JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

4/2003

Military communications technologies

Special issue edited by Wojciech Burakowski, Paul Kennedy, and Marek Amanowicz

Editorial Board

Editorial Advisory Board

ISSN 1509-4553

Circulation: 300 copies

Sowa - Druk na życzenie, www.sowadruk.pl, tel. 022 431-81-40

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

Preface

Recent world wide events have clearly demonstrated that the military response required to the changing threat has to be dynamic, and this means the any military communications capability has to be able to evolve and keep pace with these changing needs. Also, over the past years there has been a merging of the technologies that are now required by both the military to keep ahead of the threat and the commercial world where consumer demands for greater mobility and access require new and evolving technologies. The need for advanced technologies for military communications was a important feature of the 4th NATO Regional Conference on Military Communication and Information Systems that was held at the Military Communication Institute in Zegrze, Poland, between the 9th and 11th October 2002. This conference was arranged by the Military Communication Institute, the NATO C3 Agency and the Military University of Technology (Warsaw), and is formally endorsed by the First Deputy Minister of National Defence of the Republic of Poland and the NATO Assistant Secretary General for Defence Support, and this journal includes 11 papers from the conference that are seen to have a wide application and convey novel and informative approaches to the advances in new communications technologies.

The main theme of this year's conference was *The CIS (Command, Control, Communication and Information System) challenge in the new environment*, and reflecting the current NATO operational priorities there was some emphasis on mobility, interoperability and standards. Over 60 papers were presented in two parallel tracks, and a further 20 were available in a poster session where attendees could discuss in depth specific topics with the authors. The subject matter of the presentations ranged from academic papers on modelling and simulation, sensor processing technologies (e.g. the use of neural networks for the identification of vehicles from their sound signature) and military CIS network topology to papers on specific CIS requirements and the national programmes to support these requirements. However, whatever application is developed to support the military planning and command chain, an essential need is to be able to reliably and securely exchange vital information and to disseminate the orders and operational and tactical information. This requires robust and novel communications technologies, and the selected papers are aimed at providing an insight into this.

With so many excellent papers it has been difficult to single out any particular contribution for special treatment, but space does not permit more that 10 to be published here. To assist in this selection in addition to quality the chosen theme for this year's special publication is "Military communications technologies", and the selected papers are intended to give the broadest possible coverage of this theme from the papers that were presented. In every case these papers suggest either a new approach to a problem, or give a different and innovative perspective that can be used to improve the ability of NATO alliance partners to interoperate over secure and resilient communication and information system networks. Many of these papers have been drawn from the presentations given by members of the Military University in Warsaw, and the Military Communication Institute in Zegrze, and reflect the high quality and innovative nature of the research that is undertaken by these internationally respected institutions. Included also are a representative number of papers from other NATO nations and agencies.

The specific topics covered in the compendium of papers attempts to address many of the issues that now face the military communications and information systems community. The greater emphasis on, for example, mobility is placing increased demands of transmission technologies to provide improved reliability, wider bandwidth and more data and communications protection. This places increased demands on secure cryptographic technologies and transmission security, and therefore the papers on the *Error probability and error stream properties in channel with slow Rician fading* and *Remarks on improved inversion attacks on nonlinear filter generators* are recommended reading. Quality of service is a critical factor in reliable military networks, and with the introduction of IP based networks the management of reliability and assurance of delivery is of growing importance. A paper has, therefore, been included that addresses the comparison of QoS issues in ATM and IP network. Traditionally the method of assuring interoperability was through the use of agreed data standards, but with more nations joining NATO and the difference in national procurement and modernisation cycles, a means of exchanging data between systems that use different protocols is needed. NATO has been developing such a system and this is described. Finally, in addition to the provision of a secure and reliable information network, the military has a need for efficient command and control applications that do not place an unacceptable burden on valuable bandwidth. These include the co-ordination of battlefield deployment and logistic planning and assessment and simulation tools (such as MoD-SAF), and papers are also included that support such concepts; in particular technique that are optimised for use over military networks.

In the limited space available it is not possible to include all the papers from the RCMCIS'2002 conference, and in selecting only a few it is recognised that justice has not been done to the remainder. However, the full proceedings of the conference are available from the Military Communication Institute, Zegrze.

Guest Editors: Wojciech Burakowski, Paul Kennedy, and Marek Amanowicz

Paper

Error probability and error stream properties in channel with slow Rician fading

Krystyna M. Noga

Abstract — In a radio communication channel wave parameters fluctuate randomly. The signal envelope undergoes deep fades. When binary information is transmitted through such a channel, fading causes random variation of probabilities of error associated with the detection of individual elementary signals, which produces a clustering of errors. The paper presents an analytical description of the probability of bit error in the channel with very slow Rician fading and Gaussian noise for noncoherent and coherent detection. Digital systems employing error detection or error correction coding are generally based on the transmission of blocks of *N* sequential bits. Expressions are given for the probability of *n* errors occurring in *N* bits (weighted spectrum of errors) and the probability of more than *n* errors in a block of *N* bits (block error probability) for noncoherent frequency shift keying (NCFSK). Also the calculations are presented graphically.

Keywords — Rician fading channels, multipath propagation, bit error probability, stream of errors, weighted spectrum of errors, block error probability.

1. Introduction

Transmission of signal in digital radio communication systems takes place in the presence of random additive and random multiplicative disturbances; the multiplicative disturbances called ordinary as fading. We assume that the additive disturbances are represented by white Gaussian noise with zero mean value. The fading is considered as nonfrequency selective. This is valid for most cases of mobile data communications with moderate bit rates. The fading process is assumed to be stationary and slowly varying compared with the N bits duration; it is constant during data block duration. We assume that the fading is described by the Rician distribution. It is one of the double-parameter distribution of the signal envelope allowing to describe propagation conditions existing in radio channel in greater detail than a simpler and more frequently applied single parameter Rayleigh distribution. The Rayleigh and Rician distribution are only special case solutions of the random vector problem. The Nakagami distribution provides a more general solution.

The Rician distribution is an analytical model sufficient for a channel where the useful signal $s(t)$ is a sum of the stationary diffuse Gausssian signal $x(t) = a_x \cos[\omega_0 t + \varphi_x(t)]$ with zero mean value and the direct harmonic signal $A\cos_0 t$, i.e.:

$$s(t) = A\cos\omega_0 t + a_x \cos\left[\omega_0 t + \varphi_x(t)\right] = \\ = r\cos\left[\omega_0 t + \varphi_s(t)\right]. \quad (1)$$

The Rician model covers the superposition of a random Rayleigh signal with the fixed nonrandom signal. The Rician distribution can also be closely approximated by the Nakagami distribution. The Rician fading model applies in microcellular and satellite radio communication channels.

In radio communication channel radio waves parameters fluctuate randomly. Data transmission from and to mobile terminals suffers from fading effects caused by multipath propagation. The signal envelope undergoes deep fades. When binary information is transmitted through such a channel, fading causes random variation of probabilities of error associated with the detection of individual elementary signals. Deep fades cause bursts of bit errors in the transmitted data, i.e. errors in digital transmission over fading channels occur in bursts. In digital communication an important quantity is the bit error probability.

This paper presents an expression for the average probability of bit error for binary transmission in Rician channel with noncoherent frequency shift keying (NCFSK), differentially coherent phase shift keying (DPSK), coherent frequency shift keying (CFSK) and coherent phase shift keying (CPSK). Formulas for the average weighted spectrum of errors and the average block error probability for NCFSK are also presented.

2. Average probability of bit error

The static bit error probability for several common binary modulation schemes with optimum detection of nonfading signals in Gaussian noise is given by the formula

$$P_s(\rho) = \begin{cases} \dfrac{\exp(-\alpha\rho)}{2} & \text{for NCFSK, DPSK} \\ \dfrac{\operatorname{erfc}(\sqrt{\alpha\rho})}{2} & \text{for CFSK, CPSK} \end{cases} \quad (2)$$

where: ρ is the instantaneous signal-to-noise power ratio (SNR), $\operatorname{erfc}(\sqrt{x})$ denotes the error function [4], $\alpha = 0.5$ for NCFSK, CFSK and $\alpha = 1$ for DPSK, CPSK.

Since in the Eq. (2) the argument of the error function appears in the lower limit of the integral, it is analytically

difficult to perform averages of this equation. Another form for static bit error probability is presented in [2]:

$$P_s(\rho) = \frac{(a\rho)^b}{\Gamma(b)} \int_0^{\Pi/2} \frac{\cos\varphi}{(\sin\varphi)^{2b+1}} \exp\left(\frac{-a\rho}{\sin^2\varphi}\right) \mathrm{d}\varphi\,, \qquad (3)$$

where a and b are the coefficients which depend on the particular form of modulation and detection, i.e. $a = b = 0.5$ for CFSK, $a = 0.5$, $b = 1$ for NCFSK, $a = 1$, $b = 0.5$ for CPSK, $a = b = 1$ for DPSK.

Assuming Rician fading, the envelope r of the useful signal $s(t)$, described by Eq. (1), has the probability density function:

$$p(r) = \frac{r}{\sigma_x^2} \exp\left(-\frac{r^2+A^2}{2\sigma_x^2}\right) I_0\left(\frac{Ar}{\sigma_x^2}\right); \quad r \geq 0\,, \qquad (4)$$

where: A is the envelope of the direct component of the useful signal $s(t)$, σ_x^2 is the variance of the diffused component $x(t)$ of the useful signal $s(t)$, $I_0(\alpha)$ is the zero order modified Bessel function of the first kind [4].

Probability density function of the square of the envelope $r(t)$ of the useful signal is given by the formula:

$$p(r^2) = \frac{1}{2\sigma_x^2} \exp\left(-\frac{r^2+A^2}{2\sigma_x^2}\right) I_0\left(\frac{Ar}{\sigma_x^2}\right); \quad r \geq 0\,. \qquad (5)$$

The average value of the square of the envelope $r(t)$ can be written as

$$E(r^2) = A^2 + 2\sigma_x^2\,, \qquad (6)$$

where $E(x)$ denotes the expected value of the argument.

From Eq. (5) we can describe the distribution of random variable, which represents ρ defined as the instantaneous ratio power of the useful signal to average power N_0 of the additive Gaussian noise, i.e. the instantaneous signal-to-noise power ratio:

$$\rho = \frac{r^2}{2N_0}\,. \qquad (7)$$

The probability density function of ρ in Rician channel is denoted by the formula:

$$p(\rho) = \frac{N_0}{\sigma_x^2} \exp\left(-\frac{2N_0\rho+A^2}{2\sigma_x^2}\right) I_0\left(\frac{A\sqrt{2N_0\rho}}{\sigma_x^2}\right); \quad \rho \geq 0\,. \qquad (8)$$

With sufficiently slow fading, the average probability of bit error $P_D(A, \sigma_x, N_0)$, i.e. the dynamic probability of bit error, equals to $P_s(\rho)$ averaged over the distribution of SNR ρ:

$$P_D(A, \sigma_x, N_0) = \int_0^{\infty} P_s(\rho)\, p(\rho)\, \mathrm{d}\rho = E\left[P_s(\rho)\right]. \qquad (9)$$

Inserting Eqs. (2) and (8) into (9) gives the average probability of bit error in Rician channel [5]:

$$P_D(A, \sigma_x, N_0) = \frac{1}{2\Gamma(b)} \exp\left(\frac{-A^2}{2\sigma_x^2}\right) \times \times \sum_{k=0}^{\infty} \left(\frac{A^2}{2\sigma_x^2}\right)^k \frac{\Gamma(b+k+1)}{(k!)^2} B_g(k+1, b)\,, \qquad (10)$$

where: $\Gamma(b)$ is the gamma function [4], $g = \frac{N_0}{a\sigma_x^2 + N_0}$ and $B_g(x, y)$ is the incomplete beta function [4].

Let us introduce additional factors, which describe Rician channel and can be defined as

$$\rho_1 = \frac{A^2}{2\sigma_x^2}; \quad \rho_2 = \frac{\sigma_x^2}{N_0}; \quad \rho_3 = \frac{A^2}{2N_0} = \rho_1\rho_2\,. \qquad (11)$$

Then, the expected value of the SNR ρ we can express as

$$E(\rho) = \rho_0 = \rho_3 + \rho_2\,. \qquad (12)$$

In literature the factor ρ_1 (direct signal to diffused signal power ratio) sometimes is denoted as K and ρ_0 (average SNR) is denoted as Γ [1].

In the end, the formula for the average probability of bit error in Rician channel can be written as

$$P_D(\rho_1, \rho_2) = = \frac{1}{2\Gamma(b)} \exp(-\rho_1) \sum_{k=0}^{\infty} \rho_1^k \frac{\Gamma(b+k+1)}{(k!)^2} B_{g1}(k+1, b)\,, \qquad (13)$$

where $g1 = \frac{1}{a\rho_2+1}$ or using the relation (12)

$$P_D(\rho_1, \rho_0) = \frac{1}{2\Gamma(b)} \exp\left(-\rho_1\right) \times \times \sum_{k=0}^{\infty} \rho_1^k \frac{\Gamma(b+k+1)}{(k!)^2} B_{g2}(k+1, b)\,, \qquad (14)$$

where $g2 = \frac{1+\rho_1}{a\rho_0+1+\rho_1}$.

The error-rate formulas (10), (13), (14) are valid for four principal cases of binary modulation in Rician channels. The result over a Rayleigh channel can be obtained from our results when $A = 0$. However, in the border case when $\rho_2 = 0$, we have the formula for bit error probability in channel without fading. Figure 1 presents the average bit error probability for binary NCFSK modulation in Rician channel as a function of ρ_0 for various values ρ_1.

For detection in the noncoherent systems (when $b = 1$) the formula for the average probability of bit error in Rician channel can be expressed as

$$P_D(A, \sigma_x, N_0) = \frac{N_0}{2(\alpha\sigma_x^2 + N_0)} \exp\left[\frac{-A^2\alpha}{2(\alpha\sigma_x^2 + N_0)}\right] \qquad (15)$$

or

$$P_D(\rho_1, \rho_2) = \frac{1}{2(\alpha\rho_2 + 1)} \exp\left(-\frac{\alpha\rho_1\rho_2}{\alpha\rho_2 + 1}\right) \qquad (16)$$

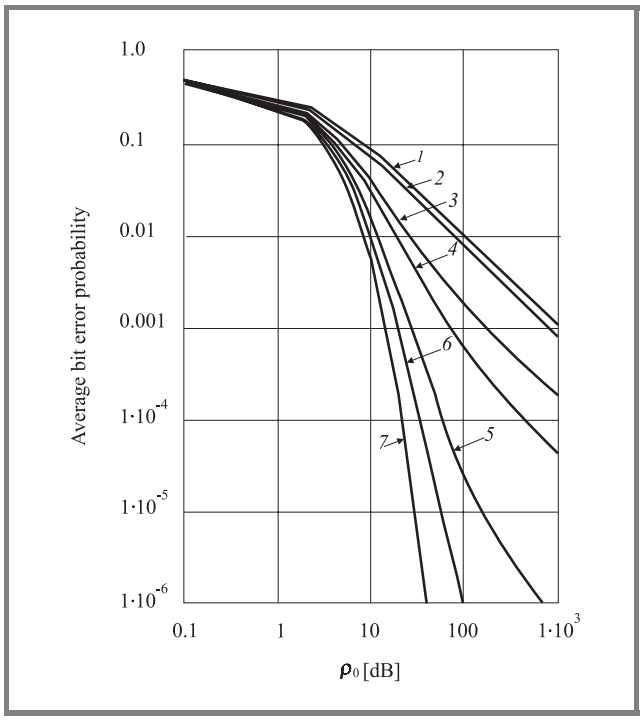


Fig. 1. The average bit error probability for NCFSK in channel with Rician fading: *1* – $\rho_1 = -10$; *2* – $\rho_1 = 0$; *3* – $\rho_1 = 5$; *4* – $\rho_1 = 7$; *5* – $\rho_1 = 10$; *6* – $\rho_1 = 12$; *7* – $\rho_1 = 16$ (*all values in dB*).

and after using the relation (12) we can rewrite it as

$$P_D(\rho_1,\rho_0) = \frac{1+\rho_1}{2(\alpha\rho_0+\rho_1+1)} \exp\left(-\frac{\alpha\rho_1\rho_0}{\alpha\rho_0+\rho_1+1}\right) \tag{17}$$

where $\alpha = 0.5$ for NCFSK and $\alpha = 1$ for DPSK.

3. Average weighted spectrum of errors

In case of digital transmission over a fading channel, time variation causes the change of bit error probability with the effect of clustering errors in the received signal. Forward error correction is often used to break up the clustering [3, 7, 8]. Digital systems employing error detection or error correction coding are generally based on the transmission of blocks of N bits. The average probability of bit error specified by Eq. (10) neither describes the number of errors nor their placement in the stream of errors. In a communication system which transmits data in blocks on N bits the probability of n errors in a block and the probability of more than n errors in a block are an important quantities which describe the stream of errors in channel with fading.

The weighted spectrum of errors, i.e., the probability of n errors occurring in a transmission of N bits, for independent bit errors is given by the binomial distribution

$$P(L_N = n) = \binom{N}{n} P_s^n(\rho)\,[1-P_s(\rho)]^{N-n} \qquad n = 0, 1, \ldots, N. \tag{18}$$

The average probability of n errors in a block of N bits (average weighted spectrum of errors) is denoted as [5, 6]

$$P_D(L_N = n) = \int_0^\infty P(L_N = n)\,p(\rho)\,\mathrm{d}\rho =$$

$$= \binom{N}{n} \sum_{j=0}^{N-n} \binom{N-n}{j} (-1)^j E\left[P_s^{j+n}(\rho)\right]. \tag{19}$$

The average in Eq. (19) is formed over the instantaneous SNR ρ which has the probability density function $p(\rho)$ described by Eq. (8).

Assuming that the Rician fading is very slow, nonselective and independent, then the instantaneous SNR remains the same over a block of N bits. For noncoherent FSK in channel with Rician fading the average weighted spectrum of errors is given by

$$P_D(L_N = n) = 2\binom{N}{n} N_0 \sum_{i=0}^{N-n} \binom{N-n}{i} (-1)^i \times$$

$$\times \left(\frac{1}{2}\right)^{i+n} \frac{1}{(n+i)\sigma_x^2+2N_0} \exp\left(\frac{-A^2(n+i)}{2\left[\sigma_x^2(n+i)+2N_0\right]}\right). \tag{20}$$

Using the relations (11) and (12) we can rewrite formula (20) as

$$P_D(L_N = n) = 2\binom{N}{n} \sum_{i=0}^{N-n} \binom{N-n}{i} (-1)^i \left(\frac{1}{2}\right)^{i+n} \times$$

$$\times \frac{1+\rho_1}{(n+i)\rho_0+2(1+\rho_1)} \exp\left(\frac{-\rho_1\rho_0(n+i)}{\rho_0(n+i)+2(1+\rho_1)}\right). \tag{21}$$

Equation (21) has been used to calculate the average probability of n errors in N bits. Figure 2 shows the $P_D(L_N = n)$ for $N = 30$ and for various values of ρ_1 and ρ_0; the values of ρ_1 and ρ_0 are as taken as $P_D(\rho_1,\rho_0) = 10^{-3}$.

Additional, Fig. 3 shows the average weighted spectrum of errors in N bits for NCFSK for different values of N, ρ_1 and ρ_0.

The probability $P_D(L_N = n)$ takes into account only the total number of errors and disregards their distribution. It is useful only for the performance evaluation of random

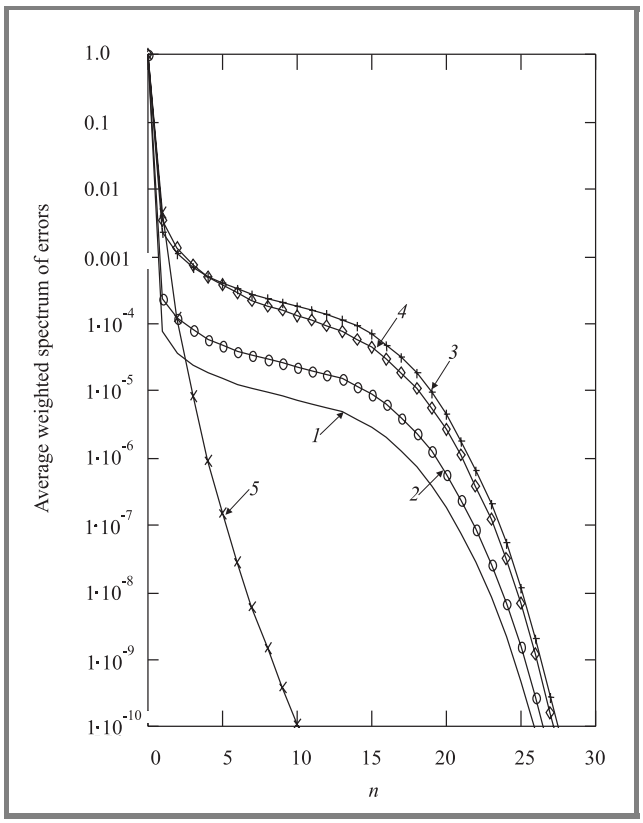


Fig. 2. Average probability of n errors in $N = 30$ bits for NCFSK: *1* – $\rho_1 = 0$, $\rho_0 = 43$; *2* – $\rho_1 = 1$, $\rho_0 = 38$; *3* – $\rho_1 = 5$, $\rho_0 = 23$; *4* – $\rho_1 = 7$, $\rho_0 = 19$; *5* – $\rho_1 = 16$, $\rho_0 = 13$ (*all values in dB*).

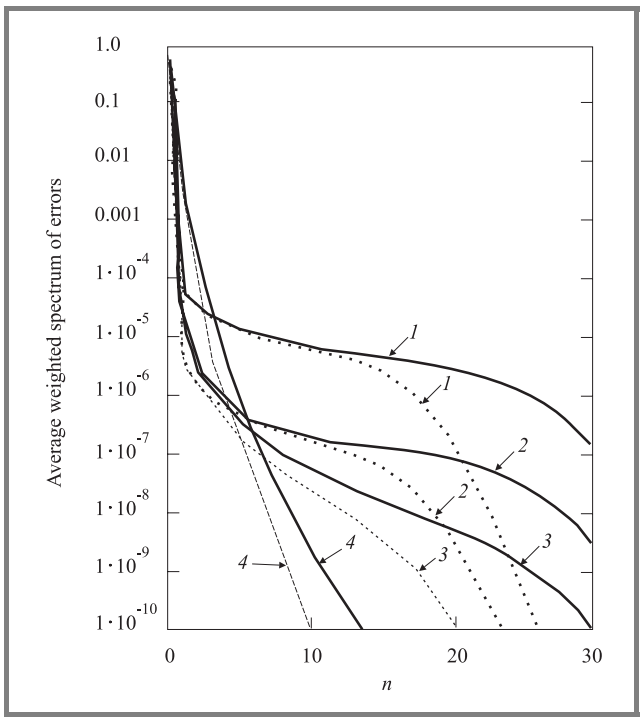


Fig. 3. Average probability of n errors in N bits for NCFSK; solid line $N = 48$, dot line $N = 30$: *1* – $\rho_1 = 0$, $\rho_0 = 43$; *2* – $\rho_1 = 10$, $\rho_0 = 28$; *3* – $\rho_1 = 12$, $\rho_0 = 20$; *4* – $\rho_1 = 16$, $\rho_0 = 13$ (*all values in dB*).

error-correcting codes. It cannot be used for burst-error-correcting codes. If the random error-correcting code is used, with the code capable of correcting up to n random errors, then the probability of correct decoding is $\sum_{i=0}^{n} P_D(L_N = i)$.

4. Average block error probability

When the system with burst-error correcting code is used the probability of correct decoding cannot be expressed only in terms of $P_D(L_N = n)$. The burst-error-correction code can correct all error vectors with length less than or equal to n. In this case the important quantity is the average probability of more than n errors in a block of N bits, i.e., average block error probability. It is denoted by

$$P_D(L_N > n) =$$

$$= 1 - \sum_{i=0}^{n} \binom{N}{i} \int_0^{\infty} P_s^i(\rho)\,[1 - P_s(\rho)]^{N-i}\, p(\rho)\mathrm{d}\rho =$$

$$= 1 - \sum_{i=0}^{n} \binom{N}{i} \sum_{j=0}^{N-i} \binom{N-i}{j} (-1)^j E\left[P_s^{j+i}(\rho)\right]. \quad (22)$$

Thus, from Eqs. (22), (2), (8), the average block error probability for NCFSK we can expressed as

$$P_D(L_N > n) = 1 - 2N_0 \sum_{j=0}^{n} \binom{N}{j} \sum_{i=0}^{N-j} \binom{N-j}{i} (-1)^i \times$$

$$\times \left(\frac{1}{2}\right)^{i+j} \frac{1}{(j+i)\sigma_x^2 + 2N_0} \exp\left(\frac{-A^2(j+i)}{2[\sigma_x^2(j+i) + 2N_0]}\right) \quad (23)$$

or in the form

$$P_D(L_N > n) = 1 - 2\sum_{j=0}^{n} \binom{N}{j} \sum_{i=0}^{N-j} \binom{N-j}{i} (-1)^i \left(\frac{1}{2}\right)^{i+j} \times$$

$$\times \frac{1+\rho_1}{(j+i)\rho_0 + 2(1+\rho_1)} \exp\left(\frac{-\rho_1\rho_0(j+i)}{\rho_0(j+i) + 2(1+\rho_1)}\right). \quad (24)$$

Figure 4 shows the average block error probability $P_D(L_{30} > 0)$ for NCFSK, i.e., the probability of at least one error in $N = 30$ bits.

When we plotted the expression (4) presented in references [1] for $D = 1$, i.e. for no diversity case, we become the same result as is shown in Fig. 4.

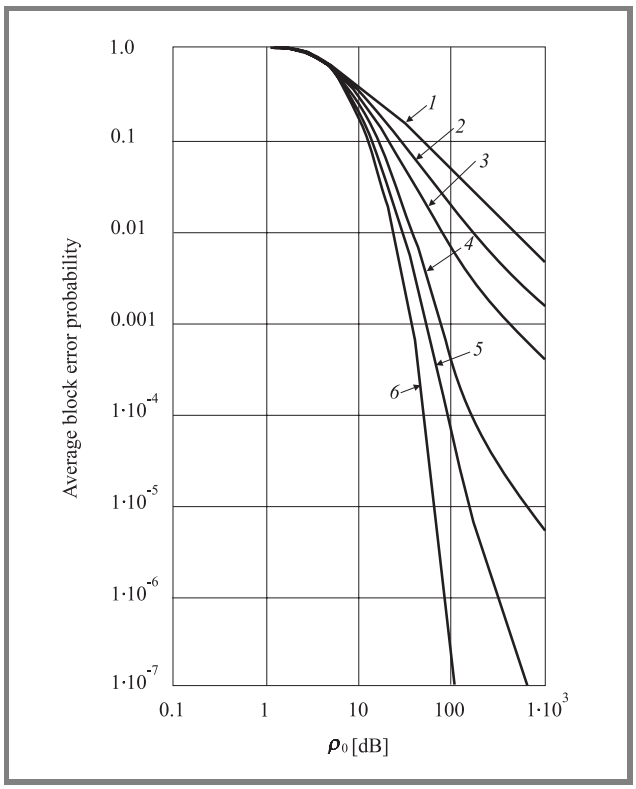


Fig. 4. Average block error probability $P_D(L_{30} > 0)$ for NCFSK and $N = 30$: *1* – $\rho_1 = 1$; *2* – $\rho_1 = 5$; *3* – $\rho_1 = 7$; *4* – $\rho_1 = 10$; *5* – $\rho_1 = 12$; *6* – $\rho_1 = 14$ (*all values in dB*).

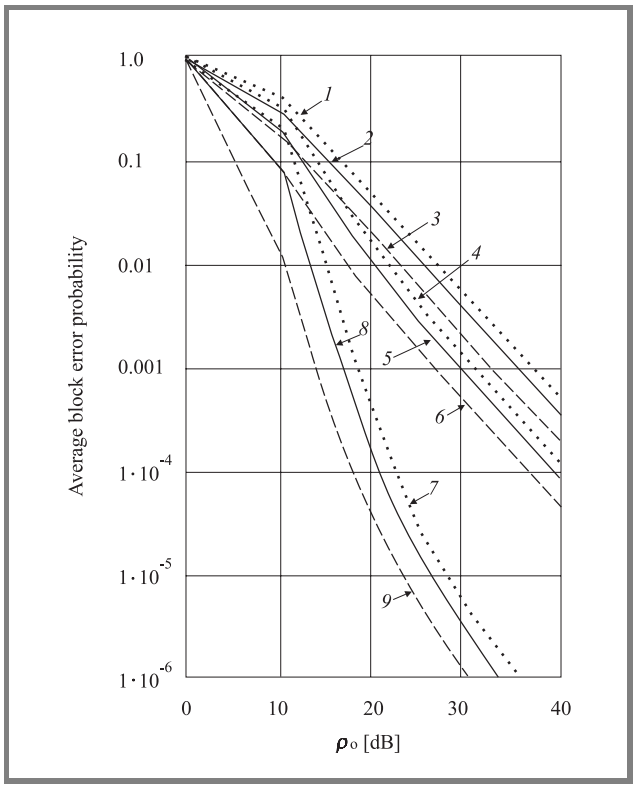


Fig. 5. Average block error probability $P_D(L_{30} > n)$ for NCFSK and $N = 30$: *1* – $\rho_1 = 0$, $n = 0$; *2* – $\rho_1 = 0$, $n = 1$; *3* – $\rho_1 = 0$, $n = 4$; *4* – $\rho_1 = 5$, $n = 0$; *5* – $\rho_1 = 5$, $n = 1$; *6* – $\rho_1 = 5$, $n = 4$; *7* – $\rho_1 = 10$, $n = 0$; *8* – $\rho_1 = 10$, $n = 1$; *9* – $\rho_1 = 10$, $n = 4$ (*all values in dB*).

Also Fig. 5 shows the average block error probability $P_D(L_{30} > 0)$ in a block of $N = 30$ bits for NCFSK and for various values of ρ_1 and n.

5. Conclusion

The presented equations are valid for nonselective, very slow and independent Rician fading. Since the presented expressions include results for the cases of Rayleigh fading and no fading, they can be widely used to evaluate the performance of error control techniques for mobile radio.

Presented results can be useful for error detection or error correction coding. In a communication system that transmits data in blocks of N bits an important quantity is the probability of more than n errors in block. If a simple automatic repeat request scheme is used, the throughput can be determined from $P_D(L_N > 0)$, i.e. from the probability of at least one error in blocks. However, if the use of forward error correction is to be investigated, the knowledge of $P_D(L_N > n)$ is required. In case of error detection a block is received correctly only if all N bits are received without error.

In Figs. 1 – 5 numerical results are presented, the influence of fading for error detection is presented. The obtained expression can easily be programmed using standard mathematical software package such as Mathcad.

References

[1] F. Adachi and K. Ohno, "Block error probability for noncoherent FSK with diversity reception in mobile radio", *Electron. Lett.*, vol. 24, no. 24, pp. 1523–1525, 1988.

[2] M. S. Alouini and M. K. Simon, "Generic form for average error probability of binary signals over fading channels", *Electron. Lett.*, vol. 34, no. 10, pp. 949–950, 1998.

[3] R. R. Eaves and A. H. Levesque, "Probability of block error for very slow Rayleigh fading in Gaussian noise", *IEEE Trans. Commun.*, vol. COM-25, no. 3, pp. 368–374, 1977.

[4] S. Gradstein and I. M. Ryzhik, *Tables of Integrals, Sums, Series and Products*. London, 1964.

[5] K. Noga, "Errors stream properties for binary transmission in channel with fading", in *Proc. Nat. Conf. Radio Diff. Radio Commun.*, Poznań, Poland, 1998, pp. 117–120.

[6] K. Noga, "The performance of binary transmission in slow Nakagami fading channels with MRC diversity", *IEEE Trans. Commun.*, vol. 46, no. 7, pp. 863–865, 1998.

[7] A. Seyoum and N. C. Beaulieu, "Semianalytical simulation for evaluation of block – error rates of fading channels", *IEEE Trans. Commun.*, vol. 46, no. 7, pp. 916–920, 1998.

[8] E. Sundberg, "Block error probability for noncoherent FSK with diversity for very slow Rayleigh fading in Gaussian noise", *IEEE Trans. Commun.*, vol. COM-29, no. 1, pp. 57–60, 1981.

Krystyna Maria Noga was born in Gdańsk, Poland. She received the M.S. degree in electrical engineering from the University of Technology in Gdańsk, in 1977 and Ph.D. degree in electronics engineering from the University of Technology in Gdańsk, in 1987. Since 1978 she has been employing in the Institute of Ship Automation in Gdynia Maritime University, Poland. Her Ph.D. thesis was entitled "Probabilistic characteristics of binary transmission in radio communications channel with slowly fading for single and diversity systems". Her research interest and technical experience include the analysis of wireless digital communications over fading channels, mobile channel modeling, multipath channels, diversity reception, signal detection and estimation, channel estimation, statistical signal processing. She is the author of numerous original research papers.

e-mail: jagat@vega.wsm.gdynia.pl
Institute of Ship Automation
Gdynia Maritime University
Morska st 81/87
81-225 Gdynia, Poland

Paper

Remarks on improved inversion attacks on nonlinear filter generators

Anna Górska and Karol Górski

Abstract — The subject of this paper are inversion attacks on stream ciphers (nonlinear filter generators), which were first introduced by Golić [3] and extended by Golić, Clark and Dawson [4]. These original attacks have computational complexity $O(2^M)$, where M is the so-called "memory size" – distance between outer taps to filter function. In [6] we have proposed improved inversion attacks which have computational complexity $O(2^{r-m})$, where r denotes the length of the shift register and m denotes the largest gap between cells with taps to filter function or to connection polynomial. In this paper we describe further extension of our previous results obtained by considering shifts of the feedback polynomial which maximise the largest gap between cells with taps to filter function or to connection polynomial. We show that the previously proposed set of design criteria [3, 6] does not prevent the new version of improved inversion attack and we propose an additional criterion based on the relationship between positions of taps to filter function and positions of taps to the multiples of the connection polynomial.

Keywords — stream cipher, shift register, nonlinear filter generator, inversion attack.

1. Introduction

Despite the growing importance of block ciphers, symmetric stream ciphers are still one of the fundamental tools in modern cryptography. Most designs are based on linear feedback shift registers (LFSR) combined by nonlinear boolean functions or filtered by nonlinear boolean functions (so-called nonlinear filter generators). Different variants exist: clock-controlled systems, multiplexed systems, memory combiners and decimated generators. Our work focuses on nonlinear filter generators (NFG) illustrated in Fig. 1. NFG can be used on its own [1] or as a building block in more complex generators.

Unfortunately there are no known, practical constructions of stream ciphers which offer unconditional security or provable computational security (the one time pad, an unconditionally secure stream cipher, cannot be regarded as practical). In practice, evaluation of the security of these ciphers is heuristic. Among the most powerful classes of attacks on stream ciphers which have to be considered are fast correlation attacks (beginning from [7]) and conditional correlation attacks [1]. The first class was initially used to attack combination generators but recently was successfully used to attack NFG [8]. In the second class best results were achieved by Golić [3] who introduced the inversion attacks (IA), which are the most powerful attacks on nonlinear filter generators. In the same paper Golić presented a set of design criteria for nonlinear filter generators which, when respected, should ensure large period, high linear complexity and good statistical properties of the output sequence as well as resistance to fast correlation attacks, conditional correlation attacks and inversion attacks.

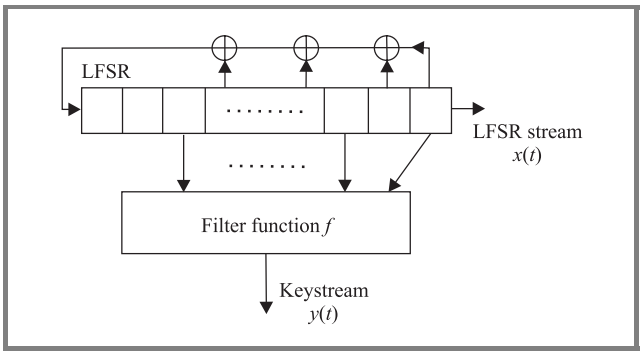


Fig. 1. Nonlinear filter generator.

In [6] we have introduced improved inversion attacks (IIA), which can have significantly lower computational complexity in comparison to basic inversion attacks. In this paper we propose an extension of this attack and a modification of the set of design criteria in order to prevent this new attack.

2. Notation and definitions

Let r be LFSR length, $n\,(n \le r)$ – denote the number of non-degenerate input variables to filter function $f(z_1, \ldots, z_n)$ and $\gamma = (\gamma_i)_{i=1\ldots n}$ denote the tapping sequence, an increasing sequence of integers specifying positions of inputs to filter function, such that $\gamma_1 = 0$ and $\gamma_n \le r-1$. Let $M = \gamma_n - \gamma_1$ denote the memory of the filter function. Let $x = (x(t))_{t=-r\ldots\infty}$ be a binary maximum-length sequence $(x = (x(t))_{t=-r+1\ldots 0}$ denotes LFSR initial state). Then the output sequence $y = (y(t))_{t=0\ldots\infty}$ is computed as:

$$y(t) = f\left(x(t-\gamma_1), \ldots, x(t-y_n)\right). \qquad (1)$$

In [3] it was proved that a filter function of one of the following forms:

$$f(z_1, \ldots, z_n) = z_1 \oplus g(z_2, \ldots, z_n) \tag{2}$$

or

$$f(z_1, \ldots, z_n) = z_n \oplus h(z_1, \ldots, z_{n-1}) \tag{3}$$

will produce (independently of the tapping sequence) a purely random output, given a purely random input, thus making the generator resistant to conditional correlation attacks.
In this case $y(t)$ takes the following form:

$$y(t) = x(t-\gamma_1) \oplus g(x(t-\gamma_2), \ldots, x(t-\gamma_n)), \tag{4}$$

or

$$y(t) = x(t-\gamma_n) \oplus h(x(t-\gamma_1), \ldots, x(t-\gamma_{n-1})). \tag{5}$$

Depending on the form of the function Golić proposed the forward inversion attack and the backward inversion attack, respectively. The average computational complexity of the attacks is $\boldsymbol{O}(2^{M-1})$ and the worst case complexity is $\boldsymbol{O}(2^M)$.

3. Inversion attack

The objective of the attack is to reconstruct the initial state of the LFSR, having a segment of keystream sequence, given the LFSR feedback polynomial, nonlinear filter function f and the tapping sequence γ. If the filter function is of the form (4) forward inversion attack is applied, which is given by the algorithm below.

Algorithm 1 (forward inversion attack):

1. Assume (not previously checked) M bits $\big(x(t)\big)_{t=-M\cdots-1}$ of unknown initial memory state.
2. By using (4), generate $\big(x(t)\big)_{t=r-M-1\ldots0}$ from a known segment $\big(y(t)\big)_{t=r-M-1\ldots0}$ of output sequence.
3. By using LFSR linear recursion, generate $\big(x(t)\big)_{t=r-M\ldots N-1}$ from first r bits of $\big(x(t)\big)_{t=-M\ldots r-M-1}$.
4. By using (1), compute $\big(y'(t)\big)_{t=r-M\ldots N-1}$ from $\big(x(t)\big)_{t=r-2M\ldots N-1}$ and compare with the known $\big(y(t)\big)_{t=r-M\ldots N-1}$. If they are the same then accept assumed initial memory state and stop. Otherwise go to 1.

When the filter function is of the form (5) backward inversion attack is applied, which is given by Algorithm 2.

Algorithm 2 (backward inversion attack):

1. Assume (not previously checked) M bits $\big(x(t)\big)_{t=-M-1\ldots0}$ of unknown initial memory state.
2. By using (5), generate $\big(x(t)\big)_{t=r-M-1\ldots0}$ from a known segment $\big(y(t)\big)_{t=r-M-1\ldots0}$ of output sequence.
3. By using LFSR linear recursion, generate $\big(x(t)\big)_{t=r-M\ldots N-1}$ from first r bits of $\big(x(t)\big)_{t=-M\ldots r-M-1}$.
4. By using (1), compute $\big(y'(t)\big)_{t=r-M\ldots N-1}$ from $\big(x(t)\big)_{t=r-2M\ldots N-1}$ and compare with the known $\big(y(t)\big)_{t=r-M\ldots N-1}$. If they are the same then accept assumed initial memory state and stop. Otherwise go to 1.

4. Improved inversion attack

The difference between the basic inversion attack and our first proposal of the improved inversion attack [6] relies on a modification of Steps 1 and 2. To make further improvement we will include an additional preprocessing phase in which we will find the largest gap between cells with taps to filter function and cells with taps to connection polynomial multiples (shifts of the polynomial). This idea was first suggested by Golić [5]. In Step 1 instead of guessing M bits of initial state we guess $r-m$ bits, where m denotes the size of the largest gap between cells of LFSR which have taps to filter function or to multiples of connection polynomial (when the largest such gap is between cells j and k, where $j<k$, then $m=k-j$). The average computational complexity of the improved attack is $\boldsymbol{O}(2^{r-m-1})$ and the worst case complexity is $\boldsymbol{O}(2^{r-m})$. The algorithm of the new attack is as follows.

Algorithm 3 (improved inversion attack with preprocessing phase):

1. (Preprocessing phase). Find the largest gap between cells of LFSR with taps to connection polynomial multiples and cells with taps to filter function. Denote the outer cells of the gap by k and $k-m$.

Rest of the attack is identical to improved inversion attack presented in [6]:

2. Assume (not previously checked) $r-m$ bits $\big(x(t)\big)_{t=-r+1\cdots-k-1,\,-k+m\ldots0}$ of unknown initial memory state.
3. By using (4), generate $\big(x(t)\big)_{t=-k\cdots-k+m-1}$ from a known segment $\big(y(t)\big)_{t=0\ldots m-1}$ of output sequence and the connection polynomial.
4. By using LFSR linear recursion, generate sequence $\big(x(t)\big)_{t=r-m\ldots N-1}$ from first r bits $\big(x(t)\big)_{t=-m\ldots r-m-1}$.
5. By using (1), compute $\big(y'(t)\big)_{t=r-m\ldots N-1}$ from $\big(x(t)\big)_{t=r-2m\ldots N-1}$ and compare with the known $\big(y(t)\big)_{t=r-m\ldots N-1}$. If they are the same then accept assumed initial memory state and stop. Otherwise go to 1.

We illustrate this attack by examples.

Example 1 (improved inversion attack (IIA) – Figs. 2 and 3). Let the connection polynomial[1] be:

$$p(x) = x^{130} + x^{63} + x^{56} + x^{35} + x^{28} + \\ + x^{21} + x^7 + x^3 + 1,$$

the tapping sequence: $\gamma = (127, 63, 31, 15, 7, 3, 1, 0)$, and let the filter function be linear in the last variable. (If the filter function would not be linear in any variable, which delimits the largest gap, we should apply an inversion attack with branching [4]). In the basic inversion attack the cryptanalyst needs to guess 127 bits, which gives computational complexity of $\boldsymbol{O}(2^{127})$ and makes the attack infeasible. In the improved version of the attack we only need to guess 66 bits $\big(x(t)\big)_{t=-129,-128,-63,\ldots,0}$ which gives the expected attack runtime of 2^{66} steps.
Let us describe how our attack works in Step 2.

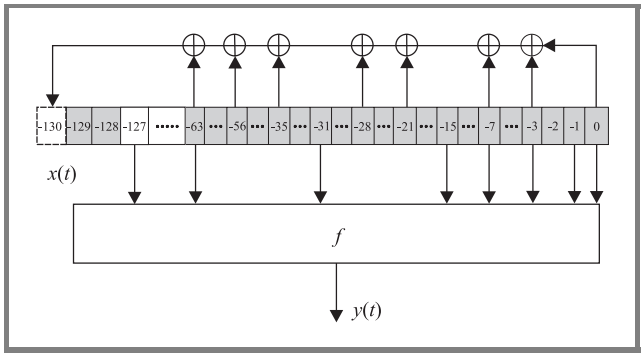


Fig. 2. Improved inversion attack on NFG (guessed (known) cells are filled with grey colour).

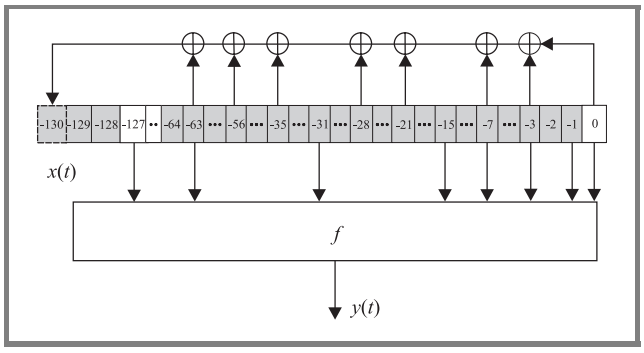


Fig. 3. Improved inversion attack on NFG cont. (guessed (known) cells are filled with grey colour).

First we calculate (identically as in IA) $x(-127)$:

$$x(-127) = y(63) \oplus g\big(x(0), x(-1), x(-3), x(-7), \\ x(-15), x(-31), x(-63)\big), \qquad (6)$$

[1]This feedback polynomial is very sparse, chosen to simplify the example.

then we clock backward the register state (so the content of cell i moves to cell $i-1$ and, after clocking we know $\big(x(t)\big)_{t=-130,-129,-128,-64\cdots-1}$).
We can calculate $x(0)$ from the connection polynomial:

$$x(0) = x(-3) \oplus x(-7) \oplus x(-21) \oplus x(-28) \oplus x(-35) \oplus \\ \oplus x(-56) \oplus x(-63) \oplus x(-130)$$

and then again calculate $x(-127)$ from the knowledge of output stream and filter function:

$$x(-127) = y(62) \oplus g\big(x(0), x(-1), x(-3), x(-7), \\ x(-15), x(-31), x(-63)\big).$$

Then again we clock the register state left (after which we know $\big(x(t)\big)_{t=-130,-129,-128,-65\cdots-1}$), calculate $x(0)$ from a connection polynomial, and so on. We continue this procedure until the LFSR state is reconstructed. Then we follow testing Steps 3 and 4.

Example 2 (improved inversion attack with preprocessing phase – Figs. 4 and 5). Let the connection polynomial be of the following form:

$$p(x) = x^{130} + x^{66} + x^{65} + x^{64} + x^{34} + x^{33} + x^{32} + x^{18} + \\ + x^{17} + x^{16} + x^{10} + x^9 + x^8 + x^6 + x^5 + x^3 + x^2 + x + 1,$$

the tapping sequence: $\gamma = (127, 63, 31, 15, 7, 3, 1, 0)$ identical to that in Example 1, and let again the filter function be linear in the last variable. The computational complexity of basic inversion attack is again $\boldsymbol{O}(2^{127})$, the complexity of IIA is $\boldsymbol{O}(2^{69})$ and the complexity of IIA with preprocessing phase is $\boldsymbol{O}(2^{66})$.

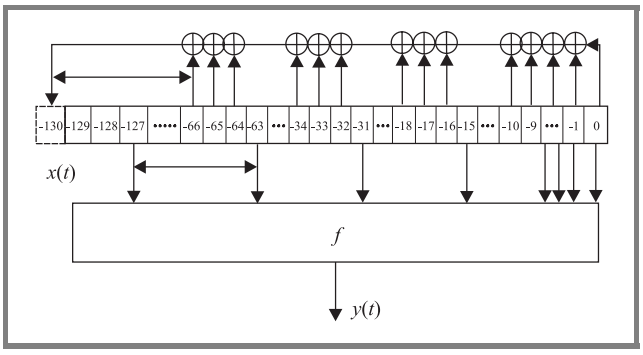


Fig. 4. Improved inversion attack with preprocessing phase on NFG.

In the preprocessing phase we find that the largest gap between taps to connection polynomial is between cells 130 and 66 and the length of that gap is equal to 64, similar as the gap between taps to filter function. The attack works as follows.
First we need to guess 66 bits $\big(x(t)\big)_{t=-63,\ldots,3}$, then we calculate $x(-127)$ from the filter function and known keystream segment:

$$x(-127) = y(66) \oplus g(x(0), x(-1), x(-3), x(-7), \\ x(-15), x(-31), x(-63)),$$

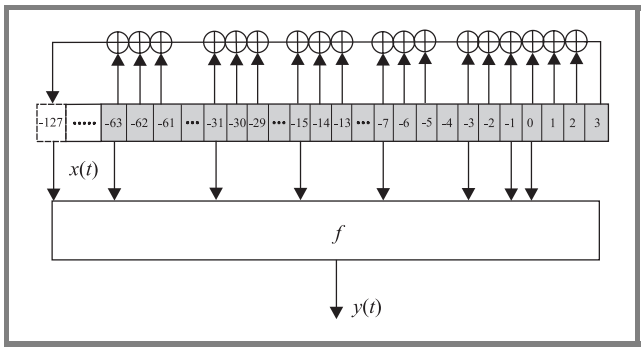


Fig. 5. Improved inversion attack with preprocessing phase on NFG (known cells are marked with grey colour).

then we clock backward the register state and, after clocking we know $\big(x(t)\big)_{t=--64\ldots2}$. We calculate $x(-127)$ from the filter function and known keystream segment:

$$x(-127) = y(65) \oplus g(x(0), x(-1), x(-3), x(-7),$$
$$x(-15), x(-31), x(-63)),$$

then we calculate $x(3)$ from the connection polynomial and so on.

5. Nonlinear filter generators design criteria

After introducing inversion attacks Golić [3] proposed a set of design criteria for NFG which were considered to ensure large period, high linear complexity and resistance to statistical attacks, inversion attacks, conditional correlation attacks and fast correlation attacks. To ensure properties from the first group (large period, high linear complexity) primitivity of connection polynomial and large algebraic order of function f are important. Good statistical properties can be ensured by the choice of filter function of the form (2) or (3).

Golić pointed out the fact that the computational complexity of inversion attack is exponential with the memory size M, rather than with the length of the register r. So, to make a cipher resistant to inversion attack he proposed to choose M as large as possible, preferably close to its maximum possible value $r-1$. Additionally, to avoid the possibility of effective reduction of memory size (by decimation technique), the tapping sequence should not be equidistant, preferably the greatest common divisor of elements of γ should be equal to one (assuming $\gamma_1 = 0$).

Resistance to conditional correlation attacks requires the number of nondegenerate inputs to f to be large enough, and the γ sequence chosen according to a full or λ-order positive difference set (with λ as small as possible for given n and r) and correlation immunity of f to be relatively large compared to λ.

To prevent fast correlation attack designers should ensure that the nonzero correlation coefficients of f to the set of linear functions are relatively small and close in magnitude. Finally, the number of nonzero terms in the feedback polynomial and in any of its low degree multiples should not be small.

The polynomial, tapping sequence and filter function used in Example 1 meet the above criteria. So, as we can see this set of design criteria does not prevent improved inversion attacks. So we propose to add the following criterion to the set:

Designers of stream ciphers should additionally minimise the largest gap between cells with taps to multiples of the connection polynomial or to the filter function.

6. Experiments

We have implemented the basic inversion attack and the improved inversion attack and we have conducted the following experiments on a typical Pentium II 400 MHz PC with 128 MB RAM:

1. Attacks on NFG with connection polynomial $p(x) = x^{33} \oplus x^{13} \oplus 1$, tapping sequence $\gamma = \{31, 15, 7, 3, 1, 0\}$ and filter function $f(x_{31}x_{15}x_7x_3x_1x_0) = x_{31} \oplus x_{15}x_3 \oplus x_7x_1 \oplus x_3x_1x_0$ for different initial states. Inversion attack on this generator takes up to few days and improved inversion attack takes up to 20 seconds (depending on initial state of the LFSR).

2. Attacks on NFG with connection polynomial $p(x) = x^{64} \oplus x^4 \oplus x^3 \oplus x^1 \oplus 1$, tapping sequence $\gamma = \{63, 31, 15, 7, 3, 1, 0\}$ and filter function $f(x_{63}x_{31}x_{15}x_7x_3x_1x_0) = x_{63} \oplus x_{31} \oplus x_{15}x_3 \oplus x_7x_1 \oplus x_3x_1x_0$. Inversion attack has computational complexity $\boldsymbol{O}(2^{63})$ so it is infeasible to conduct it on our PC. Improved inversion attack takes up to few days.

7. Conclusions and final remarks

We have proposed a powerful improvement of the inversion attacks. We have conducted several experiments which have confirmed theoretical predictions.

This attack is also effective when instead of regular LFSR, a modular LFSR is used (with inter cell feedback).

Our further research will concentrate on possible transformations of filter functions in such a way as to maximise the largest gap.

Acknowledgement

This work has been supported by grant no. 8 T11D 020 19 of the Polish Scientific Research Committee.

References

[1] R. J. Anderson, "Searching for the optimum correlation attack", in *Fast Software Encryption* – Leuven'94, *LNCS*. Springer, 1995, vol. 1008, pp. 137–143.

[2] J. Dj. Golić, "Correlation via linear sequential circuit approximation of combiners with memory", in *Advances in Cryptology – EUROCRYPT'92, LNCS*. Springer, 1993, vol. 658, pp. 113–123.

[3] J. Dj. Golić, "On the security of the nonlinear filter generators", in *Fast Software Encryption* – Cambridge'96, *LNCS*. Springer, 1996, vol. 1039, pp. 173–188.

[4] J. Dj. Golić, A. Clark, and E. Dawson, "Inversion attack and branching", in *Information Security and Privacy, ACISP'99, LNCS*. Springer, 1999, vol. 1587, pp. 88–102.

[5] J. Dj. Golić, Private communications, May 2002.

[6] A. Górska and K. Górski, "Improved inversion attacks on nonlinear filter generators", *IEE Electron. Lett.*, vol. 38, no. 16, pp. 870–871, 2002.

[7] W. Meier and O. Staffelbach, "Fast correlation attacks on certain stream ciphers", *J. Cryptol.*, vol. 1, no. 3, pp. 159–176, 1989.

[8] M. Salmasizadeh, L. Simpson, J. Dj. Golić, and E. Dawson, "Fast correlation attacks and multiple linear approximations", in *Information Security and Privacy, ACISP'97, LNCS*. Springer, 1997, vol. 1270, pp. 228–239.

Anna Górska received the M.Sc. degree in 1995 and the Ph.D. degree in 2003 both from the Faculty of Electronics and Information Technology at Warsaw University of Technology, Poland. She is currently a cryptographer in the Cryptography Division at Enigma Information Security Systems Sp. z o.o. Her research interests include the design and cryptanalysis of ciphers. She is a member of the International Association for Cryptologic Research and the Institute of Electrical and Electronics Engineers.
e-mail: ania@enigma.com.pl
ENIGMA Information Security Systems Sp. z o.o.
Cryptography Division
Cietrzewia st 8
02-492 Warsaw, Poland

Karol Górski received the M.Sc. degree in 1991 from the Institute of Telecommunications, Faculty of Electronics and Information Technology at Warsaw University of Technology, Poland. He currently heads the Cryptography Division at Enigma Information Security Systems Sp. z o.o. His duties include the management of research and development activities undertaken by the company in the area of cryptography and cryptanalysis as well as the management of software development for cryptographic devices and specialised cryptographic systems. He is a member of the Technical Committee for Information Security in IT Systems at the Polish Standardisation Committee (PKN) and an expert of the Cryptographic Algorithms and Mechanisms Working Group (WG2) in the joint ISO/IEC subcommittee on Information Technology Security Techniques (ISO/IEC JTC1 SC27). He is a also a member of the International Association for Cryptologic Research and the Institute of Electrical and Electronics Engineers.
e-mail: karol@enigma.com.pl
ENIGMA Information Security Systems Sp. z o.o.
Cryptography Division
Cietrzewia st 8
02-492 Warsaw, Poland

Paper

The tactical Intranet IPSec security concept

Mariusz Bednarczyk, Jacek Jarmakiewicz, and Jarosław Krygier

Abstract — The IPSec protocols architecture that can be applied in tactical Intranet based on the IPv6 protocol stack for wireless environment is the subject of the paper. The potential usefulness of the new version of IP protocol is very important for tactical communication systems. Additionally, Internet Engineering Task Force (IETF) security working group proposes recommendations covering the RFC 2401, 2402, 2406, that describe the security architecture for Internet Protocol. These standards, published by IETF are discussed here in military requirements context. The NATO C3 Technical Architecture model also recommends these issues. The concept of the IPSec architecture in military systems is described in the paper. The position of the security applications designed for subscriber devices with reference to layered model is also presented. The concept presented here is defined for the tactical level.

Keywords — IPSec, tactical Intranet, IP security.

1. Introduction

The modern Armed Forces need modern solutions, especially in the area of communication systems. For example, local area networks (LAN) have became just essential part of contemporary military units (command posts). Along with commercial of the shelf products (COTS) application, their security have became very important factor that have to be taken into account. The new standard covering the IPv6 protocol, that is still tested, includes mechanisms suitable for military systems. It is associated with security mechanisms (authentication, privacy and payload encryption) and mobile subscribers access to services and network resources as well as with the high quality of services requirements [6, 7].

The IPv6 have been designing as an evolution from IPv4 rather than as a major change. Useful features of IPv4 were carried over in IPv6 and less useful features were dropped. According to the IPv6 specification, the changes from IPv4 to IPv6 can be split primarily into the following categories [4]:

- **Inherent security support**. The IPv6 enables and enforces the IPSec authentication and encryption features through the extension headers. If authentication header (AH) is carried with the IP datagram, the receiving host must check the packet validity.
- **Mobility support**. The IPv6 protocol supports mobility management as an inherent function of IPv6 compared to the IPv4 that supports mobility through an additional protocol added on a top of IPv4 [5].
- **Built-in route optimization**. In IPv6 the correspondent node (CN) can learn so-called care-of-address (COA) of the mobile node (MN). The route optimization helps to prevent a problem of triangle routing. In triangle routing an incoming to MN traffic always passes through the home agent (HA), what can cause undesirable increasing the traffic load in the home network. Inherent route optimization is a major improvement compared to the IPv4 mobility protocol, which specifies route optimization as a separate extension to the mobility protocol. More importantly, all IPv6 nodes support route optimization while only mobile nodes in IPv4 can support the mobility extension.

The very large address space of IPv6 is the best-known feature of this protocol, but it is less interesting for military environment that has a relatively small dedicated Internet. The rich address formats are architecturally interesting: multicast supporting conferencing and broadcast applications; anycast supporting such activities as "use the nearest server". The attractiveness of such facilities is principally used to reduce the management of the network and bandwidth requirements as well. The auto-configuration facility could also reduce the running costs of the network as the address assignment can occur without participation of network administrator. However, this raise a security concern in loss of administrative control over the allocation of addresses in the network.

Tactical military networks are predominantly radio based and "on the move" with minimal fixed infrastructure. Also, security and mobility efficiency play essential role from military point of view.

The detailed background of the tactical Intranet structure that uses IPv6 protocol stack discussed here is clarified in [3] and [8]. The authors have limited the discussion only to the security problem that is applied to the tactical environment, treated the tactical network as a non-secure medium (denoted in the figures as a cloud).

2. The IPSec features

Security features of IPv6 have been obtained mainly by means of two dedicated extension header [1, 2]: *authentication header* (AH) and *encrypted security payload* (ESP) with complementary capabilities.

The AH header was designed to ensure authenticity and integrity of the IP packet. Its presence guards against two threats: the fixed fields illegal modification and packet

spoofing. On the other hand, the ESP header provides data encapsulation with encryption in order to ensure that only the destination node can read the payload conveyed by the IP packet. The two headers can be used together to provide all the security features simultaneously (Fig. 1).

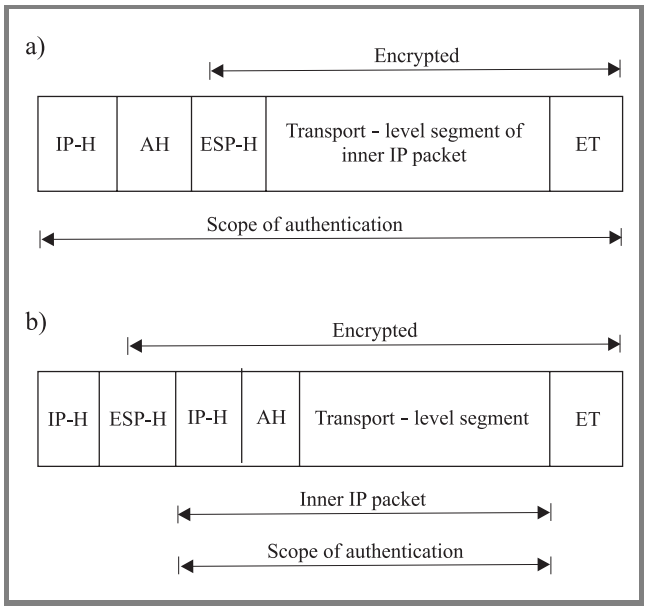


Fig. 1. Combining privacy and authentication: (a) encryption before authentication; (b) authentication before encryption. Explanations: IP-H – IP based header plus extensions headers, AH – authentication header, ESP-H – encapsulating security payload header, ET – encapsulating security payload trailing fields.

Both the AH and the ESP headers use a concept of the security association (SA) to agree on the security algorithms and parameters between the sender and the receiver. In general, each IPv6 node manages a set of SAs, one for each currently active secure communication. The *security parameters i*ndex (SPI) is a parameter contained in both the AH and ESP headers to specify which SA will be used in decryption and/or authentication the packet.

In unicast transmissions, the SPI is normally chosen by the destination node and sent back to the sender when the communication is set up. In multicast transmissions, the SPI must be common to all the members of the multicast group. Each node must be able to identify the right SA correctly by combining the SPI with the multicast address.

The negotiation of the SA (and the related SPI) is an integral part of the protocol for the exchange of security keys.

Correct application of the AH and ESP headers requires that all the communicating parties agree on a common key to be used in forming and checking the security headers. The IPv6 allows key management to occur either out-of-band or with specifically crafted protocols. However, no general agreement has been reached yet on this subject within the Internet community, with different groups stressing different needs: fast key exchange, strong authentication, lightweight protocols, and others. Key management is the area that is still mostly unsettled within the whole IPSec architecture.

The IPv6 requires each implementation to allow for manual setting of the security keys, in case of no in-line key management technique is adopted or human-based security is desired. Obviously, manual keying is possible only if the security administrators have separately agreed out-of-band on the keys to be used – for example, at a reserved meeting. This solution exhibits high personnel costs and does not scale well because it requires personal action of an operator on each network device taking part in the secure channel. Additionally, it can generate a false sense of security. The human intervention does not automatically ensure a higher level of security, due to untrusted administrators and residual problems related to hardware and software integrity of the device where the key is set. However, in spite of these disadvantages, manual key management finds application in restricted environments, with a small number of devices physically secured, that according to the security policy, can operate only when explicitly enabled by human intervention.

Within the IPSec, key management is surely the area that is less settled and the area in which much work has yet to be done before arriving at a set of protocols that completely meet the security needs at the IP level. The only decision that has already been made is that, for the sake of generality, the Internet key management protocol (IKMP) will be placed at the application layer, and it will be independent of the protocols at the lower layers.

The first proposal is to base IKMP on the coupling of the Internet security association and key management protocol (ISAKMP) and Oakley protocols, as described in the IEFT Draft, the resolution of ISAKMP with Oakley.

The ISAKMP defines a generic architecture for authenticated SA setup and key exchange, without specifying the actual algorithms to be used. In this way, it can be used with different key exchange techniques.

Oakley is a key-exchange protocol, based on a modified version of the Diffie-Hellman algorithm. Therefore, it is one of the natural partners for ISAKMP.

However, in addition to the ISAKMP-Oakley couple, different solutions are being proposed. Currently, the major competitor is simple key-management for Internet Protocols (SKIP), which bases its operations on the Diffie-Hellman algorithm. The SKIP is simple and addresses several problems of key management in high-speed networks, such as zero-message key setup and updates that permit fast dynamic rekeying (that is, frequent in-line change of the security keys to avoid analytic attacks based on accumulation of cyphertext encrypted with the same key). Moreover, although SKIP is not standardized yet, it already features many commercial-level implementations, both for UNIX workstations and for personal computers.

So the war of the key-management protocols is raging, and the likely outcome is that more than one protocol will attain

RFC status, because these protocols exhibit different merits that are valuable in different application environments.

3. Tactical Intranet security

The AH and ESP headers can be used in different ways to protect IP transmission.

In IPv6, achieving good level of security is easier and more standard than in IPv4, thanks to the AH and ESP headers. As an example, with reference to Fig. 2, let us suppose that a TCP session (channel) between host denoted H1 in network named N1 and host H2 in network N2 has to be protected only against data manipulation and origin falsification, while data privacy is not required. In this case, the AH header can be used in the following way. The firewall FW1 gets the IP packet and modifies it by adding an AH header before sending it to its partner firewall FW2. When this packet is received by the FW2, it checks the packet for integrity and origin authentication using the SPI data in the AH header. If the test is successful, then the IP header and the AH header are removed, and the remaining data (that is, the original packet) are sent to the final destination.

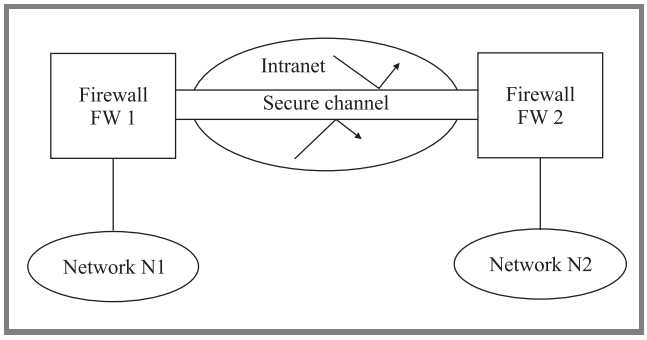


Fig. 2. An example of the tunnel between two firewalls.

If the network is implemented using only the AH header, then attackers can neither alter the transmitted packets nor insert forged packets in the channel. However, they can still read the content of the packets. To prevent disclosure of the payload, the ESP header has to be used too. Even the usage of AH in conjunction with ESP does not completely protect the traffic. Packets can be deleted by intermediate nodes or recorded and later replayed. These attacks cannot be easily contrasted at the IP level. Appropriate defenses (such as the use of unique packet identifiers and the generation of heartbeat packets) are usually placed at some upper level in the network stack. A partial solution at the IP level is likely to be offered by the new format and algorithms that are going to replace the current ones in the AH header.

In contrast to the IPv4, there is no problem with fragmentation in IPv6, because the overhead is fixed in size (the dimension of AH, or that of AH plus ESP) and fragmentation process is realized in source host.

This technique can be adopted even between the firewall and the single external host (Fig. 3). Obviously, this case is very important for guaranteed security when a mobile host is used outside the protected network, and it is a perfect complement to the mobility support features of IPv6. The firewall will act as a home agent in the neighbor discovery procedure. Mobile host will be assigned two different IP addresses: one when it is connected inside the security perimeter of the network and the other one when it is outside this perimeter. In second case, the firewall will also act as a relay, by routing packets coming from inside the corporate network to the external address, after adding the required headers (AH only, or AH plus ESP).

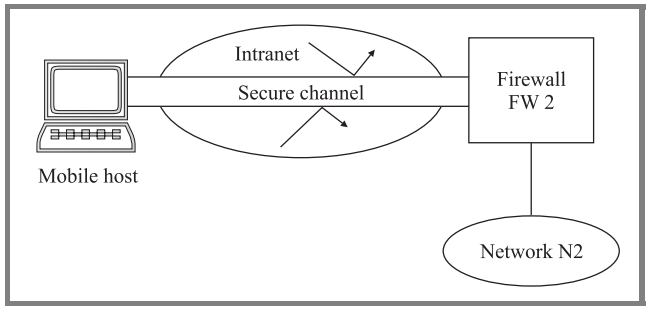


Fig. 3. Tunnel between a firewall and a single host.

This solution is not complete for application-level security because only partial protection is obtained. AH provides only host-based authentication, whereas applications usually require user-based authentication. Moreover, AH and ESP protect the data only during their transmission along the channel. After the data have been received, they are no longer protected in any way. This fact may not be relevant if the receiving host is a secure one, but there is the additional implication that origin authentication and data integrity properties are lost as well. So formal non-repudiation cannot occur after the data have been extracted from the secure channel.

The conclusion is that the security features of IPv6 do not eliminate the need for other security mechanisms, which will probably be better placed at the application level. Networked applications executing on the top of the IPv6 stack may be required in order to use the communication channel with specific features. To avoid duplication of functionality (and hence performance degradation) being able to specified at the transport layer, the security attributes of the created channel are useful.

Since IP addresses in the IPv6 are quite often dynamically assigned, it is the most importance that this process be done in a secure fashion.

Moreover, as different security properties are available through a proper combination of AH and ESP headers, it is highly desirable that they should be applied to the messages exchanged by the routers, to prevent attacks aiming to subvert the logical architecture of the network.

Following types of communications should be protected:

- The routing advertisement messages, to ensure that they are sent by an authorized router.
- The neighbor advertisement messages, to ensure that they come from authorized hosts and to avoid a risk that somebody attaches a new host to the network without proper authorization.
- The ICMP messages related to an unreachable host or network (*destination unreachable*) or to a better route (*redirect*), to ensure that these messages come from hosts or routers that were on the original path of the packets.

Securing these types of messages is surely not trivial. For example, the routing advertisements are sent to a multicast group; therefore, all the routers in the group have to know the (common) secret key to be used to verify and/or decrypt the messages. This fact implies that they can forge messages and impersonate any router in the group.
Protection of the neighbor advertisements is a serious problem. These messages can be protected only after the SA has been created between the host and the address distribution center. On the other hand, this SA can be created only after the address has been assigned to the host, so we can conclude that this is no correct solution. The break of the loop is possible. For example, priority can be given to the address assignment phase, and SA setup can be permitted only subsequently, but in this way the address assignment phase is not protected. Alternatively, public key authentication can be used. Each host is assigned a key pair (private and public key) and has to be reconfigured with the public key of the authority that signs the certificates of the routers and the address distribution centers. The last alternative is to configure the routers so that they do not advertise local prefixes. In this way, each host is forced to contact a router first.
Protection against malicious ICMP messages requires that they should be protected by the AH header, but this approach has the drawback of requiring the establishment of the SA with each router and host on the path between the source and the destination of the packets.
With respect to the messages security used by the various routing protocols, they should always be exchanged just within the frame of the SA and should be protected by the AH. For the sake of generality, this solution is highly preferable to using authentication mechanisms specific for each routing protocol.
Based on the previous analyses, we can conclude that routing security is apparently still a big problem in IPv6, but chances of solving this problem are higher than in IPv4.

4. Conclusions

The tactical military architecture is a hierarchical arrangement of mobile components. The degree of users mobility varies with the echelon and the distance from the front. The nature of military operations is such a changing that significant advantage will be achieved by the creative, timely and decisive usage of the information. Consequently, the demand for access to the information is also changing and evolving towards the new network-centric model based on more comprehensive, ubiquitous and shared information services.
From military point of view, there is no problem with address space exhaustion in the internal networks, as they are largely closed networks. Additionally, since such networks are relatively small, the problem with big routing tables does not exist.
The security features of IPv6 offer only commercial grade of security. Specific hardware encryption have to be added in military domain.
The IETF security architecture is open to apply additional encryption applications. Several solutions exist clarifying how the IPSec may be implemented to hosts in conjunction with the router or firewall. Some solutions are integrated into the native IP implementations and other ones are build-in according to bump-in-the-stack (BITS) or bump-in-the-wire (BITW) scheme [1, 9]. The second scheme enables to use an outboard crypto processor that is common designed feature of the security in military network.
Currently, the AH and ESP headers should be modified along the following guidelines:

- The AH format must be substantially changed to accommodate new and stronger authentication algorithms (HMAC – *keyed-hashing for message authentication* [10]) that support prevention of packet replay and its cancellation ([11] describes this format when used with the MD5 digest algorithm).
- The ESP specification must achieve a better orthogonality with algorithms, to simplify application of different encryption algorithms.

The benefit of these changes is that higher security will be available at the network level. Hence, applications will be able to concentrate on different security aspects, such as authorizations.

Acknowledgement

The paper is a result of the research project 0T00A047 financed by State Commitee for Scientific Research in the 2001/2002.

References

[1] R. Ramjee, T. La Porta, L. Salgarelli, S. Thuel, and K. Varadhan, "IP-based access network infrastructure for next-generation wireless data networks", *IEEE Pers. Commun.*, Aug. 2000.

[2] S. Gai, "Internetworking with IPv6 Cisco Routers", http://www.ip6.com/us/book/index.html

[3] M. Bednarczyk, J. Jarmakiewicz, and K. Maślanka, "Problems with using COTS technology in tactical communication systems based on stack of IPv6 protocol", Zegrze, 2001.

[4] D. B. Johnson and C. Perkins, "Mobility support in IPv6". Internet Draft, Apr. 2000, <draft-ietf-mobileip-ipv6-12.txt>

[5] C. Perkins, "IP mobility support", RFC 2002, Oct. 1996.

[6] W. Stallings, "IPv6: the new Internet Protocol", http://www.cs-ipv6.lancs.ac.uk/ipv6/documents/papers/stallings/

[7] W. Stallings, *Data and Computer Communications*. 5th ed. Prentice-Hall, 1997.

[8] M. Amanowicz, J. Jarmakiewicz, J. Krygier, and K. Maślanka, "Mobility management in tactical IPv6 network", in *Proc. MILCOM'2002*, Anaheim, USA, Oct. 2002.

[9] S. Kent and R. Atkinson, "Security architecture for the Internet Protocol", IETF Standard Track RFC 2401, Nov. 1998.

[10] H. Krawczyk, M. Bellare, and R. Canetti, "HMAC: keyed-hashing for message authentication", IETF RFC 2104, Feb. 1997.

[11] M. Oehler and R. Glenn, "HMAC-MD5 IP authentication with replay prevention", IETF Standards Track RFC 2085, Feb. 1997.

Mariusz Bednarczyk was born in Poland, in 1973. He received the M.Sc. degree in telecommunications domain in 1998 from Military University of Technology, Warsaw, Poland. He engages in problems of communications and information systems (CIS) modelling and simulation and advanced wireless communication technologies as well.
e-mail: mbednarczyk@wel.wat.edu.pl
Telecommunications Institute
Military University of Technology
Kaliskiego st 2
00-908 Warsaw, Poland

Jacek Jarmakiewicz received M.Sc. degree in 1989 from Military University of Technology, Warsaw, Poland. He finished postgraduate studies in the National Institute of Telecommunications in Warsaw, in the communications network management domain. He specializes in mobility management, mobile computing, communications systems modelling and simulation, computer networks and, performance evaluation.
e-mail: jjarmakiewicz@wel.wat.edu.pl
Telecommunications Institute
Military University of Technology
Kaliskiego st 2
00-908 Warsaw, Poland

Jarosław Krygier was born in Poland, in 1971. He received the M.Sc. degree in 1996 and the Ph.D. degree in 2002 from Military University of Technology, Warsaw, Poland, both in telecommunication engineering. He engages in problems of communications and information systems (CIS) modelling and simulation, wideband CDMA technology, IPng problems, CIS interoperability and telecommunication systems engineering. He is an author and co-author of over 30 scientific papers and research reports.
e-mail: jkrygier@wel.wat.edu.pl
Telecommunications Institute
Military University of Technology
Kaliskiego st 2
00-908 Warsaw, Poland

Paper

A study of differences between bent functions constructed using Rothaus method and randomly generated bent functions

Anna Grocholewska-Czuryło

Abstract — Bent functions, having the highest possible nonlinearity, are among the best candidates for construction of *S*-boxes. One problem with bent functions is the fact that they are hard to find among randomly generated set of Boolean functions already for 6 argument functions. There exist some algorithms that allow for easy generation of bent functions. The major drawback of these algorithms is the fact that they rely on deterministic dependencies and are only able to generate bent functions belonging to one specific class. In our paper we present an efficient generator of random bent functions of more than 4 arguments. Resulting functions are not bounded by constraints described above. The generator operates in algebraic normal form domain (ANF). We also present our result on comparing the performance of *S*-boxes build using our bent function generator versus a standard method of bent function construction. We also give some directions for further research.

Keywords — block ciphers, S-boxes, bent functions, construction, random generation, nonlinearity.

1. Introduction

In block ciphers based on *S*-boxes, the cryptographic strength (i.e. resistance to cryptanalysis attack) depends on the nonlinear properties of an *S*-boxes used to build the cipher. *S*-boxes are built from Boolean functions, so quality of each and every function constituting an *S*-box is of a greatest importance. Another major consideration for *S*-box construction is the way functions that form an *S*-box "interact" (behave as a group of functions) – what are the cryptographic properties of an *S*-box as a whole. These are two main factors that affect cipher's cryptographic performance.

The properties of Boolean functions have been extensively studied. The quality of a single, cryptographically strong Boolean function is measured by its cryptographic properties. The criteria against which a function quality is measured are mainly nonlinearity, balancedness, avalanche and propagation criteria.

The qualities that single Boolean functions should posses to be good candidates for *S*-box construction are very similar to those that should be characterizing a good (strong) *S*-box (taken as a linear combination of constituting functions). The nonlinear properties are by far the most important. In recent years a class of highly nonlinear functions attracted a lot of researchers' attention – a class of bent Boolean functions. These functions have in fact the highest possible nonlinearity (they also have very good propagation characteristics and are nearly balanced – another important criterion for good cryptographic function – special algorithm have been proposed by different authors for transforming bent functions into balanced Boolean functions while maintaining high level of nonlinearity).

However one major drawback is the fact that bent functions are not easily constructed (i.e. their constructions are time consuming). Trying to find bent functions by pure random search is also virtually impossible (already for 6-argument functions, only one in $2.9 \cdot 10^{-10}$ Boolean functions is bent). Also, an *S*-box that is constructed using only bent functions does not necessarily possess the best nonlinear qualities. So usually a number of different *S*-boxes should be generated and tested, and then only the best *S*-boxes should be selected for incorporating in block cipher. This approach demands fast *S*-box generation which in turn translates to fast bent function generation.

The remainder concentrates on efficient random bent function generation and using such generated function for *S*-box construction.

2. Preliminaries

We use square brackets to denote vectors like $[a_1, \ldots, a_n]$ and round brackets to denote functions like $f(x_1, \ldots, x_n)$.

Boolean function. Let GF(2) $=< \sum, \oplus, \bullet >$ be two-element Galois field, where $\sum = \{0,1\}$, $\oplus$ and $\bullet$ denotes the sum and multiplication mod 2, respectively. A function $f : \sum^n \to \sum$ is an n-argument Boolean function. Let $z = x_1 \cdot 2^{n-1} + x_2 \cdot 2^{n-2} + \ldots + x_n \cdot 2^0$ be the decimal representation of arguments $(x_1, x_2, \ldots, x_n)$ of the function f. Let us denote $f(x_1, x_2, \ldots, x_n)$ as y_z. Then $[y_0, y_1, \cdots, y_{2^n-1}]$ is called a truth table of the function f.

Linear and nonlinear Boolean functions. An n-argument Boolean function f is linear if it can be represented in the following form: $f(x_1, x_2, \ldots, x_n) = a_1 x_1 \oplus a_2 x_2 \oplus \ldots a_n x_n$.

Let L_n be a set of all n-argument linear Boolean functions. Let $M_n = \{g : \Sigma^n \to \Sigma \mid g(x_1, x_2, \ldots, x_n) = 1 \oplus f(x_1, x_2, \ldots, x_n)$ and $f \in L_n\}$. A set $A_n = L_n \cup M_n$ is called a set of n-argument affine Boolean functions. A Boolean function $f : \Sigma^n \to \Sigma$ that is not affine is called a nonlinear Boolean function.

Balance. Let $N_0[y_0, y_1, \ldots, y_{2^n-1}]$ be a number of zeros (0's) in the truth table $[y_0, y_1, \ldots, y_{2^n-1}]$ of function f, and $N_1[y_0, y_1, \ldots, y_{2^n-1}]$ be number of ones (1's). A Boolean function is balanced if $N_0[y_0, y_1, \ldots, y_{2^n-1}] = N_1[y_0, y_1, \ldots, y_{2^n-1}]$.

Algebraic normal form. A Boolean function can also be represented as a maximum of 2^n coefficients of the algebraic normal form. These coefficients provide a formula for the evaluation of the function for any given input $x = [x_1, x_2, \ldots, x_n]$:

$$f(x) = a_0 \oplus \sum_{i=1}^{n} a_i x_i \oplus \sum_{1 \le i < j \le n} a_{ij} x_i x_j \oplus \ldots \oplus a_{12\ldots n} x_1 x_2 \ldots x_n,$$

where $\sum, \oplus$ denote the modulo 2 summation.

The order of nonlinearity of a Boolean function $f(x)$ is a maximum number of variables in a product term with non-zero coefficient a_J, where J is a subset of $\{1, 2, 3, \ldots, n\}$. In the case where J is an empty set the coefficient is denoted as a_0 and is called a zero order coefficient. Coefficients of order 1 are a_1, $a_2, \ldots, a_n$, coefficients of order 2 are $a_{12}, a_{13}, \ldots, a_{(n-1)n}$, coefficient of order n is $a_{12\ldots n}$. The number of all ANF coefficients equals 2^n.

Let us denote the number of all (zero and non-zero) coefficients of order i of function f as $\sigma_i(f)$. For n-argument function f there are as many coefficients of a given order as there are i-element combinations in n-element set, i.e. $\sigma_i(f) = \binom{n}{i}$.

Hamming distance. Hamming weight of a binary vector $x \in \Sigma^n$, denoted as $hwt(x)$, is the number of ones in that vector.

Hamming distance between two Boolean functions $f, g : \Sigma^n \to \Sigma$ is denoted by $d(f,g)$ and is defined as follows:

$$d(f,g) = \sum_{x \in \Sigma^n} f(x) \oplus g(x).$$

The distance of a Boolean function f from a set of n-argument Boolean functions X_n is defined as follows:

$$\delta(f) = \min_{g \in X_n} d(f,g),$$

where $d(f,g)$ is the Hamming distance between functions f and g. The distance of a function f from a set of affine functions A_n is the distance of function f from the nearest function $g \in A_n$.

The distance of function f from a set of all affine functions is called the nonlinearity of function f and is denoted by N_f.

Bent functions. A Boolean function $f : \Sigma^n \to \Sigma$ is perfectly nonlinear if and only if $f(x) \oplus f(x \oplus \alpha)$ is balanced for any $\alpha \in \Sigma^n$ such that $1 \le hwt(\alpha) \le n$.

For a perfectly nonlinear Boolean function, any change of inputs causes the change of the output with probability of 1/2.

Meier and Staffelbach [16] proved that the set of perfectly nonlinear Boolean functions is the same as the set of Boolean bent functions defined by Rothaus [5].

Perfectly nonlinear functions (or bent functions) have the same, and the maximum possible distance to all affine functions. So their correlation to any affine function is consistently bad (minimal). Linear cryptanalysis works if it is possible to find a good linear approximation of the *S*-box.

Bent functions are not balanced. This property prohibits their direct application in *S*-box construction, however there exist numerous methods for modifying bent function in such a way so that the resulting function is balanced and still maintains the good cryptographic properties of a bent function [16]. Hamming weight of a bent function equals $2^{n-1} \pm 2^{n/2-1}$.

Differential analysis [18] can be seen as an extension of the ideas of attacks based on the presence of linear structures [3]. As perfect nonlinear Boolean function have maximum distance to the class of linear structures (equal to 2^{n-2}), they are a useful class of functions for constructing mappings that are resistant to differential attacks.

Bent functions exist only for even n. The nonlinear order of bent functions is bounded from above by $n/2$ for $n > 2$. The number of Boolean bent function for $n > 6$ remains an open problem.

3. Constructing bent functions

There exist a number of algorithms for constructing bent functions. As an example let's consider the following [8, 12].

Method 1. Let B_n denote a set of bent functions $f : \Sigma^n \to \Sigma$ with n even. Given a set of bent functions B_6, bent functions in B_8 can be constructed using the following method (Method 1).

Let $a, b \in B_6$. Then the function $f : \Sigma^8 \to \Sigma$ defined by:

$$f(x_0 \ldots, x_7) = \begin{cases} a(x_0 \ldots x_5), & x_6 = 0,\ x_7 = 0 \\ a(x_0 \ldots x_5), & x_6 = 0,\ x_7 = 1 \\ b(x_0 \ldots x_5), & x_6 = 0,\ x_7 = 0 \\ b(x_0 \ldots x_5), \oplus 1, & x_6 = 0,\ x_7 = 0 \end{cases}$$

is bent [8]. Rearrangements of the 64 blocks in the expression above also result in bent functions.

Another method for bent function construction was given by Rothaus in [5].

Method 2. Let $x = (x_1, \ldots, x_n)$ and let $a(x)$, $b(x)$ and $c(x)$ be bent functions such that $a(x) \oplus b(x) \oplus c(x)$ is also bent. Then a function $f(x, x_{n+1}, x_{n+2}) = a(x)b(x) \oplus b(x)c(x) \oplus c(x)a(x) \oplus [a(x) \oplus b(x)]x_{n+1} \oplus [a(x) \oplus c(x)]x_{n+2} \oplus x_{n+1}x_{n+2}$ is bent.

Most of the known bent function constructions take bent functions of n arguments as their input and generate bent functions of $n+2$ arguments. One major drawback of these methods is the fact that they are deterministic. Only short bent functions ($n = 4$ or 6) are selected at random and the resulting function is obtained using the same, deterministic formula every time. Even if there is some "random" element in such generation (like adding a linear term to the resulting bent function) it does not bring any new quality to the generated function.

4. Generating bent functions

To overcome some of the limitations and possible weaknesses of the bent functions construction methods described above a new algorithm of random bent functions generation have been proposed [24].

As already mentioned earlier, drawing bent functions at random is not feasible already for small number of arguments ($n > 6$). To make such generation possible, an algorithm was designed that generates random Boolean functions in algebraic normal form thus making use of some basic properties of bent functions to considerably narrow the search space. This makes the generation of bent functions feasible for $n \geq 8$ even on a standard PC machine. The algorithm for the generation of bent functions in ANF domain takes as its input the minimum and maximum number of ANF coefficients of every order that the resulting functions are allowed to have. Since the nonlinear order of bent functions is less or equal to $n/2$, clearly in ANF of a bent function cannot be any ANF coefficient of order higher then $n/2$. This restriction is the major reason for random generation feasibility, since it considerably reduces the possible search space.

The number of ANF coefficients of orders less or equal to $n/2$ can be fixed or randomly selected within allowed range (i.e. between 0 and $\sigma(f) = \binom{n}{i}$ for order i). If the number of coefficients for a given order i is fixed then all generated functions will have the same number of coefficients of that order, but the coefficients themselves will be different in each generated function. If the number of coefficients for a given order i is randomly selected then all generated functions will not only have different coefficients but also the number of coefficients of order i will vary from function to function. It is of course possible to fix the number of coefficients for some orders and have varied number of coefficients for other orders.

One important consequence of this approach is the possibility of prohibiting the generation of bent functions which are merely a linear transformations of other bent functions. This is easily achieved by setting the number of coefficients of order 0 and 1 to 0. So in the ANF of the resulting functions there will be no linear part. Bent functions of any order can be generated with this method, simply by setting any higher order coefficients to 0. Homogenous bent functions can also be generated easily.

One drawback of the method is the fact that it does not guarantee the generation of bent functions without repetitions, although the chance of generating two identical bent functions is minimal with any reasonably selected ranges of number of ANF coefficients. However, if avoiding repetitions is an absolute requirement, the set of generated bent functions must be checked for duplicates.

The limitations of this approach are twofold. First there is a feasibility limit. Number of possible functions grows with the number of coefficients of higher orders ($i > 2$) and generating a bent function quickly becomes infeasible. So the algorithm works best with the low number of higher order coefficients (e.g. < 6 for $n = 8$ and order $i = 3$ and 4). Due to the above limitation, this method does not generate all possible bent functions with equal probability. In principle, it would be possible but is not feasible for the reason described above. One has to limit the number of higher order coefficients and at the same time prohibit the generation of some bent functions.

5. Comparing pairs of bent functions

In this section some comparative results are presented. Three sets of 8 argument bent Boolean functions are analyzed: bent functions constructed using Method 1, bent functions constructed using method given in [22] (Maiorana functions with permuted inputs) and randomly generated bent functions. For random, distinct i, j the nonlinearity of $f_i \oplus f_j$ was calculated. Figures 1 and 2 show the resulting nonlinearity distribution (in percentage).

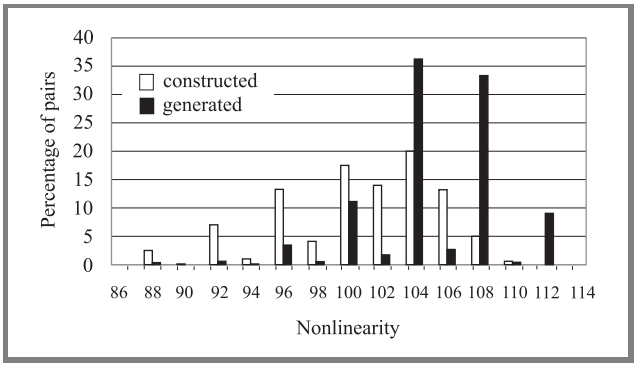


Fig. 1. Pairs nonlinearity distribution. Constructed bent (Method 1) versus generated bent.

The random bent functions were generated with the following parameters: number of 2nd order coefficients was between 7 and 14 (statistically that yields the highest number of bent functions), number of 3rd order coefficients was fixed at 2 and number of 4th order ANF coefficients was also fixed at 2. There were no coefficients of order 0 and 1 to prevent the occurrences of bent functions that would be just a linear transformations of one another.

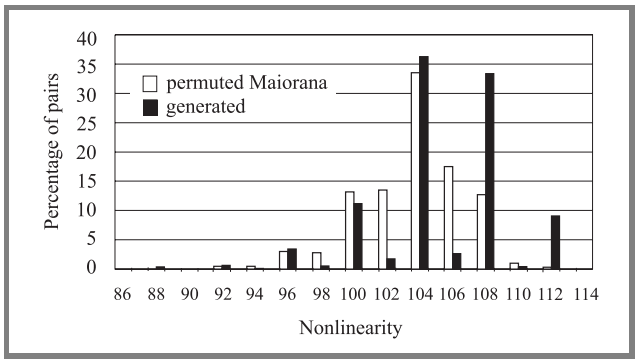


Fig. 2. Pairs nonlinearity distribution. Constructed bent (permuted Maiorana) versus generated bent.

For randomly generated functions the distribution is shifted towards higher values (i.e. pairs have better nonlinearity) and is also much more narrow - more appropriate pairs can be found in this set of functions. The results obtained in our experiments are also better then those presented in [22] for a special subsets of bent functions: Maiorana functions with permuted inputs.

6. Comparing *S*-boxes

In this section we present some properties of *S*-boxes build using randomly generated bent function. We give comparative results of the performance of *S*-boxes build of bent functions constructed using a method introduced by Rothaus [5] (Method 2 from Section 3), bent functions generated with our algorithm described in this paper and random Boolean functions (not bent). Two properties of *S*-boxes were measured: feasibility of a linear approximation which is a measure of *S*-box resistance against linear cryptanalysis and nonlinearity of the *S*-box which is one of the major criteria of cryptographic quality.

One has to note that for "real-life" applications bent functions would have to be modified to be balanced prior to their use in *S*-boxes. Such modification algorithms are beyond the scope of this paper, where main focus is kept on bent functions.

6.1. Linear approximations of S-boxes

We used a method of linear approximation as described in [23].

By linear approximation of a Boolean function $h:\sum^n \to \sum^m$, written as $Y = h(X)$, we mean any equation of the form:

$$\sum_{i\in Y'} y_i = \sum_{j\in X'} x_j, \text{ for } Y' \subseteq \{1,2,\ldots,m\},\ X' \subseteq \{1,2,\ldots,n\},$$

fulfilled with the probability of $p = N(X',Y')/2^n$, where $N(X',Y')$ denotes the number of pairs (X,Y) fulfilling the equation, and $\sum$ is a modulo 2 summation. The sets if indices X',Y' are called input and output masks.

The measure of linear approximation effectiveness is the value of a probability $\Delta p = |p - 1/2|$ called differential probability. For a fixed n a measure of effectiveness can also be defined as a value of $\Delta N(X',Y') = |N(X',Y') - 2^{n-1}|$.

In our experiment we tested linear approximations of 6 x 6 *S*-boxes, i.e. functions $Y = h(X) : \sum^6 \to \sum^6$, where sub-functions of function h were constructed bent functions, randomly generated bent functions and random functions (Fig. 3). The distribution of the best approximations was tested, i.e. maximum value of $\Delta N(X',Y')$ among all possible sets of input and output masks (except empty output mask). For each type of functions 10 000 of random *S*-boxes were tested. The number of *S*-boxes is given on Y (value) axis. The X (category) axis gives the values of the best approximations (higher value means better approximation so worse *S*-box).

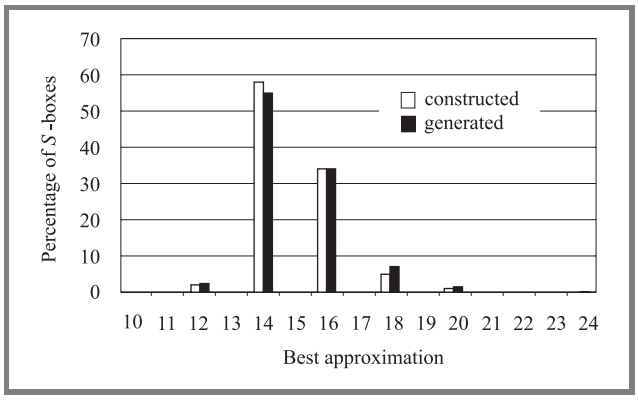


Fig. 3. Best *S*-box approximation distribution. Constructed (Rothaus) versus generated bent.

Differences between *S*-boxes build of bent functions constructed using Rothaus method (Method 2) and *S*-boxes build from randomly generated bent functions are not very evident.

6.2. Nonlinearity

Now we will show the results of testing the *S*-boxes for high nonlinearity (Fig. 4). We consider 6 x 8 *S*-boxes (each *S*-box is constructed of six 8-argument functions).

The nonlinearity of an *S*-box, so a function $F:\sum^n \to \sum^m$ such that $F(x) = \left(f_1(x), f_2(x), \ldots, f_m(x)\right)$ i $x \in \sum^n$, is calculated as minimal nonlinearity of all linear combinations of F's sub-functions. The nonlinearity of a *S*-box is then defined as follows:

$$N_F = \min\left\{N_{f_J} \mid f_J = \sum_{i\in J} f_i,\ J \subseteq (1,2,\ldots,m)\right\}.$$

To calculate a nonlinearity of a single *S*-box 2^m linear combinations have to be constructed and their distance to affine functions calculated. The lowest of all calculated nonlinearities (distances to affine functions) is the nonlinearity of the *S*-box.

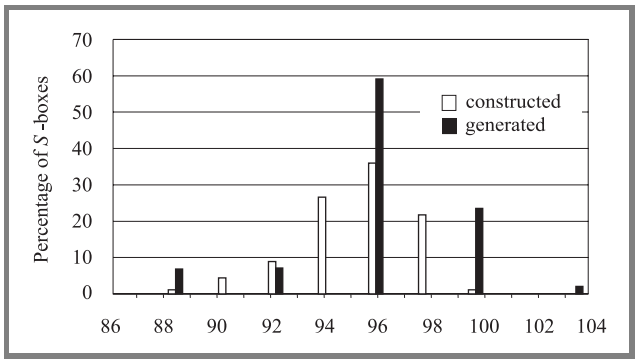


Fig. 4. *S*-box nonlinearity distribution. Constructed (Rothaus) versus generated bent.

Among *S*-boxes built from generated bent functions there exist *S*-boxes of the highest found nonlinearity of 104. There is also about 20 times more *S*-boxes of very high nonlinearity of 100 then in the case of *S*-boxes build from constructed bent functions or random balanced functions. This means that using randomly generated bent functions may lead to constructing *S*-boxes of better cryptographic qualities in less time.

However, one has to note the fact that in case of randomly generated bent functions there are also *S*-boxes of relatively poor nonlinearity (like 80). So building *S*-boxes from these functions requires (more then in other cases) careful checking the resulting *S*-boxes for possibly low nonlinearity.

7. Conclusions

From the results presented in this paper it seems that random generated bent functions offer an interesting alternative to construction methods. Not only nonlinear characteristics of these functions are equal or better then those of constructed bent functions but also generated functions have a very compact (small) algebraic normal form which can be utilized for efficient storage and fast cryptographic routines.

Next step in randomly generated bent function assessment will be checking their avalanche and propagation criteria, also when incorporated into *S*-boxes.

Perhaps a combined method of ANF generation of relatively short bent functions (i.e. $n \leq 10$) and then supplying them as an input for deterministic construction can yield some interesting results. Such functions would also have to be tested to verify their possibly superior cryptographic qualities.

References

[1] L. J. O'Connor, "An analysis of a class of algorithms for *S*-box construction", *J. Cryptol.*, vol. 7, no. 3, pp. 133–152, 1994.

[2] C. E. Shannon, "Communication theory of secrecy systems", *Bell Syst. Techn. J.*, vol. 28, pp. 656–715, 1949.

[3] K. Nyberg, "Perfect nonlinear *S*-boxes", in *Advances of Cryptology – EUROCRYPT'91, LNCS*. Springer, 1991, vol. 547, pp. 378–386.

[4] J. Seberry, X. M. Zhang, and Y. Zheng, "Systematic generation of cryptographically robust *S*-boxes", in *Proc. 1st ACM Conf. Comput. Commun. Secur.*, 1993.

[5] O. S. Rothaus, "On bent functions", *J. Combinat. Theory*, vol. 20, pp. 300–305, 1976.

[6] B. Preneel, W. Van Leekwijck, L. Van Linden, R. Govaerts, and J. Vandewalle, "Propagation characteristics of Boolean functions", in *Advances in Cryptology – EUROCRYPT'90, LNCS*. Springer, 1991, vol. 473, pp. 161–173.

[7] J. F. Dillon, "A survey of bent functions", *NSA Techn. J.*, special issue, pp. 191–215, 1972.

[8] C. M. Adams and S. E. Tavares, "Generating and counting binary bent sequences", *IEEE Trans. Inform. Theory*, vol. IT-36, pp. 1170–1173, 1990.

[9] J. A. Maiorana, "A class of bent functions", R41, 1971.

[10] C. M. Adams, "A formal and practical design procedure for substitution permutation network cryptosystems". Ph.D. thesis, Department of Electrical Engineering, Queen's University, 1990.

[11] M. Dawson and S. E. Tavares, "An expanded set of *S*-box design criteria based on information theory and its relation to differential-like attacks", in *Advances in Cryptology – EUROCRYPT'91, LNCS*. Springer, 1991, vol. 547, pp. 352–367.

[12] J. B. Kam and G. Davida, "Structured design of substitution-permutation encryption networks", *IEEE Trans. Comput.*, vol. C-28, pp. 747–753, 1979.

[13] L. O'Connor, "An analysis of product ciphers based on the properties of Boolean functions". Ph.D. thesis, Department of Computer Science, University of Waterloo, 1992.

[14] A. F. Webster and S. E. Tavares, "On the design of *S*-boxes", in *Advances in Cryptology – CRYPTO'85, LNCS*. Springer, 1986, pp. 523–534.

[15] R. Forré, "The strict avalanche criterion: spectral properties of Boolean functions with high nonlinearity", in *Advances in Cryptology – CRYPTO'88, LNCS*. Springer, 1990.

[16] W. Meier and O. Staffelbach, "Nonlinearity criteria for cryptographic functions", in *Advances in Cryptology – EUROCRYPT '89, LNCS*. Springer, 1990, vol. 434, pp. 549–562.

[17] E. Biham and A. Shamir, *Differential Cryptanalysis of the Data Encryption Standard*. Springer, 1993.

[18] M. Matsui, "Linear cryptanalysis method for DES cipher", in *Proc. EUROCRYPT'93* (abstracts), 1993.

[19] R. Yarlagadda and J. E. Hershey, "A note on the eigenvectors of Hadamard matrices of order 2^n", *Linear Algebra & Appl.*, vol. 45, pp. 43–53, 1982.

[20] R. Yarlagadda and J. E. Hershey, "Analysis and synthesis of bent sequences", *Proc. IEE*, vol. 136, pp. 112–123, 1989.

[21] K. Nyberg, "Constructions of bent functions and difference sets", in *Advances in Cryptology – EUROCRYPT'90, LNCS*. Springer, 1991, vol. 473.

[22] S. Mister and C. Adams, "Practical *S*-box design", in *Workshop on Selected Areas in Cryptography (SAC '96) Workshop Record*, Queens University, 1996, pp. 61–76.

[23] K. Chmiel, "Liniowa aproksymacja funkcji *S*-bloków". Raport nr 475. Politechnika Poznańska, Katedra Automatyki, Robotyki i Informatyki, Poznań, 2000 (in Polish).

[24] R. Wicik, "Wykorzystanie szyfrów blokowych opartych o sieci podstawieniowo-przestawieniowe o dużych *S*-boksach w specjalnych sieciach telekomunikacyjnych". Rozprawa doktorska, Wojskowa Akademia Techniczna, Warszawa, 1999 (Ph.D. thesis in Polish).

[25] A. Grocholewska-Czuryło and J. Stokłosa, "Generating bent functions", in *Proc. ACS 2001*.

Anna Grocholewska-Czuryło was born in Poznań, Poland, in 1969. In 1992 she graduated with the M.Sc. degree in computer science from the Faculty of Electrical Engineering at the Poznań University of Technology. She had been working for a year in University's Super Computer Center before she has moved on to become a teacher and a research assistant at the Laboratory of Information Technology Security. She has studied and published papers on a range of topics like natural language processing, cellular automata, neural networks and has finally focused on cryptography, and *S*-box design in particular. She has earned her Ph.D. degree in 2001.
e-mail: czurylo@sk-kari.put.poznan.pl
Laboratory of Information Technology Security
Poznań University of Technology
Marii Skłodowskiej-Curie st 5
61-542 Poznań, Poland

Paper

A comparison of ATM and IP QoS network capabilities for handling LAN traffic with QoS differentiation

Andrzej Bęben, Wojciech Burakowski, and Piotr Pyda

Abstract — Now, a network operator must choose between two packet switched technologies for providing QoS in WAN networks, which are ATM and IP QoS [3, 4, 9]. As ATM has reached the maturity with capabilities for offering a number of different network services (i.e. CBR, VBR, ABR, UBR, GFR), the IP QoS with network services like expedited forwarding, assured forwarding, etc. is still at developing phase but nevertheless is commonly regarded as capable to guarantee in near future similar QoS level as ATM. This paper tries to compare the efficiency of the mentioned technologies (in case of IP QoS network the AQUILA network concept [1, 2] is investigated) for handling traffic generated by LANs with QoS differentiation. This is extremely required since the applications running in LAN differ in QoS requirements and emitted traffic profiles (streaming, elastic). Therefore, a classification process of outgoing LAN traffic into predefined sub-streams should be performed at the entry point to WAN network (edge ATM switch or IP router). Furthermore, particular sub-streams are submitted to adequate WAN network service, available in ATM or IP QoS. The paper presents the experimental results, measured in the test bed, corresponding to QoS level and QoS differentiation provided by ATM and IP QoS core. For this purpose, a set of representative applications currently available to a LAN user was selected demanding from the core different QoS level. They correspond to streaming applications like VoIP with QoS objectives represented mainly by packet delay characteristics and elastic applications controlled by TCP protocol with minimum guaranteed throughput/goodput as target.

Keywords — traffic control, IP QoS, asynchronous transfer mode.

1. Introduction

A variety of applications in a LAN environment is now available. Apart from traditional data computer oriented applications with data transfer controlled by transmission control protocol (TCP), like file transfer protocol (FTP), Telnet, e-mail, world wide web (WWW), a user would also like to use new Internet applications, like voice over IP (VoIP), videoconferencing, etc., which are based on transferring voice or/and video. Let us remark that data transfer usually tolerates even large packet delays and, therefore, can be effectively served by e.g. IP best effort network. On the contrary, for satisfying users, the voice/video should be transferred with low packet delay and low packet losses. As a consequence, the packet flow outgoing from a LAN becomes heterogeneous with respect to quality of service (QoS) requirements for packet transfer in WAN network. Therefore, a WAN network should have capability for providing QoS differentiation. This directly leads to offering by network a number of network services (NSs), differing in QoS objectives. For instance, the file transfer should be handled by a NS aimed at throughput guarantees, while voice transfer demands a NS guaranteeing low packet delay and low packet losses.

Currently, two network technologies offering a set of NSs are available, that are ATM [7, 8, 9] and IP QoS [3, 4]. The ATM currently offers 6 native ATM NSs, i.e. constant bit rate (CBR), real time variable bit rate (rt-VBR), non-real time variable bit rate (nrt-VBR), unspecified bit rate (UBR), available bit rate (ABR), guaranteed frame rate (GFR). Each of them is designated for handling specified type of traffic (streaming, elastic) with assumed QoS objectives (concerning to cell/frame loss and/or delay) and has its own traffic control rules (traffic contract specification and policing, admission control). Among them, the UBR service only does not require traffic flow control mechanisms, since was designed as a best effort service. The rest of NSs provides QoS guarantees and requires the user to make some traffic declarations during set-up phase.

Let us remark that ATM NSs were specified with paying attention rather on types of possible traffic occured in the network while with loosely focus on the traffic generated by applications. As a consequence, since applications available in LANs are IP-oriented, a mapping between IP and ATM is needed, covering such aspects like QoS and traffic contract definitions (between packet and ATM cell level), encapsulation, connection set-up, etc.

An alternative for ATM is the IP QoS concept, which is regarded as more promising solution for seamless inter-working with IP-based applications. For the IP QoS two architectures were proposed: (1) integrated services (IntServ) [4] and (2) differentiated services (DiffServ) [3]. As IntServ architecture suffers from scalability and can be implemented in rather small networks, DiffServ approach suits well for building WAN networks. Therefore, we focus on DiffServ network concept, more specif-

ically on its representative implementation provided inside AQUILA IST European project [1, 2]. For now, four types of packet flows have been recognised as typically emitted by applications available to a LAN user and requiring QoS guarantees. They are as follows: (1) streaming constant bit rate (e.g. VoIP), (2) streaming variable bit rate (e.g. video applications), (3) elastic, produced by greedy long-live TCP or TCP-like sources (e.g. FTP), and (4) elastic, non-greedy short-live TCP sources (e.g. WWW). In this spirit, four QoS NSs have been defined and implemented in AQUILA: premium CBR (PCBR) for traffic (1), premium VBR (PVBR) for traffic (2), premium multimedia (PMM) for traffic (3), and premium mission critical (PMC) for traffic (4). Each network service is optimised for specific type of packet flows and has its own traffic control mechanisms, including admission control. In addition, standard (STD) service for best effort traffic is also provided. Obviously, one can find some similarities between NSs available in ATM and IP QoS (AQUILA). Anyway they differ in this that ATM NSs operate on cells (53 bytes packets) while NSs in IP QoS take into account packets of different length. This gives rather some advantages for ATM due to better multiplexing and simplest switching.

The investigated network is ATM/IP QoS core interconnecting a number of LAN Ethernet networks, as depicted in Fig. 1. The core offers different NSs, according to used technology. Let us assume that a LAN user is interested in getting adequate QoS from the network depending on the type of application he uses. This can be achieved only by traffic classification mechanism implemented in the edge device (edge router or edge ATM switch) and, furthermore, submitting selected traffic flow to appropriate NS, available in the core.

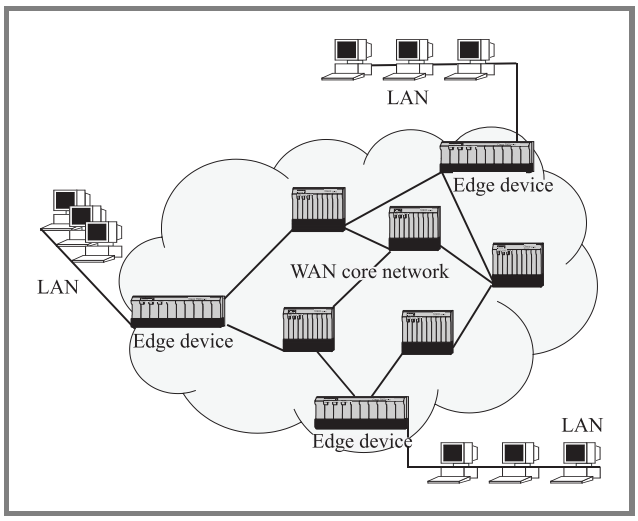


Fig. 1. Network architecture.

The paper presents the experimental results, measured in the test bed, corresponding to QoS level and QoS differentiation provided by ATM/IP QoS core. For this purpose, a set of representative applications currently available to a LAN user was selected demanding from the core different QoS. They correspond to streaming applications like VoIP with QoS objectives represented mainly by packet delay characteristics and elastic applications controlled by TCP protocol with minimum guaranteed throughput/goodput as target.

The paper is organised as follows. Characterisation of traffic profiles and QoS demands corresponding to applications available in LAN is presented in Section 3. Section 2 summarises network services available in ATM and IP QoS and compare them from the point of view of supported traffic profiles and QoS objectives. Furthermore, Section 4 introduces us to mapping rules of LAN traffic into network services with associated mechanisms like traffic classifiers, shaper and schedulers. The measurement results, showing effectiveness of ATM and IP QoS network services for handling LAN traffic with QoS differentiation are included in Section 5. Finally, Section 6 summarises the paper.

2. Types of applications in LAN

Now, a LAN user has access to a variety of applications. Table 1 shows proposed classification of applications with respect to QoS requirements [5, 6, 9] and type of emitted traffic. It assumes four types of applications classes, which are:

- Class 1: emitting elastic sporadic traffic, e.g. WWW, e-mails, etc. The applications send short messages with data flow controlled by TCP. A user is interested in short transfer time/transaction time.

- Class 2: emitting elastic bulk traffic, e.g. file transfer by FTP. The data transfer lasts relatively long (say minutes). A user wants to transfer the file in predictable time interval.

- Class 3: emitting streaming variable bit rate, e.g. video, VoIP. A user is satisfied with such application if no significant packet transfer delay and packet loses will occur. For instance, in the case of VoIP similar QoS is expected as in telephone network.

- Class 4: emitting streaming constant bit rate, e.g. virtual leased line (VLL). In this case, a circuit emulation service is required.

Concluding, a user is satisfied with applications available in LAN if the core network would guarantee adequate quality of packet transfer. This can be achieved by best effort network but only if it is significantly over-dimensioned. Other solution is to support by core a number of NSs, each of them supporting QoS level appropriate for given application.

Table 1
Application classes

Applications	Required bit rate [kbit/s]	QoS requirements		Application class
		allowed packet transfer delay	allowed packet loss rate	
WWW	Up to 100	Medium	Low	Elastic sporadic traffic
E-mails	Up to 50	High	Low	
Chatting	A few	High	Low	
Telnet	A few	Medium	Low	
Data base access	A few	High	Low	
FTP	Up to 1 000	High	Low	Elastic bulk traffic
Virtual reality environment	Up to 128	Medium	Low	Streaming variable bit rate
Video on demand	Up 512	Medium	Low	
Video broadcasting	Up to 3 000	Low	Low	
Videoconferences:				
– video	–n*128	Low	Medium	
– audio	–8–32	Low	Medium	Streaming variable or constant bit rate
IP telefony	Up to 64	Low	Medium	
VLL	Up to 2 048	Low	Low	Streaming constant bit rate

3. Application classes versus network services in ATM and IP QOS

Summarising, we have from one side a number of application classes and from other side a set of NSs supported by core. Therefore, the problem is mapping applications into adequate NSs in the way satisfying user. Let us recall, that the available NSs in ATM and IP QoS are slightly different. Table 2 shows the proposed mapping assumed for the experiments. Although in ATM we have 6 NSs, we had to limit our interest to 4 NSs only (i.e. CBR, rt-VBR, nrt-VBR and UBR), since the implemented in a switch ABR as well as GFR services are not available to applications. Corresponding to AQUILA IP QoS, the tested NSs are premium CBR, premium VBR, premium multimedia, premium mission critical and STD.

Table 2
Proposed mapping between application classes and NSs in ATM and IP QOS

Application class	ATM network service	IP QoS network service
Elastic sporadic traffic	nrt-VBR	Premium mission critical
Elastic bulk traffic	nrt-VBR	Premium multimedia
Streaming variable bit rate	rt-VBR	Premium VBR
Streaming constant bit rate	CBR	Premium CBR

The proposed mapping takes into account the traffic profiles produced by particular application classes jointly with QoS requirements and capabilities of NSs. For simplifying experiments with ATM, we have merged both elastic traffic classes into single one assigned to nrt-VBR service. A justification for doing it is that the data flow in elastic traffic is controlled by the same protocol, TCP. The nrt-VBR is designed for guaranteeing assumed cell loss ratio (and in non-direct way – TCP throughput) while cell transfer delay is not an objective. In the case of streaming variable/constant bit rate classes the mapping into adequate NSs is more obvious. The variable/constant bit rate traffic is submitted to rt-VBR/CBR in ATM or premium VBR/premium CBR in AQUILA IP QoS. The QoS objectives for the considered NSs are almost the same as QoS application requirements. Notice, that the NSs mentioned above require the user to set-up the connection with appropriate traffic declarations, corresponding to the single or double leaky/token bucket parameters. They are the peak bit rate and/or the sustained bit rate jointly with the maximum burst sizes.

4. Mechanisms for splitting LAN traffic into particular network services

Handling traffic generated by a LAN with requirements for QoS differentiation in the core, and as consequence, splitting it into appropriate NSs, demands implementation of additional mechanisms in the edge ATM switch or IP router. These mechanisms should allow us for: (1) setting up appropriate connection/reservation in a core, and (2) LAN traffic classification into sub-streams and mapping them into the established connections inside adequate NSs.

A connection in the core can be set-up by network operator or on demand by a proxy agent. Classification process will allow us for selecting the traffic sub-streams and then

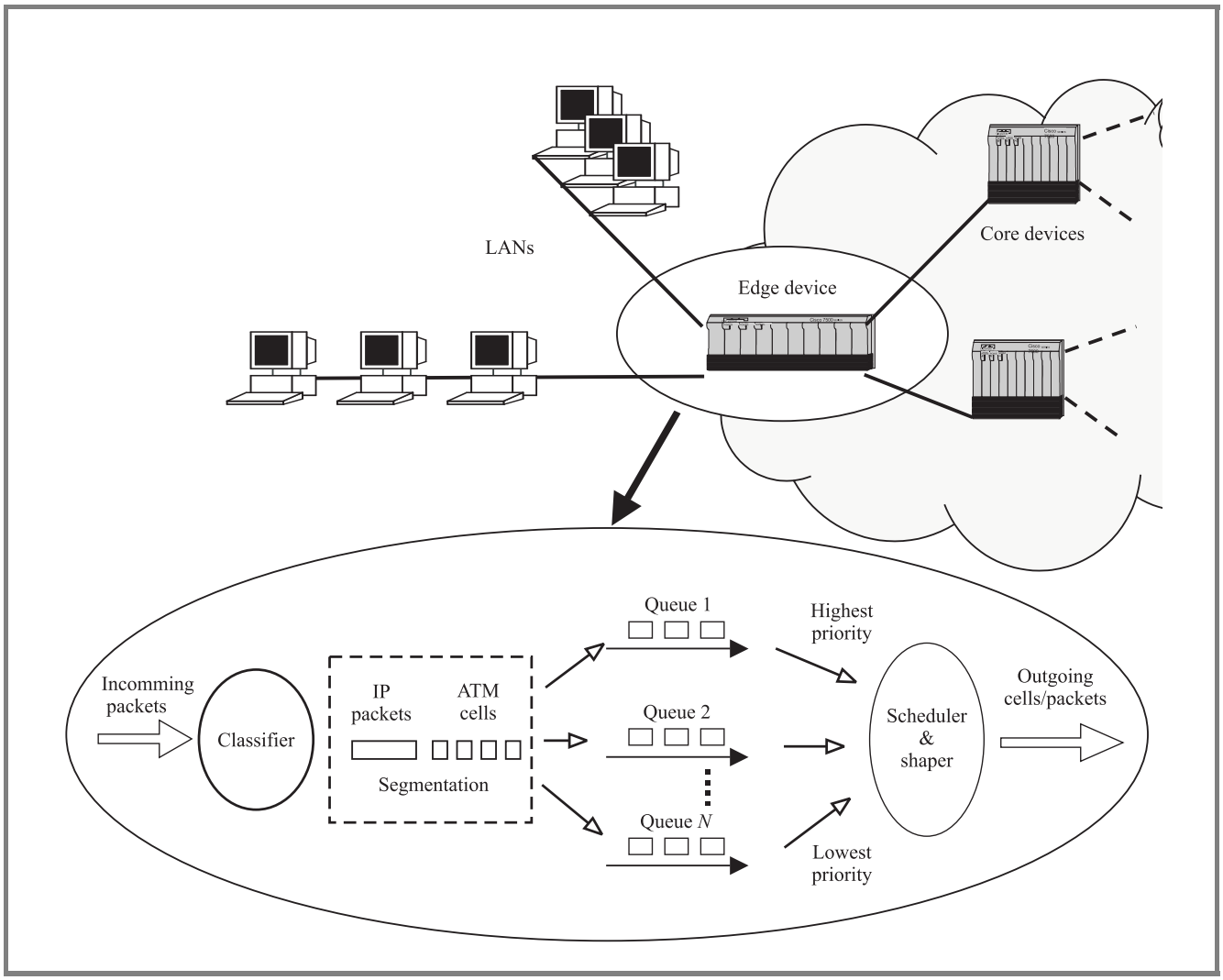


Fig. 2. Scheme for handling LAN traffic.

transferring it by the established connections belonging to a given NS. Let us remark, that each of the traffic substreams before submitting to the core should be shaped in accordance to assumed traffic contract. Figure 2 shows the scheme for traffic handling in the edge device.

First, outgoing LAN traffic is submitted to a classifier for splitting it in accordance to QoS demands (or equivalently, according to assigned NS). This process can be done on the basis of address information included in IP packet header, TCP/UDP segment header etc. Anyway, this requires implementation of a classification table, which should be updated each time a connection is set-up/released. After the classification process, the IP packets are switched to the appropriate input queues in the scheduler, governing access to the core link. Additionally, in case of ATM a packet segmentation process into ATM cell format is performed. The queues in scheduler are associated with a given NS. Usually, in ATM switches we have priority queuing (PQ) or like-PQ schedulers [9], allowing us for assigning priority in such a way that the highest is assigned to CBR, lower to rt-VBR, next to nrt-VBR and the lowest to UBR. In IP routers the most popular is the weighted fair queuing (PQ-WFQ) scheduler [1], as assumed e.g. for AQUILA IP QoS. The PQ-WFQ gives similar prioritization of traffic submitted inside premium CBR, premium VBR, premium multimedia, premium mission critical and STD.

5. Measurement results

This section presents comparative measurement results corresponding to QoS differentiating of traffic generated between LAN networks carried by ATM and IP QoS core. More specifically, we will focus on the quality perceived by a LAN user using different applications. The applications selected for the tests are VoIP, e-mails and FTP. The experiments were carried out in a test bed. The tested network is of the bottleneck type, as depicted in Fig. 3. It consists of two ATM switches (MARCONI ASX200BX), connected by direct E1 ATM link, 1.9 Mbit/s. To each

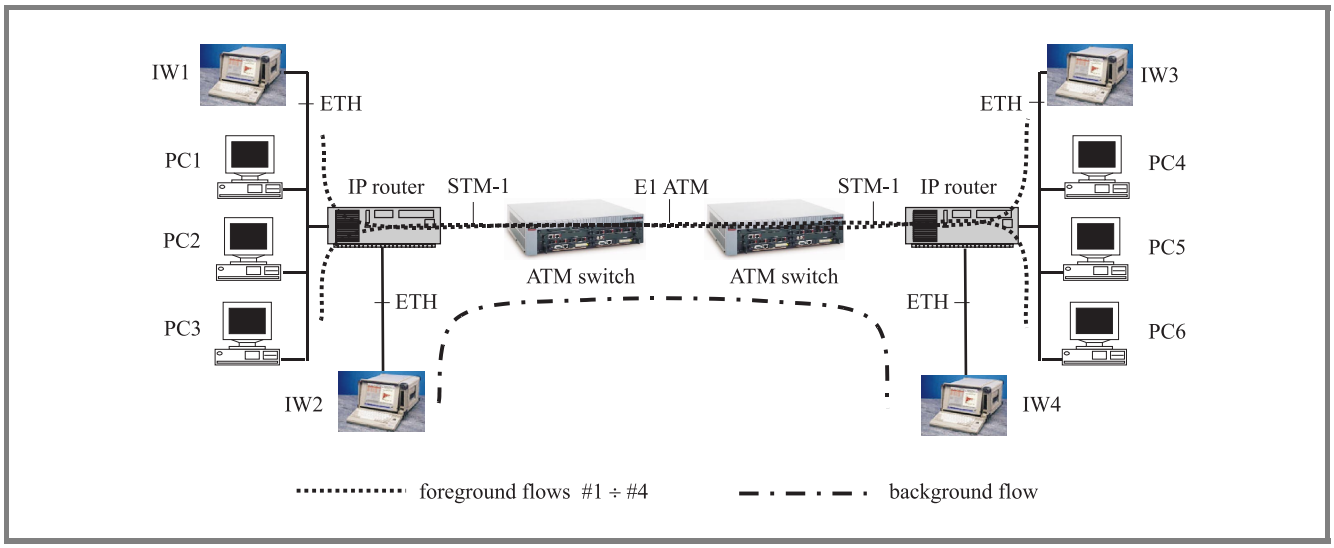


Fig. 3. Tested network.

ATM switch a LAN network is attached, containing IP router (CISCO 3640) as a gateway for 3 user terminals, PC1–PC3 (PC4–PC6), and 2 traffic generator/analysers (InterWatch 95000), IW1–IW2 (IW3–IW4).

Two network scenarios were considered, which are:

1. ATM scenario, where traffic is differentiated in ATM switch for further submitting to earlier established connections inside NSs available in ATM, i.e. CBR, VBR and UBR. In this case, the IP router performs simple packet forwarding only without applying any QoS features. It can be treated as completely transparent.

2. IP QoS scenario, where traffic is differentiated in IP router for further submitting to earlier established connections inside NSs available in IP QoS, i.e. premium CBR, premium VBR, premium multimedia, premium mission critical and STD. In this case, a single ATM connection is designated for whole IP traffic. So, the ATM is transparent.

The assumed foreground and background traffic flows carried by bottleneck link are the following:

- Flow #1, produced by UDP controlled application which requires low packet losses and low packet transfer delay, like VoIP, is established between traffic generator/analysers IW1 and IW3. This traffic is of constant bit rate type with 64 kbit/s in the peak. In this case we assume short IP packets of 53 B (bytes) size for a fair comparison between ATM and IP QoS.

- Flow #2, produced by "non-greedy" TCP source using application sending 10 kB messages, like e-mail, is established between terminals PC1 and PC4. This traffic is shaped according to contract with the peak rate PR = 100 kbit/s, the sustained rate SR = 100 kbit/s, and burst size BS = 10 kB. The packet size is 1 500 B.

- Flow #3, produced by "greedy" TCP source using application transferring large files of 5 MB, like FTP. This flow is established between terminals PC2 and PC5 and its traffic is also shaped to the same contract as for flow #2 with packet size also fixed to 1 500 B.

- Flow #4, is exactly the same traffic as flow #3, but established between terminals PC3 and PC6.

- In addition, the background traffic is submitted into best effort service, UBR in ATM and STD in IP QoS. This traffic is of constant bit rate type with the peak rate 2 Mbit/s, produced between pair of traffic generator/analysers, IW2 and IW4. The presence of this traffic produces overload in the bottleneck link 1.9 Mbit/s.

Table 3 summarises the assumed for experiments traffic flows, with specification concerning traffic contact parameters and affiliation to network services in ATM and AQUILA IP QoS. The NS affiliation follows the consideration included in Section 3 (see Table 2).

The measured parameters are:

– for TCP-controlled flows #2, #3 and #4, throughput and goodput,

– and for UDP-controlled flow #1, packet transfer delay characteristics: max packet transfer delay (max PTD), peak-to-peak packet delay variation (PDV) and packet loss rate (PLR).

The reported measured results were collected after 10 independent measurement intervals each of 5 min. They are presented with 95% confidence intervals. For each scenario, two experiments were performed, with and without background traffic.

Table 3
Types of flows assumed for experiments

Flows	Connection	Traffic contract	Assigned ATM network service	Assigned AQUILA IP QoS network service
Flow #1	IW1–IW3	Constant bit rate with peak rate 64 kbit/s	CBR	Premium CBR
Flow #2	PC1–PC4	Variable bit rate with peak rate 100 kbit/s, sustained rate 100 kbit/s, maximum burst size 10 kB	nrt-VBR	Premium mission critical
Flow #3	PC2–PC5	Variable bit rate with peak rate 100 kbit/s, sustained rate 100 kbit/s, maximum burst size 10 kB	nrt-VBR	Premium multimedia
Flow #4	PC3–PC6	Variable bit rate with peak rate 100 kbit/s, sustainable rate 100 kbit/s, maximum burst size 10 kB	UBR	STD
Background flow	IW2–IW4	Constant bit rate 2 Mbit/s	UBR	STD

Table 4
Comparative measurement results for ATM versus IP QoS core scenario

Scenario 1: ATM core									
Test case	flow #1			flow #2		flow #3		flow #4	
	max PTD [ms]	PDV [ms]	PLR [%]	throughput [kbit/s]	goodput [kbit/s]	throughput [kbit/s]	goodput [kbit/s]	throughput [kbit/s]	goodput [kbit/s]
Without background traffic	2.12	0.71	0*	73.2 ÷ 76.7	69.7 ÷ 74.2	79.8 ÷ 90.1	77.2 ÷ 85.4	77.2 ÷ 83.6	74.9 ÷ 81.4
With background traffic	2.2	0.91	0*	73.1 ÷ 76.5	68.9 ÷ 73.5	79.7 ÷ 89.7	76.8 ÷ 84.8	—	—
Scenario 2: IP QoS core									
Test case	flow #1			flow #2		flow #3		flow #4	
	max PTD [ms]	PDV [ms]	PLR [%]	throughput [kbit/s]	goodput [kbit/s]	throughput [kbit/s]	goodput [kbit/s]	throughput [kbit/s]	goodput [kbit/s]
Without background traffic	8.5	8.3	0*	69.3 ÷ 72.7	67.2 ÷ 69.5	85.6 ÷ 88.9	83.4 ÷ 86.3	78.4 ÷ 81.4	76.6 ÷ 79.1
With background traffic	8.9	8.4	0*	68.7 ÷ 71.8	65.4 ÷ 69.0	84.4 ÷ 87.5	78.8 ÷ 81.3	—	—
* No packet losses were observed, — flow starvation was observed.									

The measurement results obtained for Scenarios 1 and 2, and corresponding to foreground flows #1÷4 are collected in Table 4. One can observe that for both considered scenarios the impact of background traffic submitted to best effort service, UBR in ATM or STD in IP QoS, on traffic handled by other NSs (guaranteeing a given QoS level) is negligible, as it was expected. Comparing ATM and IP QoS, we conclude as follows:

- QoS level experienced by flow #1, related with real-time data, is worst in case of IP QoS than ATM. Let us recall that in IP QoS scenario, we mix packets of 53 bytes with packets of 1 500 bytes. Therefore, the packets from flow #1 could experience relatively large delay despite that they are handled with the highest priority. This is due to the packets multiplexing scheme applied in IP routers. So called, residual packet service time in no-preemptive service discipline, as it is in PQ or PQ-WFQ schedulers, could be essential in the presence of long size packets generated by e.g. TCP-controlled applications. This is not observed in ATM, where cells multiplexing scheme is applied.

- The values of achieved goodput in case of TCP-controlled flows #2 and #3 stay on the same level in IP QoS and ATM scenarios. This result was expected. The greater values of throughput/goodput than guaranteed by traffic contract were reached for flows #2 and #3 in both scenarios. This is due to non-dropping but marking policy for TCP-controlled flows and higher priority for nrt-VBR/premium mul-

timedia service than for UBR/STD in ATM and IP QoS, respectively. So, one can conclude that two services nrt-VBR and premium multimedia give similar QoS level.

- In both scenarios, the essential starvation of QoS experienced by flow #4 in observed, when the background traffic is on. This result was expected since for the UBR as well as STD services the lowest priority was assigned and no traffic control mechanisms are applied.

Concluding, the IP QoS can assure similar QoS level as achieved by ATM for TCP-controlled traffic. However, for streaming traffic the CBR service in ATM is more efficient than premium CBR service in IP QoS.

6. Summary

The paper reports the measurements results corresponding to a comparison between ATM and IP QoS in providing QoS differentiation for traffic generated by LAN users. In both cases, this requires implementation of additional functionality corresponding to traffic classification, shaping, connection set-up and mapping between application classes and network services at the entry point to the core network. First observation is that QoS differentiation is possible to be reached by using both considered technologies. The different QoS objectives for streaming and elastic flows can be met by using appropriate ATM or IP QoS network services. More precisely, for streaming flows the low packet transfer delay and low packet loss rate are guaranteed by CBR service in ATM or by premium CBR service in IP QoS. However, due to packet multiplexing scheme in the latter case, the observed packet delays are greater. For elastic flows, the QoS objectives expressed by TCP goodput are achieved in both cases.

References

[1] A. Bąk, W. Burakowski, F. Ricciato, S. Salsano, and H. Tarasiuk, "Traffic handling in AQUILA QoS IP network, quality of future Internet services", *Lecture Notes in Computer Science*. Springer, 2001, vol. 2156.

[2] A. Bąk *et al.*, "AQUILA network architecture: first trial experiments", *J. Telecommun. Inform. Technol.*, no. 2, pp. 3–13, 2002.

[3] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services", RFC 2475, Dec. 1998.

[4] R. Braden, D. Clark, and S. Shenker, "Integrated services in the Internet architecture: an overview", RFC 1633, June 1994.

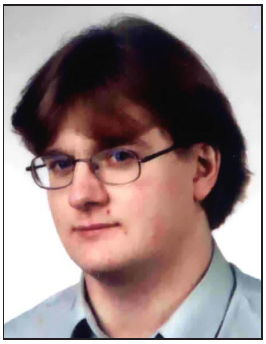
[5] P. Tran-Gia, Ed., "Impact of new services on architecture and performance of broadband networks, comptuTEAM", Final Report COST-257, Wuerzburg, Germany, 2000.

[6] S. Fahmy, R. Jain, S. Rabie, R. Goyal, and B. Vandalore, "Quality of service for Internet traffic over ATM service categories", *Comput. Commun.*, vol. 22, no. 14, 1999.

[7] ITU-T Rec., I.371, "Traffic control and congestion control in B-ISDN", 1995.

[8] J. Kenney, "Traffic management specification". Draft version 4.1, ATM-Forum BTD-TM-02.02, Dec. 1998.

[9] J. Roberts, U. Mocci, and J. Virtamo, Eds., "Broadband network teletraffic. Performance evaluation and design of broadband multiservice networks", Final Report COST 242, 1996.



Andrzej Bęben was born in Poland, in 1974. He received both M.Sc. and Ph.D. degrees in telecommunications from Warsaw University of Technology in 1998 and 2001, respectively. Now he is senior researcher at Military Communication Institute. His research interests include ATM and IP networks.

e-mail: abeben@wil.waw.pl
Military Communication Institute
05-130 Zegrze, Poland

Wojciech Burakowski was born in Poland, in 1951. He received the M.Sc., Ph.D. and D.Sc. degrees in telecommunications from Warsaw University of Technology, in 1975, 1982 and 1992, respectively. He is now a Professor at the Institute of Telecommunications, Warsaw University of Technology. Since 1990, he has been involved in the European projects COST 224, COST 242, COST 257 and 279. His research interests include ATM and IP network design as well as traffic control mechanisms.

e-mail: wb@wil.waw.pl
Military Communication Institute
05-130 Zegrze, Poland

Piotr Pyda was born in Hrubieszów, Poland, in 1972. He received the M.Sc. and Ph.D. degrees in telecommunication from Military University of Technology, Warsaw, Poland, in 1996 and 2003. He has been working for Military Communication Institute in Zegrze since 1997. Now he is an Professor Assistant. His research interests include telecommunication networks and wireless ATM.

e-mail: piotrp@wil.waw.pl
Military Communication Institute
05-130 Zegrze, Poland

Paper

The Integrated Data Environment: a new tool for interoperability and effective data integration for command and control

Jon Wilkes

Abstract — This paper describes an ongoing effort at NC3A to provide one integrated database which contains data from a number of different sources. Initially, these sources are legacy NATO systems. Later, other systems, including messaging interfaces of a wide variety, and national systems, will be added. A common data model is used as the lingua franca between systems. A COTS product has been identified that creates translator boxes to provide interfaces to and from the legacy systems.

Keywords — Integrated Data Environment, data interoperability, common data model, data modelling, data translation, ATCCIS, LC2IEDM, ADatP-32.

1. Introduction

The NATO C3 Agency has responded to customer requirements with the Integrated Data Environment (IDE) project, which has been evolving over the past three years. The intention of the effort is to provide one integrated database which contains data from a number of different sources; in the first place these will be legacy internal NATO systems. Later, other systems, including messaging interfaces of a wide variety, and national systems, will be added as requirements and political concurrence allow. It is foreseen that IDE will play a significant role in the core capability package for the Bi-SC AIS. This paper is based heavily on the work started by the late Martin Krick.

2. The problem

Many of the data exchange problems that have confronted and bedevilled NATO for the past few decades have arisen from the fact that early systems were conceived, developed and implemented as stand-alone, or "stovepipe", systems by groups of users and technicians whose requirements horizon extended no further than the immediate needs of the system on which they were engaged. In the early days, interoperability of data models was not even considered relevant.

As time progressed, and the initial desirability of being able to pass information from one system to another became a more firm requirement, many mechanisms were devised to address these issues, but always with the caveat that the software within the in-service systems, seen to be of such acquisition cost as to be untouchable for interoperability needs, could not be modified to assist in the process of bringing systems together to provide for any meaningful direct exchange of data. In addition, because early systems were so expensive, and therefore made available only to the smallest possible community of users, and because many of the more senior users had no ADP facilities at all, or at most a simple teletype, these early mechanisms were specified to be able to be used in manual environments, leading to the definition of a range of messages. Once again, these message definitions were aimed at encapsulating the specific needs of the group of users responsible for the definition of each message; correlation between messages was not a driving force in the definitions.

3. Previous studies

Many studies were carried out when the nature of the problem became so large that it could no longer be ignored; these studies stressed the need for common standards for data definition, but could not provide low-cost solutions and their conclusions were therefore ignored. In essence, they proposed a "data fusion" approach, which is nowadays seen to be both impractical and unnecessary.

4. The data fusion approach

The principle behind a data fusion approach is to define a single data model, and implement a single database, which will encompass the entire set of data currently held in all existing systems. This approach has some advantages, but also has many more major drawbacks which make it an impractical proposition. If we take two or three existing systems, and create a new database which holds all the data previously held in the three individual databases in accordance with a new all-encompassing data model, then the new database will not be the same as any of the old ones. Each application suite in the original systems must therefore be re-written.

It might be possible to create a database interface package for each system to make the new database appear as the old database, but that too would be substantial effort (and there

would be no *ab initio* guarantee of feasibility) and would represent an additional load for the original system which it might well be ill-suited to handle.
A further major, and potentially even more serious, disadvantage is that if another legacy system were to be added to the fusion set, it may impose changes on the data model which would have a knock-on effect on all current systems within the fusion set. This would lead to potentially exponentially soaring costs, and to management problems of equally soaring complexity. Little wonder that the NATO committees of the time were not persuaded to follow down this route!
The perceived advantages and disadvantages of the data fusion approach can be summarized as:

- single view of all data,
- single physical database from which all applications can draw data,

whereas the disadvantages are:

- need to agree the (large) data model between 19 nations and all NATO HQs and Agencies,
- immediate impact on all legacy systems which are required to conform to the new global data model,
- applicable only for a small number of systems (three or maybe four),
- ongoing management overhead for the fusion schema,
- complexity increases dramatically with the number of systems,
- process becomes unmanageable with large numbers of systems.

It should be noted that the advantages are not matched by any known requirement for all data to be perceived in a single view, nor that there should be a capability of providing a single database implementation which would hold all data; these advantages represent theoretical technical possibilities only. By contrast, the listed disadvantages are very real, not least the political problems associated with the first of those listed. Corresponding agreements in related areas are not famous for the speed with which such agreements have been reached nor for the technical clarity of the final agreements.

5. Other more recent studies

In the last ten years, other initiatives have been taking place on a lower profile basis, and the fruits of their endeavours are now beginning to become visible in a number of places; national implementations based on these initiatives have been put in place and have become sufficiently mature for reasonable projections to be made. Principal of these initiatives is the multi-nation ATCCIS[1] study, sponsored and led by NATO, with active participation at varying levels by eleven nations.
The major outputs of the ATCCIS study to date have been:

- a wealth of well-documented analysis,
- a fully specified data model for information exchange,
- an ATCCIS replication mechanism (ARM) for selective transfers of data between two or more ATCCIS-conformant databases.

The primary achievement of the data modellers is that they recognised that they were endeavouring to specify a data model to facilitate the exchange of information rather than for the design or development of systems; thus the level of detail of the model is appropriate to information exchange, and much low-level data, which would typically be found only in specialist systems, was not included. This separation of "local" data and "global" data has been one of the foundation links of the NC3A work on the Integrated Data Environment.

6. The integration approach

The separation of local and global data leads immediately to the concept of an IDE which addresses only some of the totality of data held in all existing (and future) systems. It also leads directly to the recognition that the IDE can be established (either as a virtual database or as a real one) for new purposes, and that the existing systems can be left with their current databases and database management systems – be they rudimentary or advanced – with the immediate benefit that no changes to those systems are required. Indeed, it became one of the design objectives of the IDE work that the IDE concept should be seen to be non-intrusive from the perspective of any legacy system.
In the integration approach, data are translated from the native (legacy) environment to the common data model of the IDE, so that the translated data subsets reside in a single database or transmission mechanism with one common data model describing all data. We may think of this common data model as a "lingua franca".
The integration approach offers as advantages:

- single view of all global data,
- no impact on legacy systems,
- no requirement to have a single database,
- all future applications can draw global data from existing databases,

[1]The common ATCCIS generic hub 4 data model was forwarded to NATO in 1999. NATO initiated a standardisation process for this data model, now called the Land C2 Information Exchange Data Model (LC2IEDM). The respective STANAG 5532 (ADatP-32) has been submitted as draft and is expected to be agreed in 2001.

- process remains manageable with large numbers of systems,
- ongoing management overhead for the integrated database is much smaller than for the fusion approach,
- technology is mature and in use in large commercial organizations,

and as disadvantage:

- as of end 2001, the technology has not been proven within a NATO operational system (but a demonstrator has been produced, and is clearly scalable to full operational use).

It may be seen that almost all of the disadvantages of the fusion approach have been stood on their heads for the IDE approach. The single view of all data, which was never supported as an operational requirement, has been scaled down to become a single view of all global data, for which operational requirements most certainly exist. The previous high impact, in terms of both cost and operational implications, of the fusion approach, has become a zero impact on those systems. The management problems remain tractable.
On the disadvantage side, the technology has not yet been tested in a full NATO operational environment, but a four-system demonstrator has been produced, and the technology is scalable to encompass a very large number of systems. In particular, the technology ensures that the management problems remain at the one-system level, and therefore do not grow as the number of systems being integrated expands.

7. Alternative techniques

There are two techniques available to implement the IDE function, data mediation and data translation. Data mediation works by first making associations of the meta-data of the data sources and the data sink, and then automatically converting source data to the sink on the basis of these pre-determined associations. In principle, this is a very powerful technique; however, at the present time the technology is still in the research stage, with academic institutions producing small-scale demonstrations. No proposals for a full-scale demonstration have come to our notice at this time. The technology is thus considered to be far too immature to be considered for introduction to NATO at the present.
By contrast, data translation is a very much more mature technology which has been in use in commerce for some time. Most of those applications have been for data warehousing applications, but some applications have been for genuine data integration applications. Where the translation process is carried out on a one-translator-per-system basis, there are very few problems about scaling to multiple systems. The scaling problems are mainly associated with the suitability of the sink data model for the spread of data types to be found in the source systems; in this respect, the highly generic nature of the ATCCIS data model is of immense benefit in minimising such risks. Finally, it must be emphasised that both techniques act on the conversion of data on a one-for-one basis. Data aggregation, data fusion and other application-level functions are outside the scope of both technologies.

8. The IDE architecture

Figure 1 gives a very simplified overview of the IDE architecture resulting in the use of translation techniques on a translator-per-system basis. Data from each legacy or national system is processed by its own local translation process to the target (sink) data model and added to the data model of the target system by normal database update techniques. The translation mechanism is a process, implemented as a software package; although for simplicity it is shown in this slide as though it were a separate system, it could equally well be hosted on the legacy system if that were to prove to be the preferred option. However, to emphasise the "No impact" concept, we always show it as a separate system.
Because the translator process will only translate data about which it has been provided with appropriate translation data (which is another form of meta-data), it acts as a simple form of guard against the accidental translation of data which is not to be released. However, the translator process makes no claims to be an approved guard, and additional security devices would normally be expected to be fitted by national authorities to protect national systems which may contain nationally-sensitive data. These would typically be positioned between the national system and the translator.
Both the initial configuration of the translator, and any subsequent upgrades or changes to a national system will require detailed analysis of the source system in order to specify the translation meta-data. For this reason, the configuration of the translators is expected always to be done by the nation concerned. Figure 1 thus shows the translators residing in the national management domain, with the exception of the specification of the output format (ATCCIS conformant) which is essentially public domain.

9. Work done by NC3A

The preliminary study on data mediation carried out in 1998 showed that the technique held potential for complex translation situations, and for the tracking of changes to databases. A simple demonstration system was created, using the most rudimentary meta-data, which was shown at JWID-99. Much interest was demonstrated by visitors at the ability to show data from three different systems out of a common database in response to a single query, with the consequential ability to provide for integrated data solutions.

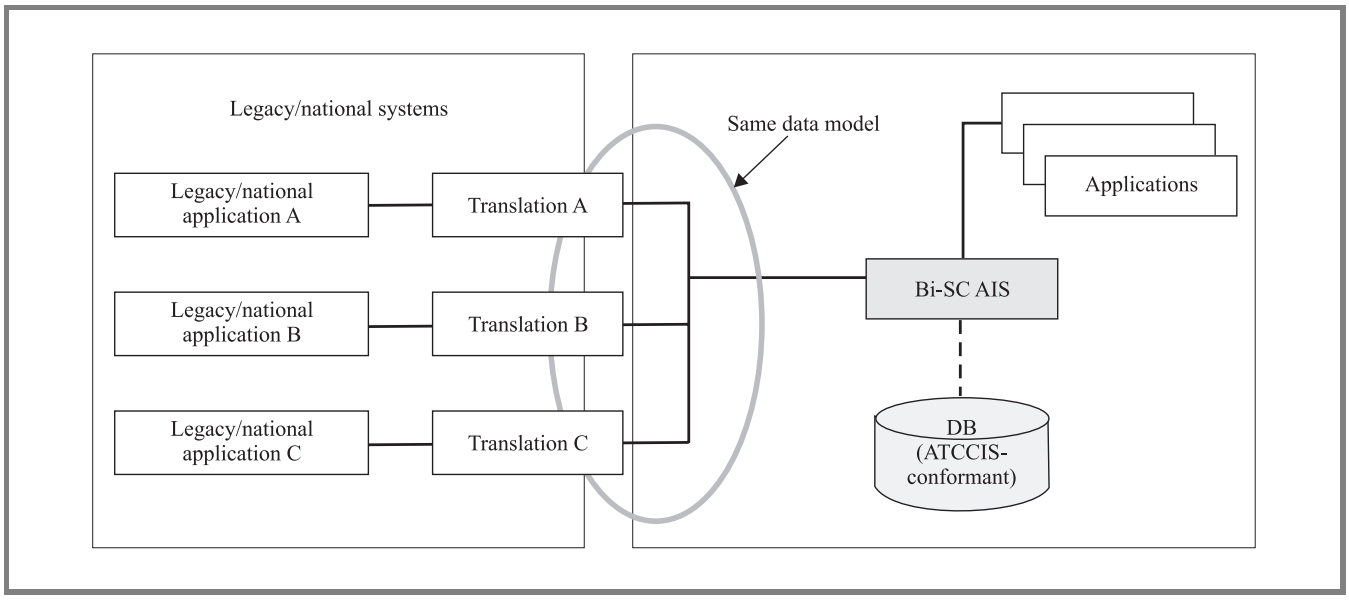


Fig. 1. The IDE architecture.

Evaluation of a contractor report made clear that, although the concepts behind the data mediation technology were both powerful and useful, the technology was very immature with no commercially available implementations of a data mediator product, and little prospect of any such products appearing in the market for some considerable time. Data mediation may have benefits for special situations in the future, yet to be assessed and proven.

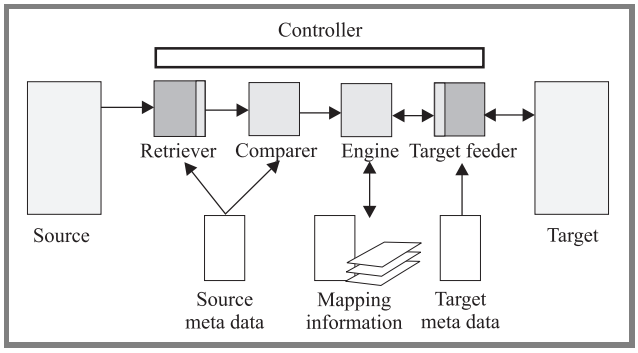


Fig. 2. The architecture of the translator box.

At the same time, an investigation was made of other products, all of which proved to be data translator systems, and it was determined that this offered a better approach for the near term. A contract was let for the development of a demonstrator using translation technology for display at JWID 2000. Problems with the suitability of the translation proposed by the contractor meant that only a very limited demonstration could be mounted at that time, but a very good tool has since been developed by the contractor as a COTS product, which has proven to be very successful and very flexible. A demonstration held at NC3A in late November 2000 showed the capabilities of this tool, and the design gives confidence for its use in many other situations, including message-oriented environments. A major demonstration was held at JWID 2001 and JWID 2002 at SHAPE.

Figure 2 shows the architecture of the translator box produced by the tool.

10. The selected data model for IDE

The selection of the ATCCIS data model, in the form known as the SHAPE Land Command and Control Information Exchange Data Model (LC2IE DM) proved to be a sound choice. The complex nature of this data model means that the specifications of the translations are themselves more complex, but no instances were found in the work on the four NATO legacy systems where translations could not be specified with alacrity and accuracy.

The NATO data administration group reference model is also based on the ATCCIS model, and is under strict configuration management; the LC2IE DM should similarly be placed under CM while it is being used as an interim measure before the full availability of the NATO reference data model. At the same time, some of the work of the NDAG could usefully be retro-fitted to the LC2IE DM to make it into a joint product, a JC2IE DM; the experience of NC3A and their contractor suggests that the minimal changes for the interim product would be small and easy to define and implement. For the November 2000 demonstration mentioned above it was necessary to add only four low-level entities (naval unit, air unit, naval facility and air facility) and to extend the range of a set of domain values to cover maritime and air factors. The total work took less than a couple of days; to repeat this work under full CM control would take less than one week. The future ATCCIS generic hub 5 may address the problem.

11. The tools used for IDE development

Mention has already been made of the shortcomings of the original analysis tools proposed by the contractor. These tools were designed for data warehousing applications where the primary focus of the tools was to analyse data – often dirty data – for which a data model did not exist. In the IDE situation, data models existed and were well documented (although there were some instances where the semantics of the data were not fully defined). Additionally, in data warehousing applications, the emphasis on fitting all source data into a single data model in the destination system does not apply. It is thus not surprising, with the benefit of hindsight, that the tools were found to be unsatisfactory for the IDE situation.

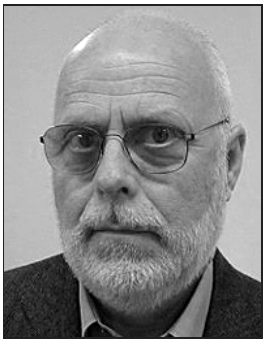
The analytical process involved in determining the translations required is both a skilled process and one which requires time. An analyst familiar with both the source system and the destination data model can complete several source tables each day if the source data model is "clean" and the semantics are fully defined and supported by exemplar data samples. Loose source data models, or a lack of semantic definition, or a lack of sample data, will slow the process to a considerable extent. The tool developed by the contractor provides considerable assistance in converting the results of the analysis into translation rules; future versions are expected to provide some additional assistance to the analysis itself, but cannot fully replace the need for analysis or the analyst.

12. Conclusions

NATO and the nations still have a plethora of incompatible data systems which are likely to remain in service for many years. A fusion approach is not appropriate, and is likely to prove unmanageable and unaffordable.

The Integrated Data Environment provides a response to this information management challenge that is both manageable and affordable, and is eminently suitable for an incremental growth approach.

Commercial off-the-shelf tools are available which support IDE and thus support coalition interoperability, NATO to NATO interoperability, NATO to nations interoperability, and coalition HQ to nations interoperability.



Jon Wilkes is a Senior Analyst Programmer in the Communications and Information Systems Division at the NATO C3 Agency in the Hague. He has been specialising in the interoperability of C2 systems, and the associated problems of the definition of data, for several years at the NC3A. More recently he has been assisting investigations into ways of implementing mechanisms for providing for interoperability between non-compatible systems which have led to the development of the IDE concept and the creation of tools to support the concept.

e-mail: jon.wilkes@nc3a.nato.int
NATO C3 Agency
Postbox 174
2501 CD The Hague, The Netherlands

Paper

NATO automated information system co-operative zone technologies

Martin Diepstraten and Rick Parker

Abstract — The core components of NATO's automated information systems (AIS) include directory services (DS), e-mail, web services, and military message handling systems (MMHS) to exchange information with similar capabilities in NATO's member nation systems or systems that are under control of multi-national coalitions. NATO has developed the concept of information exchange gateways (IEGs) to meet this requirement. This paper introduces the concept of combining symmetric co-operative zones (CZs) to form these information exchange gateways. A generic framework for the co-operative zone network and security architecture is introduced in support of co-operative zone development. It is shown how a co-operative zone network interface can be integrated with the NATO general-purpose segment communications system (NGCS). Development of the NATO co-operative zones is based on an evolutionary approach. A baseline co-operative zone configuration, supporting directory services, e-mail and web services, has been tested and validated on the allied systems interoperability testbed (ASIT). This paper reports the results of the test and validation program. The paper concludes with an overview of planned evolutionary steps for co-operative zone development. Subjects covered in this overview are extension of information services, enhancement of security architecture, and operational deployment (i.e., scalability and manageability).

Keywords — information exchange, firewall technologies, directory services, messaging services, web services, INFOSEC.

1. Introduction

NATO's changing operational environment has caused a dramatic change in the way commanders use their supporting command control and information systems (CCIS) to exchange information. As CCIS's evolve from single-purpose systems in single-level secure environments to multi-purpose systems in multiple-level secure environments, it becomes impossible to build custom interfaces for each possible permutation of information exchange and still remain flexible and responsive to change.

Within NATO's automated information system, the concept of an information exchange gateway through symmetric co-operative zones been introduced with the aim to manage and control all information exchange through a single secure entity with well-defined interfaces.

This paper will focus on the initial architecture of the IEG concept that has been tested and validated in the NC3A allied systems interoperability testbed and the evolution of the concept into operational and more advanced variations.

2. Information exchange gateway operational view

From the operational point of view an IEG can be characterised by two features:

1) the information services that are passed through the gateway;

2) the business case identifying the difference in security level that is bridged by the IEG.

Information services can be end-user related services, such as mail and web, but also be infrastructure or management services, such as domain name service (DNS) and simple network management protocol (SNMP).

Within the scope of the IEG program of work [1] three configuration cases have been identified so far:

- **Case A**. NATO-SECRET to NATO-SECRET enclaves that reside in environments under control of NATO or NATO member nations. There are no interconnections to national CCIS networks. The Case A gateway is sometimes classified as a NATO point of presence (POP). From the POP boundary, the nation has the responsibility to deliver the information service to the user.

- **Case B**. NATO-SECRET to NATO National Secret-HIGH systems that reside in environments under control of NATO member nations.

- **Case C**. NATO-SECRET to coalition secret systems that fall under the responsibility of a Combined Joint Task Force in which NATO is in the lead.

3. Information exchange gateway architecture

The basic conceptual framework of the information exchange gateway through symmetric co-operative zones is illustrated in Fig. 1.

Information exchange between a sender and receiver, both residing in separate local CCIS networks, will always take place via the sender's own CZ and through a symmetric CZ under control of the receiver. No direct traffic is allowed between two local CCIS networks other than that relayed through the source and destination CZs.

The example in Fig. 1 shows this information flow from a nation-X CCIS supporting service A to its counterpart in

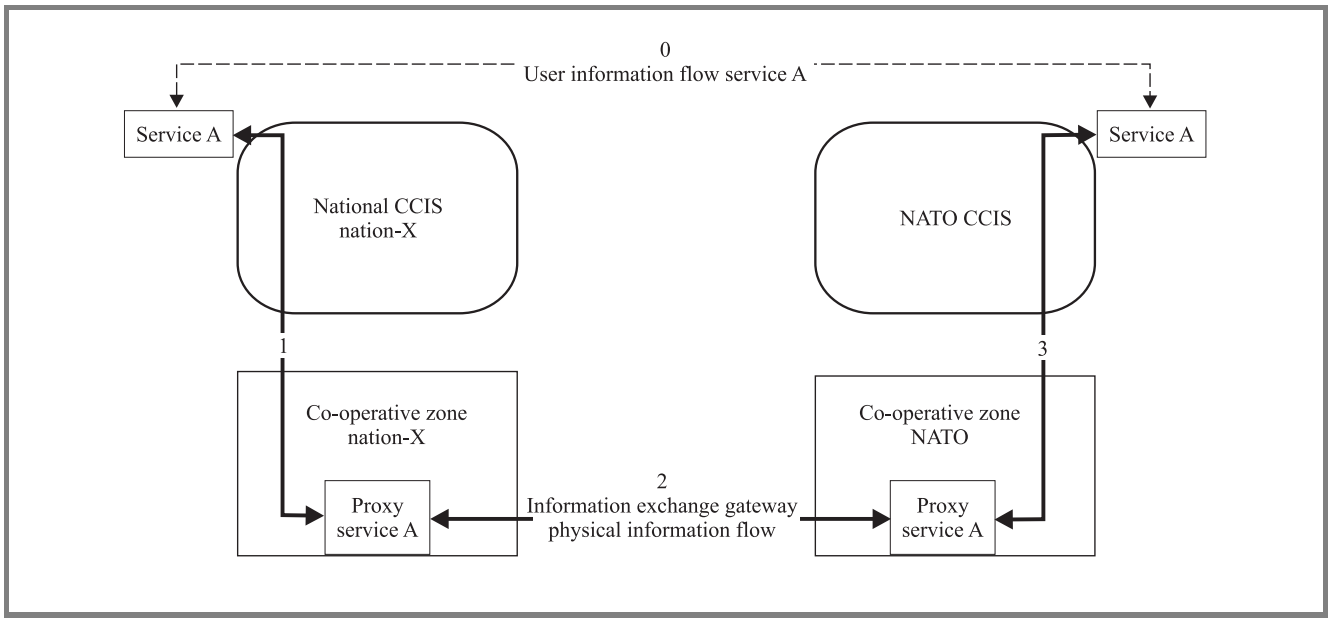


Fig. 1. Information exchange gateway through co-operative zones.

the NATO local CCIS (dashed information flow labelled 0). This information flow that consists of 3 logical connections (labelled 1 up to 3 in Fig. 1):

1) nation-X CCIS service A to nation-X CZ proxy service A;

2) nation-X CZ proxy service A to NATO CZ proxy service A;

3) NATO CZ proxy service A to NATO CCIS service A.

The information services that will be shared through a CZ are to be established on a trustworthy network and security architecture. One of the basic principles of trustworthy computing is to work with well-defined restricted interfaces. Another important principle is to avoid unnecessary complexity (keep it simple). Therefore, the architecture that has been adopted for the information exchange gateway employs **symmetric** co-operative zone modules (CZMs) at both ends of the IEG. This symmetry requirement holds the number of CZM interfaces to a minimum.

In addition to symmetry the following design principles are applied for the further development of the CZMs:

- Minimise the number of protocols that run across the IEG.
- Minimise the volume of network traffic overhead that is generated by a certain service.
- Standardise on common protocols.
- Avoid services or features that carry great risk with respect to security vulnerabilities.

The following categories of IEG architecture will be explained in more detail:

1) security architecture;

2) network architecture;

3) backbone and management services architecture.

3.1. Security architecture

The CZ security architecture most closely resembles a "screened subnet firewall configuration" based on a bastion host that provides authentication and proxy services [2]. Figure 2 illustrates the security elements comprised by a CZM. These are:

- A boundary protection device (firewall) that provides the source environment (i.e., local CCIS) with the protection required under NATO's "self-protecting node" guidance [3].
- A second boundary protection device (filtering router) that manages and secures network paths, protocols and ports to other co-operative zones between service peers at the fixed IP-to-IP-number level.
- An intrusion detection system (IDS) to detect attempted exploitation of (emerging) security flaws that occur in the CZM component systems.

The bastion host capability (providing both authentication and proxy services) has been implemented in the following fashion:

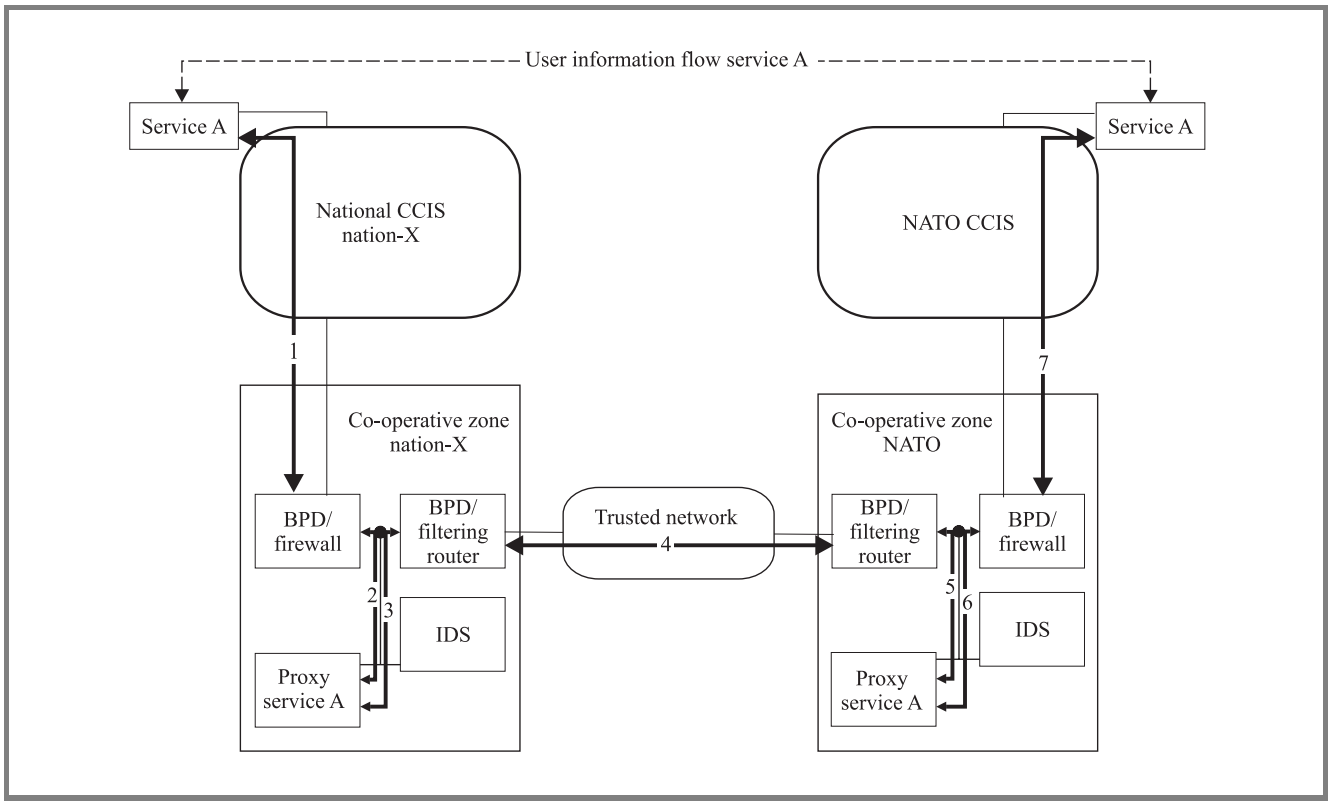


Fig. 2. Co-operative zone INFOSEC components.

- CZ proxy servers to relay traffic from/to the source/target servers in the local CCIS domain (through the BPD/firewall).
- Proxy authentication by the BPD or as part of the local CCIS system.

The chosen approach provides a high level of security, because there are three levels of defence to thwart intruders:

- The filtering router only advertises (limited) IP-numbers of the CZ subnet – the local CCIS network is invisible from the outside.
- The firewall only advertises and supports connection from local CCIS concentrator servers and counterpart proxy servers in the CZ.
- Potential emerging vulnerabilities are pro-actively detected by the IDS.

The user information flow (Fig. 2) is redirected through these three levels by seven consecutive connection steps.

3.2. Network architecture

For Cases A and B, where CZs of NATO or NATO member nation controlled entities are involved, the NATO policies for interconnection [4] prescribe the application of network encryption facilities to establish a trusted interconnected CZ WAN.

NATO will establish a "backbone" infrastructure of NATO CZs to which other, NATO national CZs, can connect. This backbone infrastructure will be based on the draft NATO general purpose communication segment architecture [5]. NATO national sites get connected to the NATO network through the nearest NGCS access router. The CZs will share one common (private) IP-space.

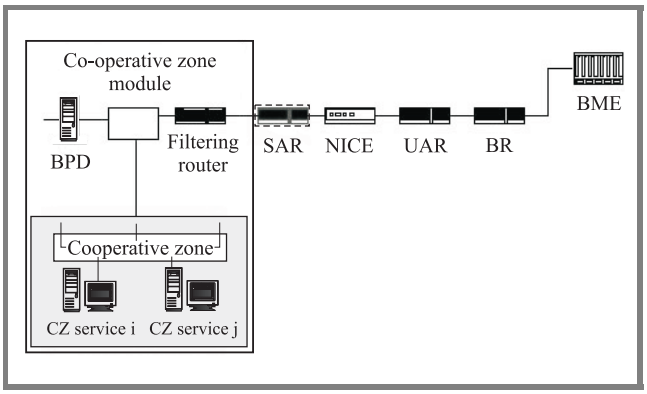


Fig. 3. Co-operative zone module NGCS interface.

Co-operative zones get integrated into the NGSC network through IP-encryption, using the NATO IP-crypto equipment (NICE) [6] and hooking into the nearest NGCS access router. Figure 3 shows which network and security devices a CZ uses to connect to the physical bearer network.

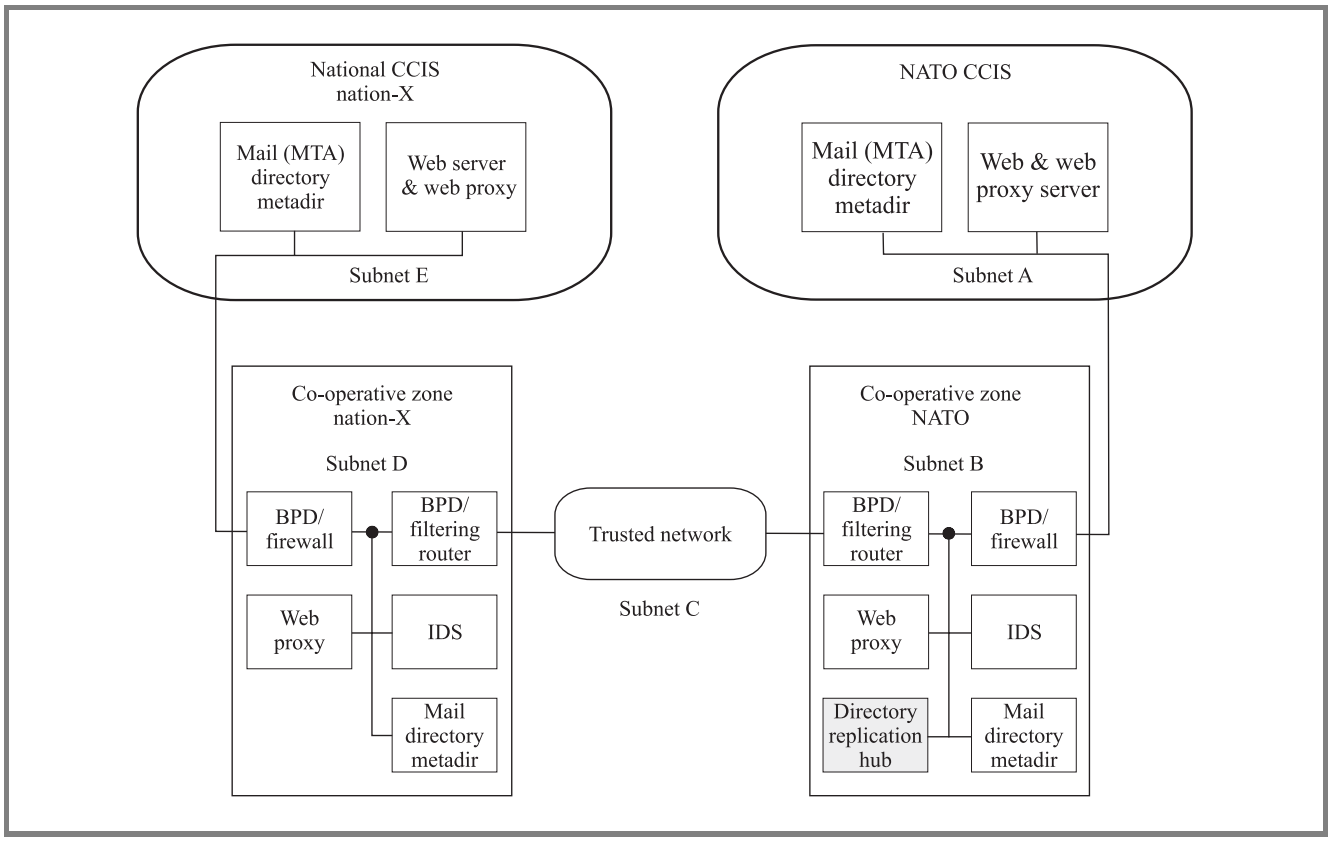


Fig. 4. Information exchange gateway ASIT system diagram.

The configuration comprises the following devices:

- secure access (Fig. 3 – dashed line) router, used for tunnelling through the IP-network;
- the NATO IP-crypto device (NICE);
- an unclassified access router (UAR);
- backbone router (BR);
- bandwidth management equipment (BME).

3.3. Backbone and management services architecture

In addition to the core IEG user services, backbone and management services are needed to maintain the potentially large network of CZM nodes. Examples of backbone services are:

- time services (based on NTP);
- distributed replication services for some of the involved user services and supporting services (e.g. directory services, DNS, etc.);
- redundant network and server backbone infrastructure.

Examples of management services are:

- naming and addressing (performed by the NATO naming and addressing authority);
- network, systems, and services monitoring (e.g. through SNMP);
- software and hardware configuration management and distribution.

4. Baseline configuration

An IEG baseline configuration has been established in the ASIT, to test and validate the concept of an IEG through symmetric CZs. The baseline configuration simulates a Case A IEG providing the following information services:

- e-mail based on X-400;
- web supporting HTTP and HTTPS;
- directory services based on LDAP version 3.

The directory service basically supports the e-mail address book capability, and directory replication is supported.
A system diagram depicting the server, network and security components of the ASIT configuration is shown in Fig. 4. The following IEG network domains were

implemented based on the IP-subnet distribution as shown in Table 1.

Tables 2–4 specify the ASIT components in further detail[1].

The remainder of this chapter describes the detailed configurations and lessons learned of the mail, web, directory, and security services.

Table 1
ASIT IP-subnets

Subnet	Specification
A	NATO CCIS LAN
B	NATO CZ
C	Routing domain in between the back-to-back filtering routers
D	Nation-X CZ
E	Nation-X CCIS LAN

Table 2
NATO/nation-X CCIS components

NATO and nation-X CCIS components	Product specification
Web server	Microsoft IIS 5.0
Web proxy server	MS ISA 2000 server
Mail server (MTA)	Microsoft Exchange 5.5
Directory server	MS Exchange 5.5 GAL
Meta-directory	Microsoft MMS 2.2

Table 3
Nation-X CZ components

Nation-X CZ	Specification
Filtering router	CISCO 2500
IDS	RealSecure
Firewall	Checkpoint Firewall-1
Web proxy	MS ISA 2000 server
Mail server (MTA)	Microsoft Exchange 5.5
Directory server	MS Exchange 5.5 GAL
Meta-directory	Microsoft MMS 2.2

Table 4
NATO CZ components

NATO CZ	Specification
Filtering router	CISCO 2500
IDS	RealSecure
Firewall	Checkpoint Firewall-1
Web proxy	MS ISA 2000 server
Mail server	Microsoft Exchange 5.5
Directory server	DCL (X500)
Directory replication hub	DCL (X500)

[1]The Microsoft Windows 2000 Advanced Server SP2 OS was used unless it is specified otherwise.

4.1. E-mail

The e-mail service is based on X-400. Each CZM contains an X-400 mail transfer agent (MTA), based on the MS Exchange 5.5 product [7], with two X-400 connectors connecting to the local CCIS MTA and a peer co-operative zone MTA. Figure 5 shows the mail flow and Exchange 5.5 site addressing as have been used in the testbed.

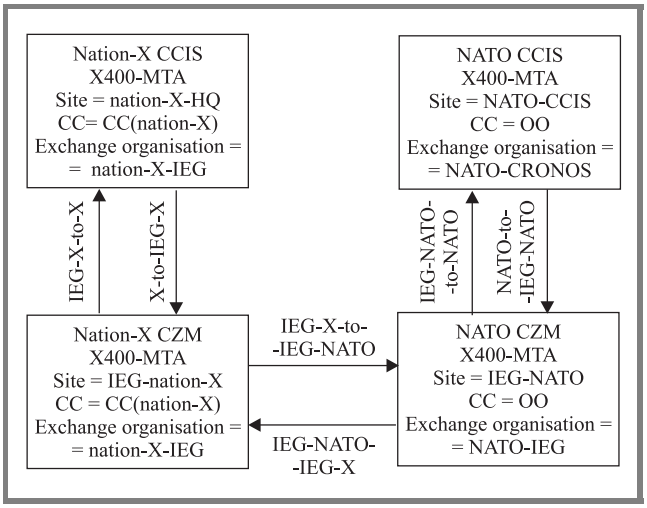


Fig. 5. Testbed e-mail configuration.

Table 5 gives an example specification of the X-400 connector labelled as "NATO-TO-IEG-NATO".

Table 5
Example X-400 connector specification

Feature	Specification
Routing	X400: C=OO;a= ;p=NATO-IEG;o= IEG-NATO;X400: C=CC(nation-X); a=;p=nation-X-CCIS;o=*
MTA conformance	1988 normal mode
X400 link options	BP-15 (in addition to BP-14) Two way alternate Allow exchange format
X400 body part	IA5
	Use the GDI from site addressing

Configuration of the e-mail infrastructure was straightforward. One issue that had to be resolved was related to passing X-400 through the firewall that is configured with NAT. X-400 connectors at both ends need to be configured as if they are communicating on a fixed IP-path. Therefore, an additional firewall rule had to be added enforcing advertisement of a fixed IP-number for incoming X-400 traffic.

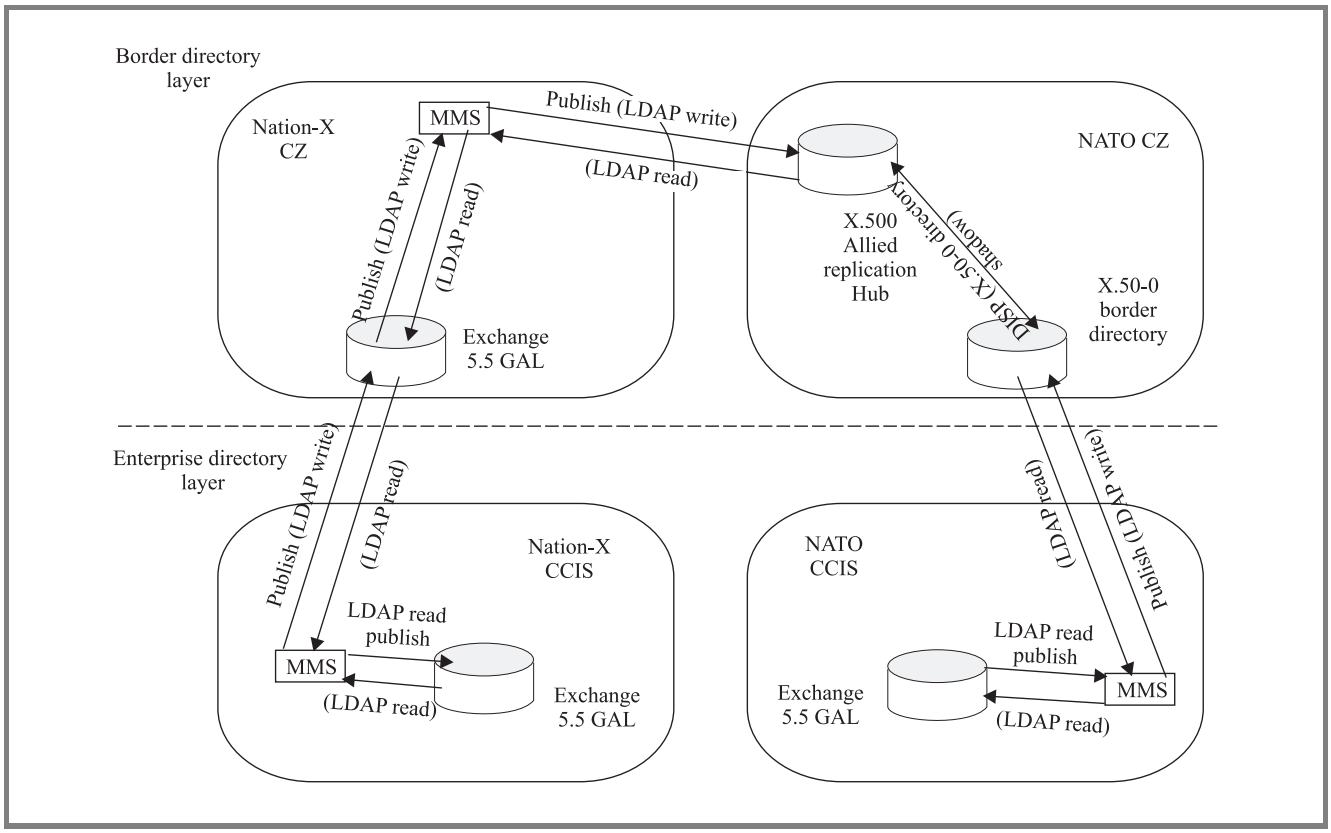


Fig. 6. Directory replication.

4.2. Directory services

The testbed's directory services closely follow the NATO directory services interoperability model [8]. This model is based on three layers of directory services:

1) custom system and application directories (lowest layer);

2) enterprise directory, supporting all enterprise common directory information (middle layer);

3) border directories, created for sharing certain directory information with other organisations (top layer).

NATO and the nations have agreed to use a schema based on ACP 133 [9] in the alliance domain. Also, the current guidance is to use X.500 replication (DISP) to ensure that all the DS data published into the alliance domain is available on every participating border directory service agent (DSA). As every nation is responsible for its own border DSA, it is very likely that they will be based on many different products, and multi-vendor X.500 DISP interoperability is not guaranteed. Some nations may therefore need to employ other replication techniques, e.g. based on meta-directory technology, to ensure that their border DSA is as well populated as those that are able to participate in the automatic X.500 replication.
The filtering and synchronisation processes that control the flow between the application and enterprise layers, and between the enterprise and border layers in the DS architecture, are commonly implemented using meta-directory technology (based on LDAP version 3).
The following mapping of DS interoperability model entities has been implemented in the testbed:

- The NATO (enterprise) and NATO nation enterprise directory are represented as the Exchange 5.5 global address list (GAL), which is an LDAP version 3 readable/writable directory.

- The NATO border directory is based on X-500 (DCL product) [10] and uses the agreed ACP133 schema.

- The nation-X border directory is based on an Exchange 5.5 GAL as a low-cost, easy to implement LDAP readable/writable directory.

- An allied replication hub directory has been implemented to facilitate directory synchronisation of border directories using the directory information shadowing protocol (DISP). This directory provides subtrees for NATO and NATO nations in which only the owner of the information has write access and all others have read access.

- Meta-directory technology has been implemented to facilitate directory synchronisation.

The test exercise that was executed on the testbed was to synchronise e-mail recipients information in support of an "allied recipients" subcontainer of the exchange e-mail address list. Figure 6 shows the directory synchronisation flow.
Information publication was achieved through:

- Publishing releasable mail recipient information from the enterprise directory layer to the border directory layer.
- Shadowing the published border directory mail recipient information (subtree) to the counterpart subtree in the hub directory.

Information download was achieved through:

- Shadowing the mail recipient information in the non-owned subtrees of the NATO replication hub directory the border directory into the equivalent subtrees in the border directory.
- Synchronising the mail recipient information in the non-owned subtrees in the border directory with the "allied recipients" container of the Exchange 5.5 (enterprise) directory.

The two protocols used for directory synchronisation were:

- DISP to synchronise the NATO replication hub X-500 directory information with the NATO border X-500 directory.
- LDAP version 3 to synchronise local CCIS directories with the border directories and to synchronise the nation-X border directory with the NATO hub X-500 directory. For configuration management and control of LDAP based directory synchronisation the Microsoft meta-directory services tool was used [11] by applying the Exchange 5.5 and generic LDAP management agents.

An important lesson learned from the directory synchronisation work is that the directory attribute-flow in between diverse systems (MMS processing rules, X-500 and exchange GAL directory schema) requires a rigorous mapping scheme of attributes and attribute translation rules.

4.3. Web services

Since Case A users require seamless web services, there must be a collaborating chain of local CCIS and CZ web proxy services for HTTP(S) traffic. The web proxy servers are responsible for routing of HTTP-traffic from/to browser to/from the target web server, through the CZ web proxy servers. No direct web server to browser traffic is allowed.
The web proxy server chain (Fig. 7) handles HTTP-requests for a "foreign" web page in the following way:

- An HTTP-request for a NATO CCIS web page hosted by a NATO CCIS web server is made by a nation-X user. It is redirected through the nation-X proxy server to the nation-X CZ proxy server and forwarded to the NATO CZ proxy server.
- The NATO CZ proxy server routes the request to the downstream NATO CCIS proxy server that will then finally connect to the target web server.
- The NATO CCIS target web server response is returned through the NATO web proxy server and the NATO CZ web proxy server to the nation-X CZ web proxy.
- The nation-X CZ proxy routes the response back to nation-X CCIS web proxy. This proxy will then finally route the response to the requesting web browser.

In the testbed experiment the Microsoft Internet acceleration server (ISA) [12] was used to implement the required web proxy server capability.

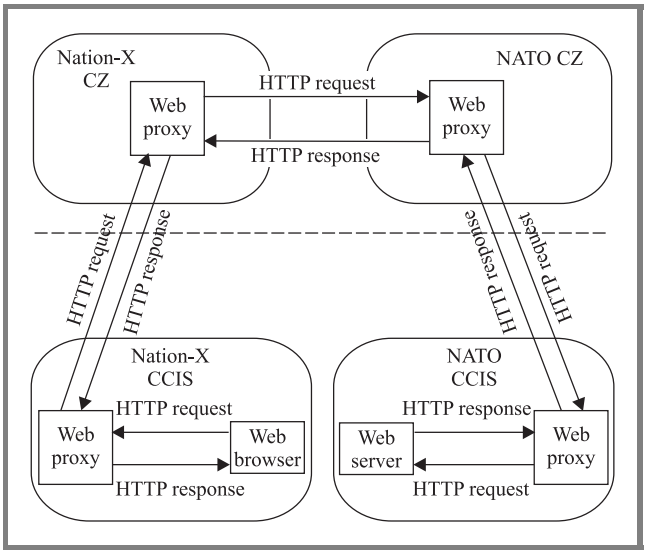


Fig. 7. Allied systems interoperability testbed web proxy chain.

Since the implementation is fully symmetric we will only give the configuration settings of the NATO local CCIS proxy server and the NATO CZ proxy server in Tables 6 and 7, respectively.
The configuration of the web proxy services (routes) was straightforward in the test configuration. It is expected that the live installation will need to be tuned (using caching, etc.) to establish acceptable response times to the end-user. It need to be noted that web browsing based on IP-addresses in "external" domains is not supported through this method.

Table 6
NATO local CCIS proxy settings

Attribute	Setting
Client sets	• NATO clients: subnet A • NATO CZ clients: IP-address NATO CZ web proxy server
Destination sets	• NATO sites: *.nato.int • Nation-X sites: *.nation-x
Protocol rules	• Access allowed to HTTP(S) for NATO clients and NATO CZ clients • Access denied for FTP/Gopher for both NATO and NATO CZ clients
Site and content rules	• Access to nation-X sites allowed to NATO clients • Access to NATO sites allowed to NATO CZ clients • Access to NATO sites allowed to NATO clients
Routing for web browser applications	• Requests to NATO sites retrieved directly from specified destination • Requests to nation-X sites are routed to a specified upstream server: IP-address NATO CZ web proxy server port: 8080

Table 7
NATO CZ proxy server settings

Attribute	Setting
Client sets	• NATO clients: NAT IP-address advertised by NATO CZ firewall (B.x) • Nation-X CZ clients: IP-address nation-X CZ web proxy server
Destination sets	• NATO sites: *.nato.int • Nation-X sites: *.nation-x
Protocol rules	• Access allowed to HTTP(S) for NATO clients and nation-X CZ clients • Access denied for FTP/Gopher for both NATO clients and nation-X CZ clients
Site and content rules	• Access to nation-X sites allowed to NATO clients • Access to NATO sites allowed to nation-X CZ clients
Routing for web browser applications	• Requests to NATO sites are routed to a specified upstream server: NATO IP-address advertised by NATO CZ firewall (B.x) • Requests to nation-X sites are routed to a specified upstream server: IP-address nation-X CZ web proxy server

4.4. Security and network services

In order to achieve the required level of security for Case A the following features have been implemented on the testbed:

- A firewall, based on an EAL-4 assurance level [13] product, checkpoint firewall-1 [14], installed on a C2 configured Windows NT-4 (service pack 3) platform [15]. In the testbed NAT has been implemented to hide the local CCIS system address space. As a measure of verification both NATO and nation-X IP subnets were assigned the same IP address range with identical IP addresses for the mail, directory, and web servers on either side of the firewall. The firewall rules support X-400, LDAP, HTTP, HTTPS.
- The filtering router has been setup to only allow connections between peer servers in the NATO and nation-X co-operative zone (e.g. MTA to MTA, proxy server to proxy server, directory server to hub directory).
- The intrusion detection system based on the realsecure product [16] has been connected to the CZ LAN as a stealth (receive-only) probe and console. The IDS monitors the traffic across the CZ in both directions, detecting incidents of known exploit attempts against the CZ proxy servers or the security components.

One important note to be made is that DNS is not required as a supporting service to resolve names to IP-numbers, because all routing is done based on fixed configured "routes" on an IP-to-IP basis through proxy routes, X-400 connectors, and DISP/LDAP connections.

5. Co-operative zone technologies evolution

The Information exchange gateway configuration that was tested and verified in the ASIT is considered to be a base line configuration. This baseline configuration will evolve in the following technology areas:

- Addition of AIS functional services.
- Adding security services by hardening security and developing Case B and Case C gateways.
- Development of backbone services.

The addition of functional services is a requirement that is very much dependent on further developments in the following areas:

- Migration of present custom interfaces from NATO to NATO nations and coalitions.
- Deployment of new allied systems.

- Operational requirements that lead to a requirement for additional functional services. For example, a new NATO system that will be deployed in the next couple of years is the NATO messaging services (NMS) system. The NMS will introduce the requirement for additional services to pass through the CZ to support a military message handling system (MMHS) [17] and, potentially, a NATO public key infrastructure (NPKI) [18].

Other services that are identified to be developed as a part of the Bi-SC AIS core capabilities [1] might also be candidates to get deployed as gateway services. Examples are:

– conferencing services (based on H323);

– real-time data streams (RMP, RAP);

– distributed database (SQL);

– middleware and XML web services communications features.

The two driving forces behind the further development of security services are:

1) optimisation of security features of the Case A baseline configuration;

2) additional security features required for Case B and Case C.

Optimisation of Case A security services includes:

- Further development of intrusion detection patterns and matching intrusion detection information at a central level.

- Configuring the server installation templates up to the C2 level [13].

- Shielding of the CZ IP-space to the local CIS network IP-space. For this it is considered to implement either NAT from CZ to local CIS or to run an IP-proxy service in front of the BPD.

- Prescribe usage of security tools such as security templates, virus checkers, vulnerability scanners for configuration and operation of information services.

- Develop a concept for centralised monitoring of intrusion detection consoles.

For Case B it is envisioned that the security services will not differ from Case A with the exception of the implementation of web publishing rules. The reverse proxy service will be restricted based on access controls. A request will be authenticated to the BPD/firewall that provides access to the local CCIS web services by impersonating a guest account in the local CCIS domain, based on the authenticated service, group, or individual user account level. The establishment of authentication services may be supported through the implementation of a NATO public key infrastructure [18] and the establishment of an allied PKI interoperability profile. Initially, though, it is anticipated that identification and authentication of authorized users will be left as a locally-selected and operated function, with support from the NATO/national AIS staff as required.

For Case C, the picture for security features looks very different from the Case A security features. Case C is only in a very early stage of concept development and it is anticipated that the services supported across the CZ will begin with 2-way messaging and directory services. Additionally, (one-way) coalition-to-NATO file transfer, much like the current interfaces between NATO and SFOR/KFOR will need to be implemented. Security options for bidirectional file transfer and web services are currently under study.

Scalability and availability are very important features that go together if the amount of interconnected CZs increases. The following solutions are considered to master scalability and availability aspects by establishing an IEG backbone infrastructure:

– redundant CZs per NATO nation;

– multiple NATO CZs (covering regions and are in hot-standby for backup);

– high availability requirements for the underlying network layer (NGCS QOS);

– redundant paths (creating secondary connectors) for MTA's;

– distribution of NATO hub directory.

6. Conclusion

A baseline configuration for an information exchange gateway has successfully been tested and validated in the allied systems interoperability testbed. Lessons learned are taken for the further evolution and the operational deployment of the information exchange concept. The NATO C3 Agency is looking forward to a very busy period with the NATO nations to test and implement the information exchange gateway concept and contribute to allied systems interoperability.

References

[1] Capability Package 5A0050/9B0020 "Provide Bi-SC Static AIS Core Capability".

[2] W. Stallings, "Network Security Essentials, Applications and Standards". Upper Saddle River, NJ: Prentice Hall, 1999.

[3] Paragraph 23 of the "Primary Directive on Security", jointly issued as AC/35-D/2004 under the NATO Security Committee and AC/322-D/0052 by the NATO C3 Board, 17 June 2002.

[4] NATO Information Management Policy (ref. a Annex II to PO(99) 189).

[5] NATO GPS Communications System Programme. Vol. II: Technical System Plan, ed. 3, Jan. 1999.

[6] NATO General Purpose Segment (GPS) Communications System (NGCS), Security Architecture, Version 1.31, 3 May 2001.

[7] Microsoft Exchange 5.5 (SP 4). Microsoft corporation, http://www.microsoft.com/exchange/default.asp

[8] NC3B ISSC DS WG, "NATO directory interoperability model", June 2001, http://nra.nacosa.nato.int/ds/zdocs/dsahwg45.zip

[9] Combined Communications Electronics Board (CCEB) Allied Message Handling (AMH) International Subject Matter Experts (ISME), "Allied Communication Publication (ACP) 133 – Common Directory Services and Procedures", March 2000.

[10] Data connection limited. DS directory version 2.4.01., http://www.dataconnection.com/dirs/diridx.htm

[11] Microsoft metadirectory services, Microsoft corporation, http://www.microsoft.com/windows2000/technologies/directory/MMS/default.asp

[12] Microsoft Internet security and acceleration server 2000 SP1, Microsoft corporation, http://www.microsoft.com/isaserver/

[13] Common criteria website, http://www.commoncriteria.org/cc/cc.html

[14] Firewall 1, checkpoint, http://www.checkpoint.com/products/protect/firewall-1.html

[15] Windows NT 4.0 security set-up for NATO classified systems (NR to NS), Version 4 (includes SP6), INFOSEC Command NACOSA, July 2000.

[16] Realsecure managed intrusion protection service (ISS), http://www.iss.net/products_services/managed_services/service_intrusion.php

[17] Standard NATO Agreement (STANAG) 4406, "Military message handling service edition 1", Dec. 1998.

[18] AC/322 (NPMA-PAC) WP-2, "NATO PKI CONOPS, v. 1.3", 5 July 2001.

Martin Diepstraten is a Principal Scientist in the Core Information Systems Technology branch of NC3A's Communications and Information System Division. Prior to joining NC3A in 2000, he worked on a number of Dutch Government and NATO communication and information system development and integration projects. His current activities focus on operating system kernel services, information systems integration and information interoperability.
e-mail: martin.diepstraten@nc3a.nato.int
Communications and Information Systems Division
NATO Consultation, Command and Control Agency
The Hague, The Netherlands

Rick Parker is a Principal Scientist in the Communications Security Techniques branch of NC3A's Communications and Information System Division. Prior to joining NC3A in 1995, he worked on a number of US and NATO Information Security projects, including R&D, standards and architecture. His current activities focus on security aspects of the CCIS and the underlying networks, vulnerability analysis and forensics.
e-mail: rick.parker@nc3a.nato.int
Communications and Information Systems Division
NATO Consultation, Command and Control Agency
The Hague, The Netherlands

Paper

Military route planning in battlefield simulation: effectiveness problems and potential solutions

Zbigniew Tarapata

Abstract — Path searching is challenging problem in many domains such as simulation war games, robotics, military mission planning, computer generated forces (CGF), etc. Effectiveness problems in military route planning are related both with terrain modelling and path planning algorithms. These problems may be considered from the point of view of many criterions. It seems that two criterions are the most important: quality of terrain reflection in the terrain model and computational complexity of the on(off)-line path planning algorithm. The paper deals with two above indicated problems of route planning effectiveness. Comparison of approaches used in route planning is presented. The hybrid, terrain merging-based and partial path planning, approach for route planning in dynamically changed environment during simulation is described. It significantly increase effectiveness of route planning process. The computational complexity of the method is given and some discussion for using the method in the battlefield simulation is conducted. In order to estimate how many times faster we can compute problem for finding shortest path in network with n big squares (b-nodes) with relation to problem for finding shortest path in the network with V small squares (s-nodes) acceleration function is defined and optimized.

Keywords — battlefield simulation, route planning, shortest paths, effectiveness problems, computational complexity.

1. Introduction

For many years in military applications a simulated battlefield is used for training military personnel. There are at least three ways to provide the simulated opponent:

- two groups of trainees in simulators may oppose each other (often used);
- human instructors who are trained to behave in a way that mimics the desired enemy doctrine (seldom used);
- computer system that generates and controls multiple simulation entities using software and possibly a human operator.

The last approach is known as a semi-automated force (SAF or SAFOR) or a computer generated force (CGF). CGF is used in military distributed interactive simulation (DIS) systems to control large numbers of autonomous battlefield entities using computer equipment and software rather than humans in simulators.
The advantages of CGF are well-known [17]:

1) they lower the cost of a DIS system by reducing the number of standard simulators that must be purchased and maintained;
2) CGF can be programmed, in theory, to behave according to the tactical doctrine of any desired opposing force, and so eliminate the need to train and retrain human operators to behave like the current enemy;
3) CGF can be easier to control by a single person than an opposing force made up of many human operators and it may give the training instructor greater control over the training experience.

As an inseparable part of CGF, modules for route planning based on the real-terrain models are used. For example in modular semi-automated forces (ModSAF) in module "SAFsim", which simulates the entities, units, and environmental processes the route planning component is located [14]. Moreover, automated route planning will be a key element of almost any automated terrain analysis system that is a component of a military command and control system. In the work [1] authors describe a combined on-road/off-road planning system that was closely integrated with a geographic information system and a simulation system. Routes can be planned for either single columns or multiple columns. For multiple columns, the planner keeps track of the temporal location of each column and insures they will not occupy the same space at the same time. In the same paper the hierarchic route planner as integrate part of predictive intelligence military tactical analysis system (PIMTAS) is discussed. In the paper [8] authors presented an on-going efforts to develop a prototype for ground operations planning, the route planning uncertainty manager (RPLUM) tool kit. They are applying uncertainty management to terrain analysis and route planning since this activity supports the commander's scheme of maneuver from the highest command level down to the level of each combat vehicle in every subordinate command. They extend the PIMTAS [1] route planning

software to accomodate results of reasoning about multiple categories of uncertainty. Authors of the paper [3] presented route planning in the close combat tactical trainer (CCTT).

Kreitzberg [11] has developed the tactical movement analyzer (TMA). The system uses a combination of digitized maps, satellite images, vehicle type and weather data to compute the traversal time across a grid cell. TMA can compute optimum paths that combine both on-road and off-road mobility, and with weather conditions used to modify the grid cost factors. The smallest grid size used is approximately 0.5 km. Author uses the concept of a signal propagating from the starting point and uses the traversal time at each cell in the array to determine the time at which the signal arrives at neighboring cells. Other researchers have chosen to decompose the map into regions that are defined by having a constant traversability across the region [1, 9, 16, 19, 20, 27]. The advantage of this approach is that the number of regions will, in general, be far fewer than the number of grid cells. The disadvantages include difficulty in defining the center of the region and the computation difficulties in determining the optimum paths between two adjacent cells. The optimum region-to-region path can be obtained by using either Dijkstra's continuous algorithm (DCA) developed by Mitchell [15]. In many cases, a multiresolution simulation modelling is used to simplify complex battlefield processes [4, 6, 16, 17].

As integrated part of route planning modules the terrain database-based model is being used. Terrain data can be as simple as an array of elevations (which provides only a limited means to estimate mobility) or as a complex as an elevation array combined with digital map overlays of slope, soil, vegetation, drainage, obstacles, transportation (roads, etc.) and the quantity of recent weather. For example in [1] authors describe heterogeneous reasoning and mediator environment system (HERMES) will allow the answering of queries that require the interrogation of multiple databases in order to determine the start and destination parameters for the route planner.

There are a few approaches in which the map (representing a terrain area) is decomposed into a graph [1, 9, 19, 20]. All of them first convert the map into regions of *go* (open) and *no-go* (closed). The *no-go* areas may be considered as obstacles and are represented as polygons. A few ways for consider the map can be used, for example: visibility diagram, Voronoi diagram, straight-line dual of the Voronoi diagram, edge-dual graph, line-thinned skeleton, regular grid of squares, grid of homogeneous squares coded in quadtree system, etc.

Effectiveness problems in military route planning are related both with terrain modelling and path planning algorithms. These problems may be considered from the point of view of many criterions. It seems that two criterions are the most important: quality of terrain reflection in the terrain model (visibility diagram, Voronoi diagram, regular grid of terrain squares, etc.) and computational complexity of the on(off)-line path planning algorithm. The paper deals with above indicated problems of route planning effectiveness.

In the next section we will discuss in details route planning approaches.

2. Comparison approaches used in route planning

It was said in the previous section that we will deal with effectiveness of two problems of battlefield simulation:

- terrain reflection in the terrain model used in battlefield simulation;
- military route planning using one of terrain models.

If terrain models are concerned a few ways for considering the map were listed in the previous section: Voronoi diagram, straight-line dual of the Voronoi diagram (the Delaunay triangulation), visibility diagram, edge-dual graph, line-thinned skeleton, regular grid of squares, grid of homogeneous squares coded in quadtree system.

The polygonal representations of the terrain are often created in database generated systems (DBGS) through a combination of automated and manual processes [19]. It is important to say that these processes are computationally complicated but are conducted before simulation (during preparation process). Typically, an initial polygonal representation is created from the digital terrain elevation data through the use of an automated triangulation algorithm, resulting in what is commonly referred to as a triangulated irregular network (TIN). A commonly used triangulation algorithm is the Delaunay triangulation. Definition of the Delaunay triangulation may be done via its direct relation to the Voronoi diagram of a set, S, of N 2D points: the straight-line dual of the Voronoi diagram is a triangulation of S.

The **Voronoi diagram** is the solution to the following problem: given a set S of N points in the plane, for each point p_i in S what is the locus of points (x, y) in the plane that are closer to p_i than to any other point of S?

The **straight-line dual** is defined as the graph embedded in the plane obtained by adding a straight-line segment between each pair of points of S whose Voronoi polygons share an edge. Figure 1 depicts an irregularly spaced set of points S, its Voronoi diagram, and its straight-line dual (i.e. its Delaunay triangulation).

The **edge-dual graph** is essentially an adjacency list representing the spatial structure of the map. To create this graph, we assign a node to the midpoint of each map edge which does not bound an obstacle (or the border). Special nodes are assigned to the start and goal points. In each non-obstacle region, we add arcs to connect all nodes at the midpoints of the edges which bound the same region. The fact that all regions are convex guarantees that all such arcs cannot intersect obstacles or other regions. Example of the edge-dual graph is presented in Fig. 2.

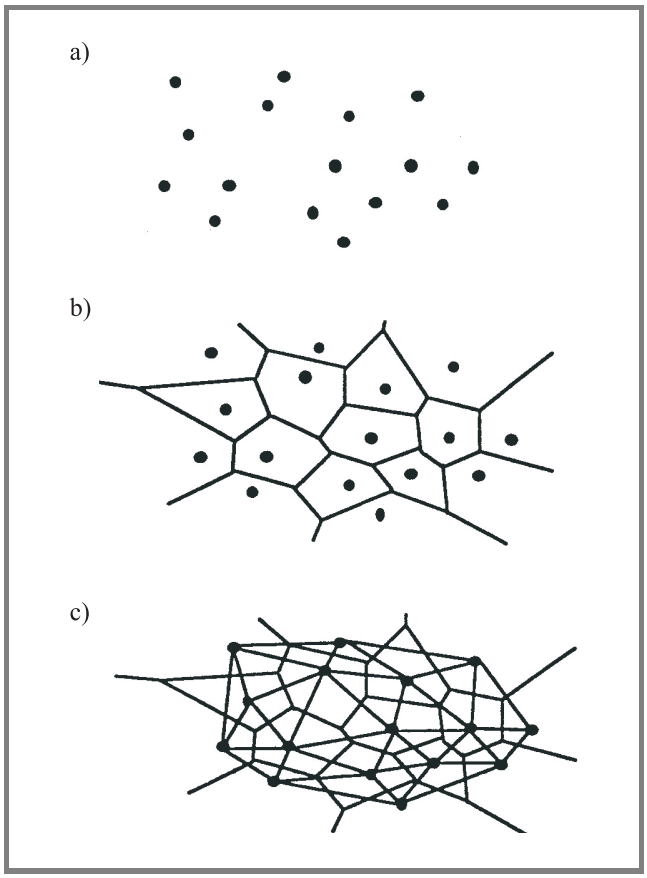


Fig. 1. Voronoi diagram and its Delaunay triangulation [19]: (a) a set S of N points in the plane; (b) the Voronoi diagram of S; (c) the straight-line dual of the Voronoi diagram (the Delaunay triangulation).

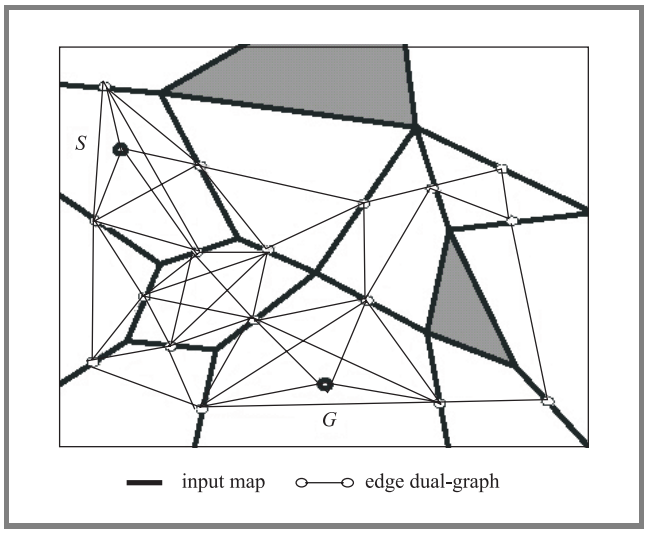


Fig. 2. Edge-dual graph. Obstacles are represented by filled polygons.

The **visibility graph**, is a graph whose nodes are the vertices of terrain polygons and whose edges joint pairs of nodes for which the corresponding segment lies inside polygon. An example is shown in Fig. 3.

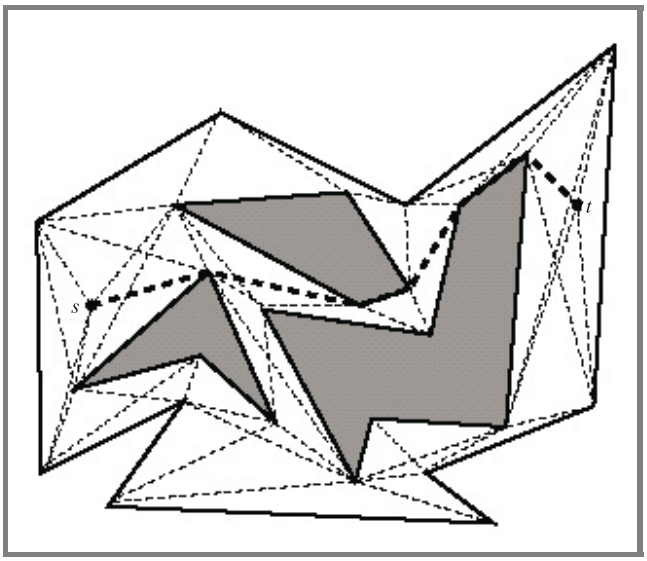


Fig. 3. Visibility graph [15]. There is marked shortest geometric path from source node s to destination t. Obstacles are represented by filled polygons.

The **regular grid of squares** divides terrain space on the squares with the same size and each square is treated as having homogeneity from the point of view of terrain characteristics. An example of this approach will present in the next sections (see Fig. 6 and Fig. 7).

The **grid of homogeneous squares coded in quadtree system** divides terrain space on the squares with heterogeneous size. The size of square results from its homogeneity according to terrain characteristics. Example of this approach was presented, e.g. in [29].

If paths planning approaches used in battlefield simulation are concerned, there are four main approaches [10]: free space analysis, vertex graph analysis, potential fields, grid based algorithms.

In the **free space approach**, only the space not blocked and occupied by obstacles is represented. For example, representing the center of movement corridors with Voronoi diagrams [19] is a free space approach (see Fig. 1).

Advantage of Voronoi diagrams is that they have efficient representation.

Disadvantages of Voronoi diagrams:

- they tend to generate unrealistic paths (paths derived from Voronoi diagrams follow the center of corridors while paths derived from visibility graphs clip the edges of obstacles);
- the width and trafficability of corridors are typically ignored;
- distance is generally the only factor considered in choosing the optimal path.

In the **vertex graph approach**, only the endpoints (vertices) of possible path segments are represented [15].

Advantages:

– this approach is suitable for spaces that have sufficient obstacles to determine the endpoints.

Disadvantages:

– determining the vertices in "open" terrain is difficult;

– trafficability over the path segment is not represented;

– factors other than distance cannot be included in evaluating possible routes.

In the **potential field approach**, the goal (destination) is represented as an "attractor", obstacles are represented by "repellors", and the vehicles are pulled toward the goal while being repelled from the obstacles.
Disadvantages:

– the vehicles can be attracted into box canyons from which they cannot escape;

– some elements of the terrain may simultaneously attract and repel.

In the **regular grid approach**, a grid overlays the terrain, terrain features are abstracted into the grid, and the grid rather than the terrain is analyzed.
Advantages:

– simplification of the analysis.

Disadvantages:

– "jagged" paths are produced because movement out of a grid cell is restricted to four (or eight) directions corresponding to the four (or eight) neighboring cells;

– granularity (size of the grid cells) determines the accuracy of terrain representation.

A many of route planners in the literature are based on the Dijkstra's shortest path algorithm, A* algorithm [7], geometric path planning algorithms [15] or its variants [12, 13, 18, 26, 31, 32]. For example, A* has been used in a number of computer generated forces systems as the basis of their planning component, to plan road routes [3], avoid moving obstacles [10], avoid static obstacles [18] and to plan concealed routes [14]. Very extensive discussion related to geometric shortest path planning algorithms was presented by Mitchell in [15] (references consist of 393 papers and handbooks). Geometric shortest paths problem is defined as follows: given a collection of obstacles, find an Euclidean shortest obstacle-avoiding path between two given points. Mitchell considers following problems:

– geodesic paths in a simple polygon;

– paths in a polygonal domain (searching the visibility graph, continuous Dijkstra algorithm);

– shortest paths in other metrics (L_p metric, link distance, weighted region metric, minimum-time paths, curvature-constrained shortest paths, optimal motion of non-point robots, multiple criteria optimal paths, sailor's problem, maximum concealment path problem, minimum total turn problem, fuel-consuming problem, shortest paths problem in an arrangement);

– on-line algorithms and navigation without map;

– shortest paths in higher dimensions.

3. Effectiveness problems in route planning

We focus one's attention on path planning algorithms and its effectiveness. Path planning algorithms used in battlefield simulation can be off-line or on-line. Off-line path planning algorithms like A* or Dijkstra's algorithm (listed in the previous section) find the whole solution before starting execution (simulation). They plan paths in advance and usually find optimal solutions. Their efficiency is not considered to be crucial and the moved object just follows the generated path. Although this is a good solution for a static environment, it is rather infeasible for dynamic environments, because if the environment or the cost functions are changed, the remaining path may need to be replanned, which is not efficient for real-time applications (e.g. real-time simulation). Let's recall that standard Dijkstra's algorithm has time complexity $O(V^2)$, where V denotes number of nodes in the graph. This complexity may be improved (if the graph is thin) implementing priority queue as binary heap, obtaining $O(E \cdot \lg V)$, or implementing priority queue as Fibonacci heap, obtaining $O(E + V \cdot \lg V)$, where E describes number of graph's edges.
In Fig. 4 we have graph of calculations time for single shortest path problem using standard Dijkstra algorithm in regular grid network with V nodes[1] (each path was calculated for the left-lower and the right-upper pair of cells (nodes) in grid network (similar to one from Fig. 6).
In Fig. 5 we have graph of calculations time for the same problem but defined as linear programming problem and solved using GAMS solver. From Fig. 4 results that when we must compute shortest path in grid network with e.g. $V = 400$ nodes (grid with size 20×20) then computational time is about 100 ms (for average case) using 1 GFLOPS processor and Dijkstra's algorithm. Let's suppose that we simulate battlefield for two-sided company level on the terrain area with size 16 km^2 (terrain square with 4×4 km size, so 4 km/20 = 200 m is side length for each of 400 cells). If we assume, that each company has 3 platoons then in the same simulation time we must plan movement, in the worst case, for $2 \times 3 = 6$ platoons (as non-divided

[1]Using computer with 1 GFLOPS processor (like PENTIUM III 800 MHz).

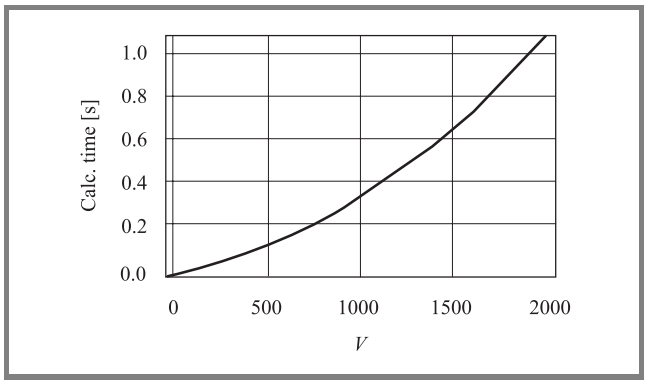


Fig. 4. Calculations time for single shortest path problem using Dijkstra algorithm in regular grid network with V nodes.

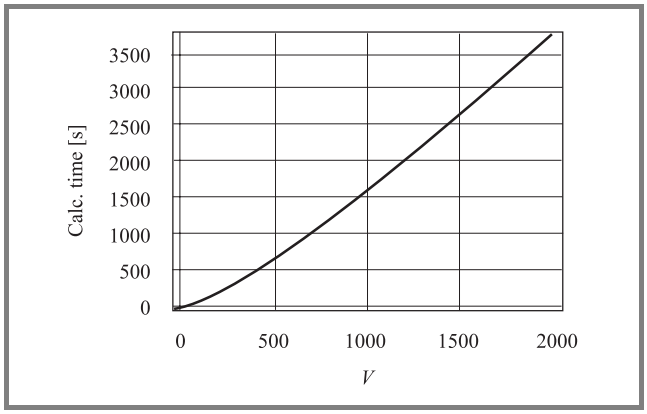


Fig. 5. Calculations time for single shortest path problem defined as linear programming problem and solved using GAMS solver in regular grid network with V nodes.

objects). Because these calculations must be done sequentially (having single processor), so estimation of computational time for all objects is about 6×100 ms = 600 ms. In this case we assumed that all processor power is used for path planning algorithm but it is some simplification, of course. Having, i.e. two-sided battalion fighting ($2 \times 3 \times 3 = 18$ platoons to plan movement in the same simulation time, in the worst case) we need 18×100 ms ≈ 2 s. This delay has significant effect on smoothness of simulation and its visualisation. And we should take into considerations that the network with 20×20 cells is small from among needed in battlefield simulation process.

There are three ways to increase effectiveness of considered problems:

- decreasing the size of terrain-based graph to decrease the computational time of paths planning algorithms [1];
- using specific on-line paths planning algorithms [10, 12, 13, 16, 32];
- using some partial path update approaches [23, 26].

Each of mentioned above ways has some advantages and disadvantages.

Advantage of the first way (**decreasing the size of graph**) is that the number of merged cells into regions will, in general, be far fewer than the number of grid cells. The disadvantages include difficulty in defining the center of the region and the computation difficulties in determining the optimum paths between two adjacent cells.

Some partial path planning algorithms (the second way) plan an off-line path, let the object follow the path, and if any new environment information is gathered, they partially re-plan the existing solution. Similar approach for multi-convoy redeployment in stochastic, dynamically changed environment, was presented in [26, 27]. Disadvantage of this approach is that some times, a small change in the environment may cause re-plan almost a complete path, which may take a long process time (when the network size is big).

The basic idea of **on-line path planning approach** (the third way), in generally, is that the object is moved step-by-step from cell to cell using some heuristic method. This approach is borrowed from movement robots path planning [13, 23, 32]. The decision about the next move (its direction, speed, etc.) depends on the current location of the object and environment status. For example, the idea of RTEF (real-time edge follow) algorithm [32] is to let the object eliminate closed directions (the directions that cannot reach the target point) in order to decide on which way to go (open directions). For instance, if the object has a chance to realize that moving to north and east will not let him reach the goal state, then it will prefer going to south or west. RTEF find out these open and closed directions, so decreasing the number of choices the object has. However, this approach has one basic disadvantage. Namely, in this approach using a few criterions simultaneously to find optimal (or acceptable) path is difficult and it is rather not possible to estimate, in advance, moment of achievement the destination. Moreover, it does not guarantee finding optimal solutions and even suboptimal ones may significanly differ from acceptable.

From this cause, we present in the next section hybrid, cells-merging-based and partial path planning approach for route planning in dynamically changed environment.

Considering route planning in the battlefield simulation we must mention multi-convoy (or multi-object) redeployment and, in consequence, multi-paths planning. Complexity of this process depends on the following conditions [29]:

- count of objects in each convoy (the convoy longer the scheduling of redeployment more complicated);
- have convoys be redeployed simultaneously?
- can convoys be destroyed during redeployment?
- can terrain-based network be destroyed during redeployment?
- have convoys be redeployed through disjoint routes?
- have convoys achieve selected places (nodes) at fixed time?
- do convoys have to start at the same time?

– have convoys determine any action strips for moving?
– can convoys be joined and separated during redeployment?
– have convoys cross through fixed nodes?, etc.

The most often problem related to multi-convoy redeployment is to move a few convoys through disjoint paths simultaneously [25, 30]. Disjoint paths condition results from safety ensuring for moved convoys. In the battlefield simulation finding disjoint paths for moved objects (e.g. tanks inside tank platoon) simplifies its movement because route for each tank do not cross route for each other and we avoid potential collisions. Disjoint paths optimization problem is NP-hard, so some heuristic or other suboptimal approaches are used [2, 21, 25, 31]. Description of some prototype module for maneuvre planning using disjoint paths approach was presented in [28].

4. A new multiresolution approach for increasing route planning effectiveness

Discussed, in the previous section, region approach for terrain included difficulty in defining the center of the region and the computational difficulties in determining the optimum paths between two adjacent cells. In this section we propose some multiresolution-based approach for finding shortest paths in the big grid networks. We assume that we have grid graph $G = \langle \boldsymbol{V}, \boldsymbol{A} \rangle$ (see Fig. 7) as representation of terrain squares (see Fig. 6), where $\boldsymbol{V}$ describes set of

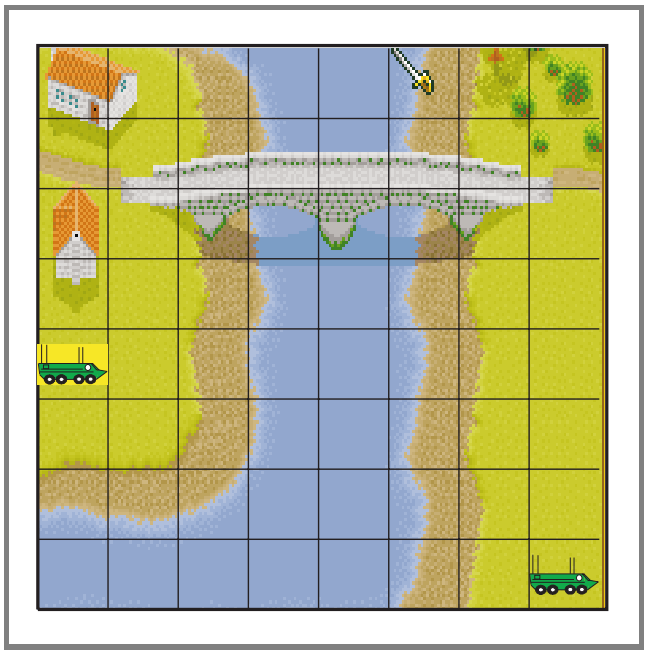


Fig. 6. Terrain space with division on regular grid squares. We want to move object from the right-lower corner to the left side.

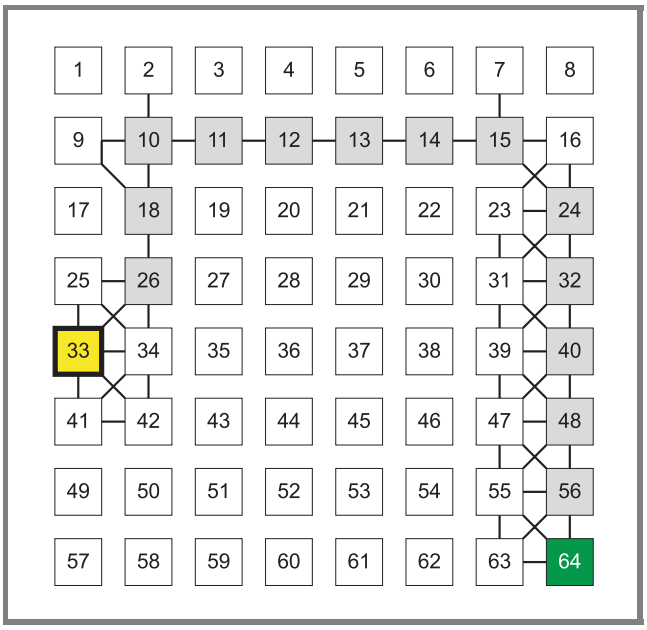


Fig. 7. Grid graph as representation of terrain squares from Fig. 6. There is marked shortest path from node 64 to node 33.

nodes (squares of terrain), $V = |\boldsymbol{V}|$, $\boldsymbol{A}$ describes set of arcs, $\boldsymbol{A} = \{\langle x,y \rangle \subset \boldsymbol{V} \times \boldsymbol{V} : \text{square } x \text{ is adjacent to square } y\}$.
In this graph we may describe some functions (as traversability, visibility, crossing time, crossing probability, detecting probability, etc.) obtaining network as model of movement environment. We assume that for each arc $\langle x,y \rangle \in \boldsymbol{A}$ we have cost $c(x,y)$. The idea of the approach is to merge geographically adjacent small squares (nodes belonging to $\boldsymbol{V}$) into bigger squares (called b-nodes, see Fig. 8) and build b-graph $\overline{G}$ (graph based on the b-nodes, see Fig. 9) using specific transformation. This transformation is based on the assumption that we set arc (b-arc) between two b-nodes $\overline{x} \subset \boldsymbol{V}$, $\overline{y} \subset \boldsymbol{V}$ when exist such two nodes $x \in \overline{x}$, $y \in \overline{y}$ that $\langle x, y \rangle \in \boldsymbol{A}$. In practice, as nodes of $\overline{G}$ graph we will consider strongly connected components of b-nodes. Cost $\overline{c}(\overline{x}, \overline{y})$ of the b-arc $\langle \overline{x}, \overline{y} \rangle \in \overline{\boldsymbol{A}}$ is set on the basis of the biggest cost of some shortest paths calculated inside the subgraph built on the nodes of $\overline{x}$. Next, in the b-graph we find shortest paths between such pairs $\overline{x_s}, \overline{y_t}$ of the b-nodes that source node s and target node t belong to sets $\overline{x_s}, \overline{y_t}$, respectively.
Formal definition of the graph $\overline{G}$ is as follows:

$$\overline{G} = \langle \overline{\boldsymbol{V}}, \overline{\boldsymbol{A}} \rangle, \tag{1}$$

where:

$$\overline{\boldsymbol{V}} = \{\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n\}\text{-set of b-nodes, } |\overline{\boldsymbol{V}}| = n,$$

$$\overline{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\} \subset \boldsymbol{V},\ i = \overline{1,n},$$

$$\underset{\substack{i,j \\ i \neq j}}{\forall} \overline{x}_i \cap \overline{x}_j = \emptyset,\ i = \overline{1,n},\ j = \overline{1,n},\ \bigcup_{i=1}^{n} \overline{x}_i = \boldsymbol{V},$$

$$\overline{\boldsymbol{A}} = \left\{ \langle \overline{x}, \overline{y} \rangle \subset \overline{\boldsymbol{V}} \times \overline{\boldsymbol{V}} : \underset{x \in \overline{x}, y \in \overline{y}}{\exists} \langle x, y \rangle \in \boldsymbol{A} \right\}.$$

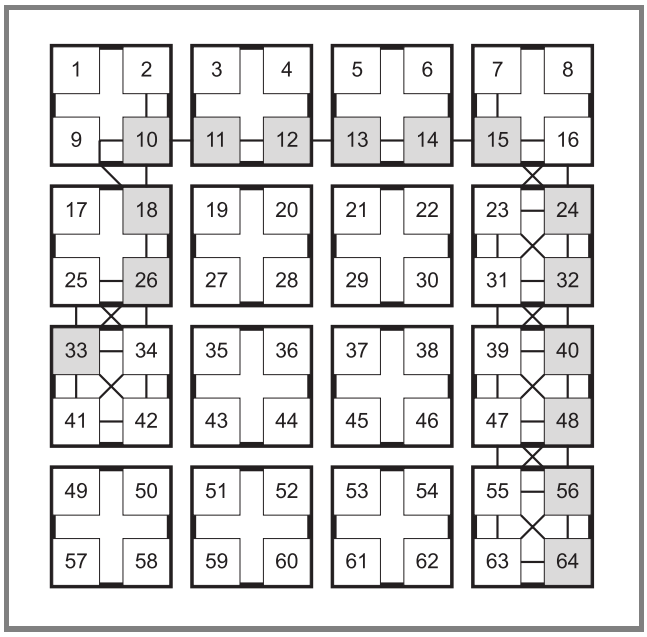


Fig. 8. Merging geographically adjacent small squares from Fig. 7 into $n = 16$ bigger squares (b-nodes).

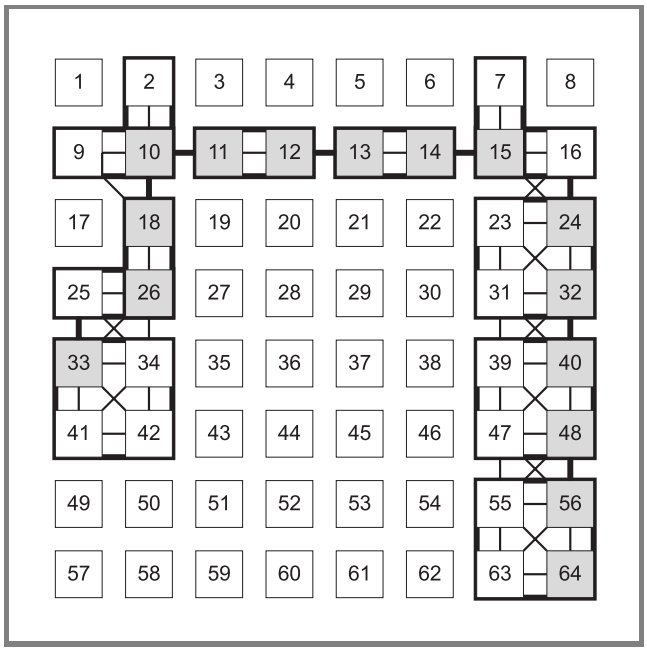


Fig. 9. b-graph for squares merging from Fig. 8. As b-nodes we use strongly connected components of b-nodes from Fig. 8.

Cost function $\overline{c}(\overline{x},\overline{y})$ for b-arc $\langle\overline{x},\overline{y}\rangle$ we determine as:

$$\overline{c}(\overline{x},\overline{y}) = \max_{\{x\in\overline{x}\}} F(x,\overline{y}), \tag{2}$$

where:

$$F(x,\overline{y}) = \min_{\left\{y\in\overline{y}:\ \exists\limits_{z\in\overline{x}}\langle z,y\rangle\in\boldsymbol{A}\right\}} L\big(P(x,y)\big),$$

$$L\big(P(x,y)\big) = \sum_{i=0}^{l(P(x,y))-1} c(x_i,x_{i+1}),$$

$$P(x,y) = (x_0 = x, x_1, x_2, \ldots, x_{l(P(x,y))} = y),$$

$$\forall_{i=\overline{0,l(P(x,y))-1}} \langle x_i, x_{i+1}\rangle \in \boldsymbol{A}.$$

The merging algorithm for b-graph-based shortest paths planning (MSP-algorithm) is following:

1. merge small squares from graph G (Fig. 7) into n bigger squares (Fig. 8) (n is parameter of the algorithm; we show in further discussion how we can set the optimal value of the n);
2. inside each of the n big squares (b-nodes) determine strongly connected components obtaining at least n subgraphs;
3. set each of subgraphs obtained from the Step 2 as b-nodes and arcs as described by (1) obtaining graph $\overline{G}$ (Fig. 9);
4. find shortest paths between each pair of nodes inside each b-node (subgraph) of $\overline{G}$ to calculate cost $\overline{c}(\overline{x},\overline{y})$ for each arc of $\overline{G}$ using Eq. (2);
5. find shortest path in $\overline{G}$ with cost function $\overline{c}(\cdot,\cdot)$ between such pairs $\overline{x_s},\overline{y_t}$ of b-nodes that source node s and target node t belong to sets $\overline{x_s},\overline{y_t}$, respectively.

It's important to explain that setting in the Step 3 strongly connected components as b-nodes assure that each node inside such component is attainable from each other, so if b-node $\overline{x}$ is connected (through b-arc) with b-node $\overline{y}$ then exist path from each node of $\overline{x}$ to each node of $\overline{y}$.
Let's estimate time complexity of MSP algorithm. We will estimate complexity of each step of the algorithm as follows (we assume that each b-node is strongly connected):

2. determination of strongly connected components in graph G: we have n b-nodes creating n merged subgraphs of G; each subgraph of G has no more than $\lceil\frac{V}{n}\rceil$ nodes, so we have complexity $O\left(n\cdot\lceil\frac{V}{n}\rceil\right) = = O(V)$;
3. we have n b-nodes so we obtain $O(n)$;
4. shortest path problem between each pair nodes in N-nodes graph has complexity $O(N^3)$; if each subgraph of G is strongly connected component of G, then has n b-nodes (creating subgraphs), so each subgraph has $\lceil\frac{V}{n}\rceil$ nodes, hence finding all-pairs shortest paths in single subgraph has complexity $O\left(\left(\frac{V}{n}\right)^3\right)$; because we must calculate it n times, so we have $O\left(n\cdot\left(\frac{V}{n}\right)^3\right)$;
5. finding shortest paths in graph $\overline{G}$: because $\overline{G}$ has n b-nodes, so using standard Dijkstra's shortest path algorithm we have $O(n^2)$.

We omit the merging Step 1 because, having n, we can prepare this step before simulation. Taking into considerations above estimations we obtain total complexity of the

algorithm as $O\left(\frac{V^3}{n^2}+n^2+V\right)$ (we have also omitted $O(n)$ because $n \ll V$).

There is very interesting and important question from the point of view of proposed approach effectiveness: how should we set n to obtain the better effectiveness than for V?

Let's notice that computational complexity of the algorithm based on the network with small squares[2] is $O(V^2)$ and for the algorithm based on the bigger squares is $O\left(\frac{V^3}{n^2}+n^2+V\right)$. It means that, in sense of complexity symbol $O(\cdot)$, the bigger squares approach is better if the following formula is satisfied:

$$\frac{V^3}{n^2}+n^2+V < V^2 \tag{3}$$

or equivalently, when

$$n^4-(V^2+V)n^2+V^3<0\,. \tag{4}$$

Solving this inequality we obtain, that $n \in [n_1, n_2]$, where

$$n_1=\sqrt{\frac{V^2+V-\sqrt{(V^2+V)^2-4V^3}}{2}}\,, \tag{5}$$

$$n_2=\sqrt{\frac{V^2+V+\sqrt{(V^2+V)^2-4V^3}}{2}}\,. \tag{6}$$

For example, for the graph from Fig. 7 ($V=64$) we obtain $n_1 \approx 8$, $n_2 \approx 64$.

In order to estimate how many times faster we compute problem for finding shortest path in the network with n big squares ($\left(\frac{V^3}{n^2}+n^2+V\right)$) with relation to problem for finding shortest path in the network with V small squares ($O(V^2)$) we may formulate acceleration function as follows[3]:

$$A(V,n)=\frac{V^2}{\frac{V^3}{n^2}+n^2+V}\,. \tag{7}$$

Exemplified graphs of $A(V,n)$ are shown in Figs. 10 and 11.

Having grid network with V squares (nodes) we can formulate following optimization problem: to find such cardinal n^*, for which

$$A(V,n^*)=\max_{n\in[n_1,n_2]} A(V,n)\,, \tag{8}$$

where n_1, n_2 are described by formulas (5) and (6).

[2]Using standard Dijkstra's shortest paths algorithm (without modifications increasing its effectiveness).

[3]Exact to complexity estimation symbol $O(\cdot)$.

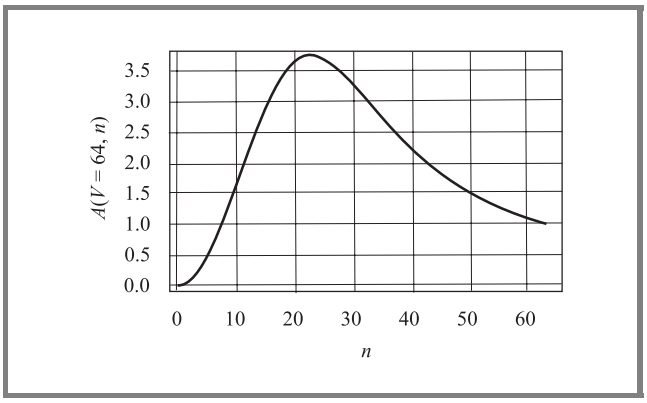


Fig. 10. Graph of $A(V, n)$ function for the network with $V = 64$ nodes.

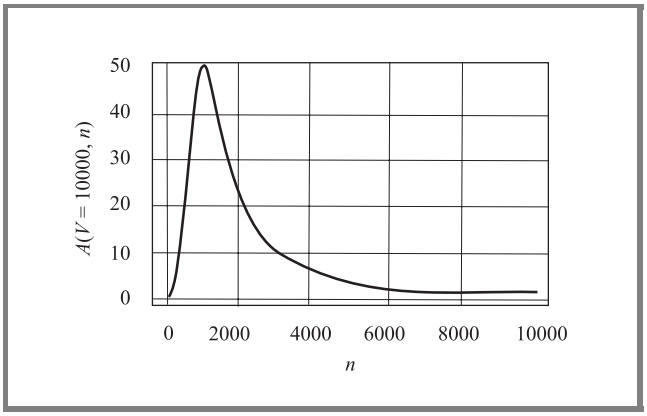


Fig. 11. Graph of $A(V, n)$ function for the network with $V = 10000$ nodes.

Let's notice, that function (7), omitting constraint for n integer, has real nonnegative maximum for the value $n^*=\sqrt[4]{V^3}$. It may be easily shown that for each $V>0$, $n^* \in [n_1, n_2]$. In practice we are interested in such value $n^{**} \approx n^*$, that square root of $\frac{V}{n^{**}}$ is cardinal number (it results from the fact that each of n^{**} big squares consist of $\frac{V}{n^{**}}$ small squares and in grid structure of the network a big square has $\sqrt{\frac{V}{n^{**}}} \times \sqrt{\frac{V}{n^{**}}}$ small squares).

In Table 1 the influence of V on n^*, n^{**} and $A(V,n^{**})$ is shown. It is easy to observe the best acceleration of shortest path algorithm using presented approach in regular grid network with V nodes may be approximated by value $A(V,n^{**}) \approx \frac{1}{2}\sqrt{V}$.

Let's notice that from presented estimations and Table 1 result that for 400-nodes grid graph considered at the beginning of the previous section movement planning for two-sided battalion fighting (for 18 platoons) will be done in time $18 \times 100/9.5$ ms ≈ 200 ms.

Table 1
Influence of V on n^*, n^{**} and $A(V,n^{**})$

V	n^*	n^{**}	$\lceil \frac{V}{n^{**}} \rceil$	$A(V,n^{**})$
100	32	25	4	4.3
400	90	100	4	9.5
900	164	225	4	12.3
1600	253	169	9	14.7
2500	354	256	9	20.4
10000	1000	1089	9	49.0
40000	2828	2500	16	96.8
90000	5196	5625	16	147.9
160000	8000	6400	25	181.4
250000	11180	10000	25	243.7

5. Conclusions

The approach presented in the paper gives possibilities to significantly decrease computational time in terrain-based route planning when the terrain environment is represented by regular grid of squares. This approach may be applied, i.e. for route planning in the simulated battlefield.

The estimations of presented algorithm effectiveness may be improved through a few ways. The first way is to use in time complexity estimations the best known shortest-path algorithm estimation $(O(E + V \cdot \lg V))$ instead complexity of standard Dijkstra's algorithm $(O(V^2))$ because the regular grid graph is thin (maximal number of direct successors for any node is 8), so $O(8V + V \lg V) < O(V^2)$ nearly for all V (exactly for $V > 11$). The second way is to improve Step 4 of the algorithm because it seems to be unnecessary determinations all-pairs shortest paths in each b-nodes (subgraphs). It's seems that is enough to determine shortest paths between "outside" nodes of b-nodes because only these nodes are used to link b-node with another. Moreover, to confirm presented estimations it is essential to conduct calculations in real grid graphs.

Presented suggestions may be contribution for further works.

References

[1] J. R. Benton, S. S Iyengar, W. Deng, N. Brener, and V. S. Subrahmanian, "Tactical route planning: new algorithms for decomposing the map", in *Proc. IEEE Int. Conf. Tools for AI*, Herndon, 1995, pp. 268–277.

[2] A. Bley, "On the complexity of vertex-disjoint length-restricted path problems", Konrad-Zuse-Zentrum fur Informationstechnik, Berlin, 1998 (see also: http://www.zib.de/PaperWeb/abstracts/SC-98-20/).

[3] C. Campbell, R. Hull, E. Root, and L. Jackson, "Route planning in CCTT", in *Proc. 5th Conf. Comput. Gener. Forc. Behav. Repres.*, Tech. Rep., Institute for Simulation and Training, 1995, pp. 233–244.

[4] C. G. Cassandras, C. G. Panayiotou, G. Diehl, W.-B. Gong, Z. Liu, and C. Zou, "Clustering methods for multi-resolution simulation modeling", in *Proc. Conf. Enabl. Technol. Simul. Sci., Int. Soc. Opt. Eng.*, Orlando, USA, 2000, pp. 37–48.

[5] C. Cooper, A. Frieze, K. Melhorn, and V. Priebe, "Average-case complexity of shortest-paths problems in the vertex-potential model", *Rand. Struct. Algor.*, vol. 16, pp. 33–46, 2000.

[6] P. K. Davis, J. H. Bigelow, and J. McEver, "Informing and calibrating a multiresolution exploratory analysis model with high resolution simulation: the interdiction problem as a case history", in *Proc. 2000 Winter Simul. Conf.*, 2000, pp. 316–325.

[7] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths", *IEEE Trans. Syst. Sci. Cybern.*, vol. SSC-4, no. 2, pp. 100–107, 1968.

[8] J. James, B. Sayrs, J. Benton, and V. S. Subrahmanian, "Uncertainty management: keeping battlespace visualization honest", http://citeseer.nj.nec.com/ 386770.html

[9] L. Joe and P. M. Feldman, "Fundamental research policy for the digital battlefield", Res. Rep. DB-245-A, RAND Co., Santa Monica, USA, 1998.

[10] C. R. Karr, M. A. Craft, and J. E. Cisneros, "Dynamic obstacle avoidance", in *Proc. Conf. Distrib. Interact. Simul. Syst. Simul. Train. Aerosp. Envir., Int. Soc. Opt. Eng.*, Orlando, USA, 1995, pp. 195–219.

[11] T. Kreitzberg, T. Barragy, and B. Nevin, "Tactical movement analyzer: a battlefield mobility tool", in *Proc. 4th Join Tactic. Fus. Symp.*, Laurel, 1990.

[12] B. Logan, "Route planning with ordered constraints", in *Proc. 16th Works. UK Plann. Schedul. Spec. Int. Group*, Durham, UK, 1997.

[13] B. Logan and A. Sloman, "Agent route planning in complex terrains", Tech. Rep. CSRP-97-30, University of Birmingham, School of Computer Science, Birmingham, 1997.

[14] M. Longtin and D. Megherbi, "Concealed routes in ModSAF", in *Proc. 5th Conf. Comput. Gener. Forc. Behav. Repres.*, Tech. Rep., Institute for Simulation and Training, 1995, pp. 305–314.

[15] J. S. B. Mitchell, "Geometric shortest paths and network optimization", in *Handbook of Computational Geometry*, J. R. Sack and J. Urrutia. Elsevier Science Publ., B.V. North-Holland, Amsterdam, 1999.

[16] D. K. Pai and L. M. Reissell, "Multiresolution rough terrain motion planning", Department of Computer Sciences, University of British Columbia, Tech. Rep. TR 94-33, Vancouver, 1994.

[17] M. D. Petty, "Computer generated forces in distributed interactive simulation", in *Proc. Conf. Distrib. Interact. Simul. Syst. Simul. Train. Aerosp. Envir., Int. Soc. Opt. Eng.*, Orlando, USA, 1995, pp. 251–280.

[18] S. Rajput and C. Karr, "Unit route planning", Tech. Rep. IST-TR-94-42, Institute for Simulation and Training, Orlando, USA, 1994.

[19] G. A. Schiavone, R. S. Nelson, and K. C. Hardis, "Interoperability issues for terrain databases in distributed interactive simulation", in *Proc. Conf. Distrib. Interact. Simul. Syst. Simul. Train. Aerosp. Envir., Int. Soc. Opt. Eng.*, Orlando, USA, 1995, pp. 89–120.

[20] G. A. Schiavone, R. S. Nelson, and K. C. Hardis, "Two surface simplification algorithms for polygonal terrain with integrated road features", in *Proc. Conf. Enabl. Technol. Simul. Sci., Int. Soc. Opt. Eng.*, Orlando, USA, 2000, pp. 221–229.

[21] A. Schrijver and P. Seymour, "Disjoint paths in a planar graph – a general theorem", *SIAM J. Discr. Math.*, no. 5, pp. 112–116, 1992.

[22] H. Sherali, K. Ozbay, and S. Subrahmanian, "The time-dependent shortest pair of disjoint paths problem: complexity, models and algorithms", *Networks*, no. 31, pp. 259–272, 1998.

[23] A. Stentz, "Optimal and efficient path planning for partially-known environments", in *Proc. IEEE Int. Conf. Robot. Automat., ICRA'94*, vol. 4, pp. 3310–3317.

[24] P. D. Stroud and R. C. Gordon, "Automated military unit identification in battlefield simulation", LAUR-97-849, *SPIE Proc.*, vol. 3069, Los Alamos National Laboratory, Los Alamos, 1997.

[25] Z. Tarapata, "Algorithm for simultaneous finding a few independent shortest paths", in *Proc. 9th Eur. Simul. Symp., ESS'97, Soc. Comput. Simul.*, Passau, Germany, 1997, pp. 89–93.

[26] Z. Tarapata, "Simulation method of aiding and estimation of transportation columns movement planning in stochastic environment", in *Proc. 13th Eur. Simul. Multiconf., Soc. Comput. Simul. Int.*, Warsaw, Poland, 1999, pp. 613–619.

[27] Z. Tarapata, "Computer simulation of individual and grouped military objects redeployment", *Bull. Milit. Univ. Technol.*, no. 1, pp. 147–162, 2000.

[28] Z. Tarapata, "Computer tool for supporting and evaluating convoys redeployment planning", *Oper. Res. Decis.*, no. 1, pp. 91–107, 2000.

[29] Z. Tarapata, "Some aspects of multi-convoy redeployment modelling and simulation", in *Proc. 21st AFCEA Eur. Symp. & Exposit.*, Prague, 2000 (compact disk publication).

[30] Z. Tarapata, "Modelling, optimisation and simulation of groups movement according to group pattern in multiresolution terrain-based grid network", in *Proc. Reg. Conf. Milit. Commun. Inform. Syst.*, Zegrze, Poland, 2001, vol. I, pp. 241–251.

[31] Z. Tarapata, "Fast method for redeploying multi-convoy in multiresolution grid network", *Bull. Milit. Univ. Technol.*, 2003 (in press).

[32] C. Undeger, F. Polat, and Z. Ipekkan, "Real-time edge follow: a new paradigm to real-time path search", SCS Publications, 2001 (see also: http://citeseer.nj.nec.com/489498.html).

Zbigniew Tarapata has graduated from Cybernetics Faculty at Military University of Technology (MUT) in Warsaw. He received his Master's degree in computer science in 1995 and Doctor's degree in the same field, in 1998. From 1995 to July 1999 he worked as an assistant and since July 1999 – as a senior lecturer at Operations Research Division of Institute of Mathematics and Operations Research of Cybernetics Faculty of MUT. His scientific interests and work are related to the following subjects: mathematical and simulation modelling of systems; transport optimization; combat modelling, simulation, optimization and prediction; graph and network optimization; algorithms effectiveness; methods and tools of multicriteria decision making and supporting.
e-mail: ztarap@isi.wat.waw.pl
Institute of Mathematics and Operations Research
Faculty of Cybernetics
Military University of Technology
Kaliskiego st 2
00-908 Warsaw, Poland

Paper

Interfacing war game simulations with tactical C2 systems – dream or reality?

Milan Šnajder and Philip W. Holden

Abstract — Decision making process in current tactical C2 systems is base on planning process of commanders and their staff. Improving tactical decision making by interfacing war game simulations with tactical C2 systems is achievable. Commander can review the results of the simulation and subsequently modify the tactical plan. Previously, the use of "training" simulations was not a viable solution to real world decision making due to the lengthy time required to input all of the combat entities, the unit organizations and personnel dispositions, the equipment configurations, status of the units and equipment, and the distribution of the available supplies. Modern C2 systems have all of this information stored in the common system databases, and this information can be used to instantiate and populate the simulation through an electronic adaptation of the data structures to match the requirements of the constructive simulation. This paper will provide description of system approach of interfacing simulation and C2 system to improve decision-making.

Keywords — C2 systems, war fighting simulation, GF-TCCS, common operatring picture, NCOE, system architecture, constructive simulations ModSAF, SAF, SIMNET CGFs, OTB, DIS, HLA, evolutionary system development, rapid prototyping, interfacing ModSAF with GF-TCCS, MOD, friendly and enemy forces, tactical operations centre.

1. Introduction

The Army of the Czech Republic (ACR) is developing an integrated, automated ground forces tactical command and control system. Previously, commanders and staffs generally performed their mission using a manual system, augmented by some commercially available hardware and software systems. Some automation and communications systems operated in an isolated manner but did not provide the mobility, functional flexibility, security, survivability, and interoperability required to support the ACR. The accelerated tempo of modern, mostly alliance-based combined-arms warfare demands rapid processing and transfer of C2 information. Ground forces combat at tactical levels requires improved battle command systems, increased capability to synchronize direct and indirect fire, faster and more comprehensive access to intelligence data, enhanced situational awareness and effective force protection. To improve agility, commanders at all echelons require the means to gain and use timely battle space information in order to make informed decisions in a manner consistently faster than that of the enemy. **The ground forces tactical command and control system** (GF-TCCS) is being developed to meet these requirements, and tools to help the commander in the decision-making process are being developed. One of the more promising is the use of constructive war game simulations to explore and analyze the effects of different courses of action.

The Czech Republic has recently installed the **war fighting simulation, ModSAF** at the Military Academy in Brno, and at the training center in Vyskov. These simulations are being used to train the tactical commanders and staff, and can be used to model various courses of action, providing valuable information to support decision making. The US Government provided the simulation software, and had a contractor, **Science Applications International Corporation** (SAIC), perform the installation of ModSAF in April, 1999. Continuous improvements to the application have been made since that time. As the simulation tools are used, new applications for the simulations are becoming evident, including being able to use the simulation during tactical operations. One of the obstacles to using simulations during a crisis or military operation is the extensive amount of time necessary to setup the simulation. To be effective, the simulation must be loaded with three categories of information:

- locations, strength and status of the friendly forces, and the presumed locations, strength, and status of the opposing forces;
- missions of the respective units;
- environment variables such as terrain and weather.

The integration of constructive simulations with the GF-TCCS will allow the electronic instantiation and parameterization of the simulation with all three categories of information, allowing it to be used in real-time decision making.

2. Ground forces tactical command and control system operational requirements

The GF-TCCS will provide seamless connectivity from the lower tactical (squad and platoon) level to the operational

commands (ground forces command and territorial forces command). GF-TCCS will be used regularly within garrison, during deployment, and in the field to maintain the soldier's proficiency at the level required to respond to the broad range of potential missions.
GF-TCCS vertically and horizontally integrates information from tactical to operational command level, and will allow the commander and staff to:

- collect process and organize large amount of battle information;
- combine information from multiple sources to create more complete and useful information;
- process information to analyze trends;
- detect unusual activities, or predict a future situation;
- develop courses of action based on situational factors;
- exchange information efficiently among and within command posts on the battlefield;
- present information as graphic displays and textual summaries.

Fundamental to the GF-TCCS operational concept and relevant **common operating picture** (COP) is a single entry, near-real-time information system, with automated interoperability between each battlefield information system. GF-TCCS provides situational information and decision sup-port to commanders and staffs in the execution of the operational/tactical battle at operational groups and below. The GF-TCCS command and control subsystems are heavily oriented toward combat operations.
The GF-TCCS command and control sub-systems are linked by **tactical area communications system** and by the **combat net radio system**. Combat forces, weapon systems and battlefield automated systems will be supported by the integrated management and control system that provides management of the tactical communications.
GF-TCCS will be linked directly to the staff information system (SIS) ACR, providing the framework for seamless connectivity from the battalion to the general staff.
The NC3 common operating environment (NCOE) is the NATO C3 standard profile (NCSP) for computing and communications infrastructure. The NCSP describes the structural foundation necessary to build interoperable and open systems. The purpose of the NCOE is to facilitate a common understanding of the concepts, constructs and methods (set up processes) required for targeted NATO systems. More detail on the NCOE component model can be found in volume 5 of the NATO C3 Technical Architecture, *NC3 Common Operating Environment*.
GF-TCCS's C2 systems are shown in Fig. 1 and include:

- maneuver control system (MCS);
- fire support control system (FSCS);
- forward area air defense control system (FAADCS);
- intelligence and electronic warfare control system (IEWCS);
- tactical logistics control system (TLCS);
- battle management vehicular information system (BMVIS).

The three supporting systems are:

- tactical area communications system (TACS);
- combat net radio system (CNRS);
- integrated management and control system (IMCS).

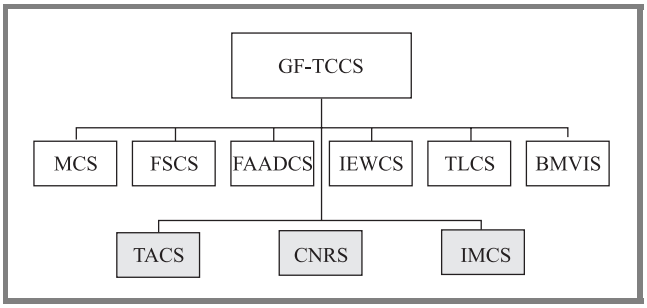


Fig. 1. Subsystems of the GF-TCCS.

The GF-TCCS provides a simultaneous, relevant picture of the battlefield at each echelon – from squad/mobile platform leader to battalion/brigade commander – based on common data collected through networks of command posts, commander's vehicles, computers, sensors and weapon platforms. This information can be used to instantiate or update the data files used by the constructive simulations.

3. GF-TCCS system architecture overview

The GF-TCCS architecture provides two complimentary methods of distributing C2 data within a single, local area command post (CP). Within a local area CP, the GF-TCCS will employ the following message distribution means:

- **event bus architecture** – for message products that are relatively small, requiring a wide distribution, and are time sensitive;
- **product server architecture** – for information products that are larger, less time sensitive, and have a limited or sequential pattern of user access or workflow.

Externally, the GF-TCCS architecture employs an external messaging architecture that provides the following data distribution means:

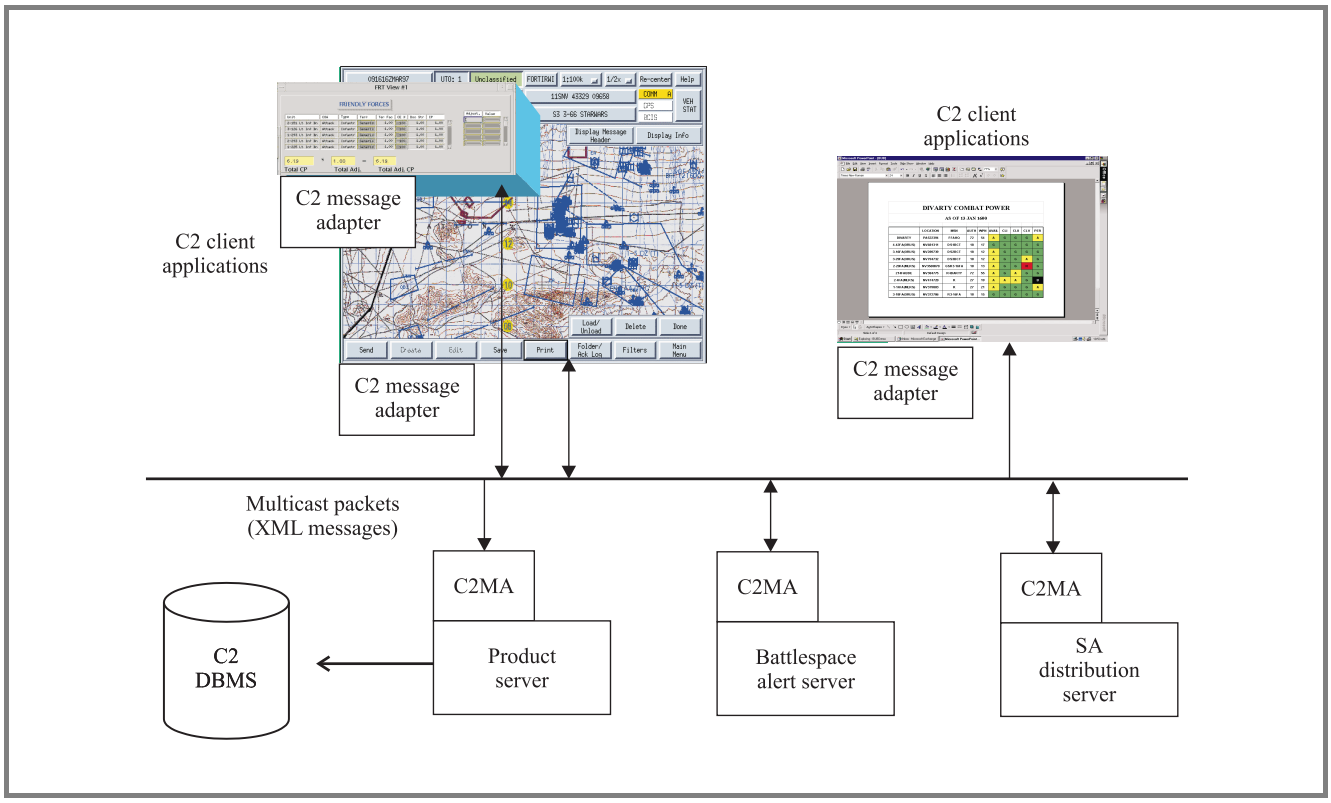


Fig. 2. GF-TCCS system architecture.

- **AdatP-3 messaging** – for NATO interoperability;
- **VMF messaging** – for inter-GF-TCCS traffic, where low bandwidth communication median is available or where low latency messaging is required;
- **XML product messaging** – for inter- GF-TCCS traffic, where a higher bandwidth communication median is available, or where larger bulk messages are required.

The event bus architecture provides a publish/subscribe mechanism to distribute information (i.e., messages) among C2 applications. In a publish/subscribe architecture, applications register themselves as producers of information (i.e., publishers) and consumers of information (i.e., subscribers). Figure 2 graphically depicts the basic connectivity of the GF-TCCS event bus.
Two general component types comprise the event bus architecture. They are:

- C2 message adapters (C2MA) that provide the interface between the C2 applications and the event bus. C2MAs distributes information (XML messages) within a tactical operation center using IP multicast.
- C2 applications that are the domain applications using the services of the event bus. C2 applications may be either client side or server side domain applications, such as the commander's information display (i.e., map) client, the task organization client, and the product data server.

4. Operational-tactical solutions

This group of applications supports all commanders and staff officer tasks. In actual solution stage, operational-tactical solutions (OTS) has about 12 specific applications for supporting typical commander and staff activities:

- TaGIS – Czech Army RETM for-mat and Aerial snap composition;
- transportation on own wheels – output – transportation plan;
- chemical situation awareness;
- planning of the radio-relay communication – schema of connection availability;
- combat power (force ratio calculation) – output;
- radio visibility awareness;
- message handling system – warning preparation;
- OWNSITREP – situational report preparation;
- DMP lifebook – lists of staff activities, software aids and active documents;

Fig. 3. TaGIS – combination of FRAGO, battalion planning overlay and actual UTO.

- TaGIS – combination of FRAGO, battalion planning overlay and actual UTO (Fig. 3);
- electronic staff lifebook – mission definition.

5. The architectural approach of the MCS

Maneuver control system (MCS), is a core tactical forces information system that provides commanders and staff with the capability to collect, coordinate, and act on real-time battlefield information.
Through the MCS, the commander transmits critical battlefield information, courses of action, schemes of maneuver, warning orders, operation orders, priorities, intelligence requests, and air operations requests. The MCS database and data files provide the information necessary to load the simulations with the friendly unit missions and objectives.

The main purpose of the event bus architecture is to distribute and receive time-sensitive battlefield events data (i.e., situation awareness data) between interested C2 applications. It connects the publishers and subscribers of messages together. The C2MA listens for a published message, validates the format of the message, and distributes it to subscribers on the bus through the use of the subscriber's C2MA. The core technology for the event bus is IP multicast, which is a standard IP network-level protocol.

The C2 message adapter is the middle-ware component that provides the event bus services to C2 applications. Its role

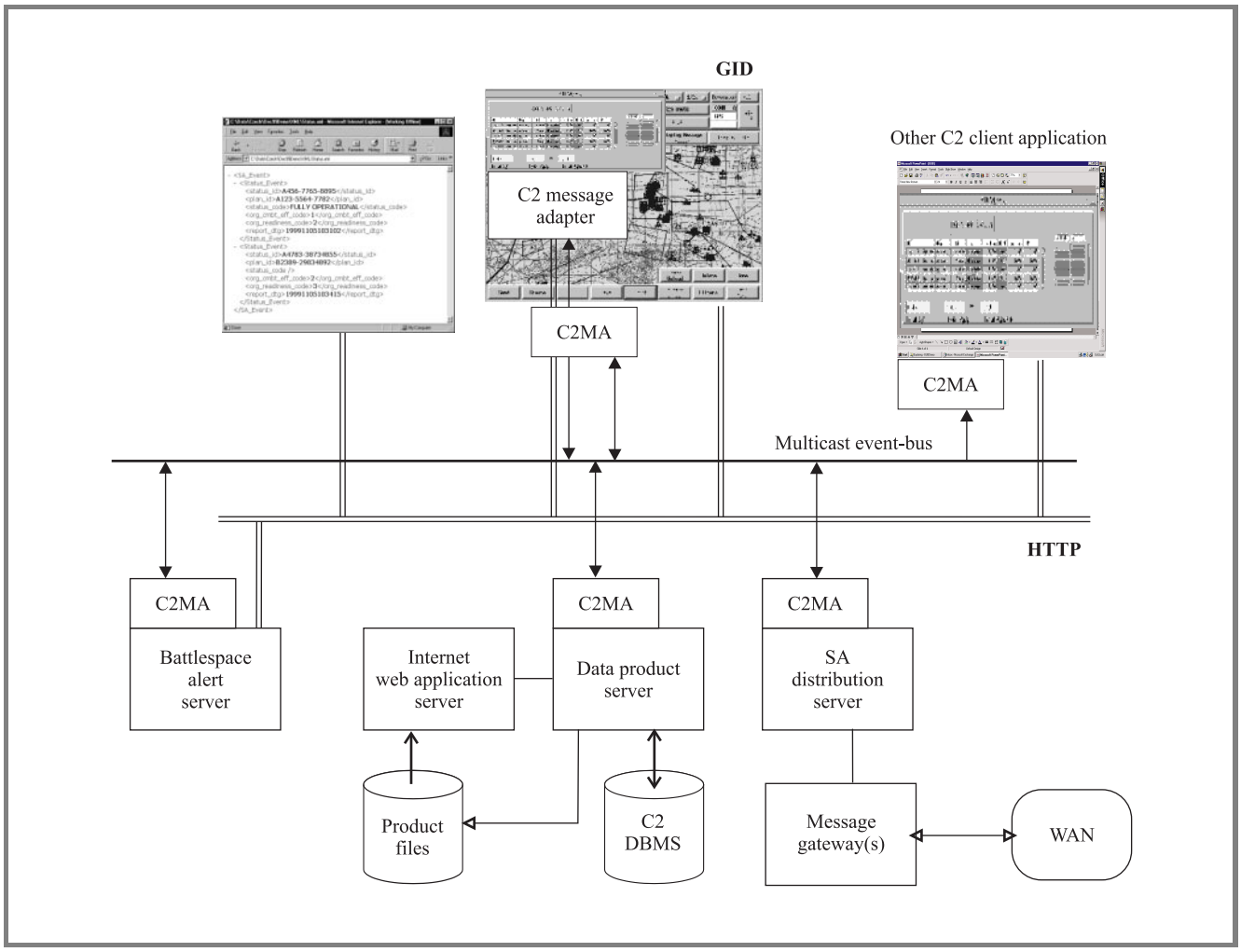


Fig. 4. GF-TCCS event bus architecture and C2 message.

is to handle subscriptions of C2 applications, listen for messages on the bus, and to publish messages from C2 applications.

6. C2 application architecture

GF-TCCS C2 functionality is provided by a set of server-side and client-side domain applications (Fig. 4). Each application component provides a unique set of C2 services that are mutually exclusive to that application component.

The applications components are:

- graphical information display (GID);
- battlefield alert server (BAS).

7. Constructive simulations

Constructive simulations include a category of computer generated forces (CGFs) that simulate battlefield entities and aggregates of those entities. The simulation of the entities includes their physical characteristics, tactical behaviors and decisions processes, and the interactions with other entities. Entities can range from an individual soldier, to ground vehicles such as tanks or armored personnel carriers, to aircraft or ships. The entities can be guided and controlled by human operators using joysticks or keyboards, or full mission simulators such as the SIMNET or CCTT manned simulators, or the entities can be completely generated and controlled by the computer. Entities entirely controlled by the computer are referred to as "automated forces". Where some human involvement in the decision process is involved, they are referred to as "semi-automated forces" (SAF). These simulations are routinely used to support army applications in three modeling and simulation domains: training, exercise, military operations (TEMO); advanced concepts and requirements (ACR); and research, development and acquisition (RDA).

ModSAF has a modular, data-driven architecture that allows the development and integration of specific detailed models that become an additive part of the overall system. ModSAF's main modules are the human interface to the simulation system (SAF station), SAF Sims, the entity, and

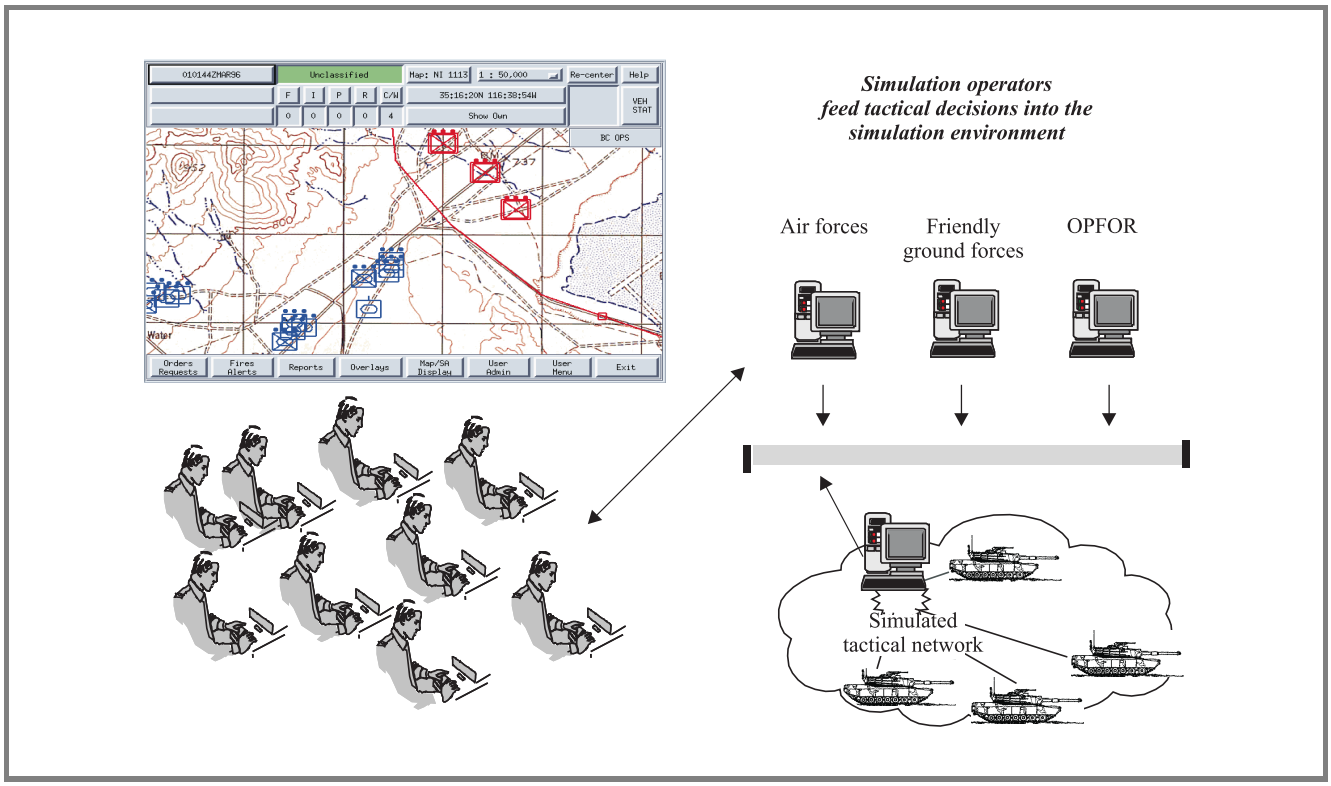


Fig. 5. The typical Czech constructive simulation system configuration.

aggregate models. SAF Sims and SAF stations use a common database to communicate with each other, and use the DIS protocol to communicate with the synthetic environment and other simulations and simulators. The common database provides command and control information for the simulation of organizations, while the DIS communications are oriented to the physical world. The simulation operator has the ability to create, modify, save or load scenario files, overlay files, minefield templates, and other setup parameters. The simulation can also start/resume, stop/pause, and restart an exercise run based on external input in the form of simulation management messages.

8. The Czech simulation system

The Army of the Czech Republic is currently using the US warfighting simulation, ModSAF, at the Military Academy in Brno, and at the training center in Vyskov. The ACR has developed expertise in using and modifying the simulations with the help of the US provided source code for the simulation and advanced developer training courses. The simulation engineers at the Military Academy have developed new capabilities for the Czech simulation system (Fig. 5). The ACR has made a number of enhancements to the simulation system, and developed the capability to build the digital terrain databases necessary to support the modular semi-automated forces (ModSAF) and its successor, the oneSAF testbed baseline (OTB) synthetic natural environment requirements. These enhanced simulations are being used to train the tactical commanders and staff, and can be used to model various courses of action, providing valuable information. The ModSAF simulation system builds upon the previous United States of America Army simulations that started with the simulator networking (SIMNET) program. The original ModSAF incorporated the software code associated with both SIMNET SAF and ODIN (73 Easting) SAF. In 1993, the US Defense Advanced Research Projects Agency (DARPA) began building ModSAF by developing an open simulation architecture, which could be used to create synthetic agents for a variety of distributed interactive simulation (DIS) applications. The initial effort fielded a system in 1993 to support the what-if simulation system for advanced research and development (WISSARD) program, which had a requirement for beyond visual range, air-to-air engagement scenarios. After the initial release, the remaining battlefield operating systems (BOS) and behaviors were added to ModSAF to fill out the synthetic battlefield. ModSAF 1.2 was released in 1994 and included the majority of systems that had been represented in the previous SIMNET SAF version. ModSAF 2.0 followed in 1995, ModSAF 2.1 in 1996, ModSAF 3.0 in 1997, ModSAF 4.0 in 1998 and finally, the last version of the simulation, ModSAF 5.0 was released in 1999. The successor to ModSAF is the one semi-automated forces testbed baseline that will be replaced by the OneSAF objective system, which is planned to be fielded in 2004.

ModSAF has been expanded to include many aspects of the modern battlefield, to include the effects of weapons of

mass destruction (WMD). ModSAF can also model the logistics aspects of the engagement, and has the capability to model support functions such as medical support to operations. The highly detailed constructive simulation has the capability to model the evaluation of medical conditions of personnel, and to model evacuation of injured individuals by various means, to include vehicles, and rotary wing aircraft (RWA). The simulation also models personnel status, including wounded in action (WIA) and killed in action (KIA) as a result of casualties caused by munitions, detonations, collisions, and non-combat illness/injury. The simulated injuries can range from smoke inhalation and other environmental affects to death by weapons of mass destruction.

9. Evolutionary systems development and rapid prototyping

Both the development of the GF-TCCS and the Mod-SAF/OTB simulation systems have benefited from the **evolutionary development process** also known as "spiral development". This development methodology is based on two main fundamentals:

1. **Incremental development**. Both whole complex and each subsystem are (and will be) developed in a sequence of increments. From the first increment, each of them is fully operable and fully integrated with all preceding increments.

2. **Rapid prototyping**. Every planned element and prepared system is prepared like a prototype and consecutively tested in development labs, in a special testbed and in the field as well.

Rapid prototyping, together with experimentation, provides an effective tool for resolving issues, experimental data collection, reducing risk early, and determining the adequacy of requirements, design, and new GF-TCCS's system capabilities before committing major resources.
The attributes of the proposed approach to evolutionary GF-TCCS development include:

- use of software development environments/tools for rapid prototyping of functionality;
- object-oriented design that allow rapid integration of COTS software;
- NATO OSE software standards and practices;
- documentation appropriate for expansion into formal specifications;
- continuous interaction and feedback from the military end-users.

10. Integrating ModSAF with the GF-TCCS

Many companies and organizations use the open architecture ModSAF development environment. The source code can be compiled using the operating system native UNIX C compilers, or the freeware toolset GNU gcc. The ModSAF distribution package includes scripts (awk, sed, lex, etc.) and makes files to support distributed development and experimentation. The US Government provided the ModSAF source code to the Czech Ministry of Defense (MoD) with a distribution agreement that allows unlimited use by the Czech MoD. The simulation engineers at the Czech Military Academy have made several modifications and enhancements to the ModSAF and OTB simulation systems. **The intent of the liberal distribution agreement is to allow the users to develop additional capabilities, as their needs require**. The simulation runs on multiple platforms, including the PC (Linux), Silicon Graphics (SGI), SUN, and DEC Alpha platforms. The computer requirements are relatively low, and the simulation can run on a PC using the Linux operating system with a 300 MHz processor, 128 Mbytes of RAM. This allows integration with the GF-TCCS without the problems normally associated with interfacing different computational platforms. Much of the Mod-SAF simulation software is written in the C language, with Java used for recent additions to the Graphical User Interface. The modular architecture and separate GUI module allow the ability to create a custom user interface using the Czech language. Although there is a strong effort within the ACR to use the English language for interoperability with NATO partners, the ability to have a Czech language interface greatly improves the ease of use for the average Czech soldier.

11. Standards compliance

In order to ensure compatibility with other NATO countries, standards compliance is crucial to the long-term success of the project. Compliance with the NATO modeling and simulation master plan is considered essential, and the simulations must provide support for a standard synthetic environment and be able to communicate using HLA and DIS 2.0.4 protocols. ModSAF meets these criteria for interoperability, although the operation in an HLA environment requires the use of a gateway such as the MaK technology HLA gateway, or the Naval Air Warfare Center – Training Systems Division (NAWCTSD) gateway. The ModSAF simulation appears to be functionally suited for integration with the GF-TCCS since the resolution of entities and the convention for identifying vehicles precludes ambiguity because the provisions for bumper numbers and task organizations is compatible with the ACR requirements and conventions. The capability to simulate dynamic environments, and the capability to assess damage to prepo-

sitioned and dynamic objects due to detonations of munitions, provides the level of detail to support the decision making process used at the tactical level.

12. Behavioral specificity

One of the traditional obstacles to using simulations to support decision making in a real-time environment is the extensive time required to create the necessary simulation configuration that reflects the current operational situation. Often, a team of multiple simulation specialists would require several days to create the data needed to represent the situation, and to load that information into the computer. For large and complex scenarios the process can require several weeks to be prepared. In order to support real-time Czech tactical requirements, the process of gathering the information on the current situation and loading the information to start a simulation run must take less than 30 minutes, and preferably less than five minutes. The ModSAF simulation is structured such that the setup and parameter information is contained in data files, and even the behaviors and unit missions can be instantiated on most entities using data file modifications. Many behaviors can be implemented with data strings, and configuration parameters can be used to customize a behavior for a unique situation. Under this approach, behaviors such as those required for a road march can be used on any ground vehicle or ground unit. However, the weapon or vehicle specific behaviors such as the tactical actions that differentiate a tank platoon road march from a mechanized platoon road march are not represented in the generic behavior.

The ModSAF simulation supports three basic types of behavior control: **Pre-planned, operator controlled**, and **automatic reaction**. The pre-planned mode of operation allows the user to specify the sequencing of behaviors for a unit, and allows control of the transitions from one behavior phase to the next. Transitions can be timephased, or caused by actions such as crossing a control measure, commands from the operator, or other simulation events. With the operator control mode, the SAF operator has the ability to modify missions during execution of a scenario, to include making immediate interventions to make temporary modifications to the mission. The automatic reaction mode allows entity reactions that are enabled by data, and when activated, a reaction overrides the current mission with a new behavior. The reaction can be resumed at the direction of the SAF operator.

Behavior models interface with the physical models through defined APIs called **generic model interfaces** (GMIs). GMI facilitates physical model extension, implementing each physical model with one distinct software unit in the form of a C language library, functionally becoming a "plug and play" system, with physical models capable of being easily inserted or deleted.

In order to make an entity such as a T-72 tank move, the movement behavior provides a proper command by calling GMIs supported by the physical model. Then the generic

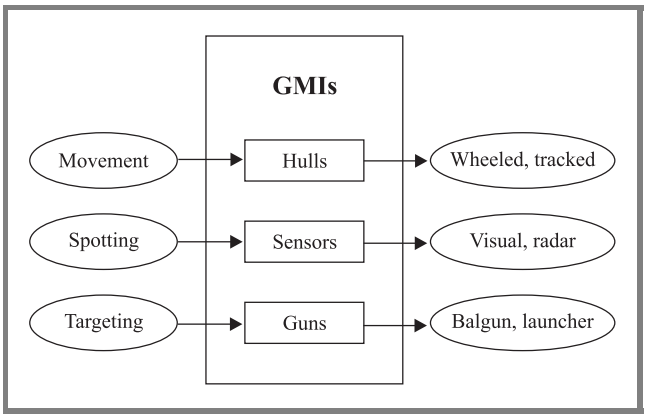


Fig. 6. The GMI provides the physical model invocation.

physical model that interfaces to a specific physical model, calls a proper physical model for implementation. In this case, it is a tracked physical model that has T-72 tracked hull characteristics. Figure 6 shows this relationship. Instead of using the specific model such as a T-72 tracked model, the SAF behaviors use an API that provides generic access to physical models. Thus, behavior models can uniformly invoke functionality regardless of the physical model that executes the invocation.

13. Modeling of tactical electronic and communications systems

The simulation used to support the GF-TCCS must have the capability to model the intelligence gathering systems such as the artillery and mortar locating radars, as well as the electronic intelligence and signal intelligence systems. The ModSAF generic sensor model (GSM) facilitates the implementation of new sensing algorithms. For example, a ground-based sensor can be created that automatically reports activity of targets within a designated area. ModSAF also has a generic radio physical model, that can be used to instantiate any of the existing and planned Czech military radio types.

14. Integration architecture

The concept for integration of the ModSAF simulation with the GF-TCCS is to build a conversion utility that will take the pertinent information from the GF-TCCS databases and insert or modify the information in the ModSAF common data-base. This conversion utility will simplify the existing laborious process of loading and instantiating the simulation, so that the entire process can be accomplished automatically with no manual user interventions. There are major information areas that require conversion and instantiation:

Friendly forces information as it is known and recorded in the GF-TCCS:

- unit table of organization and entity missions;
- unit placement on the synthetic terrain;
- unit strength and supply status;
- vehicle status and supply of fuel, crew, and ammunition.

Enemy force information according to intelligence estimates and enemy order of battle:

- unit table of organization and entity missions;
- unit placement on the synthetic terrain;
- unit strength and supply status;
- vehicle status and supply of fuel, crew, and ammunition.

The information for each of these areas for both the friendly and enemy forces is contained within the GF-TCCS databases. The integration effort is to provide the conversion between the different formats and to expand or augment the data as it is loaded into the ModSAF simulation. In addition, there will be a development process to build the user interface necessary to cause the conversion and data loading, and to integrate the display of the simulation data on the tactical situational display in the Tactical Operations Center (TOC) as the simulation is running.
The VTUE development team that is building the GF-TCCS has been examining the situation with the ModSAF simulation being used at the Czech Military Academy, and the contractor, DelInfo, is using this knowledge for the development of the interface. Presently, the GF-TCCS is in the development stage. The first increment of MCS SW package is used in the Rapid Reaction Forces of the ACR since 2001.
The ModSAF simulation is currently being used to support training exercises at the Brigade and below level at the Military Academy in Brno. The current estimate for the integration of the simulations with the command and control system is for starting the integration in mid 2003, with a demonstration capability available in the first quarter 2004, and full implementation by the end of 2004.

15. Conclusion

This paper has presented an overview of the architecture approach for integrating the ModSAF constructive simulation with the GF-TCCS. The goal of the enclosed architecture has been to present a basic decision support system infrastructure that is both flexible and presents a minimum development risk. This is achieved by leveraging both marketplace capabilities and verified market trends to simplify the information-processing approach and subsequent implementation. The backbone of the architecture is based on internal technology, in particular the HTTP and XML technologies. These technologies, combined with important architectural principals of loosely-coupled distributed computers, are molded into a baseline C2 infrastructure upon which GF-TCCS applications may be developed.
Although there are additional design efforts necessary to move forward, the GF-TCCS architecture represents a significant first step towards creating a base environment for the rapid development of C2 capabilities and software products. This architecture approach is ideally suited to spiral development and incremental upgrades.

References

[1] P. W. Holden and M. Šnajder, "Using wargame simulations to support decision making at the tactical levels", in *Proc. ITTSEC*, Orlando, USA, 2002.

[2] M. Šnajder, "The Czech approach in the development a ground forces tactical command and control system", in *Conf. CIS*, NATO HQ, Brussels, Belgium, 2002.

Milan Šnajder (Assoc. Prof., Ph.D.) is a head of IT division in Military Technical Institute of Electronics in Prague, Czech Republic. He is a project manager of Ground Forces Command and Control System of the Army of the Czech Republic. He is external lecturer in Military Academy in Brno. He is voting member of the Czech Republic delegation in the Information System Technology (IST) panel of the NATO Research Technology Organisation (RTO).
e-mail: msnajder@vtue.cz
Military Technical Institute of Electronic
Pod Vodovodem 2
158 00 Prague 5, Czech Republik

Philip W. Holden is a division manager for Science Applications International Corporation (SAIC). He is responsible for the International Training Center Program and was instrumental in the design, development, fielding and sustainment of the simulation products used at the Czech Military Academy in Brno, the Czech Training Center in Vyskov, the Slovak Military Academy in Liptovsky Mikulas, and at the Slovak Air Force Academy in Kosice. He was also part of the team that produced the first Army Training and Evaluation Programs (ARTEP) for the Field Artillery.
e-mail: holdenp@saic.com
Science Applications International
Corporation (SAIC)
12901 Science Drive
Orlando, FL 32826, USA

Paper

Web-based e-learning environment

Milan Mišovič

Abstract — Using a progressive information technologies for development of web-based courses and their administration brings a lot of practical and theoretical problems. We know web course problem is only small solution concerning construction of a large entire web-based e-learning environment. One of a practical problem is how to construct web-based electronic courses that have to meet international AICC standards. Implementation of such strict statements resulted in a lot of small or larger difficulties if we had used an elementary HTML editor. Using a special web course oriented editors can absolutely solve this problem. The second problem is construction of the web-based e-learning application that can administrate such web courses and takes into consideration AICC regulations. Development of such web applications is founded on the latest web technologies. This article introduces one approach to the modeling of the two most important components (web-based course, web-based application) of the web-based e-learning environment, convenient for Military Academy in Brno and Czech army. The article outlines the structure of web subject, e-learning environment and their implementation[1]. The LMS's structure and its functionality, based on a snaps algebra, belong to important results of the article.

Keywords — distance learning, e-learning, web-based course, web-based e-learning environment, web distributed application, learning management system, web course subjects, clients, access permission rights, communication, administration of web-based courses, e-learning functions, client interface, LMS manager, manager operations, activity snaps, algebra of snaps.

1. Introduction

According to the American Council on Education (Guiding Principles for Distance Education in a Learning Society, 1996) we can take over its definition of distant learning as: "Distance learning is a system and a process that connects all participants and resources (learners with distributed learning resources, educators with learners, learners with learners)". The term "distance learning" is very often interchanged by the second term "distance education". There are two categories of distance education delivery systems: synchronous and asynchronous. The first system requires the simultaneous participation of all students and instructors, by another words – interaction is done in "real time". There are known such implementations as: interactive TV, tele- and video-conferencing, web-conferencing, synchronous chat, virtual whiteboard etc.

[1]Some results from a complex solution for the project of web-based learning environment.

On the other hand, the second system does not require the simultaneous participation of all students and instructors. Students do not need to be gathered together in the same place at the same time. They can select their education time and learning resources that are needed. This system of education is more flexible then the former. There are known such implementations as: correspondence courses, videotaped courses, e-mail and web-based courses, etc.
The present and future possibility of distance learning is very tightly connected with the development of electronic telecomputing technology, especially with development of Internet/Intranet and multimedia technology. Modern web-based electronic learning environment can significantly integrate such tools for presentation as voice, video, and data connections between and among instructors, learners, subject matters experts, virtual libraries and Internet resources. Modern web-based e-learning consists of several mutually dependent parts, included in its strategy:

– development of web course content (subject matters);

– pedagogical approach;

– web course administration;

– web course distribution.

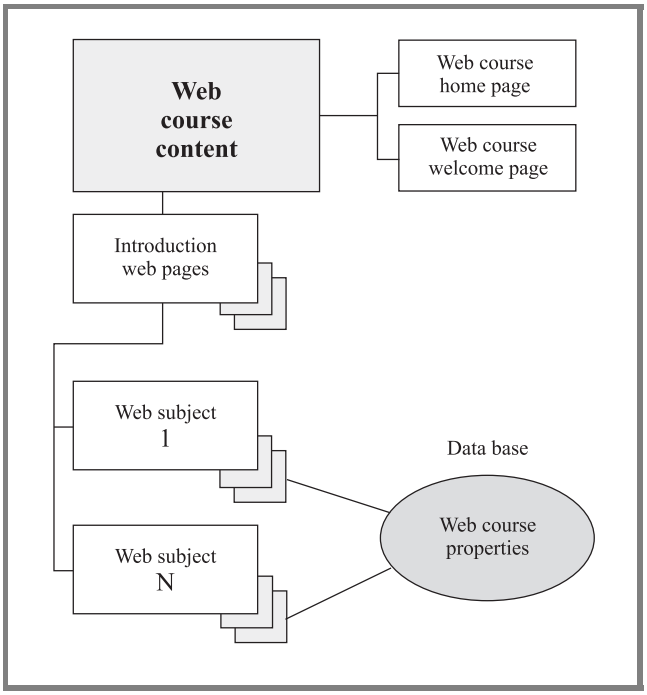


Fig. 1. Web course content structure.

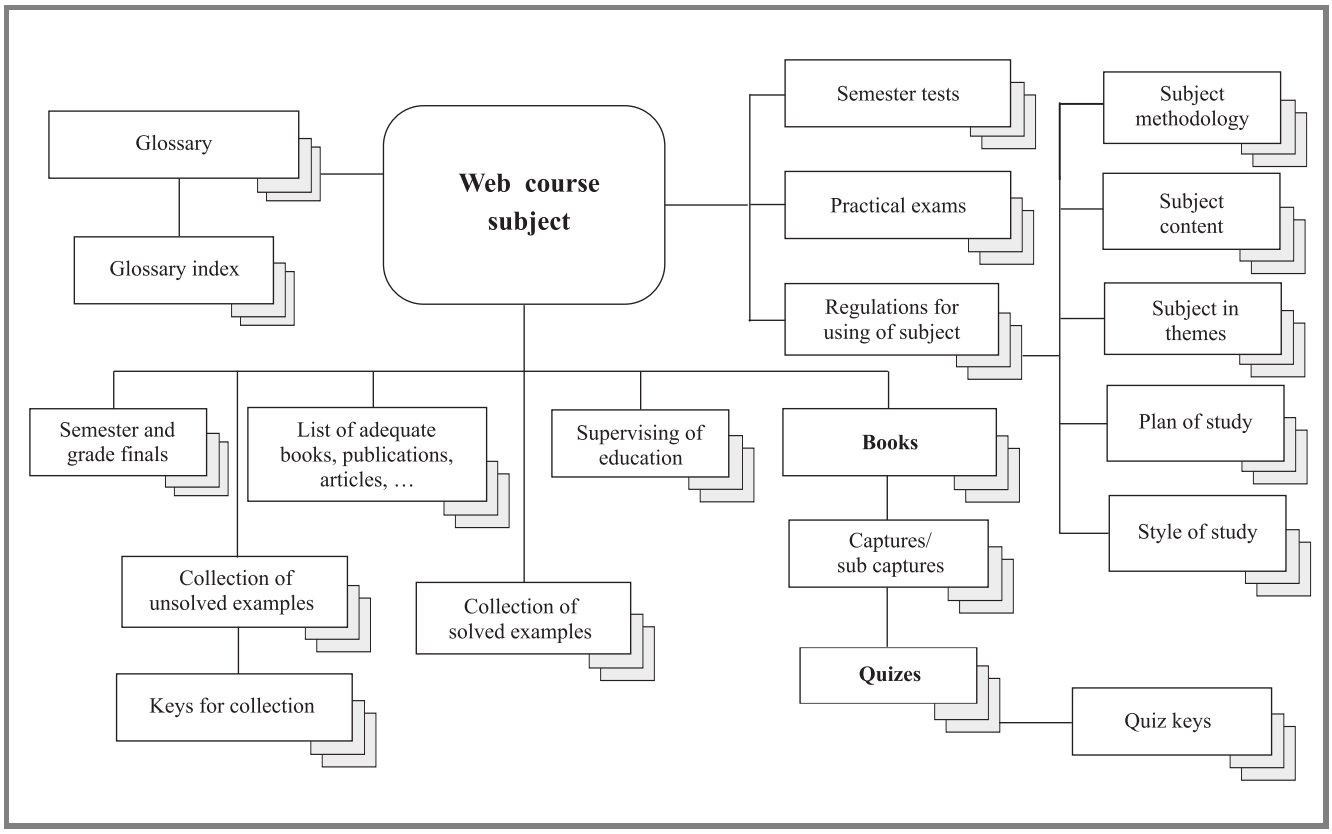


Fig. 2. Web course subject structure.

Development of web-based e-learning courses is founded on web technologies HTML and DHTML. There are very often used special course editors. These editors can usually solve problem of web course distribution, too. Pedagogical approach and web course administration are enabled by a special web-based e-learning application (learning manegement system – LMS). Modeling of such applications has to follow a lot of important rules.

2. Web course's structure

Generally, any web course can be constructed as a set of several web subjects. Even though, the common case is only one web subject.

The content of any web course can consist of the following parts (Fig. 1):

– home page, welcome page and introducing pages for all web subjects;

– web subjects, which are expressed very often by HTML or DHTML pages;

– the data base "web course properties"; this data base points to information linkages among the components of web subjects.

Figure 2 outlines possible web subject structures. This structure corresponds to the current subject in Military College Education.

3. Web-based learning environment structure

In order to model the e-learning problem domain we have to recognize all its structure, properties etc. Finding out of

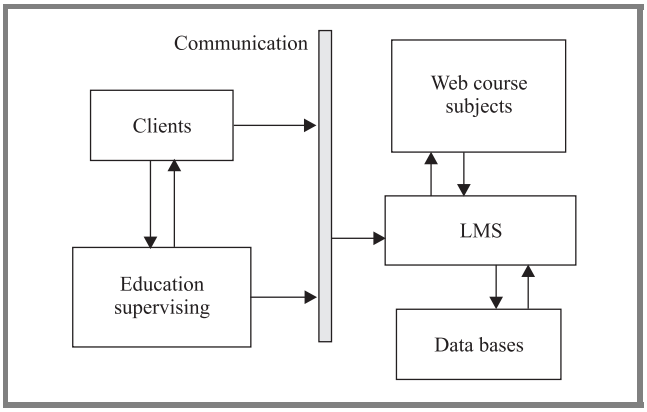


Fig. 3. Structure of web-based learning environment in subjects.

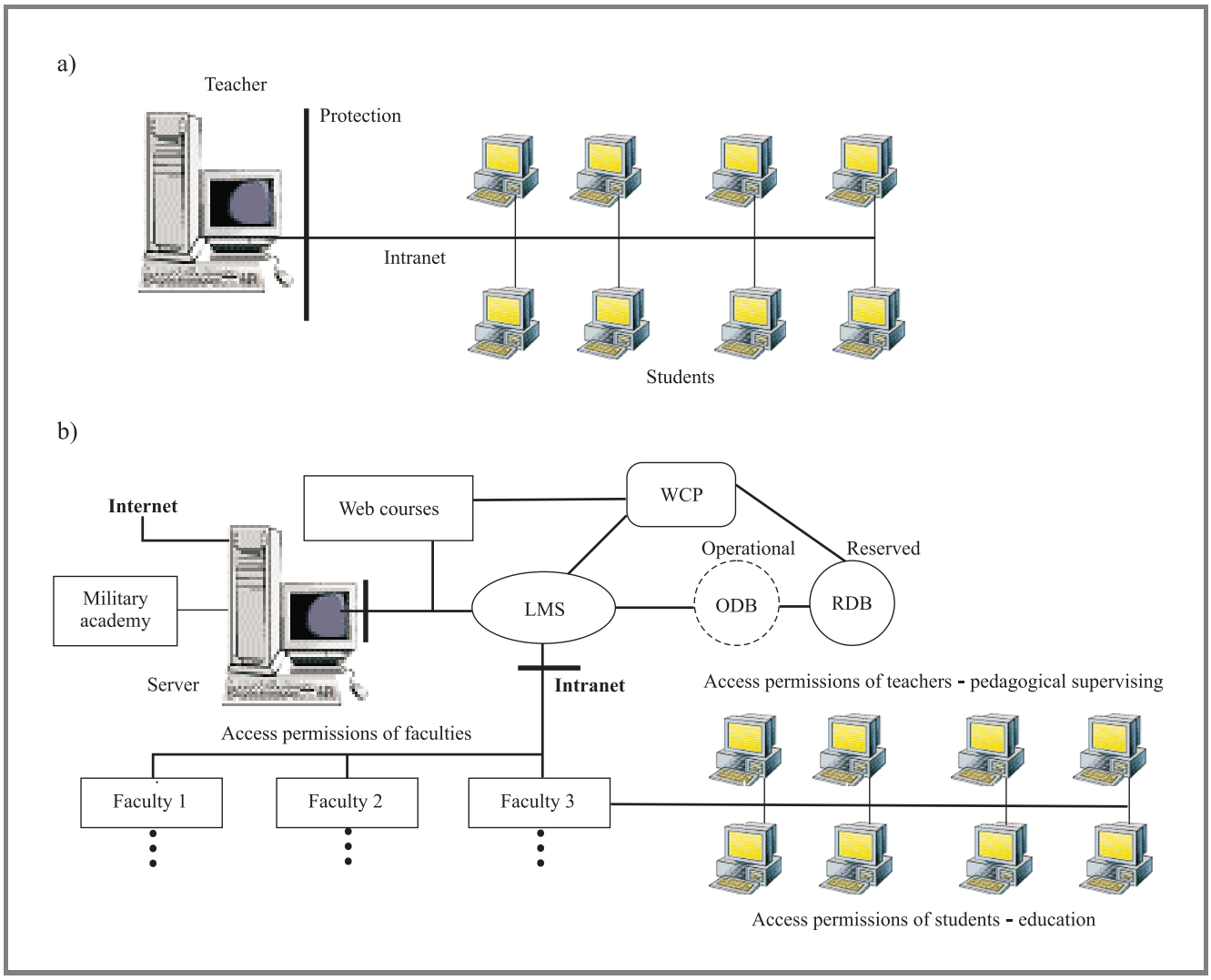


Fig. 4. Technical and software equipment of the Intranet and Internet web-based learning environment: (a) teacher's computer: Win98/Win2000, personal web server or Internet information server, learning management system, web course subjects; (b) military academy's server.

a good structure is the main problem. Next structure can be acceptable:

- web course subjects;
- clients of web course;
- education supervising;
- communication;
- data bases: "web course properties", "operational DB", "reserved DB";
- administration of web course – learning management system.

This structure enables connectivity clients – web subjects – data base only by means of the LMS (Fig. 3). This idea can influence the data model and web subject protection (code on a server farm).
The LMS will certainly appear as a web-based application with more then three layers (client code, server code, distributed data and distributed code).

Figure 4 outlines elementary implementation of the LMS on teacher's computer and on military academy's server.

4. The main e-learning's functions

The analysis of web-based e-learning environment subjects has shown that the list of the main functions should contain at least the following functions:

1. To enable client access to the web subjects according to client access permission.
2. To provide a guest show for some clients.
3. To make up possibility for a full text searching in the content of web subjects.
4. To register students activities in web subjects (monitoring).

5. To enable all teachers to get information about any student activities.
6. To give study possibilities for young teachers.
7. To organize all types of communication.
8. To provide comfort for modification of web subject content.
9. To make up a flexible communication with data base server.
10. To protect all information which has fatal meaning for LMS.
11. To record all activities in web-based e-learning environment.

5. Learning management system

The learning management system (Fig. 5) is web-based e-learning application for web course administration. Its structure corresponds to the web-based e-learning environment (Fig. 6). Generally, it should play several important roles, for example:

– administration and communication roles;
– pedagogical role;
– service role.

Special managers or assistants perform each role in the LMS. The LMS core reacts on all internal events.

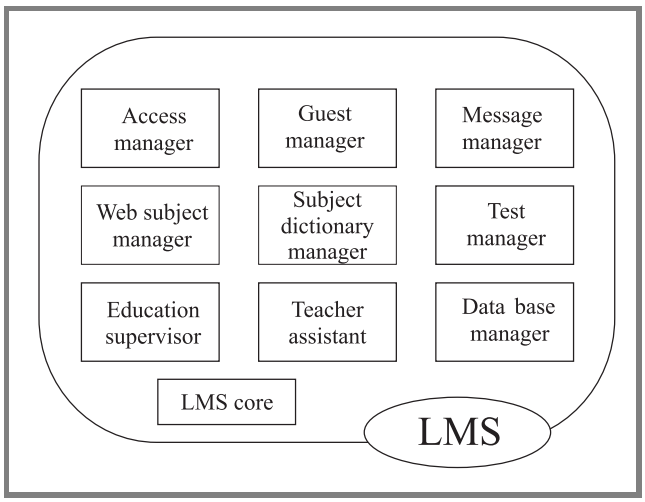


Fig. 5. The LMS's structure.

All makers of web courses have to follow AICC regulations (Fig. 7). It can be fulfilled by selection one of modern web course editors.
Interaction between the constructed web course and the LMS is performed by means of special "interaction doors".

6. Philosophic fundament of the LMS's work

Philosophy of the LMS' work is founded on producing and processing of snaps. The snap instances are the main products of the LMS. Snaps are produced in all reactions on events and their usage has a lot of possibilities for processing. Each instance of a snap is addressed by web course session-ID, web course subject-ID, client-ID, time and own snap-ID. Therefore client, web course session, web course subject, web subject test, transaction and time can describe each snap instance.
There are two types of snaps. The first snap is called registration snap. It contains only a name of performed transaction in its properties. In addition the previous possibilities, the second "full" snap contains also input and output parameters of executed transaction.
Creation of the snap is almost performed according to Fig. 8.
Internal meaning of a snap is given by its using in the LMS. All types of snaps are named in Fig. 9.
The LMS application generally accepts seven of client types:

– administrator,
– guarantee of web course,
– guarantee of web subject,
– instructor,
– student,
– external client and guest.

Valid access permission rights are defined in advance. The content of each access permission right consists of operations, those can be started by client from his interface.
By the way, snaps can be used for evaluation of learning progress quality, web subjects, pedagogical activities of instructors and convenience of the LMS.
This large using of snaps converts the LMS system to the learning quality management system (LQMS).

Definition 1. Entity "snap" is regarded as notation (C, W_c, W_s, W_t, T, t), where C, W_c, W_s, W_t, T; and t are sets of clients, web courses, web subjects, web tests, transactions and time $t_0 : t_1$ for e-learning environment session. The value of T determines a snap type.

The following text describes a special snap algebra which gives basic theoretical and practical approach to the snaps.
Operations in this algebra provide sufficient evaluation possibilities for all types of web course administrations. These operations are defined with respect to compatibility of snaps. Therefore, compatibility will be the basic property of snaps.

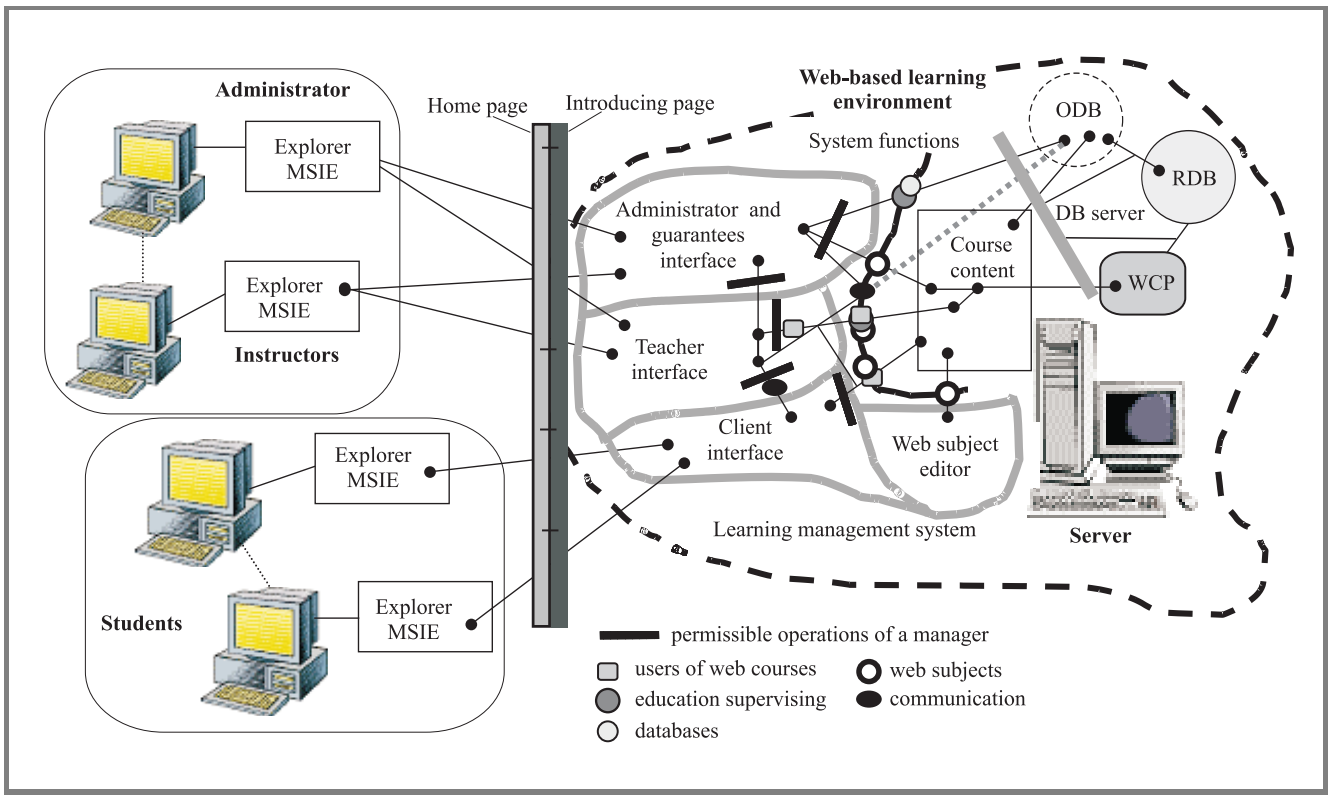


Fig. 6. Implementation of e-learning's functions in the LMS.

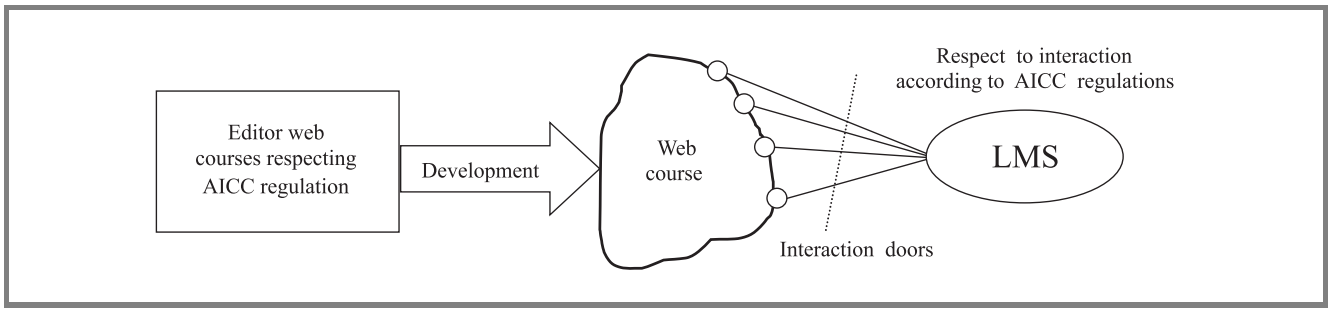


Fig. 7. Using of AICC regulations.

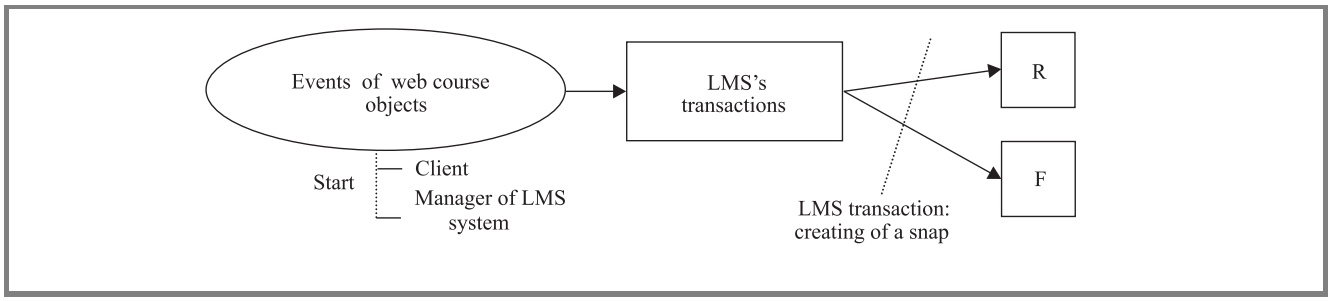


Fig. 8. Creating of snaps.

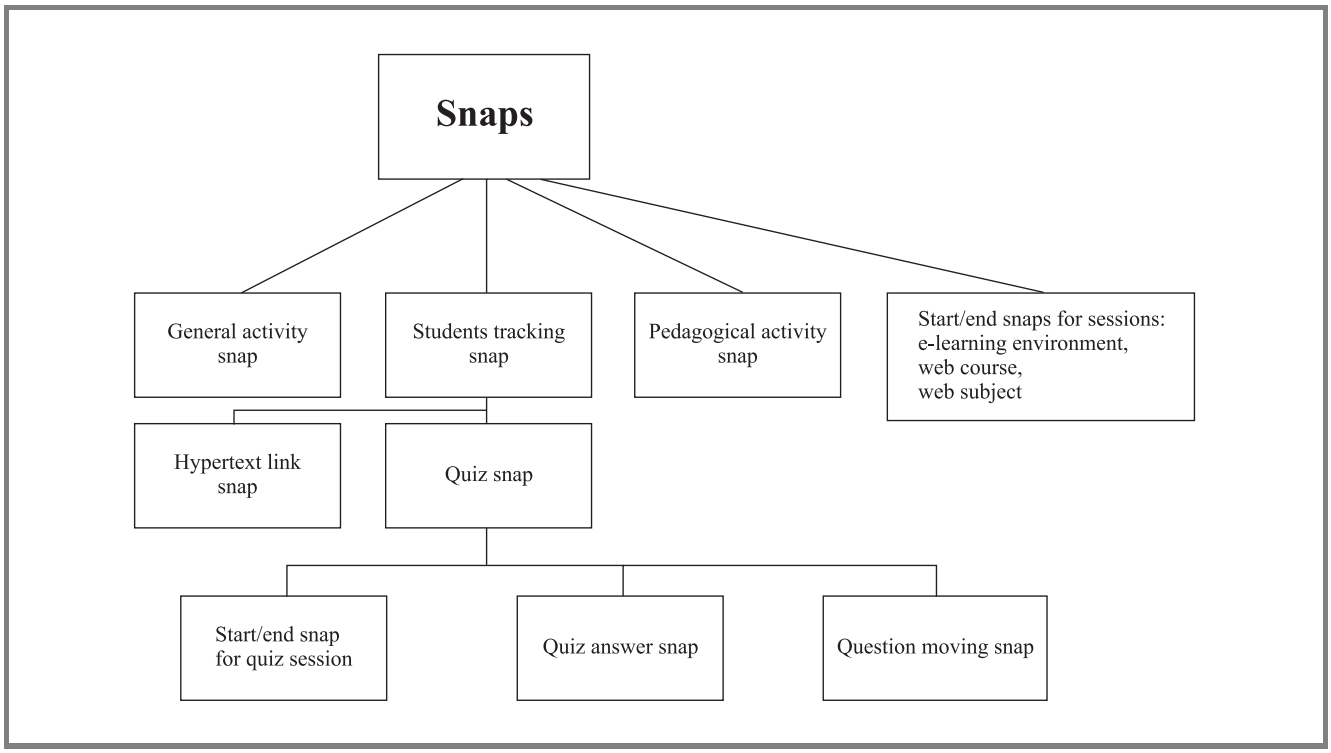


Fig. 9. Structure of web environment snaps.

7. The snap algebra

Set of elements .....Instances of snaps $\boldsymbol{x, y, z,}$...
Special element .....Empty snap $\boldsymbol{\varepsilon}$.

Definition 2. Snaps x, y are λ-compatible, if the sets
$\{C^x, W_c^x, W_s^x, W_t^x, T^x, t^x\}$,
$\{C^y, W_c^y, W_s^y, W_t^y, T^y, t^y\}$ have not empty disjunction.
λ-compatibility represents a certain number of equalities from the set
$\{C^x = C^y,\ W_c^x = W_c^y,\ W_s^x = W_s^y,\ W_t^x = W_t^y,\ T^x = T^y,\ t^x = t^y\}$.

Interpretation λ-compatibility by one equality leads to compatibility of the first level:

Client ..... $x \approx y \iff \lambda : C^x = C^y$
compatibility

Web course ..... $x \equiv y \iff \lambda : W_c^x = W_c^y$
compatibility

Web subject ..... $x \div y \iff \lambda : W_s^x = W_s^y$
compatibility

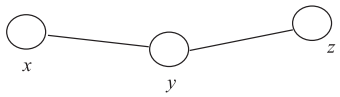
Transaction..... $x \div y \iff \lambda : T^x = T^y$
compatibility

λ-compatibilities of the second or higher level provide very useful sets of e-learning environment snaps. For example:

Client-subject $\lambda : C^x = C^y,\ \ W_s^x = W_s^y$
compatibility
generates all snaps of given client with the same web subject.
Subject-test $\lambda : W_s^x = W_s^y,\ W_t^x = W_t^y$
compatibility
generates all snaps of given web subject and web test in it.

Basic operation..... ***t*-concatenation** for $t_x < t_y$ of the both snaps x, y with the same λ-compatibility.
Notation..... $\boldsymbol{x \oplus y}$ ($y \oplus x$ is not defined).
Operation is not symmetrical but there is valid
$\boldsymbol{\varepsilon} \oplus x = x \oplus \boldsymbol{\varepsilon} = x$ and $x \oplus (y \oplus z) = (x \oplus y) \oplus z$.

Concatenation $\boldsymbol{x \oplus y \oplus z}$ of the snaps x, y, z with λ-compatibility, represents a string of the same compatibility. For graphical presentation such strings we can use following chart:



Another general operations..... $\boldsymbol{x < y},\ \ \boldsymbol{x > y}$,
(time relations)
Operation $\boldsymbol{x = y}$ is not defined and $\boldsymbol{x \neq y}$ is trivial.

Axioms.....There are not the same snaps for given client. Session with the e-learning environment is the highest snap.

There are four very important strings within learning progress administration:

1. Session with e-learning environment.
2. Session with web course.
3. Session with web subject.
4. Session with web subject test.

Each of these strings starts with "**Start snap with...**" and ends with "**End snap of...**". We can construct these strings by next operations. The snaps expressing the moving (clicking on hyperlinks) within a web subject have huge importance, too.
We can defined a specific number of convenient operations for each set of λ-compatible snaps, naturally with respect to requirements of administration types.
For example, we can suggest a list of such operations in the client-compatibility with respect to needs of learning progress administration:

E-learning environment session.....List of all e-learning
EES (C) environment sessions

Web course sessions.....List of all web course
WCSS $(C, t_0 : t_1)$ sessions

Web course session.....List of sessions
WCS $(C, W_c, t_0 : t_1)$ with given web course

Web subject sessions.....List of sessions
WSSS $(C, W_c, t_0 : t_1)$ with all subjects

Web subject session.....List of all sessions
WSS $(C, W_c, W_s, t_0 : t_1)$ with the same subject

Hyperlink web subject session.....String of used
HWSS $(C, W_c, W_s, t_0 : t_1)$ hyperlinks in subject

Hyperlink course session.....String of used
HCS $(C, W_c, W_s, t_0 : t_1)$ hyperlinks in course

Question moving string.....String of movements
QMS $(C, W_c, W_s, W_t, t_0 : t_1)$ in web test

Question answer string.....String of answers
QAS $(C, W_c, W_s, W_t, t_0 : t_1)$ in web test

By the same manner we can design a lot of useful operations in another very important sets of λ-compatible web snaps. For example, we can observe web subject using the loading of selected LMS transaction, pedagogical instructor activities etc. Naturally, if we find out that our list of snap types is not sufficient for the purposes of administration types, we can introduce another snap types. By this manner we can enlarge the LMS application.

8. Manager operations

The manager operations are derived from the main system functions of e-learning environment. The number of such operations can be larger then one hundred. Clients can start these operations from theirs interfaces only in accordance with access permission rights. These rights are different for each client type.
For example we introduce only several important manager operations:

ACCESS MANAGER	Guest show start	Login modification
Sign in/Sign out	Guest show end	Login creation
Registration	Login expiration	Login cancel
Login prologation	Personal data modification	Login prolongation prompt
Verification of access frequency	Login processing	Access permission right delegation
Access manager settings		
.....		

Message creation	Message archiving	Message canceling

MESSAGE MANAGER	Message restriction views	Message blocking
Message sending	Message symptom settings	Client calendar
View of messages	Message folder creation/canceling	Class table operations
Message view	Discussion table operations	Message manager settings
.....		

9. Conclusion

There are a lot of problems that belong to the range of web-based e-learning environment. This article has pointed only to the selected set of structural problems during its modeling. One complete modeling solution has been given in the research report "Problem Domain of eLearning" that was reviewed and accepted in January 23, 2002 (see [3]).

References

[1] J. Rosemberg, *e-Learning Strategies for Delivering Knowledge in the Digital Age*. New York: McGraw-Hill, 2001.

[2] M. Mišovič, Case study, project "Support of distance learning". Military Academy Brno, Czech Republic, 2001.

[3] M. Mišovič, Research report, "Problem Domain of eLearning", Military Academy Brno, Czech Republic, 2002.

[4] M. Mišovič, "Basics for Development of Web Courses", NATINEADS, Brno, Czech Republic, 2002.

[5] M. Mišovič, "eLearning in Military Professional Education", BELCOM, Prague, Czech Republic, 2002.

[6] L. Petras, "Contemporary state and future of military personal education", Case study, Military Academy Brno, Czech Republic, 2001.

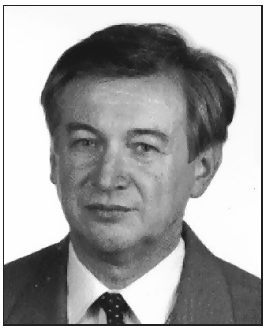
[7] J. R. Ballard and R. D. Kirkwood, "Interactive education for 21st century – the armed forces staff college and education of future decision-makers", I/ITSEC, Orlando, USA, 2001.

[8] J. Camacho, "The impact of advanced distributed learning (ADL) on joint readiness – an operational view", I/ITSEC, Orlando, USA, 2001.

[9] M. Myjak *et al.*, "The quest towards an advanced learning management system", I/ITSEC, Orlando, USA, 2001.

[10] F. Ionescu *et al.*, "Collaborative distributed learning environment for continuing education", ITEC, Lille, 2002.

[11] F. Gardiner, "Training quality system – where QMS meets LMS", ITEC, Lille, 2002.



Milan Mišovič an Associate Professor, senior lecturer, Doctor of Natural Sciences, Ph.D. of the Military Academy in Brno, Czech Republic; Head of IS Programming Tools Group. His area of scientific interest comprises: theoretical informatics, theoretical and practical aspects of software engineering and information engineering, IS methodologies, integrated development environments, development of web applications.
e-mail: milan.misovic@vabo.cz
Department of Automated Command Systems
and Informatics
Military Academy Brno
612 00 Brno, Czech Republic

Paper

An effective method of channels assignment for third generation cellular system

Marek Amanowicz, Piotr Gajewski, and Jarosław Krygier

Abstract — An original hybrid method of channels assignment for DS-CDMA system is discussed in the paper. This method combines standard PN codes assignment policy and dynamic channel assignment procedure that minimise the cost of channel assignment. OPNET simulation model DS-CDMA system was used for assessment of the hybrid method. The results of simulation presented in the paper confirmed that the proposed method importantly improves the quality of services in the third generation cellular system.

Keywords — W-CDMA, dynamic channel assignment, 3G.

1. Introduction

The work on the development of world-wide standard of third generation (3G) mobile radio system on the basis of UMTS/FPLMTS/IMT-2000 has been carried out by standardization institutes of Europe (ETSI), Japan (ARIB) and United States (ITA) under the umbrella of the International Telecommunication Union. Wideband code division multiple access (CDMA) has been chosen as the mainstream air interface solution for such networks. It is assumed that IMT-2000 will provide a wide range of services, especially multimedia and high-bit-rate packet data which require a high network capacity and appropriate quality of services.

The capacity of 3G-CDMA systems depends on the level of interference caused by each user. On the other hand the interferences depend on many factors such as user's activity, required data rate, geographical deployment of the users, power control accuracy, type of code sequences and its assignment scheme, etc. [1–3]. Most of these factors are out of control as they depend mainly on the user's behaviour. However power control procedures or channels assignment policy can be modified in a way which allows improving the system's capacity and QoS.

In our investigations we have focussed on the code sequences management method for DS-CDMA system. The hybrid channels assignment method (HCA) presented here combines two schemes, e.g. fixed and dynamic channel assignment.

The channels assignment policy for standard DS-CDMA system as well as system conditions concerning code management methods are shortly described in Section 2 of this paper. In Section 3 a general idea of HCA algorithm is presented and discussed. A general idea of HCA algorithm is presented and discussed. Simulation method was used for verification and validation of the proposed method. A description of OPNET simulation model as well as simulation results we have obtained are presented in Section 4.

2. Channel assignment for DS-CDMA systems

In widely known systems (like IS-95, IS-665, UMTS) the same sets of code sequences is used by each base station but with different phase shifts [4, 7]. The auto-correlation and cross-correlation features of such codes are specific for the defined codes family and in consequence they limit the system capacity [2, 3, 5]. The spreading codes management procedure is often based on simple choosing of one (or more) from predefined big family of codes (code shifts).

The user's requirements for services that should be provided by the mobile communications system cause the necessity of handling many services in the same time. Since such services are provided with different data rates, additional codes are used, for example orthogonal variable spreading factor (OVSF) channelisation codes in UMTS. They are designed for fitting different data rate services into the wideband radio channel with a constant bandwidth. The spreading of baseband signals is realised using so called scrambling codes (gold sequences in UMTS). OVSF codes are mutually orthogonal, so the basic co-channel interferences depend on scrambling auto and cross-correlation functions. By using different scrambling codes families with different correlation characteristics as well as by their efficient management it is possible to decrease the total level of interference and in consequence increase the system's capacity.

3. Hybrid channels assignment for 3G wireless systems

The channel assignment to a particular call could be realised by calculating minimal correlation factors at the time of assignment, with regard to each channel in the interference area or by sharing the channels into separate groups. In our case, cellular network is divided into three-cell groups like in sectorised systems. Let us assume the same radius in each sector. The proposed HCA channel assignment scheme is shown in Figs. 1 and 2. In this case

all N_k channels are divided into two groups F – "fixed" channels (can be used over the whole system) and T – "dynamic" channels (optimised channels over the sectors).

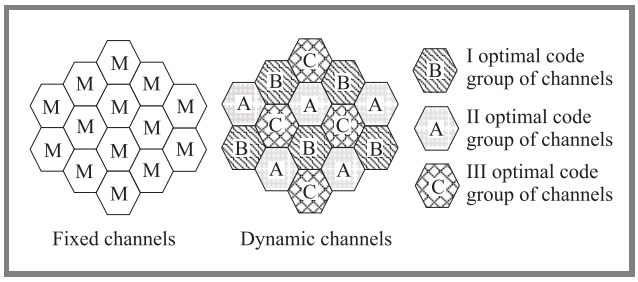


Fig. 1. Channels assignment scheme: M – number of fixed channels; A, B, C – numbers of dynamic channels.

All channels can be centrally managed. The first $F = \{1,2,\ldots,M\}$ channels are available in each cell. The last $T = \{M+1, M+2,\ldots,N_k\}$ channels are split into three sectors. In this group of channels we have $K-M$ channels optimised in each sector and $\left((N_k-M)/3\right)$ channels, which are used for priority calls and for handled over calls.

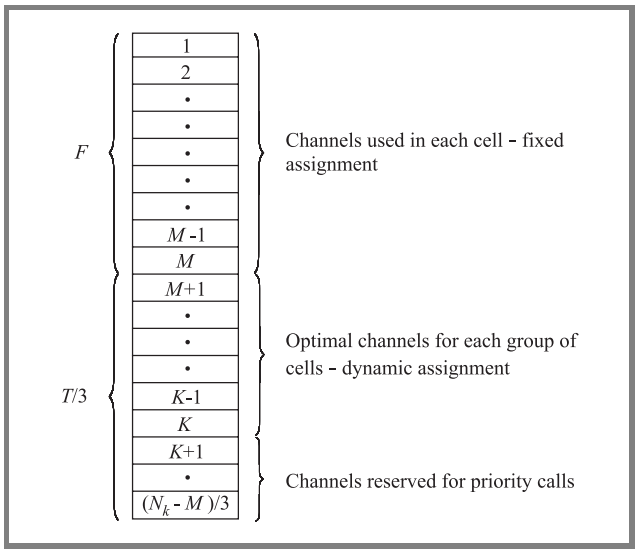


Fig. 2. Channels arrangement.

The total number of channels N_T in the system can be calculated as follows:

$$N_T = L \cdot M + [L/3] \cdot |F| + \left(L - 3 \cdot [L/3]\right) \cdot |F|/3, \qquad (1)$$

where: L is the number of cells in the system, $|F|$ – number of channels in the set F and $[L/3]$ – an integer part of $L/3$.

A channel assignment for a particular call from group F is quite simple. While a new call arrives the first free channel is used. Of course the criterion of minimal reuse distance should be fulfilled.

If all M fixed channels are already in usage then the dynamic channel is selected for a new call. For channel assignment from set T, an algorithm proposed in [6] is recommended. In this case, channels are allocated on the basis of minimisation of so-called cost function $C_x(i)$ defined for each available channel among all interference environment of cell x, e.g:

$$C_x(i^*) = \min_{i \in L(x)} \left\{C_x(i)\right\}, \qquad (2)$$

where: i^* – channel assigned to the new call in cell x, $L(x)$ – set of channels available for cell x (so excluding channels assigned to x and all cells from its interference environment), $C_x(i)$ – total cost of each channel from $L(x)$.

The cost $C_x(i)$ is calculated as a sum of weighted function $q_x(i)$ and the costs $C_x(k,i)$ of channel i assignment related to each k channel from interference neighbourhood $I(x)$, where $C_x(k,i)$ is an integer value from 0 to 3.

After releasing the dynamic channels are reallocated again in order to choose the optimal interference level using similar rule. In [6] this algorithm is proposed as a unique method for the whole cellular network. In our case, the allocation of dynamic channel is used only if there are not enough fixed channels.

4. Simulation experiments

Correctness of the above channels assignment algorithm as well as its efficiency for DS-CDMA cellular system was assessed during simulation experiments. Taking into account DS-CDMA system complexity, the OPNET simulation package was used for investigations. The basic elements of implemented network are shown in Fig. 3.

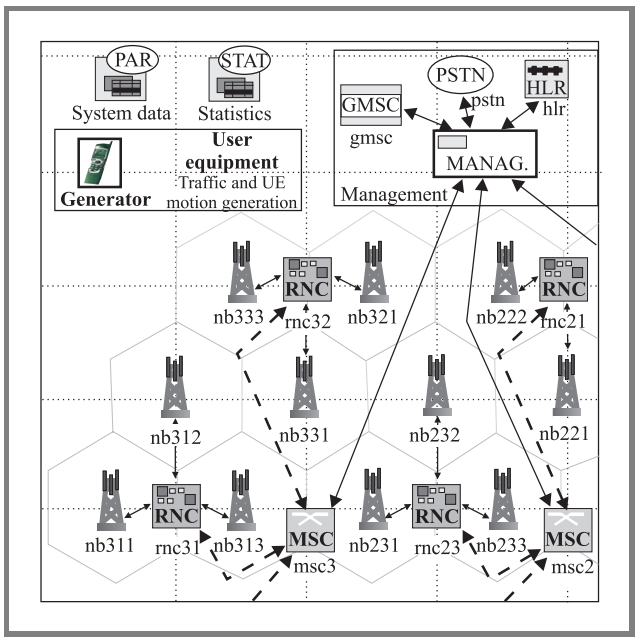


Fig. 3. An example of DS-CDMA cellular network.

The network consists of 34 base stations (nb), 10 radio network controllers (rnc), 3 mobile switching cen-

tres (msc) and other elements such as VLRs, HLR, PSTN, GMSC as well as traffic and user's movement generator.

Both calls and handover blocking probabilities were selected as the basic measures of the QoS for DS-CDMA system. These measures were also used for assessment of HCA algorithm efficiency.

The following system's parameters were assumed for simulation:

- UL/DL carrier frequency: 1922.6/2112.6 MHz;
- chip rate: 3.84 Mchps;
- propagation model: COST 231 Walfish-Ikegami;
- user's effective data rate: 12.2, 64, 144, 384;
- user's activity coefficient: 100% (0.75 Erl per user);
- mobile station (UE – user equipment) max. power: 21 dBm;
- UE antenna gain: 0 dB;
- base station (NB) max. power: 43 dBm;
- NB antenna gain: 10 dB;
- UE and NB sensitivity: -110 dBm;
- handover type: soft (2 base stations in the active set).

In CDMA systems interferences can be caused both by users and base stations, as it is shown in Fig. 4.

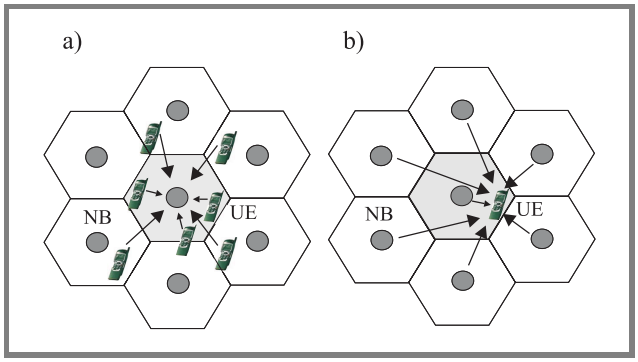


Fig. 4. Interferences caused by users UE (a) and by base stations NB (b).

The required level of E_b/N_0 in NB location (Uplink – UL) is:

$$\left(\frac{E_b}{N_0}\right)_{req} = \frac{P_{od}\cdot(SF)}{I_{\text{intra\_NB}} + I_{\text{inter NB}} + N_0}, \tag{3}$$

where: SF – spreading factor, P_{od} – received signal power, $I_{\text{intra\_NB}}$, $I_{\text{inter\_NB}}$ – intra and intercell interferences, N_0 – noise power density.

Let us assume that power control is performed ideally. It means that signal power at NB receiver input incoming from all UEs is exactly the same:

$$L^n_{b1}\cdot P^n_{nad1} = \ldots = L^n_{bm_n}\cdot P^n_{nadm_n}, \tag{4}$$

where: L^n_{b1} – signal attenuation in cell n from UE_1, P^n_{nad1} – transmitted signal power by UE_1 in cell n, m_k – number of UEs.

So, the signal power received by NB_n from UE_k is as follows:

$$P_{od} = L^n_{bk}\cdot P^n_{nadk}. \tag{5}$$

The level of intracell interferences that influence UE_k is defined as:

$$\begin{aligned} I_{\text{intra\_NB}} = {} & L^n_{b1}\cdot P^n_{nad1} + \ldots + L^n_{bk-1}\cdot P^n_{nadk-1} + \\ & + L^n_{bk+1}\cdot P^n_{nadk+1} + \ldots + L^n_{bm_n}\cdot P^n_{nadm_n} \end{aligned} \tag{6}$$

and level of intercell interferences as:

$$\begin{aligned} I_{\text{inter\_NB}} = {} & \sum_{i=1}^{m_1} L^1_{b1}\cdot P^1_{nadi} + \ldots + \sum_{i=1}^{m_{n-1}} L^{n-1}_{bi}\cdot P^{n-1}_{nadi} + \\ & + \sum_{i=1}^{m_{n+1}} L^{n+1}_{bi}\cdot P^{n+1}_{nadi} + \ldots + \sum_{i=1}^{m_L} L^L_{bi}\cdot P^L_{nadi}, \end{aligned} \tag{7}$$

where: L – number of base stations in the system, $m_1, m_2, \ldots, m_n, \ldots, m_L$ – number of active users.

A similar situation is in downlink (DL) calculations. The required level of E_b/N_0 in UE can be written as:

$$\left(\frac{E_b}{N_0}\right)_{req} = \frac{P_{od}\cdot(SF)}{I_{\text{intra\_UE}} + I_{\text{inter UE}} + N_0}, \tag{8}$$

where: $I_{\text{intra\_UE}}$, $I_{\text{inter\_UE}}$ – intra and intercell interference that can be calculated from:

$$\begin{aligned} I_{\text{intra\_UE}} = {} & L^n_{bk}\left(P^n_{nad1} + \ldots + P^n_{nadk} + \right. \\ & \left. + P^n_{nadk+1} + \ldots + P^n_{nadm_n}\right) \end{aligned} \tag{9}$$

and

$$\begin{aligned} I_{\text{inter\_UE}} = {} & L^1_{bk}\sum_{i=1}^{m_1} P^1_{nadi} + \ldots + L^{n-1}_{bk}\sum_{i=1}^{m_{n-1}} P^{n-1}_{nadi} + \\ & + L^{n+1}_{bk}\sum_{i=1}^{m_{n+1}} P^{n+1}_{nadi} + \ldots + L^L_{bk}\sum_{i=1}^{m_L} P^L_{nadi}. \end{aligned} \tag{10}$$

Above functional equations describe DS-CDMA system behaviour. During simulation, required E_b/N_0 can be calculated and on this basis the decision concerning the calls or the handovers blocking is made. Such a situation takes

place when E_b/N_0 exceeds the level shown in Table 1. In this case dynamic channel assignment procedure is running.

Table 1
Required level of E_b/N_0

Effective data rate [kbit/s]	12.2	64	144	384
Uplink	5.1	1.7	0.9	0.9
Downlink	7.9			

Figures 5 – 10 show the simulation results. Standard channels assignment scheme based on [7] and the HCA method (denoted in the figures as the modified) are compared.

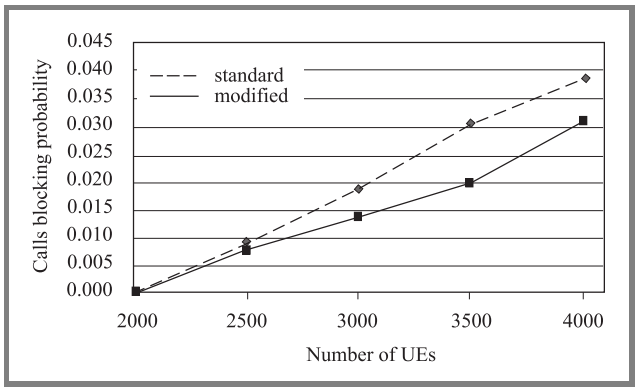


Fig. 5. Call blocking probability versus number of UEs.

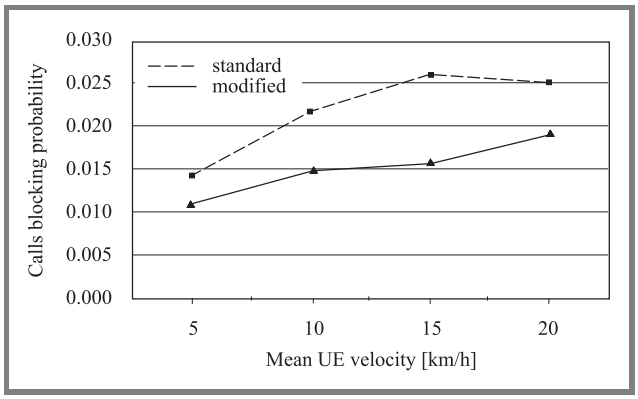


Fig. 6. Call blocking probability versus mean velocity of UEs.

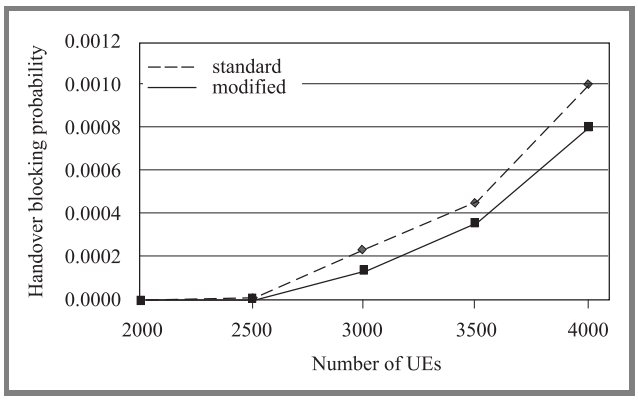


Fig. 7. Handover blocking probability versus number of UEs.

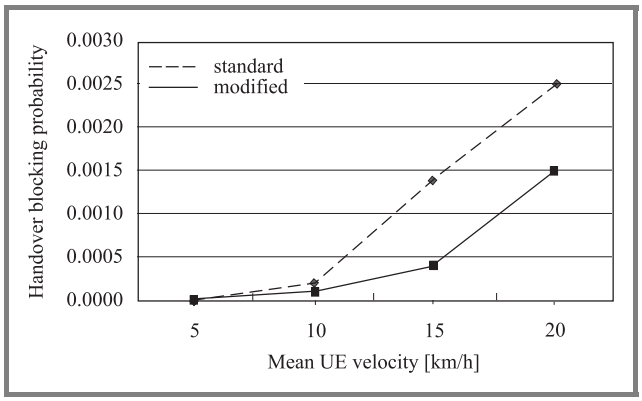


Fig. 8. Handover blocking probability versus mean velocity of UE.

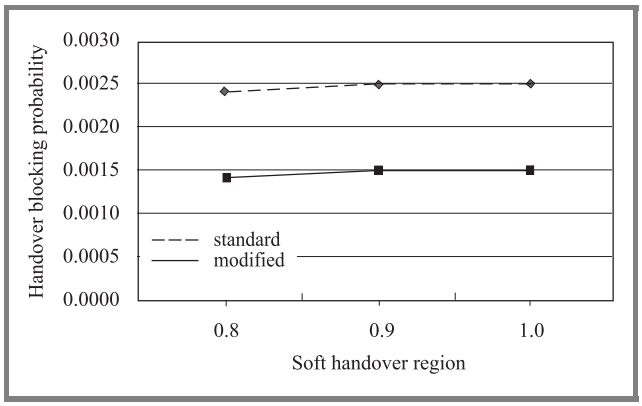


Fig. 9. Handover blocking probability versus soft handover region.

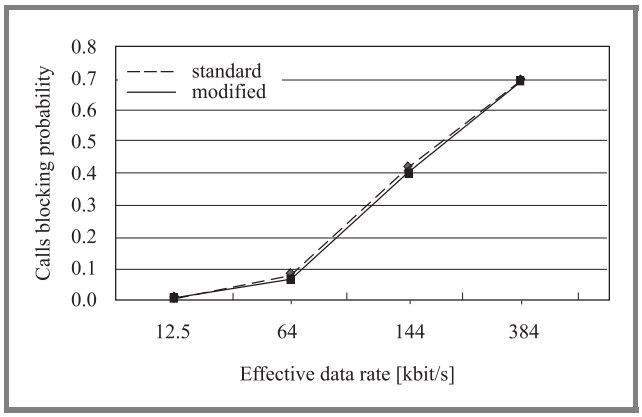


Fig. 10. Call blocking probability versus UE effective data rate.

The influence of the number of UEs on QoS measures is shown in Figs. 5 and 7. The traffic intensity in the system in this case varies from 1500 Erl to 3000 Erl.
Call and handover blocking probabilities versus mean velocity of UEs are shown in Figs. 6 and 8.
Handover blocking probability versus so-called soft handover region is shown in Fig. 9. The soft handover region

is defined here as the ratio of cell radius and radius of the circle, where UE starts soft handover process.

Calls blocking probability versus UE effective data rate is presented in Fig. 10.

5. Conclusion

On the basis of the simulation results shown in previous section, we can notice that HCA scheme significantly decreases calls and handover blocking probability in comparison with the standard method. The QoS improvement is particularly visible in case of high system load, where HCA can effectively manage the channels. The same situation is when UEs increase their velocity. In the microcellular systems, number of users crossing the cell boundary is very high. So the traffic caused by this effect is high (even higher then basic traffic). By using HCA we can see significant increase the QoS for higher user's velocity.

From Fig. 10 we can notice that the blocking probability is very high for effective data rates above 12.2 kbit/s. It is caused by high traffic (the same for all data rates) generated by the users. Summarising we can conclude that the proposed channel allocation method seems to be suitable for third generation cellular system.

References

[1] W. C. Lee, "Overview of cellular CDMA", *IEEE Trans. Veh. Technol.*, vol. 40, no. 2, 1991.

[2] H. Jeon, S. Shin, T. Hwang, and C. Kang, "Revers link capacity analysis of a CDMA cellular system with mixed cell sizes", *IEEE Trans. Veh. Technol.*, vol. 49, no. 6, 2000.

[3] P. Gajewski, J. Krygier, J. Łopatka, and J. Buczyński, "Performance of a DS-CDMA system with dynamic channel allocation and soft handover", in *Proc. ISSSTA'98*, RPA, 1998.

[4] D. E. Everitt, "Traffic engineering of the radio interface for cellular mobile networks", *Proc. IEEE*, vol. 82, no. 9, 1994.

[5] M. B. Pursley and H. F. A. Roefs, "Numerical evaluation of correlation parameters for optional phases of binary shift-register sequences", *IEEE Trans. Commun.*, vol. COM-27, no. 10, 1979.

[6] E. Del Re, R. Fantacci, and G. Gimbene, "Handover and dynamic channel allocation techniques in mobile cellular networks", *IEEE Trans. Veh. Technol.*, vol. VT-44, no. 2, 1995.

[7] 3GPP UMTS Techn. Specif. TS 25.213: "Spreading and modulation", Dec. 2001.

Marek Amanowicz was born in Poland, in 1946. He received M.Sc., Ph.D. and D.Sc. degrees from the Military University of Technology, Warsaw, Poland, in 1970, 1978 and 1990, respectively, all in telecommunication engineering. In 2001 he was promoted to the professor's title. He was engaged in many research projects, especially in the field of communications and information systems engineering, mobile and personal communications, antennas and propagation, communications and information systems modelling and simulation, communications and information systems interoperability, network management and electronics warfare. He is an author or co-author of over 180 scientific papers and research reports.
e-mail: amanowic@wil.waw.pl
Military Communication Institute
05-130 Zegrze, Poland

Piotr Z. Gajewski received the M.Sc. and Ph.D. degrees from the Military University of Technology (MUT), Warsaw, Poland, in 1970 and 1979, respectively, both in telecommunication engineering. Since 1970 he works at Electronics Faculty of Military University of Technology (EF MUT) as a scientist and lecturer in communications systems (radios, cellular, microcellular), signal processing, adaptive techniques in communication and communications and information systems interoperability. He was an Associate Professor at Telecommunication System Institute of EF MUT from 1980 to 1990. From 1990 to 1993 he was Deputy Dean of EF MUT. Currently he is the Director of Communications Systems Institute of EF MUT. He is an author (co-author) of over 80 journal publications and conference papers as well as 4 monographs. He is a member of the IEEE Vehicular Technology and Communications Societies. He is also a founder member of the Polish Chapter of Armed Forces Communications and Electronics Association.
e-mail: pgajewski@wel.wat.edu.pl
Military University of Technology
Kaliskiego st 2
00-908 Warsaw, Poland

Jarosław Krygier – for biography, see this issue, p. 18.

INFORMATION FOR AUTHORS

The *Journal of Telecommunications and Information Technology* is published quarterly. It comprises original contributions, both regular papers and letters, dealing with a broad range of topics related to telecommunications and information technology. Items included in the journal report primary and/or experimental research results, which advance the base of scientific and technological knowledge about telecommunications and information technology.

The *Journal* is dedicated to publishing research results which advance the level of current research or add to the understanding of problems related to modulation and signal design, wireless communications, optical communications and photonic systems, speech devices, image and signal processing, transmission systems, network architecture, coding and communication theory, as well as information technology. Suitable research-related manuscripts should hold the potential to advance the technological base of telecommunications and information technology. Tutorial and review papers are published by invitation only.

Papers published by invitation and regular papers should contain up to 15 and 8 printed pages respectively (one printed page corresponds approximately to 3 double-space pages of manuscript, where one page contains approximately 2000 characters).

Manuscript: An original and two copies of the manuscript must be submitted, each completed with all illustrations and tables attached at the end of the papers. Tables and figures have to be numbered consecutively with Arabic numerals. The manuscript must include an abstract limited to approximately 100 words. The abstract should contain four points: statement of the problem, assumptions and methodology, results and conclusion, or discussion, of the importance of the results. The manuscript should be double-spaced on only one side of each A4 sheet (210 × 297 mm). Computer notation such as Fortran, Matlab, Mathematica etc., for formulae, indices, etc., is not acceptable and will result in automatic rejection of the manuscript. The style of references, abbreviations, etc., should follow the standard IEEE format.

References should be marked in the text by Arabic numerals in square brackets and listed at the end of the paper in order of their appearance in the text, including exclusively publications cited inside. The **reference entry** (correctly punctuated according to the following rules and examples) **has to contain**.

From journals and other serial publications: initial(s) and second name(s) of the author(s), full title of publication (transliterated into Latin characters in case it is in Russian, possibly preceded by the title in Russian characters), appropriately abbreviated title of periodical, volume number, first and last page number, year. E.g.:

[1] Y. Namihira, "Relationship between nonlinear effective area and modefield diameter for dispersion shifted fibres", *Electron. Lett.*, vol. 30, no. 3, pp. 262–264, 1994.

From non-periodical, collective publications: as above, but after title – the name(s) of editor(s), title of volume and/or edition number, publisher(s) name(s) and place of edition, inclusive pages of article, year. E.g.:

[2] S. Demri, E. Orlowska, "Informational representability: Abstract models versus concrete models" in *Fuzzy Sets, Logics and Reasoning about Knowledge*, D. Dubois and H. Prade, Eds. Dordrecht: Kluwer, 1999, pp. 301–314.

From books: initial(s) and name(s) of the author(s), place of edition, title, publisher(s), year. E.g.:

[3] C. Kittel, *Introduction to Solid State Physics*. New York: Wiley, 1986.

Figure captions should be started on separate sheet of papers and must be double-spaced.

Illustration: Original illustrations should be submitted. All line drawings should be prepared on white drawing paper in black India ink. Drawings in Corel Draw and Postscript formats are preferred. Colour illustrations are accepted only in exceptional circumstances. Lettering should be large enough to be readily legible when drawing is reduced to two- or one-column width – as much as 4:1 reduction from the original. Photographs should be used sparingly. All photographs must be gloss prints. All materials, including drawings and photographs, should be no larger than 175 × 260 mm.

Page number: Number all pages, including tables and illustrations (which should be grouped at the end), in a single series, with no omitted numbers.

Electronic form: A floppy disk together with the hard copy of the manuscript should be submitted. It is important to ensure that the diskette version and the printed version are identical. The diskette should be labelled with the following information: a) the operating system and word-processing software used, b) in case of UNIX media, the method of extraction (i.e. tar) applied, c) file name(s) related to manuscript. The diskette should be properly packed in order to avoid possible damage during transit.

Among various acceptable word processor formats, TEX and LATEX are preferable. The *Journal's* style file is available to authors.

Galley proofs: Proofs should be returned by authors as soon as possible. In other cases, the article will be proofread against manuscript by the editor and printed without the author's corrections. Remarks to the errata should be provided within two weeks after receiving the offprints.

The copy of the "Journal" shall be provided to each author of papers.

Copyright: Manuscript submitted to this journal may not have been published and will not be simultaneously submitted or published elsewhere. Submitting a manuscript, the authors agree to automatically transfer the copyright for their article to the publisher if and when the article is accepted for publication. The copyright comprises the exclusive rights to reproduce and distribute the article, including reprints and also all translation rights. No part of the present journal may be reproduced in any form nor transmitted or translated into a machine language without permission in written form from the publisher.

Biographies and photographs of authors are printed with each paper. Send a brief professional biography not exceeding 100 words and a gloss photo of each author with the manuscript.

National Institute
of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Editorial Office

tel. +48(22) 872 43 88
tel./fax: +48(22) 512 84 00
e-mail: redakcja@itl.waw.pl
http://www.itl.waw.pl/jtit