those doing outsourcing is expanding continuously. IT has also appeared among the candidates, long ago. At the times when mainframes of room size were applied, and expensive computers could be used on time sharing basis, we could say that „outsourcing" came first which was followed by „insourcing". (for example, in Hungary the background institutions of the particular ministries organized as quasi-companies were operated as IT centres). Programming staff has been hired for over 30 years. The outsourcing of payroll activities looks back to a history of several decades. Electronic Data Systems (EDS) established by Ross Perot and the „Big Six" companies have been present in the market of IT systems for some fifty five years. Network sharing cannot be considered as a novelty, either.

Some companies have become aware that they can sell their internal services for external customers as well, and they entered the outsourcing business. Specialization and segmentation of the market started: for example, a group of the service providers, who struggled continuously with efficiency problems, targeted the government sector with outsourcing offers. The English and French companies also followed the American example (Hopkins, Cap Gemini) and by now we are free to say that the business has been globalized, as after Europe neither Asia nor Australia wanted to lag behind the others. The largest companies (among others IBM, CSC, EDS and Andersen Consulting) are offering global solutions at the moment. „If your are present at every point of the word, we are there, too, and you can leave your IT system to us„ (3).

New colours have appeared on the palette of outsourcing services. In addition to data processing services the otusourcing of processes relying on IT become increasingly popular (like business process outsourcing,) for example, the processes of logistics (Ryder), reservations in air transportation (Sabre), documents/file management (Xerox), or HR work (Hewitt)

Market – as it is normally the case- does not evaluate equally all the outsourcing business models: some of them are drop-outs, whereas others undergo rapid growth. Currently the „hottest" area of IT outsourcing is Internet, (4,5)Due to Internet the so - called „off-shore" version of IT outsourcing has emerged, (6) which makes use of the feature that geographical distance is not significant in case of certain services (like, for example, the operation of customer service call centres, or data analyzing) the service provider may sit even on the other side of the world.

Value adding by outsourcing

Lets' imagine the following situation: A company operating in machine industry operates an IT department with dozens of employees. The company is dissatisfied with the IT function: they find the costs too high and service quality too low. After some investigation the managers of the company make decision about outsourcing. They look for a service provider, agree with them, after signing the contract the partner takes over the people and machines. The earlier, internal hierarchic relations are replaced by market relations: the service provider cannot be instructed by the management, the relation is regulated by the signed market -based contract.

But what is better, what can be better than this solution:

As we have already mentioned earlier, outsourcing makes sense in case it generates new, added value, that can be shared by the customer and service provider. Naturally, no added value is created by the

mere fact of outsourcing, and transfer of control: if the same or very similar people carry out the same activities as they used to do before, there is no profit to be shared, the costs will increase, no profit is generated. Both parties make loss, or one of the parties can make benefit of the deal only on account of the other one, which is a shaky basis for a long term relationship. And we did not even talk about risks, costs that are inevitably incurred by the process of reorganization, and about payback of the investment. Obviously, the question is how to create added value, and how to shape the customer – service provider relation to ensure the growth.

Added value for the customer

Added value is generated for the outsourcing party in case their costs reduce, or the revenues increase. Let's see, what can be expected from outsourcing.

a) Most outsourcing customers expect the reduction of their costs. Where the costs accounting system operates properly, the company can see clearly the actual costs of operating the IT function in -house for a certain period, and the company will make efforts to acquire the external service cheaper.
b) The costs of IT activities performed in-house are supposed to be fixed ones, irrespective of the IT facilities used, and the volume of IT usage. As a result of the outsourcing agreement the fixed costs will be converted into variable costs: if the customers need less services, then they will use less service and also pay less: the system is flexible and scaleable. The American companies that have been able to reduce their capacities during the economic recession faced at the millennium made benefit of the scaleability (which is obviously a bad news for those who lost their jobs earlier for this reason).
c) A properly developed outsourcing agreement links service fees to indicators of usage, performance levels, Based on that the costs can be budgeted accurately, and remain controllable, cost drivers become clear, and due to the above they are easy to identify and make benefit of the cost reducing opportunities.
d) In case of outsourcing normally the service provider takes over or purchases the existing facilities and equipment of the customer, therefore he acquires capital.
e) In case the customer needs new assets, he does not have to invest funds, as he purchases service, rather than new hardware or software facilities.
f) A large, successful IT company has obviously more chance to acquire excellent IT experts than a company of other profile, where professional career opportunities are restricted. Its experiences are obviously much wider, more extensive. The customer may reasonably expect to get a professional service provider, and service quality will improve.
g) Although it is a commonplace, IT exerts deep and direct impact on corporate processes. The customer may have the expectation that the service provider should support him in the reengineering of its business processes, in the organization and implementation of process-reengineering programs (see, e.g. 7,8)
h) Where outsourcing operates properly the customer company can devote more attention to its core abilities and processes
i) The customer may gain new abilities and expertise through its outsourcing partner, and even can come up with new products, services, business lines and sales channels in the market, increasing by that its revenues.

Added value for the service provider

Let's analyze the issue of added value from the aspect of the service provider: what can the party taking over the outsourced activities expect?

a) The service provider must obviously be a professional IT company, with multiple assignments and plenty of customers. For him the economy of sale relating to provision of mass service may represent the chance for generating added value: he is able to better utilize its capacity and infrastructure in the different assignments of different volumes (a staff person can service more than one customers, a server is able to store all the data that earlier were stored on the servers of various companies, etc.,) the overhead costs of the company can be spread over more assignments, more customers. Mass scale provision of the service entails more experience and faster learning process: the company can progress quicker on the experience curve, the costs can be reduced, and the price of learning has to be paid just once in new solutions.
b) The service provider may utilize the areas where he has higher competence than that of the customer: what is service for the customer, it is the core activity for the service provider, based on its core competencies. As he has got more than one assignments, he assumes various tasks, his experiences, knowledge can be much wider than that of the customer. By utilizing its competencies he may reorganize the activities taken over, make them cheaper and more efficient. He must be better in terms of professionalism and cheaper in terms of costs than the customer.
c) He may make benefit of its professional features, size or even its geographical position: he may purchase products and software applications cheaper due to the large volumes, may develop advanced, more extensive infrastructure, can hire

better, more experienced staff, etc., In terms of the IT outsourcing services one of the most successful countries is India, where very good experts can be hired at relatively low wages, and the entire national IT strategy is built on that feature (6).

d) The service provider may also expect that with its new partner he may create something new, acquire new markets and revenue sources. The better he learns the customer, the more opportunities of cooperation and service extension can be identified, for example he may open the door for his customers towards e-commerce, and can supplement its products with e-services.

Outsourcing clearly makes the most rapid progress in those areas, where more than one of the above mentioned opportunities are available.

As the above two listings suggest that not only the process of outsourcing (decision-making about outsourcing, its approval, looking for an appropriate partner, outlining the contracts, etc.) are exciting: it is much more interesting, what happens after signature of the agreement. Without the opportunity for value adding the idea of outsourcing is not viable, one of the partners – or rather both of them – will suffer losses. The options listed above are only opportunities, they turn into actual value only in case the parties get down to their implementation. It requires a system of relations between the customer and service provider.

Development of a harmonious relationship is a hard task. What is a revenue to one partner is a cost to the other. The customer wants to reduce its costs, whereas the service provider wants to generate more revenues. The customer is persuaded by its own customers to reduce the costs continuously whereas the owners of the service provider expect the growth of revenues. In case the latter one decreases its costs, it is likely to be done on account of the quality – and so on. In this sense of the word the service provider and customers are competitors to each others, who want to acquire a bigger stake of the value that can be generated or actually generated in the deal. That is, we have to answer two questions:

1) what value can be generated jointly by the customer-service provider;
2) in what proportion is it shared by the contributors.

The ratios of sharing the revenues are subject to the bargaining positions, while bargaining position is influenced by features like, for example the intensity of competition, the extent of exposure to customers or vendors, the difficulty of changing the partners. We may put the same two questions when analyzing the outsourcing relations. We have already mentioned the opportunities for value adding, still, to understand the proportions and tactics of revenue sharing we have to analyze the relation between customer and service provider.

Lock-in

The situation when a company gets into he captivity of another company is called „lock-in" in the English language literature, which could be translated to Hungarian as the process of capturing or getting into captivity. Either the customer, or the supplier or partner may be the capturing party, which means that it is difficult to get rid of him: migration to another partner would be too costly or energy-consuming. What is more, lock-in is often not unpleasant at all, the party being locked in does not even realize that he is in a cage.

„Lock -in" is a frequent feature of IT industry, certain companies regularly apply it either as a sophisticated or as a rude method, and it features as the central element of the „lock-in" strategy (10) (with certain cynicism we may even say that everybody would like to apply this method, but not everybody is able to do so or manages to do so.) The company locked in will develop or purchase something, invest funds processes, systems, and develops the necessary channels: installs, learns, get accustomed to the products, gets to like it, and finally he is not able to give it away or has to assume substantial losses to get rid of it: the systems must be replaced, new staff must be trained, new relations have to be developed, and so on. In a marginal case the party locking in the other will get into a de facto monopoly position, and may conduct accordingly.

The bargaining position of the locked-in company will worsen as far as the other party (who – as we have already mentioned – may be either the customer, or the supplier, or partner – recognizes and abuses the exposed position of his counterpart: for example, raises the prices, decreases the quality or defines new, special conditions as requirements for maintaining the relation. The more complicated, sophisticated the product or service is, the bigger is the danger of getting locked in. The more standardized and common the service/product is (like, mass products, standard services) the easier is to change for another partner. The history of relations obviously contains actions when (a) the customer or service provider makes efforts to abuse the exposed position of the counter-party in a bargain, or where (b) the party makes efforts to get out from the cage, i.e. reinforce its bargaining position. Obviously, these strategies and actions are not always determined and planned, as getting locked-in or locking-in the other is not always deliberate, either.

The outsourcing party is locked in

Let' assume that IT is important to the outsourcing party while it is not among its core abilities, therefore he makes a decision about outsourcing. He generates fund by the transfer of facilities and people, whereas he also incurs cots, makes investments. Jérome

Barthélemy of France (11) directs the attention to the hidden costs of outsourcing. They are hidden in terms that they don not always involve direct expenditures, or expenditures are made but they are included in other costs therefore they are difficult to identify and measure. These hidden costs include, for example:

a) Looking for the service provider, contracting. According to the data of Barthélemy the average costs and expenditures amount to USD half million, while they seem to be fixed ones, they depend to a lessor extent on the value of the outsourcing campaign. The published data also suggest that even one or one-and-a half year may pass from the point where the idea comes up (IT should be outsourced) to the starting of the service, while the outsourcing party also has to take its share from the intensive work.
b) Transfer of activities to external service providers. It may take months to the service provider to acquire the same level of knowledge and skills that the internal IT unit has. According to Barthélemy the transfer of service takes one year on the average. Before the transfer of activity completed the outsourcing party will incur costs the extent of which is difficult to assess: for example, we should assess the value of the assistance provided by internal service providers to the external ones. The more sophisticated the IT service is, the higher the costs of transferring the service are. and obviously, the transfer is more difficult as well. The „human costs„ of outsourcing must be paid as well: the outsourced departments may feel that they are betrayed, their employment and wages are threatened, the affected members may resist, cut down their performance or even my leave the company. Even there were examples when the outsourced employees were on strike for weeks.
c) Most of the costs is incurred in connection with managing the outsourcing relations: we have to monitor the fulfillment of the obligations by the service provider, should he fail to do so the outsourcing company must intervene, all the aspects of amendment of the agreement must be negotiated. In the study sample of Barthélemy the average management costs amount to 8 percent of the annual costs of the outsourced IT services. Reasonably we have to take into account that most of these costs are fixed ones, that is, the companies concluding outsourcing agreements of lower value relatively spend more on managing the relations.

It would be a bad mistake to forget about the above listed „hidden costs". An American oil company acquired IT services 7-8 percent cheaper than the operating costs of the earlier, internal IT department. It was very favorable in itself, still, the management costs amounted to 17 percent of the annual costs. At the time of outsourcing they could not assess how much the simultaneous management of four external service providers would cost. What is gained on the one side, lost on the other one: the final balance was negative.

All these suggest that the outsourcing party has to be tuned to the service provider, the relationship has to be managed, problems solved and conflicts sorted out. If he decides to change for another service provider, he must start the entire process from the beginning, and earlier investments will vanish. This is where the handcuffs are closed.

The service provider is locked in

The same handcuff is on the hands of the service provider, too. He will have lots of expenses at the beginning: He has to obtain the contract, take over the facilities, equipment and staff of the customer, rationalize the capacities, and processes, has to learn the customer, understand his expectations. He may expect to generate revenues only after the service is launched. During the initial period the cash-flow balance runs to the red, the initial investments must be financed on account of the future profit while profit is generated only in smaller lots, as the difference of the service fees and the costs incurred in the following periods. That is, the service providers must think in the longer run, and obviously, they must insist on the established relations, if he invested that much.

That is, the exposure is mutual even if it is not necessarily symmetric. Both parties can do a lot to change their bargaining position. For example, the customer may decide to outsource only the „mass-type" IT activities, i.e. those, where lots of service providers are available, and it is easy to migrate from one service provider to another. He may select a service provider who considers the success of the deal a must. On the other hand the service provider may progress step by step with offers of increasing sophistication, level of complexity and specialization: in the beginning he gets foothold of an area, later on conquests it in full.

Handling the tensions

An outsourcing relationship always implies tensions of some sort. The conditions of cooperation, demands, technologies and the services themselves change rapidly and continuously, which modify the bargaining positions the direction and extent of exposure the of the parties without any deliberate action. There are methods and opportunities for reducing the tensions, out of which the following are the most important ones:

a) Careful definition of the elements and indicators of the outsourced process

The customer must clarify as early as in the request for bid the elements of the process to be outsourced, and the performance indictors he expects in respect of the particular elements. He must be aware of what is representing value for him and what performance is worthwhile to pay for. It often happens that a marginal difference of performance – which is non-significant to the customer – entails large differences in price.

The majority of the IT processes is integrated into other processes (e.g. inventory management, bookkeeping) i.e. it is varying between the customer and the service provider. In such cases it is necessary to clearly define where the liability of the former party ends and that of the latter party starts.

Normally the outsourced service is not homogeneous. Its parts should be reasonably handled separately as they may incur costs of different types and volume.

The outsourcing agreements are normally concluded for longer periods. The customers often come up with the demand for continuous improvement of the performance, as they are also required to provide improving performance indicators at the other end of the processes. The contract should reasonably set the rate of development expected from the service provider, i.e. it should not be outlined as an unclear, unidentifiable expectation.

b) Proper price for all the services

Price negotiations can come after the thorough and careful assessment of the outsourced process elements and clarification of the expectations. The price will define the ratio of sharing the generated added value (if any) between the customer and service provider. Under the liberalized market conditions the price depends on the competition and the bargaining power.

In some cases the large companies – abusing the exposed position of the suppliers – require the suppliers to agree to the „open book policy" i.e. allow insight into their books and define the purchase price subject to the acknowledged costs. If this is the case, the service provider has the only chance that he can realize a marginal profit above his costs which is considered as „acceptable„ to the customer. This might be a working solution in case of a contracting type solution where the customer does not transfer the control over the process to the external contractor, still, in case of a real outsourcing it is rather harmful. The service provider must have an opportunity to generate the largest possible added value by utilizing the scale of economy realized on mass products, utilizing his expertise and get his fair stake from the added value, to be able to generate resources for his own development.

As the goals and conditions change rapidly, it makes no sense to fix the prices in the long run. The customer must stipulate clearly the service he expects against the price, separate agreements should be concluded for eventual, extra services.

c) Flexibility, including the options of amendment

There are certain services in the area of IT outsourcing, that can be considered a mass products, that can be differentiated to smaller extent, there are lots of competitors, the costs of migration are relatively low, while more sophisticated special, individual services also exist, that require expertise which is difficult to replace.

Moving toward the latter one, the position of the service provider is gradually improving in the price negotiations. If he wants to make advantage of it, he will make regular attempts to increase his prices. In such case the major trump of the customer is the option of granting new contracts: the service provider can expect the extension of the assignment only in case he takes into consideration the interests of the other party as well.

In IT outsourcing the attempts to change the prices – initiated by any of the parties – are considered as common features. A number of conditions influence the prices and price negotiations, including technological development, changes of the demands, growth or reduction of the costs, trends of market competition, and maturation of the market. These conditions may influence the particular service types or service elements in different ways and to different extent. Therefore it is reasonable to avoid „package„ prices, when a number of service elements are bundled in one package, and a flat rate is defined for them. Prices can be changed more flexibly if the prices assigned to certain service elements are clearly defined. Cross-financing among the various service types is a similarly doubtful solution. The relationship between the parties becomes unclear, non-transparent and difficult to oversee. An increasing number of customers applies the practice of building incentives in the contract along with the prices that stimulate the service provider to improve its performance.

Long-term cooperation does not mean that the concrete agreements also have to be concluded for longer periods. For example, P. Bendor-Samuel an outsourcing expert argues for short term agreements (12) as they do not establish ever-lasting connections while the conditions change rapidly.

d) Coordination of the goals and interests

The pre-condition of a long-term, mutual prosperous relationship is the clarification and harmonization of the goals and interests. The objectives of the customer can be arranged into hierarchy, and business objectives are on the top of the hierarchy. The customer must clarify what role is played by the IT

services in the achievement of the business goals and based on that – and naturally, on the financial means – he has to define the performance levels the service provider expected to deliver. It is unnecessary to expect and pay for performance that cannot contribute significantly to the implementation of any of the business objectives.

The he service provider can see clearly his own role and tasks if he understands the business objectives of the customer, and understands the role of his own service provider performance in the chain of goals and objectives. He may think over, what values can be provide to the customers where are opportunities to the extension of the service provider relationships, what are the areas where investments are useful, what are the areas where no mistake is allowed and where can he have more freedom of movement.

Performance measuring and regulatory system

Since the outsourcing of IT activities has become common practice an increasing number of articles and books are sold in book-shops and published in journals on service level management, (SLM) One of these sets forth the following definition: SLM is the summary of the proactive methods and processes arranged into a system that ensure that all the users get appropriate, quality service in accordance with the business priorities, at a reasonable price.

This SLM philosophy is some sort of summary of the above. The customer and service provider agree about a performance level expected from the outsourced service, harmonized with the business objectives of the customer, which is affordable to the customer in financial terms. These specifications are defined in indicators, measurement figures (the document containing the result is called in English the Service Level Agreement (SLA) then a regulatory system is assigned to it, that contains the necessary values, measurement figures, benchmarks, and actions, i. e. it can be considered as some sort of controlling. One of the pre-conditions of efficiency of the system is the provision of clear, objective and reliable indicators.

To define them, it is necessary to clearly identify the services to be purchased, and naturally, they should be reasonably applied in pricing as well, as a good indicator clearly states what we buy and what we have to pay for it. We have to re-emphasize the importance of benchmarks, indicators, measurements. Normally people focus on the features that are assessed or measured by the managers or customers can measure, and the indicators and requirements are assigned to these features.

Development of the regulatory system. Lets start with the features that can be measured. Indicators (measurements) can be assigned for example a) to the outsourced activity itself, to the mode of providing service: for example, the number of employees devoted to solution of certain task by the service provider can be measured, or the number of controls they perform. This solution is closely related to the „contracting„ version of outsourcing outlined in the above, not so much to actual „outsourcing". Measurable features include (b) the results of the service, its quality, for example, availability of the equipment and facilities, or the time used for fault repair. The impact of the services on customer performance is also measurable (c): the response time to orders, productivity, quality of the products, the turnover rate of inventory. What is more, we even can make a step forward in the supply chain (d) and we also can speak about the satisfaction of our customers.

As a first approach, we think the good solution is that we aim to measure the requirements of the (c) service level, as in this case the service provider will support the achievement of the business objectives of the customer. Whoever does so, will work with business terms (cost, productivity, cycle time, inventory turnover, etc.)rather than technical ones. Do not forget that the measurement system can also serve as means of communications: there is a threat that the customer uses „business" terms while the service provider uses „technical „ language, and they misunderstand each-other. That is, measuring of business performance seems in principle a good solution, while in practice the performance of the customer is also influenced by different features, others than the service level which are even stronger than service level. Endless debates can be started about who is to be praised and who is to be blamed. The influencing abilities have to be assessed thoroughly, especially in the case of indirect, mediated solutions.

Lets go back for a moment to the definition of service level management (SLM). The term '"proactive" is contained in the definition, which suggests that a perfect regulatory system aims to prevent troubles, rather than register it on retroactive basis, „post-mortem". When shaping the system a decision must be made about the frequency of checks. If checkpoints are rare, the trends showing the problems are less apparent, and we are likely to detect only the death of the patient.

Some further problems worth mentioning in connection with the measurement system:

- Performance levels may vary in time, and they are much likely to do so. Generation of added value is no immediate process: learning the customers, reengineering of the processes, rationalization of the assets require time. Parties have to clarify when they expect the deliverables, and what changes and what extent of changes are expected in respect of the particular indicators.
- The service providers must get the information and the resources from the customer.

- As it often happens, the related activities of the service provider also have to be reengineered to increase the performance of the service provider. IT activities -as we have mentioned earlier – often appear as steps (servicing, logistic, financial etc.) of the business processes. that is, the work performed at the service provider is connected to the internal activities of the customer.
- Incentives are reasonably built in the regulatory system. It is important to provide for the fair nature of the procedure, for example, the penalty must be proportionate with the loss caused. Proactive measuring is efficient only in case both parties are stimulated to take action, that is, the necessary intervention is made in time.
- Technology offers a number of opportunities for the automation of the regulatory system, and the development of various safety /alarming systems. Naturally, it should be utilized reasonably.

The partners can do a lot to increase the value to be distributed. To achieve this goal parties should treat each other not only as enemies but as allies as well. In their relationship the struggle and alliance are present simultaneously. The latter one is based on confidence, and it is necessary to clarify the goals and expectations to create confidence and a transparent and objective system of regulations, and naturally, continuous and targeted communications are also essential.

Confidence is significant from other aspects as well. An in-depth, manifold IT outsourcing relation is a complicated and risky area, that stimulates all the affected parties to build high and strong legal walls around themselves. Legal arguments may prevail over business aspects, which is absolutely unhealthy. The stronger is the confidence between the parties, the less they need protective walls and the more they can utilize market relations replacing the hierarchy. (14)

References

1. Hamel, G.–Prahald, C.: Competing for the Future. Harvard Business School Press, 1994.
2. Bőgel György: Nyereségközpont, üzletág, divízió. Kossuth Könyvkiadó, 1999.
3. Montgomery, D.–Yip, G.: The Challenge of Global Customer Management. Stanford University, research paper # 1619
4. Hagel, J.–Brown, S.: Your Next IT Strategy. Harvad Business Review, 2001. október
5. Kerstette, J.: Software Shakeout. Business Week, 2001. március 5.
6. Bőgel György: Buddha mosolyog. CEO, 2001. 5. sz.
7. Hammer, M.–Champy, J.: Reengineering the Corporation. Nicholas Brealey Publishing, 1993
8. Tenner A. de Toro, I.: BPR – Vállalati folyamatok újraformálása. Műszaki Könyvkiadó, 1998
9. Porte, M.: Competitie Advantage. The Free Press, 1985.
10. Varian, H.–Shapiro, C.: Information Rules. Harvard Business Scholl Press, 1999.
11. Barthélemy, J.: The Hidden Costs of IT Outsourcing. MIZ Sloan Management Review, 2001. tavasz
12. Bendor-Samuel, P.: Turning Lead into Gold. Executive Excellence Publishing, 2000.
13. Sturm, R.–Morris, W.–Jander, M.: Foundations of Service Level Management. Sams, 2000.
14. Fukuyama, F.: Bizalom. Európa Könyvkiadó, 1997.

Tartalom

Editorial Office
Scientific Association for Infocommunications
HTE Budapest, V., Kossuth Lajos tér 6–8.
Phone: 00 36 353 1027 • Fax: 00 36 353 0451
e-mail: hte@mtesz.hu

Advertisement rates
1/1 (205 x 290 mm) 4C 120 000 Ft + áfa
Borító 3 (205 x 290 mm) 4C 180 000 Ft + áfa
Borító 4 (205 x 290 mm) 4C 240 000 Ft + áfa

Subscription rates for 2002
12 issues 150 USD, single copies 15 USD

Articles can be sent also to the following address
BME Department of Broadband Infocommunication System
Budapest, XI., Goldmann Gy. tér 3.
Phone: 00 36 463 1559 • Fax: 00 36 463 3289
e-mail: zombory@mht.bme.hu

Subscription
Scientific Association for Infocommunications
HTE Budapest, V., Kossuth Lajos tér 6–8.
Phone: 00 36 353 1027 • Fax: 00 36 353 0451
e-mail: hte@mtesz.hu

www.hte.hu

Publisher: MÁRIA MÁTÉ
Manager: András Dankó

Design by: Kocsis és Szabó Kft.
HU ISSN 0018-2028

Printed by: Regiszter Kft.


SAMSUNG
Digit Álomkép
©2002 Samsung Electronics Co. Ltd.
SAMSUNG
MARK
SAMSUNG
T100 TFT-LCD
Samsung T100, színes TFT kijelzővel. Valódi műremek.
Westel
a Kapcsolat
TFT colour
Élethű színes TFT-LCD kijelző
16
16 szólamú csengőhang
colour GAME
Színes játékok
Letölthető csengőhang és háttérkép
SAMSUNG DIGITall
everyone's invited
www.samsung.hu