JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

4/2004

Military communications systems

Special issue edited by Wojciech Burakowski and Luigi Bella

Developing standards for interoperability of tactical communications systems
Ch. J. Echols — Paper 3

Enterprise integration lessons learned
A. D. Ta and S. Starsman — Paper 6

Modelling and simulation of combat operations in SimCombCalculator application
T. Nowicki — Paper 14

Seamless roaming between UMTS and IEEE 802.11 networks
P. Matusz, P. Machań, and J. Woźniak — Paper 21

Boolean feedback functions for full-length nonlinear shift registers
I. Janicka-Lipska and J. Stokłosa — Paper 28

Technology solutions for coalition operations
D. Wiemer — Paper 31

User services of tactical communications in the digital age
E. Çiftçibaşı Erkan and Ş. Uzun — Paper 39

Sharing tactical data in a network-enabled coalition
J. Busch and R. Russo — Paper 45

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

This special issue is entitled "Military communications systems" and contains 15 carefully selected papers, among them 5 papers presented during *5th NATO Regional Military Communications and Information Systems (NATO RCMCIS 2003)* while the rest 10 papers are planning to be presented during this year *6th NATO RCMCIS 2004*. The NATO RCMCIS Conference is the annual event held in Zegrze since 1999. It is a joint activity of three organizations: Military Communication Institute from Zegrze (Poland), NATO NC3A Agency from Den Haag (The Netherlands) and Military University of Technology from Warsaw (Poland). The general aim of the conference is to gather the experts from NATO countries and to discuss the issues corresponding to the development of military communications and information systems.

Developing an effective military communications system is the inter-domain issue and demands good familiarity of designers with such areas as networking techniques, network management, network security and applications. Additionally, military systems should satisfy the requirements of mobility, robustness, security and fast deploying. To satisfy these expectations, a variety of different issues should be considered and this leads to co-operation of many teams working on specific topics, as mentioned.

For this special issue, we have selected papers with good technical content and which correspond to the main issues related to military communication system design.

As it was announced, the first 5 papers were presented during last year *NATO RCMCIS 2003*. Two first papers correspond to the important projects in NATO and the U.S. The first paper, by Christopher Echols, deals with the TACOMS Post 2000 architecture that is aimed at providing full interoperability between national networks. It worth to mention, that Poland actively participates in this project, and is represented by 2 organizations: Military Communication Institute from Zegrze and DGT S.A. from Gdańsk. The second paper, by Anh Ta and Scott Starsman, describes lessons learned from a U.S. Navy enterprise integration initiative called Web Enabled Navy (WEN). WEN was initiated in April 2001 with the focus of integrating navy resources and providing a single-point-of-access. The next 3 papers were prepared by the authors from different polish organizations.

The first one is addressed to the mathematical discrete problem connected with local combat operations, written by Tadeusz Nowicki. This paper represents an approach to modeling of the battlefield. The next paper, by Paweł Matusz *et al*., discusses networking techniques and seamless roaming between UMTS and IEEE 802.11 networks. Finally, a heuristic algorithm for a random generation of feedback functions for Boolean full-length shift register sequences was described by Izabela Janicka-Lipska and Janusz Stokłosa. This paper belongs to the cryptography area.

The list of selected papers from the *NATO RCMCIS 2004* opens the paper prepared by Douglas Wiemer and dedicated to technology solutions for coalition operations. The second paper, by Esra Çiftçibaşi Erkan and Şenol Uzun, describes the user services implemented in the modern TASMUS tactical system. Next, James Busch and Rita Russo propose a strategy of sharing tactical data in a network-enabled coalition. In the following paper, Pavel Eichelmann and Luděk Lukáš provide us with the introduction plan of the Network Centric Warfare concept to Czech Armed Forces. The next 3 papers correspond to different cryptography issues, which are Cryptology Laboratory, by Robert Wicik, IP-KRYPTO cipher machine for military use, by Mariusz Borowski and Grzegorz Łabuzek, and primality proving with Gauss and Jacobi sums, by Andrzej Chmielowiec. The problems of switch architectures are discussed by Grzegorz Danilewicz *et al*. Finally, we have two papers dedicated to other specific topics: Henrikas Pranevicius proposes an approach for integrated analysis of communication protocols by means of PLA formalism and Ferdinand Liedtke introduces the adaptive procedure for automatic modulation recognition.

Wojciech Burakowski (Military Communication Institute, Poland)
Luigi Bella (NATO NC3A Agency)
Guest Editors

Paper

Developing standards for interoperability of tactical communications systems

Christopher J. Echols

"The enemies of freedom have not stood still while free nations adapted to meet their threats and tactics".

Donald Rumsfield [1]
United States of America, Secretary of Defense

Abstract— The lack of interoperability in tactical communications systems has been a known fact in the North Atlantic Treaty Organization (NATO) since the end of the Cold War. This condition still exists today. The only alternative to obtain tactical communications systems with the interoperability necessitated by future operational requirements are through the establishment of NATO standards. This paper examines the tactical communications systems Post 2000 (TACOMS) project whose aim is to develop technical standards that will allow for achievement of interoperability between multinational tactical communications networks.

Keywords— TP2K, TACOMS Post 2000, NATO Tactical Communications Interoperability.

1. Introduction

Within NATO, interoperability is defined as: "The ability of Alliance Forces and when appropriate operate effectively together in the execution of assigned missions and tasks" [2]. As early as 1978, the Tri-Service Group on Area Communications (TSGCE) (SG/1) recognized the requirement for interoperable tactical communications systems. In 1985, the TSGCE formed Project Group 6 to establish standards for tactical communications in the land combat zone.

2. PG/6 Phase II Report

In 1986 NATO project Group Six, (PG/6) produced a report of their findings to the TSGCE. PG/6 divided its work into three phases. Phase I completed pre-feasibility studies and produced a report to the TSGCE documenting a recommended architectural framework. Phase II produced a second report documenting the refinement of the architectural framework upon which the draft STANAGs would be developed in Phase III [3]. The TACOMS Project Steering Group (PSG) was formed to conduct the work specified in Phase III. In April of 1998, the MOU was originally signed by 12 NATO Nations: Belgium (BE), Canada (CA), France (FR), Germany (GE), Italy (IT), Netherlands (NL), Norway (NO), Turkey (TU), Portugal (PO), Spain (SP), the United Kingdom (UK), and the United States of America (USA). Poland joined the TACOMS Project in 2004.

The TACOMS nations agreed to establish an International Project Office (IPO), and fund the project with the objective of developing draft standardization agreements or draft STANAGs to be implemented by NATO nations and Coalition Forces in both peacekeeping and in war fighting environments [4].

3. The IPO and contract team

The International Project Office (IPO) is located outside of Paris, France, within facilities maintained by the Délégation Général pour l'Armement (DGA). The IPO has overall programme and systems engineering management responsibility under the contract. Under the guidance of the 13-member nation PSG, the IPO serves as the focal point for any technical guidance and/or direction necessary during the conduct of the contract. The TACOMS contract was

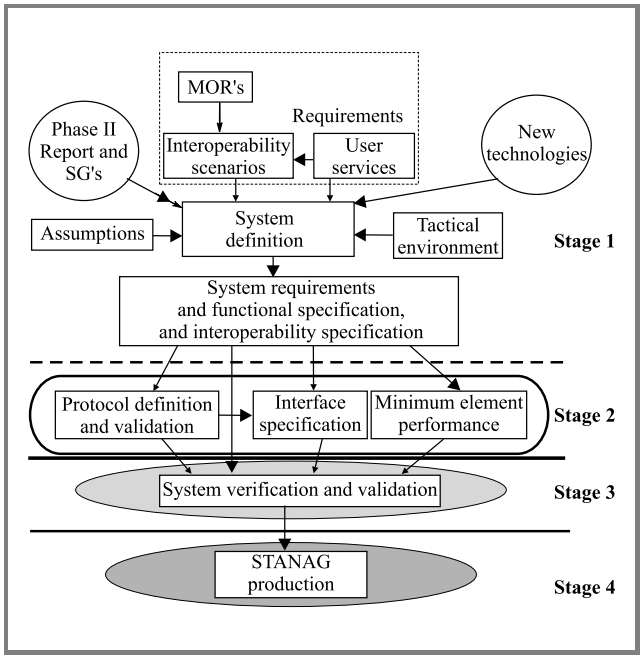


Fig. 1. TACOMS Post 2000 STANAG methodology of development.

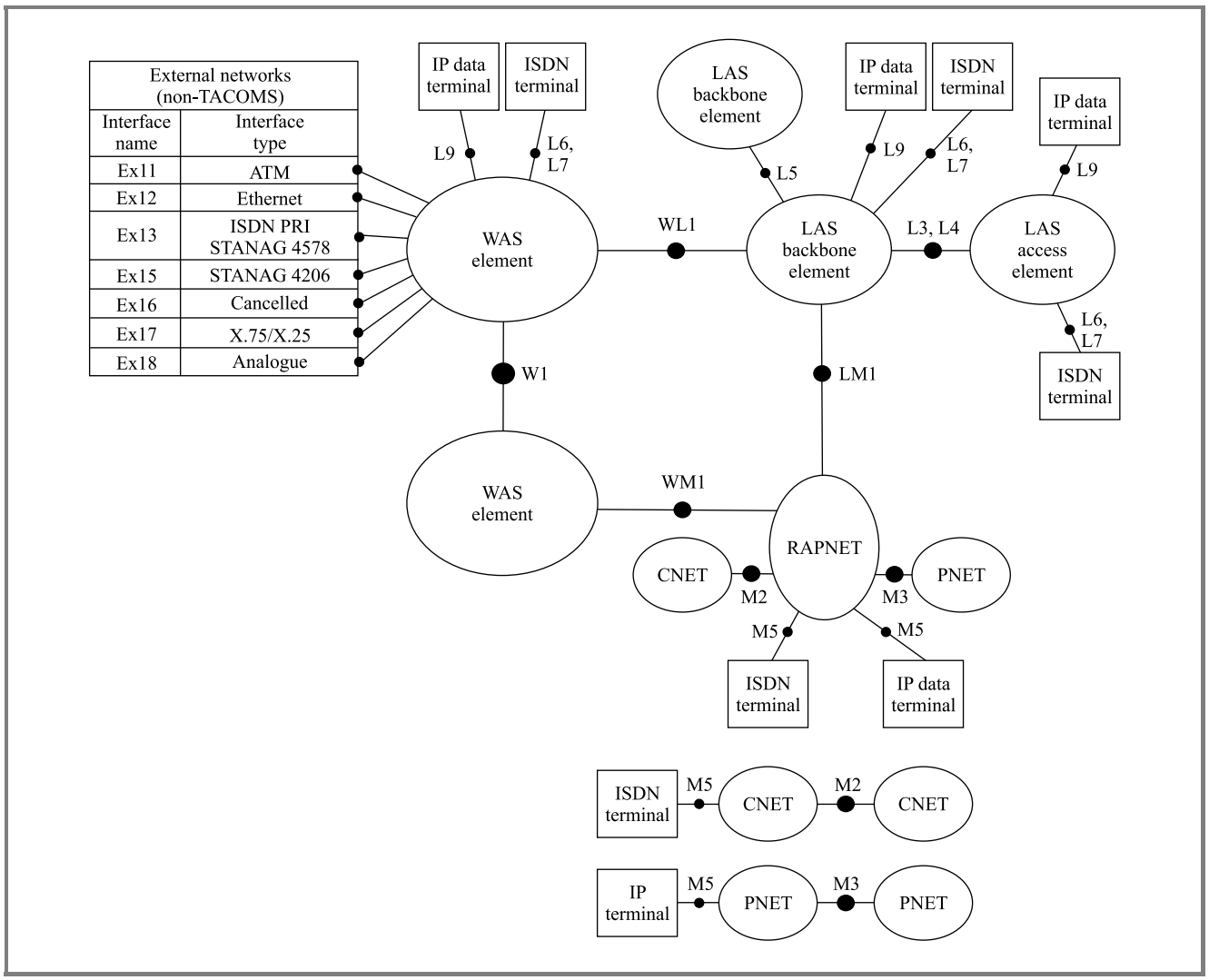


Fig. 2. TACOMS Post 2000 architecture.

awarded to the TAC ONE consortium in December of 2000. TAC ONE, is an international joint venture consortium with shareholders from five major defense telecommunications companies: International Telephone and Telegraph, British Aerospace Systems, Marconi-Selenia, EADS, and THALES-France. In addition, there are 13 sub contractors representing the national industry of the TACOMS MOU signatory nations. The initial contract period was for three years. However, it was extended until March 2005 after the reorientation of the TACOMS network was changed from technology specific, to be technology independent.

4. TACOMS STANAG development methodology

The TACOMS standards will be based to the maximum extent possible on existing civil standards in order to facilitate the use of civil technology, and on commercial off the shelf (COTS) equipment in the design of defense communications systems. The methodology being used in the development of the standards specified in the Phase II report is being conducted in four stages as illustrated in Fig. 1.

5. TACOMS Post 2000 architecture

The TACOMS Post 2000 architecture is based on the following four subsystems: Local Area Subsystem (LAS), Wide Area Subsystem (WAS), Mobile Subsystem (MS), and System Management and Control Subsystem (SMCS). These subsystems are described in Fig. 2.

- **Local Area Subsystem (LAS)**. The LAS provides the means for local communications services in the local area (Head Quarters) and provides its users the connectivity to the Wide Area and Mobile Subsystem as well as to other local area subsystems.

- **Wide Area Subsystem (WAS)**. The WAS provides the backbone links for communications for the longer distances. The WAS allows for transit and interconnection of other subsystems and external systems. Interfaces with civilian and military strategic networks are also available in the WAS. The WAS shall also include a minimal access for users.
- **Mobile Subsystem (MS)**. The MS consisting of various radios (MRR's) and connectivity to the WAS and LAS that provide users the capability to communicate reliable while stationary or mobile in a physically and electronically hostile environment. The radios themselves and the Over the Air (OTA) interoperability are not part of the TACOMS project.
- **System Management and Control Subsystem (SMCS)**. Each of the subsystems have inherent network management and control requirements. A management and control subsystem will implement the network management protocol.

6. Implementing interoperability

In the 2002 Prague Capabilities Commitment it states that individual NATO nations have now made firm commitments to improve capabilities in more than 400 specific areas. Included among them are command control and communications, and combat effectiveness. "The many reforms, initiatives and programs agreed in Prague are the beginning of a transformation process essential to guaranteeing the security of the territory, populations and forces of NATO members against all threats and challenges" [5]. In 2005, the draft STANAGs will be presented to the PSG then forwarded to the NATO Communications Network Sub Committee (CNSC) to begin the promulgation process. Depending on the time allowed for promulgation and ratification of the STANAGs, TACOMS is not expected to be ratified until the year 2008. The TACOMS STANAGs will be sufficiently detailed to allow national industries of NATO nations to produce their own compliant systems. The STANAGs will however not define the switching technologies to be used. That decision is based on national preferences. The STANAGs will define user services, network elements and their minimum required performance, interoperability points (IOPs), the directory, naming and addressing structure, etc. With this capability, each nation should then evolve their existing and soon-to-be-fielded systems towards the common standards, thus progressively enhancing interoperability in the tactical communications systems of NATO nations and coalition forces.

7. Conclusion

Achieving true interoperability will require full support from all nations to support the sharing of security information among coalition nations, and the development of national systems. In addition to contributing the funding for prototyping and testing of the draft STANAGs to validate the TACOMS standards. It is desired to have the TACOMS Interopability standards included in the NATO Response Forces (NRF) requirements, just like the Multilateral Interoperability Program (MIP) standards have been. The inclusion into the NRF requirements will help to ensure implementation by all nations, and provide multi national coalition forces the capability necessary to interoperate and communicate real-time without the interoperability problems they are faced with today. The PG/6 Phase II report helped to established the TACOMS project which can be used as a model for developing not only land but also air and maritime interoperability standards.

References

[1] Defense News, Jan. 2003.
[2] NATO Project Group SC-6/PG/ 6 Phase II Report, Executive Summary, issue 2, Sept. 1994.
[3] T. D. Johnson, Canadian Army, TACOMS Project, Draft Executive Summary, 2000.
[4] TACOMS Post 2000, Memorandium of Understanding, Apr. 1998.
[5] NATO After Prague, NATO Office of Information and Press, 2003, online: http://www.nato.int

Christopher J. Echols born in the United States. He holds a M.Sc. degree in project management, and a B.Sc. degree in computer science. He is a graduate of the U.S. National Defense University Chief Information Officers (CIO) Certification Course and the Information Resource Management College Advanced Management Program. Since October 2002, he has served as the Project Manager, Tactical Communications (TACOMS) Post 2000 Project, Paris, France. The TACOMS project is a 22 million dollar multi-national project to develop interoperable tactical communication standardization agreements (STANAGs) for use by NATO and its coalition forces. As the PM he is responsible for the overall project management of the International Project Office (IPO), located on Fort d'Issy, Paris.
e-mail: christopher.echols@tacoms.dga.defense.gouv.fr
TACOMS Post 2000, International Project Office
18 Rue du Docteur Zamenhoff
92131 Paris, France

Paper

Enterprise integration lessons learned

Anh D. Ta and Scott Starsman

Abstract—This document describes the lessons learned from a United States Navy enterprise integration initiative called Web Enabled Navy (WEN). WEN was initiated in April 2001 with the foci of integrating navy resources and providing a single-point-of-access, Web environment to all business and operational applications. The navy's applications operate within a complex network environment that spans commands afloat, ashore, and overseas. The challenges addressed are similar to those faced by large, multi-national corporations and include some unusual characteristics including islands of intermittently, bandwidth-limited, connected information consumers. Both technical and management lessons learned will be described and denoted as either prerequisite or success factors.

Keywords— web services, service oriented architecture, navy, military, web enablement, Task Force Web.

1. Web Enabled Navy initiative

As the United States Department of Defense (DoD) is going through its *Force Transformation* in response to the Quadrennial Defense Review, the United States Navy is also going through a fundamental *transformation* into an enterprise with integrated information and knowledge resources – *one integrated navy*. In April 2001, an initiative called Web Enabled Navy (WEN) was started to integrate navy resources and provide a single-point-of-access, Web environment to all business and operational applications. The tasking was in two parts: the first issued to the Navy's System Command's to convert applications to web services; the second to Task Force Web (TFWeb) to plan, design, and implement a strategy that would allow the navy to realize the WEN vision [1]. Additionally, TFWeb was tasked to provide vision and guidance to both the ashore and afloat system integration government agencies and contractors.

In striving for the goal of one integrated navy, the WEN initiative must deliver solutions addressing the following issues:

- heterogeneous IT environments,
- multiple network architectures,
- diverse user communities,
- technological volatilities,
- distributed authority,
- multiple parallel IT efforts,
- limited financial resources.

These issues will be elaborated further on.

Heterogeneous IT environments. At the start of WEN initiative, it was estimated that there were one hundred thousand applications across the U.S. Navy implemented in a variety of programming languages and technology products. The U.S. Navy IT environment is analogous to a multi-national commercial enterprise. For example, a typical ship might have thousands of applications, hundreds of databases, and vendor unique solutions that didn't support innovation, plus users with diverse needs ranging from administrative activities (e.g., writing a performance review) to mission-critical activities (e.g., to tracking potential threats). Furthermore, most applications are poorly integrated across functional boundaries. While ad hoc solutions to cross-application integration issues have been implemented in some instances, they tend to be slow, inefficient, and required a complex series of agreements between application owners to maintain. Furthermore, as the navy modernizes its fleets, there is a growing dependency on information technologies to monitor and operate the machinery that sustains the ships in their vital missions. Finally, significant authoritative data originates from ships, which needs to be accommodated by the heterogeneous environment. Extending the integration problem at the ship level to the fleet level and further scaling it to the navy enterprise level involving both the afloat and ashore communities describes the scope of the technical challenges faced by the Web Enabled Navy initiative – o*ne navy with connected communities*.

Multiple network architectures. The WEN initiative is required to leverage the existing U.S. Navy infrastructures. Currently, there are four major information technology infrastructures. In no particular order, they are the Navy-Marine Corps Intranet (NMCI) infrastructure for the Continental US (CONUS) shore community, the Base Level Information Infrastructure (BLII) for the Outside Continental US (OCONUS) shore community, the Marine Corps Enterprise Network (MCEN), and the Integrated Shipboard Network System (ISNS; commonly referred to as "IT-21") for the afloat community. Each of these infrastructures has a unique constituency, operational characteristics, and programmatic along with business support. At present, combat systems must retain their separate network architectures. The challenge is to integrate these disparate physical infrastructures into a single logical infrastructure – *one enterprise services layer*.

Diverse user communities. The targeted U.S. user communities for the WEN initiative include active-duty personnel, reserve personnel, family members, retirees and U.S. Navy contractors. Each user community has different

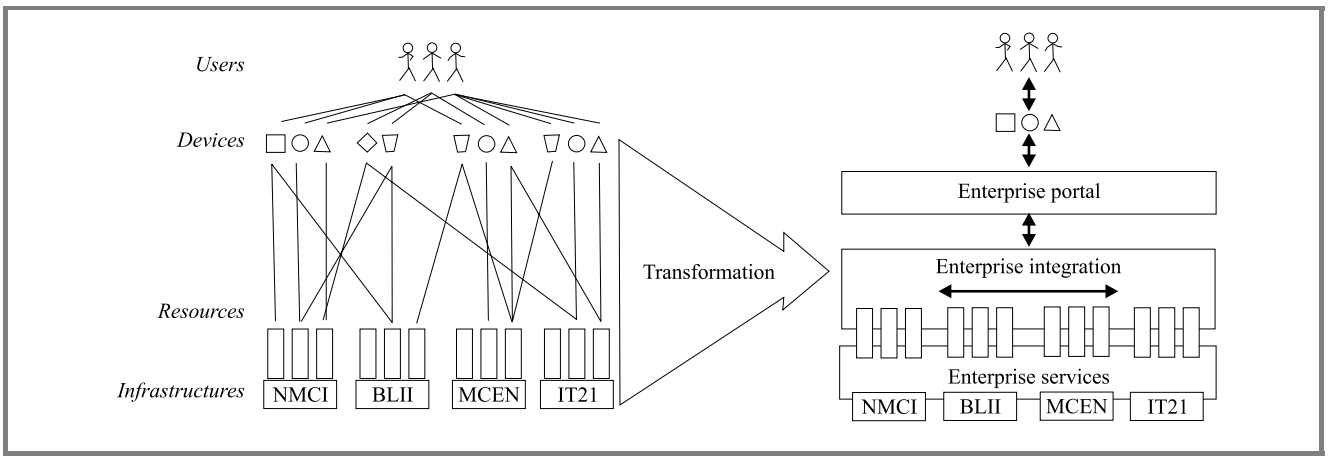


Fig. 1. WEN transformation vision.

technical and security requirements matrixed across the various network environments. Additionally, information is shared with key coalition partners during multi-national operations.

Technology and market volatilities. At the start of the WEN initiative, the major integration technologies available were middleware and enterprise portal products. While middleware technology has been available for many years, the industry sector lacks open interface standards resulting in numerous vendor-specific and proprietary solutions. As for the enterprise portal sector, the technology was just then emerging into its own identity. As characteristic of an emerging market place, the available portal engines had propriety interfaces and there were numerous mergers of enterprise portal vendors. For a large initiative such as the WEN, the risks associated with a single vendor solution was unacceptable from both technical and acquisition perspectives. Furthermore, a single vendor approach would limit the ability to share and innovate.

Limited financial resources. In addition to the technical integration challenges, a large organization such as the U.S. Navy has to seriously consider the financial effects of any initiative because of the vast number of applications in use across the U.S. Navy enterprise. Further complicating the issue, each application was at a different lifecycle stage, supported by varying logistics infrastructures, and with budgets already allocated to tasks spanning a multi-year timeframe. Because the allocated budgets did not contain resources explicitly designated to comply with the WEN requirements, program managers were required to identify existing application development and maintenance funding to support Web enablement.

2. Navy enterprise architecture

The Task Force Web team has developed an enterprise architecture and an integration strategy to realize the WEN vision (see Fig. 1). These products address the enterprise integration problem from two dimensions. The *application-to-user* delivery dimension focuses on methods for integrating content across the U.S. Navy into a single-point-of-access, Web environment. The *application-to-application* integration dimension focuses on the integration of U.S. Navy resources within an end-to-end business process perspective.

As an additional constraint on the TFW team, the resulting solution must comply with the architectural guiding principles:

- Maintain a flexible architecture that can accommodate both mission changes and technological changes.
- Implement an architecture consisting of *best-of-breed* components, where possible.
- Implement a standard-based, vendor-neutral architecture that minimizes ripple effects to back-end applications when architecture components are changed and allows innovation at the back end.
- Implement an architecture that minimizes information overload with automated information flow.
- Implement an architecture that enhances "shared awareness" through seamless integration of data sources.
- Implement a cost-effective architecture that maximizes acces to information and services.
- Implement an architecture that leverages XML-based industry standards to abstract from platform and programming language peculiarities.

The success of the WEN initiative depends on a balanced solution for both dimensions.

2.1. Prerequisites

A holistic approach was taken that addressed authoritative data sources and interoperable infrastructure issues.

Authoritative data sources. The U.S. Navy's approach to the establishment of authoritative data sources is through an organizational structure called the Functional Data Managers (FDMs). Each FDM is responsible for rationalizing the large number of databases into an orthogonal subset that more closely meets the navy's needs within a functional area. The expected benefits are:

1) the reduction of duplication among navy applications and databases,

2) the cost saving from items such as licensing costs, management of change overhead (such as revisions and updates), user training, etc.

A long-term goal is to align functional applications managed by the Functional Application Managers (FAM), an organization analogous to FDM, as a source for composite application frameworks (CAF) [2].

Interoperable infrastructures. TFWeb is leveraging the exiting infrastructures (i.e., NMCI, IT-21, BLII, and MCEN) to provide many basic support services required to support WEN functionality. However, some enterprise requirements are so critical that distributed development and/or deployment cannot meet target requirements. Two of these critical enterprise services are global directory services for identify management and enterprise data replication/synchronization providing ubiquitous data presentation.

The integration of the navy's directory services into a unified Navy Global Directory Service (NGDS) is a key enabler for many important capabilities such as personalization, enterprise single sign on, and enterprise rolebased access control. To date, there have been few standards that support cross-domain authentication. The Security Assertion Markup Language (SAML) standard promises to begin addressing this issue and is seen as a critical standard in building a true enterprise-wide single-sign on solution. These traditional infrastructure issues will have an unusually large effect on the user experience.

Currently, the data replication and synchronization strategy is predominantly being addressed through existing file-based replication solutions such as Collaboration at Sea (CAS). A more comprehensive solution providing true data synchronization between Relational Database Systems is planned but has yet to be implemented.

2.2. Vendor-neutral portal architecture

Given the challenges described and the recognition of the rapid rates of technological changes, Task Force Web focuses on bringing to bear multiple parallel efforts by government and vendors to create the desired environment. When possible we select best of breed recognizing that speed and facilitating vendor neutrality are key elements to be considered. The Navy Enterprise Portal (NEP) architecture has the following key characteristics:

- Presentation layer:
  - a standard-based, component based, modular enterprise portal architecture to provide U.S. Navy users with a single-point-of-access, Web environment that supports and integrates both the afloat and the ashore communities;
  - XML-centric design to support multiple display formats (e.g., HTML, Wireless Markup Language – WML, VoiceXML).
- Portal abstraction layer:
  - a portal connector component to provide an abstraction between the portal and content providers that minimizes ripple effects to content providers from replacement of the portal;
  - an external portlet registry component to provide a registry that further enhances independent from the portal.
- Business logic layer and data store layer:
  - XML-centric design to support capturing metadata and flexible customization (e.g., CSS, XSD, XSL);
  - an XML Web services integration strategy to provide a flexible environment for *composite application* leveraging resources across the heterogeneous IT environments.

The portal architecture focuses on standards-based interfaces, which will enable the U.S. Navy to develop its enterprise architecture and Web-enabled systems incrementally along with maintaining vendor neutrality. For areas where an industry standard is not available, TFW has developed an abstraction layer that provides vendor neutrality. For example, in the absence of a portlet API standard (e.g., JSR-168 and WRSP), TFWeb developed a "portal connector" component that provides an abstraction between the portal and content producers' applications. The portal connector satisfies several key requirements:

- It provides URL rewrite to ensure the seamless access of resources across the navy, where many of the resources are behind firewall with strict security filtering rules (e.g., IP filtering at the proxy server).
- It invokes an XSL engine to provide a consistent, server-based XSL transformation.
- It provides access to the portal context information such as user name and selected style template. This information is similar to the information defined in emerging portlet standards such as Web Services for Remote Portals (WSRP) and Java Portlet Specification JSR-168.

As the result of having a portal connector component, the navy enterprise architecture can integrate any portal with minimal ripple effects to content producers' applications and services.

2.3. Pragmatic integration strategy

Redesigning all U.S. Navy applications as Web applications is an imposing task in the near term given financial, operational, and organizational constraints. TFWeb encourages content producers to either develop their application from the beginning as Web application with Web services interface or to provide a Web services "wrapper" for key functionality of the existing application. This Web services-enablement of applications is referred to as a TFWeb *content integrated* application (see Fig. 2). Via the Web services interface, the returned data and metadata should be formatted in XML along with the associated eXtensible Stylesheet Language (XSL) stylesheet specifying the default display format (e.g., an HTML XSL stylesheet for desktop Web browsers or a WML XSL stylesheet for mobile Web browsers). The associated content provider can be referenced/accessed either through a URL over HTTP protocol (e.g., REST style Web services) or as Web services over SOAP/HTTP protocol. Display content is rendered by transforming the XML data using the provided XSL stylesheet.

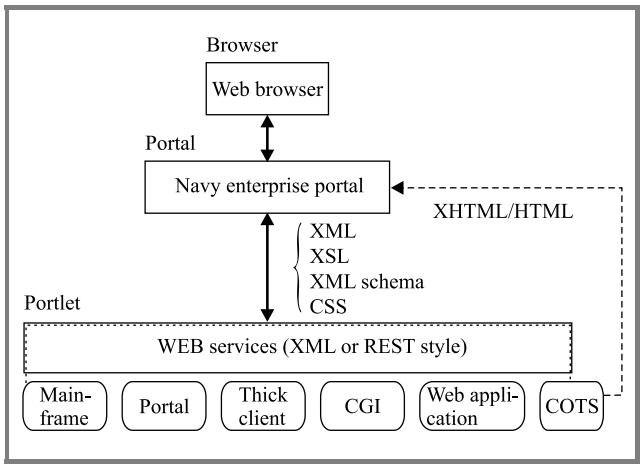


Fig. 2. Content integration requirements.

The expected benefit from TFWeb's content integration approach is a foundation for a Service-Oriented Architecture (SOA) with *composite applications* based on Web services. In addition, the approach focuses on an Enterprise Application Integration (EAI) strategy that allows leveraging of existing investments through existing Web standards.

For Commercial off the Shelf (COTS) products, the integration requirement is content integration. To support application-toapplication integration, the vendor shall provide supporting data showing how its Web services interface can be used to achieve the equivalent capabilities as the "out-of-the-box" HTML interface. The granularity of the Web services interface should correspond to the granularity of business logic codified in the product. If a COTS product does not have a Web services interface then an estimate of the level of effort required to develop such an interface (e.g., using a vendor toolkit) must be provided. "Out-of-the-box" HTML interfaces can be integrated into the NEP via URL references (i.e., *reference integration*). However, the HTML interface needs to be tested in the NEP because some HTML constructs that work in a stand alone Web browser mode (i.e., accessing the COTS' HTML interface directly) will not work within a portal environment (e.g., dynamically generated relative URL, frameset, etc.).

3. Lessons learned

3.1. Management

This section presents organizational and process lessons learned from the management perspective.

Identify and overcome social resistance. In general, management commitment on all IT initiatives is important. However, management commitment is essential for *enterprise-wide* IT initiatives. Voluntary compliance with strategic business changes will generally not produce the speed or depth or change required. Program managers have previously defined assumptions, program requirements, schedule, and budget. Furthermore, the Web services architecture fundamentally changes the business model within which program managers work. The reason many of our systems are built as vertical stovepipes is that it is easier to design, deploy, justify, and support a system that is entirely controlled by a single program manager. The complexity of operations and synchronization of business data are transferred to the operational user or decision maker. Interoperability woes are suffered by virtually every user of these systems. Web services technology breaks this stovepipe approach by forcing a separation of data, business logic, and presentation and the accompanying decomposition of customary monolithic applications.

The TFWeb team has commitment and support at two levels. Senior navy leadership consisting of flag officers and Senior Executive Service (SES) civilians have all been given and acknowledged the mandate from the Chief of Naval Operations (CNO) that all navy applications be Web-enabled. The other layer demonstrating support and commitment for the enterprise Web enabling effort is the layer of technicians, engineers, and programmers who implicitly understand the technical drivers for Web enablement. Indeed, this layer did not need to be "sold" on the value of this initiative as they've been battling with the consequences of stovepiped systems for years. The most challenging layer to convince has been the middle layer of program managers who have specific program and budget goals and have incentives to continue the existing system of funding and development.

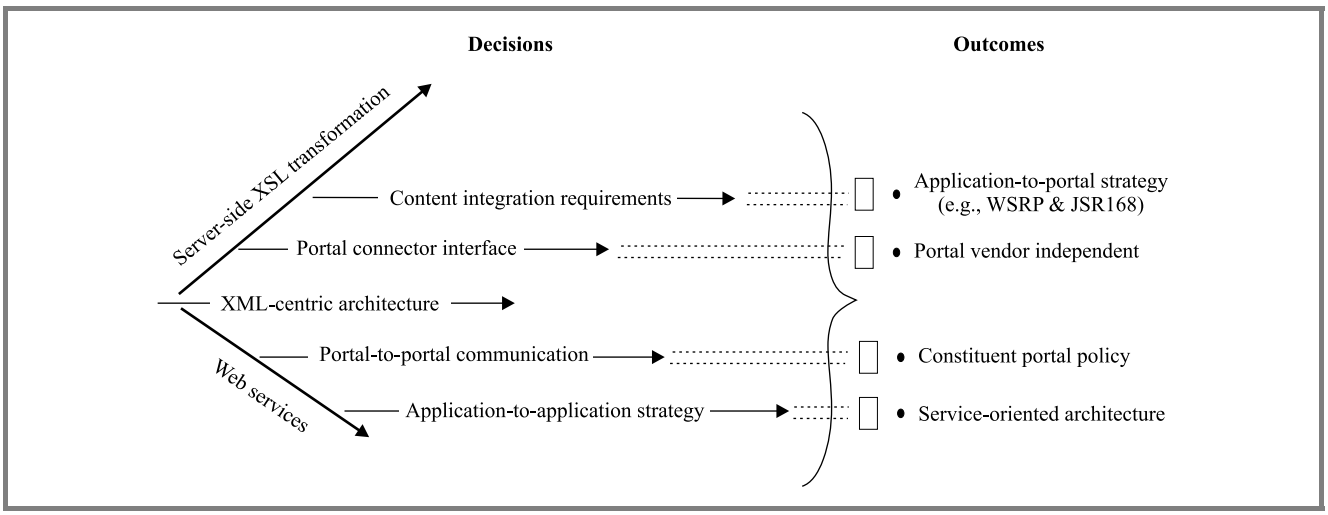


Fig. 3. Architectural decisions.

Simplify integration. A majority of the resistance to any enterprise-wide IT initiative is motivated by the perceived complexity and level of effort required to comply. This resistance is compounded if the developers and/or program managers perceive that the standards to which they must develop is in flux or is incomplete.

TFWeb began with a very aggressive schedule and released a pilot portal and development documentation within 180 days of its establishment. Unfortunately, this resulted in the external developers bearing the brunt of the learning experience as the Web-enabled environment took shape. By the time the NEP architecture had stabilized, many developers had become wary of the changing development standards and continued development at a much more cautious pace.

In order to overcome this problem, the architecture should be developed and deployed to a very limited audience consisting largely of developers eager to make the system work. Their feedback should be incorporated into design revisions. External developers should not be included until the design has stabilized to the point where backward compatibility is reasonably assured as the architecture moves forward. Future architecture releases should be publicly scheduled far in advance and adequate backward compatibility must be ensured over a reasonable period of time.

Once the design has stabilized, a set of small, completed examples should be developed and distributed. "Learning through examples" is a popular learning approach among software developers. In addition to distributing examples, the process of compliance should be simplified.

Identify and integrate key enterprise content/services. The first production version of the NEP was deployed as a piece of the infrastructure where content was to be provided by the existing application owners. Initial acceptance of the NEP instance would have been accelerated had it initially included a set of tools in common use across the enterprise. In this case, basic collaborative tools such as chat, message boards, content management, and white boarding are excellent examples of enterprise content that proved compelling to our user base.

The initial non-availability of these services arises from the fact that the navy enterprise does not acquire and deploy applications in an enterprise fashion. Instead, each navy program supporting a specific user community will purchase implementations of these functions. This has yielded a costly, duplicative, and noninteroperable suite of tools that were difficult to individually integrate into the NEP. Ultimately, TFWeb elected to use some of the broadest existing licenses to deploy an initial capability and address the need for enterprise procurement to fulfill longer-term requirements.

3.2. Technical

This section presents architecture and/or integration and content lessons learned from the technical perspective.

In reviewing the progress of TFWeb, there were several architectural decisions that were fundamental to the quality of the current WEN architectural. Figure 3 illustrates the important architectural decisions and their outcomes. First, the recognition of the importance of XML as an interface technology among the U.S. Navy's heterogeneous IT environments laid the foundation for adopting enterprise focused technology such as Web services. Second, the decision to adopt server-side XSL transformation led to the development of the portal connector component that provides an abstraction interface between the portal engine and the back-end applications. In addition, the server-side XSL transformation decision led to the emphasis on integrating content as XML formatted data. This emphasis is the motivation for the TFWeb's content integration requirements. The TFWeb's content integration requirements are the foundation of the application-to-application integration strategy allowing data to be accessed without requiring parsing through HTML or propriety formats. Overall,

the emphasis of Web services in the TFWeb's content integration requirements provide a simple yet powerful solution to issues such as portal-to-portal communication, where portlet content can be syndicated to other portals (e.g., syndicating portlets content from constituent portals to the enterprise portal). (*NOTE. It is important to note that while emerging portlet standards such as WSRP and JSR-168 provide a standard interface for integrating content into portal addressing TFWeb's application-toportal integration dimension, they do not solve the TFWeb's application-to-application integration dimension because their interface can only return data embedded in HTML fragments for display. Thus, if the content consumers want the raw data, an HTML parser is required. As for TFWeb's content integration approach, the data is always available in XML.*)

Develop architectural guiding principles. Architectural guiding principles convey the criteria, aligned with program vision or requirements, for selecting candidate solutions or technology products. This aspect is essential in an IT initiative that involves many organizations with often-conflicting objectives.

Develop a technology roadmap. In general, a technology roadmap document promotes an *active* technology management approach that helps organizations identify and track candidate technologies for adoption. Specifically, a technology roadmap is the best place to provide the rationale for selecting a candidate technology/product because it contains comparisons at both the technology and technology product levels (e.g., identifying dead-end technology). With a technology roadmap and its taxonomy of technology, organizations can quickly evaluate emerging technologies or available technology products. In addition, proof-of-concept prototypes should be performed to gain insights into potential implementation risks. A technology roadmap document should be developed as a companion document to an architecture document. While a technology roadmap captures the rationale for adopting technologies, an architecture document mainly focuses on describing the architecture with emphasis on describing the interactions and functionalities of architectural components.

Emphasize server-side implementation. In any enterprise-wide deployment, the mandate of one client device is not realistic. Hence, server-side implementations of candidate technology should be considered to facilitate the rapid technology insertion while avoiding software distribution problem such as Web browser upgrade. An example is the adoption of XSL technology. Through prototypes, the TFWeb team discovered that XSL client-side processing is only available starting with Microsoft Internet Explorer 5.5. In addition, there were several server-side XSL engines with different interfaces. To ensure consistent XSL engine interface and behavior, TFWeb team adopted a specific server-side XSL engine.

Emphasize portal abstraction. The enterprise portal market is still evolving. In fact, many of the risks that TFWeb team faced during the beginning of WEN initiative are still valid (i.e., market volatility). TFWeb developed two architectural components to provide an abstraction around the portal engine to minimize ripple effects when the portal engine is replaced. The first component is the portal connector. Its function (e.g., portlet context info) is similar to the set of interfaces being defined in emerging portlet standards such as WSRP and JSR-168. The other component is an external portlet registry. The original intention for the external portlet registry was primarily management of portlet information, where the portlet many not be Web services- based. However, as the NEP architecture evolved to adopt Web services the external portlet registry has evolved to become a Web services registry that complies with the Universal Description, Discovery and Integration (UDDI) standard.
In the striving for maintaining an abstraction layer between the portal and the content providers, NEP use of the portal has been limited to just presenting the portlets' content. A desirable goal is to access more of the portal's functionalities (e.g., virtual community, federated search) via a standard interface.

Develop a comprehensive content integration strategy. The major challenge in any integration effort is implementing a common protocol and data format for disparate systems to communicate. Figure 4 illustrates the design of a Web services complying with the TFWeb's content integration strategy. TFWeb's approach of using XML to format data along with metadata and document-centric Web services as the common protocol provide a flexible foundation for integration with some unexpected benefits:

- XML-enabled database. Most available database products provide support for formatting result set data in XML, thus, providing content producers an efficient method to produce XML-formatted data.

- Web services interoperability. Since TFWeb adopted an XML-centric design before adopting Web services, most of the content providers were already producing XML-formatted data. As a result, their migration to a Web services interface was easy. Specifically, the Web services method parameter for accessing the data can be string type because the data is already in XML. The use of string type eliminates a common Web services interoperability problem between different Web services implementation. In fact, it is consistent with the recommendations in WS-I basic profile [3].

- Multiple XML namespace data. One of the profound benefits of the TFWeb's content integration strategy is the ability to produce the same data for different XML namespaces. Although most databases can format the result set into XML, the XML tag names are often mapped to the database column names. One can build custom Web services serializer to format the XML tag but this effort is not trivial. A recom-

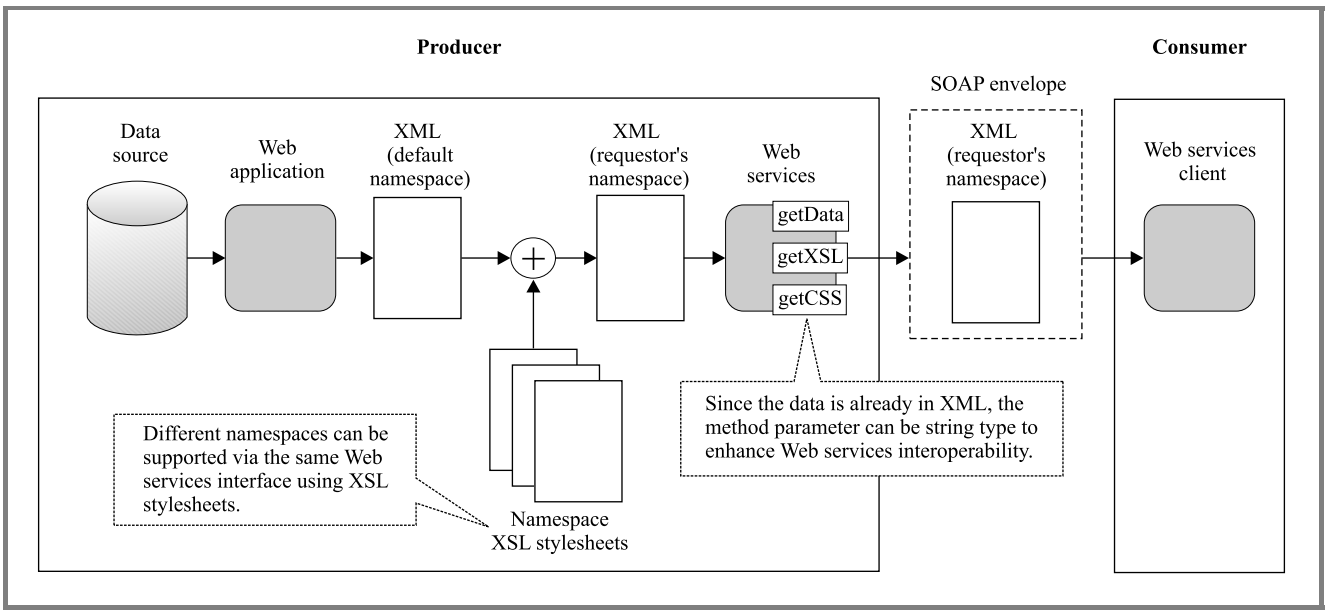


Fig. 4. Benefits of content integration requirements.

mended approach is to apply an XSL transformation to the generated data in order to obtain XML-formatted data complying with the requester's XML namespace.

- Support both integration dimensions. The approach of using Web services offers a solution that simultaneously addresses both the *application-to-portal* (i.e., vertical integration – integrated presentation) and *application-to-application* (i.e., horizontal integration – automated business process) integration dimensions. For example, portlet content in a constituent portal can be syndicated to an enterprise portal using Web services. As the result, the portlet content is available at both the enterprise and local levels. In addition, the presentation of the portlet content can be easily transformed to comply with the enterprise "look and feel" because the data is already in XML.

Emphasize XML-based formats for content. Given that the main objectives of the WEN initiative are to provide access to all resources and to support application-to-application integration, content formatted in using an XML standard (e.g., Rich Site Summary – RSS) is more efficient to process and repurpose for different presentation devices. In addition, XML-formatted data when use with compression produced bandwidth efficient content.

Emphasize bandwidth efficient content. Communication bandwidth is always a resource in demand and increasing bandwidth for the whole enterprise is often financially infeasible. Hence, content producers should be encouraged to review their implementations with an emphasis on reducing file and storage size through the use alternative formats. For example, the width and height attributes of an HTML $<$ IMG/ $>$ tag should not be used to present a small image of a picture. Instead, the HTML $<$ IMG/ $>$ tag should reference a file of an image that was reduced using an image editor. Beyond the scope of HTML, graphical content can be represented more efficient in terms of file size and display quality using new vector-based graphics standards such as Scalable Vector Graphics (SVG) and Extensible 3D (X3D) Graphics.

TFW has developed several proof-of-concepts demonstrating that using vector-based graphic formats with compression can dramatically reduce file size of graphical content while not compromising the content quality. Furthermore, it was discovered that the combination of JavaScript, XML-based formats (e.g., SVG, X3D), and Synchronized Multimedia Integration Language (SMIL) support offline content manipulation via the XML Document Object Model (DOM) interface and only require burst communication with the server.

4. Summary

Enterprise-wide IT initiative requires a flexible, innovative, holistic solution that addresses both the technical and management perspectives. Furthermore, the candidate solutions have to be evaluated in the larger enterprise context to avoid selecting local-optimized solutions. While the TFWeb team has learned many lessons and solved many problems, there are still some remaining challenges:

– enhance identify management with emphasis on single sign on capability;

– enhance portal personalization with an enterprise-wide replication and synchronization strategy;

– increase *core enterprise services* to reduce duplication in implementations among content providers;

– ensure common "portlet behavior" in addition to common "look and feel" at the enterprise presentation level;

– increase adoption and usage of metadata across the enterprise through the use of ontologies.

One remaining challenge for WEN is a typical struggle for IT initiatives; it is the struggle between developing a long-term, enterprise-wide solution and developing a solution that produces results immediately whether or not the solution is consistent with the long-term, enterprise-wide direction. From the beginning, TFWeb has chosen to focus on developing a long-term vision emphasizing flexibility and robustness of the architecture over portal content. While a vendor-specific stovepipe approach would have been much more rapidly fielded, TFWeb opted to deal with interoperability issues proactively rather than reactively. I hope you will remember Aesop's fable of the hare and tortoise when you consider the results of this effort. We may be viewed as the tortoise in some minds but our version of a tortoise is scalable and reliable while moving at the Internet speed.

References

[1] Task Force Web, Web Enabled Navy Architecture (Forward), Version 2.0, 31 Jan. 2001.

[2] Aberdeen Group, "Using Composite Applications to Lower Integration Costs", Executive White Paper, Apr. 2003.

[3] K. Ballinger *et al*., "Basic Profile", Version 1.0a, Web Services Interoperability Organization (WS-I), 8 August 2003, http://www.ws-i.org/Profiles/Basic/2003-08/BasicProfile-1.0a.htm (accessed 29 August 2003).

Anh D. Ta is a technical manager with consulting experiences with executive management, software system development, and enterprise integration initiatives. In addition, he has taught software engineering and computer science courses at several universities. Dr. Ta has a doctorate in Information Technology from George Mason University.
e-mail: ata@mitre.org
MITRE Corporation
McLean, Virginia, USA

Scott Starsman has served in a variety of operational and command and control positions in the U.S. Navy for 17 years. He has been responsible for the development, deployment, maintenance, and operation of satellite communication systems, advanced networking systems, and complex C2 systems. He has a doctorate in electrical engineering from Old Dominion University.
e-mail: scott.starsman@navy.mil
United States Navy
Norfolk, Virginia, USA

Paper

Modelling and simulation of combat operations in SimCombCalculator application

Tadeusz Nowicki

Abstract—The basis of the article is mathematical discrete problem connected with local military battle. The problem consists of two dependent problems. One of them shows how to find our troops allocation to enemy's troops. The battle damage assessment function for the problem is proposed. The second problem is connected with finding global strategy of attack. Local battle is described by modified Lanchester model. Two level logistic model is built for describing materials and munitions resupply management processes. This idea was used in computer simulator SimCombCalculator. The paper shows how to use this simulator for finding attack strategy from the platoon to brigade level.

Keywords*— *battlefield model, decision support, mathematical modelling, computer simulation.

1. Introduction

This article extends the previous research on mathematical models of military battlefield [1, 2, 4–6] and proposition of logistic support during the battle [3]. Many up-to-date mathematical and simulation models of military battlefield are dedicated to operation and strategy levels of decision processes. The idea of building battlefield simulator for lower levels of military structures is interesting from the practical point of view. Mathematical problems formulated for finding strategy of attack are usually complicated. Methods for solving these problems are complicated as well. They need high computer speed and memory. For lower levels of decision processes on the battlefield we have not so complicated problems. It comes from number of troops, not so large terrain, rather poor logistic processes, etc. Implication from these assumptions is chance to design computer simulator for commanders which will be used and perform on for example palmtop.

In the paper the problem of mathematical designing the attack against enemy in the local battle is shortly described. In order to prepare the best plan of attack in a short time there should be investigated special methodology. Mathematical or other formal models, optimization problems and methods for solving these problems are very important in that methodology.

In the model a battlefield area is divided into sectors small, quadratic and nearly homogeneous in the sense of size. Our site in the local battle has finite number of military units used in strategy of attack. Each military unit should be displaced to appointed target. The route of military unit to target is described by sequence of sectors.

Military units are described by many characteristics. We can divided them into subgroup: localization, military equipment, weapon, warfare, combat material, petrol and munitions, and so one. Each sector on the battlefield is characterized by enemy defense power in that sector. From this point of view routes of military units to targets are difficult for covering. This paper is prepared to show a few formal problems that should be solved during an attack designing:

- model of battlefield terrain reducing into sectors which depends on defense abilities;
- the problem of military units allocation to selected targets;
- the analysis of probabilistic characteristics of sectors crossing by military units;
- the problem of solving local battle;
- the problem of dynamic route modification;
- the problem of start and finish point and circumstances in simulation model.

It was built computer simulator SimCombCalculator (simulation of combat, calculator – SCC). Simulator is based on primitive database that contains information about both sides of conflict and terrain. Procedures connected with methods finding strategy of attack and user interface is made in Delphi environment.

2. The model of attack against enemy

We assume that strategy of attack made by single military unit against enemy is represented by i'ts target, located in particular sector, and sequence of many small and nearly homogeneous sectors as a way to this target.

Let

$$\overline{S} = \{1, 2, 3, \ldots, s, \ldots, S\} \tag{1}$$

is set of sector indexes (numbers). Information about sectors neighborhood is defined and described by matrix

$$\mathrm{D} = |d_{ij}|_{s \times s} \tag{2}$$

where $d_{ij} = 1$ when ith sector is in neighborhood (is so called "next") of jth sector and $d_{ij} = 0$ otherwise.

If we assume that number of our military units taking part in attack is N then: a_n is index of initial (starting) sector for nth military unit ($n = 1, \ldots, N$), b_n is index of final sector for nth military unit and there is sector in which target of nth military unit attack is ($n = 1, \ldots, N$), t_n is the moment of nth military unit start to mission ($n = 1, \ldots, N$), v_{ns} is mean time passed by nth military unit to cross the distance of sth sector area, ($n = 1, \ldots, N,\ s = 1, \ldots, S$).
So that

$$a = (a_1, a_2, \ldots, a_n, \ldots, a_N) \tag{3}$$

is vector of starting sectors for our military units,

$$b = (b_1, b_2, \ldots, b_n, \ldots, b_N) \tag{4}$$

is vector of final sectors for our military units (in which there are targets for military units),

$$t = (t_1, t_2, \ldots, t_n, \ldots, t_N) \tag{5}$$

is vector of starting moments of our military units.
We decide to make discrete intervals of time during an attack performance. So, let

$$T = \{1, 2, 3, \ldots, t, \ldots, T\} \tag{6}$$

is set of essential moments considered in our model.
The schedule of an attack strategy can be described by matrix

$$\mathrm{x} = \left[x_{nst}\right]_{N \times \overline{S} \times T}, \tag{7}$$

where $x_{nst} = 1$, when nth military unit is in sth sector while tth moment and $x_{nst} = 0$ otherwise.
In general point of view the schedule of an attack strategy described by matrix have to satisfy conditions given below:

$$x_{nst} \in \{0, 1\}, \quad n = \overline{1, N}, \quad t = \overline{1, T}, \quad s = \overline{1, S}, \tag{8}$$

in one moment each military unit should take place only in one sector

$$\sum_{s=1}^{S} x_{nst} = 1, \quad t = \overline{1, T}, \quad n = \overline{1, N}, \tag{9}$$

after obtaining final sector military unit don't move in our model

$$x_{nb_nt} \leq x_{nb_n(t+1)}, \quad t = \overline{1, T-1}, \quad n = \overline{1, N}, \tag{10}$$

before beginning the mission military unit is continuously in start sector

$$x_{na_nt} = 1, \quad t = \overline{1, t_n}, \quad n = \overline{1, N}, \tag{11}$$

military unit should cross borders only of next sectors

$$x_{ns_1t} + x_{ns_2(t+1)} - \mathrm{d}_{s_1s_2} \leq 1 \tag{12}$$

for $s_1, s_2 \in \overline{S},\ t = \overline{1, T-1},\ n = \overline{1, N}$, military unit cross sector not longer than v_{ns}

$$\sum_{t=1}^{T} x_{nst} \leq v_{ns}, \quad s = \overline{1, S}, \quad n = \overline{1, N} \tag{13}$$

military unit should cross sector continuously

$$\sum_{k=1}^{v_{ns}} x_{ns(t+k)} \geq v_{ns}\left(x_{ns(t+1)} - x_{nst}\right) \tag{14}$$

for $t = \overline{t_n, T-v_{ns}},\ s = \overline{1, S},\ n = \overline{1, N}$.

3. Probabilistic parameters of sectors

The probability $\zeta_s(m)$ of a single military unit down and out in the sth sector when m enemy's military units are in this sector must be determined. It is done in two steps. At the first step, the probability of single military unit down and out is considered separately. At the second step, the allocation of targets (may be enemy's units) in every sectors to our military units is considered.
Let assume, that the one type of enemy's military units are in the sector. Than the probability of down and out the single military unit is known from for example Lanchester model or each other. We can estimate the number of required enemy's units that are needed for destroying our single military unit. The rule of military unit destroying is determined in dependence on type of our military unit and enemy's unit type.
We can estimate probability $\zeta_s(m)$ even from models presented in [1] and [2] where many schemes like Bernoulli scheme, Poisson scheme are used or described on the basis of semi-regenerative process $\varphi(t) = \eta_{v(t)}\left(t - S_{N(t)}\right)$ $t \geq 0$, where $v(t)$ is semi-Markov process with state space $Y = \{0, 1, 2, \ldots, m\}$, which describes a process of number military units changing in the sector, $S_{N(t)}$ is the last moment of $v(t)$ changing and $N(t)$ is the number of the process state changing in the $[0,t)$ interval. The $\eta_i(t)$, $(i \in Y)$ processes are independent with states, denoting of military units numbers, which are ready to fight. The $\{\eta_i(t), \tau_i,\ i \in Y\}$ is a class of death processes $\eta_i(t)$ on the interval $0 \leq t < \tau_i$, particularly, there are homogenous Markov chains, which describe ways of regeneration and decreasing of units military means in dependence on semi-Markov process $v(t)$ state.

4. The Battle Damage Assessment Function (BDA-F)

Another problem is to determine the allocation targets to our military units. We suppose that two sides are on the battlefield. We consider situation in which weapon of our N military units are used to destroy M units of enemy's side. It is possible to accept such simplify conditions:

- weapons which belongs to decision maker are homogeneous in the sense of its destroying and additive potential;
- argument of BDA-F function is the destroying power potential which decision maker decide to attach in order to damage a specific target;
- BDA-F function of targets is increase function;
- zero value of destroying power potential causes zero-value of BDA-F function of target;
- BDA-F function accumulate target's destruction obtained during the local combat because it is a closed process;

– above a certain value of destroying power potential attached to a target it's BDA-F function is nearly equal to it's total value;

– we know the probability of destroying effectiveness; it depends on the number of elements of side A which try to destroy side B and depends on the number of element of side B being damaged;

– for every element we know that the value of BDA-F function indicate the level of element's losses adequate to a certain value of destroying power potential attached to this element;

– we can accept an assumption that each kind of destroying power potential can be represented by real number as multiplicity of a certain standard destroying power potential.

The last assumption comes form a real military method connected with a certain standard representation for every destroying power potential.
Assumptions listed above enable us to approximate BDA-F function of target, for the side A, by real function as follows [2, 5]:

$$h_n(y) = C_n\left(1 - \mathrm{e}^{-\alpha_n p_n y_n}\right), \quad n \in \overline{1, N}, \tag{15}$$

where: y_n – destroying power potential of our side allocated to the nth target of enemy's side, C_n – value of mth target, p_n – probability of successful attack, α_n – function which represents sensitiveness of nth target of side B to unitary destroying power potential allocated to it.

5. The military units allocation to targets

Let we assume that there are M enemy's objects as targets for our military units. For enemy's looses estimation we assume that we know global destroy potential of our units. We are looking for optimal allocation vector of destroy potential of our military units to targets

$$y = \left[y_m\right]_M, \tag{16}$$

where y_m is the global destroy potential of our military units attached to mth enemy's target in order to damage it. In the first step we assume that vector y has continuous variables as coordinates.
It is easy to show that the problem of finding the strategy of fighting is as follows: if K is the global destroying potential K of our military units we are looking for such $y^* \in Y$ which meets

$$\sum_{n=1}^{N} C_n \mathrm{e}^{-\alpha_n p_n y_n^*} = \min_{y \in Y} \sum_{n=1}^{N} C_n \mathrm{e}^{-\alpha_n p_n y_n}, \tag{17}$$

where

$$Y = \left\{ y \in R^N : \sum_{n=1}^{N} y_n \le K, y_n \ge 0, \ \ n = \overline{1, N} \right\}. \tag{18}$$

Lagrange function for this problem is as follows:

$$L(y, u) = \sum_{n=1}^{N} C_n \mathrm{e}^{-\alpha_n p_n y_n} + u_0 \left(\sum_{n=1}^{N} y_n - K \right) - \sum_{n=1}^{N} u_n y_n . \tag{19}$$

Kuhn-Tucker differential conditions for the problem above can be presented as below:

$$\nabla_{y_n} L(y, u) = -\alpha_n p_n C_n \mathrm{e}^{-\alpha_n p_n y_n} + u_0 - u_n = 0, \quad n = \overline{1, N}, \tag{20}$$

$$\nabla_{u_0} L(y, u) = \sum_{n=1}^{N} y_n - K \le 0, \tag{21}$$

$$\nabla_{u_n} L(y, u) = -y_n \le 0, \quad n = \overline{1, N}, \tag{22}$$

$$\left(\nabla_u L(y, u), u\right) = u_0 \left(\sum_{n=1}^{N} y_n - K \right) - \sum_{n=1}^{N} u_n y_n = 0, \ n = \overline{1, N}. \tag{23}$$

We can show [1] that the problem can be solved by very effective method that uses the final formula (it is effect of solving Kuhn-Tucker differential conditions) [2, 5]:

$$y_n^* = \frac{-1}{\alpha_n p_n} \ln \frac{\left(\mathrm{e}^{-K} \prod_{m=1}^{N} (\alpha_m p_m C_m) \right)^g}{\alpha_n p_n C_n}, \tag{24}$$

where

$$g = \frac{1}{\sum_{m=1}^{N} \frac{1}{\alpha_m p_m}}. \tag{25}$$

If all elements of vector y are nonnegative than this vector is optimal for our problem. If not, we put all negative elements of vector y into zero and for the other elements we use the formula given above.
Now we should allocate our military units to targets that give us similar destroying potential as we calculate from solution above. So, we modify vector y to discrete vector.

This provides us additionally to obtain values of vector b coordinates.
It is worth to remark that parameter p_n we can calculate as product of values $\zeta_s(m)$ from the shortest paths from starting sectors to target sectors. The problem of finding the shortest paths we can solve with one of the well-known methods in theory of graphs.

6. The criterion function for strategy of attack

We propose objective function for attack evaluation in the following form:

$$F(x) = \sum_{n=1}^{N} w_n \sum_{s=1}^{S} \sum_{t=2}^{T} x_{nst} \frac{1}{v_{ns}} \zeta_s \left(\sum_{k=1}^{N} x_{ks(t-1)} \right), \tag{26}$$

where

$$w_n \in [0, 1], \quad n = \overline{1, N}, \quad \sum_{n=1}^{N} w_n = 1, \tag{27}$$

are coefficients connected with values of targets which will be destroyed by particular military unit, $\zeta_s(m) : \{1, 2, 3, \ldots, N\} \to \Re$, is the function connected with probability of a single military unit (of m military units being in sth sector) success crossing.
We are looking for such x^*, which gives us

$$F(x^*) = \max_{x \in X} F(x). \tag{28}$$

7. Recurrent method for solving the problem

The recurrent algorithm for searching the best attack schedule is proposed. At first we should have the shortest routes for each target. Then we can solve the problem of military units allocation to targets. Having solution of allocation problem we are able to looking for first estimation of optimal attack schedule. We solve problem (28) with conditions (8)–(14) for decision matrix (7). It is easy to show that this problem can be solve by well known dynamic programming method, because of recurrent formula

$$x_{ns(t+1)} \frac{1}{v_{ns}} \zeta_s \left(\sum_{k=1}^{N} x_{kst} \right). \tag{29}$$

The solution obtained is input data for improving last military units allocation. After that, we can improve last attack schedule, etc. Finally, if we achieve attack schedule good enough, for global criterion, we finish our searching.

8. Simulator

Computer simulator SimCombCalculator has possibility to choose number and kind of troops. It was built special editor that help us to define every military units on battlefield. For defining concrete or only type of military troops for both sides we can use windows given in Fig. 1.

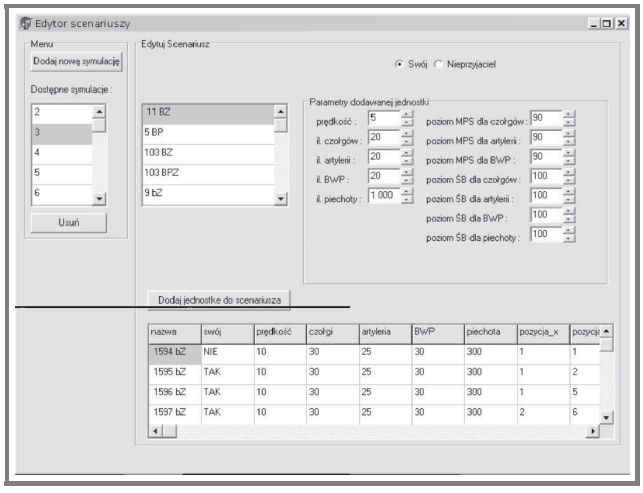


Fig. 1. Interface of SimCombCalculator scenario editor.

We can edit basic characteristics for every military units: number of unit, level of military structures, kind of unit, average speed, etc. (Fig. 2).

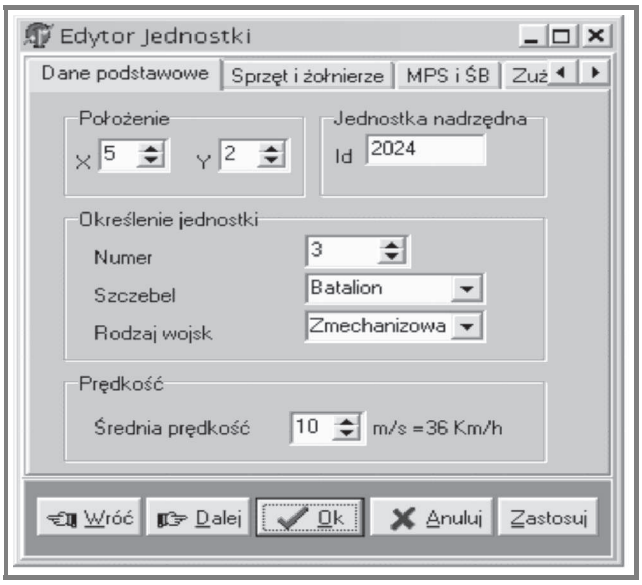


Fig. 2. Interface for setting basic characteristics of troops.

It possible to describe damage potential connected with weapons used by military units. Even level of materials can be defined for each military unit (Fig. 3).

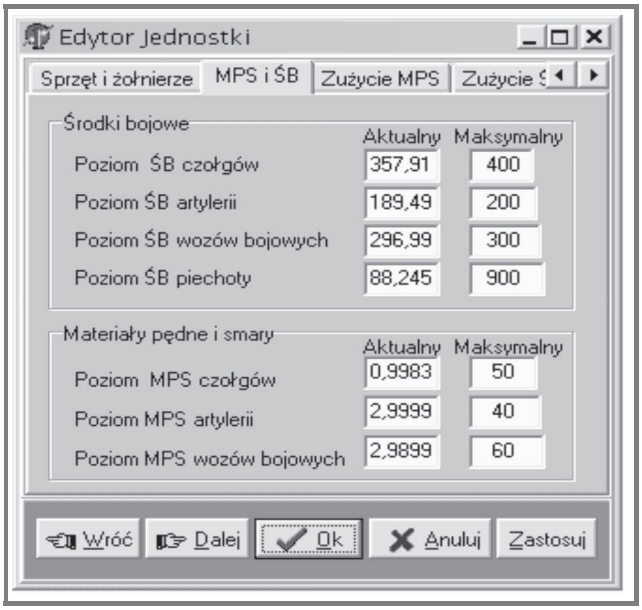


Fig. 3. Interface for any troop materials defining.

Number of tanks, artillery, infantry fighting vehicle, and people can be described (Fig. 4).
Attrition processes of petroleum oil and lubricants can be described (Fig. 5).
Attrition processes of munitions can be described as well (Fig. 6).
The palette in which troops can be located on the battlefield has following form (Fig. 7).

Palette is scalable so it is possible to make sectors more smaller (Fig. 8).
If digital map is necessary to use than simulator offer such possibility (Fig. 9).
If it is necessary basic information about troops can be displayed on windows (Fig. 10).
It is possible to put several parameters for simulation, for example: length of time quantum, frequency of steps, frequency of database recording, etc. (Fig. 11).
Results during the simulation processes are shown on the window (Fig. 12).
Many characteristics connected witch results of local battles can be displayed. We can see how many peoples, materials, and weapons has given military unit in selected moment in time (Fig. 13).

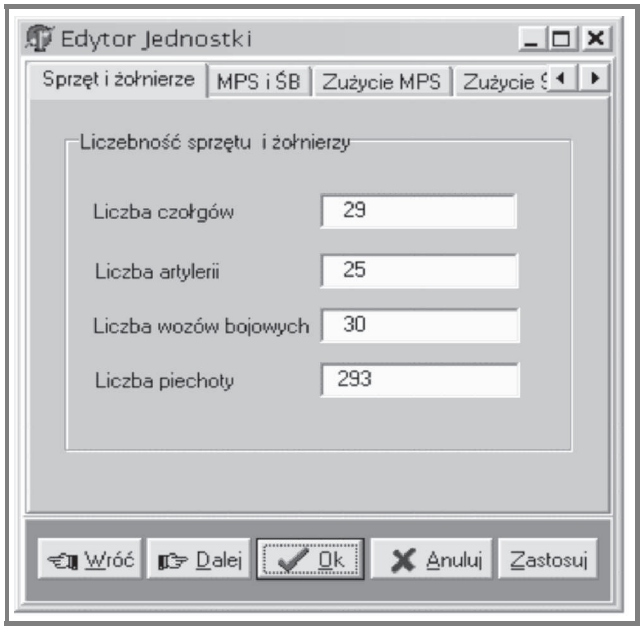


Fig. 4. Interface for any troop armament defining.

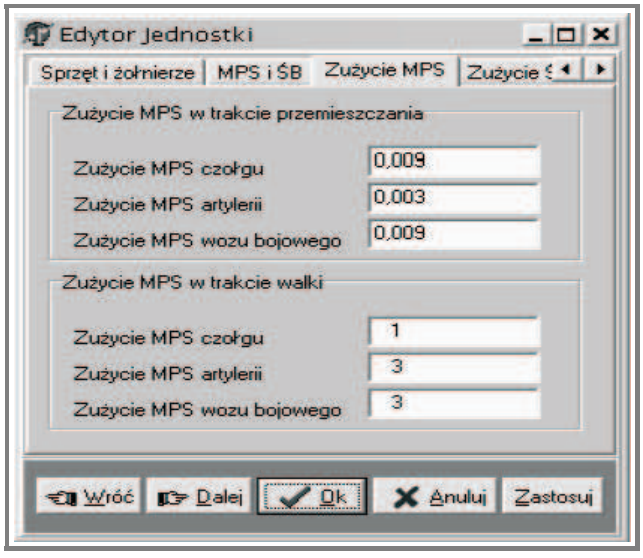


Fig. 5. Interface for main material attrition processes defining.

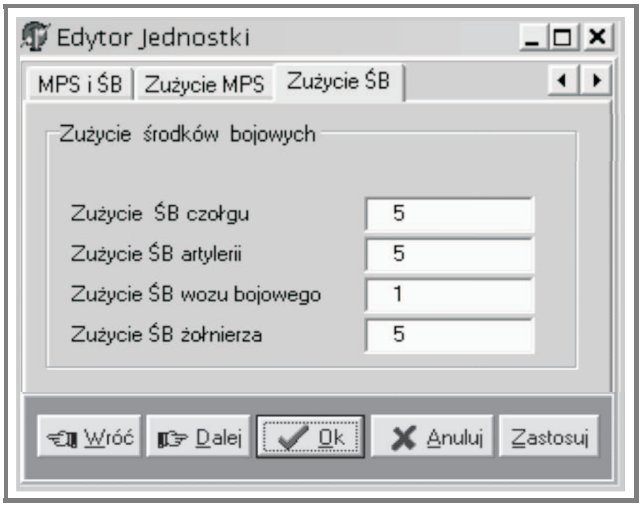


Fig. 6. Interface for people and munition attrition processes defining.

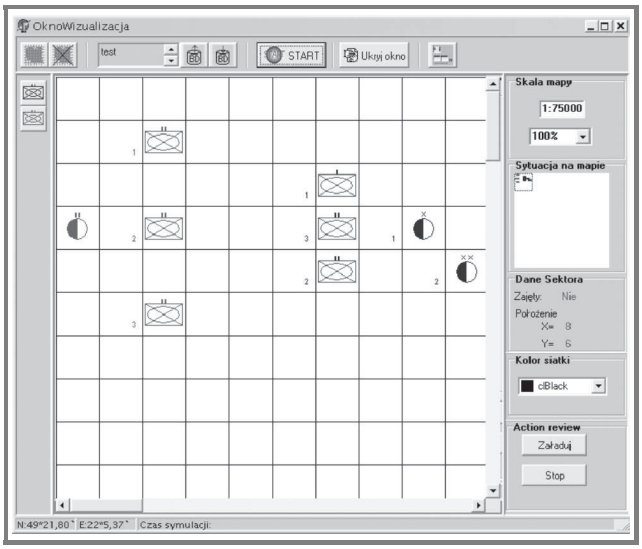


Fig. 7. Palette for locating military troops.

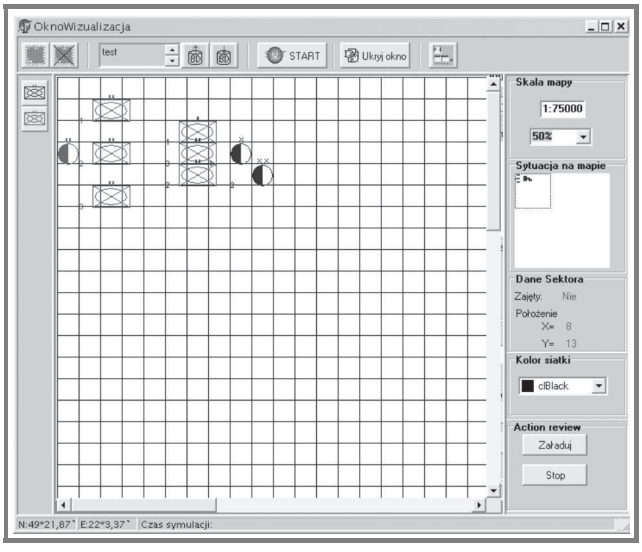


Fig. 8. Scalability of SimCombCalculator palette.

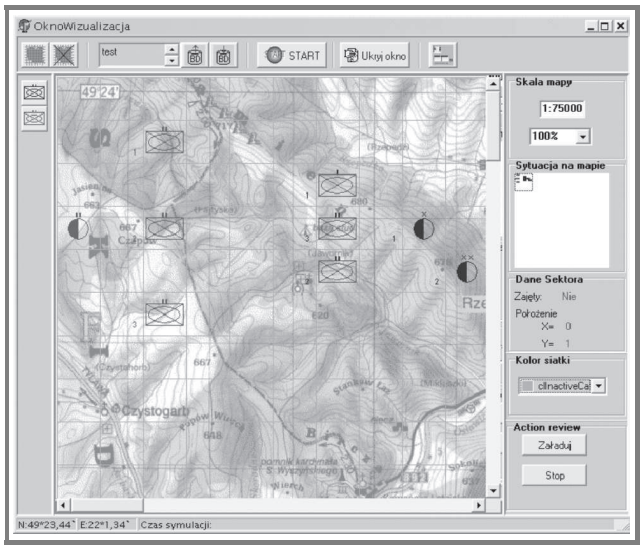


Fig. 9. Digital map as background of SimCombCalculator.

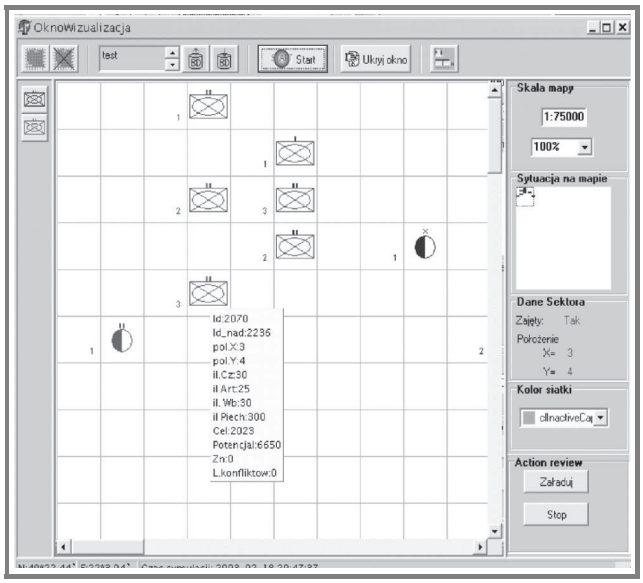


Fig. 10. Monitoring of information about individual troop.

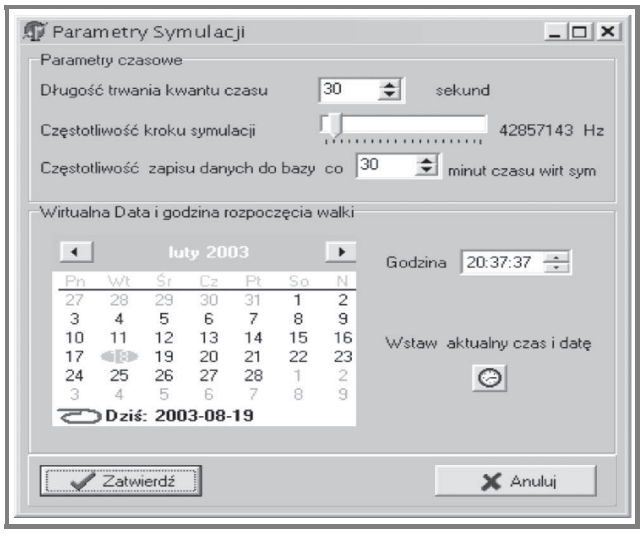


Fig. 11. Palette for simulation parameters setting.

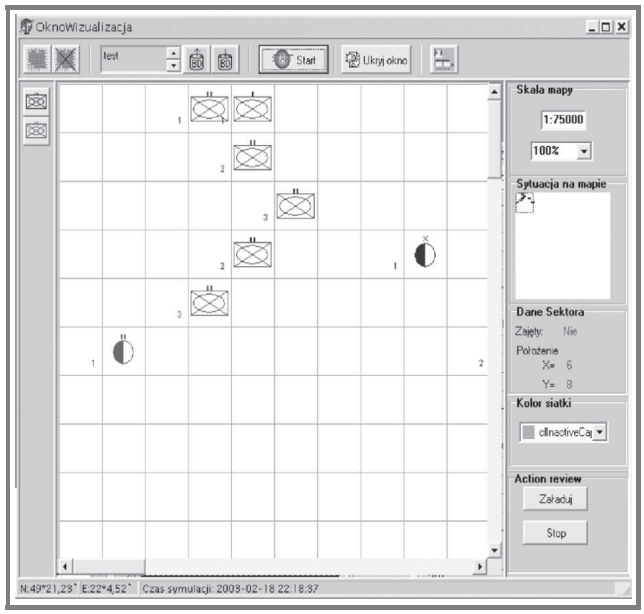


Fig. 12. The fight of troops.

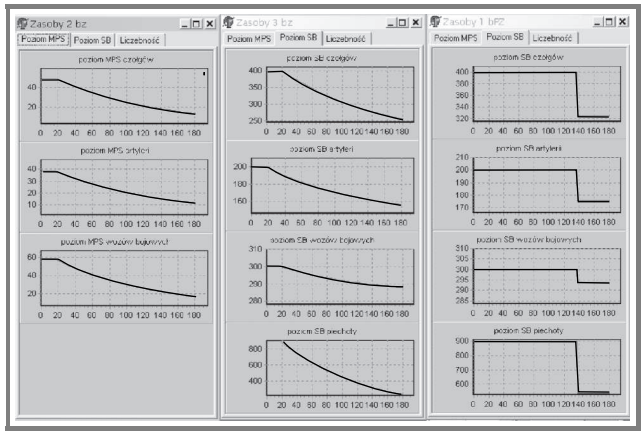


Fig. 13. Monitoring of troops material attrition.

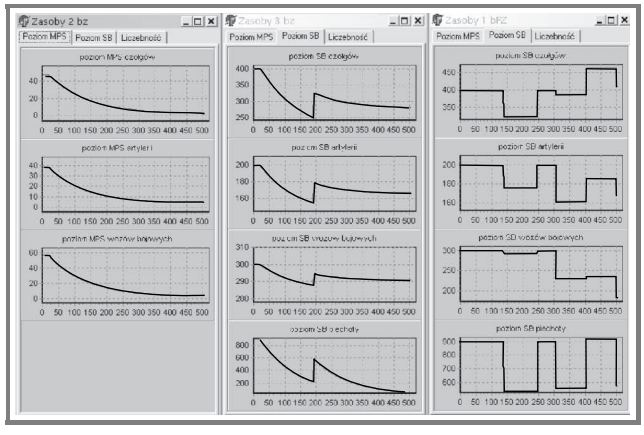


Fig. 14. Monitoring of material attrition in material bases.

Processes of resupply military units with people, materials and other things from company, battalion and brigade loading points can be shown during simulation (Fig. 14).

We can see elementary characteristics of military units during simulation (Fig. 15).

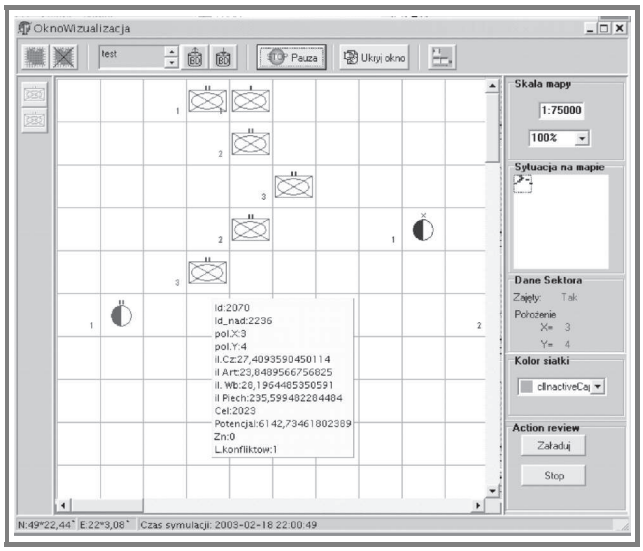


Fig. 15. Characteristics of military units during simulation.

It is possible to change every selected details during simulation.
The time to obtain the solution is good enough for it's practical use in simulation environment for battle description and analyse.

9. Conclusions

It is proven that simulation application for lower level of military processes on battlefield can be designed. Even complicated mathematical problems that are used in simulator procedures are perform in not too long time. Our application is rather prototype and professional equipment for lower level commanders should be built from the beginning. Thus the idea of building mathematical models of the battlefield and using computer environment for testing these models is important, necessary and possible at all.

References

[1] A. Najgebauer and T. Nowicki, "The methodology of modelling and interactive simulation of combat processes for CAX and DSS", in *Proc. SPIE, Int. Conf. Enabl. Technol. Simul. Sci. IV AeroSense 2000*, Orlando, USA, 2000.

[2] T. Nowicki, "Modelling and simulation of local battle under two-level control", in *RCMCIS'2001 Int. NATO Reg. Conf. Milit. Commun. Inform. Syst.*, Zegrze, Poland, 2001.

[3] T. Nowicki, "Optimal strategy of attack made on unreliable path network", in *RCMCIS'2002 Int. NATO Reg. Conf. Milit. Commun. Inform. Syst.*, Zegrze, Poland, 2002.

[4] T. Nowicki and A. Najgebauer, "Modelling and analysis of conflict situation in a distributed interactive simulation environment", in *Proc. Syst. Sci. Conf.*, Wrocław, Poland, 1998.

[5] T. Nowicki and A. Najgebauer, "The local combat model in a simulation environment", in *Proc. Syst. Sci. Conf.*, Wrocław, Poland, 1998.

[6] T. Nowicki, "Modelling and simulation of a local many-of-many combat situation", in *IASTED Int. Conf. Model. Simul. MS'99*, Philadelphia, USA, 1999.

Tadeusz Nowicki was graduated from the Cybernetics Faculty of Military University of Technology in 1980. He finished his Ph.D. thesis in 1984 and D.Sc. thesis in 1992. He is still working at the Military University of Technology. Now he is Associate Professor and Head of Computer Science Institute in this University as well. His main scientific interest is strongly connected with mathematical modelling, military battlefield modelling, theory of reliability, computer systems effectiveness and computer simulation.
e-mail: nowicki@isi.wat.waw.pl
Military University of Technology
Kaliskiego st 2
00-908 Warsaw, Poland

Paper

Seamless roaming between UMTS and IEEE 802.11 networks

Paweł Matusz, Przemysław Machań, and Józef Woźniak

Abstract—Mobile Internet access is currently available mainly using 2G/3G cellular telecommunication networks and wireless local area networks. WLANs are perceived as a local complement to slower, but widely available cellular networks, such as existing GSM/GPRS or future UMTS networks. To benefit from the advantages offered by both radio access networks, a mobile user should be able to seamlessly roam between them without the need to terminate already established Internet connections. The goal of this paper is to present an overview of the profitability of performing vertical handovers between UMTS and IEEE 802.11b using Mobile IP. Several simulations have been carried out using NS-2, which prove that handovers from IEEE 802.11b to UMTS can, under certain circumstances, be profitable not only when there is no more IEEE 802.11b coverage. Simulation results show that a mobile user should be able to roam between these networks depending on the current available channel bandwidth and quality, generated traffic type and number of users in both of them.

Keywords*— *UMTS, IEEE 802.11b, handover, roaming, Mobile IP.

1. Introduction

Both universal mobile telecommunication system (UMTS) and wireless local area network (WLAN) technologies enable fast Internet access. While UMTS is generally still in the phase of development with only a few existing installations, WLANs are already widely deployed. Devices supporting the IEEE 802.11a and IEEE 802.11b [10] standards are being manufactured by many companies and are widely available. Hot spots are installed at most large airports and in other public places such as hotels, train stations, and restaurants. Users of PDAs or notebooks with IEEE 802.11 network interface cards can easily access the Internet in such places, benefiting from the relatively high bandwidth of WLANs (11 Mbit/s in case of IEEE 802.11b and 54 Mbit/s in case of IEEE 802.11a).

On the other hand, the coverage offered by WLANs is quite limited. Those who need to have access to the Internet from almost everywhere must use a cellular network such as GSM/GPRS or UMTS. GPRS offers very low bit rates (theoretically up to 170 kbit/s, practically about 50 kbit/s [6]), which is often not satisfactory. 3G networks, such as UMTS, offer higher bit rates, theoretically up to 2 Mbit or even over 10 Mbit using HSDPA and 20 Mbit additionally using MIMO. Practically, for slow moving mobile users (pedestrians) the available bit rate should be about 384 kbit/s, although higher bit rates can be achieved depending on some conditions, such as good radio conditions or below-average cell load [9].

To enable switching between two different radio access networks, mobile users should have a terminal equipped with two network interface cards or a dual network interface card, e.g., supporting IEEE 802.11b and UMTS [1]. The terminal should be able to seamlessly switch (roam) between both networks without the user noticing it. Such a mechanism can be provided by Mobile IP [5]. In general, when available, the mobile terminal should connect to the Internet via WLAN to benefit from a higher bit rate and when it leaves the area covered by WLAN, it should automatically switch to UMTS. This is an obvious solution when the user roams between areas with WLAN coverage, while constantly being in the range of UMTS.

But, under certain circumstances, the efficiency of Internet access using WLAN could become much worse than using UMTS. Congestions may occur at the radio interface (multiple terminals trying to access the same access point at the same time), in the LAN connecting all APs with an Internet gateway or on the link connecting the gateway to the Internet (for example an often used 2 Mbit DSL connection). It would then be profitable for a terminal to switch from WLAN to UMTS, despite the still available WLAN coverage.

There has already been some research done in the field of using Mobile IP to switch between IEEE 802.11 and 2G/2.5G (GSM/GPRS) networks. Some of the conclusions, such as TCP-related issues, apply to UMTS [11]. But, mainly due to higher available data rates, lower packet delays and usage of WCDMA in UMTS [9] there are many issues that have never been discussed before.

In this paper handovers between UMTS and IEEE 802.11b using Mobile IP are analysed and discussed. The goal is to determine whether switching from IEEE 802.11b to UMTS can be profitable when there is both IEEE 802.11b and UMTS coverage. In the proposed scenario, a mobile user, equipped with a dual network interface card, accesses the Internet from a place with overlapping IEEE 802.11b and UMTS coverage. The mobile terminal may seamlessly switch between the two available radio access networks using Mobile IP. The profitability of performing such handovers is analysed, depending both on radio and network conditions – number of WLAN users, volume and type of traffic generated by those users, available UMTS channel bandwidth, and channel BLER. It is proven that, in case of UMTS, handovers between IEEE 802.11b and UMTS can be profitable.

The situation when the terminal leaves the range of the IEEE 802.11b access point is not analysed, because in such a case the only possibility is to switch to UMTS, regardless of available conditions. Such analysis has already been done for GPRS and in general applies to UMTS.

2. Mobile IP handover overview

Handover describes a mechanism when a user moves through the coverage of different wireless cells. A handover between wireless cells of the same type is referred to as horizontal handover, while a handover between cells of different type is known as vertical handover [4]. Because IP protocols were designed for stationary systems, some extensions have been proposed to introduce mobility support.

The main problem of a handover is that an IP address uniquely identifies both the end point and host locations. Because the mobile host can change its localization, there is a need to update the host's IP address and route packets to the mobile host's new subnetwork. Because of this, all active connections using the mobile host's previous IP address, e.g., TCP connections, would be broken.

There are some solutions to the mobility problem in IP networks, e.g., IETF Mobile (MIP) IPv4 [5], IETF Mobile IPv6 [7], Cellular IP [8] and HAWAII [13, 14]. Because of hierarchical network division into domains the mobility can be divided into Inter-domain mobility and Intra-domain mobility. Inter-domain mobility (also called Macro mobility) is related to a movement from one domain to another. A domain is defined as a large wireless network under a single authority. On the other hand Intra-domain mobility (also called Micro mobility) refers to user's movement within a particular domain.

Almost all solutions that address Micro mobility (e.g., Cellular IP and HAWAII) assume that Mobile IP is only used for Macro mobility. Because IPv6 is not often used in today's networks, Mobile IPv4 is perceived as a appropriate current solution. The protocol aims at continuous TCP connections even though the IP address changes when the handover occurs. The mobile host is assigned a Home Address that identifies the host in its home network. To solve the problem of IP addressing, Mobile IP introduces a temporary Care-of-Address (CoA) in a foreign network. Two new functions are added to the network infrastructure: a Home Agent (HA) and Foreign Agent (FA). After the mobile host moves to the new IP domain it obtains a Care-of-Address from a Foreign Agent (Foreign Agent Care-of-Address) or through some external means (Co-located Care-of-Address) such as DHCP. In the next step the mobile host registers the new address with its Home Agent. From now on, the Home Agent tunnels all packets for the mobile host through the Foreign Agent.

An important issue concerning handover performance is movement detection. Mobile IP supports three movement detection schemes: Lazy Cell Switching, Prefix Matching, and Eager Cell Switching [10]. In the Lazy Cell Switching scenario the mobile host waits until the lifetime of its registration expires and then tries to reregister again or to discover a new Foreign Agent to register with. If Agent Advertisement messages are not received, then the station attempts to solicit an advertisement using an Agent Solicitation message. In the Prefix Matching scheme the mobile host uses the "prefix extension" to determine whether a newly received Agent Advertisement is from the same subnet. If the prefix is different, the mobile host knows it is connected to a new subnet and registers. Eager Cell Switching is based on the mobile host receiving beacons from multiple FAs simultaneously. Once the current FA is no longer available (e.g., because the mobile has moved) then it selects a new one form this list.

There are additional movement issues concerning vertical handover. When the mobile is registered with the FA at the higher level (with higher cells) and moves into the cell coverage of the lower level (downwards handover) Mobile IP advertisements can be continuously received. In that case Eager Cell Switching cannot be used because the mobile is connected to the previous Foreign Agent.

3. Simulation setup

To simulate handovers between IEEE 802.11b and UMTS a detailed and realistic simulation environment was created using Network Simulator 2 (NS2) [12]. In addition to the already available components such as mobility management and IEEE 802.11 MAC, support for UMTS radio access has been implemented. This support takes into account all significant features of WCDMA and the UMTS radio protocol stack.

The simulated network architecture is presented in Fig. 1. It consists of two radio access networks, the UMTS and IEEE 802.11b access networks, connected via Internet.

The IEEE 802.11b radio access network consists of a single access point (AP), connected to a 100 Mbit Ethernet LAN, which is in turn connected to the Internet through a gateway, using a 2 Mbit/s link (e.g., DSL). Other APs can be connected to the same LAN (and the same gateway, which is not simulated), forming a wireless radio access network. Moreover, the particular number of mobile terminals are simulated, all associated with the same AP.

The UMTS architecture adheres to UMTS Release 4 specifications. The simulated UMTS terrestrial radio access network (UTRAN) consists of a single radio access network (RAN) controlled by a radio network controller (RNC). One Node-B (working in FDD mode) connected to the RNC via a 155 Mbit/s (STM-1) ATM link is simulated. A number of mobile terminals can be simulated, all located in the same cell and therefore using the same Node-B. The RNC is connected to a serving GPRS support node (SGSN) in the UMTS core network (CN) via a 655 Mbit/s (STM-4) ATM link. The SGSN is connected via a 655 Mbit/s ATM link to a gateway GPRS support node (GGSN) that connects the CN to the Internet via

a 2 Mbit/s link. Only a part of the packet switched (PS) domain is simulated. Other network elements in the CN (such as the whole circuit switched (CS) domain, HLR, VLR, etc.) are neglected, because they do not affect the simulation. AAL2 is used for transport between RNC and Node-B and AAL5 between RNC, SGSN and GGSN.

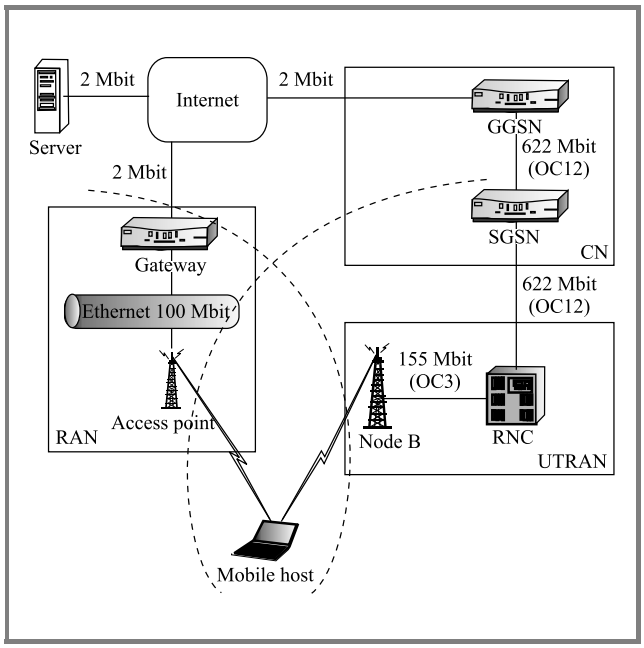


Fig. 1. Simulated network architecture.

Mobility management is performed by the Mobile IP mechanism, as described in [1, 2] and [5]. The Home Agent (HA) is located somewhere on the Internet (where the mobile host obtained its address) and Foreign Agents (FA) are located in the WLAN's gateway and in the SGSN. When a mobile terminal is using the FA, all traffic must first be sent from the server to the HA where it is encapsulated and then transferred to the FA, which performs decapsulation and routes it to the terminal. TCP acknowledgements generated by the terminal are routed directly to the server, without having to pass through the HA.

A scenario with overlapping UMTS and WLAN coverage, which is the most common scenario in urban environments, is considered. It is assumed that the mobile terminal is equipped with either two wireless interface cards, or a dual UMTS/IEEE 802.11b interface card. Both interfaces operate independently, i.e., UMTS and IEEE 802.11b connections can be active at the same time. This facilitates seamless handovers, because there is no time wasted for a new connection setup (assuming overlapping network coverage). During a handover from WLAN to UMTS, the terminal is authenticated and authorized in the UMTS network and a dedicated channel (DCH) is allocated while data transfer is performed by the IEEE 802.11b interface. If DCH with satisfactory QoS parameters (generally throughput and bit error rate) is allocated and the handover is assumed profitable, the actual handover takes place. During a handover from UMTS to WLAN, the terminal is first registered, authenticated and authorized by the AP in the WLAN while the data is still transferred using UMTS. After successful registration, the handover (when assumed profitable) takes place and UMTS channels are released. Such a scenario is simulated, because traffic delivery delays caused by the actual handovers are not additionally prolonged by authentication, authorization, and resource allocation mechanisms. User billing, authentication, radio network ownership (the same or different owners of both the UMTS and WLAN networks) and similar problems do not directly affect the simulation and are out of scope of this paper.

During simulations, the mobile user downloads a file from a FTP server located on the Internet. This implies the need for the best available bandwidth, which directly affects download time. The user's terminal can perform a handover either when it leaves or enters the area covered by WLAN range, or when a network congestion occurs in WLAN. The average delay between the WLAN gateway (or the GGSN) and the server have been set to 25 ms, which is the average value of inter-Europe packet delay in February 2003 according to Internet traffic measurements performed by Stanford University [15].

4. IEEE 802.11b and UMTS configurations

To carry out simulation experiments one AP and a number of mobile terminals (MT) were configured. Every MT was associated with the AP. The medium access control (MAC) sublayer operated in distributed coordination function (DCF) mode. In that mode the medium access algorithm is fully distributed and every MT uses carrier sense multiple access with collision avoidance (CSMA/CA) algorithm to access the shared medium. The DCF function is the basic and obligatory mode

Optional request-to-send and clear-to-send (RTS/CTS) function was used for all frame lengths, to test the worst case scenario by generating additional control traffic. RTS/CTS handshake also alleviates the hidden node problem, that is, when two or more MTs associated with the same AP cannot hear each other.

In the simulated scenario every mobile station set up an FTP connection with the server located outside the current subnetwork. The node utilizes a 2 Mbit/s Internet connection to reach the FTP server. The change of network conditions has been simulated by increasing the number of mobile stations. When a mobile station experiences the lack of sufficient network resources or estimates that the UMTS network can offer better resources, it can decide to switch to UMTS.

The constant, one-way processing delay introduced by all UMTS network elements in CN and UTRAN and by user equipment (UE, the mobile terminal) is estimated as 60 ms,

as specified in [3] and [9]. An additional delay is introduced by link buffering and signal propagation, but because of fast ATM links (STM-1 and STM-4) this delay is insignificant compared to the processing delay and delay introduced by the radio protocol stack, and the radio interface.

The configuration of the radio protocol stack in UTRAN partially determines the delays that occur on the radio interface between Node-B and UE. It is assumed that one dedicated transport channel (DCH) is allocated for the user in downlink and one in uplink to guarantee the required bandwidth. Such a guarantee cannot be made when using common or shared channels. In UMTS, dedicated channels with bandwidths up to about 2 Mbit/s can be allocated for a single user, but theoretically only in a fixed (indoors) environment. UMTS is required to support data rates of 144 kbit/s for mobile terminals moving with vehicular speeds and 384 kbit/s for terminals moving with pedestrian speeds (up to about 5 km/h). Because the simulation scenario may include user movement within the area covered by both networks, it is assumed that a 384 kbit/s DCH can be allocated most of the time. The available bit rate can change depending on the load of the cell and on radio conditions. This is why simulations have also been performed for downlink channels with bit rates lower than 384 kbit/s, although all have been done for a 32 kbit/s uplink channel. The uplink channel does not need to provide high bandwidth, because it conveys mainly TCP acknowledgements and RLC status messages.

Table 1
Radio protocol stack configuration

Parameters	Downlink	Uplink
Channel rate [kbit/s]	384	32
PDCP mode	No-header	
RLC mode	AM	
RLC block size [bits]	320	
Logical channel	DTCH	
Transport channel	DCH	
TTI [ms]	10	
Transport formats	0×320 1×320 2×320 3×320 4×320 8×320 12×320	0×320 1×320

Table 1 presents the radio protocol stack configuration used in simulations for both the uplink (32 kbit/s) and downlink (384 kbit/s) channels. Channels with other bandwidths differ only by the number of transport blocks defined in transport formats for those channels.

5. Simulation results

In UMTS, when using dedicated channels, the effective channel throughput depends only on the radio channel quality, described by the block error rate (BLER) parameter. This parameter represents the percentage of transport blocks which encounter bit errors on the radio link and therefore require retransmission. For BLER = 0% (no retransmissions) the effective channel utilization is about 95% because of the addition of UMTS radio protocol stack headers [1]. This figure can slightly change depending on the radio protocol stack configuration.

Figures 2, 3, and 4 depict results of simulations in a situation when the mobile host is connected to the Internet via UMTS. Average packet delay has been measured between the server and the mobile host as a function of packet length, allocated channel bandwidth, and channel BLER. Because the channel bandwidth, once assigned to a user, does not change, the actual packet delay is one of the variables that should be considered while making a handover decision.

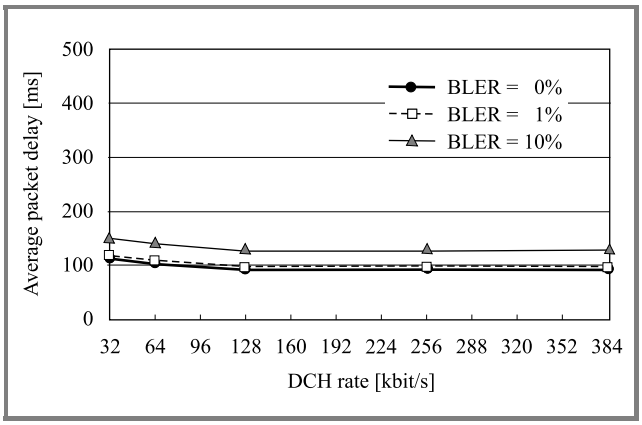


Fig. 2. 100-byte packet delay for UMTS.

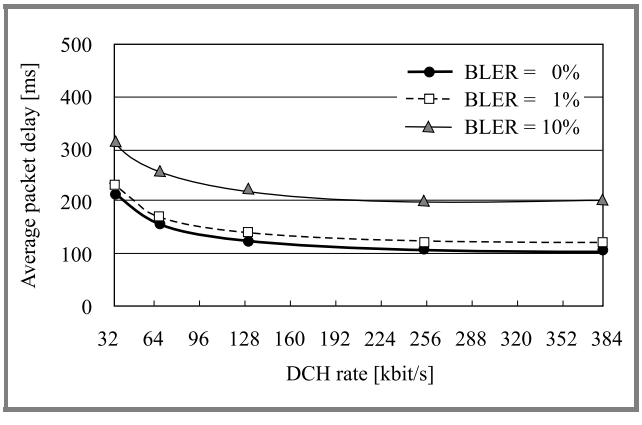


Fig. 3. 500-byte packet delay for UMTS.

As can be observed in Figs. 2, 3, and 4, the shorter the average packet size, the less the average delays that the packets encounter. This is caused by the fact that even for small channel bandwidths short packets can be sent

in just a few TTIs and do not have to be spanned over several TTIs, as larger packets do. Additionally, shorter packets fits in the smaller number of transport blocks, so the probability of packet retransmission (as described by [3]) is less than for large packets. Applications requiring small packet delays should use short packets (at least when using dedicated channels with small throughput) to minimize the delays.

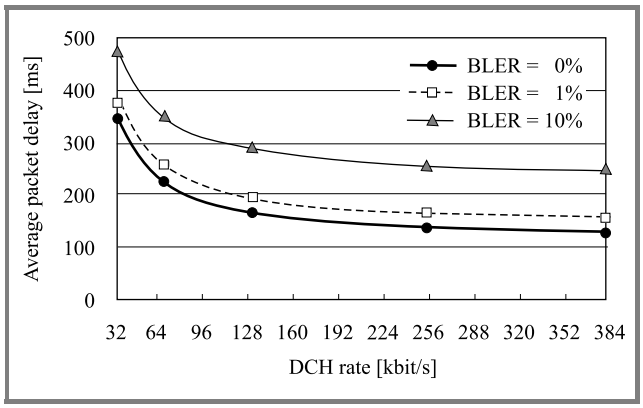


Fig. 4. 1000-byte packet delay for UMTS.

Throughput and packet delay experienced by a mobile station in WLAN are depicted in Figs. 5 and 6. According to the simulation scenario, in Fig. 1 the maximum throughput is limited by the bandwidth of the leased line connecting the WLAN to the Internet (2 Mbit/s). Generally, as the number of mobile hosts in the current subnetwork increases, the network conditions deteriorate. This is because of limited radio resources that must be shared by all stations.

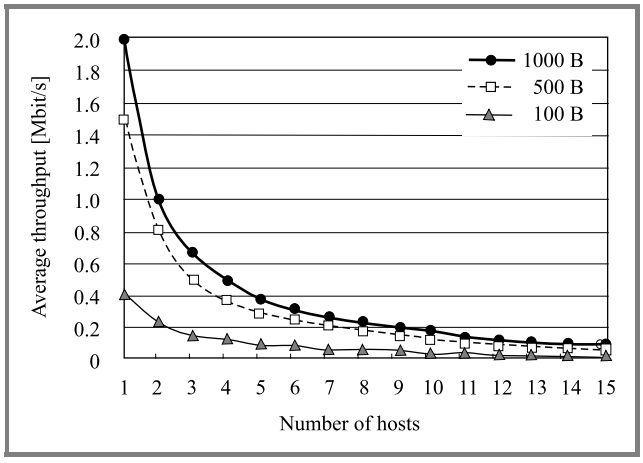


Fig. 5. Average throughput per station in WLAN.

Bandwidth utilized by the mobile host depends on the average size of transmitted packets. For shorter packets the MAC protocol overhead becomes substantial and average throughput deteriorates. This is mainly because of RTS/CTS handshake. When packets are short the data transmission time is small in comparison to the control frames exchange period.

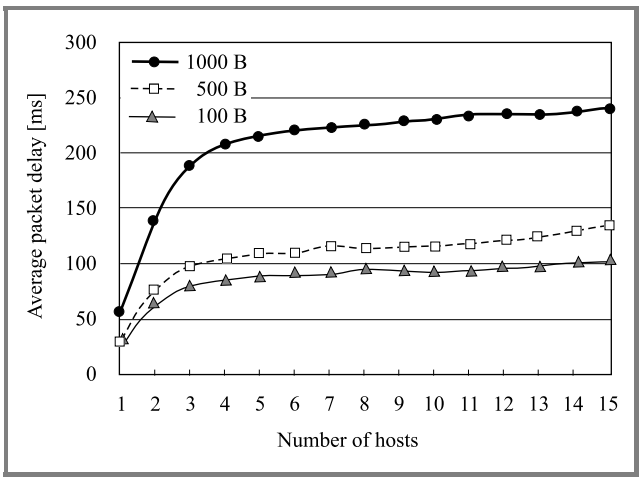


Fig. 6. Average packet delay in WLAN.

Average packet delay increases as network load increases and then saturates at a certain level. Packet delay is limited because the TCP protocol does not send more packets than can fit in the TCP window.

The handover itself does not cause any additional packet delay nor interrupts packet delivery – a route to the mobile host is always known by the Home Agent. This is because two independent network interface cards for UMTS and IEEE 802.11b are assumed to be used. Before performing the handover by the Mobile IP mechanism, the mobile host is already connected to both radio access networks and can access both Foreign Agents. These connections must be already established, because the mobile host has to have some knowledge about the available throughputs and delays in each network before making the handover decision. After the decision is made, the Mobile IPv4 handover mechanism [5] is invoked to switch controlling the Mobile Host from one Foreign Agent to another, in the other radio access network.

During the Mobile IP message exchange packets are sent continuously through one Foreign Agent and, after the Home Agent receives the new Registration Request, they immediately start being sent through the other Foreign Agent. The only delay in receiving packets may be caused by the difference in packet delays in both access networks and can be estimated at the time of making the handover decision (assuming that the average packet delays are known).

6. Conclusions

Mobile users can roam between UMTS and IEEE 802.11b not only when there is no WLAN coverage (which is the typical reason), but also when resources offered by UMTS are better than those offered by a reachable IEEE 802.11b network. It has been proven through simulations that, depending on experienced network conditions, handovers

between IEEE 802.11b and UMTS can be profitable. A mobile user equipped with two network interface cards or a dual network interface card able to manage UMTS and IEEE 802.11b radio connections independently can benefit from the possibility of seamless roaming between the two available radio access networks. UMTS, unlike 2G systems, offers satisfactory channel throughput and QoS for most applications. Depending on QoS requirements of the generated traffic, a mobile user can choose to switch to a radio access network that offers the most suitable QoS conditions, e.g., guaranteed throughput or average packet delay.

Currently, work is being done to specify an optimal criterion that the mobile host can use to switch between available access networks. It should take into account network conditions, which can be hard to accurately measure or estimate. Simulation results presented in this paper may help by providing some reference values.

References

[1] 3GPP TR 22.934, "Network (WLAN) interworking", V6.0.0, 09.2002.

[2] 3GPP TR 23.923, "Combined GSM and Mobile IP Mobility Handling in UMTS IP CN", V.3.0.0, 05.2000.

[3] 3GPP TR 25.853, "Delay Budget within the Access Stratum", V4.0.0, 03.2001.

[4] A. Fenstag, H. Karl, and G. Schäfer, "Current developments and trends in handover design for ALL-IP wireless networks", Internal Technical Report, Technical University Berlin, Berlin, 08/18/2000, Version 1.3.

[5] C. Perkins, "IP Mobility Support for IPv4", RFC 3344, 08.2002.

[6] C. Smith and D. Collins, *3G Wireless Networks*. New York: McGraw-Hill Telecom, 2002.

[7] D. Johnson, C. Perkins, and J. Arkko, "Mobility Support in IPv6", IETF Internet Draft, 03.2003.

[8] J. Gomez, C-Y. Wan, S. Kim, Z. Turanyi, and A. Valko, "Cellular IP", Internet Draft, 2000, draft-ietf-mobileip-cellularip-00.txt

[9] H. Holma and A. Toskala, *WCDMA for UMTS*. Chichester: Wiley, 2002.

[10] "IEEE Standard for Information Technology – LAN/MAN – Specific requirements – Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: High-Speed Physical Layer Extension in the 2.4 GHz Band", IEEE 802.11b: Supplement to International Standard ISO/IEC 8802-11: 1999(E) ANSI/IEEE Std 802.11, 1999.

[11] L. Morand and S. Tessier, "Global mobility approach with mobile IP in all IP networks", *IEEE Pers. Commun.*, Febr. 2002.

[12] "The network sumulator – ns-2", http://www.isi.edu/nsnam/ns/

[13] R. Ramjee and T. La Porta, "Paging support for IP mobility", Internet Draft, July 2000, draft-ietf-mobileip-paging-hawaii-01.txt

[14] R. Ramjee, T. La Porta, S. Thuel, K. Vardhan, and L. Salgarelli, "IP micro-mobility support using HAWAII", Internet Draft, July 2000, draft-ietf-mobileip-hawaii-01.txt

[15] "SLAC's PingER history tables ", http://www.slac.stanford.edu/cgi-wrap/table.pl

Paweł Matusz received his M.Sc. degree in computer science from Gdańsk University of Technology (GUT), Poland, in 1999. Since then he has been working at Intel as a researcher and tester, mainly in the field of 3G systems and wireless networks. In 2002 he started Ph.D. studies at Gdańsk University of Technology, Department of Information Systems. His research and scientific interests focus on performance analysis, optimization and interoperability of high speed wireless and cellular networks, mainly UMTS and IEEE 802.16.

e-mail: pmatusz@eti.pg.gda.pl
Faculty of Electronics, Telecommunications
and Informatics
Gdańsk University of Technology
G. Narutowicza st 11/12
80-952 Gdańsk, Poland

Przemysław Machań received M.Sc. in computer science (2001) and M.Sc. in information management (2003) degrees from Gdańsk University of Technology (GUT), Poland. Currently, he is studying towards Ph.D. at GUT. His research work includes IP and WLAN mobility, QoS in WLANs and WLAN architectures. He is also working as a software engineer in Intel Corporation, at R&D networking site located in Gdańsk.

e-mail: przemac@thenut.eti.pg.gda.pl
Faculty of Electronics, Telecommunications
and Informatics
Gdańsk University of Technology
G. Narutowicza st 11/12
80-952 Gdańsk, Poland

Józef Woźniak received his M.Sc., Ph.D. and D.Sc. degrees in telecommunications from the Faculty of Electronics, Gdańsk University of Technology (GUT), Poland, in 1971, 1976 and 1991, respectively. In 2001 he became a Professor. Prof. J. Woźniak has rich industrial and scientific experience. In February 1984 he participated in research work at the Vrije Universiteit, Brussel. From December 1986 to March 1987 he was a Visiting Scientists at the Dipartimento di Elettronica, Politecnico di Milano. In both cases he was working on modelling

and performance analysis of packet radio networks. In 1988/89 he was a Visiting Professor at the Aalborg University Center, lecturing on computer networks and communication protocols. Prof. Józef Woźniak is author or co-author of more than 170 scientific papers and co-author of four books. He is also co-editor of 4 conference proceedings and co-author of 4 student textbooks and great number of unpublished scientific reports. His scientific and research interests include network architectures, analysis of communication systems, network security problems, mobility management in WATM as well as LAN and MAN operational schemes together with VLANs analysis.

e-mail: jowoz@pg.gda.pl
Faculty of Electronics, Telecommunications
and Informatics
Gdańsk University of Technology
G. Narutowicza st 11/12
80-952 Gdańsk, Poland

Paper

Boolean feedback functions for full-length nonlinear shift registers

Izabela Janicka-Lipska and Janusz Stokłosa

Abstract—In the paper a heuristic algorithm for a random generation of feedback functions for Boolean full-length shift register sequences is presented. With the help of the algorithm one can generate *n*-stage Boolean full-length shift register sequences for (potentially) arbitrary $n \geq 6$. Some properties of the generated feedback functions are presented.

***Keywords*— cryptography, shift registers, Boolean functions.**

1. Introduction

Nonlinear shift registers generating full-length sequences, also referred to as de Bruijn sequences, have many applications in modern communications systems, especially in cryptography as components of complex devices and algorithms in cipherment and decipherment processes. There exists a number of methods for the generation of full-length sequences (*cf.* [1, 2, 6]).

In the paper we present some results of experiments done on nonlinear Boolean functions. The functions used as feedback functions of shift registers give full-length sequences generated by these shift registers. We generated all functions for *n*-stage shift registers, for $n = 3$, 4, 5 and 6. The experiments led us to the heuristic algorithm for generating *n*-stage full-length shift registers, where the number n of stages is sufficiently great.

2. Preliminaries

Let Z_2^n be n-dimensional vector space over the finite field GF(2). An n-argument Boolean function is a mapping $f : Z_2^n \to Z_2$. Let A_n be the set of all n-argument affine Boolean functions. Every Boolean function which is not affine is said to be nonlinear.

Let $f(x_{n-1}, x_{n-2}, \ldots, x_0) = y_z$ and let z be the decimal equivalent of the function's argument, i.e., such a positive integer that for each argument $(x_{n-1}, x_{n-2}, \ldots, x_0)$ we have

$$z = x_{n-1} \cdot 2^{n-1} + x_{n-2} \cdot 2^{n-2} + \ldots + x_0 \cdot 2^0 .$$

Then $[y_{2^n-1}, \ldots, y_1, y_0]$ is called the truth table of f. The value y_0 is the least significant value in the truth table and y_{2^n-1} is the most significant value. We divide the table $[y_{2^n-1}, \ldots, y_1, y_0]$ into two subtables: least significant

$$[y_{2^{n-1}-1}, \ldots, y_1, y_0]$$

and most significant

$$[y_{2^n-1}, \ldots, y_{2^{n-1}+1}, y_{2^{n-1}}] .$$

For each truth table of an n-argument Boolean function we compute the decimal value of its 4-bit codes in the following way:

1) divide the truth table into 2^{n-2} words; each word is composed of 4 bits;
2) compute the decimal equivalent for every word;
3) compute the algebraic sum of all decimal equivalents.

As an example let the truth table of 5-argument Boolean function be given:

$$[00000101100101001111101001101011].$$

There are eight 4-bit words in the table:

$$0000\ 0101\ 1001\ 0100\ 1111\ 1010\ 0110\ 1011$$

The decimal equivalents of the words and their sum are as follows:

$$\begin{array}{cccccccccc} [0000 & 0101 & 1001 & 0100 & 1111 & 1010 & 0110 & 1011] \\ 0 & +5 & +9 & +4 & +15 & +10 & +6 & +11 & = 60. \end{array}$$

The definitions presented below are taken from [4]. The Boolean function is said to be balanced if in its truth table the number of ones equals the number of zeros. An n-argument Boolean function f is a function with linear structure if there exists $a \in Z_2^n$ such that $a \neq (0, 0 \ldots, 0)$ and for every $x \in Z_2^n$ either $f(x) \oplus f(x \oplus a) = 0$ or $f(x) \oplus f(x \oplus a) = 1$. The Hamming distance of two n-argument Boolean functions f and g, presented with the help of their truth tables, is the number of positions in which the two truth tables differ. The distance of a function f to the set A_n is defined as the minimum of the Hamming distances to all functions of A_n. The nonlinearity of f, denoted by N_f, is the minimal Hamming distance between f and A_n. If f is n-argument, $n \geq 3$, and it is balanced then [5]:

$$\mathrm{N}_f \leq \begin{cases} 2^{n-1} - 2^{\frac{1}{2}n-1} - 2, & \text{for } n \text{ even} \\ \lfloor \lfloor 2^{n-1} - 2^{\frac{1}{2}n-1} \rfloor \rfloor, & \text{for } n \text{ odd,} \end{cases}$$

where $\lfloor \lfloor x \rfloor \rfloor$ denotes the maximum even integer less than or equal to x.

An n-stage nonlinear feedback shift register over GF(2) (nNFSR for short) consists of n cells ($n \geq 1$) joined as in Fig. 1, where symbols of a nonempty alphabet $\{0, 1\}$ may be put in as a Boolean function f of n arguments. The content of all n cells is said to be a state of nNFSR. The nNFSR works in the discrete time. The state of nNFSR at

a given moment $t+1$ $(t \geq 0)$ is determined by its state at the moment t and results from shifting the content of the cell number r to the cell number $r+1$ $(0 \leq r \leq n-2)$, and putting the value $f(x_{n-1}, x_{n-2}, \ldots, x_0)$ of the function f for the state of this nNFSR at the moment t into the cell number 0.

Let us mention that the numbering of register cells is crucial for the algorithm of feedback functions choosing presented later.

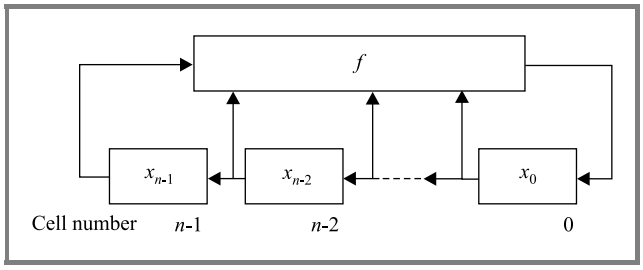


Fig. 1. An n-stage shift register with the feedback function f.

Let $s = (x_{n-1}, x_{n-2}, \ldots, x_0)$ be the state of an nNFSR. The state s_0 from which the nNFSR starts the work is said to be initial. A sequence (s_i) of states of an nNFSR is periodic with period equal to T if T is the smallest positive integer such that for each $i = 1, 2, \ldots$ the condition $s_{i+T} = s_i$ holds. An nNFSR generates periodic sequences of the period $T \leq 2^n$. A sequence of states generated by nNFSR is called a full-length sequence if $T = 2^n$. Each nNFSR generates $B(n) = 2^{2^{n-1}-n}$ full-length sequences [2].

In the sequel a sequence (s_i) of states generated by an nNFSR with a nonlinear n-argument Boolean function f will be called a sequence generated by the function f. An nNFSR that generates a full-length sequence is called full-length shift register.

3. Properties of feedback functions of full-length shift registers

Using the exhaustive search in the set of all n-argument (for $n = 3, 4$ and 5) Boolean functions we chose all functions which give, when used as feedback functions, full-length sequences. We can state the following facts.

Fact 1: The least and the most significant bits of the most significant truth subtable equal 0.

Fact 2: The most significant truth subtable has an odd number of 1s.

Fact 3: The least significant truth subtable is the negation (i.e., respective bits are negated) of the most significant truth subtable.

Fact 4: The sum of decimal equivalents of 4-bit words in the truth tables equals $2^{n+1} - 2^{n-3} = 15 \cdot 2^{n-3}$ (i.e., 15, 30, 60 for $n = 3, 4, 5$, respectively).

Fact 5: Nonlinear n-argument (for $n = 3, 4, 5$) Boolean function generating full-length sequence is balanced (follows from the Fact 3) and is of linear structure.

Fact 6: The greater number of 1s in the truth table the greater value of nonlinearity.

Fact 7: The nonlinearity N_f of obtained function has one of the values: 2, 6, 10, ..., and in general it is equal to $2+4i$ for $i = 0, 1, 2, \ldots$. The nonlinearities never have the maximum value $2^{n-1} - 2^{\frac{1}{2}n-1} - 2$ for n even and $\left\lfloor \left\lfloor 2^{n-1} - 2^{\frac{1}{2}n-1} \right\rfloor \right\rfloor$ for n odd.

There is 2^{26} 6-argument nonlinear Boolean functions generating full-length sequences. Choosing all of them by exhaustive search in the set of 2^{64} 6-argument Boolean functions is difficult with respect to the time needed for computation. Therefore, it was assumed that Facts 1–4 are true also for n-argument $(n \geq 6)$ functions. This assumptions leads to the following algorithm.

4. Algorithm for choosing all n-argument $(n \geq 6)$ nonlinear Boolean functions generating full-length sequences

Input: The most significant truth subtable of n-argument nonlinear Boolean function given by $[y_{2^n-1}, \ldots, y_{2^{n-1}+1}, y_{2^{n-1}}]$.

Output: The set of all n-argument nonlinear Boolean functions (given by the truth tables) generating full-length sequences, $n \geq 6$.

Method:

1. Let $y_{2^n-1} = 0$ and $y_{2^{n-1}} = 0$.
2. For $i = 1, 3, 5, 7, 9, \ldots, 2^{n-1} - 3$ generate in the lexicographical order the words $y_{2^n-2}, \ldots, y_{2^{n-1}+2}$, $y_{2^{n-1}+1}$ having i 1s; for each word:
   a. Construct the most significant subtable.
   b. Construct the least significant subtable by the negation of all bits in the most significant subtable.
   c. Concatenate the tables constructed in steps 2a and 2b.
   d. Verify whether the sum of decimal equivalents of all 4-bit words equals $15 \cdot 2^{n-3}$; if not then process for the next i.
   e. Verify whether the n-stage shift register with the feedback function given with the help of the truth table constructed in the step 2c generates the sequence of period 2^n; if so then store the truth table.

The number of all words of the form $y_{2^n-2}, \ldots, y_{2^{n-1}+2}$, $y_{2^{n-1}+1}$ with i 1s equals $\binom{2^{n-1}-2}{i}$.
Hence, the number of all words having the odd number $1, 3, 5, 7, 9, \ldots, 2^{n-1}-3$ of 1s is equal to

$$s(n) = \binom{2^{n-1}-2}{1} + \binom{2^{n-1}-2}{3} + \binom{2^{n-1}-2}{5} + \ldots + \binom{2^{n-1}-2}{2^{n-1}-3} = 2^{2^{n-1}-3}.$$

If by the efficiency of the algorithm we understand the quotient η of the number $B(n)$ of all n-arguments nonlinear Boolean functions generating full-length sequences and the number $s(n)$ of all examined functions, then

$$\eta = \frac{B(n)}{s(n)} = 2^{-n+3}.$$

For example, if $n = 16$ the efficiency $\eta = 2^{-13}$. It means that on average one function in the set of 8192 functions has the required property. The efficiency of the algorithm is quite satisfactory.
If in the algorithm instead of "generate in the lexicographical order the words $y_{2^n-2}, \ldots, y_{2^{n-1}+2}, y_{2^{n-1}+1}$ having i 1s" we allow "generate randomly the words $y_{2^n-2}, \ldots, y_{2^{n-1}+2}, y_{2^{n-1}+1}$ having i 1s" we can use the algorithm for random generation of Boolean functions generating full-length sequences for an arbitrary n.
The computational experiment confirmed the efficiency of the algorithm for 6-argument functions; for all 6-argument functions Facts 1–7 are true. We verified the randomly generated functions for n from 7 to 20 and in every case Facts 1–7 were true.
The algorithm were successfully used in the synthesis of nNFSRs for FSR-255 family of hash functions [3].

References

[1] T. Etzion and A. Lempel, "Algorithms for the generation of full-length shift register sequences", *IEEE Trans. Inform. Theory*, vol. IT-30, no. 3, pp. 480–484, 1984.

[2] H. Fredricksen, "A survey of full length nonlinear shift registers cycle algorithms", *SIAM Rev.*, vol. 24, no. 2, pp. 195–221, 1982.

[3] T. Gajewski, I. Janicka-Lipska, and J. Stokłosa," The FSR-255 family of hash functions with variable length of hash result", in *Artificial Intelligence and Security in Computing Systems*, J. Sołdek and L. Drobiazgiewicz, Eds. Boston: Kluwer, 2003, pp. 239–248.

[4] W. Meier and O. Staffelbach, "Nonlinearity criteria for cryptographic functions", in *Advances in Cryptology – EUROCRYPT'89*, J.-J. Quisquater and J. Vandewalle, Eds., *LNCS*. Berlin: Springer, 1990, vol. 434, pp. 549–562.

[5] J. Seberry, X.-M. Zhang, and Y. Zheng, "Nonlinearly balanced Boolean functions and their propagation characteristics", in *Advances in Cryptology – CRYPTO'93*, D. R. Stinson, Ed., *LNCS*. Berlin: Springer, 1994, vol. 773, pp. 49–60.

[6] J.-H. Yang and Z.-D. Dai, "Construction of m-ary de Bruijn sequences (extended abstract)", in *Advances in Cryptology – AUSCRYPT'92*, J. Seberry and Y. Zheng, Eds., *LNCS*. Berlin: Springer, 1993, vol. 718, pp. 357–363.

Izabela Janicka-Lipska is an adjunct at Poznań University of Technology, Poland. Her research interest includes data security in information systems and cryptology, especially methods of designing Boolean functions, hash functions and shift registers. Her doctoral thesis is "Nonlinear feedback functions of maximal shift registers and their application to the design of a cryptographic hash function" (2001, in Polish).
e-mail: Janicka-Lipska@sk-kari.put.poznan.pl
Institute of Control and Information Engineering
Poznań University of Technology
Marii Skłodowskiej-Curie Sq. 5
60-965 Poznań, Poland

Janusz Stokłosa is a Professor at Poznań University of Technology, Poland. His research interest includes data security in information systems and cryptology, especially methods of designing cryptographic algorithms. He is author of a number of publications, also books: *Algebraic and Structural Automata Theory* (North Holland, 1991, coauthor), *Cryptographic Method of Data Protection* (1992, in Polish), *Cryptographic Algorithms* (1994, in Polish), *Data Security in Information Systems* (2001, in Polish, coauthor), *Data Protection and Safeguards in IT Systems* (2003, in Polish, coauthor).
e-mail: Stoklosa@sk-kari.put.poznan.pl
Institute of Control and Information Engineering
Poznań University of Technology
Marii Skłodowskiej-Curie Sq. 5
60-965 Poznań, Poland

Paper

Technology solutions for coalition operations

Douglas Wiemer

Abstract— As computer technology has advanced, information processing in command, control, communications, computers, intelligence, surveillance and reconnaissance (C4ISR) systems has become highly complex. The information processed by these systems is usually of a very highly sensitive nature and is entered into specific systems that are physically isolated from each other. The physical isolation of these systems makes it cumbersome to exchange information between systems. The result is inefficient sharing of sensitive information in situations where timeliness of exchange could be a life or death reality. Since the mid 1990's, increasing efforts have been placed on improving coalition operations. Many systems have been created with the goal to improve the sharing of information and collaborative planning across coalition boundaries. The usability of these systems have had mixed levels of success and improvements will always be necessary. This paper will briefly describe three advances in telecommunications technology that could be leveraged to significantly improve coalition operations. These technologies are; the session border controller (SBC), advances in pattern matching technology, and multi-protocol label switching (MPLS).

Keywords— C4ISR, coalition, CCIS, command and control, information technology, military, defence, telecommunications, operations, SBC, session border controller, content inspection, MPLS, multi-protocol label switching, pattern matching, trusted downgrade, explicit route.

1. Introduction

As computer technology has advanced, information processing in command, control, communications, computers, intelligence, surveillance and reconnaissance (C4ISR) systems has become highly complex. The information processed by these systems is usually of a very highly sensitive nature and is often sensitive in both hierarchical (top secret, secret, etc.) levels and non-hierarchical (CANUS, CANUK, AUSCANUK, etc.) levels. Often, this information is entered into specific systems that are physically isolated from each other so that mandatory access controls can be maintained. The physical isolation of these systems makes it cumbersome to exchange information between systems of overlapping security policy. The result is inefficient sharing of sensitive information in situations where timeliness of exchange could be a life or death reality.

Since the mid 1990's, increasing efforts have been placed on improving coalition operations. Many systems have been created with the goal to improve the sharing of information and collaborative planning across coalition boundaries. The usability of these systems have had mixed levels of success.

In general, information domains are separated based on sensitivity of the information that is processed within the domain. In some cases, such as in the case of multi-level secure (MLS) systems, information of differing sensitivity may be processed within a single system. However, in the case of MLS systems, the information is still contained by the technology to prevent the inadvertent release of information from a higher security domain to a lower security domain. In each case, most systems still follow the Bell-LaPadula security policy model [1] for control of access to information of particular sensitivity levels. Basically, the Bell-LaPadula model allows access to information objects based on a "write up" and "read-down" policy. This means that a subject at a lower sensitivity level can write into an equivalent or higher sensitivity level, while a subject of a higher sensitivity level can have read access to information of an equivalent or lower level sensitivity level.

While the Bell-LaPadula model is a key policy model for information security, considerations beyond strict hierarchy of information causes complications. For example, many national military systems maintain additional caveats on information that are not strictly hierarchical. These caveats may be "eyes only" caveats like CANUS, CANUK, etc. – or particular operational codewords. Another good example is the NATO caveat that is placed on information generated within that information domain. These caveats often create subsets of information sensitivity (sometimes termed "non-hierarchical") equivalence that become complicated to control in a coalition environment.

Furthermore, regardless of the presence or absence of particular sensitivity levels, each nation within a coalition has national sovereignty considerations that must be handled in a coalition environment. In most cases, each nation connecting to a coalition information domain will place a firewall between their national systems and the coalition domain. The role of the firewall is to establish access and information flow control between information domains. In addition to firewalls, depending on the nature of the information, additional cryptographic mechanisms will be used to ensure the confidentiality of information in transit.

In addition, cryptographic mechanisms provide the means to verify the integrity of information when received. While firewalls and cryptography systems provide significant measure of control over access to and exchange of information, they create a complex set of intermediary systems between two operational users. A basic coalition connectivity picture is provided in Fig. 1. This diagram represents the conceptual connectivity between any nationally

sovereign operations environment to a coalition operational environment.

In the diagram in Fig. 1 the national system operations represent systems that are under the sole control of a single participating nation. These would be systems that are within the sovereignty of a particular nation. In the coalition operation, there may be many such national systems connected to the coalition operations domain. The coalition operations domain is a shared domain of many participating nations and may often be created for specific operational purposes.

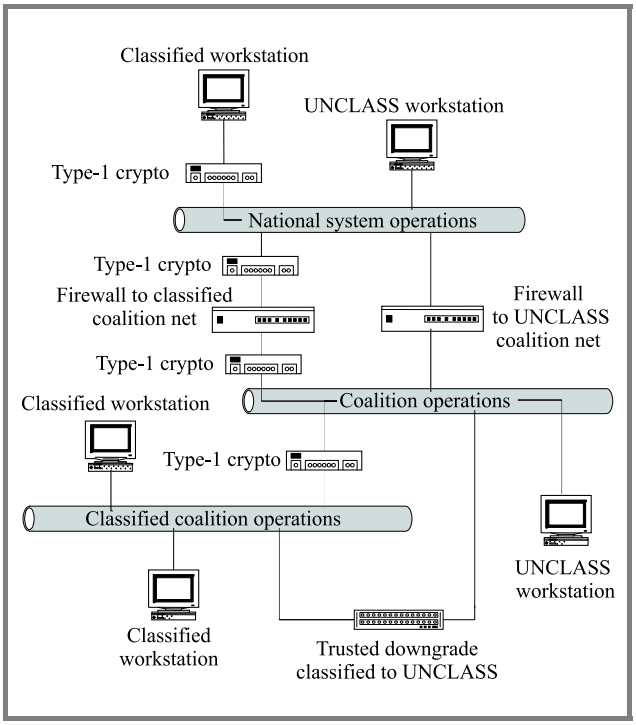


Fig. 1. Basic coalition connectivity.

Throughout the remainder of the paper, this basic coalition connectivity diagram will be used to highlight the issues related to operational needs, created as a result of the presence of intermediary systems. For the issues described, this paper will also describe three advances in telecommunications technology that could be leveraged to significantly improve current coalition operations. These technologies are:

- use of session border controller (SBC) technology to improve interworking of collaborative planning tools across coalition boundaries;
- use of advanced pattern matching engines for improved information sharing across national domain boundaries;
- use of multi-protocol label switching (MPLS) explicit routes (ER) to monitor/control packet flow of sensitive traffic.

2. Collaborative planning tools

2.1. Connectivity and protocol issues

Referring to Fig. 1, systems operating in the UNCLASS national operational domain that are connected to the coalition operational domain must pass traffic through a firewall. Note that these information domains may alternatively be at some other operationally equivalent sensitivity level (i.e., sensitive but unclass (SBU)). The firewall controls access and information flow between these information domains. In general, the firewall policy will be set, such that all traffic is blocked, except particular types of traffic between particular individuals or machines. In general, the policy will, as a minimum, place controls based on a 5-tuple (IP source, IP destination, TCP source, TCP destination, protocol type). Beyond the policies established by the firewall, network address translation (NAT) compounds the problem due to the need for mapping and manipulation of IP addresses at the information domain boundaries.

Classified workstations operating in the national system domain are connected in a similar manner via a firewall. However, note also that Type-1 cryptographic systems are used to protect the information as it passes through the UNCLASS (or other lower operational domains). As a result, the traffic must be decrypted prior to the firewall in order for the firewall to take appropriate policy decisions and then be re-encrypted in the coalition environment until it reaches the classified coalition domain.

In general, security policies are established such that connections between national systems and coalition systems must be initiated from within the national system and must use well known (pre-defined) ports. Any unused port is explicitly blocked. An example of such a policy has been used by the Canadian Forces in the Joint Warrior Interoperability Demonstrations (JWID) and has been described in [2].

The strict nature of such a policy ensures a strong measure of control over the flow of information; however implementations limit capabilities for true collaborative planning. Many collaborative planning tools use a signaling channel, on a known port, to negotiate a data or session channel port. Since the data or session port is not known a priori, the firewall policy is usually configured to block such traffic. The alternative is to leave a block of ports open such that the negotiated data port is allowed. While leaving these ports open solves an operational issue, it also creates additional security risk. This additional risk is usually deemed unacceptable and therefore the ports are closed and the application is denied.

Collaborative planning tools that fall into this category include voice over IP (VoIP), whiteboarding, chat, video teleconferencing (VTC), instant messaging (IM), etc. All of these applications serve a significant role in coalition collaborative planning and most remain a technology challenge to allow the connection of these tools to national systems. Thus, the utility of the collaborative planning tools is limited when national systems specifically deny connection.

While it may be expected that these limitation have been solved in recent years since publication of [2], this is not the case. Since the beginning of 2004, the US government has released at least two separate solicitations regarding the issues surrounding collaborative planning tools [3, 4]. It is recognized that "efficient, seamless ways to share information of varying classification levels and political sensitivities over a single network do not currently exist" [4].

2.2. Session border controller

In 2003, SBCs were voted the number 1 hottest new technology by Telecom Magazine [5]. Used in the telecommunications industry, SBCs exist "to provide a demarcation point between two service providers' VoIP networks, allowing them to manage signalling and control routing for VoIP traffic" [5]. The key ideas behind the SBC are the management of signalling traffic and the routing of the data traffic. Originally designed to facilitate call setup for VoIP traffic, the concepts behind the SBC create a controlled interface between two network domains that are ideally suited for multi-media applications and protocols – the types of protocols that support coalition collaborative planning tools.

Sitting at the interface between two information domains, the SBC intercepts the signalling protocol between two systems. Taking the example of VoIP, the SBC will monitor for either of two dominant standards, the session initiation protocol (SIP) or H.323. When intercepted, the SBC either directly manages and controls the connection using an internal application level gateway (ALG), or it uses a separate control interface protocol to communicate with an external ALG system. The mechanisms required to support the external ALG system are being defined by the Middlebox Communications Working Group (MIDCOM WG) of the Internet Engineering Task Force (IETF).

While the details of SIP and H.323 differ significantly, the goal is the same, to establish a voice connection between two VoIP end points (phones). This paper will focus on SIP as "some observers believe that SIP will become dominant" [6]. SIP is "an application-layer control (signalling) protocol for creating, modifying, and terminating sessions with one or more participants. These sessions include Internet telephone calls, multimedia distribution, and multimedia conferences" [7]. One of the key benefits of SIP in the context of coalition interoperability is the fact that it is media independent. This means that the SIP itself is not tied to a particular media type (i.e., voice), but can be used for virtually any type of media traffic (i.e., video, instant messaging, whiteboarding). This is because negotiation of the media type and the parameters of the session are negotiated during the call setup process.

SIP supports five facets of establishing and terminating multimedia communications [7]:

- user location: determination of the end system to be used for communication;
- user availability: determination of the willingness of the called party to engage in communications;
- user capabilities: determination of the media and media parameters to be used;
- session setup: "ringing", establishment of session parameters at both called and calling party;
- session management: including transfer and termination of sessions, modifying session parameters, and invoking services.

A simplified illustration of SIP being used to initiate and control a session between two end points is provided in Fig. 2. The initiating end point sends an "Invite" request to the recipient. Among other fields, the "Invite" will contain several fields that relate to the routing of the call to the recipient. These fields are the "Via" containing the address expected for response; "To" containing the destination universal resource identifier (URI) and "From" field containing the source URI.

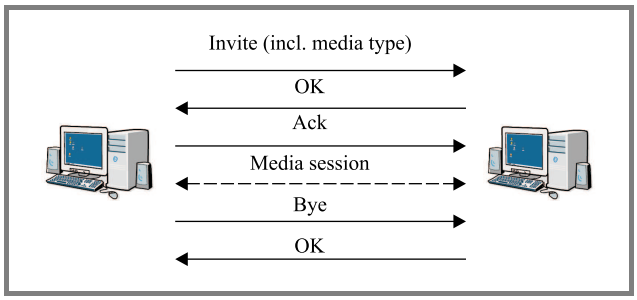


Fig. 2. Session initiation protocol.

In addition, the SIP invite will contain a "content-type" field that is used to identify the media application type that will be described in the body of the SIP invite message. For example, the content-type field may identify "application/SDP" to identify the session description protocol (SDP), used for voice. The body of the SIP invite would contain parameters associated with the SDP that could be processed by the recipient to enable the media type. The "content-type" and the body of the SIP payload are critical to the SIP message as it will contain the IP address and dynamic port assignments for any collaborative planning applications. This information is essential for the operation of the session border controller, as will be described below.

Assuming that the parameters of the invite are accepted, and then the recipient of the invite responds with an "OK" message. This message is acknowledged by the originator using an "Ack" and the media session is established.

During the course of a session, either end may add or remove other media sessions or types as required, by negotiating these sessions through the existing SIP session. When either end point terminates the session, a "Bye" message is sent to the other party, which is acknowledged by an "OK" message.

Not included in the illustration of SIP in Fig. 2 is the SIP use of proxies. Within the definition of SIP, the protocol allows for the use of intermediate proxies that are used to

relay the messages from the initiator to the recipient. Typically a SIP call will not go directly between two end points but will instead pass through proxies at the boundary between different network domains. In the case of a connection to a coalition network, a proxy would sit at the boundary between the national domain and the coalition domain. It is possible; however, that additional proxies will be required as the call request is passed through various points in the network. This would depend on the overall network connectivity.

A session border controller acts as a firewall that uses an application level gateway programmed to understand SIP. ALGs go deep into the data in the SIP packet and parse the "content-type" and payload. This allows the ALG to determine the IP addresses and dynamic ports that are required to enable the data ports of the collaborative planning applications. By understanding which ports need opening, the SBC dynamically opens only those ports needed by the application, leaving all others securely closed. This technique of opening small numbers of ports in the firewall dynamically is called "pinholing". One of the key advantages of the ALG is that the constant monitoring of the session ensures immediate knowledge of call termination, allowing the "pinhole" to be closed immediately as well.

As described earlier, NAT causes difficulties in the use of collaborative planning tools. A SIP proxy is used to provide NAT traversal. The proxy has knowledge of the IP domain on both sides of the proxy, and separates the SIP call into two separate calls: one from the end point in the national domain to the proxy and one from the proxy to the coalition domain. The proxy is an intermediary control point and resolves the NAT issue. In most cases, the ALG will also incorporate a proxy and therefore is able to handle both NAT issues and the dynamic assignment of ports.

In some instances, the ALG of an SBC will be implemented in a separate device from the firewall. In this case, the SIP messages will be routed to the separate ALG for processing. The separate device will then dynamically control the firewall by telling it the IP address and UDP (or TCP) port information determined from the SIP payload. This approach using a separate device is being promoted by the MIDCOM WG in the IETF and is illustrated in Fig. 3.

The advantage of the MIDCOM scheme is that the firewall is not burdened by the impact of processing the SIP messages. Once the session is established, and pinholes are created, only the media streams are processed by the firewall and the impact of media dependent characteristics for latency, jitter and quality of service are minimized. In addition, once implemented with a MIDCOM interface, the firewall no longer needs to be upgraded for each new application service. Instead, the ALG can be upgraded separately, thereby minimizing the operational impact on the firewall.

In summary, the SBC provides a dynamic firewall solution that can be used to improve coalition operations. The nature of the connectivity to a coalition network creates difficulties due to NAT and the firewall policies that block applications using dynamic port assignments. The SBC uses proxy enabled ALGs programmed to understand SIP that dynamically open ports in the firewall. This dynamic control of ports is termed "pinholing". The ALG may be implemented directly on the firewall, or it may be a separate device that communicates with the firewall using a MIDCOM scheme. The use of the ALG could allow collaborative planning tools to span the national and coalition information domain while minimizing the risks previously associated with the dynamic port assignments.

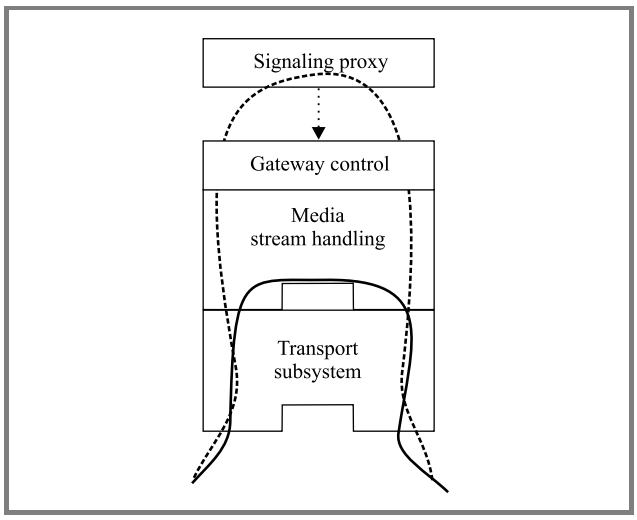


Fig. 3. ALG as separate signaling proxy.

3. Trusted downgrade systems

Another area of connectivity that causes difficulties to coalition operations is the interface between higher and lower level sensitivity domains. As shown in Fig. 1, a trusted downgrade system would sit at the interface between the classified operational domain and the UNCLASS or SBU domain. The role of a trusted downgrade system is to allow the approved release of information from the higher level domain to the lower level domain. Recalling the Bell-LaPadula model, it is specifically denied by policy to "write-down", that is, to pass information from the higher domain to the lower domain is forbidden.

However, this security model does not account for the "system high" nature of the classified operational domain. "System high" refers to the concept that all information in an information domain is treated as of the highest classification of information processed on the system. This means that despite the fact that some information may not be SECRET, if it is contained in a system that operates at the SECRET level then it is treated as if it is SECRET. Consequently, strict enforcement of the Bell-LaPadula model prevents the valid transfer of information from a higher domain to a lower information domain.

Recognizing this limitation, there have been several products developed and approved for operation that perform a trusted downgrade operation. The Radiant Mercury [8] is one such system developed by Lockheed Martin that has been in service at least since the mid 1990's. There are other similar systems in operation such as the ISSE Guard [9]. In general, these systems operate on a trusted computing base (TCB) that has been evaluated and approved in accordance with one of the trusted computing evaluation programs, such as the common criteria for information technology security evaluation (published as ISO standard 15405). As well, these systems enforce the control of information release from high to low in one of two manners. Either they use automated methods to approve release based on a review of highly formatted messages such as USMTF messages, or they rely on the approval of a release authority (i.e., approved email from a valid user with appropriate rights for release). A concise list of current MLS systems in use, including systems used for trusted downgrade is available at [10].

Taking the RM as an example, the RM release 3.0 serves two main roles and operates on a Sun platform. The RM has the capability to automatically review and approve the release of highly formatted message according to predefined rules. Due to the highly formatted nature of the message and the extensive rule base, it is possible for effective controls to be established such that only information appropriate to the lower sensitivity domain is released to that lower domain. The second feature of the RM is to approve the release of imagery files. In this case, the Radian Mercury operates in a manner similar to most trusted downgrade systems; an authorized release authority must identify the file as approved for release and pass the file to the RM. The RM examines a header that has been applied to the file by the release authority and passes the approved image into the lower security domain.

The RM and similar products operate on a TCB that is based on a trusted operating system running on a general purpose processor. The operating system is responsible for the secure containment of information on both the high and low side of the system and is also responsible for the trusted transfer of information from the high side to the low side. This architecture has proved very useful, secure and is able to meet the needs of some operational requirements. However, with the growing scope of coalition operations, these systems may not be able to handle the increased demands placed on them. The architecture used forces a complete reassembly of the information content in order for the application to scan, parse, review, modify, approve and release all messages. This is a highly processor and memory intensive process that can be impacted by increased demands on the system.

On the other hand, modern networking equipment uses datapath technology that is designed to scan, parse, review and modify information at the packet level. For example, as a packet is received by a router, the packet header information must be scanned and parsed to extract the common 5-tuple information (IP source, IP destination, TCP source, TCP destination, protocol port). This information is used as a lookup key for access into a forwarding information base (FIB). Effectively, the FIB provides instructions to the datapath to inform it of the actions that should be taken on the packet. Often, these actions may include the modification of the packet (i.e., in the case of NAT) prior to forwarding to the release interface.

At first glance, the technology used in the switch and router datapath may seem ideally suited to the problem of trusted downgrade operations at very high rates. However, the technology widely used in switches and routers has been engineered for the specific problem of fast header inspection and forwarding. On the other hand, they are not well suited to scanning and parsing of information deep within the payload of a packet or where information spans across multiple packets. However, the basic technology used in switches and routers has formed the basis for advanced pattern matching engines (PME) that support inspection at higher layers.

3.1. Pattern matching engines

Pattern matching engines have been developed to support a variety of applications. These applications range from content switching to intrusion detection and prevention systems (IDS/IPS) to automated anti-virus platforms. PMEs have also been called content inspection engines (CIE) or simply search engines (SE).

There are two main types of PMEs, those that are based on a Ternary Content Addressable Memory (TCAM) technology and those that are algorithmic based. PMEs may optionally support both exact match and regular expression matching criteria. The TCAM type of PME has a direct relationship to the technology in use in the switch and router datapath for packet forwarding.

PMEs provide considerable potential for the improvement of the capability of trusted downgrade systems. The matching engine provides the baseline technology needed to perform high speed parsing and review of received information and the technology is intended to provide pattern matching capability deep into packet contents and across packet boundaries. An anti-virus scanning feature is a typical example where the entire payload of a data stream may need to be scanned. These applications exist in current products available on the market.

In some cases, PMEs are available directly from component suppliers. The PAX.port™ line of devices from IDT [11] is an example of PMEs available from a component supplier. According to the Linley Group, IDT, Inc. is currently the market leader in SE components, with Cypress ranked second [12]. PME components hold promise for the future of trusted downgrade platforms since they offer a flexible program matching language that can support a wide range of applications. Unfortunately, this also remains their biggest challenge. Following development of a system, considerable effort may be required to program the device for particular application needs of trusted down-

grade applications. Also, the technology is generally used for matching against relatively short "keys" or "strings". Additional research is required to determine the suitability of such devices for the multiple match criteria that may be required, for example, in applications supporting military message formats (i.e., USMTF).
In addition to the general device supplier category, there are system vendors that have performed extensive research into PME technology and use this technology in their platforms. In this case, the technology is often tuned to the particular application space. For example, the FortiGate™ line of products from FortiNet™ makes use of the FortiASIC™ to perform fast pattern matching for anti-virus applications [13].
At this stage, additional research into the use of PME technology to support trusted downgrade applications is required. However, it appears that this technology could provide better baseline platform capability than the general purpose processors architectures in use today.

4. Controlled flow of sensitive traffic

4.1. Nature of provider networks

In a multinational coalition operation, connectivity for both tactical and strategic networks is established through network paths that are likely not a part of the normal grid used by these nations. As new network paths are created, it is often a requirement to lease the network from providers. This leaves the network connectivity paths outside the control of the national military force requesting the service. This creates a situation where the provider may route the traffic through other nations where the owner of the data may not want traffic to pass.
In fact, even using the standard strategic networks that provide for normal national operations, the provider may route traffic through areas where the national military may not want the traffic to go. Fortunately, in the case of normal day-to-day operations, the provider provisioned network is often controlled by strict contractual agreements that preclude routing of traffic through particular areas of the world.

Before considering the mechanisms available to control the flow of sensitive information, it is important to highlight why the traffic routes are a concern. On the one hand, all sensitive operational traffic will be protected from disclosure by some cryptographic means. In the case of classified operational traffic, Type-I cryptography is used, thereby providing the assurance that even if intercepted, the traffic is unreadable and is therefore protected. This being the case, there it can be argued that the route taken by the traffic is not a concern.
On the other hand, information system security needs to account for the integrity, availability and accountability of the information, not just the confidentiality. Integrity protection ensures that any modification of the traffic is detected. In addition, it is desirable that opportunities to modify the traffic be minimized. It may be beneficial to know that the traffic has been tampered with, but this doesn't help with the fact that the correct data has not been received. Also, availability concerns highlight the importance of ensuring that that there are no interruptions to service guarantees from the provider.
Given these concerns, an argument can be made that national bodies may still desire to have added control of the path that traffic takes within a provider network. The nature of routed data networks does not really support this type of control. Routing protocols are designed to negotiate best path options for traffic. While the network provider can establish controls of the paths, the path options are based on the provider considerations for best path, not the considerations of the data owner. For example, a provider will establish paths to maximize bandwidth utilization and meet quality of service (QoS) guarantees. In some cases, this may result in the passing of traffic over links that reside in hostile locations. It would be beneficial for the coalition operations partners to have some measure of control over the approval of traffic.

4.2. Multiprotocol label switching and explicit routes

In traditional routed networks, the routing of traffic is based on the address of the destination of the packets. In the case of most networks today, this address is an Internet Protocol (IP) address. By contrast, in label switching, "instead of a destination address being used to make the routing decision, a number (a label) is associated with the packet... a label is placed in a packet header and is used in place of an address (an IP address usually), and the label is used to direct traffic to its destination" [14].
Label switching provides several advantages to the network provider [14]; speed and delay, scalability, simplicity, resource consumption, and route control.
Route control is the key consideration in the context of control over coalition traffic paths. Route control allows the system to designate a specific route path from among many that may lead to the same destination. This is sort of like placing an "Air Mail" label on a letter. With the "Air Mail" label, the letter will take a non-standard path that, one would hope, has an improved delivery time. In the same way, a label can be used in a system to control the route taken. The provider can engineer the network to route high priority traffic to one set of resources, while lower priority traffic takes a different path. The removal of lower priority traffic from the high priority resources reduces congestion and ensures guaranteed service levels can be met.
Multiprotocol label switching is published under RFC3031 [15]. MPLS combines label swapping and forwarding with network layer routing. "The idea of MPLS is to improve the performance of network layer routing and the scalability of the network layer" [14]. Within MPLS, a label switch path (LSP) is established either through a route negotiation protocol (i.e., link determination protocol), or through constraint-based routing. In constraint-based routing, the LSP is established manually.

It is possible to combine both automated route protocol establishment and constraint-based LSP configuration. In this case, the automated routes would be restricted by the configured constraints. These constraints are often associated with quality of service. A router that is aware of MPLS is termed a label switch router (LSR).
There are two main methods that MPLS uses to choose an LSP between nodes. In the first method, the LSR is free to independently select the next hop LSR based on knowledge it has in its routing table. The second method is called explicit routing. ER is used to define constraints on the LSP by identifying specific LSRs that are must be used in the LSP. Assuming that provider networks are MPLS capable, the concept of ERs could be extended to define constraints on the LSP that prevent coalition traffic from passing through routers that reside on undesired traffic paths.

Note that to control traffic based on sensitivity, MPLS would need to be extended. As described earlier, ER is generally used to provide QoS guarantees. Within the various methods to negotiate an LSP, there is no real concept of LSR location information, nor is there any notion of an "approval to process" identifier. This information would be critical to the extended use required to control the flow of sensitive information. Furthermore, the constraint-based link determination protocol (CR-LDP) used to establish LSPs would need to be extended to include an authentication mechanism that includes both location and approval to process criteria.

5. Evaluation considerations

The methods described in this paper are not specifically related to security products. However, the use of these methods to support coalition interoperability is identified to ease the burden of constraints placed on coalition operations due to security concerns. Therefore, it is important to note that incorporating these methods into systems supporting coalition interoperability will require trusted product evaluations and certification and accreditation for operation. Current research into the use of these technologies has not included any consideration for product security evaluation, though there is no known reason to believe that these techniques could not be included in a secure system design.

6. Conclusions

This paper has examined several issues related to coalition interoperability. These issues related to:

- the denial of collaborative planning tools across the national to coalition boundary;
- processing requirements for trusted downgrade platforms;
- the controlled flow of sensitive traffic.

Despite development and operational deployment of many systems, each of these topics remains a challenge to coalition interoperability.
This paper has identified three technology advances that could be used to improve coalition interoperability. These technology advances are:

- the session border controller;
- advances in pattern matching technology;
- use of multi-protocol label switching explicit routes.

Integration of any of these technologies will require trusted product evaluation and certification and accreditation for operational approval.

References

[1] D. E. Bell and L. J. Lapadula, "Secure Computer Systems: Unified Exposition and Multics Interpretation", MTR-2997, Rev. 1, MITRE Corp., Bedford Mass., March 1976.

[2] D. Wiemer, "Wiemer-Murray domain security policy model for international interoperability", in *Proc. NISSC*, Arlington, USA, 1998, http://csrc.nist.gov/nissc/1998/proceedings/paperF20.pdf

[3] Defense Information Technology Contracting Office (DITCO), "DRAFT Next Generation Collaboration Service, Statement of Objectives", published as pre-solicitation, 27 Jan. 2004.

[4] Department of the Navy, Naval Supply Systems Command, "Multinational Information Sharing Environment", published as presolicitation notice, no. H1967, 23 Jun. 2004, http://www1.eps.gov/spg,

[5] Telecommunications Magazine, "10 hottest technologies", Apr. 2003, http://www.telecommagazine.com/default

[6] National Institute of Standards and Technology, Special Publ. 800-58, "Security Considerations for Voice Over IP Systems", Apr. 2004, http://csrc.nist.gov/publications/drafts/NIST_SP

[7] J. Rosenberg *et al.*, "SIP: Session Initiation Protocol", RFC3261, Internet Engineering Task Force (IETF), June 2002, http://www.ietf.org/rfc/

[8] J. Pike, "FAS intelligence resource program: Radiant Mercury (RM)", Jan. 2000, http://www.fas.org/irp/program/disseminate/radiant_mercury.htm

[9] ISSE Guard Program Office, "ISSE Guard, suite of products for secure multi-domain information exchange", Nov. 2003, http://www.rl.af.mil/tech/programs/isse/

[10] D. Zellmer, "Multi-level security: reality or myth", As part of GIAC practical repository, SANS Institute, March 2003, http://www.giac.org/practical/GSEC/Douglas_Zellmer.pdf

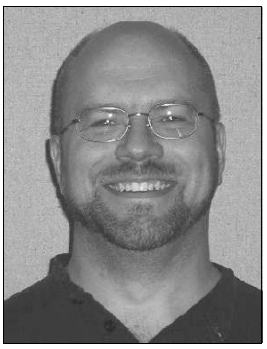
[11] Integrated Device Technology, Inc., "IDT™ PAX.port™ 1200 Content Inspection Engine", 2003, http://www.idt.com/docs/76T1200BH_BR_88384.pdf

[12] The Linley Group, "Search engine market maturing", *The Linley Wire*, vol. 4, issue 11, June 2004, http://www.linleygroup.com/npu/newsletter/wire061004.html#2

[13] FortiNet™, "Comprehensive solutions for real time network protection", 2004, http://www.fortinet.com/doc/FortinetBroch.pdf

[14] U. Black, "MPLS and Label Switching Networks". Prentice Hall, 2002.

[15] E. Rosen, A. Viswanathan, and R. Callon, "Multiprotocol Label Switching Architecture", RFC3031, IETF, Jan. 2001, http://www.ietf.org/rfc/



Douglas Wiemer is a retired Canadian Armed Forces Captain. While serving, he was employed in the Air Force as a Communications and Electronics Engineer and specialized in information security on data networks for Command, Control and Intelligence Systems (CCIS). Among other roles, he served as the System Security Engineer for the planning of the Canadian participation in the Joint Warrior Interoperability Demonstration 1997 (JWID'97). Douglas Wiemer is currently employed in the Alcatel Networks, Research and Innovation (R&I) group on the Intelligent Switch Routers (ISR) project. His research is focused on architecture studies of the switch and router datapath. These studies are used to assess the capabilities and technology needed in the datapath to enable advanced services like the Session Initiation Protocol (SIP).

e-mail: douglas.wiemer@alcatel.com
Alcatel Networks
Research and Innovation Group (R&I)
Intelligent Switch Routers Project (ISR)
600 March Rd, Ottawa, Ontario, Canada

Paper

User services of tactical communications in the digital age

Esra Çiftçibaşi Erkan and Şenol Uzun

Abstract—Increasing demands on an extensive amount of digital data and information flows for C4I is forcing the modern armies to freeze and omit EUROCOM based tactical area communications system and to develop new concepts based on the adoption of modern communication systems such as ISDN, ATM described in the TACOMS Post 2000 Final Report II in NATO. ASELSAN, as a leading company in the military electronics arena, is following and participating all the activities of TACOMS Post 2000 together with Turkish Ministry of Defense. Turkey's tactical area communications system TASMUS is a mature and fielded system, which will satisfy future communication needs of the 21st century C4I systems. In this paper, we describe the basic features and user services of TASMUS. With the support of simultaneous voice and data capabilities, TASMUS aims to form mobile, survivable, flexible and secure network to support all the present and future communication requirements of the tactical commanders. Using the near real time data communications feature, TASMUS is also significant for the network-centric warfare applications such as tactical sensor and weapon systems, besides the communication needs of the Turkish Army.

Keywords— tactical communications, IP, X.25, C4I systems.

1. Introduction

Tactical battlefield is now becoming a ground for extensive digital data exchange where many sensors, weapons, computers and command centres need to exchange high-speed data in order to perform effectively and coherently. More so, these units need to carry out their data exchange while on the move because the new military doctrines heavily emphasize mobility and flexibility.

Future battlefield, no doubt, will be a digitized one. The emerging issues related to the implementation of the digitized battlefield are listed below.

1. Common picture of the battlefield in near-real time.
2. Shared data among battlefield operating systems.
3. Ability to concentrate on combat power effectively and decisively.
4. High-speed exchange of data.
5. Fusion and display of intelligence information to commanders at all levels.
6. Rapid exchange of targeting data from sensor to weapon.

All of these issues are related to reliable and efficient exchange of information on the tactical field.

On the battlefield, several command and control functions such as fire support, manoeuvre control, intelligence, electronic warfare, and logistics support need to be executed simultaneously. Command and control functions rely on rapid and reliable exchange of information on the tactical field. These command control functions may have different communication requirements, however in consequence what needed is an integrated solution that will provide the necessary communication support to all of these command and control applications.

Existing tactical area communication systems, such as EUROCOM, are not sufficient to meet the demands of the future battlefield. Even NATO, the originator of EUROCOM standards has an ongoing project the Tactical Communications Post 2000. The aim of this project is to define the next generation tactical area communication systems for NATO.

So, looking at the today's tactical picture, new concepts and state-of-the-art technologies need to be utilized in order to meet the communication requirements of the future battlefield. This is exactly what the TASMUS project is all about. The TASMUS project has been realized to meet the challenges of the future battlefield.

2. Digital information age and TASMUS

The tactical communication system TASMUS aims to form: mobile, survivable, flexible, secure network to support all the present and future communication requirements of the commanders, and also to provide the commanders with the communications background to form the real time picture of the battlefield. TASMUS makes the crucial near real time data communications needed by the tactical sensor and weapon systems available.

TASMUS is deployed in area (theatre) of military operations such that, seamless communication between the army command and battalion/company commander level is achieved. TASMUS provides interfaces to the strategic systems above the army level, while providing connection to the existing CNR systems via CNRI.

TASMUS brings together the state of the art technologies in military communications. It incorporates ATM and ISDN switching technologies, together with digital ISDN terminals with built-in crypto module, enabling simultaneous voice and data capability with synchronous, asynchronous, X.25 data, IP data and video interfaces. Video conferencing is also supported over the TASMUS network.

Mobile subscribers in TASMUS acquire service thorough TDMA radios, namely, iSTAR (Integrated Services Tactical Radio) acting as the mobile access terminal. The iSTAR concept is based on radio networking and packet communications. On the tactical field, TDMA radios use time division multiple access technique for accessing the medium and automatically form a radio network where all the network management functions are carried out in a distributed fashion. The TDMA radio network can provide multiple simultaneous voice and data connections to the mobile users. The TDMA radios have built-in encryption/decryption, ECCM and LPI/LPD features enable secure communications in the battlefield. The TDMA radios are also equipped with GPS receivers. The TDMA radio system automatically distributes the GPS information of each mobile unit over the entire TASMUS network.

3. TASMUS features

In the tactical field:

1. TASMUS responds to all of the voice, data, fax and video communication requirements of the battle groups, with an integrated system solution.

2. TASMUS presents a communication infrastructure, providing survivable and reliable communication services with prompt response to all kinds of sudden changes using distributed routing algorithms and flexible system topology upgrades.

3. ATM technology, incorporated in TASMUS nodal points, enables efficient use of link capacities between switches and makes the support of flexible and distinct services possible to its users.

4. ISDN technology, incorporated in TASMUS access switches, furnishes integrated communication service support such as simultaneous voice and data (asynch., synch., X.25, IP, video, video conference) to its users through the tactical ISDN terminals.

5. ADSL technology, incorporated in TASMUS access switches, gives the support of IP service to its users through the tactical ADSL terminals. Each ADSL terminal has 3 Ethernet interfaces for generating LANs.

6. In terms of communication services, TASMUS provides plain and encrypted voice, async data up to 38.4 kbit/s, synch. data up to 64 kbit/s, X.25 packet data up to 64 kbit/s, IP packet data up to 640 kbit/s, video up to 64 kbit/s and video conference up to 384 kbit/s to its wired digital subscribers by using tactical ISDN terminals and ADSL terminals. TASMUS also provides plain and encrypted voice, async data up to 38.4 kbit/s, synch. data up to 64 kbit/s, X.25 packet data up to 9.6 kbit/s, IP data up to 64 kbit/s depending on the application and video up to 64 kbit/s to its mobile subscribers by using iSTAR TDMA radios.

7. All the voice traffic in TASMUS is carried with 4.8 kbit/s CELP voice coding.

8. In terms of information services TASMUS provides communication background to figure out near real time picture of the battlefield to the tactical commanders. At any time during the battle, commanders are provided with the geographical positions of all of their subordinate units on the tactical field. Visualizing the real time picture of the battlefield to the tactical commanders will enhance the tactical decision-making process will also reduce the voice communication traffic.

9. Near real time data communications is required by tactical sensor and weapon systems. This kind of sensor to weapon communication involves exchanging target data and is intolerable to delays. The communication system must deliver the information to the destination in time. TASMUS provides near-real time communications required by tactical systems.

10. TASMUS supports all of the teleservices, bearer services and supplementary services including military features such as “priority”, “pre-emption”.

11. TASMUS network management and planning system, “SYSCON” meets military network management and planning requirements by using both ITU-T M.3000 TMN system control concepts and NATO TACOMS Post 2000 system control concepts.

There is also a tactical data bank in TASMUS, which provides digital maps, geographical data, meteorological info, intelligence reports, and logistics info.
TASMUS forms a rapid, flexible, reliable, survivable and secure tactical communication network to meet the current and future’s armed forces mobility needs.

4. TASMUS network architecture

As shown in Fig. 1 TASMUS has a layered architecture.
Highest layer is the Wide Area Subsystem (WAS), which carries out the backbone switching, constituted of the nodal points. Interfaces to the strategic systems and PTT are located on the WAS.
The middle layer is the Local Area Subsystem (LAS), which is formed by access points (AP) connected to the nodal points (NP). The APs contain the access switches through which the users access the system. The APs also constitute a gateway for the mobile users.
The lowest layer is the Mobile Subsystem (MS). In the Mobile Subsystem mobile subscribers use mobile subscriber terminals (MST) to access the TASMUS switching backbone.

In addition to those layered subsystems, the system control SYSCON carries out all the necessary control functions such as system planning, control and management.

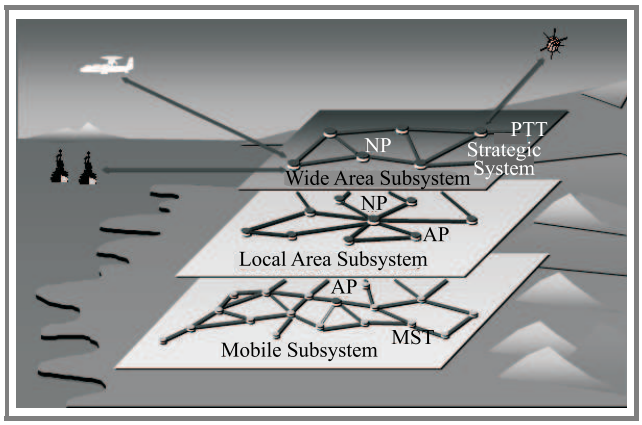


Fig. 1. TASMUS layered architecture.

Detailed architecture, interconnection of the subsystems and interfaces to the other systems such as CNR, PTT and strategic system is shown in Fig. 2.

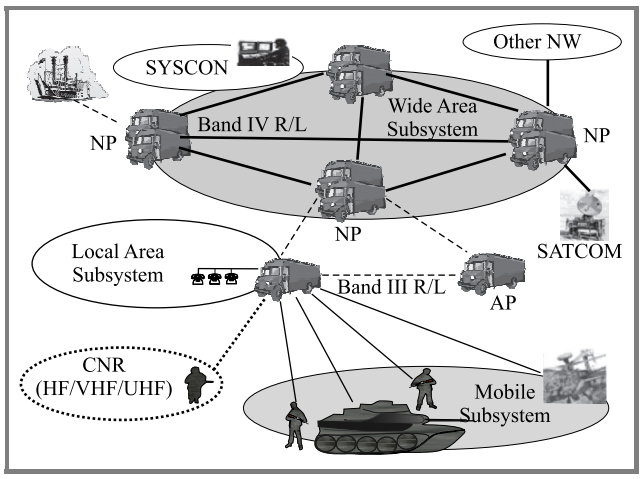


Fig. 2. Detailed architecture of TASMUS.

For different radio applications at WAS, LAS and MS, TASMUS uses multi-band multi-role iSTAR radios.
iSTAR is a new generation radio family which combines both single channel radio access (SCRA) and packet radio concepts.
iSTAR is capable to be used as mobile terminal (both for voice and data) and mobile terminal access equipment to switching backbone. Those two types of applications are given below.

1. iSTAR radio functioning as radio access point (RAP) for the access of mobile subscriber (AP switch connection to mobile subscriber).

2. iSTAR radio functioning as mobile subscriber terminal (MST) meets all the communication requirements of the mobile subscribers and forms independent network among the MSTs.

5. User terminals in TASMUS

TASMUS has digital subscriber terminals, ADSL IP terminals, iSTAR mobile subscriber terminals (iSTAR-MST) and iSTAR personal subscriber terminal (iSTAR-PST) all of which have build-in encryption.
The digital subscriber terminals have ISDN S_0 interface with the ISDN switch. They have voice and data capability at the same time. They are capable of performing asynchronous data transfer up to 38.4 kbit/s, synchronous data transfer up to 64 kbit/s, and near real time X.25 transfer up to 64 kbit/s. They also have IP packet data transfer capability. Each digital subscriber terminal can serve IP addresses and IP capability, via the switch which it is connected, to maximum 5 external host computers. They can also send SMS messages to the other user terminals among the TASMUS and to the ordinary GSM phones via a GSM modem.
ADSL IP terminals have ADSL interface with the ISDN switch. They are used for IP packet data communication. They have 3 Ethernet ports and via these 3 ports totally 510 host computers can be connected to TASMUS IP network or any other connected strategic/tactical IP networks. They act as IP routers.
iSTAR-MST and iSTAR-PST terminals are the devices of TDMA radio family. They have voice and data capability at the same time. They are capable of performing asynchronous data transfer up to 38.4 kbit/s, synchronous data transfer up to 64 kbit/s, and near real time X.25 data transfer. They also have IP packet data transfer capability. iSTAR-MST's can serve IP addresses and IP capability to maximum 5 external host computers. iSTAR-MST's and iSTAR-PST's can also send SMS messages to the other user terminals among the TASMUS and to the ordinary GSM phones via a GSM modem.

6. User services in TASMUS

TASMUS provides several services to the tactical area users such as secure voice, file transfer (using near-real time X.25, IP packet data service, synchronous and asynchronous protocols), web access, e-mail, short message service, video conferencing, digital fax, and wireless LAN applications.

Secure voice service. TASMUS provides encrypted and non-encrypted voice services using the end-terminals such as iSTAR-MSTs, iSTAR-PSTs, digital subscriber terminals and analog subscriber terminals. TASMUS secure voice service is 4.8 kbit/s CELP encoded. These terminals are located in every point of TASMUS, such as the nodal points, access points or the system control point.

File transfer service. TASMUS has capability to transfer files between hosts connected to the end-point terminals. Asynchronous, X.25 and IP packet data services can be used for file transfer. File transfer protocol (FTP) is used for file transfer using IP.

For FTP applications, an FTP server is located in the system control point in TASMUS as shown in Fig. 3.

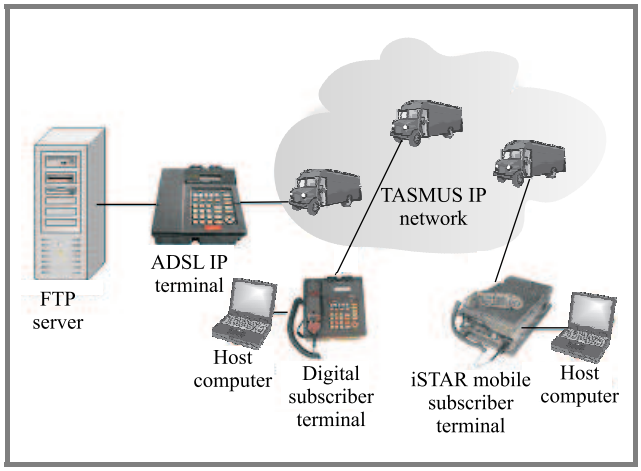


Fig. 3. TASMUS FTP service.

The file transfer can also be performed using the near-real time X.25 service in TASMUS via the user terminals digital subscriber terminals, iSTAR-MST and iSTAR-PST.

Web service. With the use of an HTTP server TASMUS has capability to support web services in the tactical area. Via the end-terminals digital subscriber terminal, ADSL IP terminal, iSTAR-MST and iSTAR-PST; the hosts can be connected to the TASMUS IP network and can access the web applications through the HTTP server, which is located in the system control point of TASMUS. A similar web access application is shown in Fig. 4.

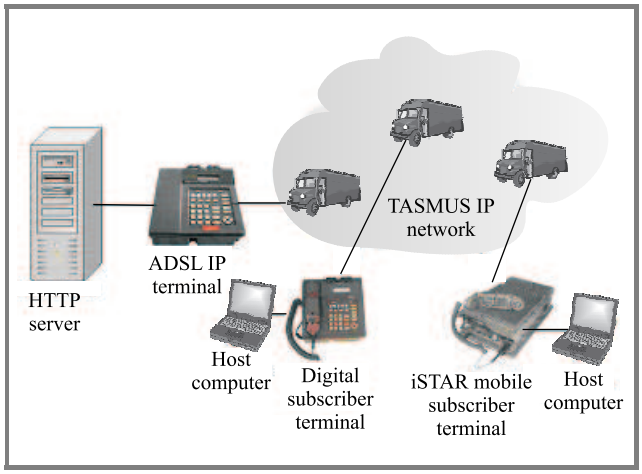


Fig. 4. TASMUS HTTP service.

E-mail service. TASMUS can serve e-mail service to its hosts that are connected to the TASMUS IP network. The end-points, digital subscriber terminal, ADSL IP terminal, iSTAR-MST and iSTAR-PST can be used to be connected to the TASMUS IP network.

For an e-mail application, the SMTP is used between the hosts and the mail server. The mail server (SMTP server) is again located in the system control point and the hosts can create accounts and send e-mails to each other via this server. A typical application of this service is shown in Fig. 5.

In this service the mail that host computer-1 creates, is first sent to the e-mail server by the network. The server then processes the mail and sends it back to its original destination host computer-2.

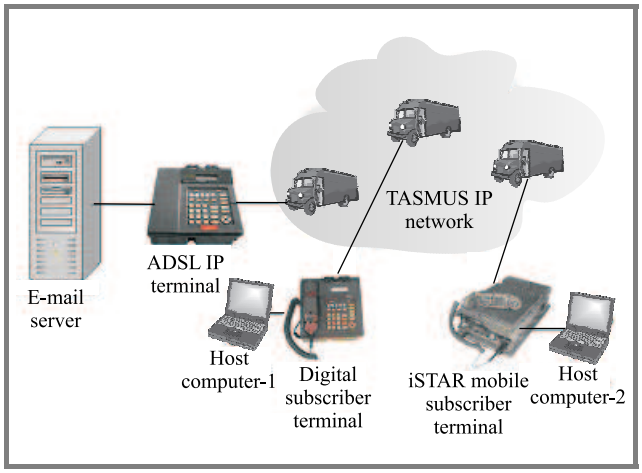


Fig. 5. TASMUS e-mail service.

SMS service. TASMUS end-terminals digital subscriber terminals, iSTAR-MST's and iSTAR-PST's have capability to send and receive SMS messages like commercial GSM phones but all are encrypted. This facility is handled with the use of an SMS server, again located in the system control point of TASMUS. The created SMS messages are first sent to the SMS server. They are logged in the server and forwarded to the original destination terminal afterwards. A typical application is shown in Fig. 6.

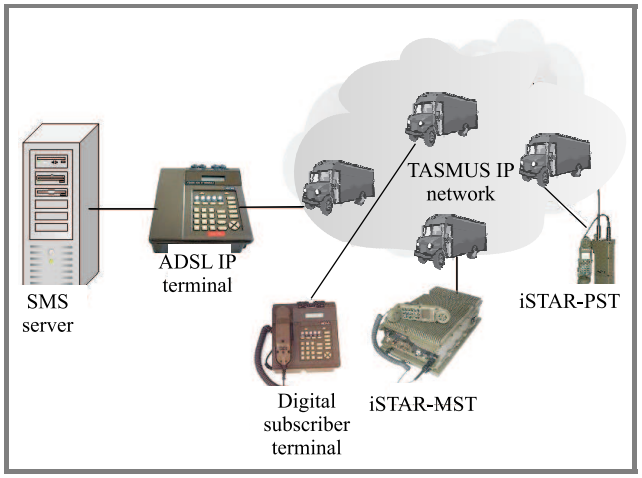


Fig. 6. TASMUS SMS service.

Video conferencing service. The video conference units (VCU) of TASMUS can be connected to the access point shelters through the switch's ISDN-BRI interfaces.

The video coding standard H.263 is used and the rate of the video session depends on the number of BRI lines used. The supported rates are 128, 256 and 384 kbit/s, corresponding to 1, 2 and 3 BRI lines. Two VCU units can establish point-to-point video sessions. A video server, which is located in the nodal point, is used to arrange the conference between more than two VCU. The video server does not join the conference; instead it decides the rate of the conference. It has a 2 Mbit/s. ISDN PRI interface with the nodal point switch. A video server can serve a video session with maximum 6 VCU. A typical application is shown in Fig. 7.

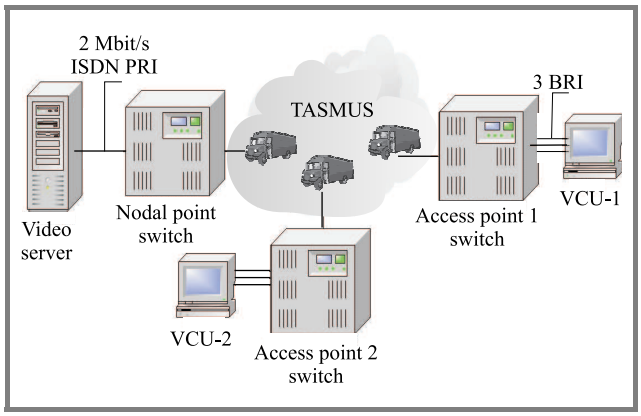


Fig. 7. TASMUS video conference service.

Fax service. Using the end-terminals digital subscriber terminals, iSTAR-MST's and iSTAR-PST's; up to 9600 bit/s. Digital fax applications can be performed in TASMUS. A typical application is shown in Fig. 8.

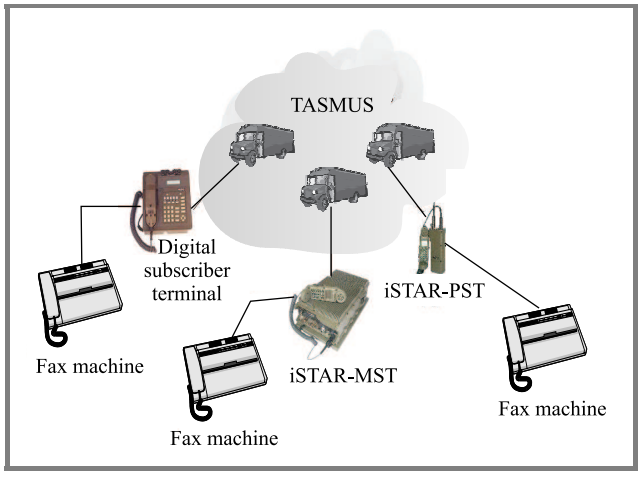


Fig. 8. TASMUS fax service.

Wireless LAN service. With the secure access point units placed in the access point shelters, TASMUS supports WLAN service to its tactical area users. The secure access point units are connected to the high capacity ADSL IP terminals, or the digital subscriber terminals in the access point. The standard IEEE 802.11b is implemented and the 2.412–2.472 GHz frequency band (the ISM band) is used. With this service, the tactical area users who are located in the 500 meter range of the access point have the TASMUS IP service. The configuration of TASMUS wireless LAN service is shown in Fig. 9.

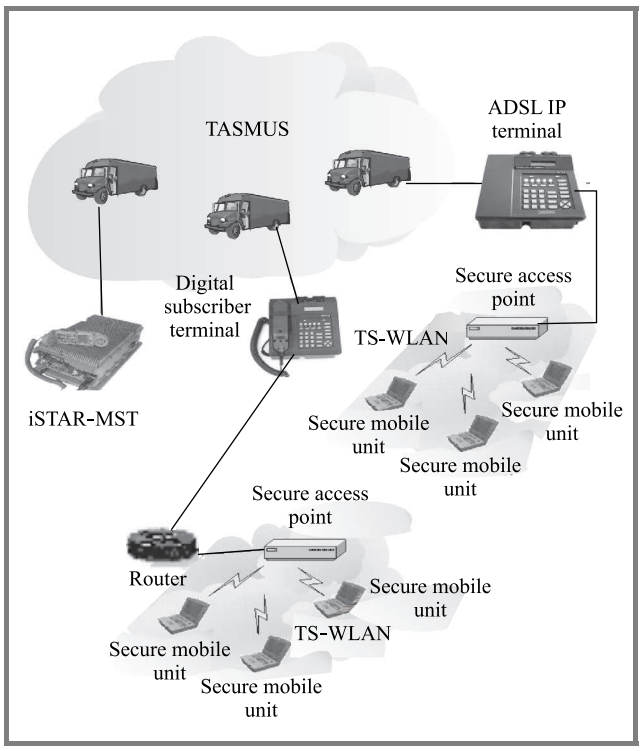


Fig. 9. TASMUS WLAN service.

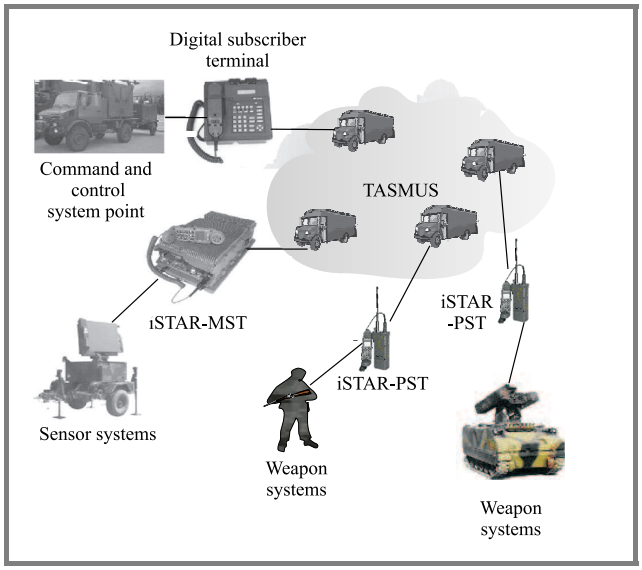


Fig. 10. Near-real time X.25 service.

Near-real time X.25 service. Near real time data communications is supported in TASMUS for the interoperability with the tactical sensor and weapon systems. The sensor and weapon systems involves exchanging target data and is their communication is intolerable to delays.

The end-terminals digital subscriber terminal, iSTAR-MST's and iSTAR-PST's are used for near-real time X.25 data communication. A typical TASMUS-sensor and weapon system integration using iSTAR radio is shown in Fig. 10.

7. Conclusion

TASMUS aims to form mobile, survivable, flexible and secure network to support all the present and future communication requirements of the tactical commanders. TASMUS also conveys the real time picture of the battlefield to the commanders. TASMUS system provides the crucial near real time data communications needed by the tactical sensor and weapon systems.
TASMUS brings together the state of the art in military communication technologies. It incorporates ATM and ISDN switching technologies used with digital ISDN terminals with built-in crypto module, simultaneous voice and data capability with synchronous, asynchronous, X.25, IP data and video interfaces. Video conferencing and digital fax are also supported over the TASMUS network. TASMUS supports all of the teleservices, bearer services and supplementary services that the TASMUS subscribers need. The communication subsystem is the base of all activities in the battlefield. No sophisticated device or system would be useful if the communication requirements are not satisfied. TASMUS, which is a mature and fielded system, will satisfy future communication needs of the 21st century C4I systems.

References

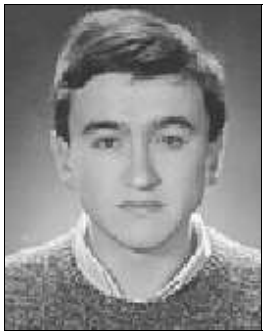
[1] F. Eken and Ş. Uzun, "iSTAR radio network for tactical use", in *NATO RTA Symp. Tact. Mob. Commun.*, Lillehammer, Norway, 1999.

[2] F. Eken and Ş. Uzun, "TASMUS, Turkish Armed Forces tactical communications system", in *AFCEA 19th Eur. Symp.*, Brno, Czech Republic, 1998.

[3] A. Akay and F. Eken, "New generation digital radio for tactical mobile communications", *Eur. Def. Technol.*, no. 27, 1996.

[4] O. Başbuğoğlu, Ş. Uzun, and F. Eken, "Packet radio for tactical communication", in *AFCEA Symp. Dig. Revol. Milit.*, 1995.

Esra Çiftçibaşi Erkan received her B.Sc. degree from Electrical and Electronics Engineering Department of the Hacettepe University, in 1998 and her M.Sc. degree from Electrical and Electronics Engineering Department of the Middle East Technical University in 2001. In May 1998, she started working in ASELSAN Inc. as a software engineer at Crypto and Information Security Group. Since May 2002 she has been working as a senior systems engineer in the Systems Engineering in Military Communications Department of ASELSAN Inc.
e-mail: eciftcibasi@aselsan.com.tr
Aselsan Inc.
PK 101 Yenimahalle-Ankara, Turkey



Şenol Uzun received his B.Sc. degree from Gaziantep Electrical and Electronics Engineering Department of the Middle East Technical University, in 1988 and his M.Sc. degree from Electrical and Electronics Engineering Department of the Middle East Technical University in 1992. In March 1988, he started working in ASELSAN Inc. as a research engineer at Telecommunications Systems Division. He has worked in Civilian Communications Department, in Engineering and Product Quality Department and in Electronics Desing Department in Aselsan in the consecutive years. He is currently the manager of the Systems Engineering in Military Communications Department of ASELSAN Inc. His current interests lie in the areas of tactical communication systems, digital mobile communications and and C4I Systems.
e-mail: SUzun@aselsan.com.tr
Aselsan Inc.
PK 101 Yenimahalle-Ankara, Turkey

Paper

Sharing tactical data in a network-enabled coalition

James Busch and Rita Russo

Abstract—The NATO Command, Control and Consultation Agency (NC3A) is a participant in a coalition project called the Shared Tactical Pictures (STP). The aim of STP is to develop methods and techniques to enable the sharing of a wide variety of information – e.g., ground surveillance sensors, airborne sensor platforms, recognized pictures, and much more – across a widely distributed network. As NATO changes its war-fighting paradigm from a well-known and stable alliance configuration to more flexible, coalition-based operations, solving the problem of information-sharing has never been more important. This paper discusses the technical and operational developments being explored in STP.

Keywords— service oriented architecture, Web services, coalition, concept of operations, shared picture.

1. Introduction

As NATO changes its war-fighting paradigm from a well-known and stable alliance configuration to more flexible, coalition-based operations, solving the problem of information-sharing has never been more important.

There is a need to take an overall architectural view in order to produce an operational doctrine for coalition operations. This will facilitate developing some common rules in the deployment of command and control (C2) systems. This doctrine will be critical in guiding the IT part of coalition deployments in the future.

At the technical level, there is a need to ensure that authorized (but not unauthorized) users are able get the data they need – where they need it, when they need it. This information must be available regardless of where it actually resides and regardless of who "owns" it.

At the tactical level, the problem becomes even more acute, as one must carefully consider the network situation: the possibility of disadvantaged (slow, low bandwidth) links, communication failures and other problems that may affect the availability and reliability of this information.

Attempting to address these issues, the NATO Command, Control and Consultation Agency (NC3A) is a participant in a coalition project called the Shared Tactical Pictures (STP). The aim of STP is to develop methods and techniques to enable the sharing of a wide variety of information – e.g., ground surveillance sensors, airborne sensor platforms, recognized pictures, and much more – across a widely distributed network. And because it is being developed for a variable-profile coalition environment, the composition of the user group (data consumers) and the set of available information sources (data providers) are not necessarily known in advance and may change quickly over time. Thus it is critical that the environment be designed to be flexible enough to allow dynamic registration of data providers and the dynamic search for assets by data consumers. There is also a need to provide a smart data-fusion capability to merge information in a reasonable way and help the users make sense of all the information that is available.

2. Operational doctrine

With the advent and the formalization of new types of alliance missions and the complexity involved in the conduct of modern military operations, new challenges are outlined for NATO. Just to mention some of the central issues: considering the wide range of possible coalition scenarios[1], the "doctrine" adopted within the specific type of coalition is essential in the definition of the nature of command relationships, both between the assigned national/multinational forces and HQs and also between HQs. The policy and rules have to be defined on a case by case basis, since forces from partner and other non-NATO nations may be invited to participate and each of them have their own internal doctrines.

The implications of these rules closely affect the operational concepts and processes involved, not only as far as the deployment of C2 systems is concerned, but also the upstream process of collection and prioritization of commander's requirements, the subsequent assignment and control of (national) assets and the nature of orders to subordinates finally generated.

As a consequence, within NATO, a process of transformation and adaptation to the new emerging scenarios has been undertaken and significant effort is being put in the direction of systems interoperability achievement. To this purpose, an architectural approach to system design, through the implementation of agreed standards and products, is followed during the development of new C3 systems, which will then undergo a rigorous interoperability testing programme. This new approach is part of the so-called *NATO C3 System Interoperability Process* [1].

In line with the above, an *overarching NATO Interoperability Policy* [2] is under development. This policy must formalize processes in support of present and future execution of a full range of NATO missions and tasks and provide guidance for the harmonization of interoperability requirements. Among these is the need for single and joint

[1]Possible types can be: joint, allied multinational, lead nation coalition, and ad hoc coalition.

service capabilities – supplied by all participants – to cooperate and, in some cases, be coordinated to provide support for the achievement of a single goal.

Considering the huge amount of different facilities, in terms of network infrastructures, communication and information systems, in use by different nations, it can be reasonably assumed that this is a very challenging task, which requires a lot of work to set the premises and the environment in which the operational actions and missions will be conducted.

The Shared Tactical Pictures (STP) initiative – of which the first phase has been the shared tactical ground picture (STGP) – is attempting to solve the problem of information sharing in a coalition environment. In the words of the STP vision statement:

"In future coalition operations, all available information that may be relevant to the production of a decision-quality tactical ground picture, irrespective of source and type, is made available to all eligible participants to provide them with actionable information consistent with their military requirement and level of command."

STP is not the development of a new system, nor does it attempt to supplant existing national systems. Rather, STP is a process that defines short-term and low-cost tasks ("quick wins") in order to develop concepts, methods and standards that will extend utilization of existing information; share data in an interoperable environment; leverage national operational picture capabilities; and enable progressive development of interoperability of data, databases, applications, systems and networks.

In the context of the STP project, an activity of architecture modelling is under way, which at first stage is being characterized by the collection of information on the systems in use or under development by the nations, in terms of *policy, process* and *product*. This does not simply mean an inventory activity, but also the examination of the state of the art as far as national facilities/capabilities are concerned, which are supposed to be used in future in an interoperable environment.

As a matter of fact, there are some reasonable issues that concur to slow down the course of this activity. One is the releasability of sensitive information by nations, in a context in which other participating nations, possibly not known in advance, can access that information. As far as *policy* is concerned, this has a very high implication in assembling the concepts of the adopted national doctrines. On the other hand, the STP community has agreed that currently existing rules and policies should constitute the basis for establishing more general high-level rules, applicable in a dynamic coalition environment.

The reluctance by nations to release information also affects the clear understanding of "what" can be shared within the coalition, i.e., national owned assets and systems along with their capabilities and products. Addressing this issue is a key goal of the STP initiative.

Under the term *process*, all the envisaged CONOPS[2] and TTP[3] are to be considered. Focus points that need to be described in detail are relationships between different level of commands, information cycles/flows between different operational nodes and the coordination of the operational staff itself. All these are key points for the effective and prompt tasking of available systems and for the provision of as much appropriate and timely support as possible to satisfy the original requirements during a mission.
An overall knowledge of the associated components, *products* and services of the existing systems taken into account are also of great importance to see how much is covered so far. Some emphasis has been placed on four different types of system products: BA[4], C2-BFT[5], ISR[6] and NC[7]. An appropriate analysis should also lead to the detection of possible gaps in the wide range of system capabilities required by the users at all LOCs[8].
Even if a number of the mentioned issues are still at the investigatory phase and will probably get no satisfactory analysis results, the ongoing activity is intended to serve as vehicle for reaching a common understanding on what a general coalition doctrine might be in order to lay the basis for building up a real concept of interoperability amongst heterogeneous environments and finally enable the achievement of a common operating picture, in order to speed-up the decision process and the course of actions.

3. Service oriented architecture

Service oriented architecture (SOA) is an architectural style whose goal is to achieve loose coupling among interacting software agents. A service is a unit of work done by a service *provider* to achieve desired end results for a service *consumer*.
Consuming a service is usually "cheaper" and more effective than doing the work ourselves. This is called "separation of concerns", and it is regarded as a principle of software engineering.
SOA achieves loose coupling among interacting software agents – which can be systems, users or devices – by employing two architectural constraints:

- A small set of simple and ubiquitous interfaces to all participating software agents. Only generic semantics are encoded at the interfaces. The interfaces should be universally available for all providers and consumers.

[2]Concept of operations, they provide the vision for users on how systems/capabilities are operated and utilized.

[3]Tactics, techniques and procedures for the operation and exploitation of assets. They are usually aimed at for commanders, staff and operators directly involved in the planning and tasking of interoperating assets, at both the operational and tactical level.

[4]Battlespace awareness.

[5]Blue force tracking.

[6]Intelligence, surveillance and reconnaissance.

[7]Net-centricity – the idea behind it is the flexible integration of command posts and decision centres, sensors and sensor systems, warfighters and commanders in a netwsork, to enable an operation.

[8]Level of commands.

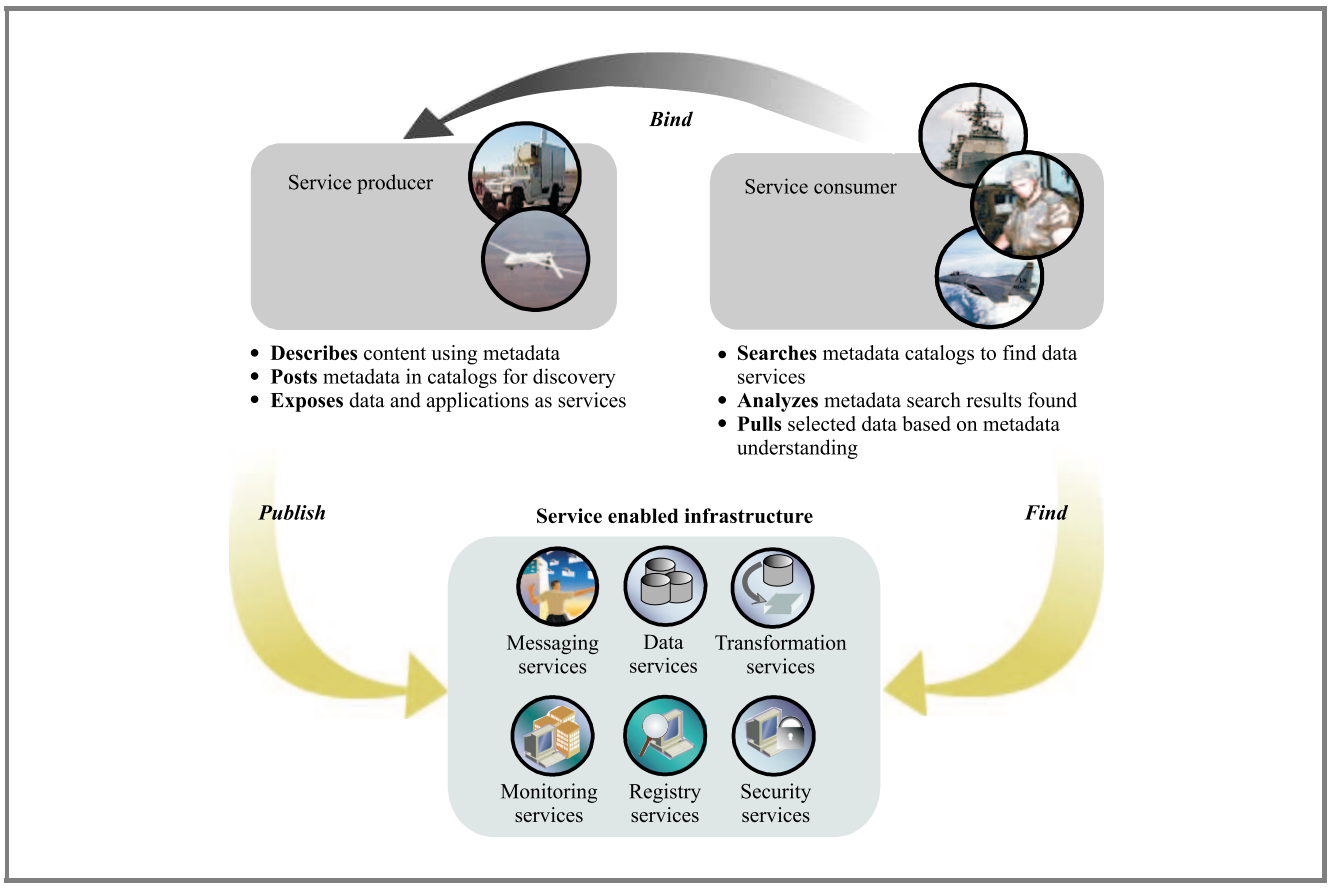


Fig. 1. Service oriented architecture (diagram courtesy of Booz Allen Hamilton [6]).

- Descriptive messages constrained by an extensible schema delivered through the interfaces. No, or only minimal, system behaviour is prescribed by messages. A schema limits the vocabulary and structure of messages. An extensible schema allows new versions of services to be introduced without breaking existing services. This schema is based on XML[9], the de facto standard language of inter-system communication.

3.1. SOA roles and operations

Any SOA contains three roles: service consumers, service providers, and a service registry (Fig. 1).

- A **service provider** is responsible for creating a service description, publishing that service description to one or more service registries, and receiving invocation messages from one or more service consumers.
- A **service consumer** is responsible for finding a service description published to one or more service registries and is responsible for using service descriptions to bind to or invoke service providers.
- A **service registry** is responsible for advertising service descriptions published to it by service providers and for allowing service consumers to search the collection of service descriptions contained within the service registry.

Each of these roles can be played by any program, software agent or network node. In some circumstances, a single software agent might fulfil multiple roles; for example, a program can be a service provider, providing a service to downstream consumers as well as a service consumer itself consuming services provided by others.
An SOA also includes three operations: *publish, find*, and *bind* (or *invoke*). These operations define the contracts between the SOA roles:

- The **publish** operation is an act of service registration or service advertisement. When a service provider publishes its service description to a service registry, it is advertising the details of service to a community of service consumers.
- The **find** operation is the logical dual of the publish operation. With the find operation, the service consumer states a search criterion, such as type of service, various other aspects of the service such as

[9]XML is the extensible markup language. An XML document is simply ASCII text that follows certain standard structural principles. XML is a "metamarkup" language. Unlike its cousin HTML – the language of Internet web pages – XML does not have a pre-defined set of tags and elements. Rather, an XML document is self-describing, allowing virtually unlimited types of content.

quality of service guarantees, and so on. The service registry matches the find criteria against its collection of published service descriptions. The result of the find operation is a list of service descriptions that match the find criteria.

- The **bind** operation embodies the relationship between the service consumer and the service provider. When the consumer attempts to invoke the publisher's service, a bind operation takes place.

The key to SOA is the service description. It is the service description that is published by the service provider to the service registry. It is the service description that is retrieved by the service consumer as a result of the find operation. It is a service description that tells the service consumer everything it needs to know in order to bind to or invoke the service provided by the service provider. (A popular analogy is the telephone book. A human customer uses the telephone book to learn how to access a business service, e.g., phone number, address; a service consumer uses a service registry to learn how to access an SOA service, e.g., location, invocation method). The service description also indicates what information (if any) is returned to the service consumer as a result of the service invocation [3, 5, 6].

4. The Shared Tactical Pictures

4.1. The STP concept

STP is all about sharing information in a coalition, without the need to develop expensive, time-consuming new systems.

A key element of STP is that it is a true multi-national project. Teams from the US, UK, Norway and NC3A have already been involved in the development process; other nations including Sweden and Italy are expected to begin contributing in 2005.

The ultimate goal of STP is to create convergence amongst coalition interoperability initiatives. To do this, STP is developing an open, scalable architecture that will enable each nation to implement its unique solution while maintaining effective interoperability. This is illustrated in Fig. 2 with some representative data sources that will be described later in this paper.

This is where the service oriented architecture concept described above comes in. An SOA implementation, as described above, is perfectly suited for a highly heterogeneous, highly dynamic environment such as a variable-profile coalition. As a result, this is the design that has been chosen for the STP initiative.

The various data sources that have been integrated into the first phase, STGP (listed below), have exposed their core functionality as services. To do this, a set of Web service "wrappers" have been written as interfaces into the underlying systems. These services can then be accessed by any service-enabled client by issuing a standard SOAP[10] request, which results in data being returned in standard XML format. The existing systems have not themselves changed at all; rather, their functionality has been made available to the STP environment by means of these service interfaces.

The initial results have been promising. Data providers representing a wide variety of information have been integrated. These information sources share one common goal: each attempts to provide some kind of "picture" of ground activity in a certain area. The systems being used for STP include:

– unattended ground sensors (UGS);

– airborne surveillance and reconnaissance (SAR) systems that produce ground tracks (e.g., JSTARS, ASTOR, U-2);

– SAR systems that produce high-resolution images;

– existing systems developed to the multinational interoperability programme (MIP) standard.

These systems contain complementary data that are stored very differently; under normal circumstances would be quite difficult or even impossible for them to interoperate. To take one simple example, the airborne (SAR) sensors produce either ground moving track indicator (GMTI) or link-16 formatted data, while the SAR cameras generate images in binary format. Clearly, sharing this information in its native form presents a huge challenge.

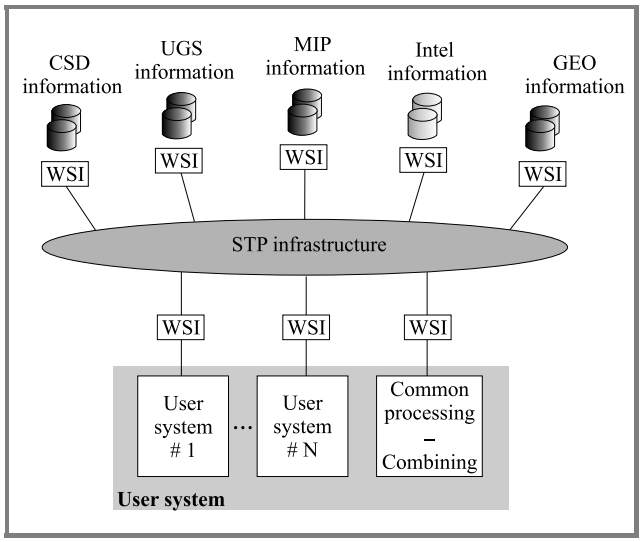


Fig. 2. The Shared Tactical Pictures.

[10]SOAP is the simple object access protocol. It is basically an "envelope" for an XML message in a SOA environment, containing routing and other header information for the message.

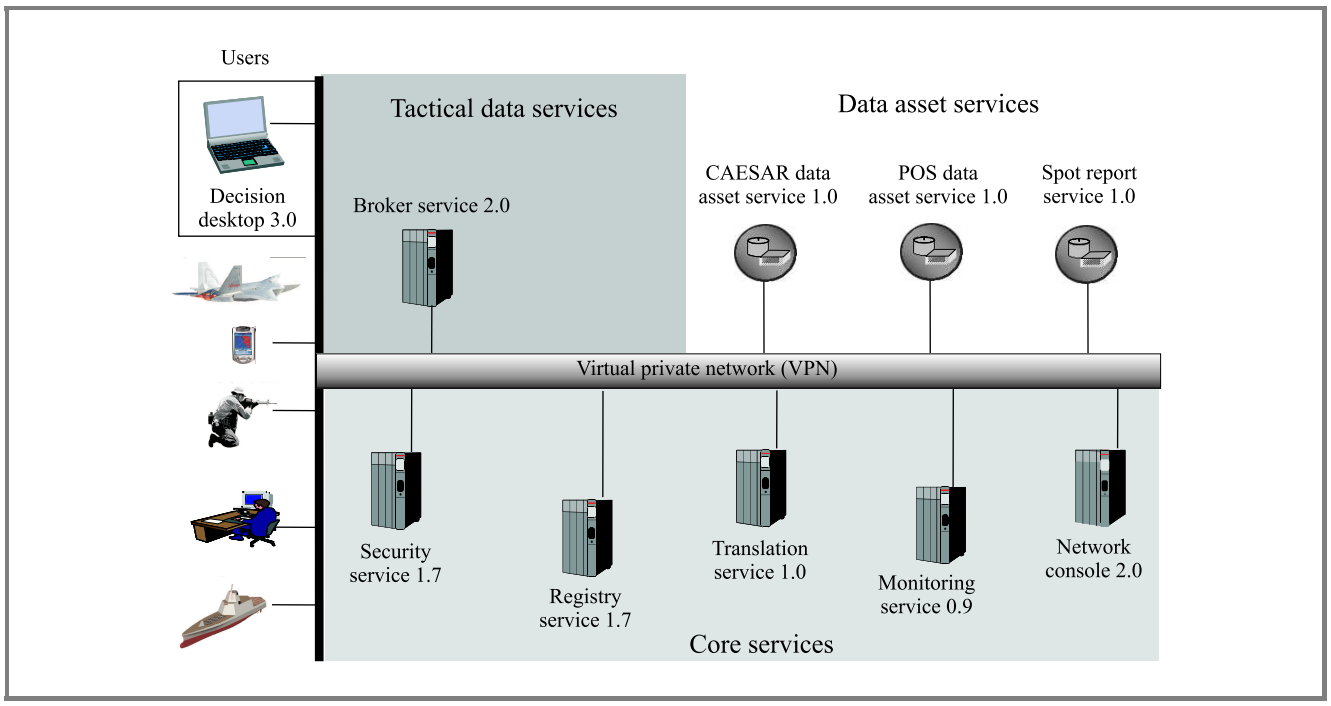


Fig. 3. The set of services developed for STP domain.

However, by developing a Web services[11] interface for each of these data sources – a "window" into their functionality – and by packaging the data as common XML, it becomes possible to share data among the disparate systems using an agreed-upon format, and it becomes possible for a user to make use of the different types of information offered by each of these platforms.

4.2. The STP Web services implementation

4.2.1. Producers and consumers

The set of services that have been developed for the STP domain is represented in Fig. 3. At the heart of the STP implementation are the so-called "Core services": services that represent some foundational functionality and which are available to all producers and consumers. The two main core services are the *registry* (based on UDDI) and *security*. (The registry, as described in the previous section, maintains knowledge of the location and access procedures for each service on the network. The security service, through the use of a public key infrastructure (PKI) issues and validates certificates to ensure secure transactions between consumer and provider.) In addition, there are a set of *translation* services and a *monitoring* service, which will be discussed later.

There is also a set of *data asset* services: providers of information to the coalition. In this case there are three primary service providers: the coalition aerial surveillance and reconnaissance (CAESAR) shared database (CSD), which aggregates ground track and imagery information coming from the various SAR platforms described above; the passive observation sensor (POS), which is the unattended ground sensor generating GMTI data; and spot reports, which give a human observer in the field the ability to enter text reports about what is being observed.

Finally, the *broker* service allows users to subscribe to data from certain sources. This will be described further later in this document.

The *consumer* for all this information can be almost anything, from a user with a web browser to a network-capable PDA to another system. For the purposes of the STP exercises, a visualization application called decision desktop (DD) has been developed, which has the ability to render all the different types of data (ground tracks, images, textual observations, etc.) being produced.

4.2.2. Dynamic data providers

When a service become available on the network – for example, when one of the airborne SAR platforms begins generating data – it communicates with the registry, providing the registry with three key items: *where it is located on the network, what services it provides*, and *how to access these services*. This is the **publish** operation described earlier, and it enables the other, data consuming services (such as the end user's system) to discover and make use of the service (the **find** operation described earlier). Finally, the consumer invokes the service (the **bind** operation) and receives the data.

[11] The term "Web services" refers to a specific instantiation of an SOA, one that is based on XML messages being transported via HTTP over TCP/IP networks. This is in fact the technology being employed by STP; therefore the terms "SOA" and "Web services" are often used interchangeably throughout this document.

4.2.3. Helper services

The SOA paradigm is largely a pull mechanism: the consumer (user) requests information from the producer and receives a response. The flaw in this approach is that it puts a burden on the consumer to keep up with the status of the producer; in other words, the only way the user will receive the latest information from a data source is if he continuously asks for it, and the only way the user will know if a new data source becomes available is if he looks for it.

This is where the *broker* service mentioned in the previous section comes into action. The broker acts on behalf of the user to check each data source for updated information; it also continually scans for new data sources of relevance to the user. For example, assume that a particular user is interested in all SAR imagery that is produced in a certain geographic area, regardless of the source, and wants it as soon as it is available. The user can set up a "subscription" with the broker service to continually poll the various data asset services and return up-to-date information as soon as it becomes available. The user thus no longer has to be concerned when services are dynamically added or removed, or when the data being offered by a producer changes, because the broker takes care of the interactions and automatically forwards relevant images to him.

A similarly valuable service is the *monitor*. In an SOA, the only way to know for sure if a registered service is indeed available is to issue a query to it; if the query fails then the service is unavailable. Clearly this is inefficient, especially if there are many users (or brokers on behalf of users) constantly doing this. Therefore, STP has developed a monitoring service that constantly evaluates the state of all services on the network. When a user (or broker on behalf of a user) wants to see which services are currently available, it merely issues a request to the monitor for the latest status.

Finally, the *translation* services give the capability to translate data from the format offered by the data provider into one preferred by the data consumer. For example, as discussed earlier the airborne SAR platforms (e.g., JSTARS) produce ground tracks in a tactical data link format called link-16. The web service that acts as the interface to these systems presents the data as an XML representation of link-16. However, the default data consumer in the STP environment (decision desktop) requires its information to be delivered in a common data specification known as resource description framework (RDF). Therefore, STP provides a link-16-to-RDF translator service, which is automatically invoked when the decision desktop consumer accesses the JSTARS provider. By the time the data reaches the consumer, it has been translated from an XML representation of link-16 to an XML representation of RDF. This same process is available for all of the other types of data, including GMTI-to-RDF, POS-to-RDF, spot report-to-RDF, and NSIF-to-JPEG[12] .

[12]NSIF is a NATO-standard format for images. As it is not widely supported by, for example, web browsers, the STP environment offers translations from NSIF to more common image formats such as JPEG.

4.2.4. Information flow

So how does it all work together?

The following very simple example will help to illustrate the process. Assume that a command officer in the field wants to see all relevant ground tracks for a particular area of interest (AOI). He wants to find any providers of this information on his network, query them for all relevant detections in his AOI, and get the data back in a form he can use. The following steps will be taken in the STP environment (Fig. 4).

1. The data provider(s) and translation services come online, and register with the registry service by each sending a SOAP message (formatted as XML) to identify what it can provide and where it is located.

2. The user starts the decision desktop visualization tool.

3. The user logs into the system, thereby providing credentials which will be validated against the various data sources. He also enters into DD the specific type of information and geographic area in which he is interested.

4. By taking the previous steps, the user creates a "subscription" with the broker service. The user also can indicate the frequency at which he wants updates.

5. The broker sends a SOAP message (XML) to the registry to find all data providers that offer the chosen type of information (ground tracks).

6. The broker sends a SOAP message (XML) to the security service verifying that this user has the rights to the data provider(s). If so, then ...

7. The broker issues SOAP (XML) requests to each of the data providers. In each case, if there is new information available, the broker receives ground tracks(s) in an XML message formatted as, for example, link-16.

8. Knowing that the user on whose behalf it is working needs data in RDF format, the broker sends a SOAP message (XML) to the registry inquiring about a translation service on the network that provides link-16-to-RDF translation.

9. Once the service is located, the broker sends a SOAP message (XML) to the translator requesting translation on the enclosed link-16 data.

10. The returned message containing ground track(s), now in XML formatted as RDF, is returned to the user's system which issued the initial query.

11. The user's system makes use of the returned message in whatever way it requires; in this case, decision desktop plots the returned ground track(s) on a map.

12. Steps 5–11 continue until the user cancels the subscription. At any time, Step 1 can be repeated (a new service provider comes onto the network); when this happens the broker "learns" about it as soon as it re-queries the registry in Step 5.

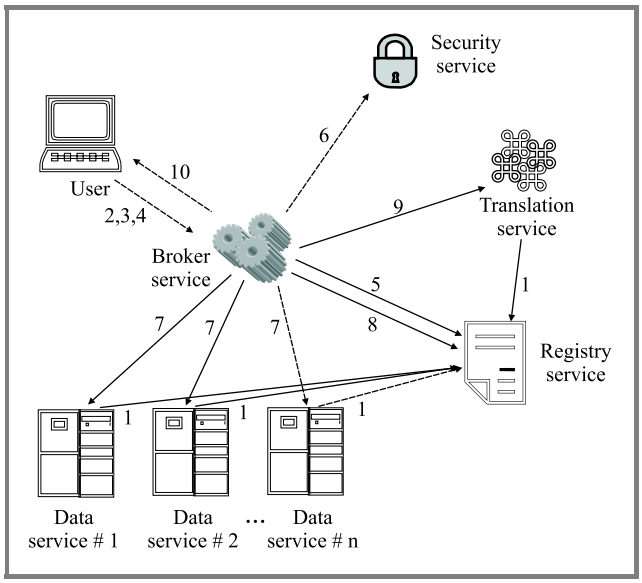


Fig. 4. The Shared Tactical Pictures environment.

These steps are illustrated in the adjacent diagram (Fig. 4). The arrows point from the initiators to the recipients of the messages.

Although this sequence of steps seems fairly simple, what's being accomplished is very powerful.

By developing a Web service interface, each of the data providers described earlier has made its information available in a common format. The flexibility of the service oriented architecture allows data sources to become available and be dynamically "discovered" by customers of that type of information. By standardizing on XML, information exchange is facilitated amongst disparate entities. Making use of the broker and translation services, the user can have information sources found and queried on his behalf, and the resultant data delivered in a format that he can use.

4.2.5. Future goals

There are some exciting additions to the STP environment in the coming months. This will include integration with some of the command and control (C2) systems taking part in the multinational interoperability programme (MIP) as they add Web services interfaces to their systems. The special requirements of deploying services to tactical users – users with low bandwidth and possibly limited viewing facilities – are being explored.

In addition, an important step will be the development of intelligent data fusion capability. It is currently possible to correlate information from a single data source. However, it would be very powerful to be able to correlate the information coming from multiple data sources, e.g., inform the user that the individual ground track being reported by sensor X is in fact the same as the ground track being reported by sensor Y. This will also be valuable in the area of blue force tracking, as different systems working together can help to identify to whom various entities belong.

Finally, there will be efforts in the future to offer a full "picture" in addition to just the ground situational awareness developed so far. This may include recognized air, maritime and environmental pictures; all Web service-enabled to take advantage of the power and flexibility of the service oriented architecture.

5. Conclusions

The shared Tactical Pictures (STP) is an important and exciting initiative in two ways.

First, it is attempting to define the policies and doctrines involved in making information available across a coalition, regardless of the source of the data.

Second, it is at the forefront of investigating the technologies of the future – service oriented architectures, XML and Web services – which will help make heterogeneous system interoperability a reality. The prototype work that has been done has already shown that ground status information can be dynamically shared from multiple, disparate systems.

The on-going work on the STP project is expected to contribute to coalition efforts for many years to come.

References

[1] "NATO C3 System Interoperability Directive", AC/322-D(2004)0001, Jan. 2004.

[2] "NATO Interoperability Policy", EAPC(NCSREPS)WP(2004)0010-REV4, June 2004.

[3] H. He, "What is service oriented architecture?", Sept. 2003, www.xml.com

[4] S. Graham, S. Simeonov, T. Boubez *et al.*, *Building Web Services with Java*. SAMS Publ., 2002.

[5] N. Aarup, J. Busch, B. Christiansen *et al.*, "Inter-systems exchange mechanism, NC3A view", NC3A Techn. Note, Aug. 2004.

[6] Booz Allen Hamilton, "Service oriented architecture concepts", presentation by E. Yuan in *STGP Techn. Meet.*, Malvern, UK, Sept. 2003.

James Busch is a Principal Scientist at the NATO C3 Agency (NC3A). He came to NC3A in 2002 from the IT Industry, where he spent fifteen years working on distributed computing and enterprise application integration projects for a variety of industries. He is currently involved in several initiatives for NC3A, including the architecture of rapidly deployable information systems modules and the exploration of new technologies such as Web services.
e-mail: James.Busch@nc3a.nato.int
NATO Command, Control and Consultation Agency
P.O. Box 174, 2501 CD The Hague
The Netherlands

Rita Russo is a Scientist at the NATO C3 Agency (NC3A). She has six years of experience in Telecommunications and IT Industry, during which she was involved in the development of system specifications and acquisition of standards' directives for 2G and 3G Mobile Radio systems and networks. Since joining the Agency in 2003, she has become involved in several initiatives related to architectural design within coalition environments and the information security area.
e-mail: Rita.Russo@nc3a.nato.int
NATO Command, Control and Consultation Agency
P.O. Box 174, 2501 CD The Hague
The Netherlands

Paper

Introduction of the Network Centric Warfare concept to Czech Armed Forces

Pavel Eichelmann and Luděk Lukáš

Abstract—One of the features determining the strength of troops is command and control. The quality of command and control is determined by the quality of the command and control systems (C2S). The Czech Armed Forces are developing tactical command and control systems (TC2S) for the Ground and Air Forces. The TC2S were a little bit separately developed. Now, we want to use the benefit of shared situation awareness. The concept of Network Centric Warfare is a solution of this problem. The integration of both TC2S's is its main objective. To apply this concept, separate C2S's of different units are integrated into one logical system, into one joint C2S. Thus it is possible to create the common operation picture for all units of the task force.

Keywords—Network Centric Warfare, command and control, tactical command and control system, situation awareness.

1. Theoretical background of Network Centric Warfare

The Czech Armed Forces are under the process of reform; they reduce their structure to become smaller, modern and mobile. The modern armed forces have developed intensively. They change their mission, capabilities and structure according to assigned political tasks. There is a growth of their dynamic abilities, swing operation, range and number of fulfilled activities. The above-mentioned complicated problems need new manners to solve. One of the possible manners is to improve the command and control process. It is directed into improvement of situation evaluation, commander's intent formulation and task assignment to subordinated units. It is important for each element of battle formation to know its task and its contribution to fulfilment of the whole mission.

To solve this problem, the command and control system (C2S) is improved according to Network Centric Warfare (NCW) concept. The Network Centric Warfare is a concept that is based on integration of separate C2S's of different units into one logical system, into one common C2S. Thus, it is possible to create the common operation picture for all units of the task force.

NCW focuses on combat power that can be generated from effective linking or networking the engaged military troops. It is characterized by the ability of the geographically dispersed forces to create high-level shared battle space awareness that can be exploited via self-synchronization and other network centric operations to achieve commander's intent.

NCW concerns about networking rather than networks. It concerns increased combat power that can be generated by a network centric force. The power of NCW is derived from effective linking or networking knowledgeable entities that are geographically or hierarchically dispersed. NCW recognizes the centrality of information and is potentially a source of power. Networking of entities enables them to share information and to collaborate to develop shared awareness, and also to collaborate with one another to achieve certain degree of self-synchronization. The net result is the increased combat power.

NCW provides opportunities to improve both command and control (C2) and execution at each echelon. NCW offers the opportunity to not only be able to develop and execute highly synchronized operations, but also to explore C2 approaches based upon horizontal coordination or self-synchronization of battle entities.

NCW is based on using new abilities of the information and communication technologies (ICT). There is a new view on strength of information. By sharing information we can improve our situation awareness and understanding. New ICT has the potential to improve this process.

Information is compared to glue that bonds organization into one unit stuck together to accomplish the object function. Higher dynamic and swing of operation needs new quality of the glue. Correct information flows provide activity synchronization, information sharing, situation awareness and understanding.

The commander and staff who understand to battle situation can better formulate the battle objective and direction of main effort. Like view of the chess-board is important for a chess-player, knowledge of friendly troops and enemy situation, terrain and weather conditions is important for a commander. The data collection was limited in the past. This status of uncertainty was named "fog of war".

Nowadays we have new better communication and information systems (CIS) and this restriction was reduced. We can collect position and status data in real time. The information system reduces "fog of war". It provides situation awareness that is same like real battle situation. The common shared knowledge of situation is important for success of task force. The Air Force and Ground Forces need common operation picture that is created as an intersection of their separate operation pictures. Common operation

picture improves objective formulation, task assignment, activity synchronization of the task force.

NCW is based on robust network environment. The separate C2S's are integrated into one logical system. The sensors, decision makers and action elements are interconnected together. Technological borders and bottlenecks of information exchange are reduced. Each force element has access to all needed information. Provided information creates a real situation picture. There is a common database that maintains data for all command levels. All command levels have the same situation picture.

2. NCW importance for Czech Armed Forces

The Czech Armed Forces battle capability has been improved by introduction and usage of new C2S and CIS. The Ground Forces' Tactical Command and Control System (GF-TCCS) has been developed and used by the Ground Forces. It is predetermined for the division or lower levels. The Air Forces' Tactical Command and Control System (AF-TCCS) has been developed and used by Air Forces, too. Developing these C2S's have followed the battlefield digitisation concept. However, in the development phase these TC2S were a little bit separated. To apply NCW concept, we would like to integrate GF-TCCS and AF-TCCS into one logical system (Fig. 1). Thus, our Joint Forces will have a C2S of new quality. We will create better potential for joint actions of the Ground and Air Forces.

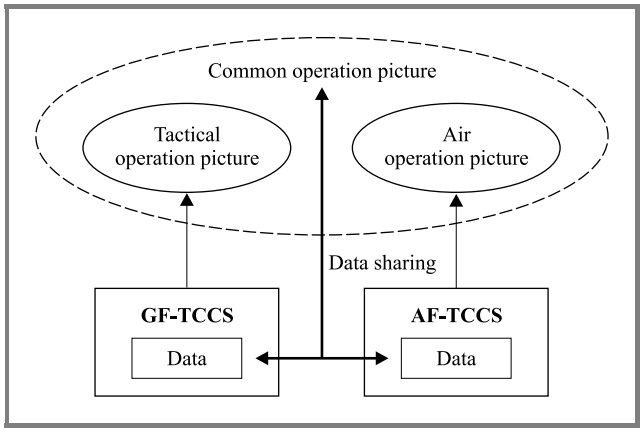


Fig. 1. Common operation picture based on integration of GF-TCCS and AF-TCCS.

It is important to improve our joint doctrine and procedures based on NCW. We have to realize new possibilities and to design new tactics and procedures. These tactics and procedures will be tested and improved to be a part of the new doctrine. We improve abilities of formulating battle objective, of synchronising activities and fulfilling tasks. Thus NCW will have influence on mission capability package of the Czech Armed Forces.

3. NCW program description

It should be noted that the Czech Armed Forces now have no real concept detailing the NCW program. Several partial studies have been made on this subject with analyses of those topics. The Military Technical Institute of Electronics and Military Technical Institute of Air Forces, both from Prague, have been the leaders in solution of this problematic area.

There is no doubt that each military organization improving its own performance must, sooner or later, implement the NCW principles. The notions like information superiority, knowledge superiority, decision superiority, sharing information, common operation picture, etc., aren't only phrases but they must be implemented in C2S to improve C2. Last military conflicts have reflected that achievement of these funds makes the operations much more effective.

We know that NCW implementation will have expressive impact to the Czech Armed Forces. This concept causes to force a broad organizational measures and C2S system architecture changes. Each of these changes is important and crucial, however C2S system architecture and suitable ICT implementation is important for the first phase of the NCW concept.

We will study and analyse good foreign experience in this area. There is a good example of Sweden, Norway and Finland cooperation. Together they have provided an analysis of their current C2S's and their networking. They received what they have and what not. Important result was what they would have to complete to obtain the robust NCW C2S.

At the beginning, the Czech Armed Forces will necessitate carrying out a thorough analysis of NCW abilities. Thus we will locate the positive part of next effort. It isn't possible to invest an astronomical sum and purchase technology of all sorts, but rather to choose the fundamental pillar of future architecture. Thus, the system can be realized through gradual evolutional steps. This progress has two benefits. Financial area is the first. We can accomplish correct technology acquisitions. Personnel area is the second. The NCW C2S will have new capabilities. Its using will need a change of users thinking to take advantage of NCW capabilities. Our commanders and operational personnel have to realize step by step these capabilities for C2 improvement.

Our military experts will contribute to the NATO NCW program to receive state-of-art knowledge and experience. Other cooperation will be done on commercial ICT firm level. Therefore, we prepare and constitute our solution teams that will be capable to design and develop the NCW C2S.

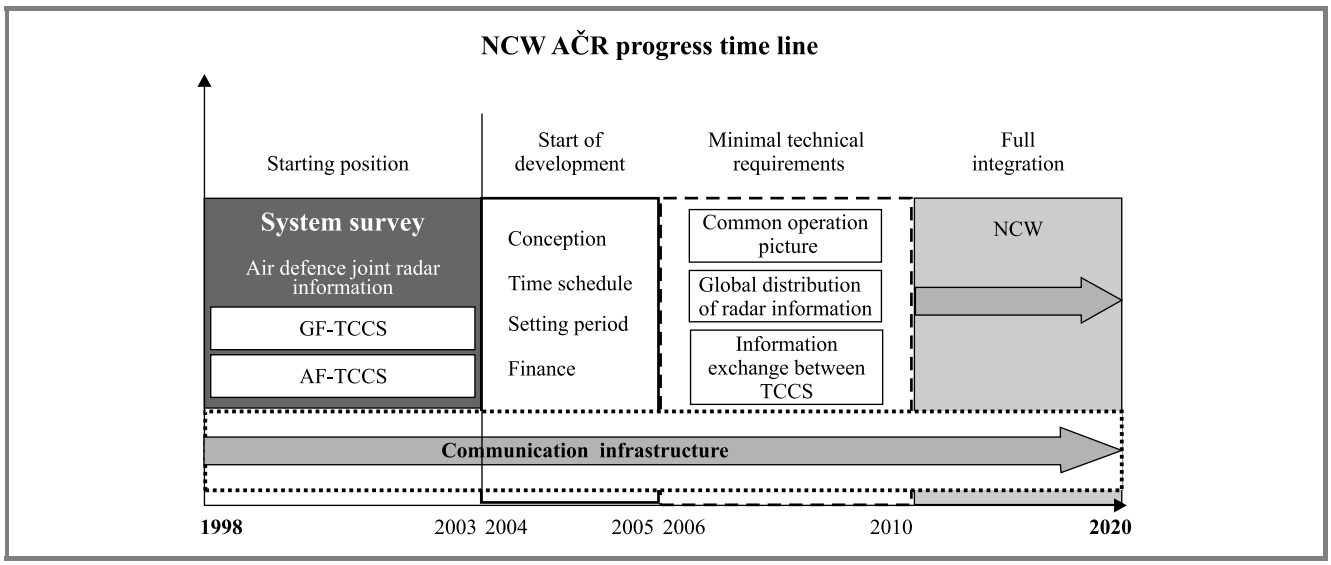


Fig. 2. The time line of the Czech Armed Forces NCW program.

To successfully solve the NCW problems, it will be necessary to make up conceptual study that would cover all the important issues of above NCW. The analyses of GF-TCCS and AF-TCCS will create one of the main parts of this conceptual study. We have to analyse all the important IT problems and find the solution for integration of AF-TCCS and GF-TCCS. The integration of TCCS and Stationary Information System will be next step. It will be done by integration of the MoD Overarching Information System and TCCS. Thus we will have military information environment for seamless information sharing and exchange. Figure 2 depicts a time line of Czech Armed Forces NCW program.

The key is in using NATO standards, MIP for example. IT standards used for development of other Czech Armed Forces Information Systems (Staff Information System for example) should also be concerned so that necessary level of information sharing is provided.

At first, GF-TCCS and AF-TCCS integration will be done on brigade level. We assume information sharing and creation of common operation picture. The fundamental set of formalized messages will be defined that will mediate information between GF-TCCS and AF-TCCS. The chosen technology will be demonstrated and verified. Thus we can pick suitable HW and SW platforms and technologies. We hope that we can exchange the basic set of formalized messages at first. This is important for information sharing. Creation of a simplified version of common operation picture will be the next step. Thus we can provide for common sharing of situation information through all command levels and authorized users.

We assume that the main integration problem will be in the area of information security. Each of the information systems has its own security rules. The National Security Agency plays an important role here because it is the certification authority in this area. We will respect NATO security standard too. Information security is concerned as a very sensitive area and it is necessary to look for global solution suitable for all NCW C2S.

It is important to determine several crucial issues that will have the key role in our process of C2S integration. The following are some of these problems:

- NCW capability analyses and suitable selection of proper function;
- methodology assessment to provide for an information system integration;
- C2S system architecture change;
- suitable SW and HW technology selection to provide for function and data integration;
- technology implementation providing information exchange;
- common operation picture creation;
- security proposal and realization of integrated systems;
- migration of the already established and operated information systems into the final NCW shape;
- modernization of the action elements, namely to accomplish the NCW requirements;
- to design a model solution of internetworking separate C2S to optimise effectiveness of the whole NCW C2S.

The set of above mentioned problems reveals that this project won't be short-term, but it will be strategic. We can say that it is the gradual reconstruction of the modern

armiesthat shall enable variety of missions. We hope that the Czech Armed Forces will be among these armies.

4. NCW technological groundwork

The Czech Armed Forces have already modernized several years. Our forces are equipped with the weapon systems and CIS coming from 1970s. We have problems with our off road vehicles, light trucks, armoured vehicles, aviation and tactical CIS.

The situation seems unsatisfactory for NCW implementation. However contrary is the true. We can prepare all process of NCW implementation utilising our coalition partner experience. We can prepare the research background of NCW, choose the suitable tactic of NCW implementation, prepare appropriate projects. We have time for thinking about NCW in the Czech Armed Forces. But there is short time for thinking.

There is a good position for NCW in the area of stationary communication system. Our microwave radio relay network provides the data transmission circuits for information systems and computer networks. The data networks are made of CISCO products. It is possible to say that we have homogeneous environment for stationary data transmission. The situation was convenient in the area of mobile tactical communication too. For tactical needs we have developed and fielded new tactical communication system TAKOM, which was presented on this conference last year. It is a digital ISDN communication system that uses 2 Mbit/s channels. CISCO products like in the stationary systems provide data transmission.

But we have faced a problem to provide data transmission for mobile users by Tactical Radio System. It is composed of VHF and HF radios. This system can provide data transmission or voice traffic but not simultaneous data and voice transmission. The essential disadvantage of this solution is in preemption of data transmission by voice traffic. Situation awareness message delay may be tens of seconds or minutes.

The above-mentioned solution is based on using Czech VHF radios RF-13. However there are some considerations about using software-defined radios for Tactical Radio System. Our Military Technical Institute of Electronic, Prague tested Rohde & Schwarz M3TR and Harris Falcon. The question of TRS implementation is open. Thus we have to solve problem to provide simultaneous voice and data traffic. It is an imperative request for NCW.

To determine fitness for NCW we will have to test communication and information technology features. There are many possibilities but we have to choose perspective technologies able of providing for interoperability and simplicity of our solutions.

Computer networksfielded in the Czech Armed Forces are largely based on INTEL (WINDOWS OS) or SPARC (SOLARIS OS) technologies. There is continual exchange of HW and SW platforms to satisfy information system requirements.

The Air Force use several computer based information systems namely BOIS and SEKTOR-VS. The basic objective of these systems is command and control support. These systems are produced by Czech firms. Similar situation is in the Ground Forces. They use Battle Vehicle Information System and Movement Control System to provide situation awareness and message transmission. The process of broad usage of informatics within military structure started seven years ago.

Evidently better is the situation in aviation because the Czech government has approved leasing of JAS 39 GRIPEN supersonic aircraft. It is equipped with state-of-art CIS technology.

For object identification we use sensors of multiple types. In the area of active radars we use old Soviet and Czechoslovak products of various age. Mainly the Soviet radars need upgrading especially for digital processing of radar information and largely implementation of IFF. The passive surveillance systems are represented by the older type (TAMARA) and the new type (VERA). For both of them it is necessary to solve the problem of data fusion to get common air picture.

5. Czech Armed Forces NCW status

Brigadier general Jiri Baloun, chief of J-6/General Staff, has presented his vision of the Czech Armed Force NCW [6]. He considered NCW was along with technology also doctrine, information superiority, new style of command and control and military thinking. The Czech Armed Forces have elaborated several fundamental studies that will present the solution of our NCW problems. We will cooperate in the appropriate NATO working group to solve these problems. The elementary problem is providing for CIS interoperability and data sharing to ensure the situation awareness. The problem is simple to define but difficult to solve. For solution we need a good vision, coalition technology experience and standards, plan and time to ensure evolution and integration of systems into one, into information age environment.

The NCW C2S is potential for better C2. We can benefit only from employing that potential. The solution is beyond signal corps responsibility taking all the Czech Armed Forces. It is an issue of new style of command and control. The change of thinking is extremely long route with many obstacles. But we have to finish this journey to have better Armed Forces.

6. Conclusion

The Czech Armed Forces are under the process of reform, reduction of structure to be smaller, modern and mobile. The key point is the growth of their dynamic abilities,

swing operation, range and number of fulfilled activities. One of possible manners to fulfil the above-mentioned requests is improvement of the command and control process. We can solve this problem to improve the command and control system according to Network Centric Warfare concept. The network centric warfare is a concept that is based on integration of separate digital C2S's of different units into one logical system, into one common C2S, exploiting the ICT benefit. NCW offers the opportunity to not only be able to develop and execute highly synchronized operations, but also to explore C2 approaches based upon horizontal coordination or self-synchronization of battle entities.

We know that NCW implementation will have an immense impact on the Czech Armed Forces. Our military experts will cooperate on NATO NCW program to receive state-of-art knowledge and experience. One of the major integration problems will be in the area of information security.

The NCW C2S is potential for better C2. We can benefit only from employing that potential. It is an issue of new style of command and control. The change of thinking is extremely long route with many obstacles. But we have to finish this journey to have better Armed Forces.

References

[1] D. S. Alberts and J. Gartska, "Understanding Information Age Warfare", in *CCRP*, Washington, USA, 2001, p. 312.

[2] J. V. Mc Gee and L. Prusak, *Managing Information Strategically*. Wiley, 1993.

[3] M. Snajder and P. Zadina, "Technological support of NCW, project OPER-SÍŤ", in *VTUE*, Praha, Czech Republic, 2003, pp. 48–95.

[4] L. Lukáš and P. Hruza, "The concept of C2 communication and information support", in *CCRTS*, San Diego, USA, 2004, p. 7.

[5] L. Lukáš, P. Tomecek, and P. Hruza, "The synergic integrated concepts of C2S", in *ICCRTS*, Copenhagen, Denmark, 2004, p. 7.

[6] J. Baloun, "Network enabled capabilities", in *C2 Conf.*, Brno, Czech Republic, 2004.

Pavel Eichelmann currently holds the post of research officer of TCCS department of Information Technology Development Agency Prague. In 1998 he graduated from Military Academy (MA) in Brno. After graduation from the MA he fulfilled his regimental duties in the signal brigade as the commander of signal company. Since 2001 he works like researcher in area of Air Forces CIS. His research and publication activities are directed to the field of TCCS, Air Forces Information Systems and computer networks.

e-mail: pavel.eichelmann@army.cz
Information Technology Development Agency
Pod vodovodem st 2
158 00 Prague, Czech Republic

Luděk Lukáš currently holds the post of chief of COMSYS section of the Military Technology Faculty of University of Defence in Brno. In 1981 he graduated from Military Technical University in Liptovsky Mikulas (Slovakia). After graduation from the MTU he fulfilled his regimental duties in the signal brigade as the commander of troposcatter company and chief of staff of signal battalion and commander of Army signal center. Since 1991 as Military Academy personal he gradually held appointments as lecturer. He is a member of national board for interoperability and the technical program committee of RCMCIS conference in Poland. His research and publication activities are directed to the field of C2 communication and information support, C4I systems and tactical communication system.

e-mail: ludek.lukas@vabo.cz
University of Defence/K-302
Kounicova st 65
612 00 Brno, Czech Republic

Paper

Cryptology Laboratory – its quality system and technical competence according to the ISO/IEC 17025 standard

Robert Wicik

Abstract—A laboratory is an organization which operates a quality system, has technical competence, generates valid results and its quality system and technical competence are conformed and recognized. A cryptology laboratory operates in information technology security area, where cryptographic methods of information protection play main role. Appropriate confidence, correctness and effectiveness of security services is needed and may be achieved trough development, evaluation, accreditation and certification processes performed by competent and commonly recognized organizations like: laboratories, certification and accreditation bodies. We describe in the paper the accreditation and certification structure and the IT security framework and also the role of the cryptology laboratory in this structure and framework.

Keywords*— *cryptology laboratory, certification, accreditation, quality system, ISO/IEC 17025 standard.

1. Introduction

A laboratory is an organization [1] which:

- operates a quality system – according to the ISO/IEC 17025[1] international standard;
- is technically competent – has competent personnel, sufficient equipment and appropriate test methods and procedures;
- is able to generate technically valid results;
- has conformed and recognized competence (for example) by government bodies.

A cryptology laboratory is a specific kind of laboratory which operates in information technology (IT) security area, where cryptographic methods of information protection play main role. Each IT system or a product has its own requirements [3] for maintenance of confidentiality, integrity and availability and implement a number of technical security enforcing functions to meet these requirements. Appropriate confidence, correctness and effectiveness of these functions are needed and may be achieved through development, evaluation, accreditation and certification processes performed by competent and commonly recognized organizations: laboratories, certifications bodies and accreditation bodies.

[1] "Ogólne wymagania dotyczące kompetencji laboratoriów badawczych i wzorcujących", PN-EN ISO/IEC 17025, 2001 (in Polish).

The main area of activity of the Cryptology Laboratory is testing, analyzing and evaluating of cryptographic:

- systems and devices;
- transformations (ciphers, integrity functions, generators, etc.);
- protocols;
- other crypto mechanisms.

Results of these tests, analysis and evaluations are used in design and certification processes of:

- secure IT products;
- products that bring security features to general systems
- or in design and accreditation of secure IT systems.

In this paper, we explain evaluation processes and evaluation criteria of secure IT products and systems during the certification of type and the certification of conformity. We show the role in certification processes of the Cryptology Laboratory and we emphasize importance of its quality system and the set of technical procedures.

2. Product certification and system accreditation

Information technology security [4] means:

- confidentiality – prevention of the unauthorized disclosure of information;
- integrity – prevention of the unauthorized modification of information;
- availability – prevention of the unauthorized withholding of information or resources.

An IT system or product has its own requirements for maintenance security services and it implements a number of technical security enforcing functions to meet these requirements. Appropriate confidence, correctness and effec-

tiveness of these functions is needed and may be achieved through development, evaluation, accreditation and certification processes performed by competent and commonly recognized organizations like: laboratories, certification and accreditation bodies (see Fig. 1).

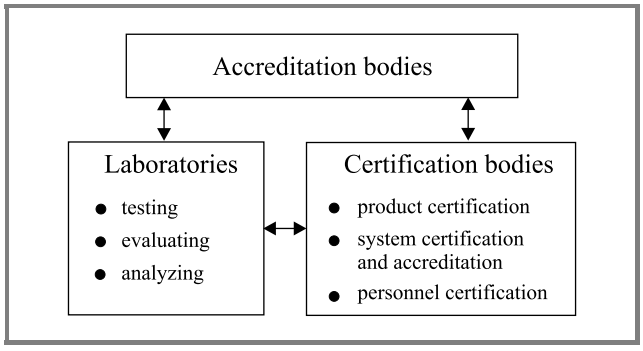


Fig. 1. Accreditation and certification structure.

Accreditation bodies recognize and verify competence of laboratories and certification bodies. It is a national organization responsible for accrediting laboratories and certification bodies according to the standards, for example to the ISO/IEC 17025 standard or another requirements. Accreditation is a procedure by which an authoritative accreditation body gives formal recognition that the laboratory or certification body is competent to carry out specific conformity assessment tasks.
Certification bodies perform a product and system certification, a system accreditation and a personnel certification. A certification body is a national organization, often the National Security Authority, responsible for administering evaluations of products and systems within that country. The certification body issue certificates (certification reports) – public documents, which are formal statements confirming the results of the evaluation and that the evaluation criteria, methods and procedures were correctly applied.
Laboratories perform testing also analyzing and evaluating of devices and systems. Accredited laboratories issue evaluation technical reports which are submitted to the certification body detailing the findings of an evaluation and forming the basis of the certification.

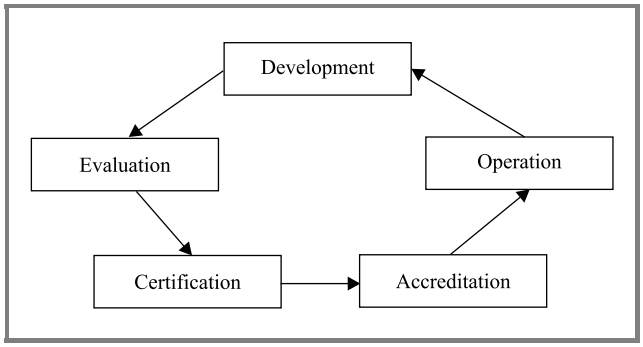


Fig. 2. IT security framework.

IT security framework covers [2]: development, evaluation, certification, accreditation and operation of a security device or system (as in Fig. 2).
In the development process an IT system or product is built.
In the evaluation process it is assessed against defined security evaluation criteria.
In the certification process it is confirmed that the results of an evaluation are valid and the evaluation criteria have been applied correctly.
In the system accreditation process it will be confirmed that the use of an IT system is acceptable within a particular environment and for a particular purpose.
In the secure operation process an accredited system is operated according to approved procedures.

3. General requirements for testing laboratories

A testing laboratory should meet requirements according to the ISO/IEC 17025 standard issued in 1999, which covers: "General requirements for the competence of testing and calibration laboratories". The standard [1] includes:

– management requirements concerning quality system and documentation;

– technical requirements concerning personnel, test methods and equipment.

This standard is applicable to all organizations performing test also laboratories where testing forms part of product certification. It is for use in developing quality, administrative and technical systems. Laboratory clients, regulatory authorities and accreditation bodies may also use it in confirming or recognizing the competence of laboratories.
Management requirements [1] for a laboratory include:

1. Organization – the description of the laboratory organization and the identification of potential conflicts of interest, when a laboratory is a part of larger organization.

2. Quality system – introduces some specific directions what must be in a quality policy statement to comply with ISO/IEC 17025 standard.

3. Document control – demonstrates how laboratory's documents are issued, identified, changed and approved.

4. Review of requests, tenders and contracts – it is to resolve any differences between the request or tender and the contract before any work commences. Each contract shall be acceptable both to the laboratory and the client.

5. Subcontracting of tests and calibrations – when the laboratory subcontracts work, this work shall be placed with a competent subcontractor (complies with ISO/IEC 17025 standard).

6. Purchasing services and supplies – the laboratory should have procedures for the selection purchasing of services and supplies that affect the quality of the tests.

7. Service to the client – describes laboratory's cooperation with the client and principles of a client's access to relevant areas of the laboratory's work.

8. Complaints – the laboratory should have procedures for the resolution of complaints received from clients.

9. Control of nonconforming testing work – there are specific requirements for dealing with nonconforming testing results and reference to corrective action in such cases.

10. Corrective action – there are specific procedures defined for cause analysis, selection and implementation of corrective action, subsequent monitoring and follow-up audits.

11. Preventive action – it should be undertaken, if improvements and potential sources of nonconformances, either technical or concerning the quality system are identified.

12. Control of records – quality and technical records from internal audits and management reviews as well as records of corrective and preventive actions.

13. Internal audits – internal auditors periodically conduct internal audits of activities of the laboratory to verify its quality system and the testing activities.

14. Management reviews – the laboratory's executive management periodically conduct a review of the laboratory's quality system and testing activities to ensure their continuing suitability and effectiveness, and to introduce necessary changes or improvements.

Technical requirements for a laboratory include [1]:

1. General – describes which factors determine the correctness and reliability of the tests performed by a laboratory.

2. Personnel – the laboratory should have competent personnel as well as plans and procedures of education and training.

3. Accommodation and environmental condition – laboratory facilities and environmental conditions should help testing and do not adversely affect the required quality of tests.

4. Test methods and method validation – the laboratory uses methods and procedures for testing which should be validated if they are not standardized.

5. Equipment – the laboratory shall possess sampling, measurement and test equipment which should be regularly calibrated or checked.

6. Measurement traceability – it is traceability to the International System of Units (SI). It is possible and desirable in some areas and not in others.

7. Sampling – it is a procedure whereby a part of a substance, material or product is taken to provide for testing.

8. Handling of test items – the laboratory shall have procedures for the transportation, receipt, handling, protection, storage and retention of test items to protect the interests of the laboratory and the client.

9. Assuring quality of test results – the laboratory shall have quality control procedures for monitoring the validity of tests undertaken.

10. Reporting the results – the results of each test carried out by the laboratory shall be reported in the test report accurately, clearly, unambiguously and objectively and shall include all the information requested by the client and necessary for the interpretation of the test results.

Many factors determine the correctness and reliability of the tests performed by a laboratory. These factors include contributions from: human factors, accommodation and environmental conditions, test methods and method validation, equipment, measurement traceability, sampling and handling of test items. The laboratory shall take account of these factors in developing test methods and procedures, in the training and qualification of personnel and the selection of the equipment it uses.

4. Documentation of a laboratory

There are quality and technical documentations maintained in the laboratory. Laboratory's documentation should:

- be prepared according to the requirements included in the ISO/IEC 17025 standard;
- strict correspond with the real laboratory's activities;
- be updated, approved, legible, and accessible for the laboratory's personnel.

Documentation of the laboratory should include:

- The Quality Manual,
- The Management Procedures Manual,
- The Technical Procedures Manual,
- Test Instructions,
- other documents and records describing every important properties and activities of the laboratory.

Basic laboratory's document is "The Quality Manual". In the quality manual there are defined the laboratory's quality system policies and objectives. The quality manual outlines the structure of the documentation used in the quality system of the laboratory. The quality manual includes or makes references to the supporting management procedures and technical procedures. The roles and responsibilities of quality and technical management as well as the rest of the personnel are defined in the quality manual.
"The Management Procedures Manual" is created if such procedures are not included in the quality manual. This manual covers supporting procedures of managing quality system and not covers technical (sampling and testing) procedures. Some aspects and data of laboratory's quality system can be covered by other documents, for example: rights and responsibilities of the laboratory's personnel and the organization and management structure.
"The Technical Procedures Manual" consists of laboratory-developed procedures which describe non-standard methods, laboratory-designed/developed methods, amplifications and modifications of standard methods. Developed methods and procedures should have been validated before use. Validation is the confirmation by examination and the provision of objective evidence that the particular requirements for a specific intended use are fulfilled. Each developed procedure should include records with results obtained form validation process.
"The Test Instructions" are supplements to the technical procedures. Test instructions are issued, if technical procedures insufficiently describe testing methods. A test instruction describes in details each steps to perform a technical procedure, such as: preparing samples, checking equipment, preparing workspace, performing tests, recording observations and results, other details needed to perform testing procedure.
Other documents and records maintained in the laboratory:

- definitions and terminology;
- a organization scheme of the laboratory;
- personal cards (function descriptions, duties, responsibilities);
- training (annual plans and programs of trainings);
- client complaints (register of complains, laboratory's judgments);
- approved suppliers (cards, problems, reviews);
- measuring equipment (register of equipment and their service, repairs, checks and calibrations);
- audits (schedule, programmes and reports);
- internal problems and correction actions;
- projects (proposals, realizations, approvals and introductions of documents of quality and testing system).

5. Test methods and procedures of the Cryptology Laboratory

The main area of activity of the Cryptology Laboratory is testing, analyzing and evaluating of cryptographic:

- systems and devices;
- transformations (ciphers, integrity functions, generators, etc.);
- protocols and other crypto mechanisms.

Results of these tests, analysis and evaluations are used in design and certification processes of:

- secure information technology products,
- products that bring security features to general systems;

or in design and accreditation of:

- secure information technology systems.

In the cryptology area, there are dominating non standard methods, which base on the common recognized evaluation criteria of ciphers, devices and systems. There are well known (or not) methods for determining crypto security parameters and features, which can by applicable as laboratory developed methods and procedures.
The cryptology laboratory can perform methods and procedures for:

- type certification;
- certification of conformity for objects or systems;
- evaluation of parameters for ciphers and other crypto transformations, binary sequences, etc.

Type certification bases on the national law and regulations and on the established criteria, for example on the Information Technology Security Evaluation Criteria – ITSEC [2] or Common Criteria for Information Technology Security Evaluation – CC [3]. Methods and procedures for type certification describes major actions taken during this process. Such methods and procedures can cover evaluations of crypto transformations and mechanisms used in an assessed object.
Technical procedures for type certification can cover procedures such as:

- Inspection of the documentation – includes methods for determining whether the documentation of evaluated object is complete, precisely and exhaustively describes all aspects, which are essential for certification and for the goals, which was put in.
- Checking correctness of implementation – includes methods for determining whether the security enforcing functions are correctly implemented in the evaluated object.

- Effectiveness analyzing – includes methods for determining whether the security measures implemented in the evaluated object are effective against the identified threats.
- Vulnerabilities analyzing – includes methods for determining how recommended countermeasures prevent evaluated object from successfully attacking using construction, operational and exploitable vulnerabilities detected in the object.
- Other technical procedures written according to the established criteria (ITSEC, CC, etc.).

Certification of conformity authenticate that evaluation object is consistent with type of object, which obtained type certificate. Certification of conformity process can cover all of objects or samples taken from production line. Technical procedures for certification of conformity make use of documentation and prototype objects, which were basis of the type certification. Each assessed object can require new procedures developed on the start of certification process.

Technical procedures for certification of conformity can cover procedures such as:

- Examination of conformity for the software – includes methods for determining whether the tested software is compatible with prototype of this software.
- Examination of conformity for the device – includes methods for determining whether the tested object is compatible with prototype of the object, for which certificate of type was issued.

Technical procedures for evaluation of crypto-transformation can cover procedures such as:

- Randomness testing of binary sequences produced by stream ciphers, block ciphers, hash functions, etc.
- Independence testing of pairs of binary sequences produced by stream ciphers, block ciphers, hash functions, etc.
- Avalanche testing of block ciphers and other crypto transformations, for example S-boxes.
- Nonlinearity testing of crypto transformations, for example S-boxes.

5.1. Randomness testing

Binary sequences produced by ciphers should be statistically random in order to achieve high security level of cryptographic system. We examine randomness of sequences using statistical tests [8, 10]. These tests use statistics of binary samples and also chi-square and normal distributions. We use following statistical tests: frequency test, serial test, poker tests, runs test and autocorrelation test. During statistical testing of binary sequence we count appropriate statistics for each test. Obtained statistics we split into classes, which identify them from the best to the worst.

In the Cryptology Laboratory we have a technical procedure, which describes in details how to get samples of binary sequence, how to perform statistical testing of samples and how to interpret results of testing. This procedure uses laboratory-developed software, which implement statistical tests used in the method. The procedure and software are validated in the laboratory that gives expected results of testing. We use this procedure for testing binary sequences taken from:

– random and pseudorandom generators;

– stream and block ciphers;

– password and key generators;

– other cryptographic functions, where randomness is crucial.

5.2. Independence testing

Independence testing is similar to randomness testing but concern pairs of binary sequences. Binary sequences produced by ciphers should be statistically independent and we examine independence of pairs of them using statistical tests [9, 10]. These tests use statistics of pairs of binary samples and also chi-square and normal distributions. We examine independence of a pair of binary sequences using three statistical tests for appropriate bit-length pairs. For each test we count suitable statistics. Obtained statistics we split into classes, which identify statistics from the best to the worst.
In the Cryptology Laboratory we have a technical procedure, which describes in details how to get samples of pairs of binary sequence, how to perform statistical testing of samples and how to interpret results of testing. This procedure uses laboratory-developed software, which implement statistical tests used in the method. The procedure and software are validated in the laboratory that gives expected results of testing.

5.3. Avalanche testing

There are a few criteria of avalanche testing. The basic are: avalanche effect and strict avalanche criterion (SAC). Full avalanche effect and full SAC should appear after a few rounds in a properly constructed block cipher. Full avalanche effect will occur in the cipher, if the average number of changed bits in cryptograms is equal to the half-length of ciphered block, as a result of any bit of plain text or key changes. The cipher will fulfil the strict avalanche criterion, if the average probability of bit changing in cryptograms is equal to 0.5, as a result of any bit of plain text or key changes.
In the Cryptology Laboratory we implemented software calculating avalanche effect and strict avalanche effect which occur in the evaluated block cipher. This software is used by procedures which describes in details how to prepare cipher for testing, how to perform testing of avalanche effects in the cipher and how to interpret results of testing.

The procedure and software are validated in the laboratory that gives expected results of testing. We use this procedure for testing block ciphers and other cryptographic functions, where avalanche effects are crucial.

5.4. Nonlinearity testing

Nonlinearity is a basic criterion in achieving resistance of ciphers to cryptanalysis. We can calculate nonlinearity for Boolean functions and for complex functions composed of Boolean functions. Nonlinearity of a function is the Hamming distance to the nearest affine function (it is called classical nonlinearity) or to the nearest function, which have linear structure (it is called strict avalanche criterion). Cryptographically strong functions should have high nonlinearity – it means – large distance to cryptographically weak affine functions and function with linear structure.
In the Cryptology Laboratory we implemented software and a procedure for nonlinearity testing. We use this procedure for testing: elements of block ciphers (for example S-boxes), Boolean functions (used in S-boxes and stream ciphers) and other cryptographic functions, where nonlinearity is crucial.

5.5. Examination of conformity

In the Cryptology Laboratory we have testing procedures used in certification of conformity processes of cryptographic devices and software. We examine evaluated object, whether it is consistent with type of the object, which obtained type certificate. We process all of the objects or samples taken from production line. Results of our testing are basis for issuing certificate of conformity for examined objects by certification authority. Security devices and software with such certificates (and with certificate of type) can be used in IT security systems which process sensitive information (if obtain accreditation certificate).
Each assessed types of objects can require new procedures developed on the start of certification process.

6. Summary

Products and systems used for processing sensitive and classified information should be evaluated, certificated and accredited by competent and common recognized organizations like laboratories and certifications bodies. Certification bodies conduct certification processes, commission laboratory's testing, evaluations and analysis, and on this basis issue certificates for products and systems. Laboratories and certification bodies should operate according to the standards and be recognized by accreditation bodies.
The Cryptology Department has been operating in the Military Communication Institute for many years. The Cryptology Laboratory operates within the whole Cryptology Department using the quality system according to the ISO/IEC 17025 standard. The Cryptology Laboratory bases on the resources of the Cryptology Department, its personnel, knowledge and experience. The quality system according to the ISO/IEC 17025 standard was introduced in the Cryptology Laboratory in 2002. The Cryptology Laboratory is now recognized by the Products Certification Unit of the Military Security Authority.
Main object of interest of the Cryptology Laboratory is testing, analyzing and evaluating military crypto devices and systems. The Cryptology Laboratory performs works commissioned and financed by producers and vendors of cryptographic devices and systems. Results of the Cryptology Laboratory works are used in the certification and accreditation processes performed by the Products Certification Unit.

References

[1] "General requirements for the competence of testing and calibration laboratories", ISO/IEC 17025 Standard, 1999.

[2] "ITSEC (and ITSEM) – Information Technology Security Evaluation Criteria (and Methodology)", United Europe Commission, 1991–1993.

[3] "CC (and CEM) – Common Criteria (and Evaluation Methodology) for Information Technology Security Evaluation", ISO/IEC 15408 Standard, 1999.

[4] "UK Scheme Publications – UK IT Security Evaluation and Certification Scheme", Certification Body – CESG, Cheltenham, UK, 1999–2002.

[5] "Quality Manual and Procedures Manuals of Cryptology Laboratory", Military Communication Institute, Zegrze, 2002–2004.

[6] J. Pieprzyk, T. Hardjono, and J. Seberry, *Fundamentals of Computer Security*. Berlin: Springer-Verlag, 2003.

[7] A. Menezes , P. van Oorschot, and S. Vanstone, *Handbook of Applied Cryptography*. Boca Raton Florida: CRC Press, 1996.

[8] R. Wicik, "Properties of a block cipher based on extended Feistel network with large S-boxes", in *Proc. Conf. RCMCIS'2000*, Zegrze, Poland, 2000.

[9] R. Wicik, "The statistical test for determining independence of pseudorandom bit sequences used in cryptographic systems", in *Proc. Conf. RCMCIS'01*, Zegrze, Poland, 2001.

[10] M. Borowski and R. Wicik, "How to speed up a stream cipher", in *Proc. Conf. RCMCIS'02*, Zegrze, Poland, 2002.

Robert Wicik was born in Iłża, Poland, in 1970. He received the M.Sc. degree in computer science and the Ph.D. degree in telecommunications – cryptographic data security from the Military University of Technology, Warsaw, Poland, in 1994 and 2000, respectively. He has been working in the Military Communication Institute since 1994. His main interests include cryptology and data security. He has been leading the Cryptology Laboratory since 2002.
e-mail: wicik@wil.waw.pl
Military Communication Institute
05-130 Zegrze, Poland

Paper

IP-KRYPTO cipher machine for military use

Mariusz Borowski and Grzegorz Łabuzek

Abstract—Polish military IP networks can be effectively and cheaply secured by IP-KRYPTO cipher machines which are developed in the Military Communication Institute. The co-operation with Polish manufacturers – Optimus and ABA Kraków and the usage of COTS elements and ideas speed up the research and development works. The IP-KRYPTO cipher machine will be used for securing the "SECRET" data so it must be certified to fulfill E3 ITSEC evaluation criteria. This requirement generates additional challenges in the development process when the COTS elements are to be implemented.

Keywords— IPsec standards, ITSEC evaluation criteria.

1. Introduction

Military computer networks, among other things, need the security cryptographic services which can be implemented at several layers in a network infrastructure. Military institutions have used link-level encryption for years. In the method every communications link is protected with a pair of encrypting devices – one on each end of the link. While this method provides excellent data confidentiality between two crypto devices, other cryptographic services are inaccessible. Of course, this method does not work at all in the public networks, where few of the intermediate links are accessible or trusted to the user.

2. Usage of IPsec standards as a COTS development idea

The Internet Protocol security (IPsec) is a framework of open standards for ensuring secure private communications over IP networks. Based on standards developed by the Internet Engineering Task Force (IETF), IPsec ensures confidentiality, integrity, and authenticity of data communications across the public IP network.

The concept of a "Security Association" (SA) is fundamental to IPsec [2]. SA is a simple "connection" that affords security services to the traffic carried by it. Two types of SA are defined: transport mode and tunnel mode.

The transport mode SA is a security association between two hosts. In IPv4, a transport mode security protocol header appears immediately after the IP header and any options, and before any higher layer protocols (e.g., TCP or UDP).

The tunnel mode SA is essentially an SA applied to an IP tunnel. Whenever either end of a security association is a security gateway, the SA must be the tunnel mode. For the tunnel mode SA, there is an "outer" IP header that specifies the IPsec processing destination, plus an "inner" IP header that specifies (apparently) the ultimate destination for the packet.

The Internet Protocol security is based on the following two protocols:

- Authentication Header protocol (AH) [3],
- Encapsulation Security Payload protocol (ESP) [4].

The AH protocol provides integrity and authentication services, while the ESP protocol delivers mainly confidentiality. These protocols may be applied alone or in combination with each other to provide desirable set of security services. Each protocol can be used in the transport mode and the tunnel mode.

The encrypted packets look like ordinary IP packets (Fig. 1), they can easily be routed through any IP network, such as the Internet, without any changes to the intermediate networking equipment.

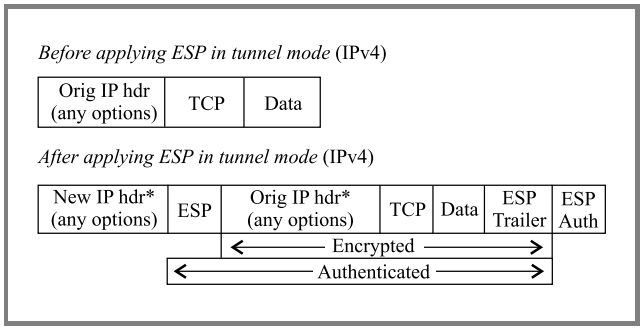


Fig. 1. Appling ESP to a packet in tunnel mode.

The only devices that know about the encryption are the end points. This feature greatly reduces both the implementation and management costs.

3. The TCE 621 IP Crypto Device – the prototype

According to the COTS strategy the IPsec standards were implemented in military IP crypto devices by many manu-

facturers. One of such approach is the TCE 621 IP Crypto Device (Fig. 2) developed by THALES Communications company which was approved for securing NATO IP networks.

Fig. 2. The TCE 621 IP Crypto Device.

The TCE 621 is inserted between a host (end system) or a company network and the IP network. This is used to establish virtual private networks (VPN) solutions, or to provide end-to-end protection of the communications between single hosts. The TCE 621 protects the communications between hosts by adding end-to-end security services to the IP protocol. All security services are provided by the IPsec ESP protocol [4] as specified by IETF.

The TCE 621 provides:

- NATO approved cryptographic algorithms;
- a 10 megabits per second Ethernet interface;
- security for between 500 and 900 IP datagrams per second, depending on pocket size;
- up to 1000 security associations;
- work in a net consisted of up to 1000 such devices;
- audit support and centralized management;
- protection against electromagnetic emanation according to AMSG 720B;
- ruggerized, temper resistant shelter.

The Polish army is not equipped with the IP crypto cipher machines of such functionality but it needs them. The Military Communication Institute, which has long tradition in building the cryptographic devices, starts research and the development process. For efficiency and the time and money saving we decided to use some of the COTS components and ideas.

4. The Optimus ABA IPSec Gate – the first Polish approximation

Polish computer companies, Optimus and ABA Kraków, developed and introduced the Optimus ABA IPSec Gate crypto devices which are based on IPsec standards.
The Optimus ABA IPSec Gate (Fig. 3) uses exclusively well-tried standard PC hardware which is cost-effective and deliverable in the long term. Moreover, the cipher machine can be integrated into every underlying IP-based IT infrastructure.
The Optimus ABA IPSec Gate is based on a specially minimized and hardened Linux operating system with implementation of IPsec standards named FreeS/Wan. All software is stored on flash RAM with manipulation protection. Moreover the software integrity is checked during the system start up, and in small regular time periods.

Fig. 3. The Optimus ABA IPSec Gate.

The Optimus ABA IPSec Gate provides the strongest encryption technology available on the commercial market, using concurrent 3DES (168-bit key) and Rijndael – AES (256-bit) algorithms as standard offerings. The unit employs the Secure Hash Algorithm (SHA-1 and SHA-2), as well as Diffie-Hellman and Internet Key Exchange (IKE) protocols [6] to perform authentication and automatic exchange of a key material.

Supported standards:

- [2], [3], [4] – main IPsec standards;
- [5], [6], [7] – IPsec standards supported session keys establishing methods;
- RFC 2403, The Use of HMAC-MD5-96 within ESP and AH;
- RFC 2404, The Use of HMAC-SHA-1-96 within ESP and AH;
- RFC 2104, HMAC: Keyed-Hashing for Message Authentication Code;
- RFC 2451, The ESP CBC Mode Cipher Algorithms;

- RFC 2408, Internet Security Association and Key Management Protocol;
- RFC 3173, IP Payload Compression Protocol (IPComp);
- RFC 2394, IP Payload Compression Using DEFLATE.

The throughput of data depends on the selected cryptographic algorithm. The data throughput of the Optimus ABA IPSec Gate is scaled with the clock rate of the underlying processor hardware and achieves up to 90 Mbit/s with AES on a Intel Celeron 1.7 GHz processor. It supports dial-up functions (PPP) via ISDN terminal adapters, analogue modems, DSL and GSM/GPRS mobile phones.
The Optimus ABA IPSec Gate was certified for securing the "RESTRICTED" data by the national security authority according to the "Protection of Classified Information Act".

5. The IP-KRYPTO cipher machine research process

The Military Communication Institute in cooperation with Optimus and ABA Kraków companies is developing an IP-KRYPTO cipher machine for use by the Polish army. The cipher machine is designated for securing the "SECRET" data in the military and government IP networks; therefore it must be worked out and produced to fulfill the ITSEC evaluation criteria.
The IP-KRYPTO cipher machine has to fulfill the E3 ITSEC evaluation criteria, but the cryptographic algorithms have to fulfill E6 criteria. For rapid development we decided to use COTS technologies and equipments. The target of the evaluation (TOE) [8] which satisfies E3 criteria must be prepared in the development process with:

- security target for the TOE;
- informal description of the architecture of the TOE;
- informal description of the detailed design;
- test documentation;
- library of test programs and tools used for testing the TOE;
- source code or hardware drawings for all security enforcing and security relevant components;
- informal description of correspondence between source code or hardware drawings and the detailed design.

While the use of IPsec standards as an example of COTS ideas is consistent with the shown criteria in regard to the deliberated specification, the implementation of software modules without the source code is unacceptable.
The main tasks in the research process:

a) designing fast and secure national cryptographic algorithms with a suitable documentation;
b) implementation of algorithms in the COTS implementation of IPsec standards named FreeS/WAN;
c) designing a secure and effective key establishing method;
d) designing a secure and temper resistant keying module;
e) designing an IP-KRYPTO motherboard with an appropriate powerful processor and solved thermal problems;
f) designing a temper resistant and satisfying lack of electromagnetic emanation (AMSG 720B) shelter;
g) designing a shelter and a motherboard which satisfy environmental and mechanical demands;
h) preparing a network planning and key generation station.

Some of the aforementioned challenges will be discussed more precisely here.
The IP-KRYPTO cipher machine demands fast and secure national cryptographic algorithms. The IPsec standards are open and suitable for the implementation of any national algorithms instead of those shown for the commercial use. The new national algorithms were designed in the Military Communication Institute. In addition to the security features also their speed of work was important because the IP-KRYPTO must support at least a 10 Mbit/s Ethernet interface.
The IP-KRYPTO cipher machine needs a motherboard with an appropriate powerful processor and the solved thermal problems. The Optimus ABA IPSec Gate manufactured by our partners uses exclusively the well-tried standard PC hardware which is cost-effective but this commercial equipment cannot be implemented in the military products. We decided to use a hardened industrial standard PC motherboard. The motherboard provides 10/100 Mbit/s Fast Ethernet Ports, a socket 478 processor and DiskOn-Module like flash memory on the board. It also supports optional IPsec encryption accelerator which uses commercial algorithms according IPsec standards. We did not decide to design national VLSI IPsec module because of small scale of production, costs and absence of trusted manufactures (national algorithms will have to be implemented in it). All IPsec operations are software implemented and executed by a fast processor. Computational power of the processor is very important but it causes problems with freezing all modules on the motherboard. Freezing

demands are additionally complicated by the needed lack of electromagnetic emanation of the cipher machine. Mechanical and environmental exposures also must be taken into consideration.

The source code or hardware drawings for all security enforcing and the security relevant components have to be shown in the certification process so the IP-KRYPTO cipher machine operates under control of the Linux operating system. The Linux OS is the open source operating system and everyone wise has an access to its code. The Linux OS used in the IP-KRYPTO crypto machine is integrated with software implementation of IPsec standards named FreeS/WAN. All security services are provided by the IPsec ESP protocol in the tunnel mode. New functions have been added to the standard FreeS/WAN software. The functions concern the ability of connecting with the device located behind a Network Address Translator (NAT-Traversal) and Dead Peer Detection (DPD). Additionally the IP-KRYPTO software is integrated with a software firewall based on iptables with an automatic configuration option. This configuration is created upon configuration parameters for maximizing security and does not reduce functionality. By default the firewall blocks all packets except packets moving in the IPsec tunnel. All applications have been compiled with the use of Gnu Compiler Collection (GCC) with stack-smashing protector. It is an extension for protecting applications from stack-smashing attacks. The protection is realized by buffer overflow detection and the variable reordering feature to avoid the corruption of pointers.

Key establishing methods for IPsec are shown in [6]. Phase 1 is where two ISAKMP peers establish a secure, authenticated channel with which to communicate. In the phase the Diffie-Hellman protocol is used for establishing a common secret. The common secret is used in Phase 2 for establishing cryptographic keys. The IP-KRYPTO cipher machine will secure the "SECRET" data, so the data must stay secret by the next fifty years. In that case we have a well known cryptographic theorem: "If the Diffie-Hellman protocol is secure than the scheme is secure". A well known method of securing the Diffie-Hellman protocol is using more numerous $GF^*(p)$ or transforming the protocol to the group of points on elliptic curves. These examples are shown in [7] and they are suitable for the commercial use where such a long time for securing the data is not required. The security of the Diffie-Hellman protocol is based on the difficulty of solving a discrete logarithms problem, other well known public key algorithms (for example RSA) are based on the difficulty of number factoring. The security of the transformations is based on mathematical problems and computational complexity. But an adversary is able to record all exchanges between two IP-KRYPTO devices and wait (it has fifty years) until new mathematical or computational methods are achieved. Thus a simple copy of the standard [6] in the developing the IP-KRYPTO cipher machine is not available.

The IP-KRYPTO cipher machine needs a secure and temper resistant keying module. The Optimus ABA IPSec Gate used a Flash RAM for hiding the keys and configuration. All data on the Flash RAM are ciphered. This is an excellent example of using the COTS technology which is fast, cost effective and suitable for securing the "RESTRICTED" data. The requirements for securing the "SECRET" data demand that the key module must have a temper resistant shelter and support an emergency clearing function executing without outer power supply.

The last step in the developing the IP-KRYPTO cipher machine will be creating a network planning and key generation station. The main goal of the station is preparation of the configuration data and cryptographic keys for all IP-KRYPTO machines and filling their keying modules.

6. Summary

In the paper we showed the IP-KRYPTO research process. If a crypto device must be used for securing the "SECRET" data, then special requirements must be fulfilled. Usage of COTS technologies and equipments is suitable in developing crypto devices for securing the "RESTRICTED" data. Such techniques implemented in the devices dedicated for securing higher clause data need an extremely cautious, well prepared and experienced developing team. The simplest way is with the use of COTS ideas and standards but military equipments must fulfill additional mechanical, environmental and electromagnetic emanation criteria.

IP-KRYPTO cipher machines demand also a network planning and key generation station. At the end of the research process the prototype of the cipher machine with the station must be certified by the national security authority according to the "Protection of Classified Information Act".

References

[1] "Cryptel®-IP. The TCE 621 IP Crypto Device", htpp://www.thalesgroup.no

[2] RFC 2401, "Security Architecture for the Internet Protocol".

[3] RFC 2402, "IP Authentication Header".

[4] RFC 2406, "Encapsulating Security Payload (ESP)".

[5] RFC 2407, "Internet IP Security Domain of Interpretation for ISAKMP".

[6] RFC 2409, "Internet Key Exchange (IKE)".

[7] RFC 2412, "OAKLEY Key Determination Protocol".

[8] "ITSEC evaluation criteria", htpp://www.cesg.gov.uk

Mariusz Borowski was born in Lublin, Poland, in 1968. He studied at the computer science faculty of the Military University of Technology, Warsaw. Graduated in 1992 and in 1995 he received a Ph.D. in the cryptographic data security at the Military University of Technology. Since his graduating he has been working at the Military Communication Institute. His main objects of interests: cryptology, computer and network security, computer aided computation.
e-mail: borow@wil.waw.pl
Military Communication Institute
05-130 Zegrze, Poland

Grzegorz Łabuzek was born in Opole, Poland, in 1978. He graduated from the Technical University of Opole in 2002. From March 2002 to March 2004 he worked in the ABA Kraków company. Since March 2004 he has been working at the Military Communication Institute. His main objects of interests: computer and network security, open source operating system, cryptology.
e-mail: labuzek@wil.waw.pl
Military Communication Institute
05-130 Zegrze, Poland

Paper

Primality proving with Gauss and Jacobi sums

Andrzej Chmielowiec

Abstract— This article presents a primality test known as APR (Adleman, Pomerance and Rumely) which was invented in 1980. It was later simplified and improved by Cohen and Lenstra. It can be used to prove primality of numbers with thousands of bits in a reasonable amount of time. The running time of this algorithm for number N is $O((\ln N)^{C \ln\ln\ln N})$ for some constant C. This is almost polynomial time since for all practical purposes the function $\ln\ln\ln N$ acts like a constant.

Keywords— prime numbers, primality proving, cyclotomic ring, Gauss sum, Jacobi sum, APR.

1. Introduction

Probabilistic primality tests are in fact compositness tests. That kind of test gives us correct answer only if the number has nontrivial factor. In other words this test gives two possible answers:

– number is composite,

– number may be prime.

In the former case we are absolutely sure that the number is composite, but in the latter one we receive only statistical information. That kind of unsertainty can be accepted in RSA users keys. Situation is completely different with certificate authority keys and elliptic curve parameters. In those cases wrong verification of primality may compromise all keys in cryptosystem.

We can avoid this problem using algorithm which proves primality and always gives correct answer. There are two very effective methods of primality proving. One of them gives primality certificate and is based on elliptic curves [2]. The second one – Gauss and Jacobi sums primality test [3] – is the topic of this paper. The test based on Gauss sums [4, 7] is interesting only from theoretical point of view and it doesn't have practical implications. Its improvement – Jacobi sums test [5, 6] – is much more efficient and can be used to prove primality of numbers with thousands of bits in a reasonable amount of time.

In the following sections we will show how the theoretical results can be interpreted in terms of computer programming. It will allow to understand the basic idea of Jacobi sums method and will be helpful for those who will try to implement this test in practice. The article does not contain any proof, as it can be easily found in references. Probably the best theoretical description of this algorithm can be found in Henri Cohen's book [6].

2. Theoretical background

In the whole article we will use the following notation:

N	–	number which is tested for primality,
p, q	–	small prime numbers.

2.1. Cyclotomic fields

We start from definition of algebraic structure in which the test operations are performed.

Definition 1: If $\zeta_n \in \mathbb{C}$ is such that $\zeta_n^n = 1$ and for all $k < n$ we have $\zeta_n^k \neq 1$, then ζ_n is a primitive nth root of unity, and field extension $\mathbb{Q}(\zeta_n)$ is the nth cyclotomic field. □

Proposition 1: Let $K = \mathbb{Q}(\zeta_n)$ be nth cyclotomic field.

1. The extension $K/\mathbb{Q}$ is a Galois extension, with Abelian Galois group given by

$$G = \mathrm{Gal}(K/\mathbb{Q}) =$$

$$\{\sigma_a : \ (a,n) = 1, \ \text{where } \sigma_a(\zeta_n) = \zeta_n^a\}.$$

   In particular, the degree of $K/\mathbb{Q}$ is $\phi(n)$, where ϕ is the Euler function.

2. The ring of integers of K is $\mathbb{Z}_K = \mathbb{Z}[\zeta_n]$. ■

Such a definition is good from a theoretical point of view, but is also completely impractical. We need to find a way that would allow us to represent elements of $\mathbb{Q}(\zeta_n)$ and $\mathbb{Z}[\zeta_n]$ in a computer. This is possible by the definition of a cyclotomic polynomial:

$$\Phi_n(X) = \prod_{(a,n)=1, \ 0<a<n} (X - \zeta_n^a).$$

It is possible to show that $\Phi_n(X) \in \mathbb{Z}[X]$ and the following lemma is true.

Lemma 1: If ζ_n is a primitive nth root of unity, and $\Phi_n(X)$ is the nth cyclotomic polynomial, then

$$\begin{aligned} \mathbb{Q}(\zeta_n) &\simeq \mathbb{Q}[X]/\Phi_n(X), \\ \mathbb{Z}[\zeta_n] &\simeq \mathbb{Z}[X]/\Phi_n(X). \end{aligned}$$

This isomorphism fixes elements of $\mathbb{Q}$, and sends ζ_n on to X. ■

In this approach we only need to compute the nth cyclotomic polynomial. The following formula gives us a very effective way to do this for small values of n

$$\Phi_n(X) = \prod_{d|n} \left(1 - X^{\frac{n}{d}}\right)^{\mu(d)}, \qquad (\star)$$

where $\mu(d)$ is the Möbius function

$$\mu(d) = \begin{cases} 1 & \text{if } d = 1, \\ (-1)^k & \text{if } d \text{ is product of } k \text{ distinct primes,} \\ 0 & \text{in other cases.} \end{cases}$$

Example 1: Using formula $(\star)$ for a prime or a power of prime, we can easily compute cyclotomic polynomials:

$$\begin{aligned} \Phi_p(X) &= \frac{1-X^p}{1-X} = \sum_{i=0}^{p-1} X^i \\ &= X^{p-1} + \cdots + X + 1, \\ \Phi_{p^k}(X) &= \frac{1-X^{p^k}}{1-X^{p^{k-1}}} = \sum_{i=0}^{p-1} X^{ip^{k-1}} \\ &= X^{(p-1)p^{k-1}} + \cdots + X^{p^{k-1}} + 1. \end{aligned}$$

□

The main part of Jacobi sums test will be based on computations in the ring $\mathbb{Z}[\zeta_{p^k}]$. It is known from Lemma 1 that $\mathbb{Z}[\zeta_{p^k}] \simeq \mathbb{Z}[X]/\Phi_{p^k}(X)$, and $\Phi_{p^k}(X)$ can be computed from $(\star)$. The next example shows how to do arithmetic operations in such a ring.

Example 2: Let $p = 2$, and $k = 2$. The previous considerations lead us to

$$\mathbb{Z}[\zeta_4] \simeq \mathbb{Z}[X]/\Phi_4(X),$$

where $\Phi_4(X) = X^2 + 1$. Of course every element of $\mathbb{Z}[X]/\Phi_4(X)$ may be represented as polynomial of degree $< \deg(\Phi_4)$. Suppose that

$$\begin{aligned} a(X) &= a_1 X + a_0, \\ b(X) &= b_1 X + b_0 \end{aligned}$$

are such representations. Addition and subtraction can be done by coordinates:

$$a(X) \pm b(X) = (a_1 \pm b_1)X + (a_0 \pm b_0).$$

Multiplication is a little bit more complicated. To do this we first have to compute $a(X)b(X)$ in $\mathbb{Z}[X]$, and then reduce the product modulo $\Phi_4(X)$. Since $(a_1X + a_0)(b_1X + b_0) \mod (X^2+1) = a_1b_1X^2 + (a_1b_0 + a_0b_1)X + a_0b_0 \mod (X^2+1) = (a_1b_0 + a_0b_1)X + (a_0b_0 - a_1b_1)$, then we have

$$a(X)b(X) = (a_1b_0 + a_0b_1)X + (a_0b_0 - a_1b_1).$$

□

2.2. Group rings

In the previous section Galois group of extension $\mathbb{Q}(\zeta_n)/\mathbb{Q}$ was defined as

$$G = \{\sigma_a : \ (a,n) = 1, \text{ where } \sigma_a(\zeta_n) = \zeta_n^a\}.$$

If we interpret $\mathbb{Q}(\zeta_n)$ as $\mathbb{Q}[X]/\Phi_n(X)$ then every element $\sigma_a \in G$ can be written as $\sigma_a(X) = X^a \mod \Phi_n(X)$. Group G is in fact the group of automorphisms of the field $\mathbb{Q}(\zeta_n)$ which fixes the base field $\mathbb{Q}$ [9]. There is a natural action of G on $\mathbb{Q}(\zeta_n)$ and $\mathbb{Z}[\zeta_n]$.

Example 3: Consider $\mathbb{Q}(\zeta_4) \simeq \mathbb{Q}[X]/(X^2+1)$ a Galois group of extension $\mathbb{Q}(\zeta_4)/\mathbb{Q}$ is defined as

$$G = \{\sigma_1, \sigma_3\} \simeq (\mathbb{Z}/4\mathbb{Z})^*.$$

Let $b(X) = b_1X + b_0$ be the representation of element from $\mathbb{Q}[X]/(X^2+1)$. Since every $\sigma_a \in G$ fixes $\mathbb{Q}$ elements, we have

$$\begin{aligned} \sigma_1(b(X)) &= b_1\sigma_1(X) + b_0 \\ &= b_1X + b_0, \\ \sigma_3(b(X)) &= b_1\sigma_3(X) + b_0 \\ &= b_1X^3 + b_0 \mod (X^2+1) \\ &= -b_1X + b_0. \end{aligned}$$

The same is true if $b(X) \in \mathbb{Z}[X]/(X^2+1)$. □

Now we will introduce the ring of $\mathbb{Z}$-linear combinations of elements of G. This structure plays a crucial role in the extension of Gauss sums test to Jacobi sums test, and it allows to apply the latter in practice.

Definition 2: Define group ring $\mathbb{Z}[G]$ as the set of elements $f = \sum_{\sigma \in G} f_\sigma \sigma$, where all $f_\sigma \in \mathbb{Z}$. Operations in $\mathbb{Z}[G]$ are defined in the following way:

$$\begin{aligned} f \pm g &= \sum_{\sigma \in G} (f_\sigma \pm g_\sigma)\sigma, \\ f \cdot g &= \sum_{\sigma,\tau \in G} (f_\sigma g_\tau)(\sigma\tau). \end{aligned}$$

□

Addition and subtraction is in $\mathbb{Z}[G]$ defined by coordinates. Multiplication is a little bit more complicated. The following example explains how to do this operation.

Example 4: Consider group ring $\mathbb{Z}[G]$ with G defined as in previous example $G = \{\sigma_1, \sigma_3\} \simeq (\mathbb{Z}/4\mathbb{Z})^*$. The following table presents group law for G

$\cdot$	σ_1	σ_3
σ_1	σ_1	σ_3
σ_3	σ_3	σ_1

If $f = f_1\sigma_1 + f_3\sigma_3$ and $g = g_1\sigma_1 + g_3\sigma_3$ are elements of $\mathbb{Z}[G]$, then we have:

$$\begin{aligned} f \pm g &= (f_1 \pm g_1)\sigma_1 + (f_3 \pm g_3)\sigma_3, \\ f \cdot g &= (f_1g_1)(\sigma_1\sigma_1) + (f_1g_3)(\sigma_1\sigma_3) + \\ &\quad (f_3g_1)(\sigma_3\sigma_1) + (f_3g_3)(\sigma_3\sigma_3) \\ &= (f_1g_1)\sigma_1 + (f_1g_3)\sigma_3 + \\ &\quad (f_3g_1)\sigma_3 + (f_3g_3)\sigma_1 \\ &= (f_1g_1 + f_3g_3)\sigma_1 + (f_1g_3 + f_3g_1)\sigma_3. \end{aligned}$$

□

Now we are ready to extend group action given by G to action given by $\mathbb{Z}[G]$.

Definition 3: Let G be a Galois group of extension $\mathbb{Q}(\zeta_n)/\mathbb{Q}$, and $\mathbb{Z}[G]$ denote its group ring. If $f \in \mathbb{Z}[G]$ and $x \in \mathbb{Q}(\zeta_n)$, then we define action of f on x by

$$x^f = \prod_{\sigma \in G} \sigma(x)^{f_\sigma}$$

for $x \neq 0$, and $0^f = 0$. □

This definition is quite natural and it has very nice properties. One can immediately check, that for all $x, x_1, x_2 \in \mathbb{Q}(\zeta_n)$ and $f, f_1, f_2 \in \mathbb{Z}[G]$ we have:

1. $x^{f_1+f_2} = x^{f_1} x^{f_2}$,
2. $x^{f_1 f_2} = \left(x^{f_1}\right)^{f_2} = \left(x^{f_2}\right)^{f_1}$,
3. $(x_1 + x_2)^f = x_1^f + x_2^f$,
4. $(x_1 x_2)^f = x_1^f x_2^f$.

Example 5: Let $a(X) = a_1 X + a_0$ represent an element of $\mathbb{Q}(\zeta_4) \simeq \mathbb{Q}[X]/(X^2+1)$, then for $f = \sigma_1 + \sigma_2$ we can obtain

$$\begin{aligned} a(X)^f &= \sigma_1(a_1X + a_0)\sigma_3(a_1X + a_0) \\ &= -a_1^2X^2 + a_0^2 \mod (X^2+1) \\ &= a_1^2 + a_0^2, \end{aligned}$$

and for $g = 2\sigma_1 + \sigma_3$ we have:

$$\begin{aligned} a(X)^g &= a(X)^{\sigma_1 + f} \\ &= \sigma_1(a_1X + a_0)(a_1^2 + a_0^2) \\ &= (a_1^3 + a_0^2 a_1)X + (a_0 a_1^2 + a_0^3). \end{aligned}$$

□

The most interesting is the case of $n = p^k$, where p is prime. The following proposition shows the relation which will be usful in the next section.

Proposition 2: If $n = p^k$ and $G = \mathrm{Gal}(\mathbb{Q}(\zeta_n)/\mathbb{Q})$, then the set

$$\mathfrak{p} = \{f \in \mathbb{Z}[G] : \ \zeta_p^f = 1\}$$

is a prime ideal of group ring $\mathbb{Z}[G]$. ■

2.3. Dirichlet characters

Dirichlet character χ modulo q is a group homomorphism from $(\mathbb{Z}/q\mathbb{Z})^*$ to $\mathbb{C}^*$. If q is prime then character χ can be defined by choosing value $\chi(g)$ for some generator g of $(\mathbb{Z}/q\mathbb{Z})^*$.

Example 6: If $q = 5$, then $g = 2$ is a generator of $(\mathbb{Z}/5\mathbb{Z})^*$, and all Dirichlet characters may be defined by choosing value for $\chi(g)$. But χ must be a homomorphism, so its image has to be a multiplicative subgroup of order four in $\mathbb{C}^*$. There are only four such possibilities:

1. $\chi_1(g) = 1$ (trivial character),
2. $\chi_2(g) = -1$,
3. $\chi_3(g) = i$,
4. $\chi_4(g) = -i$.

□

It can be shown, that the set of all characters modulo q forms a group.

Proposition 3: All characters from $(\mathbb{Z}/q\mathbb{Z})^*$ to $\mathbb{C}^*$ form a group with neutral element χ_0 such that $\chi_0(x) = 1$ for all $x \in (\mathbb{Z}/q\mathbb{Z})^*$. ■

Since χ is a homomorphism and $|(\mathbb{Z}/q\mathbb{Z})^*| < \infty$ one can show that the set of character values forms a multiplicative group which is a subgroup of $\langle \zeta_q \rangle = \left\langle e^{\frac{2\pi i}{q}} \right\rangle$. Definition of character may be extended to a multiplicative map from $\mathbb{Z}/q\mathbb{Z}$ to $\mathbb{C}$ by taking $\chi(x) = 0$ for all $x \notin (\mathbb{Z}/q\mathbb{Z})^*$. It can then be lifted to map from $\mathbb{Z}$ to $\mathbb{C}$. More information about characters can be found in [8].

2.4. Gauss and Jacobi sums

We are now ready to give the definition and some basic properties of Gauss and Jacobi sums.

Definition 4:

1. Let χ be a character modulo q. The Gauss sum $\tau(\chi)$ is defined by
$$\tau(\chi) = \sum_{x \in (\mathbb{Z}/q\mathbb{Z})^*} \chi(x)\zeta_q^x,$$
where $\zeta_q = e^{\frac{2\pi i}{q}}$.
2. Let χ_1, χ_2 be two characters modulo q. The Jacobi sum $j(\chi_1, \chi_2)$ is defined by
$$j(\chi_1, \chi_2) = \sum_{x \in (\mathbb{Z}/q\mathbb{Z})^*} \chi_1(x)\chi_2(1-x).$$

□

There is a nontrivial connection between those two objects. It allows us to implement our primality proof in a very effective way.

Proposition 4: Let χ_1, χ_2 be characters modulo q such that $\chi_1\chi_2 \neq \chi_0$. Then

$$j(\chi_1, \chi_2) = \frac{\tau(\chi_1)\tau(\chi_2)}{\tau(\chi_1\chi_2)}.$$

■

It is clear that if χ is a character modulo prime number q, then its values belong to some group $\langle \zeta_n \rangle$, where $n \mid q-1$. But it means that $\tau(\chi) \in \mathbb{Z}[\zeta_n, \zeta_q]$ and $j(\chi_1, \chi_2) \in \mathbb{Z}[\zeta_n]$. The second ring is simpler and has smaller cost of arithmetic operations. The next section will show how to use it to effectively implement the primality test.

3. Primality test

The previous section presented the theoretical basis of Jacobi sums test concept. Now we will try to sum up the theory and show how it can be used in the construction of a primality proving algorithm. It will be done in two steps. The first step describes an impractical algorithm based on Gauss sums (that are located in a large ring). The second one uses particular properties of Jacobi sums to move computations into a smaller ring, where Gauss sums are replaced by Jacobi sums.

We assume that N has already passed the Rabin-Miller test and it is highly improbable that N is composite. Our aim is the proof of primality of N. In this section we fix prime numbers p,q such that $p^k \mid q-1$ and $p^{k+1} \nmid q-1$. Let χ be character modulo q of order $n = p^k$ in the group of characters.

3.1. Basic test

The fundamental concept is to prove a generalization of Fermat's little theorem. It allows us to verify many congruences that are satisfied by prime numbers and together imply primality of the tested number.

Proposition 5: Let $\beta \in \mathbb{Z}[G]$. Then if N is prime, there exists $\eta(\chi) \in \langle \zeta_n \rangle$ such that

$$\tau(\chi)^{\beta(N-\sigma_N)} \equiv \eta(\chi)^{-\beta N} \mod N, \qquad (\star_\beta)$$

where $\eta(\chi) = \chi(N)$. ■

Note that $\mathbb{Z}[G]$ acts not only on $\mathbb{Z}[\zeta_n]$ but also on $\mathbb{Z}[\zeta_n, \zeta_q]$. But it doesn't matter because action on ζ_q is trivial (identity action). In the final version of the test, the congruence $(\star_\beta)$ in $\mathbb{Z}[\zeta_n, \zeta_q]$ will be transformed to equivalent condition in $\mathbb{Z}[\zeta_n]$. So results of this section are important only from the theoretical point of view and it is unnecessary to give examples of operations in $\mathbb{Z}[\zeta_n, \zeta_q]$.

In order to present the main result of this section, first we have to define the so called $\mathfrak{L}_p$ condition.

Definition 5: Condition $\mathfrak{L}_p$ is satisfied iff for all prime divisors r of N and all positive integers a we can find $l_p(r,a)$ such that

$$r^{p-1} \equiv N^{(p-1)l_p(r,a)} \mod p^a.$$

□

Now we are ready to formulate the fundamental theorem which allows us to give primality proof of number N.

Theorem 1: Let t be an even integer. Define

$$e(t) = 2 \prod_{q \text{ prime},\ (q-1)\mid t} q^{v_q(t)+1}.$$

Assume that $(N, te(t)) = 1$ and $e(t) > \sqrt{N}$. For each pair of primes (p,q) such that $(q-1) \mid t$ and $p^k \parallel (q-1)$, let $\chi_{p,q}$ be a character modulo q of order p^k (if g_q is a generator modulo q, then we can take $\chi_{p,q}(g_q) = \zeta_{p^k}$). If the following conditions are satisfied:

1. all $\chi_{p,q}$ satisfy $(\star_\beta)$ for some $\beta_{p,q} \notin \mathfrak{p}$,
2. condition $\mathfrak{L}_p$ is true for all primes $p \mid t$,
3. for every $0 \le i < t$ and $r = N^i \mod e(t)$ if $r \neq 1$, then $r \nmid N$,

then N is prime. ■

This theorem is interpreted as follows. If congruence $(\star_\beta)$ is false for some $\chi_{p,q}$, then we have that N is not prime (just like in Fermat test). But if $(\star_\beta)$ is true for all defined characters then we get some extra information about possible divisors of N. This allows to prove primality of N or gives its nontrivial factor. So the only problem is to verify the $\mathfrak{L}_p$ condition. The following proposition gives a practical method for checking it.

Proposition 6: Suppose that χ is a character modulo q of order p^k which satisfies $(\star_\beta)$ for some $\beta \notin \mathfrak{p}$. If one of the following conditions is true, then $\mathfrak{L}_p$ is satisfied:

1. $p \ge 3$,
2. $p = 2$, $k = 1$ and $N \equiv 1 \mod 4$,
3. $p = 2$, $k \ge 2$ and $q^{\frac{N-1}{2}} \equiv -1 \mod N$. ■

3.2. Jacobi sums

The test based on Gauss sums is asymptotically fast, however it is far from being practical. Main reason for this situation is the computation of $\tau(\chi)^{\beta(N-\sigma_N)}$. One needs to work in $\mathbb{Z}[\zeta_n, \zeta_q]$ and this is very slow in practice.

Example 7: If we want to test number $N < 10^{100}$ then we can take $t = 5040$. In this case $n = p^k$ will be very small, more precisely $p^k \le 16$. Unfortunately q will be much larger, the largest value being $q = 2521$. This forces us to consider polynomials of degree $> 1,5 \cdot 10^4$ and coefficients reduced modulo N. Multiplying such polynomials takes about $2 \cdot 10^8 \approx 2^{27}$ multiplications modulo N and makes this completely unpractical. □

The above example shows that using Gauss sums is computationally infeasible. One of possible ways to make the test practical, is to replace the $(\star_\beta)$ congruence by some condition depending only on Jacobi sum which lies in a smaller ring $\mathbb{Z}[\zeta_n]$. Fortunately, it is possible and the next three propositions give complete description of this construction.

First we present a very nice result which gives equivalent condition for all practically considered odd primes p.

Proposition 7: Let $3 \le p < 6 \cdot 10^9$ and $p \neq 1093, 3511$. If we denote by E the set of all integers $1 \le x < p^k$ coprime to p, then condition $(\star_\beta)$ is equivalent to congruence:

$$j(\chi,\chi)^\alpha \equiv \eta(\chi)^{-cN} \mod N,$$

where

$$\alpha = \sum_{x \in E} \left\lfloor \frac{Nx}{p^k} \right\rfloor \sigma_x^{-1}$$

and $c = 2(2^{(p-1)p^{k-1}} - 1)/p^k$. ■

Note that the restriction on p in above proposition is completely irrelevant in practice. Even if we want to test the primality of numbers having 10^9 decimal digits, we would never need primes larger than 1093. This means that the practical problem of testing $(\star_\beta)$ for $p \geq 3$ is solved. The next two propositions decribe the case $p = 2$.

Proposition 8: Let χ be a character modulo q of order 2^k with $k \geq 3$. Denote by E the set of all integers $1 \leq x < 2^k$ that are congruent to 1 or 3 modulo 8. Set $\delta_N = 0$ for N congruent to 1 or 3 modulo 8, $\delta_N = 1$ if N is congruent to 5 or 7 modulo 8. The $(\star_\beta)$ condition can be replaced by

$$\left(j(\chi,\chi)j(\chi,\chi^2)\right)^\alpha j\left(\chi^{2^{k-3}},\chi^{3\cdot 2^{k-3}}\right)^{2\delta_N} \equiv$$

$$(-1)^{\delta_N}\eta(\chi)^{-cN} \mod N,$$

where

$$\alpha = \sum_{x\in E}\left\lfloor \frac{Nx}{2^k}\right\rfloor \sigma_x^{-1}$$

and $c = 3(3^{2^{k-2}}-1)/2^k$. ■

Proposition 9: For $p = 2$, $k = 1$ and $\beta = 1$ condition $(\star_\beta)$ is equivalent to the congruence

$$(-q)^{\frac{N-1}{2}} \equiv \eta(\chi) \mod N.$$

For $p = 2$, $k = 2$ and $\beta = 1$ condition $(\star_\beta)$ is equivalent to the congruence

$$j(\chi,\chi)^{\frac{N-1}{2}} q^{\frac{N-1}{4}} \equiv \eta(\chi)^{-1} \mod N$$

if $N \equiv 1 \mod 4$, and to the congruence

$$j(\chi,\chi)^{\frac{N+1}{2}} q^{\frac{N-3}{4}} \equiv -\eta(\chi) \mod N$$

if $N \equiv 3 \mod 4$. ■

4. Implementation and results

4.1. Description of the algorithm

This subsection is based on algorithm given by Henri Cohen in his book [6] and gives pseudocode of Jacobi sums primality test.

Algorithm 1 (Precomputations): Let B be an upper bound on the numbers we want to test. This algorithm makes precomputations of values that don't depend on N.

1. Find such t that $e(t)^2 > B$ (see Theorem 1 for definition).
2. For every prime q dividing $e(t)$ with $q \geq 3$ do as follows.
   (a) Find a primitive root g_q modulo q, and a table of the function $f(x)$ defined for $1 \leq x \leq q-2$ by $1 - g_q^x = g_q^{f(x)}$ and $1 \leq f(x) \leq q-2$.
   (b) For every prime p dividing $q-1$ let k be such number that $p^k \mid q-1$ and $p^{k+1} \nmid q-1$. Let $\chi_{p,q}$ denote the character defined by $\chi_{p,q}(g_q^x) = \zeta_{p^k}^x$.
   (c) If $p \geq 3$ or $p = 2$ and $k = 2$, compute

$$J(p,q) = j(\chi_{p,q},\chi_{p,q}) = \sum_{1\leq x\leq q-2}\zeta_{p^k}^{x+f(x)}.$$

   If $p = 2$ and $k \geq 3$, compute $J(2,q)$ as above and then

$$J_3(q) = j(\chi_{2,q},\chi_{2,q})j(\chi_{2,q},\chi_{2,q}^2) =$$

$$J(2,q)\left(\sum_{1\leq x\leq q-2}\zeta_{2^k}^{2x+f(x)}\right),$$

   and

$$J_2(q) = j\left(\chi_{2,q}^{2^{k-3}},\chi_{2,q}^{3\cdot 2^{k-3}}\right)^2 =$$

$$\left(\sum_{1\leq x\leq q-2}\zeta_8^{3x+f(x)}\right)^2.$$

The above algorithm shows how to compute the set of Jacobi sums for tested numbers that are smaller than some upper bound B. The key step of this part is to compute a very large table $f(x)$ of $q-1$ elements, which allows us to determine Jacobi sums. Of course this part doesn't have to be precomputed as it was suggested before. Experiments show that it takes about $3-5\%$ of total time of the testing procedure, so the described step can be done during every test. The next algorithm combines all previous theoretical results and it proves primality or compositeness of the tested number.

Algorithm 2 (Jacobi sums primality test): Suppose that $N \leq B$ and precomputation step has been done.

1. If $(te(t),N) > 1$, then N is composite.
2. For every prime $p \mid t$ set $l_p \leftarrow 1$ if $p \geq 3$ and $N^{p-1} \not\equiv 1 \mod p^2$, $l_p \leftarrow 0$ otherwise.
3. For each pair (p,q) of primes such that $p^k \parallel (q-1) \mid t$ execute 4a if $p \geq 3$, 4b if $p = 2$ and $k \geq 3$, 4c if $p = 2$ and $k = 2$, 4d if $p = 2$ and $k = 1$. Then go to Step 5.

4a. (Based on Proposition 7): Let E be the set of all positive integers smaller than p^k and coprime to p. Set $\Theta \leftarrow \sum_{x\in E} x\sigma_x^{-1}$, $r \leftarrow N \mod p^k$, $\alpha \leftarrow \sum_{x\in E}\left\lfloor \frac{rx}{p^k}\right\rfloor \sigma_x^{-1}$, and compute $s_1 \leftarrow J(p,q)^\Theta \mod N$, $s_2 \leftarrow s_1^{\lfloor N/p^k\rfloor} \mod N$, and $S(p,q) \leftarrow s_2 J(p,q)^\alpha \mod N$.

If p^kth root of unity η such that $S(p,q) \equiv \eta \mod N$ doesn't exist, then N is composite. If η exists and is a primitve root, then set $l_p \leftarrow 1$.

4b. (Based on Proposition 8): Let E be the set of all positive integers smaller than 2^k that are congruent to 1 or 3 modulo 8. Set $\Theta \leftarrow \sum_{x\in E} x\sigma_x^{-1}$,

$r \leftarrow N \mod 2^k$, $\alpha \leftarrow \sum_{x \in E} \left\lfloor \frac{rx}{2^k} \right\rfloor \sigma_x^{-1}$, and compute $s_1 \leftarrow J_3(q)^{\Theta} \mod N$, $s_2 \leftarrow s_1^{\lfloor N/2^k \rfloor} \mod N$, and $S(2,q) \leftarrow s_2 J_3(q)^{\alpha} J_2(q)^{\delta_N} \mod N$, where $\delta_N = 0$ if $r \in E$, $\delta_N = 0$ otherwise.

If 2^kth root of unity η such that $S(2,q) \equiv \eta \mod N$ doesn't exist, then N is composite. If η is a primitve root and in addition $q^{(N-1)/2} \equiv -1 \mod N$, then set $l_2 \leftarrow 1$.

4c. (Based on Proposition 9): Compute $s_1 \leftarrow J(2,q)^2 \cdot q \mod N$, $s_2 \leftarrow s_1^{\lfloor N/4 \rfloor} \mod N$, and $S(2,q) \leftarrow s_2$ if $N \equiv 1 \mod 4$, $S(2,q) \leftarrow s_2 J(2,q)^2$ if $N \equiv 3 \mod 4$.

If the fourth root of unity η such that $S(2,q) \equiv \eta \mod N$ doesn't exist, then N is composite. If η is a primitve root and in addition $q^{(N-1)/2} \equiv -1 \mod N$, then set $l_2 \leftarrow 1$.

4d. (Based on Proposition 9): Compute $S(2,q) \leftarrow (-q)^{(N-1)/2} \mod N$.

If $S(2,q) \not\equiv \pm 1 \mod N$ then N is composite. If $S(2,q) \equiv -1 \mod N$, and $N \equiv 1 \mod 4$, then set $l_2 \leftarrow 1$.

5. (Based on Proposition 6): For every $p \mid t$ such that $l_p = 0$ do as follows. Take random number q such that $q \nmid e(t)$, $p \mid (q-1)$ and $(q,N) = 1$. Execute Step 4a–4d according to the value of pair (p,q).

If after a reasonable number of tries, some l_p is still equal to 0, then send message saying that the test failed (this is highly improbable).

6. (Based on Theorem 1): For $i = 1, \ldots, t-1$ compute $r_i \leftarrow N^i \mod e(t)$. If for some i, r_i is a nontrivial factor of N then N is composite. Otherwise N is prime.

Presented algorithm works well both in theory and in practice. Pomerance and Odlyzko have shown that the complexity of the Jacobi sums test is

$$O(\ln N^{C \ln \ln \ln N})$$

for some constant C. This is almost polynomial time.

4.2. Choosing good t

We see that in Step 6 Algorithm 2 needs to do $O(t)$ divisions. That means that the chosen t shouldn't be too large. On the other hand, if t will be small, then $e(t)^2$ may be smaller than N and this will not allow to use the algorithm. Steps 4a–4d have small complexity only if factors of t are small. All the above considerations tell us that t shouldn't be too small and too large, and it should have small factors. Unfortunately there is no good description of this problem. So we present values of t that are based on intuition and experiments. Table 1 presents the sample values.

Table 1
Sample values of t for testing numbers

$\log_2 N$	t
≤ 101	$180 = 2^2 \cdot 3^2 \cdot 5$
≤ 152	$720 = 2^4 \cdot 3^2 \cdot 5$
≤ 204	$1260 = 2^2 \cdot 3^2 \cdot 5 \cdot 7$
≤ 268	$2520 = 2^3 \cdot 3^2 \cdot 5 \cdot 7$
≤ 344	$5040 = 2^4 \cdot 3^2 \cdot 5 \cdot 7$
≤ 525	$27720 = 2^3 \cdot 3^2 \cdot 5 \cdot 7 \cdot 11$
≤ 774	$98280 = 2^3 \cdot 3^3 \cdot 5 \cdot 7 \cdot 13$
≤ 1035	$166320 = 2^4 \cdot 3^3 \cdot 5 \cdot 7 \cdot 11$
≤ 1566	$720720 = 2^4 \cdot 3^2 \cdot 5 \cdot 7 \cdot 11 \cdot 13$
≤ 2082	$1663200 = 2^5 \cdot 3^3 \cdot 5^2 \cdot 7 \cdot 11$
≤ 3491	$8648640 = 2^6 \cdot 3^3 \cdot 5^2 \cdot 7 \cdot 11 \cdot 13$

4.3. Running times of the algorithm

We have implemented Algorithm 2 and Rabin-Miller test using the same standard modular arithmetic. Table 2 presents comparison of running time between Jacobi sums test and 320 iterations of Rabin-Miller test. Presented times suggest that the Jacobi sum test can not be used for fast generation of prime numbers. But it can be used for single operations such as generation of

- cryptosystem parameters,
- certificate authority key.

The benefit is the certainty that the base of our cryptosystem satisfies theoretical requirements.

Table 2
Sample runnig times for Jacobi sums test
and 320 iterations of Rabin-Miller test on 430 MHz PC

$\log_2 N$	Jacobi sums	Rabin-Miller
64	0.06 s	0.06 s
128	0.16 s	0.15 s
256	2.09 s	0.75 s
512	37.6 s	5.02 s
1024	907 s	37.1 s

Comparsion of times from Table 2 is also presented in Fig. 1. The logarithmic scale of time was taken to show how close is the runnig time of Jacobi sums test to polynomial time. We can see that the complexity is bounded by $C_1 \ln^3 N$ for Rabin-Miller test and by $C_2 \ln^{4.5} N$ for Jacobi sums primality prooving method. Last remark shows that

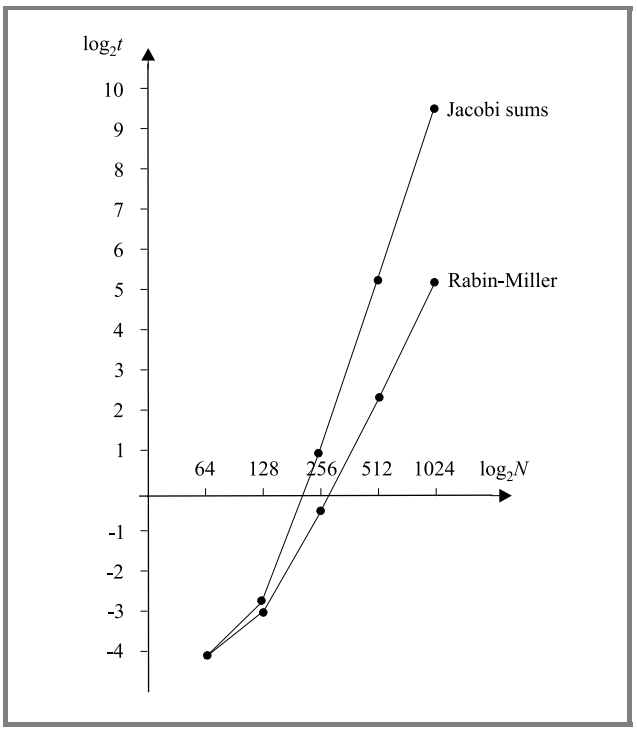


Fig. 1. Comparsion of runnig times for Jacobi sums test and 320 iterations of Rabin-Miller test (based on data from Table 2).

for practical applications Jacobi sums method is faster than deterministic, polynomial algorithm proposed by Agrawal, Kayal and Saxena [1].

5. Conclusions

This article presented a primality proving algorithm based on Jacobi sums. Described algorithm is about 2400 times slower than a single iteration of Rabin-Miller test for 512-bit numbers. This means that it can't be used in applications where time is one of the most critical resourses and where high speed is necessary. On the other hand there exist situations where security is much more important than speed. Then Jacobi sums test can be successfully used to verify primality of strong pseudoprime numbers.

References

[1] M. Agrawal, N. Kayal, and N. Saxena, "Technical report", Department of Computer Science and Engineering Indian Institute of Technology, Kanpur, 2002.

[2] O. Atkin and F. Morain, "Elliptic curves and primality proving", *A.M.S.*, vol. 61, pp. 29–68, 1993.

[3] L. Adleman, C. Pomerance, and R. Rumely, "On distinguishing prime numbers from composite numbers", *Ann. Math.*, vol. 117, pp. 173–206, 1983.

[4] E. Bach and J, Schallit, *Algorithmic Number Theory*. Cambridge: MIT Press, 1996.

[5] W. Bosma and M. van der Hulst, "Primality proving with cyclotomy", Ph.D. thesis, Amsterdam, University of Amsterdam, 1990.

[6] H. Cohen, *A Course in Computational Algebraic Number Theory*. Berlin: Springer-Verlag, 1993.

[7] R. Crandall and C. Pomerance, *Prime Numbers: A Computational Perspetive*. New York: Springer-Verlag, 2001.

[8] K. Ireland and M. Rosen, *A classical Introduction to Modern Number Theory*. New York: Springer-Verlag, 1990.

[9] J. Rotman, *Galois Theory*. New York: Springer-Verlag, 1990.

Andrzej Chmielowiec is a graduate of the Warsaw University, Faculty of Mathematics, Informatics and Mechanics. For the last two years he has been working for Enigma Information Security Systems Sp. z o.o. as cryptographer. His major job activity is to implement and introduce new cryptographic algorithms based on modern algebra methods. At the top of the list of author's interests there are applications of computational algebra. He participates in calling informal seminars dedicated to that subject at the Mathematics Faculty of UW.

e-mail: achmielowiec@enigma.com.pl
Enigma Information Security Systems Sp. z o.o.
Cietrzewia st 8
02-492 Warsaw, Poland

Paper

Packet switch architecture with multiple output queueing

Grzegorz Danilewicz, Mariusz Głąbowski, Wojciech Kabaciński, and Janusz Kleban

Abstract— In this paper the new packet switch architecture with multiple output queuing (MOQ) is proposed. In this architecture the nonblocking switch fabric, which has the capacity of $N \times N^2$, and output buffers arranged into N separate queues for each output, are applied. Each of N queues in one output port stores packets directed to this output only from one input. Both switch fabric and buffers can operate at the same speed as input and output ports. This solution does not need any speedup in the switch fabric as well as arbitration logic for taking decisions which packets from inputs will be transferred to outputs. Two possible switch fabric structures are considered: the centralized structure with the switch fabric located on one or several separate boards, and distributed structure with the switch fabric distributed over line cards. Buffer arrangements as separate queues with independent write pointers or as a memory bank with one pointer are also discussed. The mean cell delay and cell loss probability as performance measures for the proposed switch architecture are evaluated and compared with performance of OQ architecture and VOQ architecture. The hardware complexity of OQ, VOQ and presented MOQ are also compared. We conclude that hardware complexity of proposed switch is very similar to VOQ switch but its performance is comparable to OQ switch.

Keywords— high-speed packet switching, output queueing, buffer, switch fabric, switching node, multicast.

1. Introduction

The transmission capacity of optical fibers has caused a tremendous increase in data transmission speed. The development of broadband access technologies resulted in the need for next generation routers with high-speed interfaces and large switching capacity. One of constrains that limits the switching capacity is the speed of memories used for buffering packets to resolve contention resolution in packet switches. Buffers can be placed on inputs, outputs, inputs and outputs, and/or within the switch fabric. Depending on the buffer placement respective switches are called input queued (IQ), output queued (OQ), combined input and output queued (CIOQ) and combined input and crosspoint queued (CICQ) [1].

In the OQ strategy all incoming cells (i.e., fixed-length packets) are allowed to arrive at the output port and are stored in queues located at each outlet of switching elements. The cells destined for the same output port simultaneously do not face a contention problem because they are queued in the buffer at the outlet. To avoid the cell loss the system must be able to write N cells in the queue during one cell time, N is the total number of inlets of the switch. No arbiter is required because all the cells can be switched to respective output queue. The cells in the output queue are served using FIFO discipline to maintain the integrity of the cell sequence. In OQ switches the best performance (100% throughput, low mean time delay) is achieved, but every output port must be able to accept a cell from every input port simultaneously or at least within a single time slot (a time slot is the duration of a cell). If more cells will request access to a particular output port than the switch fabric output buffer can support, the excess cells must be discarded. An output buffered switch can be more complex than an input buffered switch because the switch fabric and output buffers must effectively operate at a much higher speed than that of each port to reduce the probability of cell loss. The bandwidth required inside the switching fabric is proportional to both the number of ports N and the line rate. This speed is necessary when all inputs simultaneously transfer a cell to the same output port. Such case is called "hot spot" and often occurs when a popular server is connected to a single switch port. The internal speedup factor is inherent to pure output buffering, and is the main reason of difficulties in implementing switches with output buffering. It is no longer possible to find RAMs with sufficiently fast access time taking into account an increasing line rate. Since the output buffer needs to store N cells in each time slot, its speed limits the switch size.

The IQ packet switches have the internal operation speed equal to (or slightly higher) than the input/output line speed, but the throughput is limited to 58.6% under uniform traffic and Bernoulli packet arrivals because of head-of-line (HOL) blocking phenomena [2]. This problem can be solved by selecting queued cells other than the HOL cell for transmission, but it is difficult to implement such queueing discipline in hardware. Another solution is to use speedup, i.e., the switch's internal links speed is greater than inputs/outputs speed. However, this also requires a buffer memory speed faster than a link speed. To increase the throughput of IQ switches space parallelism is also used in the switch fabric, i.e., more than one input port of the switch can transmit simultaneously [3].

One of the proposed solution for IQ switches, which is recently widely considered in papers, is a virtual output queuing (VOQ) [4, 5]. In this solution an input buffer in each input port is divided into N parallel queues, each storing packets directed to different output port. When a new cell arrives at the input port, it is stored in the destined queue and waits for transmission through a switch fabric. In this architecture, the memory speed remains compati-

ble with the line rate, but a good matching algorithm between inputs and outputs is needed so that it can achieve high throughput and low latency. The performance of the switch can be improved when the internal switch fabric operates a few times faster than the line rate, but faster memories are also needed in this case. Different scheduling algorithms for VOQ switches were considered in the literature [5–9], most of them achieve 100% throughput under uniform traffic, but the throughput is usually reduced under non-uniform traffic. The arbitration scheme should be realized slot by slot, therefore the arbiter's speed also limits the capacity of the switch.

In this paper we propose a new switch architecture which uses multiple output queuing (MOQ). In this architecture buffers are located at output ports and are divided into N separate queues. Each of N queues in one output port stores packets from one input port. We assume, that fixed-length switching technology is used, i.e., variable-length packets are segmented into fixed-length packets, called time slots or cells, at inputs and reassembled at the outputs. We will use terms cell and packet interchangeably further on. In the proposed architecture at most one packet is to be written to the one output queue in one time slot. Therefore, the memory speed is equal to the line speed, but the performance of the switch is very similar to those of OQ switch.

The rest of the paper is organized as follows. In Section 2 the general switch architecture is proposed. The possible switch fabric structures are described. Centralized and distributed switch fabrics structures and possible buffer arrangements are considered. In the next section performance evaluation of the proposed switch architecture using simulation is done. Then some comparison between implementation complexity of the proposed switch architecture and a VOQ switch are given, followed by conclusions.

2. The switch architecture

2.1. General architecture

In this paper we propose the new switch architecture which uses output queueing. To reduce the memory speed an output buffer at each output port is divided into N separate queues. Each queue stores packets directed to the output only form one input. In this way this architecture is similar to the VOQ switch, but multiple buffers are located at output ports not at input ports. We will call this architecture the multiple output queueing switch. The general architecture of the switch is shown in Fig. 1. The switch consists of N input ports, N output ports and the switch fabric. Input and output ports can be implemented on separated ingress and egress cards, as it is shown in Fig. 1, or they may be placed on one line card, as it will be shown latter. Each ingress card is connected to the switch fabric by one line, while N outputs from the switch fabric are connected to one egress card. At the output port buffer memory is divided into N separate queues. Each queue stores packets directed from one input port. The output queue denoted by $OQ_{j,i}$ at the output port j stores packets directed to this output port from input i. At the given time slot each input port can send at most one packet and each output port can receive up to N packets, each from different input ports. Therefore, these packets can be simultaneously written to N queues.

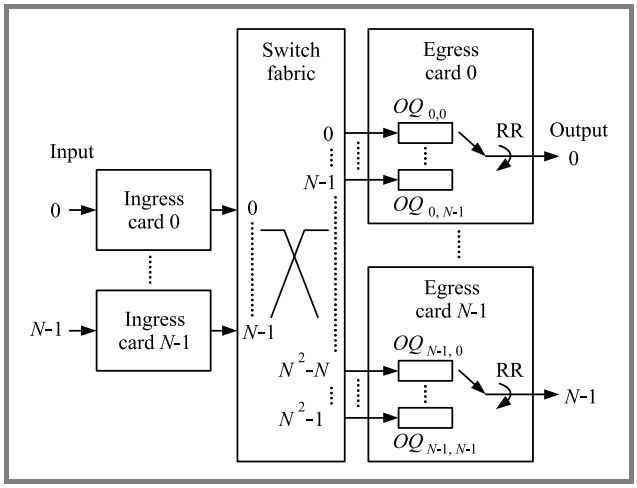


Fig. 1. The switch architecture with multiple output queueing.

The main advantage of these architecture is that it can operate at the same speed as input and output ports, and the lack of arbitration logic, which decides which packets from inputs will be transferred through the switch fabric to output ports (this arbitration mechanism is needed in VOQ switches). However, since we have N queues in each output port, it is necessary to use an output arbiter, which chooses a packet to be sent to the output line. We propose to use round-robin scheme, which is widely used because of its fairness. The buffer management algorithms will be discussed in more details later on.

2.2. The switching fabric architecture

We will now consider some possible switch fabric implementations. The switch fabric in the proposed switch should have a capacity of $N \times N^2$ and should be nonblocking at the packet level. It should be noted that there is no need to support full connectivity in the switch fabric. Any input should only have a possibility to send packets to N different switch fabric's outputs, each of these N outputs should be connected to the different output port. In general, input i, $0 \leq i \leq N-1$ should be able to transfer a packet to the switch fabric output $jN+i$, $0 \leq j \leq N-1$, when this packet is directed to output port j. The packet will be then stored in $OQ_{j,i}$.

The switch fabric can be organized either in the centralized mode or the distributed mode. In the first case the switch fabric constitutes one module produced on one board (or several boards). This architecture is shown in Fig. 2. We assume here that one input port and one output port

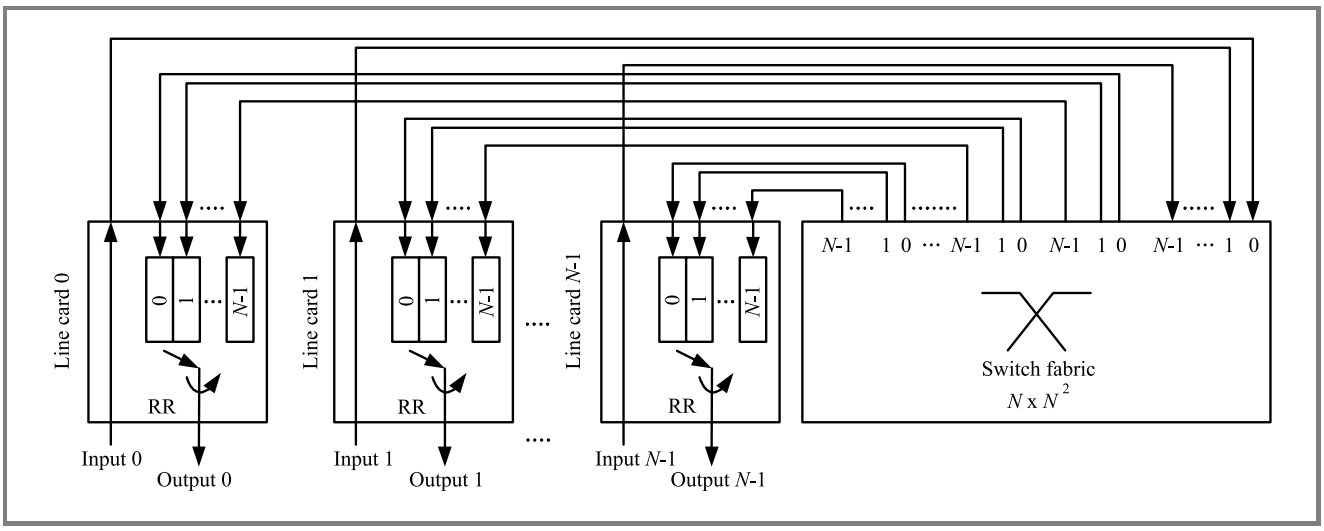


Fig. 2. The switch architecture with the centralized switch fabric.

are arrange on one line card. Buffers are placed at output ports. The switch fabric may be realized for instance using the tree architecture, or may be based on the crossbar architecture as it is shown in Figs. 3 and 4, respectively. The stacked-banyan switch fabric proposed in [10] can also be used. In all these solutions $N^2 + N$ lines is needed for connecting input and output ports to the switch fabric. Each line card is connected by means of $N+1$ lines to the switch fabric. For switches of greater capacity number of connectors will limit the switch size. The capacity of the switch may be increased by using fiber connections with wavelength multiplexing. The other solution is to combine output buffers within the switch fabric. In this case line cards will be connected with the switch fabric by two lines, but the switch fabric will require more boards with buffer memories.

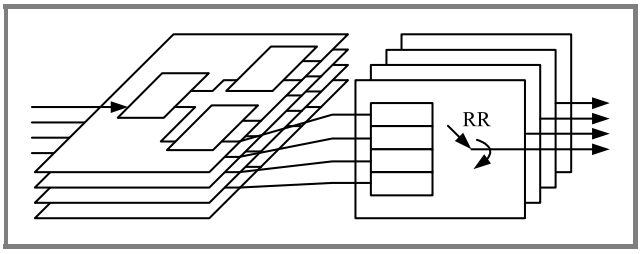


Fig. 3. The switch fabric using the tree architecture.

In the distributed mode the switch fabric is distributed over line cards (or ingress/egress cards). In this case each line card comprises also a segment (or a part) of the switch fabric. The capacity of such segment is $1 \times N$, and for each line card there is N outgoing lines to connect outputs of the $1 \times N$ switch fabric to buffers located on the same and other line cards, and N incoming lines to N output queues (see Fig. 5). The switch fabric based on the tree architecture can be decentralized by putting each $1 \times N$ segment on one line card (compare Fig. 3). The crossbar architecture can be also decentralized in such a way that each row of the crossbar switch fabric (which corresponds to one input – see Fig. 4) is placed on one line card.

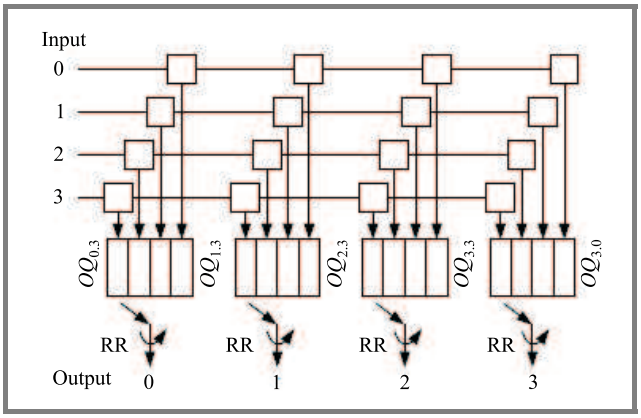


Fig. 4. The switch fabric based on the crossbar architecture.

The drawback of the decentralized architecture described above is the number of outputs from line cards. We need $2N$ lines (N incoming and N outgoing) for each line card. This number may be reduced by putting the switch fabric on the output side of the line card, as it is shown in Fig. 6. Connection lines between cards work as busses and arriving packets are broadcast from inputs to all outputs. Address filters AF at each line card determine whether respective packets are destined to the output. Cells directed to the given output are passed through the address filters to the output queues. The advantage of this architecture is the reduced number of connecting lines between line cards, which is now equal to N. The number of address filters is N^2, and they should operate with the line speed. The speed of connecting lines between line cards is also equal to the line speed.

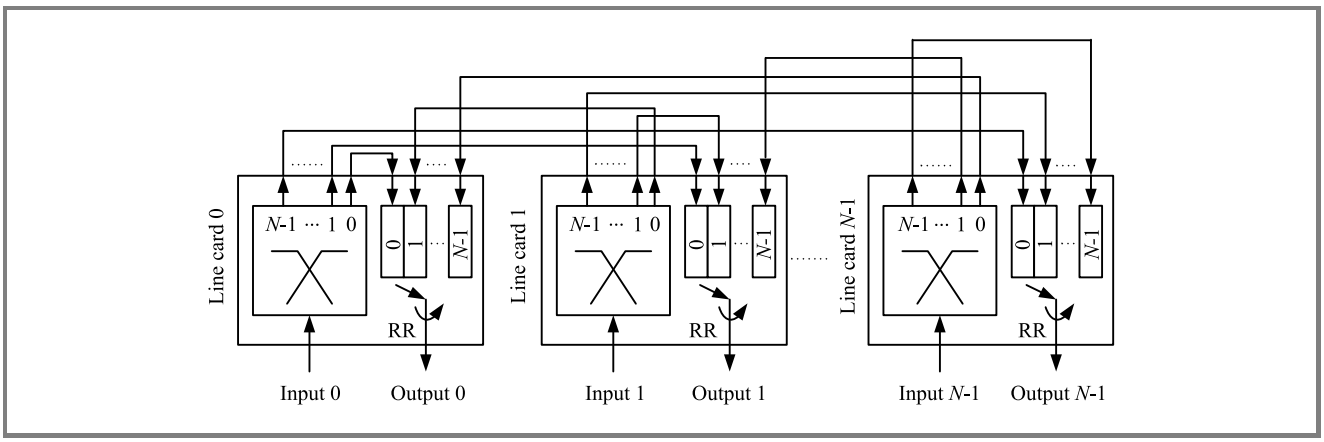


Fig. 5. The switch architecture with the decentralized switch fabric – version 1.

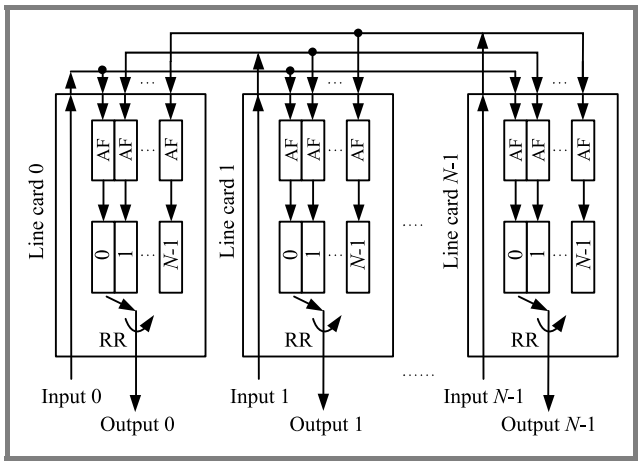


Fig. 6. The switch architecture with the decentralized switch fabric – version 2.

2.3. Buffer arrangements

Buffers in output ports are arranged into N separate queues. When N packets from N input ports are directed to one output port in the same time slot, each packet is written to the different queues. Therefore, the memory speed is the same as the line speed. Buffers may be arranged as separate queues with independent write pointers or as a memory bank with one pointer which points the same memory cells in each queue. Packets from N queues in each output port are read out using the round-robin (RR) algorithm. When independent write pointers are used, the round-robin pointer, denoted by RR, is moved to the queue next to those read out in the previous time slot. When packets are written to the same position of the buffers (one write pointer is used), the operation of RR is modified in such a way, that when all packets from the same position (i.e., which were simultaneously written to the buffers) are already read out, the RR is set back to 0. The operation of these two arrangements will be described by means of the following example.

In the first case the separate pointer is assign to each queue. This pointer, denoted by $MP_{j,i}$, points the end of queue $OQ_{j,i}$, where the next incoming packet to output j from input i will be written to. The example for output x is shown in Fig. 7. It is assumed that all queues are empty at the beginning of the first time slot. Pointers are shown by arrows which shows the state of the pointers at the end of respective time slots. In the first time slot two packets (numbered 1 and 2) from inputs 0 and 1 arrive to the considered output x. The round-robin pointer is set to 0 (the HOL packet from $OQ_{x,0}$ has the highest priority). Since buffer $OQ_{x,0}$ is empty, the packet from input 0 is immediately directed to the output, the RR pointer is set to 1, and packet 2 is stored in $OQ_{x,1}$. The state of RR at the end of the time slot is shown in Fig. 7. The pointer of $OQ_{x,1}$ is moved to the next memory cell. In the next time slot packets from inputs 0, 1 and 3 arrive (numbered 3, 4, and 5, respectively). They are stored in respective queues, while packet 2 from $OQ_{x,1}$ is sent out. During the third time slot packets 6, 7 and 8 arrive from inputs 1, 2, and 3, respectively. Since RR is now set to 2 and buffer $OQ_{x,2}$ is empty, packet 7 is sent directly to the output, while packets 6 and 8 are stored in $OQ_{x,1}$ and $OQ_{x,3}$. In the next time slot packet 5 will be sent out from $OQ_{x,3}$. In this example packet 7 is sent before packet 5, but these packets arrive to the considered output from different inputs. The sequence of packets from the same input port is preserve.

In the second case there is one pointer for all queues. This pointer, denoted by MP_j, points to the memory cells in all queues of output j, where the next incoming packets will be written to. The example is shown in Fig. 8. In the first time slot two packets (numbered 1 and 2) from inputs 0 and 1 arrive to the considered output x. The round-robin pointer is set to 0 (the HOL packet from $OQ_{x,0}$ has the highest priority). Since buffer $OQ_{x,0}$ is empty, the packet from input 0 is immediately directed to the output, packet 2 is stored in $OQ_{x,1}$, the MP_x is moved to the next memory cells in all queues (shown by arrows in Fig. 8), and the RR pointer is set to 1 (here also the state of RR is

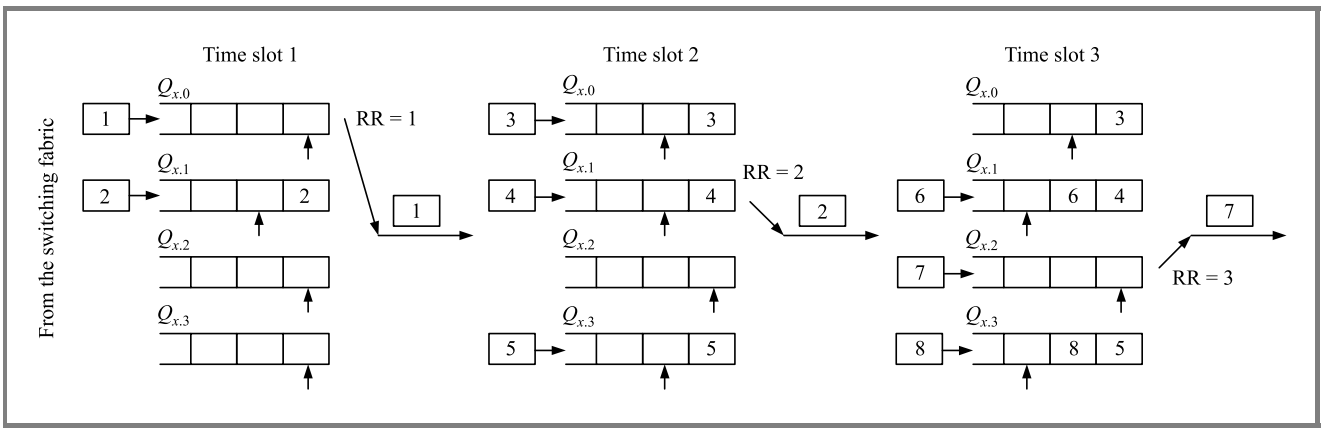


Fig. 7. The example of buffer operation with separate pointers.

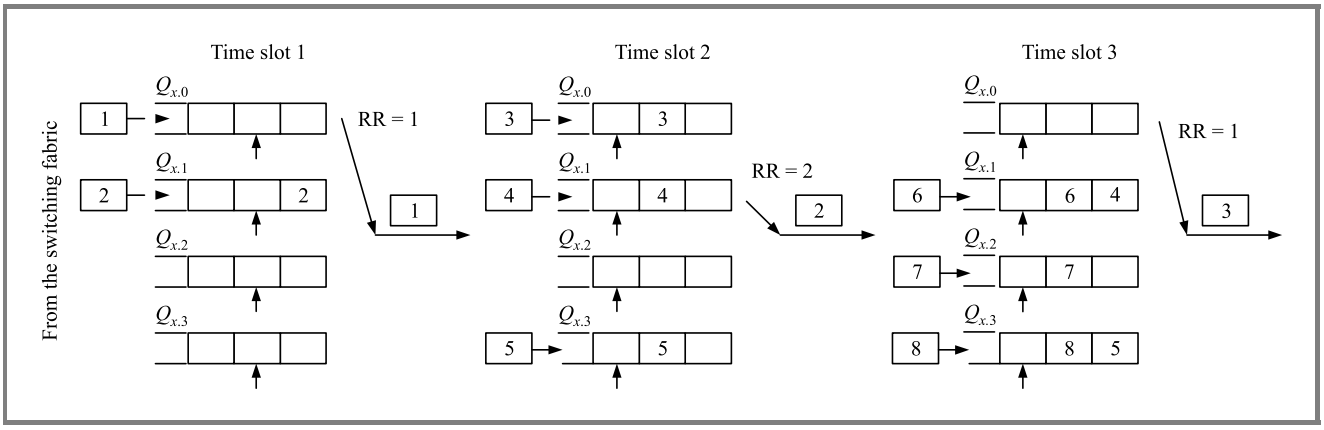


Fig. 8. The example of buffer operation with one pointer.

shown at the end of the time slot). In the next time slot packets from inputs 0, 1 and 3 arrive (numbered 3, 4, and 5, respectively). They are stored in the second memory cell of respective queues, while packet 2 from $OQ_{x,1}$ is sent out. After this packet is read out, there is no any packet in the first memory cell in all queues. Therefore, the next cells in the queues are moved to the HOL position, and the RR is set to 0. During the third time slot packets 6, 7 and 8 arrive from inputs 1, 2, and 3, respectively. Since RR is now set to 0, packet 3 from $OQ_{x,0}$ is sent to the output, while new packets are written to the buffer. In the next three time slots packets 4, 5, and 6 will be sent out from $OQ_{x,1}$, $OQ_{x,3}$, and $OQ_{x,1}$, respectively.

In this second approach all packets which arrive to the given output are written in the same position of each buffer. So we can use only such positions where all memory cells are empty. When in the given time slot less than N packets arrive to the output, some memory cells will be empty and they could not be used to store packets until all packet in the same position of all buffers are read out. Therefore, the memory is not used as efficiently as in the first approach. In the next section only the performance of this first approach will be evaluated.

3. Performance evaluation

In order to evaluate performance measures for the proposed MOQ switch architecture, the corresponding simulation researches have been conducted. The researches have been carried out for the switch with a size of $N \times N$ ($N = 8$) for the following values of traffic load for an input port: $p = 0.6; 0.7; 0.8; 0.9$. We have assumed that offered traffic is uniformly distributed for N outputs (uniform traffic). We have further assumed that the service time of each cell is deterministic and equal to one.

The results of the simulations are shown in the charts (Figs. 9 and 10) in the form of marks with 95% confidence intervals that have been calculated after the t–student distribution for the five series with 10,000,000 time slots. For each of the points of simulation the value of the confidence interval is at least one order lower than the mean value of the results of the simulation. In many cases the value of the confidence interval is lower than the height of the sign used to indicate the value of the simulation experiment.

We have evaluated two performance measures for switch architectures, i.e., mean waiting time (mean cell delay) and cell loss probability (CLP). The obtained performance

measures of MOQ architecture have been compared with performance of OQ architecture [6] and VOQ architecture (algorithm iSLIP with one and four iterations) [7].

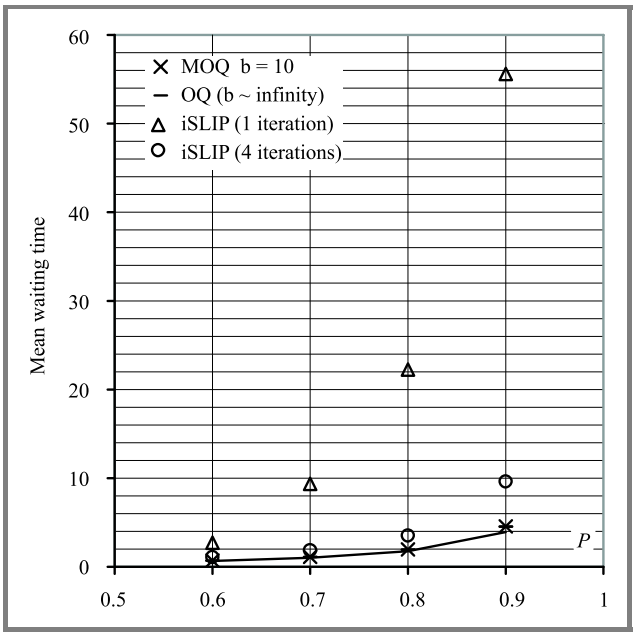


Fig. 9. Mean waiting time (in time slots), $N = 8$.

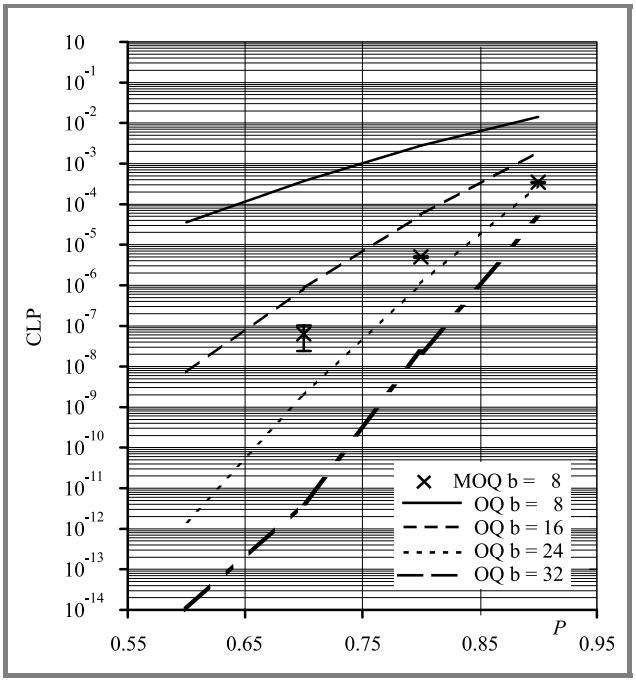


Fig. 10. Cell loss probability, $N = 8$.

Figure 9 plots the mean waiting time (in time slots) of the MOQ switch as a function of input load p. The presented results have been obtained for the switch in which the size of each of $N = 8$ output buffers (of the tagged output port) was limited to $b = 10$. The adopted buffer size assures – for each value of traffic load – stable values of mean waiting times (the application of larger buffers do not lead to increase in values of waiting time). The simulation results enabled us to compare the MWT values in the proposed MOQ switch architecture with the results obtained for OQ architecture. For OQ switch we have assumed that the buffer size is large enough to get stable values of MWT parameter. We can notice that both architectures are comparable. This phenomenon results from similar characteristics of both FIFO discipline for single queue and round robin discipline for cyclic-service set of queues. Additionally, Fig. 9 shows the performance of iSLIP algorithm in virtual output queuing architecture. It is evident from the presented results that – regardless of the number of iterations in iSLIP algorithm – the VOQ switch architecture is charcterised by higher values of MWT than the proposed MOQ switch architecture.

Another important performance measure for packet switches is the cell loss probability. Figure 10 compares the results of CLP obtained for MOQ switch with the results calculated for the switch with output queuing. It is intuitively clear, that the proposed switch architecture requires greater total number of memory cells (N buffers for each output port) in order to keep the same value of CLP parameter as in the case of switches with single output queue for each output port.

4. Comparison

In the previous section MWT and CLP in MOQ, OQ and VOQ switches were compared. Now we compare a hardware complexity of these architectures. This comparison is summarized in Table 1. The MOQ switch uses the same number of buffers as VOQ switch and of the same speed as the line rate. The OQ switch comprise only N buffers, but they have to be N times faster than the line speed.

Table 1
The hardware complexity of different buffering strategies

Parameters	OQ	VOQ	MOQ
Number of buffers	N	N^2	N^2
Memory speed (in line speed)	N	1	1
Switch fabric capacity	$N \times N$	$N \times N$	$N \times N^2$
Switch fabric speed	N	1	1
Switch fabric hardware	N^2	N^2	N^2
Number of schedulers	–	$2N$	N
Wiring complexity	N	N^2	N^2

The switch fabric speed in MOQ is also the same as line speed, and the same is true for VOQ switch, provided that no speed-up is used to increase the performance of the VOQ switch. In OQ switch the switch fabric is N times faster. However, in MOQ architecture the switch fabric has the capacity of $N \times N^2$ instead of $N \times N$. But this greater capacity does not result in greater hardware complexity, since MOQ switch require the same number of switching elements (when crossbar architecture is considered) as OQ or VOQ switches.

The important issue is the packet scheduling mechanism and wiring complexity. The OQ switches do not need packet schedulers and the wiring complexity is $O(N)$. On the other hand, VOQ switches need $2N$ schedulers when iterative maximal matching algorithm is used (one scheduler in each input port and one in each output port). These schedulers are to be connected between themselves so the wiring complexity is $O(N^2)$. The similar complexity is needed when one centralized scheduler is used, since each VOQ has to be connected with the scheduler to send request signal when it has a HOL packet. The MOQ architecture has the same wiring complexity but lines are used to connect MOQs with the switch fabric, instead of connecting VOQs and schedulers. The MOQ switches need also N schedulers, one for each output, but there is no need to connect schedulers between themselves. The MOQs of the given output are to be connected to the scheduler of this output and this is done inside the output port (a line card or an egress card).

Comparing the hardware complexity and the performance of the switches we can say, that the MOQ architecture is attractive and worth considering in constructing high-speed and high-capacity switches. The hardware complexity is very similar to VOQ switches but the performance of the MOQ architecture is much better, at least when uniform traffic is considered. Simulation results shows, that the performance of the MOQ switch is very similar to the OQ switch.

5. Conclusions

We have proposed the new packet switch architecture which uses multiple output queueing. This architecture looks attractive for constructing high-speed packet switches. The hardware complexity of this architecture is very similar to VOQ switch but its performance is comparable to OQ switch. This paper contains the first considerations and the first results we obtained for this architecture. The architecture is also very promising since it can naturally support multicast traffic. Further research are needed to evaluate the performance of the MOQ switch under other traffic types (non-uniform, hot-spot), the buffer length evaluation, as well as the practical buffer implementation in either separate chip or in the switch fabric.

References

[1] K. Yoshigoe and K. J. Christensen, "An evolution to crossbar switches with virtual ouptut queuing and buffered cross points", *IEEE Network*, vol. 17, no. 5, pp. 48–56, 2003.

[2] M. K. Karol, M. Hluchyj, and S. Morgan, "Input versus output queuing on a space-division packet switch", *IEEE Trans. Commun.*, vol. 35, pp. 1347–1356, 1987.

[3] J. Xie and Ch.-T. Lea, "Speedup and buffer division in input/output queuing ATM switches", *IEEE Trans. Commun.*, vol. 51, no. 7, pp. 1195–1203, 2003.

[4] Y. Tamir and G. Frazier, "High performance multi-queue buffers for VLSI communications switches", in *Proc. Comput. Archit.*, Honolulu, Hawaii, United States, 1988, pp. 343–354.

[5] T. Anderson *et al.*, "High-speed switch scheduling for local-area networks", *ACM Trans. Comput. Syst.*, vol. 11, no. 4, pp. 319–352, 1993.

[6] H. J. Chao, C. H. Lam, and E. Oki, *Broadband Packet Switching Technologies: A Practical Guide to ATM Switches in IP Routers.* New York: Wiley, 2001.

[7] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, "Achieving 100% throughput in input-queued switches", *IEEE Trans. Commun.*, vol. 47, no. 8, pp. 1260–1267, 1999.

[8] H. J. Chao, "Saturn: a terabit packet switch using dual round-robin", *IEEE Commun. Mag.*, vol. 38, no. 12, pp. 78–84, 2002.

[9] E. Oki, R. Rojas-Cessa, and H. J. Chao, "A pipeline-based approach for maximal-sized matching scheduling in input-buffered switches", *IEEE Commun. Lett.*, vol. 5, no. 6, pp. 263–265, 2001.

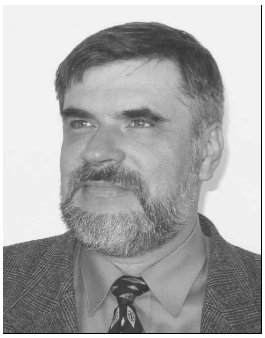
[10] B. Kraimeche, "Design and analysis of the stacked-banyan ATM siwtch fabric", *Comput. Netw.*, vol. 32, no. 2, pp. 171–184, 2000.

Grzegorz Danilewicz was born in Poznań, Poland, in 1968. He received the M.Sc. and Ph.D. degree in telecommunication from the Poznań University of Technology (PUT), Poland, in 1993 and 2001, respectively. Since 1993 he has been working in the Institute of Electronics, Poznań University of Technology, where he currently is an Assistant Professor. His scientific interests cover photonic broadband switching systems with special regard to the realization of multicast connections in such systems. He is a member of the IEEE Communication Society. He has published 25 papers.

e-mail: gdanilew@et.put.poznan.pl
Institute of Electronics and Telecommunications
Poznań University of Technology
Piotrowo st 3A
60-965 Poznań, Poland

Mariusz Głąbowski was born in Turek, Poland, in 1973. He received the M.Sc. and Ph.D. degrees in telecommunication from the Poznań University of Technology (PUT), Poland, in 1997 and 2001, respectively. Since 1997 he has been working in the Institute of Electronics and Telecommunications, Poznań University of Technology, where he currently is an Assistant Professor. He is engaged in research and teaching in the area of performance analysis and modelling of multiservice networks and switching systems. He has published papers.

e-mail: mglabows@et.put.poznan.pl
Institute of Electronics and Telecommunications
Poznań University of Technology
Piotrowo st 3A
60-965 Poznań, Poland



Wojciech Kabaciński received the M.Sc., Ph.D., and D.Sc. degrees in communication from Poznań University of Technology (PUT), Poland, in 1983, 1988 and 1999, respectively. Since 1983 he has been working in the Institute of Electronics and Telecommunications, Poznań University of Technology, where he currently is an Associate Professor. His scientific interests cover broadband switching networks and photonic switching. He has published three books, 78 papers and has 10 patents. Prof. Kabaciński is a member of the IEEE Communication Society and the Association of Polish Electrical Engineers.
e-mail: kabacins@et.put.poznan.pl
Institute of Electronics and Telecommunications
Poznań University of Technology
Piotrowo st 3A
60-965 Poznań, Poland

Janusz Kleban was born in Pobiedziska, Poland. He received the M.Sc. and Ph.D. degrees in telecommunications from the Poznań University of Technology (PUT) in 1982 and 1990, respectively. From August 1982 to November 1983 he was with Computer Centre for Building Industry in Poznań where he worked on data transmission systems. He has been with Institute of Electronics and Telecommunications at PUT, where he currently is an Assistant Professor, since December 1983. He is involved in research and teaching in the areas of computer networks, switching networks, broadband networks and various aspects of networking. He is author and co-author of many publications and unpublished reports.
e-mail: jkleban@et.put.poznan.pl
Institute of Electronics and Telecommunications
Poznań University of Technology
Piotrowo st 3A
60-965 Poznań, Poland

Paper

Integrated analysis of communication protocols by means of PLA formalism

Henrikas Pranevicius

Abstract—Aggregate approach and its possibilities for specification and analysis of computer network protocols are presented. The theoretical basis of the aggregate approach is a piece-linear aggregate (PLA) for formal specification of systems. The advantage of that approach is that it permits to create models both for analysis correctness of specifications and simulation. Some methods that can be used for validation and verification of aggregate specifications are presented also.

Keywords—piece-linear aggregates, ESTELLE/Ag specification language, validation, simulation, communication protocols.

1. Introduction

The stage of formal specification is one of the most important during the design of software of communication protocols. Such formal specification is usually used for analysis and implementation purposes. In the stage of analysis it is necessary to resolve two tasks: analysis of logical correctness and evaluation of the system functioning parameters.

Different mathematical schemes are used for creating formal descriptions of systems, such as: different automate models, Petri-nets, data flow and state transition diagrams, temporal logic technique, abstract communicating methods and other [1, 2].

When a formalization method is chosen, it is desirable that both above mentioned analysis tasks could be resolved on the bases of a single formal description. The aggregate approach has such property and it has been successfully used both for correctness analysis and for simulation of computer network protocols [3–5]. Specification language ESTELLE/Ag and the specifications analysis tool PRANAS-2 have been created on the base of the aggregate method (Ag). There are some differences between ESTELLE/Ag and the ESTELLE standard ISO: the piece-linear aggregate model is used in ESTELLE/Ag. The use of such a model instead of a finite-state automate, which is the formal background of the standard ESTELLE, enables to create models both for validation and simulation. This is possible due to the special structure of the piece-linear aggregate. Apart from the discrete components describing the state of the modules, there are also continuous components to control event-sequences in the module. These continuous components are called operations. By means of operators, sequences of actions are described, the intermediate results of which are invisible on the outside. If such operation sequence is being performed at a given instance of time the corresponding operation is called "active". Thus, an individual module involves two types of events: arrival of an input signal and completion of an active operation. The specification analysis system PRANAS-2 consists of the following software tools: a specification editor, a validation subsystem and a simulation subsystem. The editor provides the capability to create a specification in ESTELLE/Ag. The validation subsystem permits to construct a validation model for the program generating the reachability graph. After completing the construction of the reachability graph, it is possible to verify the following specification characteristics: completeness, deadlock freeness, boundedness, absence of static deadlock, absence of dynamic deadlock, termination.

The same specification changes are carried out when the simulation model is creating. This is necessary in order to define the duration of operations and to introduce additional variables for gathering statistics about the evaluated system parameters.

Section 2 describes the general principles of piece-linear aggregates (PLA) formalism. Methods used for correctness analysis of PLA specification are presented in Section 3. Section 4 illustrates the use PLA formalism for formal specification and integrated analysis of event driven local computer network protocol.

2. General principles of the aggregate approach

In the application of the aggregate approach for system specification, the system is represented as a set of interacting piece-linear aggregates. The PLA is taken as an object defined by a set of states Z, input signals X, and output signals Y. The aggregate functioning is considered in a set of time moments $t \in T$. The state $z \in Z$, the input signals $x \in X$, and the output signals $y \in Y$ are considered to be time functions. Apart from these sets, transition H and output G operators must be known as well.

The state $z \in Z$ of the piece-linear aggregate is the same as the state of a piece-linear Markov process, i.e., $z(t) = \big(\upsilon(t), z_\upsilon(t)\big)$, where $\upsilon(t)$ is a discrete state component taking values on a countable set of values; and $z_\upsilon(t)$ is a continues component comprising of $z_{\upsilon 1}(t)$, $z_{\upsilon 2}(t), \ldots, z_{\upsilon k}(t)$ co-ordinates.

When there are no inputs, the state of the aggregate changes in the following manner:

$$\upsilon(t) = \text{const}, \frac{dz_\upsilon(t)}{dt} = -\alpha_\upsilon,$$

where $\alpha_\upsilon = (\alpha_{\upsilon 1}, \alpha_{\upsilon 2}, \ldots, \alpha_{\upsilon k})$ is a constant vector.

The state of the aggregate can change in two cases only: when an input signal arrives at the aggregate or when a continuous component acquires a definite value. The theoretical basis of piece-linear aggregates is their representation as piece-linear Markov processes.

Aggregate functioning is examined on a set of time moments $T = \{t_0, t_1, \ldots, t_m, \ldots\}$ at which one or several events take place, resulting in the aggregate state alternation. The set of events E which may take place in the aggregate is divided into two non-intersecting subsets $E' = E' \cup E''$. The subset $E' = \{e'_1, e'_2, \ldots, e'_N\}$ comprises classes of events (or simply events) e'_i, $i = \overline{1,N}$ resulting from the arrival of input signals from the set $X = \{x_1, x_2, \ldots, x_N\}$. The class of events $e''_i = \{e''_{ij}, j = 1,2,3,\ldots\}$, where e''_{ij} is an event from the class of events e''_i taking place the ***j***th time since the moment t_0. The events from the subset E' are called external events. A set of aggregate input signals is unambiguously reflected in the subset E', i.e., $X \to E'$. The events from the subset $E'' = \{e''_1, e''_2, \ldots, e''_f\}$ are called internal events, where $e''_i = \{e''_{ij}, j = 1,2,3,\ldots\}$, $i = \overline{1,f}$ are the classes of the aggregate internal events. Here, f determines the number of operations taking place in the aggregate. The events in the set E'' indicate the end of the operations taking place in the aggregate.

The events of the subsets E' and E'' are called the evolutionary events of the aggregate. The main evolution events are sufficient for unambiguous determination of the aggregate evolution. Apart from the basic evolutionary events, auxiliary evolutionary events may be considered, which are simultaneous to the basic ones and determine the start of the operations.

For every class of events e''_i from the subset E'', control sequences are specified $\{\xi_j^{(i)}\}$, where $\xi_j^{(i)}$ – the duration of the operation, which is followed by the event e''_{ij} as well as event counters $\{r(e''_i, t_m)\}$, where $r(e''_i, t_m)$, $i = \overline{1,f}$ is the number of events from the class e''_i taken place in the time interval $[t_0, t_m]$.

In order to determine start and end moments of operation, taking place in the aggregate the so-called control sums $\{s(e''_i, t_m)\}$, $\{w(e''_i, t_m)\}$, $i = \overline{1,f}$ are introduced, where $s(e''_i, t_m)$ – the time moment of the start of operation followed by an event from the class e''_i. This time moment is indeterminate if the operation was not started; $w(e''_i, t_m)$ is the time moment of the end of the operation followed by the event from the class e''_i. In case of no priority operations, the control sum $w(e''_i, t_m)$ is determined in the following way: $w(e''_i, t_m) = s'(e''_i, t_m) + \xi_{r(e''_i, t_m)+1}$, if at moment t_m an operation is taking place, which is followed by the event e_i; in the opposite case $w(e''_i, t_m) = \infty$. The infinity symbol (∞) is used to denote the undefined values of the variables.

Control sums determine only the possibility conditions for the events after the moment t_m, while the event occurrence moments are not determined.

Let us specify the meaning of the co-ordinates of the aggregate state. The discrete component of the state, $\upsilon(t_m) = \{\upsilon_1(t_m), \upsilon_2(t_m), \ldots, \upsilon_p(t_m)\}$, presents the system state:

$$z_\upsilon(t_m) = \{w(e''_i, t_m), w(e''_2, t_m), \ldots, w(e''_f, t_m)\}$$

are control co-ordinates specifying the moment of evolutionary events occurrence.

The control co-ordinate $w(e''_i, t_m)$ corresponds to every each e''_i from the subset of events E'', while always $w(e''_i, t_m) \geq t_m$.

The state co-ordinates $z(t_m)$ can change their values only at discrete time moments $\boldsymbol{t_m}$, $m = 1,2,\ldots$ of event occurrence, remaining fixed in each interval $[t_m, t_{m+1}), m = 0,1,2,\ldots$ where t_0 – the initial moment of system functioning.

When the state of the system $z(t_m)$, $m = 0,1,2,\ldots$, is known, the moment t_{m+1} of the following event is determined by a moment of input signal arrival to the aggregate or by the equation:

$$t_{m+1} = \min\{w(e''_i, t_m)\},\ 1 \leq i \leq f.$$

Class of the next event e_{m+1} is specified by an input signal if it arrives at the time moment t_{m+1} or is determined by the control co-ordinate, which acquire minimal value at the moment t_m, i.e., if $w(e''_i, t_m)$ acquires minimal value, then $e_{m+1} = e''_i$.

The operator H states the new aggregate state:

$$z(t_{m+1}) = H[z(t_m), e_i],\ e_i \in E' \cup E''.$$

The output signals y_i from the set of output signals $Y = \{y_1, y_2, \ldots, y_m\}$ can be generated by an aggregate only at occurrence moments of events from the subsets E' and E''. The operator G determines the content of the output signals:

$$y = G[z(t_m), e_i],\ e_i \in E' \cup E'',\ y \in Y.$$

Further transition and output operators will be denoted $H(e_i)$ and $G(e_i)$.

3. Correctness analysis of aggregate specifications

3.1. Reachable states approach for aggregate model validation

An essence of the reachable states method is a use of the global state which is considered as a joint state of a system after aggregate system composition. A graph of the reachable states is created as oriented one: its nodes stand for global states of the system, its arcs indicate the possible transitions from one state to another. Initial and final states must be specified in working out the graph. The resulting

states graph is used for an analysis of defined properties of a system, as some of them are closely related with the graph structure. The given validation method allows to investigate general properties of a system such as boundedness, absence of redundancy in specification, completeness, absence of static deadlocks, absence of dynamic deadlocks, termination.

3.2. Invariant approach for aggregate model validation

A system invariant (I) is the assertion, which describes correct system functioning and it must remain true in spite of the events taking place and system transition from one state to another.

The essence of the method is as follows: assertions are formulated in relation to the co-ordinates of the aggregate model so as to express the requirements for the system functioning.

On the base of a conceptual model of an analysed system we can describe system functioning by the event sequence, which may be represented by the graph G(V), where V is a set of vertices and $\mathbf{A} = \{a_{ij}\}$ is an adjacency matrix. In this case $V = \{e_1, e_2, \ldots, e_n\}$, where e_i is ith event, n is a number of events. $(e_i e_j) \neq (e_j e_i)$, i.e., the graph is oriented.

The set of states, which the system may enter after the event e1, is called as the ith set of possible states (SS_i – symbolic state). $SS_i = \left\{ z \in Z | (\exists z') \left((z' \in Z) \wedge EP_i(z') \wedge (z = H_i(z', P)) \right) \right\}$, where Z is a set of all possible system states, $EP_i(z')$ is an enabling predicate of the event $\boldsymbol{e_i}$ in the state z', P is a set of probabilistic parameters of the system and H_i is a transition operator determining the new system state when the event e_i occurs.

The system considered being in the symbolic state SS_i only if it is in the state z and $z \in SS_i$. Relying this SS_i definition, every event e_i is related to the symbolic state SS_i, therefore replacing the set of vertices V in the graph G(V) by $V' = \{SS_1, SS_2, \ldots, SS_n\}$ while the adjacency matrix $\mathbf{A}$ remains unchanged. We obtain the graph of symbolic states G(V') which describes the system operation by determining the possible set of states and transitions from one symbolic state to another.

The presented formalization and analysis method will be illustrated by example of specification and integrated analysis of timed protocol with slot reuse.

4. Specification, validation and simulation of event-driven local computer network protocol

4.1. Conceptual model of on event-driven local computer network protocol

There are many computer communication applications requiring high bandwidth and high reliability in operation, which still allow simple and low cost implementation. This type of network exists in robotics, vehicles, homes, etc. These applications set restrictions on the system in terms of usable hardware, cost, and cabling. Such networks are in many cases meant for one special application and not for a general purpose use. The number of stations is generally small compared with typical LAN applications, and the variation in the number of stations is small during the life cycle of the network.

Typical requirements for the media access protocols in these applications are: high reliability of the environment, where the electrical disturbance level is a high scalable bandwidth: self-stabilizing properties; and simplicity combined with low cost of implementation. Solutions based on the existing media access standards do not meet these requirements in many cases.

The protocol described by Sintonen is design to offer high bandwidth while keeping the structure simple. The configuration is a physical bus, where stations form a logical ring. The algorithm is based on the noticeable events on the bus (hence the name event-driven bus protocol). The protocol is distributed, except in the initialization phase. Every station listens to the bus and receives both the destination address and the source address, and stores them in the registers *DA* and *SA* respectively. A station is also capable of sending the bus and detecting the event frame ended. The algorithm for sending and receiving is as follows.

Receiving:
When a station notices it's own address in the *DA* field, it receives the frame.

Sending:
When a station has a frame to send, it waits until it receives the address of it's predecessor in the *SA* of the frame. Then it waits for the event frame ended. After that event, it waits a time period $D', D' \leq 2d$, where d is the end to end delay of, the bus. Then it sends its frame, and waits for a time delay $D', D' > 2d$ to hear the next station begin sending. When this happens, the sending phase is ended. If a station has nothing to send its turn comes, it sends an empty no data frame, a kind of a token, to pass the turn to next station in sequence.

There is one station which initializes the ring, known as the fixed control station. The control station can also detect a failed station and is capable of executing a reconfiguration algorithm to restore the normal operation of the ring.

5. Aggregate specification of on event-driven local computer network protocol

An aggregate schemes of a specification of an analyzed event oriented protocol is depicted in Fig. 1. The aggregates $Station\_0, Station\_1, \ldots, Station\_(n-1)$ depict the stations

which are switched on to the network, and the aggregate *Bus* describes the performance channel. *Station_0* is the controlling one. The signals that are transmitted between the aggregates have also been shown in Fig. 1.

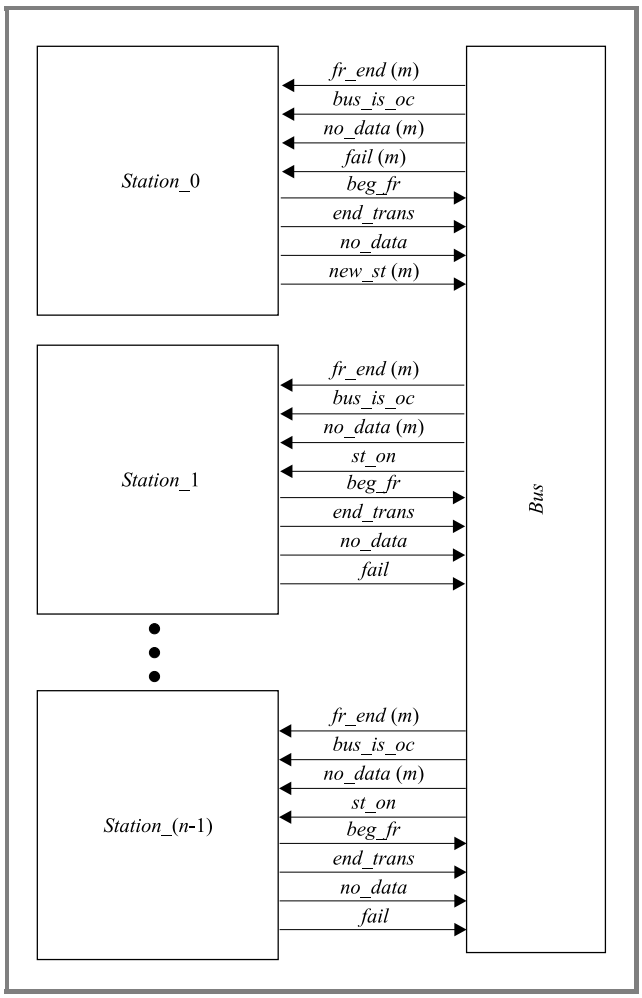


Fig. 1. Aggregate scheme of a model.

Aggregate *Station_nr*, $nr = \overline{1, n-1}$

1. Set of input signal

$$X_{nr} = \{fr\_end(m),\ bus\_is\_oc,\ no\_data(m),\ fail\};$$

where: $fr\_end(m)$ – end of the transmitting; *bus_is_oc* – bus is occupied; *no_data_*(m) – no data for transmission; *st_on* – switching on of the station; n – number o station; m – the number of station where packet is sending.

2. Set of output signals $Y_{nr} = \{y\}$,

$$y \in \{beg\_fr,\ end\_trans,\ no\_data,\ fail\};$$

where: beg_fr – beginning of the frame transmitting; *end_trans* – end of the frame transmitting; *no_data* – no data for transmitting; $fail$ – station is switched off.

3. Set of internal events

$$E''_{nr} = \{e''_1(taim\_DI),\ e''_2(taim\_D),\ e''_3(trans\_fr),$$
$$e''_4(arr\_fr),\ e''_5(swit\_of)\};$$

where: $e''_1(taim\_DI)$ – end of timer *DI*; $e''_2(taim\_D)$ – end of timer *D*; $e''_3(trans\_fr)$ – end of the frame transmitting; $e''_4(arr\_fr)$ – moment of a frame arrival; $e''_5(swit\_of)$ – moment of the station switching.

4. Controlling sequences:

$$e''_i(\ldots) \to \{\xi_{ij}\},\ i = \overline{1, 5},\ j = \overline{1, \infty};$$

where ξ_{ij} – duration of an operation, followed by the event $e''_i(\ldots)$.

5. Discreet component of state

$$\upsilon(t_m) = \{st(t_m),\ actD(t_m),\ sw(t_m)\};$$

where: $st(t_m) \in \{0, 1\}$; 0 – no frame for transmitting, 1 – there is a frame for transmitting;

$$actD(t_m) = \begin{cases} 0, & \text{timer } D \text{ is switched off;} \\ 1, & \text{timer } D \text{ is switched on;} \end{cases}$$

$$sw(t_m) = \begin{cases} 0, & \text{station is switched off;} \\ 1, & \text{station is switched on;} \end{cases}$$

6. Initial state: $st(t_0) = 0$; $act\_D(t_0) = 0$; $sw(t_0) = 0$;
$w\left(e''_1(taim\_DI), t_0\right) = \infty$;
$w\left(e''_2(taim\_D), t_0\right) = \infty$;
$w\left(e''_3(trans\_fr), t_0\right) = \infty$;
$w\left(e''_4(arr\_fr), t_0\right) = t_0 + \xi_{4j}$;
$w\left(e''_5(swit\_of), t_0\right) = t_0 + \xi_{5j}$.

7. Transfer operators:
$H\left(e'(fr\_end)\right)$: (The end of packet sending)
$w\left(e''_1(taim\_DI), t_{m+1}\right) = t_m + \xi_{1j}$
if $sw(t_m) = 1 \wedge m = nr$.
$H\left(e'(bus\_is\_oc)\right)$: (Bus is busy)
$$\left.\begin{array}{l} w\left(e''_2(taim\_D), t_{m+1}\right) = \infty, \\ w\left(e''_4(arr\_fr), t_{m+1}\right) = t_m + \xi_{4j}, \\ act\_D(t_{m+1}) = 0 \end{array}\right\},$$
if $sw(t_m) = 1 \wedge act\_D(t_m) = 1$.
$H\left(e'(no\_data)\right)$: (There are no data for sending)
$w\left(e''_1(taim\_DI), t_{m+1}\right) = t_m + \xi_{1j}$
if $sw(t_m) = 1 \wedge m = nr$;
$$\left.\begin{array}{l} w\left(e''_2(taim\_D), t_{m+1}\right) = \infty, \\ w\left(e''_4(arr\_fr), t_{m+1}\right) = t_m + \xi_{4j}, \\ act\_D(t_{m+1}) = 0 \end{array}\right\},$$
if $sw(t_m) = 1 \wedge act\_D(t_m) = 1$.
$H\left(e''_1(taim\_DI)\right)$: (Timer *DI* has expired)
$$\left.\begin{array}{l} w\left(e''_3(trans\_fr), t_{m+1}\right) = t_m + \xi_{3j}, \\ y = beg\_fr \end{array}\right\},$$
if $st(t_{m+1}) = 1$;

$$\left.\begin{array}{l} w\left(e_2''(taim\_D), t_{m+1}\right) = t_m + \xi_{2j}, \\ act\_D(t_{m+1}) = 1, \\ y = no\_data \end{array}\right\},$$
if $st(t_{m+1}) \neq 1$.

$H\left(e_2''(taim\_D)\right)$: (Timer D has expired)
$y = fail$.

$H\left(e_3''(trans\_fr)\right)$: (The end of packet sending)
$st(t_{m+1}) = 0$;
$w\left(e_2''(taim\_D), t_{m+1}\right) = t_m + \xi_{2j}$;
$act\_D(t_{m+1}) = 1$;
$y = end\_trans$.

$H\left(e_4''(arr\_fr)\right)$: (The packet has arrived)
$st(t_{m+1}) = 1$.

$H\left(e_5''(swit\_of)\right)$: (The station is seething of)
$sw(t_{m+1}) = 0$;
$w\left(e_1''(taim\_DI), t_{m+1}\right) = \infty$;
$w\left(e_2''(taim\_D), t_{m+1}\right) = \infty$;
$w\left(e_3''(trans\_fr), t_{m+1}\right) = \infty$;
$w\left(e_4''(arr\_fr), t_{m+1}\right) = \infty$;
$act\_D(t_{m+1}) = 0$;
$st(t_{m+1}) = 1$.

Aggregate *Station_0*

The functioning of this aggregate is similar to that of the aggregate ***Station_nr***. Therefore, only the differences are presented in respect to the agregate ***Station_nr***.

1. Set of input signals:
   $X_0 = X_{nr} \setminus \{st\_on\} \cup \{fail(m)\}$;
   where: X_{nr} – set of input signal of aggregate *Station_nr*; m – is the number of the stations switched on.

2. Set of output signals:
   $Y_0 = Y_{nr} \setminus \{fail\} \cup \{new\_st(m)\}$;
   where: Y_{nr} – set of output signal of aggregate *Station_nr*; m – the number of the switched on station.

3. Set of internal events:
   $E_0'' = E_{nr}'' \setminus \left\{e_5''(swit\_off),\ e_7''(taim\_T)\right\}$
   $\cup \left\{e_{8i}''(swit\_on), \ldots, e_{8,n-1}''(swit\_on)\right\}$;
   where: $e_7''(taim\_T)$ – end of timer T; $e_{8i}''(swit\_on)$ – ith station switched on.

4. Controlling sequences for the events are introduced $e_7''(\ldots)$ and $e_{8i}''(\ldots)$:
   $e_7''(taim\_T) \mapsto \{T\}$;
   $e_{8i}''(swit\_on) \mapsto \{\xi_{ij}\},\ i = \overline{1, n-1},\ j = \overline{1, \infty}$;
   where: ξ_{8ij} – the operation duration after finishing of which the ith station is switched on; T – the duration of timer T.

5. Discrete component of state
   $\upsilon(t_m) = \left\{st(t_m),\ actD(t_m)\right\}$.

6. Initial state:
   $act\_D(t_0) = 1;\ st(t_m) = 0$;
   $w\left(e_7''(taim\_D),\ t_0\right) = t_0 + T$;
   $w\left(e_{8i}''(swit\_on), t_0\right) = \infty,\ i = \overline{1, n-1}$.

7. Transfer operators:
   $H\left(e'(fr\_end)\right)$: (Bus is busy)
   $$\left.\begin{array}{l} w\left(e''(taim\_DI),\ t_{m+1}\right) = t_m + \xi_{1j}, \\ act\_DI(t_{m+1}) = 1 \end{array}\right\},$$
   if $m = nr$;
   $w\left(e_7''(taim\_T),\ t_{m+1}\right) = t_m + T$,
   if $m \neq nr$.

   $H\left(e'(bus\_is\_oc)\right)$: (Bus is occupied)
   $$\left.\begin{array}{l} w\left(e_7''(taim\_T)\right) = t_m + T \\ w\left(e_2''(taim\_D),\ t_{m+1}\right) = \infty, \\ w\left(e_4''(arr\_fr),\ t_{m+1}\right) = t_m + \xi_{4j} \\ act\_D(t_{m+1}) = 0 \end{array}\right\},$$
   if $act\_D(t_{m+1}) = 1$.

   $H\left(e'(no\_data)\right)$: (There are no data for sending)
   $$\left.\begin{array}{l} w\left(e_1''(taim\_DI),\ t_{m+1}\right) = t_m + \xi_{1j}, \\ w\left(e_7''(taim\_T),\ t_{m+1}\right) = \infty, \\ w\left(e_2''(taim\_D),\ t_{m+1}\right) = \infty, \\ act\_DI(t_{m+1}) = 1, \\ act\_D(t_{m+10}) = 0 \end{array}\right\},$$
   if $m = 0$;
   $$\left.\begin{array}{l} w\left(e_2''(taim\_D),\ t_{m+1}\right) = \infty, \\ w\left(e_4''(arr\_fr),\ t_{m+1}\right) = t_m + \xi_{4j}, \\ act\_D(t_{m+1}) = 0 \end{array}\right\},$$
   if $m = 0 \wedge act\_D(t_m) = 1$;
   $w\left(e_7''(taim\_T), t_{m+1}\right) = t_m + T$.

   $H\left(e'(fail)\right)$: (The station is)
   $w\left(e_{8m}''\left(swit\_on(m)\right)\right) = t_m + \xi_{mj}$.

   $H\left(e_1''(taim\_DI)\right)$: (End of timer DI)
   $act\_DI = 0$;
   $$\left.\begin{array}{l} w\left(e_3''(trans\_fr),\ t_m\right) = t_m + \xi_{3j}, \\ y = beg\_fr \end{array}\right\},$$
   if $st(t_{m+1}) = 1$;
   $$\left.\begin{array}{l} w\left(e_2''(taim\_D),\ t_m\right) = t_m + \xi_{2j}, \\ act\_D(t_{m+1}) = 1, \\ y = no\_date \end{array}\right\},$$
   if $st(t_{m+1}) \neq 1$.

   $H\left(e_2''(taim\_D)\right)$: (The end of timer D)
   $act\_D(t_{m+1}) = 0$;
   $w\left(e_1''(taim\_DI),\ t_{m+1}\right) = t_m + \xi_{1j}$,
   $act\_DI(t_{m+1}) = 1$;
   $w\left(e_4''(arr\_fr),\ t_{m+1}\right) = t_m + \xi_{4j}$;
   $y = fail$.

   $H\left(e_3''(trans\_fr)\right)$: (The transmission of packet has ended)
   $st(t_{m+1}) = 0$;
   $w\left(e''(taim\_D),\ t_{m+1}\right) = t_m + \xi_{2j}$,
   $act\_D(t_{m+1}) = 1$;
   $y = end\_trans$.

   $H\left(e_4''(arr\_fr)\right)$: (The packet has arrived)
   $st(t_{m+1}) = 1$.

$H\left(e''_7(taim\_T)\right)$: (The timer T has expired)
$w\left(e''_1(taim\_DI),\ t_{m+1}\right) = t_m + \xi_{1j}$;
$act\_DI(t_{m+1}) = 1$.

$H\left(e''_{8k}\left(swit\_on(k)\right)\right)$: (The station is switching on)
$y = new\_st(k+1)$.

Aggregate *Bus*

1. Set of input signals:
   $X = \{[beg\_fr,\ end\_trans,\ no\_data,\ new\_st(m)]_0,$
   $[beg\_fr,\ end\_trans,\ no\_data,\ fail]_1, \ldots,$
   $[beg\_fr,\ end\_trans,\ no\_data,\ fail]_{n-1}\}$.
2. Set of output signals:
   $Y = \{[fr\_end(m),\ bus\_is\_oc,\ no\_data(m),\ fail(m)]_0,$
   $[fr\_end(m),\ bus\_is\_oc, st\_on,\ no\_data(m)]_1, \ldots,$
   $[fr\_end(m),\ bus\_is\_oc, st\_on,\ no\_data(m)]_{n-1}\}$.
3. Set of internal events $E'' = \varnothing$.
4. State $\upsilon(t_m) = \{q_i(t_m),\ i = \overline{1,N},\ kan(t_m)\}$;
   where: $q_i(t_m) \in \{1, 2, \ldots, N\}$; $q_i(t)$ – the number of successor for the ith station;
   $kan(t_m) = \begin{cases} 0, \text{ channel is idle;} \\ 1, \text{ channel is occupied.} \end{cases}$
5. Initial state:
   $kan(t_0) := 0$; $i := 1$;
   while $i < n$ do begin $q_i(t_0) := i+1$; $i := i+1$; end.
6. Transfer operators:
   $H\left[e'_{1k}\left(new\_st(p)\right)\right] : k = \overline{2,n}$; (New station)
   $i := p$;
   if $i = n$ then $i := 0$;
   while $q_i(t_m) = 0$ do begin $i := i+1$;
   if $i = n$
   then $i := 0$; end;
   $j := 1$;
   while $q_j(t_m) \neq i+1$ do $j := j+1$;
   $q_j(t_{m+1}) := p$; $q_p(t_{m+1}) := i+1$;
   $y_p := st\_on$.

   $H\left[e'_{2k}(beg\_fr)\right] : k = \overline{1,n}$; (The start of packet sending)
   $kan(t_{m+1}) := 1$;
   for $i := 1$ to n do
   if $i \neq k$ and $q_i(t_m) > 0$ then
   $y_i := bus\_is\_oc$.

   $H\left[e'_{3k}(end\_trans)\right] : k = \overline{1,n}$ (The end of packet transmission)
   $kan(t_{m+1}) := 0$;
   for $i := 1$ to n do
   if $i \neq k$ and $q_i(t_m) > 0$ then
   $y_i := fr\_end\left[q_k(t_m)\right]$.

   $H\left[e'_{4k}(no\_data)\right] : k = \overline{1,n}$ (There are no data for sending)
   for $i := 1$ to n do
   if $i \neq k$ and $q_i(t_m) > 0$ then
   $y_i := no\_data\left[q_k(t_m)\right]$.

   $H\left[e'_{5k}(fail)\right] : k = \overline{2,n}$ (The station is switching of)
   $y_1 := fail[k]$;
   $i := 1$
   while $q_j(t_m) \neq k$ do $i := i+1$;
   $q_i(t_{m+1}) := q_k(t_m)$;
   $q_k(t_{m+1}) := 0$.

5.1. Results of validation and simulation

The correctness of the created specification was investigated by means of protocol analysis system PRANAS-2. This system permitted one to investigate general protocol properties such as: completeness; deadlock freeness; boundedness; cyclic behavior; termination.

Table 1
Example of validation

```
{32}      L:  2 3 1 0
          MO:  1 0 1 1 0 1 3
            Tim_T Arr_fr Taim_DI
          M[1]:  2 0 0 0 1 1
            Swit_of Arr_fr
          M[2]:  3 0 0 0 1 1
            Swit_of Arr_fr

  ↓       Taim_DI in MO
{58}      L:  2 3 1 1
          MO:  1 0 0 1 1 1 3
            Tim_T Arr_fr Trans_fr
          M[1]:  2 0 0 0 1 1
            Swit_of Arr_fr
          M[2]:  3 0 0 0 1 1
            Swit_of Arr_fr

  ↓       Trans_Fr in MO
{104}     L:  2 3 1 0
          MO:  1 1 0 1 0 0 3
            Tim_T Arr_fr Taim_D
          M[1]:  2 0 0 1 1 1
            Swit_of Arr_fr Taim_DI
          M[2]:  3 0 0 0 1 1
            Swit_of Arr_fr

  ↓       Taim_DI in M1
{79}      L:  2 3 1 1
          MO:  1 0 0 1 0 0 3
            Tim_T Arr_fr Taim_D
          M[1]:  2 0 1 0 1 1
            Swit_of Arr_fr Trans_fr
          M[2]:  3 0 0 0 1 1
            Swit_of Arr_fr

  ↓       Trans_fr in M1
{150}     L:  2 3 1 0
          MO:  1 0 0 1 0 0 3
            Tim_T Arr_fr Taim_D
          M[1]:  2 1 0 0 1 0
            Swit_of Arr_fr Taim_D
          M[2]:  3 0 0 1 1 1
            Swit_of Arr_fr Taim_DI

  ↓       Taim_DI in M1
{92}      L:  2 3 1 1
          MO:  1 0 0 1 0 0 3
            Tim_T Arr_fr
          M[1]:  2 0 0 0 1 0
            Swit_of Arr_fr
          M[2]:  3 0 1 0 1 1
            Swit_of Arr_fr Trans_fr
```

In Table 1, some validation results are represented. The numbers included in brackets $\{\ldots\}$ refer to the number of the state. Numbers written after `L`, `MO` and `M[i]` have

the following meanings of discrete and continuous coordinates of state:

```
L:    q1; q2; q3; kan;
MO:   nr; act_D; act_DI; act_T;
      act_trans_fr; st; n_act;
M[i],i=1,2:  nr; act_D;
      act_trans_fr; act_DI;sw;st.
```

5.2. Simulation results

Simulation results are represented in Table 2. The parameters of the model are the following: `Taim_Frame` – duration of frames; `Taim_Head` – duratin of the head of frames; `Taim_D` – duration of the timer *D*; `Taim_DI` – duration of the timer *DI*; `Taim_T` – duration of timer *T*; `V` – velocity of the channel; `n` – number of stations; `Arr_Frame` – parameter of a puasonian input stream; `T_swit_on` and `T_swit_off` – intensity of operations `swit_on` and `swit_off`, which have exponential distributions.
Characteristics of the model: `T_Wait` – the mean value of transmitting a frame including the waiting time; `L_Wait` – mean value of the waiting time; `K_Useful` – coefficient utilization of a channel; `K_Full` – coefficient of full utilization of a channel.

Table 2
Simulation results

Taim_Frame = 800 bit, Taim_Head = 160 bit, Tau_Data = 4 bit, Taim_D = 0.0000025 s, Taim_DI = 0.0000012 s, Taim_T = 100 s, T_swit_on = T_swit_of = 0.				
1. V = 10000000 bit/s, Arr_Frame = 0.001 s				
n	*T_Wait*	*L_Wait*	*R_Useful*	*K_Full*
2	0.00011	0.00001	0.1418	0.8323
4	0.00013	0.00003	0.2806	0.8639
6	0.00016	0.00006	0.4118	0.8939
8	0.00019	0.00010	0.5297	0.9207
10	0.00025	0.00016	0.6304	0.9437
2. V = 50000000 bit/s, Arr_Frame = 0.001 s				
n	*T_Wait*	*L_Wait*	*R_Useful*	*K_Full*
2	0.00002	0.00000	0.0307	0.4639
4	0.00002	0.00001	0.0611	0.4831
6	0.00003	0.00003	0.0917	0.5025
8	0.00003	0.00001	0.1219	0.5217
10	0.00003	0.00001	0.1529	0.5413
3. V = 50000000 bit/s, Arr_Frame = 0.000135 s				
n	*T_Wait*	*L_Wait*	*R_Useful*	*K_Full*
2	0.00002	0.00000	0.0212	0.5921
4	0.00003	0.00001	0.3848	0.6882
6	0.00004	0.00002	0.5399	0.7864
8	0.00006	0.00004	0.6513	0.8569
10	0.00009	0.00007	0.7160	0.8979

6. Conclusions

The presented method of formal specification permits on the base of single specification to carry out validation general and individual properties and simulation. It permits to investigate the analysed system more thoroughly.

References

[1] G. I. Holzmann, "The model checker SPIN", *IEEE Trans. Softw. Eng.*, vol. 23, no. 5, pp. 279–295, 1997.

[2] B. P. Zeigler, *Theory of Modelling and Simulation*. New York: Academic Press, 2000.

[3] H. Pranevicius, "Aggregate approach for specification, validation, simulation and implementation of computer network protocols", in *LNCS*, Berlin: Springer-Verlag, 1991, vol. 502, pp. 433–477.

[4] H. Pranevicius, V. Pilkauskas, and A. Chmieliauskas, "Aggregate approach for specification and analysis of computer network protocols", Technologija, Kaunas University of Technology, 1994.

[5] H. Pranevicius, "Formal specification and analysis of distributed systems", in Lecturer Notes "Applications of AI to Production Engineering", Technologija, Kaunas, 1997, pp. 269–322.

[6] H. Pranevicius, "Formal specification and analysis of distributed systems", *J. Intell. Manuf.*, no. 9, pp. 559–569, 1998.

Henrikas Pranevicius is a Professor of the Kaunas University of Technology and the Head of Business Informatics Department. He is habilituated doctor of Technical Sciences at Ryga Electronic and Computer Technics Institute sinice 1984 and doctor of science from Kaunas Politechnical Institute since 1970. Area of his research activity is: formal specification, validation and simulation of distributed systems including telecommunication and logistic systems. The theoretical background of investigation is piece-linear aggregate formalism, which permits to use the single formal specification for models development both for performance and behaviour analysis.
e-mail: hepran@if.ktu.lt
Kaunas University of Technology
Studentu st 50
LT-51368 Kaunas, Lithuania

Paper

Adaptive procedure for automatic modulation recognition

Ferdinand Liedtke

Abstract—An adaptive procedure for automatic modulation recognition is described. With it the automatic modulation classification and recognition of radio communication signals with a priori unknown parameters is possible effectively. The results of modulation recognition are important in the context of radio monitoring or electronic support measurements. The special features of the procedure are the possibility to adapt it dynamically to nearly all modulation types, and the capability to recognize continuous phase modulation (CPM) signals like Gaussian minimum-shift keying (GMSK) too. A time synchronization to the symbol rate is not necessary.

Keywords— modulation recognition, modulation classification, signal analysis, radio monitoring.

1. Introduction

The results of modulation recognition are important in the context of radio monitoring or electronic support measurements. Originally the research was conducted for short wave or high frequency (HF) communication signals in the radio frequency (RF) range between 1 and 30 MHz. In this range there exist many different signal modulation types or wave forms, and often the systems or signal characteristics are not standardized. Furthermore, modern HF radio systems are able to change their parameters and their modulation types continuously, dependent on the quality of the transmission channel. The developed modulation recognition procedure can be used not only for the analysis of HF signals, but also for radio communication signals from higher RF ranges, e.g., VHF or UHF. In those ranges lots of military and civil mobile radios are running. Though many of those systems are standardized with known signal characteristics, it will be of interest to detect the activity of specific systems and their currently used modulation types. The additional challenge in the higher frequency ranges are the often used CPM waveforms, e.g., GMSK with varying B*T values. For those soft-keyed wave forms it is more difficult to find relevant signal values suited for modulation recognition than for wave forms with hard keying. A modulation recogniser developed for digitally modulated signals with hard keying is described in [1]. In this context it is important to remember that in general the symbol rate or the time points for symbol synchronization are not known for the non-authorized receiver. In many papers about modulation recognition the a priori knowledge of the exact values for centre frequency, bandwidth, and symbol rate of every interesting signal is assumed, e.g., [2, 3]. This will be appropriate for an application like the adjustment of a software radio which expects several well defined signal wave forms. In contrary to that, this presupposition cannot be maintained in general for applications like radio monitoring or electronic support measurements. For these applications a special robustness against parameter inaccuracies is necessary. A further challenge for the modulation recognition described here is the quantity of different wave forms, several of which are often totally unknown from the beginning. Therefore it is desirable to have a recognition procedure which adapts easily to the various wave forms.

For the future, some further development of the recognition procedure is planned.

2. Principle of the procedure

The necessary pre-processing of the received signal is represented in Fig. 1. From the down converted and digitised signal a spectrum is computed with the aim to accomplish a *spectral segmentation*, i.e., to estimate the centre frequency f_c and the bandwidth B of a significant spectrum part.

With the determined parameters the signal is appropriately shifted in frequency and filtered to get the complex base band signal z. On this occasion the segmentation procedure is not discussed in detail. The signal z is fed into the *modulation recognition* module. The essential parts of this module are represented in Fig. 2. The figure parts marked in grey designate the differences to the formerly used modulation recogniser [1]. The further processing is performed in three branches: The upper two branches with the *squaring of absolute values*, the *squaring of complex values*, and the following digital Fourier transforms (DFTs) are used for exploitation of the symbol rate information, which perhaps may be included in the signal, by means of the *spectral line detection*. The *squaring of absolute values* could reveal the appropriate spectral line, indicating the symbol rate for digitally modulated signals with hard keying while the *squaring of complex values* is provided for CPM. Within the described modulation recogniser the symbol rate estimation is not necessary for obtaining a time synchronization, the sole aim is the estimation of the approximate symbol dwell time T. Time T is needed for the calculation of the appropriate values for the difference phase $D\varphi$ within the *feature extraction* module in the third branch. In future, the spectral symbol rate information, which differs in dependence on the modulation type, could perhaps be utilized for the classification process too, but in the currently used recogniser it is not yet used for this purpose. Before the *feature extraction* is carried out, a *coordinate resolving* module provides the signal in polar coordinates a and φ.

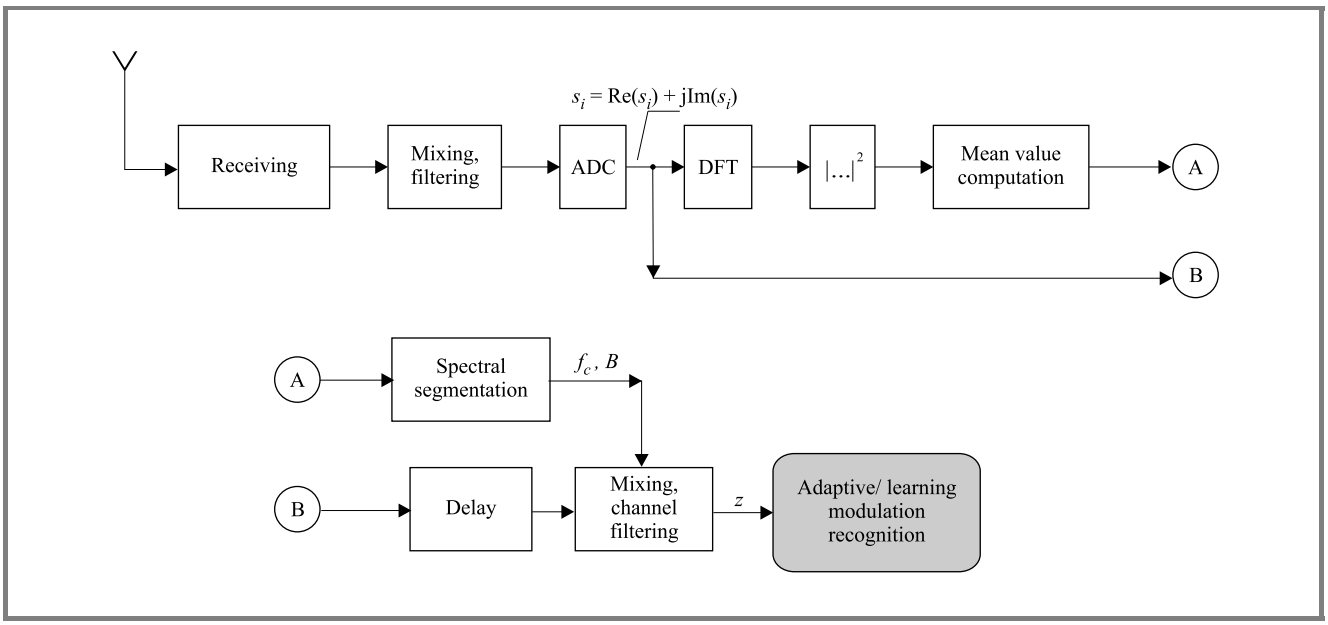


Fig. 1. Signal analysis with modulation recognition.

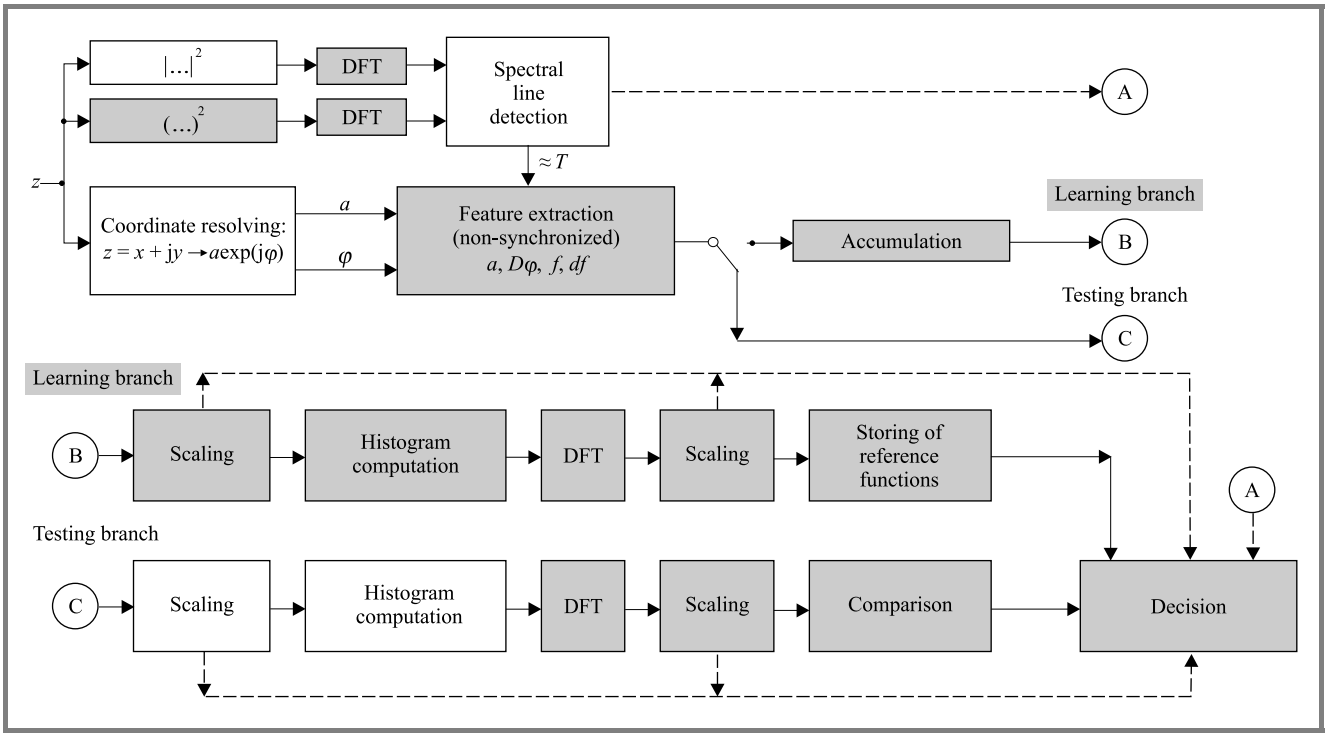


Fig. 2. Adaptive modulation recognition, non-synchronized.

The *feature extraction* module provides relevant values of the following signal parameters: amplitude a, difference phase $D\varphi$, instantaneous frequency f, and derivation of the instantaneous frequency df. Concerning the parameter φ, difference values and not the phase values themselves are used for the purpose to be resistant to a non-exact frequency tuning.

The time difference for extracting every pair of phases for calculating $D\varphi$ is T. The parameter f is essential for recognizing simple frequency modulated signals like frequency-shift keying and the derivative of f is provided for chirp signals. The procedure for extracting the relevant parameter values is discussed in the last paragraph of this section.

After *feature extraction* there is a splitting into two branches, a *learning branch* and a *testing branch*. By means of the *learning branch*, it is possible to adapt to the various modulation types or wave forms. For this aim the relevant parameter values are accumulated in a learning phase

and further processed as follows: the parameter values are appropriately scaled and written into *histograms*, one histogram for each of the four parameters. The histograms are then transformed into a picture domain by means of the DFT. The transformed functions or picture functions are a kind of characteristic functions with the property that the interesting information is concentrated in the first part of each function. It was found out that the use of the first quarter of the respective picture function values is sufficient. In the following text this part of a picture function is still called "picture function". The picture functions are appropriately *scaled* and *stored as reference functions*, for every modulation type a set of four functions. After having finished the learning of all interesting wave forms, the change-over switch following the *feature extraction* module (see Fig. 2) is switched into the lower position and the testing or working phase begins. The signal processing in the testing branch is the same as that in the learning branch up to the *scaling* of the picture functions. In the following *comparison* the actual set of picture functions are compared to all stored function sets and the set with the least deviation determines the modulation type. The deviation measure used here is the least mean square (LMS) and the fusion of the results obtained from the processing of the individual parameters is done by simply adding the LMS results. The dashed lines in Fig. 2 refer to the intention to search (in future) for additional features, obtained from spectral or scaling parameters.

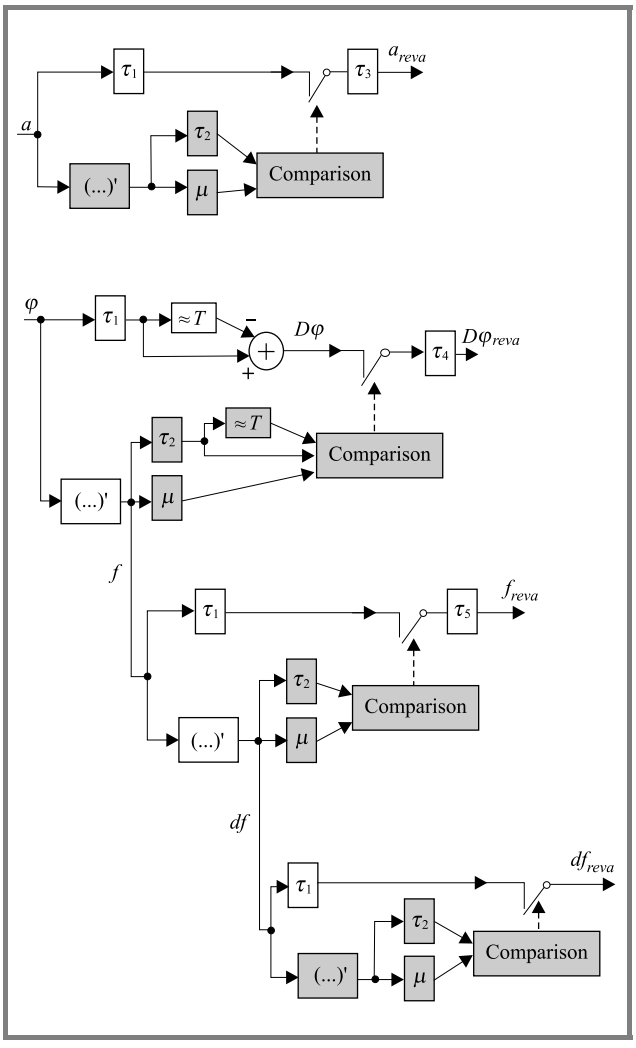


Fig. 3. Feature extraction, non-synchronized.

The extraction of the relevant parameter values for a, $D\varphi$, f, and df is represented in Fig. 3. Because the extraction was not intended to need the synchronization to the symbol rate, another method had to be worked out. The **re**levant **va**lues are those where the eyes of their so-called eye patterns are wide open, i.e., the parameters have reached their steady states. The way to find these values is to continuously compute the *time derivative* of each parameter (in Fig. 3 depicted with $(\ldots)'$), to compute the *sliding mean* μ of these derived values, and to continuously perform a *comparison* of μ with the corresponding original parameter value. It was found out that the relevant values are those which are near to the mean of the derived values. This is a similar procedure as the search for extreme values of a function by means of its derivative. Here, the *comparison* is not carried out with the derivative value zero, but with the mean of the derived values. The intention is to compensate parameter trends, which normally are still existent. Concerning the phase it has to be remembered that always difference phase values have to be considered.

3. Experimental results

The whole procedure including signal generation, anti-aliasing filtering, noise addition, reception filtering, and modulation recognition was developed and tested on a PC with the software tools of MATLAB. Successful tests were also carried out with real signals obtained from the mobile radio systems GSM, TETRA, and TETRAPOL.

A class space of 15 modulation types was defined, most of them digitally modulated wave forms. The frequently used modulation types amplitude-shift keying with two states (ASK2), frequency-shift keying (FSK2), minimum-shift keying (MSK) including different GMSK forms, and phase-shift keying (PSK) with different numbers and positions of relevant phase states are considered among others. The class space was supplemented with a white Gaussian noise class (WGN) and a reject class (REJ). All classes are represented in Table 1. The listed signal to noise ratios (SNRs) are explained in the next paragraph. The arrangement with altogether 17 classes was chosen to challenge the recognition procedure. In realistic scenarios the class number could be reduced in many cases, which alleviates the classification task.

To learn the class specific reference functions 16 learning experiments per class were carried out, whereby for every experiment a signal segment with a length corresponding to 256 symbols was used. The number of individual signal samples per symbol dwell time T was chosen to be 8. In the next step the appropriate decision levels were learned

Table 1
Used classes/wave forms

No.	Name of class/wave form	Comments	SNR_{start} [dB]
0	REJ	Reject class	—
1	WGN	White Gaussian noise	—
2	ASK2	Amplitude-shift keying, 2 states	15
3	FSK2	Frequency-shift keying, 2 states, $\beta = 0.85$	12
4	MSK	Minimum-shift keying	12
5	GMSK05	Gaussian MSK, B*T = 0.5	12
6	GMSK03	Gaussian MSK, B*T = 0.3	12
7	PSK2	Phase-shift keying, 2 states	9
8	$\pi/2$ DPSK2	Differential PSK2, keying with $+/-\pi/2$	9
9	PSK4	Phase-shift keying, 4 states	14
10	$\pi/4$ DPSK4	Differential PSK4, keying with $+/-\pi/4$	14
11	OPSK4	Offset PSK4	14
12	PSK8	Phase-shift keying, 8 states	20
13	ASK2/PSK8	Hybrid digital modulation type, 2 amplitude states for each of the 8 phase states	25
14	CLOVER	ASK2/PSK8 with Chebyshev pulse weighting	25
15	LICHIRP	Linear chirp	20
16	CHIRPKEY	Chirp keying with linear up and down chirp	20

with 16 additional experiments. The used SNRs are related to those values with which an authorized receiver could reach a symbol error probability of 10^{-4} with non-coherent or differential demodulation, whereby a coding gain was not considered. These SNRs differ in dependence on the modulation type and they are called SNR_{start}. The use of different SNR_{start} values for the different modulation classes was realized because the authorized receiver expects different SNRs for different modulation types too, compare Table 1. For example the SNR_{start} value for the PSK4 types is 14 dB and that for the PSK2 types is 9 dB. A quarter of the experiments was carried out with SNR_{start}, the next quarter was performed with $SNR_{start} - 2$ dB, the third quarter has $SNR_{start} - 4$ dB and the last quarter was performed with $SNR_{start} - 6$ dB, i.e., the mean SNR was $SNR_{start} - 3$ dB. In that way it was possible to take into account different SNRs with comparatively low experiment numbers. The testing or working phase was carried out with 64 experiments per class, each experiment with a signal length corresponding to 256 symbols too. The principle of choosing the SNRs was the same as that used in the learning phase.

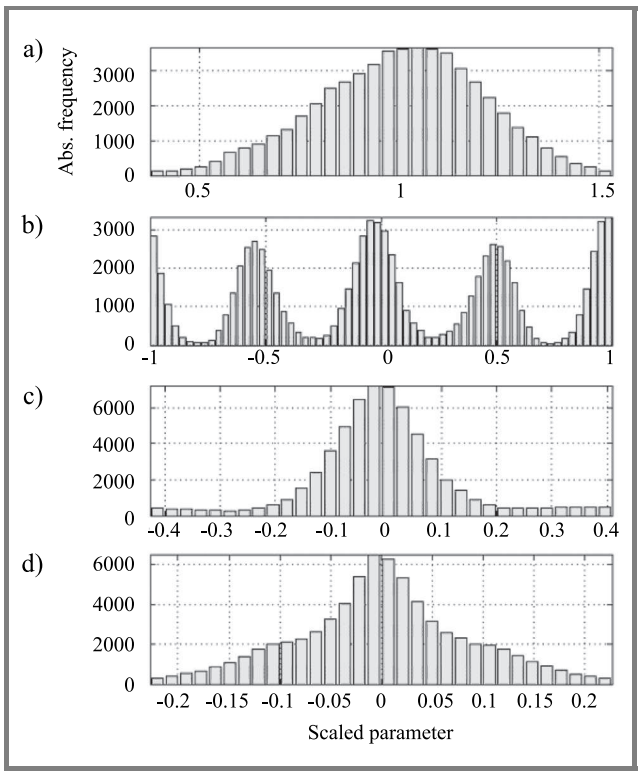


Fig. 4. Parameter histograms for PSK4: (a) amplitude a; (b) difference phase $D\varphi$; (c) instantaneous frequence f; (d) differential instantaneous frequency df.

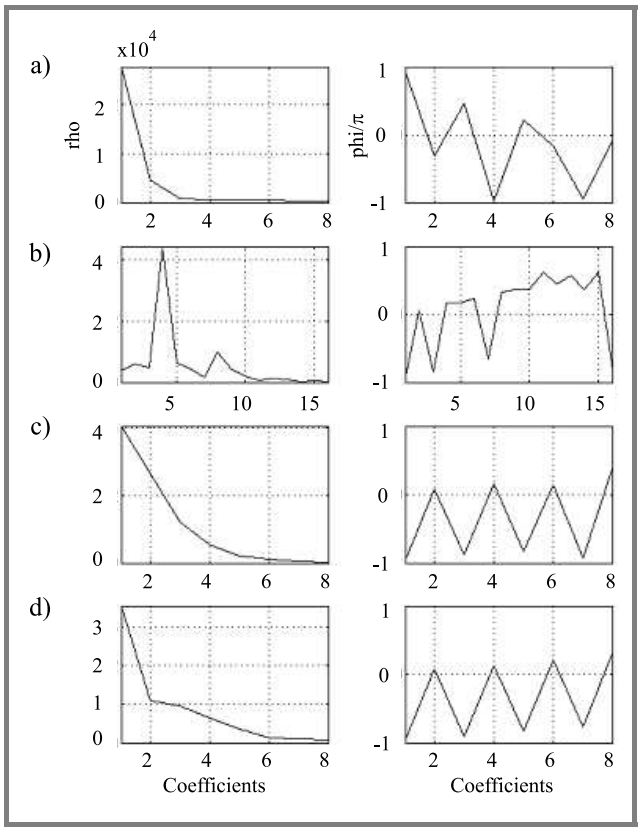


Fig. 5. Picture functions for PSK4: (a) amplitude a; (b) difference phase $D\varphi$; (c) instantaneous frequence f; (d) differential instantaneous frequency df.

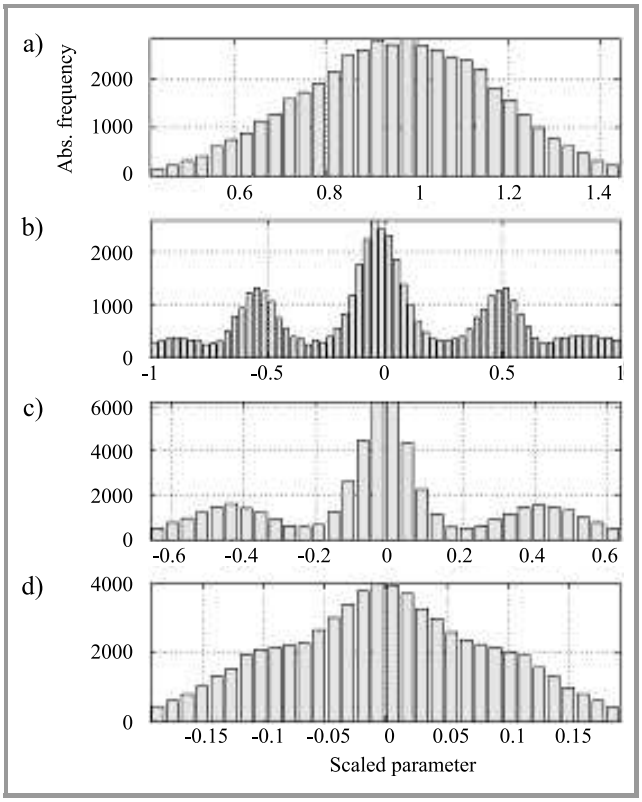


Fig. 6. Parameter histograms for OPSK4: (a) amplitude a; (b) difference phase $D\varphi$; (c) instantaneous frequence f; (d) differential instantaneous frequency df.

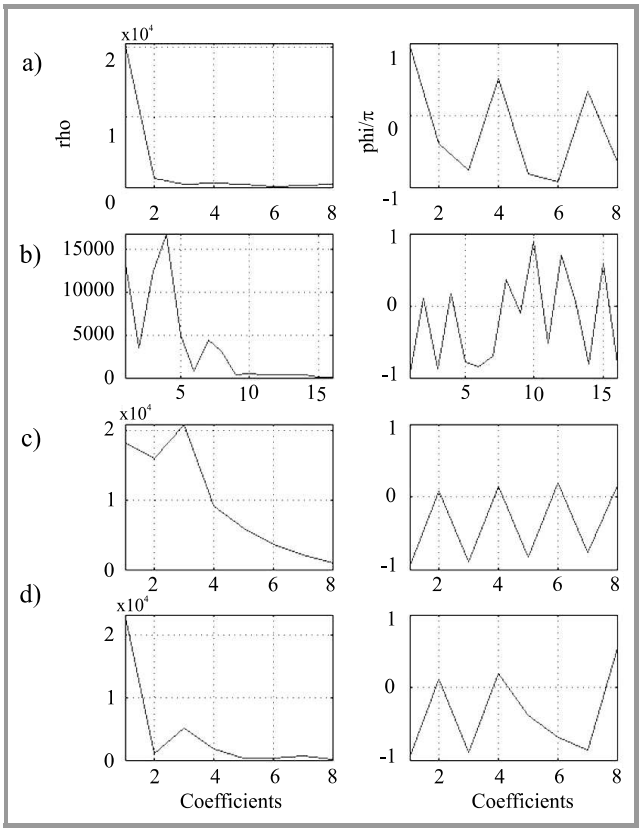


Fig. 7. Picture functions for OPSK4: (a) amplitude a; (b) difference phase $D\varphi$; (c) instantaneous frequence f; (d) differential instantaneous frequency df.

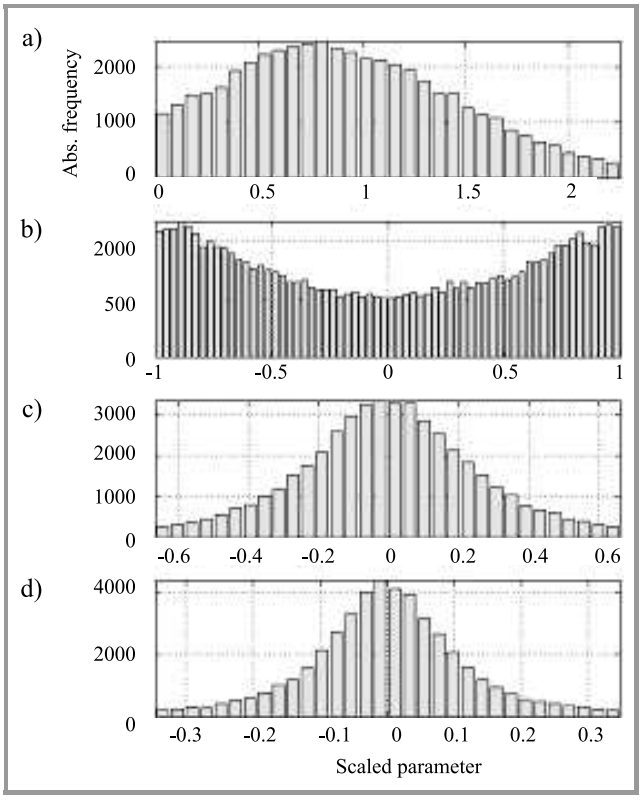


Fig. 8. Parameter histograms for WGN: (a) amplitude a; (b) difference phase $D\varphi$; (c) instantaneous frequence f; (d) differential instantaneous frequency df.

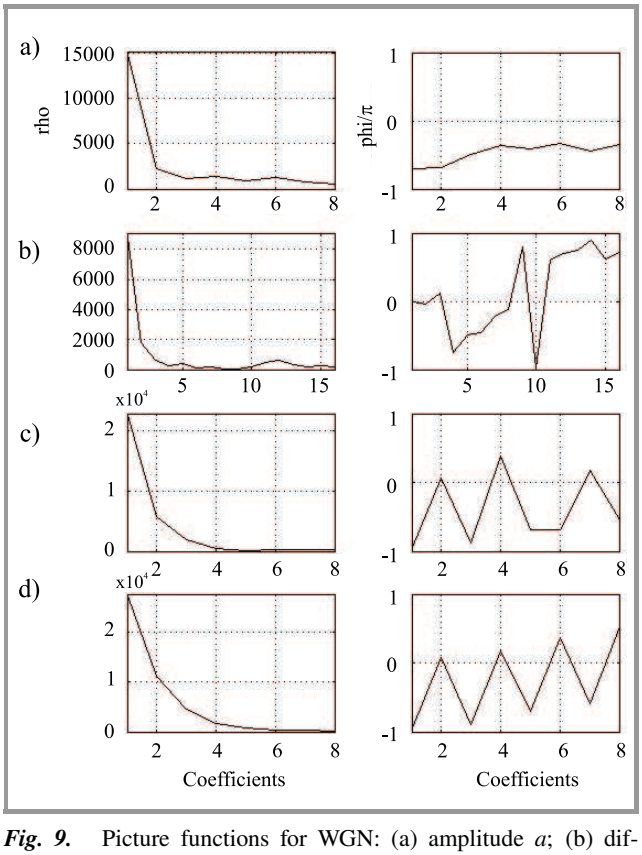


Fig. 9. Picture functions for WGN: (a) amplitude a; (b) difference phase $D\varphi$; (c) instantaneous frequence f; (d) differential instantaneous frequency df.

Table 2
Confusion matrix, $\mathrm{SNR_{start}} - 6\ \mathrm{dB} \leq \mathrm{SNR} \leq \mathrm{SNR_{start}}$

Input	Output class																
	ASK2	FSK2	MSK	GMSK05	GMSK03	PSK2	$\pi/2$ DPSK2	PSK4	$\pi/4$ DPSK4	OPSK4	PSK8	ASK2/PSK8	CLOVER	LICHIRP	CHIRKEY	WGN	REJ
ASK2	98.4																1.6
FSK2		100															
MSK			**70.3**	**23.4**	**4.7**												1.6
GMSK05			**32.8**	**54.7**	**12.5**												
GMSK03				**21.9**	**75**												3.1
PSK2						98.4											1.6
$\pi/2$ DPSK2							100										
PSK4								89.1									10.9
$\pi/4$ DPSK4									98.4								1.6
OPSK4							3.1			87.5							9.4
PSK8											96.9						3.1
ASK2/PSK8												96.9					3.1
CLOVER													95.3				4.7
LICHIRP														95.3			4.7
CHIRKEY															100		
WGN																85.9	14.1

Table 3
Confusion matrix, $\mathrm{SNR_{start}} - 9\ \mathrm{dB} \leq \mathrm{SNR} \leq \mathrm{SNR_{start}} - 3\ \mathrm{dB}$

Input	Output class																
	ASK2	FSK2	MSK	GMSK05	GMSK03	PSK2	$\pi/2$ DPSK2	PSK4	$\pi/4$ DPSK4	OPSK4	PSK8	ASK2/PSK8	CLOVER	LICHIRP	CHIRKEY	WGN	REJ
ASK2	67.2																32.8
FSK2		84.4														14.1	1.6
MSK			**39.1**	**26.6**	**32.8**												1.6
GMSK05			**6.3**	**40.6**	**51.6**												1.6
GMSK03				**17.9**	**82.8**												
PSK2						68.8										4.7	26.6
$\pi/2$ DPSK2							78.1									6.3	15.6
PSK4						1.6	1.6	64.1								1.6	31.3
$\pi/4$ DPSK4									78.1								21.9
OPSK4						9.4			3.1	60.9							26.6
PSK8										4.7	60.9						34.4
ASK2/PSK8												71.9					28.1
CLOVER													64.1				35.9
LICHIRP				4.7	3.1				1.6					78.1			12.5
CHIRKEY					34.4										64.1		1.6
WGN																85.9	14.1

As a first demonstration example, histograms of the four examined parameters a, $D\varphi$, f, and df for a PSK4 signal are depicted in Fig. 4. The results are obtained from 16 experiments with 256 symbols per experiment, i.e., from 4096 symbols altogether. This comparatively high symbol number was chosen to clearly depict the typical characteristics. For a normal recognition process the exploitation of 256 symbols is entirely sufficient. It can be seen from Fig. 4 that the $D\varphi$ histogram has 4 peaks. This is the main characteristic indicating the PSK4 wave form.

The picture functions are depicted in Fig. 5. In the left column the results for the absolute values rho and in the right column the phase results phi$/\pi$ of the histogram transforms are depicted. The important result is the peak position at the abscissa value 4 in Fig. 5 on the left hand side, second row. The corresponding results of an OPSK4 signal are depicted in Figs. 6 and 7. The peaks in the $D\varphi$ histogram (see Fig. 6) are less evident than those of the corresponding PSK4 histogram. This becomes noticeable in Fig. 7, left hand side, second row, too. However, for sufficiently good SNRs, OPSK4 can be classified correctly. For lower SNRs some classifications as PSK2 or $\pi/2$ DPSK2 are possible. For comparison with the WGN situation the corresponding results for WGN are depicted in Figs. 8 and 9. It can be observed that neither the histograms nor the picture functions give any indication of a digitally modulated signal (with two or more characteristic parameter positions). Therefore it could be understood that the WGN class can be easily

discriminated from the other classes. The principal considerations concerning the other classes, not yet discussed, are similar.

The final classification results are represented in form of so-called confusion matrices (see Tables 2 and 3). The names of all treated classes were entered into the left column and into the uppermost row. The left column depicts the classes fed into the modulation recogniser and the classification results were entered into the columns with the corresponding output class names indicated in the uppermost row. For an ideal recogniser all results, each 100%, are contained in the diagonal matrix elements. These elements are indicative of correct classifications.

The learning of the reference picture functions was performed with 16 experiments per class. For every experiment a signal segment length corresponding to 256 symbols was used. Another 16 experiments per class were carried out to learn the necessary variances of the deviations of the test picture functions from the stored reference functions. These values were needed for arranging the appropriate decision levels. The SNRs were chosen in the range $\mathrm{SNR}_{\mathrm{start}} - 6\ \mathrm{dB} \leq \mathrm{SNR} \leq \mathrm{SNR}_{\mathrm{start}}$ as discussed above. The testing phase consisted of 64 experiments for each class. The results depicted in the confusion matrices are indicated in percent related to 64. The difference between Tables 2 and 3 is the different SNR selection in the test experiments. While the tests for Table 2 were performed with $\mathrm{SNR}_{\mathrm{start}} - 6\ \mathrm{dB} \leq \mathrm{SNR} \leq \mathrm{SNR}_{\mathrm{start}}$ the tests for Table 3 were carried out with SNRs 3 dB worse. The results in Table 2 show that most of the classification results are contained in the diagonal elements or in their neighbourhood. The results for the CPM signals MSK, GMSK05, and GMSK03 are marked with bold numbers to indicate the close relationship between the three modulation classes. In other words, the spread of classification results is caused by the similarity of the wave forms and could not really assessed as false classifications. Most of the non-successfully classified signals were assigned to the reject class REJ. This is a desired result because a rejection or a classification as WGN is preferred to a false classification. The classification of two OPSK4 signals (3.1%) as $\pi/2$ DPSK2 is caused by experiments with comparatively low SNR. Both wave forms are similar in certain aspects. The classification results presented in Table 3 are worse because the SNRs were 3 dB lower. The false classification of 34.4% of the CHIRKEY signals as GMSK03 is not really critical because CHIRKEY is a here synthesized artificial signal wave form, which was chosen very similar to GMSK03. The false classification of 9.4% of the OPSK4 signals as PSK2 is not surprising because for the smaller SNRs these wave forms are also similar. The numbers of the other false classifications are still comparatively moderate.

For assessing the discussed experimental results one has to realize that for real scenarios the class number could be reduced in many cases. This alleviates the classification task. For the future, further work is planned on the following items: Test and integration of additional classification features, appropriate evaluation and fusion of the feature parameter values, adaptation of the classification procedure, and tests.

4. Conclusions

An automatic modulation recognition procedure is described which is suited for many modern signal wave forms including CPM. The automatic modulation recognition is interesting in the context of radio monitoring or electronic support measurements. For this application the signal parameters are often not known a priori, e.g., the centre frequency, the bandwidth, and the symbol rate. As a typical pattern recognition procedure the modulation recognition has a learning and a testing phase. During the learning phase an adaptation to the new wave forms is easily possible. The procedure was tested with 16 wave forms with different SNRs. The results are comparatively good. For the future some further improvements are planned.

References

[1] F. Liedtke, "Computer simulation of an automatic classification procedure for digitally modulated signals with unknown parameters", *Sig. Proc.*, vol. 6, no. 4, pp. 311–323, 1984.

[2] J. Palicot and C. Roland, "A new concept for wireless reconfigurable receivers", *IEEE COM Mag.*, pp. 124–132, July 2003.

[3] M. L. D. Wong and A. K. Nandi, "Automatic digital modulation recognition using artificial neural network and genetic algorithm", *Sig. Proc.*, vol. 84, pp. 351–365, 2004.

Ferdinand Liedtke is working in the department for Radio Communications and Electronics (FE) of the Research Institute for Communications, Information Processing, and Ergonomics (FKIE) of the FGAN in Germany. He is working on the topics detection, segmentation, classification, and automatic modulation recognition of signals with a priori unknown parameters for many years. He has written papers and got patents. Several of the modulation recognisers offered by national and international companies are based on the concepts worked out in the FGAN-FKIE-FE. Dr.-Ing. Liedtke is further interested in modern radio systems and their susceptibilities.

e-mail: liedtke@fgan.de
FGAN-FKIE-FE
Neuenahrer Str. 20
D-53343 Wachtberg-Werthhoven, Germany

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. Orłowska, "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 proof-read 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.

Introduction of the Network Centric Warfare concept to Czech Armed Forces

P. Eichelmann and L. Lukáš — Paper — 53

Cryptology Laboratory - its quality system and technical competence according to the ISO/IEC 17025 standard

R. Wicik — Paper — 58

IP-KRYPTO cipher machine for military use

M. Borowski and G. Łabuzek — Paper — 64

Primality proving with Gauss and Jacobi sums

A. Chmielowiec — Paper — 69

Packet switch architecture with multiple output queueing

G. Danilewicz et al. — Paper — 76

Integrated analysis of communication protocols by means of PLA formalism

H. Pranevicius — Paper — 84

Adaptive procedure for automatic modulation recognition

F. Liedtke — Paper — 91

National Institute
of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

Editorial Office

tel. +48(22) 512 81 83
tel./fax: +48(22) 512 84 00
e-mail: redakcja@itl.waw.pl
http://www.itl.waw.pl/jtit