

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

2010
Special edition

ITU Plenipotentiary Conference

Guadalajara, Mexico, 4-22 of October, 2010

Broadband Challenges

H. Touré, ITU Secretary General

3

International Co-operation Needed to Face Information Age Challenges

M. Gaj, UnderSecretary of State, Ministry of Infrastructure, Poland

8

National Institute of Telecommunications – International Partnership: Yesterday, Today and Tomorrow

W. Halka, General Director, National Institute of Telecommunications, Poland

10

Rural Telecommunications Infrastructure Selection Using the Analytic Network Process

Y. Gasiea, M. Emsley and L. Mikhailov

15

Telecommunications, Universal Service and Poverty in Mexico: a Public Policy Assessment (1990-2008)

C. Casanueva-Reguart and A. Pita S.

30

Performance Analysis of a Bi-Objective Model for Routing and Wavelength Assignment in WDM Networks

C. Simões et al.

43

Computational Methods for Two-Level 0-1 Programming Problems through Distributed Genetic Algorithms

K. Niwa, T. Hayashida and M. Sakawa

56

(Contents Continued on Back Cover)

Editorial Board

Editor-in Chief: ***Paweł Szczepański***

Associate Editors: ***Krzysztof Borzycki***
Marek Jaworski

Managing Editor: ***Maria Łopuszniak***

Technical Editor: ***Ewa Kapuściarek***

Editorial Advisory Board

Chairman: ***Andrzej Jajszczyk***
Marek Amanowicz
Daniel Bem
Wojciech Burakowski
Andrzej Dąbrowski
Andrzej Hildebrandt
Witold Hołubowicz
Andrzej Jakubowski
Alina Karwowska-Lamparska
Marian Kowalewski
Andrzej Kowalski
Józef Lubacz
Tadeusz Łuba
Krzysztof Malinowski
Marian Marciniak
Józef Modelski
Ewa Orłowska
Andrzej Pach
Zdzisław Papir
Michał Pióro
Janusz Stokłosa
Wiesław Traczyk
Andrzej P. Wierzbicki
Tadeusz Więckowski
Józef Woźniak
Tadeusz A. Wysocki
Jan Zabrodzki
Andrzej Zieliński

ISSN 1509-4553 on-line: ISSN 1899-8852

© Copyright by National Institute of Telecommunications
Warsaw 2010

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 of the *Journal of Telecommunications and Information Technology* provides selected articles that have been published in this quarterly in recent years, with three exceptions. The first one is the acceptance speech by Hamadoun Touré, ITU Secretary General, at the ceremony of awarding him the title of Doctor Honoris Causa of the Wrocław University of Technology (Poland), on 14 September 2010 (reproduced with the kind consent of the author). The second exception is the foreword by Magdalena Gaj, Deputy Minister, Undersecretary of State in the Ministry of Infrastructure, Poland. The third one is the text by Wojciech Hałka, Director General, National Institute of Telecommunications, Poland, focused on international co-operation of the Institute, especially with the ITU.

This special issue is published with the delegates to the 19th ITU Plenipotentiary Conference in mind. Like the previous such conferences, this one is the key event for the development of information and communication infrastructure and services, worldwide. The ITU Member States gathered at Guadalajara Mexico, from 4 to 22 October 2010, decide together upon the current problems and future role of the Union. At this occasion the delegates discuss informally also other issues related to the conference agenda but not listed explicitly in it. Bilateral collaboration is one of such issues that contributes significantly to the success of the Union.

Dr. Touré in his acceptance speech focuses on broadband services and universal access to them as a key issue. Minister Gaj underlines that Poland is open to collaboration. As an old ITU Member country, Poland has been well known to all delegates, but it is now undergoing fast changes. Minister Gaj and Director Hałka offer brief information on the status and progress made recently. The remaining texts cover various topics within the scope of the Journal. Their selection offered here may help in giving the reader an idea on what topics deserve such collaboration and with whom.

The *Journal of Telecommunications and Information Technology* is published in English by the National Institute for Telecommunications, Poland. It publishes original contributions related to telecommunication and information issues produced by its staff and/or submitted from around the world. All papers are subject to peer review. The electronic version of the Journal is available at <http://www.itl.waw.pl/publikacje/kwartalnik-jtit>.

Ryszard Strużak
Guest Editor

Broadband Challenges

Acceptance Speech Presented at the Ceremony of Awarding Dr. Hamadoun Touré the Title of Doctor Honoris Causa of Wrocław University of Technology, Poland, 14 of September 2010

Dr. Hamadoun Touré
ITU Secretary General

Excellencies,
Distinguished guests,
Ladies and gentlemen,

It is a great pleasure to be here with you in Wrocław today and it is a tremendous honour for me to have been made an Honorary Doctor of this fine institution.

I am here today to accept this honour, but also to share with you my passion, as leader of the International Telecommunication Union, to Connect the World – and in particular to ensure that the digital divide is not allowed to become a broadband divide.

Here in Poland, tremendous progress has been made in the adoption and use of information and communication technologies, ICTs, especially over the past decade.

Indeed, with mobile cellular penetration of 117% at the beginning of 2010, it would be hard to find a Polish person of any age today without access to a mobile phone.

And an emerging knowledge society is clearly taking shape here, with close to 60% Internet penetration in Poland by the end of last year, and almost all access to the Internet now taking place over broadband connections.

Ladies and gentlemen,

Is broadband really that important? It is that important! I cannot over-stress this. Imagine trying to do any skilled work or apply for any skilled job in the modern world today without a broadband connection. Imagine how difficult things would be for any of us in the room here today, without email or web access. No online shopping. No access to the vast wealth of information resources on the internet. And this is just the very beginning.

We cannot predict how the development of the information society, built on broadband, will enrich the lives of everyone on the planet, but we can already see, as the second decade of the 21st century unfolds, that broadband is an extraordinary tool for social and economic development.

Indeed, it is perhaps the greatest opportunity we have ever known for human progress.

It is my firm belief that broadband is – and will continue to be – an extraordinary enabler; especially in the developing world, and especially in countries with large rural and remote populations.

Let me give you just a few examples:

With broadband networks, health services can be delivered far more effectively to ageing or isolated populations. Re-

mote monitoring of patients, for example, has proven to be far more effective than bringing people in to clinics or hospitals.

With broadband networks, we can better educate the next generations of children, wherever they live. With wireless and mobile networks we can reach out to them wherever they are.

With broadband networks, traffic networks can be streamlined, government services can be delivered more efficiently, and energy supplies can be properly monitored, controlled and conserved.

With broadband networks, we can create the right environment for applications like mobile banking, which can improve the lives of hundreds of millions of people around the world.

With broadband networks we can help to ensure environmental sustainability and help to manage and mitigate climate change.

And with broadband networks, progress can be accelerated towards meeting the Millennium Development Goals.

Distinguished guests,

Two things need to change if broadband is to become a ubiquitous resource for all the world's people.

Firstly, governments need to raise broadband to the top of the development agenda so that rollout is accelerated and the benefits are brought to as many people as possible.

And secondly, broadband needs to become much more affordable around the world.

Looking at the second point first, there are quite extraordinary disparities today between the affordability of broadband access in different countries around the world.

Let me give you some dramatic figures from ITU's 'Measuring the Information Society 2010' report.

In the top 21 countries included in the fixed broadband Internet sub-basket, published in the report, broadband subscriptions cost less than 1% of average monthly income – and under 3% of average monthly income in a further 22 countries. Here in Poland, for example, the figure is just 1.4%.

At the other end of the scale, in the most expensive 28 countries in ITU's list – most of which are UN-designated Least Developed Countries, LDCs – a monthly broadband subscription costs over 100% of average monthly income.

What a terrible irony that is!



Fig. 1. Dr. Hamadoun Touré receives his certificate of Ph.D. HC awarded him by the Wrocław University of Technology, Poland (handed by prof. Daniel Bem). © Wrocław University of Technology.



Fig. 2. Dr. Hamadoun Touré in the Hall of Wrocław University of Technology offering his formal Acceptance Speech. © Wrocław University of Technology.



Fig. 3. Dr. Hamadoun Touré opening the commemorating inscription with his name accompanied by prof. Tadeusz Więckowski (Rector of Wrocław University of Technology). © Wrocław University of Technology.

The people who can least afford access to broadband are being asked to pay the most, relative to their income.

This is scandalous. But there are grounds for optimism, and broadband access is getting more affordable every year. Worldwide, broadband prices dropped by 42% between 2008 and 2009, and broadband became more affordable in almost every market across the globe last year.

There are many reasons for this, and I am confident that broadband will continue to get more affordable as time goes on. One of the most important drivers in terms of bringing down costs in both the developed and the developing world has been an ongoing increase in capacity, itself driven by a newly competitive environment.

In Africa, for example, we are seeing new broadband capacity coming on-stream fast, and I was delighted to be personally present in Kenya when a new submarine cable was brought onshore in March, and then to be in French Guiana just last month to see new satellites for Africa being launched.

These are truly inspirational events, and signs of the very positive times in which we live.

Ladies and gentlemen,

Let me turn now to the need for government leaders to promote broadband at the national level.

Broadband networks deliver benefits across society as a whole, delivering improvements in services and reductions in costs – particularly in sectors such as healthcare, education, energy and transport.

This means that they can quickly pay for themselves. Indeed, estimates show that in most developed countries, cost savings of just 0.5% to 1.5% over ten years, in these four key sectors alone, can justify the entire cost of building national point-to-point, fibre optic networks.

Research also consistently shows that investment in any sort of ICTs has a direct positive effect on GDP growth – and that higher-end technologies, such as broadband networks, deliver the greatest benefits.

So that while a 10% increase in mobile teledensity seems to increase GDP by some 0.7 percentage points, the same increase in broadband penetration can boost GDP by an average of 1.3%.

This is why we launched our 'Build on Broadband' initiative earlier in the year, and why we are so proud to be playing a key leadership role in the 'Broadband Commission for Digital Development', which will deliver its report to the UN Secretary-General, Ban Ki-moon, this coming Sunday, just ahead of the 2010 MDG Summit in New York. The Broadband Commission is chaired by President Kagame of Rwanda and Carlos Slim Helú, Honorary Life-



Fig. 4. Prof. Tadeusz Więckowski, Magdalena Gaj (UnderSecretary of State, Poland) and Dr. Hamadoun Touré. © Wrocław University of Technology.

time Chairman of Grupo Carso. The Vice-Chairs are myself, and Irina Bokova, Director General of UNESCO, and we have a total of over 50 very high-profile Commissioners from the public and private sectors.

Our Commissioners come from many different walks of life and many different backgrounds, but they all share a common vision: the desire to see the vast benefits of broadband brought to all the world's people.

Together, we are determined to demonstrate broadband's truly transformational nature.

Together, we believe we have a unique, once-in-a-generation opportunity to drive social and economic progress.

Distinguished guests,

What will these broadband networks of the future look like? From an infrastructure perspective, it is certain that no single technology will be able to provide all the answers. Optical fibre is certainly desirable at the core of the Internet, and for the majority of backhaul traffic, to achieve a high-capacity backbone. But at the edges of the network, and in particular in the hands of end-users, it seems likely that mobile devices will deliver the majority of broadband applications and services to most people.

Indeed, this is already the case, with nearly 900 million mo-



Fig. 5. Dr. Hamadoun Touré and Wojciech Halka (General Director of NIT, Poland). © Wrocław University of Technology.

bile broadband subscriptions forecast to be achieved globally in 2010 – even if many mobile broadband subscribers do not yet use their mobile devices for Internet access.

Depending on local conditions – such as geographic location, economic prosperity, etc. – there is a role for a host of different technological solutions for providing broadband access: from cable to fixed wireless; from satellite to microwave; from xDSL to mobile technologies; and many more.

As a result, policy-makers should seek to adopt a technology-neutral approach, as they need to plan and accommodate new upgrades of current fixed and wireless technologies, as well as brand new technologies which do not yet exist.

Policy-makers also need to recognize that demand for radio-frequency spectrum is likely to grow rapidly – bearing in mind that as a precious resource, spectrum allocation must be based on costs, efficiency of use, and the needs of users.

Ladies and gentlemen,

Broadband can bring the world's riches within reach of all the world's people. But it also opens up new opportunities for wrong-doers, and in particular cybercriminals.

It is estimated that each year cybercriminals steal up to

a trillion dollars worth of intellectual property from businesses worldwide. In addition, many millions of individuals have their privacy violated, suffer identity theft or have their hard-earned savings stolen from them.

Governments constantly face cyberattacks – and terrorists increasingly rely not just on their weapons, but on the power of cyberspace technologies like GPS and VoIP to sow destruction.

As a result, the next major war will probably begin in cyberspace rather than on the ground.

Given the scale of the threat – and the phenomenal harm that can be caused by even a single cyber attack – we cannot rely on ad hoc solutions or hope to survive by strengthening our defences only after attacks have occurred.

No – we must work together, to ensure a coordinated response.

This is why ITU is playing a lead role in coordinating global efforts in this area, and why we launched the Global Cybersecurity Agenda in 2007.

The GCA is now in its operational phase, with a physical home in Malaysia at the headquarters of IMPACT – the International Multilateral Partnership Against Cyber-Threats – and I strongly encourage all ITU Members to take advantage of these facilities.

We are also increasingly seeing children being targeted online by all manner of wrong-doers, including paedophiles.

As a result we also launched – within the framework of the GCA – the Child Online Protection (COP) initiative, which has been established by ITU and other stakeholders as an international collaborative network for action to promote the online protection of children worldwide.

Cybersecurity is a global issue, which can only be solved with global solutions. So one of the most important actions we must take to enhance security in cyberspace is to increase cooperation and coordination at the global level.

Existing frameworks, while good, are not in fact global. The best example is the Budapest Convention on Cyber-crime, which has been ratified by only 26 countries. The nature of the convention makes it hard for other countries to join.

ITU therefore offers to host international discussion – identifying aspects we can agree on (there are many); recognizing that there are many different viewpoints; and working with all stakeholders, from government, industry and civil society.

Distinguished guests,

In closing, I would like to mention one further issue which is of importance to every single person on the planet: I refer of course to the environment, and the prospect of catastrophic climate change.

I am personally proud and pleased to be working in a sector where we can act as a strong force for good in this area.

ICTs are the single most powerful tool humankind has at its disposal to avoid potential climate catastrophe. While technology contributes around 2.5% of global greenhouse

gas emissions, more efficient use of modern technologies could cut global power consumption by 15%.

ICTs help to reduce waste, cut business travel and make industry more efficient. And new technologies being developed within ITU – such as Next-Generation Networks – can reduce network and data centre power consumption by up to 40%.

The universal charger, which was standardized by ITU last year, will deliver a 50% reduction in standby energy consumption, eliminate up to 80,000 tonnes of redundant chargers, and cut GHG emissions by at least 13 million tonnes annually.

Satellites are also tremendously important – not just in monitoring, but in helping increase food output and reduce emissions. Satellite-based intelligence services for farmers, for example, which cost less than US \$ 15 per hectare annually, can increase yields by as much as 10%. And using satellite monitoring produces 98% fewer emissions than ordinary ground monitoring.

Satellites also play a vital role when natural disasters and catastrophes occur, often providing the only reliable communications links when other networks have been destroyed. ITU is proud to play a very active role wherever we can – in Haiti and Chile, earlier this year, for example, where we were quick to provide emergency communications services after the earthquakes there.

To close, let me thank you once again for the honour you have bestowed upon me here today – and let me encourage each and every one of you to act as an ‘ICT advocate’, and in particular as a broadband advocate!

Dr. Hamadoun Touré is the ITU Secretary-General since 2006, after serving eight years as Director ITU/BDT. Prior to his election as Director of BDT he worked for ICO Global Communications. Earlier he was employed by the International Telecommunications Satellite Organization (INTELSAT). On the national scene, he was with the national telecommunication operator in Mali. He holds a Masters Degree in Electrical Engineering from the Technical Institute of Electronics and Telecommunications of Leningrad (USSR) and Doctor of Philosophy Degree (Ph.D.) from the University of Electronics, Telecommunications and Informatics of Moscow (Russia). He is proficient in English, French and Russian. Dr Touré contributed to many important projects and conferences. As Director BDT, he played a key role in the WSIS process. As the Secretary General, he is leading the ITU toward an innovative, forward looking organization satisfying the needs of ITU Members and meeting new challenges. Dr. Touré has been awarded the highest national decorations and honorary doctorates of six universities. He is a member of IEEE, of International Telecommunication Academy, and of the Royal Swedish Academy of Engineering Sciences.

International Co-operation Needed to Face Information Age Challenges

Magdalena Gaj

UnderSecretary of State, Ministry of Infrastructure, Poland

I am delighted to present to the delegates of the ITU 2010 Plenipotentiary Conference in Guadalajara Mexico a summary of recent changes in telecommunication sector in my country. Poland has become a member of the International Telecommunication Union in 1921 when it gained independence after the 1st World War. Earlier it ceased to exist as an independent entity for more than a hundred years. From the very beginning, in close collaboration with other ITU Members, the Polish telecommunication sector was growing fast. That growth was stopped during the 2nd World War after which the sector was completely disrupted. Its reconstruction started immediately at the war end, but accelerated rapidly only after the transformation of our political and economic system in 1989, as result of the famous Solidarity Movement.

My first task after the ministerial appointment in 2008 was to identify urgent issues to focus on. Although Poland is a Member of the European Union from 1 May 2004, it joined EU when the other Members were already in an advanced state of development. In the telecommunication sector we have suffered a number of problems:

- Low broadband penetration rate in comparison with more advanced EU countries.
- Strong position of the incumbent operator on the market.
- Newly build housing without broadband infrastructure.
- Disproportionate high access charges
- Barriers in the investment process.
- Insufficient IT knowledge among population.
- Generation exclusion.
- Exclusion of poor social strata and inhabitants of rural regions.

To solve them, we have joined the “i2010” and other European programs and started “Digital Poland” initiative. We are increasing our involvement in various ITU and other international initiatives. We are seeking assistance from our own National Institute of Telecommunications and we welcome collaboration with governmental and private

sector entities from abroad. The challenges of information and knowledge society listed above, are common for a number of countries. We strongly believe that our problems can successfully be solved all over the world only in close international co-operation and wide exchange of national experience of each country. The ITU has created the best framework for such a joint effort. My country is ready to share our experience with other ITU Members interested.

In Poland, our Ministry of Infrastructure, together with the regulatory Office of Electronic Communication, have started with deep modifications of our telecommunication legislation and regulation structure. A new Act on “Support of the Development of Telecommunication Services and Networks” has been prepared and came into force just in July this year. New legislation authorizes local governments and authorities for telecommunication activities, such as network and services provision. The new “rights of way” impose obligation on commodities enterprises, such as public ways, electrical, water and sewing companies, to share their infrastructures with telecommunication operators. The construction law is modified implementing simplified procedures of construction licensing and projects approvals. New competencies are given to electronic communication regulator, including right to appeal any local prohibition against new telecommunication investments, including radiocommunication masts and stations. The common aim of these rules is to faster and make easier investment processes in telecommunication sector, especially in new regional broadband infrastructure, new generation access networks (including wireless networks) and new service platforms. We believe these new regulations will encourage wider investments in our telecommunication sector and will end with higher penetration of broadband access and new services in our country. We invite all partners of international ITU community to share their experiences in similar efforts.

New telecommunication and information technologies are indispensable vehicles in our way for better life. They promote social and economical development, but they create numerous challenges in each country. International exchange of best practices can make this way easier. We are ready to co-operate with the ITU community on that way.



Magdalena Gaj, born in 1974 in Chełm, Poland. Graduated from Faculty of Law (MA degree) at University of Maria Curie-Skłodowska in Lublin; applicant at District Chamber of Solicitors in Lublin, 1998–2002. Applicant at Solicitors Office till 2001. Worked in the Office for Telecommunication

Regulations, Office for Postal and Telecommunications regulation since 2001, and now – Office of Electronic Communications (UKE) in Legal Department, since 2005 – Director of Legal Department. Scope of competence: legal service of UKE President within telecommunications, postal, administration, civil and EU laws. From 2008 in the Ministry of Infrastructure, as Under-Secretary of State responsible for telecommunication sector's matters in Poland.

National Institute of Telecommunications – International Partnership: Yesterday, Today and Tomorrow

Wojciech Hałka

General Director of NIT, Poland

The ITU 2010 Plenipotentiary Conference in Guadalajara, Mexico, is top policy-making body focused on future role of the organization in coming years. This meeting of the delegates from the countries all-over the world, creates an excellent opportunity for the exchange of experiences. In Poland there is a hot public discussion on-going on the future challenges related to the implementation of new information and communication technologies (ICT) and services. People are expecting fast modernization of their living, economic growth and social promotion. New governmental projects such as “Digital Poland” or “Terrestrial Digital Television” have been launched facing these challenges. The Digital Poland project covers innovative initiatives in education sector (ICT skills), full digitalization of National Archives, Art and Historic Collections, e-government, and support for new e-services providers and telecommunication infrastructure builders. Although the goals of those projects are of common agreement, details of their realization ways are of public discussions. These projects required the legislative and regulatory structure to be reviewed and modified where necessary, see the accompanying paper of Magdalena Gaj. Experience of other countries in similar projects and the best practice cases are monitored and followed carefully by social partners, organizations, involved business and administration.

The National Institute of Telecommunications, NIT, a 75-years old research institution based in Warszawa (with branches in two other cities, Gdańsk and Wrocław) is deeply involved in all those undertakings. With our highly qualified staff of 250 R&D employees, we are a significant partner in most of those projects. Being a public institution, we serve as an independent Research Center working for business partners as well as for the Administration. Let me present a short selection of our recent projects:

- frequency planning for digital TV broadcasting, in many variants of single and multiple frequency plans. Those projects have been done for the Administration and for individual broadcasters;
- electromagnetic compatibility studies and spectrum management proposals, for the National Regulator (Office of Electronic Communication) and for business partners;
- broadband development strategies proposals for Regional Governments;

- information data base system development for telecommunication broadband networks inventory in Poland (on-going project);
- telecommunication services quality measurement system (automatic, for PSTN and GSM networks) for inspection service of the National Regulator;
- data mines and telecommunication decisions support systems for various telecom operators;
- reports, studies and expertise on numbering plans, signalization systems, data retention, billing systems, telecommunication frauds, market and regulatory analysis for governmental agencies and business partners;
- conformance tests and authorized laboratories certificates for equipment suppliers and producers.

We have developed close co-operation with many partners in the framework of a number of scientific and industrial consortiums. Our main partners in Poland are Technical Universities in Warszawa, Gdańsk, Wrocław, Poznań, Kraków as well as R&D units of telecommunication operators, broadcasters, producers (e.g., Orange Labs, Poland) and independent consulting companies. A number of our research projects have been awarded grants by the Ministry of Science and Higher Education. Of this type, gained in public concurs, are “New Generation Services and Networks” (2008–2010) and “Future Internet Engineering” (2010–2012), both projects.

The National Institute of Telecommunications has developed also international co-operation in three areas:

- 1) scientific and technical projects,
- 2) international conferences, and
- 3) training/education.

As concerns the first area, NIT participates actively in a number of European Projects, funded from EU Regional and Scientific Funds, such as Framework Projects (NEMO, BReATH, CRUISE), InterReg (EfficienSea) and COST (Actions 270, 291, 299, MP0702, 2100). In the second area, NIT organizes and co-organizes a number of international conferences. Examples are the International Conference on Transparent Optical Networks (ICTON) and

the conference on Decision Support for Telecommunication and Information Society (DSTIS). In 1972, NIT co-founded the Wrocław International Symposium on EMC that continues until now, and in 2010 was held under the auspices of Dr. Hamadoun Touré, Secretary General of ITU. It has been awarded the Special Symposium Recognition Award by the IEEE Electromagnetic Compatibility Society: *“In recognition of the vision and contribution of the founders of the Wrocław International Symposium and Exhibition on Electromagnetic Compatibility [...] Held for the first time September 14 to 16, 1972 [...] it was the first regular International Symposium and Exhibition on Electromagnetic Compatibility held in Europe.”* The IEEE is the world’s largest professional association dedicated to advancing technological innovation and excellence for the benefit of humanity. Its grants and recognitions honor endeavors that have made a lasting impact on humanity, technology, and the profession. In 2010, the Wrocław EMC symposium has been transformed into “EMC Europe”.



Fig. 2. The lesson of antenna measurements.



Fig. 1. Memorandums of Understanding: NIT-ICTP and NIT-ITU.

Our international co-operation is organized within framework agreements, such as Letter of Intent or Memorandum of Understanding (MoU). In that form we co-operate with a number of well-known international and national institutions. Among international ones are International Institute for Applied Systems Analysis (IIASA) in Luxemburg, Austria, and the Abdus Salam International Center for Theoretical Physics (ICTP) in Trieste, Italy. The list of our national partners abroad contains the Japan Advanced Institute of Science and Technology (JAIST) in Nomi, Japan, Instituto de Estudos Sócios-Econômicos (INECS-Coimbra) in Coimbra, Portugal, University Stellenbosch in Republic of South Africa and others. The co-operation agreements define the scope of common studies, conferences, workshops, and seminars as well the exchange of the staff, etc.

As an example, NIT cooperates with ICTP in organizing the ICTP ITU/BDT schools, workshops, and awareness conferences for young scientists. Each year, some 25 young scientists, mostly from the developing countries, participate in these activities that usually are lasting for a few weeks. In that way NIT contributes to the advancement of

human resources and research and development capabilities in the developing countries in the field of radio science and technology. These activities, held in the ICTP in Trieste, are often complemented by activities and projects abroad. For instance, two young scientists from Technical University of Malawi spent in 2008 3 months in Poland, in the frame of the NIT-ICTP MoU. They were finishing their Advanced Research Studies on Low-Cost Antennas for Wireless Networking in Developing Countries in the NIT EMC&Antenna Laboratories in Wrocław. This was made possible thanks to the grant offered them by the ICTP.

Since its very beginning NIT maintains special relations with the ITU. Most of contributions submitted to ITU on behalf of our country have been based on the works done in NIT and/or authored by our staff. Many of them have been included in final texts of several official ITU documents. Our staff members have participated in technical and stan-



Fig. 3. The certificates of completed studies in Electromagnetic Compatibility Department of NIT in Wrocław of Mr. Timothy Chadza and Mr. Mayamiko Nkoloma, scientists from Malawi.



Fig. 4. The group of students of Advanced Research Studies on Low-Cost Antennas for Wireless Networking in ICTP, Trieste.

dardization works of selected Study Groups in ITU-T and ITU-R sectors. Some of them served as Working/Study Group Raporteurs, V-Chairs and/or Chairs. In the past, for instance, the ITU/CCIR Plenary Assembly New Delhi 1974 elected Ryszard Strużak for the post of V-Chair of CCIR Study Group on Spectrum Management and Monitoring. He resigned in 1985, and left NIT to serve further the ITU community with his knowledge in a high post in the Geneva Headquarters. Earlier, Jerzy Rutkowski, also

our staff member, served successfully at that post for several years. There is no place here to list all our staff members, who contributed to various activities, including ITU Expert Missions. Myself, I was as an ITU expert working on an ITU project in Burundi. From among our NIT colleagues, perhaps Ryszard Strużak and Władysław Moroń have gained the widest recognition within the ITU serving for many years as the elected members of the Radio Regulation Board.

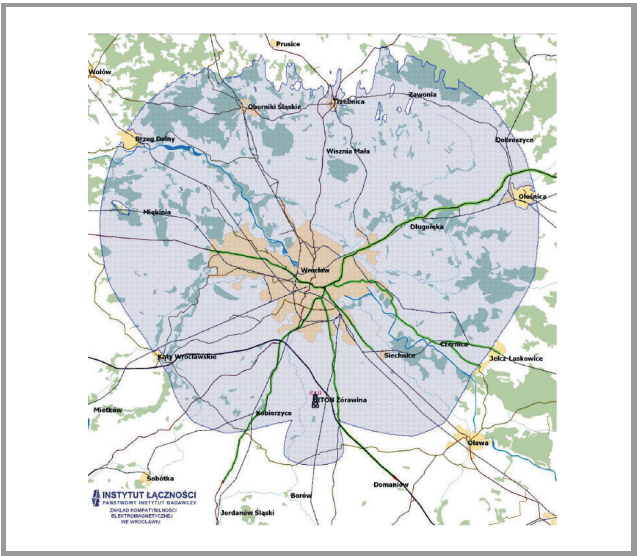


Fig. 5. Example of the coverage map of the first in Poland trials of the mobile DAB+ system.



Fig. 6. Ground Conductivity Map of Poland.

More recently, the NIT's experts have prepared also our in-kind contribution for ITU-R SG III: reports on experimental tests of Digital Radio Broadcasting DAB+ emissions in Wrocław region as well as the updated Map of Ground Conductivity of Poland. With the support from ITU-D, we have organized seminars and workshops, i.e., the 2007 ITU Workshop and National Conference on New Generation Networks in close cooperation with the experts from Canada. We co-operate also with countries of Central and Eastern Europe, within the framework of ITU regional activities. In February this year NIT has established a Node of the Centers of Excellence Network Project in the European Region, after signing the Memorandum of Understanding with Sami Al Bashir Al Morshid, Director of Telecommunication Development Bureau (BDT). In this connection we are organizing later this year two workshops on Long Term Evolution (LTE) systems and on Future Internet technologies. We invite all interested participants to join us at these events.

We believe our co-operation with international partners adds a significant “networking” value and is beneficial to all parties involved. The benefits are of different nature, from sharing (i.e., reducing) the costs of common endeavors, to gaining new knowledge, skills, and experiences, invaluable in today's rapid changes of telecommunication technologies, services and applications.



Wojciech Hałka, born in 1949, graduated from Warsaw University of Technology (1971, BScEE), postgraduate studies CITCOM in Ecole Supérieure d'Electricité, Antenne de Rennes (1978) and European Studies MATRA in Universiteit Maastricht (1978). He worked as researcher in National Institute of Telecommunications in War-

saw, Poland, till 1991, at Spie Batignoles (FR) for National Power Center in Egypt (1986–1987) and as ITU expert in Burundi (1991). In 1991–1998 he worked in the Ministry of Communications, as director of Regulatory and Development Department. He was a delegate of Polish administration in Kyoto ITU Plenipotentiary Conference (1994) and in WTO, CEPT, ECTRA conferences. Next, till 2003 at Netia Telekom S.A. responsible for data transmission network and interconnections agreements. In 2003–2005 at the Ministry of Infrastructure, as Under-Secretary of State responsible for telecommunication and postal sectors he implemented the 2nd EU Regulatory Package into Polish Telecommunication Act. In March 2006 he joint back the NIT, as deputy director for development, and then in July 2009 he was appointed as General Director of the NIT.

Rural Telecommunications Infrastructure Selection Using the Analytic Network Process

Yousef Gasiea, Margaret Emsley, and Ludmil Mikhailov

Abstract—The decisions involved in rural settings are of complex nature, with some aspects compounded by the presence of intangible criteria. Hence, a suitable approach is needed that can produce effective solutions. This paper describes the applicability of a multicriteria decision-making method, specifically the analytic network process (ANP), to model the selection of an appropriate telecommunications infrastructure technology, capable of deploying e-services in rural areas of developing countries. It aims to raise awareness among telecommunication planners about the availability of ANP, and to demonstrate its suitability to enhance the selection process. The proposed model is constructed based on concerned experts' views of relevant selection criteria and potential technology alternatives. Its network structure caters for all possible dependencies and interactions among criteria and alternatives.

Keywords—*analytic hierarchy process, analytic network process, multicriteria decision making, rural telecommunications, technology selection.*

1. Introduction

Telecommunications technology is evolving rapidly and offers information links between urban and rural areas that can overcome distance barriers and provide e-services to these hardly accessible areas. Recent technological advances in transmission systems like fiber optics, wireless and satellite can now supply services to these locations at affordable prices. However, with different criteria for technology evaluation and various telecommunications infrastructure alternatives available nowadays, the selection process becomes complicated; there is uncertainty and multiple conflicting objectives with sociological, demographical, environmental, political, cultural, economic and technical aspects. This raises the need for some kind of structure or model, based on a suitable multicriteria decision making (MCDM) method.

Some relevant papers cited in literature tackling problems from rural telecommunications field using such methods, with particular focus on the application of analytic hierarchy process (AHP) to rural telecommunications include: Nazem *et al.* [1] use the AHP, to develop a two-phased decision support system to aid the design of rural area telecommunication networks and in [2] examines ways of building an effective rural telecommunications network to facilitate rural development in an information-intensive society. Lee and Kim [3] present a methodology using analytic network

process and zero-one goal programming (ZOGP) for information systems projects selection problems that have multiple criteria and interdependence property. In another paper, Nazem *et al.* [4] develops a specific multicriteria decision support mathematical programming model for dealing with the definition of a “hub structure” that is the selection of a number of “nucleus cities” in the context of a rural network planning process. Chemane *et al.* [5] use DecidIT tool based on MCDM to improve the quality of decisions in selecting internet access technologies. Sasidhar and Min [6] use AHP to select the optimal access technology for a rural community under a multiple number of criteria such as cost quality and speed. Nepal [7] applies AHP to the evaluation of rural telecommunications infrastructure. Finally, Andrew *et al.* [8] present a model regarding the applicability of using the AHP for enhancing the selection of communication technologies for rural areas.

While significant decision models are being presented in these papers, but, very few studies have considered all criteria relevant to rural telecommunications, and most of them obviously apply no factor interactions. For example, if a model's emphasis is mainly technical, then the economic, social, regulatory and environmental criteria are probably not adequately addressed. Basically, the AHP is a suitable method when optimization is not pursued, resources are not restricted, and interdependencies between factors do not exist [9]. However, such models do not consider important issues such as interaction among and between decision making levels/clusters as well as dependency among qualitative factors. These are important issues in rural telecommunications decision problems which cannot not be structured hierarchically because they involve many interactions and dependencies requiring a MCDM method to holistically deal with qualitative and quantitative data, with different conflicting objectives, to arrive at a consensus decision in relation to the choice of a suitable rural telecommunication technology.

To the best knowledge of the authors, applications of the analytic network process (ANP) to the selection of rural telecommunications infrastructure technologies have not been cited in the published literature. This paper therefore attempts to fill this gap in the literature to particularly allow for the explicit consideration of dependencies and interactions in the decision making process and still maintains the acknowledged advantages of the AHP method. The ANP is chosen in this paper because of its several advantages

over the AHP and other MCDM methods, such as its holistic approach, in which all the factors and criteria involved are laid out in advance in a network system that allows for dependency. Its power lies in its use of special ratio scales to capture interactions for making accurate predictions and reach better decisions [10]. Moreover, its suitability in offering solutions in a complex multicriteria decision environment, together with the availability of software supporting its functions, further acknowledge its applicability to tackle such a problem. It has also proved to be successful in utilizing expert knowledge to tackle several selection problems, e.g., [3] and [11].

The remainder of this paper is organized as follows. Section 2 articulates the selection of rural telecommunications infrastructure problem. The underlying methodology of the proposed approach, the ANP, is briefly introduced in Section 3. The development of the proposed model is explained in Section 4. The pairwise comparisons are described in Section 5. The results are discussed in Section 6 and the paper ends with conclusions in Section 7.

2. The Choice of Technology for Rural Telecommunication Infrastructure

There is a need to provide access to the main telecommunications network and expand connectivity to such areas, thus enabling the rollout of the appropriate telecommunication services. However, the choice of appropriate telecommunications infrastructure technology that will provide the required e-services within various constraints is a challenge. Typically, technology selection is based on a mixture of different criteria, one of which is the remoteness of a village. If the village is within 35 km of the nearest local exchange, telecommunication services can be provided to that village using a one-hop last-mile link. However, if the village is further away, at least two transmission hops must be established [12]. Hence, two types of telecommunications infrastructure technologies are needed to provide rural telecommunication services, namely backbone network (core) and access network (last mile).

The backbone network provides the long-haul signal transmission from the country's main telecommunication centre to the remote access network, i.e., trunking services [12]. This network may be wireless or wireline, including analogue and digital transmission technology over fiber optic, wireless or satellite transmission media [5]. The access network provides the connectivity between the end-user and the backbone network and may be based on wireless or wireline technologies, e.g., copper wires or wireless, connected to network nodes at the edge of the backbone network. Technologies in both networks can be circuit-switched or packet-switched. Any decision made for each of these two segments must take into account the characteristics of rural settlements.

The primary focus of this paper is mainly on the backbone network by attempting to provide a structure of the deci-

sion problem and proposing a technology selection model of such an infrastructure. The telecommunications backbone is, in general, a key problem for rural information infrastructure, as low population density is linked to high cost of service for any communications technology, especially for wireline services. It poses the greatest challenge to bringing affordable telecommunication services to rural residents. However, once it is in place and running, it will be possible to connect other nearby rural villages with a wide range of telecommunication technologies and needed services. The infrastructure technology selection process, especially in the case of rural telecommunications in developing countries, is a multi-faceted, multi-criteria decision making problem, requiring consideration of some wide-ranging qualitative factors related to socio-economic and political issues. These are hard to quantify and will have great impact on the selection process, in respect of the social, environmental, regulatory and demographical concerns, etc.

Furthermore, in order to incorporate other tangible factors, in the absence of past statistical data to analyze, such as technical and economic related factors, etc., it is necessary to use a suitable multicriteria method for analysis and synthesis by a group of experts rather than an individual. A telecommunication operator usually receives several technology solutions from external vendors. The challenge of matching the parameters of an engineering problem to the available solutions becomes a challenge to the telecommunications engineer in this particular selection phase [8]. A typical conceptual rural telecommunications infrastructure selection model is illustrated in Fig. 1.

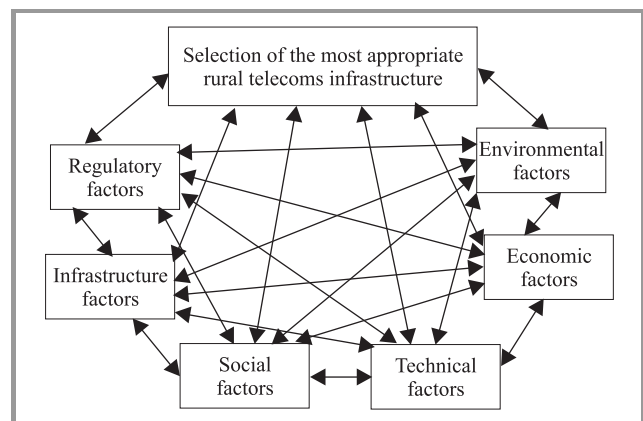


Fig. 1. A conceptual model for the selection of rural telecommunication infrastructure (revised and adapted from [13]).

The obvious significant implication of this conceptual model is that the technical factors are only one subset among others when selecting rural telecommunication technologies, albeit a necessary part. The other factors, such as the sociological, environmental, economic, regulatory and the infrastructure-related are regarded as essential factors that also need to be considered. This can be envisaged as a holistic approach in which the outcome of the selection process is not only dependent on the technical factors, but

arises out of the interactions among the various factors. An ANP-based decision model is therefore proposed as a suitable methodology because “*decisions obtained from a network can be significantly different from those obtained from a more complex hierarchy*” [14]. It is constructed to include an in-depth and comprehensive examination of all pertinent factors and will be dependent on the perceptual weightings, provided by telecommunications experts.

3. The Analytic Network Process

The ANP is a multi-attribute decision making approach developed by Thomas L. Saaty and was originally called the supermatrix technique [15]. It is a generalization of the AHP decision methodology where hierarchies are replaced by networks, allowing the capturing of the outcome of dependence and feedback within and among the clusters of elements. Its network structure differs from a hierarchy as illustrated in Fig. 2 [10]. The hierarchy has a goal, levels of elements and connections between the elements.

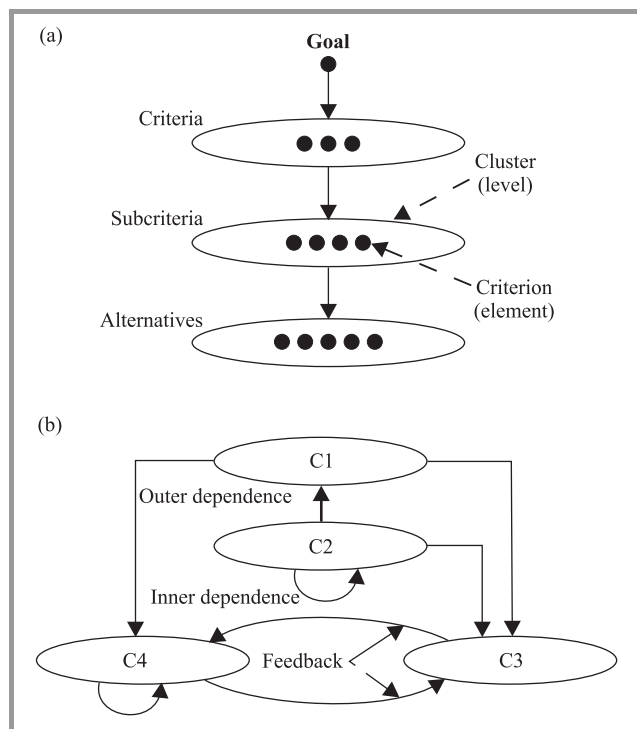


Fig. 2. Examples of a hierarchy (a) and a network (b).

It has no inner dependence and no feedback from lower to higher levels. Unlike the hierarchy, the network structure has no levels but clusters of elements where every element can depend on any other element. The influence is transmitted from one cluster to another (outer dependence) and back, either directly from the second cluster, or, by transiting through intermediate clusters along a path which sometimes can return to the original cluster forming a cycle [10]. The existence of feedback indicates there is mutual outer

dependence of criteria in two different clusters, which prevents the problem from being modeled hierarchically due to the difficulty in deciding which cluster is higher/lower than the other. Also, because of inner dependence, the relationships between same level criteria are not represented hierarchically.

The specific ANP model is based on the reasoning, knowledge and experience of experts in the field and relies on the process of eliciting managerial inputs, allowing for a structured communication among decision makers, so that it can act as a qualitative tool for strategic decision-making problems. “*It is a relatively new methodology that is still not well-known to the operations research community and practitioners*” [9]. With its capability to deal with dependence and feedback, it is the most general framework for a detailed analysis of societal, governmental and corporate decisions that is available today to the decision-maker [15]. Therefore, in recent years, there has been an increased use of the ANP in a variety of decision making problems and numerous applications have been published in literature [16].

The ANP is a coupling of two parts. The first part consists of a control network of criteria that controls the interactions in which the criteria should be identified, organized and prioritized in the framework of a control network. The second part is to derive a network of influences among the factors and clusters, i.e., the influence of elements in the feedback system with respect to each of these criteria. Paired comparison judgments of homogeneous elements are performed and synthesized to obtain the priorities of these criteria. The ANP then joins all possible outcomes together in its structures and both judgement and logic are used to estimate the relative influence from which the overall answer is to be derived [15]. The *SuperDecisions* software can be used to perform matrices computation and solve AHP/ANP problems [17].

4. The Development of the Decision Model

In this section, we introduce an ANP model and its development to show how the ANP can be used in the rural telecommunications environment. As each telecommunications infrastructure provider will have its own set of criteria. The attempt here is to present a generalized model based on factors and alternatives identified from the published literature, best practices and telecommunications experts that could then be adapted or extended to support a particular context or a situation of a developing country.

4.1. Setting Selection Criteria

To adapt the ANP methodology for such a technology selection process, it is the foremost activity of the researcher to examine the relevant issues involved. Hence, the first task is the definition of the criteria that will be used for

the selection of the appropriate technology for rural connectivity. The activities used to consolidate the final list of the selection criteria, were:

- an intensive literature survey of past studies on similar problems, including: [5], [6], [8], [12] and [18]; the outcome of this activity was the consolidation of an initial list of criteria that comprises the most important factors for the problem at hand;
- interactions with telecommunication experts both from industry and academia from all over the world, who were contacted through e-mail to provide feedback on the initial list of criteria; from the aforementioned activities, a consolidated list of 31 selection criteria, deemed to affect the planners' decision in the choice of rural telecommunications backbone infrastructure, can be observed in Table 1.

Table 1
Ordering and clustering of criteria according to relevance and importance

Cluster	Criteria	Mean
(A) Technical	(A1) Reliability	4.00
	(A2) Ease of maintenance	3.94
	(A3) Remote network management	3.88
	(A4) Compatibility	3.81
	(A5) Ease of installation	3.72
	(A6) Scalability	3.54
	(A7) Bandwidth	3.53
	(A8) Flexibility	3.52
	(A9) Latency	3.30
(B) Infrastructure	(B1) Coverage range	3.80
	(B2) Security of physical infrastructure	3.73
	(B3) Proposed usage	3.40
	(B4) Availability of skilled technicians	3.34
	(B5) Access to existing telecoms infrastructure	3.32
	(B6) Remoteness of area	3.26
	(B7) Rollout time	3.11
	(B8) Parallel infrastructure	2.97
(C) Economic	(C1) Operating cost	4.13
	(C2) Funding sources	4.11
	(C3) Capital cost	3.98
	(C4) Return on investment	3.63
	(C5) Economic development of area	3.32
(D) Social	(D1) Demand	3.77
	(D2) Affordability	3.73
	(D3) Population density	3.48
	(D4) Community of interest	3.42
(E) Regulatory	(E1) Spectrum availability	3.74
	(E2) Licensing constraints	3.52
	(E3) Rights of way	3.30
(F) Environmental	(F1) Terrain topography	3.24
	(F2) Climatic conditions	3.00

4.2. The Online Survey

In order to rank the selection criteria according to their relative importance, an online questionnaire was designed. Telecommunication experts were asked to rate the importance of each factor using a five-point Likert-type scale, ranging from: *not important* = 1, *moderately important* = 2, *strongly important* = 3, *very strongly important* = 4 and *extremely important* = 5. A pilot survey was conducted before posting the questionnaire online and subsequently the questionnaire was slightly modified.

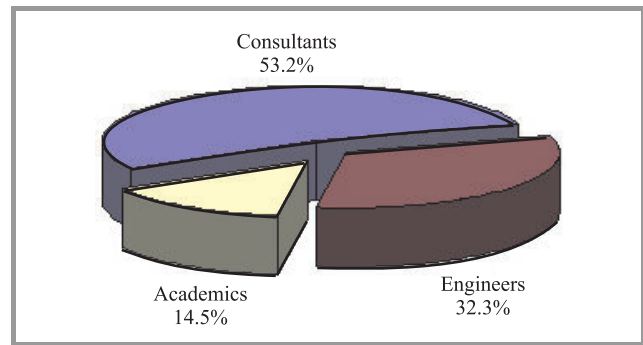


Fig. 3. Categorization of respondents by their professional background.

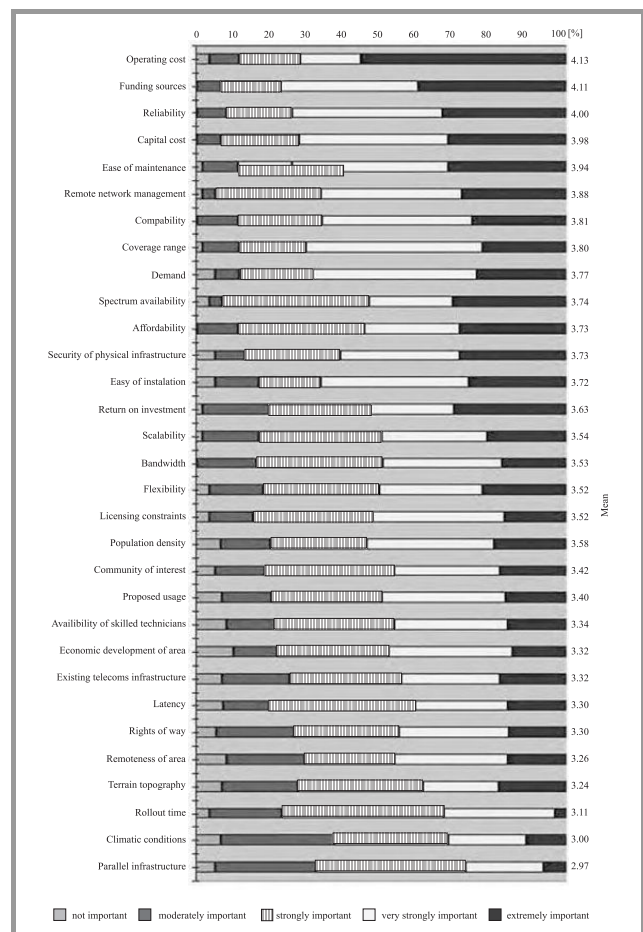


Fig. 4. A graphical representation of the online survey results in percentages.

Table 2
Comparisons of some features of potential alternatives

Alternative	Advantages	Disadvantages
(G1) Fiber optic cable	High speed More reliability High flexibility	High cost Long rollout time Most difficult to deploy
(G2) Power line communication	Simplicity Low cost Use of power lines	Less reliability Data signal disruption Noise and interference
(G3) Microwave link	High speed Low cost equipment Fast deployment	Low reach and line of sight Licensing constraints Less bandwidth and flexibility
(G4) Satellite communication	Wide coverage Ease of deployment Overcomes topography	High latency High cost Limited bandwidth

The obtained responses effectively reached 62 responses, which is considered adequate because the purpose of the survey was mainly to obtain a range of diversified expert opinions with respect to each particular selection factor.

The respondents' profiles showed that all of them are generally involved in telecommunications field, where some of them are particularly dealing with rural telecommunications projects. They can be categorized by their professional backgrounds into three categories as shown in Fig. 3: of the 62 respondents 20 (32.3%) of them work as telecommunication engineers, 33 (53.2%) as consultants and 9 (14.5%) as academics. This mix up of the respondents' expertise confirms their familiarity with the selection factors and also indicates that they were very well placed to provide useful data for such a survey.

The results were then analyzed using SPSS (SPSS Inc., 2006), and univariate descriptive statistics were generated, including the relative importance index for each factor. Figure 4 summarizes the obtained results and shows that all proposed criteria are mostly within the *strongly important* and *very strongly important* categories, the only exception being the results of the "operating cost" criterion which is inclined more towards the *extremely important* grade.

4.3. Grouping of Criteria into Clusters

The mean rating values were used to group the criteria into six clusters coded A through F according to relevance, in this order: (A) Technical, (B) Infrastructure, (C) Economic, (D) Social, (E) Regulatory, and (F) Environmental. Each cluster only includes criteria that are comparable or do not differ by orders of magnitude [10]. Table 1 shows the coding and the ordering of criteria for all clusters.

4.4. Alternatives Identification

The activities abovementioned in Subsection 4.1 were repeated in order to identify potential technology alternatives. The published literature, e.g., [19], identified four techno-

logical solutions to provide rural backbone infrastructure to promote e-services in rural areas of developing countries that include two wireline technologies: fiber optic cable and power line communication, and two wireless technologies: fixed wireless and satellite, which were initially highlighted as candidate decision alternatives for this research. After consultation with telecommunication experts, the alternatives finally selected for this research are (G1) fiber optic cable, (G2) power line communication, (G3) microwave link and (G4) satellite communication. Table 2 briefly summarizes some characteristics of alternatives.

4.5. Assessing Dependencies

After structuring the decision problem, the next step is to examine the dominance of influence among criteria. In order to fulfil this task, a new survey questionnaire was distributed to experts who had an overview of the research, were interested and actually involved in the field of rural telecommunications, who were asked to identify the dependencies among criteria. Seven completed questionnaires were collected. The majority rule was then used to aggregate the responses into a single matrix, which was developed using a zero-one matrix of criteria against criteria using a binary value of 1 to signify dependence of one criterion on another, and zero otherwise [20]. A majority condition of 4 out of 7 (4/7) experts' consensus (i.e., 57%) was considered as a minimum requirement for any entry that indicates the existence of a direct relationship between any pair of criteria.

Table 3 shows all possible connections, where the entries can take the following values:

- 0 indicates no relationship exists based on 7 experts' consensus;
- 0 indicates the entries have obtained < 4 experts' consensus;
- 1 indicates the entries have obtained ≥ 4 experts' consensus.

Table 3
The aggregated dependency matrix showing connections among all elements

Clusters Criteria		with respect to																																			
		A1	A2	A3	A4	A5	A6	A7	A8	A9	B1	B2	B3	B4	B5	B6	B7	B8	C1	C2	C3	C4	C5	D1	D2	D3	D4	E1	E2	E3	F1	F2	G1	G2	G3	G4	
A	A1	0	1	0	0	0	0	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
	A2	0	0	1	1	1	0	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	
	A3	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	
	A4	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	
	A5	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	
	A6	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	1	0	1	1	0	1	1	1	1	1	1	
	A7	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	1	0	1	0	1	1	0	0	0	1	1	1	1	
	A8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	1	1	1	
	A9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1	1	
B	B1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1	1	
	B2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	
	B3	1	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1	1	1	1	1	0	1	0	1	1	1	
	B4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	
	B5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	1	1	
	B6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	1	1	1	1	
	B7	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1	0	1	1	1	1	1	1	1	1	1	
	B8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	1	1	
C	C1	1	1	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1	1	1	1	
	C2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	
	C3	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	1	1	
	C4	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	1	0	1	0	1	1	1	1	1	1	0	0	0	1	0	1	1	1	
	C5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	1	1	1	
D	D1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	1	1	1	
	D2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	1	1	1	
	D3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	
	D4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	
E	E1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	1	1	1	
	E2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1
	E3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1
F	F1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	F2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
G	G1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	
	G2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	
	G3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	
	G4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	0	

As a result, the entries represented by 1 indicate the existence of a direct relationship from criterion i to criterion j based on the consensus of at least 4 experts, i.e., if criterion i depends on criterion j , the entry a_{ij} will take 1. The criteria in the rows are evaluated with respect to the criteria in the columns, i.e., the 1 in the columns will determine which criteria in the rows are to be pairwise compared with respect to that column. Subsequently, a pairwise comparison matrix will be constructed only for the dependent criteria. Using the *Design module* of the *SuperDecisions* software [17], the network model was constructed according to Table 3, the connections between clusters are illustrated in Fig. 5. A cluster is connected to another cluster when at least one element in it is connected to at least two elements in another cluster. It should be noted that two-way arrows connecting the clusters represent interdependencies among elements, where an arrow direction signify depen-

dence and starts from an element to another that may influence it [17].

Figure 5 contains the entire inner dependence – the parent element and the elements to be compared are in the same cluster so that the cluster is linked to itself and a loop link appears – among elements within each cluster except in the environmental and alternative clusters. It indicates that the connections between the elements are in the same cluster. For example, column A8 means A2, A4, A5, A6 are interrelated with respect to A8.

The proposed model also contains outer dependence which is the relationship between elements in one cluster with others in other clusters [15]. For example, in Table 3, when considering A8, the elements G1, G2, G3 and G4 in the (G) *Alternative cluster* are interconnected and pairwise compared with respect to A8 in the (A) *Technical cluster*. The exception is the regulatory and environmental

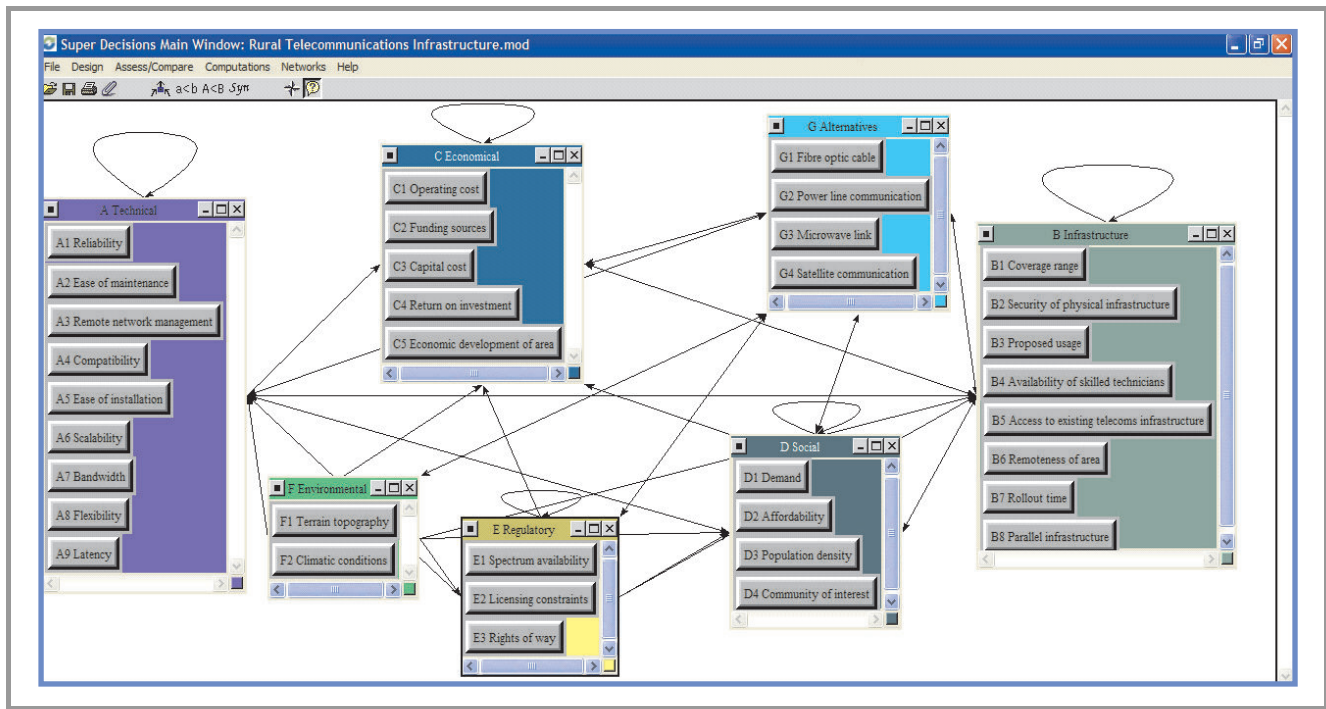


Fig. 5. The ANP network model with connections among elements/clusters.

clusters, i.e., none of the elements in both clusters depend on elements from other clusters with respect to a common attribute within a cluster. Also, the technical and social clusters have no outer dependence on the economic cluster.

Feedback links in which one compares the alternatives with respect to criteria, as in a hierarchy, and also compares the dominance of one criterion versus another for each alternative exist in this structure. Table 3 illustrates that there is mutual outer dependence of criteria in two different clusters as can be seen between the alternative cluster and all other clusters; technical and social clusters and infrastructure and economic clusters. For example, G1 is the parent element and all elements in other clusters except in cluster G are its children elements, which indicates that criteria may be compared with respect to an alternative. This is the strength of the ANP approach because dependence and feedback are incorporated in real life problems, in which a decision process not only compares alternatives with respect to criteria but also vice versa. For instance, in addition to separately comparing G1, G2, G3 and G4 with respect to A1 and A7, A1 and A7 must also be compared with respect to G1. A pairwise question to be asked is: what is a more dominant characteristic of fiber optic cable technology, its reliability or its bandwidth? However, since feedback involves cycles, and cycling can be an infinite process, the operations needed to derive the priorities become more demanding than with hierarchies [20].

Based on the above analysis, it is obvious that the developed inner and outer dependence and feedback among the network structure shown in Table 3 excludes the hierarchy form and calls for the network form to model

the selection of rural telecommunications infrastructure technology.

5. Pairwise Comparisons

After constructing the ANP network, the next phase is the measurement and data collection stage which involves compiling a list of experts to provide judgements for pairwise comparisons. Both the AHP/ANP derive ratio scale priorities by making paired comparison of elements on common elements. The subjective judgements are to be entered and assigned a numerical value based on the nine-point scale suggested by Saaty [21] to obtain the corresponding pairwise judgment matrices. A score of 1 indicates the equality between the two elements whereas score 9 represents the dominance of the row element in the matrix over the column element. A reciprocal value is automatically assigned in the opposite position in the matrix, i.e., $a_{ij} = 1/a_{ji}$.

In this model, pairwise comparisons are identified according to the connections developed in Table 3 and then relevant pairwise comparison matrices are created accordingly. The columns in the table present the parent elements, while the rows present the children elements in the structure. For example, G1 is a parent element and A1 through F2 are its children elements. The elements that are to be pairwise compared are always all in the same cluster. They are compared with respect to their parent element, the element from which they are connected.

There are a number of comparison matrices for every parent element, and one comparison matrix for elements in the same cluster originating from the same parent element.

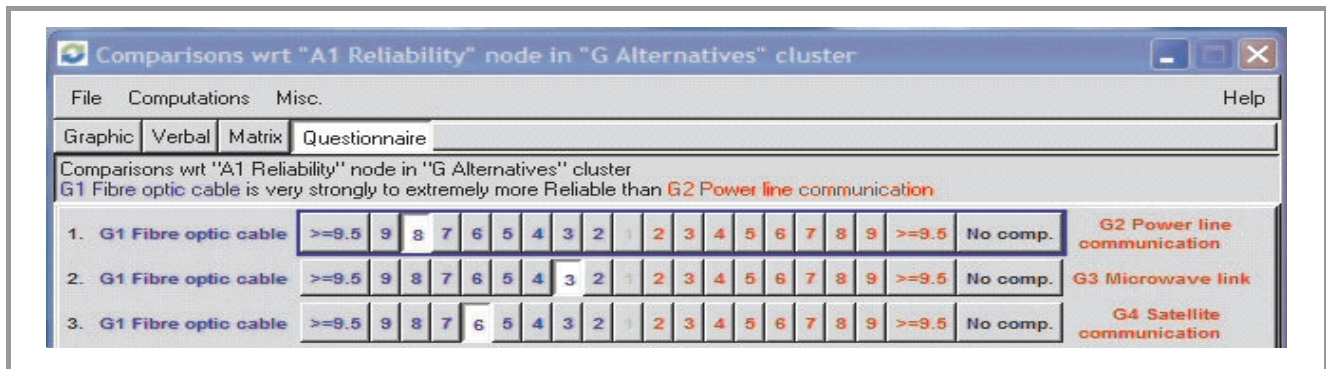


Fig. 6. An example of *SuperDecisions* pairwise comparison process.

For example, there are four comparison matrices for criterion B6, one for each of clusters A, B, C and G. Elements within D cluster cannot be compared with respect to B6 because there should be at least two entries of 1 available within any cluster to perform pairwise comparisons. Therefore, A1, A2, and A5 through A9 are pairwise compared with respect to B6; B1 through B5, B7 and B8 are pairwise compared with respect to B6; C1 and C5 are pairwise compared with respect to B6; and G1 through G4 are pairwise compared with respect to B6. This results in local priorities of the children elements with respect to the parent element. It is only necessary to make $n(n-1)/2$ comparisons to establish the full set of pairwise judgements, where n denotes the number of elements (nodes). For example, six pairwise comparison questions are required for A1 because $n = 4$ for the alternatives outer dependence on A1, while for A8, twelve pairwise comparison questions are needed because $n = 4$ for the inner dependence within the technical cluster, and also $n = 4$ for the alternatives outer dependence on A8.

In this developed structure; there are a total of 92 judgement matrices which include 674 pairwise comparison questions for both inner and outer dependences developed within the network. It is obvious that the task of asking such a large number of questions would be very enormous and would require intensive efforts and extended time. Hence, in order to establish a more rational approach to collect pairwise comparison judgements from qualified telecommunication experts, and also to economize efforts, it was decided to design and use several online questionnaires to gather data from experts. The questionnaires included all required pairwise questions to assess expert judgments in relation to the relative influence of affecting elements on the affected ones.

An example of such pairwise question is: "In selecting an appropriate backbone infrastructure technology in rural areas of developing countries, which influences fiber optic cable technology more, ease of installation or ease of maintenance? Conversely, given the ease of installation, which of these technologies are more dominant, fiber optic cable or satellite?"

Since the clusters in this network are not equally important, their weights in the cluster matrix are obtained by

pairwise comparisons. Each cluster is taken in turn as a parent cluster, and the other clusters connected to it are pairwise compared for importance with respect to their influence on it [17]. It should be noted that the pairwise comparisons to assess the influence of some cluster on all other clusters is actually what distinguishes the ANP from the AHP.

For example, one of the cluster comparison questions addressed to the experts is: "Which influences the selection of rural telecommunications backbone infrastructure more, economic or technical issues?" The obtained cluster weights are used in a later stage to weight all the elements in the unweighted supermatrix. The individual expert pairwise comparisons are aggregated into a representative group judgment, by applying geometric means.

A score corresponding to the group judgment regarding this question is then clicked to highlight the technology providing more reliability relative to the technology providing less reliability. While, a score of 1 indicates the equality between the two technologies, the blue scores represent the dominance of the row element in the matrix (e.g., G1) over the column element (e.g., G2) and the red scores are vice versa. A reciprocal value is automatically assigned in the opposite position in the matrix.

An example of the comparison process used in *SuperDecisions* is shown in Fig. 6. It presents the pairwise comparisons between alternatives G1 and G2, regarding the reliability factor. The question being asked is "With respect to reliability, which technology is more reliable: fiber optic technology or power line communication?" The group judgment was that G1 is between *very strongly* and *extremely* more reliable than G2, therefore the comparison value of 8 is entered.

The comparison between all other alternatives regarding different criteria is done in the same way.

The next stage of the process includes the computations of the relative importance of the elements. For each comparison matrix a local priority vector (also referred as an eigenvector) is computed, by applying the eigenvector approach [14], provided that the inconsistency ratio (IR) of this matrix is less than 0.1.

Table 4
Comparison matrix of alternatives with respect to reliability, $IR = 0.0958$

Alternative	G1	G2	G3	G4	Eigenvector
(G1) Fiber optic cable	1.000	8.240	3.350	5.730	0.580
(G2) Power line communication	0.121	1.000	0.178	0.217	0.044
(G3) Microwave link	0.299	5.630	1.000	3.830	0.262
(G4) Satellite communication	0.175	4.610	0.261	1.000	0.114

Table 5
Ordering and clustering of criteria according to relevance and importance

Cluster	Criteria	Priorities [%]	
		Normalized	Limiting
(A) Technical	(A1) Reliability	18.73	2.42
	(A2) Ease of maintenance	21.70	2.81
	(A3) Remote network management	17.05	2.21
	(A4) Compatibility	2.62	0.33
	(A5) Ease of installation	2.80	0.36
	(A6) Scalability	11.98	1.55
	(A7) Bandwidth	13.42	1.74
	(A8) Flexibility	3.50	0.45
	(A9) Latency	8.21	1.06
(B) Infrastructure	(B1) Coverage range	58.32	7.96
	(B2) Security of physical infrastructure	5.04	0.69
	(B3) Proposed usage	12.07	1.65
	(B4) Availability of skilled technicians	3.24	0.44
	(B5) Access to existing telecoms infrastructure	4.74	0.65
	(B6) Remoteness of area	3.22	0.44
	(B7) Rollout time	5.84	0.80
	(B8) Parallel infrastructure	7.53	1.03
(C) Economic	(C1) Operating cost	16.39	7.53
	(C2) Funding sources	34.70	15.94
	(C3) Capital cost	7.84	3.60
	(C4) Return on investment	37.32	17.15
	(C5) Economic development of area	3.75	1.72
(D) Social	(D1) Demand	64.96	1.74
	(D2) Affordability	20.49	0.55
	(D3) Population density	9.91	0.27
	(D4) Community of interest	4.65	0.12
(E) Regulatory	(E1) Spectrum availability	56.99	0.67
	(E2) Licensing constraints	27.45	0.32
	(E3) Rights of way	15.56	0.18
(F) Environmental	(F1) Terrain topography	56.81	0.71
	(F2) Climatic conditions	43.19	0.54

The *SuperDecisions* can also deal with the issue of improving consistency of the matrices, by identifying the most inconsistent judgments. The matrix consistency can then be improved by providing more consistent judgements by the decision makers so that $IR \leq 0.12$. For further explanation of inconsistency and how to calculate it, one can refer to [22].

An example of an aggregated comparison matrix within the alternative cluster (G) with respect to reliability (A1),

and the corresponding values of the eigenvector, is shown in Table 4.

From this table we can see that in terms of reliability, the fiber optic cable technology has the highest priority (0.580) followed by microwave links and satellite with (0.262) and (0.1244), respectively. The less reliable technology is the power line communication (0.044). Since IR is less than 0.12, this matrix is considered of acceptable consistency.

The eigenvector derived in this way is then entered as a part of some column of a supermatrix. It represents the impact of a given set of elements in a component on another element in the system, where a component in a supermatrix is the block, defined by a cluster name at the left and a cluster name at the top. If an element has no influence on another element, its influence priority is assigned zero [20]. The formation of a supermatrix in the ANP allows for the resolution of the effects of the interdependence that exists between the elements of the system.

The *SuperDecisions* performs necessary matrix operations for structuring of the three supermatrices, associated with this model, as shown in the Appendix. Table A1 illustrates the unweighted supermatrix that contains the local priorities derived from pairwise comparisons throughout the network; they can be read directly from this matrix. The weighted supermatrix shown in Table A2; is obtained by multiplying all the elements in a component of the unweighted supermatrix by the corresponding cluster weight, i.e., each block of column eigenvectors belonging to a component is weighted by the priority of influence of that component. This makes the entire columns sum to unity exactly, i.e., the weighted supermatrix is said to be “column stochastic”. Finally, the limit supermatrix is obtained by raising the weighted supermatrix to the power k , where k is an arbitrarily large number, to allow for convergence of the interdependent relationships.

The final values of priorities of all the elements are obtained by normalising each block, so that the columns of the limit supermatrix become identical. The values of the priorities of all elements can be read from any column [15] as can be seen in Table A3.

The *SuperDecisions* has also been used to produce the priorities shown in Table 5. It contains the relative importance of all criteria considered in the model. For example, under the limiting priorities’ column, one can observe that the most important factors among all are the *Return on investment* criterion with a priority of 17.15% followed by the *Funding sources* criterion with 15.94%. According to the *Normalized* priorities column, the most important criterion is the *Demand* with a priority of 64.96%, followed by the *Coverage range* with 58.32%. Among the technical criteria; the *Ease of maintenance*, *Reliability* and *Remote network management* criteria have the highest priorities of 21.71%, 18.73% and 17.05%, respectively. The *Spectrum availability* and *Terrain topography* factors are regarded as the most important within regulatory and environmental clusters, with priorities of 56.99% and 56.81%, respectively.

The relative importance of all other criteria considered in the model can be seen in Table 5.

6. Conclusions

This research paper reports on the applicability of using a MCDM method to enhance the selection process of an essential rural infrastructure technology. An ANP model incorporates both qualitative and quantitative approaches to a decision problem. The qualitative part includes:

- identification of the decision problem;
- ensuring the suitability of ANP to solve the problem;
- decomposing the unstructured problem to a set of manageable and measurable levels;
- compiling a list of experts to provide judgements for making the decision.

The quantitative part includes:

- designing a questionnaire to collect input data through pairwise comparison;
- estimating the relative importance between any two elements in each matrix and calculating the relevant eigenvectors;
- measuring the inconsistency of each matrix by employing the consistency ratio;
- eventually constructing the supermatrix using the eigenvectors of the individual matrices.

Based on the performed analysis, it is shown that the problem has inner and outer dependences and feedback among the elements, which excludes the hierarchy form (AHP) and requires a network form to model the selection process. The paper illustrates the use of the ANP method, but no real life conclusions should be drawn from it, as each telecommunication infrastructure provider will have its own set of criteria. The attempt here is to present a generic model based on factors and alternatives identified from the published literature, best practices and telecommunications experts that could then be adapted or extended to support a particular context or a situation of a developing country. Planners may therefore augment this model with their own company-specific factors that might change the priorities.

The obtained results reflect the preferences of experts who made the judgments, therefore, they cannot be considered as an objective assessment of the relative suitability of the four technologies as backbone infrastructure in rural areas. Final alternatives scores should, therefore, be thought of as an input to the decision-making process rather than its end. This process would be refined with experience, optimising the accuracy and time taken to reach proper decisions regarding the choice of telecommunication infrastructure in rural surroundings.

Table A2
The weighted supermatrix

	A								B								C					D					E					F				G			
	A1	A2	A3	A4	A5	A6	A7	A8	A9	B1	B2	B3	B4	B5	B6	B7	B8	C1	C2	C3	C4	C5	D1	D2	D3	D4	E1	E2	E3	F1	F2	G1	G2	G3	G4				
A1	0.000	0.403	0.000	0.000	0.000	0.000	0.428	0.000	0.000	0.000	0.160	0.000	0.023	0.000	0.020	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.019	0.082	0.023	0.008	0.025	0.003				
A2	0.000	0.000	0.869	0.495	0.219	0.000	0.000	0.362	0.000	0.000	0.020	0.000	0.125	0.000	0.025	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.194	0.066	0.051	0.025	0.007	0.016	0.005				
A3	0.000	0.000	0.000	0.375	0.000	0.000	0.197	0.000	0.000	0.141	0.000	0.000	0.000	0.534	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.017	0.007	0.014	0.008				
A4	0.000	0.000	0.000	0.000	0.184	0.000	0.000	0.079	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.007	0.012	0.038	0.004	0.008				
A5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.070	0.000	0.000	0.016	0.000	0.026	0.000	0.084	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.059	0.066	0.030	0.008	0.030	0.005	0.007			
A6	0.000	0.000	0.000	0.000	0.000	0.000	0.053	0.359	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.816	0.000	0.000	0.000	0.000	0.000	0.162	0.000	0.035	0.000	0.037	0.053	0.000	0.006	0.007	0.010	0.041	0.004	0.019					
A7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.055	0.000	0.178	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.000	0.000	0.136	0.000	0.219	0.000	0.194	0.107	0.000	0.000	0.000	0.062	0.005	0.021	0.018					
A8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.028	0.000	0.000	0.051	0.000	0.005	0.000	0.004	0.024	0.009	0.040					
A9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.015	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.020	0.042	0.000	0.000	0.000	0.003	0.003	0.066	0.054					
B1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.346	0.049	0.000	0.023	0.000	0.000	0.000	0.360	0.142	0.028	0.000	0.099	0.170	0.091	0.116	0.000	0.073	0.000	0.046	0.016	0.079	0.006				
B2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.015	0.000	0.000	0.000	0.000	0.000	0.000	0.022	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.016	0.012	0.026	0.045				
B3	0.152	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.683	0.000	0.000	0.000	0.018	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.020	0.000	0.240	0.046	0.033	0.012	0.054	0.000	0.004	0.000	0.018	0.034	0.009	0.015				
B4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.006	0.021	0.030					
B5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.018	0.000	0.113	0.000	0.000	0.000	0.000	0.044	0.000	0.000	0.000	0.000	0.000	0.000	0.137	0.000	0.000	0.005	0.020	0.031	0.007				
B6	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.015	0.037	0.000	0.000	0.000	0.015	0.030	0.042	0.021	0.018	0.004					
B7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.094	0.000	0.007	0.000	0.000	0.000	0.000	0.000	0.000	0.154	0.000	0.013	0.000	0.076	0.010	0.043	0.028	0.102	0.060	0.004	0.007	0.010					
B8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.008	0.000	0.000	0.000	0.000	0.000	0.000	0.131	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.085	0.005	0.080				
C1	0.625	0.536	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.674	0.000	0.520	0.000	0.200	0.000	0.000	0.000	0.000	0.000	0.000	0.137	0.000	0.460	0.463	0.225	0.486	0.262	0.022	0.162	0.036						
C2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.661	0.569	0.245	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.072	0.072	0.188	0.165				
C3	0.000	0.000	0.000	0.000	0.536	0.000	0.000	0.000	0.000	0.201	0.760	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.060	0.000	0.025	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.053	0.237	0.020	0.030				
C4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.559	0.000	0.691	0.000	0.000	0.000	0.000	0.616	0.888	0.000	0.227	0.000	0.324	0.408	0.614	0.028	0.614	0.000	0.000	0.000	0.049	0.000	0.059	0.067	0.048	0.132				
C5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.100	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.251	0.000	0.000	0.000	0.000	0.253	0.071	0.019	0.069	0.048	0.103					
D1	0.152	0.000	0.000	0.000	0.000	0.000	0.220	0.000	0.000	0.000	0.000	0.090	0.000	0.000	0.081	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.079	0.057	0.000	0.000	0.000	0.000	0.000	0.037	0.029	0.045	0.014					
D2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.079	0.065	0.000	0.000	0.000	0.021	0.028	0.011	0.038					
D3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.093	0.098	0.005	0.006	0.008	0.005				
D4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.005	0.003	0.010				
E1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000					
E2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000					
E3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000					
F1	0.000	0.000	0.000																																				

[illegible]

References

- [1] S. M. Nazem, H. Lee, and Y. Shi, "Designing rural area telecommunication networks via hub cities", *Omega*, vol. 22, pp. 305–314, 1994.
- [2] S. M. Nazem, Y. H. Liu, H. Lee, and Y. Shi, "Implementing telecommunications infrastructure: a rural America case", *Telemat. Inform.*, vol. 13, pp. 23–31, 1996.
- [3] J. W. Lee and S. H. Kim, "Using analytic network process and goal programming for interdependent information system project selection", *Comput. Oper. Res.*, vol. 27, pp. 367–382, 2000.
- [4] S. M. Nazem, Y. Shi, H. Lee, K. Sung Yeol, P. Tae Ho, and S. Myung Ho, "Multicriteria hub decision making for rural area telecommunication networks", *Eur. J. Oper. Res.*, vol. 133, no. 3, pp. 483–495, 2001.
- [5] L. Chemane, L. Ekenberg, O. Popov, and T. Cossa, "MCDM model for selecting internet access technologies – a case study in Mozambique", in *IEEE EUROCON 2005*, Belgrade, Serbia & Montenegro, 2005.
- [6] M. Sasidhar and K. J. Min, "Decision support models for the selection of internet access technologies in rural communities", *Telemat. Inform.*, vol. 22, no. 3, pp. 201–219, 2005.
- [7] T. Nepal, "Evaluation of rural telecommunications infrastructure in South Africa", *Probl. Nonlin. Anal. Eng. Syst.*, vol. 3, no. 24, pp. 138–149, 2005.
- [8] T. N. Andrew, P. Rahoo, and T. Nepal, "Enhancing the selection of communication technology for rural telecommunications: an analytic hierarchy process model", *Int. J. Comput. Syst. Sig.*, vol. 6, no. 2, 2005.
- [9] J. S. Shang, Y. Tjader, and Y. Ding, "A unified framework for multicriteria evaluation of transportation projects", *IEEE Trans. Eng. Manage.*, vol. 51, no. 3, 2004.
- [10] T. L. Saaty, *Theory and Applications of the analytic Network Process: Decision Making with Benefits, Opportunities, Costs, and Risks*. Pittsburgh: RWS Publications, 2005.
- [11] L. M. Meade and A. Presley, "R&D Project Selection using the analytic network process", *IEEE Trans. Engin. Manage.*, vol. 49, pp. 59–66, 2002.
- [12] M. Pipattanasomporn, "A study of remote area internet access with embedded power generation", Ph.D. thesis. Alexandria, University of Virginia, 2004.
- [13] T. N. Andrew and D. Petkov, "The need for a systems thinking approach to the planning of rural telecommunications infrastructure", *Telecommun. Policy*, vol. 27, no. 1–2, pp. 75–93, 2003.
- [14] T. Saaty, *Decision Making with Dependence and Feedback, the Analytic Network Process*. Pittsburgh: RWS Publications, 1996.
- [15] T. L. Saaty, *Decision Making with Dependence and Feedback, the Analytic Network Process*, Pittsburgh: RWS Publications, 2001.
- [16] H. Lee, C. Kim, H. Cho, and Y. Park, "An ANP-based technology network for identification of core technologies: a case of telecommunication technologies", *Expert Syst. Appl.* (in press).
- [17] R. W. Saaty, *Decision making in complex environments: a tutorial for the SuperDecisions software*. Creative Decisions Foundation, 2003.
- [18] C. Douligeris and J. Pereira, "Telecommunications quality study using the analytic hierarchy process", *IEEE J. Selec. Areas Commun.*, vol. 12, no. 2, pp. 241–250 1994.
- [19] Y. Kawasumi, Y. Miyoshi, and K. Koizumi, "How to provide broadband access for rural and remote areas of developing countries", in *Proc. PTC'07: Telecom with Vision*, Honolulu, Hawaii, 2007.
- [20] T. L. Saaty, *Decision Making with the Analytic Network Process: Economic, Political, Social and Technological Applications with Benefits, Opportunities, Costs and Risks*. New York: Springer, 2006.
- [21] T. L. Saaty, *The Analytical Hierarchy Process: Planning, Priority Setting, Resource Allocation*. New York: McGraw-Hill, 1980.
- [22] T. L. Saaty, *Multicriteria Decision Making: The Analytic Hierarchy Process*. Pittsburgh: University of Pittsburgh Press, 1988.

Note: this text is reprinted from JTIT 2/2010



Yousef Gasiea received the B.Eng. degree in communications engineering from the Higher Institute of Electronics, Ben Walid, Libya, in 1988, and the M.Sc. degree in engineering management from Al-Fateh University, Tripoli, Libya, in 1999. He has worked as an electronic and telecommunications engineer, supervising a field study to measure the signal strength received from various geo-stationary satellites, and installation and maintenance of PABX, TVRO and SMATV systems. Currently he works towards his Ph.D. in engineering management at the University of Manchester, UK. His current research interests include planning of rural telecommunications infrastructure, multicriteria decision methods, decision support systems and analytical decision processes. yousef.gasiea@postgrad.manchester.ac.uk
School of Mechanical, Aerospace and Civil Engineering
University of Manchester
M60 1QD Manchester, United Kingdom



Margaret Emsley is civil and structural engineering graduate whose early career involved working on construction sites in the UK and a 12 month design office secondment. She was then a Research Assistant at Loughborough University, working on productivity measurement, before commencing her Ph.D. which was concerned with modelling the location of tower cranes. For the past 24 years she has worked at the University of Manchester where she is a Senior Lecturer in the School of Mechanical, Aerospace and Civil Engineering. Her research interests are in the field of construction and project management and include the application of neural networks (where she has carried out funded research into their use in the prediction of construction cost), game theory and simulation modelling. margaret.emsley@manchester.ac.uk
School of Mechanical, Aerospace and Civil Engineering
University of Manchester
M60 1QD Manchester, United Kingdom



Ludmil Mikhailov received the M.Sc. degree in automatic control from the Technical University in Sofia, Bulgaria, and the Ph.D. degree in technical cybernetics from the Bulgarian Higher Certifying Commission, in 1976 and 1982, respectively. He was an Associated Professor in the Institute of Control and System Research at the Bulgar-

ian Academy of Sciences. Currently he is a Senior lecturer at the Manchester Business School, the University of Manchester. He is the author of about 90 technical papers in peer-reviewed journals and international conferences. His current research interests include multiple criteria decision analysis, fuzzy logic systems, decision-making under uncertainty, and intelligent decision support systems.

e-mail: ludi.mikhailov@manchester.ac.uk

Manchester Business School

University of Manchester

M15 6PB Manchester, United Kingdom

Telecommunications, Universal Service and Poverty in Mexico: a Public Policy Assessment (1990–2008)

Cristina Casanueva-Reguart and Antonio Pita S.

Abstract—This article analyzes the design and implementation of telecommunications service policies targeted at the poorest regions of Mexico (1990–2008). It begins by defining universal access and service policies, their economic and social rationale. Secondly, it discusses the scope of public policies on universal service provision designed by Mexican authorities to achieve the goal of universal access. Thirdly, the paper analyzes the distributive effects of this set of policies among the poorest sectors of the population. The sources on which this research was based were two national surveys: the *Household Income and Expenditure Survey* (2008), and the *Household Survey of the Access and Use of Information Technologies* (2007). The additional information on regional economic development was based on the poverty indexes by the national population council and economic information given by Mexico's Census Bureau. Additional use was made of the annual reports prepared by Ministry of Communications, statistics published by the Federal Telecommunications Commission and official documents prepared by the government agencies. Finally, a series of in-depth interviews was conducted with the former representatives of the Office of Rural Telephony. Finally, the article discusses, in the light of available evidence, possible explanations for the apparent failure of the universal service policy that was implemented to bring at least basic voice services to Mexico's neediest.

Keywords—development, digital divide, market power, regulatory capture, social inclusion, universal service.

1. Introduction

The telecommunications services infrastructure is an important factor for economic development and social inclusion, and a crucial component leading to greater equality when services are available to and affordable to any person, irrespective of levels of income and geographic location.

In developing countries gaps remained in the market mainly because of regulatory failure, combined with exceptionally challenging geography and extremely low population densities, isolation and extreme poverty [1]. Government, institutions and public policy design have been faced with a situation where authorities had a social obligation to ensure that their people had access to basic telecommunications, but the ability to enforce these obligations entailed increasing difficulties. Some of these difficulties are related to the asymmetry between the government regulatory bodies and the power of the operators, particularly in the case

of an incumbent operator with market power in the supply of almost every telecommunications service¹.

This article analyzes the design and implementation of telecommunications service policies targeted at the poorest regions of Mexico (1990–2008). It begins by defining universal access and service policies, as well as their economic and social rationale. The article then discusses the scope of public policies on universal service provision designed by Mexican authorities to achieve the goal of universal access, namely:

- Monitoring by the regulatory authorities of compliance on the part of the incumbent telecommunications operator, Teléfonos de México (Telmex), in fulfilling its social obligations as the dominant operator, following the company's privatisation (1990);
- Government policy aimed at providing connectivity – basic telephony services – to rural communities. This article presents the main research results, which focus on basic (voice) telecommunications services, both land-line and wireless². This policy was deployed by private operators, mainly the incumbent. The participation of operators was based on a public tender processes organized by the government, which has resulted in an additional profitable business for the incumbent operator, with limited results in terms of access to these services by the poor.

Thirdly, the paper analyzes the distributive effects of this set of policies among the poorest sectors of the population. Finally, the article discusses, in the light of available evidence, possible explanations for the apparent failure of the universal service or social coverage policies that were implemented to bring at least basic voice services to Mexico's neediest, as well as the difficulties faced by the regulatory bodies behind the design and implementation of these policies.

¹Almost 20 years after Telmex's privatization (1990) in 2009, Telmex's operated the 84.8% of the total the number of fixed line connections, and 72.3% of cellular lines. "The Competitive Intelligence Unit" (CIU), <http://octavioislas.wordpress.com/2009/07/23/3236-mexico-the-competitive-intelligence-unit-competencia-en-mexico-%C2%BFque-20-anos-no-es-nada-telecomunicaciones/>

²The results and the discussion, presented in this article, are part of a larger project on Universal Access and Service which include telecommunications services and bandwidth in the provision of Internet, data and video transmission.

The sources on which this research was based were two national surveys: the *Household Income and Expenditure Survey* (2008), and the *Household Survey of the Access and Use of Information Technologies* (2007). The additional information on regional economic development reported in this article was based on the poverty indexes drawn up by the national population council and economic information given by Mexico's Census Bureau National Institute for Statistics, Geography and Information (INEGI), Bank of Economic Information (BIE), and housing and population counting, 2005. Additional use was made of the annual reports prepared by the Communications and Transport Secretary (Ministry of Communications), statistics published by the Federal Telecommunications Commission (COFETEL) and official documents prepared by the government agencies in charge of designing and monitoring the universal service telecommunications policies, principally the Ministry of Communications. Finally, a series of in-depth interviews was conducted with the former representatives of the Office of Rural Telephony, who were in charge of monitoring the implementation of social and universal telecommunications policies.

2. Access to Telecommunications Services as a Fundamental Right; Definitions of Universal Access and Universal Service: its Economic and Social Rationale

The recent literature on the universality of telecommunications services states, as a fundamental principle, that every citizen has the right to access telecommunications services with high quality standards. This implies fulfilling three basic conditions: universality, equality and continuity, defined as follows:

- universality: every citizen has the right to have access to telecommunications services with a high standard of quality;
- equality: universal access, irrespective of income levels and geographic location;
- continuity: ensuring that the service continues to maintain high quality standards, in an uninterrupted manner.

Regarding universality, the literature on this subject distinguishes between universal access and universal service:

- Universal access is when everyone has the right to use the service somewhere, in a public place. This service is also defined as communal or shared access. In general there would be at least one point of access per settlement over a certain population size, within a convenient and reasonable distance.

- Universal service is when every individual, household, business or institution can be provided with a service, using it privately, either at home or in an increasingly mobile variant, carried with the individual through wireless devices [2].

The rationale for a universal access or universal service policy is both economic and social. The economic rationale is based on the market's inability to provide infrastructure, connectivity and services on a universal basis. Thus, universal access and service policies are justified in the face of market gaps, so as to guarantee equality of economic opportunities, since telecommunications services are a critical component in the production of goods and services, as well as for social inclusion and, increasingly so, for political participation.

The social rationale consists of the will of the public policy makers who, as representatives of the state and its citizens, must guarantee social inclusion and avoid the exclusion of parts of the population, irrespective of income levels and geographic location.

Furthermore, the three defining characteristics of universal access and universal service are:

- availability: the service is accessible to inhabited parts of the country through public, community, shared or personal devices;
- accessibility: all citizens can use the service, regardless of location, gender, disabilities and other personal characteristics;
- affordability: the service is affordable to all citizens.

Thus, the policy of a universal telecommunications service consists of an explicit, direct and focused public policy, aimed at offering telecommunications services at prices that are affordable to the poorest sector of the population. The fulfilment of this policy requires a subsidy, since this sector of the population cannot afford these services at market prices. The subsidy may be applied on the supply or demand side of these services. On the supply side, it is usually applied through development and/or optimisation of the infrastructure (investment), which allows connectivity and thus takes into account the difficulty of infrastructure provision given the level of geographical dispersion of these communities, as well as the cost of providing the service.

On the demand side, the affordability of the provision of these services must be addressed, because the population currently without service generally lives in conditions of poverty, sometimes extreme, and thus their income does not allow them access to these services at the market price. Hence, when supplying these services is not profitable for the service operators, provision depends on various forms of subsidy [3].

3. Public Policies on Universal Service Provision in Mexico under the Responsibility of the Telecommunications Regulatory Bodies

According to the population counting of 2005, there were 184,748 rural communities in Mexico with a population of fewer than 2,500 inhabitants, and 197,479 communities of fewer than 5,000. These communities are inhabited by over 30 million men and women, which represented 29.1% of the Mexican population. In addition, a notable feature of these communities was their high level of dispersion, with 92.5% having fewer than 500 inhabitants.

Thus, in order to increase the social coverage of telecommunications services, a set of policies was drawn up and implemented. They represented the main public policies that have been put into practice to provide access and universal service in Mexico and among them, the most important were:

- Monitoring of Teléfonos de México's (Telmex, the incumbent operator) compliance with its obligations to provide a universal service, as set out in Telmex's licence of 1990, at the time that the public telecommunications operator was privatized. This licence was granted by the Ministry of Communications and Transport (Ministry of Communications³), which has acted as a regulator and has the authority to monitor Telmex's compliance with its obligations regarding universal service, rural telephony, as well as the modernization and expansion of the public network, contained in Telmex's licence.
- Rural telephony (1995–2007), aimed at offering services to communities with fewer than 500 inhabitants, with a direct subsidy from the Ministry of Communications.
- The creation of the *Social Coverage Fund* (Fondo de Cobertura Social, FONCOS: 2002–2007) by the Ministry of Communications aimed at increasing the coverage of rural telephony in communities with a population of between 400 and 2,499 inhabitants. For this purpose the regulator organized went out to public tender, calling for bids from telecommunications operators. The conditions of the tender included

³In Mexico there are two main regulatory agencies directly involved with telecommunications, the Subsecretaría de Comunicaciones (Ministry of Communications), which is part of the Secretary of Communications and Transports (Ministry of Communications and Transport). The second regulatory agency is the Comisión Federal de Telecomunicaciones (Federal Telecommunications Commission), which is an autonomous government agency. In the case of universal service and universal access, the agency in charge of designing these policies is mainly the Ministry of Communications. This ministry has performed the major role in the design of universal telecommunications policy and in the surveillance of the implementation universal services by the operators.

the provision of non-returnable monetary resources, which were originally allocated by the Ministry of Finance. In addition, the regulator allocated frequency bandwidth resources that were reserved for the purpose of social and universal service coverage, with a ten year licence (renewable) to use these frequency bands.

3.1. *Telmex Licence: Universal Service Obligations and Network Growth*

The design of the universal service obligations included in the incumbent operator's licence should have addressed the main challenges posed by the gaps in the market, mainly in small rural communities generally located in remote and isolated areas and where the poorest people of the country live. This section deals with the context in which Telmex was privatized, which explains the scope of the clauses relating to universal service and rural telephony included in the licence. The scope of both the content and schedule for its implementation, as well as their impact on bringing connectivity to the neediest communities, was shaped to some extent by the context in which the privatization of Telmex took place. This section begins by describing this context, then briefly presents the content of the clauses relating to universal service obligations, namely, regarding network expansion, rural telephony and public telephones. It then presents the main results of an analysis of the impact of Telmex's fulfilment of these licence clauses.

In order to understand the government policy makers' limited leverage on the definition of the clauses relating to universal obligations or social coverage in Telmex's franchise agreement(1990), it is useful to analyze the specific juncture at which this process took place. During the 1980s, with the economy severely indebted, the burden of foreign debt and fiscal deficit had a major effect on the process of privatization of public companies and specifically on the approach adopted in the privatization of Telmex. Between 1965 and 1980 the economy had been growing at an average rate of 6.7%, but during the 80s this slowed to a yearly average of 1.8%. In this context, government policy makers embarked on an aggressive privatization program of public companies, with two purposes in mind: to increase the efficiency of Mexico's economy and to improve Mexico's public finances.

In the case of the public telephone company in Mexico, the potential revenue gain from this privatization and the public finance argument prevailed. The decision to privatize the public telecommunications operator in Mexico took place in 1989. During the privatization of Telmex, emphasis was placed on expected revenues. In fact, the privatization process was chaired by the Ministry of Finance and not by the Ministry of Communications and Transport [4]–[6].

To maximize revenues from privatization, the government sold to a single set of investors a package that included Telmex and Telnor, in addition to the only nationwide mobile network franchise, as mentioned above, the Federal

Microwave Network, as well as an ample bandwidth allocation. Thus, overnight, the emerging company became a formidable player in the sector. It was allowed to offer all types of telecommunications services with the exception of television broadcasting services. Thus, Telmex became a horizontally and vertically integrated telecommunications service provider with a nationwide network for all its services. Furthermore, it was guaranteed little or no competition in key services for several years. The new private owners of Telmex were given a *de jure* monopoly over the long distance markets (national and international) for six years. They inherited the monopoly over local telephone services. By creating a horizontally and vertically integrated telecommunications company, the government could receive a higher price for privatizing the firm and reach a short-term public finance goal. For potential buyers, the company was very attractive [4], [5]. They were allowed to buy a stream of excess profits sustained by a monopoly, more valuable than a stream of revenue generated under competitive conditions [7].

In the negotiation process that accompanied the privatization of Telmex, government policy makers let the collection of revenues from privatization preside over other goals such as economic efficiency, well-being and social inclusion. This context explains the lack of leverage or bargaining power of the authorities, specifically on the subject of social coverage. Although Telmex's license included clauses governing universal service obligations, rural telephony and public telephone booths, as well as network expansion, Telmex's commitment to these clauses ended in 1994 and the results lagged very much behind the objective of providing a basic universal telecommunications service to rural areas.

The following paragraphs give a summary of the main clauses relating to Telmex's universal service and network growth obligations. The result of an empirical analysis is also presented, showing the outcomes of the implementation of these policies, reflecting the achievements of network growth, rural telephony and public telephone services or telephone booths (1990–1998).

From a regulatory perspective, Telmex's licence made this company operate as a regulated monopoly. The company was given a set of operational goals that it was required to meet:

- to expand the number of basic telephone lines by a minimum of 12% per annum, until the end of 1994;
- to continually reduce the waiting period for the basic telephone service in localities with automatic switching capabilities to a maximum of one month by the year 2000;
- to provide a public payphone service to every locality with more than 500 inhabitants by the end of 1994, and to increase the penetration of public telephone booths from 0.5 per thousand inhabitants, to five per thousand by the end of 1998;

- to provide a public payphone service to every locality with more than 2,500 inhabitants (less than 5,000 inhabitants, according to the definition of rural community), if there were at least 100 applications from potential end users, and an up front payment equivalent to three months' line rental. *"After these conditions were fulfilled (...) Telmex would deliver the service within a time frame of not more than 18 months"*⁴;
- to publish a four year (network) expansion and modernization programme, in accordance with the goals set by the licence, and to agree with the Ministry of Communications on programmes for rural telephony and public telephone booths.

3.1.1. To Expand the Number of Basic Telephone Lines by a Minimum of 12% per Annum

The 12% telephone line expansion requirement ended only four years after the 1994 privatization of the sector. Quantitative goals for a longer period would have conflicted with the goal of revenue maximization at the time of privatization. The results shown in Table 1 tend to support the view that the quantitative line expansion requirement set out in the concession or Telmex licence was an effective regulation. In fact, during the years 1991 to 1994, Telmex's average annual line expansion was 11.8%, and thus close to meeting the 12% requirement. However, once such an expansion requirement ceased to exist, average line expansion fell to only 6.8% in the period 1994–2000 (see Table 1).

Table 1
Five year average growth in telephone lines and GDP (1965–2000)

Period	Telephone lines [%]	GDP [%]
1965–1970	12.8	6.9
1970–1975	12.9	6.5
1975–1980	10.5	6.7
1980–1985	6.4	1.9
1985–1990	7.6	1.7
1990–1994	11.8	3.6
1994–2000	6.8	3.5
Source: SCT, Anuarios Estadísticos (1965–2000).		

3.1.2. Rural Telephony: Basic Telephone Service to Communities with More than 500 Inhabitants (1990–1998)

As a result of the negotiations between government policy makers in the field of telecommunications and the group of investors, Telmex's licence freed them from their obligation to serve communities with fewer than 500 in-

⁴SCT, Modificación a Teléfonos de México, 1990, http://www.cft.gob.mx/work/sites/Cofetel_2008/resources/LocalContent/3964/1/10ago90.pdf

Table 2
Telephone service to communities with more than 500 inhabitants

State	Lines per 1,000 inhabitants towns 500 to 2,499	Lines per 1,000 inhabitants towns 500–4,999	Rural communities communicated by Telmex, 1990–1994	Total population towns 500–2,499	Total population towns 500 to 4,999	GDP per capita
Chiapas	1.00	0.44	950	951,521	2,136,825	37.8
Oaxaca	1.24	0.65	1,362	1,095,547	2,102,278	39.8
Tabasco	1.31	0.78	678	517,227	863,855	47.6
Guerrero	1.27	0.63	932	732,388	1,470,855	47.8
Tlaxcala	0.77	0.37	109	141,396	294,861	47.9
National	1.35	0.65	16,738	13,339,307	27,937,529	83.2
Campeche	1.43	0.70	135	94,653	193,781	121.7
Quintana Roo	1.54	0.74	122	79,123	164,691	126.4
Coahuila	1.65	0.70	238	144,448	337,934	129.8
Nuevo León	3.30	0.79	232	70,211	293,812	173.5
Distrito Federal	0.00	0.00	0	13,268	26,550	188.0

Source: SCT, Annual Reports, INEGI, Census 1990.

habitants, which according to the census of 1990 represented 21.16 million people or 47.2% of the inhabitants in rural communities in Mexico [8].

An analysis of the impact of Telmex's rural telephony operations points to very limited results. The impact on telephone density, following Telmex's compliance with requirements on basic telephone service provision in rural towns, was extremely low, even when telephone density was estimated as the number of lines per thousand inhabitants⁵. This estimate shows that the country's average telephone density was 1.35 lines per 1,000 inhabitants in rural communities (500 to 2,499 inhabitants), and if the definition of rural communities included "enlarged rural communities" (from 500 to 4,999 inhabitants), the telephone density estimate drops to half the previous figure, or 0.65 lines per thousand inhabitants. Based on this analysis it is possible to assert that fulfilment of overall requirements, provision of rural telephony and the installation of public telephone booths in rural areas, had very much fallen behind in relation to the goals set by Telmex's licence (see Table 2, which presents the telephone density, according to our definition, in the five more prosperous states and in the five poorest states in Mexico).

Thus, in spite of the fact that public telephone booths were the strategy mostly used by Telmex to fulfil its universal or social obligations, compliance with the commitment of providing public access through public telephone booths was insufficient. At the end of 1998, Telmex admitted that it had only installed 3.19 public booths per 1,000 inhabitants. Taking into account that Mexico's population at the time was 96 million, Telmex would have had to install at least 480 thousand public booths in order to comply with the social obligation dictated by its licence. Unfortunately,

⁵See Caslon analytics, metrics and statistics, <http://www.caslon.com.au/metricsguide8.htm>

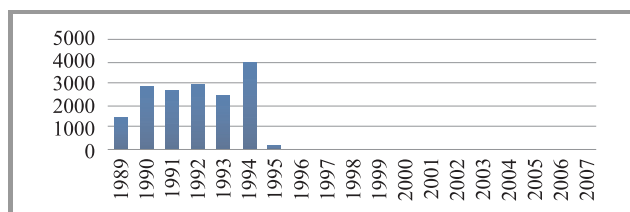


Fig. 1. Number of rural communities with more than 500 inhabitants, communicated by Telmex with at least one telephone line (1990–2007).

according to the definition of universal service set out in Telmex's licence and because of the government's lack of leverage at the time of privatization, the universal service obligation ceased and areas that were served with at least one public booth increased slightly in 1995 and 1996 and ceased to grow indefinitely from 1997 (see Figure 1 and Table 3).

Table 3
Telephone lines provided by Telmex in communities with a population of 500 or more inhabitants

Acumulated	Annual [number]	Annual [%]
4,350	2,854	190.8
16,542	4,006	32.0
Average annual growth: 1990–1994		93.4253
16,735	193	1.2
16,738	0	0.0
16,738	0	0.0
16,738	0	0.0
Average annual growth: 1994–2007		0.0014

Sources: SCT, Annual Reports (2000 and 2007).

Table 4
Teledensity and level of development in different states in Mexico, 2008

State	Lines per households [%]	Non residential lines ¹ [%]	Mobile lines ²	GDP per capita	Poverty index
National	52.2	12.7	71	96.8	
Chiapas	19.1	4.2	41.1	37.8	Very high
Oaxaca	26.4	4.6	39.8	39.8	Very high
Tabasco	28.1	6.7	75	47.6	High
Guerrero	40.6	7	44.8	47.8	Very high
Tlaxcala	40.3	4.4	44.1	47.9	Medium
Campeche	32.7	6.4	71	121.7	High
Quintana Roo	38.0	16.1	95.9	126.4	Low
Coahuila	63.3	12	86.6	129.8	Low
Nuevo León	79.9	22.2	91.8	173.5	Very low
Distrito Federal	104.6	40.5	102.7	188	Very low
¹ lines per one thousand employed people, ² lines per one hundred people. Sources: Cofetel, 2008; INEGI, 2005 and 2008, National Population Council, 2005.					

According to Telmex's license, the commitment to provide a basic service under the premise of universal access through public booths in Mexico is far outweighed by the challenge of providing services to the poorest communities

Table 5
Average revenue per capita in selected developing countries, 2007

Latin America	Country	ARPU	GDP per capita	ARPU/GDP per capita ¹
	Colombia	131.0	7,400	1.8
	Mexico	178.0	12,400	1.4
	Brazil	17.2	9,500	0.2
	Chile	17.0	14,300	0.1
	Argentina	12.2	13,100	0.1
Emerging Europe	Hungary	26.5	19,300	0.1
	Czech R.	28.5	24,500	0.1
	Poland	17.0	16,200	0.1
	Ukraine	7.0	7,000	0.1
	Russia	10.0	14,800	0.1
Africa/Middle East	Turkey	13.7	12,000.00	0.1
	Iraq	12.3	3,700.00	0.3
	South Africa	20.7	9,700.00	0.2
	Egypt	9.7	5,000.00	0.2
Emerging Asia	India	8.6	2,600	0.3
	China	10.8	5,400	0.2
	Korea	45.0	25,000	0.2
	Taiwan	23.0	30,100	0.1
	Singapore	348.0	49,900	0.7
	Hong Kong	22.5	42,000	0.1
¹ ARPU – average revenue per user: the estimation gives ARPU as a percentage of GDP per capita, in each country. Sources: Merrill Lynch, 2008, CIA World Factbook, 2008, and Office of Rural Telephony, 2009.				

of Mexico. This is true more particularly in the context of a large deficit of telecommunications services in the country as a whole: on average, there is provision to only 5 households out of every ten (52.2%) and 12.7 non residential lines per one thousand employed people (see Table 4).

Most likely the deficit of connectivity has been compensated for by the use of cellular or mobile services, at market prices. A comparative analysis of revenue per minute from wireless services suggests that Mexico has one of the highest tariffs for mobile services (see Table 5).

The following section presents information relating to the “agreements” between Telmex and the Ministry of Communications, which according to its licence, were supposed to continue after 1994.

3.1.3. Telmex Agreements with the Ministry of Communications on Programmes for Network Expansion, Rural Telephony and Public Telephone Booths (1995 to date)

In 1995, according to Telmex's licence, the company was required every four years to establish a programme for network expansion and provision of rural telephony and public telephone booths. However, in spite of the fact that an exchange of documents took place between Telmex and the Ministry of Communications in 1995, this exchange did not materialize into an action programme. It was not until 1998 that this took place, when Telmex set up a rural telephony programme, supposedly in an agreement with the Ministry of Communications, retroactive to 1995 (1995–1998).

In December 1998 Telmex sent a report on the fulfilment of the goals of this programme. The main goal achieved was the additional coverage, compared to 1994, of 4,288 communities, through public telephone lines or public booths.

However, strictly speaking, the service provided to these additional communities was part of a rural telephony project undertaken by the Ministry of Communications. Telmex was involved as the winner of a public tender in which they were chosen as the provider, but the project was financed by the Ministry of Communications (the results are presented below of the universal service policy undertaken by the Ministry of Communications).

It was only in July 2006, when Telmex delivered its results on network expansion and rural telephony for the periods 1995–1998, 1999–2002 and 2003–2006. The documents handed over by Telmex to the Ministry of Communications were accompanied by an appeal for confidentiality and the Ministry of Communications accepted the condition of confidentiality requested by Telmex. Nevertheless, the annual reports published by the Ministry of Communications (2001–2006) showed that since 1994, Telmex made a very small contribution in this area.

The privatization of Telmex and the attempts to regulate the monopoly through this company's license did not bring a strengthening of the regulatory authority's monitoring and enforcement capabilities. This had a severe effect on achieving the goals set by the universal service or universal obligations defined in Telmex's licence following privatization. Studies on telecommunications reform suggest that privatization by itself, without a strong regulator, does not yield significant performance improvements in the telecommunications sector [5], [7], [9].

4. Rural Telecommunications Services for Communities of Fewer than 500 Inhabitants: Direct Government Subsidy

This section analyzes the policy directly implemented by the Ministry of Communications, aimed at providing telecommunications services to rural communities of between 100 and 499 inhabitants. These programmes originally focused on small towns and villages with fewer than 500 inhabitants (1990–2002), but later on, with the establishment of the *Social Coverage Fund* (FONCOS), the focus of these programmes shifted to communities of between 400 and 2,500 inhabitants.

Table 6
Rural telephony supplier

Cellular company	Lines	%
Telcel	8,358	25.1
Iusacell	11,012	33.1
Telecomm	13,772	41.4
Others	100	0.3
Total	33,242	100.0
Source: Office of Rural Telephony.		

The programme was directly financed by the Ministry of Communication (1995–2006) and was aimed at communities with fewer than 500 inhabitants. It involved the main providers of mobile telecommunications services, including Telcel, the mobile company part of the same group as Telmex (25%), and the public satellite company, Telecomm (41.4%, see Table 6).

The size of these targeted communities confirmed that the policy successfully focussed on the poorest towns of Mexico and even in cases where regions of higher income were served, the subsidy focussed on the neediest rural communities that are generally located in remote and isolated areas (see Table 3). Nevertheless, telephone density in these small towns remained extremely low, as was the case of the services provided by Telmex in larger communities, where the estimated telephone density is 0.45 lines per 100 inhabitants when population data for these communities is taken from the 2000 census and 0.44 with population data is from 2005 (counting of population and housing, 2005, see Table 7). It is also likely that the actual telephone density in these small towns was higher because of the use of mobile technology, when it was available in these regions.

An additional source of information consisted of in-depth interviews with the former representatives of the Office of Rural Telephony, part of the aforementioned Ministry of Communications, where data compiled by this office is based on their fieldwork aimed at verifying the correct operation of the installed lines. This information complements that of the previous analysis⁶.

The information provided by the Office of Rural Telephony showed that 33,242 lines were installed between 1995 and 2006. An analysis of this information also showed a very rapid growth in the number of installed lines between 1995 and 2000 (135.42% yearly average growth) and that the pace of growth declined considerably over the following years, where the yearly average growth observed between 2001 and 2005 was only 1.16%. There was no evidence of growth between 2006 and 2009.

The results of the former analysis are even more dramatic considering the outcome of the fieldwork aimed at verifying the correction operation of the equipment. Here the data showed that only 41.5% of the installed lines were in operation and out of these, 58.5% of the lines were out of service and abandoned. The interviews with the former representatives of this office suggested that the people of these communities were gradually shifting to mobile technology, where this service was available, in spite of the higher costs involved in the use of mobile communication.

The former observations raise doubts over the commitment made by the government for bridging the connectivity gap in the smallest and poorest communities of Mexico. It also

⁶Is worth mentioning that there is an inconsistency between the information accounted by the Annual Report (2007) and the Office of Rural Communication, the first source reported 34,676 installed lines, and the Office of Rural Telephony accounted for 33,242, the difference between the two sources is 1,434 installed lines.

Table 7

Rural telephony, lines installed by Secretary of Communications, towns between 100 and 499 inhabitants, 1995–2007

	Communities	Lines 2007	Teledensity /100, 2000	Teledensity /100, 2005	GDP per capita
Total	184,748	34,676	0.45	0.44	70.88
Chiapas	19,237	3,560	0.48	0.42	28.6
Oaxaca	10,025	2,540	0.41	0.37	32.5
Tlaxcala	1,138	117	0.41	0.41	37.3
Michoacán	8,965	1,861	0.45	0.45	39.9
Chihuahua	12,095	896	0.43	0.53	102.9
Quintana Roo	1,800	177	0.46	0.44	107.5
Campeche	2,595	240	0.49	0.35	121.7
Baja California	3,918	248	0.56	0.56	93.0
Campeche	2,595	240	0.49	0.35	121.7
Nuevo León	5,169	561	0.50	0.57	133.1
Source: INEGI, Censo de Población y Vivienda, 2000; Conteo Población y Vivienda, 2005; Ministry of Communications, <i>Annual Report</i> (2000–2007) and Bank of Economic Information (BIE), 2007.					

raises questions over the nature of the agreements signed by the Ministry of Communications and the operators undertaking the installation of the telephone lines, specifically operation and maintenance in accordance with acceptable quality standards. The personnel interviewed agreed

Table 8

Number of new telephones installed in communities with less than 500 inhabitants 1995–2008

1995	4,000
1996	9,369
1997	10,545
1998	20,208
1999	23,063
2000	31,083
Average annual growth 1995–2000	135.42
2001	31,083
2002	31,453
2003	31,820
2004	32,326
2005	32,841
2006	33,240
2007	33,242
Average annual growth 2001–2008	1.16
Source: Ministry of Communications, Office of Rural Telephony.	

that the contracts included maintenance and quality clauses, which pose additional questions on the strength of the ministry as a regulator capable of enforcing these clauses.

5. Social Coverage Fund (FONCOS)

In 2002 the *Social Coverage Fund* (FONCOS) was established as a trust fund with an allocation of 75 million

US dollars⁷ provided by the Ministry of Finance to the Ministry of Communications. Its main purpose was for the funding of social telecommunications services, focussed on serving communities of between 400 to 2,500 inhabitants. The Ministry of Communications designed two different public tender processes. The first, *Basic Telephony 1* (STB-1 to use its Spanish acronym), was aimed at installing public telephone lines in communities of extreme poverty. The second, *Basic Telephony 2* or STB-2, focussed on communities with higher levels of income. For STB-1, the subsidy for the chosen operator consisted of both financial and bandwidth resources for 10 years (renewable), which were reserved by the government for social coverage purposes. The subsidy to the end user included all expenses relating to the installation and rental of the equipment, so that the end user had only to pay for call traffic, charged for via prepaid cards.

For STB-2, the subsidy to the winning operator consisted of bandwidth resources only. The end user was charged for installation costs and call traffic, exonerating them from payment for the rented equipment. In this case, the subsidy for the chosen company consisted only of the licence to operate bandwidth resources for 10 years (also renewable). Although in the first round four companies participated, two of them were Telmex and Telcel. The latter is a mobile service provider belonging to the same consortium as Telmex. In the second round of the tender process, Telmex was the only bidder. In both public tenders Telmex, the incumbent operator, was chosen. The contract with this incumbent operator was signed on February 2005 with the target of serving 5,979 communities.

There were two changes to the terms of the contract signed between Telmex and the Ministry of Communications. The first was related to the inability to serve 737 communities

⁷The total sum accounted by 750 million pesos, the exchange rate between Mexican pesos and US dollars at the time was around 10 Mexican pesos per US dollar.

due to the fact that these towns lacked an electricity infrastructure or due to difficulties imposed by weather contingencies. The settlement consisted of a time extension granted to Telmex in order to serve 506 communities.

The second change to the original contract consisted of exchanging bandwidth resources reserved by the government for social coverage purposes for bandwidth with high commercial value for Telmex. This change had severe implications for both the implementation of the universal service process and in terms of the dominant control of infrastructure on the part of the incumbent operator. This latter implication had negative consequences due to the lack of competition in the telecommunications services markets, thus affecting the economy and society as whole.

In November 2006, a few weeks before the end of the presidential and ministerial administration of 2000–2006, an exchange of frequency bands took place: its 21 MHz allocation in the 1.5 GHz band, which was originally allocated by the Ministry of Communications to Telmex as part of the Social Coverage Fund, was exchanged for 10 MHz in the 450 MHz band. The Ministry of Communications did not exercise its power to monitor the use of these frequency bands.

The exchange of frequency bands turned out to be commercially convenient for Telmex, since the 450 MHz band was the most appropriate for the provision of wireless services with technology known as CDMA 450. Among the advantages of the use of frequency resources with this technology are:

- The ability to digitalize and interleave calls with a code attached to each one, allowing a large number of simultaneous calls without interference.
- An additional advantage consisted of having a larger coverage per cell, which requires a smaller number of cells, resulting in a more cost-effective technology. Also, the possibility of supplying a wide variety of services, such as Internet, telephony, data transmission, videoconferencing and connectivity between local networks.
- This frequency band also makes use of CDMA 2000 1X and CDMA 2000 1xEV-DO technologies, which allow for high speed data transmission, equivalent to the digital service line (DSL).

The exchange of bandwidth resources dedicated to social telephony for resources with ten years of high commercial value was carried out by the Ministry of Communications and allowed Telmex access to and use of these resources without going through an open public tender. This raised questions over Telmex's interest in participating in the social coverage tender process.

Former representatives of the Office of Rural Telephony argued that Telmex's true interest was to acquire the use of the frequency bandwidth resources, with a potentially high financial return, thus evading the higher transactional and monetary costs involved in taking part in an open public

tender, which has been the allocation mechanism for radio bandwidth resources for commercial use drawn up by the government in accordance with the federal law on telecommunications (1995)⁸.

The former analysis leads us to consider the role of the government authorities in organizing tender processes and allocating public financial and bandwidth resources for social coverage. In this case, the Ministry of Communications played a different role by granting valuable infrastructure resources to be used commercially, at a very low cost for the incumbent operator.

Additionally, and based on fieldwork and remote monitoring performed by the Office of Rural Telephony, the supervision of the services offered by Telmex under the *Social Coverage Fund* (FONCOS) showed that, out of the programme objective of 109,016 telephone lines (75,797 lines under the STB-1 program and 33,219 under the STB-2 program), only 88,791 were actually installed, which implies that 20,225 lines were never installed.

There was a brief period, after 2006, when the new administration of the Ministry of Communications verified the services delivered by Telmex, under the *Social Coverage Fund*. The Ministry Office of Rural Telephony identified numerous irregularities, for example the installation of two land line connections in the same household, which proved less costly for Telmex (19,397 lines). A similar discovery was made of lines that were not connected to any specific household, which obstructed the verification of their operation (6,983 lines). In contrast, before 2006, the Ministry of Communications had punctually paid Telmex, based on the invoices that the company presented. During a brief period of time the Ministry of Communications initiated a process to impose sanctions on Telmex and to suspend payments to the company. However, this process never transcended the boundaries of the ministry because different groups within the ministry restricted the sanctioning process. Furthermore, the group that initiated this process no longer serves in the Ministry of Communications.

Here again, the analysis reveals the role of the regulator, firstly in the tender process, and specifically, in the process of allocating bandwidth resources with a high potential return for Telmex. Furthermore, the regulator did not supervise the use of these resources, which were specifically allocated for social communications coverage. This finding suggests regulatory capture and corruption on the part of the regulator [10], taken to a serious extreme since the regulator did not exercise its power in preventing the re-allocation of resources originally targeted at the poorest people, which in turn strengthened the market power of Telmex. Secondly, the regulator did not impose sanctions on Telmex for its breach of the agreement on social coverage. The role of the regulator was eclipsed, most probably by numerous instances of lobbying, resulting in a failure to

⁸Federal Telecommunications Law: Article 14. The licenses for the use of radiofrequency bands for determined uses will be granted through an open public auction. The Federal Government has the right to receive the agreed monetary resources; <http://www.diputados.gob.mx/LeyesBiblio/pdf/118.pdf>

Table 9
Telecommunication services distribution according to households' income (deciles), 2008

Service	Deciles									
	1	2	3	4	5	6	7	8	9	10
Line connection	22.6	45.8	56.4	66.5	76.2	78.5	83.4	87.7	91.7	92.6
Mobile service ¹	22.1	42.2	52.0	65.2	70.4	77.7	82.1	87.0	91.1	86.7
Cable or satellite TV	5.0	12.7	17.5	26.2	32.6	41.5	50.3	62.3	72.3	75.8
Internet	0.2	1.7	3.4	7.2	11.4	17.4	27.2	41.4	56.1	60.1

¹ Mobile services are accounted when at least one member of the household has a mobile line.
Source: INEGI: ENIGH, 2008.

consider the well being and social inclusion of the poorest sector of the population.

So far the limited achievements of the different public policies aimed at providing universal service have been presented as being due firstly to the limited implementation of the clauses set out in Telmex's licence and secondly to the violation of various agreements, including the FONCOS contract with Telmex.

There are then at least three main findings that can be drawn from the previous analysis: firstly, that the provision of universal service or universal access has been extremely limited in addressing the market gaps in Mexico's rural areas and telephone density in the different services continues to be very low. Twenty years after privatization of the public telephone company and 15 years after the liberalization of the telecommunications markets in Mexico, connectivity and telephone density remain a major challenge for public policy in Mexico. The second major finding is the continual breaching by Telmex of its universal access or service commitments, not only as was originally stated as part of its licence, but later as the result of a contract that was signed with the Ministry of Communications making it the main supplier of these services. Finally, the third finding is the limited leverage of the telecommunications authorities and their difficulty in enforcing contracts and agreements, as well as in imposing sanctions.

There is an extremely low density telecommunications infrastructure in rural areas, with the exception of the mobile infrastructure existing in some. This has led end users to rely increasingly on mobile services, which are more costly than regular services. The following section analyzes the distributive effects of this set of policies among the poorest sectors of the population.

6. Access and Expenditure in Telecommunications Services by the Poorest Sectors of the Population and their Relation to Income Distribution

In this section, we analyze the distribution of telecommunications services as a function of different levels of household income. The starting point for this analysis is

the decile distribution of households by level of income and their expenditure on telecommunications services as a proportion of their income. The source of information is the *Household Income and Expenditure Survey* for 2008, based on a nationwide representative sample. Each decile comprises the same number of households, which are ranked from the lowest to highest income. Comparing the lowest income decile with the highest, the latter is 30 times higher.

6.1. Access to Services

Those households within the lowest income decile (the poorest) have a significantly lower level of access to telecommunications services compared with households with higher income. This finding is consistent with the results previously presented on universal service provision and on the low density of infrastructure presented before. In the lowest decile, only two households in every 10 have a home telephone line connection, while in the highest income decile, nine out of every 10 have a land line connection in their homes. Very similar figures can be found for mobile services (see Table 9).

In the case of cable or satellite TV services, which have the technical capability for supporting telecommunications services and are currently used by many countries for that purpose, including urban areas of Mexico, the distribution of these services is even more skewed than land line connections or mobile services. Thus, in the lowest decile only 5% of households have access to TV based on this infrastructure, while 75% of households in the highest income decile have access to this service. The most dramatic case of this unequal distribution is found in relation to Internet access, where 60.1% of households in the highest income decile have access to the Internet, while in the lowest, only 0.02% have this service in their homes (see Table 9).

6.2. Expenditure on Telecommunications Services

The expenditure of the poorest households on telecommunications services as a percentage of their total outgoings is twice as high as the expenditure of the wealthiest households: 4.2% in the lowest deciles and 1.9% in the highest

deciles. These results suggest that the demand for telecommunications services tends to be inelastic, which means that people tend to demand and use these services regardless of their income. The larger proportion of the expenditure of poorest families is explained by the fact that they live in remote and isolated areas and depend more on public telephone booths and mobile services, which tend to be more expensive (see Table 10). As mentioned before, mobile services in Mexico have one of the highest prices compared to other developing countries (see Table 5).

Table 10

Expenditure in telecommunications as a percentage of total expenses by decil (2006=100)

Decil	Average households income per quarter MEX pesos	Expenditure in telecommunications as a percentage of total expenses
1	3,320	4.2
2	7,174	4.1
3	10,042	4.3
4	12,739	4.3
5	15,845	4.4
6	19,506	4.5
7	24,246	4.2
8	31,472	3.8
9	43,796	3.2
10	99,215	1.9

Source: INEGI: ENIGH, 2008.

An additional factor that induces greater expenditure for the poorest areas is related to the outdated definition of the areas of local service, which has no technical or economic (cost-related) basis and artificially classifies a call as long distance, incurring higher charge. This particularly affects people in rural areas, where the largest share of their traffic consists of long distance calls. The higher expenditure on telecommunications services has an impact on the opportunity for the poorest sectors of the population to spend in other areas like health, nutrition, education, home maintenance, or make a higher investment in productive activities.

These results also suggest that the provision of telecommunications services under the aegis of “universal access” or “universal service” is lagging behind the unfulfilled demand for these services. This leaves the poorest sectors of the population dependent on the supply of services at market prices, and mostly wireless services which are more expensive and difficult for them to afford.

7. Conclusions

Almost 20 years after Mexico’s privatization of the telecommunications services by the incumbent operator, Telmex (Teléfonos de México), the premise of universal service is far from being fulfilled. Thus on average, only five out of

every 10 households have access to a basic telephone service and in some states such as Tabasco and Chiapas only three, while in Oaxaca only two (1.9) out of 10 households have access to a telephone line. Similarly, when analyzing expenditure on telecommunications services, the collected data has demonstrated that in contrast to what has been the goal of the universal service policy in Mexico, namely social inclusion and overcoming poverty, the telecommunications policies have become a regressive tax for Mexico’s poorest. These dramatic results are in direct contradiction with the fact that Telmex cheaply acquired radio bandwidth for social services by continually winning their public tenders, because regrettably, they later used them exclusively for commercial purposes.

In explaining the reasons for these poor results, the paper has also pointed out the shortcomings of the regulation in place and of the implementation of this regulation by the authorities, who are directly responsible for the failure to comply with the telecommunications service policies originally targeted at the poorest regions of Mexico. These shortcomings consist essentially of:

- restricting the application of services to communities with more than 500 inhabitants;
- allowing Telmex to choose between serving rural communities either by a public telephone booth or by a land line connection to households;
- allowing a policy far below the standard provided by the International Telecommunication Union;
- providing insufficient direct government subsidies for rural telecommunications services for communities with fewer than 500 inhabitants located in the poorest states in Mexico, who not only lack a basic telecommunications infrastructure but also present the lowest telephone density, seriously jeopardizing their right to use adult distant learning education;
- imposing strong limits on competition by establishing high barriers of entry for competitors in the mobile services market, which has allowed very high prices to be set for the end user and some of the highest among developing countries (as shown in Table 5 above).

These shortcomings show that universal access to telecommunications services in Mexico is a representative case of “regulatory capture” [10], [11], where the regulators and government authorities have been “captured” by the incumbent operator and have subordinated their regulatory power to a monopoly-based profit-seeking behavior.

Finally, although it is impossible to go back to the time of privatization, there is a long list of different regulations that have already been implemented by both developed and developing countries, which were created before the described policies for efficient universal access and service were imposed upon the incumbent operator’s licence permit. These policies have achieved a larger density and better distribu-

tion of services and some of the following regulations could thus now be applied with greater chances of success:

- An obligation to provide interconnection on a non-discriminatory basis, according to high quality standards and establishing charge estimates on the basis of long term incremental cost. This applies particularly to those networks with market power: Telmex and Telcel.
- Unbundling the local loop, thus allowing non-discriminatory access to sections of the incumbent operator's network infrastructure.
- To guarantee a free flow of information on network capabilities, specifically on the points of presence and network architecture. This also applies particularly to Telmex and Telcel. This will contribute to creating incentives for new investment and the participation of new players, increasing coverage in regions previously lacking provision.
- To provide services across networks on a non-discriminatory basis, to high quality standards. It is known that international networks get better roaming services from Mexican networks than the cross-network services that national networks get between one another. This has been a barrier to entry for new players that have prevented investment and coverage in regions lacking coverage. It is worth reiterating that Telcel is the operator holding control of 75% of lines and operating the largest mobile infrastructure.
- To allow mobile virtual network operators (MVNO), which enable new players in the market to provide mobile phone services without necessarily having their own licensed bandwidth allocation, nor does it necessarily require them to have the entire infrastructure required to provide mobile telephone services.
- Re-defining the domain of local services, whose definition currently incurs an artificial increase in prices for so-called long distance calls. Without a technical basis for such a definition, this particularly affects those rural areas whose traffic is mainly long-distance based.
- Enforcing the declaration of Telmex as an operator with (monopolistic) market power and imposing upon it special requirements regarding quality, prices and information, so as to level the playing field by allowing other operators to enter the market and promote healthy competition.
- Closely monitoring Telmex and Telcel, in order to guarantee the proper delivery of telecommunications services to the poorer areas of Mexico.
- To ensure accountability and transparency in all legal processes relating to telecommunications services, regulation and competition. This not only provides legal certainty, but is a potential antidote for regulatory capture and corruption.

References

- [1] S. Maddens, *Trends in Universal Access and Service Policies: Changing Policies to Accommodate Competition and Convergence*, in "Support for the Establishment of Harmonized Policies for the ICT Market in the ACP", Geneva, ITU, 2005 [Online]. Available: http://www.itu.int/ITU-D/treg/Events/Seminars/GSR/GSR09/doc/USPolicy_ITUEC.pdf
- [2] ITU, "Universal access and service", Module 4. ICT Regulation Toolkit, Section 1.1.2., 2008 [Online]. Available: <http://www.ictregulationtoolkit.org/en/Sections.html>
- [3] UIT, Informe sobre las soluciones innovadoras en materia de gestión y financiación de las políticas de servicio y acceso universales. Ginebra: UIT-D, Comisión de Estudio 1, 3er Periodo de Estudios (2002–2006), 2006 [Online]. Available: <http://www.itu.int/opb/publications.aspx?lang=es&parent=D-STGSG01.07.1&folder=D-STG-SG01.07.1-2006>
- [4] R. del Villar, "Competition and equity in telecommunications", in *No Growth without Equity? Inequality, Interests and Competition in Mexico*, S. Levy and M. Walton, Eds. Washington: Palgrave MacMillan and The World Bank, 2009, pp. 321–364.
- [5] C. Casanueva and R. del Villar, "Infrastructure regulation difficulties: the basic telecommunications industry in Mexico (1990–2000)", in *Critical Infrastructures State of the Art in Research and Application*, Wil A. H. Thissen and P. M. Herder, Eds. Dordrecht: Kluwer, 2003, pp. 179–207.
- [6] C. Casanueva, "The opening of Mexico's telecommunications markets", in *Telecommunications Reform in India*, R. Dossani. Connecticut: Quorum Books, 2002, pp. 277–296.
- [7] R. G. Noll, "Notes on privatizing infrastructure industries", in *Conf. Telecommun. Reform in India*, Asia/Pacific Research Center, Stanford University, November 9 and 10, 2000 [Mimeo].
- [8] "Censo de Población y Vivienda de 1990" [Online]. Available: <http://www.inegi.org.mx/est/contenidos/proyectos/ccpv/cpv1990/default.aspx>
- [9] S. Wallsten, S. "Telecommunications privatization in developing countries: the real effects of exclusivity periods", in *Conf. Telecommun. Reform in India*, Asia/Pacific Research Center, Stanford University, November 9 and 10, 2000 [Mimeo].
- [10] F. Boehm, "Corrupción y captura en la regulación de los servicios públicos", *Revista de Economía Institucional*, vol. 7, no. 13, pp. 245–262, 2005.
- [11] S. Haber, *Crony Capitalism and Economic Growth in Latin America: Theory and Evidence*. Stanford: Hoover Institution, 2002.

Note: this text is reprinted from JTIT 2/2010



Cristina Casanueva-Reguart

received her Ph.D. in Development Studies from Stanford University in 1988 and holds an MA from Harvard University and a B.A. in sociology from the Universidad Iberoamericana, Mexico City. She worked in the Federal Government (1992–1998) where she was part of the Federal

Competition Commission, Mexico's anti trust agency. Working in the Economic Regulation Department, she focused on the analysis of telecommunications markets, participating in the interministerial group that defined the rules of interconnection. In the following administration she joined the Ministry of Communications, where she was involved in drafting the Federal Law on Telecommunications. In 1998 she returned to an academic position at the Monterrey Institute of Technology, in Mexico City, where she led the creation of an Interdisciplinary Masters Pro-

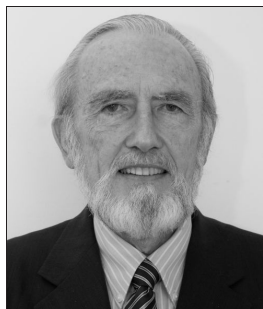
gram on the Management of Telecommunications. In 2001, Dr Casanueva joined the Universidad Iberoamericana where she has written numerous articles on telecommunications, regulation and public policy. In 2007–2008, she worked with the Ministry of Communications as a consultant in the field of universal telecommunications service provision.

e-mail: ccasanueva@stanfordalumni.org

Universidad Iberoamericana

Prolongación P. Reforma 880

01210 Mexico D.F., Mexico



Antonio Pita S. obtained his Ph.D. and M.Sc. in solid and fluid mechanics from the University of Iowa in 1967 and 1962, and his B.S.E. from the Johns Hopkins University in 1960. He spent 15 years in teaching and research at the Tecnológico de Monterrey's campus in Monterrey, Mexico

and 25 years in R&D at Vitro, a Multinational Mexican-based glass company, including five years as Director of the Corporate R&D Center, 10 years as Technology Director for one of Vitro's Business Units and 10 years as a Senior Researcher. His last assignment was the establishment of Vitro's new R&D centre in Fribourg, Switzerland. Since then he has been Manager of University Relations for the Global Centre for Technology and Innovation that Cemex, a Multinational Mexican-based cement company, has in Brugg, Switzerland and is also Director of the Swiss Office for International Relations for the Tecnológico de Monterrey. He has published 12 articles, 25 Technical reports and holds 3 US and International Patents on Glass and Pre-Reacted Batches of Raw Materials for the Production of Glass Formulas.

e-mail: antonio.pita@itesm.mx

Cemex, Manager of University Relations

for the Global Centre for Technology and Innovation

Römerstrasse 13

2555 Brugg, Switzerland

Performance Analysis of a Bi-Objective Model for Routing and Wavelength Assignment in WDM Networks

Carlos Simões, Teresa Gomes, José Craveirinha, and João Clímaco

Abstract—Establishing end-to-end connections on wavelength division multiplexing (WDM) networks requires setting up lightpaths, defining the sequence of optical fibres and the wavelength in each fibre (the routing and wavelength assignment problem) for traffic flow. This paper reviews a bicriteria model for obtaining a topological path (unidirectional or symmetric bidirectional) for each lightpath request in a WDM network, developed by the authors, and presents a performance analysis of the model by considering important network performance measures. An extensive performance analysis of the two bicriteria model is presented, comparing the performance metrics obtained with the monocriterion models using the same objective functions, in five different reference networks commonly used in literature.

Keywords—multicriteria optimization, routing in WDM networks.

1. Introduction

1.1. Background Concepts

All-optical networks based on wavelength division multiplexing (WDM) have emerged as a promising technology for network operators to respond to an increased demand for broadband services, exploiting the huge bandwidth of optical fibres. All-optical networks based on wavelength division multiplexing consist of optical fibre links and nodes, and the WDM scheme divides the optical bandwidth into independent channels, each one with a different wavelength, operating at transmission rates compatible with the lower capacity of the end user's devices. Each node in a all-optical network has a dynamically configurable optical switch or router which supports wavelength based switching or routing. Configuring these optical devices across the network enables that node pairs can establish point-to-point all-optical connections, or lightpaths, for information transfer. A lightpath may span several fibre links and consist of wavelength channels in the sequence of this links, interconnected at the nodes by means of optical routing. In order to establish a lightpath, the network needs to decide on the topological route and the wavelength(s) for the lightpath. In the absence of wavelength converters, a lightpath must use the same wavelength on all the links of its route (the *wavelength continuity constraint*), but wavelengths can be reused by different lightpaths in the network, as long as they do not share any fibre link.

Given a set of connection requests, the problem of setting up lightpaths by defining a path and assigning a wavelength to each of its links for every connection is called the routing and wavelength assignment (RWA) problem.

If the network nodes have wavelength converters, it is possible to assign different wavelengths on the multiple links of the lightpath. As a result, the wavelength continuity constraint is relaxed, thereby increasing the possible number of lightpaths that can simultaneously be established in the network. However, since wavelength converters are costly and may cause signal quality degradation, often no wavelength converters are used or only some nodes have this capability. The converter configuration of the network is called full if all nodes have wavelength converters and sparse if only a part of the nodes have them. Obviously, wavelength conversion leads to lower blocking probabilities, but, in practice, some works have shown that with only a small number of converters placed in strategic locations, a significant performance improvement can be achieved [1]. On the other hand, when a node is capable of converting a wavelength to any other wavelength, the node is said to have complete conversion capability. If a node is able to convert an incoming wavelength to only a subset of available wavelengths, the node is said to have limited or partial conversion capability. A wavelength converter is said to have a conversion degree D , if it can shift any wavelength to one of D wavelengths.

Multi-fibre networks use several fibres per link. Considering that the nodes have no wavelength converters, the possibilities of finding a lightpath satisfying the wavelength continuity constraint is higher than in single fibre networks. A multi-fibre network with F fibres per link and W wavelengths per fibre is functionally equivalent to a single-fibre network with $F \times W$ wavelengths and conversion degree of F [2].

Connection requests are usually considered to be of three types: static, incremental, and dynamic [3]. Regarding static traffic, the entire set of connections is known in advance, and the problem is then to set up lightpaths for these connections seeking to minimize network resources such as the number of wavelengths or the number of fibres in the network. In this case the RWA problem is known as the static lightpath establishment (SLE) problem. Concerning incremental-traffic, connection requests arrive sequentially, a lightpath is established for each connection, and

the lightpath remains in the network indefinitely. In the case of dynamic traffic, a lightpath is set up for each connection request as it arrives, and the lightpath is released after some finite amount of time. The objective in the incremental and dynamic traffic cases is to set up lightpaths and assign wavelengths seeking to minimize the amount of connection blocking, or to maximize the number of connections that are established in the network at any time. This problem is designated as the dynamic lightpath establishment (DLE) problem. The SLE problem can be formulated as an integer linear program [4], which is known to be NP-complete [5].

In order to make the RWA problem more tractable, it can be divided into two sub-problems – routing and wavelength assignment, so that each sub-problem can be solved separately. Nevertheless note that each sub-problem is still NP-complete [5].

Routing methods in WDM networks can be classified into two types: static routing and adaptive routing. Static routing encompasses fixed routing and fixed-alternate routing. In fixed routing the same fixed route for a given source-destination pair remains unchanged throughout time. The fixed-alternate routing scheme pre-computes a set of paths between each source-destination pair, and for each request, a path from this pre-computed set is chosen. Adaptive routing involves a dynamical search for a path when a connection request arrives, taking into account the current state of the network. Therefore in general, adaptive routing gives better blocking performance than fixed-alternate routing [3].

A wavelength assignment algorithm is used to determine the wavelengths in the arcs along the path chosen in the routing step. Many wavelength assignment algorithms have been proposed such as random, first fit, most-used, least-used, least-loaded, max-sum, min-product, and relative capacity loss schemes [3].

In most approaches presented in the literature, routing and wavelength assignment are *optimized* separately by considering a decomposition of the global RWA problem through heuristic algorithms, because these problems are NP-complete. However, some algorithms consider the routing and the wavelength assignment jointly [6], [7].

Many different integer linear programming (ILP) formulations have been proposed for the RWA problem in WDM optical networks, under different objectives. However, although those formulations lead to exact solutions, most of the times they have not been used for developing solution schemes except for very small networks or for some rounding off procedures [8]. Examples of monocriterion approaches to the RWA problem, considering different metrics as objective functions are in the references: [3], [4], [6], [7], [9]–[12].

1.2. Multicriteria Models

In general routing protocols only optimize one metric, typically using some variant of a shortest path algorithm.

However all-optical WDM networks can be characterized in terms of performance by multiple metrics. Also the design of real networks usually involves multiple, often conflicting objectives and various constraints. The development of multicriteria models that explicitly represent the different performance objectives, enabling to treat in a consistent manner the trade off among the various criteria, seems to be potentially advantageous in face of the inherent limitations of single objective approaches.

Note that in models involving explicitly multiple and conflicting criteria, the concept of optimal solution is replaced by the concept of non-dominated solutions. A non-dominated solution is a feasible solution such that no improvement in any criterion may be achieved without sacrificing at least one of the other criteria.

A state-of-art review on multicriteria approaches in communication networks was presented in [13], including a section dedicated to routing models. A recent review on multicriteria routing models can be seen in [14].

In [15] two different criteria, path length and congestion in the network, are considered and applied sequentially (the second metric is only used if a tie occurs in the first one). Two algorithms for dynamic traffic were proposed: least congested shortest hop routing where priority is given to efficient resource utilization (the algorithm selects the least congested path among all shortest hop paths currently available); and shortest hop least congested routing in which priority is given to maintaining the load balance in the network (it selects the shortest hop path among all the least congested paths). This type of models may be considered as a first-tentative multicriteria approach as analyzed in [13].

In [16] a set of link disjoint routes for each node pair, in a network with dynamic traffic, is pre-computed. Then the weighted least-congestion routing and first-fit wavelength assignment algorithm, which includes two criteria (hop count and free wavelengths), combined in a single weighted metric, is used to rank the paths. The same approach was proposed earlier in [17], which tries to minimize the resource utilization while simultaneously balancing the traffic load in the links. Nonetheless [17] only considers networks with full wavelength conversion. These models, which consider as solution of the bicriteria problem the optimal solution of the single objective function resulting from the weighted sum of the considered criteria, do not take full advantage of the possibilities of multicriteria approaches and may lead to less effective solutions.

A bicriteria model for obtaining a topological path (unidirectional or symmetric bidirectional) for each lightpath request in a WDM network with multi-fibre links and an exact resolution approach for that model was presented by the authors in [18]. The first criterion is related to bandwidth usage in the links (or arcs) of the network. The second criterion is the number of links (hops) of the path. The resolution approach [18] uses an exact procedure to calculate non-dominated topological paths based on a k -shortest

path algorithm [19] which is based on an adaptation of the MPS algorithm [20]. Furthermore, preference thresholds, defined in the objective function's space, combined with a Chebyshev distance to a reference point [21] are used for selecting the final solution. The solution of this bicriteria model is a non-dominated topological (optically feasible) path. A heuristic procedure is then used to assign wavelengths to the links.

The focus of this paper is to present an extensive and systematic performance analysis study of the bicriteria routing model [18] with respect to certain network performance measures by comparison of their results with the results of the associated single objective models, one related to the bandwidth usage and another that minimizes the number of used links (hop count). An incremental traffic model (where the duration of the connections is assumed unlimited) will be considered in several benchmark networks used in previous works in the area of WDM networks. The selected network performance measures are: the frequency of rejected requests (global blocking probability estimate), total used bandwidth, mean hop count of accepted requests, percentage of links with minimal free bandwidth, the average CPU time per request, and the percentage of non-optimal solutions.

The paper is organized as follows. In Section 2 the model without protection is described, together with the resolution approach of the bicriteria model. Performance analysis of the results obtained using several network topologies are presented and discussed in Section 3. Finally, some conclusions are drawn in Section 4.

2. The Bicriteria Routing Model

2.1. Model Description

In this section we review the features of the bicriteria routing model associated with the dynamic lightpath establishment problem with incremental traffic, in a WDM network, proposed in [18]. The model was developed for application in large WDM networks, with multiple wavelengths per fibre and multi-fibres per link. The bicriteria routing model considers the DLE problem with incremental traffic, and a mixture of unidirectional and bidirectional (symmetric) connections. In order to cover a wide variety of networks, different types of nodes are considered (with complete wavelength conversion capability, limited range conversion or no wavelength conversion capability) in the model. Due to the real-time nature of the intended application, solutions should be obtained in a short time. This requirement lead to the separation of the routing and wavelength assignment problems, having in mind an automatic selection of the solution (among the non-dominated solutions, previously identified). The wavelength assignment problem is solved separately, after the bicriteria routing problem.

Let $R = \{N, L, C, T_N\}$ represent the WDM network where:

- N is the set of nodes, $N = \{v_1, v_2, \dots, v_n\}$, $n = \#N$.
- L is the set of directed arcs, $L = \{l_1, l_2, \dots, l_m\}$, $m = \#L$.
- Set of wavelengths, $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_W\}$, $W = \#\Lambda$.
- Set of fibres, $F = \{f_1, f_2, \dots, f_k\}$, $k = \#F$.
- Let $l_i = (v_a, v_b, \bar{o}_{l_i})$, $\bar{o}_{l_i} = (o_{l_i1}, o_{l_i2}, \dots, o_{l_ik})$, $v_a, v_b \in N$.

If $o_{l_ij} = (1, \bar{a}_j)$ ($j = 1, 2, \dots, k$), then fibre f_j belongs to arc l_i and contains the wavelengths signalled in $\bar{a}_j$, $\bar{a}_j = (a_{j1}, a_{j2}, \dots, a_{jW})$ where $a_{ju} = 0, 1, 2$ ($u = 1, 2, \dots, W$):

$$a_{ju} = \begin{cases} 0, & \text{if } \lambda_u \text{ does not exist in fibre } f_j \\ 1, & \text{if } \lambda_u \text{ exists and is free in fibre } f_j \\ 2, & \text{if } \lambda_u \text{ exists but is busy in fibre } f_j \end{cases} \quad (1)$$

If $o_{l_ij} = (0, \bar{a}_j)$ ($j = 1, 2, \dots, k$), fibre f_j does not belong to arc l_i .

- C is the arc capacity, $C(l_i) = (\bar{n}_{l_i}, \bar{b}_{l_i})$, with $\bar{n}_{l_i} = (n_{l_i1}, n_{l_i2}, \dots, n_{l_iW})$ and $\bar{b}_{l_i} = (b_{l_i1}, b_{l_i2}, \dots, b_{l_iW})$ where n_{l_ij} is the total number of fibres in arc l_i with wavelength λ_j and b_{l_ij} is the number of fibres where that wavelength is free in arc l_i .
- $T_N(v_i)$ is a table for each node $v_i \in N$ which represents the wavelength conversion capability of the nodes, that is the possibility of transferring the optical signal from one input λ_i to an output λ_j in the node:

$$T_N(v_i) = [t_{uv}], \quad \forall v_i \in N; u, v = 1, 2, \dots, W, \quad (2)$$

where $t_{uv} = 1$ (0) whether (or not) λ_u can be converted into λ_v , in node v_i .

A *topological path*, p in R , is described by: a source node, a destination node ($v_s, v_t \in N$) and the ordered sequence of nodes and arcs in the path, $p = \langle v_1, l_1, v_2, \dots, v_{i-1}, l_{i-1}, v_i \rangle$, such that the tail of arc l_k is v_k and the head of l_k is v_{k+1} , for $k = 1, 2, \dots, i-1$ (all the v_i in p are different).

Besides the ordered sequence of nodes and arcs, a *lightpath* p^λ also comprises the fibre used in each arc and the wavelength on the fibres:

$$p^\lambda = \langle l_c^*, \dots, l_d^* \rangle = \langle (v_s, v_u, f_i, \lambda_\alpha), \dots, (v_x, v_t, f_j, \lambda_\beta) \rangle \quad (3)$$

where $f_i, \dots, f_j \in F$, $\lambda_\alpha, \dots, \lambda_\beta \in \Lambda$, represent fibres and wavelengths, respectively.

Note that l_c^* corresponds to $l_c = (v_s, v_u, \bar{o}_{l_c})$ which implies $o_{l_ci} = (1, \bar{a}_i)$ and if $a_{i\alpha} = 1$ then $a_{i\alpha}$ will change from 1 to 2 if p^λ is selected.

A *bidirectional lightpath* $p^\lambda = (p_{st}^\lambda, p_{ts}^\lambda)$ is supported by a *bidirectional topological path* $p = (p_{st}, p_{ts})$, which is a pair of symmetrical topological paths.

Firstly we will describe the bicriteria model used for calculating topological paths.

The first objective function, $c_1(p)$ is related to the bandwidth usage in the links of the path p and is expressed in the inverse of the available bandwidth in the links:

$$\min_{p \in D} \left\{ c_1(p) = \sum_{l \in p} \frac{1}{b_l^T} \right\}, \quad (4)$$

where D is the set of topological paths for the origin-destination node pair and b_l^T is the total available capacity in link l , in terms of available wavelengths. This criterion seeks to avoid already congested links, favoring a balanced distribution of traffic throughout the network, and hence decreasing the blocking probability and therefore increased the expected revenue.

Note that the values of the available bandwidths b_l^T to be used in each instance of the resolution of the bi-objective optimization problem are directly calculated from the vector b_l in $C(l)$:

$$b_l^T = \sum_{j=1}^W b_{lj}, \quad \forall l \in L. \quad (5)$$

The second objective consists of minimizing the number of arcs of the path, $h(p)$, seeking to avoid bandwidth waste, hence favouring global efficiency in the use of network resources as well the reliability of optical connections (longer paths are more prone to failure).

$$\min_{p \in D} \{c_2(p) = h(p)\}. \quad (6)$$

Note that in many cases there is no feasible solution which optimizes the two objective functions, $c_1(p)$ and $c_2(p)$, simultaneously. A certain amount of conflict is therefore expected between c_1 and c_2 , and no optimal solution (in most cases) will exist for this problem. Therefore the candidate solutions to the topological RWA multicriteria model are topological paths which are non-dominated solutions to the bi-objective problem:

$$(\mathcal{P}) \quad \left\{ \begin{array}{l} \min_{p \in D_T} c_1(p) \\ \min_{p \in D_T} c_2(p) \end{array} \right. \quad (7)$$

Given two paths p_1 and p_2 , from s to t in R , path p_1 dominates p_2 , denoted by $p_1 D p_2$, if and only if $c_i(p_1) \leq c_i(p_2)$ ($i = 1, 2$) and at least one of the inequalities is strict. A path p is a non dominated solution if no other feasible path dominates it.

The set of admissible solutions, D_T , consists of all topological paths between the source-destination node pair which correspond to *viable lightpaths* p^λ , that is, lightpaths with the same arcs as p and with a free and usable wavelength (according to T_N) in every arc. The topological paths in these conditions (elements of D_T) will be designated as *viable topological paths*, for the given origin-destination node pair. For obtaining D_T firstly the free wavelengths in each arc will have to be identified, taking into account the wavelength conversion capabilities specified in T_N , then the set of viable paths p^λ for each pair of origin-destination nodes becomes implicitly defined.

This model was extended to bidirectional connections between nodes s and t by considering a bidirectional lightpath $p^\lambda = (p_{st}^\lambda, p_{ts}^\lambda)$ supported by a bidirectional topological path $p = (p_{st}, p_{ts})$ which is a pair of symmetrical topological paths. In this case the set D_T^b of feasible solutions to the bicriteria model will be the set of viable bidirectional topological paths p , i.e., characterized by the fact that both (unidirectional) topological paths p_{st} and p_{ts} are viable. Therefore the bi-objective bidirectional routing optimization problem is formulated as:

$$\min_{p \in D_T^b} \left\{ c_1(p) = \sum_{l \in p_{st}} \frac{1}{b_l^T} + \sum_{l \in p_{ts}} \frac{1}{b_l^T} \right\} \quad (8)$$

$$\min_{p \in D_T^b} \{c_2(p) = h(p) = h(p_{st}) + h(p_{ts})\} \quad (9)$$

We will assume the most common situation in real networks where the two paths p_{st}, p_{ts} are topologically symmetrical, thence $h(p) = 2h(p_{st})$. Note that this does not imply that the wavelengths used in the two opposite directions are necessarily symmetrical.

2.2. Resolution Approach

The first stage of the resolution approach is an exact algorithm enabling the calculation of non-dominated viable topological paths and the selection of a path according to an automatic procedure that uses preference thresholds defined in the objective function's space and reference points obtained from those thresholds. This algorithmic approach will be reviewed in this subsection.

The second stage is the assignment of wavelengths (and corresponding fibres) along the arcs of the selected path p . For this purpose we will use the maximization of the *wavelength bottleneck bandwidth*, $b_j(p)$ ($\lambda_j \in \Lambda$):

$$\max_{\lambda_j \in \Lambda} \left\{ b_j(p) = \min_{l \in p \wedge b_{lj} > 0} b_{lj} \right\} \quad (p \in D_T). \quad (10)$$

Note that this procedure is equivalent to the choice of the least loaded wavelength (LL) along the arcs of the path. Moreover, if all the nodes of the network enable full wavelength conversion, once a viable topological path is selected, the choice of the wavelength(s) in the arcs is irrelevant in terms of network performance. When the nodes have no conversion capability the proposed scheme of wavelength selection is known to give good results (see, e.g., [3]). In any case it can be concluded from many studies that the critical factor in terms of network performance is the selection of topological paths, the choice of wavelength having a minor impact.

In the present model this choice of wavelength will correspond to specify λ_{j^*} in arc l^* :

$$b_{l^*j^*} = \max_{\lambda_j \in \Lambda} \left\{ b_j(p) = \min_{l \in p \wedge b_{lj} > 0} b_{lj} \right\} : \exists \text{ viable } p^\lambda \text{ which uses } \lambda_{j^*} \text{ in } l^* \in p \quad (11)$$

For bidirectional connections, once a non-dominated solution $p \in D_T^b$ has been selected, the wavelengths (and fibres) to be used along p_{st} and p_{ts} are chosen applying the same procedure to each path. Note that the chosen wavelength(s) in each path can be different.

The aim of the resolution procedure is to find a *good* compromise path from the set of non-dominated solutions, according to certain criteria, previously defined. It must be stressed that path calculation and selection have to be fully automated, as part of a telecommunication network routing mechanism, therefore the use of an interactive decision approach is precluded.

The candidate solutions are computed according to an extremely efficient k -shortest path algorithm, MPS [20], [22], by using a version adapted to paths with a maximum number of arcs (length constrained k -shortest paths) as described in [19]. The algorithm is applied to the convex combination of the two objective functions:

$$f(p) = \alpha c_1(p) + (1 - \alpha)c_2(p) \quad 0 \leq \alpha \leq 1. \quad (12)$$

Note that the value of α just determines the order in which the solutions are found, and its choice is purely instrumental, since all non-dominated solutions can be calculated.

The selection of a solution is based on the definition of preference thresholds for both functions in the form of requested and acceptable values for each of them. These thresholds enable the specification of priority regions in the objective function's space, as illustrated in Fig. 1.

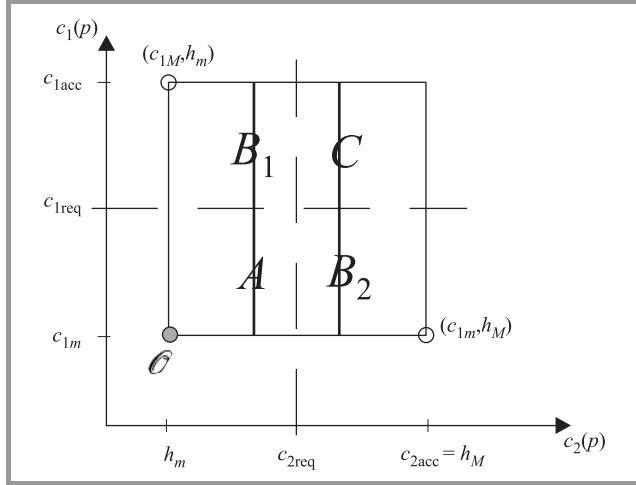


Fig. 1. Preference regions.

In the first step, vertex solutions p^{c1} and p^{c2} (viable topological paths) which optimise each objective function, $c_1(p)$ and $c_2(p) = h(p)$, respectively, are computed by solving the associated shortest path problems. This leads to the ideal solution, $\mathcal{O}$, in the objective functions' space.

$$p^{c1} = \arg \min_{p \in D_T} c_1(p) \quad (13)$$

$$p^{c2} = \arg \min_{p \in D_T} \{c_2(p) = h(p)\} \quad (14)$$

The preference thresholds c_{1req} , c_{2req} (requested values) and c_{1acc} , c_{2acc} (acceptable values) for the two metrics are given, taking into account the discrete nature of $c_2(p) = h(p)$, according to the following expressions:

$$c_{1m} = c_1(p^{c1}) \quad \wedge \quad c_{2M} = h_M = c_2(p^{c1}), \quad (15)$$

$$c_{1M} = c_1(p^{c2}) \quad \wedge \quad c_{2m} = h_m = c_2(p^{c2}), \quad (16)$$

$$c_{1acc} = c_{1M}, \quad (17)$$

$$c_{2acc} = h_M, \quad (18)$$

$$c_{1req} = \frac{c_{1m} + c_{1M}}{2}, \quad (19)$$

$$c_{2req} = \left\lfloor \frac{h_m + h_M}{2} \right\rfloor. \quad (20)$$

Priority regions are defined in the objective functions' space according to Fig. 1, in which non-dominated solutions are searched for. Region A (Fig. 1) is the first priority region where the requested values for the two functions are satisfied simultaneously. In the second priority regions, B_1 and B_2 , only one of the requested value is satisfied while the acceptable value for the other function is also guaranteed. Concerning these two regions we will give preference to solutions with less arcs, that is, preference is given to solutions in B_1 over solutions in B_2 . In region C only the acceptable values, c_{1acc} and c_{2acc} , are satisfied and it is the last priority region to be searched for.

Finally it will be necessary to select a solution among the non-dominated solutions in the highest priority region, with at least one non-dominated solution, $S \in \{A, B_1, B_2, C\}$. This implies that if no such solutions were found in A, then non-dominated solutions in B_1 , B_2 , C, in this order would be searched for.

Concerning the selection of a solution when there is more than one non-dominated solution in a given region S, the used method for ordering such solutions is a reference point type approach that considers that the 'form' of the region where solutions are located 'reflects' in some manner the user's preferences. At this step a reference point based procedure of the type proposed in [23] is used, by considering as reference point the 'left bottom corner' of region S. This point coincides with the ideal optimum if $S = A$.

Reference type approaches minimize the distance of the solutions to a specific point by using a certain metric, recurring to a scalarizing function [21]. In the present context a weighted Chebyshev metric proportional to the size of the "rectangle" S (see Fig. 2) is used. Therefore, one will select the solution p^* :

$$p^* = \arg \min_{p \in S_N^c} \max_{i=1,2} \{w_i |c_i(p) - \underline{c}_i|\}, \quad (21)$$

where S_N^c is the set of non-dominated paths which correspond to points in S and $(\underline{c}_1, \underline{c}_2)$ is the considered

reference point which corresponds to the ‘left bottom corner’ of region S . The weights w_i of the metrics are chosen in order to obtain a metric with dimension free values:

$$w_i = \frac{1}{\bar{c}_i - \underline{c}_i}, \quad (22)$$

where $(\bar{c}_1, \bar{c}_2)$ is the ‘right top corner’ of S , so that $\underline{c}_i \leq c_i(p) \leq \bar{c}_i$ ($i = 1, 2$) for all p such that $(c_1(p), c_2(p)) \in S$. An illustrative example is in Fig. 2, where the number assigned to each bullet is the computation order of the corresponding solution (according to Eq. (12)), and solution (2) would be the one to be selected, since it has the shortest distance to the reference point.

Details of this selection procedure can be seen in [18], [23].

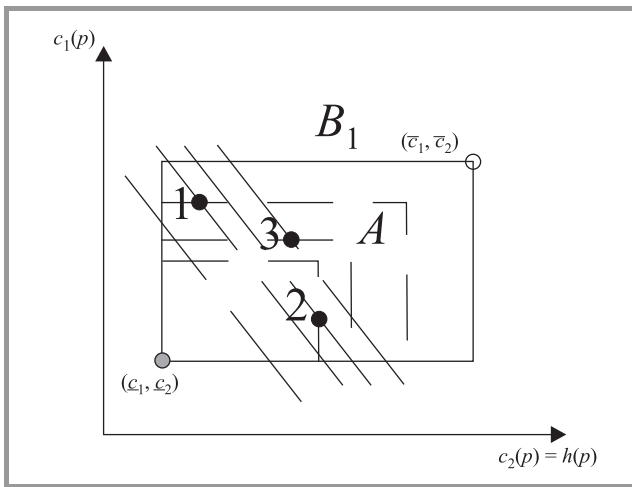


Fig. 2. Choosing the final solution.

In this resolution method, we combine a weighted sum procedure to obtain candidate solutions with a reference-point based method to select a solution in a higher priority region. In this form we sought to make the most of the very great efficiency of the used shortest path ranking procedure, based on the MPS algorithm [22] and of the inherent superiority of the use of a reference point-based procedure, as a solution selection method. Furthermore, note that in the present context, the computational efficiency is a very important aspect taking into account the automated nature of the routing mechanism, which requires a solution in a very short time period. This factor becomes more critical in networks of higher dimension.

The final step of the resolution method is the choice of wavelengths along the arcs of the selected path. This step follows the procedure described in Subsection 2.2 which is based on the maximization of the wavelength bottleneck bandwidth.

The described resolution method can be applied straightforwardly to the calculation and selection of bidirectional lightpaths, considering the necessary adaptations to the objective functions, specified in Eqs. (8) and (9).

3. Performance Analysis of the Model

Extensive simulations with the model were made on several typical WDM networks found in literature. This section presents the simulation results for five such net-

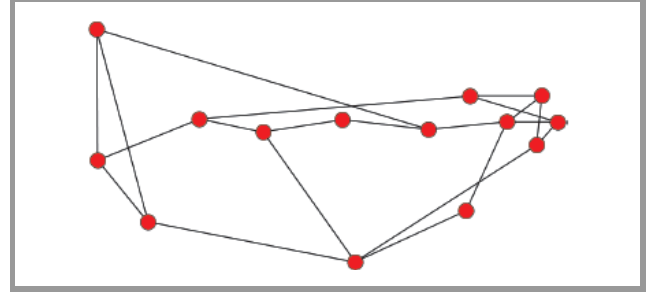


Fig. 3. NSFNET network (14 nodes and 21 links) [24].

works, namely, the NSFNET [24] (see Fig. 3), the Pan-European network COST 266BT [24], [25] (Fig. 4), the denser version of this network [25] – COST 266TT network (see Fig. 5), a typical core network presented in [26] – KL network (Fig. 6), and a typical network provider network presented in [27] – ISP network (Fig. 7). Table 1 summarizes the main characteristics of these networks. All the networks were dimensioned for about one thousand bidirectional lightpaths (1084 for NSFNET, 1008 for both COST 266BT and Cost 266TT, 1050 for KL network, and 918 for ISP network) and each fibre has 16 wavelengths.

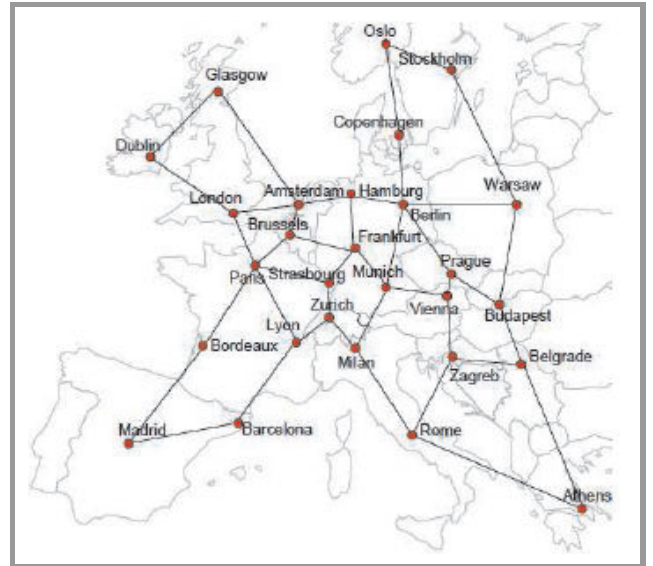


Fig. 4. COST 266BT Pan-European network (28 nodes and 41 links) [24], [25].

Concerning the wavelength conversion aspect, simulations were conducted considering two different scenarios: all nodes without conversion capability and five nodes with total conversion capability (central nodes were chosen with this capability).

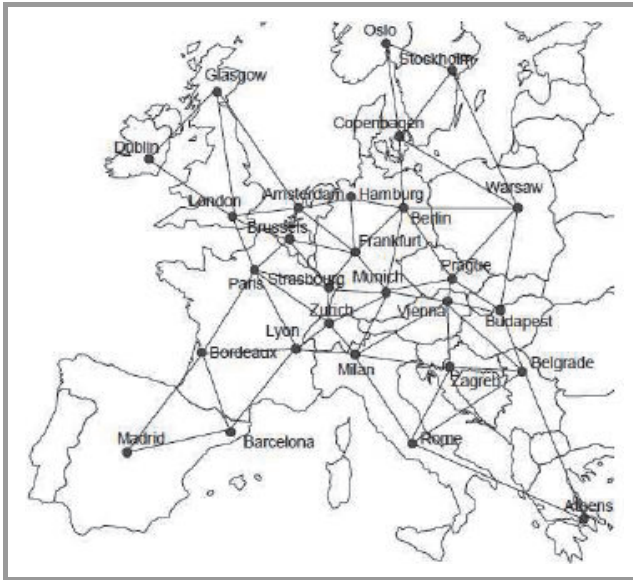


Fig. 5. COST 266TT network (28 nodes and 61 links) [25].

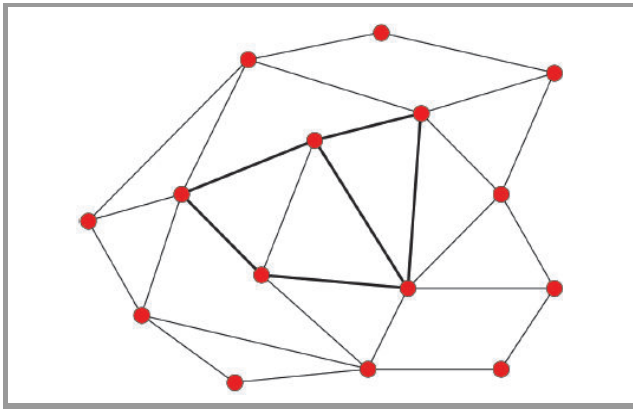


Fig. 6. KL network (15 nodes and 28 links) [26].

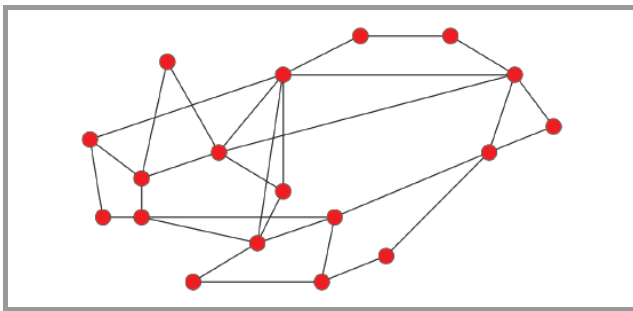


Fig. 7. ISP network (18 nodes and 30 links) [27].

Table 1
Networks characteristics

Network	Number of		Nodal degree
	nodes	links	
NSFNET [24]	14	21	3.00
COST266BT [24, 25]	28	41	2.93
COST266TT [25]	28	61	4.36
KL [26]	15	28	3.73
ISP [27]	18	30	3.33

Simulations considered 1200 connection requests (incremental traffic) in two different cases: with 100% bidirectional requests and with 5% unidirectional requests (because most of the connection requests for lightpaths are bidirectional).

Simulation results showed that the performance of the networks where five nodes have total conversion capability is nearly the same as for the networks without conversion. Therefore, from now on, we only present the “no conversion” scenario. Discussion and conclusions remain true for the second scenario.

For performance assessment purposes, results obtained using the bicriteria (BiC) model will be compared with the corresponding results using the single objective formulations, namely, the first objective function related with the bandwidth usage (SP_c1), and the shortest path concerning hop count (SP_c2). Several relevant network performance measures will be used in this comparison.

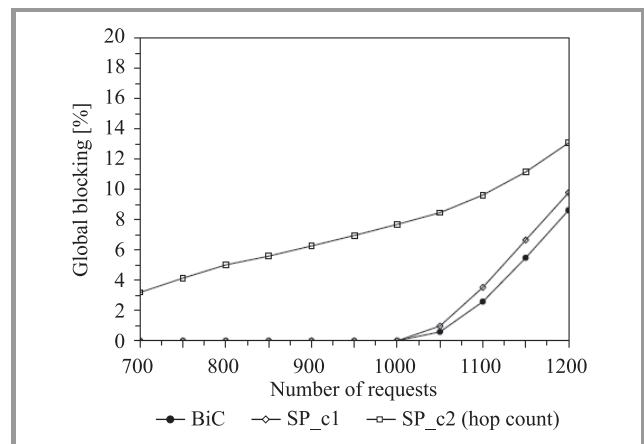


Fig. 8. Global blocking – NSFNET network.

Figure 8 shows the global blocking probability in the NSFNET network, for 100% bidirectional requests. As we can see, the bicriteria approach leads to a blocking probability lower than that in the single objective formulations. The difference is significantly higher with the shortest path approach SP_c2. Also note that until 1000 requests both BiC and SP_c1 do not exhibit any blocking. The SP_c2 model rejects requests much earlier, and this clearly confirms that choosing the shortest path based only on the hop count is a poor strategy.

In Fig. 9, the number of accepted requests is shown together with the used bandwidth (BW). When the number of requests is high, the used bandwidth exceeds 95%, but this corresponds to approximately 1100 accepted requests in BiC, a value that surpasses the number of connections for which the network was dimensioned (1084 for the NSFNET).

Although the BiC model uses more bandwidth than the SP_c2, it should be noted that BiC supports a significantly higher number of connections. In fact, as it can be seen in Fig. 10, BiC allows a lower average number of hops per connection. Another interesting conclusion that emerges from the analysis of Fig. 9 is that the BiC, while

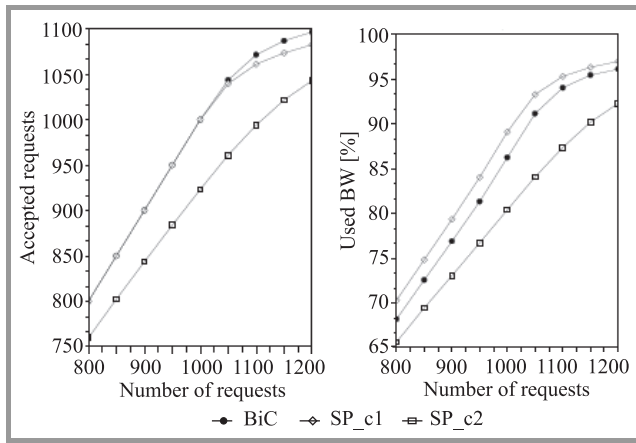


Fig. 9. Accepted requests versus used bandwidth – NSFNET network.

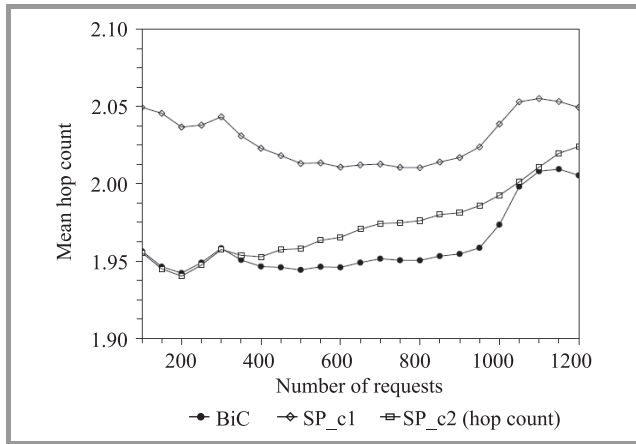


Fig. 10. Mean hop count – NSFNET network.

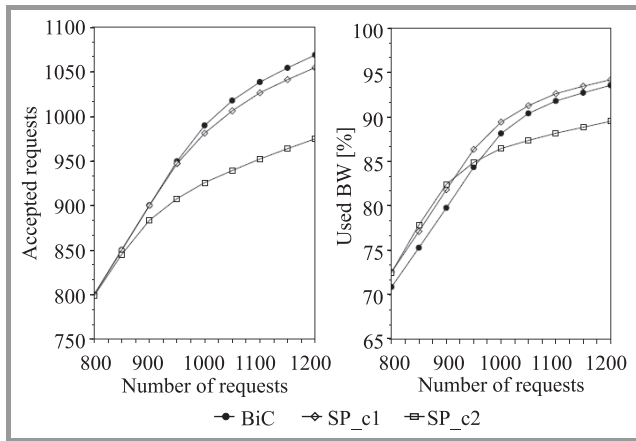


Fig. 11. Accepted requests versus used bandwidth – COST 266BT network.

accepting more requests than SP_c1, when traffic load is high, always uses less bandwidth, which shows its superior performance.

Although not shown here, the results obtained when 5% of the requests were unidirectional are rather similar to the ones with 100% bidirectional connections.

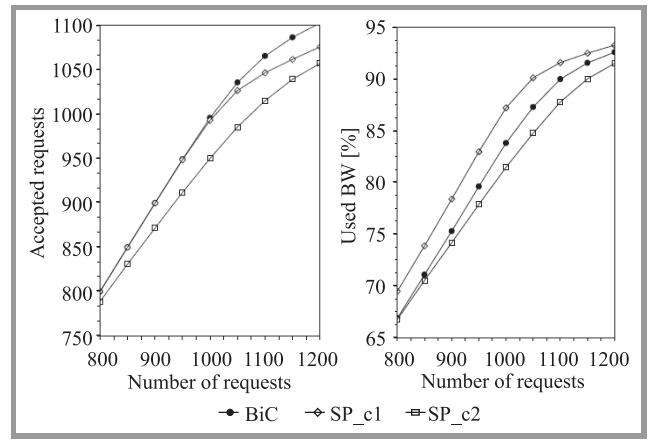


Fig. 12. Accepted requests versus used bandwidth – COST 266TT network.

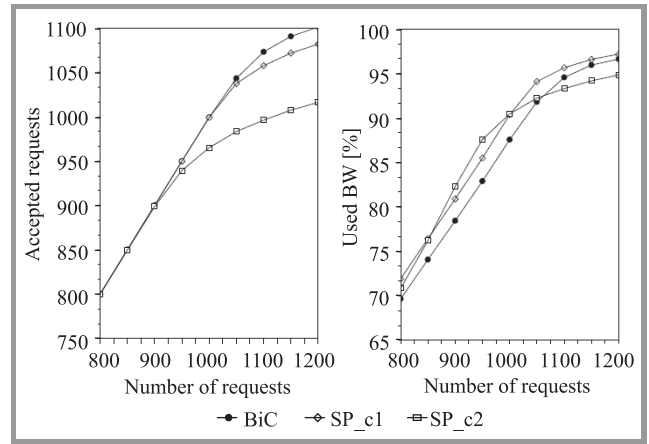


Fig. 13. Accepted requests versus used bandwidth – KL network.

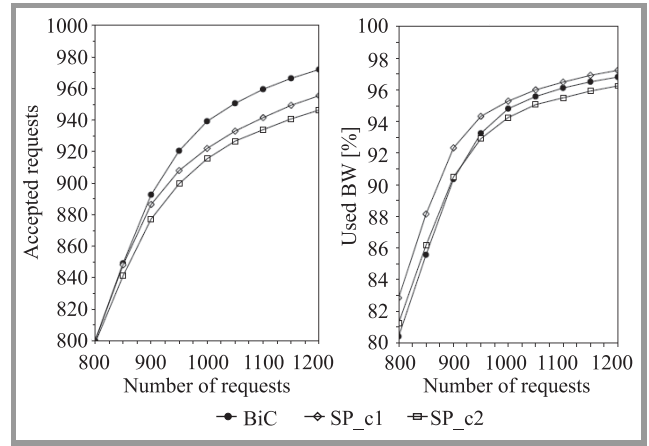


Fig. 14. Accepted requests versus used bandwidth – ISP network.

The results in the other networks exhibit the same behavior. In the COST 266BT network, BiC also has lower blocking than SP_c1 and SP_c2 (more accepted requests) but always uses an amount of bandwidth smaller than SP_c1 (see Fig. 11). For moderated traffic loads (until 950 connection requests), BiC even uses less bandwidth

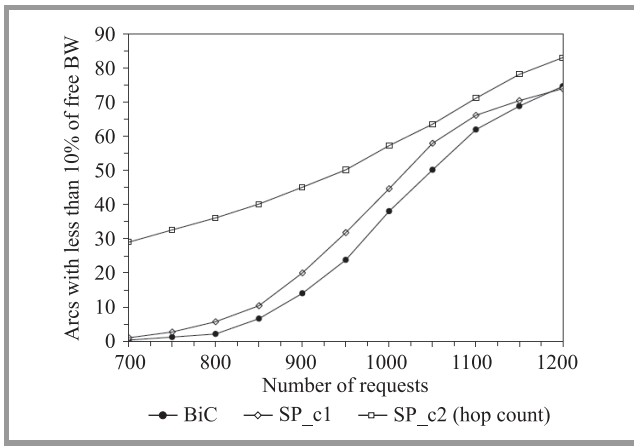


Fig. 15. Arcs with less than 10% of free bandwidth – COST 266TT network

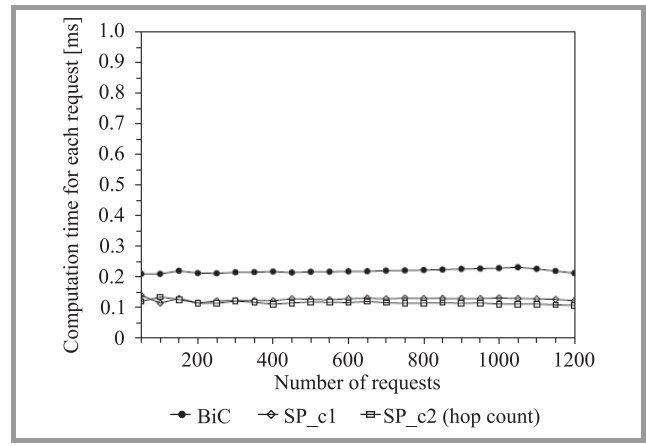


Fig. 18. Computation time for each request – NSFNET network.

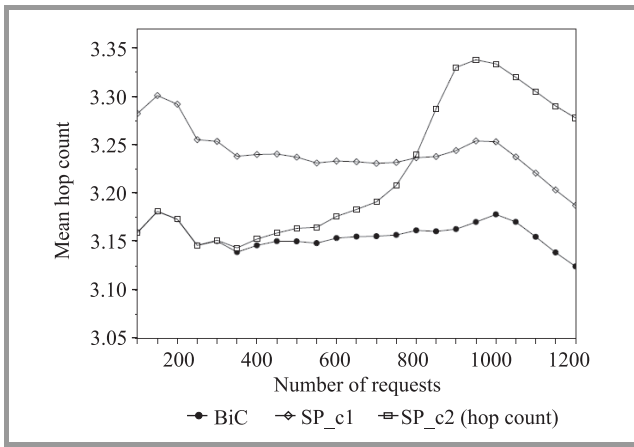


Fig. 16. Mean hop count – COST 266BT network.

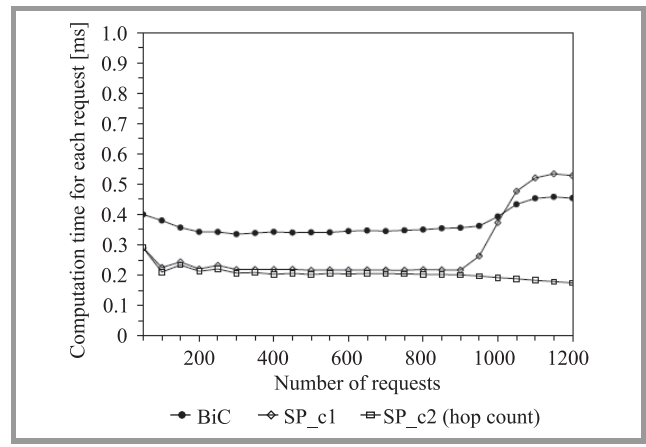


Fig. 19. Computation time for each request – COST 266BT network.

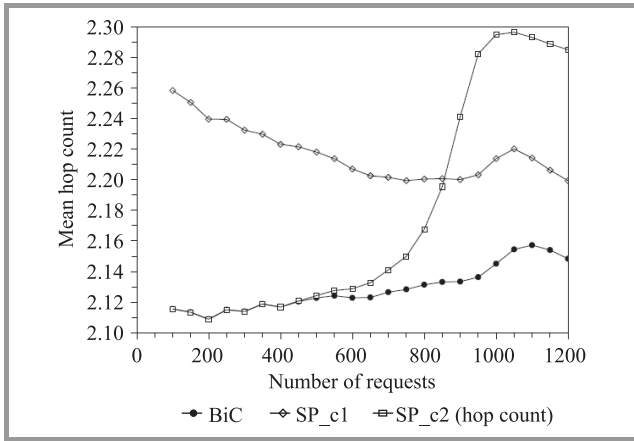


Fig. 17. Mean hop count – KL network.

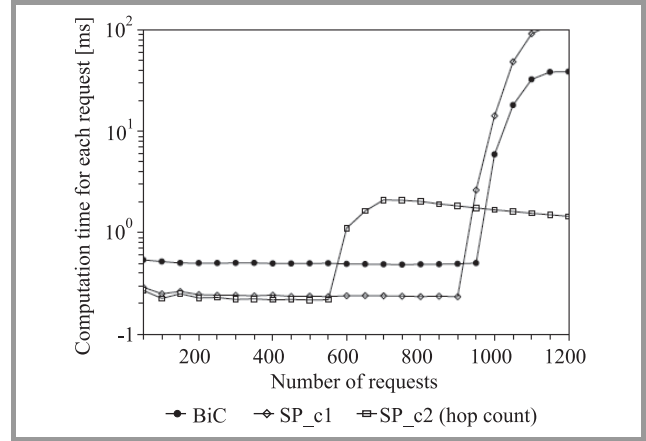


Fig. 20. Computation time for each request – COST 266TT network.

than SP_c2, despite allowing the establishment of many more lightpaths.

In Figs. 12, 13, and 14 the same performance measures are shown for COST 266TT, KL, and ISP networks, respectively. The superior performance of the BiC model is consistent in all tested networks.

It is also interesting to analyze the ability of the different approaches to distribute traffic over the network. Figure 15 plots the number of links in the COST 266TT network with less than 10% of free bandwidth. BiC leads to a lower number of links with less than 10% free bandwidth than SP_c1 and SP_c2. Knowing that the number

of accepted request is also higher, we can conclude that BiC has a performance significantly better than the single objective counterparts. The same behavior was observed in the remaining simulated networks (although not shown here).

Another interesting network performance measure is the average number of hops per established lightpath. As it can be seen in Figs. 10, 16, and 17 (for NSFNET, COST 266BT and KL networks, respectively), BiC uses in average a smaller number of links in all ranges of traffic loads. This happens because it takes advantage of the characteristics of the two metrics. For light traffic, the BiC model chooses shorter connections and, in fact, achieves paths as short as SP_c2. But, unlike SP_c2, BiC is concerned with the load already present in the network links. On the other hand, SP_c1 does not take into account the hop count, leading to longer paths, even when the network is nearly “empty”. As the traffic load increases, the worst choice of the initial paths in SP_c2 leads to bottlenecks in some links. This results in the selection of longer paths, and in higher blocking probability. Above approximately 800 requests, the average number of hops per lightpath in SP_c2 is even greater than in SP_c1 – the traffic distribution is more effective in this model. Also note that when the number of connection requests exceeds approximately 1000, the mean hop count decreases in the three approaches. With this traffic load the network is already congested and node pairs topologically distant are experiencing greater difficulties in establishing a successful connection. So only some “short” connection requests obtain a service, lowering the mean hop count.

Regarding CPU time, the BiC approach requires more CPU, as would be expected, but CPU times are still very low, not exceeding 0.25 ms in NSFNET, 0.45 ms in COST 266BT, 0.3 ms in KL network, and 0.35 ms in ISP network. In the COST 266TT network CPU time remains under 0.6 ms until 950 connection requests. Figures 18, 19, and 20 show the CPU time per connection request for NSFNET, COST 266BT and COST 266TT networks, respectively. The CPU times remain stable as the traffic load grows in the NSFNET, ISP, KL and COST 266BT networks. In COST 266TT network (see Fig. 20), above 950 connection requests CPU time for BiC and SP_c1 approaches increases considerably, and can be as high as 40 ms. This effect occurs when the traffic load is very high and coincides with the starting of visible blocking in the network. Note that this substantial increase in CPU time is even larger in SP_c1.

In order to assess the degree of conflict between the two objective functions used in BiC approach, the number of requests without an optimal solution was calculated. Figures 21 and 22 show the percentage of non-dominated non-optimal solutions in COST 266BT and COST 266TT networks. Although the number of non-dominated solutions is relatively low this does not compromise the interest in using a bicriteria model. In fact many of the ideal solutions of the bicriteria model might possibly have not

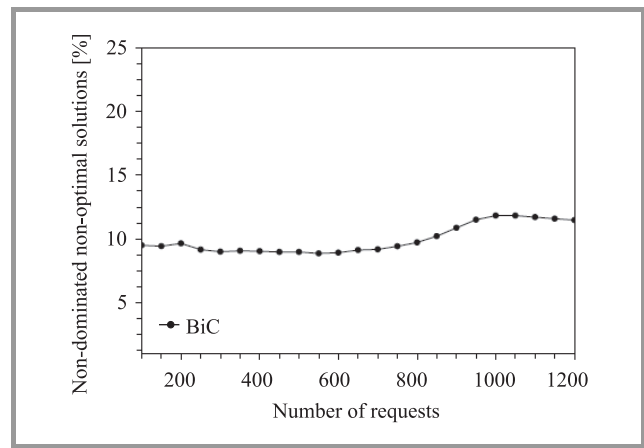


Fig. 21. Non-dominated non-optimal solutions – COST 266BT network.

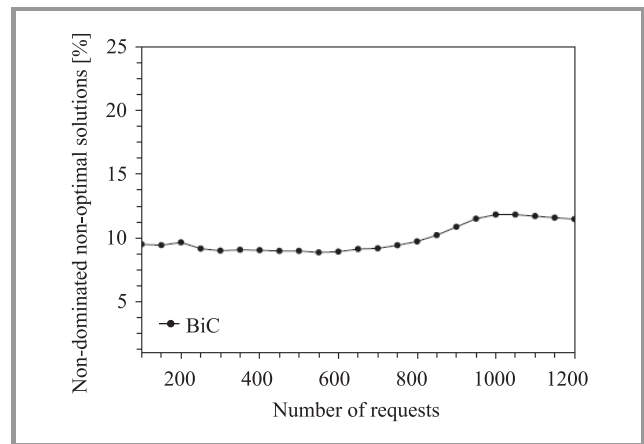


Fig. 22. Non-dominated non-optimal solutions – COST 266TT network.

been found by the single objective models because they correspond to alternative optimal solutions in one of the objective functions.

4. Conclusions

The routing and wavelength assignment problem in WDM networks, as seen from a full traffic engineering perspective, involves multiple metrics, to be optimized, and specific constraints. Therefore multicriteria approaches like the one described in this paper enable to explicitly represent the various performance objectives and to address, in a consistent manner, the trade offs among the various criteria.

A bicriteria model for obtaining a topological path (unidirectional or symmetric bidirectional) for each lightpath request in a WDM network was reviewed. The model considers two criteria – the first one takes into account the bandwidth usage in the links of the network and the second one the number of links of the path. The automated resolution approach uses a k -shortest path algorithm, as well as preference thresholds defined in the objective function’s space,

combined with a Chebyshev distance to a reference point (which changes with the analyzed preference region). Having obtained a non-dominated topological path, a heuristic procedure was then used to assign wavelengths to the links. The performance of this bicriteria model was analyzed using several benchmark networks, and considering a comparison with the results of the two single criterion approaches corresponding to each of the criteria used in the BiC model. Clearly, the BiC approach resulted in lower global blocking than the single criterion models SP_{c1} and SP_{c2}. This is due to an initial better choices of paths and a more balanced distribution of traffic load. At moderate load, although BiC approach accepts more requests, BiC uses less bandwidth than SP_{c1}; SP_{c2} uses less bandwidth than the BiC but it leads to a significant lower number of successful connections.

The impact of having five nodes with wavelength conversion capability was negligible in the simulated situations.

Although the BiC approach uses more CPU time per request its performance was nevertheless quite good – below 0.5 ms except in the denser network (COST 266TT).

In a following paper we will address the network performance analysis issues of an extension of the bicriteria model that provides dedicated path protection, in the event of failures. This is an issue of great importance having in mind the great amount of traffic carried in these optical networks. To provide the necessary network resiliency, while preserving the multicriteria nature of the developed model, this extension (see [28]) enables to obtain a topological pair of node disjoint paths for each connection request. The developed performance analysis study will consider the same type of experimentation and performance measures as in this paper and will present interesting conclusions concerning the potentialities (put in evidence in this paper) and limitations of the use of multicriteria approaches in this context.

Acknowledgements

This study was partially supported by PT Inovação and by programme POSC of the EC programme cosponsored by national funds.

Carlos Simões also thanks the FCT and IPV by the grant of the Ph.D. scholarship SFRH/BD/49739/2009.

References

- [1] X. Chu, J. Liu, and Z. Zhang, "Analysis of sparse-partial wavelength conversion in wavelength-routed WDM networks", in *Proc. IEEE Conf. INFOCOM'2004*, Hong Kong, China, 2004, vol. 2, pp. 1363–1371.
- [2] S. Bandyopadhyay, A. Jaekel, A. Sengupta, and W. Lang, "A virtual wavelength translation scheme for routing in all-optical networks", *Phot. Netw. Commun.*, vol. 4, pp. 391–407, 2002.
- [3] H. Zang, J. P. Jue, and B. Mukherjee, "A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks", *Opt. Netw. Mag.*, vol. 1, pp. 47–60, 2000.
- [4] R. Ramaswami and K. Sivarajan, "Routing and wavelength assignment in all-optical networks", *IEEE/ACM Trans. Netw.*, vol. 3, pp. 489–500, 1995.
- [5] I. Chlamtac, A. Ganz, and G. Karmi, "Lightpath communications: An approach to high bandwidth optical WAN's", *IEEE Trans. Commun.*, vol. 40, pp. 1171–1182, 1992.
- [6] D. Banerjee and B. Mukherjee, "A practical approach for routing and wavelength assignment in large wavelength-routed optical networks", *IEEE J. Sel. Area Commun.*, vol. 14, pp. 903–908, 1996.
- [7] A. E. Ozdaglar and D. P. Bertsekas, "Routing and wavelength assignment in optical networks", *IEEE/ACM Trans. Netw.*, vol. 11, pp. 259–272, 2003.
- [8] B. Jaumard, C. Meyer, and B. Thiongane, "Comparison of ILP formulations for the RWA problem", *Opt. Switch. Netw.*, vol. 4, pp. 157–172, 2007.
- [9] R. M. Krishnaswamy and K. N. Sivarajan, "Algorithms for routing and wavelength assignment based on solutions of LP-relaxations", *IEEE Commun. Lett.*, vol. 5, pp. 435–437, 2001.
- [10] M. Saad and Z.-Q. Luo, "On the routing and wavelength assignment in multifiber WDM networks", *IEEE J. Sel. Area Commun.*, vol. 22, pp. 1708–1717, 2004.
- [11] R. Krishnaswamy and K. Sivarajan, "Design of logical topologies: A linear formulation for wavelength-routed optical networks with no wavelength changers", *IEEE/ACM Trans. Netw.*, vol. 9, pp. 186–198, 2001.
- [12] B. Jaumard, C. Meyer, and B. Thiongane, "ILP formulations for the routing and wavelength assignment problem: symmetric systems", in *Handbook of Optimization in Telecommunications*. Springer Science + Business Media, 2006, pp. 637–677.
- [13] J. Clímaco and J. Craveirinha, "Multicriteria analysis in telecommunication planning and design – problems and issues", in *Multiple Criteria Decision Analysis: State of the Art Surveys*, vol. 78 of *International Series in Operations Research & Management Science*. New York: Springer Science, 2005, pp. 899–951.
- [14] J. Clímaco, J. Craveirinha, and M. Pascoal, "Multicriteria routing models in telecommunication networks – overview and a case study", in *Advances in Multiple Criteria Decision Making and Human Systems Management*. IOS Press, 2007, pp. 17–46.
- [15] A. Todimala and B. Ramamurthy, "Congestion-based algorithms for online routing in optical WDM mesh networks", in *Proc. IASTED Int. Conf. CII'03*, Scottsdale, USA, 2003, pp. 43–48.
- [16] X. Chu and B. Li, "Dynamic routing and wavelength assignment in the presence of wavelength conversion for all-optical networks", *IEEE/ACM Trans. Netw.*, vol. 13, pp. 704–715, 2005.
- [17] C.-F. Hsu, T.-L. Liu, and N.-F. Huang, "On adaptive routing in wavelength-routed networks", *Opt. Netw. Mag.*, vol. 3, pp. 15–24, 2002.
- [18] T. Gomes, J. Craveirinha, J. Clímaco, and C. Simões, "A bicriteria routing model for multi-fibre WDM networks", *Phot. Netw. Commun.*, vol. 18, pp. 287–299, 2009.
- [19] T. Gomes, L. Martins, and J. Craveirinha, "An algorithm for calculating the k shortest paths with a maximum number of arcs", *Investigação Operacional*, vol. 21, no. 2, pp. 235–244, 2001.
- [20] E. Martins, M. Pascoal, and J. Santos, "Deviation algorithms for ranking shortest paths", *Int. J. Found. Comput. Sci.*, vol. 10, no. 3, pp. 247–263, 1999.
- [21] A. P. Wierzbicki, "The use of reference objectives in multiobjective optimization", *Theory and Application, Proceedings, Lecture Notes in Economics and Mathematical Systems*, vol. 177, pp. 468–487, 1980.
- [22] E. Martins, M. Pascoal, and J. Santos, "An algorithm for ranking loopless paths", Tech. Rep. 99/007, CISUC, 1999 [Online]. Available: <http://www.mat.uc.pt/marta/Publicacoes/mps2.ps>
- [23] J. Clímaco, J. Craveirinha, and M. Pascoal, "An automated reference point-like approach for multicriteria shortest path problems", *J. Syst. Sci. Syst. Engin.*, vol. 15, pp. 314–329, 2006.
- [24] A. Betker, C. Gerlach, M. Jäger, M. Barry, S. Bodamer, J. Späth, C. M. Gauger, and M. Köhn, "Reference transport network scenarios MultiTeraNet Report", 2003, [Online]. Available: <http://iospress.metapress.com>

- [25] S. Maesschalck, D. Colle, M. Lievens, P. Demeester, C. Mauz, M. Jaeger, R. Inkret, B. Mikac, and J. Derkacz, "Pan-European optical transport networks: an availability-based comparison", *Phot. Neww. Commun.*, vol. 5, pp. 203–225, 2003.
- [26] M. Kodialam and T. Lakshman, "Dynamic routing of bandwidth guaranteed tunnels with restoration", in *Proc. INFOCOM'2000 Conf.*, Tel-Aviv, Israel, 2000, vol. 2, pp. 902–911.
- [27] G. Apostolopoulos, R. Guérin, S. Kamat, and S. Tripathi, "Quality of service based routing: a performance perspective", in *Proc. ACM SIGCOMM '98 Conf.*, Vancouver, Canada, 1998, pp. 17–28.
- [28] T. Gomes, C. Simões, J. Craveirinha, and J. Clímaco, "A bi-objective model for routing and wavelength assignment in resilient WDM networks", in *Safety, Reliability and Risk Analysis: Theory, Methods and Applications*, vol. 4, pp. 2627–2634, 2008.

Note: this text is reprinted from JTIT 3/2010



Carlos Simões obtained an undergraduate diploma in electrical engineering science- telecommunications and electronics in 1995 and a M.Sc. degree in systems and automation-telecommunications in 1999, both from the University of Coimbra, Portugal. Since 1998, he has been an Assistant at the Polytechnic Institute of Viseu,

Portugal, and a researcher at INESC-Coimbra R&D Institute. He is currently working towards his Ph.D. degree in electrical engineering and informatics at the University of Coimbra and is involved in the research of multicriteria models for routing in optical networks.

e-mail: csimoes@ipv.pt

School of Technology and Management of Viseu
Polytechnic Institute of Viseu
Campus Politecnico de Repeses
P-3504-510 Viseu, Portugal

Institute of Computers and Systems Engineering
of Coimbra (INESC-Coimbra)
Rua Antero de Quental, 199
P-3000-033 Coimbra, Portugal



Teresa Gomes is Assistant Professor in telecommunications at the Department of Electrical Engineering and Computers of the Faculty of Sciences and Technology of the University of Coimbra, Portugal, since 1998, and a researcher at the INESC-Coimbra. She obtained the following degrees: undergraduate diploma in electrical engineering

science (E.E.S.)-informatics at the Coimbra University (1984), M.Sc. in computer science (1989) and Ph.D. in E.E.S.-telecommunications and electronics (1998), both at

the University of Coimbra. Her main present interests are routing, protection and reliability analysis models and algorithms for optical and MPLS networks.

e-mail: teresa@deec.uc.pt

Department of Electrical Engineering and Computers
University of Coimbra
Pinhal de Marrocos
3030-290 Coimbra, Portugal



José Craveirinha is Full Professor in telecommunications at the Department of Electrical Engineering and Computers of the Faculty of Sciences and Technology of the University of Coimbra, Portugal, since 1997. He obtained the following degrees: undergraduate diploma in electrical engineering science (E.E.S.)-telecommunications

and electronics at IST, Lisbon Technical University (1975), M.Sc. (1981) and Ph.D. in E.E.S. at the University of Essex (UK) (1984) and Doct. of Science (Agregado) in E.E.S.-telecommunications at the University of Coimbra (1996). Previous positions were: Associate Professor and Assistant Professor at the FCUT, Coimbra University, telecommunication R&D engineer (at the CET-Portugal Telecom). He coordinated a research group in teletraffic engineering and network planning at the INESC-Coimbra R&D Institute since 1986 and was Director of this Institute during 1994–1999. He is author/co-author of more than 100 scientific and technical publications in teletraffic modeling, reliability analysis, planning and optimization of telecommunication networks. His main present interests are in multicriteria routing and reliability analysis models and algorithms for optical and multiservice-IP/MPLS networks.

e-mail: jcrav@deec.uc.pt

Department of Electrical Engineering and Computers
University of Coimbra
Pinhal de Marrocos
3030-290 Coimbra, Portugal



João Clímaco is Full Professor at the Faculty of Economics of the University of Coimbra and President of the Scientific Committee of the INESC-Coimbra. He obtained the M.Sc. degree in control systems at the Imperial College of Science and Technology, University of London (1978); the "Diploma of Membership of the Imperial College of Science and Technology" (1978); the Ph.D. in optimization and systems theory, Electrical Engineering

Department, University of Coimbra (1982) and the title of “Agregação” at the University of Coimbra (1989). He served in the past as the Vice-President of ALIO- Latin Ibero American OR Association and Vice-President of the Portuguese OR Society. He belongs to the editorial board of the following scientific journals: *Investigação Operacional* (Journal of the Portuguese OR Society), *Journal of Group Decision and Negotiation*, *International Transactions in Operational Research* (ITOR), *ENGEVISTA* and *Rio’s International Journal on Sciences of Industrial and Systems Engineering and Management*. He is also member of the editorial board of the University of Coimbra Press. His current interests of research include multicriteria decision

aiding, multiobjective mathematical programming, location analysis and telecommunication network planning and management.

e-mail: jclimaco@inescc.pt

Faculty of Economics University of Coimbra

Av. Dias da Silva 165

P-3004-512 Coimbra, Portugal

Faculty of Economics

Institute of Computers and Systems Engineering
of Coimbra (INESC-Coimbra)

Rua Antero de Quental, 199

P-3000-033 Coimbra, Portugal

Computational Methods for Two-Level 0-1 Programming Problems through Distributed Genetic Algorithms

Keiichi Niwa, Tomohiro Hayashida, and Masatoshi Sakawa

Abstract—In this paper, we consider a two-level 0-1 programming problem in which there is not coordination between the decision maker (DM) at the upper level and the decision maker at the lower level. We propose a revised computational method that solves problems related to computational methods for obtaining the Stackelberg solution. Specifically, in order to improve the computational accuracy of approximate Stackelberg solutions and shorten the computational time of a computational method implementing a genetic algorithm (GA) proposed by the authors, a distributed genetic algorithm is introduced with respect to the upper level GA, which handles decision variables for the upper level DM. Parallelization of the lower level GA is also performed along with parallelization of the upper level GA. The proposed algorithm is also improved in order to eliminate unnecessary computation during operation of the lower level GA, which handles decision variables for the lower level DM. In order to verify the effectiveness of the proposed method, we propose comparisons with existing methods by performing numerical experiments to verify both the accuracy of the solution and the time required for the computation.

Keywords—distributed genetic algorithm, Stackelberg solution, two-level 0-1 programming problem.

1. Introduction

In the real world, we can often encounter situations that there are multiple decision makers (DMs) in hierarchically structured organizations, and decisions may be taken serially or simultaneously in order to optimize each of the objectives. This kind of problem has been formulated as a two-level programming problem [1]. In two-level programming problems, the upper level DM makes his/her decision first, and then, with full knowledge of the decision of the upper level DM, the lower level DM makes his/her decision in order to optimize his/her own objective function. According to this rule, the upper level DM also makes a decision so as to optimize the objective function of self. The solution defined as the above mentioned procedure is a Stackelberg solution. In this paper, both the upper level and the lower level have one DM, and the problem is treated as a two-level 0-1 programming problem in which both DMs treat all of their decision variables as 0-1 variables.

As an overview of research dealing with two-level programming problems that include discrete variables, Bard *et al.* presented an algorithm based on the branch-and-bound

approach in order to derive the Stackelberg solution for two-level 0-1 programming problems [2] and two-level mixed integer programming problems [3]. Wen *et al.* [4] have presented a computation method for obtaining the Stackelberg solution to two-level programming problems which have 0-1 variables for the decision variables in the upper level and real variables for the decision variables in the lower level.

On the other hand, the adaptive process of systems in the natural world has been explained, and genetic algorithms (GAs) which imitate the evolution occurring in living organisms have been receiving attention at international conferences related to GAs, publications by Goldberg [5], as have methodologies for optimization, adaptation and learning. GAs have also been adopted for a variety of combinatorial optimization problems, and their effectiveness has been reported [6].

An example of research related to two-level programming problems using GAs is given by Anandalingam, *et al.* [7] which presents a method for deriving a Stackelberg solution for two-level linear programming problems. Also, Nishizaki, *et al.* presented an algorithm based on GAs in order to derive the Stackelberg solution for two-level integer programming problems [8] and two-level mixed integer programming problems [9]. In order to derive a Stackelberg solution for 0-1 programming problems related to two-level decentralized systems, the authors [10] have also proposed a computational method that adopts the double string proposed by Sakawa, *et al.* as the individual representation. In order to improve the computational accuracy of approximate Stackelberg solutions, the authors have proposed computational methods that implement sharing [11] and cluster analysis [12] methods. Furthermore, the authors have proposed a computational method using parallel genetic algorithm [13]. Use of these methods allows for the derivation of approximate Stackelberg solutions with relatively high precision and in a relatively short time, but there is still room for improvement, particularly with regards to calculation times.

Therefore, this paper focuses on two-level 0-1 programming problems, and proposes an improved computational method that addresses problems related to the computational method proposed by the authors for deriving the Stackelberg solution. Specifically, a distributed genetic algorithm is introduced with respect to the upper level GA, which handles decision variables for the upper level DM, in order to improve the computational accuracy of approxi-

mate Stackelberg solutions and decrease the computational time of a computational method implementing a genetic algorithm proposed by the authors. Also, parallelization of the lower level GA is performed along with parallelization of the upper level GA. The proposed algorithm is also improved in order to eliminate unnecessary computation during operation of the lower level GA, which handles decision variables for the lower level DM. In order to verify the effectiveness of the proposed method, we propose comparisons with the existing method and the computational method using parallel GA by performing numerical experiments to verify both the accuracy of the solution and the time required for the computation.

2. Two-Level 0-1 Programming Problem

For the sake of brevity, we denote the upper and lower level DMs by DM1 and DM2, respectively. The two-level 0-1 programming problem is expressed as

$$\left. \begin{array}{l} \text{maximize}_{\mathbf{x}} \quad z_1(\mathbf{x}, \mathbf{y}) = \mathbf{c}_1 \mathbf{x} + \mathbf{d}_1 \mathbf{y} \\ \text{where } \mathbf{y} \text{ solves} \\ \text{maximize}_{\mathbf{y}} \quad z_2(\mathbf{x}, \mathbf{y}) = \mathbf{c}_2 \mathbf{x} + \mathbf{d}_2 \mathbf{y} \\ \text{subject to} \quad A\mathbf{x} + B\mathbf{y} \leq \mathbf{b} \\ \mathbf{x} \in \{0, 1\}^{n_1}, \mathbf{y} \in \{0, 1\}^{n_2}, \end{array} \right\} \quad (1)$$

where $\mathbf{x} = (x_1, \dots, x_{n_1})^T$, and $\mathbf{y} = (y_1, \dots, y_{n_2})^T$ are the vectors of decision variables for DM1 and DM2; $z_1(\mathbf{x}, \mathbf{y})$, and $z_2(\mathbf{x}, \mathbf{y})$ respectively represent the objective functions of DM1 and DM2; $\mathbf{c}_1 = (c_{11}, \dots, c_{1n_1})$, $\mathbf{d}_1 = (d_{11}, \dots, d_{1n_2})$, $\mathbf{c}_2 = (c_{21}, \dots, c_{2n_1})$, and $\mathbf{d}_2 = (d_{21}, \dots, d_{2n_2})$ denote the coefficient vectors of the objective functions; A and B are $m \times n_1$ and $m \times n_2$ coefficient matrices in the constraints, respectively; $\mathbf{b} = (b_1, \dots, b_m)^T$ is a coefficient vector of the right hand side of the constraints; the superscript T means transposition of a vector.

For the sake of simplicity, in this paper, it is assumed that each component of A , B , $\mathbf{b}$, $\mathbf{c}_1$, $\mathbf{c}_2$, $\mathbf{d}_1$, and $\mathbf{d}_2$ is positive.

It is possible to express the process for choosing the Stackelberg solution for a two-level 0-1 programming problem in the following manner. Each decision maker completely knows objective functions and constraints of the opponent and self, and DM1 first makes a decision and then DM2 makes a decision in order to maximize the objective function with full knowledge of the decision of DM1. That is to say, when the decision by DM1 is denoted $\hat{\mathbf{x}}$, DM2 solves the 0-1 programming problem (2) with parameters $\hat{\mathbf{x}}$, choosing the optimal solution $\mathbf{y}(\hat{\mathbf{x}})$ as the rational reaction to $\hat{\mathbf{x}}$.

$$\left. \begin{array}{l} \text{maximize}_{\mathbf{y}} \quad z_2(\hat{\mathbf{x}}, \mathbf{y}) = \mathbf{d}_2 \mathbf{y} + \mathbf{c}_2 \hat{\mathbf{x}} \\ \text{subject to} \quad B\mathbf{y} \leq \mathbf{b} - A\hat{\mathbf{x}} \\ \mathbf{y} \in \{0, 1\}^{n_2} \end{array} \right\} \quad (2)$$

Under this premise, DM1 also determines $\mathbf{x}$ by choosing the value which maximizes its own objective function. For

problems which adopt the Stackelberg solution to conceptualize their solution, it is assumed that there is no consensus among DMs that might mutually constrain decisions. Putting it another way, their relationship may be described as non-cooperative.

3. GA Based Computational Method

We propose a computational method through GA in order to obtain Stackelberg solutions to the two-level 0-1 programming problems. In Subsections 3.1, 3.2 and 3.3, first, we describe fundamental elements of GA, which are coding procedure, a decoding procedure and genetic operators, used in the computational method using GA [10]. In this paper, we call the computational method using GA [10] normal GA (NGA). Furthermore, we show additional elements used in the proposed computational method using distributed genetic algorithm and the computational method using parallel genetic algorithm [13] in Subsections 3.4 and 3.5. Finally, the algorithm used in the proposed computational method using distributed genetic algorithm is described in Subsection 3.6.

3.1. Coding and decoding

When solving 0-1 programming problems using GAs, binary strings are usually adopted to express individuals [5], [14]. However, under this representation it is possible that infeasible individuals that do not satisfy the constraints may be generated, so there is a danger that the performance of the GAs may degrade. Thus, in this paper, a double string [6] is used which is composed of the substring corresponding to the decision of DM1, $\mathbf{x}$, and the substring corresponding to the decision of DM2, $\mathbf{y}$, as shown in Fig.1 in order to derive only feasible solutions. The decisions of DM1 and DM2 are handled by performing genetic operators on each sub-individual. We call the GA operating to the decision $\mathbf{x}$ of DM1 the upper level GA, and the GA operating to the decision $\mathbf{y}$ of DM2 the lower level GA.

← Individual for $\mathbf{x}$ →			← Individual for $\mathbf{y}$ →		
$i_x(1)$	...	$i_x(n_1)$	$i_y(1)$	...	$i_y(n_2)$
$S_{i_x(1)}$	...	$S_{i_x(n_1)}$	$S_{i_y(1)}$	...	$S_{i_y(n_2)}$

Fig. 1. Double string.

In Fig. 1 $s_{i_x(m)} \in \{0, 1\}$, $i_x(m) \in \{1, \dots, n_1\}$, and for $m \neq m'$ it is assumed that $i_x(m) \neq i_x(m')$. Similarly, $s_{i_y(m)} \in \{0, 1\}$, $i_y(m) \in \{1, \dots, n_2\}$, and for $m \neq m'$ it is assumed that $i_y(m) \neq i_y(m')$. Also, in the double string, $i_x(m)$, $i_y(m)$ and $S_{i_x(m)}$, $S_{i_y(m)}$ express indexes of the elements of each solution vector respectively, and their values.

In order to generate only feasible solutions, a decoding algorithm proposed by the authors [10] is also applied to the upper level and the lower level GA.

3.2. Reproduction

We describe the reproduction operator of the lower level GA. Substituting the given value $\mathbf{x}$ of the decision variable in the upper level GA and the value of $\mathbf{y}$ obtained by decoding individuals in the lower level GA into the objective function of DM2, $z_2(\mathbf{x}, \mathbf{y})$, the value of the evaluation function for each individual is obtained. Next, the fitness value for each individual is derived using linear scaling, and the individuals remaining in the next generation are determined by applying elitist expected value selection.

We describe the reproduction operator of the upper level GA. Substituting the value of $\mathbf{x}$ obtained by decoding individual in the upper level GA and the value of the rational reaction $\mathbf{y}(\mathbf{x})$ obtained by applying the lower level GA into the objective function of DM1, $z_1(\mathbf{x}, \mathbf{y}(\mathbf{x}))$, the value of the evaluation function for each individual is obtained. Next, the fitness value for each individual is calculated by applying linear scaling and adopting a clustering method. The individuals remaining in the next generation are determined by applying elitist expected value selection based on these fitness values.

3.3. Crossover and Mutation

For double strings, if single-point or multi-point crossover operators are performed then there is a possibility that infeasible individuals may be generated because the indexes occurring in the offspring, $i_x(m)$, $i_x(m')$, $m \neq m'$ or $i_y(m)$, $i_y(m')$, $m \neq m'$, may have the same number. When solving the traveling salesman problem or the scheduling problem through GAs, this kind of violation occurs. In order to circumvent such violation, partially matched crossovers (PMX) have been devised. In this paper, a modified version of PMX is used in order to handle the double strings proposed by Sakawa *et al.* [6]. Also, when determining whether or not to apply the crossover operator, a probability p_c is used. Its value is set in advance.

PMX procedure

- Step 1:* For two individuals expressed using double strings, s_1 and s_2 , two crossover points are set at random.
- Step 2:* According to PMX, the upper strings of s_1 and s_2 , along with the corresponding lower strings are reordered, generating s'_1 and s'_2 .
- Step 3:* For double strings, the offsprings, s''_1 and s''_2 , resulting from the application of the revised PMX are obtained by exchanging the lower strings between the two crossover points s'_1 and s'_2 .

It is well recognized that the mutation operator plays a role of local random search in genetic algorithms. In this paper,

the mutation operator is applied to each string, and inversion is used for index strings. For binary strings, mutation of bit-reverse type is adopted. When applying the mutation operator to individuals, it is first determined whether or not the mutation operator will be applied to an individual according to the mutation probability p_m . In the case that mutation is applied, it is then determined whether to apply inversion or bit-reverse according to the mutation selection constant M_{Pum} .

Mutation procedure

- Step 1:* For an individual s , expressed using a double string, a random number r_m is generated. If $r_m \leq M_{Pum}$, a point on the 0-1 string is chosen at random and bit-reverse is performed, yielding s'_1 . Otherwise, Step 2 is adopted.
- Step 2:* Two points on the index string are chosen at random, and inversion is applied to the substring between the two points, yielding s'_2 .

3.4. Application of the parallel genetic algorithm

In genetic algorithms, it is possible to perform parallel processing in the greater part of the operations included in the algorithm. In reproduction operations, however, because it is necessary to calculate evaluation values for each individual in a population, and based on that value determine the fitness of each individual, direct application of parallel processing is difficult. Research related to the parallelization of GAs started with improvements to such barriers to the implementation of parallelization, and a variety of types of models have been proposed and their effectiveness noted by numerous researchers [15], [16], [17]. Today, GAs that implement parallel processing have come to be called parallel genetic algorithms.

The computational method proposed by the authors [13] divides the lower level GA operations and assigns them across multiple processors. Also, the computational method adopts the single-population master-slave GAs as the upper level GA. By assigning the calculation of individual fitness values, crossover operator and mutation operator to multiple processors, calculation times are reduced. We call the computational method proposed by the authors the computational method using parallel GA (PGA). However, while the computational method using parallel GA succeeds in obtaining good approximate solutions and reducing the amount of computational time, there is still likely much more room for improvement. In this study, therefore, we aim for further improvements of the precision of approximate solutions and further reductions in computational time, and consider parallelization of the upper level GA and the lower level GA implemented by the computational method using parallel GA.

The multiple-population genetic algorithms performs parallel processing by dividing the population and assigning the partial populations (sub-populations) to multiple processors. If we adopt this model as our computational method,

it is possible to use the computational method using GA proposed by authors [12] without a lot of modifications. In that case, the multiple-population genetic algorithm is adopted as the upper level GA. The upper level GA operations proposed by authors are applied for partial population assigned to each processor. Additionally, migration operator is performed in every migration interval. Also, it is able to divide the lower level GA operations and assign them across multiple processor in the same way as PGA. In this paper, we employ a multiple-population genetic algorithms (distributed genetic algorithms) so as to obtain good approximate Stackelberg solutions and reduce calculation times. We call the proposed computational method using distributed GA (DGA).

3.5. Lower Level GA Avoidance Procedures

In this paper we introduce a storage region as shown in Fig. 2.

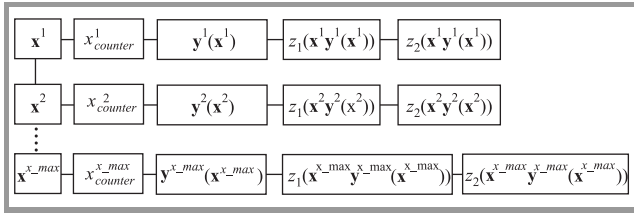


Fig. 2. Storage for saving $\mathbf{x}$ and $\mathbf{y}(\mathbf{x})$.

Here, $\mathbf{x}^i, i = 1, \dots, x_max$ indicate those values of $\mathbf{x}$ that were used in the past for handling individuals of the upper level GA, and $\mathbf{y}^i(\mathbf{x}^i), i = 1, \dots, x_max$ indicate the values of the rational reactions saved with $\mathbf{x}^i$ obtained by the lower level GA. $x_counter \in \{1, 2, \dots, y_max\}, i = 1, \dots, x_max$ indicate the number of times that the lower level GA was used to find the rational reaction $\mathbf{y}^i(\mathbf{x}^i)$ for $\mathbf{x}^i$. x_max indicates the maximum number of DM1 decisions $\mathbf{x}$ saved, and y_max indicates the maximum number of times that the lower level GA can be repeatedly used to find the rational reaction $\mathbf{y}^i(\mathbf{x}^i)$ for $\mathbf{x}^i$. $z_1(\mathbf{x}^i, \mathbf{y}^i(\mathbf{x}^i))$ and $z_2(\mathbf{x}^i, \mathbf{y}^i(\mathbf{x}^i))$ are stored $\mathbf{x}^i, \mathbf{y}^i(\mathbf{x}^i)$ values used in place of DM1 and DM2 objective functions. By using the following algorithm, the number of applications of the lower level GA is reduced, and unnecessary calculation times eliminated.

Storage of the rational reaction $\mathbf{y}(\mathbf{x})$ and lower level GA avoidance procedures

Step 1: If there exists in $\mathbf{x}^i$ an upper level GA individual $\bar{\mathbf{x}}$, proceed to Step 2. If one does not exist, then check if the number of $\mathbf{x}^i$ has reached x_max , and if so continue on to Step 3. If not, proceed to Step 4.

Step 2: If $x_counter$ has reached y_max , then the saved $\mathbf{y}^i(\mathbf{x}^i)$ is returned to the upper level GA as the rational reaction and the algorithm terminates. If not reached, proceed to Step 4.

Step 3: Select the least of the values $z_1(\mathbf{x}^i, \mathbf{y}^i(\mathbf{x}^i))$ from the saved $\mathbf{x}^i$, and take that $\mathbf{x}^i$ value as $\mathbf{x}^k$. After applying the lower level GA and thus obtaining the rational reaction $\mathbf{y}(\bar{\mathbf{x}})$ for $\bar{\mathbf{x}}$, if $z_1(\mathbf{x}^k, \mathbf{y}^k(\mathbf{x}^k)) \leq z_1(\bar{\mathbf{x}}, \mathbf{y}(\bar{\mathbf{x}}))$, save $\bar{\mathbf{x}}, \mathbf{y}(\bar{\mathbf{x}}), z_1(\bar{\mathbf{x}}, \mathbf{y}(\bar{\mathbf{x}})), z_2(\bar{\mathbf{x}}, \mathbf{y}(\bar{\mathbf{x}}))$ in the storage region $\mathbf{x}^k$, and terminate the algorithm.

Step 4: After obtaining the rational reaction $\mathbf{y}(\bar{\mathbf{x}})$ for $\bar{\mathbf{x}}$ by applying the lower level GA, save $\bar{\mathbf{x}}, \mathbf{y}(\bar{\mathbf{x}}), z_1(\bar{\mathbf{x}}, \mathbf{y}(\bar{\mathbf{x}})), z_2(\bar{\mathbf{x}}, \mathbf{y}(\bar{\mathbf{x}}))$, and terminate the algorithm.

Implementation of the algorithm described above improved upon previous methods.

3.6. The Algorithm for the Improved Computational Method

The algorithm used in the computational method after improvement can be described as follows, N_p denotes the number of processors.

Step 1: For each processor $q, q = 1, \dots, N_p$, apply the upper level GA operations on Step1 through Step 7. Taking the generation of the upper level GA as $t_{uq} := 0$, N_u initial individuals are randomly generated.

Step 2: For each individual $\mathbf{x}$ in the upper level GA, determine whether or not to apply the lower level GA, and find the number of lower level GA to apply, N_{ul} . For those individuals N_{ul} , apply the lower level GA operations in Step 2-1 through Step 2-3, and obtain the rational reaction $\mathbf{y}(\mathbf{x})$. For those $N_u - N_{ul}$ individuals to which the lower level GA will not be applied, take the saved $\mathbf{y}(\mathbf{x})$ as the rational reaction, and proceed to Step 4.

Step 2-1: Set $t_l := 0$. Randomly generate N_l lower level GA individuals $\mathbf{y}$, and take these as the initial population of the lower level GA. Proceed to Step 2-2.

Step 2-2: By using $\mathbf{x}$ given as the upper level GA individual and $\mathbf{y}$ generated by the lower level GA, the DM2 objective function value is calculated. After applying linear scaling to the value, the reproduction operator is applied. Proceed to Step 2-3.

Step 2-3: If t_l has exceeded the previously defined a maximum number of generation M_l , take the individual with the best fitness value as the optimal individual $\mathbf{y}(\mathbf{x})$, and proceed to Step 3. Otherwise, apply crossover operator and mutation operator to each lower level GA individual, let $t_l = t_l + 1$, and proceed to Step 2-2.

Step 3: By using the lower level rational reaction $\mathbf{y}(\mathbf{x})$ obtained by operation of the lower level GA and the individual $\mathbf{x}$ of the upper level GA, calculate the values for the DM1 and DM2 objective functions. Perform the procedures required to save $\mathbf{x}$ and its rational reactions $\mathbf{y}(\mathbf{x})$ to the storage region, and proceed to Step 4.

Step 4: Calculate the DM1 objective function for each upper level GA individual $\mathbf{x}$, and after performing linear scaling, apply the clustering method to measure the level of convergence of the individuals. Depending upon the degree of convergence, calculate the fitness value of each individual. Proceed to Step 5.

Step 5: If t_{uq} has exceeded the previously set a maximum number of generation M_u , then terminate the algorithm. In that case, the individual obtained up to that generation with the best fitness value is taken as the optimal individual $(\mathbf{x}, \mathbf{y})$. Otherwise, proceed to Step 6.

Step 6: Reproduction operator is performed using the fitness values of each individual of the upper level GA. Apply crossover operator and mutation operator to each upper level GA individual, and proceed to Step 7.

Step 7: If $t_{uq} \bmod m_i$ (migration interval) = 0, after performing synchronization between the processors, apply migration. Return to Step 2 with $t_{uq} := t_{uq} + 1$.

4. Numerical Experiments

Numerical experiments are carried out in order to demonstrate the feasibility and the effectiveness of DGA. We apply DGA, PGA, and NGA to twelve types of two-level 0-1 programming problems as shown in Table 1. Each problem has five constraints.

Table 1
Problems used in the numerical experiments

Problem	DM1 variables	DM2 variables	Constraint strength
A	15	15	I (50%)
			II (70%)
			III (90%)
B	20	20	I (50%)
			II (70%)
			III (90%)
C	25	25	I (50%)
			II (70%)
			III (90%)
D	30	30	I (50%)
			II (70%)
			III (90%)

In this case, the elements A , B , $\mathbf{c}_1$, $\mathbf{c}_2$, $\mathbf{d}_1$, and $\mathbf{d}_2$ of the two-level 0-1 programming problem are selected at random from the closed interval $[10, 99]$, and the b_i element of $\mathbf{b}$ is set according to the equation

$$b_i = r_i \left(\sum_{j=1}^{n_1} a_{ij} + \sum_{k=1}^{n_2} b_{ik} \right), i = 1, \dots, m. \quad (3)$$

Furthermore, a_{ij} represents the ij element of matrix A , and b_{ik} represents the ik element of matrix B . Here, r_i represents the strength of the constraint. In the strong constraint problem (I), a random number is determined at random from the closed interval $[0.45, 0.55]$, in the middle constraint problem (II), a random number is selected at random from the closed interval $[0.65, 0.75]$, and in the weak constraint problem (III), a random number is determined at random from the closed interval $[0.85, 0.95]$, respectively. The decimal portion of the b_i value is rounded off, and the result is stored as an integer value.

Next, the GA parameters are set for NGA, PGA, and DGA as follows. First, there are parameters that are used in common by all three methods, namely the population sizes of the upper level GA and the lower level GA, the crossover rate, the mutation rate, and the maximum number of generation, and those values are set as 120 for the population size, 0.9 for the crossover rate, 0.02 for the mutation rate, and 300 for the maximum number of generation. Next, for PGA and DGA, α is set to 0.25. The initial number of clusters, k , is set to 5, and d_{max} and d_{min} , used to measure the distance between individuals, are set to 2.5 and 1.0, respectively. In addition to these parameters, x_{max} and y_{max} are set to 100 and 5, respectively. The number of processors is set to 3. Finally, for DGA, we apply the random ring model to the communication topology and adopt Best-Hole model as the selection of emigrants and immigrants. The number of migration interval m_i is set to 15. The migration rate is set to 2.5%.

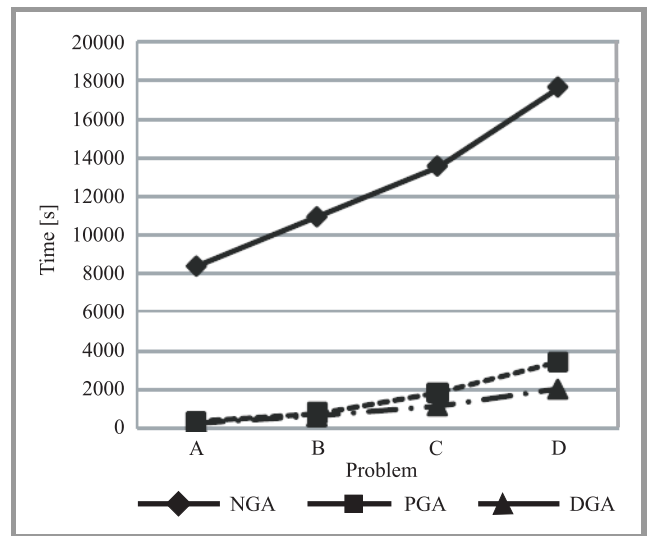


Fig. 3. Comparison of calculation times (the strong constraint problem (I)).

Table 2
Comparison of solution precisions

Problem	Constraint strength	DGA				PGA			
		best	worst	average	variance	best	worst	average	variance
A	I	1078	1078	1078.0	0.00	1078	1078	1078.0	0.00
	II	1377	1377	1377.0	0.00	1377	1377	1377.0	0.00
	III	1727	1727	1727.0	0.00	1727	1727	1727.0	0.00
B	I	1342	1342	1342.0	0.00	1342	1342	1342.0	0.00
	II	1764	1764	1764.0	0.00	1764	1764	1764.0	0.00
	III	2157	2157	2157.0	0.00	2157	2157	2157.0	0.00
C	I	1713	1713	1713.0	0.00	1713	1713	1713.0	0.00
	II	2221	2207	2214.6	43.84	2221	2207	2215.7	20.81
	III	2680	2680	2680.0	0.00	2680	2680	2680.0	0.00
D	I	1864	1857	1859.8	11.76	1864	1851	1856.8	12.36
	II	2480	2444	2469.6	253.44	2480	2439	2449.7	233.01
	III	2930	2930	2930.0	0.00	2930	2930	2930.0	0.00
Problem	Constraint strength	NGA				Enumeration			
		best	worst	average	variance				
A	I	1078	1068	1077.0	9.00	1078			
	II	1377	1377	1377.0	0.00	1377			
	III	1727	1727	1727.0	0.00	1727			
B	I	1342	1328	1340.4	17.24	-			
	II	1764	1754	1763.0	9.00	-			
	III	2157	2157	2157.0	0.00	-			
C	I	1713	1713	1713.0	0.00	-			
	II	2221	2207	2210.7	23.81	-			
	III	2680	2672	2678.5	9.05	-			
D	I	1864	1849	1856.9	11.29	-			
	II	2480	2444	2470.0	80.80	-			
	III	2930	2915	2927.0	20.00	-			

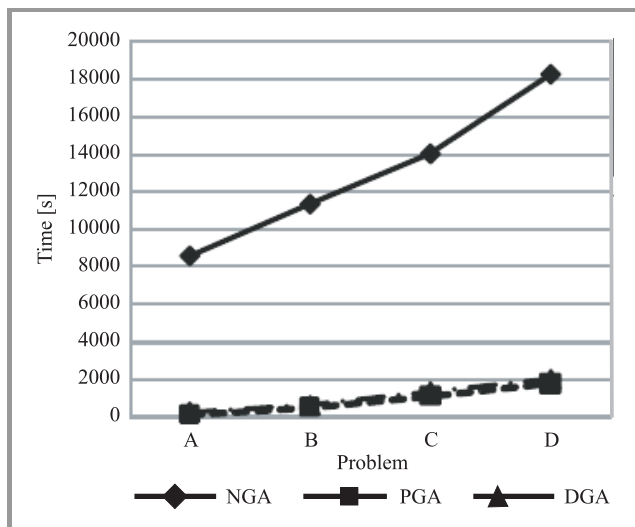


Fig. 4. Comparison of calculation times (the middle constraint problem (II)).

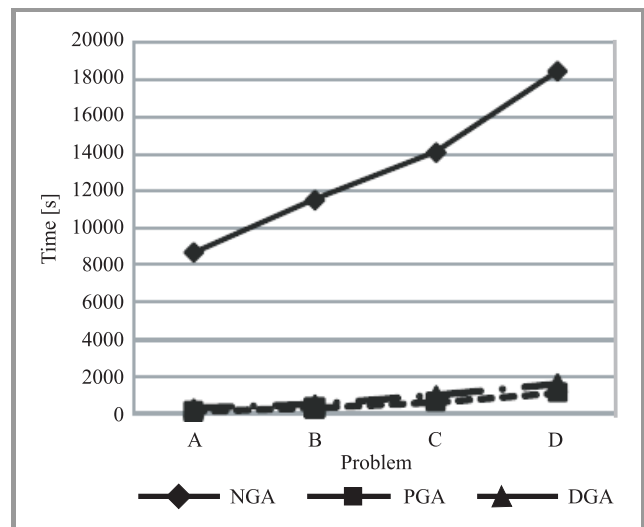


Fig. 5. Comparison of calculation times (the weak constraint problem (III)).

Next we will describe the experimental environment. The experiment is run on a personal computer with a 2.80 GHz CPU and running Windows XP. The compiler used is Microsoft Visual C++ 6.0.

For each problem, NGA, PGA, and DGA are run ten times each. The results are given in Table 2 through Table 5. In Table 2, for all trials of Problem A, DGA, PGA, and NGA derive precise Stackelberg solutions. But, for the re-

Table 3
Comparison of calculation times

Problem	Constraint strength	DGA				PGA			
		best	worst	average	variance	best	worst	average	variance
A	I	220.78	351.25	269.82	1138.83	213.64	434.43	319.70	3611.94
	II	247.30	320.80	275.36	482.07	108.05	166.72	128.41	316.94
	III	215.46	281.27	253.94	427.14	62.20	99.46	82.38	191.60
B	I	394.09	801.84	599.33	12700.54	599.11	856.95	770.43	4550.59
	II	464.08	772.94	610.84	7149.91	419.83	569.50	504.90	2619.66
	III	345.83	633.73	473.91	5558.42	248.95	360.14	293.30	1186.14
C	I	975.14	1326.10	1149.18	10536.46	1728.47	1913.19	1824.04	4479.17
	II	1092.68	1569.18	1337.10	25015.18	1067.88	1199.30	1116.03	1888.83
	III	772.44	1191.86	965.45	16951.74	537.94	665.39	605.52	1484.93
D	I	1640.06	2426.27	2000.36	53181.01	3347.20	3547.55	3428.30	3115.52
	II	1568.27	2515.28	1980.26	81371.51	1753.85	1781.79	1767.90	97.18
	III	1012.63	2268.36	1619.37	89178.03	1073.29	1169.15	1125.58	848.64
Problem	Constraint strength	NGA				Enumeration			
		best	worst	average	variance				
A	I	8379.75	8613.52	8414.59	4578.06	566.73			
	II	8550.34	8668.06	8618.90	842.89	652.38			
	III	8708.25	8714.92	8711.78	3.91	665.81			
B	I	10882.38	11006.64	10954.14	1946.30	–			
	II	11296.80	11467.94	11360.96	2888.93	–			
	III	11447.75	11907.14	11520.65	17290.82	–			
C	I	13467.95	13651.13	13573.35	1886.21	–			
	II	14022.60	14067.24	14032.10	147.58	–			
	III	14060.64	14090.92	14081.75	80.04	–			
D	I	17532.97	17794.47	17646.13	4659.97	–			
	II	18208.95	18263.34	18250.02	235.16	–			
	III	18424.03	18541.39	18461.96	1912.68	–			

Table 4
Lower level GA avoidance counts

Problem	Constraint strength	DGA				PGA			
		best	worst	average	variance	best	worst	average	variance
A	I	35558	34898	35283.7	46901.01	32008	27418	29742.4	1589311.44
	II	35539	35093	35291.8	24252.76	31094	26980	29362.6	1272588.24
	III	35634	35133	35371.4	24639.24	32006	29261	30431.2	948838.16
B	I	34983	33060	33995.6	265319.84	21852	14956	17355.0	3164457.60
	II	34757	32693	33959.8	317217.36	20845	14153	17304.8	4949162.76
	III	35274	34007	34667.5	121787.45	23258	17347	20721.5	3514280.45
C	I	32740	30118	31504.7	518636.01	8626	5577	6963.4	1213648.44
	II	32705	30017	31094.1	677375.69	5164	1599	3707.9	1823801.69
	III	33408	31591	32853.0	390069.20	16092	10170	13038.9	3552493.09
D	I	30696	26108	27780.8	1576362.56	1205	83	490.9	110786.89
	II	30819	26494	28875.6	2235743.84	847	114	292.5	46420.85
	III	32731	27108	30240.6	1688052.64	5939	3469	4376.3	514510.01

sults of Problem A-I, NGA is not possible to obtain precise Stackelberg solutions in some trials. Performance in deriving the best solutions to Problem B is equal for all three methods, but comparing the worst values and the average values shows that DGA and PGA present the best performance. Observe that the best values and the worst values obtained by DGA for Problem C and Problem D are

equal or superior to the corresponding values obtained by NGA. Also, for Problem C and Problem D without Problem D-II, comparing the average values shows that DGA is superior to NGA. Finally, for Problem C and Problem D, DGA is superior or approximately equivalent to PGA in all results. When comparing the three methods with regards to the solution precision obtained, DGA is superior.

Table 5
Comparison of generations in which the best solution was found

Problem	Constraint strength	DGA				PGA			
		best	worst	average	variance	best	worst	average	variance
A	I	5	147	37.8	2210.96	3	42	13.9	130.09
	II	6	20	11.7	16.81	2	13	9.0	8.80
	III	5	20	11.0	20.80	3	16	9.3	18.81
B	I	5	158	54.4	1845.04	13	147	62.2	2204.76
	II	9	53	16.7	155.01	7	22	16.2	27.56
	III	9	29	16.6	35.24	9	30	18.9	35.69
C	I	15	71	23.0	258.60	15	59	38.3	180.01
	II	14	165	61.1	3067.89	50	292	141.1	6644.29
	III	13	247	75.6	6204.84	26	53	40.8	75.76
D	I	12	256	68.4	7015.44	62	291	157.3	5343.41
	II	23	299	117.7	9811.21	36	249	159.9	4262.09
	III	19	70	39.6	275.04	30	239	74.9	3441.69
Problem	Constraint strength	NGA				Enumeration			
		best	worst	average	variance				
A	I	5	208	48.0	6130.80	–			
	II	4	16	8.4	10.64	–			
	III	3	126	22.4	1205.84	–			
B	I	13	133	45.7	1728.41	–			
	II	9	44	16.3	91.61	–			
	III	9	63	19.5	223.25	–			
C	I	11	35	20.5	34.85	–			
	II	16	66	28.2	352.76	–			
	III	13	26	18.8	17.36	–			
D	I	19	108	34.7	666.61	–			
	II	14	103	51.1	721.69	–			
	III	23	284	64.1	5833.29	–			

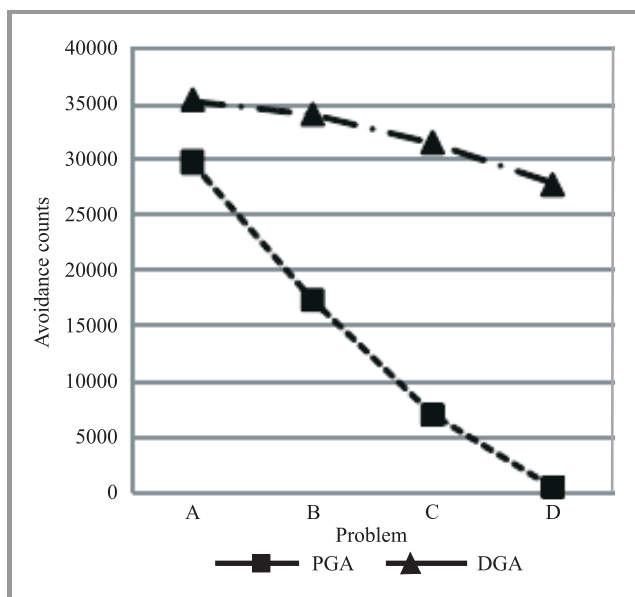


Fig. 6. Comparison of average avoidance counts (the strong constraint problem (I)).

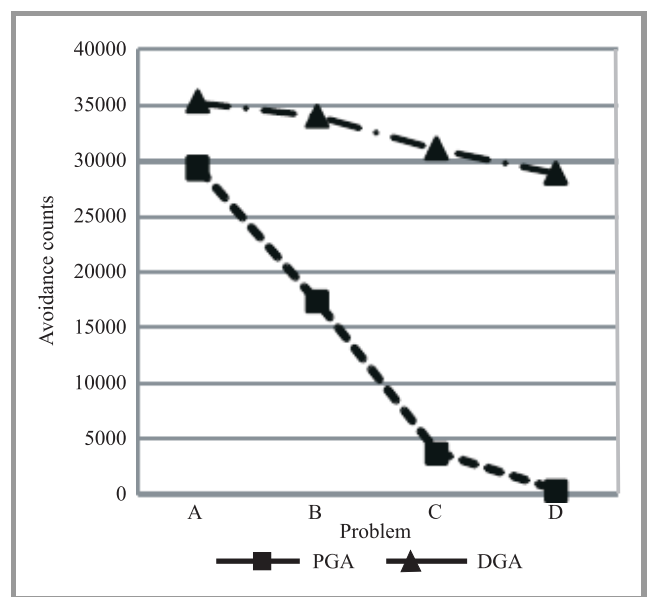


Fig. 7. Comparison of average avoidance counts (the middle constraint problem (II)).

Examining the calculation time results displayed in Table 3, for all trials of all problems, DGA and PGA are superior to NGA. Also, the calculation times of DGA are less than 15% of the calculation times of NGA.

Figure 3 shows the average calculation times for each size of the strong constraint problems (I). We can see that as compared to PGA, DGA is superior.

The average calculation times for each size of the middle constraint problems (II) are illustrated in Fig. 4. When comparing DGA and PGA, DGA is approximately equivalent or slightly inferior.

The average calculation times for each size of the weak constraint problems (III) are shown in Fig. 5. From Fig. 5, DGA is approximately equivalent or slightly inferior to PGA.

The results of the avoidance count for the lower level GA are shown in Table 4. From the results listed in this table, we see that with both DGA and PGA, as the scale of the problem increases the number of lower level GA avoidances is reduced.

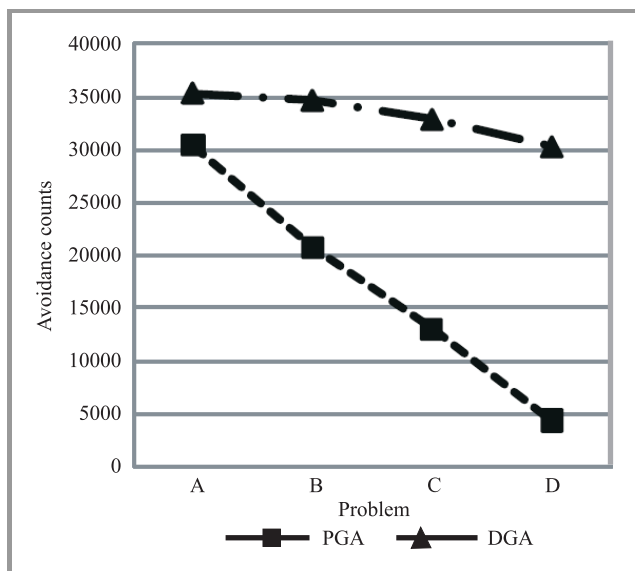


Fig. 8. Comparison of average avoidance counts (the weak constraint problem (III)).

The average avoidance counts for each size of the strong constraint problems (I), the middle constraint problems (II), and the weak constraint problems (III) are shown in Figs. 6–8. Examining the results displayed in these figures, we see that the avoidance counts of DGA gradually decrease as the size of the problem increases. On the other hand, the avoidance counts of PGA rapidly decrease.

Finally, the results of the comparison of generations in which the best solution was found are illustrated in Table 5. The results listed in Table 5 show that with both DGA and PGA, the generation deriving the best result is even later than with NGA, and so it is likely that by introducing cluster analysis methods the diversity of individuals within the population of the upper level GA would be maintained, and rapid population convergence avoided.

From the above results, compared to the other two computational methods, DGA is the superior computational method, both from a standpoint of solution precision and required calculation time.

5. Conclusion

This paper has focused on a two-level 0-1 programming problem in which there is not coordination between the decision maker at the upper level and the decision maker at the lower level. The authors have proposed a modified computational method that solves problems related to computational methods for obtaining the Stackelberg solution. Specifically, in order to improve the computational accuracy of approximate Stackelberg solutions and shorten the computational time of a computational method implementing GA proposed by the authors, a distributed genetic algorithm has been introduced with respect to the upper level GA, which handles decision variables for the upper level DM. Also, parallelization of the lower level GA has been performed along with parallelization of the upper level GA. The proposed algorithm has been improved in order to eliminate unnecessary computation during operation of the lower level GA, which handles decision variables for the lower level DM. In order to verify the effectiveness of the proposed method, numerical experiments have been carried out. From the results, we have shown that the proposed method is the superior to the other two computational methods, both from a standpoint of solution precision and required calculation time.

References

- [1] K. Shimizu, Y. Ishizuka, and J. F. Bard, *Nondifferentiable and Two-Level Mathematical Programming*. Norwell: Kluwer, 1997.
- [2] J. Bard and J. Moore, "An algorithm for the discrete bilevel programming problem", *Nav. Res. Log.*, vol. 39, pp. 419–435, 1992.
- [3] J. Bard and J. Moore, "The mixed integer linear bilevel programming problem", *Oper. Res.*, vol. 38, pp. 911–921, 1990.
- [4] W. P. Wen and Y. H. Yang, "Algorithms for solving the mixed integer two-level linear programming problem", *Comput. Oper. Res.*, vol. 17, pp. 133–142, 1990.
- [5] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Norwell: Addison Wesley, 1989.
- [6] M. Sakawa and M. Tanaka, *Genetic Algorithms*. Tokyo: Asakura Publishing, 1995.
- [7] G. Anandalingam, R. Mathieu, C. L. Pittard, and N. Sinha, "Artificial intelligence based approaches for solving hierarchical optimization problems," in *Impacts of Recent Computer Advances on Operations Research*, R. Sharda, B. L. Golden, E. Wasil, O. Balci and W. Stewart, Eds. New York: Elsevier Science Publishing, 1989, pp. 289–301.
- [8] I. Nishizaki and M. Sakawa, "Computational methods through genetic algorithms for obtaining Stackelberg solutions to two-level integer programming problems", *Cybernet. Syst. Int. J.*, vol. 36, pp. 565–579, 2005.
- [9] I. Nishizaki and M. Sakawa, "Computational methods through genetic algorithms for obtaining Stackelberg solutions to two-level mixed zero-one programming problems", *Cybernet. Syst. Int. J.*, vol. 31, pp. 203–221, 2000.
- [10] K. Niwa, I. Nishizaki, and M. Sakawa, "Decentralized two-level 0-1 programming through genetic algorithms with double strings", in *Proc. 2nd Int. Conf. Knowl. Int. Electron. Syst.*, 1998, vol. 2, pp. 278–284.

- [11] K. Niwa, "Revised computational methods for using genetic algorithms for obtaining Stackelberg solutions to two-level 0-1 programming problems", in *Essays and Studies in Commemoration of the 40th Anniversary of the Founding of Hiroshima University of Economics*, Hiroshima University Economics, 2007, pp. 771–794 (in Japanese).
- [12] K. Niwa, I. Nishizaki, and M. Sakawa, "Two-Level 0-1 Programming Using Genetic Algorithms and a Sharing Scheme Based on Cluster Analysis", in *Proc. Int. MultiConf. Eng. Comput. Sci.*, Hong Kong, China, 2008, pp. 1931–1936.
- [13] K. Niwa, I. Nishizaki, and M. Sakawa, "Two-level 0-1 programming through parallel genetic algorithms", in *Proc. Int. MultiConf. Eng. Comput. Sci.*, Hong Kong, China, 2009, vol. 2, pp. 2000–2005.
- [14] D. E. Goldberg and R. Lingle, "Alleles, loci, and the traveling salesman problem", in *Proc. First Int. Conf. Genet. Algorithms Appl.*, Hillsdale, USA, 1985, pp. 154–159.
- [15] D. Abramson and J. Abela, "A parallel genetic algorithm for solving the school timetabling problem", in *Proc. 15th Austr. Comput. Sci. Conf.*, Hobart, Australia, 1992, vol. 14, pp. 1–11.
- [16] E. Cantú-Paz, *Efficient and Accurate Parallel Genetic Algorithms*. Norwell, Massachusetts: Kluwer, 2000.
- [17] T. C. Fogarty and R. Huang, "Implementing the genetic algorithm on transputer based parallel processing systems", in *Proc. First Int. Conf. Parallel Prob. Solv. Nat.*, Dortmund, Germany, 1990, pp. 145–149.

Note: this text is reprinted from JTIT 2/2010



Keiichi Niwa received the Ph.D. degree in systems engineering from Hiroshima University in 2000. Since 2000, he has jointed Hiroshima University of Economics where he is currently a Associate Professor at Department for Information Systems in Business. His main research interests include meta-heuristics and inte-

ger programming.

e-mail: ki-niwa@hue.ac.jp

Department for Information Systems in Business

Faculty of Economics

Hiroshima University of Economics

Hiroshima, 731-0192, Japan



Tomohiro Hayashida is currently Assistant Professor at Department of Artificial Complex Systems Engineering, Graduate School of Engineering of the Hiroshima University, Japan. His current research interests are agent-based simulation analysis, game theory, and decision making theory.

e-mail: hayashida@hiroshima-u.ac.jp

Department of Artificial Complex Systems Engineering

Graduate School of Engineering

Hiroshima University

Higashi-Hiroshima, 739-8527, Japan



Masatoshi Sakawa received B.Sc., M.Sc., and Ph.D. degrees in applied mathematics and physics at Kyoto University, in 1970, 1972, and 1975, respectively. From 1975 he was with Kobe University where, since 1981, he was an Associate Professor in the Department of Systems Engineering. From 1987 to 1990, he was

a Professor of Department of Computer Science at Iwate University. At present he is a Professor of Hiroshima University, Japan, and is working with the Department of Artificial Complex Systems Engineering, Hiroshima University. His research and teaching activities are in the area of systems engineering, especially, mathematical optimization, multiobjective decision making, fuzzy mathematical programming and game theory.

e-mail: sakawa@hiroshima-u.ac.jp

Department of Artificial Complex Systems Engineering

Graduate School of Engineering

Hiroshima University

Higashi-Hiroshima, 739-8527, Japan

A Framework for Evaluation of Communication Bandwidth Market Models

Wojciech Stańczuk, Piotr Pałka, Józef Lubacz, and Eugeniusz Toczyłowski

Abstract—The article presents a method of analysis of market-based models for resource allocation in communication networks. It consists of several stages: classification of a market model, generation of input data, data adaptation to a tested model, test calculations and, finally, presentation and interpretation of results. A set of general criteria to assess various models has been proposed. Tests are run using dedicated computer applications, data is stored in open XML-based format originated in the multicommodity market model. Network topologies are derived from the SNDlib library.

Keywords—*auctions, bandwidth trading, resource allocation.*

1. Introduction

A dominant form of trading on the market of resources in telecommunications networks is bilateral contracting. Time of negotiations is undesirably long in relation to high dynamics of business processes in the telecommunications market. In addition, the bilateral nature of negotiations reduces the transparency of trade rules. This often enables network operators or service providers with significant market power to obtain better trading conditions than it is justified. Therefore, the research is conducted on innovative mechanisms for trading of transport resources in networks to enhance the efficiency of their usage and the quality of conditions of competing for them, particularly in the form of auctions and exchanges. Potential benefits of the introduction of such multilateral trading patterns on the market for telecommunications network capacity are discussed in [1], problems of the organization of network bandwidth exchanges are presented in [2].

The variety of possible models and their variants creates a problem of objective evaluation and the feasibility of mutual comparison. The models may implement many different resource allocation algorithms and apply multiple optimization criteria taking into account economic and technical constraints. The chances of a simple quantitative assessment of one model in comparison to others and indicating its advantages and disadvantages are hindered. This paper is an attempt to develop a methodological approach to testing and comparing models of market-based allocation of capacity in communication networks. The method can be helpful in choosing trade models adequate to specific markets segments in the telecommunications sector.

The structure of the article is as follows: Section 2 briefly describes the survivable network design library (SNDlib)

and the multicommodity market model (M^3). Section 3 shows the successive stages of testing. Section 4 presents a set of comparative criteria. Section 5 shows an example of the application of the proposed framework for a selected bandwidth trading model. Section 6 summarizes the results of the research.

2. A Method of Analysis of Market-Based Models

The proposed approach to evaluating market models results in a multi-stage framework. The stages, shown in Fig. 1, are as follows:

- model classification,
- test data generation,
- data adaptation for a model,
- running tests,
- output data analysis.

Each step can be performed independently, using separate tools. Data passed between the successive stages is stored in text files saved in a extensible markup language (XML)-based format.

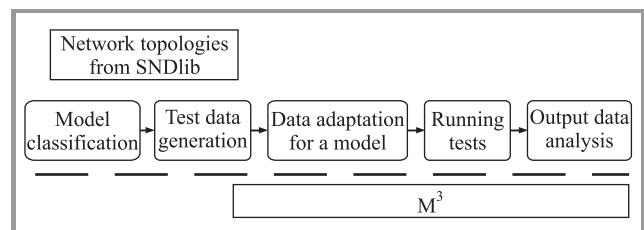


Fig. 1. Schema of the bandwidth model testing framework.

The proposed methodology integrates the results of other research: the data of the SNDlib and the model of a multicommodity market process (M^3).

2.1. The SNDlib Library

The survivable fixed telecommunication network design library [3] is a scientific library sharing exemplary data

for problems of the design/dimensioning of communication transport networks.

Examples stored in the library reflect the topology of real networks. All of them are saved in a standardized XML dialect. The main purpose of the library is to collect actual data on research problems and create a platform for exchange of information between scientists and engineers involved in network design. The library comprises network topologies with the structure of network links and bandwidth demands, best solutions and their dual bounds, an up-to-date bibliography and a list of conferences on the subject. The scope of solving methods is broad and includes models of linear/integer programming, branch-and-bound algorithms, column generation, a Lagrangian relaxation and meta-heuristics such as evolutionary algorithms, simulated annealing or taboo search.

2.2. The M^3 Model

Consistent description of a broad scope of potential trade processes on the bandwidth market needs an application of an adequately flexible information model. Such flexibility and a high degree of openness can be obtained by using an information model based on the M^3 [4]. The M^3 is a set of formal models that describe data and communication messages on multi-commodity infrastructure markets. The M^3 model has been adopted in our research for a market of transport resources in telecommunication networks.

The M^3 model enables a generic description of the trading information exchanged between market participants. The data is stored in a special M3-XML dialect allowing the expression of: the existing network infrastructure, the time scale in which the trade is accomplished, the entity structure (sellers, buyers, brokers, leaseholders, etc.), the trade object structure (description of traded goods) and the offers submitted by individual market participants. It contains, in particular, descriptions of market offers: elementary (singlecommodity), integrated (multicommodity), and also grouping: describing more complex relations between elementary commodities or integrated offers with common conditions or resource constraints. The introduction of the mechanism of grouping of the offers facilitates the formulation of non-trivial constraints for each individual market participant, and proper valuation/quoting of the offers.

In practical applications one may use all or just selected elements of the M^3 data model. It is worth noting that the M^3 model and its M3-XML data format do not specify how the trade itself and the allocation of resources are realized.

3. Stages of Preparing and Running Tests

This section presents the sequential stages of preparation and running tests of bandwidth market models.

3.1. Classification of the Model

The properties of models originate in economics, game theory principles of mechanism design and in technical features of traded bandwidth. Selecting distinctive features of the models creates a space for their classification and grouping. The set of the classifying features have been developed upon analysis of the network resource structure and the organization of trade processes. The categories of the network design problems from [3] have also been taken into account.

Type of commodities. The number of types of resources in telecommunication networks is significant: there are many technologies used in transmission systems (switching and multiplexing techniques). The resources can have a physical or a virtual nature and their detailed specification depends on a layered architecture of modern networks.

Our research is focused on transport communication networks and in this context the general elementary commodity is a bandwidth of a point-to-point connection between a pair of nodes in a specified network layer. It is defined as a network capacity enabling transmission of specified data amount from a source node to a destination node during specified quantum of time.

The bandwidth of a network link can be offered for sell. The bandwidth needed to serve a traffic demand can be purchased – the existence of a direct connection between the pair of nodes of the demand is not necessary, the demand can be realized with one or more paths (consisting of sequences of links).

The trade models can take into account many of characteristic features of bandwidth commodities. The most significant ones are following:

- direction of bandwidth: bandwidth commodities can be directed, undirected (data flow is possible in both directions) or asymmetric (capacity depends on the direction of data flow); bidirectional bandwidth commodities can be modeled with two oppositely directed commodities;
- divisibility of bandwidth: bandwidth can be fully divisible, modular (divisible with the accuracy of units), unit (a particular case of modular bandwidth modeled in combinatorial auctions), predefined (divisible only within specified volumes);
- commodity structure: some trade models omit structural relations between commodities and concern trading only separate elementary commodities; other models enable trading complex structures of commodities, particularly on the demand side: the structures can refer to a set of predefined network paths, a specified set of network links or whole subnetworks (e.g., for purposes of building virtual private networks).

Relations between offers and commodities. The models can take into account different kinds of assignment between

traded commodities and related trade offers. One can distinguish:

- elementary offers: a single offer concerns a single elementary commodity;
- bundled offers: a single offer concerns a bundle of elementary commodities, the commodities can be sold/bought in equal or different amounts.

Market participants. The models can describe many organizational forms of trade. The basic division takes into consideration the number and the roles of market participants. One can distinguish:

- single-side models: a market operator sell bandwidth capacity to its clients, or a single client buys services from many competing network operators;
- double-sided models: there are many buyers and sellers.

The important role in trade models is a market operator – the entity that balances the market: allocates resources and sets prices. In centralized models there is one market operator. In distributed models there can be many of them.

Quality of service constraints. Market contracts concerning telecommunication resources usually define a set of parameters describing the quality of service (QoS). The trade models can take those constraints into account, e.g., specifying the maximal length of a path or the maximal delay of a packet.

Resource allocation rule. The network resources are allocated by means of a defined algorithm operating on the available market offers. In the context of transport network resources one can distinguish the basic allocation rules:

- path setting: the required point-to-point bandwidth connections are served with paths consisting of a sequence of communication links;
- allocation of single resources: buy offers concern separate resources that is not bundled without any explicitly expressed relation.

Pricing rule. Market models set contract prices according to a specified rule. Some popular examples of such rules are: English auction, second-price auction, Vickrey-Clarke-Groves auction, dual pricing, etc.

Exchanged messages. The essential feature of a trade model is the type of signals exchanged between market participants and a market operator. The signals can have various forms:

- point characteristics: a traditional market offer indicating commodities, their amounts and offer prices;
- partial characteristics of preferences, e.g., a set of points from utility function or a stepwise offer;
- full characteristics of preferences, e.g., in the form of an utility function.

Market balancing dynamics. The models can be divided according to the time schedule of the market balancing process. One can distinguish two basic classes:

- one-time auctions: market participants submit their offers and then a market operator allocates resources and sets prices taking into account all submitted offers;
- iterated auctions: the final market balance is achieved in a sequence of steps, in which market participants can modify their signals submitted to a market.

Implementation. The models can be implemented with many different techniques and tools originated in operations research, computer programming and mathematics. Examples of the implementation types are following: linear programming, mathematical programming, dynamic programming, parametric equations, heuristics, etc.

3.2. Input Data Generation

One of the purposes of the research presented in this paper is to create a library of test examples for the scientific community involved in the design of market algorithms and models for communication transport networks. Input data for test cases should reflect the size of demand and supply observable in the real network bandwidth market. The test data used for the research may come from the following sources:

- examples of network design problems;
- economic models of supply and demand;
- real data from the telecommunications market.

Trade patterns for transport resources of networks and balancing market offers are conceptually similar to the problems of network design – the relationship between trade mechanisms and designing the network have been discussed in [5]: the demand for network bandwidth and communication links between nodes can be interpreted as a market offer to buy/sell network resources. In order to generate a test network topology and market offers one can therefore use the data from the SNDlib library. The examples from the database cannot be used directly due to their important constraints. Firstly, there is only one link between pairs of nodes in a network topology, while the market models assume the existence of multiple offers on the bandwidth segment between a pair of nodes (a similar requirement applies also to buy offers). Secondly, any pricing information in the examples of the SNDlib is only expressed by determining the cost of installation and expansion of links, while in the trade models one requires the price of bids submitted by buyers and sellers. All the missing elements can be added to the original network design examples, e.g., by means of the pseudo-random generation adjusted to the specific test case.

The second source of the data on the potential offers on the bandwidth market is the use of economic models of supply and demand. Such models in the context of resources in telecommunication networks are under research, especially in the field of modeling network traffic and demand for services. Some analysis of supply and demand can be transferred to the ground of telecommunications from other infrastructure markets such as energy or transportation.

The test data can also come from the analysis of the actual data on transactions accomplished in the real telecommunications market. Acquiring such data is difficult in practice because there are not any network bandwidth exchanges operating on a larger scale, and the information about bilateral contracts between telecommunications companies is generally private and not publicly available. However, there are some internet sites showing examples of bandwidth prices in certain local markets, e.g., the U.S., and reports on global trends in the development of the telecommunications market, such as the work [6].

3.3. Data Adaptation – Conversion to M^3 Format

The adopted information model for the framework is the M^3 – any data of test cases should be saved in the M^3 -XML format. This can be achieved in two ways.

The first one is the direct generation of input data in the desired format. This requires dedicated computer tools allowing editing or automatic generation of the test data.

The second approach assumes the use of examples contained in the SNDlib library enriched with the entity structure, the object structure and trade offers. Choosing an XML dialect as a data format results in the opportunity to use widely available read/write software libraries for many programming languages, which facilitate the development of new tools for test data generation. The natural way to transform XML data into another format is the use of a extensible stylesheet language transformation (XSLT) [7]. One can also use the query language XQuery [8].

For the purpose of this research a mixed approach has been adopted. A dedicated computer program with a graphical user interface has been designed and implemented. It has been used to enrich the SNDlib examples with lacking elements and save them in the simple dedicated bandwidth market XML (BM-XML) format describing the network nodes and the offers for sale and purchase of network capacity. A series of XSLT transformations processing the BM-XML format into M^3 -XML format have been used. In this way we have obtained a convenient set of computer tools used for importing and converting examples of network topology to the new library of complete test cases for bandwidth market models.

3.4. Running Tests

The next step is to run required calculations of network capacity allocation and pricing for a tested model with the prepared data. It is assumed that all the input data is stored in the M^3 -XML format. Therefore any market

mechanism, which is a subject to the tests, should be able to read input data and generate output data compatible with the M^3 information model. The flow of data during the execution of the tests is illustrated in Fig. 2.

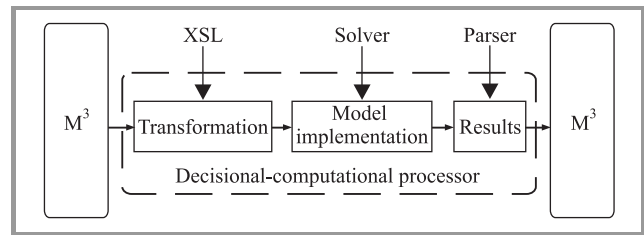


Fig. 2. M^3 -XML format data flow in running tests.

The concept of a universal decisional-computational processor has been developed [9] and the initial implementation has been made: it solves the tasks described by the data stored in M^3 -XML format. The corresponding XSLT transformation converts the data to the internal representation of the appropriate model. The processor returns the data in the M^3 -XML format. Such modular architecture allows independent implementation of the trade models from the evaluation framework: compatibility with the M^3 information model can be achieved by the input and output interfaces. The XSLT transformations can be applied to convert the input data into the format required by the particular computing processor, e.g., a standard mathematical programming solver or a dedicated implementation of the algorithm of a trade model. Examples of such transformations have already been created [4] and they are further extended in our research on new trading models. They convert the data in M^3 -XML format into GMPL format (gnumath programming language) [10], which can be used by optimization solvers, e.g., GlpSol and AMPL (a mathematical programming language). Linear models can also be easily converted to other LP (linear programming) formats, e.g., the one used by the CPLEX solver.

3.5. Analysis of Output Data

The next step to take after having completed series of computational experiments is to analyze the obtained results. The application of the M^3 information model to describe the output of the tested models facilitates the analysis. If one had different data formats for various trade models the comparison of the results would be much more difficult and time-consuming. The adopted XML-based format allows for simple data conversion to other formats for presentation purposes, such as exporting to spreadsheets or graphical visualization applications.

4. Comparative Criteria

A fundamental part of the evaluation framework is the proper selection of indicators. Trade models are often designed for a specific market context and defined aspects

of resource allocation, so the comparative criteria should be general enough to be applicable to most of them. We propose a set of such indicators, they represent measures originating in economics, game theory and technical efficiency.

A properly designed trade model should strive to meet several desirable properties that make it attractive for a wide range of market participants (traders, decision makers, regulators). These properties can be used as evaluation criteria providing information about the "quality" of a given model. The set of criteria is divided into three categories: global, individual and technical. The division into the individual and global ones expresses the natural market game between individual participants' interests and global interests. The third category aims in evaluation of the models in terms of their technical efficiency.

4.1. Global Criteria

Economic efficiency. The main measure of economic efficiency is the market surplus. It is defined as the aggregated economic benefits derived from the market exchange of goods. If the market mechanism encourages participants to submit truthful bids, then economic prosperity could be determined accurately using actual bids. If the market offer does not comply with the participant's preference profile, it can be used only as an imprecise measure known as economic benefits. It is the difference between total value of goods purchased and the total value of goods sold on the market, as follows:

$$Q = \sum_{m \in B} d_m e_m - \sum_{l \in S} p_l s_l, \quad (1)$$

where e_m is the unit offer price of buy offer m , d_m is realized bandwidth of buy offer m , s_l is the unit offer price of sell offer l , and p_l is realized bandwidth of sell offer l .

Incentive compatibility. The incentive compatibility property holds if no market participant has incentives to report signal different from its type/preferences: no agent has incentives to report an untruthful offer.

Incentive compatible mechanism prevents strategic actions of the participants. The measure of the effectiveness of the mechanism against strategic players is the allocative inefficiency (AI) defined as follows:

$$AI = \frac{Q^0 - Q}{Q^0} 100\%, \quad (2)$$

where Q^0 is the economic surplus if every market participant submits a truthful bid, Q is the actual (achieved) value of the economic welfare, where participants can submit bids incompatible with their preferences. It should be noted that the participants may submit bids incompatible with their profile of preferences in order to achieve higher profits.

Budget balance. A trade model has balanced budget if sum of sellers' expenses is equal to sum of buyers' incomes:

there is no need to surcharge the trading mechanism, and the mechanism does not earn additional profits.

The first quantitative measure describing this criterion is the value of the difference between total sellers' income SI and total buyers' expense BE related to the total market value. This measure has got the term budget imbalance in relation to total turnover ($RBIT$):

$$RBIT = \frac{SI - BE}{SI + BE} 100\%. \quad (3)$$

The second measure describing the budget balance criterion is the value of the difference between total sellers' income and total buyers' expense related to the market surplus. This measure has got the term budget imbalance in relation to market surplus ($RBIS$):

$$RBIS = \frac{SI - BE}{Q} 100\%. \quad (4)$$

Pareto-efficiency. The results given by the trading model are Pareto-efficient, if one can not improve the result for one market participant without making some other participants worse off: the results are Pareto-efficient if such results are not Pareto-dominated by other results.

4.2. Individual Criteria

Individual economic benefits. From the perspective of an individual market participant the trading model should allow to obtain the highest possible value of the individual economic benefits. The measure is defined as the value of the individual utility function for each market participant.

Absolute individual fairness – individual rationality. A trade model is individually rational, if no market participant loses from the participation in the trade. This property is also called the voluntary participation (if it may lose then it can choose not to participate in the trade).

Relative individual fairness. A trade model is fair in a relative sense, when from the perspective of each participant no other offer is favored in relation to its offer.

Other criteria related to fairness have been outlined below:

- anonymity: a participant remains anonymous if renumbering of participants does not affect the obtained outcome;
- symmetry: two participants with the same parameters, in the same market situation (the same utility functions) should obtain the same individual results;
- an equal price: each participant receives the same volume of a commodity for the same unit price.

4.3. Technical Efficiency

The possibility of a practical technical implementation of the trade model is evaluated with technical efficiency indi-

cators dependent on the computational complexity of the model and on its reliability. Exemplary measures are:

- duration of single market balancing;
- total duration of the market balancing (for iterative mechanisms);
- a number of exchanged messages (total, average).

5. Example

The evaluation framework has been applied to test the balancing communication bandwidth trade (BCBT) model discussed in [5]. It is a multicommodity trade model for the market of network transport resources with many buyers and many sellers. The model assumes complete divisibility of offered bandwidth (sell offers) and the ability to allocate any fraction of bandwidth to any bandwidth demand (buy offers). The market balancing process consists of setting the paths serving buy offers with a combination of sell offers. The formal format of the model is a linear programming problem representing a double auction with the maximization of market welfare.

The evaluation of the properties of the BCBT model has been carried out according to the framework principles outlined in the paper. Below a brief description of this test case is presented.

5.1. Classification of the Model

Type of commodities. The BCBT model assumes that both links (sell offers) and paths (buy offers) concern unidirectional network capacity. The division model of the offered bandwidth is continuous: the seller and buyer may indicate bandwidth amount of their bids with any positive real number. The bandwidth can be traded in the range from 0 to a maximal admissible volume specified for the offers.

Relations between offers and commodities. All market offers in the model are elementary offers concerning elementary point-to-point bandwidth connections in the network.

Market participants. The model represents a double auction: there are many buyers and many sellers. It is a centralized trading model – a single market operator balances the market.

Quality of service constraints. There are not any QoS parameters included in the model (e.g., there is no limit on the maximum length of paths).

Resource allocation rule. The balancing mechanism can freely allocate bandwidth in the network, in particular a single sell offer can be used to serve multiple buy offers, a single buy offer can be realized with many paths. There are no restrictions on the possibility of setting paths in the network: there are no predefined paths, they are set during

the market balancing process taking into account of all available offers.

Pricing rule. The pricing rule adopted is the dual pricing rule – it is based on dual prices of the balance constraints. In the BCBT model the balance constraints exist only for the links – so the prices are determined directly for the sell offers only. The prices for demands (buy offers) are determined as a sum of the values of bandwidth bought on links in the related paths.

Exchanged messages. The messages exchanged in the model are point characteristics of market participants' preferences in the form of a traditional offer expressing the amount of commodities and their unit prices.

Market balancing dynamics. The model is an example of a one-time double auction – the market participants submit offers and then the market is balanced – resources are allocated and prices are calculated.

Implementation. The model has been implemented as a linear program maximizing total market welfare. The linear program is written in GMPL and AMPL.

5.2. Input Data

The network topology is derived from the example of the NOBEL-EU network from the SNDlib developed in the IST NOBEL project “*Next Generation Optical Networks for Broadband European Leadership*” [11]. The nodes of the network are located in 28 major European cities. The network topology (nodes and potential links) is illustrated in Fig. 3.

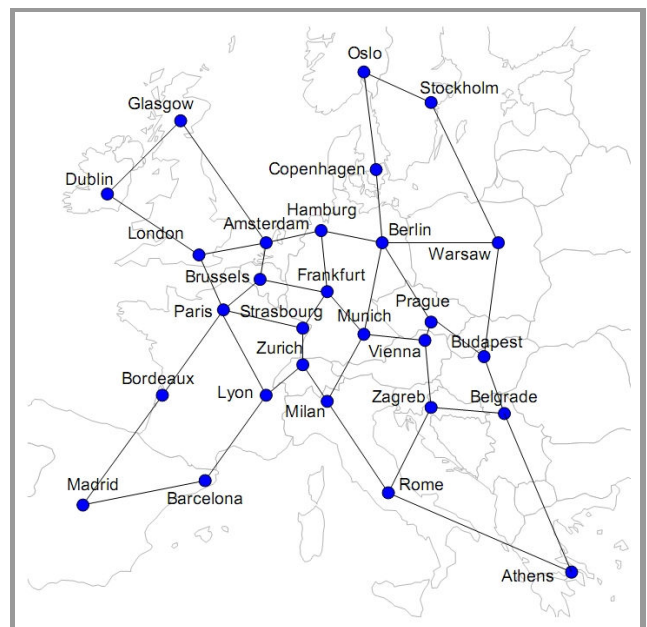


Fig. 3. Pan-European network of EU project IST NOBEL [11].

Bandwidth buy offers are independent from the traffic demands included in the original SNDlib example. It is assumed that the offered prices rise along with the geo-

graphical distance between a pair of nodes, and offered bandwidth capacity decreases with the distance. 150 buy offers have been generated by the following algorithm:

1. Draw randomly a source and a destination network node (different from each other).
2. Set pseudo-randomly offer price according to the expression (5) with dependency on the geographical distance between the nodes:

$$\text{Gauss}\left(\text{AvgP} \cdot \text{Dist}, \frac{\text{AvgP}}{3\text{Dist}}\right). \quad (5)$$

3. Set pseudo-randomly offer bandwidth capacity according to the expression (6) with dependency on the geographical distance between the nodes:

$$\text{Uniform}(0.5 \text{ AvgC}, 1.5 \text{ AvgC}) \left(1 - \frac{\text{Dist}}{\text{Dist}_{\max}}\right) \quad (6)$$

with the following denotation:

- $\text{Gauss}(a, b)$: Gaussian distribution with the expected value a and the standard deviation b ;
- $\text{Uniform}(a, b)$: uniform distribution in the range $[a, b]$;
- AvgP : average unit bandwidth price related to a unit distance;
- AvgC : average bandwidth capacity for adjacent nodes (distance equals 0);
- Dist : distance between nodes;
- $\text{Dist}_{\max}$: the maximal distance between the nodes in the network.

Following parameter values has been used for buy offer generation: $\text{AvgP} = 1$, $\text{AvgC} = 15$.

Sell offers have been also generated randomly (they have replaced links in the original SNDlib example: the primary links have been interpreted as bandwidth segments on which sell offers are submitted). The assumption made is similar to the case of buy offers: the offered prices rise along with the geographical distance between a pair of nodes, and offered bandwidth capacity decreases with the distance. 5 separate sell offers have been generated on every bandwidth segment using the same mathematical expression for setting prices and capacities as for the buy offers (with the following parameter values: $\text{AvgP} = 0.5$, $\text{AvgC} = 10$) according to the algorithm:

1. Set pseudo-randomly the direction of the offer (the probability of each direction equals 0.5).
2. Set pseudo-randomly offer price according to the expression (5) with dependency on the geographical distance between the nodes.
3. Set pseudo-randomly offer bandwidth capacity according to the expression (6) with dependency on the geographical distance between the nodes.

The test data generated for the Nobel-EU network including 150 buy offers and 205 sell offers has been saved in a BM-XML format file. The data set is a simple but complete test case, and it can be used also for testing the properties of other models than the BCBT.

5.3. Adaptation and Running Tests

A single test procedure consists of three phases. The first one is the adaptation of the data stored in BM-XML format to the M3-XML format. The second phase is the solution of the problem in accordance with a given mathematical model of the BCBT. The third phase is the collection of the results and their analysis. The detailed description of each step is as follows:

- The data in the BM-XML format has been converted to the M³ information model using several XSLT transformations. The result of this procedure is a set of files in the M3-XML format.
- The obtained data set is then passed to the decisional-computational processor performing the following steps: the data is converted to a GMPL model implementing the BCBT allocation model using an XSLT transformation; the complete GMPL model with numeric data is passed to a linear programming solver (e.g., AMPL, GlpSol) returning results.
- The results are parsed to the M3-XML format.

Calculations have been performed on a PC (Intel Core 2 Duo 2.60 GHz, 2 GB RAM).

5.4. Output Data Analysis

The numerical data and short comments for the framework indicators have been presented below.

Economic efficiency. From a global perspective the BCBT model is economically efficient in the sense that it maximizes the global economic welfare. Thus, for given market offers, no better allocation of bandwidth resources is possible. The market welfare achieved in the test case is 13861.

Incentive compatibility. The obtained value of the allocative inefficiency equals 0.026. It is a very good outcome – it indicates that the possibility for speculation on the market is limited.

Budget balance. The RBIT and RBIS benchmarks are 0, so the model fulfills the requirements for the budget balance property.

Pareto-efficiency. The Pareto-efficiency property holds for the family of BCBT models. Pareto-efficiency is ensured if

the aggregation function of the individual objective functions is strictly increasing. The family of BCBT models uses objective function as the sum of individual objective functions, which is strictly increasing with respect to every coordinate, so the whole BCBT family is Pareto-efficient.

Individual economic benefits. The calculated contract prices have been used to determine the individual benefits of the market participants. The list of their values has not been included here because of the limited space of the paper.

Absolute individual fairness – individual rationality. The individual benefits for those market participants, whose offers have been successfully traded, are positive. So the requirements of the individual rationality property are fulfilled.

Relative individual fairness. The requirements for relative fairness of individual market participants are partially met in the test case. In general their fulfilment depends on the implementation of solvers (e.g., the order of identical market offers may have impact on the volume of their realization). The BCBT model itself does not have any additional constraints for complying with these requirements.

Duration of market balancing. The total time of the market balancing has been very short: 0.95 s. There have not been any additional computation for resource pricing because the prices have been derived from the parameters of the linear program solution. The LP formulation of the BCBT model does not comprise integer variables, to it is feasible even for large networks.

Number of exchanged messages. The market participants have submitted 150 buy and 205 sell offers – there are total 355 bid messages (the same number of messages covers announcing the results of the trade).

6. Summary

The paper presents an attempt to develop a methodological approach to the problem of comparing different models for trading capacity in the communication transport networks. Several stages of test preparation and running computations have been distinguished and described. The use of standardized data format is essential: the task of comparison of results becomes easier and the same test cases can be reused for many models. The proposed framework will be used to study and compare the properties of different trade models, such as c-SeBiDA (combinatorial sellers' bid double auction) [12], MIDAS [13], BCBT [5], NSP (network service provider) [14], and other new models developed by the authors in their research. The described framework and its tools can be a part of a broader research platform forming an advanced computing environment for testing

the market models of resource allocation in communication networks.

The evaluation framework is in its early stage and still there is a need to refine many of its elements. The most important areas perceived for improvement include:

- The development and extension of economic models to describe the bandwidth demand and supply enabling generation of test data reflecting the real market conditions.
- Refine the set of criteria enabling to study and compare specific characteristics of trading models, e.g., resulting from the market game between trade participants.

These and other extensions of the basic methodology proposed in this paper will be presented in the authors' future papers.

Acknowledgements

The authors acknowledge the Ministry of Science and Higher Education of Poland for partially supporting the research through project PBZ-MNiSW-02/II/2007. The results of the research have been discussed within the JRA3.3 activity of the FP7 Euro-NF Network of Excellence.

References

- [1] S. S. Chiu and J. P. Crametz, "Taking advantage of the emerging bandwidth exchange market", White paper, RateXchange Ltd., 1999.
- [2] W. Stańczuk, J. Lubacz, and E. Toczyłowski, "Trading transport resources of communication networks on bandwidth exchanges", in *Proc. 46th FITCE Congr.*, Warsaw, Poland, 2007.
- [3] "SNDlib: survivable fixed telecommunication network design" [Online]. Available: <http://sndlib.zib.de/>, 2010
- [4] "M³ – multicommodity market model" [Online]. Available: <http://www.openm3.org/>, 2010
- [5] W. Stańczuk, J. Lubacz, and E. Toczyłowski, "Trading links and paths on a communication bandwidth market", *J. Univers. Comput. Sci.*, vol. 14, no. 5, pp. 642–653, 2008.
- [6] "Bandwidth pricing report", Tech. Rep., TeleGeography, PriMetrica Inc., 2007.
- [7] "XSL transformations" [Online]. Available: <http://www.w3.org/TR/xslt>, 2010
- [8] "Xquery 1.0: an XML query language" [Online]. Available: <http://www.w3.org/TR/xquery/>, 2010
- [9] P. Kacprzak, M. Kaleta, P. Pałka, K. Smolira, E. Toczyłowski, and T. Traczyk, *Procesor decyzyjno-obliczeniowy dla rynkowego modelu danych M³*. Warsaw: WKŁ, 2008 (in Polish).
- [10] "GLPK (GNU linear programming kit)" [Online]. Available: <http://www.gnu.org/software/glpk>, 2010
- [11] "Report on multilayer traffic engineering mechanisms for NOBEL solutions in medium and long term scenarios, deliverable – d2.2", Tech. Rep., IST NOBEL "Next Generation Optical Networks for Broadband European Leadership", Phase 2, 2007.
- [12] R. Jain and P. P. Varaiya, "An efficient incentive-compatible combinatorial market mechanism", in *Proc. Allerton Conf.*, Chicago, USA, 2004.

- [13] M. Dramitinos, G. D. Stamoulis, and C. Courcoubetis, "An auction mechanism for allocating the bandwidth of networks to their users", *Comput. Netw.*, vol. 51, iss. 18, pp. 4979–4996, 2007.
- [14] R. Jain and J. Walrand, "An efficient mechanism for network bandwidth auction", in *IEEE Netw. Oper. Manage. Symp. Worksh. 2008*, Salvador da Bahia, Brazil, 2008, pp. 227–234.

Note: this text is reprinted from JTIT 2/2010



Wojciech Stańczuk received the M.Sc. degree in telecommunications from Warsaw University of Technology (WUT), Poland, in 2001. Currently he is a Ph.D. student and a researcher in the Institute of Telecommunications (WUT). His research interests lie in modeling technoeconomic aspects of communications networks: resource allocation, pricing, infrastructure investments.

e-mail: w.stanczuk@tele.pw.edu.pl

Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Piotr Palka is an Assistant Professor of the Operations and Systems Research Division in the Institute of Control and Computation Engineering at the Warsaw University of Technology (WUT), Poland. He received the M.Sc. and Ph.D. degrees in computer science in 2005 and 2009, respectively, both from WUT. His research

interest focus on the market mechanisms, especially the incentive compatibility on the infrastructure markets and on the multi-agent systems. His current research is focused on application of multicommodity turnover models.

e-mail: P.Palka@ia.pw.edu.pl

Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Józef Lubacz received his Ph.D., and D.Sc. at Warsaw University of Technology (WUT), Poland. He was the Head of the Department of Teleinformatics (WUT, 1987–2002), Dean of the Faculty of Electronics and Information Technology (WUT, 2002–2005). Currently he works as a Full Professor at Warsaw University of Technology (WUT). He is a member of the General Council for Higher Education of Poland (since 2006) and the Director of the Institute of Telecommunications (WUT, 2008). His scientific interests focus on: networks planning methodology, non-technological issues of information society and philosophy of technology.

e-mail: jl@tele.pw.edu.pl
Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Eugeniusz Toczyłowski is a Professor, the Head of Operations and Systems Research Division, at the Institute of Control and Computation Engineering at the Warsaw University of Technology (WUT), Poland. He received the M.Sc. degree in 1973, Ph.D. in 1976, D.Sc. in 1989, and the title of Full Professor in 2004. His

main research interests are centered around the operations research models and methods, including structural approaches to large scale and discrete optimization, auction theory and competitive market design under constraints, multicommodity trading models, and design of management information systems.

e-mail: E.Toczyłowski@ia.pw.edu.pl

Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland

Performance Analysis of Hybrid Phase Shift Keying over Generalized Nakagami Fading Channels

Mahmoud Youssuf and Mohamed Z. Abdelmageed

Abstract—In addition to the benefits of hybrid phase shift keying (HPSK) modulation in reducing the peak to average power ratio of the transmitted signal to reduce the zero crossings and the 0° -degree phase transmissions, HPSK enhances the bit error rate (BER) measure of the signal performance. The constellation of the HPSK is analyzed, and an expression for the conditional probability of HPSK modulation over additive white Gaussian noise (AWGN) is derived. This BER measure of HPSK is shown to outperform quadrature phase shift keying (QPSK) modulation. HPSK performance through Nakagami – m fading channel is also considered.

Keywords—bit error rate, hybrid phase shift keying, quadrature phase shift keying.

1. Introduction

Hybrid phase shift keying (HPSK) is used in wideband code division multiple access (WCDMA) systems thanks to its low peak-to-average power ratio. This low ratio of peak-to-average power results in reducing the number of zero crossing of phase transitions of the output transmitted signal. In this paper we prove that HPSK outperforms other quadrature modulation techniques, such as offset quadrature phase shift keying (OQPSK) by 3 dB.

The paper is organized as follows: Section 2 describes the HPSK constellation in case of two channels at different amplitudes. Section 3 derives an expression for the conditional probability of error of HPSK modulated signal over an additive white Gaussian noise (AWGN). In Section 4 we apply the obtained expression in evaluating the performance of HPSK modulated signal over a generalized Nakagami – m channel. Finally, Section 5 includes numerical results and comments.

2. The Hybrid Phase Shift Keying Constellation

The HPSK has been proposed as the spreading technique for WCDMA to eliminate the zero crossings for every other signal transition and to eliminate the 0° -degree phase shift transitions for every other chip point, as well as to improve the bit error rate (BER) measure of the direct sequence WCDMA (DS-WCDMA) system performance.

In 3G systems the mobile station (MS) can transmit more than one channel. The different channels are assigned to either I or Q path.

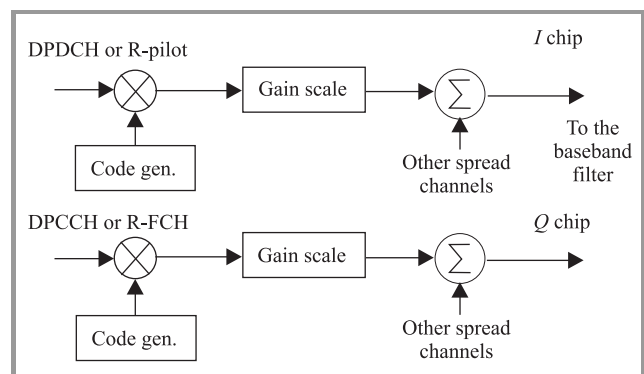


Fig. 1. The basic reverse channel structure of 3G system. Explanations: DPCCH – dedicated physical control channel, DPDCH – dedicated physical data channel, R-FCH – reverse fundamental channel.

In the case of transmitting only two channels, as in Fig. 1, one of the channels (DPDCH or R-pilot) is applied to the I path and the other channel (DPCCH or R-FCH) is applied to the Q path [1]. Additional, high data rate channels are combined alternatively on the I or Q paths. Each channel is spread by a different orthogonal even-numbered Walsh code. In the general case the channels can be at

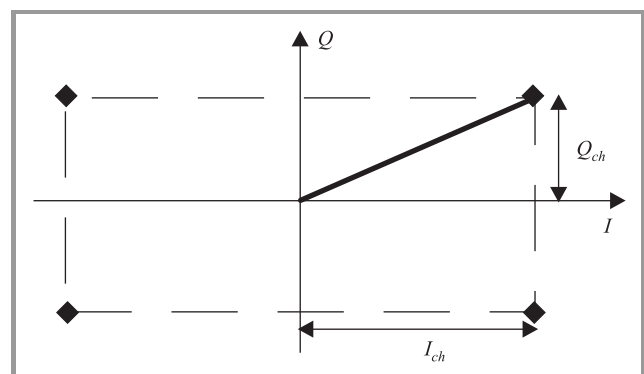


Fig. 2. The 4-QAM constellation for two channels at different amplitudes.

different power levels, as in Fig. 2 which maps onto a rectangular four quadrature amplitude modulation (4-QAM) constellation.

The QAM signal waveforms may be expressed as

$$\begin{aligned} S_m(t) &= \text{Re}\{Ae^{j\theta_m}g(t)e^{j2\pi f_c t}\} \\ &= A_{ch}g(t)[\cos(2\pi f_c t)\cos(\theta_m) \\ &\quad - \sin(2\pi f_c t)\sin(\theta_m)], \end{aligned} \quad (1)$$

where: $A_{ch} = \sqrt{I_{ch}^2 + Q_{ch}^2}$, where I_{ch} and Q_{ch} are the information bearing signals amplitudes of the I path channel and Q path channel, respectively, $g(t)$ is the pulse shape signal, $\theta_m = \tan^{-1} \frac{Q_{ch}}{I_{ch}}$ is the phase of the signal vector and it varies with $m = 1, 2, 3, 4$.

So, the QAM signal waveform may be viewed as a linear combination of two orthogonal wave forms $f_1(t)$ and $f_2(t)$ such that [2]:

$$S_m(t) = S_{m1}f_1(t) + S_{m2}f_2(t), \quad (2)$$

where S_{m1} and S_{m2} are the component of the two dimensional vector S_m :

$$S_m = [S_{m1} \quad S_{m2}] = [I_{ch} \quad Q_{ch}]. \quad (3)$$

In the 4-QAM according to the position of the vector point and according to the value of I_{ch} and Q_{ch} :

$$\therefore S_m = [A_{ch} \cos \theta_m \quad A_{ch} \sin \theta_m]. \quad (4)$$

In the reverse link of DS-CDMA systems the I_{ch} and Q_{ch} are complex scrambled with a complex scrambling signal ($I_s + jQ_s$) as in Fig. 3.

The final I and Q signals are produced mathematically by the multiplication of the two complex signals; the complex data signal ($I_D + jQ_D$) which has already spread into chips ($I_{ch} + jQ_{ch}$), and the complex scrambling signal ($I_s + jQ_s$) so the final I and Q signals are:

$$I + jQ = (I_{ch}I_s - Q_{ch}Q_s) + j(I_{ch}Q_s + Q_{ch}I_s) = A_{ch}A_s e^{j(\phi_{ch} + \phi_s)},$$

$$I = A_{ch}A_s \cos\left(\frac{\pi}{M}(2n-1) + \theta_m\right) = A \cos\left(\frac{\pi}{M}(2n-1) + \theta_m\right), \quad (5)$$

$$Q = A_{ch}A_s \sin\left(\frac{\pi}{M}(2n-1) + \theta_m\right) = A \sin\left(\frac{\pi}{M}(2n-1) + \theta_m\right). \quad (6)$$

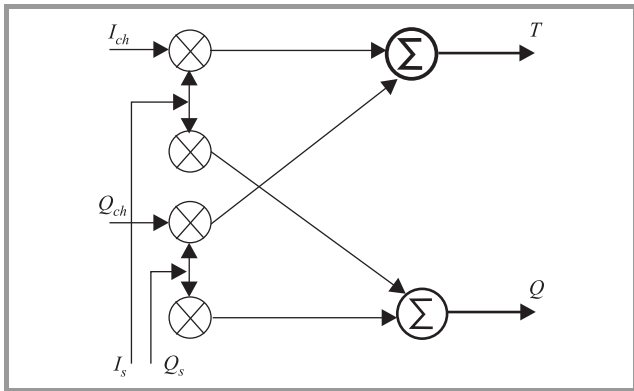


Fig. 3. Complex scrambling of HPSK.

Since final constellation is formed by complex multiplication of the two signals of chip constellation and scrambling constellation which is always QPSK constellation as in Fig. 4, then: from Eqs. (5) and (6) we conclude that the final signal ($I + jQ$) is a QPSK constellation with two dimensional vector representation S_{mn} , where:

$$S_{mn} = \begin{bmatrix} A \cos\left(\frac{\pi}{M}(2n-1) + \theta_m\right) & A \sin\left(\frac{\pi}{M}(2n-1) + \theta_m\right) \end{bmatrix}, \quad (7)$$

where $A = A_{ch}A_s$; $n = 1, 2, 3, 4$; $m = 1, 2, 3, 4$ and $M = 4$.

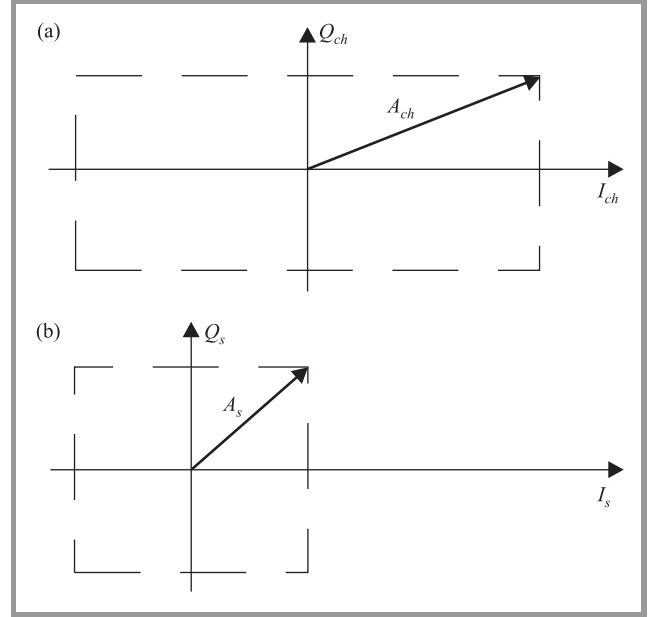


Fig. 4. The chip constellation (a) and the scrambling constellation (b).

This new constellation has points that rotate according to the angle: $\frac{\pi}{4}(2n-1) + \theta_m$, while θ_m changes according to the value of I_{ch} and Q_{ch} , for example, if $n = 1$, the point (I_{ch}, Q_{ch}) will be transferred by the angle equal to $(45^\circ + \theta_m)$. So the new constellation is really an eight point constellation with two independent QPSK constellation as shown in Fig. 5 according to the value of θ_m . One of these two constellations corresponds to $\theta_m > 45^\circ$ and it rotates by angle equal to $\phi_1 = (\theta_m - 45^\circ)$ from the original axes. The other QPSK constellation corresponds to $\theta_m < 45^\circ$ and rotates with angle equal to $\phi_2 = -(\theta_m - 45^\circ)$ from the original axes. So the new final constellation consists of two independent QPSK constellations with complex scrambling:

$$S_n = \begin{bmatrix} A \cos\left[\frac{2\pi}{M}(n-1) + \phi_1\right] & A \sin\left[\frac{2\pi}{M}(n-1) + \phi_1\right] \end{bmatrix}, \quad (8)$$

$$S'_n = \begin{bmatrix} A \cos\left[\frac{2\pi}{M}(n-1) + \phi_2\right] & A \sin\left[\frac{2\pi}{M}(n-1) + \phi_2\right] \end{bmatrix}. \quad (9)$$

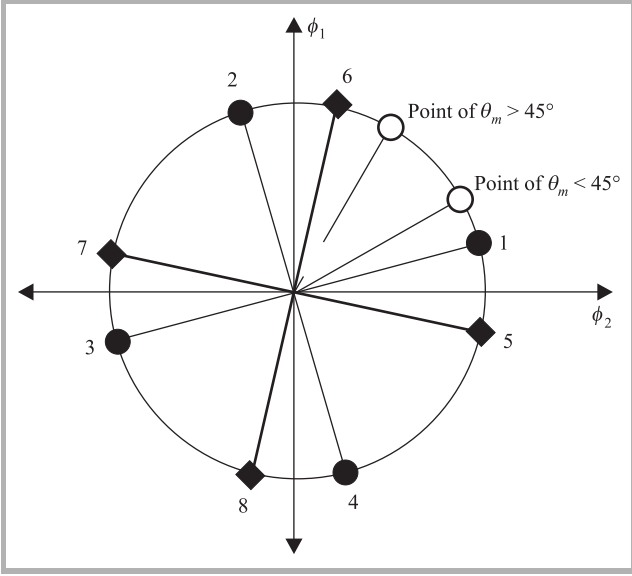


Fig. 5. The final constellation of the scrambled chip of different channel amplitudes.

So, the final constellation has eight points, as in Fig. 5, with the angular distribution determined by the relative levels of the two channels signals.

3. Probability of Error of Hybrid Phase Shift Keying Modulated Signal over Additive White Gaussian Noise

We concluded in the previous section that the final constellation of the two channels at different amplitudes consists of two independent QPSK constellations with complex scrambling. The average value of the amplitude of the new constellation is $\sqrt{2}$ times the value of the amplitude of the traditional QPSK. To obtain the BER of the HPSK modulated signal as a measure of its performance we will assume that this signal is transmitted over AWGN channel. The received signal is demodulated with correlated demodulator or a matched filter demodulator and the decision for the received observation vector $r = [r_1, r_2, r_3, \dots, r_N]$ among all the transmitted signals S_m is based on the maximum of the conditional probability distribution function (pdf) $P(r/S_m)$, which is the maximum likelihood (ML).

The optimum ML detector computes a set of M correlation metrics $C(r, S_m) = -2rS_m$ and selects the signal corresponding to the largest correlation metric [3].

Applying this metric in our study case, r is the received signal vector $r = [r_1, r_2]$ which is projected onto each of the four possible transmitted signals vectors S_m for $m = 1, 2, 3, 4$, where $M = 4$. We can consider that

the correlation detector in the case of HPSK modulated signal is equivalent to a phase detector that computes the phase of the received signal from r and selects the signal S_m whose phase is closest to r , where the phase of r is $\theta_r = \tan^{-1} \frac{r_2}{r_1}$.

The probability of error can be computed if we determine the power density function of θ_r $P_{\theta_r}(r)$. We consider the case in which the transmitted signal phase is equal to zero, as in Fig. 6.

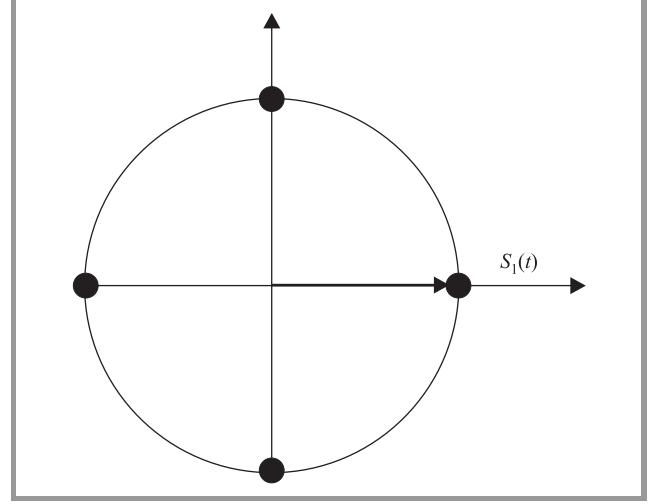


Fig. 6. The HPSK vector constellation.

The transmitted vector $S_1 = [\sqrt{\epsilon_s}, 0]$, where ϵ_s is the energy of the transmitted HPSK signal $S_1(t)$. The received signal vector is $r = [r_1, r_2] = [\sqrt{\epsilon_s} + n_1, n_2]$, where n_1 and n_2 are jointly Gaussian random variables with mean and variances $E(r_1) = \sqrt{\epsilon_s}$, $E(r_2) = 0$ and $\sigma_{r_{12}} = \sigma_{r_{22}} = \sigma_{r_2} = \frac{1}{2}N_0$. Consequently, the joint pdf of r_1 and r_2 is:

$$P(r_1, r_2) = \left(\frac{1}{\pi N_0} \right) \exp \left[-\frac{(r_1 - \sqrt{\epsilon_s})^2 + r_2^2}{N_0} \right]. \quad (10)$$

The pdf of the phase θ_r is obtained by a change in variables from (r_1, r_2) to (A, θ_r) , where:

$$A = \sqrt{r_1^2 + r_2^2}, \quad \theta_r = \tan^{-1} \frac{r_2}{r_1},$$

$$P(A, \theta_r) = A \left(\frac{1}{\pi N_0} \right) \exp \left[-\frac{(A^2 + \epsilon_s - 2\sqrt{\epsilon_s}A \cos \theta_r)}{N_0} \right], \quad (11)$$

$$P_{\theta_r}(\theta_r) = \int_0^\infty P(A, \theta_r) dA = \frac{1}{2\pi} e^{-2\gamma_s \sin^2 \theta} \int_0^\infty A e^{-\frac{A - \sqrt{4\gamma_s} \cos \theta}{2}} dA, \quad (12)$$

where $\gamma_s = \frac{\epsilon_s}{N_0}$ is the signal-to-noise ratio (SNR).

For large values of $\gamma_s \gg 1$ and $|\theta_r| \leq \frac{\pi}{2}$, $P_{\theta_r}(\theta_r)$ is well approximated as

$$P_{\theta_r}(\theta_r) = \frac{1}{2\pi} e^{-2\gamma_s \sin^2 \theta_r} I, \quad (13)$$

where

$$I = \int_0^\infty A \exp\left(-\frac{(A - \cos \theta_r \sqrt{4\gamma_s})^2}{2}\right) dA, \quad (14)$$

$$I = \cos \theta_r \sqrt{2\pi} \sqrt{4\gamma_s}, \quad (15)$$

$$P_{\theta_r}(\theta_r) = \sqrt{\frac{2\gamma_s}{\pi}} \cos \theta_r e^{-2\gamma_s \sin^2 \theta_r}. \quad (16)$$

When $s_1(t)$ is transmitted a decision of error is made if the noise causes the phase to fall outside the range $[-\frac{\pi}{4}, \frac{\pi}{4}]$, and hence the probability of a symbol error is:

$$P_4 = 1 - \int_{-\frac{\pi}{4}}^{\frac{\pi}{4}} P_{\theta_r}(\theta_r) d\theta_r. \quad (17)$$

By substituting for $P_{\theta_r}(\theta_r)$ and performing the change of variables from θ_r to μ , where $\mu = \sin \theta_r \sqrt{2\gamma_s}$, we find

$$P_4 = 2Q\left(\sqrt{2\gamma_s \sin^2\left(\frac{\pi}{4}\right)}\right) = 2Q(\sqrt{\gamma_s}). \quad (18)$$

In case of HPSK

$$\varepsilon_s = 2(k\varepsilon_b) = 4\varepsilon_b \therefore P_4 = 2Q\left(\sqrt{\frac{4\varepsilon_b}{N_0}}\right). \quad (19)$$

Since the transmitted signals represented by the vector points of the final constellation are equally likely to be transmitted and since the 8-points are distributed around a circle consisting of two independent QPSK constellations with complex scrambling then the average probability of the symbol error in the case of two channels at different amplitudes is:

$$P_{\text{symbol}} = \frac{1}{2} [P_1(e) + P_2(e)], \quad (20)$$

where $P_1(e)$ is the probability of symbol error of the first QPSK constellation, $P_2(e)$ is the probability of symbol error of the second constellation:

$$P_s = 2Q\left(\sqrt{\frac{4\varepsilon_b}{N_0}}\right). \quad (21)$$

The bit error probability in this case is:

$$P_b = Q\left(\sqrt{\frac{4\varepsilon_b}{N_0}}\right) = Q(\sqrt{4\gamma}), \quad (22)$$

where γ is the SNR $= \varepsilon_b/N_0$.

It is simply interesting to compare the performance of HPSK with that of QPSK since both types of signals are two dimensional. Since the error probability is dominant by the arguments of the Q function, we may simply compare the arguments of Q for the two signal formats of HPSK and QPSK. The ratio of the two arguments is equal $R = 2$. So, HPSK has SNR advantage of $10 \log 2 \approx (3 \text{ dB})$ over QPSK.

4. The Performance of HPSK Signal over a Generalized Nakagami – m Channel

The mobile communication channel is noisy, multipath and is subjected to fading. The channel fading conditions depend on the propagation conditions and the clutter types the waves propagate through. In some cases the fading can be more severe than Rayleigh, while in other cases where line of sight or near line of sight conditions is available, the signal will be more stable. The Nakagami distribution is shown to fit results more generally than other distributions [4]. In this section we will evaluate the average BER of HPSK systems subjected to Nakagami fading. We will evaluate the expected value of the conditional P_b as given by Eq. (22) over Nakagami distribution. In Nakagami channel the path amplitude probability density function is given by

$$f_r(r) = \frac{2}{\Gamma(m)} \left(\frac{m}{\Omega}\right)^m r^{2m-1} \exp\left(-\frac{m}{\Omega} r^2\right), \quad (23)$$

where m is the fading parameter and it describes the fading severity which is defined as the ratio of moments or it is the ratio of the square of the mean signal power to the variance of the signal power

$$m = \frac{\Omega^2}{E[r^2 - \Omega]}, \quad m \geq 0.5, \quad \Omega = E[r^2]. \quad (24)$$

The received signal power γ follows gamma distribution and its pdf is given by

$$f_r(\gamma) = \left(\frac{m}{\Omega}\right)^m \frac{\gamma^{m-1}}{\Gamma(m)} \exp\left(-\frac{m}{\Omega} \gamma\right); \quad \gamma \geq 0, \quad m \geq 0.5.$$

The average probability of error will be given by

$$P(e) = \int_0^\infty P(e/\gamma) f_r(\gamma) d\gamma = \int_0^\infty Q(\sqrt{4\gamma}) f_r(\gamma) d\gamma. \quad (25)$$

The integral of the average probability is evaluated in [5]. The average probability of error is finally given by

$$P_e = \left(\frac{1}{2\sqrt{\pi}}\right) \frac{\Gamma\left(m + \frac{1}{2}\right)}{\Gamma(m+1)} \left(\frac{2m}{2m+4\Omega}\right)^m \times {}_2F_1\left(\frac{1}{2}, m; m+1; \frac{2m}{2m+4\Omega}\right), \quad (26)$$

where ${}_2F_1(a, b; c, x)$ is the Gauss hypergeometric function [6] defined by

$${}_2F_1(a, b; c, x) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{x^n}{\Gamma(n+1)} \quad (27)$$

or with integral representation:

$${}_2F_1(a, b; c, x) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 t^{b-1} (1-t)^{c-b-1} (1-xt)^{-a} dt, \quad (28)$$

where $c > b > 0$.

The integral representation of ${}_2F_1$ is valid under the assumption that $|x| \leq 1$. In our case, it reduces to:

$${}_2F_1\left(\frac{1}{2}, m; m+1; x\right) = \frac{\Gamma(m+1)}{\Gamma(m)} \int_0^1 \frac{t^{m-1}}{\sqrt{(1-xt)}} dt. \quad (29)$$

The average BER can then be given by

$$P_e = \left(\frac{1}{2\sqrt{\pi}}\right) \frac{\Gamma\left(m+\frac{1}{2}\right)}{\Gamma(m)} x^m \int_0^1 \frac{t^{m-1}}{\sqrt{(1-xt)}} dt, \quad (30)$$

where

$$x = \frac{2m}{2m+4\Omega}. \quad (31)$$

The substitution $(t = \sin 2(\theta)/x)$, is useful to solve the integral. Finally, we have for arbitrary value of m :

$$P_e = \left(\frac{1}{\sqrt{\pi}}\right) \frac{\Gamma\left(m+\frac{1}{2}\right)}{\Gamma(m)} \int_0^{\arcsin(\sqrt{x})} \sin^{2m-1} \theta \cdot d\theta \quad (32)$$

and for integer m , Eq. (32) will result into:

$$P_e = \left(\frac{1}{\sqrt{\pi}}\right) \frac{\Gamma\left(m+\frac{1}{2}\right)}{\Gamma(m)} \sum_{n=0}^{m-1} \binom{m-1}{n} \frac{(-1)^n}{2n+1} \left[1 - y^{(n+\frac{1}{2})}\right], \quad (33)$$

where

$$y = \sqrt{1-x} = \sqrt{\frac{4\Omega}{2m+4\Omega}}. \quad (34)$$

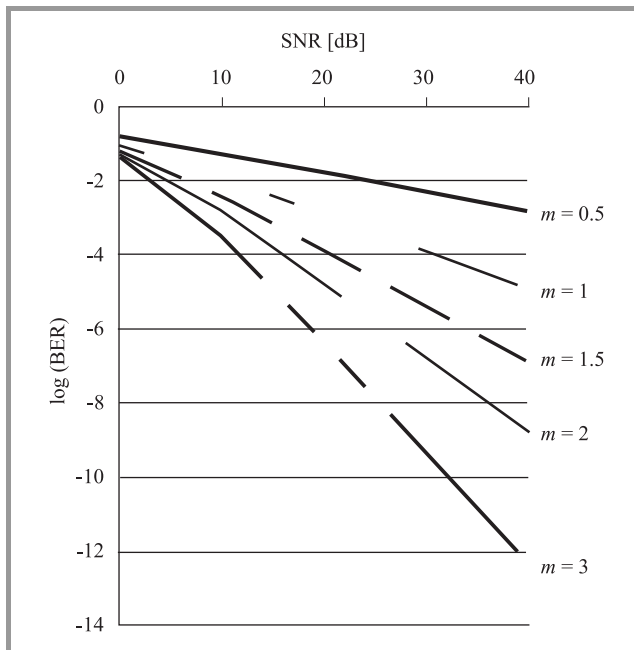


Fig. 7. The BER versus SNR of HPSK in Nakagami fading channels.

Special cases. The probability of error is computed for different value of m (Fig. 7). For the case of severe fading:

- $m = \frac{1}{2}$ (the half Gaussian distribution) Eq. (32) reduces to:

$$P_e = \arcsin(\sqrt{x})/\pi, \quad (35)$$

- $m = 3/2$, Eq. (33) will reduce to:

$$P_e = \left(\arcsin(\sqrt{x}) - \sqrt{x(1-x)}\right)/\pi, \quad (36)$$

- $m = 1$ (the case of Rayleigh fading), the average probability of error in Eq. (33) reduces to:

$$P_e = (1-y)/2, \quad (37)$$

- $m = 2$, Eq. (33) will result to:

$$P_e = (2-3y+y^3)/4, \quad (38)$$

- $m = 3$ (very close to Rician fading), Eq. (33) will result to:

$$P_e = (8 - (15y - 10y^3 + 3y^5))/16. \quad (39)$$

5. Conclusion

In this paper we come to an easy to evaluate expressions for the BER of HPSK performance in Nakagami fading channel as shown in the previous section. The cases of more severe fading than Rayleigh ($m = 1$) and Rician ($m = 3$) fading are considered. Figure 7 shows the probability of error of HPSK for different values of m . Also we proved that HPSK has SNR advantage of nearly 3 dB over QPSK.

References

- [1] "HPSK spreading for 3G", Application Note. Agilent Technol. Inc., 2003.
- [2] J. G. Proakis, *Digital Communications*. New York: McGraw-Hill, 2001.
- [3] Lowell Hoover, "Deriving the Quadratic Modulators", PAN101A & AN102A. Bloomington: Polyphase Microwave Inc., 2005.
- [4] U. Charach, "Reception through Nakagami fading multipath channels with random delays", *IEEE Trans. Commun.*, vol. Com-27, no. 4, pp. 657-670, 1979.
- [5] E. K. Al Hussaini and A. Al-Bassioni, "Performance of MRC diversity systems for the detection of signals with Nakagami fading", *IEEE Trans. Commun.*, vol. Com-33, no. 12, pp. 1315-1319, 1985.
- [6] L. C. Andrews, *Special Functions for Engineers and Applied Mathematicians*. London: Macmillan Publ., 1985.

Note: this text is reprinted from JTIT 3/2010



Mahmoud Youssuf is a Ph.D. candidate received his Master of science degree in electrical engineering at Alazhar University, Egypt. He is a Lecturer at the Department of Systems and Computers at Alazhar University. His research interests include issues related to wireless communications, federal networks, satellite commu-

nication applications. He is an author of research studies published at national and international journals, conference proceedings.

e-mail: mahmoud.youssuf@gmail.com

Alazhar University
Faculty of Engineering
Department of Systems and Computers
Republic Palace Square
Alawkaf Buildings A-1, 4th floor
Hadyek Elqubba, Cairo, Egypt

Mohamed Zaki Abdelmageed received his Ph.D. degree in electrical engineering. He is the chief of the Department of Systems and Computers at Alazhar University, Egypt. His research interests are related to networks, wireless communications and computer systems. He has published research papers at different national and international journals, conference proceedings.

e-mail: azhhar@mail.eun.eg

Alazhar University
Faculty of Engineering
Department of Systems and Computers
Republic Palace Square
Alawkaf Buildings A-1, 4th floor
Hadyek Elqubba, Cairo, Egypt

Charging Phenomena at the Interface Between High- k Dielectrics and SiO_x Interlayers

Olof Engström, Bahman Raeissi, Johan Piscator, Ivona Z. Mitrovic, Stephen Hall,
Heinrich D. B. Gottlob, Mathias Schmidt, Paul K. Hurley, and Karim Cherkaoui

Abstract—The transition regions of $\text{GdSiO}/\text{SiO}_x$ and $\text{HfO}_2/\text{SiO}_x$ interfaces have been studied with the high- k layers deposited on silicon substrates. The existence of transition regions was verified by medium energy ion scattering (MEIS) data and transmission electron microscopy (TEM). From measurements of thermally stimulated current (TSC), electron states were found in the transition region of the $\text{HfO}_2/\text{SiO}_x$ structures, exhibiting instability attributed to the flexible structural molecular network expected to surround the trap volumes. The investigations were focused especially on whether the trap states belong to an agglomeration consisting of a single charge polarity or of a dipole constellation. We found that flat-band voltage shifts of MOS structures, that reach constant values for increasing oxide thickness, cannot be taken as unique evidence for the existence of dipole layers.

Keywords—defects, dielectrics, high- k , metal oxide semiconductor.

1. Introduction

The gate function of future metal oxide semiconductor (MOS) transistors has attracted a large scientific community to an expedition into the periodic system for tracking the Dielectric Grail. Wanted is a material with acceptable energy offset values, ΔE , between the energy bands of the dielectric and the silicon crystal while, in addition, having a high enough dielectric constant, k . Yet, to fulfill the demands of low current leakage and high capacitive coupling between gate metal and transistor channel, the crucial property is the product $k \times \Delta E$ of these two quantities [1]. So far, for CMOS applications most of the efforts have been limited to metal oxides. The change from the extremely well mastered thermal SiO_2 material, to an oxide based on metals among the transition or rare earth series, has disclosed obstacles that were unnoticeable for traditional technology. Beside the problems of chemical stability between these new “high- k ” oxides and the silicon substrate, crystallization, sensitivity to humid environment, higher concentrations of oxide traps and interface states are properties, not uncommon among these materials. Driven by technology, this has given rise to needs for understanding their microscopic properties from chemical, physical and electrical point of view.

A common attribute of high- k oxide films deposited on silicon is the occurrence of an SiO_x interlayer between

the high- k material and the silicon crystal. This evokes interface electron state properties similar to those at thermal SiO_2/Si interfaces [2], [3]. Even if the interlayer lowers the effective k value of the film, it often gives better conditions for a transistor channel than those offered by a direct interface due to lower charge carrier scattering by the former. However, it must be paid for by an extra interface occurring between SiO_x and the high- k material [4]. As the total physical thickness of the film is in the range of 5 nm or smaller, on this length scale the transition from SiO_x to the high- k material can hardly be considered abrupt. It is found to include a transition region with undefined stoichiometry and thus with possible structural instabilities [5]–[11]. Recently, the occurrence of traps either singular or in dipole configurations have been noticed as the potential origin of charge sources decreasing the quality of presumptive gate insulators [11]–[16]. In the present paper, we will describe the physical and electrical properties of transition regions at $\text{GdSiO}/\text{SiO}_x$ and $\text{HfO}_2/\text{SiO}_x$ interfaces and demonstrate how charge carrier traffic at such positions can be interpreted in order to characterize the trap properties. Based on this reasoning, the possible existence of a dipole layer in the transition region will be addressed.

2. Properties of Transition Regions

The dominating defect causing charge carrier traps in the bulk of transition and rare-earth metal oxides is commonly considered to be the oxygen vacancy. It has similar properties as the E' center in SiO_2 and has been the object of a rich theoretical [15], [16] and experimental [10], [11] literature. In HfO_2 it has been predicted to be an amphoteric center with four [15] or five [16] charge states ranging from double donor to double acceptor behavior with the latter states positioned in the range 1–2 eV from the HfO_2 conduction band edge. Furthermore, the double acceptor level is argued to be connected with large lattice relaxation [16]. It has even been proposed that the double negatively charged energy level falls below the single negatively charged level, thus exhibiting negative-U property [17]. These trap properties will be further discussed below in relation to the results from electrical measurements.

As one example of the complex high- k oxide/ SiO_x interface, we will discuss data from GdSiO . This dielectric was

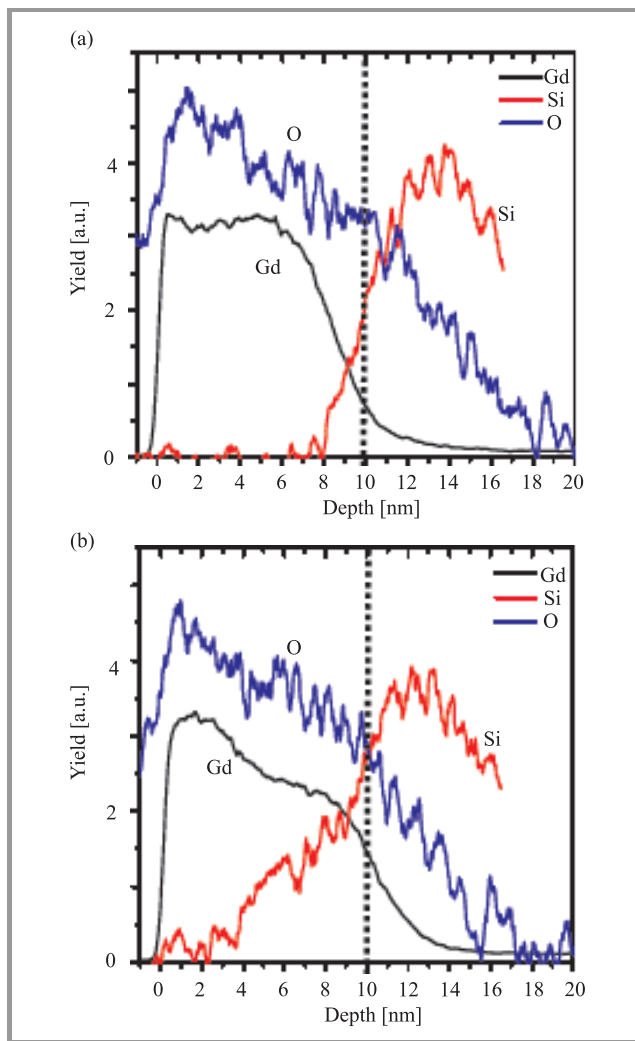


Fig. 1. MEIS data for samples with GdSiO evaporated on top of a 4 nm thermal oxide: (a) as deposited and (b) after RTA at 900°C for 1 s.

prepared by evaporating Gd_2O_3 on a 4 nm thick thermal SiO_2 layer on silicon [18]–[21]. The double layer was partly transformed into GdSiO by rapid thermal anneal (RTA) for 1 s at 900°C. The depth distribution of elements before and after the thermal anneal is shown by medium energy ion scattering (MEIS) data in Fig. 1. A steeply decreasing silicon concentration from the silicon side into the oxide directly after deposition (Fig. 1(a)), reflects a complicated diffusion process taking place already at this stage. At the 10 nm mark, the concentration of oxygen is a factor of 2 higher than the concentration of silicon. This indicates a reminiscence of SiO_2 which quickly becomes a suboxide at larger distances from this interface, where a nearly stoichiometric Gd_2O_3 takes over. The extremely high gradient of Si at the 10 nm point would be expected to stage structural instability. After the RTA at 900°C, silicon has penetrated the whole oxide layer (Fig. 1(b)) creating a GdSiO with varying concentration of Si. This structure has been demonstrated to fulfill the industrial target for low standby power, 22 nm double gate SOI transistors [18].

A second example is given by a transmission electron microscopy (TEM) picture showing the interface between HfO_2 and SiO_x in Fig. 2. This sample was prepared by reactive sputtering of Hf and O_2 followed by an anneal at 800°C for 10 min [12]. Between the white string showing the SiO_x layer and HfO_2 , appearing as a black area,

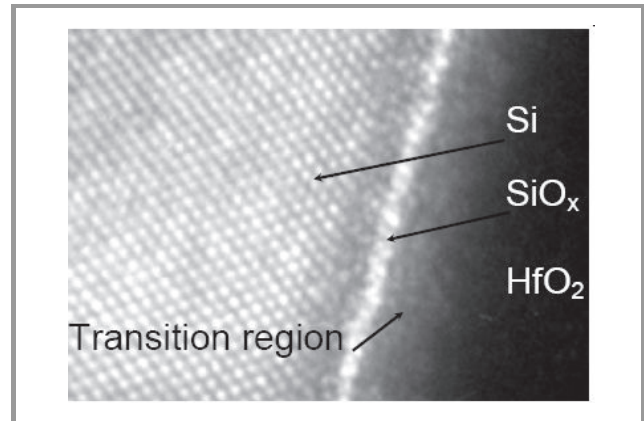


Fig. 2. TEM graph from a cross section of $\text{Si}/\text{SiO}_x/\text{HfO}_2$, where the HfO_2 layer was deposited by reactive sputtering.

a milky band is shown with a width of about 2 nm. Here, one may expect the same type of diffusion taking place as in the case of GdSiO above. As will be shown below, traps formed in this region consequently have an unusual electric behavior.

3. Charge Carrier Traffic at Transition Regions

3.1. Energy Relations and Charge Exchange

We consider an interface between a high- k oxide and an SiO_x interlayer on silicon as schematically shown in Fig. 3. The transition region is marked by a wide patterned band and includes a trap level at an energy distance ΔE below the silicon conduction band edge. A negative net charge is present in the interlayer, such that an electric field F is directed from the silicon crystal towards the high- k oxide. Assuming that the thickness of the interlayer is about 1 nm, tunneling through this layer is non-negligible and can be described as assisted by a decay of the effective density of states, N_C , from the silicon conduction band determined by $N_C \exp[-x_0/\lambda]$, where x_0 is the distance from the SiO_x/Si interface to the trap and λ is a damping factor. A captured electron can be transferred into the Si bulk by a two step process, starting with a thermally driven mechanism to the tunneling states at the energy level of the silicon conduction band edge. As long as the electric field is high enough, this process is followed by tunneling into the silicon conduction band as depicted in Fig. 3. Emitting electrons from the trap position will lower the electric field. The system

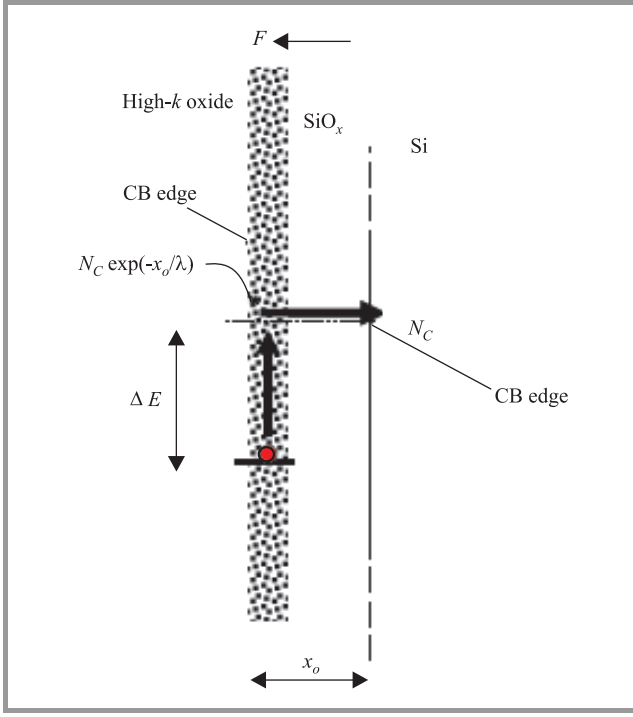


Fig. 3. Two step emission path of electrons captured in traps positioned in the transition region. The first step is thermal, to decaying states from the silicon conduction band followed by tunneling to the silicon.

may, therefore, reach a situation where the electric field becomes too low for the second transfer step to occur, which means that the emission process stops. As the tunneling probability is very sensitive to the magnitude of the electric field, this termination may be very abrupt as we will see in the experimental data following.

As long as the tunneling rate dominates, the bottleneck of this process is the thermal electron emission rate, e_n , from the trap to an effective density of states decreased by the tunneling probability

$$e_n = \sigma_n v_{th} N_C \exp\left(-\frac{x_0}{\lambda}\right) \exp\left(-\frac{\Delta E}{k_B T}\right), \quad (1)$$

where σ_n is the capture cross section representing a local transition from the tunneling state to the trap state, v_{th} is the average thermal velocity of electrons in the silicon conduction band, k_B is Boltzmann's constant and T is absolute temperature. As will be described in detail below, using thermally stimulated current (TSC) technique to determine the emission rate, the temperature is linearly increased with time during the measurement cycle. As the electric field, F , decreases due to the decrease of negative charge in the SiO_x layer, the energy distance ΔE will increase. In addition, the temperature increase will move the Fermi level in the silicon bulk to deeper energy in the band gap. For a given applied voltage across the structure, this will lower the position of the silicon conduction band edge at the SiO_x/Si interface which will increase the voltage drop across the interlayer and tend to decrease ΔE . Assuming, as a first

order approximation, that both these processes are linear with temperature, we write

$$\Delta E = \Delta E_0 - \alpha T, \quad (2)$$

where ΔE_0 is the energy distance between the silicon conduction band edge and the trap level extrapolated to $T = 0$ K and α is a constant determined by the two competing processes described above. Using Eq. (2) in (1), we get

$$e_n = \sigma_e v_{th} N_C \exp\left(-\frac{\Delta E_0}{k_B T}\right), \quad (3)$$

where

$$\sigma_e = \sigma_n \exp\left(\frac{\alpha}{k_B}\right) \exp\left(-\frac{x_0}{\lambda}\right) \quad (4)$$

can be considered as an effective capture cross section, influenced by the changing energy level position and the tunneling probability. We notice that, measuring the emission rate as a function of temperature and plotting this quantity in an Arrhenius graph would give an activation energy corresponding to an extrapolation of ΔE to $T = 0$ K under the assumption of linear conditions. However, this reasoning does not take into account the properties related to local molecular dynamics of the trap volume.

For a case like the oxygen vacancy, where the transition is argued to be connected with a strong lattice relaxation [16], the emission process would be influenced also by the vibronic properties of the trap [21] and characterized by “hysteretic tunneling”, adding a simultaneous trap relaxation and tunneling into the picture [22]. This would lead to two additional pre-exponential factors in the expressions for the thermal emission rate one originating from a possible thermally activated σ_n :

$$\sigma_n = \sigma_0 \exp\left(-\frac{\Delta U}{k_B T}\right), \quad (5)$$

where σ_0 depends on a combination of matrix elements including electronic and atomic wave functions involved in the process while ΔU is an activation energy originating from the vibrational properties of the trap system [21], [23]. The second effect comes from the entropy factor [21], [23], [24]:

$$X_n = \exp\left(\frac{\Delta S}{k_B}\right), \quad (6)$$

where ΔS is the change in entropy due to the change in vibrational frequency of the ionic part of the trap when the electron is released. Hence, combining the effects of the change in ΔE due to de-charging and those of a vibrating electron-ion trap system, we find from Eqs. (1)–(6):

$$e_n = \sigma_0 \exp\left(-\frac{\Delta U}{k_B T}\right) \exp\left(\frac{\Delta S}{k_B}\right) \exp\left(\frac{\alpha}{k_B}\right) \times \exp\left(-\frac{x_0}{\lambda}\right) v_{th} N_C \exp\left(-\frac{\Delta H_0}{k_B T}\right), \quad (7)$$

where the activation energy now is represented by an enthalpy, $\Delta H_0 = \Delta E_0 + \Delta S T$, including the heat, $\Delta S T$, stored by the local vibrational modes [21], [23], [24].

For this case the first factors in Eq. (7) make up an “effective” capture cross section, σ_e , which would be obtained from an Arrhenius plot of the thermal emission rate, e_n :

$$\sigma_e = \sigma_0 \exp\left(-\frac{\Delta U}{k_B T}\right) \exp\left(\frac{\Delta S}{k_B}\right) \exp\left(\frac{\alpha}{k_B}\right) \exp\left(-\frac{x_0}{\lambda}\right). \quad (8)$$

As will be shown below, TSC results demonstrate instabilities among the traps investigated in the $\text{HfO}_2/\text{SiO}_x$ transition region. This is most probably originating from restructuring of the molecular arrangement around the trap volume, thus changing the matrix elements for charge carrier transition involved in σ_0 as well as quantities of the vibrational properties reflected by ΔU , ΔS , and ΔH_0 in Eqs. (7) and (8).

3.2. Thermally Stimulated Current

When measuring TSC from traps in a MOS system [12], the sample is first brought into accumulation at room temperature followed by a temperature decrease to about 50 K. During this procedure, traps at the interface and in the oxide are filled by charge carriers. At the low temperature point, the system is biased into deep depletion and a temperature increase, linear as a function of time, is applied. For an n-type semiconductor this gate bias is negative. As long as the temperature is lower than about 200 K and the total scanning time up to that temperature is shorter than

about 10 min, it should be noticed that, due to the low temperature, an extremely low concentration of holes is expected in the valence band of the silicon crystal. Therefore, all de-charging processes observed as a TSC can be expected to originate from electron exchange at the insulator/silicon interface. The current created by emitted electrons is expressed by [12]

$$i(T) = \frac{C_{ox}}{C_s + C_{ox}} q N_T e_n \exp\left[-\int_{T_0}^{T_1} \beta e_n(u) dU\right], \quad (9)$$

where C_{ox} and C_s are the capacitances of the oxide and the depleted semiconductor, respectively, N_T is the surface concentration of captured carriers and β is the scanning rate of the temperature, linear in time, t , such that $T = \beta t$. The function given by Eq. (9) is plotted in Fig. 4(a) for an activation energy of 0.13 eV and an effective capture cross section of 10^{-24} cm^2 . Using the rather complicated expression in Eq. (8) for parameter extraction from experimental data is not practical. However, calculating the integral factor in this equation as a function of temperature, for the same input data as used above, one finds the graph shown in Fig. 4(b). It is noticed that the integral takes a value close to 1 for the initial part of the TSC curve in Fig. 4(a). Furthermore, as $C_{ox} \gg C_s$ for the deep depletion conditions used in the experiment, this part of the experimental data is proportional to the thermal emission rate e_n . Estimating the total concentration, N_T , of trap levels from the area under the TSC peak, therefore, gives a possibility to find effective capture cross sections from Arrhenius plots of e_n [12].

3.3. Experimental TSC Results on the $\text{HfO}_2/\text{SiO}_x$ Interface

Figure 5 shows experimental TSC data from electron emission at the transition region of an MOS capacitor with an $\text{Al}/\text{HfO}_2/\text{SiO}_x/\text{Si}$ structure like the one shown by TEM in Fig. 2. Two important features can be observed in Fig. 5:

- 1) repeated measurement gives a different TSC peak position on the temperature scale;
- 2) the TSC curves are terminated before they reach the maximum point of the theoretical curves (solid curves) fitted to the experimental data.

These results are typical and occur for a large majority of the present samples with HfO_2 prepared by reactive sputtering as well as for samples prepared by atomic layer deposition (ALD) [25]. The observation (1), reveals an instability of the traps as mentioned above in relation to Eq. (9). Such instability after voltage stress under similar conditions as in the present experiment, was observed also by Bersuker *et al.* in [10]. The observation (2), can be interpreted as a result of the tunneling process involved in the electron emission. During the emission process, the negative charge in the transition region, born by the captured

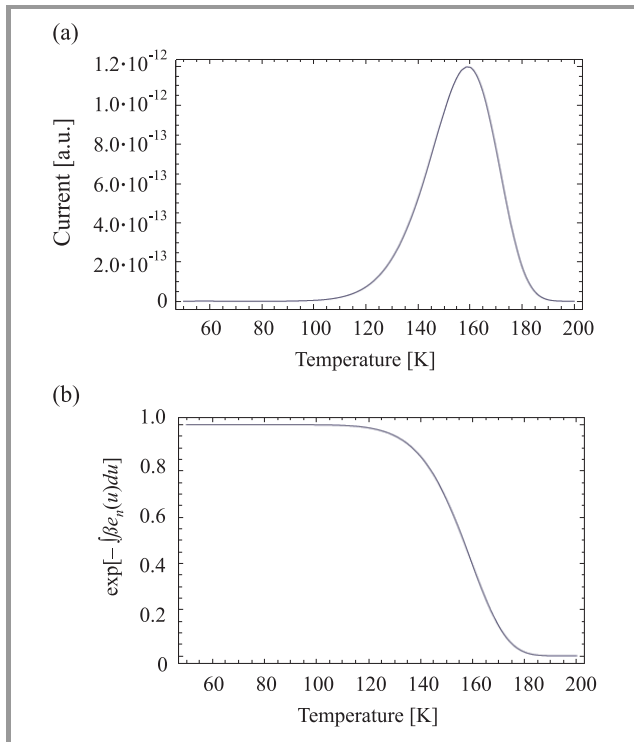


Fig. 4. (a) Theoretical plot of thermally stimulated current for an activation energy of 0.13 eV, an effective capture cross section of 10^{-24} cm^2 and a temperature scan rate of 20 K/minute. (b) The integral factor in Eq. (9) as a function of temperature. This function is close to 1 for the initial part of the curve in (a).

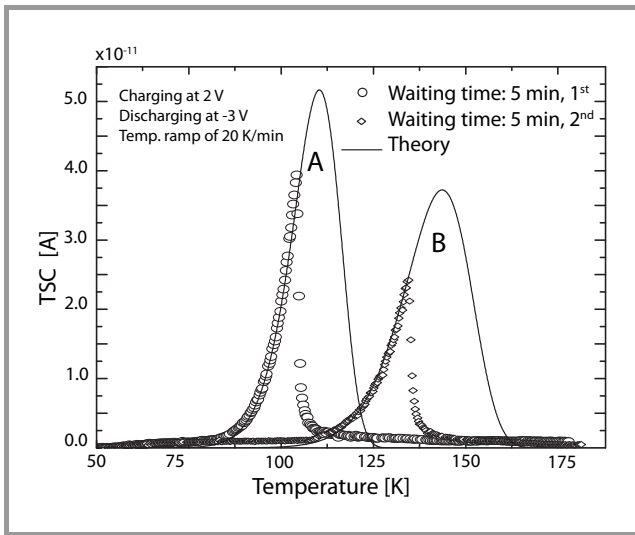


Fig. 5. Experimental TSC data (points) for MOS structures with HfO₂ prepared by reactive sputtering compared with theoretical calculations (solid curves) calculated from Eq. (9). Repeated measurement gave rise to a shift of the TSC peak along the temperature axis, reflecting structural changes of the emitting electron traps. The activation energy for the curves A and B is 0.13 eV and 0.16 eV, respectively, and the corresponding capture cross sections are $6 \cdot 10^{-22} \text{ cm}^2$ and $9 \cdot 10^{-21} \text{ cm}^2$, respectively.

electrons, will decrease. This will decrease the electric field driving the tunneling until the field strength is too weak for continued emission. As the tunneling probability is very sensitive to a change in electric field, the TSC will get an abrupt termination as shown for the two curves in Fig. 5.

4. Single Charge Versus Dipole Charge

The recent increasing interest in the properties of high-*k*/SiO_x transition layers includes novel ideas about a possible occurrence of closely separated charge planes with different polarities within this region [13]. It has been described as dipole planes occurring as a result of oxygen transfer across the interface between the high-*k* and the SiO_x material [14]. According to an idea proposed in [14], such transition would take place from the material with the highest surface density of oxygen atoms to that with lower density. For transition metal oxides, like HfO₂, this would imply that oxygen is transferred from this material into SiO_x, creating interstitials and leaving oxygen vacancies behind. As the relation between oxygen surface densities of rare-earth metal oxides, like Gd₂O₃ and SiO_x is the opposite, such transfer would instead go from the SiO_x to the high-*k* side.

The influence of a single negative charge and a dipole surface on the conduction band relations for the metal/HfO₂/SiO_x/n-type Si structure is depicted in Fig. 6. The geometries are shown in Fig. 6(a), while Fig. 6(b) and Fig. 6(c) show the charge relations and the conduction band relations for an open circuit case and a short circuit case, respectively. In order to demonstrate the specific influences of these charge planes, we assume that no other charges

are present in the structure. Considering first the single charge case in the left column of Fig. 6 for an open circuit case in Fig. 6(b), where the opposite positive charge is assumed to exist at a long distance, the electron energy of the whole structure is lifted in parallel in relation to an earth plane. Short circuiting, as shown in Fig. 6(c), will cause positive charge to appear at the metal gate and in the depletion region occurring in the n-type semiconductor. Compared with an ideal structure without charge, an increased positive voltage, V_{FB} , on the gate would be needed to obtain flat-band condition for this case. Furthermore, V_{FB} would increase linearly with increasing thickness of the HfO₂ layer.

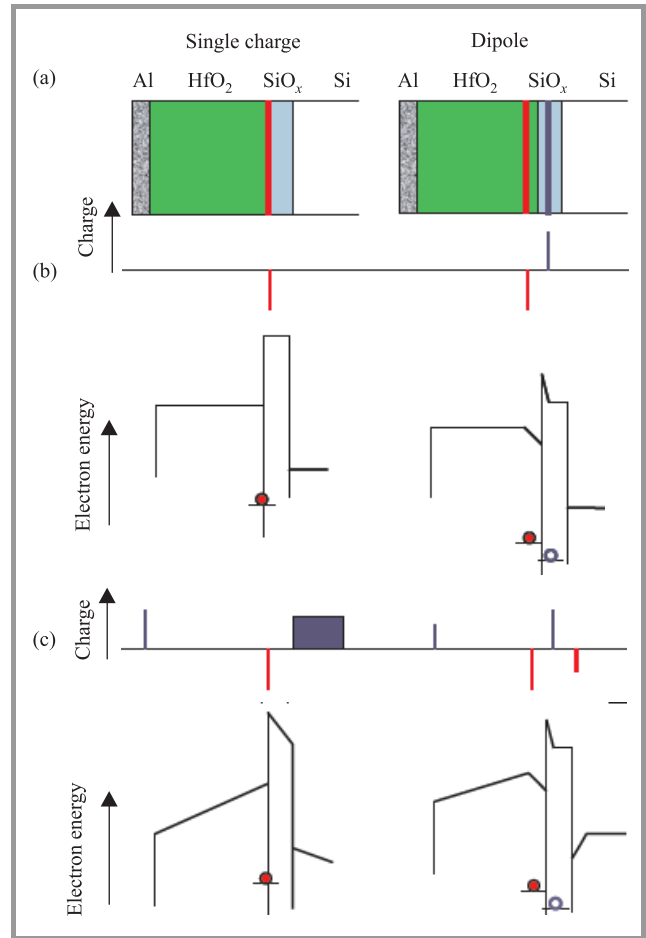


Fig. 6. Illustration of (a) charge configuration, and conduction band relations for (b) open circuit and (c) short circuit conditions of a metal/high-*k*/SiO_x/Si structure. The columns demonstrate the conditions for a single charge plane (left) and a dipole plane (right), respectively.

For a dipole layer at the HfO₂/SiO_x interface, as shown in the right column of Fig. 6, under open-circuit conditions, one would expect an electric field between the two charge planes only (Fig. 6(b)). This would create a potential drop in that domain, while constant potentials would occur outside the two planes separated by a voltage created inside the dipole. Short-circuiting this structure gives rise to a positive charge on the gate and a negative charge in

the semiconductor, forcing the latter into accumulation condition. These two charges are exactly equal and vary depending on the HfO_2 thickness in such a way that the sum of the energy drops in HfO_2 and in the semiconductor is constant and equal to qV_{FB} . In this case a negative V_{FB} , equal to the potential drop inside the dipole, is needed to obtain flat-band conditions. However, contrary to the situation with a single negative charge, the dipole combination would give a V_{FB} which is independent of the thickness of HfO_2 . Such a behavior was indicated in [14] for HfO_x/Si_x interfaces.

The single charge case demonstrated by Fig. 6(a) offers a model for straightforward explanation of the anomalous TSC data as discussed in Section 3. The dipole case in Fig. 6(b) likewise gives rise to an energy relation of the conduction bands supporting electron injection into the silicon and thus a TSC with similar behavior as that from a single negative charge plane. Therefore, TSC does not give direct information on which of these two charge constellations is the source of current.

5. Discussion

In order to explain the saturating flat-band voltage and its positive sign as observed in [12] when increasing the thickness of HfO_2 , the idea put forward in [13] requires injection of negatively charged traps from the high- k side into SiO_x . This would give rise to a dipole directed in the opposite way to that discussed in relation to Fig. 6. Such a model is not completely unproblematic. First, injecting an oxygen ion from the HfO_2 into the SiO_x layer might be expected to give rise to an interstitial, leaving behind a vacancy in the HfO_2 . According to recent theoretical results [15], [16], vacancies are amphoteric and act as acceptors with energy levels at about 1.5 eV from the HfO_2 conduction band. They may therefore tend to be filled by electrons from the silicon conduction band and thus become negatively charged. On the other hand, the donor levels connected with oxygen vacancies are expected close to the middle of the bandgap of HfO_2 and are most probably occupied by electrons and neutral at voltages of the flat-band values at about 0.5 V in [13]. The origin of positive charge required on the HfO_2 side, therefore, is questionable.

A second problem with a dipole model may be the quantity of charge needed to obtain V_{FB} shifts in the range of 0.3–0.5 V as observed in [13]. As the thickness of the SiO_x layer normally is about 1 nm, the maximum value for the distance between the two charge sheets of an assumed dipole would be about that value. In order to achieve a shift of 0.3 V, this requires a charge density of more than $5 \cdot 10^{15} q \text{ As cm}^{-2}$. It can be compared with the charge contained in the TSC curves of Fig. 5, which is in the range of $10^{13} q \text{ As cm}^{-2}$. Finally, the strongest argument for a dipole model might be the saturating flat-band voltage. However, this result does not uniquely lead to a dipole configuration as shown in the following. Taking into account the continuous character of the transition region, a satu-

rating V_{FB} may result also from single charge condition. Contrary to the assumption made in [13], the transition from HfO_2 to SiO_x cannot be considered abrupt on length scales in the range of a few nm. As noticed in Fig. 1, the oxygen concentration increases with distance from the silicon side due to diffusion into the silicon crystal. Similar behavior of oxygen concentration has been observed by elastic recoil detection analysis (ERDA) studies on HfPrO samples [26]. Therefore, it is reasonable to assume that the concentration of oxygen vacancies decreases with distance from the silicon side. Furthermore, assuming that thinner HfO_2 layers, due to the out-diffusion of oxygen to the surface, also as observed in Fig. 1, have a lower decay of vacancies, it is probable that they have a smaller decay in concentration.

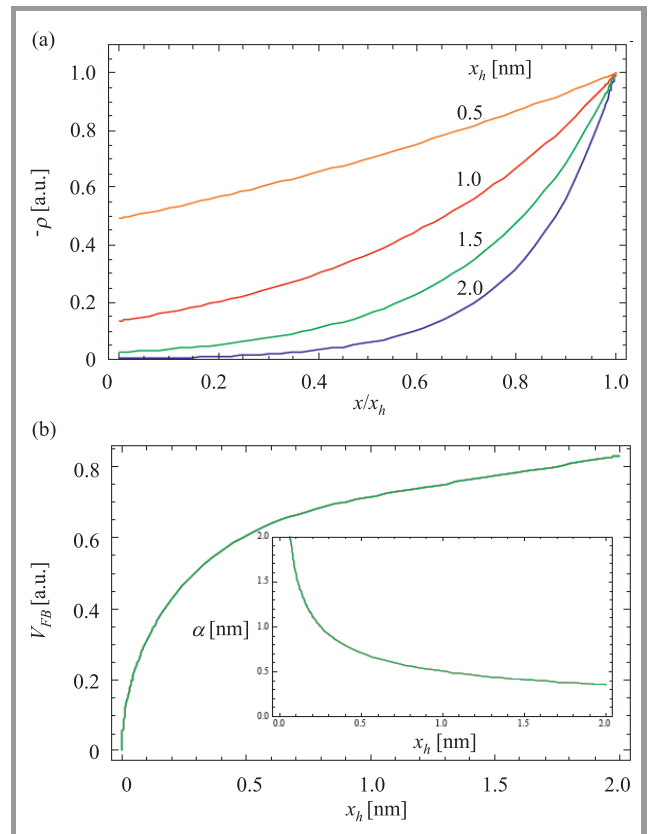


Fig. 7. (a) Assumed depth distribution of acceptor state volume concentrations in the transition region for HfO_2 with thicknesses of 0.5, 1.0, 1.5 and 2.0 nm. The point $x = 0$ is at the HfO_2 surface, while $x = x_h$ is at a reference point where SiO_2 transfers to suboxides. (b) The flat-band voltage as a function of HfO_2 thickness for acceptor concentrations as shown in (a) with the decay constant, α , varying as a function of x_h as shown in the inset of (b).

A theoretical example is shown in Fig. 7(a), where the concentration of traps is plotted for different thicknesses as a function of normalized distance from a defined reference point in the transition region. Such a point could be the one, where SiO_2 transfers to suboxides. For simplicity, the curves in Fig. 7(a) are exponential functions with decay factors, α , decreasing with the thickness x_h of HfO_2 , as

shown in the inset of Fig. 7(b). Solving Poisson's equation for the functions in Fig. 7(a):

$$\frac{d^2V}{dx^2} = -\frac{\rho_0}{k\epsilon_0} \exp\left(-\frac{x_h - x}{\alpha}\right), \quad (10)$$

where V is electrical potential, ρ_0 is the volume charge density at the reference point, x_h , ϵ_0 is the dielectric permittivity of vacuum and x is distance from the HfO_2 surface, we find V_{FB} from the potential at $x = 0$ as

$$\frac{k\epsilon_0}{\rho_0} V_{FB} = -\alpha \left\{ x_h + \alpha \left[1 - \exp\left(-\frac{x_h}{\alpha}\right) \right] \right\}. \quad (11)$$

The V_{FB} normalized by the pre-factor in Eq. (11) is plotted in Fig. 7(b) as a function of HfO_2 thickness, x_h . The shape of this curve, with a tendency to saturate for increasing, x_h , was interpreted in [12] as a result of a dipole layer. Figure 7(b) demonstrates that under assumptions based on MEIS data, it is possible to obtain similar behavior from a singly charged layer.

6. Conclusions

The theoretical TSC curves shown in Fig. 5 were calculated assuming an ensemble of electron states with one common discrete energy level. The agreement between the initial curvature of these graphs and the experimental data indicates that the majority of traps observed by TSC have similar properties of their charge carrier statistics and are positioned with limited spread in distance from the silicon/ SiO_x interface. This excludes bulk traps in the HfO_2 as candidates and suggests a trap distribution concentrated at the $\text{HfO}_2/\text{SiO}_x$ interface as depicted in Fig. 7(a) for delivering electrons to the TSC. Also, TSC peaks from states at the Si/SiO_x interface are ruled out by earlier experiments as they have been demonstrated to occur at temperatures above 200 K, outside the scale of Fig. 5 [12]. The extremely low values of capture cross sections in the range $10^{-26} - 10^{-19} \text{ cm}^2$, normally found by TSC for this kind of experiments [11], [27], are explained by the tunneling process needed for carrier injection into the silicon substrate and the thermally activated process expected for carrier capture into oxygen vacancies [16]. Also, taking into account the flexible structural network anticipated in the transition region, the shift of the TSC peaks as shown in Fig. 5 may be assigned to the restructuring of oxygen vacancy defects as discussed in Section 2.

The argument in [13], for the existence of a dipole plane at the interface between HfO_2 and SiO_2 layers, was based on the observation of flat-band voltage as a function of the thickness of HfO_2 . It was shown that V_{FB} saturated to a near constant value when this thickness increases as expected for a dipole charge and discussed above in connection with Fig. 6. However, such a result is not unique for a dipole. By assuming an oxygen vacancy distribution in the high- k layer as estimated from MEIS data, we have demonstrated in the present work that similar V_{FB} dependence may be the result of a single charge distribution. We conclude, therefore, that more evidence is necessary before

the existence of a dipole layer occurring in high- k/SiO_x transition regions is confirmed.

Acknowledgment

The authors collaborate under the banner of the "High- k -Gang" (<http://www.high-k-gang.eu/>). The work has benefited from funding provided by the European Network of Excellence NANOSIL within FP7 (ICT-216171).

References

- [1] O. Engström, B. Raciassi, S. Hall, O. Bui, M. C. Lemme, H. D. B. Gottlob, P. K. Hurley, and K. Cherkaoui, "Navigation aids in the search for future high- k dielectrics: physical and electrical trends", *Solid-State Electron.*, vol. 51, iss. 4, pp. 622–626, 2007.
- [2] B. Raciassi, J. Piscator, O. Engström, S. Hall, O. Bui, M. C. Lemme, H. D. B. Gottlob, P. K. Hurley, K. Cherkaoui, and H. J. Osten, "High- k -oxide/silicon interfaces characterized by capacitance frequency spectroscopy", *Solid-State Electron.*, vol. 52, iss. 9, pp. 1274–1279, 2008.
- [3] P. K. Hurley, K. Cherkaoui, E. O'Connor, M. C. Lemme, H. D. B. Gottlob, M. Schmidt, S. Hall, Y. Lu, O. Bui, B. Raciassi, J. Piscator, O. Engström, and S. B. Newcomb, "Interface defects in HfO_2 , LaSiO_x and Gd_2O_3 high- k -metal-gate structures on silicon", *J. Electrochem. Soc.*, vol. 155, no. 2, pp. G13–G20, 2008.
- [4] M. A. Quevedo-Lopez, P. D. Kirsch, S. Krishnan, H. N. Alshareef, J. Barnett, H. R. Harris, A. Neugroschel, F. S. Aguirre-Tostado, B. E. Gnade, M. J. Kim, R. M. Wallace, and B. H. Lee, "Systematic gate stack optimization to maximize mobility with HfSiON EOT scaling", in *Proc. ESSDERC 2006 Conf.*, Montreux, Switzerland, 2006, pp. 113–116.
- [5] G. Lucovsky, Y. Wu, H. Niimi, V. Misra, and J. C. Phillips, "Bonding constraints and defect formation at interfaces between crystalline silicon and advanced single layer composite gate dielectrics", *Appl. Phys. Lett.*, vol. 74, no. 14, pp. 2005–2007, 1999.
- [6] G. Lucovsky, J. P. Maria, and J. C. Phillips, "Interfacial strain induced self-organization in semiconductor dielectric gate stacks. II. Strain relief at internal dielectric interfaces between SiO_2 and alternative dielectrics", *J. Vac. Sci. Technol. B*, vol. 22, iss. 4, pp. 2097–2104, 2004.
- [7] G. Lucovsky and J. C. Phillips, "Defects and defect relaxation at internal interfaces between high- k transition metal and rare earth dielectrics and interfacial native oxides in metal oxide semiconductor (MOS) structures", *Thin Solid Films*, vol. 486, iss. 1–2, pp. 200–204, 2005.
- [8] F. Giustino, A. Bongiorno, and A. Pasquarello, "Equivalent thickness of thin oxide interlayer in gate insulator stacks on silicon", *Appl. Phys. Lett.*, vol. 86, iss. 19, pp. 192901–1–3, 2005.
- [9] P. Broqvist and A. Pasquarello, "Band gaps and dielectric constants of amorphous hafnium silicates: a first principle calculation", *Appl. Phys. Lett.*, vol. 90, iss. 8, pp. 082907–1–3, 2007.
- [10] G. Bersuker, C. S. Park, J. Barnett, P. S. Lysaght, P. D. Kirsch, C. D. Young, R. Choi, B. H. Lee, B. Foran, K. van Benthem, S. J. Pennycook, P. M. Lenahan, and J. T. Ryan, "The effect of interfacial layer properties on the performance of Hf-based gate stack devices", *J. Appl. Phys.*, vol. 100, p. 094108, 2006.
- [11] J. T. Ryan, P. M. Lenahan, G. Bersuker, and P. Lysaght, "Electron spin resonance observations of oxygen deficient silicon atoms in the interfacial layer of hafnium oxide based metal-oxide-silicon structures", *Appl. Phys. Lett.*, vol. 90, p. 173513, 2007.
- [12] B. Raciassi, Y. Y. Chen, J. Piscator, Z. H. Lai, and O. Engström, "Electron traps at $\text{HfO}_2/\text{SiO}_x$ interfaces", in *Proc. ESSDERC 2008 Conf.*, Edinburgh, Scotland, 2008, pp. 130–133.
- [13] K. Iwamoto, Y. Kamimuta, A. Ogawa, Y. Watanabe, S. Migita, W. Mizubayashi, Y. Morita, M. Takahashi, H. Ota, T. Nabatame, and A. Toriumi, "Experimental evidence for the flatband voltage shift of high- k metal-oxide-semiconductor devices due to the dipole formation at the high- k/SiO_2 interface", *Appl. Phys. Lett.*, vol. 92, iss. 13, pp. 132907–1–3, 2008.

- [14] K. Kita and A. Toriumi, "Origin of electric dipoles formed at high- k /SiO₂ interface", *Appl. Phys. Lett.*, vol. 94, iss. 13, pp. 132902-1--3, 2009.
- [15] K. Xiong, J. Robertson, M. C. Gibson, and S. J. Clark, "Defect energy levels in HfO₂ high-dielectric constant gate oxide", *Appl. Phys. Lett.*, vol. 87, iss. 18, pp. 183505-1--3, 2005.
- [16] J. L. Gavartin, D. Muñoz Ramo, A. L. Shluger, G. Bersuker, and B. H. Lee, "Negative oxygen vacancies in HfO₂ as charge traps in high- k stacks", *Appl. Phys. Lett.*, vol. 89, iss. 8, pp. 082908-1--3, 2006.
- [17] Y. P. Feng, A. T. Lim, and M. F. Li, "Negative-U property of oxygen vacancy in cubic HfO₂", *Appl. Phys. Lett.*, vol. 87, iss. 6, pp. 062105-1--3, 2005.
- [18] H. D. B. Gottlob, M. Schmidt, A. Stefani, M. C. Lemme, H. Kurz, I. Z. Mitrovic, W. M. Davey, S. Hall, M. Werner, P. R. Chalker, K. Cherkaoui, P. K. Hurley, J. Piscator, O. Engström, and S. B. Newcomb, "Scaling potential and MOSFET integration of thermally stable Gd silicate dielectrics", *Microelectron. Eng.*, vol. 86, iss. 7-9, pp. 1642-1645, 2009.
- [19] H. D. B. Gottlob, A. Stefani, M. Schmidt, M. C. Lemme, H. Kurz, I. Z. Mitrovic, M. Werner, W. M. Davey, S. Hall, P. R. Chalker, K. Cherkaoui, P. K. Hurley, J. Piscator, O. Engström, and S. B. Newcomb, "Gd silicate: a high- k dielectric compatible with high temperature annealing", *J. Vac. Sci. Technol. B*, vol. 27, iss. 1, pp. 249-252, 2009.
- [20] I. Z. Mitrovic and S. Hall, "Rare earth silicate formation – a route towards high- k for the 22 nm node and beyond", *J. Telecommun. Inform. Technol.*, no. 4, pp. 51-60, 2009.
- [21] O. Engström and A. Alm, "Energy concepts of insulator-semiconductor interface traps", *J. Appl. Phys.*, vol. 54, no. 9, pp. 5240-5244, 1983.
- [22] W. B. Fowler, J. K. Rudra, M. E. Zvanut, and F. J. Fiegl, "Hysteresis and Franck-Condon relaxation in insulator-semiconductor tunnelling", *Phys. Rev. B*, vol. 41, no. 12, pp. 8313-8317, 1990.
- [23] O. Engström and H. G. Grimmeiss, "Vibronic states of silicon dioxide interface traps", *Semicond. Sci. Technol.*, vol. 4, no. 12, pp. 1106-1115, 1989.
- [24] O. Engström, T. Gutt, and H. M. Przewlocki, "Energy concepts involved in MOS characterization", *J. Telecommun. Inform. Technol.*, no. 2, pp. 86-91, 2007.
- [25] M. Johansson, "Silicon device substrate and channel characteristics influenced by interface properties", Ph.D. thesis, Chalmers University of Technology, Göteborg, Sweden, 2005.
- [26] B. Raeissi, "Charge carrier traffic at interfaces in nanoelectronic structures", Ph.D. thesis, Chalmers University of Technology, Göteborg, Sweden, 2010.
- [27] M. Y. A. Yousif, M. Johansson, and O. Engström, "Extremely small hole capture cross sections in HfO₂/Hf_xSi_yO_z/p-Si structures", *Appl. Phys. Lett.*, vol. 90, iss. 20, pp. 203506-1--3, 2007.

Note: this text is reprinted from JTIT 1/2010



Olof Engström received the Ph.D. degree in solid state physics from the University of Lund in 1975 and was later employed by ASEA AB for research on high power thyristors, by AB Rifa for development of MOS technology and by the Swedish Defence Research Institute for sensor research. In 1984, he came to Chalmers Uni-

versity of Technology as a Professor in solid state electronics. Between 1996 and 1999, he served as Dean of Chalmers School of Electrical and Computer Engineering and 1999-2002 he was the Director of the Microtechnology Center at Chalmers (MC2). From 2003 he is back in research as a Professor at the Department of Microtechnology and Nanoscience of MC2. His present research interest is in, high- k -materials and quantum dots and Schottky barriers on silicon nanowires. In 1991, he founded Samba Sensors AB, a company for production of fiberoptical pressure sensors. He is a member of the Royal Swedish Academy of Engineering Science, the Finnish Society of Science and the High- k -Gang.

e-mail: olof.engstrom@chalmers.se

Chalmers University of Technology

Department of Microtechnology and Nanoscience

SE-412 96 Göteborg, Sweden



Bahman Raeissi received the B.Sc. degree in physics from Isfahan University of Technology, Isfahan, Iran, in 2002 and M.Sc. degree in nanoscale physics and engineering, Chalmers University of Technology, Göteborg, Sweden, in 2005. He is currently a Ph.D. student at the Department of Microtechnology and Nanoscience (MC2),

Chalmers University of Technology. His main research interests are fabrication and characterization of high- k dielectrics, semiconductor wafer bonding and characterization of quantum dots.

e-mail: bahman.raeissi@chalmers.se

Chalmers University of Technology

Department of Microtechnology and Nanoscience

SE-412 96 Göteborg, Sweden



Johan Piscator received the M.Sc. degree and the Ph.D. degree in electronic engineering from Chalmers University of Technology, Göteborg, Sweden, in 2003 and 2009, respectively. His main research interests are fabrications and characterization of silicon nanowires and characterization of high- k oxides for CMOS applications.

e-mail: johan.piscator@gmail.com

Chalmers University of Technology

Department of Microtechnology and Nanoscience

SE-412 96 Göteborg, Sweden



Ivona Z. Mitrovic received the Ph.D. degree in electronic engineering from the University of Liverpool, UK, in 2007, the M.Sc. degree in materials science from the University of Belgrade in 2002, and Dipl.-Ing. degree in microelectronics from the Faculty of Electronic Engineering, University of Nis, Serbia, Yugoslavia, in 1997. She

took part in a research project concerning BaTiO₃ ceramics (1997–2001), worked as a Research Assistant (2001–2007) and a Research Associate at the University of Liverpool (2000–2009). Since June 2009, she is a Lecturer in the Solid State Electronics Research Group, Department of Electrical Engineering and Electronics, University of Liverpool. Her research interests span materials for beyond 22 nm technological node targeting energy harvesting products for medical, automotive and aerospace applications, as well as emerging technologies for energy conversion and storage.

e-mail: ivona@liverpool.ac.uk

Department of Electrical Engineering and Electronics
University of Liverpool
Brownlow Hill
Liverpool L69 3GJ, United Kingdom



Stephen Hall has been Head of the Department of Electrical Engineering and Electronics at the University of Liverpool, UK, from 2001 to the present date. He has interests spanning materials characterization, device physics and innovative device design and gate level circuits. He has over 200 conference and journal papers in the

area of silicon technology, devices and circuits. These include novel measurements and contributions to the understanding of MOS related interfaces and materials quality. He has successfully designed and built novel MOS and bipolar devices in silicon for about 20 years. More recently, his work encompasses gate level circuits relating to low voltage/low power SOI, micro-power and biologically inspired concepts. He was Technical Programme Chair of ESSDERC 2008, and currently sits on the Steering Committee of ESSDERC/ESSCIRC and INFOS, for which he was vice-Chair in 2009 and is a member of the Steering Committee from 2009.

e-mail: S.Hall@liverpool.ac.uk

Department of Electrical Engineering and Electronics
University of Liverpool
Brownlow Hill
Liverpool L69 3GJ, United Kingdom



Heinrich D. B. Gottlob received his Dipl.-Ing. degree in electrical engineering and information technology from the RWTH Aachen University, Germany, in 2003 and his Dr.-Ing. degree in 2007. In 2003 he joined AMO GmbH as R&D engineer and he is currently leader of the nanoelectronics group and manager of AMO's

nanotechnology lab AMICA. Research topics include novel high-*k* dielectrics and metal gate stacks as well as fully depleted ultrathin body and non-planar SOI devices. He is author and co-author of more than 30 journal papers.

e-mail: gottlob@amo.de

Advanced Microelectronic Center Aachen (AMICA)

AMO GmbH

Otto-Blumenthal st 25

D-52074 Aachen, Germany



Mathias Schmidt received his Dipl.-Ing. degree in electrical engineering from the RWTH Aachen University, Germany, in 2004. He is currently working towards his Dr.Ing. degree at the Advanced Microelectronic Center Aachen (AMICA), AMO GmbH Aachen. His research topics include ultrathin body silicon devices with planar and

non-planar channels with special respect to experimental mobility extraction and integration of novel high-*k*/metal gate stacks and strained channel materials.

e-mail: schmidt@amo.de

Advanced Microelectronic Center Aachen (AMICA)

AMO GmbH

Otto-Blumenthal st 25

D-52074 Aachen, Germany



Karim Cherkaoui received the Ph.D. degree in 1998 from the Institut National des Sciences Appliquées of Lyon, France. During his Ph.D. he gained experience in the spectroscopy of point defects in semi-insulating materials. In 1999 he joined the Tyndall National Institute (formerly NMRC), University College Cork, where he has established and developed several low temperature electrical metrology techniques. His current research interests include the process development and characterization of high dielectric constant materials on Si and III-V materials for future MOS devices.

e-mail: karim.cherkaoui@tyndall.ie
Tyndall National Institute
University College Cork, Lee Maltings
Prospect Row, Cork, Ireland



Paul Hurley received his Ph.D. (1990) and B.Eng. (1985 – first class honors) in electronic engineering at the University of Liverpool, UK. He is a senior research scientist at the Tyndall National Institute, University College Cork, Ireland, where his work focuses on high dielectric constant (high- k) ma-

terials intended for use as gate level insulators in transistors for future integrated circuits. The emphasis of his current research is the formation and characterization of high- k films on silicon and III–V semiconductor substrates. He is a member of the Technical Committee of the Insulating Films on Semiconductors (INFOS) conference and the International Workshop on Dielectrics in Microelectronics (WoDiM). In addition to research activities, he is a part time lecturer in the Department of Micro-electronic Engineering at University College Cork. He has published over eighty papers in the field of microelectronics.

e-mail: paul.hurley@tyndall.ie
Tyndall National Institute
University College Cork, Lee Maltings
Prospect Row, Cork, Ireland

Channel Identification Using Chaos for an Uplink/Downlink Multicarrier Code Division Multiple Access System

Miloud Frikel, Said Safi, Boubekur Targui, and Mohammed M'Saad

Abstract—A scheme of chaotic spreading sequence for multicarrier code division multiple access system (MC-CDMA) is proposed to estimate the transmission channel. This system spreads spectrum and identifies the channel, simultaneously. The proposed scheme uses a chaotic sequence generated by a logistic map as a training signal and estimate channel parameters according to dynamics of the chaotic sequence. Encoding data by chaotic sequences is first built and then the orthogonal codes are used to spread the encrypted data for multiusers scheme. At the reception, first the channel parameters are identified using a training chaotic sequence in order to equalize the received data, and then the encrypted information is decoded for the desired user. The studies reveal that the proposed system (chaos plus orthogonal codes) significantly outperforms the Walsh-Hadamard code spreading in MC-CDMA system. The performance of the system is considered in the multiuser case by means of simulation. The simulation result shows that the proposed chaotic code spreading approach for channel identification achieves significant improvement in the channel identification, comparing to using others training sequence or the least square method.

Keywords—channel identification, chaos code, equalization, MC-CDMA system.

1. Introduction

The chaotic signals have some properties such as broadband, orthogonality and complexity aspects, which motivate researches in the area of communication and signal processing to investigate if chaos based communication offers advantages over classical communication systems in the last years.

Potential applications of chaos resulting directly from those aspects are spread-spectrum, multiuser communication, and cryptography [1]–[4]. In chaotic communications, a chaotic sequence is transmitted through the transmission channel [5]. If the channel is not ideal, which is often the case in practice, the transmitted signal is corrupted before it reaches the receiver. Hence, channel equalization is required to reduce the bit error rate of the receiver as small as possible.

In many practical cases, channel parameters are unknown. Hence, channel equalization must be performed from the corrupted signal alone, and this is called the blind channel

equalization [6], [7]. In classical communication systems, most of the channel equalization algorithms are based on the statistical properties of the transmitted signal [8]. However, since a chaotic sequence is a deterministic signal, the statistics-based equalization techniques will not achieve optimum estimation accuracy for chaotic communication systems because they do not take into account the inherent properties specific to a chaotic signal. Various chaotic blind identification and equalization techniques based on different properties of the transmitted chaotic signal have been developed recently [9].

The multicarrier code division multiple access (MC-CDMA) system is based on the combination of code division multiple access (CDMA) and orthogonal frequency division multiplexing (OFDM) [10]–[13] which is potentially robust to channel frequency selectivity. Furthermore, it has good spectral efficiency, multiple access capability and it's easy to be implemented with fast Fourier transform (FFT). For the scheme based on a combination of CDMA and multicarrier technique spreads the original data stream over different subcarriers using a spreading code in the frequency domain [14]. Indeed, in the past decade, there is growing need for technological innovations to satisfy the increase demand for personal wireless radio communications. This technology must be able to allow users to efficiency share common resources, whether it involves the frequency spectrum and computational load. That is why the MC-CDMA [14], the one of representative of the multicarrier techniques, has been considered as a promising system for the next generation of wireless communication. One large advantage of this technology is its robustness in case of multi-path propagation, and it's capable to combat frequency selective fading, flexible to generate different data rates and provides bandwidth efficiency [12].

Multicarrier systems are very sensitive to synchronization errors such as carrier frequency offset and phase noise. Synchronization errors cause loss of orthogonality among subcarriers and considerably degrade the performance especially when large number of subcarriers presents. There have been many approaches on synchronization algorithms in literature [15]–[17]. MC-CDMA designed for multiusers using the same channel, so the problem of channel identification appears, in this paper we propose to encode chaotic

cally data before its spreading in order to use this sequence for system identification.

2. Time and Frequency Domain MC-CDMA System Description

In order to transmit data of a large number of users, the frequency band should be, optimally, used. The objective is to transmit, in simultaneous over the same channel the maximum of information, so the use of multiplexing. Indeed, in CDMA [15], the users have access, in the same time, to the totality of the frequency band, in the receiver, to distinguish between them, we use a different codes affected for each user. That was possible thanks to the technique of spectral spreading, in condition that the emitted signals by different users have some proprieties allowing them to separate.

In opposition of the other techniques of multi-access such frequency division multiple access (FDMA) and time division multiple access (TDMA), where the capacity of the number of users is limited by the frequency and time resources, respectively, the number of users in CDMA is fixed by the proprieties of used spreading codes [18]. That is why the CDMA is an alternative to the others multiplexing techniques to increase the re-use frequency factor and eventually the spectral efficiency of communication systems. A different approach to further increase the system capacity without allocating additional frequency spectrum is the use of code multiplexing.

The MC-CDMA modulator spreads the data of each user in frequency domain (we consider a binary phase shift keying (BPSK) data), the complex symbol g_j of each user j is, firstly, multiplied by each chips $c_{j,k}$ of spreading code SC_j , and then applied to the modulator of multicarriers. Each subcarrier transmits an element of information multiplied by a code chip of that subcarrier.

We consider, for example, the case where the length L_c of spreading code is equal to the number N of subcarriers. The optimum space between two adjacent subcarriers is equal to inverse of duration T_c of chip of spreading code in order to guarantee the orthogonality between subcarriers. The MC-CDMA signal is given by

$$s(t) = \sum_{k=0}^{N-1} a_k c_{m,k} e^{2i\pi f_k t}, \quad (1)$$

where m is the user number.

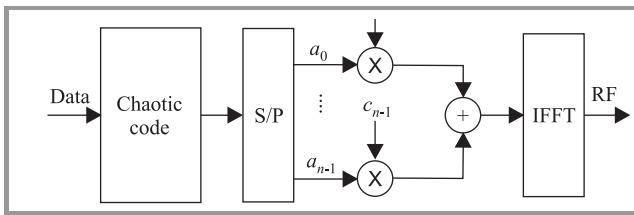


Fig. 1. Chaotic MC-CDMA transmitter. Explanation: S/P – serial/parallel.

We consider the channel invariant in time and characterized by P paths of magnitudes β_p and phase θ_p . The impulse response is given by

$$h(\tau) = \sum_{p=0}^{P-1} \beta_p e^{i\theta_p} \delta(\tau - \tau_p).$$

The relationship between the emitted signal $s(t)$ and the received signal $r(t)$ is given by

$$r(t) = h(t) * s(t) + n(t),$$

where $*$ is the convolution product and $n(t)$ is the additive white Gaussian noise (AWGN):

$$\begin{aligned} r(t) &= \int_{-\infty}^{+\infty} \sum_{p=0}^{P-1} a_k \beta_p e^{i\theta_p} \delta(\tau - \tau_p) s(t - \tau) d\tau + n(t) \\ &= \sum_{p=0}^{P-1} a_k \beta_p e^{i\theta_p} s(t - \tau) + n(t), \end{aligned} \quad (2)$$

where P is the number of paths.

The transmitter of the MC-CDMA scheme using both time domain and frequency domain is shown in Fig. 1.

The discrete chaotic sequence is: $\chi = [a_0, a_1, \dots, a_{N-1}]$, where N is the chaotic sequence duration.

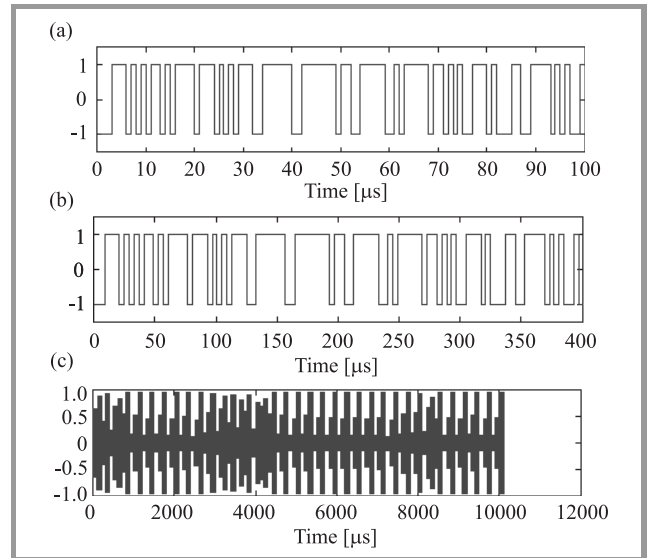


Fig. 2. Binary phase shift keying data (a), spread by the Walsh-Hadamard code (b), and by the chaotic code (c).

In Fig. 2, we represent an example of BPSK data spread by the Walsh-Hadamard code and by the chaotic code. In the following equation we represent an example of the Walsh-Hadamard code C (e.g., for four users):

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

In a system of M users the emitted signal through a channel is given by

$$s(t) = \sum_{m=0}^{M-1} \sum_{k=0}^{N-1} a_m[k] c_{m,k} e^{2i\pi f_k t}. \quad (3)$$

The received signal after passing through the channel is:

$$r(t) = \sum_{m=0}^{M-1} \sum_{p=0}^{P-1} \sum_{k=0}^{N-1} a_m[k] \beta_p e^{i\theta} g_{m,k} c_{m,k} e^{2i\pi(f_0+k/T_s)} \times \delta(t - \tau_p) + n(t). \quad (4)$$

At the reception, we demodulate the signal according to N subcarriers, and then we multiply the received sequence by the code of the desired user (Fig. 3). Equalization is, then, applied to estimate the frame $g_{m,k}$.

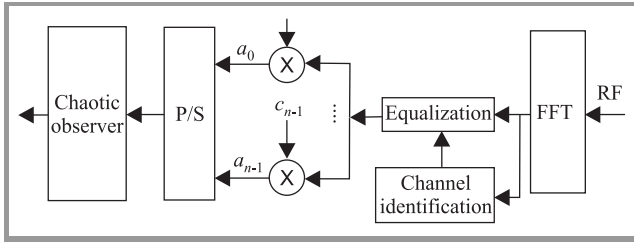


Fig. 3. Chaotic MC-CDMA receiver.

When there are M active users, the received signal is:

$$r(t) = \sum_{m=0}^{M-1} \sum_{k=0}^{N-1} h_{m,k} c_m[k] a_m[k] \cos(2\pi f_c t + 2\pi k \frac{F}{T_b} t + \theta_{m,k}) + n(t), \quad (5)$$

where the effects of the channel have been included in $h_{m,k}$ and $\theta_{m,k}$, $n(t)$ is AWGN with a one-sided power spectral density of N_0 .

We assume that $m = 0$ corresponds to the desired signal. With this model, there are N matched filters with one matched filter for each subcarrier. The output of each filter contributes one component to the decision variable, ϑ_0 . Each matched filter consists of an oscillator with a frequency corresponding to the frequency of the particular BPSK modulated subcarrier that is of interest and an integrator. In addition, a phase offset equal to the phase distortion introduced by the channel, $\theta_{m,k}$, is included in the oscillator to synchronize the receiver to the desired signal in time.

To extract the desired signal's component, the orthogonality of the codes is used. For the k th subcarrier of the desired signal, the corresponding chip, $c_0[k]$, from the desired user's code is multiplied with it to undo the code. If the signal is undistorted by the channel, the interference terms will cancel out in the decision variable due to the orthogonality of the codes. As the channel will distort the subcarrier components, equalization gain, $g_{0,k}$, may be included for each matched filter branch of the receiver.

Applying the receiver model to the received signal given in Eq. (4) yields the following decision variable for the k th data symbol assuming the users are synchronized in time:

$$\vartheta_0 = \sum_{m=0}^{M-1} \sum_{k=0}^{N-1} h_{m,k} c_m[k] a_m[k] \frac{2}{T_b} \int_{kT_b}^{(k+1)T_b} \cos(2\pi f_c t + 2\pi k \frac{F}{T_b} t + \theta_{m,k}) \cos(2\pi f_c t + 2\pi k \frac{F}{T_b} t + \hat{\theta}_{0,k}) dt + \eta, \quad (6)$$

where $\hat{\theta}_{0,k}$ denotes the receiver's estimation of the phase at the k th subcarrier of the desired signal and the corresponding AWGN term, η is given as

$$\eta = \sum_{k=0}^{N-1} \int_{kT_b}^{(k+1)T_b} n(t) \frac{2}{T_b} \cos(2\pi f_c t + 2\pi k \frac{F}{T_b} t + \hat{\theta}_{0,k}) dt. \quad (7)$$

Assuming perfect phase correction, $\hat{\theta}_{0,k} = \theta_{0,k}$, the decision variable reduces to

$$\vartheta_0 = a_0[k] \sum_{k=0}^{N-1} h_{0,k} + \sum_{m=0}^{M-1} \sum_{k=0}^{N-1} a_m[k] c_m[k] c_0[k] h_{m,k} \cos(\hat{\theta}_{m,k}) + \eta, \quad (8)$$

where $\hat{\theta}_{m,k} = \theta_{0,k} - \theta_{m,k}$. If $\theta_{0,k}$ and $\theta_{m,k}$ are i.i.d. uniform r.v.'s on the interval $[0, 2\pi]$, then $\hat{\theta}_{m,k}$ is also uniformly distributed on the interval $[0, 2\pi]$. Note that the decision variable consists of three terms.

The first term corresponds to the desired signal's component, the second corresponds to the interference and the last corresponds to a noise term:

$$\vartheta_0 = \xi_{inf} + \beta_{int} + \eta, \quad (9)$$

where ξ_{inf} and β_{int} are the terms of information and interferences, respectively, defined by

$$\xi_{inf} = a_0[k] \sum_{i=0}^{N-1} h_{0,k}, \quad (10)$$

$$\beta_{int} = \sum_{m=0}^{M-1} \sum_{k=0}^{N-1} a_m[k] c_m[k] c_0[k] h_{m,k} \cos(\hat{\theta}_{m,k}),$$

$$\eta = \sum_{k=0}^{N-1} \int_{kT_b}^{(k+1)T_b} n(t) \frac{2}{T_b} \cos(2\pi f_c t + 2\pi k \frac{F}{T_b} t + \hat{\theta}_{0,k}) dt.$$

The noise can be approximated by a zero-mean Gaussian random variable with the following variance:

$$\sigma_\eta^2 = N \frac{N_0}{T_b} E[g_{0,k}^2]. \quad (11)$$

3. Chaotic Codes

Chaos based communication systems qualify as broadband systems in which the natural spectrum of the information

signal is spread over a very large bandwidth [19]. This class of systems is called spread spectrum communication systems since they make use of a much higher bandwidth than that of the data bandwidth to transmit the information. Nowadays, pseudo-noise sequences such as Gold sequences and Walsh-Hadamard sequences are so far the most popular spreading sequences and have good correlation properties, limited security and can be reconstructed by linear regression attack for their short linear complexity [20]. A chaotic sequence generator can visit an infinite number of states in a deterministic manner and therefore produce a sequence which never repeats itself [4].

There is the flexibility in choosing the spreading gain as the sequences can be truncated to any length. The search for the best set of codes contributing reduced multiple access interference (MAI) is still one of the severe requirements of future MC-CDMA systems.

This chaotic spreading code is used before the use of an orthogonal code for multiuser transmission. These chaotic and Walsh-Hadamard codes have produced good result in utility and reducing MAI [21].

A single system described by its discrete chaotic map can generate a very large number of distinct chaotic se-

quences, each sequence being uniquely specified by its initial value [22]. This dependency on the initial state and the nonlinear characteristic of the discrete map make the MC-CDMA system highly secure. A chaotic map is a dynamic discrete-time continuous-value equation that describes the relation between the present and next value of chaotic system. Let x_{n+1} and x_n be successive iterations of the output x and M is the forward transformation mapping function. The general form of multidimensional chaotic map is $x_{n+1} = M(x_n, x_{n-1}, \dots, x_{n-m})$. A simple logistic map is given in the following equation:

$$x_{n+1} = \mu(1 - x_n)x_n \quad (12)$$

and $1 \leq \mu \leq 4$, where μ is the bifurcation parameter and the system exhibits a great variety of dynamics depending on the value of μ ($3.6 \leq \mu \leq 6$).

In Fig. 4 we represent logistic map for different constant number μ .

Using logistic map the chaotic spreading sequences for the BPSK system is generated. After assigning different initial condition to each user, the chaotic map is started with the initial condition of the intended receiver and iterated repeatedly to generate multiple codes.

4. Channel Identification Using Chaos

The goal of the equalization techniques is to reduce the effect of the fading and the interference while not enhancing the effect of the noise on the decision of what data symbol was transmitted. Whenever there is a diversity scheme involved, it may involve receiving multiple copies of a signal from time, frequency or antenna diversity, the field of classical diversity theory can be applied. These equalization techniques may be desirable for their simplicity as they involve simple multiplications with each copy of the signal. However, they may not be optimal in a channel with interference in the sense of minimizing the error under some criterion.

Before the equalization, the receiver estimates the channel parameters using a chaotic sequence χ of length N as a training sequence, we assume that the channel is invariant during a time T_p .

The information term is then given by

$$\mathbf{h} = \chi^{-1}r(t). \quad (13)$$

To simplify the formulation, we consider one user in this system, instead of $h_{m,k}$ we use h_k , the same with the equalizer parameters $g_{m,k}$; so $\mathbf{h} = \{h_0, h_1, h_2; \dots; h_{N-1}\}^T$ and $r(t)$ is the received data from the chaotic sequence transmitted. We suppose that the receiver knows the transmitted chaotic sequence χ by the transmitter based on the model of logistic map Eq. (12). It is shown, from Eq. (13), that the channel parameters can be easily estimated. So the equalization term using equal gain combining (EGC) is given by

$$g_k = \frac{h_k^*}{|h_k|^2}.$$

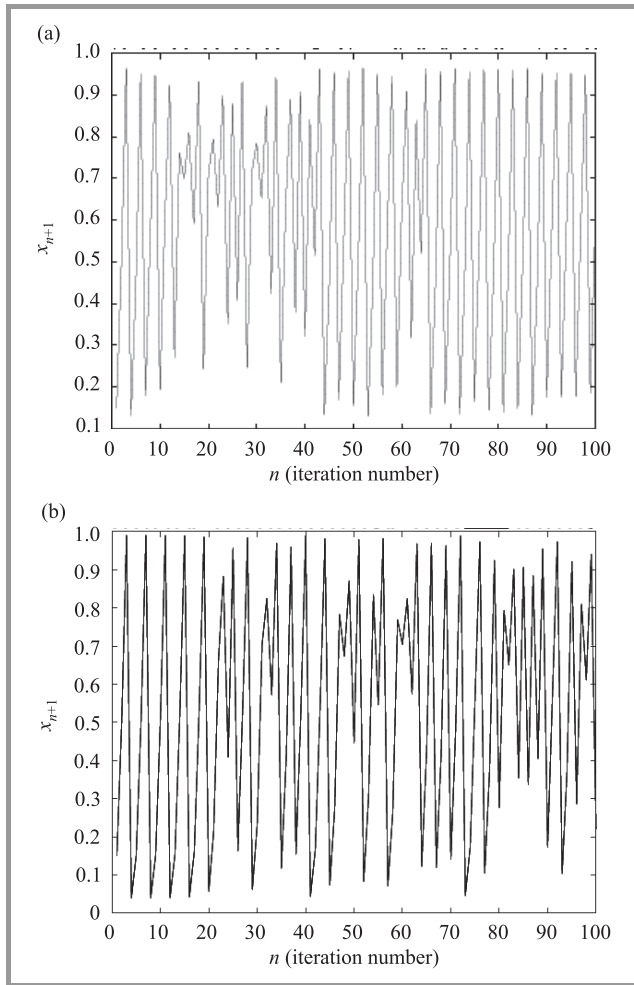


Fig. 4. The logistic map for (a) $\mu = 3.86$ and $x_0 = 0.15$, and (b) $\mu = 3.96$ and $x_0 = 0.15$.

The information is then equalized, from Eq. (10) we have:

$$\xi_{inf} = a_m \sum_{k=0}^{N-1} h_k g_k. \quad (14)$$

This technique does not attempt to equalize the effect of the channel distortion in any way. This technique may be desirable for its simplicity as the receiver does not require the estimation of the channel's transfer function. Using this scheme, the decision variable of Eq. (9) is given as

$$\vartheta_0 = \xi_{inf}^{egc} + \beta_{int}^{egc} + \eta^{egc} \quad (15)$$

with $\xi_{inf}^{egc} = a_m$ and $\beta_{int}^{egc} = \sum_{m=1}^{M-1} a_m[k] c_m[k] c_0[k] \cos(\hat{\theta}_{m,k})$, where the noise can be approximated by a zero-mean Gaussian random with a variance of: $\sigma_\eta^2 = N \frac{N_0}{T_b}$.

5. Theoretical Analysis

5.1. Uplink Transmission

We analyze bit error rate (BER) of the proposed chaotic MC-CDMA system. We have calculated the theoretical BER for the classical equalizer EGC.

Below, some of the theoretical performance results obtained are given:

$$\sigma_{\beta_{int}}^2 = (M-1) \bar{P}_m \text{ (variance of interferences)}$$

and

$$\sigma_\eta^2 = N \frac{N_0}{T_b} \text{ (variance of noise).}$$

We have the general form of BER [14] :

$$BER = \frac{1}{2} \text{erfc} \left(\frac{0.5 \left(\sum_{k=0}^{N-1} h_k g_k \right)^2}{\sigma_{\beta_{int}}^2 + \sigma_\eta^2} \right)^{1/2}. \quad (16)$$

In the case of EGC we have :

$$BER = \frac{1}{2} \text{erfc} \left(\frac{\frac{1}{2} \left(\sum_{k=0}^{N-1} h_k \right)^2}{(M-1) \bar{P}_m + N \frac{N_0}{T_b}} \right)^{1/2}. \quad (17)$$

The objective is to find an approximation of $h = \sum_{k=0}^{N-1} h_k$.

In the limiting case of a large number of subcarriers, $\left(\sum_{k=0}^{N-1} h_k \right)$ can be approximated by the law of large number (LLN) to be the constant $NE[h_k]$. The advantage of

using the LLN is that it requires low computational complexity. Using the LLN simplifies the expression for the probability of error to [14]:

$$N \text{ is large} \implies \gamma_0 = \sum_{k=0}^{N-1} h_k \simeq NE[h_k],$$

$$BER = \frac{1}{2} \text{erfc} \left(\frac{\frac{1}{2} N^2 E^2[h_k] T_b}{(M-1) \bar{P}_m T_b + N N N_0} \right)^{1/2},$$

$$BER = \frac{1}{2} \text{erfc} \left(\frac{\frac{\pi}{4} N SNR}{(M-1) SNR + N} \right)^{1/2}, \quad (18)$$

where N is the number of subcarriers and M is the number of users.

5.2. Downlink Transmission

Transmissions in the downlink, i.e., the transmission from the base station to the terminals through the same channel.

In this section, we'll use the notation of Eq. (8), and we assume perfect phase correction for interference. The generalized decision variable given in Eq. (8), simplifies to

$$\vartheta_0 = a_0[k] \sum_{k=0}^{N-1} h_k g_k + \sum_{m=0}^{M-1} \sum_{k=0}^{N-1} a_m[k] c_m[k] c_0[k] h_k g_k + \eta.$$

The used codes are orthogonal and the product, $c_{k,i} c_{k,j}$ is, then, equal to 1 with the probability 1/2 and -1 with the same probability for $i \neq j$:

$$\vartheta_0 = a_0[i] \sum_{k=0}^{N-1} h_k g_k + \sum_{m=0}^{M-1} a_m[k] \left(\sum_{k=0}^{N/2-1} h_k g_k - \sum_{k=0}^{N/2-1} h_k g_k \right) + \eta. \quad (19)$$

6. Performance Analysis

In this section, the approximations for the BER using the LLN is evaluated numerically. Using the expressions for the BER obtained for uplink transmissions in a Rayleigh fading channel, the average BER versus the number of co-channel interferers with a spreading factor $N = 128$ is shown in Fig. 5. To calculate the BER, it is assumed that the local-mean power of each interferer is equal to the local-mean power of the desired signal. The signal-to-noise ratio (SNR), which is assumed to be 10 dB, is defined as

$$SNR = \frac{\bar{P}_0 T_b}{N_0},$$

where $\bar{P}_0$ is the power of each user supposed equal for all users.

From Fig. 5, we remark that we have, approximately, the same BER results if we use both codes such as: orthogonal

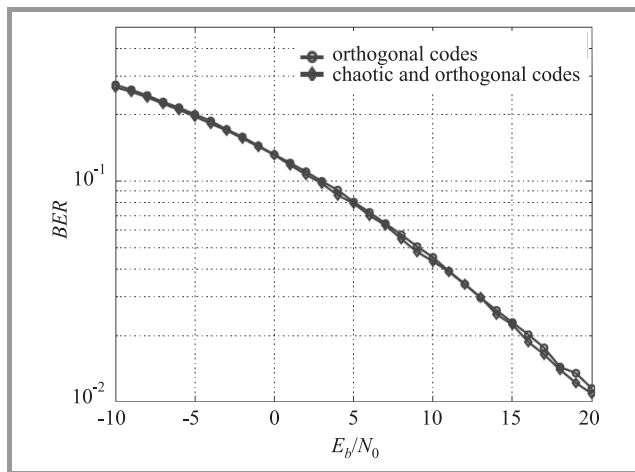


Fig. 5. Simulated BER of equalizers for chaotic MC-CDMA system.

code and chaotic code, comparing to the results obtained if we use only the orthogonal code. This result is very interesting because the use of the chaotic code allows first channel identification and this transmission is more secure than the classical MC-CDMA as combine both the Walsh-Hadamard and the chaotic codes.

7. Conclusion

Channel identification of a digital modulation technique MC-CDMA was proposed. This technique is based on the use of the chaotic sequence as a pilot frame. In fact, the chaotic code given by a nonlinear system (in this paper we have used the logistic map but other chaotic's systems could be used such Hénon map, Rössler map, Lyapunov fractal, Horseshoe map, ...) is known by the transmitter and the receiver. These codes are very sensitive to the initial conditions, so security depends on difficulty in finding these parameters.

The EGC equalization technique was used after the channel identification to correct the channel's distortion. The performance of this technique, gauged by the average bit error rate, was analytically and numerically evaluated. It is demonstrated that the performance of MC-CDMA system employing chaotic binary sequences can be superior to that for conventional codes in terms of bit error probabilities.

References

- [1] S. Papadimitriou, T. Bountis, S. Mavroudi, and A. Bezerianos, "A probabilistic symmetric encryption scheme for very fast secure communications based on chaotic systems of difference equations", *Int. J. Bifurc. Chaos*, vol. 11, pp. 3107–3115, 2001.
- [2] N. K. Pareek, V. Patidar, and K. K. Sud, "Discrete chaotic cryptography using external key", *Phys. Lett. A*, vol. 309, pp. 75–82, 2003.

- [3] A. T. Parker and K. M. Short, "Reconstructing the keystream from a chaotic encryption scheme", *IEEE Trans. Circ. Syst.*, vol. 48, pp. 624–630, 2001.
- [4] V. Stefano and S. Gianluca, "Improving PA efficiency by chaos based spreading in MC/DS-CDMA systems", in *Proc. IEEE ISCAS Conf.*, Kos, Greece, 2006, pp. 1195–1198.
- [5] A. Abel and W. Schwarz, "Chaos communication-principles schemes and system analysis", *Proc. IEEE*, vol. 90, no. 5, pp. 691–710, 2002.
- [6] S. Kaiser, "Multi-carrier CDMA mobile radio system-analysis and optimization of detection, decoding, and channel estimation", Ph.D. thesis, University of Munich, Germany, 1998.
- [7] S. Safi, M. Frikel, M. M'Saad, and A. Zeroual, "Blind impulse response identification of frequency radio channels: application to BRAN a channel", *Int. J. Sig. Proc.*, vol. 4, no. 1, pp. 201–206, 2007.
- [8] S. Safi and A. Zeroual, "Blind parametric identification of linear stochastic non-Gaussian FIR systems higher order cumulants", *Int. J. Syst. Sci.*, vol. 35, no. 15, pp. 855–867, 2004.
- [9] W. Xingang, M. Zhana, G. Xiaofeng, L. Heng, and L. Ying-Cheng, "Spread-spectrum communication using binary spatiotemporal chaotic codes", *Phys. Lett. A*, vol. 334, pp. 30–36, 2005.
- [10] "Broadband Radio Access Networks (BRAN); High Performance Radio Logical Area Network (HIPERLAN) Type 2: Requirements and architectures for wireless broadband access", ETSI, TR 101 031 V2.2.1 (1999–01).
- [11] "Broadband Radio Access Networks (BRAN); HIPERLAN Type 2; Physical (PHY) layer", ETSI, TS 101 475 V1.3.1 (2001–12).
- [12] F. Horlin, S. De Rore, E. Lopez-Estraviz, F. Naessens, and L. Van der Perre, "Impact of frequency offsets and IQ imbalance on MC-CDMA reception based on channel tracking", *IEEE J. Select. Areas Commun.*, vol. 24, no. 6, pp. 1179–1188, 2006.
- [13] K. Fazel and L. Papke, "On the performance of convolutionally-coded CDMA/OFDM for mobile communication system", in *Proc. IEEE Int. Symp. Pers. Ind. Mob. Radio Commun. PIMRC'93*, Berlin, Germany, 1993, pp. 468–472.
- [14] N. Yee, J. M. G. Linnartz, and G. Fettweis, "Multi-carrier-CDMA in indoor wireless networks", in *Proc. IEEE Int. Symp. Pers. Ind. Mob. Radio Commun. PIMRC'93*, Yokohama, Japan, 1993, pp. 109–113.
- [15] J. M. G. Linnartz, "Performance analysis of synchronous MC-CDMA in mobile Rayleigh channel with both delay and Doppler spreads", *IEEE Trans. Veh. Technol.*, vol. 50, no. 6, 2001, pp. 1375–1387.
- [16] M. Frikel, B. Targui, M. M'Saad, and F. Hamon, "Bit error rate analysis of the controlled equalization for MC-CDMA", in *IEEE ICSPC'07 Conf.*, Dubai, United Arab Emirates, 2007.
- [17] S. Kaiser, "OFDM-CDMA versus DS-CDMA: performance evaluation for fading channels", in *Proc. IEEE Int. Conf. Commun.*, Seattle, USA, 1995, pp. 1722–1726.
- [18] D. Mottier and D. Castelain, "SINR-based channel pre-equalization for uplink multi-carrier CDMA systems", in *Proc. IEEE Int. Symp. Pers. Ind. Mob. Radio Commun. PIMRC'02*, Lisbon, Portugal, 2002, pp. 1488–1492.
- [19] S. Azou, G. Burel, and C. Pistre, "A chaotic direct-sequence spread-spectrum system for underwater communication", in *Proc. OCEANS'02 MTS/IEEE Conf.*, Biloxi, USA, 2002, vol. 4, pp. 2409–2415.
- [20] C. Ling and S. Li, "Chaotic spreading sequences with multiple access performance better than random sequences", *IEEE Trans. Circ. Syst. I Fund. Theory Appl.*, vol. 47, no. 3, pp. 394–397, 2006.
- [21] S. L. Georgoulis and D. G. M. Cruickshank, "Transmitter-based inverse filters for reducing MAI and ISI in CDMA-TDD downlink", *IEEE Trans. Wirel. Commun.*, vol. 3, no. 2, pp. 353–358, 2004.

- [22] M. P. Kennedy, R. Rovatti, and G. Setti, "Chaotic electronics in telecommunications", in *Proc. CRC Conf.*, Boca Raton, USA, 2000.

Note: this text is reprinted from JTIT 1/2010



Miloud Frikel received his Ph.D. degree from the Center of Mathematics and Scientific Computation CNRS URA 2053, France, in array processing. Currently he is with the GREYC laboratory (CNRS URA 6072) and the ENSICAEN as Assistant Professor. From 1998 to 2003, he was

with the Signal Processing Lab, Institute for Systems and Robotics, Institute Superior Tecnico, Lisbon, Portugal, as a researcher in the field of wireless location and statistical array processing, after been a research engineer in a software company in Munich, Germany. He worked in the Institute for Circuit and Signal Processing of the Technical University of Munich, Germany. His research interests span several areas, including statistical signal and array processing, cellular geolocation (wireless location), space-time coding, direction finding and source localization, blind channel identification for wireless communication systems, and MC-CDMA systems.

e-mail: mfrikel@greyc.ensicaen.fr

GREYC UMR 6072 CNRS

Ecole Nationale Supérieure d'Ingénieurs de Caen (ENSICAEN)

6, B. Maréchal Juin

14050 Caen, France



Said Safi was born in Beni Mellal, Morocco, in 1971, received the B.Sc. degree in physics (option electronics) from Cadi Ayyad University, Marrakech, Morocco, in 1995, M.Sc. and Ph.D. degrees from Chouaib Doukkali University and Cadi Ayyad University, in 1997 and 2002, respectively. He has been a Professor of information theory and telecommunication systems at the National School for Applied Sciences, Tangier, from 2003 to 2005. Since 2006, he is a Professor of applied mathematics and programming at the Faculty of Science and Technics, Beni Mellal. In 2008 he received the Ph.D. degree in telecommunication and informatics from the Cadi Ayyad University. His general interests span the areas of communications and signal processing, estimation, time-series analysis, and system identification – subjects on which he has published 10 journal papers and more than 40 conference papers. Current research topics focus on transmitter and receiver diversity techniques for single- and multiuser fading commu-

nication channels, and wide-band wireless communication systems.

nication channels, and wide-band wireless communication systems.

e-mail: safi.said@gmail.com

Polydisciplinary Faculty

Sultan Moulay Slimane University

P.O. Box 523

Beni Mellal, Morocco



Boubekour Targui was born in Abadla, Algeria. He received the B.Sc. degree in the electronic engineering from the University of Sidi Bel Abbess, Algeria, in 1994, and the M.Sc. and Ph.D. degrees in control engineering from the University of Claude Bernard Lyon I, Lyon, France, in 2000. Since 2002, he has held a position of Associate

Professor of networking and telecommunication in the IUT of Caen, University de Caen, France. His current research interests are in equalization and synchronization in MC-CDMA systems, nonlinear observers and synchronization.

e-mail: boubekour.targui@greyc.ensicaen.fr

GREYC UMR 6072 CNRS

Ecole Nationale Supérieure d'Ingénieurs de Caen (ENSICAEN)

6, B. Maréchal Juin

14050 Caen, France



Mohammed M'Saad was educated at Mohammadia School of Engineering in Rabat, Morocco, where he held an Assistant Professor position in 1978. He held a research position at the Laboratoire d'Electronique et d'Etude des Systèmes in Rabat, where he prepared his doctor engineering degree in process control. In November 1982, he joined the Laboratoire d'Automatique de Grenoble to work on theory and applications of adaptive control. He received his Doctorat d'Etat-Sciences Physiques from the Institut National Polytechnique de Grenoble in April 1987. In March 1988, he held a research position at the Centre National de Recherche Scientifique. In September 1996, he held a Professor position at the Ecole Nationale Supérieure d'Ingénieurs de Caen, where he is the Head of the Control Group at the GREYC. His main research areas are adaptive control theory, system identification and observation, advanced control methodologies and applications, computer aided control engineering.

e-mail: mohammed.msaad@greyc.ensicaen.fr

GREYC UMR 6072 CNRS

Ecole Nationale Supérieure d'Ingénieurs de Caen (ENSICAEN)

6, B. Maréchal Juin

14050 Caen, France

The Design of 4×4 Multimode Interference Coupler Based Microring Resonators on an SOI Platform

Trung-Thanh Le and Laurence W. Cahill

Abstract—This paper would like to propose a novel microring resonator based on 4×4 multimode interference (MMI) couplers. The device acts as two separate microring resonators just in one structure. The transfer matrix method and the three dimensional beam propagation method (3D-BPM) are used to verify the working principle of the device. The device is then designed on silicon on insulator (SOI) technology. This device may be a very promising building block for optical switches, filters, add-drop multiplexers, delay lines and modulators.

Keywords—integrated optics, multimode interference couplers, optical logic gates.

1. Introduction

Microring resonators have been used as a basic building block for optical signal processing applications such as optical switches, filters, modulators, and add-drop multiplexers [1]. Almost all of the reported works on microring resonator structures have used directional couplers as the coupling element [2]. In order to meet a variety of requirements for high-speed signal processing, the coupling coefficient needs to be adjusted arbitrarily and the directional couplers can meet this requirement. However, for applications requiring high quality factor Q of the resonators, i.e., for high speed operation, the separations between two waveguides in the directional coupler must be very small. As a result, high loss due to conversion loss of modes is occurred as shown in [3] recently. Moreover, the directional coupler has a large size and small fabrication tolerance. Therefore, multimode interference (MMI) couplers are used in such structures instead of directional couplers due to their advantages of compactness, ease of fabrication, large fabrication tolerance and ease of cascaded integration [4].

A microring resonator based on a 2×2 MMI coupler was demonstrated on silicon on insulator (SOI) channel waveguides for the first time in [2]. However, the coupling ratios of the conventional MMI couplers are very limited and there are only four available coupling ratios 0:100, 50:50, 85:15, and 72:28 if using only one conventional MMI coupler [5]. Therefore, it is highly desired to implement the couplers with variable coupling ratios. To do so, the common approach is to use Mach-Zehnder interferometer (MZI) structures, in which phase shifters are added to the MZI arms to control the phase of the propagation signals as shown in [6], [7].

In this paper, a novel microring resonator based on 4×4 MMI couplers is given for the first time. The most interesting characteristic of the proposed device is that the device acts as two separate microring resonators based on 2×2 MMI couplers. The coupling ratios can be varied by using two separate phase shifters for the two microring resonators. In order to verify the working principle of the device, the transfer matrix and beam propagation method are used. The device is designed on an SOI platform.

2. Microring Resonators Based on 4×4 MMI Couplers

A microring resonator based on 4×4 MMI couplers (MMI-MZI structure) is shown in Fig. 1. The two 4×4 MMI couplers have the same width W_{MMI} and length $L_{MMI} = \frac{3L_\pi}{2}$, where $L_\pi = \frac{\pi}{\beta_0 - \beta_1}$ is the beat length of the MMI coupler. In order to make tunable devices, two phase shifters $\Delta\phi_1$ and $\Delta\phi_2$ based on the thermo-optic effect are used in the two arms. Alternatively, passive phase shifters could be used to provide a desired (fixed) power splitting ratio.

In order to analyze the device, the transfer matrix of the identical MMI couplers needs to be derived first and then the total transfer matrix of the device can be determined. The 4×4 MMI coupler can be described by a transfer matrix [8]:

$$M_{4 \times 4} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{bmatrix}, \quad (1)$$

where m_{ij} ($i, j = 1, \dots, 4$) are complex coefficients calculated by using the modal propagation method. At the length $L = \frac{3L_\pi}{4}$, the phases of the equal output signals at the output waveguides can be calculated from

$$\phi_{ij} = \phi_0 + \pi + \frac{\pi}{16}(j-i)(8-j+i), \text{ for } i+j \text{ even}$$

$$\text{and } \phi_{ij} = \phi_0 + \frac{\pi}{16}(i+j-1)(8-j-i+1), \text{ for } i+j \text{ odd.}$$

Here, input ports i ($i = 1, \dots, 4$) are numbered from down to up and the output ports j ($j = 1, \dots, 4$) are numbered from up to down in the MMI coupler and $\phi_0 = -\beta_0 L_{MMI} - \frac{\pi}{2}$ is a phase constant factor that is associated with the MMI ge-

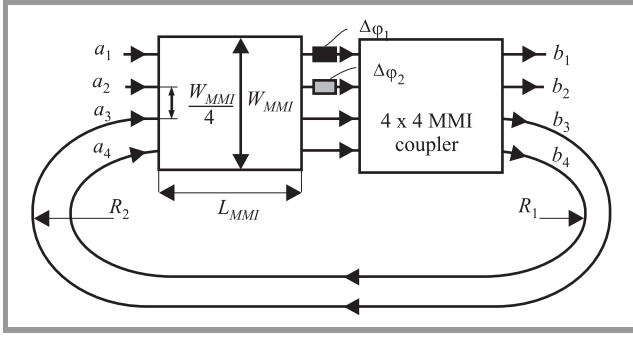


Fig. 1. The structure of a microring resonator based on 4×4 MMI couplers.

ometry and therefore can be neglected in the following analyses.

The 4×4 MMI coupler at a length of $L_1 = \frac{3L_\pi}{4}$ is described by the following transfer matrix:

$$M_{4 \times 4} = \frac{1}{2} \begin{bmatrix} -1 & -e^{j\frac{3\pi}{4}} & e^{j\frac{3\pi}{4}} & -1 \\ -e^{j\frac{3\pi}{4}} & -1 & -1 & e^{j\frac{3\pi}{4}} \\ e^{j\frac{3\pi}{4}} & -1 & -1 & -e^{j\frac{3\pi}{4}} \\ -1 & e^{j\frac{3\pi}{4}} & -e^{j\frac{3\pi}{4}} & -1 \end{bmatrix}. \quad (2)$$

If the length is doubled to $L_{MMI} = 2L_1 = \frac{3L_\pi}{2}$, a new 4×4 MMI coupler is formed and its transfer matrix is

$$M_{new} = (M_{4 \times 4})^2 = \frac{1}{2} \begin{bmatrix} 1-j & 0 & 0 & 1+j \\ 0 & 1-j & 1+j & 0 \\ 0 & 1+j & 1-j & 0 \\ 1+j & 0 & 0 & 1-j \end{bmatrix}. \quad (3)$$

This matrix can be considered as consisting of two separate submatrices which describe two 2×2 the 3 dB MMI couplers, both having the transfer matrix:

$$M_2 = \frac{1}{2} \begin{bmatrix} 1-j & 1+j \\ 1+j & 1-j \end{bmatrix} = \frac{1}{\sqrt{2}} e^{-j\frac{\pi}{4}} \begin{bmatrix} 1 & j \\ j & 1 \end{bmatrix}. \quad (4)$$

If two 4×4 MMI couplers are connected in the MZI structure of Fig. 1, then the relations between the complex amplitudes at the input ports and output ports can be expressed in terms of the transfer matrices of the 3 dB MMI couplers and the phase shifters as follows:

$$\begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = e^{j\frac{\Delta\phi_1}{2}} \begin{bmatrix} \tau_1 & \kappa_1 \\ \kappa_1^* & -\tau_1^* \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}, \quad (5)$$

$$\text{where } \tau_1 = \sin\left(\frac{\Delta\phi_1}{2}\right) \text{ and } \kappa_1 = \cos\left(\frac{\Delta\phi_1}{2}\right), \quad (6)$$

$$\begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = e^{j\frac{\Delta\phi_2}{2}} \begin{bmatrix} \tau_2 & \kappa_2 \\ \kappa_2^* & -\tau_2^* \end{bmatrix} \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}, \quad (7)$$

$$\text{where } \tau_2 = \sin\left(\frac{\Delta\phi_2}{2}\right) \text{ and } \kappa_2 = \cos\left(\frac{\Delta\phi_2}{2}\right). \quad (8)$$

Therefore, the whole device can be viewed as consisting of two independent microresonators having different power coupling ratios as shown in Fig. 2. This means that two independent switches and filters can be made by using this structure.

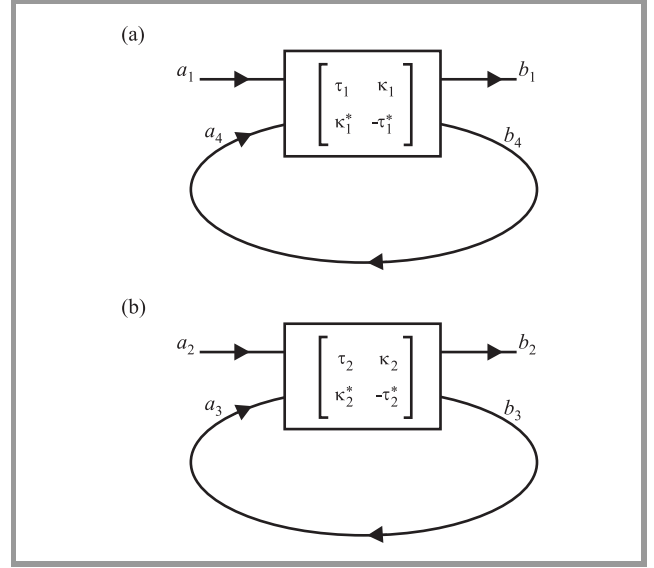


Fig. 2. Two separate microresonators created from the 4×4 MMI structure: (a) microresonator using input ports 1 and 4, and (b) microresonator using input ports 2 and 3.

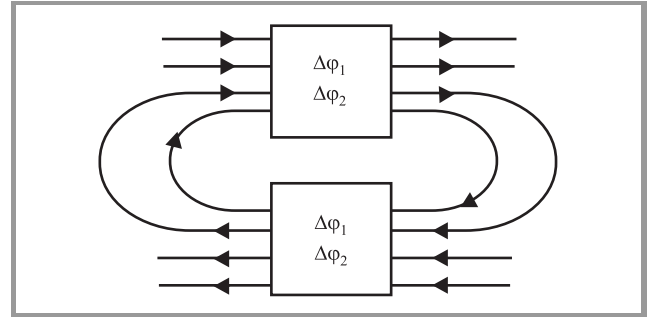


Fig. 3. Two separate add-drop multiplexers based on microresonators using 4×4 MMI-MZI structures.

Another useful structure for implementing add-drop multiplexing functions is shown in Fig. 3, where the microresonator is coupled to a second MMI-MZI structure. The coupling ratios can be controlled by the phase shifters. Similarly, two independent add-drop filters can be obtained from this structure.

3. Simulation Results and Discussions

In this section, in order to verify the working principle of the devices, the three dimensional beam propagation method (3D-BPM) [9], [10] is used to optimize the designs for MMI devices. It is well known that the finite difference time-domain (FDTD) method is a general method

to solve Maxwell's partial differential equations numerically in the time domain. Simulation results for devices on the SOI channel waveguide using the 3D-FDTD method can achieve a very high accuracy. However, due to the limitation of computer resources and memory requirements, it is difficult to apply the 3D-FDTD method to the modeling of large devices on the SOI channel waveguide. Meanwhile, the 3D-BPM was shown to be a quite suitable method [11], [12] that has sufficient accuracy for simulating devices based on SOI channel waveguides [13], [14]. The waveguide structure used in the designs is shown in Fig. 4. Here, SiO₂ ($n = 1.46$) is used as the upper cladding material. An upper cladding region is needed for devices using the thermo-optic effect in order to reduce loss due to metal electrodes. Also, the upper cladding region is used to avoid the influence of moisture and environmental temperature.

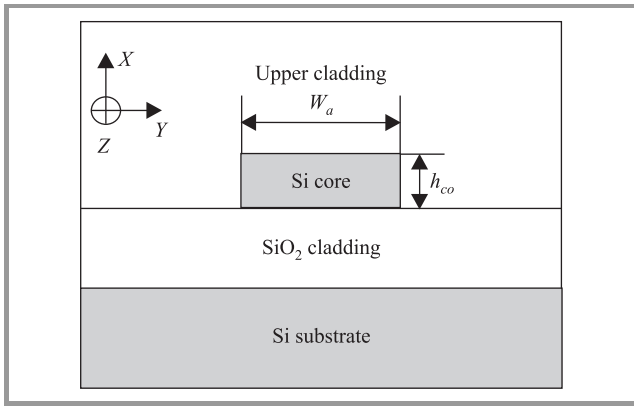


Fig. 4. Waveguide structure used in the simulations.

The parameters used in the designs are as follows: the waveguide has a standard silicon thickness of $h_{co} = 220$ nm and access waveguide widths are $W_a = 0.48$ μm for single mode operation. The refractive index of the silicon core is $n_{\text{Si}} = 3.45$. It is assumed that the designs are for the TE (transverse electric) polarization at a central optical wavelength $\lambda = 1550$ nm. The width of the MMI is $W_{\text{MMI}} = 6$ μm . The access waveguide is tapered to a width of $W_{tp} = 800$ nm to improve device performance.

The 3D-BPM simulations for a 4×4 MMI coupler having a length of $L_1 = \frac{3L_\pi}{4}$ are shown in Fig. 5a for a signal presented at input port 1 and Fig. 5b for a signal at input port 2. The optimized length of the MMI coupler calculated by using 3D-BPM is $L_1 = 71.70$ μm . The excess loss calculated is 0.35 dB for both cases.

If two 4×4 MMI couplers with the same length of $L_{\text{MMI}} = \frac{3L_\pi}{4}$ are cascaded together, then a 4×4 MMI coupler having a length of $L_{\text{MMI}} = \frac{3L_\pi}{2}$ is formed. The transfer matrix of this new 4×4 MMI coupler is given by Eq. (4). The 3D-BPM will now be used to verify this prediction. The 3D-BPM simulations for this 4×4 MMI coupler are shown in Fig. 6. The 3D-BPM simulations show that the power splitting ratios for each MMI coupler used in these microresonators are 0.42/0.43. The optimized length of

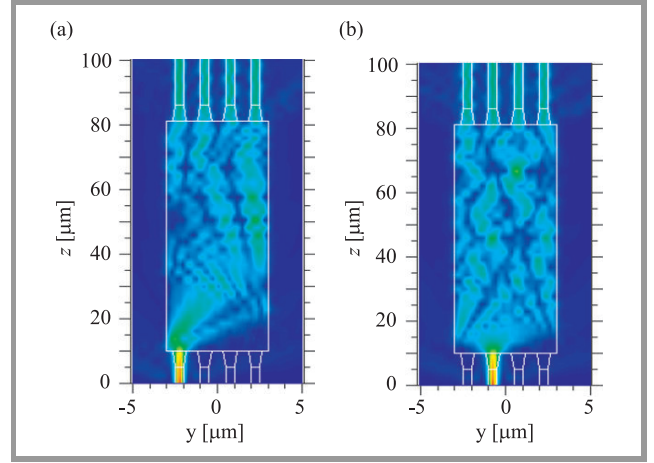


Fig. 5. The 3D-BPM simulations for a 4×4 MMI coupler at length $L_1 = \frac{3L_\pi}{4}$ with input signal at (a) port 1 and (b) port 2.

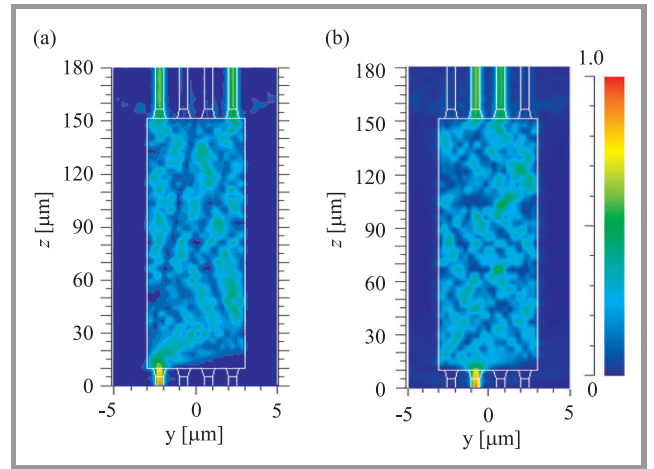


Fig. 6. The 3D-BPM simulations for a 4×4 MMI structures with the signal at (a) input port 1 and (b) input port 2.

each MMI coupler is found to be $L_{\text{MMI}} = 141.7$ μm . For both cases, the excess losses are 0.7 dB and the imbalances are 0.1 dB.

By connecting the above MMI structures (each MMI with length of $L_{\text{MMI}} = \frac{3L_\pi}{2}$) together as shown in Fig. 1, double-microresonators can be achieved. Phase shifters can be introduced in the linking arms of the MZI structure in order to vary the coupling coefficients of the couplers. For example, if phase shifts of $\Delta\phi_1 = \Delta\phi_2 = 0$ are introduced at the linking arms of the MZI, then the 3D-BPM simulation (Fig. 7a) shows that the normalized output powers (for signal at input port 1) at output ports 1, 2, 3 and 4 are 0, 0, 0 and 0.8, respectively. The computed excess loss is 0.96 dB.

If phase shifts $\Delta\phi_1 = \frac{\pi}{2}$ and $\Delta\phi_2 = 0$ are introduced at the linking arms, the 3D-BPM simulation for the signal at input port 1 is shown in Fig. 7b. The normalized output powers at output ports 1, 2, 3 and 4 calculated to be 0.41, 0, 0 and 0.39, respectively. The excess loss is 0.9 dB and the imbalance is 0.2 dB in this case.

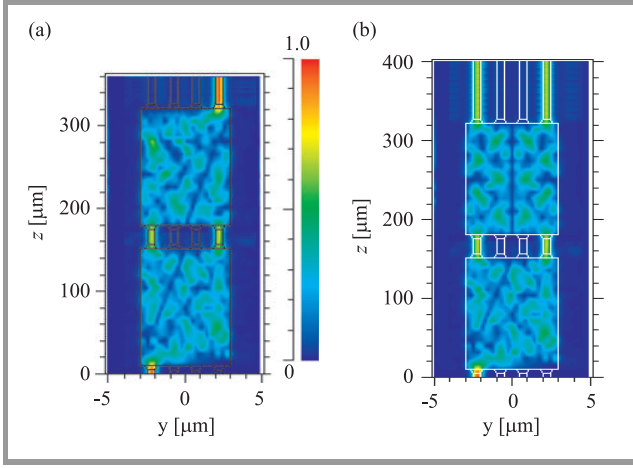


Fig. 7. The 3D-BPM simulations of cascaded MMI couplers used for the microresonator structure when the signal is at input port 1 with (a) phase shifts $\Delta\phi_1 = \Delta\phi_2 = 0$ and (b) $\Delta\phi_1 = \frac{\pi}{2}$ and $\Delta\phi_2 = 0$.

The quality factor (Q) of a single microresonator based on a 2×2 MMI coupler is given by [2]

$$Q \approx \frac{\pi n_g L_R \sqrt{\alpha\tau}}{\lambda(1 - \alpha\tau)}. \quad (9)$$

Here, n_g , L_R , and λ are the group index, length of the racetrack waveguide and wavelength, respectively. τ is the transmission coefficient of the coupler. In order to achieve a high Q microresonator, the coupling coefficient $|\kappa| = \sqrt{\alpha^2 - \tau^2}$ needs to be small. Note that the bend radius also has a strong effect on the Q -factor. However, if the bend radius increases, then the transmission loss will increase, while the bend loss does not significantly decrease. Here α^2 is the power loss factor introduced by MMI couplers and racetrack waveguides (including both bend loss and transmission loss). By varying the phase shifts at the linking arms, the power coupling ratios can be tuned.

The complete device presented in this paper is equivalent to two separate 2×2 MMI-based microresonators. Each microresonator may have different transmission characteristics such as different quality (Q), different free spectral range (FSR), finesse (F) and different bandwidth (BW).

In the following examples, the Q -factors of the microring resonators will be determined. Consider a 4×4 microring resonator with following parameters: bend radii of the first microring resonator and second microring resonator (see Fig. 1) are $R_1 = 5 \mu\text{m}$ and $R_2 = 20 \mu\text{m}$. If the 3 dB MMI couplers with the same length of ($L_{MMI} = 141.7 \mu\text{m}$ (Fig. 6) are used in the structure, then the calculated power transmission coefficients are $|\tau_1|^2 = 0.42$ and $|\tau_2|^2 = 0.42$. The estimated Q -factor for the first microring resonator is $Q_1 \approx 1000$ and for the second microring resonator is $Q_2 = 1650$.

If two 4×4 MMI couplers having the same length of $L_{MMI} = 141.7 \mu\text{m}$ are connected together in order to produce 2×2 tunable MMI couplers (Fig. 7), then power transmission ratios of the two MMI couplers $|\tau_1|^2$ and $|\tau_2|^2$

can be varied by adjusting the phase shifts $\Delta\phi_1$ and $\Delta\phi_2$ in the linking arms, respectively. For example, if phase shifts $\Delta\phi_1 = 0.9\pi$ and $\Delta\phi_2 = 0.5\pi$, then the 3D-BPM shows that the normalized output powers at output ports 1, 2, 3 and 4 are 0.7, 0, 0 and 0.1, respectively. The calculated power transmission coefficients are $|\tau_1|^2 = 0.7$ and $|\tau_2|^2 = 0.41$. Therefore, the estimated Q -factors are $Q_1 \approx 4300$ for the first microring resonator with bend radius $R_1 = 5 \mu\text{m}$ and $Q_2 \approx 2500$ for the first microring resonator with bend radius $R_2 = 20 \mu\text{m}$.

It is interesting to note that provided that the phase shifters operate with sufficient speed, functions of switching, modulating and add-drop multiplexing can be achieved in both microresonators.

4. Conclusions

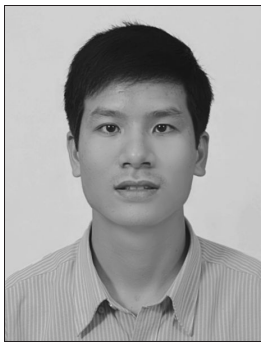
In this paper, we have presented a novel microring resonator structure based on 4×4 MMI couplers. This device acts as two separate microring resonators. The design of the device on the SOI platform has been presented by using the transfer matrix method and the working principle is verified using the 3D BPM method. If the phase shifters operate fast enough, then the device may be a very promising building block for optical switches, optical modulators, and optical add-drop multiplexers.

References

- [1] D. G. Rabus, *Integrated Ring Resonators – The Compendium*. Berlin: Springer-Verlag, 2007.
- [2] D.-X. Xu, A. Densmore, P. Waldron, J. Lapointe, E. Post, and A. Delage, “High bandwidth SOI photonic wire ring resonators using MMI coupler”, *Opt. Expr.*, vol. 15, no. 6, pp. 3149–3155, 2007.
- [3] F. Xia, L. Sekaric, and Y. A. Vlasov, “Mode conversion losses in silicon-on-insulator photonic wire based racetrack resonators”, *Opt. Expr.*, vol. 14, no. 9, pp. 3872–3886, 2006.
- [4] T. T. Le and L. W. Cahill, “The modeling of MMI structures for signal processing applications”, in *Integr. Opt. Dev. Mater. Technol. XII Proc. SPIE*, San Jose, USA, 2008, vol. 6896, pp. 68961G–68961G-7.
- [5] P. A. Besse, E. Gini, M. Bachmann, and H. Melchior, “New 2×2 and 1×3 multimode interference couplers with free selection of power splitting ratios”, *IEEE J. Lightw. Technol.*, vol. 14, no. 10, pp. 2286–2293, 1996.
- [6] A. Yariv, “Critical coupling and its control in optical waveguide-resonator systems”, *IEEE Photon. Technol. Lett.*, vol. 14, no. 4, pp. 483–485, 2002.
- [7] J. M. Choi, R. K. Lee, and A. Yariv, “Control of critical coupling in a ring resonator-fiber configuration: application to wavelength-selective switching, modulation, amplification, and oscillation”, *Opt. Lett.*, vol. 26, no. 16, pp. 1236–1238, 2001.
- [8] M. Bachmann, P. A. Besse, and H. Melchior, “General self-imaging properties in $N \times N$ multimode interference couplers including phase relations”, *Appl. Opt.*, vol. 33, no. 18, pp. 3905–3911, 1994.
- [9] W. P. Huang, C. L. Xu, and S. K. Chaudhuri, “A finite-difference vector beam propagation method for three-dimensional waveguide structures”, *IEEE Photon. Technol. Lett.*, vol. 4, no. 2, pp. 148–151, 1992.
- [10] W. P. Huang, C. L. Xu, W. Lui, and K. Yokoyama, “The perfectly matched layer (PML) boundary condition for the beam propagation method”, *IEEE Photon. Technol. Lett.*, vol. 8, no. 5, pp. 649–651, 1996.

- [11] D. Dai and S. He, "Design of an ultrashort Si-nanowaveguide-based multimode interference coupler of arbitrary shape", *Appl. Opt.*, vol. 47, no. 19, pp. 38–44, 2008.
- [12] D. Dai and S. He, "Optimization of ultracompact polarization-insensitive multimode interference couplers based on Si nanowire waveguides", *IEEE Photon. Technol. Lett.*, vol. 18, no. 19, pp. 2017–2019, 2006.
- [13] E. Dulkeith *et al.*, "Group index and group velocity dispersion in silicon-on-insulator photonic wires", *Opt. Expr.*, vol. 14, no. 9, pp. 3853–3863, 2006.
- [14] J. I. Dadap *et al.*, "Nonlinear-optical phase modification in dispersion-engineered Si photonic wires", *Opt. Expr.*, vol. 16, no. 2, pp. 1280–1299, 2008.

Note: this text is reprinted from JTIT 2/2009



Trung-Thanh Le received the B.E. and M.E. degrees in electronic and telecommunication engineering from the Hanoi University of Technology, Vietnam, in 2003 and 2005, respectively. Since 2003, he has been a lecturer at the National University of Transportation and Communications, Hanoi. He works towards the Ph.D. degree

in electronic engineering at the La Trobe University, Melbourne, Australia, since 2006. His research interests are

integrated optical devices and photonic signal processing, especially multimode interference based devices.

e-mail: thanh.latrobe@gmail.com

Department of Electronic Engineering

La Trobe University

Melbourne, Vic 3086, Australia



Laurence W. Cahill received the B.E., M.Sc. and Ph.D. degrees from the University of Melbourne, Australia. He is a Professor in the Department of Electronic Engineering, La Trobe University, Australia. He is a senior member of the IEEE, member of SPIE and a Fellow of Engineers Australia. His research interests lie in the area of

optical and photonics. Particular interests are semiconductor laser dynamics, high speed circuits for optical sources and receivers, mid infrared detectors, photonic switching, fibre sensors, computer aided design, image processing, and signal processing.

e-mail: L.Cahill@latrobe.edu.au

Department of Electronic Engineering

La Trobe University

Plenty road, Bundoora Campus

Melbourne, Vic 3086, Australia

Percolation Driven Flooding for Energy Efficient Routing in Dense Sensor Networks

Gergely Vakulya and Gyula Simon

Abstract—Simple flooding algorithms are widely used in ad hoc sensor networks either for information dissemination or as building blocks of more sophisticated routing protocols. In this paper a percolation driven probabilistic flooding algorithm is proposed, which provides large message delivery ratio with small number of sent messages, compared to traditional flooding. To control the number of sent messages the proposed algorithm uses locally available information only, thus induces negligible overhead on network traffic. The performance of the algorithm is analyzed and the theoretical results are verified through simulation examples.

Keywords— *flood routing, percolation, sensor network.*

1. Introduction

Energy efficiency is a key question in wireless sensor networks made of large number of inexpensive nodes with limited energy reserves and restrained processing and communication capabilities. Such networks have been used in a wide range of applications with various goals, using different platforms and technology [1]–[4]. Independently of the nature of the application, however, certain middleware services are always present, one of them being routing. Limited resources on the sensor nodes require that routing protocols be simple, energy-conserving and robust.

Depending on the actual deployment scenario and available hardware, several ad hoc routing algorithms have been proposed. Location aware routing schemes, e.g., greedy perimeter stateless routing [5], location aided routing [6], distance routing effect algorithm [7] use position information of the nodes obtained from global positioning system (GPS) or other localization services. Other data centric schemes do not use location information, e.g., directed diffusion [8], ad hoc on demand distance vector routing (AODV) [9], dynamic source routing [10], temporarily ordered routing [11], or zone routing [12], just to name a few. To allow sensor networks better cope with scaling, hierarchical protocols were designed, e.g., low energy adaptive clustering hierarchy (LEACH) [13] and threshold sensitive energy efficient sensor network protocol (TEEN) [14].

The simplest routing protocol is flooding [15], which is a useful simple (but inefficient) protocol in itself, and also is a building block for several more sophisticated algorithms, e.g., directed flood routing [16], controlled flood routing [17], tiny ad hoc routing [18], or AODV [9], which is used in Zigbee's routing protocol [19].

The basic flooding protocol is the following:

1. *Start*. When a source node intends to send a message to all other nodes in the network, it broadcasts the message to its neighbors.
2. *Relay*. When a node receives a message for the first time, it rebroadcasts the same message to its neighbors. All other copies of the same message, received later, will be discarded by the recipient node.

Flooding naturally can be used in its basic form for network-wide dissemination of information. In this context the important performance criteria of the algorithm are delivery ratio, latency, and the energy efficiency. Flooding is also used for route discovery [9]. Route discovery protocols find a route from a source node to a destination node in two phases. First the network is flooded by a discovery request message originated from the source, and then a reply message is sent back from the destination to the source node. In this way the network learns the path from the source to the destination and this route will be used in the subsequent communication sessions. In this context the quality – primarily the length – of the discovered route is an important performance criterion as well.

Apart from its simplicity the main advantage of the flood routing protocol lies in its robustness: the implicit redundancy built in the algorithm provides resistance against high degree of message loss and node failures as well. The drawback of the algorithm is the large number of packets transmitted in the network, referred to as broadcast storm: whether it is necessary or not, each node will retransmit each message. In a dense network the unnecessarily high number of messages can cause frequent collisions (and thus performance degradation) and it wastes network energy as well [20].

To reduce the number of routing messages, probabilistic variations of the flood routing were proposed. The main idea behind these algorithms is that a node randomly decides whether to forward a message or not. Several variants of random flood algorithms were proposed. According to the simplest protocol, a node will rebroadcast a message after its first reception with probability p and discards it with probability $1 - p$ (clearly, $p = 1$ results in basic flooding).

Test results verify what intuition suggests: higher p values provide higher network coverage than small p values. Moreover, in sufficiently large and sufficiently dense random networks there is an interesting bimodal behavior of

the algorithm: if $p > p_c$, where the critical probability p_c depends on the network topology, the message reaches practically all nodes in the network, otherwise only a small portion of the network receives the message. Thus the optimal choice clearly would be $p = p_c$. Modified algorithms try to further increase the performance by various modifications; e.g., premature message death can be avoided by varying p as a function of hop-distance from the source: nodes close to the source rebroadcast messages with higher probability, while distant nodes use smaller p . A comprehensive study on random flooding algorithms can be found in [15]. The common problem with these algorithms is that the design parameters (probabilities) depend on actual network layout, no automatic or adaptive solution is known. Especially in a network with varying node density the probabilistic algorithms tuned to work reliably will be suboptimal.

More sophisticated adaptive schemes can successfully handle networks with varying properties as well. Location based algorithms use node positions to optimize retransmissions [21], graph-based algorithms utilize connectivity information up to 2-hop neighbors to construct a dominating set in a distributed way [22]. Other heuristic rules, e.g., message counters and neighbor coverage schemes were also proposed [20]. These schemes can adapt automatically to changing topology at the price of higher complexity.

Sensor networks are often modeled by the Poisson-Boolean model. In this model nodes are scattered on the plane by a Poisson point process with constant density λ , and each node has a fixed communication radius r (disc model). Two nodes can communicate with each other if their distance is less than r . The Poisson point process applies reasonably well to controlled, but still random deployments. The simplistic disc model, however, is far from reality, where communication range is not circularly symmetric, but rather has an anisotropic shape. Speaking of connectivity, however, the disc model has an important property: it can be considered the worst case model, since other transmission models provide easier percolation under similar circumstances [23].

Percolation theory [24] has important impact on wireless communication networks. Results in continuum percolation theory show that if the density of transmitting stations (or alternatively, their communication power) in a Poisson-Boolean network is greater than a critical value λ_c then an unbounded connected component is formed with probability one (the network percolates) [23], [25]. On the other hand, if $\lambda < \lambda_c$ then all connected components are finite with probability one. Thus percolation is an important property of a network if long distance multi-hop communication is required.

In this paper a new percolation inspired solution will be proposed based on probabilistic flood routing algorithms. The new algorithm adaptively sets the rebroadcast probability based on the network topology, using locally available neighborhood information only. The proposed algorithm provides high performance with a low number of

messages and can adapt its behavior dynamically to the network properties. The proposed algorithm is only marginally more complex than the basic flood routing thus it can be successfully used in applications with limited resources as well.

The rest of the paper is organized as follows. In Section 2 related works will be summarized. In Section 3 the proposed algorithm will be introduced. An important property of the algorithm will be proven based on percolation theory: every message broadcast in the network will reach infinite number of nodes almost surely. In practice this property ensures that the message reaches almost all of the nodes in a finite-sized network. The performance evaluation of the algorithm through simulation can be found in Section 4. Conclusions are drawn in Section 5.

2. Related Work

Results of percolation theory have been applied to wireless networks, forming the theory of continuum percolation [26]. More realistic network models, i.e., unreliable communication links and anisotropic radiation patterns were studied in [23]. An important result of [23] shows that the debatable disc model actually provides a conservative estimate of the network connectivity: percolation in networks with disc communication shapes is more difficult than in networks with any other convex communication shape with the same surface ("discs are hardest" conjecture). This property justifies the usage of the disc model when connectivity issues are studied.

Results from percolation theory have been infiltrating various recent sensor networking algorithms. As a practical application of this phenomenon, the distributed minimum-link-degree rule was proposed to counterbalance local spatial inhomogeneities in ad hoc networks [25]. To provide network wide connectivity, the transmission powers of the nodes are tuned so that each node has at least a minimum number of neighbors.

In [27] a distributed energy saving mechanism is used by switching the nodes in the network on and off, while providing percolation in the network. The simple rule for the algorithm is to keep the active time ratio of the nodes larger than λ/λ_c . This algorithm does not scale well since each node in the network has to know the global parameter λ . The idea was further elaborated in [28], where a practically useable distributed algorithm was proposed for the control of the active time ratio. In this algorithm only the local density, namely the number of neighbors is used. For a node with degree k the active ratio η_k is defined as follows:

$$\eta_k = \begin{cases} \frac{\phi}{k} & k > \phi \\ 1 & k \leq \phi \end{cases},$$

where ϕ is a density independent design parameter. A similar idea will be used in the proposed flooding algorithm.

3. Percolation Driven Flooding

The idea behind the percolation driven flood routing is the following: use probabilistic flooding and set the retransmission probability adaptively at every node so that with high probability the message reaches almost all of the nodes in the network.

The algorithm is the following:

1. The source node broadcasts the message with probability one.
2. After the first reception of a message node n rebroadcasts it with probability p_n , and discards it with probability $1 - p_n$. Copies of the same message received multiple times are discarded with probability one. Probability p_n is defined as

$$p_n = \begin{cases} \frac{K_{\min}}{K_n} & K_n > K_{\min} \\ 1 & \text{otherwise} \end{cases}, \quad (1)$$

where: K_n is the degree of node n (i.e., the number of neighbors of node n), and the $K_{\min}$ design parameter is the required minimum number of neighbors. Two nodes are neighbors if they can hear each other (i.e., a symmetric link exists between them).

The algorithm requires neighbor discovery to determine the number of neighbors. Each node can gather this local information by transmitting and receiving *hello messages*. In practical situations parameter $K_{\min}$ is chosen around 7, as will be justified by the simulation examples.

Theorem 1: If in the infinite random network, generated by a Poisson-Boolean process with density λ and communication radius r , there exists design parameter $K_{\min} < \infty$ independent of λ , such that when node n with degree K_n transmits a message with probability according to Eq. (1), then each message reaches infinite number of nodes with probability one.

Proof: Let us denote the graph representing the network by $G(\lambda, r, 1)$ where λ is the density, r is the communication radius, and the third parameter is a probabilistic value used in the generator process (used and discussed later) and is simply set to one at the moment. Using scaling, we define another process with density $\lambda' = \lambda r^2$ and communication radius 1. This process yields the same graph $G(\lambda', 1, 1)$. Now let us use the following Theorem 2 [28].

Theorem 2: Given $G(\lambda', 1, 1)$ with $\lambda' > \lambda'_c$, there exists $2 < \varphi < \infty$, independent of λ , such that when each node with degree k is active with probability:

$$p(k) = \begin{cases} \frac{\varphi}{k} & k > \varphi \\ 1 & k \leq \varphi \end{cases},$$

then $G(\lambda', 1, p(\cdot))$ is percolated. The proof of Theorem 2 can be found in [28].

If in the network we decide *before* a particular message is broadcasted which nodes will retransmit the message (when

they receive it) by activating/deactivating nodes with probability (1) we get $G(\lambda', 1, p(\cdot))$. In this way we construct a virtual network for each message and we can apply Theorem 2 to it. Thus there exist $K_{\min} = \varphi < \infty$ so that G percolates. From this it follows that the message will reach infinite number of nodes [24]. ■

In real-world finite-size networks Theorem 1 means that with an appropriate choice of $K_{\min}$ each message reaches practically all nodes in the network. Further properties of the algorithm will be analyzed through simulation examples.

4. Performance Evaluation

In this section the performance of the algorithm will be evaluated through simulation results. First the performance metrics and the network/communication models will be introduced, followed by the simulation environment and the simulation scenarios. Finally, the results of the simulations will be presented to illustrate the behavior of the proposed algorithm.

4.1. Performance Metrics

The main performance criteria of broadcast algorithms are the message delivery ratio and the number of sent messages. If the number of nodes in the network is N and the number of nodes receiving the message is N_{rec} then the coverage ratio is defined as $C_r = \frac{N_{rec}}{N}$.

Clearly, C_r close to 1 is required for a good quality of service.

Another quality metric is the total number of sent packets M while a message is propagated in the network. Trivially, for basic flood routing $M = N$, if $C_r = 1$. We expect lower number of messages and still high delivery ratio for a good performance algorithm. For easier comparison, we will use the normalized number of messages defined as

$$M_{norm} = \frac{M}{N}.$$

If flooding is used for route discovery, an important performance metric is the length L of the discovered route.

Other possible metrics are message delay/latency, length of flood period, and number of collisions. The actual low level details, e.g., implementation, hardware and media access control (MAC) layer properties are not investigated in detail in this paper, but their effect is examined through high level parameters they have impact on, i.e., C_r , M_{norm} , and L will be studied.

4.2. Network Model

To model random deployment of nodes or possibly mobile networks, in the simulations nodes are placed at random, according to a uniform distribution on a two-dimensional area. Finite communication distances are represented by the disc model, where a communication link is assumed to exist between two nodes if they are less than the communication radius apart. According to the results published

in [24], this – otherwise rather simple and idealistic – model can be considered a worst case model.

Our communication channel model does not deal with specific details of the physical layer or the MAC layer; rather we use a probabilistic model: a message can reach a neighbor within the communication radius with probability $p_{rec} < 1$. This high-level model represents message losses due to collision, fading, or other disturbances as well. The model does not distinguish between individual phenomena but rather incorporates the different sources in one parameter, thus some aspects of reality (e.g., inter-message dependencies) are neglected, but the model is faithful enough to provide useful and easy means for testing.

4.3. Simulation Environment

To perform high-speed simulation, a simulator was written in C to validate the efficiency of the proposed algorithm. The program places nodes randomly according to the different test scenarios considered. The number of nodes N , the communication radius r , and K_{min} are input parameters.

In each simulation, a new placement is generated, and a source node starts transmission. The simulator returns the size of connected component containing the source node, the number of nodes receiving the packet, and the total number of messages. Optionally, the software can visualize the topology of the network, the active data paths, and the nodes' reception status.

4.4. Test Scenarios

In the test we used 3 different scenarios to model sensor network setups. The first scenario is a random uniform distribution (with constant density), while in the second

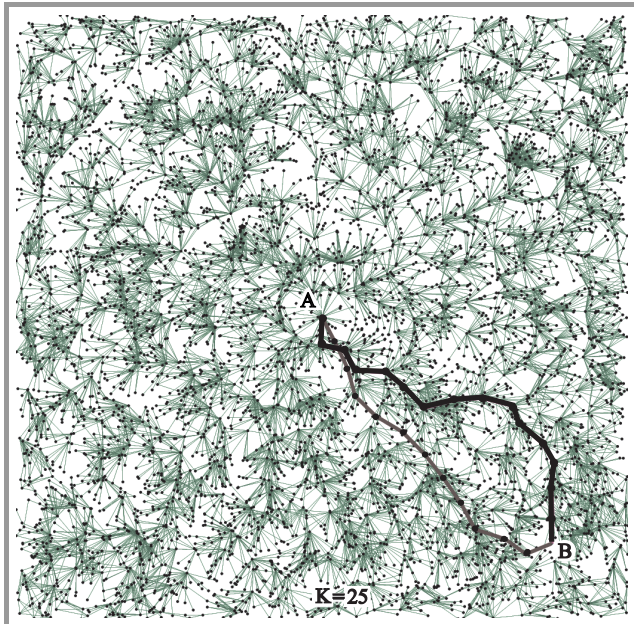


Fig. 1. Uniform random distribution scenario with 3000 nodes. The average number of neighbors is 25. Thin and thick lines show examples for the message paths between nodes A and B, discovered by flood and percolation driven flood, respectively.

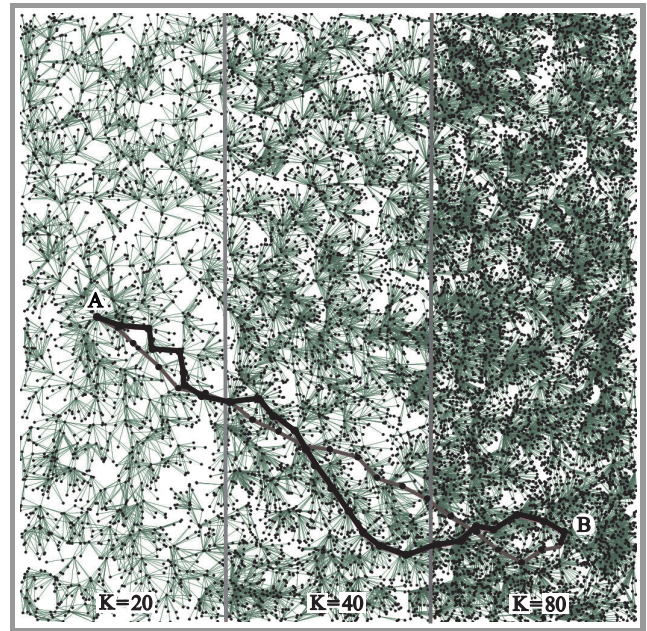


Fig. 2. The 3-density scenario. The average numbers of neighbors in the regions from left to right are 20, 40, and 80. Thin and thick lines show examples for the message paths between nodes A and B, discovered by flood and percolation driven flood, respectively.

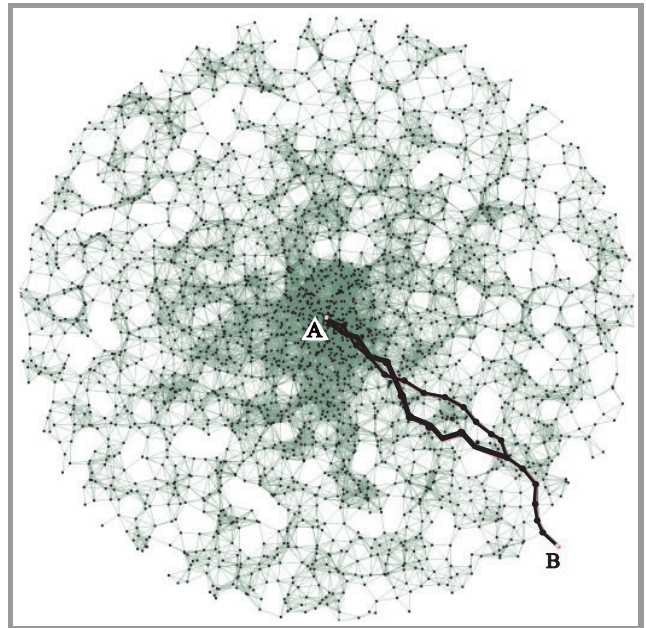


Fig. 3. The circular scenario with 3000 nodes. The average number of neighbors is 100 in the centre and 2 at the perimeter. Thin and thick lines show examples for the message paths between nodes A and B, discovered by flood and percolation driven flood, respectively.

scenario we used three regions; in each region nodes were placed with uniform random distribution, but with different densities. In the third scenario the density is increasing towards the centre, thus there is continuous change and extreme deviation in local density. Typical examples for the test networks can be seen in Figs. 1, 2 and 3.

4.5. Test Results

To test coverage ratio C_r and the number of messages M we placed 3000 nodes in a square-shaped area, as shown in Fig. 1.

We varied the communication radius to provide different network densities: the average number of neighbors K was set to 10, 30, and 100. The $K_{\min}$ value varied from 1 to 20 in 0.25 steps. We run the simulation 100 times for each communication radius, $K_{\min}$, and p_{rec} values, thus each point in the subsequent graphs is the average of 100 experiments.

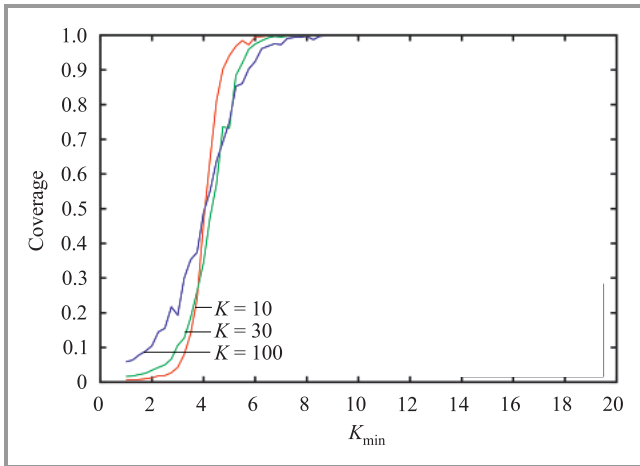


Fig. 4. Coverage versus $K_{\min}$ for different network densities in the uniform distribution scenario (constant network density), $p_{rec} = 1$.

Figure 4 presents coverage versus $K_{\min}$, while p_{rec} was set to constant 1 (error-free communication). The graphs for different network densities are similar: when $K_{\min}$ is low ($K_{\min} < 3$), the message delivery ratio is very low. If $K_{\min}$ is increased, the coverage is increasing very quickly, the critical value being around $K_{\min} \approx 4.5$. Coverage reaches 90% at $K_{\min} \approx 5$. If $K_{\min} > 7$, practically all nodes in the network receive the message. Figure 4 illustrates that percolation driven flood can indeed be used in networks with different densities: $K_{\min}$ is independent of the actual network density.

In Fig. 5 the normalized number of sent messages M_{norm} is shown versus $K_{\min}$, for the previous experiments. In case of the conventional flood routing $M_{norm} = 1$, because all nodes relay the received message. As Fig. 4 illustrates, $K_{\min} > 7$ gives almost perfect delivery ratio. Around this value $K_{\min} \approx 7$ the total number of messages is greatly decreased, as presented in Fig. 5. Depending on the node density, in the tests the total number of messages were reduced by 30..95%, higher reduction rates belonging to higher densities.

Clearly, in networks, where the average node degree K is only slightly higher than $K_{\min}$ the number of messages can be decreased only moderately, while in dense networks a much lower number of messages is enough to provide good delivery ratio, as shown in Fig. 5. According

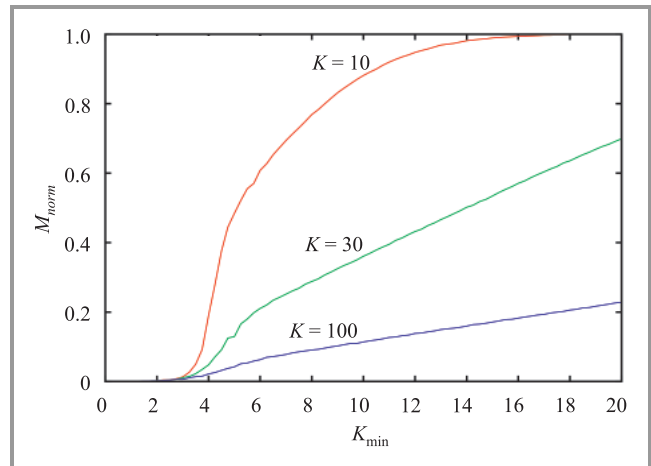


Fig. 5. Total number of messages versus $K_{\min}$ for different network densities in the uniform distribution scenario (constant network density), $p_{rec} = 1$.

to Figs. 4 and 5, in dense networks the percolation driven flood algorithm can effectively reduce the number of message while maintaining good coverage.

Figure 6 illustrates the effect of unreliable communication links. In the experiment constant density ($K = 30$) was used, while the p_{rec} reception probability was varied between 0.3 and 1. The figure shows that the algorithm

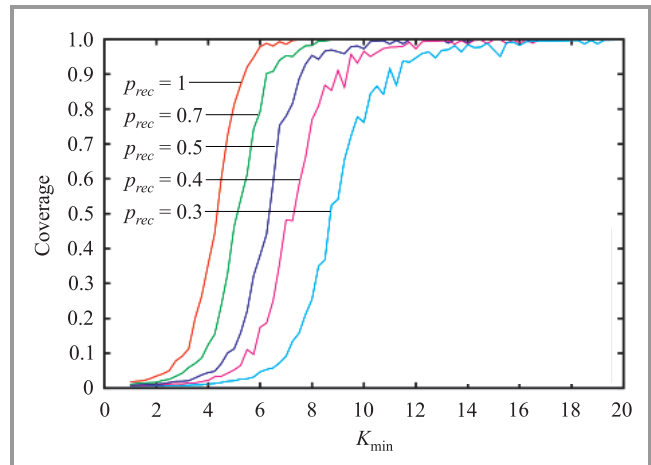


Fig. 6. Coverage versus $K_{\min}$ for different p_{rec} in the uniform distribution scenario (constant network density), $K = 30$.

can provide high coverage even in the presence of bad quality communication links, but as intuitively expected, higher $K_{\min}$ is necessary as the communication channel degrades.

Figure 7 presents the associated number of messages for the unreliable communication experiment.

Percolation driven flood routing algorithm is capable of handling varying network densities. To test this property we used the scenario illustrated in Fig. 2, where three areas with different densities are present. Clearly, a constant retransmission probability would either be suboptimal (set to provide sufficient coverage in the less dense area

as well) or would not provide good coverage in the sparse regions of the network.

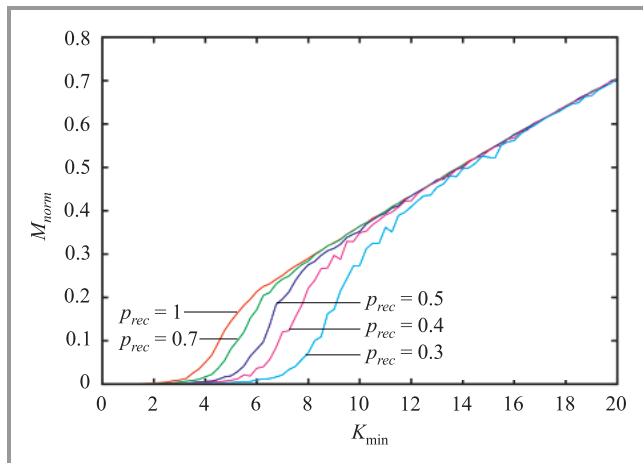


Fig. 7. Number of total messages versus $K_{\min}$ for different p_{rec} in the uniform distribution scenario (constant network density), $K = 30$.

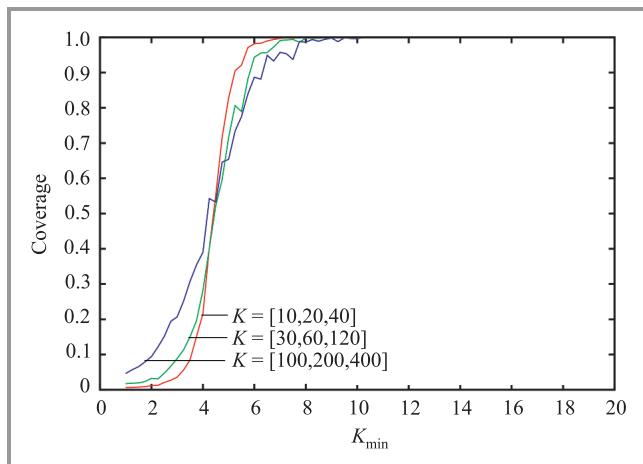


Fig. 8. Coverage versus $K_{\min}$ for different network densities in the varying network density scenario, $p_{\text{rec}} = 1$.

Figure 8 illustrates the effectiveness of the percolation driven flood algorithm, showing the coverage versus $K_{\min}$ when the densities of regions from left to right were set to [10, 20, 40], [30, 60, 120], and [100, 200, 400]; and $p_{\text{rec}} = 1$. The behavior is quite similar to that of the first scenario (constant density): percolation happens around the same $K_{\min} \approx 5$, and practically full coverage can be provided if $K_{\min} > 7$.

The associated total numbers of messages are shown in Fig. 9, where the gain with respect to the basic flooding algorithm is apparent.

In Fig. 10 the number of coverage versus $K_{\min}$ is illustrated in the circular scenario. In the experiments the number of neighbors varied between [120..4], [220..20] and [525..50], respectively. The figure clearly presents the excellent performance of the algorithm even if there are extreme differences in the number of neighbors of each node. The ideal value for $K_{\min}$ is around 7, in this case the algorithm gives

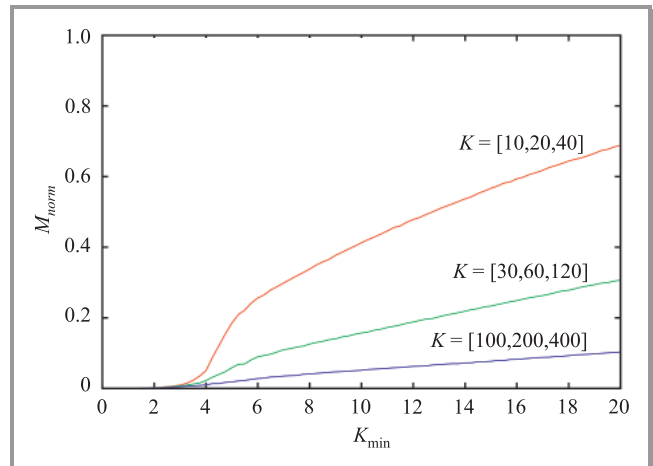


Fig. 9. Number of total messages versus $K_{\min}$ for different network densities in the varying network density scenario, $p_{\text{rec}} = 1$.

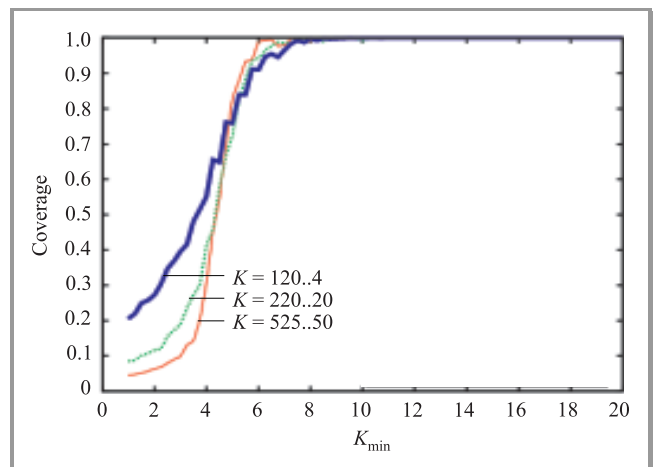


Fig. 10. Coverage versus $K_{\min}$ for different network densities in the circular scenario, $p_{\text{rec}} = 1$.

at least 95% delivery ratio. In case of $K_{\min} = 8$ practically full coverage is provided.

In Fig. 11 the normalized number of messages is shown for the three different density intervals. In case of $K_{\min} = 8$,

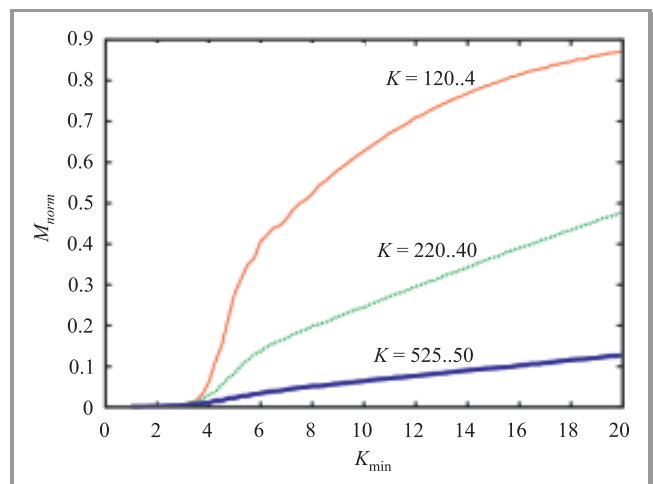


Fig. 11. Number of total messages versus $K_{\min}$ for different network densities in the circular scenario, $p_{\text{rec}} = 1$.

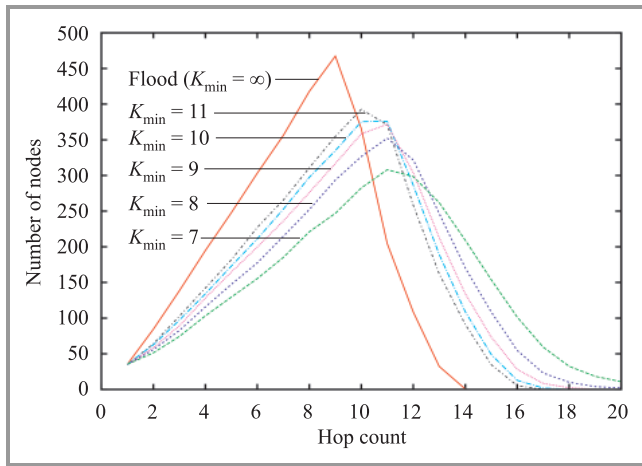


Fig. 12. Distribution of the hop counts in the uniform distribution network for different $K_{\min}$ values, $p_{rec} = 1$.

the algorithm gives a gain of 55–95%, with respect to the original flooding.

The length L of the discovered route is important when flooding is used for route discovery. To test the properties of percolation driven flooding in this respect we measured hop distances discovered by flooding and percolation driven flooding, for different $K_{\min}$ values. In the tests the uniform distribution (constant density) scenario was used with $K = 35$ and $p_{rec} = 1$. The source node was placed to the center of the field and hop distances to all other nodes were measured. The histogram created from averaging 100 independent experiments is illustrated in Fig. 12.

The distribution is linear for small hop distances. The explanation is shown in Fig. 13 for the ideal case: messages are propagated in belt-shaped increments. The areas of the subsequent belts increase linearly, thus the number of nodes in each belt is linearly increasing. If the node density is smaller, the number of nodes in each belt is smaller, thus the slope of the histogram will be smaller. According to Fig. 12, smaller $K_{\min}$ also causes decrease in the slope: e.g., for $K_{\min} = 9$, the slope is 40% smaller than in the basic flooding case. This means that the detected route is 40% longer than that of the basic flooding.

The declining part at higher hop distances is due to the finite size of the network: the messages reach the network

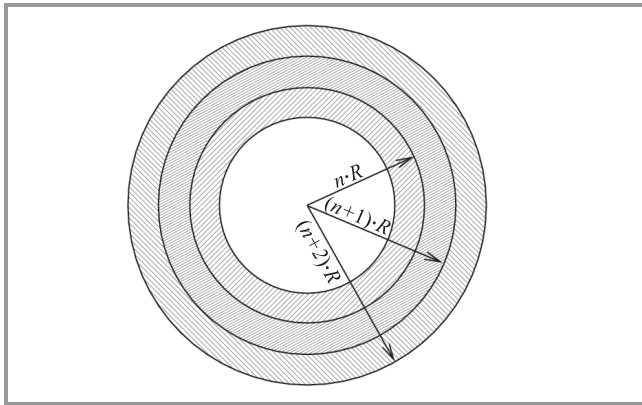


Fig. 13. Ideal propagation of a message in a dense network.

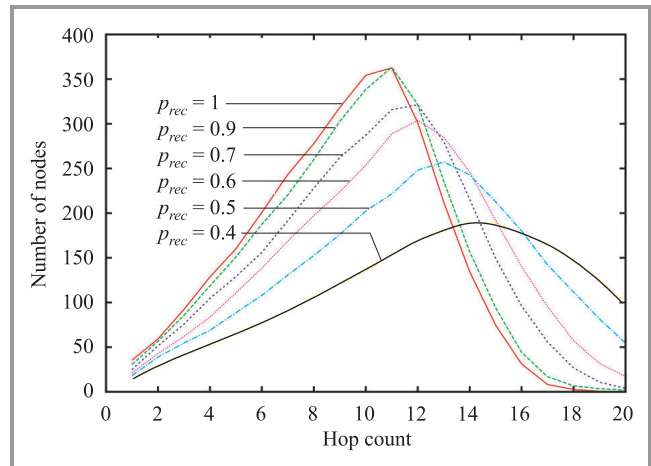


Fig. 14. Distribution of the hop counts in the uniform distribution network for different p_{rec} values, $K_{\min} = 9$.

perimeter and eventually die. The tail of the distribution presents the maximum hop distance from the source, which is in the square setup ideally $\sqrt{2}$ times the hop distance measured at the maximum of the histogram.

The effect of unreliable links is shown in Fig. 14. The histogram shows the average of 100 experiments conducted in the uniform test scenario with $K = 35$, and $p_{rec} = 0.4..1$. As intuitively expected, unreliable links cause smaller slope, i.e., higher hop distances. For example, link quality $p_{rec} = 0.7$ results in the increase of hop distance by 18%, while for the rather unreliable link quality $p_{rec} = 0.4$ the increase was as high as 63%.

As simulation examples show, percolation driven flooding results in considerably higher hop distances than basic flood routing when small $K_{\min}$ parameters are used. Naturally, low link quality also causes larger hop distances. This effect may be an undesired side effect in route discovery protocols, thus in these applications basic flooding may be more advantageous.

5. Conclusions

A percolation driven flood routing algorithm was proposed, which uses only locally available neighborhood information to reduce broadcast storm. The algorithm is able to massively reduce the number of messages in the network and maintaining high delivery ratio at the same time. Theoretical results prove the usefulness of the algorithm: it is able to provide high coverage, if the network density is high enough.

Simulation tests were performed to validate the performance of the algorithm. The proposed algorithm reduced the total number of messages in the network by 30–95%, depending on the network density, while the coverage was almost 100%, with appropriate choice of $K_{\min}$ ($K_{\min} > 7$). The percolation driven flood algorithm is adaptive to changing node density thus provides high coverage with low number of messages in all scenarios. The algorithm is robust in the presence of unreliable links as well.

According to test result, the percolation driven flood routing algorithm results considerably (even 50%) longer hop distances when used for route discovery purposes.

The overhead of the algorithm is very low since only the number of neighbors must be known by each node. This information can be gained by simple neighborhood discovery protocols, e.g., sending and receiving *hello messages*.

For route discovery purposes the application of the traditional flooding is more advantageous, if the side effect of the longer routes is undesirable. On the other hand, the proposed algorithm is a superior and robust alternative to traditional flooding algorithm in dense networks for message dissemination purposes.

References

- [1] A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson, "Wireless sensor networks for habitat monitoring", in *Proc. WSN Conf.*, Atlanta, USA, 2002, pp. 88–97.
- [2] A. Arora *et al.*, "A line in the sand: a wireless sensor network for target detection, classification, and tracking", *Comput. Netw.*, vol. 46, no. 5, pp. 605–634, 2004.
- [3] A. Ledeczi *et al.*, "Countersniper system for urban warfare", *ACM Trans. Sen. Netw.*, vol. 1, no. 2, pp. 153–177, 2005.
- [4] T. He *et al.*, "VigilNet: an integrated sensor network system for energy-efficient surveillance", *ACM Trans. Sen. Netw.*, vol. 2, no. 1, pp. 1–38, 2006.
- [5] B. Karp and H. T. Kung, "Greedy perimeter stateless routing (GPSR) for wireless networks", in *Proc. ACM MobiCom Conf.*, Boston, USA, 2000, pp. 243–254.
- [6] Y. B. Ko and N. H. Vaidya, "Location-aided routing (LAR) in mobile ad hoc networks", in *Proc. ACM MobiCom Conf.*, Dallas, USA, 1998, pp. 66–75.
- [7] S. Basagni, I. Chlamtac, V. R. Syrotiuk, and B. A. Woodward, "A distance routing effect algorithm for mobility (DREAM)", in *Proc. ACM MobiCom Conf.*, Dallas, USA, 1998, pp. 76–84.
- [8] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva, "Directed diffusion for wireless sensor networking", *IEEE/ACM Trans. Netw.*, vol. 11, no. 1, pp. 2–16, 2003.
- [9] C. E. Perkins and E. M. Royer, "Ad-hoc on-demand distance vector routing", in *Proc. 2nd IEEE Worksh. Mob. Comput. Syst. Appl.*, New Orleans, USA, 1999, pp. 90–100.
- [10] D. B. Johnson and D. A. Maltz, *Dynamic Source Routing in Ad Hoc Wireless Networks*. Boston: Kluwer, 1996.
- [11] V. Park and M. S. Corson, "A highly adaptive distributed routing algorithm for mobile wireless networks", in *Proc. IEEE Infocom Conf.*, Kobe, Japan, 1997, pp. 1405–1413.
- [12] Z. Haas and M. Pearlman, "The performance of query control schemes for the zone routing protocol", in *Proc. ACM SIGCOMM Conf.*, Vancouver, Canada, 1998, pp. 167–177.
- [13] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless sensor networks", in *Proc. Hawaii Int. Conf. Syst. Sci.*, Hawaii, USA, 2000.
- [14] A. Manjeshwar and D. P. Agrawal, "TEEN: a protocol for enhanced efficiency in wireless sensor networks", in *Proc. 1st Int. Worksh. Paral. Distrib. Comput. Iss. Wirel. Netw. Mob. Comput.*, San Francisco, USA, 2001.
- [15] Z. J. Haas, J. Y. Halpern, and L. Li, "Gossip-based ad hoc routing", *IEEE/ACM Trans. Netw.*, vol. 14, no. 3, pp. 479–491, 2006.
- [16] M. Maróti, "Directed flood-routing framework for wireless sensor networks", in *Proc. 5th ACM/IFIP/USENIX Int. Conf. Middlew.*, Toronto, Canada, 2004, pp. 99–114.
- [17] L. H. Costa, M. D. De Amorim, and S. Fdida, "Reducing latency and overhead of route repair with controlled flooding", *Wirel. Netw.*, vol. 10, no. 4, pp. 347–358, 2004.
- [18] P. Gburzynski and W. Olesinski, "On a practical approach to low-cost ad hoc wireless networking", *J. Telecommun. Inform. Technol.*, no. 1, pp. 29–42, 2008.
- [19] Zigbee Alliance [Online]. Available: <http://www.zigbee.org>
- [20] Y. C. Tseng, S. Y. Ni, and E. Y. Shih, "Adaptive approaches to relieving broadcast storms in a wireless multihop mobile ad hoc network", in *Proc. 21st Int. Conf. Distrib. Comput. Syst.*, Phoenix, USA, 2001, pp. 481–488.
- [21] J. Yang, B. Kim, M.-T. Sun, and T.-H. Lai, "Location-aided broadcast in wireless ad hoc networks", *J. Inform. Sci. Eng.*, vol. 23, no. 3, pp. 869–884, 2007.
- [22] J. Wu, F. Dai, M. Gao, and I. Stojmenovic, "On calculating power aware connected dominating set for efficient routing in ad hoc wireless networks", *J. Commun. Netw.*, vol. 5, no. 2, pp. 169–178, 2002.
- [23] M. Franceschetti, L. Booth, M. Cook, R. Meester, and J. Bruck, "Percolation in multi-hop wireless networks", in *Caltech Paradise, ETR055*, 2003 [Online]. Available: <http://www.paradise.caltech.edu/papers/etr055.pdf>
- [24] G. Grimmet, *Percolation*. New York: Springer, 1989.
- [25] I. Glauche, W. Krause, R. Sollacher, and M. Greiner, "Continuum percolation of wireless ad hoc communication networks", *Physica A*, vol. 325, no. 3–4, pp. 577–600, 2003.
- [26] J. Jonasson, "Optimization of shape in continuum percolation", *Ann. Probab.*, vol. 29, no. 2, pp. 624–635, 2001.
- [27] O. Dousse, P. Mannersalo, and P. Thiran, "Latency of wireless sensor networks with uncoordinated power saving mechanisms", in *Proc. 5th ACM Int. Symp. MobiHoc'04*, Tokyo, Japan, 2004, pp. 109–120.
- [28] Z. Kong and E. M. Yeh, "Distributed energy management algorithm for large-scale wireless sensor networks", in *Proc. 8th ACM Int. Symp. MobiHoc'07*, Montreal, Canada, 2007, pp. 209–218.

Note: this text is reprinted from JTIT 2/2009



Gergely Vakulya received his M.Sc. in information technology from the University of Pannonia, Hungary, in 2008. Currently he is pursuing his Ph.D. in information science and technology at the University of Pannonia. His research interest is sensor networking.

e-mail: vakulya@dcs.uni-pannon.hu
University of Pannonia
Egyetem u. 10
8200 Veszprém, Hungary



Gyula Simon received his M.Sc. and Ph.D. in electrical engineering from the Budapest University of Technology, Hungary, in 1991 and 1998, respectively. Currently he is an Associate Professor at the University of Pannonia. His main research interest includes adaptive signal processing and sensor networks.

e-mail: simon@dcs.uni-pannon.hu
University of Pannonia
Egyetem u. 10
8200 Veszprém, Hungary

A Microscopic Approach for THz Intersubband Challenges

Mauro F. Pereira

Abstract—The main candidate to be a practical and low cost high power THz source is the intersubband-based quantum cascade laser, which can have a tremendous impact in many practical applications, including last mile and indoor telecommunication systems. In this review we discuss current challenges for THz intersubband device development from a microscopic point of view. Next summarize the search for new mechanisms and structure designs that can lead to intersubband gain without population inversion. This is a very important topic of current research, since is both an extremely elegant phenomenon from the basic physics of view and crucial for effective lasing in the THz range. The reason is that scattering phenomena can lead to level broadenings of the same order of magnitude of the lasing transitions, making population inversion by carrier injection in upper lasing subbands extremely difficult. Previous work in the literature is compared and contrasted with a new scheme that may lead to high temperature lasing by engineering the nonequilibrium population inversion with a combination of band structure and many body effects mediated by a k-space filter.

Keywords—band structure engineering, coupled valence bands, intersubband laser, intersubband transitions, lasing without inversion, terahertz radiation.

1. Introduction

Intersubband optics and the terahertz (THz) range of the electromagnetic spectrum are the current frontiers in semiconductor science from both fundamental and applications points of view. To date the exploitation of THz waves has been largely underdeveloped due to the lack of a compact coherent source; providing high output power preferably with continuous wave (cw) operation. This is despite the huge potential THz technology has in a varied list of applications; detecting tumours and skin cancers, pharmaceutical applications, detecting and discriminating different explosive threat materials, environmental sensing and gas monitoring, industrial process control, as well as applications in astronomy, semiconductor imaging, security and medical imaging and telecommunication applications.

Due to the increasing demand for bandwidth it is expected that THz communication systems will be developed in a few years time. American and Japanese companies have already taken the first steps in this direction. Short-range indoor communication systems which work at carrier frequencies of a few hundred gigahertz will represent the wireless local area network (LAN) systems of the future. Such systems could also solve the current “last mile telecommunication” problem.

As a matter of fact, the demand for bandwidth in wireless short-range communication systems has doubled every 18 months over the last 25 years and there is no reason to expect that this trend would come to an end [1].

There are several different meanings for bandwidth and here it means the rate at which information can be transmitted over a given medium (telephone cables, cable TV, microwave relays, fiber optics, satellite links, etc.) In general the permissible bandwidth is about 0.1% to 1% of the carrier frequency. This means that the available bandwidth grows with the carrier frequency. In other words, the higher the frequency, the greater the volume of information that can be transmitted.

The current short-range communications systems Bluetooth and wireless LANs will not be able to deliver the bandwidth needed in 10 years because they operate with carrier frequencies of only a few gigahertz and the same happens for the currently emerging ultra wide band technology. Those systems are limited to rates below 1 Gbit/s. Future systems delivering data rates above 10 Gbit/s will have to work at several ten or preferably a few hundred GHz and this will require higher carrier frequencies. The use of bandwidth is regulated in the USA up to frequencies of 300 GHz with a window from 275 to 300 GHz which is reserved for communications [2]. In Europe the allocation is the same but ends at 275 GHz. In other words, there is free communication bandwidth available above 275 GHz. Systems operating at those frequencies are already in the THz range. Thus, compact and inexpensive devices emitting in this range are required. The quantum cascade laser (QCL) is a potential candidate.

The first mid infra red (MIR) QCL was demonstrated in 1994 by Faist *et al.* at Bell Labs [3] and the extension of the QCL design to emission frequencies below the Reststrahlen band was achieved in 2002 when Köhler *et al.* [4] demonstrated the first THz QCL, at 4.4 THz. Since then they have achieved a maximum operating temperature of 186 K [5].

Even though intersubband emitters are the best candidate as compact high power sources for THz radiation, a current review of the field seems to indicate that conventional designs for QCLs emitting at a given energy $\hbar\omega$ appear to have an maximum operating temperature $T_{\max} \approx \hbar\omega/k_B$ [6]. Thus attainment of room temperature operation seems unlikely without external cooling. Dramatic progress can only be achieved with deep understanding and unconventional manipulation of microscopic mechanisms.

Figure 1 illustrates some of the microscopic difficulties that prevent current THz QCLs to operate at room temperature at high output power. The most critical issue currently

limiting THz QCLs as widely exploitable imaging sources is achieving population inversion at high temperatures depicted in Fig. 1(a)–(d).

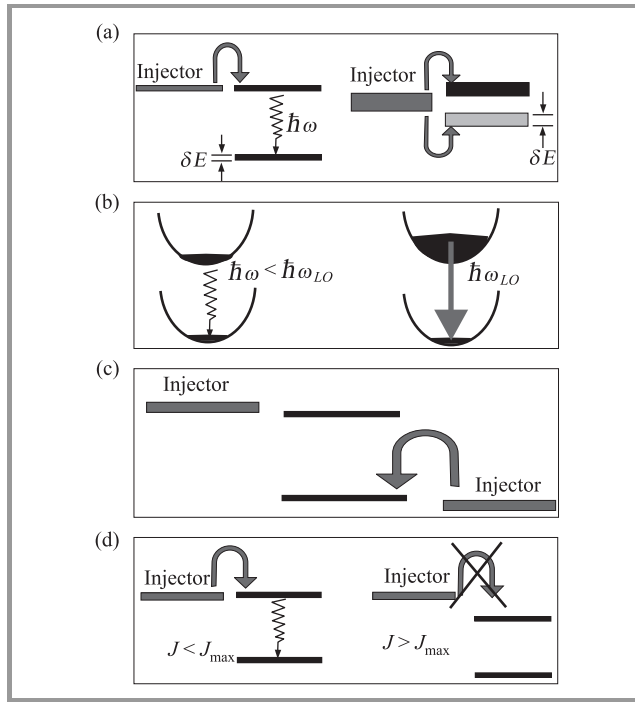


Fig. 1. Cartoon of the most critical issue currently limiting THz QCLs as widely exploitable imaging sources is achieving population inversion at high temperatures.

In Fig. 1(a), in contrast to the mid-IR case (left diagram), for THz transitions (right diagram) the energy difference between the lasing levels is not much larger than the level broadening. The electrons can tunnel from the injector to the upper or lower levels reducing population inversion. In Fig. 1(b), in the right-hand diagram the temperature is high. The upper subband is occupied up to electronic states with sufficient energy to emit longitudinal-optic (LO) phonons, and thus decay nonradiatively to the lower lasing level, reducing the population inversion. In Fig. 1(c), thermal backfilling of the lower radiative state by carriers from the injector states reduces population inversion at high temperatures. In Fig. 1(d), at high temperatures beyond the peak bias level when the maximum achievable current $J_{\max}$ is reached, the injector and upper levels get misaligned and the device ceases to operate.

2. Lasing Without Population Inversion

Lasing without inversion (LWI) is a new solution for producing laser light and it has been achieved in gas and semiconductor lasers, opening new possibilities for cleverly sidestepping traditional difficulties of producing radiation in both extremes of the spectrum: ultraviolet and x-ray lasing on the high energy side and far infrared (TERA-MIR) in the low energy range. In a gas of atoms, laser light

buildup begins when a single photon, emitted by an atom in a high-energy (excited) state, stimulates other excited atoms to emit photons with identical attributes. Ordinary lasers normally require the energy-intensive process of “population inversion”, in which a majority of the atoms must be excited into a high-energy state. Promoting atoms into excited states prepares them for participating in the laser process, but it also serves to prevent them from soaking up the light and thereby sabotaging the laser process.

However, maintaining a population inversion in ultraviolet and x-ray lasers is extremely difficult because the high-lying excited states necessary to produce such light are so short-lived. To the best of the author’s knowledge, the first demonstration of lasing without inversion appeared in atomic optics and was based on quantum interference in Rb atoms [7]. Previous experiments had produced nanosecond bursts of light without population inversion, but the paper [7] was the first to report a sustained laser beam through LWI. In these experiments, an external laser beam essentially creates two pathways for the atoms to get from the ground state ($|a\rangle$) to the excited state ($|b\rangle$). In the rubidium experiment, for example, the probability of getting atoms from ($|a\rangle$) to ($|b\rangle$) becomes the overlap of the likelihood of getting from state ($|a\rangle$) to state ($|b\rangle$) directly and going from state a to an even higher excited state ($|c\rangle$) then decaying to state ($|b\rangle$). Under the proper conditions, the overlapping likelihoods can interfere so as to cancel each other out, preventing absorption. In other words, effective LWI in atomic systems has been based on quantum interference. Future goals are to achieve LWI in inexpensive diode lasers (like those in CD players) and to produce x-ray and UV light through LWI. Here we analyse a very different scenario in intersubband emitters [8]–[10]. This may turn out crucial in the search for room temperature THz QCLs. The main reason for this is that dephasing and scattering phenomena can lead to level broadenings of the same order of magnitude of the lasing transitions, making population inversion by carrier injection in upper lasing subbands extremely difficult, as depicted in the cartoons of Fig. 1.

The first experimental realization of intersubband lasers without inversion exploited the nonparabolicity of the conduction subbands and local population inversion near $k = 0$ even though the lowest subband may have larger global occupation [10].

Later on, valence-band-based designs have been proposed [11]–[13]. Intervalence band emitters based on Si-Ge structures have been investigated. However, lasing on quantum cascade structures has never been demonstrated. Only electroluminescence has been measured so far [14]–[16]. A complete set parameters required for predictive calculations of optical properties of Si-Ge devices is still not known, although progress in this direction has been recently achieved [17]. On the other hand, the material parameters for the III-V system investigated here are very well known. This review complements the results given recently in [18], [19] comparing and contrasting the results for two different well widths.

The following features are unique of the approach introduced in [18] and were not found in previous studies in the literature:

- The strongly k -dependent transverse-electric (TE) transition is used to create a k -space filtering effect that enhances the gain due to local population inversion that arises due to the strongly nonparabolicity of the valence bands.
- The nonequilibrium Keldysh Green's function method is used allowing the consideration of dephasing effects.
- Detrimental cross-absorption due to multiple transitions is taken into account.

Another useful feature of the approach is that simple surface emitting designs can be constructed due to the TE polarization of the emitted field.

3. Numerical Results and Discussion

Necessary conditions to obtain lasing without inversion exploiting nonparabolicity are that the upper conduction subband should either have smaller effective mass or cooler electrons. For unstrained GaAs-AlGaAs wells, even though the conduction bands can be characterized by effective masses (parabolic) "nonparabolicity" appears with different effective masses per subband. And usually the lowest band has a lighter effective mass. The valence bands in contrast have typically lower averaged effective masses in the upper lasing subband which is an advantage. However, as the nonparabolicity can be in some cases so strong that an effective mass does not make sense and the full dispersions must be used, which can complicate the numerical calculations enormously, specially if many particle and dephasing effects are taken into account [18], [19].

The model system investigated in this paper is globally out of equilibrium but the holes are assumed to be thermalized within each subband with occupation functions characterized by different temperatures. The temperatures of the electrons (or holes) in different subbands can be different. This fact has as determined experimentally for conduction band designs by means of microprobe photoluminescence experiments [20]. It is a design challenge to cool the upper subband electrons for more efficient lasing. Furthermore the number of carriers in each subband in an intersubband system depends on the injection scheme and can be chosen independently of the temperature in simulations, see, e.g., [18].

The results shown next are all for non-global inversion conditions and GaAs/Al_{0.3}Ga_{0.7}As quantum wells. Only the top two valence subbands 1 and 2 are occupied and in all cases the lowest hole band has equal or more total carrier density than the upper (lasing) hole subband, i.e., $n_2 \leq n_1$. Detrimental cross absorption due to higher (empty) hole subbands is taken into account. Before proceeding further the following point should be highlighted: very good

agreement with experiments has been obtained by considering electron-phonon and electron-impurity scattering as the main origin of dephasing in mid infrared systems [21]–[23]. Interface roughness scattering also plays a role even in THz systems, however, this effect depends on sample quality [24], [25] and the inclusion of sample quality considerations would be beyond the scope of this paper.

The nature of the main non-radiative channel between subbands, which is crucial to the device performance, remains a matter of controversy for THz systems. The typical energy spacing for one active region period is lower than the optical phonon energy (36 meV in GaAs). Thus, optical phonon emission may not always be the dominant non-radiative mechanism unless the electrons are very hot. Moreover, many-body effects (electron-electron scattering resonances) have been observed in recent experiments for THz QCLs operating in the quantum Hall regime [26]. Thus in the theoretical limit analyzed in this paper, the dephasing is due only to electron-electron mechanisms [27]. Consideration of other scattering channels will be the subject of future research.

The optical properties in which we are interested are calculated in our approach through the optical susceptibility, χ , which can be directly obtained from the Green's function G . The absorption $\alpha(\omega)$ and gain spectra $g(\omega) = -\alpha(\omega)$ are calculated from the imaginary part of the optical susceptibility $\chi(\omega)$ (see, [18]):

$$\begin{aligned}\alpha(\omega) &= \frac{4\pi\omega}{cn_b} \Im\{\chi(\omega)\}, \\ \chi(\omega) &= 2 \sum_{\mu \neq \nu, \vec{k}} \wp_{\mu\nu}(k) \chi_{\nu,\mu}(k, \omega).\end{aligned}\quad (1)$$

Here n_b denotes the background refractive index, c is the speed of light, $\wp_{\nu\mu}(k) = ed_{\nu\mu}(k)$ is the transition dipole moment between the subbands ν and μ , which are labelled $\mu = 1, 2, \dots$ from the top valence band. Thus in the discussion that follows, if holes are injected in the subband 2 and make a transition to subband 1 creating a photon, this actually means that an electron made a transition from valence subband 1 to subband 2 and will be called a (2,1) transition.

The nonequilibrium steady-state susceptibility function $\chi_{\nu\mu}(k, \omega)$ is evaluated through the carriers Keldysh Green's function G , whose time evolution is described by a Dyson equation. The resulting integro-differential equation for $\chi_{\nu\mu}(k, \omega)$ is solved numerically in this paper including many body effects at the Hartree-Fock level, complex non-parabolic band structure, correlation and dephasing mechanisms.

The numerical scheme used here can be summarized as follows: the first step is the solution of the $8 \times 8 \mathbf{k} \cdot \mathbf{p}$ Hamiltonian [27]. The Green's functions and self-energies are expanded using eigenstates and eigenvalues of this Hamiltonian. Next, by assuming thermalized holes, the full nonequilibrium Green's function (NEGF) scheme is simplified and reduces to the self-consistent evaluation of chemical potentials and self-energy matrix elements which lead

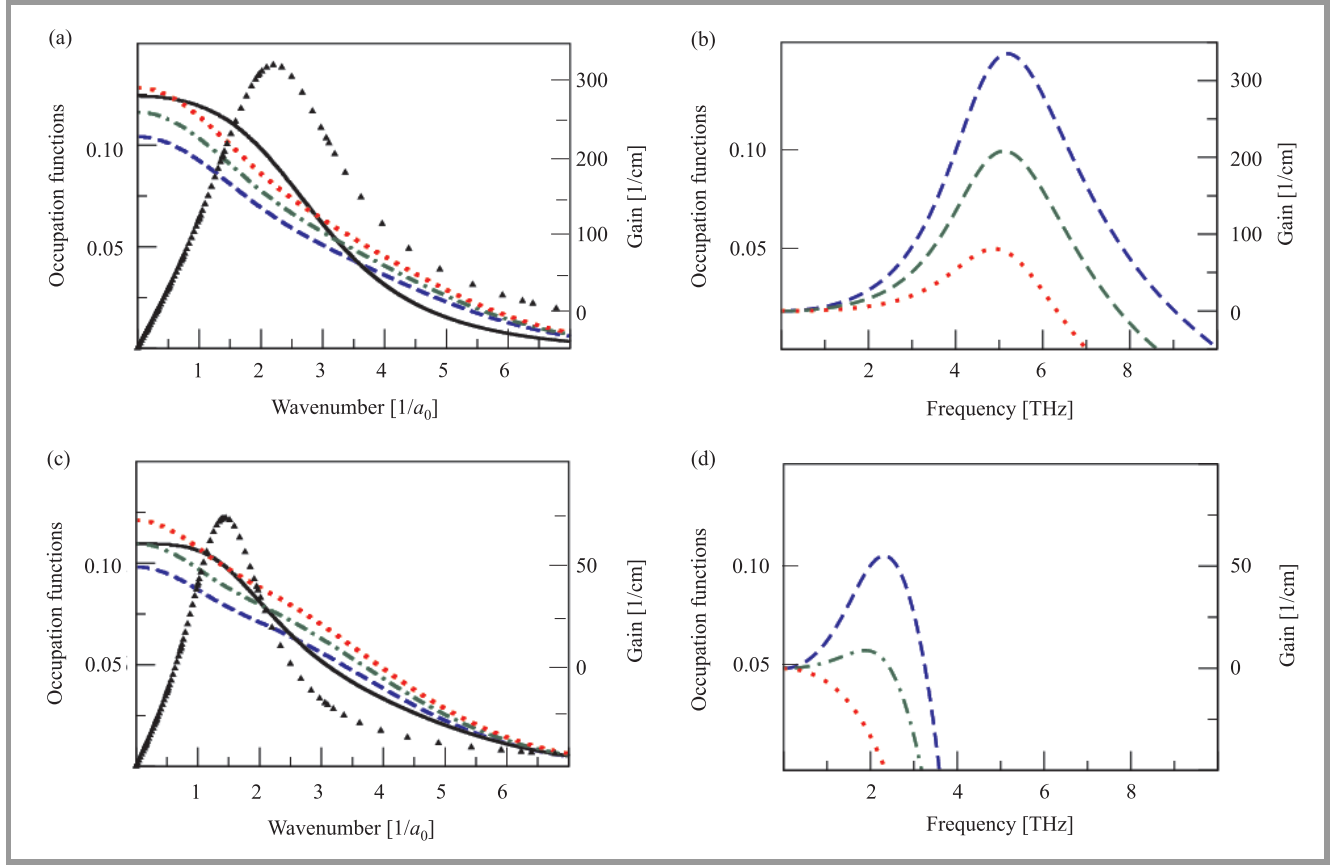


Fig. 2. Comparison of occupation functions and transition dipole moments (symbols) in (a), (c) against gain spectra in (b), (d). The two quantum wells considered have widths of 5 nm (a), (b) and 10 nm (c), (d). The wavenumber is presented in units of $1/a_0$, where a_0 is the 2D Bohr radius.

to subband energy renormalizations, dephasing constants and occupation functions. Only carrier-carrier scattering is considered here and details of the corresponding self-energy are given in [27].

Finally, absorption and gain are given by the solution of the integro-differential equation obtained from the carriers Green's function by numerical matrix inversion. The susceptibility function can be written as

$$\chi_{\nu\mu}(k, \omega) = \sum_{\mathbf{k}' \neq \mathbf{k}} (M_{\mathbf{k}\mathbf{k}'}^{\nu\mu})^{-1} \wp_{\nu\mu}(\mathbf{k}') \delta n_{\nu\mu\mathbf{k}'}, \quad (2)$$

where $(M_{\mathbf{k}'\mathbf{k}}^{\nu\mu})^{-1}$ is the inverse of

$$M_{\mathbf{k}\mathbf{k}'}^{\nu\mu} = (\hbar\omega - e_{\nu\mu}(\mathbf{k}) + i\Gamma_{\nu\mu}) \delta_{\mathbf{k},\mathbf{k}'} + (1 - \delta_{\mathbf{k},\mathbf{k}'}) \delta n_{\nu\mu\mathbf{k}} \tilde{V}_{\mathbf{k}-\mathbf{k}'}^{\nu\mu}. \quad (3)$$

Here $\delta n_{\nu\mu\mathbf{k}} = n_{\nu}(\mathbf{k}) - n_{\mu}(\mathbf{k})$ denotes the nonequilibrium population difference between subbands ν and μ . Further details of the renormalized energies $e_{\nu\mu}$, electron-electron scattering broadening $\Gamma_{\nu\mu}$ and the Coulomb matrix elements $\tilde{V}_{\mathbf{k}-\mathbf{k}'}^{\nu\mu}$ are given in [21], [27]. The role of the dipole function $\wp_{\nu\mu}(\mathbf{k}')$ as a $\mathbf{k}$ -space filter is then clear in Eq. (2). The overlap of a strongly peaked dipole function exactly

where local population inversion $\delta n_{\nu\mu\mathbf{k}'}$ takes place enhances the gain without inversion effect as illustrated by Fig. 2. More details of the numerical method are given in [18].

Figure 2 compares and contrasts the occupation functions and gain spectra for two different quantum wells. An important remark should be made before analyzing the numerical results. The gain spectra are inversely proportional to the period length. The period used here is the quantum well width, $L_c = 5$ nm or 10 nm consistently with the model. In actual QCLs, the period is extended to include barriers, injector, and collectors, and it can easily be at least 5 as large. Since the number of photons emitted per period remains constant this means that the gain spectra calculated here are very high in comparison to what should be expected in actual QCL structures. Note, however, that the qualitative analysis that follows is fully consistent with the single quantum well model.

The occupation functions are given on the left and corresponding gain spectra on the right panels for 5 nm (top) and 10 nm (bottom) GaAs-Al_{0.3}Ga_{0.7}As quantum wells. Only the two top subbands are occupied and in both subbands the electrons are thermalized at $T = 300$ K. In both panels in the left the triangle symbols are the corresponding transition dipole moments for the (1,2) transition respon-

sible for the gain. The dipoles have been scaled in the y-axis to fit in the plot. The solid curve is the occupation of the upper subband, fixed in all gain calculations at $n_2 = 4 \cdot 10^{11}$ carriers/cm². From bottom to top, the dashed, dot-dashed and dotted curves are for increasing lower hole subband total occupation density $n_1 = 4, 4.5$ and $5 \cdot 10^{11}$ carriers/cm².

The corresponding gain curves are easily identified on the right and correspond to the densities in the second subband. In other words, from top to bottom, the dashed, dot-dashed and dotted curves are for increasing lower hole subband total occupation density $n_1 = 4, 4.5$ and $5 \cdot 10^{11}$ carriers/cm². The gain decreases as the lower occupation increases. The transition dipole moment, which has a strong filtering effect in k-space is scaled to fit in the plots. Following intuition, the gain decreases as the global occupation of the lower band increases. The 5 nm quantum well leads to better local inversion assuming that the same 2D density can be achieved in both samples and leads to more robust gain against increases in the lower subband carrier density.

However, in an actual structure it may be more difficult to inject the same global density of electrons for 5 nm then for 10 nm. More realistic calculations will be performed with a new generation of our NEGFs simulator, which will be capable of describing valence band cascaded structures and will be the subject of future publications.

In summary, this manuscript reviewed a possible solution for intersubband emitters that can lead to revolutionary devices for the completely open field of THz telecommunications and many other applications.

References

- [1] S. Cherry, "Edholm's law of bandwidth", *IEEE Spectr.*, vol. 41, pp. 58–59, 2004.
- [2] M. Koch, "IN-door THz communications: a vision for 2020", in *Terahertz Frequency Detection and Identification of Materials and Objects*, R. E. Miles et al., Eds. Nato Security Through Science Series. Berlin: Springer, 2007, pp. 325–338.
- [3] J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A. L. Hutchinson, and A. Y. Cho, "Quantum cascade laser", *Science*, vol. 264, no. 5158, pp. 553–556, 1994.
- [4] R. Köhler, A. Tredicucci, F. Beltran, H. E. Beere, E. H. Linfield, A. Davies, D. A. Ritchie, R. Iotti, and F. Rossi, "Terahertz semiconductor-heterostructure laser", *Nature*, vol. 417, no. 6885, pp. 156–159, 2002.
- [5] S. Kumar, Q. Hu, and J. Reno, "186 K operation of terahertz quantum-cascade lasers based on a diagonal design", *Appl. Phys. Lett.*, vol. 94, no. 13, pp. 131105-1–131105-3, 2009.
- [6] B. S. Williams, "Terahertz quantum-cascade lasers", *Nat. Phot.*, vol. 1, no. 9, pp. 517–525, 2007.
- [7] A. S. Zibrov, M. D. Lukin, D. E. Nikonov, L. Hollberg, M. O. Scully, V. L. Velichansky, and H. G. Robinson, "Experimental demonstration of laser oscillation without population-inversion via quantum interference and in RB", *Phys. Rev. Lett.*, vol. 75, no. 8, pp. 1499–1502, 1995.
- [8] A. Wacker, "Coexistence of gain and absorption", *Nat. Phys.*, vol. 3, no. 5, pp. 298–299, 2007.
- [9] R. Terazzi, T. Gresch, M. Giovanni, N. Hoyler, F. Faist, and N. Sekine, "Bloch gain in quantum cascade laser", *Nat. Phys.*, vol. 3, no. 5, pp. 329–333, 2007.
- [10] J. Faist, F. Capasso, C. Sirtori, D. L. Sivco, A. L. Hutchinson, M. S. Hybertsen, and A. Y. Cho, "Quantum cascade lasers without intersubband population inversion", *Phys. Rev. Lett.*, vol. 76, no. 3, pp. 411–415, 1996.
- [11] G. Sun, A. Liu, and J. B. Khurgin, "Valence intersubband lasers with inverted light-hole effective mass", *Appl. Phys. Lett.*, vol. 72, no. 12, pp. 1481–1483, 1998.
- [12] L. Friedman, G. Sun, and A. Soref, "SiGe/Si THz laser based on transitions between inverted mass light-hole and heavy-hole subbands", *Appl. Phys. Lett.*, vol. 78, no. 4, pp. 401–403, 2001.
- [13] R. A. Soref and G. Sun, "Terahertz gain in a SiGe/Si quantum staircase utilizing the heavy-hole inverted effective mass", *Appl. Phys. Lett.*, vol. 79, no. 22, pp. 3639–3641, 2001.
- [14] G. Dehlinger, L. Diehl, U. Gennser, H. Sigg, J. Faist, K. Ensslin, D. Grützmacher, and E. Müller, "Intersubband electroluminescence from silicon-based quantum cascade structures", *Science*, vol. 290, no. 5500, p. 2277, 2000.
- [15] L. Diehl, S. Mentese, E. Müller, D. Grützmacher, H. Sigg, U. Gennser, I. Sagnes, Y. Campidelli, O. Kermarrec, D. Bensahel, and J. Faist, "Electroluminescence from strain-compensated Si_{0.2}Ge_{0.8}/Si quantum-cascade structures based on a bound-to-continuum transition", *Appl. Phys. Lett.*, vol. 81, no. 25, pp. 4700–4702, 2002.
- [16] R. Bates, S. A. Lynch, D. Paul, Z. Ikonik, R. W. Kelsall, P. Harrison, S. L. Liew, D. J. Norris, A. G. Cullis, W. R. Tribe, and D. D. Arnone, "Interwell intersubband electroluminescence from Si/SiGe quantum cascade emitters", *Appl. Phys. Lett.*, vol. 83, no. 20, pp. 4092–4094, 2003.
- [17] D. J. Paul, "8-band k · p modeling of the quantum confined Stark effect in Ge quantum wells on Si substrates", *Phys. Rev. B*, vol. 77, no. 15, pp. 155323-1–155323-7, 2008.
- [18] M. F. Pereira Jr., "Intervalence transverse-electric mode terahertz lasing without population inversion", *Phys. Rev. B*, vol. 78, no. 24, pp. 245305-1–245305-5, 2008.
- [19] M. F. Pereira, "Valence intersubband gain without population inversion", *Centr. Eur. J. Phys.*, DOI: 10.2478/s11534-009-0061-5 (in press) [Online]. Available: <http://www.springerlink.com/content/j9nm0k6179g142n6/>
- [20] M. S. Vitiello, G. Scamarcio, V. Spagnolo, B. S. Williams, S. Kumar, Q. Hu, and J. L. Reno, "Measurement of subband electronic temperatures and population inversion in THz quantum-cascade lasers", *Appl. Phys. Lett.*, vol. 86, no. 111115–111117, 2005.
- [21] M. F. Pereira Jr., S.-C. Lee, and A. Wacker, "Controlling many-body effects in the midinfrared gain and terahertz absorption of quantum cascade laser structures", *Phys. Rev. B*, vol. 69, p. 205310, 2004.
- [22] R. Nelander, A. Wacker, M. F. Pereira Jr., D. G. Revin, M. R. Soulby, L. R. Wilson, J. W. Cockburn, A. B. Krysa, J. S. Roberts, and R. J. Airey, "Fingerprints of spatial charge transfer in quantum cascade lasers", *J. Appl. Phys.*, vol. 102, no. 11, pp. 113104-1–113104-5, 2007.
- [23] M. F. Pereira Jr., R. Nelander, A. Wacker, D. G. Revin, M. R. Soulby, L. R. Wilson, J. W. Cockburn, A. B. Krysa, J. S. Roberts, and R. J. Airey, "Characterization of intersubband devices combining a nonequilibrium many body theory with transmission spectroscopy experiments", *J. Mater. Sci. Mater. Electron.*, vol. 18, no. 7, pp. 689–694, 2007.
- [24] A. Wacker, "Coherence and spatial resolution of transport in quantum cascade lasers", *Phys. Stat. Sol. C*, vol. 5, no. 1, p. 215, 2008.
- [25] T. Kubis, "Quantum theory of transport and optical gain in quantum cascade lasers", *Phys. Stat. Sol. C*, vol. 5, no. 1, pp. 232–235, 2008.
- [26] G. Scalari, S. Blaser, J. Faist, H. Beere, E. Linfield, D. Ritchie, and G. Davies, "Terahertz emission from quantum cascade lasers in the quantum Hall regime: evidence for many body resonances and localization effects", *Phys. Rev. Lett.*, vol. 93, no. 23, pp. 237403-1–237403-4, 2004.
- [27] M. F. Pereira Jr. and H. Wenzel, "Interplay of Coulomb and non-parabolicity effects in the intersubband absorption of electrons and holes in quantum wells", *Phys. Rev. B*, vol. 70, no. 20, pp. 205331-1–205331-8, 2004.

Note: this text is reprinted from JTIT 4/2009



Mauro F. Pereira was born in Rio de Janeiro, Brazil. He received the B.Sc. in physics at PUC/RJ (1983) and the M.Sc. in physics (1985). He completed the Ph.D. in optical sciences at the Optical Sciences Center in Tucson/AZ in 1992 and received an equivalent Dr. Sci. degree in physics from UFRJ in 1993. He was a Re-

search Associate at PUC/RJ, CBPF, Uni-Rostock, and the TU-Berlin, an invited lecturer in Bremen, an Associate Professor at UFBA and a Senior Researcher at Tyndall National

Institute before joining the Materials and Electrical Engineering Research Institute of Sheffield Hallam University as a Professor. His expertise includes nonlinear and quantum optics, exciton and polariton effects, band structure engineering, many-body effects, semiconductor lasers including quantum cascade structures, nonequilibrium Green's functions, quantum transport and numerical methods. The results are aimed at fundamental understanding and as input for the design and simulation of novel optical and electrooptical devices.

e-mail: M.Pereira@shu.ac.uk

Materials and Engineering Research Institute
Sheffield Hallam University
S1 1WB, Sheffield, United Kingdom

Impact of Signaling System Performance on QoE in Next Generation Networks

Jordi Mongay Batalla, Jarosław Śliwiński, Halina Tarasiuk, and Wojciech Burakowski

Abstract—The first experience of quality by multimedia applications' users takes place during the setup phase of a new connection. If the setup phase is not accepted or "slowly accepted", the confidence of the user decreases. The user becomes more sensitive when he/she pays the connections with assured quality of service (QoS). In this case, the process of call request should be also accomplished with QoS guarantees. This paper presents the signaling sub-system implemented within the EuQoS system. The EuQoS signaling process follows main assumptions of next generation networks (NGN) architecture and performs tasks related with codec agreement between multimedia end users, admission control and resource reservation functions. In this paper, we present analytical, simulation and experimental results showing the impact of signaling system performance on quality of experience (QoE) for the potential users of multi-layer EuQoS system. In particular, the presented approach aims at ensuring user QoE of the connection setup phase by ensuring QoS for transferring signaling messages by the network.

Keywords—*call setup delay, class of service, heterogeneous networks, next generation networks, quality of experience, quality of service, signaling system.*

1. Introduction

This paper deals with an impact of the signaling system on user's quality of experience (QoE) in next generation networks (NGN). In particular we focus on signaling system and its procedures implemented within the EuQoS¹ project [1]. The aim of the EuQoS system [2], [3] is to guarantee end-to-end quality of service (QoS) in heterogeneous multi-domain networks. For this purpose the EuQoS system combines a complete architecture and a full framework [3] to provide absolute end-to-end QoS guarantees for a number of end-to-end classes of service (E2E CoS) [4], [5]. The architecture considers heterogeneous network scenario, addressing the multi-core IP network, as well as, the current access network technologies, such as WiFi (wireless fidelity), xDSL (digital subscriber line), LAN/Ethernet (local area network/Ethernet) and UMTS (universal mobile communication systems). Moreover, the generic architecture is open to add new network technologies. The applications, for which finally QoS

is provided are, among others, voice over IP, video on demand, data transfer, and interactive game. All of them require call setup phase before data transfer.

The call setup phase is accomplished by the EuQoS signaling system, which has been designed in compliance with the ITU-T recommendations about signaling requirements for IP QoS networks [6] and requirements for resource and admission control functions, as defined in scope of NGN activities [7]. The signaling system follows a "push mode" approach of NGN architecture [8], i.e., the system requires that in order to start the setup procedure [9], the applications must send resource reservation requests in an explicit way to the control plane architecture.

The setup procedure starts at sending the calling user's request to the system and finishes when receiving the corresponding response indicating that the new call can be admitted and the called user is ready to initiate the connection. The time between these two events is well known as call setup delay [10]. In [11], ITU-T imposes limitations on expected values of setup delay. Among other requirements, it states that for international connections and under normal load conditions [12], the setup delay for the 95% of the setup procedures should be less or equal to 11 s.

In fact, the first experience of quality by the multimedia application users takes place during the request of a new call. If the request is not accepted or "slowly accepted", the confidence of the user decreases. The notion of QoE is already evolving. In the last years, the ITU-T Standardization Group 12 (SG12) specified the QoE requirements as definable end-to-end parameters, which provide information not only about the network layer (as performed by the QoS parameters) but also about the transport and application layers behavior [13].

We follow this notion of QoE and in our studies the definable end-to-end parameter is the call setup delay as suggested in [14]. The aim of this paper is to show the impact of signaling system performance on setup delay in EuQoS system based on NGN architecture.

In order to guarantee target values of call setup delay, the system should dedicate some resources to the signaling process (setup and release procedures). These resources are both server and network resources. The server resources are intended for the processing of the signaling messages within the signaling servers, which manage new calls. Whereas, the network resources are intended

¹EuQoS – IST 6 FR EU project "End-to-end Quality of Service Support over Heterogeneous Networks".

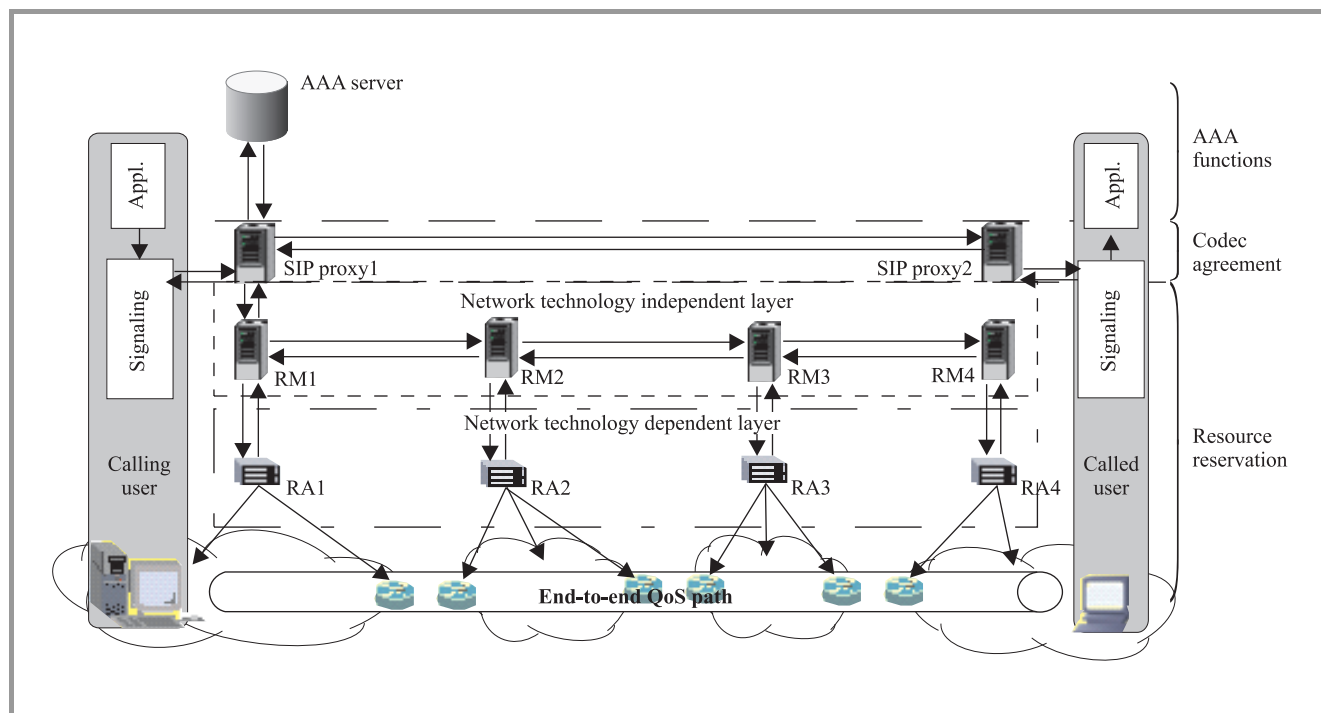


Fig. 1. The EuQoS system.

for transferring the signaling messages by the network between signaling servers. For reserving the necessary server and network resources, we should correctly dimension the signaling system. We propose to use the decomposition approach, which separately considers three phases of the signaling procedure in EuQoS system – all the operations performed:

- at the application layer, it means in session initiation protocol (SIP) proxies and authentication, authorization and accounting (AAA) servers for, e.g., voice over IP application;
- at the technology independent/technology dependent layer (TI/TD layer), for performing the admission control and resource allocation functions;
- at the network layer for transferring the signaling messages by dedicated signaling CoS.

Since the above three phases are distributed in time, i.e., they do not carry out simultaneously, we may enforce delay constraints for each one, so that the sum of the delay constraints of all the phases is less or equal to 11 s for the 95% of setup procedures. For illustration purposes, in our analysis we assume that the target setup delay value introduced by each of considered phases should be not greater than 3.5 s for the 95% of setup procedures.

Let us focus on the first phase of setup procedure, i.e., the operations performed at application layer. In previous studies of the EuQoS project [15], we obtained required performance parameters for the involved SIP proxy devices, which actually are less demanding than commercial ones.

In fact, based on the literature we can observe that the SIP proxies are capable of running a high number of calls per second. For example, CISCO SIP proxies is able to handle about 100 call/s [16] and the direct routing approach allows even higher rates [17]. In the same way, the current AAA servers may manage until thousands of calls per second, e.g., the host intrusion detection system (HIDS) v3.1 of HP [18]. Measurements of mean establishment times of voice over IP calls related with AAA and SIP proxies functions were presented in [19]. Under normal load conditions, the mean establishment time is 0.94 s. Taking into account the state of the art, in our studies we focus on performance evaluation of the TI/TD layer and the transfer of signaling messages by the network (signaling class of service), which has been studied with less attention in the literature. These processes are the center of our studies in Sections 3 and 4, respectively. Previously, we present main features of the EuQoS system in Section 2. Finally, Section 5 summarizes the paper.

2. Overview of EuQoS System

The EuQoS system consists of three main layers, which are application layer, technology independent/technology dependent layer, and network layer as indicated in Fig. 1. The signaling procedure to set up a new connection (e.g., voice over IP) in the network is the following.

When any user sends a new call request to the EuQoS system, this initiates security functions with the AAA server.

These functions check if the user sending a call is authorized to use the EuQoS system. Next, the SIP proxy initiates the codec agreement with the called user. After the codec agreement, the first resource manager (RM) in the way checks whether the end-to-end QoS path to the called user exists and could provide the QoS guarantees required by the associated end-to-end CoS. The RM in the access domain periodically receives from the QoS routing protocol, the information about the QoS paths, i.e., the end-to-end paths with predefined QoS guarantees (in the form of target values for QoS parameters as IPTD, IPDV, IPLR). The QoS routing protocol is an enhanced version of border gateway protocol called EQ-BGP [2], [20], which builds end-to-end QoS paths on multi-domain network.

Afterwards, the system checks if, currently, there are available network resources (bandwidth, buffer capacities) to accept the call in consecutive domains. This is performed by the admission control functions implemented in resource managers and resource allocators (RA) as it will be indicated in the next section. If the process positively concludes, the system (SIP proxy) sends an affirmative response to the calling user allowing the communication.

Figure 1 presents the EuQoS signaling system, which is divided into the servers involved in AAA functions, the servers (SIP proxy) involved into the codec agreement and the servers (RM and RA) involved into the admission control function and resource allocation. The last ones, in turn, are divided into the servers, which are independent of the network technology (RM) and the servers, which functionalities depend on the network technology (RA). The purpose of this specialization is to allow each domain for custom implementation of its own QoS mechanisms without requiring other domains to be aware of their specific details.

Using the ITU-T terminology for NGN [21], [22], the SIP proxy would fulfill the service control function (SCF) at the service stratum; whereas, the resource manager and resource allocator would carry transport control function (TCF) at the transport stratum. The resource manager we would call policy decision functional entity (PD-FE), which is independent of the network technology and the resource allocator we would call the transport resource control functional entity (TRC-FE), which depends on the network technology.

A set of signaling elements (functions, databases, interfaces, etc.), as well as the signaling protocols involved in the different signaling layers were defined and implemented in the system. The diameter base protocol [23] communicates the SIP proxy with the AAA server in the access domain; the SIP protocol communicates end users and proxies, next steps in signaling (NSIS) protocol [24] connects the RMs and, at last, common open policy service (COPS [25]) communicates the RMs with the RAs and these ones with the network devices.

3. Performance Evaluation of TI/TD Layer

In this part, we analyze the impact of call handling scenario and the processing of messages at TI/TD layer on the part of the setup delay related to the TI/TD layer, which we denote as T_{ti-td} .

We focus on results of T_{ti-td} only for simulation studies, because the prototype implementation of RM and RA servers in the project was not optimized to take adequate conclusions. We present details about the call handling scenario at the TI/TD layer, a methodology to calculate T_{ti-td} for single and multi-domain scenario as well as the assumed simulation model. The aim of our studies is to show how many calls the system can handle respecting target values of T_{ti-td} (3.5 s for the 95% of setup procedures). In [15] we presented simulation results of performance evaluation on TI/TD layer. In this paper we enhance these studies showing detailed simulation results of analyzed queuing network.

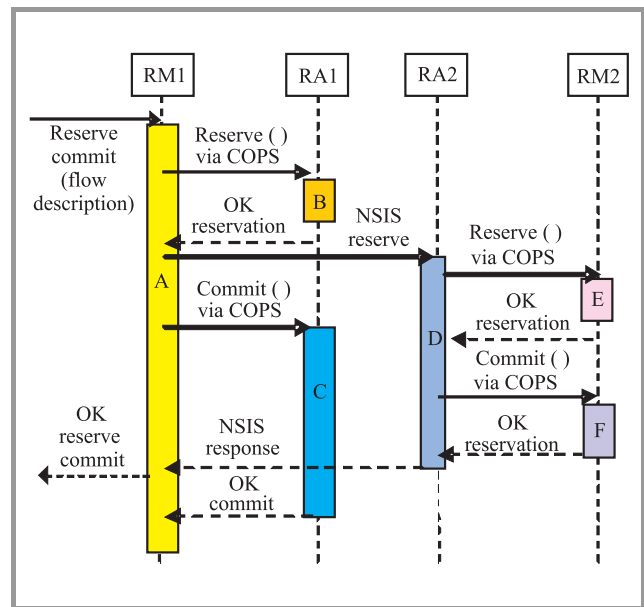


Fig. 2. Call scenario for two domains (domain 1: RA1, RM1; domain 2: RA2, RM2), one direction.

Figure 2 shows the horizontal (between RMs) and vertical (between RM and RA) signaling exchange for the successful setup of a call traversing two domains (domain 1 and domain 2). The call handling scenario is as follows: RM1 receives a QoS request and asks to the connection admission control (CAC) module in RA1 if there are enough resources to handle the new connection. If so, the requested resources are reserved and RA1 sends a confirmation to RM1. Next, RM1 forwards the QoS request to its peer RM2 and in parallel sends a request to RA1 to actually allocate the reserved resources in the associated access network equipment. When RM1 receives the confirmation from both RA1 and RM2, it replies to SIP proxy that the new call

can be admitted. As we can see in Fig. 2 the call handling scenario in RM2 and RA2 (i.e., in the egress domain) is the same as in RM1 and RA1 (ingress domain).

When the path has more than two domains, the resources are reserved also in transit domains. However, only the ingress domain RM checks whether an end-to-end path with feasible QoS exists. For bi-directional calls, the call handling process is performed in parallel during the setup procedure. So, in order to calculate the T_{ti-id} for a successful scenario, it is sufficient to simulate it only in one direction.

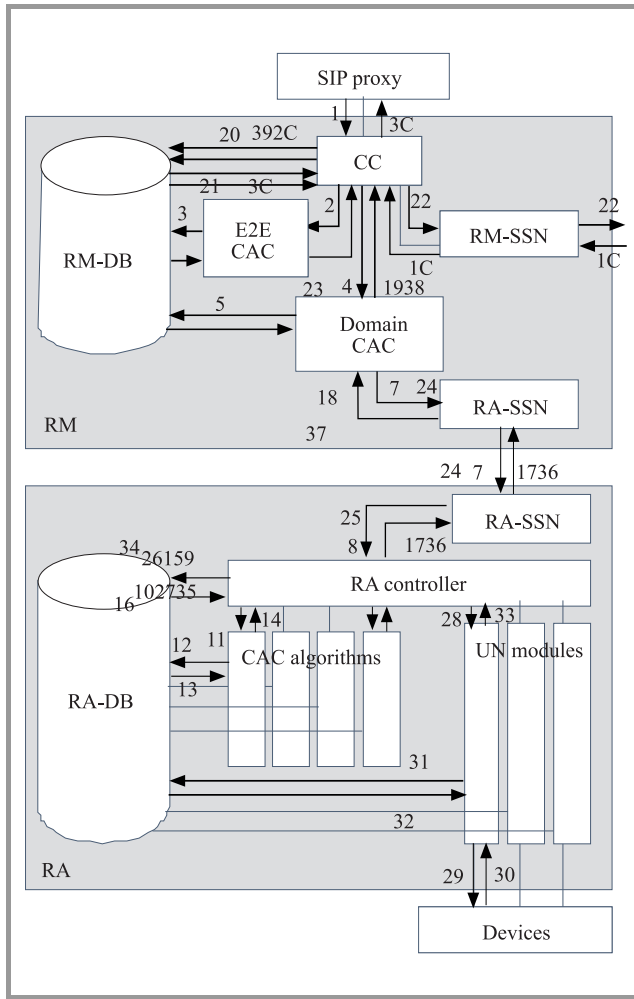


Fig. 3. The RM and RA architecture and call handling scenario for ingress domain calls and confirmations. Explanations: DB – data base, UN – underlining network, 1,2...39 task numbers, C – confirmation task.

The detailed architecture of the RM and RA, together with the tasks identifications (task ID) are presented in Fig. 3. In the RM, the call controller (CC) receives QoS requests and controls all the CAC submodules. The end-to-end CAC (E2E CAC) is in charge of checking whether an end-to-end path with feasible QoS characteristics actually exists (by looking at its EQ-BGP routing information base). Then, the CC asks the domain CAC to check the operator policies and to look for an intra-domain path. The domain CAC

has two submodules, one for the intra-domain part and the other one for the inter-domain link. Finally, the RA controller receives the QoS request via signaling module interconnecting the RM with the RA signaling and service negotiation (RA SSN). The RA controller runs a different CAC algorithm for each CoS. The RA CAC algorithms check the amount of available resources by querying the RA data base (RA-DB) and decide about the acceptance of the new call. If the call is accepted, the RA controller asks the appropriate access network UN module to configure its network devices for handling the new call. Different mechanisms must be configured depending on the type of network device (IP router, LAN/Ethernet switch, WiFi access point, etc.). Note that UN modules are involved in the call handling process only in *access* domains. For *transit* domains, resources are reserved basing only on the decision of the RA CAC algorithm, and there is no dynamic configuration of resources in inter-domain network devices.

We model the RM and RA architecture (Fig. 3) as the queuing network shown in Fig. 4. The simulation model is composed of a chain of servers, each one associated to an infinite-length FIFO (first in first out) queue. We distinguish eight types of servers: RM-CC (1), RM-DB (2), RM-RA link (3), RM-RM link (4), RA-C (5), RA-DB (6), RM-RA link (7), devices (8).

We assume that new calls and call acceptance confirmations independently arrive to the RM according to Poissonian processes with given mean arrival rates. Service processing times are the following deterministic values:

- RM-CC, RA-CC: $t_{RM-CC} = t_{RA-CC} = 100 \mu s$;
- database access in the RM and RA: $t_{RM-DB} = t_{RA-DB} = 1 \text{ ms}$;
- RM-RA link, RA-RM link, RM-RM link: $t_{RM-RA} = t_{RA-RM} = t_{RM-RM} = 1 \text{ ms}$;
- access to the device: $t_{DEV} = 600 \text{ ms}$ (WiFi) and $t_{DEV} = 100 \text{ ms}$ (IP).

The processing times for WiFi and IP devices are based on test bed network equipment (LINKSYS WiFi access point WRT54G, CISCO 1841 router or Linux router kernel 2.6.18 IMQ). For other elements as call controllers or data bases, we tried to infer the expected processing time based on today's high speed servers.

We consider three event types in the simulation model: *new call arrival*, *new confirmation arrival*, and *departure of a task from a server*. Arrivals and confirmations invoke a sequence of tasks, whereas, the departure of a task from one server determines its arrival to another server. Furthermore, we distinguish call events processed at *ingress*, *transit* or *egress* domains. The maximum number of tasks is performed when processing calls at the ingress domain, since the E2E CAC is checked. Fewer tasks are required to

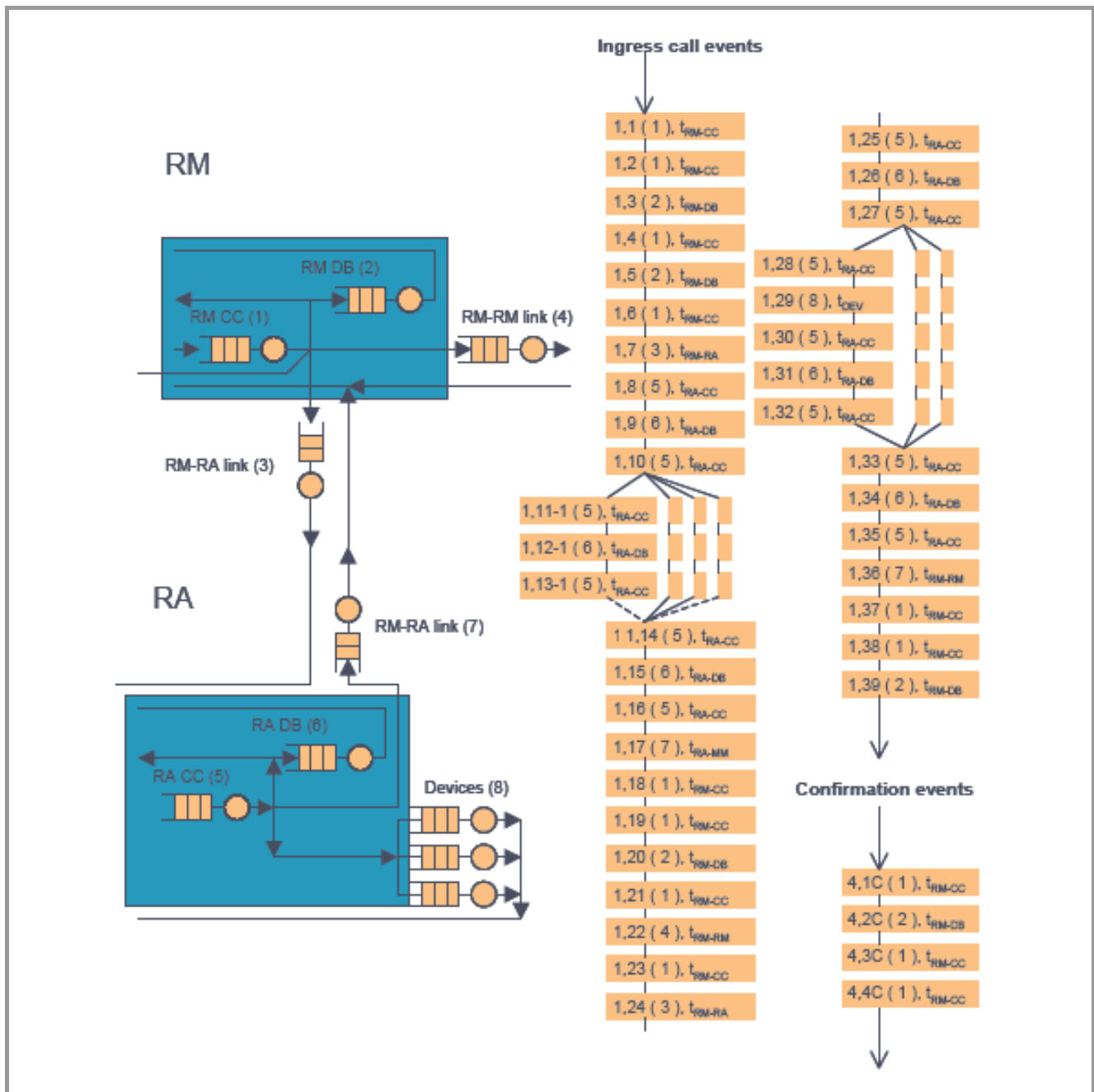


Fig. 4. The RM and RA for the ingress domain – simulation model.

process confirmation arrivals. The mean call arrival rates are $\lambda_{ingress}$, λ_{egress} , $\lambda_{transit}$, and the mean arrival rate of confirmations for ingress and transit domains are $\lambda_{conf-ingress}$ and $\lambda_{conf-transit}$. Due to space limitations, we only present events related to ingress domain.

Note that, in these simulations, we do not model the state machine of NSIS or COPS (see Fig. 2). The detailed message exchange of NSIS is modeled in simulation studies of signaling class of service (see Section 4). For COPS protocol, we model its performances by t_{RM-RA} , t_{RA-RM} times only.

The above simulation model has been coded in a discrete event simulator written in C++. Each event is represented by the quadruple $\langle \text{Sequence ID}, \text{Task ID}, \text{Server ID}, \text{Processing time} \rangle$. The Sequence ID denotes whether the event is a call or a confirmation arrival.

We begin our studies with single domain performance evaluation. In particular, we simulate two scenarios. First, we assume that the single access domain handles both ingress and egress calls and confirmations assuming $\lambda_{ingress} = \lambda_{egress} = \lambda_{conf}$. We simulate WiFi and IP access domains in isolation. For each test we vary the call and confirma-

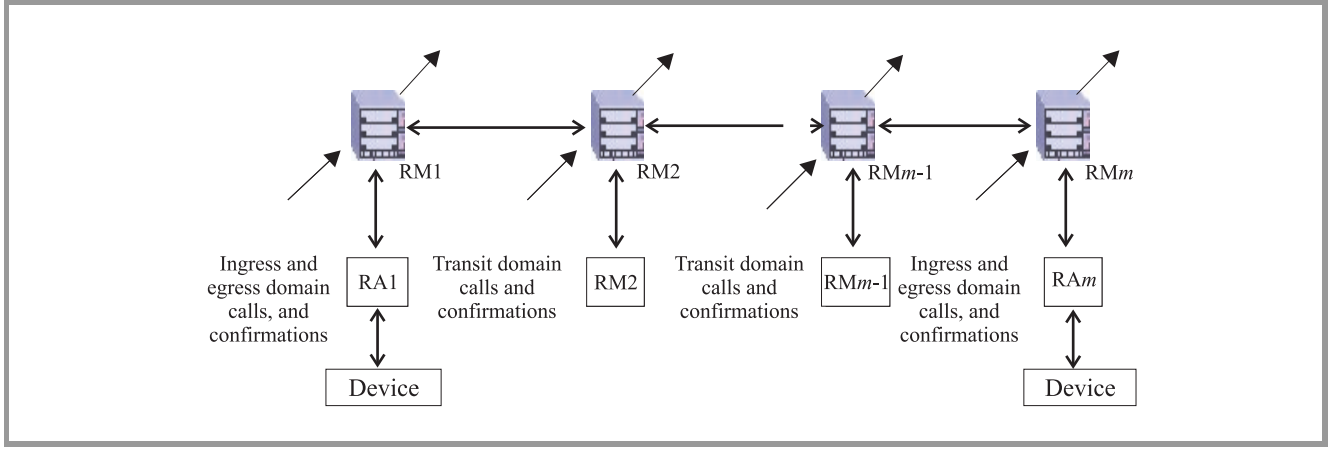


Fig. 5. Multi-domain call handling scenario with m domains, where $m = 2 + n$, 2 access domains and n transit domains.

tion arrival rates up to the 95% resource utilization of the RM-RA elements (in this case the bottleneck element is the device itself).

In the next step, we analyze performances of single transit domain assuming that RM and RA handle only transit calls and confirmations assuming $\lambda_{transit} = \lambda_{conf}$; in this case the bottlenecks are RM-DB and RA-DB, it means the access to the databases.

Finally, based on the results obtained for single ingress/egress and transit domains, we are able to calculate the quantiles of T_{ti-d} for a multi-domain scenario (see Fig. 5).

For this purpose we distinguish among the following sequences of tasks performed by RM and RA in the ingress domain:

- $S1_{ingress}$: resource reservation tasks in $RM1 \leftrightarrow RA1$ performed before forwarding the request to the next RM;
- $S2_{ingress}$: tasks in $RM1 \leftrightarrow RA1 \leftrightarrow DEV1$, i.e., resource allocation tasks;
- $S3_{ingress}$: RM tasks invoked by confirmation arrival from neighboring domain.

Thus, we define the delay T_{ti-d} as follows:

$$T_{ti-d} = t_{ingress}^{S1} + \max \left\{ t_{ingress}^{S2}, (n \cdot t_{transit} + t_{egress} + t_{ingress}^{S3}) \right\}, \quad (1)$$

where $t_{ingress}^{S\#}$ is the delay introduced by sequence $S\#$, t_{egress} is the delay in the egress domain, n is the number of transit domains, $t_{transit}$ is the delay introduced by a transit domain.

Based on the detailed simulation results obtained for single domain scenarios (for ingress/egress and transit) we can derive:

$$t_{ingress}^{S2} < t_{ingress}^{S3} + t_{egress} + n \cdot t_{transit}. \quad (2)$$

By applying (1) and (2) the T_{ti-d} is calculated:

$$T_{ti-d} = t_{ingress}^{S1} + t_{ingress}^{S3} + t_{egress} + n \cdot t_{transit}. \quad (3)$$

In our studies, we calculate results when the access domain are both WiFi or both IP. The T_{ti-d} quantiles for 2, 10 and 20 domains is presented in Figs. 6–8.

In particular, Fig. 6 shows the results for two WiFi and two IP access domains. For both scenarios, the curves show a significant difference between the delays obtained for the two technologies. With reference to the 3.5 s target for the 0.95-quantile of T_{ti-d} , we observe that, in WiFi, $\lambda_{ingress}$, λ_{egress} , and λ_{conf} should not exceed about 0.6 call/s. In IP domain, the limit is about 4.75 call/s. For scenarios with 10 and 20 domains we assume that two domains are access domains while the others are transit domains.

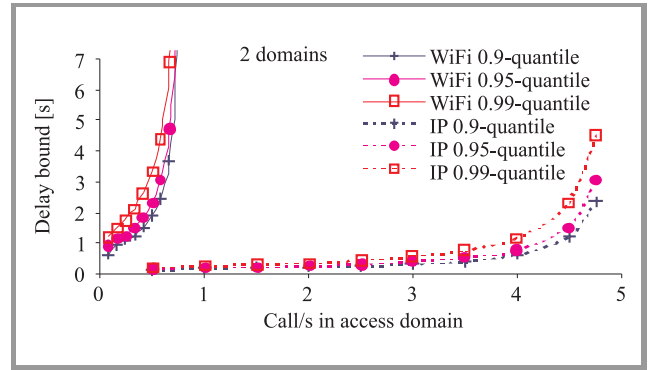


Fig. 6. Quantiles of T_{ti-d} for 2 WiFi or 2 IP domains versus $\lambda_{ingress}$ ($\lambda_{ingress} = \lambda_{egress} = \lambda_{conf}$).

Figures 7 and 8 show the quantile values of T_{ti-d} for WiFi and IP access domains, respectively. The transit call rates are 233 or 300 call/s in each transit domain. The results show that the number of transit domains has smaller impact on the T_{ti-d} than the transit call arrival rate. Moreover,

acceptable call arrival rates are larger for IP access domains than for WiFi.

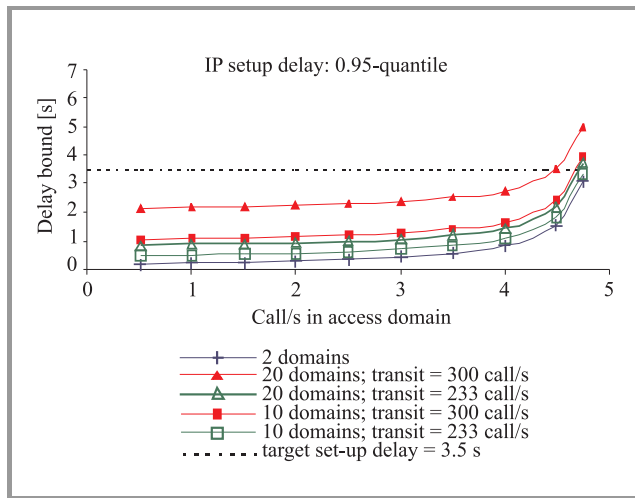


Fig. 7. The 0.95-quantile of T_{ti-td} for 2 IP access domains versus $\lambda_{ingress}$ ($\lambda_{ingress} = \lambda_{egress}$ and $\lambda_{ingress} = \lambda_{conf}$).

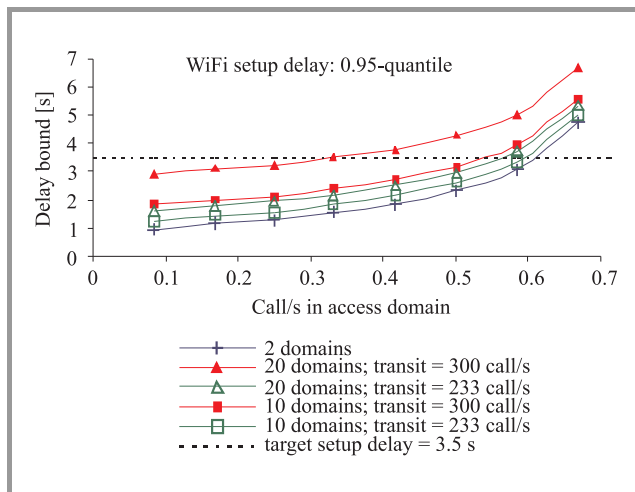


Fig. 8. The 0.95-quantile of T_{ti-td} for 2 WiFi access domains versus $\lambda_{ingress}$ ($\lambda_{ingress} = \lambda_{egress}$ and $\lambda_{ingress} = \lambda_{conf}$).

Figure 9 shows some characteristics of the transit RM and transit RA. In particular, we collected characteristics of server utilization, mean delay in the queues, and mean queue sizes. As we can observe, the main bottlenecks of RM and RA elements for transit domains are RM-DB and RA-DB servers. The characteristics of these two servers are very close to each other.

Concluding, the signaling system at the TI/TD level is able to handle about 300 call/s in transit domains and it correctly scales with the number of transit domains (for values of 10 and 20 transit domains). An important open point is the call arrival process for NGN multi-service networks.

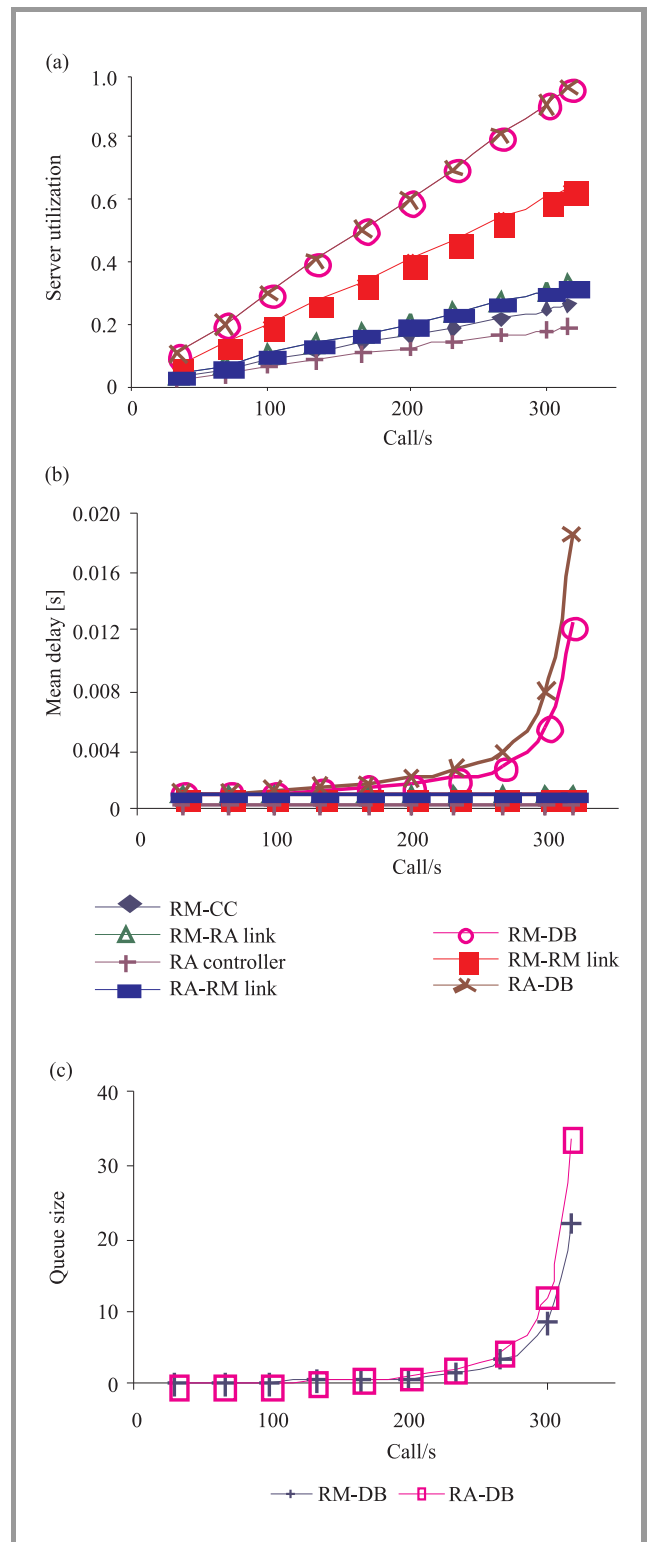


Fig. 9. Transit domain characteristics: (a) utilization of RM and RA servers, (b) mean delay, and (c) mean queue size for RM and RA elements versus transit call arrival rates ($\lambda_{transit}$) $\lambda_{transit} = \lambda_{conf} - \text{transit}$.

In these studies, we commonly considered telephony arrival model only but, in further studies, we are also facing up multi-service models.

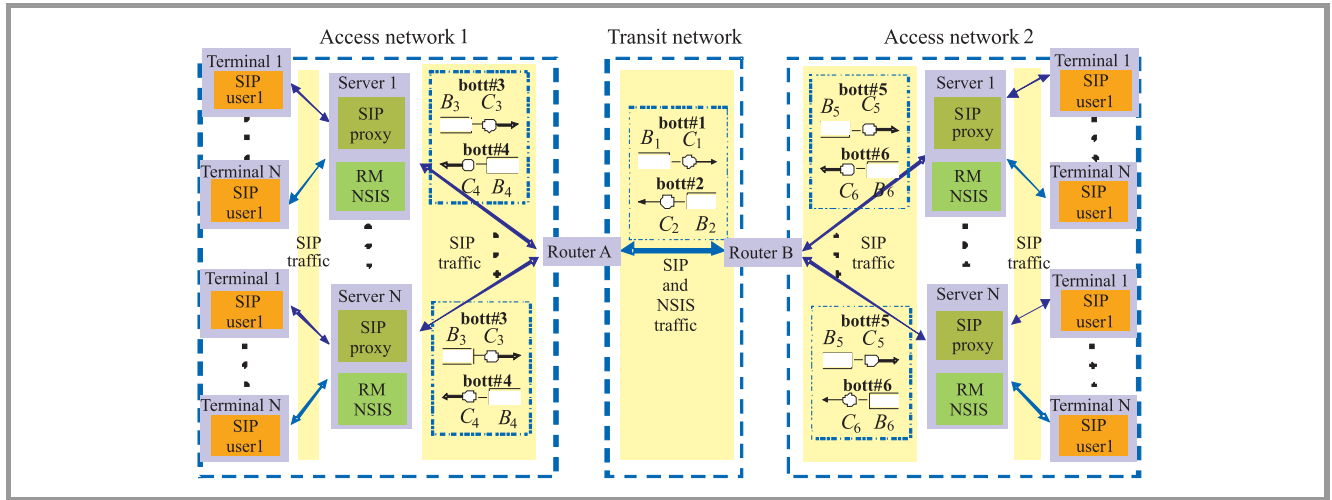


Fig. 10. Model of the EuQoS signaling system.

4. Performance Evaluation of the Signaling Class of Service

At the network layer we propose to implement the signaling class of service. The IETF document [4] defines, among others, the signaling class of service (S-CoS) designated to carry signaling traffic related with setup and release procedures. The objective of the S-CoS is to ensure target values of the part of the call setup delay related to the transfer of the signaling packets. We call this delay as *transfer packet call setup delay* or briefly T_s delay. We expect to obtain target values of T_s delay by transferring the signaling packets with adequate quality of service expressed in target maximum IPTD. In fact, according to [4] we can assume that the S-CoS tolerates the delay variation of the signaling traffic. One of the solutions to guarantee the QoS objectives for the signaling traffic is a correct resource provisioning [26].

For the aim of provisioning the S-CoS, we investigate the bottlenecks within the signaling path. The bottlenecks are the slowest links [27] in whose entrance packets of different setup procedures gather and where we should provision appropriate resources for each of the setup procedures.

To compute the necessary amount of bandwidth within the S-CoS for one setup procedure, we propose to ensure the same maximum IPTD (maxIPTD) in all the bottlenecks along the path. From the knowledge of the sizes of packets submitted to each bottleneck, and more precisely from the length of the longest burst of packets of the setup procedure submitted to each bottleneck we obtain the maxIPTD for 1 isolated setup procedure as (4)

$$\begin{aligned} \max IPTD &= \frac{\text{length of longest packet burst}_{bott\ 1} \times 8}{C_{bott\ 1}} = \dots \\ &= \frac{\text{length of longest packet burst}_{bott\ i} \times 8}{C_{bott\ i}}. \end{aligned} \quad (4)$$

The value of $C_{bott\ i}$ indicates the necessary amount of bandwidth for one setup procedure in the bottleneck i of S-CoS.

The equation, which completes the consistency of the system of linear equations presented in (4), comes from the value of T_s delay. In fact, we defined that the setup procedure should finish in a time equal to 3.5 s as presented in equation:

$$T_s \text{ delay} = \sum_{\text{all bottleneck } s} \frac{\sum_{\text{all packets in bottleneck}} \text{packet length} \times 8}{C_{bott\ i}}. \quad (5)$$

In the EuQoS system we assume that bottlenecks only may appear at the exit of SIP proxies and next inter-domain border routers as presented in the model of Fig. 10. Therefore, we consider only SIP and NSIS traffic, which is the unique signaling traffic in these bottlenecks.

As we may deduce from Eqs. (4) and (5), the necessary resources for signaling traffic ($C_{bott\ i}$) depend on the message sequence. Figure 11 shows the setup message sequence in the EuQoS system for which we implement the S-CoS. The lengths of the packets referred in Fig. 11 do not consider user data protocol (UDP), IP and link layer headers. Based on the above setup message sequence and assumptions and by the way of example, we will present further down simulation and measurement results.

Since packet losses cause retransmissions and, as an undesired result, increase delay of the whole setup procedure, we provision the buffer resources to ensure no losses in the queues (IPLR = 0) over normal operation of the network (no link failures). Since the biggest burst of packets in one setup procedure (EuQoS system) contains 2 packets (see Fig. 11) in the direction from calling to called and 3 packets from called to calling, then, the buffer size for M simultaneous setup procedures should equal $2 \times M$ packets in the calling-to-called direction and $3 \times M$ packets in the opposite direction.

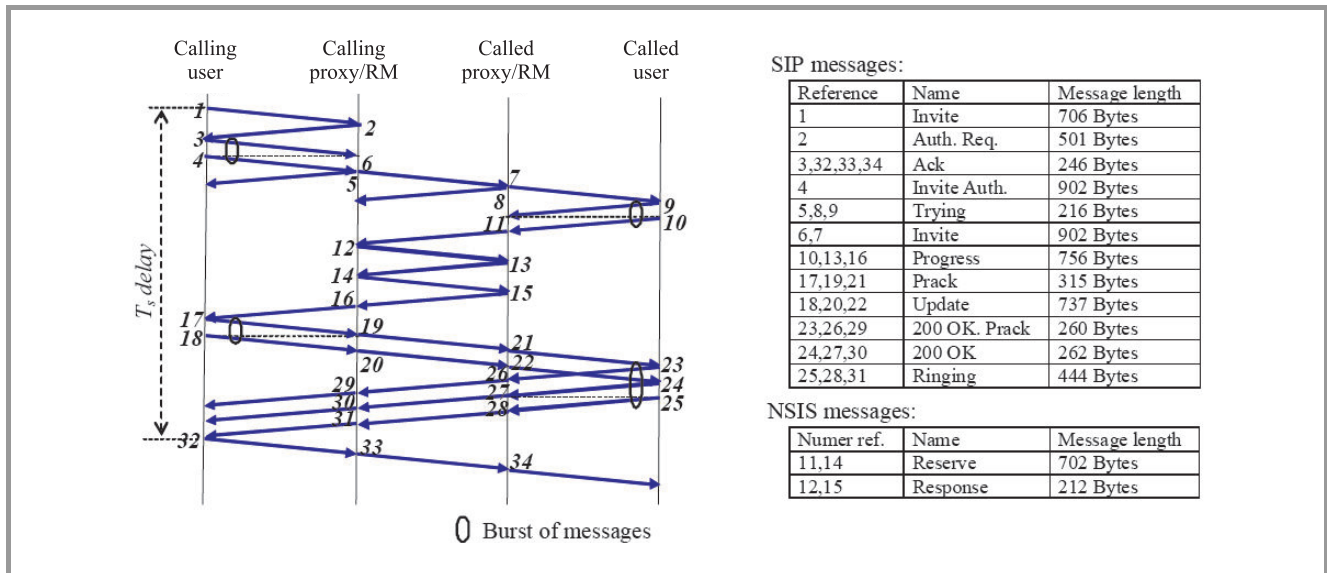


Fig. 11. SIP and NSIS message sequence in EuQoS system.

Let us remark that the transport protocol used to carry the signaling messages is not excessively important since, by provisioning the buffers, we assume no losses. Anyway, if we transport the signaling messages using the UDP protocol, we have to pay attention to the retransmission timers of the signaling protocols because these values are manually set and do not consider the round trip time (RTT) values of the network. The SIP protocol defines the default value of the retransmission timer equal to 500 ms when we use UDP protocol [28; Section 17.1.1.1]. The timers in the calling user will stop when the messages and its responses cross the whole network. This time is usually higher than 500 ms, so, such a value of retransmission timer results in the unnecessary retransmission of signaling packets (packets are not lost). This could cause a dangerous overload of the S-CoS. Therefore, in EuQoS implementation, we set the default values of SIP timers over UDP to 3 s.

4.1. Evaluation by Simulations

The performed tests aim at validating the proposed provisioning method to ensure target T_s delay. In our simulations we assume the same model as presented in Fig. 10. Moreover, we assume that the value N is equal to 5, this implies 5 servers and 25 terminals in each access network. The terminal i of the access network 1 (AN1) initiates a setup procedure with the terminal i of the access network 2 (AN2) and also the terminal i of the AN2 initiates another setup procedure with the terminal i of the AN1. All the terminals (50) initiate setup procedures at the same time and when the setup procedure corresponding to the terminal i finishes, then, the terminal i instantly initiates a new setup procedure. This is the worst-case scenario because finally, any new setup procedure finds the system full. Note that, in a real scenario, the setup procedures randomly arrive to

the system and may find the system not completely full. Therefore, the values of T_s delay presented in these simulations are the upper bound.

Table 1
Bottleneck link capacity for 1 unique setup procedure

Initiated in AN1						
C [kbit/s]	C ₁	C ₂	C ₃	C ₄	C ₅	C ₆
bott#1–bott#6	38.3	32.3	49.6	32.3	38.3	42.3
bott#1–bott#4	28.7	24.2	37.2	24.2	–	–
bott#1–bott#2	15.8	13.3	–	–	–	–
Initiated in AN2						
bott#1–bott#6	32.3	38.3	42.3	38.3	32.3	49.6
bott#1–bott#4	21.4	25.3	28.0	25.3	–	–
bott#1–bott#2	13.3	15.8	–	–	–	–

Table 2
Bottleneck buffer size for 1 unique setup procedure

Initiated in AN1						
B [packets]	B ₁	B ₂	B ₃	B ₄	B ₅	B ₆
bott#1–bott#6	2	3	2	3	2	3
bott#1–bott#4	2	3	2	3	–	–
bott#1–bott#2	2	3	–	–	–	–
Initiated in AN2						
bott#1–bott#6	3	2	3	2	3	2
bott#1–bott#4	3	2	3	2	–	–
bott#1–bott#2	3	2	–	–	–	–

Table 3

Link capacities for 25 setup procedures initiated in AN1 and 25 initiated in AN2 (values set in the simulation scenario)

$\frac{C}{B}$ [kbit/s] [packets]	$\frac{C_1}{B_1}$	$\frac{C_2}{B_2}$	$\frac{C_3}{B_3}$	$\frac{C_4}{B_4}$	$\frac{C_5}{B_5}$	$\frac{C_6}{B_6}$
bott#1–bott#6	$25 \times (38.3 + 32.3) =$ 1765.0 $25 \times (2 + 3) = 125$	$25 \times (32.3 + 38.3) =$ 1765.0 $25 \times (2 + 3) = 125$	$5 \times (49.6 + 42.3) =$ 459.5 $25 \times (2 + 3) = 125$	$5 \times (32.3 + 38.3) =$ 353.0 $25 \times (2 + 3) = 125$	$5 \times (38.3 + 32.3) =$ 353.0 $25 \times (2 + 3) = 125$	$5 \times (42.3 + 49.6) =$ 459.5 $25 \times (2 + 3) = 125$
bott#1–bott#4	$25 \times (28.7 + 21.4) =$ 1252.5 $25 \times (2 + 3) = 125$	$25 \times (24.2 + 25.3) =$ 1237.5 $25 \times (2 + 3) = 125$	$5 \times (37.2 + 28.0) =$ 326.0 $25 \times (2 + 3) = 125$	$5 \times (24.2 + 25.3) =$ 247.5 $25 \times (2 + 3) = 125$	100 000 125	100 000 125
bott#1–bott#2	$25 \times (15.8 + 13.3) =$ 727.5 $25 \times (2 + 3) = 125$	$25 \times (13.3 + 15.8) =$ 727.5 $25 \times (2 + 3) = 125$	100 000 125	100 000 125	100 000 125	100 000 125

The simulation environment is ns-2 [29] where we developed modules to simulate the signaling process and integrated them into the EuQoS simulation model [30]. The integration of these modules permits its use in any network scenario (even in environments with different network techniques). In our own-implemented modules we model the signaling protocols: SIP and NSIS.

We introduce a bottleneck in the respective link by setting the appropriate value of the link capacity obtained from the provisioning method presented above. The links that are not considered bottlenecks in a certain test are set to 100 Mbit/s. Also the links between terminals and servers are 100 Mbit/s links. Next, we perform three tests, each one with 10 000 setup procedures. The number of bottlenecks change from one test to another. The first test considers 6 bottlenecks, from bott#1 and bott#6 as indicated in Fig. 10. The next test considers 4 bottlenecks (from bott#1 to bott#4) and the last one considers only the bottlenecks in the both directions of the link between router A and router B, i.e., bott#1 and bott#2.

We expect that each bottleneck will require a different provisioning, since the volume of signaling traffic submitted to them is different (see Fig. 11). Tables 1 and 2 show the necessary values of capacity in the bottleneck links and size of the bottleneck buffers for one unique setup procedure initiated in the AN1 and one unique setup procedure initiated in AN2 for the cases of 6, 4 and 2 bottlenecks in the signaling path. Values of this tables we exploit to complete Table 3, which presents the values set in the links and buffers of the simulation scenario considering that all the 50 terminals initiate setup procedures.

For each test, we calculate the time that packets last on transferring the 100 Mbit/s links. This time is not considered within the provisioning method and we subtract it from the total T_s delay value obtained in each test. The T_s delay values of the test (after subtraction) are presented in Table 4. Specifically, Table 4 shows the T_s delay of the shortest (min) and longest (max) setup procedure. In our simulations, we discard the first setup procedures of

the tests because, at the beginning, the system is empty and these first setup procedures finish earlier, i.e., we only consider the T_s delay values of the setup procedures after striking the balance of the simulation process. We may observe that the values of T_s delay respect the target value equal to 3.5 s in all the cases. In opposition to the simulations performed in [26], in the presented simulation approach there is no multiplexing gain in the network because

Table 4

Value of T_s delay for the scenarios with 6, 4 and 2 bottlenecks (simulation results)

Bottleneck	bott#1–bott#6 min/max	bott#1–bott#4 min/max	bott#1–bott#2 min/max
T_s delay [s]	3.499/3.500	3.499/3.500	3.500/3.500

the system is permanently occupied. In fact, the setup procedures arriving to the system synchronize ourselves and transfer the network in the same order. This simulation approach allowed us to validate the provisioning method presented above.

4.2. Evaluation by Measurements

In this section we present measurement results of S-CoS performance evaluation. The aim of the measurements is to demonstrate that the S-CoS provisioned by the method presented in this paper may ensure target T_s delay in the test bed environment. On the other hand, we will also demonstrate that we cannot ensure delay requirements for the signaling traffic not carried by the S-CoS, if the network is heavily loaded.

Figure 12 presents the measurement test bed, where we emulate a scenario with two domains that are interconnected by an inter-domain link. The capacity of the inter-domain link equals 8 Mbit/s, while links inside each domain offer 100 Mbit/s capacity.

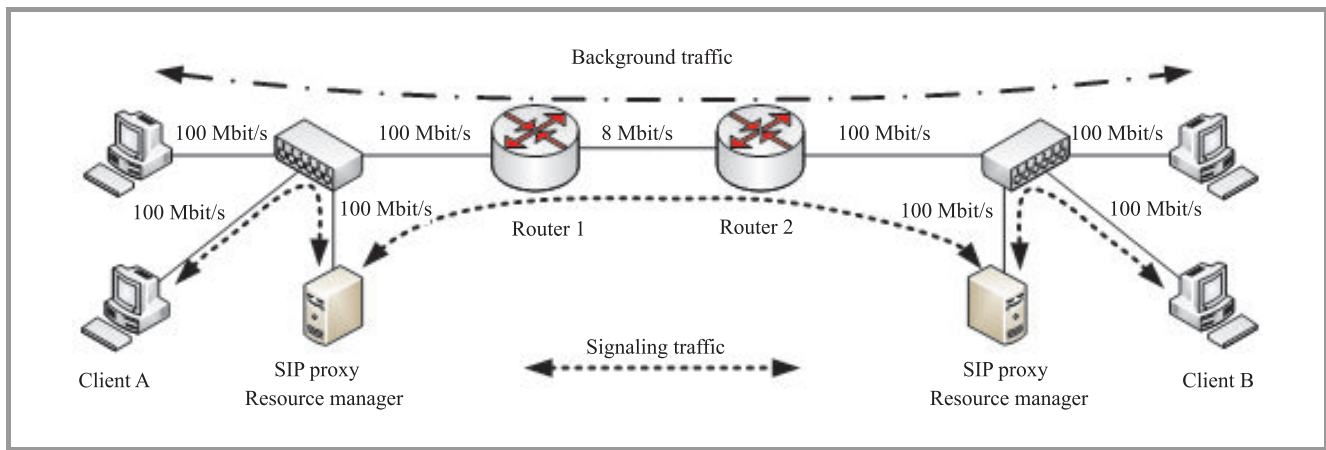


Fig. 12. Measurement test bed.

For implementing the S-CoS, all the signaling servers must “mark” their own packets before sending them into the network. This is done by setting the differentiated services code point (DSCP) field in the IP header of packets. The DSCP value assumed for the signaling packets is equal to binary value 101000 [4].

Figure 13 presents the model of the outgoing port of two routers: router 1 and router 2. In router 1 it is implemented in the direction from client A to client B and, in router 2, in the opposite direction.

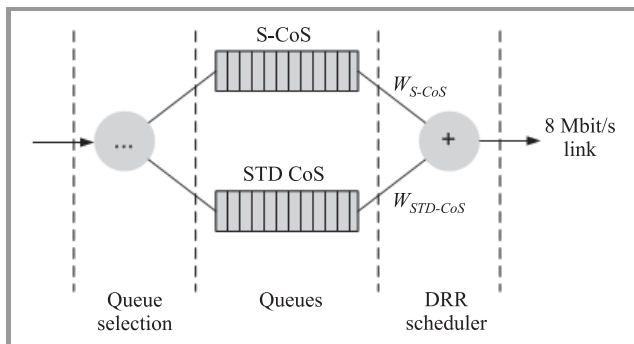


Fig. 13. Model of the outgoing port of the routers.

For our measurements, we consider only two classes of service: the S-CoS and the standard CoS (STD CoS), which is the best effort service. To ensure bandwidth separation between considered classes in IP router we use the deficit round robin (DRR) scheduler. We provision the S-CoS with 1000 and 2000 kbit/s in the two tests, respectively. The scheduler’s weight for S-CoS traffic (w_{S-CoS}) is set to guarantee assumed amount of bandwidth, whereas, the scheduler’s weight for STD CoS traffic ($w_{STD-CoS}$) complements the link capacity up to 8 Mbit/s.

By the provisioning method, we may calculate how many simultaneous setup procedures can be admitted, as a maximum, in the S-CoS not to exceed 1000 and 2000 kbit/s. When applying the provisioning method, we consider one unique bottleneck between proxy/RMs in the direction from client A to client B since the setup procedures are initiated

only in client A. Moreover, we do not consider bottlenecks between clients and proxy/RMs because there is only signaling traffic. Table 5 presents the maximum number of simultaneous setup procedures for the provisioned S-CoS capacities, as well as the necessary buffer size set in the S-CoS to avoid losses for this number of simultaneous setup procedures.

Table 5
Class of service capacities and buffer sizes
(experiment configuration)

S-CoS capacity	1000 kbit/s	2000 kbit/s
Maximum number of simultaneous setup procedures	131	262
S-CoS buffer size	262 packets	524 packets

An artificial call generator in the client A initiates the setup procedures and maintains a constant number of them running in the system. When a setup procedure finishes, the call generator instantly initiates a new one.

On the other hand, the background traffic (STD CoS) is composed by 10 TCP connections in each direction, which fill the 8 Mbit/s link. We set the buffer size in the STD CoS equal to 1000 packets, which is sufficiently large to prevent any packet loss and introduces a high delay in the STD CoS traffic.

All measurements that we take, are performed at the signaling application level (T_s delay). We simplified the RM implementation with main focus on processing of signaling messages in order to reduce the delay introduced by state management operations, e.g., database access and storage and policy algorithms.

In the first experiment we compare the performance of signaling system with and without the S-CoS. We measure the T_s delay for cases with 131 or 262 running setup procedures, with capacity assigned to S-CoS equal to 1 Mbit/s or 2 Mbit/s, respectively. In Fig. 14 we show the maximum and minimum values of T_s delay measured in the test bed. The measurement consisted of at least 1000 setup procedures after warm up period.

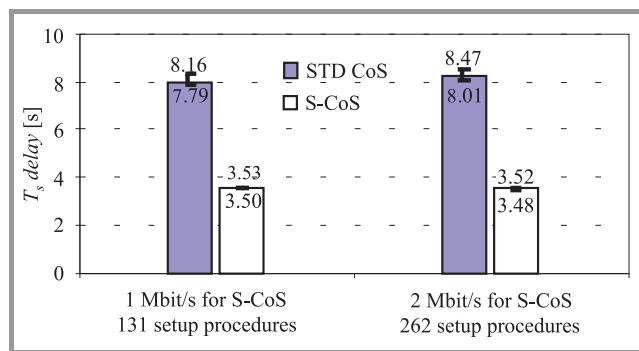


Fig. 14. Measurements of T_s delay within the EuQoS test bed.

By introducing the S-CoS in the network we are able to guarantee the value of the T_s delay; for both considered cases it is lightly higher than assumed target of 3.5 s. The tiny differences are due to the transfer of signaling packets by the not-bottleneck links. On the other hand, when signaling messages shares the capacity with the background traffic in STD CoS, the observed average T_s delay exceeds 8 s. As we expected, without the implementation of the S-CoS, we are not able to ensure the setup delay requirements for Internet telephony presented in [11]. Therefore, we argue that signaling traffic and best effort traffic must not be handled by the same network resources when the network is heavily loaded.

In the next experiment we evaluate the characteristics of T_s delay when the number of simultaneous setup procedures is lower than the calculated maximum value (for given S-CoS capacity). Figure 15 shows this characteristic for the case when 2 Mbit/s capacity is dedicated for S-CoS; we may see maximum and minimum measured values of T_s delay. The relation between T_s delay and the number of simultaneous setup procedures is linear, as we could expect. Moreover, for given number of simultaneous setup procedures, the T_s delay of different setup procedures does not vary so much (intervals are very small).

On the basis of the above results, we strongly recommend to introduce the dedicated class of service to ensure QoS

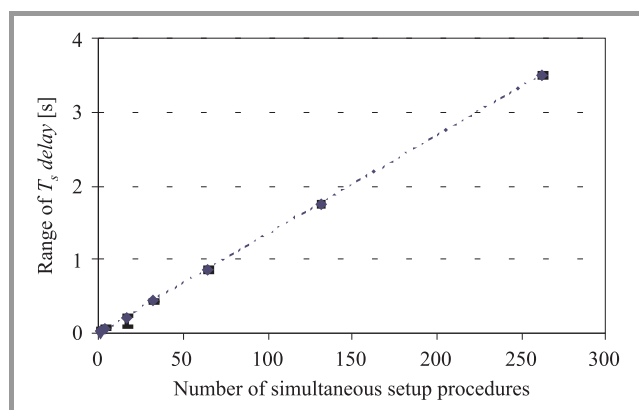


Fig. 15. Measured characteristics of T_s delay within the EuQoS test bed.

at the packet transfer of signaling messages. Unfortunately, the inherent characteristics of the signaling traffic (bursts of packets, sporadic traffic, etc.) do not allow to effectively control it and, because of this, it is not possible to carry the signaling traffic by other CoSs on aggregation with other kinds of traffics.

5. Conclusions

In this paper we studied the signaling performance of one example of next generation network system: the EuQoS system. We considered main processes inside the signaling system, as call handling scenario and transferring of the signaling messages by the network. Both these factors may affect into the call setup delay and, as a consequence, into the user quality of experience. We decided to confront them separately, using decomposition approach. As a result, we achieved a system capable of establishing end-to-end connections with the same requirements for call setup delay as assured in the ISDN network [10].

The key for the correct design of the system lies in reserving enough resources, both for signaling servers (processing power) and for the transfer of signaling messages in the network (bandwidth and buffer). For the signaling servers, we calculated the maximum call arrival rate (number of setup procedures per second) that the system can handle to assure given call processing delay. We considered all the signaling servers involved into the signaling system, paying special attention to the signaling servers designed to the resource reservation and allocation process over exemplary technologies. Following the analysis and relations between signaling servers, we evaluated system performance by means of dedicated simulation and measurement tools.

To complete the provisioning of resources, we investigated how the traffic generated by the signaling servers is handled in the network. We assumed that the number of simultaneous setup procedures handled by the signaling system is limited. Thanks to this assumption, we were able to provide dimensioning rules for signaling class of service. The rules provide a value of bandwidth and buffer space that should be dedicated for signaling traffic at the bottleneck points in order to guarantee given setup delay requirements. The devised provisioning method was verified by the simulation of the signaling system and by measurements in the test bed scenario.

Though we strived to do general analysis of the signaling system features, some results may be particular for the EuQoS system. Anyway, the presented proposals and methods can be used as a good guideline for the design of other signaling systems.

Future works in this field should be directed to find solutions for signaling congestion control, i.e., how to ensure a maximum number of simultaneous setup procedures within the system. It is also important to consider traffic models at call level, however, it requires an insight into measurements in multiservice networks. Other problem is related with the inherent complexity of the signaling systems in next generation networks. In this work, we con-

sidered only the main signaling servers in each domain, but in reality we should expect multiple ways of system decomposition and distribution.

Acknowledgements

We would like to thank all the partners of the EuQoS consortium for their cooperation and their work on developing the EuQoS system.

References

- [1] "The EuQoS Consortium" [Online]. Available: <http://www.euqos.eu/>
- [2] X. Masip-Bruin *et al.*, "The EuQoS system: a solution for QoS routing in heterogeneous networks", *IEEE Commun. Mag.*, vol. 45, no. 2, pp. 96–103, 2007.
- [3] W. Burakowski *et al.*, "Provision of end-to-end QoS in heterogeneous multi-domain networks", *Ann. Telecommun. Springer*, vol. 63, iss. 11, p. 559, 2008.
- [4] J. Babiarez and F. Baker, "Configuration Guidelines for DiffServ Service Classes". Internet RFC 4594, Aug. 2006.
- [5] K. Chan, J. Babiarez, and F. Baker, "Aggregation of Diffserv Service Classes". Internet RFC 5127, Febr. 2008.
- [6] "Signaling Requirements for IP QoS", ITU-T TR Q-Series Supplement 51 (10/2004).
- [7] "Resource and admission control functions in next generation networks", ITU-T Rec. Y.2111 (09/2006).
- [8] "General overview of NGN", ITU-T Rec. Y.2001 (10/2004).
- [9] "Framework for the network management of IP-based networks", ITU-T Rec. E.417 (02/2001).
- [10] "Network grade of service parameters and target values for circuit-switched services in the evolving ISDN", ITU-T Rec. E.721 (05/1999).
- [11] "Network post-selection delay in PSTN/ISDN networks using Internet telephony for a portion of the connection", ITU-T Rec. E.671 (03/2000).
- [12] "Network grade of service parameters and target values for maritime and aeronautical mobile services", ITU-T Rec. E.774 (10/1996).
- [13] "Quality of experience requirements for IPTV services", ITU-T Rec. G.1080 (06/2008).
- [14] A. Moorsel, "Metrics for the Internet age: quality of experience and quality of business", in *Fifth Perform. Worksh.*, Nuremberg, Germany, 2001.
- [15] H. Tarasiuk *et al.*, "Designing the simulative evaluation of an architecture for supporting QoS on a large scale", in *Proc. QoS 2008 Conf.*, Marseille, France, 2008 (to appear ACM Digital Library).
- [16] CISCO systems document [Online]. Available: http://www.cisco.com/en/US/tech/tk652/tk701/technologies_white_paper09186a00800a9818.shtml.
- [17] V. Matic, I. Franicevic, and D. Sekalec, "Parallel SIP proxy servers using direct routing approach", in *Proc. Int. Conf. Softw. Telecommun. Comput. Netw. SoftCOM 2006*, Split-Dubrovnik, Croatia, 2006.
- [18] "Host Intrusion Detection System (HIDS) v3.1. Sizing Guidelines and Tuning Primer", HP Company internal document, Sept. 2005.
- [19] Y. Bai *et al.*, "A novel distributed wireless VoIP server based on SIP", in *Proc. Multimed. Ubiquit. Eng. Conf. MUE'07*, Seoul, Korea, 2007.
- [20] A. Beben, "EQ-BGP: an efficient inter-domain QoS routing protocol", in *Proc. 20th IEEE Int. Conf. Adv. Inform. Netw. Appl.*, Los Alamitos, USA, 2006, pp. 560–564.
- [21] N. Morita, H. Imanaka, O. Kamatani, T. Oba, and K. Tanida, "Overview and status of NGN standardization activities at ITU-T", *NTT Tech. Rev.*, Nov. 2007.
- [22] J. Song *et al.*, "Overview of ITU-T NGN QoS control", *IEEE Commun. Mag.*, vol. 45, no. 9, pp. 116–123, 2007.
- [23] P. Calhoun *et al.*, "Diameter Base Protocol". Internet RFC 5127, Sept. 2003.
- [24] R. Hancock *et al.*, "Next Steps in Signaling (NSIS): Framework". Internet RFC 4080, June 2005.
- [25] D. Durham *et al.*, "The COPS (Common Open Policy Service) Protocol". Internet RFC 2748, Jan. 2000.
- [26] J. Mongay Batalla and R. Janowski, "Provisioning dedicated class of service for reliable transfer of signaling traffic", in *Proceedings 20th International Teletraffic Congress, June 2007, Ottawa, Canada*, Lecture Notes in Computer Science, vol. 4516. Berlin-Heidelberg: Springer: 2007, pp. 853–864.
- [27] P. Rodriguez and E. W. Biersack, "Dynamic parallel access to replicated content in the Internet", *IEEE/ACM Trans. Netw.*, vol. 10, no. 4, pp. 455–465, 2002.
- [28] J. Rosenberget *et al.*, "SIP: Session Initiation Protocol". Internet RFC 3261, June 2002.
- [29] "The Network Simulator ns-2" [Online]. Available: <http://www.isi.edu/nsnam/ns/>
- [30] "The ns-2 signaling modules" [Online]. Available: <http://tnt.tele.pw.edu.pl/include/tools/sim-euqos-ptl.tgz>

Note: this text is reprinted from JTIT 4/2009



Jordi Mongay Batalla was born in Barcelona, Spain, in 1975. He received the M.Sc. degree in telecommunications from Universitat Politècnica de València in 2000. He worked one year in Centro Nazionale di Astrofisica in Bologna, Italy, as research scientist specializing in qualities and capacities of diff-serv networks. Nowadays, he is

finishing Ph.D. studies in the Warsaw University of Technology, Poland. He is with Telecommunications Network Technologies (TNT) Group from 2004. His research interest focus mainly on quality of service in diffserv networks and next generation network architecture.

e-mail: jordim@tele.pw.edu.pl

Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Jarosław Śliwiński received the M.Sc. and Ph.D. degrees from the Warsaw University of Technology, Poland, in 2003 and 2008, respectively. His research interests cover traffic control, systems' design and implementation methodology.

e-mail: jareks@tele.pw.edu.pl

Institute of Telecommunications
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland



Halina Tarasiuk received the M.Sc. degree in computer science Szczecin University of Technology, Poland, in 1996 and Ph.D. degree in telecommunications from the Warsaw University of Technology, in 2004. From 1998 she is with Telecommunication Network Technologies Group at the Institute of Telecommunications, War-

saw University of Technology. In 2003 as a member of the group she received Rector's Award for scientific achievements. From 2004 she is an Assistant Professor at the Warsaw University of Technology. From 1999 to 2003 she was collaborated with Polish Telecom R&D Centre. She participated in several European and national projects (2004–2008). Her research interests focus on NGN and NWGN architectures, node and network virtualization, signaling system performance, admission control and resource allocation methods and queueing mechanisms. She is the author and co-author of more than 40 research papers presented in books, journals, and conference proceedings and about 30 technical reports.

e-mail: halina@tele.pw.edu.pl

Institute of Telecommunications

Warsaw University of Technology

Nowowiejska st 15/19

00-665 Warsaw, Poland



Wojciech Burakowski was born in Warsaw, Poland, in 1951. He received his M.Sc., Ph.D. and D.Sc. degrees in telecommunications from the Warsaw University of Technology in 1975, 1982 and 1992, respectively. Now he works as Full Professor at the Institute of Telecommunications, Warsaw University of Technology and the National Institute of Tele-

communications, Warsaw. He leads TNT research group. Since 1990 he has been involved in many COST and EU Framework Projects. He is a member of Telecommunications Section of the Polish Academy of Sciences and an expert in 7 FR Programme. He was a chairman and a member of many technical programme committees of national and international conferences. He is the author or co-author of about 170 papers published in books, international and national journals and conference proceedings and about 70 technical reports. His research areas include new networks techniques, ATM, IP, heterogeneous networks (fixed and wireless), network architecture, traffic engineering, simulation techniques, network mechanisms and algorithms.

e-mail: wojtek@tele.pw.edu.pl

Institute of Telecommunications

Warsaw University of Technology

Nowowiejska st 15/19

00-665 Warsaw, Poland

Information for Authors

Journal of Telecommunications and Information Technology (JTIT) is published quarterly. It comprises original contributions, dealing with a wide range of topics related to telecommunications and information technology. **All papers are subject to peer review.** Topics presented in the JTIT report primary and/or experimental research results, which advance the base of scientific and technological knowledge about telecommunications and information technology.

JTIT 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, voice communications devices, image and signal processing, transmission systems, network architecture, coding and communication theory, as well as information technology.

Suitable research-related papers should hold the potential to advance the technological base of telecommunications and information technology. Tutorial and review papers are published only by invitation.

Manuscript. TEX and LATEX are preferable, standard Microsoft Word format (.doc) is acceptable. The author's JTIT LATEX style file is available:

<http://www.itl.waw.pl/publ/jtit/doc/jtitaut.zip>

Papers published should contain up to 10 printed pages in LATEX author's style (Word processor one printed page corresponds approximately to 6000 characters).

The manuscript should include an abstract about 150–200 words long and the relevant keywords. The abstract should contain statement of the problem, assumptions and methodology, results and conclusion or discussion on the importance of the results. Abstracts must not include mathematical expressions or bibliographic references.

Keywords should not repeat the title of the manuscript. About four keywords or phrases in alphabetical order should be used, separated by commas.

The original files accompanied with pdf file should be submitted by e-mail: redakcja@itl.waw.pl

Figures, tables and photographs. Original figures should be submitted. Drawings in Corel Draw and PostScript formats are preferred. Figure captions should be placed below the figures and can not be included as a part of the figure. Each figure should be submitted as a separated graphic file, in .cdr, .eps, .ps, .png or .tif format. Tables and figures should be numbered consecutively with Arabic numerals.

Each photograph with minimum 300 dpi resolution should be delivered in electronic formats (TIFF, JPG or PNG) as a separated file.

References. All 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. Samples of correct formats for various types of references are presented below:

- [1] Y. Namihiro, "Relationship between nonlinear effective area and mode field diameter for dispersion shifted fibres", *Electron. Lett.*, vol. 30, no. 3, pp. 262–264, 1994.
- [2] C. Kittel, *Introduction to Solid State Physics*. New York: Wiley, 1986.
- [3] S. Demri and E. Orłowska, "Informational representability: Abstract models versus concrete models", in *Fuzzy Sets, Logics and Knowledge-Based Reasoning*, D. Dubois and H. Prade, Eds. Dordrecht: Kluwer, 1999, pp. 301–314.

Biographies and photographs of authors. A brief professional author's biography of up to 200 words and a photo of each author should be included with the manuscript.

Galley proofs. Authors should return proofs as a list of corrections 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 one week after receiving the offprint.

Copyright. Manuscript submitted to JTIT should not be published or simultaneously submitted for publication elsewhere. By submitting a manuscript, the author(s) 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 all translation rights. No part of the present JTIT should not be reproduced in any form nor transmitted or translated into a machine language without prior written consent of the publisher.

For copyright form see:

<http://www.itl.waw.pl/publ/jtit/wskazowki.html>

A copy of the JTIT is provided to each author of paper published.

(Contents Continued from Front Cover)

A Framework for Evaluation of Communication Bandwidth Market Models

W. Stańczuk et al.

66

Performance Analysis of Hybrid Phase Shift Keying over Generalized Nakagami Fading Channels

M. Youssuf and M. Z. Abdelmageed

75

Charging Phenomena at the Interface Between High- k Dielectrics and SiO_x Interlayers

O. Engström et al.

81

Channel Identification Using Chaos for an Uplink/Downlink Multicarrier Code Division Multiple Access System

M. Frikel et al.

91

The Design of 4 x 4 Multimode Interference Coupler Based Microring Resonators on an SOI Platform

T.-T. Le and L. W. Cahill

98

Percolation Driven Flooding for Energy Efficient Routing in Dense Sensor Networks

G. Vakulya and G. Simon

103

A Microscopic Approach for THz Intersubband Challenges

M. F. Pereira

111

Impact of Signaling System Performance on QoE in Next Generation Networks

J. M. Batalla et al.

117

Editorial Office

National Institute
of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland

tel. +48 22 512 81 83
fax: +48 22 512 84 00
e-mail: redakcja@itl.waw.pl
<http://www.itl.waw.pl/jtit>