ΕΘΝΙΚΟΝ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟΝ
ΠΑΝΕΠΙΣΤΗΜΙΟΝ ΑΘΗΝΩΝ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΑΠΟΤΙΜΗΣΗ ΤΟΥ ΕΡΓΟΥ ΤΟΥ ΤΜΗΜΑΤΟΣ

ΣΤΟΧΟΙ ΓΙΑ ΤΗΝ ΠΕΡΙΟΔΟ 2008 - 2013

ΠΑΡΑΡΤΗΜΑ Χ

ΠΕΡΙΛΗΨΗ ΔΙΔΑΚΤΟΡΙΚΩΝ ΔΙΑΤΡΙΒΩΝ - ΤΟΜΟΣ 2008

ΜΑΡΤΙΟΣ 2009

ΠΑΝΕΠΙΣΤΗΜΙΟΥΠΟΛΗ - ΑΘΗΝΑ 15784
Τηλ.: 210 727 5161 , FAX: 210 727 5214 , e-mail: secret@di.uoa.gr

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
DEPARTMENT OF INFORMATICS & TELECOMMUNICATIONS

ABSTRACTS OF DOCTORAL DISSERTATIONS

Athens 2008

Volume 3

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS
DEPARTMENT OF INFORMATICS & TELECOMMUNICATIONS

ABSTRACTS OF DOCTORAL DISSERTATIONS

The Committee of Research and Development

Ioanis Emiris (chair)
Dimitris Gunopulos
Elias Manolakos
Panos Rondogiannis

ISSN

Copyright © 2008

Volume 3

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
Panepistimiopolis, 15784 Athens, Greece

PREFACE

This volume contains the extended abstracts of 12 Doctoral Dissertations conducted in the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens and completed in the time period 12/2006 to 12/2007.

The goal of this volume is to demonstrate the breadth and quality of the original research conducted by our Ph.D. students and to facilitate the dissemination of their research results. We are happy to present the third collection of this kind and expect this initiative to continue in the coming years. The submission of an extended abstract in English is required by all graduating Ph.D. students. We would like to thank all Ph.D. students who contributed to this volume and hope that this has been a positive experience for them.

Finally, we would like to thank Mr. Demetris Manatakis for his help in putting together the material included in this publication. The image in the cover is a painting “Amatory” by Yannis Moralis.

The Committee of Research and Development

Ioannis Emiris (Chair)
Dimitris Gunopulos
Elias Manolakos
Panos Rondogiannis

Athens, December 2008

Table of Contents

Generic object-oriented information models for reconfigurable communication subsystems in beyond 3G mobile systems

Vangelis Gazis[1]
National and Kapodistrian University of Athens
Department of Informatics and Telecommunications

gazis@di.uoa.gr

Abstract. Global consensus on the beyond 3G mobile era sketches a heterogeneous system that combines different wireless access systems in a complementary manner and is vested with reconfiguration capabilities that enable the flexible and dynamic adaptation of the network infrastructure to meet the ever-changing service demands. For protocol stack reconfiguration to become commonplace, a language suitable for modeling, expressing and circulating metadata essential to reconfiguration, including reconfigurable device capabilities and semantic properties of reconfigurable protocol stacks, is necessary. We outline related standardization initiatives in the mobile domain, summarize existing work in reconfiguration architectures and identify key shortcomings that hinder the advent of ubiquitously reconfigurable systems. Further on, we outline the major limitations of existing standards for the representation of capability information pertaining to protocol stacks. To support reconfigurable communication systems, we identify essential metadata classes and introduce an associated object-oriented UML model. We elaborate on the design rationale of the UML model, presenting and discussing the alternative metadata representation standards and suitable encoding formats. Finally, we demonstrate the suitability of our UML model by using it to describe the standardized protocol stacks of 3G cellular network elements.

Problem statement

The wide disparity in the technical characteristics of network devices suggests that reconfigurable wireless systems will need a common set of mechanisms capable of identifying and triggering reconfiguration actions on the network infrastructure and/or the mobile devices. Fundamentally, this calls for *a common vocabulary for describing the architecture of a reconfigurable system*, discovering the feasible reconfiguration options and, finally, selecting the change to be applied upon it.

[1] Dissertation advisor: Lazaros Merakos, Professor

Related work and motivation of research

SDR Forum has defined a Software Radio Architecture (SRA) for mobile devices based on a variant of the CORBA CCM specification. SRA defines OMG IDL interfaces for installing and using waveform (i.e., software radio) applications within a single device and a set of (CCM) XML profiles to describe the hardware/software components of an SDR system, their properties and their interconnections (i.e., metadata about its software architecture). These XML profiles concern (SDR) component packaging and deployment issues and are bound to a specific deployment setting. Therefore, they cannot express important deployment invariants (e.g., protocol inter-dependencies). Most importantly, XML lacks the semantics that ensure unambiguous descriptions, thus falling short of applications where preservation of semantic integrity is sine qua non. Such applications include the unanticipated on-the-field assembly of protocols in a protocol stack that satisfies the inter-dependencies of all its constituent protocols and can provably realize the services it is intended to.

CCM treats a component and all its possible implementations as a specific named collection of features described by an OMG IDL component definition or a corresponding entry in a CORBA Interface Repository, i.e., a CCM component is assumed to comply with some well-defined behavior. However, the CCM standard does not prescribe a particular association between a CCM component and a formal semantic descriptor of that behavior, nor does it define any mechanisms to establish such an association at development-time and/or at runtime. Without an unambiguous definition of component behavior semantics, independently developed component implementations may be semantically incompatible, thus undermining the interoperability of CCM applications such the dynamic assembly of protocol stacks implemented with SRA technology.

Based upon SRA but without its dependencies on CCM, the PIM/PSM specification developed by the OMG Software-Based Communication (SBC) group also overlooks the semantic aspects arising in multiple protocol stacks and communication standards. These semantics depend solely on the different valid ways that individual protocol layers can be combined in a protocol stack; unfortunately, the PIM/PSM proposal is based on the original SRA model and provides no such modeling instruments.

The (now joint) Parlay/OSA initiative has been a major step forward towards flexible mobile service provision but did not anticipate the case of reconfigurable systems; Parlay/OSA considers the network functionality as immutable and defines technologically agnostic (i.e., OMG IDL, W3C WSDL) interfaces for accessing it. Although their logical architecture does not preclude it, the case of communication systems capable of dynamically adapting their internal instrumentation (e.g., their protocol stack) and behavior whilst operational is beyond their current scope.

To summarize, the semantic aspects of reconfiguration, particularly in applications where independently developed protocol layers with various inter-dependencies must be assembled into a set of protocol stacks that is guaranteed to function as intended to, are generally overlooked by existing initiatives.

Current standards for equipment capabilities

3GPP standards

The 3GPP network management specifications define the Network Resource Model (NRM), a protocol-independent model describing information objects that represent 3GPP network resources (e.g., an RNC network element). A generic NRM defines information object classes and interfaces independent of any protocol solution set (e.g. CORBA/IDL, CMIP/GDMO) and network domain (e.g. UTRAN, GERAN), thus providing the largest subset of information object classes that are common to all NRM instances to be defined by 3GPP (e.g., Core Network NRM, UTRAN NRM). The generic NRM specifies logical interfaces for a network management agent to retrieve the attributes of a network element, to navigate any containment relations to information objects contained therein and to manage subscription to particular events of interest so as to receive future notifications concerning those events. The 3GPP UMTS NRM specification builds on the generic NRM and extends it with additional information object classes modeling the functional entities located in UMTS network elements (e.g., RNC function) along with their possible containment configuration (e.g., RNC functions contains zero or more Iub functions).

OMA standards

The Open Mobile Alliance (OMA) User Agent Profile (UAProf) specification is concerned with capturing classes of device capability and user preference information for the purpose of customizing content delivery. UAProf achieves interoperation to standards for Composite Capability / Preference Profile (CC/PP) distribution over the Internet by leveraging mechanisms standardized by W3C for capability description and negotiation, namely:

The Resource Description Framework (RDF) standard for the definition of the UAProf data (i.e., information) model.

The Resource Description Framework Schema (RDF Schema) specification for the definition of the User Agent Profile vocabulary.

The Composite Capability / Preference Profile (CC/PP) specification as a high-level structured framework for describing capability and preference information using RDF.

The capability and preference information is represented as collections of properties (i.e., attribute-value pairs) that are classified into one of several components, each of which represents a distinguished set of characteristics. The current UAProf specification includes (but is not limited to) the following components:

HardwarePlatform, describing the hardware characteristics of the device (e.g., device type, model number, display size, input and output capabilities, etc)

SoftwarePlatform, describing the application environment, operating system and installed software of the device (e.g., operating system vendor and version, MexE support, list of audio/video encoders, etc)

BrowserUA, describing the HTML browser application.

NetworkCharacteristics, dealing with network properties and settings (e.g., supported network bearers, etc).

WAPCharacteristics, pertaining to the WAP capabilities supported by the device (e.g., WML script libraries, WAP version, WML deck size, etc).

PushCharacteristics, dealing with the push capabilities supported by the device (e.g., supported MIME types, maximum size of push message sent to the device, etc)

Profile attributes may have composite and/or multiple values and the final value of each profile attribute is resolved according to the resolution semantics prescribed for that particular attribute in the UAProf specification. The latter is reused in the 3GPP Mobile Station Application Execution Environment (MExE) specification for 3G mobile terminals.

Limitations of existing standards

From a modeling viewpoint, the generic NRM specification supports arbitrary type attributes and containment hierarchies, and the granularity of the event detection and notification mechanism is adequate for basic object-level events (e.g., a change in the attribute value of an object). Unfortunately, the generic NRM is of little practical value in describing the reconfiguration capabilities of 3GPP UMTS network elements, as it lacks a precise definition for object classes and attribute types pertaining to reconfiguration in a 3GPP UMTS network context.

The UAProf schema was designed to express immutable device capability information in strata above the network layer, where a sufficient level of abstraction from underlying network technologies is the de facto working assumption. Network and/or link layer properties (e.g., multiple RAT capability) tend to be technology specific and thus, are either unsupported or insufficiently addressed by the current UAProf specification. Although it is possible to express capability information for reconfigurable protocol stacks in suitable UAProf extensions (i.e., UAProf components) that can be integrated in the current UAProf schema relatively easy, the applicability of these solutions is hampered by the flat component model of UAProf. In the current UAProf standard, nesting of components within components is not possible, practically ruling out the representation of inherently hierarchical structures, which are fundamental building blocks of software architectures and commonplace in reconfiguration applications (e.g., protocol graphs).

Design rationale in modeling reconfigurable protocol stacks

Stratification (i.e., layering), the basic structure mechanism for protocol stacks, renders each protocol layer impervious to the functionality within other protocol layers. The functionality embodied in a protocol layer offers a particular set of services to higher protocol layers and expects a particular set of services from protocol layers. Typically, the specification of some protocol's functionality includes only the offered

services and the associated Service Access Point (SAP) primitives to invoke them; semantic information and functional dependencies to the set of services expected from other protocol layers is considered well-known and omitted from the specification.

An important issue concerns the specification of a suitable (abstract) model for reconfigurable software architectures – particularly protocol stack architectures. The software architecture of a computing system refers to its structure, which comprises software components, their externally visible properties and the established relationships among them. The introduction of reconfigurable mobile systems will require a suitable object-oriented model to describe their internal software architecture and structure in an implementation agnostic way. Such a 'reconfiguration vocabulary' provides the unified view required to define the capabilities of reconfiguration systems and to develop reconfiguration algorithms independent of implementation technologies.

Functional requirements of protocol stack reconfiguration

To support out-of-the-box reconfiguration, a reconfigurable protocol stack must be based on a modular (i.e., component-based) architecture and support structural hierarchies of arbitrary depth through component composition. Furthermore, it must adhere to an information model for reconfiguration-related metadata that describes the reconfigurable (software) architecture, its constituent components and their properties (e.g., usage semantics and component inter-dependencies) using what effectively constitutes a *reconfiguration ontology*. Hence, a reconfigurable protocol stack must be adaptable at two levels:

The base level that includes the software implementations of protocol functionality.
The meta-level comprising the (abstract) specifications of protocol functionality.
Thereupon, we propose that reconfigurable protocol stacks are built upon abstractions of (protocol) behavior specifications and implementations of those specifications.

Reference points for protocol stack reconfiguration

Reconfiguration of communication personalities and protocol stacks entails a certain degree of exchange functionality manifesting at a certain reference point. When reconfiguration is about switching between different implementations of the same protocol, the exchange reference point is virtual in the sense that it exists between an abstract specification of the respective protocol's functionality and all of its available implementations – as opposed to being an actual reference point in the protocol stack architecture. When reconfiguration entails changes to the stratification of protocol instances in a communication device, then the exchange reference point lies in the reconfigurable protocol stack realm, specifically at the boundaries of adjacent protocol instances (Fig 1).

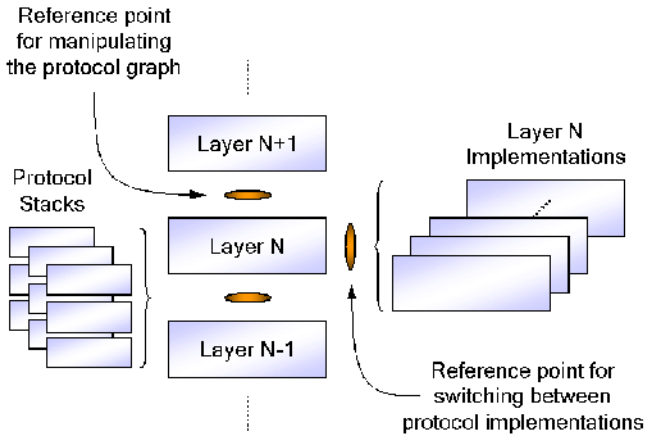


Fig 1. Essential reference points in protocol stack reconfiguration.

Metadata support for compositional definitions

Observing that different standards may reference common protocols but stratify them in different ways, we realize that the analysis granularity must support modeling of (protocol stack) standards independently of modeling of (protocol) specifications – and vice versa. To support discovery and resolution of protocol interdependencies, we postulate that protocol definitions include navigable associations to the services provided and required by each individual protocol, where each service is defined by an unambiguously identified (possibly formal) descriptor.

An object-oriented model for reconfiguration metadata

Metadata classes

`Product`, the root abstract class in our model, specifies an 'exchangeable' item that is identified through a textually represented name (e.g., by querying a name registry service). To align our model to the W3C Semantic Web work and its Resource Description Framework (RDF) that considers anything that can be identified via a URI as a resource, we include a (URI-convertible) URL attribute that provides a unique identifier of each individual `Product` instance as a Semantic Web resource.

`Service` is a subclass of `Product` that refers to some precisely defined functionality and has a textual description property. A `Service` instance provides an unambiguous

placeholder for a service definition accompanied by a textual descriptor that might be associated with arbitrary formal semantics, provided those semantics have a textual representation (e.g., OMG IDL, ITU SDL). It is not particularly important that a unique formal format is employed for the service descriptor, since adaptation mechanisms may be used to identify the appropriate handler for each available format (e.g., for syntax validation purposes). However, it is paramount that the service descriptor identifies the service *unambiguously* – an issue that is further developed in the subsection entitled "Metadata encoding".

`Specification` is a subclass of `Product` with additional (textual) attributes (author, version, release, description and summary) that represents behavioral and/or functional specifications (e.g., the specification of a session protocol). It is meant to provide a first-class abstraction for standards developed and published by authoritative bodies[2], like the UMTS standards developed by 3GPP (e.g., the GPRS Tunneling Protocol (GTP) specification). Currently, such specifications are recorded in various human-readable formats (e.g., the IETF Request for Comments (RFC) text format). The lack of a common machine-interpretable format for specifications published by different standardization bodies rules out the possibility of having those specifications parsed and interpreted by a computational agent (e.g., one in control of a reconfigurable protocol stack). `Standard` is a subclass of `specification` that provides a generic container for related specification instances, to facilitate modeling of specifications that reference other specifications, possibly published by a different authoritative body (to the one publishing the standard). For example, the 3GPP UMTS IP Multimedia Subsystem (IMS) is a standard that leverages specifications developed and published by a different authoritative body (i.e., the IETF SIP specification). We stress that, through the `Specification` and `Standard` classes, inheritance-based and composition-based modeling of communication standards is possible, thus allowing significant modeling flexibility. `Implementation` is a subclass of `Product` that refers to a software artifact that may realize multiple specifications. It is meant to model the real-life software implementation of a specification's associated functionality but may also represent functionality that is not associated to a particular specification (e.g., utility functionality). Given that an implementation may be developed in different programming languages and supporting technologies (e.g., C, C++,.NET) and packaged in various deployment formats, the `DeploymentArtifact` class and its sublcasses are used to model the different deployment artifacts (

Fig 2) an implementation may have.

`Binding` is an association class that provides a first-class abstraction for an association between a `Service` instance and a `Specification` instance. Its design purpose is to model a particular stratification of `Specification` instances in the context of a `Standard` instance. To facilitate the reuse of `Binding` instances, a `Binding` instance may be referenced by multiple `Standard` instances. During processing of a `Standard` instance, a computational agent may discover the referenced `Binding` instances

[2] The term authoritative bodies is not restricted to public bodies vested with specification development authority but includes all legal entities (e.g., private enterprises, physical persons) with a legitimate right to develop and publish the specification for a product or service.

by navigating the `binding` association(s) of the particular `Standard`. The purpose and use of the `Binding` class is demonstrated in subsequenet sections.

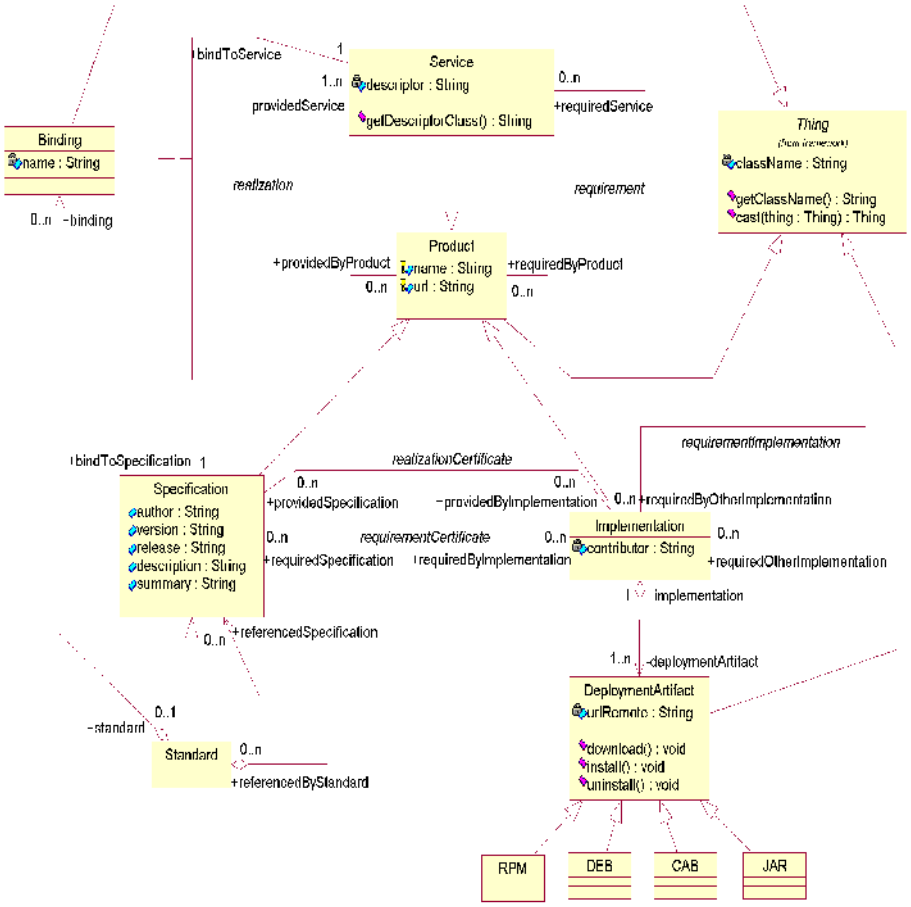


Fig 2. The object-oriented UML model for reconfiguration metadata.

Metadata associations

A `Specification` instance may depend on multiple services much as it may render multiple services. Similarly, a particular implementation, in addition to the set of services that its associated specifications collectively require and realize, may depend on additional services to function properly and may realize additional services. Because they apply to specifications and implementations alike, these concerns are expressed

through a pair of associations between the `Product` and `Service` classes named `requirement` and `realization`, respectively. This degree of modeling flexibility with regard to required and realized services facilitates arbitrary implementations (e.g., from third parties) that do, however, comply to a specification. Finally, the `requirementImplementation` association can be used to model an implementation depending on other implementations to function properly.

It is not mandatory that an `Implementation` instance be associated a `Specification` instance; it might as well be an implementation of utility functionality not subject to standardization yet required by other implementations. Thus, the case of an `Implementation` unassociated to a `Specification` instance is considered valid. Typically, the association between a `Specification` instance and an `Implementation` instance is handled through the `realizationCertificate`, `requirementCertificate` (multilateral) associations. The former signifies that the `Implementation` instance realizes the behavior of a set of `Specifications` and the latter marks its dependence upon a set of `Specification` instances.

Metadata encoding

Considering that reconfiguration metadata will be subject to processing and exchange in different administrative domains, it should be represented in an instrumentation-independent format that promotes interoperability. Two W3C standards, XML and RDF are considered as prime candidates for this task. XML is easier to use and manipulate, but RDF has a far greater capacity for expressing semantically rich information. Most importantly, only RDF is capable of unambiguous representation, as the RDF Model Theory on which it is based defines an explicit unique interpretation of any RDF data Consequently, a particular piece of information can be represented in RDF in exactly one unique way, while in XML many different representations with the same underlying meaning are possible. This advantage of RDF comes at the cost of being more verbose and significantly more complex, making it less attractive for the vast majority of users and developers.

In our approach, all reconfiguration metadata are represented in RDF, while the vocabulary used in the RDF representation is a combination of the standard RDF vocabulary and an extension vocabulary defined in an RDF Schema document, all using XML as the serialization format. The extension vocabulary is named RCM and is derived from an isomorphic mapping of the introduced UML model to an RDF Schema document. Reconfiguration metadata are expressed in the RCM vocabulary that integrates seamlessly to the standard RDF vocabulary through the RDF namespace mechanism. The primary reason for choosing RDF is its ability for unambiguous representations. Furthermore, RDF models can be serialized in XML, an interoperable, machine-interpretable textual format that is easily circulated across different administrative domains.

Support for reconfiguration option discovery

The process by which an intelligent agent discovers the possible combinations of known communication personalities and associated protocol implementations that render an integral and usable system is termed reconfiguration option discovery. Thanks to the graph model theory foundation of RDF, one can query an RDF graph of reconfiguration metadata for entries with particular properties and get unambiguous results. W3C has developed the SPARQL query language that establishes the definitive grammar for querying an RDF graph for statements matching a given pattern. Through formulation and submission of appropriate SPARQL queries, an intelligent agent can navigate a knowledge base (i.e., a RDF graph) of reconfiguration metadata and thus support reconfiguration option discovery. Regarding protocol stack reconfiguration, the application of SPARQL facilitates the discovery of the set of services required at a particular protocol strata as well as the set of specifications and/or implementations providing those services. The use of RDF and SPARQL greatly simplifies the consistency checking of reconfigurations, thus preserving the protocol stack's semantic integrity across reconfigurations.

Application scenarios: 3G network elements and protocol stacks

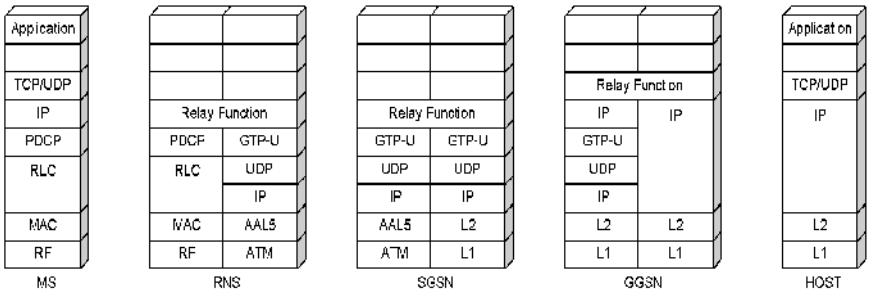
Application of the reconfiguration metadata model in 2G/3G protocol stacks

In this section we use the introduced reconfiguration ontology to model the protocol stratifications in the user plane of the packet switched domain of the GPRS and UMTS access networks (Fig 3). The (tentative) list of services in Table 1 servers mostly illustration purposes; alternative identification and naming of service entities (e.g., in further detail) is possible. Fig 4 provides a simplified form of the RDF graph for the reconfiguration metadata pertaining to the Fig 3 protocol stacks.

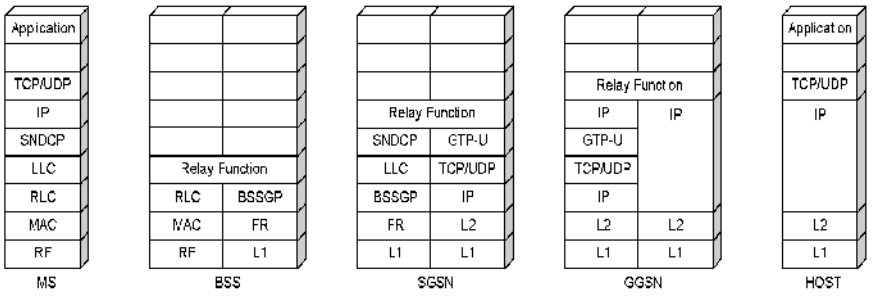
Table 1. Analysis of 2G/3G user plane protocol stacks and identification of Service classes.

Providing Specifications	Service	Requiring Specifications
RF	"Layer_1"	MAC, FR
AAL5, FR	"Layer_2"	IP, BSSGP
BSSGP	"Layer_2_BSSGP"	LLC
LLC	"Layer_2_LLC"	SNDCP
MAC	"Layer_2_MAC"	RLC
RLC	"Layer_2_RLC"	PDCP, LLC
IP	"Layer_3"	UDP
UDP	"Layer_4"	GTP

IP	"Layer_Tunneling"	IP, GTP
PDCP, SNDCP	"Layer_Convergence"	IP



UMTS (3G) Packet-Switched (PS) domain user plane protocol stacks



GPRS (2G) Packet-Switched (PS) domain user plane protocol stacks

Fig 3. User plane protocol stacks for the 3GPP GPRS and UMTS cellular access standards.

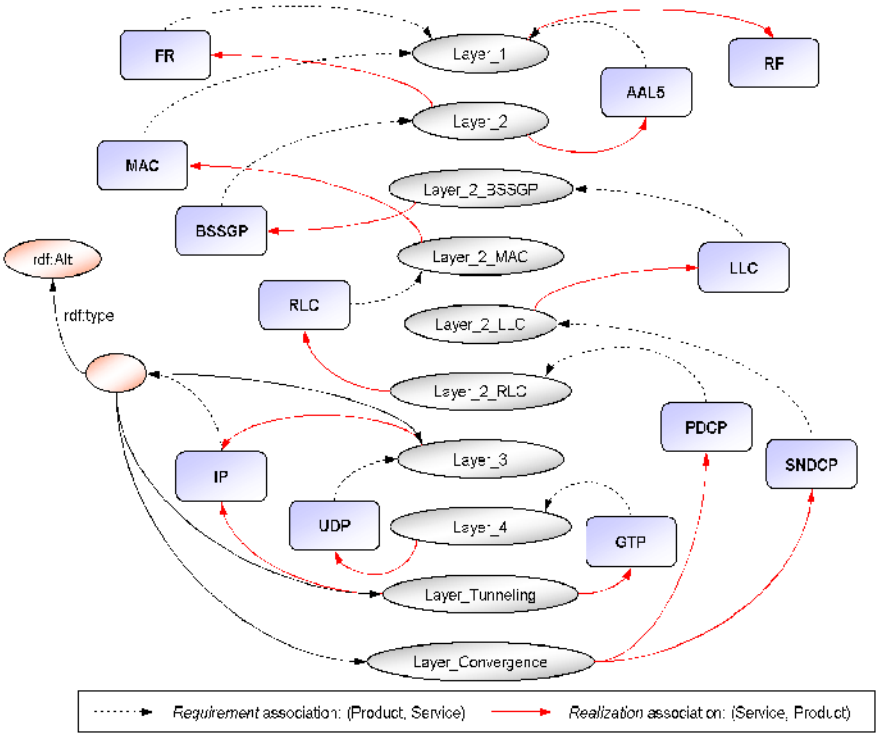


Fig 4. Simplified RDF graph for reconfiguration metadata pertaining to 2G/3G protocol stacks.

A communication standard may include several protocol stacks each with a specific stratification of protocol instances. Different communication personalities may employ some protocol instances in common but stratify them in radically different ways, thus requiring additional metadata classes and/or associations to capture and express the stratification differentia among them. For example, let's consider the user plane protocol stack of the Serving GPRS Support Node (SGSN) and Radio Network Controller (RNC) network elements in the 3GPP UMTS cellular network. The SGSN and RNC exhibit significant similarities in their protocol stacks, referencing the same protocols (e.g., GTP, UDP, IP, etc) but stratifying them differently, depending on the particular interface (Iu-PS, Gn, Iub, Iur, etc) the protocol stack concerns. While reconfiguration option discovery based on requirement and realization associations supports the identification of all valid alternative stratifications, it cannot contribute to the decision regarding which particular alternative to employ in a given situation. To support switching between entire communication personalities and their associated protocol stacks, it is required to explicitly model the differences in the involved protocol stratifications (if any) and record them in the reconfiguration metadata knowledge base.

The `Binding` class in our metadata model undertakes this role, by modeling a specific association between a `Service` instance and a `Specification` instance. Each `Binding` instance associates a `Service` to a `Specification` and applies in the context of all `Standard` instances referencing it. The primary purpose of the `Binding` class is to restrict the applicable `Specification` instance for a `Service` instance, effectively guiding reconfiguration option discovery through a specific edge in the RDF graph of the reconfiguration metadata. During processing of a `Standard` instance, an intelligent agent may discover the `Binding` instances contained therein in order to narrow the set of valid `Specification` instances for a specific `Service` and to select the appropriate `Specification` instance among multiple alternatives.

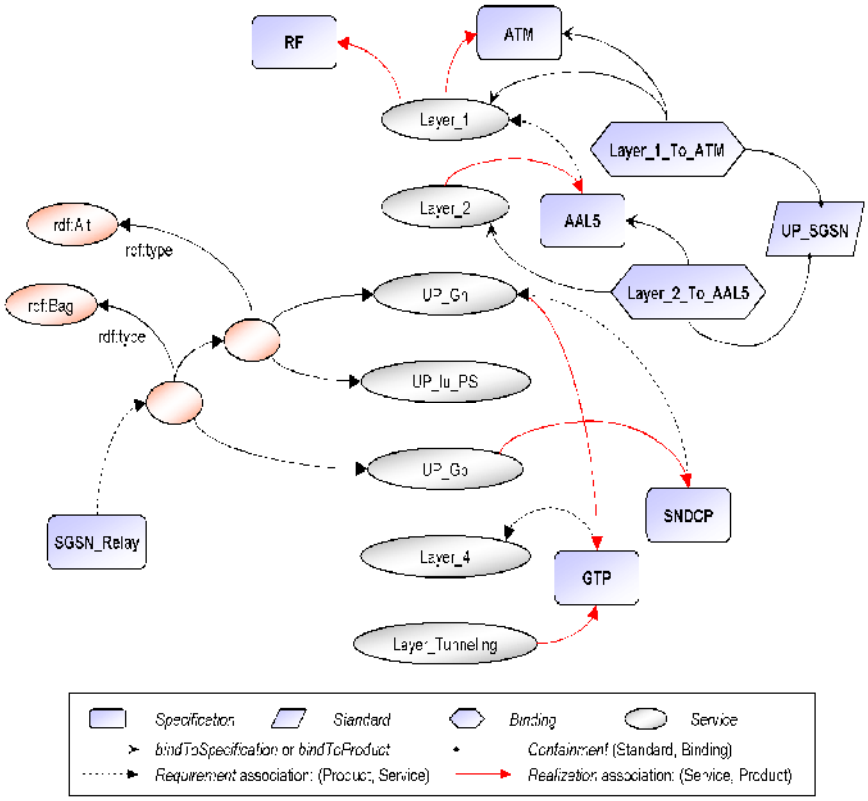


Fig 5. Simplified RDF graph for reconfiguration metadata pertaining 2G/3G SGSN network element.

Modeling and representation of the SGSN UMTS network element

We now illustrate modeling of the user plane protocol stacks and their associated dependencies for the SGSN network element for its Iu-PS, Gb and Gn interfaces, as follows:

Each of the standardized SGSN user plane interfaces (i.e., "Iu_PS", "Gb" and "Gn"), is modeled as a `Service` instance (prefixed by "UP"). The protocol specifications providing each `Service` instance are indicated through `providedByProduct` associations.

The SGSN network exposes different logical interfaces and their associated protocol stacks to different network elements (e.g., HLR, RNC, GGSN). The SGSN user plane relay functionality is modeled as a Specification instance named "SGSN_Relay" that depends on the "Gn" service and either one of the "Iu_PS" and "Gb" services. This realistically models the packet switched domain user plane architecture of a SGSN network element that interfaces to a GGSN network element over the Gn logical interface and may also interface to an RNC network element over the Iu-PS interface and/or a BSC network element over the Gb interface.

The collection of user plane protocol stacks of the SGSN network element are modeled as a `Standard` instance named "SGSN" that contains at least one "SGSN_Relay" Specification instances. This caters for SGSN network elements with multiple simultaneous interfaces to both RNC and BSC network elements (i.e., a 2G/3G SGSN).

A protocol stratification that is specific to the SGSN network element is modeled as a distinct `Binding` instance between a `Service` instance and a particular `Specification` instance. In the SGSN case, the stratification of the IP protocol over the AAL5 protocol in the Iu-PS interface is modeled as a `Binding` between the "Layer_2" Service and the "IP" `Specification`. In a similar manner, the stratification of the AAL5 protocol over the ATM protocol in the Iu_PS interface is also modeled as a `Binding` between the "Layer_1" Service instance and the "ATM" `Specification` instance.

According to the above formulation, each "SGSN_Relay" instance will be dependent upon either of the ("Gb", "Gn") and ("UP_Iu_PS", "Gn") Service pairs. By modeling the SGSN as a Standard instance that contains "SGSN_Relay" Specification instances, it is possible to support both the Gb and Iu-PS interfaces by multiple "SGSN_Relay" instances. Fig 5 shows the RDF graph for the SGSN protocol stack in a simplified form.

Reconfiguration option discovery must be generic so as to support inter-standard and intra-standard reconfigurations with minimal runtime adjustments. Our model effectively supports that capability through the `Standard` and `Binding` classes. If `Binding` instances contained in a `Standard` instance are treated as invariants to be preserved during reconfiguration, then reconfiguration is classified as an intra-standard reconfiguration, i.e., a reconfiguration affecting only those parts of the protocol stack that the respective standard allows. If, however, `Binding` instances are ignored by reconfiguration option discovery, then reconfiguration is an inter-standard reconfiguration, i.e.,

a reconfiguration that may radically modify the current communication standard, possibly resulting in a different communication personality, perhaps even one not described by a `Standard` instance.

Conclusions and directions for future work

Consensus on the vision of mobile systems beyond 3G mandates system support for the reconfiguration of individual protocol layers, entire protocol stacks and communication personalities (e.g., cellular, ad-hoc) through common procedures. That poses major challenges in all aspects of reconfiguration research, from designing an expressively sufficient model for reconfiguration metadata to engineering the functionality that supports reconfiguration of protocol stacks within operating network equipment. The work presented herein identified the essential classes and associations to support the envisaged reconfiguration capability for protocol stacks, regardless of what its supporting functional architecture is. The basic merit of our information model is its support for unambiguously specifying the associations that may occur between the services realized throughout an arbitrary stratification of protocols and the specifications and/or implementations (of protocols) requiring and/or providing those services, however complex those associations may be. This includes associations with requirement and realization semantics that are essential to the preservation of the protocol stack's integrity across reconfigurations. Not being tied to a specific reconfiguration architecture (e.g., SRA), the introduced reconfiguration ontology may also serve as a common language to describe reconfigurable protocol stacks in a uniform way that promotes the interoperability of the different architectures supporting reconfiguration. In this respect, the contribution of the present thesis is twofold: At a research level, it identifies the pivotal aspects of reconfiguration, enumerates and classifies its manifestations, and, points out the relation to software architecture research. At a technical level, it provides a minimal yet expressive object-oriented UML model for reconfiguration metadata that can be employed to describe the capabilities of reconfigurable protocol stacks for beyond 3G systems. The higher complexity associated with the choice of RDF as our model's realization technology is the price to pay for semantic univocality – although collateral benefits, such as seamless plug-in to the Semantic Web infrastructure and setting the foundation of a reconfiguration knowledge base to support the development of self-aware, cognitive adaptive systems, probably offset the cost in the long run.

References

1. V. Gazis, N. Alonistioti, and L. Merakos, "A generic model for reconfigurable protocol stacks in beyond 3G", IEEE Wireless Communication Magazine, vol. 13, no. 3, 2006.

2. V. Gazis, N. Alonistioti, and L. Merakos, "Metadata design for reconfigurable protocol stacks in beyond 3G", Kluwer Wireless Personal Communications Magazine, vol. 36, no. 1, pp. 1–28, 2006.
3. V. Gazis, N. Alonistioti, and L. Merakos, "Toward a generic Always Best Connected capability in integrated UMTS/WLAN mobile networks (and beyond)", IEEE Wireless Communications Magazine, vol. 12, no. 3, 2005.
4. V. Gazis, N. Alonistioti, and L. Merakos, "A generic architecture for Always Best Connected UMTS/WLAN mobile networks", International Journal of Wireless and Mobile Computing (IJWMC), δεκτό προς δημοσίευση.
5. V. Gazis, N. Alonistioti, N. Houssos, M. Koutsopoulou, S. Gessler, and J. Quittek, "Intelligent network provisioning for dynamically downloadable applications in beyond 3G mobile networks", Kluwer Journal of Network and System Management (JNSM), vol. 14, no.2, pp. 221–241, 2006.
6. N. Houssos, K. Kafounis, V. Gazis, and N. Alonistioti, "Application-transparent adaptation in wireless systems beyond 3G", International Journal of Management and Decision Making (IJMDM), vol. 6, no. 1, 2005.
7. N. Houssos, V. Gazis, and N. Alonistioti, "Enabling delivery of mobile services over heterogeneous converged infrastructures", Kluwer Information System Frontiers Journal, vol. 6, no. 3, 2004.
8. D. Wisely, H. Aghvami, S. L. Gwyn, T. Zahariadis, J. Manner, V. Gazis, N. Houssos, and N. Alonistioti, "Transparent IP radio access", IEEE Wireless Communication Magazine, vol. 10, no. 4, 2003.
9. S. Panagiotakis, M. Koutsopoulou, N. Houssos, V. Gazis, and N. Alonistioti, "An advanced service provision framework for reconfigurable mobile networks", International Journal of Mobile Communications, vol. 1, no. 4, 2003.
10. Σε βιβλία:
11. N. Alonistioti, S. Panagiotakis, M. Koutsopoulou, V. Gazis, and N. Houssos, "Open APIs for Flexible Service Provision and Reconfiguration Management", in Software Defined Radio: Architectures, Systems and Functions, M. Dillinger, K. Madani, and N. Alonistioti, Eds.: John Wiley & Sons, Ltd, 2003, pp. 165-189.
12. S. Panagiotakis and V. Gazis, "Adaptive Protocols", in Software Defined Radio: Architectures, Systems and Functions, M. Dillinger, K. Madani, and N. Alonistioti, Eds.: John Wiley & Sons, Ltd, 2003, pp. 73-92.
13. Σε συνέδρια:
14. V. Gazis, N. Alonistioti, and L. Merakos, "Metadata design for introspection-capable reconfigurable systems", in Lecture Notes in Computer Science: NETWORKING 2004, Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications, Vol. 3042. Springer, Athens, Greece (2004) 1318-1325.
15. N. Houssos, V. Gazis, M. Koutsopoulou, and N. Alonistioti, "Middleware platform support for the realization of advanced business models in beyond 3G environments", in proceedings of the 8th International Workshop on Mobile Multimedia Communications, Munich, Germany, 2003.
16. N. Houssos, V. Gazis, and N. Alonistioti, "Application-transparent adaptation in wireless systems beyond 3G", in proceedings of the 2nd International Conference on Mobile Business (M-Business) Vienna, Austria, 2003.
17. N. Houssos, F. Foukalas, V. Gazis, K. Kafounis, and N. Alonistioti, "Adaptability issues in reconfigurable environments", in proceedings of 2nd ANWIRE workshop on reconfigurability, Mykonos, Greece, 2003.

18. V. Gazis, N. Houssos, M. Koutsopoulou, S. Pantazis, and N. Alonistioti, "Towards reconfigurable 4G mobile environments", in proceedings of the 2nd ANWIRE workshop on reconfigurability, Mykonos, Greece, 2003.
19. V. Gazis, N. Houssos, N. Alonistioti, and L. Merakos, "Service provision and reconfiguration management in 4G mobile networks", in proceedings of the IEEE 5th International Conference on Mobile Wireless Communication Networks (MWCN), Shangri-La Hotel, Singapore, 2003.
20. V. Gazis, N. Houssos, N. Alonistioti, and L. Merakos, "Generic system architecture for 4G mobile communications", in proceedings of the IEEE 57th Semiannual Vehicular Technology Conference (VTC), Jeju, Korea, 2003.
21. V. Gazis, N. Houssos, and N. Alonistioti, "Reconfiguration management In 3G/4G mobile environments: Requirements, process and architecture", in proceedings of the SDR Forum Technical Conference, Orlando, Florida, USA, 2003.
22. V. Gazis, N. Houssos, A. Alonistioti, and L. Merakos, "On the complexity of Always Best Connected in 4G mobile networks", in IEEE 58th Semiannual Vehicular Technology Conference (VTC), Orlando, Florida, USA, 2003.
23. N. Houssos, V. Gazis, S. Panagiotakis, M. Koutsopoulou, and A. Alonistioti, "Advanced business models and flexible service provision for reconfigurable mobile systems", in proceedings of the SDR Forum Technical Conference, San Diego, California, USA, 2002.
24. N. Houssos, V. Gazis, S. Panagiotakis, S. Gessler, A. Schuelke, and S. Quesnel, "Value-added service management in 3G networks", in proceedings of the IEEE/IFIP 8th Network Operations and Management Symposium (NOMS), Florence, Italy, 2002.
25. N. Houssos, V. Gazis, and A. Alonistioti, "A flexible management architecture for the support of advanced business models in 3G mobile service provision", in proceedings of the 1st International Conference on Mobile Business (M-Business), Athens, Greece, 2002.
26. V. Gazis, N. Houssos, N. Alonistioti, and L. Merakos, "Service provision evolution in mobile communication networks", in proceedings of the EU-China Post Conference on Beyond 3G Beijing, China, 2002.
27. V. Gazis, N. Houssos, A. Alonistioti, and L. Merakos, "Evolving perspectives on 4th generation mobile communication systems", in proceedings of the IEEE 13th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Lisbon, Portugal, 2002.
28. N. Alonistioti, N. Houssos, S. Panagiotakis, M. Koutsopoulou, and V. Gazis, "Intelligent architectures enabling flexible service provision and adaptability", in proceedings of the Wireless Design Conference (WDC), London, UK, 2002.
29. S. Panagiotakis, V. Gazis, M. Koutsopoulou, N. Houssos, Z. Boufidis, and N. Alonistioti, "Roaming issues for service provisioning over 3rd Generation mobile networks", in proceedings of the Advanced Technologies Applications and Market Strategies for 3G (ATAMS), Krakow, Poland, 2001.
30. K. Molnar, Z. Nagy, S. Panagiotakis, V. Gazis, N. Houssos, and M. Koutsopoulou, "Location features in the MOBIVAS project", in proceedings of the Mobile Location Workshop (MLW), Espoo, Finland, 2001.
31. M. Koutsopoulou, V. Gazis, and A. Kaloxylos, "A novel billing scheme for UMTS networks", in proceedings of the International Symposium on 3rd Generation Infrastructure and Services, Athens, Greece, 2001.
32. M. Koutsopoulou, C. Farmakis, and V. Gazis, "Subscription management and charging for value added services in UMTS networks", in proceedings of the Vehicular Technology Conference (VTC), Rhodes, Greece, 2001.

33. M. Koutsopoulou, N. Alonistioti, V. Gazis, and A. Kaloxylos, "Adaptive charging accounting and billing system for the support of advanced business models for VAS provision in 3G systems", in proceedings of the IEEE 12th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), San Diego, California, USA, 2001.
34. N. Alonistioti, V. Gazis, M. Koutsopoulou, and S. Panagiotakis, "An Application platform for downloadable VAS provision to mobile users", in proceedings of the IST Mobile Communications Summit, Galway, Ireland, 2000.

Concept Mapping in Didactics of Informatics. Assessment as a Tool for Learning in Web-based and Adaptive Educational Environments

Evangelia Gouli*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
lilag@di.uoa.gr

Abstract. Regarding assessment as an integral and essential part of the processes of teaching and learning, in the context of this thesis, alternative assessment methods (i.e. self-, peer- and collaborative-assessment) and tools (i.e. concept maps) are studied, aiming to promote learning, to evaluate cognitive skills and to cultivate/develop meta-cognitive and social skills. Furthermore in the direction of promoting meaningful learning through assessment, computer-based learning environments are developed, which exploit these methods and tools and have as basic unit the concept of the activity.

Keywords: assessment, concept mapping, feedback, adaptation, self-assessment, peer-assessment, collaborative-assessment.

1 Introduction

Assessment is central to the practice of education and one of the most powerful educational tools for promoting and motivating effective learning. Whereas in the past, assessment is considered as a means to determine measures and thus certification, there is now a notion of assessment as a tool for learning and a realization that the potential benefits of assessing are much wider and impinge on in all stages of the learning process [6]. Birenbaum [1] has made a useful distinction between two cultures in the measurement of achievement and relates them to developments in the learning society. In traditional so-called *testing culture*, instruction and assessment (testing) are considered to be separate activities and the testing culture fits well with the traditional approach to education where teaching is seen as an act of depositing the content which students receive, memorize and reproduce [1], [6]. The changing learning society has generated the so-called *assessment culture* which strongly emphasizes the integration of instruction and assessment and assessment culture is in accord with the constructivist approach to education where learning is viewed as a process through which the student creates meaning [1], [6]. Assessment culture can be used to change

* Dissertation Advisor: Maria Grigoriadou, Associate Professor

instruction from a system that transfers knowledge into students' heads to one that tries to develop students who are capable of learning how to learn.

In many cases, poor assessment practices can often be held responsible for low quality instruction and learning and may lead to undesirable consequences such as learning difficulties and reduction of students' motivation for learning. Many researchers argue that sound assessment practices can be used to improve instruction [6]. The exploitation of alternative methods and tools may make the assessment process a valuable learning experience, contribute to changing the culture amongst students from a testing culture to an assessment culture, foster a deep approach to learning and encourage students to engage continuously and change their learning methods. In this context and regarding the assessment of student learning as an integral and essential part of the processes of teaching and learning, the main goal of the research was: (i) to study alternative tools and assessment methods which aim to promote learning, to evaluate cognitive skills and to cultivate/develop meta-cognitive and social skills, and (ii) to develop computer-based learning environments, which have as basic unit the concept of the activity and exploit alternative assessment tools such as concept maps and assessment methods such as peer-, self- and collaborative-assessment.

Concept maps are considered to be a valuable tool of an assessment and learning toolbox, as they provide an explicit and overt representation of learners' knowledge structure and promote meaningful learning [17]. A concept map is comprised of nodes (representing concepts), and links, annotated with labels (representing relationships between concepts), organized in a structure (hierarchical, cyclic or hybrid) to reflect the central concept of the map. The triple Concept-*Relationship*-Concept constitutes a proposition, which is the fundamental unit of the map. Concept mapping, the process of constructing a concept map, is considered to be a creative activity, in which students must exert effort to clarify concept meanings in specific domain knowledge, by identifying important concepts, establishing the concepts relationships, and denoting their structure [17]. Various applications of concept maps in education and a number of concept mapping software tools are presented in [4]. Towards the direction of exploiting the value of concept map as assessment and learning tool, an adaptive web-enabled concept mapping environment, referred to as COMPASS (COncept MaP ASSessment and learning environment) was developed. The aim of COMPASS is twofold: to assess learners' understanding as well as to support the learning process.

Regarding the exploitation of alternative assessment methods, contemporary educational theories indicate that self-, peer- and collaborative-assessment enable students to actively participate in the assessment process, think more deeply, develop important cognitive skills such as critical thinking, evaluative abilities, teamwork, decision-making, self-monitoring and regulation, get inspiration from their peers' work, learn to collaborate, criticise constructively and suggest improvements, and reflect on the amount of effort they put into their work and judge the appropriateness of the standards they set for themselves [6], [19], [20]. However, students require exerting more effort than in traditional assessment methods as they undertake multiple roles such as the role of author and assessor and have to be trained and understand their role in the assessment process. An overall overview of studies of self-, peer- and collaborative-assessment can be found in [19], [21]. In an attempt to elaborate and contribute to the realization of these assessment methods, a web-based environment, referred to as

PECASSE (PEer and Collaborative ASSessment Environment) was developed, which engages learners in self-, peer- and collaborative-assessment activities.

The paper is organized as follows. In Section 2, a description of the web-enabled adaptive concept mapping environment COMPASS is provided. Afterwards, in section 3, the PECASSE environment is presented and, the paper ends, in section 4, with the main points of the research and its contribution.

2 An Overview of COMPASS

COMPASS (available at http://hermes.di.uoa.gr/compass) is a web-enabled concept mapping assessment and learning environment, which aims to assess learner's understanding as well as to support the learning process by employing a variety of concept mapping activities, applying a scheme for the qualitative and quantitative estimation of learner's knowledge and providing different informative, tutoring and reflective feedback components, tailored to learner's individual characteristics and needs [11].

Based on the learning goal that student selects, which corresponds to a fundamental topic/concept of the subject matter, COMPASS provides various activities, addressing specific learning outcomes. Depending on the outcomes, the activities may employ different concept mapping tasks, such as the construction of a map, the evaluation/correction, the extension and the completion of a given map; each of these tasks provides a different perspective of learner's understanding [18]. The concept mapping tasks are characterized along a directedness continuum from high-directed to low-directed, based on the context of the task and the support provided to students; students may have at their disposal a list of concepts and/or a list of relationships to use in the task and/or may be free to add the desired concepts/relationships. In Fig. 1, the main screen of COMPASS is shown. It consists of (i) the menu and toolbar, which provide direct access to several facilities such as the provision of feedback and the analysis of the map, and (ii) the Working Area, on which the central concept (in case of the construction) or the working map (constructed by the teacher) (e.g. the map that students have to evaluate/correct, or extend or complete or comment) are presented.

In the following, we discuss the assessment scheme applied for the evaluation of students' concept maps and the feedback process followed.

Evaluating a Concept Map in COMPASS. Concept maps have been extensively used, especially in science education, to assess learners' knowledge structure, in large-scale as well as in classroom assessment. The assessment is usually accomplished by comparing learner's map with the expert's one [18]. Two most commonly investigated assessment methods are the structural method and the relational method. The structural method [17] is limited to hierarchical maps and takes into account only the valid map components (e.g. propositions, examples, links/cross-links). The relational method focuses on the accuracy of each proposition, presents a high degree of inter-rater reliability and the evaluation results correlate well with both classroom and standardized tests [18]. Furthermore, most of the assessment schemes proposed in literature either have been applied to studies where the assessment of concept maps is hu-

man-based [18], or constitute a theoretical framework [15]. Regarding the computer-based assessment of concept maps, it seems that it is in its infancy as the number of systems that have embedded a scheme for automated assessment and for feedback provision is minimal. For example, the system proposed by [3] takes into account only the valid components, ignoring the invalid ones, which may contribute to the overall knowledge structure, whilst the assessment in Reasonable Fallible Analyzer (RFA) [5] is based on the identification of quite a few errors.

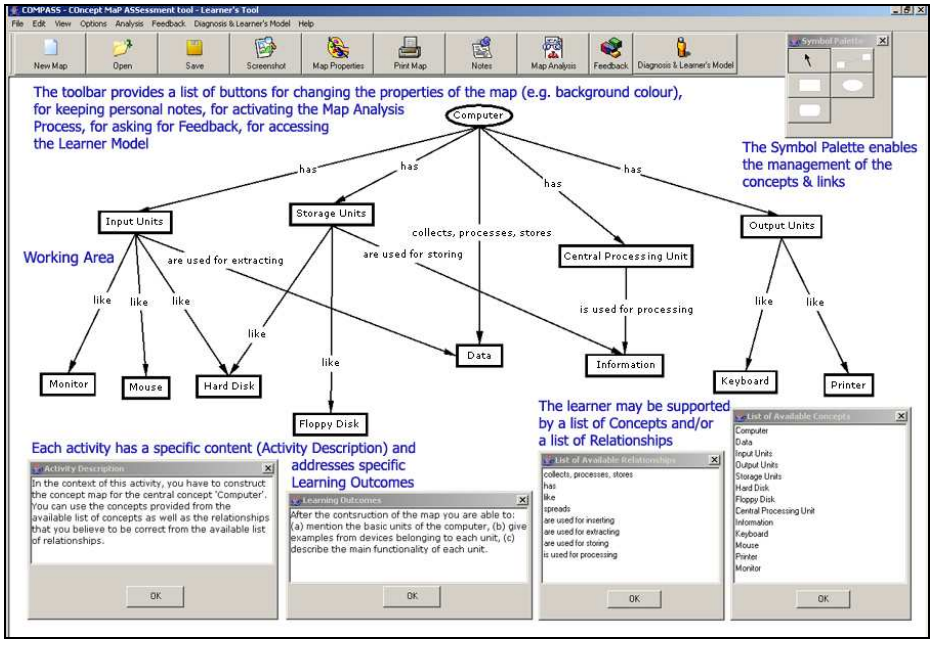


Fig. 1. The main screen of COMPASS. The Working Area presents a concept map constructed by a student in the context of a construction task supported with a list of concepts and relationships. The specific task is one of the activities provided in the context of the learning goal "The Computer Architecture"

Our work is an extension of this line of research. We propose a scheme for the assessment of concept maps and subsequently for the evaluation of learner's knowledge level on the central concept of the map. The proposed scheme adopts the relational method by examining the accuracy and completeness of the presented propositions on students' map and taking into account the missing ones, with respect to the propositions presented on the expert map. The analysis of the map (i) is based on the assessment of the propositions according to specific criteria concerning completeness, accuracy, superfluity, missing out and non-recognizability, (ii) results into the identification of specific error categories, and (iii) is discriminated in the qualitative and quantitative analysis. The qualitative analysis is based on the qualitative characterization of the errors and aims to contribute to the qualitative diagnosis of student's knowledge; that is student's incomplete understanding/beliefs and false beliefs. The quantitative analysis aims to evaluate learner's knowledge level on the central concept of the map

and is based on the weights assigned to each error category as well as to each concept and proposition that appear on expert map. The weights are assigned by the teacher and reflect the degree of importance of the concepts and propositions as well as of the error categories, with respect to the learning outcomes addressed by the activity. In this way, the teacher has the possibility to personalize the assessment process. An analytical description of the assessment scheme incorporated into COMPASS is given in [10]. The results derived from the map analysis are represented to students in an appropriate form during the feedback process.

The Feedback Process in COMPASS. Recently developed computer-based concept mapping environments attempt to embed a scheme for feedback provision [3], [5], [14]. The feedback has mainly an informative and guiding orientation and is tailored to specific common errors identified on student's concept map after the comparison of the map with the expert one. For example, in the RFA [5], feedback is provided about the quantitative score of student's map accompanied with explanation of how the score is obtained. For concepts and propositions that student believes that have not been properly credited, a dialogue between the RFA and the student could begin. Also, hints concerning missing concepts and links as well as incorrect relationships are provided. The system proposed by [2] provides hints (feedback strings defined by the expert) about specific errors such as missing propositions. Moreover, none of the systems takes into account students' individual differences.

In this line of research, COMPASS provides feedback aiming to serve processes of assessment and learning by (i) informing students about their performance and their "current" state, (ii) guiding and tutoring students in order to identify their false beliefs, focus on specific errors, reconstruct their knowledge and achieve specific learning outcomes addressed by the activity/task, and (iii) supporting reflection in terms of encouraging students to "stop and think" and giving them hints on what to think about, indicating potentially productive directions for reflection [11], [12]. To this end, different forms of feedback are supported with respect to the addressed learning outcomes and student's preferences (text-based, graphical-based and dialogue-based form) and multiple Informative, Tutoring and Reflective Feedback Components (ITRFC) are available during the feedback process in an attempt to serve processes of informing, guiding/tutoring and reflection. The Tutoring Feedback Components (TFU) supply students with learning material for the concepts represented on expert map and/or the concepts included in the provided list of concepts. The TFU are structured in two levels (the learning goal level and the activity level) and are associated with various types of knowledge modules (e.g. description or a definition of the concept under consideration, an image, an example, a counterexample, a task or a case) which aim to serve students' individual preferences and cultivate skills such as critical thinking, ability to compare and combine alternative feedback units etc. The ITRFC are structured in multiple layers and their stepwise presentation supports the gradual provision of feedback and enables students to elaborate on the feedback information and return to their map in order to correct any errors. The adaptive functionality of COMPASS is reflected to the personalization of the provided feedback in order to accommodate a diversity of students' individual characteristics and is implemented through (i) the technology of adaptive presentation that supports the provision of vari-

ous alternative forms of feedback and feedback components, and (ii) the stepwise presentation of the feedback components in the dialogue-based form of feedback. Specific student's characteristics (i.e. knowledge level, preferences, interaction behaviour), which are maintained in learner model and recorded either through student's interaction with the system or defined by the student explicitly, are used as a source of adaptation. COMPASS gives students the possibility to have control over the feedback presentation process at any time during the interaction with the environment by selecting the preferred form of feedback and by intervening in the stepwise presentation process of the dialogue in order to activate the desired stage and select the desired feedback components.

Evaluation of COMPASS. During the formative evaluation of the COMPASS environment, two empirical studies were conducted. The aim of the first study was to investigate the validity of the assessment scheme incorporated into COMPASS, as far as the quantitative estimation of students' knowledge level is concerned. In particular, we investigated the correlation of the quantitative assessment results obtained from COMPASS with the results derived from two other approaches; the holistic assessment of concept maps by a teacher and the assessment of concept maps based on the similarity index algorithm [9]. The results revealed that there is a high degree of convergence on the three scores assigned to the students' concept maps. Also, the estimation of student's knowledge level generated by COMPASS correlates closely with the similarity index, which is considered a valid indication of the quality of students' knowledge and has been taken as evidence of validity of the assessment of concept maps in other studies [16].

The second study was conducted in order to examine the hypotheses that COMPASS would help students positively on learning. In particular, the aim of the study was to investigate the effects on students' learning that have different instructional methods (concept mapping with COMPASS vs. traditional teaching) and to record the students' opinions of the COMPASS environment. Prior to the intervention, all students were administered pre-tests in achievement. After the pre-test, students were randomly assigned to one of the groups (experimental vs. control). At the conclusion of the intervention, all participants completed the post-achievement test and the students of the experimental group were asked to fill a questionnaire for COMPASS. The concept of 'Peripheral Storage Units' was used as the experimental content. The experimental group studied the concept of 'Peripheral Storage Units' by using the COMPASS environment. They were asked to construct a concept map concerning the specific central concept. They had at their disposal a list of concepts, a list of relationships and the feedback material provided by COMPASS. The control group participated in a lecture, where the instructor introduced the specific central concept and a traditional classroom teaching was followed.

The results shown that although the difference on pre-test performance is not significant (t=-0.255, df=63, 2-tailed p=0,799), the average performance after the intervention for the experimental group was significantly higher (t= 4.179, df=63, 2-tailed p<0.001) than that of the control group. Moreover, for the experimental group as well as for the control group, the difference on the performance between the two time-conditions was significant (*experimental group*: t=-24.035, df=32, 2-tailed p<0.001,

control group: t=-10.080, df=31, 2-tailed p<0.001). The results indicated that both groups improved their performance after following one of the instructional methods, but the participants of the experimental group following the instructional method with COMPASS significantly outperformed the participants who followed the traditional teaching method. This is an indication that the COMPASS environment had a better learning impact on students than the traditional teaching method. Moreover, the students of the experimental group were able to represent more accurate concepts on their maps and construct more accurate relationships among these concepts. This provides evidence supporting the inference that experimental group students were able to achieve overall higher measures of performance than control group students.

From the analysis of students' responses to the questionnaire was found that all of the students enjoyed their activity with COMPASS and found the process of constructing a concept map with COMPASS interesting. Most of the students reported that they were able to use all the supported functions immediately with minimal difficulty and they found the environment pleasant and enjoyable. The available list of concepts, the structure/steps of the dialogue-based form of feedback and the educational material stood high in most of the students favour. Among the facilities that were characterized as most useful were the explanation of the expert for the false/accurate beliefs, the educational material, the reflective questions and the performance feedback. Most of the students reported that the provided feedback helped them to learn the concepts, understand their errors and construct their concept map. All of the students reported that their activity with the COMPASS environment helped them to understand most of the underlying concepts and learn the central concept of 'Peripheral Storage Units'.

3 An Overview of PECASSE

PECASSE is a web-based environment, which engages students in self-, peer- and collaborative-assessment activities and can be used for distance education or blended learning or distance learning modes of study (available at http://hermes.di.uoa.gr:8080/pecasse). In PECASSE, students may act as (i) "*authors*" being able to submit their work/activity, which has been carried out either individually or collaboratively, (ii) "*assessors*" being responsible to evaluate (a) their own work in a brief way or according to specific criteria (self-assessment), and/or (b) their peers' work on their own or by collaborating with other learners (peer-assessment) and/or by collaborating with the instructor (collaborative-assessment), and (iii) "*feedback evaluators*" being able to evaluate the quality of feedback, provided by their assessors.

The literature review of systems developed to support self-, peer- and collaborative-assessment reveals that most of the systems focus mainly on peer-assessment and there is a lack of a system that supports all the assessment methods (self-assessment, peer-assessment and collaborative-assessment) and their possible combinations (e.g. peer- and collaborative-assessment, self- and collaborative-assessment). In most systems, authors are individuals and just a few systems support group of learners as authors. Moreover, the possibility of assessors to be group of learners is limited. The grouping of learners (in systems that authors/assessors are group of learners) as well

as the assignment of assessors is mainly done randomly; none of the systems takes into consideration learners' individual differences such as knowledge level or ability to evaluate peers' work. Regarding the review process, alternative approaches for setting the standards of the review and the form of scoring are not supported; the assessors do not have the possibility to set their own criteria/questions, enrich the criteria/questions set by the instructor and define the form of scoring.

Having as an objective to extend this line of research, we developed PECASSE, which is a discipline-independent web-based environment. In addition to the basic functions such as the uploading of assignments, the scoring/commentary of the work assessed and the presentation of the results to authors, PECASSE supports self-assessment, peer-assessment, collaborative-assessment and their combinations, individual and collaborative elaboration of the activities, review of the activities by one or group of learners, grouping of learners and assignment of assessors following alternative strategies and taking into consideration learners' individual differences, collaboration of authors and/or assessors in a synchronous and asynchronous way, alternative review methods (i.e. commentary letter or assessment form) and a variety of strategies for setting the assessment scheme applied in the review process (i.e. the instructor sets the assessment scheme or the instructor defines a template of the assessment scheme and the assessor has the possibility to modify the proposed template or the assessor proposes the criteria/questions and the form of scoring and collaborates/discusses with the instructor in order to result in an acceptable scheme or the assessor defines his/her own criteria and questions as well as the form of scoring).

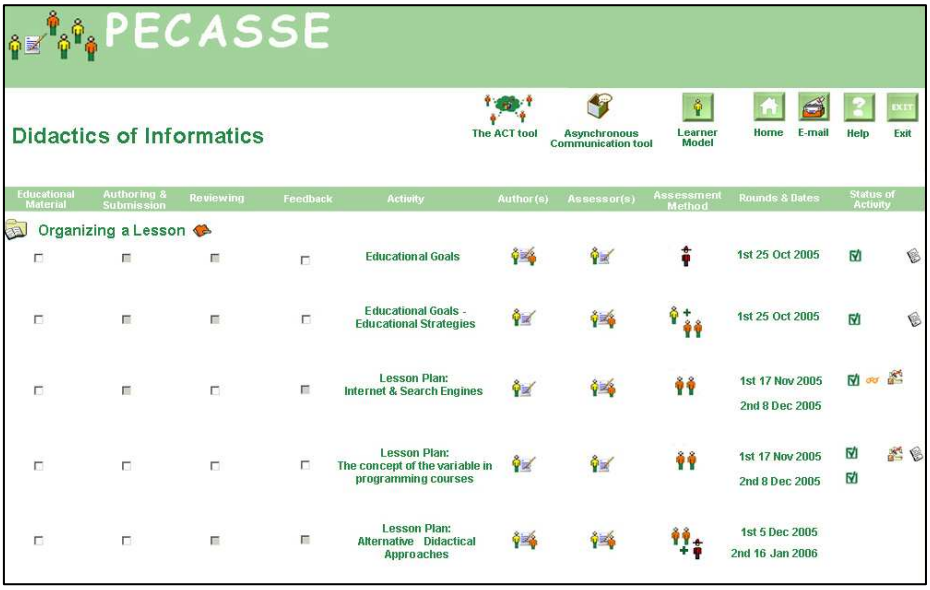


Fig. 2. A screen shot of the main screen of the PECASSE environment

In PECASSE, students have the possibility to actively participate in the assessment process which involves the following steps and can be carried out in three consecutive rounds at most, that is Step 1, 2 and 3 can be repeated up to three rounds:

- *Step 1 - Authoring & Submission*: The author is responsible to submit the activity until the deadline and proceed to self-assessment by filling a brief form.
- *Step 2 - Reviewing*: After the deadline of the submission phase, assessors are informed about the activities that they have to review. The assessors have the possibility to be anonymous or eponymous with respect to their preference. Different strategies can be followed for setting the assessment scheme with respect to the learning outcomes of the activity. In case of collaborative-assessment, the instructor collaborates with assessors in order to clarify objectives and negotiate details of the assessment process. The assessors (i.e. individual learner or group of learners) of the same activity have the possibility to collaborate in order to discuss their comments regarding the activity under review.
- *Step 3 - Feedback*: This step includes the provision of feedback to authors, the revision of the initial submitted work and the evaluation of assessors. After the deadline of the review process, the activities accompanied with grades and/or comments are returned to authors. The 'best' activities with respect to the grades assigned by the assessors and the instructor are published. Authors have the possibility to revise their work submitted to the 1st Step, taking into account their assessors' comments and the 'best' activities. Moreover, authors can communicate with assessors in order to clarify any non-understandable comments. Furthermore, authors are asked to evaluate their assessors through an evaluation form.

Fig. 2 presents the main screen of the environment after student's selection of a specific learning goal. More specifically, the learning goal of "Organizing a Lesson" in the context of the subject matter "Didactics of Informatics" and a set of five activities are presented. The first activity entitled "Educational Goals" is a collaborative one (see icon for author(s)), it is going to be assessed by one assessor (see icon for assessor(s)) and the collaborative-assessment method is followed (see icon for assessment method). Students have the possibility to access their learner model, which is dynamically updated during their interaction with PECASSE in order to keep track of their "current state". Students can see the information held in their learner model concerning their progress and communication. Furthermore, students can communicate with the instructor and their peers in the context of the subject matter in a synchronous (icon "The ACT tool" [8]) or asynchronous way (icon "Asynchronous Communication tool"). For each activity appearing in Fig. 2, students have the possibility to select the available steps of the assessment process with respect to the deadlines defined.

The group formation of students and the assignment of assessors (that is the construction of groups "authors-assessors") is facilitated by a group formation tool, referred to as OmadoGenesis [7]. OmadoGenesis enables the following strategies: (i) random assignment by the system, (ii) assignment by the instructor on the basis of his/her preferences or learners' demands, and (iii) assignment by the system on the basis of learners' individual characteristics. In any case, the instructor has the possibility to intervene and rearrange the group members in cases where conflicts are encountered and undesirable groups are formed. The instructor defines the strategy that will be followed, the group size (i.e. the desired number of learners in a group or the desired number of activities for review) and the students that will be grouped for a specific activity. In case of the third strategy, the group formation of students as well as the assignment of assessors is based on learners' model. The instructor selects stu-

dents' characteristics (up to 4) that wish to be taken into consideration such as learner's learning style and knowledge level. Then, for each selected characteristic, the instructor defines (i) if the group members will have similar values (homogeneity of the group) or dissimilar (heterogeneity of the group) and (ii) the algorithm that is going to be used and its parameters in order to find an optimal solution (for a description of the algorithms see [7]).

Evaluation of PECASSE. The study for the evaluation of PECASSE [13] showed that the majority of the participant-students were satisfied with the usefulness and the usability of the available facilities and the realization of the assessment methods. Most of the students asserted that PECASSE promotes and enhances the learning process. However, students characterized the process followed in PECASSE as time and effort consuming. In line with other researches in the area [19], the majority of the students had a positive attitude towards peer-assessment, asserting that they had received a great benefit from assessing their peers' work. More specifically, they commented that their involvement in peer-assessment made them work at a deeper level of understanding and they benefited both from the experience and the wide range of comments they received. In the context of the collaborative-assessment, most of the students characterized the role of the expert-assistants as necessary, guiding and encouraging. Moreover, they consider that the assistants' participation gave them the possibility to share a good mutual understanding of the assessment scheme through discussions and negotiations. Regarding self-assessment, most of the students did not understand the importance of self-evaluating their own activity.

As far as students' role as assessors is concerned, the quality of their work was rather high. Most of the students managed to construct the assessment form including a number of new and correct-defined criteria and question items, apply the criteria in a successful way and provide quality feedback. Moreover, most students suggested that the feedback they received from their peers was valuable for the revision of their initial work. Students also consider that the template of the assessment form and the support provided by expert-assistants helped them to design their own assessment form, provide useful feedback and cope with their role as assessors. The major problem of the review process was the difficulties that students encountered in identifying all the problems and weaknesses of the work under review. Probably, this is due to students' limited experience in the underlying learning task concerning the design and evaluation of lesson plans. In the future, we intend to use additional subjective measures such as interviews in order to analyze students' perspectives and clarify the specific problem. Two important issues revealed from the particular study that is worthwhile to mention are the need for instructor/assistant participation in the whole process and the training of students before undertaking the role of assessor.

4 Conclusions

The research presented contributes to the fields of educational assessment, didactics of informatics, concept mapping and design of computer-based adaptive learning envi-

ronments. The main contribution of the work lies in the development of learning environments that exploit alternative assessment tools such as concept maps and assessment methods such as self-, peer- and collaborative-assessment and aim to support the learning and assessment processes.

COMPASS is a web-enabled discipline-independent concept mapping environment, which aims to assess learner's understanding as well as to support the learning process. The discriminative characteristics of COMPASS are the provision of various concept mapping activities, the proposed scheme for the qualitative and quantitative estimation of learner's knowledge, the different forms of feedback supported (text-, graphical- and dialogue-based), the provision of multiple ITRFC, which serve processes of informing, guiding/tutoring and reflection, the adaptivity of the feedback process that interweaves the gradual provision of the ITRFC with the adaptive presentation of alternative forms of feedback and feedback components, accommodating learners' knowledge level, preferences and interaction behaviour, and the learner support and control offered over the feedback process.

PECASSE provides a web-based assessment environment for learners to criticize others' work, review and revise their own ideas/work, collaborate with the instructor and their peers and share their ideas. The discriminative characteristics of the PECASSE environment are the support of self-assessment, peer-assessment and collaborative-assessment as well as their combinations with respect to the learning outcomes of the activity, the options offered for the definition of authors and assessors, (i.e. the author and/or the assessor of an activity could be an individual or a group of learners), the variety of strategies offered for the assignment of assessors and the group formation of students, taking into account learners' individual differences, and the variety of strategies offered for the setting of the assessment scheme applied.

COMPASS and PECASSE could be valuable tools of instructor's toolbox, aiming to foster a learning approach to assessment. Possible enhancements of the research could be the development of facilities that support collaborative concept mapping and the exploitation of the environments within the daily educational practice.

References

1. Birenbaum, M., Dochy, F.: Alternatives in assessment of achievements, learning processes and prior knowledge. Boston:Kluwer (1996)
2. Cimolino, L., Kay, J., Miller, A.: Incremental student modelling and reflection by verified concept-mapping. In: Aleven, V., Hoppe, U., Kay, J., Mizoguchi, R., Pain, H., Verdejo, F., Yacef, K. (eds.) Supplementary Proceedings of the AIED2003: Learner Modelling for Reflection Workshop, Sydney, Australia, pp. 219-227 (2003)
3. Chang, K., Sung, T., Chen, S-F.: Learning through computer-based concept mapping with scaffolding aid. Journal of Computer Assisted Learning 17(1), 21-33 (2001)
4. Coffey, J., Carnot, M., Feltovich, P., Feltovich, J., Hoffman, R., Cañas, A., Novak, J.: A Summary of Literature Pertaining to the Use of Concept Mapping Techniques and Technologies for Education and Performance Support. (Technical Report submitted to the US Navy Chief of Naval Education and Training). Pensacola, FL: Institute for Human and Machine Cognition, (2003) Available online at: http://www.ihmc.us/users/acanas/Publications/ConceptMapLitReviewIHMCLiteratureReviewonConceptMapping.pdf

5. Conlon, T.: Formative assessment of classroom concept maps: the Reasonable Fallible Analyser. Journal of Interactive Learning Research 17(1), 15-36 (2006)
6. Dochy, F., McDowell, L.: Assessment as a tool for learning. Studies in Educational Evaluation 23(4), 279-298 (1997)
7. Gogoulou, A., Gouli, E., Boas, G., Liakou, E., Grigoriadou, M. : Forming Homogeneous, Heterogeneous and Mixed Groups of Learners. In: Brusilovsky, P., Grigoriadou, M., Papanikolaou K. (eds.) Proceeding of the Workshop on Personalization in E-learning Environments at Individual and Group Level (PING) held in conjunction with 11th International Conference on User Modeling (UM2007), Corfu, Greece (2007)
8. Gogoulou, A., Gouli, E., Grigoriadou, M. : Adapting and Personalizing the Communication in a Synchronous Communication Tool. Journal of Computer Assisted Learning (2008) (to appear)
9. Goldsmith, T., Johnson, P., Acton, W.: Assessing structural knowledge. Journal of Educational Psychology 83, 88-96 (1991)
10. Gouli, E., Gogoulou, A., Papanikolaou, K., Grigoriadou, M.: Evaluating learner's knowledge level on concept mapping tasks. In: Goodyear, P., Sampson, D., Yang, D., Kinshuk, Okamoto, T., Hartley, R., Chen N-S. (eds.) Proceedings of the 5th IEEE International Conference on Advanced Learning Technologies (ICALT 2005), Kaohsiung, Taiwan, pp. 424-428 (2005)
11. Gouli, E., Gogoulou, A., Tsakostas, C., Grigoriadou, M.: How COMPASS supports multi-feedback forms & components adapted to learner's characteristics. In: Cañas A., Novak J.(eds.) Concept Maps: Theory, Methodology, Technology, Proceedings of the Second International Conference on Concept Mapping, San José, Costa Rica, Vol.1 pp. 255-262 (2006)
12. Gouli, E., Gogoulou, A., Papanikolaou, K., Grigoriadou, M.: An Adaptive Feedback Framework to Support Reflection, Guiding and Tutoring. In: Magoulas, G., Chen S.(eds.) Advances in Web-based Education: Personalized Learning Environments, pp. 178-202 (2006)
13. Gouli, E., Gogoulou, A., Grigoriadou, M.: Supporting Self-, Peer- and Collaborative-Assessment in E-Learning: the case of the PECASSE environment. Journal of Interactive Learning Research, (2008) (to appear)
14. Hsieh, I.-L., O'Neil, H.: Types of feedback in a computer-based collaborative problem-solving group task. Computers in Human Behavior 18, 699-715 (2002)
15. Lin, S-C., Chang, K-E., Sung, Y-T., Chen, G-D. A new structural knowledge assessment based on weighted concept maps. In: Proceedings of the International Conference on Computers in Education, pp. 679-680 (2002)
16. McClure, J., Sonak, B., Suen, H.: Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. Journal of Research in Science Teaching, 36, 475-492 (1999)
17. Novak, J., Gowin, B.: Learning How to Learn. New York: Cambridge University Press (1984)
18. Ruiz-Primo, M., Shavelson, R.: Problems and issues in the use of concept maps in science assessment. Journal of Research in Science Teaching 33 (6), 569-600 (1996)
19. Sluijsmans, D., Dochy, F., Moerkerke, G.: Creating a learning environment by using self-, peer- and co-assessment. Learning Environments Research 1, 293-319 (1999)
20. Sung, Y-T., Chang, K-E., Chiou, S-K., Hou, H-T.: The design and application of a web-based self- and peer-assessment system. Computers & Education 45, 187-202 (2005)
21. Topping, K.: Peer assessment between students in colleges and universities. Review of Educational Research 68(3), 249-276 (1998)

Telecommunications Networks Planning and Evaluation with Techno-Economic Criteria

Dimitris Katsianis

Department of Informatics and Telecommunications, University of Athens
Panepistimiopolis Ilissia, Athens, Greece, GR-15784*

dkats@di.uoa.gr

Abstract. In this thesis a new tool for technoeconomic analysis has been developed. This tool is analytically presented as well as the methods that have been followed. Thesis methodology has been applied in case studies for 3G networks as well as fixed wireless and wireline access including real option and game theory approach and sensitivity and risk analysis.

Keywords: Techno-economic Analysis, Telecommunications, Demand Forecast, Real Options, Game Theory, Investments, Risk Analysis

Introduction

In this thesis we illustrate a complete methodology approach for the technoeconomic analysis of telecommunication networks that hereafter is implemented in specific detailed telecommunication case studies, wireless or wireline technologies. In the first part of this Thesis, following this methodology approach, a new tool for technoeconomic analysis has been developed. This tool is analytically presented as well as the methods that have been followed. Within this tool analysis, the theoretical approach for the cost evolution of the telecommunication components, the methodology for the demand forecast for telecommunication services and products, the approach for calculating the operation, administration and maintenance cost of the telecommunication network as well as the integration of the risk analysis model that quantified the influence of the critical parameter of the problem have been analytically presented. In the second part of this thesis the methodology has been applied in case studies for 3G networks as well as fixed wireless and wireline access (including FTTx solutions). For these case studies the real option theory has been applied, in order to clearly define the uncertainty of the investment. In addition, a game theory approach for a competition model between an incumbent and a newcomer operator has been analyzed. Finally the technoeconomic tool has been used, for the definition of a viable approach, for the provision of broadband services

* Dissertation Advisor. Thomas Sphicopoulos, Professor

in less competitive areas in Greece, including sensitivity and risk analysis (not including in this paper).

Techno-economic methodology

The techno-economic evaluation of the case studies has been carried out using the methodology introduced and the tool presented hereafter. The TITAN project [1] developed a model that predicts the cost evolution of the network components and is based on a combination of learning curves and logistic models. In addition, for each network component, the prediction uncertainties have been specified as a function of time. The learning curve model, which presents the cost of a component as a function of production volume, can be transferred to a model predicting the costs as a function of time, by the introduction of a logistic model. The original methodology and tool have been enhanced to be able to cope with complex multimedia service and network structures. Furthermore, the methodology has been improved especially in the definition of services and assessment of operations, administration and maintenance costs. As for the maintenance cost, it is defined separately and is automatically included in the model. The operation and administration cost of the Network elements are user-defined. The life-cycle cost (LCC) of the network is then produced by adding OA&M (Operation Administration & Maintenance) costs and IFCs (Installed First Costs). Finally, the overall financial budget is calculated for the various architectures by comparing the LCC to the overall revenue. This method has been followed by several telecommunication operators in Europe [2][3][4] (e.g. Deutsche Telecom, France Telecom, Telenor, Swisscom, Telecom Italia, KPN) affecting their investment policy for new services.

Structure of the Tool for Techno-Economic Evaluations.

Fig. 1 analyses the main principles of the methodology used in this thesis [5]. The cost figures for the network components have been collected in an integrated cost database, which is the “heart” of the model. This database is frequently updated with data obtained from the major telecommunication operators, suppliers, standardization bodies and other available sources. These data concerns the initial prices for the future commercial networks components as well as a projection for the future production volume of them. The cost evolution of the different components derives from the cost in a given reference year and a set of parameters which characterizes the basic principles of the component. For each component in the database, the cost evolution is estimated according to the model described in the next paragraph. In addition, estimations for the OA&M cost and the production volume of the component are incorporated in the database. As a next step in the network evaluation a services specification is needed which will be provided to the consumers. The network architectures for the selected set of services will be defined, and a geometrical model or a radio model, will be used in order to calculate the length of the cables (or the

number of the Radio stations) as well as the civil works for their installation (database data). The future market penetration of these services and the tariffs associated with them, according to each operator's policy, will be used for the construction of the market evolution model. The operator tariff policy could be taken into account by modifying the tariff level in conjunction with the expected penetration of the offered services. Results from statistics or surveys can be easily integrated into the tool when formulas measuring the impact of tariff level to the saturation of the services are available.

By entering the data into a financial model we calculate the revenues, investments (and IFC) cash flows and profits (or other financial results) of the study network architectures for each year of a project's study period. In the final evaluation of the techno-economic model, critical indexes are calculated in order to decide about the profitability of the investment.

This tool has been proved that is able to evaluate project of different scale as well as completely different and independent telecommunication technologies. The adoption of alternative financial (real options approach) and strategic methods (game theory) can be included in the tool as will be illustrated in this paper.

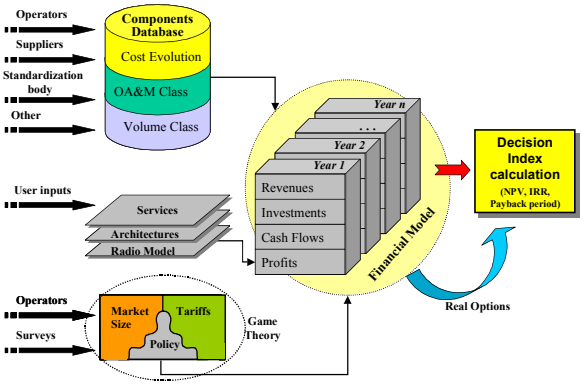


Fig. 1. Techno-economic methodology and Tool [5].

Cost Evolution Of The Network Components.

The cost prediction curve is dependent on a set of parameters such as reference cost at a given time, the learning curve coefficient that reflects the type of component, penetration at the starting time and penetration growth in the component's market. The cost database contains estimation on these parameters for all components and generates cost predictions based on the extended learning curve. The forecast function for the evolution of the relative accumulated volume $n_r(t)$ is illustrated in Eq. (1)

$$n_r(t) = \left(1 + e^{\left\{ \ln\left[n_r(0)^{-1} - 1 \right] - \left[\frac{2 \cdot \ln 9}{\Delta T} \right] \cdot t \right\}} \right)^{-1} \quad (1)$$

The expression for $n_r(t)$ can be substituted into a learning curve formula Eq.(2) yielding the final expression for price versus time in the cost database.

$$P(t) = P(0) \cdot \left[n_r(0)^{-1} \cdot n_r(t) \right]^{\log_2 \cdot K} \quad \text{or}$$

$$P(t) = P(0) \cdot \left[n_r(0)^{-1} \cdot \left(1 + e^{\left\{ \ln\left[n_r(0)^{-1} - 1 \right] - \left[\frac{2 \cdot \ln 9}{\Delta T} \right] \cdot t \right\}} \right)^{-1} \right]^{\log_2 \cdot K} \quad (2)$$

where $n_r(0)$ is the relative accumulated volume in year 0. The value of $n_r(0)$ should be equal to 0.5 for components that exist in the market and their price is expected to be further reduced due to aging rather than due to the production volume (i.e. very old products-many years in the market). From estimations in industrial telecommunication network components, $n_r(0)$ could be 0.1 for mature products and 0.01 for new components in the market. P(0) is the price in the reference year 0, ΔT is the time for the accumulated volume to grow from 10 % to 90 %, and K is the learning curve coefficient. K is the factor that causes reduction in price when the production volume is doubled. The K factor can be obtained from the production industry, mainly the suppliers. For a component (with constant nr(0)=0.1) that the ΔT is equal to 10 years and K is equal to 0.98, Eq. (2) gives almost 2% of reduction in the price of the component per year for the first 10 years. If ΔT is 5 years, this reduction is almost 4% per year for the first 5 years. All the above described values have been extensively used for the evaluation of telecommunications investment projects.

OA&M Approach.

The Operation Administration and Maintenance (OA&M) approach is divided into three separate components as follows:

1. The cost of repair parts
2. The cost of repair work
3. The Operation and Administration cost for each service cross-related to the number of customers or to the number of critical network components.

The formula for calculating OA&M cost is given by:

$$(OA\&M)_i = \frac{V_{i-1} + V_i}{2} \cdot \left(P_i \cdot R_{class} + P_l \cdot \frac{MTTR}{MTBR} \right) + OA \quad (3)$$

where V_i is the equipment volume in year i, P_i is the price of cost item in year i, R_{class} is the maintenance cost percentage for every cost component, P_l is the cost of a single working hour, MTTR is the mean time to repair and MTBR is the mean time between repairs for the cost item in question. The first term into the parenthesis

represents the cost of repaired parts, the second term represents the cost of repair work, while OA represents the Operation and Administration cost. In order to implement the calculation of the OA&M cost, classes for MTTR and MTBR are defined in the database of the technoeconomic tool as well as values for P_l and P_i.

Market Analysis

The demand modelling and broadband forecasts are essential inputs to all business case analyses. Therefore, demand models and forecasts for different access technologies in the fixed and mobile network have been developed in this thesis. In addition, models have been developed for forecasting the total broadband penetration in Europe. Forecasts have been made based on a four-parameter logistic diffusion model [6] which is recommended for long-term forecasts as well as for new services [2]. The model and the values employed are based on a compilation.

The demand model is defined by the following expression:

$$Y_t = \frac{\mathrm{M}}{(1+\exp(a+bt))^c} \qquad (4)$$

where *Yt* is the demand forecasted at time *t* and *M* is the saturation level of the penetration which is estimated a-priori. The parameters *a*, *b* and *c* are be estimated by a stepwise procedure, attempting to value these parameters using non-linear regression and data from of external reports and market surveys.

Selected Cases Studies

In the second part of this thesis the methodology has been applied in case studies for 3G networks as well as fixed wireless and wireline access (including FTTx solutions). For these case studies the real option theory has been applied, in order to clearly define the uncertainty of the investment. In addition, a game theory approach for a competition model between an incumbent and a newcomer operator has been analyzed. Finally the technoeconomic tool has been used, for the definition of a viable approach, for the provision of broadband services in less competitive areas in Greece, including sensitivity and risk analysis.

The financial perspective of the mobile networks in Europe

This case study presents a techno-economic evaluation of 3G roll-out scenarios in two "typical" European countries with contrasting profiles, analyzing both the incumbent and newcomers business cases. The analysis is based on the techno-economic methodology developed within this thesis. Market and tariff forecasts as well as the technological evolutionary paths are discussed and financial figures are analyzed. Sensitivity analysis follows these basic results in order to identify the impact of

uncertainties and risks. The success of such an investment project is mainly depended on the regulatory framework, demand and tariff structure and the market share.

Fig. 2a illustrates these different evolution paths. While the intermediate steps are overlaid onto a GSM network, UMTS requires full buildout of the radio access subsystem. Incumbent operators may, however, re-use existing GSM sites. This is a major advantage for an incumbent operator in order to provide advanced multimedia mobile services.

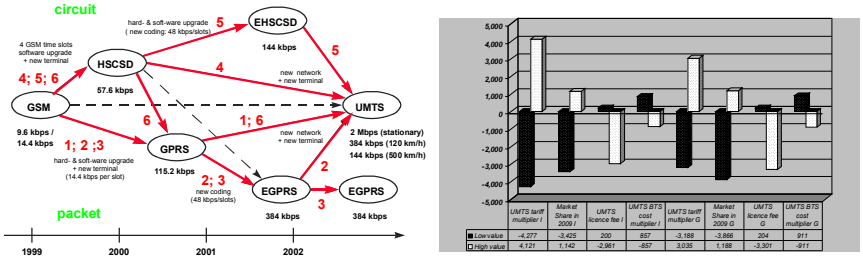


Fig. 2. a) Mobile evolution steps and b) Sensitivity analysis results. Change in NPV compared to base value (4,961 M€) in case of large country both for Incumbent (I) and Greenfield (G) case [5]

The techno-economic prospects for a new entrant and especially for an incumbent operator planning to deploy the UMTS technology are found to be positive according to the base scenarios of this study. However, after investigating the sensitivity of NPV (**Fig. 2** b) to factors such as market share, tariffs, license costs, and base station costs, we draw the reader's attention to potential pitfalls. Specifically, the following elements were identified to have major consequences on the profitability of this new business:

- Regulatory decision to promote competition: By deciding to open the UMTS market to at least four competing operators, regulators are hoping that the competitive dynamics will work to offer the widest range of services to the most customers possible at least cost. However, overcrowding leading to an end market share of 10% results in negative NPV for both the incumbent and greenfield operators. Conversely, NPV is improved by 1,200 M€ for a 5% increase in market share in 2009.
- Cost of licenses: License fee and therefore the license assignation mechanism (auction or comparative hearings) can seriously deteriorate the business case since the payback period can be delayed by more than a year, together with significantly decreasing NPV. License fee increasing from 10 to 150 €/inhabitant decreases the NPV by 66% for the incumbent.
- Tariffing of voice and data services: The tariff level ranks first over service penetration and market share as the most significant factor for UMTS profitability. This result seems logical since, in NPV calculation, the tariff level directly impacts total revenues, whereas the other parameters affect the number of customers, and hence the costs. Nonetheless, it must not be construed that operators are free to hike tariffs as they wish to achieve a positive result. Indeed, the competitive

context and dropping prices for fixed network services will severely limit their room to maneuver in this area.

- Investment schedule: Since operators deploy the radio network, using a coverage rather capacity approach (mainly due to license obligation) the cost of BTS equipment incurs a heavy financial burden. Although increased BTS cost has limited impact, it leads to larger investments in the pre-service year.

In conclusion, UMTS operators will have not very much latitude to roll out their networks. Heavy investments are required early on in order to cover the most dense areas, and then once again for the suburban areas. Competitive pressure will keep tariff levels low, and operators will need to consolidate their market assumptions with extreme care in order to evaluate the payback period. Lastly, they must have enough financial resources to stay in debt for a long period of time.

Advanced Access Networks – Use of Real Options Approach

The following case is examined: an incumbent operator offering existing services over twisted copper pairs such as POTS, ISDN and dial-up based Internet service, starts offering wideband and broadband services in 2000 using his existing copper plant. At some future date, the operator may decide to install fibre closer to the customers and thereby be able to increase customer reach and offer more advanced services over VDSL. The investment problem is illustrated in **Fig. 3**a.

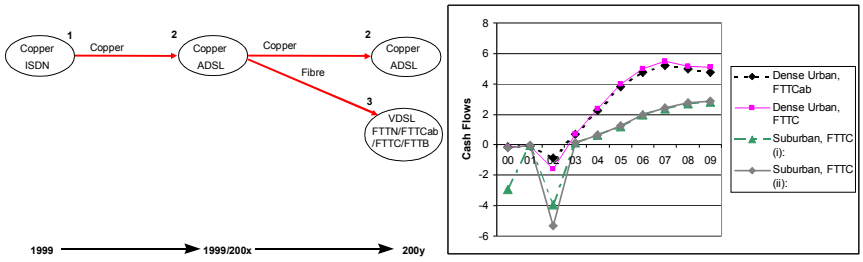


Fig. 3 a) Broadband access network investment problem and b) Cash Flows of the selected cases

The four selected strategies are: Dense Urban, FTTCab, Dense Urban, FTTC Suburban, FTTExch-FTTCab-FTTC (Suburban FTTC (i)) and Suburban, FTTN-FTTC (Suburban FTTC (ii)) The calculated Cash flows of the four operator strategies are shown in **Fig. 3**b.

In a dense urban area, the initial coverage of high-speed ADSL and VDSL services is quite high due to the short average loop length compared to a suburban area. The difference between the investments for the FTTCab and FTTC strategy is balanced by the difference in revenues. The NPVs of the four rollout strategies, when using a traditional approach, are calculated as 13.1, 13.2, -0.1 and 1.6 MEURO respectively. The Suburban FTTC (i) would therefore be rejected. Moving to the discussion in the

introduction, we now turn to the assessment of the alternatives using the methodology of Real Options. First, the cash flows and their constituencies (revenues, investments etc.) are divided into two phases as in explained in [7][8][9]: the cash flows that stem from the initial service offering and the cash flows that result directly from a fibre upgrade. These phases will be denoted Phase 1 and Phase 2 respectively in the following discussion. An adjusted NPV for Phase 2 (and therefore the whole project) that includes the value of the flexibility in timing of the upgrade and the time value of money is calculated by the use of Black-Scholes [8] formula for European call option as described in [10]. The total project NPV is then calculated as the sum of the NPV of Phase 1 and the adjusted NPV of Phase 2. In order to find the value of the option of investment deferral, five parameters are required. These five parameters which are normally used in the calculation of financial call options are summarized in Table 1 along with their "interpretation" in an investment context.

Table 1 Analogy between investment opportunity and call option

Investment Opportunity	Variable	Call option
Present value of a project's operating assets to be acquired	S	Stock price
Expenditure to acquire the project's assets	X	Exercise price
Length of time the decision may be deferred	T	Time to expiration
Time value of money	r_f	Risk-free rate of return
Riskiness of the project assets	σ^2	Variance of returns on stock

The Suburban FTTC (i) strategy was used for the example. As seen, the adjusted NPV is positive compared to the negative NPV obtained from the "traditional" NPV method! The decision not to go for the investment is therefore changed to: invest in FTTCab now – then wait a few years – and then invest in FTTC! **Fig. 4** shows the traditional NPVs compared to adjusted NPVs with volatilities of 40% and 60% for the Suburban FTTC strategies:

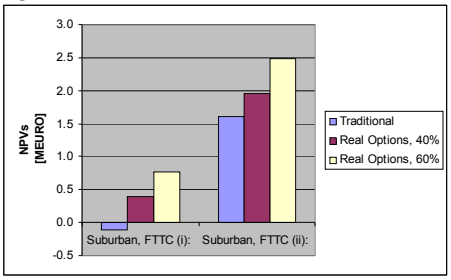


Fig. 4 NPVs for Suburban area strategies

In the FTTC (i) strategy, the decision is changed as already mentioned. For a volatility of 60% the NPV is 0.77 MEURO compared to the –0.1 MEURO using the

traditional approach! In the FTTC (ii) strategy, the change in NPV is 22% and 55% respectively for the two values of volatility. If a traditional NPV methodology had been used, the broadband project would not have been initiated at all with the given assumptions because the NPV of Phase 1 is negative. The uncertainty, here described by the volatility, always has a "good" and a "bad" side. If for example the demand for new services turn out to grow faster than expected, equipment prices drop faster than expected etc., investment is considered. Otherwise, the investment decision is deferred. In cases where the sign is not changed when using Real Options, significant improvements in NPV can still be obtained due to value of the built in option. In the FTTC (ii) case with volatility of 60%, the built in option of this project (the value of the call option minus the traditional Phase 2 NPV) is even exceeding the traditional NPV of the whole project.

A Game Theory modeling approach for 3G Operators

This case study presents the technoeconomic evaluation of a 3G rollout scenario followed by the identification of the market conditions for two operators in a simple game theory model. The considered scenarios reflect the point of view of both dominant operators and new entrants. Technoeconomic results are presented in terms of net present value (NPV), acting like the pay-off function in the proposed theoretical Game Theory model.

The calculations were based on two main inputs from the TE model results (case 1 of this paper) and the hypothetical (empirical) market parameters. TE model served as the main root for calculating inputs of our calculations. So we accepted all the suppositions built into the TE model, and different models for "incumbent" and "newcomer" were prepared. We followed the 10-years time horizon, and naturally all the investments, costs and revenue figures. All the other basic parameters were set to illustrate an "average western European case".
To be able to calculate the pay-off, two important parameters were picked up, namely:

- The market share and
- the price

Within the Eurescom P901 (EURESCOM 2001) an improved model for the definition of the market behavior have been proposed. A general "S-like" curve was supposed, with three (3) sort of market behavior regarding price-reaction of customers. This kind of function stands for both companies, but with different parameters. In our calculation market function is built from the newcomer operator (player 2) point of view. All the parameters refer to the newcomer. The Market Share of the new comer is given by:

$$MS = MS_{Start} + \Delta MSD_{Min} + \Delta MSD_{Max} \cdot e^{-e^{\left(a+b\frac{TM1-TM2}{\sqrt{1+TM2-TM1}}\right)}} \quad (4)$$

Where,
MS : Market share

MS_{Start} : Market share at the beginning, where decision is made

$\Delta MSD_{Min}, \Delta MSD_{Max}$: Minimum and maximum of market share changes (delta)

TM_1, TM_2 : Tariff multiplier of player 1 and player 2 respectively.

Analytically in Fig. 5 Data Flow between TE Tool and Game Theory Tables. the data flow between the Technoeconomic tool (3G operator model) and the Game theory tables are illustrated.

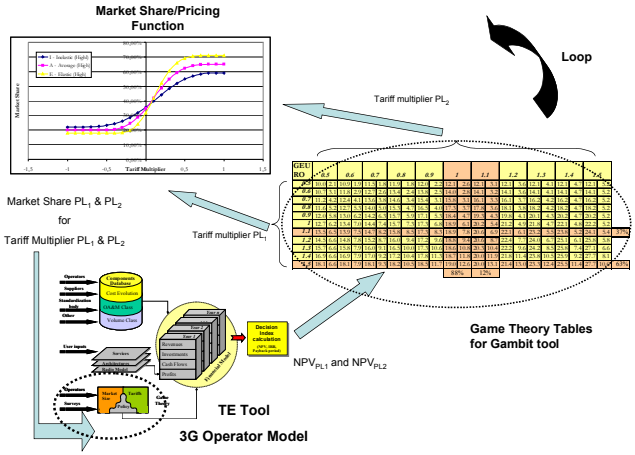


Fig. 5 Data Flow between TE Tool and Game Theory Tables.

For each point define by a specific tariff policy via the tariff multipliers exist one market share for each player (player1 – incumbent, player 2 – new comer). In the next step for each value of the tariff multiplier and market share one value of NPV per player can been calculated by the TE tool (NPV_{PL1} and NPV_{PL2}). These NPVs values are stored in a cell, consist as a step for the players for Gambit Tool. The process continues until all cells completed with NPVs values, (last step [TM_{PL1}, TM_{PL2}]=[1.5, 1.5]). These results are basically all the sensitivity results for all combinations. The process will be repeated for all defined market types Insensible-Small, Medium, High], [Average-Small, Medium, High] και [Sensible-Small, Medium, High], 9 times in total.

The following tables (in B€) show the results reached by calculating pay-off functions and in addition the Nash-equilibrium points are illustrated. Each cell contains two figures. The first column is the tariff multiplier of player 1 (PL_1) and the first row belongs to player 2 (PL_2). Each cell defined by the tariff multipliers contains two values. The first value is the NPV of Operator 1 (NPV_{PL1}) and the second is the NPV of Operator 2 (NPV_{PL2}) for a specific tariff policy. (i.e. First cell [0.5, 0.5] means 50% reduction in tariff compared to the base tariff for both players and NPV_{PL1}=9.9B€, NPV_{PL2}=2.2 B€.). The "position" of each player can be improved if its financial result

is higher. The crossings of the vertical and horizontal colored differently cells show the Nash-equilibrium points. The Nash-equilibrium points are bordered differently and so do strategies of competitors bound to the equilibrium. In case of having solutions with mixed strategies probabilities of specific strategies, these are put at the end of rows or columns (i.e. Table 2 82% and 12% probability). Important to notice that these strategy combinations and equilibrium points should be seen as being Nash-sense, and even if we might speak about strategy likely to be followed, the general meaning of this, is different from i.e. "dominant strategy".

Table 2: Pay-off matrix of HA version (High market share, Average market)

	0.5		*0.6*		*0.7*		*0.8*		*0.9*		*1*		*1.1*		*1.2*		*1.3*		*1.4*		*1.5*		
0.5	10.0	2.1	10.9	1.9	11.5	1.8	11.9	1.8	12.0	2.2	12.1	2.6	12.1	3.1	12.1	3.6	12.1	4.1	12.1	4.7	12.1	5.2	
0.6	10.7	3.1	11.8	2.9	12.7	2.6	13.4	2.4	13.8	2.5	14.0	2.8	14.1	3.2	14.1	3.6	14.1	4.1	14.1	4.7	14.1	5.2	
0.7	11.2	4.2	12.4	4.1	13.6	3.8	14.6	3.4	15.4	3.1	15.8	3.1	16.1	3.3	16.1	3.7	16.2	4.2	16.2	4.7	16.2	5.2	
0.8	11.6	5.2	12.7	5.3	14.0	5.0	15.3	4.7	16.5	4.0	17.3	3.7	17.8	3.6	18.1	3.8	18.2	4.2	18.2	4.7	18.2	5.2	
0.9	12.0	5.8	13.0	6.2	14.2	6.3	15.7	5.9	17.1	5.3	18.4	4.7	19.3	4.3	19.8	4.1	20.1	4.3	20.2	4.7	20.2	5.2	
1	12.7	6.2	13.4	7.0	14.4	7.4	15.7	7.3	17.3	6.8	18.9	6.1	20.2	5.4	21.2	4.9	21.8	4.7	22.1	4.8	22.2	5.2	
1.1	13.5	6.5	13.9	7.5	14.7	8.2	15.8	8.5	17.3	8.3	18.9	7.8	20.6	6.9	22.1	6.1	23.2	5.5	23.8	5.2	24.1	5.4	37%
1.2	14.5	6.6	14.8	7.8	15.2	8.7	16.0	9.4	17.2	9.6	18.8	9.4	20.6	8.7	22.4	7.7	24.0	6.7	25.1	6.1	25.8	5.8	
1.3	15.7	6.6	15.8	7.9	16.0	9.1	16.5	10.0	17.3	10.6	18.6	10.8	20.3	10.4	22.2	9.6	24.2	8.5	25.8	7.4	27.1	6.6	
1.4	16.9	6.6	16.9	7.9	17.0	9.2	17.2	10.4	17.8	11.3	18.7	11.8	20.0	11.9	21.8	11.4	23.8	10.5	25.9	9.2	27.7	8.1	
1.5	18.1	6.6	18.1	7.9	18.1	9.3	18.2	10.5	18.5	11.7	19.0	12.6	20.0	13.1	21.4	13.0	23.3	12.4	25.5	11.4	27.7	10.0	63%
											88%		12%										

Having Nash-sense equilibrium practically means both competitors play their best strategy, related to the other strategies selected. This can be done only in such a way, that players know each other's strategy in advance. Within limits, this could be a real situation. Formation of mixed strategies as solutions means that unambiguous or simple strategies do not exist in order to reach equilibrium. Of course, the reader might define other kind of equilibrium condition than Nash-sense, e.g. managers may be interested in strategy being "the best" independently of other players' strategy-selection, or what happens if all the other players work the ruin of the others. In the first case we can reach "dominant strategy" and in the second we would have in our hands "the safe strategy", both being a kind of equilibrium.

Falling back on our Nash-equilibrium points, and having mixed strategy solutions, the operators should play on a "statistical base", and choose strategies relying on numerical possibilities. As this seems to be rather "unrealistic", operators would likely follow a strategy of higher probability value. Quite independently of initial market share of the newcomer and type of market, newcomer has to have always smaller prices, as the market analyzer might feel that this is "natural". In most of our results this offers quite small tariffs compared to the incumbent operator's ones.

As initial market share of newcomer increases, both competitors should decrease their prices, but regarding our inelastic total market models, incumbent always has a chance of having "the highest" price, being viable at the same time (quite positive NPV). It is interesting that with increasing sensibility of the market the solutions exist only with mixed strategies. This means that the mutual impact of competitors on each other, assumes that players have investigated what the other player does, deeply into statistical decision making "process".

Conclusions

Concluding, investments in telecommunication technology can develop a new market area and expand traditional options for new players. Any business modelling should be accompanied by technoeconomic evaluation in order to give readers insights into the financial perspective and viability of a telecommunication investment project. In addition real options approach should be a complement to existing capital-budgeting systems and NOT a substitute of them. Game theory should be followed in order to have a better understanding of the competition as it effect on the financial perspective of the telecommunication projects.

References

[1] B. T. Olsen, et al., "Technoeconomic Evaluation of Narrowband and Broadband Access Network Alternatives and Evolution Scenario Assessment," *IEEE J. Sel. Areas Comm.*, vol. 14, no. 8, 1996.

[2] L.A. Ims, "Broadband Access Networks Introduction strategies and techno-economic evaluation, Telecommunications Technology And Applications Series, Chapman & Hall 1998, ISBN 0 412 828200

[3] IST-TONIC available on-line http://www-nrc.nokia.com/tonic/

[4] ACTS-TERA available on-line http://www.telenor.no/fou/prosjekter/tera/index.htm

[5] D. Katsianis *et al*, "The financial perspective of the mobile networks in Europe", *IEEE Personal Commun. Mag., Dec. 2001 Vol 8, No 6, pp 58-64.*

[6] Th. Monath, N.K. Elnegaard, Ph. Cadro, D. Katsianis and D. Varoutas, "Economics of fixed broadband access network strategies" *IEEE Comm. Mag.*, Sep 2003

[7] T. A. Luehrman, "Investment opportunities as real options: Getting started on the numbers", *Harvard Business Review*, vol. 76, (July-August), 1998, pp. 51-67.

[8] F. Black and M. Scholes, "The pricing of options and corporate liabilities", *Journal of Political Economy* no 81, pp.637-659

[9] A. K. Dixit and R. S. Pindyck, "Investment under uncertainty", *Princeton University Press*, Princeton, N.J., 1994, ISBN 0-691-03410-9

[10] J. C. Hull, "Options, futures, and derivatives" *Prentice Hall, (3rd ed.)*, 1997, ISBN 0-13-264367-7

Method for Predicting the Perceived Quality of Service for Digital Video as a Function of the Encoding Bit Rate and the Content Dynamics

Harilaos G. Koumaras*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
koumaras@di.uoa.gr

Abstract. This paper presents a novel method for fast and quantified estimation of the Perceived Quality of Service (PQoS) for MPEG-4 video content, encoded at constant bit-rates. Taking into account the instant PQoS variation due to the Spatial and Temporal (S-T) activity within a given MPEG-4 encoded content, this paper introduces the Mean PQoS (MPQoS) as a function of the video encoding rate and the picture resolution, and exploits it as a metric for objective video quality assessment. The validity of this metric is assessed by comparing PQoS experimental curves to the theoretical benefit functions vs. allocated resources. Based on the proposed metric, and taking into account the qualitative similarity between theoretical and experimental curves, the paper presents a prototype method for pre-encoding PQoS assessment based on the fast estimation of the S-T activity level of a video signal.

Keywords: Perceived Quality of Service (PQoS), Mean Perceived Quality of Service (MPQoS), Benefit function, Objective measurement of PQoS

1 Introduction

Multimedia applications that distribute audiovisual content over 3G/4G (3rd/4th generation) networks (such as video on demand (VOD) and real time entertainment streaming services) will be based on digital encoding techniques (e.g. MPEG-4 standard [6]), which achieve high compression ratios, by exploiting the spatial and temporal redundancy in video sequences. However, digital encoding causes image artifacts, which result in perceived quality degradation. Due to the fact that the parameters with strong influence on the video quality are normally those, set at the encoder (with most important the bit rate, the frame rate and the resolution), the issue of the user satisfaction in correlation with the encoding parameters has been raised.

* Dissertation Advisor: Dimitris Martakos, Assoc. Professor

One of the 3G/4G visions is the provision of audiovisual content at various quality and price levels [17]. There are many approaches to this issue, one being the Perceived Quality of Service (PQoS) concept. The evaluation of the PQoS for audiovisual content will provide a user with a range of potential choices, covering the possibilities of low, medium or high quality levels. Moreover the PQoS evaluation gives the service provider and network operator the capability to minimize the storage and network resources by allocating only the resources that are sufficient to maintain a specific level of user satisfaction.

The evaluation of the PQoS is a matter of objective and subjective evaluation procedures, each time taking place after the encoding process (post-encoding evaluation). Subjective quality evaluation processes of video streams (PQoS evaluation) require large amount of human resources, establishing it as a time-consuming process (e.g. large audiences evaluating video/audio sequences) [14]. Objective evaluation methods, on the other hand, can provide PQoS evaluation results faster, but require large amount of machine resources and sophisticated apparatus configurations. Towards this, objective evaluation methods are based and make use of multiple metrics [18], which are related to the content's artifacts (i.e. tilling, blurriness, error blocks, etc.) resulting from the encoding process [19].

This paper presents a quantified PQoS assessment method for MPEG-4 video encoded sources, which provides pre-encoding PQoS estimation based on a single metric experimentally derived from the Spatial and Temporal (S-T) activity level of a given video content. The pre-encoding nature of the proposed method alleviates both the machine resource requirements and the time consumption of the already existing post encoding methods, making PQoS evaluation quick, easy and economically affordable for 3G/4G commercial implementations.

Towards this, a quality meter tool was used [9], providing objective PQoS results (based on multiple metrics) for each frame within a video clip. Initially, such objective PQoS results were obtained for a short homogeneous MPEG-4 video of specific encoding parameters (i.e. encoding bit-rate, resolution). The graphical representation of these results vs. time, demonstrated the instant PQoS of each frame within the video clip, besides indicating the Mean PQoS (MPQoS) of the entire video (for the whole clip duration). Similar experiments were conducted for the MPQoS calculation of the same video content, each time applying different encoding parameters. The results of these experiments were used to draw-up experimental curves of the MPQoS of the given video content, as a function of the encoding parameters. The same procedure was applied for a set of video sequences, each one with different S-T activity level. Comparison of these experimental curves with those resulting from the theoretical algebraic benefit functions [10], [16] indicated a qualitative similarity among them, proving therefore the validity of the MPQoS as a metric for objective quality evaluation. A generalized approach to the above theoretical model is given in [10], where the algebraic benefit function is used to represent the user satisfaction in correlation with the allocated resources of competitive multimedia services. The term benefit function was introduced in [16] and represents the grade of the user satisfaction resulting from the use of a specific set of QoS and resource parameters.

Furthermore, this paper shows that the experimental MPQoS vs. bit rate curves can be successfully approximated by a group of exponential functions, which confine

the QoS characteristics of each individual video test sequence to three parameters that form the Quality Vector (QV) of the specific clip. Showing that these parameters are correlated, it can be concluded that the experimental measurement of just one of them, for a given short video clip, is sufficient for the determination of the other two. In this way, a single measurement of the MPQoS is sufficient for the analytical determination of the MPQoS vs. Bit rate curve for a given video clip. As a result, the proposed metric can be also used as a criterion for pre-encoding decisions, concerning the encoding parameters to be set for satisfying a certain PQoS, in respect to a given S-T activity level of a video sequence.

Following this introductory section, the rest of the paper is organised as follows: In section II the bibliographic background of the PQoS evaluation is presented. In section III the Perceived Quality Meter tool is presented, while Section IV describes the variation of the MPQoS (obtained by the quality meter tool) as a function of the encoding bit rate. Section V presents the exponential approximation of the MPQoS vs. Bit rate curves, and Section VI describes the proposed method for objective PQoS evaluation based on a single metric. Section VII tests the proposed method on non-homogeneous media clips and finally, section VIII concludes the paper.

2 BACKGRROUND & RELATED WORK

Over the last years, with the increased popularity of multimedia applications (i.e. video on demand, streaming services, multimedia conference), emphasis has been put on developing methods and techniques for evaluating the perceived quality of video content.

The methods and techniques that have been proposed in the bibliography can be sorted into two groups:

- The assessment methods that their scope is the determination of the encoding settings (i.e. resolution, frame rate, bit rate), which are required in order to carry out successfully the communication task of a multimedia application (i.e. video conference). In other words, the scope of these methods is the estimation of the adequate video quality level for a particular multimedia communication task.
- The assessment methods that their aim is the evaluation of the quality level of a media clip based on the detection of artifacts on the signal caused by the encoding process. In contrast with the methods of the previous category, the scope of these methods is not the determination of the adequate level, but the classification of a video content at a perceived quality scale.

The methods of the first group in order to determine the adequate quality level for a specific multimedia application, take under consideration a great number of parameters and metrics that depend on the task nature and the user emotional behavior [12]. For example the classification of the task as foreground or background in correlation with its complexity [3], is a parameter that differentiates the quality demands of a multimedia application. On the other hand, the emotional content of a multimedia

communication task alters the required quality level of the specific communication service [13]. Due to this, various parameters are measured in order to estimate the appropriate minimum quality level of a multimedia application. Such parameters are:

- The user characteristics (i.e. knowledge background, language background, familiarity with the task, age)
- The situation characteristics (i.e. geographical remoteness, simultaneous number of users, distribution of users)
- The user cost (i.e. heart rate, blood volume pulse)
- The user behavior (i.e. eye tracking, head movement)

However, these methods have still some issues to solve on technical, theoretical and practical level. A user that participates in such an assessment procedure is wired at so many points on the body (even on the head may wear the eye tracking equipment), which causes uncomfortable feelings and affects its behaviour. Technical issues, such as the eye tracking loss and the manual calibration/correction by a human operator, affect the reliability of the methods in real time environments [12].

The methods of the second group, which aim at ranking the video quality of a media clip based on the detection of visual artifacts caused by the encoding process, are mainly categorized into two classes: The subjective and objective ones.

The subjective test methods, which have mainly been proposed by International Telecommunications Union (ITU) and Video Quality Experts Group (VQEG), involve an audience of people, who watch a video sequence and score its quality as perceived by them, under specific and controlled watching conditions. Afterwards, the statistical analysis of the collected data is used for the evaluation of the perceived quality. The Mean Opinion Score (MOS) is regarded as the most reliable method of quality measurement and has been applied on the most known subjective techniques: The Single Stimulus Continue Quality Evaluation (SSCQE) and the Double Stimulus Continue Quality Evaluation (DSCQE) [7], [1], [14]. However the MOS method is inconvenient due to the fact that the preparation and execution of subjective tests is costly and time consuming and its implementation today is limited to scientific purposes, especially at VQEG experiments.

For this reason, a lot of effort has recently been focused on developing cheaper, faster and easier applicable objective evaluation methods. These techniques successfully emulate the subjective quality assessment results, based on criteria and metrics that can be measured objectively. The objective methods are classified, according to the availability of the original video signal, which is considered to be in high quality.

The majority of the proposed objective methods in the literature requires the undistorted source video sequence as a reference entity in the quality evaluation process, and due to this are characterized as Full Reference Methods [18], [26]. These methods are based on an Error Sensitivity framework with most widely used metrics the Peak Signal to Noise Ratio (PSNR) and the Mean Square Error (MSE).

$$\mathrm{PSNR} = 10\log_{10}\frac{L^2}{\mathrm{MSE}},$$

where L denotes the dynamic pixel value (i.e. equal to 255 for 8bits/pixel monotonic signal) (1)

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 ,$$

where N denotes the number of pixels, and x_i /y_i the i^{th} pixel value in the original/distorted signal (2)

However, these overused metrics have seriously been criticized that they do not provide reliable measurements of the perceived quality [21]. For this reason, a lot of effort has been focused on developing assessment methods that emulate characteristics of the Human Visual System (HVS) [25], [4], [2], [8] using Contrast Sensitivity Functions (CSF), Channel Decomposition, Error Normalization, Weighting and finally Minkowski error pooling for combining the error measurements into a single perceived quality estimation. An analytical description of the framework, which these methods use, can be found in [22].

However it has been reported [21], [20] that these complicated methods do not provide more accurate results than the simple mathematical measures (such as PSNR). Due to this some new full reference metrics that are based on the video structural distortion, and not on error measurement, have been proposed [23], [24].

On the other hand, the fact that these methods require the original video signal as reference deprives their use in commercial video service applications, where the initial undistorted clips are not accessible. Moreover, even if the reference clip is available, then synchronization predicaments between the undistorted and the distorted signal (which may have experienced frame loss) make the implementation of the Full Reference Methods difficult and impractical.

Due to these reasons, the recent research has been focused on developing methods that can evaluate the PQoS level based on metrics, which use only some extracted structural features from the original signal (Reduced Reference Methods) [5] or do not require any reference video signal (No Reference Methods) [11], [9].

However, due to the fact that the 3G/4G vision is the provision of audiovisual content at various quality and price levels [17], there is great need for developing methods and tools that will help service providers to predict quickly and easily the PQoS level of a media clip. These methods will enable the determination of the specific encoding parameters that will satisfy a certain quality level. All the aforementioned post-encoding methods may require repeating tests in order to determine the encoding parameters that satisfy a specific level of user satisfaction. This procedure is time consuming, complex and impractical for implementation on the 3G/4G multimedia mobile applications.

In this context, this paper presents a novel objective evaluation method, which will enable the pre-encoding estimation of the PQoS level for MPEG-4 coded video clips, alleviating therefore the time and procedure requirements of the already existing methods.

3 PERCEIVED QUALITY METER TOOL

A software implementation, which is representative of the non-reference objective evaluation class, is the Quality Meter Software (QMS) that was used in this paper [9]. The QMS tool measures objectively the instant PQoS level (in a scale from 1 to 100) of digital video clips. Since it belongs to the non-reference class, its use is quick and easy. The evaluation algorithm of the QMS is based on vectors, which contain information about the averaged luminance differences of adjacent pixels.

The high compression during the MPEG-4 encoding process, results in loss of high frequency Discrete Cosine Transformation (DCT) coefficients. Within an MPEG-4 block (8x8 pixels), the luminance differences and discontinuities between any pair of adjacent pixels are reduced, by the encoding and compression process. On the contrary, for all the pairs of adjacent pixels, which are located across and on both edge sides of the border of adjacent DCT blocks, the luminance discontinuities are increased, by the encoding process.

More specifically, the average luminance differences of the previously referred pixel pairs depend on the encoding parameters (mainly on the bit rate). This means that low bit rate results in significant tiling of the video clip, which finally causes PQoS degradation. Based on this fact, the QMS tool uses these luminance differences as an objective metric.

The average luminance $L(x, y)_{average}$ of a pixel, having plane coordinates (x,y), can be computed by the surrounding K x K adjacent pixels [18] using the following equation (3):

$$L(x, y)_{average} = \frac{1}{KxK} \sum_{i=-K/2}^{K/2} \sum_{j=-K/2}^{K/2} L(x+i, y+j) \qquad (3)$$

In the case of QMS, the above equation is specialized setting K equal to 2, which results in taking in consideration the luminance values of the first neighbouring pixels only.

The validity of the specific QMS has been assessed by comparing quality evaluation results, derived from the QMS, to corresponding subjective quality assessment results, which were deduced by a Single Stimulus Continues Quality Evaluation (SSCQE) test procedure. This comparison [9] proved that the QMS tool, despite the fact that it is based on a simple algorithm, emulates successfully the corresponding subjective quality assessment test.

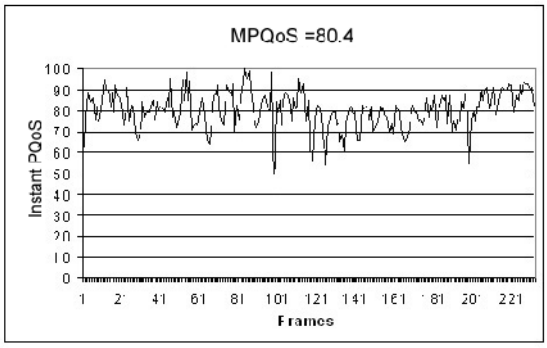


Figure 1. The instant PQoS of the Mobile & Calendar clip with CIF resolution derived from the Quality Meter Software

Figure 1 depicts an example measurement of the instant PQoS, derived from the specific QMS for the well known video clip Mobile & Calendar, which was encoded using the MPEG-4 standard (Simple Profile) at 800 Kbps (Constant Bit Rate) with Common Interface Format **(**CIF) resolution, key-frame period equal to 100 frames and 25 frames per second (fps). The instant PQoS vs. time curve (where time is represented by the frame sequence) varies according to the S-T activity of each frame. For frames with high complexity the instant PQoS level drops, while for frames with low S-T activity the instant PQoS is higher. Such instant PQoS vs. time curves, derived by the above QMS, can be used to characterize and categorize a short video clip according to its content. Introducing the concept of the Mean PQoS (MPQoS), the average PQoS of the entire video sequence, over the whole duration of a short clip, can be used as a metric for ranking it into a perceived quality scale.

$$\text{MPQoS} = \frac{\sum_{i=1}^{N} \text{Instant PQoS}_i}{N}, \quad \textit{where N denotes the total frames of the test signal} \tag{4}$$

For example, considering three quality categories, defined as low, medium and high (the corresponding ranges can be set at 70-80, 80-90 and 90-100), the video clip with instant PQoS curve being that of figure 1, has MPQoS equal to 80.4 and can be ranked and categorized as a medium quality video clip. The limits of the quality categories can be specified according to the needs of the service provider.

4 VARIATION OF MPQoS AS A FUNCTION OF THE ENCODING BIT RATE.

In order to specify the variation of the MPQoS vs. the encoding bit rate and the Spatial and Temporal activity level (as is indicated by the graphical representation of the instant PQoS vs. time derived from the QMS software tool), four short in duration test sequences, which are representative of specific Spatial and Temporal activity levels, were used. These well known video clips are shown in table 1.

Clip 1	Low Spatial & Temporal Activity Level	Suzie
Clip 2	| Medium Spatial & Temporal Activity Level	Cactus
Clip 3		Flower Garden
Clip 4	| High Spatial & Temporal Activity Level	Mobile & Calendar

Table 1. The test video sequences

The Spatial and Temporal (S-T) activity level of a video clip is crucial for the encoding efficiency and the achieved perceived quality, because video coding methods exploit both temporal and spatial redundancy in order to achieve compression of the video data. Due to the fact that temporally adjacent frames are quite similar and therefore highly correlated (temporal correlation), the video encoder attempts to compress video data by exploiting this temporal redundancy. In the spatial domain, the encoder exploits the high correlation between neighbouring pixels (spatial correlation), and makes prediction of them based on neighbouring samples. [15]

Therefore, in this paper the term Spatial and Temporal Activity level is used in order to express the dynamics of the video content, which affect the correlation level on the Spatial and Temporal domain. Media clips with static content (i.e. talk shows, debates etc.) have low Spatial and Temporal activity level in contrast with media clips with active, quick and complex scenes (i.e. sport events, action scenes), which correspond to high Spatial and Temporal activity level. The test signals of Table 1 cover a wide range of the Spatial and Temporal activity scale.

For the experimental needs of this paper, each test video clip of Table 1, was transcoded from its original MPEG-2 format at 12 Mbps with PAL resolution and 25 fps to ISO MPEG-4 (Simple Profile) format, at different constant bit rates (spanning a range from 50kbps to 1.5Mbps for CIF (Common Intermediate Format) and 20kbps to 800kbps for QCIF (Quarter Common Intermediate Format), with key-frame period equal to 100 frames in both cases). For each corresponding bit rate, a different ISO MPEG-4 compliant file with CIF resolution (352x288) and QCIF resolution (176x144) respectively was created. The frame rate was set at 25 frames per second (fps) for the transcoding process in all test videos.

Each ISO MPEG-4 video clip was then used as input in the QMS tool. From the resulting instant PQoS vs. time graph (like the one in figure 1), the MPQoS value of each clip was calculated. This experimental procedure was repeated for each video clip in CIF and QCIF resolution.

The results of these experiments for the test signals with CIF resolution are depicted in figure 2, where PQ_L denotes the lowest acceptable MPQoS level (corresponding to 70 in the scale from 1 to 100 for CIF resolution) and PQ_H denotes the best MPQoS level that each video can reach. Respectively, figure 3 depicts the results for the test sequences with QCIF resolution, where PQ_L corresponds to 40 in the hundred scale (the PQ_L value in the QCIF case corresponds to approximately 40% quality degradation comparing to the PQ_L value of the CIF case, because of the lower resolution). Comparing the curves of figures 2 and 3, it is deduced that lower resolution (QCIF) results in MPQoS curves that reach faster and at lower bit rates their PQ_H values, which are degraded in comparison with the corresponding PQ_H values of higher (CIF) resolution curves.

Referring to the curves of figure 2 (or 3), the following remarks can be made:

1. The minimum bit rate (BR_L), which corresponds to the lowest acceptable MPQoS level (PQ_L), depends on the S-T activity level of the video clip.

2. The variation of the MPQoS vs. bit rate is an increasing function, but non linear. Moreover, the quality improvement of an encoded video clip is not significant for bit rates higher than a specific threshold. This threshold depends on the S-T activity of the video content.

Comparing the experimental curves of figures 2 and 3 to those resulting from the theoretical algebraic benefit functions, described in [10], qualitative similarity among them is noticed. Thus, the experimental curves, which were derived from the QMS tool, are qualitatively very similar to the theoretically expected, proving therefore their validity. A quantitative comparison is not possible, because benefit function is very general and refers to a number of different parameters in both the horizontal and vertical axes. Mapping user satisfaction and allocated resources of the general algebraic benefit function model to MPQoS level and encoding bit rate respectively, the experimental curves offer a quantitative approach of the theoretical ones, which can be useful in practical and commercial applications.

Moreover, it is of great importance (based on the above mapping) the fact that the algebraic benefit function is not identical for all the types of audiovisual (AV) content, but it comprises a set of curves that follow the same basic shape. This provides a multi-dimensional characteristic to the benefit function. The differentiation among these curves comes from their slope and position on the benefit-resource plane, which depend on the S-T activity of the video content. Thus, the curve has low slope and transposes to the lower-right area of the benefit-resource plane, for AV content of high S-T activity. On the contrary, the curve has high slope and transposes to the upper-left area, for low S-T activity content.

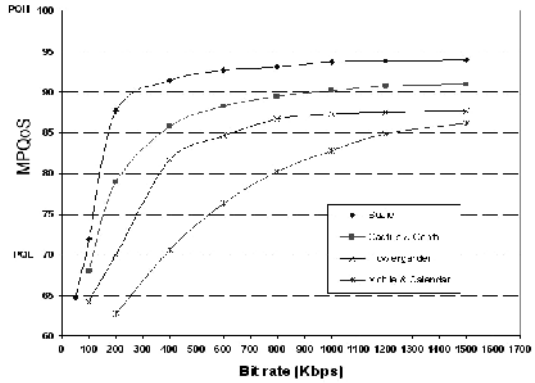


Figure 2. The MPQoS vs. Bit rate curves for CIF resolution

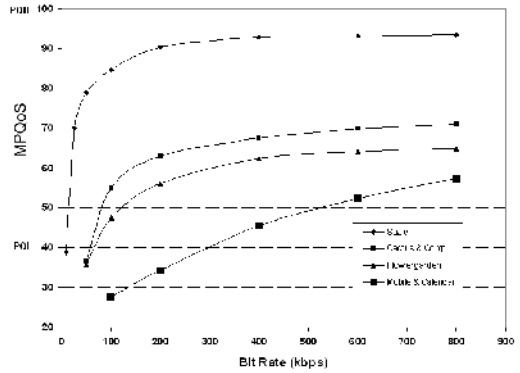


Figure 3. The MPQoS vs. Bit rate curves for QCIF resolution

Practically, the transposition of the curve to the upper-left area means that content with low S-T activity (e.g. a talk show) reaches a better PQoS level at relatively lower bit rate in comparison with a video content with high S-T activity. In addition, when the encoding bit rate decreases below a threshold, which depends on the video content, the PQoS practically "collapses". On the other hand, the transposition of the curve to the lower-right area means that content with high S-T activity (e.g. a football match) requires higher bit rate in order to reach a satisfactory PQoS level. Nevertheless, it reaches its maximum PQoS value more smoothly than in the low S-T activity case.

In this context, the MPQoS vs. Bit rate curves were also drawn for a set of media clips, which were captured from common television programs in DV (Digital Video) PAL format and encoded at CIF resolution following exactly the same encoding procedure as described previously. The video clips had relatively homogeneous content (i.e. talk show, football, swimming, speech etc.) with duration spanning from 15

seconds up to 60 seconds. Performing numerous experiments, it was deduced that the shape of the derived MPQoS vs. Bit rate curves was similar for a specific content (independently of the clip duration), having maximum matching error in all cases below 3%.

Moreover, according to the S-T level of each real media clip, the derived MPQoS vs. Bit rate curves followed the corresponding shape and inclination of the reference curves of figure 2. Therefore, real video clips with low S-T activity level (i.e. talk shows) produced curves similar to the one derived from Suzie test signal, while clips with high S-T activity level (i.e. football and sports) produced curves similar to the one derived from Mobile & Calendar clip. Figure 4 and 5 illustrate the experimental results of two representative real-captured clips (talk-show and football).

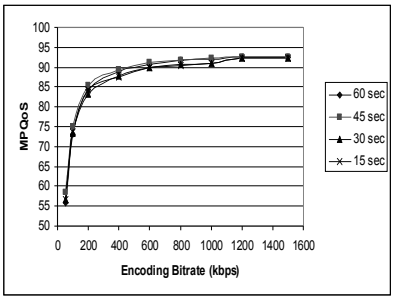


Figure 4. The MPQoS vs. Bit rate curves for real video clips with talk show content.

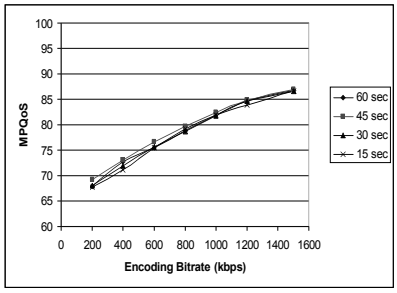


Figure 5. The MPQoS vs. Bit rate curves for real video clips with football/sports content.

Therefore, it can be deduced that the MPQoS provides sufficiently distinguishable curves also for real video clips according to their content, independently of their duration. For feature films it is not suggested the use of the proposed MPQoS metric, because the concept of the mean quality for so long media is pointless. Given that in 3G/4G mobile communication systems, the offered multimedia clips will not endure more than a couple of minutes, the proposed method can be valuable in 3G/4G mobile communication applications and services.

5 EXPONENTIAL APPROXIMATION OF MPQoS vs. BIT RATE CURVES.

Referring to figures 2 and 3, each MPQoS vs. bit rate curve can be described by the following three parameters:

(a) The minimum bit rate (BR_L) which corresponds to the lowest acceptable PQoS value (e.g. 70 for CIF)

(b) The highest reached PQoS level (PQ_H)

(c) A parameter α that defines the shape and subsequently the slope of the curve.

In [10] it is proposed that the QoS characteristics of a specific multimedia service can be described by a Quality of Service Vector (QoS Vector). So, each application is specified by a QoS Vector = (q_1, q_2,…,q_n), which can be used for determining the necessary resource allocation that corresponds to a specific level of user satisfaction. Adapting the general approach of the QoS Vector to the needs of this paper, a Quality Vector (**QV**) can be defined as :

$$\mathbf{QV} = (\alpha, BR_L, PQ_H) \qquad (5)$$

The experimental curves of figure 2 (or 3) can be approximated by a group of exponential functions. In this respect, the MPQoS level of a MPEG-4 video clip, encoded at bit rate BR, can be analytically estimated by the following equation:

$$MPQoS = [PQ_H - PQ_L] (1 - e^{-\alpha [BR-BRL]}) + PQ_L, \alpha>0 \text{ and } BR>BR_L \qquad (6)$$

where the parameter α is the time constant of the exponential function, which determines the shape of the curve.

Since the maximum deviation error between the experimental and the proposed exponential MPQoS curves was measured to be less than 4% in the worst case (for all the test signals), the proposed exponential model of MPQoS vs. bit rate can be considered that approximates successfully the corresponding experimental curves.

So each **QV** contains the QoS parameters, which are necessary for describing analytically the dependence of the MPQoS level on the encoding bit rate and subsequently the resolution, according to the proposed exponential approximation model. The experimental curves of figures 2 and 3 can be approximated successfully by specific **QV**s, which are shown in Table 2. Furthermore, PQ_L =70 for CIF resolution and PQ_L =40 for QCIF is assumed.

Experimental curves of MPQoS vs. bit rate and their corresponding exponential approximations were compared not only for the above four reference video clips, but also for non-reference AV content. For this purpose, short video clips of 30 second duration (approximately), were captured from common TV programs in DV PAL format and encoded according to MPEG-4 standard, following again the same experimental procedure that was described in Section III. The AV content varied from talk shows to sport events, spanning a wide range of S-T activity. The results showed that the experimental curves of MPQoS vs. bit rate were successfully approximated by exponential functions, with a deviation error less than 4%. Moreover, the element values of the corresponding **QV**s were in the range of those in table 2.

Test Sequence	α	BR_L (Kbps)	PQ_H (Quality Units)
Suzie (MPEG-4 CIF)	0.0083	95	93.91
Cactus (MPEG-4 CIF)	0.0063	110	90.89
Flower (MPEG-4 CIF)	0.0056	200	87.62
Mobile (MPEG-4 CIF)	0.0045	400	86.20
Suzie (MPEG-4 QCIF)	0.013	22	93.50
Cactus (MPEG-4 QCIF)	0.007	55	71.04
Flower (MPEG-4 QCIF)	0.006	65	64.79
Mobile (MPEG-4 QCIF)	0.005	300	57.32

Table 2. Quality Vector elements that correspond to test sequences for CIF and QCIF cases

6 FAST EVALUATION OF THE QV ELEMENTS

The accurate determination of the bit rate that results in a desired MPQoS level enables the better utilization of the storage capacity and also of the bandwidth allocation during the transmission of AV content. Due to the fact that the specified encoding bit rate is exactly the one that corresponds to a certain quality level, there is no waste in the storage or bandwidth resources. part from this, methods for estimating the variation of MPQoS vs. bit rate are very important to the 3G/4G mobile communication systems, because they help towards the evolution of a consumer mass market, where the service provider will offer AV content at various quality levels, among which the consumer will be able to choose the one, at which he/she prefers to watch it.

Practically, in order to achieve this, and given a short video clip, first it must be categorized according to its content. Afterwards, it must be encoded at the appropriate bit rates that satisfy the diverse perceived quality levels and finally stored in a server. Today, the determination of the bit rates, which correspond to the various quality levels, can be achieved only by multiple repeating post-encoding measurements of the MPQoS at various bit rates. Since this is a complicated and time consuming process, an alternative simple and fast pre-encoding evaluation method is proposed, based on the use of the **QV** elements (BR_L, PQ_H and α) of a specific video clip.

Furthermore, showing that these elements are correlated, the evaluation/determination of only one of them for a given video clip is sufficient to accurately determine the other two and ultimately deduce the corresponding exponentially approximated PQoS vs. Bit rate curve. he correlation among the three **QV** elements can

be derived experimentally. Considering the four test video clips of table 1, which cover a wide range of S-T activity level, the variation of their **QV** elements vs. the S-T activity level is depicted in figure 6, for the case of MPEG-4 (Simple Profile) and CIF resolution. Similar curves can be derived for the case of QCIF resolution. According to figure 6, it is obvious that there is interdependence between the elements, so if one out of the three **QV** elements is specified for a given video clip, then the other two can be accurately determined.

Among the three elements, PQ_H is the most convenient to be experimentally calculated, given that the variation of MPQoS vs. Bit rate is exponentially approximated. Using the QMS tool, which was described in Section III, one only measurement/estimation of the MPQoS at a high encoding bit rate is sufficient for the accurate determination of the PQ_H value for a given video clip. The consequent steps are simple: Using the estimated PQ_H value and the reference curves of figure 6, the corresponding values of BR_L and α can be graphically extrapolated. Thus, having defined the three **QV** elements, the analytical exponential expression of the MPQoS vs. Bit rate can be deduced using equation (6).

In order to succeed an analytical approach, the experimental dependence of Parameter α, BR_L and PQ_H on the S-T level (figure 6) can be successfully described by power series of the polynomial form $\sum_{k=0}^{\infty} b_k x^k = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \ldots$, where b_0, b_1,… are real number constants. For the purposes of this paper, only the first three terms of the power series are used, which provide a satisfying degree of accuracy for the approximation of the experimental data (bold curves of figure 6).

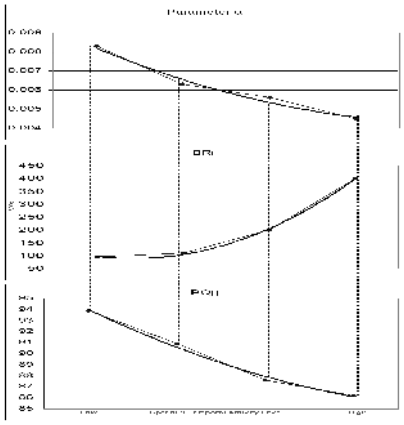


Figure 6. The variation of the **QV** elements

$$\begin{cases} \text{Parameter\_}\alpha(x) = \sum_{k=0}^{\infty} b_k x^k = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \ldots \approx 0.0103 - 0.0023x + 0.0002x^2 & (7) \\ BR_L(x) = \sum_{k=0}^{\infty} c_k x^k = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + \ldots \approx 181.25 - 130.75x + 46.25x^2 & (8) \\ PQ_H(x) = \sum_{k=0}^{\infty} d_k x^k = d_0 + d_1 x + d_2 x^2 + d_3 x^3 + \ldots \approx 98.255 - 4.64x + 0.4x^2 & (9) \\ \text{where x is related to the S-T activity level of the media clip} \end{cases}$$

By this way, the complexity of the proposed expressions (7), (8) and (9) is maintained in low level, making possible the practical use of them. As described previously, one only measurement/estimation of the MPQoS (using the QMS tool), at a high encoding bit rate is enough for the accurate determination of the PQ_H value for a given video clip. Substituting the measured PQ_H value in equation (9), the corresponding x variable can be accurately calculated, by solving this equation. From the two roots, the smaller positive one is accepted and used as input to the other two equations (8) and (7), from where the BR_L and Parameter α can be accurately calculated. Thus, having defined the triple elements (Parameter α, BR_L, PQ_H), the analytical exponential expression of the MPQoS vs. Bit rate can be deduced using equation (6), enabling the pre-encoding MPQoS evaluation for the specific video clip.

Variable x is strongly related to the S-T activity level of the test signal. It was experimentally measured that as x increases, S-T activity level increases, too. From the couples of the roots derived from equation (9), the lower ones are analogous to the S-T activity level, while the higher ones are reverse analogous. So, the lower ones are retained and further used, in order to achieve agreement with the experimental measurements.

7 TEST OF THE METHOD ON NON-HOMOGENEOUS CONTENT

Multimedia applications of 3G/4G mobile communication systems will be based on the provision of short in duration video content at various quality and price levels, among which the consumer will be able to choose. The already described proposed method enables the pre-encoding estimation and determination of the encoding parameters that satisfy a specific PQoS level. This section tests the proposed method on non-homogeneous content.

Due to the fact that it is difficult to capture real video clips that are representative of various non-homogeneous levels, processed media clips with controlled level of non-homogeneity, were created using two real captured sequences with contrary content and S-T level: A talk show and an active scene from a football game. The non-homogeneous media clips were created using interchanging portions of these two sequences with specific ratio of high and low S-T level. Table 3 epicts the characteristics of the media clips that were derived by this procedure.

Clips	Total Talk Show Duration (sec)	Total Football Duration (sec)	Ratio Talk/Football
Clip 1	120	0	∞
Clip 2	105	15	7.00
Clip 3	90	30	3.00
Clip 4	75	45	1.67
Clip 5	60	60	1.00
Clip 6	45	75	0.60
Clip 7	30	90	0.33
Clip 8	15	105	0.14
Clip 9	0	120	0.00

Table 3 The characteristics of the non-homogeneous media clips

The two real captured sequences (talk show and football) were edited in their original format (DV PAL) in order to produce the final non-homogeneous clips of Table 4. Afterwards, the edited DV clips were encoded with ISO MPEG-4 (Simple Profile) format, at different constant bit rates (spanning a range from 200kbps to 1.5Mbps and key-frame period equal to 100 frames). For each corresponding bit rate, a different ISO MPEG-4 compliant file with CIF (Common Intermediate Format) resolution (352x288) was created. The frame rate was set at 25 frames per second (fps) for all the test signals.

Each ISO MPEG-4 video clip was then used as input in the QMS tool. From the resulting instant PQoS vs. time graph (like the one in figure 1), the MPQoS value of each clip was calculated, following exactly the same procedure, like the one that was described in section IV. The derived MPQoS vs. bit rate curves of this procedure are depicted on figure 7 and are very similar to the reference curves of figure 2, considering similar level of S-T activity level.

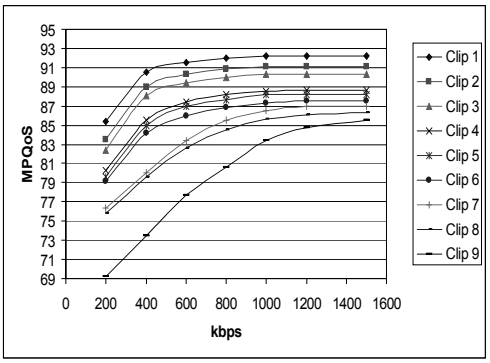


Figure 7. The experimental MPQoS vs. Bit rate curves for non-homogeneous media clips

Clip Name	Mean Error %
Clip 1	0.434
Clip 2	0.708
Clip 3	0.942
Clip 4	1.884
Clip 5	0.836
Clip 6	1.062
Clip 7	1.934
Clip 8	4.364
Clip 9	0.718

Table 4 Mean errors of the predicted MPQoS vs. Bit rate curves for non-homogeneous media clips

Afterwards for each clip of table 3 the proposed technique was applied. The derived estimated MPQoS vs. Bit rate curves were compared to the experimental ones of figure 7 and table 4contains the corresponding mean error for each clip.

According to Table 5, it is shown that the proposed method predicted successfully the MPQoS vs. bit rate curves, even for media clips with non-homogeneous video content with a worst case mean error equal to 4.364%. Therefore, the proposed method is valid and provides reliable results also for video clips with non-homogeneous content.

Moreover, the proposed technique was also tested on a set of 20 real captured video clips, containing various non-homogeneous video contents, with duration spanning from 2 minutes up to 10 minutes. These video clips were captured in DV PAL format from common TV programs. Following again the same encoding procedure as previously, ISO MPEG-4 compliant files were produced for each real captured DV test clip. Afterwards, the experimental and theoretical (according to the proposed method) MPQoS vs. Bit rate curves were derived for each media clip.

The worst case mean error between the experimentally and theoretically derived MPQoS curves for the twenty real captured videos was measured to be equal to 4.08%. This error is lower than the worst case error (4.364%) of the specially edited non-homogeneous media clips, proving that the proposed method can be also applied successfully on real video clips with non-homogeneous content.

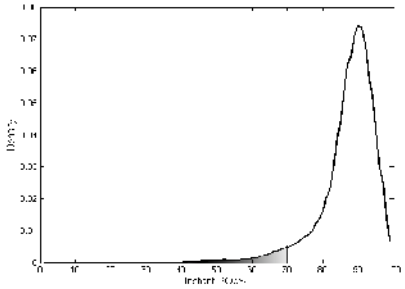


Figure 8. The PDF of Instant PQoS for non-homogeneous clips

In order to examine the variation of the instant PQoS for non-homogeneous media clips, the Probability Density Function (PDF) for all the non-homogeneous clips of Table 4 was drawn. Figure 8 illustrates the corresponding PDF, where it is observed that the values of instant PQoS are highly concentrated around the MPQoS (i.e. 86.93), with a Standard Deviation equal to 8.7. The probability of unacceptable quality, i.e. instant PQoS values below 70 (grey area in figure 8), is approximately equal to 0.0375. Therefore, the fluctuation of instant PQoS around MPQoS does not significantly affect the accuracy of the MPQoS metric. However, it must be noted that in the proposed method the MPQoS is used for bit rates that generate relatively high/accepted PQoS levels (i.e. MPQoS>70). The case of lower encoding bit rates, which correspond to low/unaccepted MPQoS values, is not examined in this paper, because such low quality levels are not commercially worthy and are not expected to be offered in the upcoming 3G/4G services.

8 CONCLUSIONS

Existing hardware/software perceived quality meters provide post encoding measurements of instant PQoS vs. time variation for a video content. In this paper, the mean PQoS (MPQoS), for the whole duration of a video clip, is proposed as a metric that characterizes a video clip as a single entity. Experimental MPQoS vs. Bit rate curves (derived from experimental measurements of the instant PQoS) compared qualitatively to the theoretical curves of benefit function vs. allocated resources, showing similarity in their shape and therefore proving the validity of the experimental ones. Furthermore, a mapping of the user satisfaction and allocated resources of the theoretical benefit function model to MPQoS and bit rate respectively, reveals that the algebraic benefit function is not identical for all the types of AV content. Instead of this, the benefit function is a multi-dimensional entity, which can be analyzed in a set of curves, all following the same basic shape. This differentiation depends on the S-T activity level of the video content.

Moreover, the experimental MPQoS curves can be successfully approximated by a group of exponential functions, with a deviation error of less than 4%. This enables the analytical description of the MPQoS dependence on the encoding bit rate. Based on this, a method for fast pre-encoding estimation of the MPQoS level of a video clip is proposed, which allows the ranking of the clip according to the S-T activity of its content, enabling an optimized utilization in the corresponding storage and bandwidth resources.

References

[1] Th. Alpert and L. Contin, “DSCQE Experiment for the Evaluation of the MPEG-4 VM on Error Robustness Functionality”, ISO/IEC – JTC1/SC29/WG11, MPEG 97/M1604, 1997.
[2] A. P. Bradley, “A Wavelet Difference Predictor”, IEEE Transactions on Image Processing. Vol. 5, pp. 717-730, 1999.

[3] W. Buxton, “Integrating the periphery and context: A new taxonomy of telematics”, in Proceedings of Graphics Interface 1995, pp. 239-246, 1995.
[4] S. Daly, “The Visible Difference Predictor: An algorithm for the Assessment of Image Fidelity”, in Proceedings SPIE, Vol. 1616, pp. 2-15, 1992.
[5] I. Pr. Guawan and M. Ghanbari, “Reduced-Reference Picture Quality Estimation by Using Local Harmonic Amplitude Information”, London Communications Symposium 2003.
[6] ISO-IEC 14496 “MPEG-4 Coding of Audio Visual Objects”
[7] ITU “Methology for the subjective assessment of the quality of television pictures”, Recommendation ITU-R BT.500-10, 2000.
[8] Y. K. Lai and J. Kuo, “A Haar Wavelet Approach to Compressed Image Quality Measurement”, Journal of Visual Communication and Image Understanding, Vol. 11, pp. 81-84, 2000.
[9] J. Lauterjung, “Picture Quality Measurement”, Proceedings of the International Broadcasting Convention (IBC), Amsterdam, 1998, pp. 413-417.
[10] W. Lee and J. Srivastava, “An Algebraic QoS-Based Resource Allocation Model for Competitive Multimedia Applications”, International Journal of Multimedia Tools and Applications, Kluwer Editions, Vol. 13, pp. 197-212, 2001.
[11] L. Lu, Z. Wang, A. C. Bovik and J. Kouloheris, “Full-reference video quality assessment considering structural distortion and no-reference quality evaluation of MPEG video”, IEEE International Conference on Multimedia, 2002.
[12] J. Mullin, L. Smallwood, A. Watson and G. Wilson, “New techniques for assessing audio and video quality in real-time interactive communications” Third International Workshop on Human Computer Interaction with Mobile Devices, Lille, France, 2001.
[13] J. Olson, “In a framework about task-technology fit, what are the tasks features”, Proceedings of CSCW ’94: Workshop on video mediated communication: Testing, Evaluation & Design Implications, 1994.
[14] F. Pereira and T. Alpert, “MPEG-4 Video Subjective Test Procedures and Results”, IEEE Transactions on Circuits and Systems for Video Technology. Vol. 7(1), pp. 32-51, 1997.
[15] I. G. Richardson, H.264 and MPEG-4 Video Compression : Video Coding for Next Generation Multimedia, Wiley, 2003.
[16] B. Sabata, S. Chatterjee and J. Sydir, “Dynamic Adaptation of Video for Transmission under Resource Constraints”, International Conference of Image Processing, Chicago, October 1998.
[17] P. Seeling, M. Reisslein and B. Kulapala, “Network Performance Evaluation Using Frame Size and Quality Traces of Single Layer and Two Layer Video: A Tutorial”, IEEE Communications Surveys, Volume 6, No. 3, Third Quarter 2004.
[18] K. T. Tan and M. Ghanbari, “A Multi-Metric Objective Picture Quality Measurements Model for MPEG Video”, IEEE Transactions on Circuits and Systems for Video Technology, Vol.10(7), pp. 1208-1213, 2000.
[19] S. Voran and S. Wolf, “Objective Estimation of Video and Speech Quality to Support Network and QoS Efforts”, 2nd Quality of Service Workshop, Houston, Texas, February 2000.
[20] VQEG. “Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment”, http://www.vqeg.org. 2000.
[21] Z. Wang, A. C. Bovik and L. Lu, “Why is Image Quality Assessment so Difficult”, Proceedings IEEE International Conference in Accoustics, Speech and Signal Processing, Vol. 4, pp. 3313-3316, 2002.
[22] Z. Wang, H. R. Sheikh and A. C. Bovik, Objective Video Quality Assessment. The Handbook of Video Databases: Design and Applications, B. Furht and O. Marqure, CRC Press, pp. 1041-1078, 2003.
[23] Z. Wang, L. Lu, A. C. Bovik, “Video Quality Assessment Based on Structural Distortion Measurement”, Signal Processing: Image Communication, special issue on Objective video quality metrics, Vol. 19(2), pp. 121-132, 2004.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity", IEEE Transactions on Image Processing, Vol. 13(4), pp. 1-14, 2004.
[25] A. B. Watson, J. Hu and J. F. McGowan, "DVQ: A Digital Video Quality Metric Based on Human Vision", Journal of Electronic Imaging, Vol. 10(1), pp. 20-29, 2001.
[26] St. Wolf and M. H. Pinson, "Spatial – Temporal Distortion Metrics for in-service Quality Monitoring of any Digital Video System", SPIE International Symposium on Voice, Video, and Data Communications, Boston, 1999 pp. 11-22

Analysis of a New Signaling Method at the Physical Layer for Optical Packet Switched Networks

Efthymios N. Lallas*

Department of Informatics and Telecommunications,
University of Athens[1], GR - 15784, TYPA Buildings, Ilisia, Greece
thymios@di.uoa.gr

Abstract. For the implementation of IP over fiber concept, the All Optical Label Swapping (AOLS) technique is proposed, where the packet routing and forwarding functions are carried out directly in the optical domain. The proposed method is based on the combination of the optical frequency shift keying (OFSK) modulated label with the on–off keying (OOK) modulated payload on the same optical carrier. This orthogonal modulation scheme, for the label encoding onto the intensity modulated payload is studied for the first time via extensive simulation of a network system where the abovementioned functions are taken place.

Keywords: AOLS, MPLS, FWM, SOA, XPM, optical FSK, ER, MZI wavelength converter, 2R regenerator

1 Introduction

1.1 Overview of the problem

The rapid growth of packet based Internet traffic, has fully overtaken circuit switched traffic and has imposed the need for ultrahigh link capacities and ultrahigh packet switching speeds, at network nodes. Multiprotocol label switching (MPLS) technology has been introduced, as a solution for packet routing and forwarding functions, and it is based on label swapping mechanisms. Instead of reading huge route lookups, a single label is read, on each packet [1]. In today's approach, IP packets are mapped to ATM cells, which in turn mapped to SONET frames. However, by carrying the IP packets directly over the WDM layer, we overcome the need of transportation over the two intermediate layers, resulting in an all optical

Dissertation Advisor: Dimitris Syvridis, Professor

process and increased network throughput. The above mentioned trend is supported by the generalized MPLS (GMPLS) protocol, where λ (wavelength) switched channels play the role of the label switched paths in MPLS protocol[2]. Apart from the λ labeling, additional label information can be encoded and attached to the IP packet, via various label encoding methods, at the edge nodes before entering the WDM core network, thus creating an adaptation-encapsulation layer, lying between IP and WDM layers. All optical label swapping (AOLS) is the method of coding the optical label onto the packet, after having removed the old one, for all optical packet routing and forwarding. It directly determines the structure and performance of the optical core node (router), and it is strongly related to the channel bandwidth efficiency and the transmission quality of the packet and the label.
Our method relies a proposed IM/FSK (Intensity modulated-Frequency shift keying) scheme [3], with the label-payload encoding based on four wave mixing (FWM) process in a semiconductor optical amplifier (SOA).
A detailed numerical simulation analysis is carried out, for the investigation of the limits of the method, concerning the propagating distance, the extinction ratio of the IM signal, the modulation index of the FSK header, the number of successive label swapping nodes with their corresponding fiber spans, as well as an optimization of the system critical parameters in order to maximize the above limits. The architecture of the intermediate node assumed in this work is based on two main units, a typical Mach Zender interferometer MZI- SOA based module for the FSK header removal and payload regeneration, and a SOA based FWM module for the label encoding on the IM payload.

1.2 Alternative label coding techniques

In the past few years, many methods have been proposed and studied for all optical labeling. According to the way optical label is attached to the datagram, there are four main categories: i) optical subcarrier multiplexed (SCM) header ii) bit serial header iii) optical orthogonal modulation and iv) wavelength labeled WDM.
The optical SCM method accommodates both, the label and the data payload on the same wavelength, considering the payload as the baseband and modulating the label on an RF frequency subcarrier channel [4]. However, during the propagation of the DSB (double side band) signal through a dispersive fiber, upper and lower SCM side bands will undergo different phase shifts, due to different phase velocities. There are two solutions that can handle it, one concerns the carrier suppressed label extraction, via a fiber Bragg grating (FBG) or a fiber loop mirror [5-6], while the other concerns the single sideband (SSB) transmission via a notch filter [7]. Another drawback of the above mentioned modulation method is that SCM can not support, adequately, high bitrate systems (40Gbps and above) due to electronic components limitations.
The label wavelength method uses a separate wavelength for the transmission of the optical label [8], making inefficient use of the bandwidth and underutilizing the label channel capacity. Moreover, as payload and label propagate through a dispersive

fiber, on separate wavelengths, they would have different speeds due to chromatic dispersion, resulting in a walk off between them.
As far as bit serial method is concerned, many interesting approaches have been proposed, for the optical label processing (extraction and reinsertion) from the payload. These concern the usage of time to wavelength, or inversely mapping, via FBG optical correlation [9], the usage of time gated, wavelength shifting PPLN waveguides [10], a XOR logic [11], a continuous wave tag [12], or other interesting optical pulse code correlation methods[13]. However, bit serial method may require strict synchronization and timing control. Moreover, it sometimes demands different power levels or coding formats (eg RZ payload with NRZ header), in order to distinguish between label and payload.
Two more, interesting approaches concern header separation, using different states of polarization and separate header and payload generation on two symmetrical beat longitudinal modes, caused by original carrier suppression[14].
Finally, the orthogonal modulation and its binary representative (IM/DPSK or IM/FSK) has received major interest. The data payload is intensity modulated, while the label is represented by either the phase or the frequency information of the optical carrier[15]. The crucial point here, is the low extinction ratio, required for proper operation of the label receiver. It has been shown that strong intensity modulation of the payload introduces crosstalk and deteriorates the label quality. As a solution, Manchester coding (instead of NRZ) of the payload pulses is strongly recommended, to suppress the crosstalk term, thus providing better results [16].

2 Orthogonal Label Coding Technique Analysis

This section covers a description of the system, its components, and its importance in an AOLS structure, followed by a theoretical description of the SOA model, used for the implementation of the FWM scheme. Finally an individual investigation and optimization of the critical parameters of the orthogonal IM/FSK scheme, and the limitations imposed on the AOLS network system for a high bit rate propagating IM/FSK signal, are analyzed via numerical simulations.

2.1 Description of the IM/FSK orthogonal scheme and the AOLS network

The whole network system as can be seen in fig.1 is consisted of three kind of nodes: The starting node, the intermediate node and the end node. The starting node is responsible for the generation of the IM payload signal and the FSK modulated header signal, by the corresponding transmitter units, and their combination onto a common wavelength carrier. The most crucial block for the system operation, is the intermediate node which is responsible for the old label extraction and removal, the payload wavelength conversion and regeneration, and the new label generation and insertion with the bare payload, onto an optical carrier.

Fig.1 shows analytically the components that implement each of the above functional blocks. At the starting node, a 625Mbs or a 2,5Gbs NRZ, optical FSK label is combined with the intensity modulated IM 10Gbs or 40Gbs NRZ data payload

respectively. This FSK modulated label, has been realized by chirping, through direct modulation, a laser transmitter, at a low modulation index, according to a typical optical FSK scheme. An FSK compensation scheme has been added on the FSK transmitter. According to the scheme, an electroabsorbsion (EA) modulator,(any AM modulator would do),accepts the optical FSK data, while at the same time is driven with the inverse electrical data, hence the intensity variations at the laser output have been completely removed and furthermore the residual IM has been minimized. At the same time, IM packet payload is generated by externally modulating a MZ amplitude modulator at low extinction ratio (3dB). Both signals enter the SOA after being amplified in such a way, that payload is the pump and label is the signal, according to a typical FWM scheme. Spectra of the IM payload, the FSK modulated header and the conjugate signal at the output of the FWM module are shown in fig 2. In fact, the starting node accomplishes the IP packet and header encapsulation function, generating an optical labeled packet, while the end node strips the packet from the label, thus giving back the pure IP packet.

Before entering the intermediate node, the signal propagates over a span of single mode (SMF) fiber, followed by the proper dispersion compensation (DCF) fiber. Dispersion compensation is required for proper FSK operation, due to the walk off effect between the two FSK tones. The FSK tone spacing is also a crucial matter.

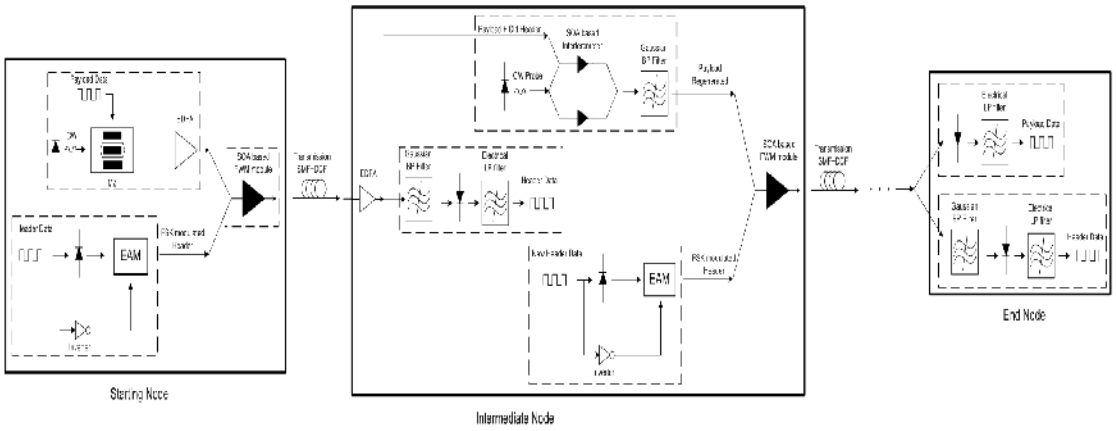


Fig. 1. Detailed schematic representation of the proposed method.

The intermediate node which possesses the functionalities of the AOLS core router consists of three subunits: the label extraction unit, the label removal and payload regeneration unit, and the label insertion unit.

The label extraction module, is a typical optical FSK receiver, composed by a Gaussian 10GHz optical bandpass filter, a PIN photodiode and the appropriate electrical lowpass filter. The same module also applies to the end node, for the final label extraction. The label removal and 2R regeneration module consists of a typical co propagating Cross Phase modulation (XPM) based MZI-SOA model followed by

the appropriate optical Gaussian bandpass filter. Due to the XPM process, all the coherently encoded information (FSK modulation) is not transferred to the output of the module, while the IM payload is not only preserved but also 2R regenerated.

Finally, it is worth mentioning that, sometimes, core routers need to change only the wavelength of the labeled packet and not the encoded label itself, since it is about a two level label encoding. In such case, XPM based interferometric wavelength converter IWC-SOA devices cannot be applied, and instead some transparent wavelength conversion scheme, that preserves FSK modulation, should be used. The appropriate device is then the SOA based FWM scheme, where the IM/FSK signal is the pump and a CW laser is the signal input. Hence, the IM/FSK input will be copied, without a change in the encoded label.

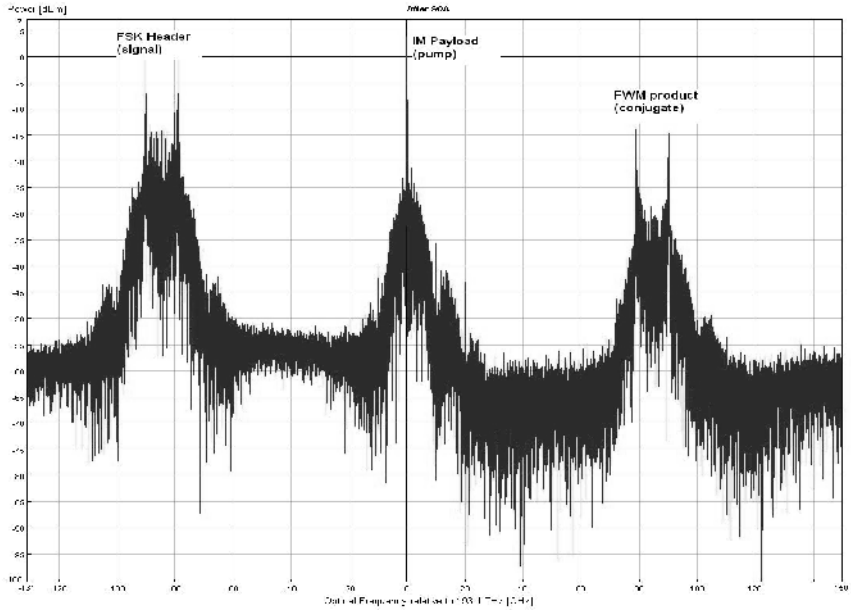


Fig. 2. Spectra of the IM payload (pump), the FSK modulated header (signal) and the conjugate signal at the output of the FWM module.

2.2 SOA model

In the model presented [17], interband and intraband carrier dynamics mechanisms, such as carrier density pulsation (CDP), carrier heating (CH) and spectral hole burning (SHB) have been taken into account. SOA model is described by the following coupled mode rate equations:

$$\frac{\partial A_1}{\partial z} = \frac{1}{2}\left[g\left(1 - i\alpha\right) - \alpha_l\right]A_1 \tag{1}$$

$$-\frac{1}{2}\left(n_{1,2}\left|A_2\right|^2 + n_{1,3}\left|A_3\right|^2 + n_{1,4}\left|A_4\right|^2\right)A_1$$

$$-\frac{1}{2}\left(n_{2,1} + n_{3,1}\right)A_2 A_3 A_1^* - \frac{1}{2}n_{2,4}A_2^2 A_4^* + ASE_1$$

$$\frac{\partial A_2}{\partial z} = \frac{1}{2}\left[g\left(1 - i\alpha\right) - \alpha_l\right]A_2 \tag{2}$$

$$-\frac{1}{2}\left(n_{2,1}\left|A_1\right|^2 + n_{2,3}\left|A_3\right|^2 + n_{1,4}\left|A_4\right|^2\right)A_2$$

$$-\frac{1}{2}\left(n_{1,2} + n_{4,2}\right)A_1 A_4 A_2^* - \frac{1}{2}n_{1,3}A_1^2 A_3^* + ASE_2$$

$$\frac{\partial A_3}{\partial z} = \frac{1}{2}\left[g\left(1 - i\alpha\right) - \alpha_l\right]A_3 \tag{3}$$

$$-\frac{1}{2}\left(n_{3,1}\left|A_1\right|^2 + n_{3,2}\left|A_2\right|^2 + n_{3,4}\left|A_4\right|^2\right)A_3$$

$$-\frac{1}{2}\left(n_{1,4} + n_{2,4}\right)A_1 A_2 A_4^* - \frac{1}{2}n_{1,2}A_1^2 A_2^* + ASE_3$$

$$\frac{\partial A_4}{\partial z} = \frac{1}{2}\left[g(1-i\alpha)-\alpha_l\right]A_4 \tag{4}$$

$$-\frac{1}{2}\left(n_{4,1}\left|A_1\right|^2 + n_{4,2}\left|A_2\right|^2 + n_{4,3}\left|A_3\right|^2\right)A_4$$

$$-\frac{1}{2}\left(n_{1,3}+n_{2,3}\right)A_1 A_2 A_3^* - \frac{1}{2}n_{2,1}A_2^2 A_1^* + ASE_4$$

where Ai (i =1,2,3,4) is the propagating optical field, ASEi is the Amplified Spontaneous Emission (ASE) noise, α_l the internal linear loss, τ_s is the carrier lifetime of the SOA and g the signal gain, given by:

$$\frac{\partial g}{\partial t} = \frac{g_s - g}{\tau_s} - g\frac{P_{tot}}{P_{sat}\tau_s} \tag{5}$$

where gs is the small signal gain, Ptot the total power, Psat the SOA saturation power. The coefficients nij represent the three diffusion contributions, from the nonlinear processes, CDP, CH and SHB given by the following:

$$n_{i,j} = n_{i,j}^{CDP} + n_{i,j}^{CH} + n_{i,j}^{SHB} \tag{6}$$

$$n_{i,j}^{CDP} = \frac{g(1-i\alpha)}{P_{sat}} \times \frac{1}{\left[1-i(\omega_i-\omega_j)\tau_s\right]\cdot\left[1-i(\omega_i-\omega_j)\tau_{SHB}\right]} \tag{7}$$

$$n_{i,j}^{CH} = \varepsilon_{CH}\, g(1-i\alpha_{CH}) \times \frac{1}{\left[1-i(\omega_i-\omega_j)\tau_{CH}\right]\cdot\left[1-i(\omega_i-\omega_j)\tau_{SHB}\right]} \tag{8}$$

$$n_{i,j}^{SHB} = \varepsilon_{SHB}\, g(1-i\alpha_{SHB}) \times \frac{1}{1-i(\omega_i-\omega_j)\tau_{SHB}} \tag{9}$$

where ε and α are the gain compression factors and linewidth enhancement factors respectively, for the CH and SHB phenomena.

2.3 Numerical results and discussion

At first in this section, a proper value for the FSK modulation index, (spectral spacing), between the two FSK tones of the header, at the FSK transmitter, is determined. At second, system's limitations are examined, without including intermediate label swapping nodes, only the starting node with its transmitters, the FWM based combination scheme, the transmission module, and the end node with its corresponding receivers. This investigation is carried out for the determination of the optimum values of critical system parameters, such as the extinction ratio of the signal for different bit streams, the dynamic range of the FWM module and the transmission distance effect on the propagating signal. The above research is done for 10Gbs NRZ IM payload signal, with 625Mbs NRZ FSK modulated header. Next, we continue on testing the system limitations by searching for the optimum values of the above critical parameters, when successive intermediate label swapping nodes with their corresponding transmission spans, are included, according to a typical network route. Hence it's the number of cascaded label swapping nodes that determines system limitations. Finally, the possibility of the system to operate at 40Gbs NRZ IM payload with a 2.5Gbs NRZ FSK modulated header, is investigated.

Optical FSK header transmitter investigation. Advantages and disadvantages for large and small frequency spacings are mentioned. On one hand, a large frequency spacing, with a proper optical bandpass filter would seem to be desired for proper FSK demodulation at the receiver. It is common truth that the more we modulate the signal, the larger frequency deviations we achieve, and the better results we get at the receiver for a constant bit rate. Another benefit of the large frequency spacing is related to the chirp characteristics of the total signal (IM combined with FSK). Chirping generally results in a broadening of the signal spectrum, so if this broadening is too large, the FSK modulated label will be influenced. Luckily, there's no degradation of the FSK signal as long as chirp falls within the bandwidth of the filters used for direct detection of the FSK tone.

On the other hand, in the case of dense WDM networks, with small channel spacing would impose a small frequency deviation between the two FSK tones. Secondly, there is the residual amplitude modulation of the payload, imposed by strong FSK modulation, which has also been removed, due to the applied FSK compensation scheme at the FSK transmitter, otherwise it would require a weak modulation and consequently small spacing.

It is obvious that, the greater FSK tone deviation we get, the better performance we achieve, since it is easier and more efficient to demodulate the header. On the contrary, payload's performance is deteriorating, as the FSK tone deviation increases because the residual intensity modulation effect imposed by the stronger FSK modulation becomes more pronounced. FSK values around 12-20GHz could be the

perfect choice for combined acceptable payload and header performances, at the receivers.

Investigation on system limitation without intermediate label swapping nodes. There are some critical parameters, that directly or indirectly, affect system's performance, such as the extinction ratio (ER) of the externally modulated payload, the width of the optical filter of the label receiver, and the dynamic range of the FWM. The IM modulated payload's ER trade off, is one of the most crucial points, for the proposed encoding method. High ER ensures high performance for the payload, but it is catastrophic for the header while, low ER continues to support optical FSK, even during the zeros of the payload at the expense of the lower IM modulated payload performance. Good performance above BER threshold can be achieved for the 10Gbs payload and 625Mbs header in the back to back configuration with ER values not higher than 4dB. Such an ER is not sufficiently high to ensure high performance in a network configuration and is one of the weak points for the IM/FSK technique as well as for other orthogonal techniques proposed in the literature.

It is also shown by our simulation runs, that the header sustains a good performance for transmission distances up to 80km, while the corresponding maximum transmission distance for the payload is not higher than 60km, which therefore is the maximum distance between two successive nodes for the complete system.

Finally, one of the most important operating parameters to characterize, is the dynamic range of the SOA based FWM module. The data signals do not always have the same power level when they reach the FWM module. The dynamic range, for various pump power levels at the input of FWM module, is around 8.84dB.

Investigation on system limitation when intermediate label swapping nodes are included. At first, system performance is examined, as concerns the number of intermediate nodes, for a constant ER of the signal, of 3dB, and 50km, dispersion compensated SMF fiber span. The maximum number of label swapping nodes that can be supported by the system is five. It is worth noticing that, header preserves an almost steady behavior, due to the constant ER, while payload decreases gradually as the number of the spans and the nodes across the route increases.

System limitation, concerning the number of intermediate nodes, continues by varying the distance per span, for a constant ER, and inversely, varying the ER, for a constant 50km span distance. Three different span values have been assumed of 30, 50 and 70km, for a constant ER value of 3dB. The total length of the network path reduces as the number of the intermediate nodes increases, thus the payload signal of the system survives after a four node route of 70km span distance, in between nodes, or a five node route of 50km span distance, or finally a six node route of 30km span distance. As concerns header performance, all the above mentioned system configurations, have the same behavior. The reason is that, at each node the header is extracted and a new one is inserted.

Finally system's limitation is also examined, concerning the number of cascaded intermediate nodes, for a constant distance span of 50km, with the ER signal values of 3dB, 8dB, and 12dB. The payload signal of the system, survives after a ten node

route, when it is IM modulated with 12dB ER, or after a six node route, when it is IM modulated with 8dB ER, or after a five node route with 3dB ER.

40Gbs IM payload with 2.5Gbs FSK header for AOLS applications. The two critical functions which should respond at these high bit rates are the header erasure XPM unit and the FWM based combiner of IM modulated payload with FSK modulated header. As already mentioned above both these modules use as a key element a SOA. The ability of XPM unit to operate at 40Gbs has been already proved experimentally[18]. Concerning the FWM unit, it would be expected to have no problem in operating at high bit rates, since the FWM phenomenon based on ultrafast nonlinear processes in the SOA has much higher frequency limitations. Unfortunately, the numerical simulations proved that this is not the case, mainly for the FWM unit. The reason is due to the pump modulation scheme needed for the FSK/IM combination, and is related to the long free carrier relaxation times.

3 Conclusion

An orthogonal IM/FSK encoding scheme, with the label–payload coupling based on FWM in a SOA, and the label removal with payload wavelength conversion and regeneration, based on XPM in a SOA-MZI module, has been investigated on its limits, via extensive simulation. Many critical parameters of the system have been tested and optimized, in order to determine these limits, and concern almost all the functional blocks of the configuration, starting from the FSK transmitter module individually, the back to back edge node configuration with and without propagation DCF fiber span, and the complete configuration, with the inclusion of intermediate label swapping nodes and their corresponding transmission spans, according to a typical network route. The system has been tested mostly for 10Gbs NRZ IM payload with 625Mbs FSK modulated header, which appear to be the upper limits for the bit rate of the two bit streams (payload and header).

References

1. A. Viswanathan, N. Feldman, Z. Wang and R. Callon “Evolution of multiprotocol label switching”, IEEE Commun. Mag, Vol.36, pp165-173, May 1998.
2. A. Banerjee et al. “Generalized multiprotocol label switching: An overview of routing and management enhancements”, IEEE Commun. Mag, pp144-150, Jan. 2001.
3. E.N.Lallas, N.Skarmoutsos and D.Syvridis “An optical FSK based label coding technique for the realization of the all optical label swapping” IEEE Photon. Technol. Lett. vol. 14. Pp. 1472-1474, Oct.2002.
4. D. J. Blumenthal, et al. “All optical label swapping networks and technologies” J. Lightwave Technol, Vol. 18, pp 2058-2075, Dec.2000.
5. H.J. Lee, S.J.B. Yoo, V.K. Tsui and S.K.H. Fong “A simple all optical label detection and swapping technique incorporating a fiber Bragg grating filter” IEEE Photon. Technol. Lett, vol.13, pp 635-637, June 2001.

6. G. Rossi, O. Jerphagnon, B. Olsson and D.J. Blumenthal "Optical SCM data extraction using a fiber loop mirror for WDM network systems" IEEE Photon. Technol. Lett. Vol.12, pp 897-899, July 2000.
7. Y.M.Lin, W.I.May and E.K.Chang "A novel optical label swapping technique using erasable optical single sideband subcarrier label" IEEE Photon. Technol. Lett. Vol.12, pp 1088-1090, Aug. 2000.
8. A.Okada "All optical packet routing in AWG based wavelength routing networks using an out of band filter" in Proc. OFC'02, vol. 1, Washington, DC, 2002, Paper WG1, pp. 213-215.
9. X.Jiang, K.M.Feng, M.Cardakli, J.X.Cai, A.E.Willner, V.Grubsky, D.S.Stavrodubov and J.Feinberg "Control monitoring of routing bits and data packets in WDM networks using wavelength to time mapping " IEEE Photon. Technol. Lett. Vol.11, pp 1186-1188, Sept. 1999.
10. D.Gurkan,et al. "Simultaneous label swapping and wavelength conversion of multiple independent WDM channels in all optical MPLS using PPLN waveguides as wavelength converters" IEEE J. Lightwave Technol. vol. 11, pp. 2739-2745, Nov.2003.
11. T.Fjelde, A.Kloch, D. Wolfson, B.Dagens, A. Coquelin, I.Guillemot, F.Poingt and M.Renaud "Novel scheme for simple label swapping employing XOR logic in an integrated interferometric wavelength converter" IEEE Photon. Technol. Lett. Vol.13, pp. 750-752, July 2001.
12. X.Jiang, X.P.Chen and A.E.Willner "All optical wavelength independent packet header replacement using a long CW region generated directly from the packet flag" IEEE Photon. Technol. Lett. Vol.10, pp. 1638-1640, Nov.1998.
13. X.Jiang, X.P.Chen and A.E.Willner "All optical wavelength independent packet header replacement using a long CW region generated directly from the packet flag" IEEE Photon. Technol. Lett. Vol.10, pp. 1638-1640, Nov.1998.
14. J.Yu, and G.K.Chang "A novel technique for optical label and payload generation and multiplexing using optical carrier suppression and separation" IEEE Photon. Technol. Lett. vol. 16, pp. 320-322, Jan.2004.
15. K.Vlachos et al. "An optical IM/FSK coding technique for the implementation of a label controlled arrayed waveguide packet router" IEEE J. Lightwave Technol. vol. 21, pp. 2617-2628, Nov.2003.
16. J.Zhang, N.Pablo, V.Nielsen C.Peucheret and P.Jeppesen "Performance of Manchester coded payload in an optical FSK labeling scheme" IEEE Photon. Technol. Lett. vol. 15. Pp. 1174-1176, Aug. 2003.
17. H.Simos, A.Bogris and D.Syvridis "Investigation of a 2R all optical regenerator based on four wave mixing in a semiconductor optical amplifier" IEEE J. Lightwave Technol. vol.22, pp.1-9, Jan.2004.
18. C. Joergensen et. Al. "All optical wavelength conversion at bit rates above 10Gb/s using semiconductor optical amplifiers" IEEE J. Lightwave Technol. Vol.3, pp. 1168-1180, Oct 1997

Signal Processing Techniques in Cryptography

Konstantinos Limniotis*

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
TYPA Buildings, University Campus, 15784 Athens, Greece
klimn@di.uoa.gr

Abstract. Security of cryptographic symmetric primitives is studied in this thesis. Pseudorandomness characteristics of cryptographic sequences are analyzed, resulting in new methods for constructing sequences with high linear complexity. Connections between nonlinear complexity and other cryptographic criteria are also established, whereas a new recursive algorithm for efficiently computing the minimal feedback shift register which generates a given sequence is provided. Furthermore, security issues of cryptographic Boolean functions that are used in cryptographic systems as components of sequence generators are studied; on this direction, new efficient formulas for determining best quadratic approximations of several classes of Boolean functions are derived, leading to new design principles that should be considered in the construction of secure cryptosystems.

Keywords: Boolean functions, complexity, Discrete Fourier Transform, feedback shift registers, sequences, stream ciphers.

1 Introduction

Cryptographic algorithms are categorized into two families, namely symmetric or secret-key algorithms and public-key algorithms [22]. Symmetric algorithms are the only ones that achieve several important functionalities such as high speed and low-cost encryption and are used in conjunction with public-key techniques in order to safely distribute the secret key among the members of a group. Symmetric algorithms, being further classified into block ciphers and stream ciphers, are used in many applications. Especially stream ciphers are widely used to provide confidentiality in environments characterized by a limited computing power or memory capacity, and the need to encrypt at high speed. Typical examples of stream ciphers are the A5/1 and E0 algorithms, employed in GSM communications and Bluetooth protocol respectively.

In general, a stream cipher consists of a binary keystream generator, whose output $k_1k_2\ldots$ is added modulo 2 to the original message $m_1m_2\ldots$, leading to the encrypted message (ciphertext) $c_1c_2\ldots$ (Fig. 1). *Shift registers* of linear (LFSR) or nonlinear (NFSR) feedback are a basic building block of keystreams generators in stream ciphers. The security of such systems is strongly contingent on the pseudorandomness characteristics of the keystream. The pseudorandomness is attributed to several factors; amongst others, an important cryptographic

* Dissertation Advisor: Prof. Nicholas Kalouptsidis.

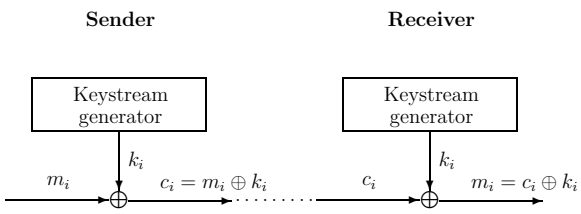


Fig. 1. Basic functionality of a stream cipher

feature of a sequence is its *nonlinear complexity* or simply *complexity*, defined as the length of the shortest feedback shift register that produces the sequence. Especially the *linear complexity*, i.e. the length of the shortest LFSR generating a given sequence, is important for assessing resistance to cryptanalytic attacks, like the *Berlekamp–Massey algorithm* (BMA) [19]. However, determining the connections and trade-offs between several cryptographic criteria of sequences remains an open problem. Since most constructions are ad-hoc, finding good generators is of great theoretical and practical value.

High linear complexity keystreams are generated by applying Boolean functions either as *filters* [7][26] or *combiners* [26], to one or several LFSRs respectively. In any case, the highest value attainable by linear complexity depends on the degree of the function [7]. The problem of determining the exact linear complexity attained by filterings is open. Two classes of filters have been introduced, namely *equidistant* [26] and *normal* [8], that allow to derive lower bounds on the linear complexity. Given a filter of degree k and a LFSR of length n whose characteristic polynomial is primitive over $\mathbb{F}_{2^n}$, the lower bound on the linear complexity of keystreams is $\binom{n}{k}$ for both types of filters. These results rely on the so-called Rueppel's *root presence test* [26].

On the contrary, the general case of nonlinear complexity has not been studied to the same extent. In [4], a directed acyclic word graph is used to exhibit the complexity profile of sequences over arbitrary fields. An approximate probability distribution for the nonlinear complexity of random binary sequences is derived in [3]. Recent results are provided in [24], where the minimal nonlinear FSR generating a given sequence is computed via an algorithmic approach, and [25] where the special case of a quadratic feedback function of the FSR is treated.

Apart from the pseudorandomness characteristics of keystream, many attacks on conventional cryptographic algorithms are related to some properties of the underlying Boolean functions. The formalization of well-known attacks against LFSR-based stream ciphers have led to the definitions of some relevant quantities related to Boolean functions. These quantities measure the resistance of a cryptosystem to classical attacks. For instance, high algebraic degree of nonlinear filters or combiners is prerequisite for constructing sequences achieving high linear complexity. Furthermore, the *nonlinearity* of Boolean functions is one of the most significant cryptographic properties; it is defined as the minimum distance from all affine functions, and indicates the degree to which attacks based on linear cryptanalysis [21] can be prevented. With the appearance of more recent attacks, such as algebraic [2], and low order approximation attacks [11],

Boolean functions need also have the property that they cannot be approximated efficiently by low degree functions. Hence, the *rth order nonlinearity* characteristics of Boolean functions need also be analyzed. This is known to be a difficult task for $r > 1$, whereas even the second order nonlinearity is unknown for all Boolean functions, with the exception of some special cases, or if the number of variables n is small [1].

In this thesis, state-space representations are employed as vehicle to the study of complexity of binary sequences. System theoretic concepts, namely controllability and observability, are used to characterize minimal sequence generators [6]. Jordan canonical forms are used for the complete analysis of sequences whose Fourier transform is not defined [15]. A new *generalized discrete Fourier transform* (GDFT) is proposed that presents the same properties with the GDFT defined in [20]. In addition, connections of this new GDFT with a new vectorial trace representation of sequences are established that *facilitate the generation of sequences with prescribed linear complexity*. Furthermore, nonlinearly filtered maximal length sequences with period $N = 2^n - 1$ are studied under this framework, resulting in new general classes of nonlinear filters of degree k which *generalize Rueppel's equidistant filters and guarantee the same lower bound* $\binom{n}{k}$ *on the linear complexity* [15]. The connections between the nonlinear and Lempel-Ziv complexity are also studied, which is a well-known open problem [23]. It is shown that the eigenvalue profile of a sequence, which determines the Lempel-Ziv complexity, also determines its nonlinear complexity profile [14]. Furthermore, for any periodic binary sequence, we establish the dependence of the minimum achievable compression ratio on its nonlinear complexity by deriving a lower bound depending on the complexity [14]. Based on the properties of the nonlinear complexity profile, a new efficient recursive algorithm producing the minimal FSR of a binary sequence is developed, thus *generalizing the BMA to the nonlinear case* [14],[16]. Finally, explicit formulas are proved that compute all best quadratic approximations of a class of functions with degree 3 or 4 [9]. These results are based upon Shannon's expansion formula of Boolean functions and *hold for an arbitrary number n of variables*. The derived method reveals new design principles for cryptographic functions. An analysis of contemporary constructions of functions is also performed, indicating potential weaknesses if construction parameters are not properly chosen.

This summary is organized as follows; First, Section 2 introduces the basic definitions and settles the notation. Section 3 provides the basic results regarding the linear complexity of sequences obtained by state space generators, while Section 4 presents the new results regarding the nonlinear complexity, as well as its connections with Lempel-Ziv complexity. The algorithmic method of computating the best quadratic approximations of Boolean functions is described in Section 5. Finally, concluding results are given in Section 6.

2 Preliminaries

Let $f : \mathbb{F}_2^n \to \mathbb{F}_2$ be a Boolean function, where $\mathbb{F}_2 = \{0, 1\}$. The set of Boolean functions on n variables is denoted by $\mathbb{B}_n$. The complement of a binary variable x will be denoted by $x' = x + 1$, where "+" represents addition modulo 2.

Boolean functions are expressed in their *algebraic normal form* (ANF) as

$$f(x_1, \ldots x_n) = \sum_{e \in \mathbb{F}_2^n} a_e \, x_1^{e_1} \cdots x_n^{e_n}, \qquad a_e \in \mathbb{F}_2 \tag{1}$$

where the sum is taken modulo 2, $e = (e_1, \ldots, e_n)$, while $x_i^1 = x_i$ and $x_i^0 = 1$. The *degree* of f equals $\deg(f) = \max\{\mathrm{wt}(e) : a_e = 1\}$, and $\mathrm{wt}(e)$ is the *Hamming weight* of vector e. If $\deg(f) = 1, 2, 3$, then f is called *affine* (or *linear* if its constant term is zero), *quadratic*, *cubic*; terms with degree $k \leq \deg(f)$ in ANF comprise its *kth degree part*. The *distance* of $f, g \in \mathbb{B}_n$ is $\mathrm{wt}(f + g)$.

Another representation , the so-called *Exclusive-or Sum-Of-Products* (ESOP), occurs if the variables in (1) are taken to be in either complemented or uncomplemented form.

The *Shannon's expansion formula* of $f \in \mathbb{B}_n$ with respect to x_j is

$$f(x_1, \ldots, x_n) = f_0 \parallel_j f_1 \triangleq (1 + x_j) f_0 + x_j f_1 \,, \quad 1 \leq j \leq n$$

where *sub-functions* $f_0, f_1 \in \mathbb{B}_{n-1}$ do not depend on x_j; they are the restriction of f in $x_j = 0, 1$.

The Walsh transform of $f \in \mathbb{B}_n$ at $a \in \mathbb{F}_2^n$ is the real-valued function

$$\widehat{\chi}_f(a) = \sum_{x \, \in \, \mathbb{F}_2^n} \chi_f(x) (-1)^{\phi_a(x)} = 2^n - 2 \, \mathrm{wt}(f + \phi_a) \tag{2}$$

with $\chi_f(x) = (-1)^{f(x)}$ and $\phi_a(x) = \sum_i a_i x_i$. The minimum distance between f and all affine functions is determined by

$$\mathcal{NL}_f = \min_{v \, \in \, \mathfrak{R}(1,n)} \big\{ \mathrm{wt}(f + v) \big\} = 2^{n-1} - \frac{1}{2} \max_{a \, \in \, \mathbb{F}_2^n} |\widehat{\chi}_f(a)| \tag{3}$$

and is called *nonlinearity* of f. An affine function v such that $\mathrm{wt}(f + v) = \mathcal{NL}_f$ is a *best affine approximation* of f, denoted by λ_f, and $\mathcal{A}_f$ is the set of all its best affine approximations. The above can be extended to *best quadratic approximations* of f, denoted by ξ_f, which are quadratic functions u satisfying $\mathrm{wt}(f + u) = \min_{u : \deg(u) \leq 2} \{\mathrm{wt}(f + u)\} \triangleq \mathcal{NQ}_f$.

A sequence $y = \{y_i\}_{i \geq 0}$ with elements in the finite field $\mathbb{F}_2$ is said to be *ultimately periodic* if there exist integers $T > 0$ and $t_0 \geq 0$ such that $y_{i+T} = y_i$ for all $i \geq t_0$. The least integer T with this property is called *period* of y, and t_0 is its *preperiod*. If $t_0 = 0$, then the sequence y is said to be *periodic*. If y has finite length N, then $y^N \triangleq y_0^{N-1}$ denotes the whole sequence. For any $0 \leq j < N$, the tuple y_0^j is a *prefix* of y^N; for the special case that $j < N - 1$, such a prefix is called *proper prefix*. A *suffix* of y^N is any tuple y_j^{N-1}, $0 \leq j \leq N - 1$; a proper suffix is similarly defined. Such sequences are typically generated by FSRs satisfying a recurring relation of the form $y_{i+n} = h(y_{i+n-1}, \ldots, y_i)$, $i \geq 0$, where $n > 0$ equals the number of stages of the FSR. The feedback $h : \mathbb{F}_2^n \to \mathbb{F}_2$ is a nonlinear function, usually having a zero constant term, mapping elements of the nth-dimensional vector space $\mathbb{F}_2^n$ onto $\mathbb{F}_2$. In the case of a linear feedback, i.e. $y_{i+n} = a_{n-1} y_{i+n-1} + \cdots + a_1 y_{i+1} + a_0 y_i$, each LFSR is associated with its characteristic polynomial $f(z) = z^n + a_{n-1} z^{n-1} + \cdots + a_0$.

Definition 1 *The length of the shortest FSR generating a sequence y^N is referred to as* nonlinear complexity *of y^N, and is denoted by $c(y^N)$. The integer-valued sequence $c(y^1), \ldots, c(y^N)$ is called* nonlinear complexity profile. *The linear complexity* $\mathrm{lc}(y^N)$ *and the linear complexity profile are similarly defined.*

When the LFSR generating y has the minimum possible length, then its characteristic polynomial $f(z)$ is called *minimal polynomial* of y.

Let the vector $\boldsymbol{y} = (y_0\ y_1\ \cdots\ y_{N-1})$ contain the first N elements of a binary periodic sequence y of period N. Let N be a divisor of $2^n - 1$ for some positive integer n. The Fourier transform of $\boldsymbol{y}$ is the vector $\boldsymbol{Y} = (Y_0\ Y_1\ \cdots\ Y_{N-1})$ of length N whose elements are given by

$$Y_j = \sum_{i=0}^{N-1} y_i \alpha^{ij}, \qquad 0 \le j < N \tag{4}$$

where $Y_j \in \mathbb{F}_{2^n}$ and $\alpha \in \mathbb{F}_{2^n}$ is an element of order N in the extension field $\mathbb{F}_{2^n}$. Vector $\boldsymbol{y}$ is reconstructed from $\boldsymbol{Y}$ by means of the inverse Fourier transform

$$y_i = \sum_{j=0}^{N-1} Y_j \alpha^{-ij}, \qquad 0 \le i < N\,. \tag{5}$$

A direct consequence of (4) is that the Fourier coefficients satisfy the *conjugacy property*, i.e. $Y_{2j \bmod N} = Y_j^2$ for all $0 \le j < N$. The linear complexity of a periodic sequence over $\mathbb{F}_2$, with period N a divisor of $2^n - 1$ for some integer n, equals the Hamming weight of its Fourier transform (*Blahut's theorem*). A generalized Fourier transform, that describes sequences of arbitrary period, is defined in [20].

For any positive integer N such that $\gcd(N, 2) = 1$ and each $j \in \mathbb{Z}_N = \{0, \ldots, N-1\}$, we define the set of distinct elements $I_j = \{j, 2j, \ldots, 2^{n_j - 1} j\}$ to be the *cyclotomic coset* of j, where all elements are taken modulo N and $n_j = |I_j|$. The least element in I_j is referred to as *coset leader* and the set containing all coset leaders modulo N will be denoted by I. From the definition of cyclotomic cosets and the conjugacy property satisfied by (4) and (5), the Fourier transform can be equivalently written as

$$y_i = \sum_{j \in I} \mathrm{tr}_1^{n_j}(Y_j \alpha^{-ij}) \tag{6}$$

where $Y_j \in \mathbb{F}_{2^{n_j}}$ and the function $\mathrm{tr}_1^{n_j}(z) = z + z^2 + \cdots + z^{2^{n_j - 1}}$ is the *trace function* that maps elements of $\mathbb{F}_{2^{n_j}}$ onto its prime subfield $\mathbb{F}_2$ [13]. The above is called *trace representation* of the sequence y.

3 Linear complexity of sequences obtained by state-space generators

In this section we focus on linear state space generators, described by

$$x_{i+1} = \boldsymbol{A}\, x_i \tag{7a}$$

$$y_i = c^T\, x_i \tag{7b}$$

where x_i, c are $n \times 1$ vectors, c^T denotes the transpose of c, and $\boldsymbol{A}$ is an $n \times n$ matrix. The integer n defines the *dimension* of the system $\mathfrak{L}$. Clearly, any LFSR can be described by (7). Any such system is denoted by $\mathfrak{L} = \langle \boldsymbol{A}, c, x_0 \rangle$.

Proposition 1 ([5]) *A linear realization $\mathfrak{L} = \langle \boldsymbol{A}, c, x_0 \rangle$ of a periodic sequence y with dimension n is minimal (i.e. there is no other linear realization of lower dimension generating y) if and only if it is both controllable and observable.*

Proposition 2 ([5]) *Let $\mathfrak{L} = \langle \boldsymbol{A}, c, x_0 \rangle$ and $\mathfrak{L}' = \langle \boldsymbol{A}', c', x_0' \rangle$ be two minimal linear realizations of a periodic sequence y. Then, $\mathfrak{L}$ and $\mathfrak{L}'$ are necessarily isomorphic (or equivalent) since there exists a change of coordinates $\boldsymbol{P}$ such that it holds $\boldsymbol{A}' = \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}$, $(c')^T = c^T\boldsymbol{P}^{-1}$, and $x_0' = \boldsymbol{P}x_0$.*

It is well-known that any matrix with coefficients in an algebraically closed field can be put into the so-called *Jordan canonical form*. Thus, among all isomorphic minimal linear realizations $\mathfrak{L}$ of a given sequence y, there exists one with state transition matrix $\boldsymbol{A}$ in the Jordan canonical form.

Theorem 1 ([15]) *Let $\mathfrak{L} = \langle \boldsymbol{A}, c, x_0 \rangle$ be a linear realization of sequence y with matrix $\boldsymbol{A}$ in the Jordan canonical form. The generator $\mathfrak{L}$ is controllable (resp. observable) if and only if there is one Jordan block associated with each eigenvalue and all elements of the initial state vector x_0 (resp. output vector c) corresponding to the last row (resp. first column) of each Jordan block are nonzero.*

Theorem 2 ([15]) *With the above notation, let the state transition matrix $\boldsymbol{A}$ be diagonal. Then, the generator $\mathfrak{L}$ is minimal if and only if the eigenvalues of $\boldsymbol{A}$ are pairwise distinct and all elements of x_0 and c are nonzero.*

In this thesis it is proved (by using the above results) that, if y is a periodic binary sequence with least period N, then it admits a diagonal realization $\mathfrak{L}$ over the splitting field of $z^N - 1$ if and only if its Fourier transform exists, or equivalently $\gcd(N, 2) = 1$. In this case, the initial state of the diagonal realization with dimension N equals its Fourier transform. Hence, we directly prove that the dimension of a minimal realization of any such sequence equals its linear complexity. Similar results for the more interesting general case of sequences whose Fourier transform is not defined are also proved. Let $z = (z_1 \ z_2 \ \cdots \ z_m)^T$ be an $m \times 1$ vector with elements over $\mathbb{F}_{2^n}$. We define the *block-trace function*

$$\mathbf{tr}_1^n(z) = \left(\mathrm{tr}_1^n(z_1) \ \mathrm{tr}_1^n(z_2) \ \cdots \ \mathrm{tr}_1^n(z_m)\right)^T \tag{8}$$

that maps vectors of $\mathbb{F}_{2^n}^m$ to elements of the vector space $\mathbb{F}_2^m$. Then, we prove the following.

Theorem 3 *Let y be a binary sequence of period $N = 2^e m$, with m odd and $e > 0$, and minimal polynomial $f(z)$ factored as*

$$f(z) = f_1(z)^{d_1} f_2(z)^{d_2} \cdots f_r(z)^{d_r} \tag{9}$$

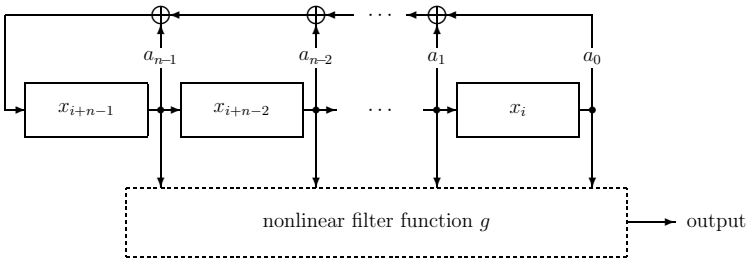


Fig. 2. A nonlinear filter applying to a LFSR

where $d_s > 0$ and f_s is irreducible of degree n_s, $1 \leq s \leq r$. Further let α be a primitive mth root of unity over $\mathbb{F}_2$ lying in the splitting field of $z^N - 1$, and let α^{j_s} be a root of f_s. Then y is always written in the vectorial trace representation

$$y_i = \sum_{1 \leq s \leq r} \mathbf{1}_{d_s}^T \, \mathbf{tr}_1^{n_s}(\boldsymbol{J}_{j_s}^i z_s), \qquad i \geq 0 \tag{10}$$

where $z_s = (z_{s,1} \; z_{s,2} \; \cdots \; z_{s,d_s})$ is a vector over the splitting field of $f(z)$, and the Jordan block $\boldsymbol{J}_{j_s}$ has dimension d_s and diagonal element α^{j_s}.

Equation (10) provides a novel generalized trace representation of any sequence y, including those whose minimal polynomial does not have distinct irreducible factors. The practical importance of the above resides in the fact that we can generate sequences of prescribed linear complexity by appropriately selecting the initial state *for any LFSR*.

Based on (10), which generalizes (6), we introduce below a new *generalised discrete Fourier transform* (GDFT) for sequences of arbitrary period $N = 2^e m$, with m odd and $e > 0$.

Definition 2 ([15]) *Let y be a binary sequence of period $N = 2^e m$ generated by $\mathfrak{L} = \langle \boldsymbol{J}, \mathbf{1}, \boldsymbol{Y} \rangle$ of dimension N, where $\boldsymbol{J}$ is a Jordan matrix. Then, the initial state*

$$\boldsymbol{Y} = (Y_0 \; Y_1 \; \cdots \; Y_{N-1})^T \tag{11}$$

is defined as the generalized discrete Fourier transform of y.

The above definition is easily generalized to sequences over fields with an odd prime characteristic p, since it can be easily seen that the vectorial trace representation (10) holds for sequences over any field. Clearly, if $e = 0$ the above generalized discrete Fourier transform and the vectorial trace representation (10) coincide with their ordinary counterparts. The advantage of the proposed GDFT is that while being a natural generalization of the usual DFT from a system theoretic point of view, it also allows the easy computation of the linear complexity of sequences by means of the Günther weight, like the GDFT proposed in [20].

Next, we consider a LFSR of length n with a primitive characteristic polynomial f, where a nonlinear filter function g of degree $k \leq n$ is applied to its stages

(Fig. 2). The realization $\mathfrak{N} = \langle \boldsymbol{A}, g, x_0 \rangle$ of the sequence $y = \{y_i\}_{i \geq 0}$ generated by the above automaton is described by

$$x_{i+1} = \boldsymbol{A} x_i \tag{12a}$$

$$y_i = g(x_i) \tag{12b}$$

where $\boldsymbol{A}$ is the companion matrix of the characteristic polynomial f and $x_0 = (x_{0,1}\ x_{0,2}\ \cdots\ x_{0,n})^T$ is the initial state of $\mathfrak{N}$. The study of such generators is simplified if they are linearly described; this is accomplished by treating each product term in the ANF of g as a single variable by suitably extending the state space. This procedure is called *linearization* of generator $\mathfrak{N}$ and is based upon properties of Kronecker products. Hence, minimality characteristics of the equivalent linearized system are also determined by controllability and observability arguments. It is proved in this thesis that the Rueppel's root presence test can be re-derived by this analysis. Furthermore, based on the analysis via Kronecker products, the following result is proved.

Theorem 4 ([15]) *Let $\mathfrak{N} = \langle \boldsymbol{A}, g, x_0 \rangle$ be a generator of dimension $n > 0$ and let the characteristic polynomial f of the state transition matrix $\boldsymbol{A}$ be primitive. For some positive integer δ with $\gcd(\delta, 2^n - 1) = 1$ assume the nonlinear filter g consists of only one product of degree $k < n$*

$$g(z_1, z_2, \ldots, z_n) = z_{t_1} z_{t_2} \cdots z_{t_k}$$

where $t_1 > \delta$ or $t_k \leq n - \delta$. If there exists an integer $1 \leq i \leq k$ such that the product $g_i(z_1, \ldots, z_n) = z_{t_1} \cdots z_{t_{i-1}} z_{t_{i+1}} \cdots z_{t_k}$ is equidistant, with distance δ, then the linear complexity of the generated sequence is lower bounded by $\binom{n}{k}$.

Theorem 4 generalizes Rueppel's results since it defines filter functions that are slightly different from equidistant filters but admit the same lower bound on the linear complexity. It is also proved that these results can be used to generalize other classes of nonlinear filters that generate sequences achieving a prescribed lower bound on the linear complexity - such as filters based on normal bases. Clearly, the above new construction provides greater flexibility in designing nonlinear filters that output sequences of high linear complexity.

4 Nonlinear complexity and Lempel-Ziv complexity

This section studies the nonlinear conplexity of sequences and its connections with Lempel-Ziv complexity; the latter is defined by the number of words occuring via a specific parsing procedure of the sequence described in [12]. As it is shown in [12], the Lempel-Ziv complexity is determined by the the eigenvalue profile $k(y^1), k(y^2), \ldots, k(y^N)$ of the sequence y^N. The value of $k(y^i)$ equals $i - s_i$, where s_i is the length of the longest suffix of y^i that is present at least twice within y^i. By proving a series of results, we get the following.

Theorem 5 ([14]) *Let $c(y^{n-1}) = m$ and assume the minimal FSR of y^{n-1} does not generate y^n. Then, it holds*

$$c(y^n) = \max\{c(y^{n-1}), n - k(y^{n-1})\} \tag{13}$$

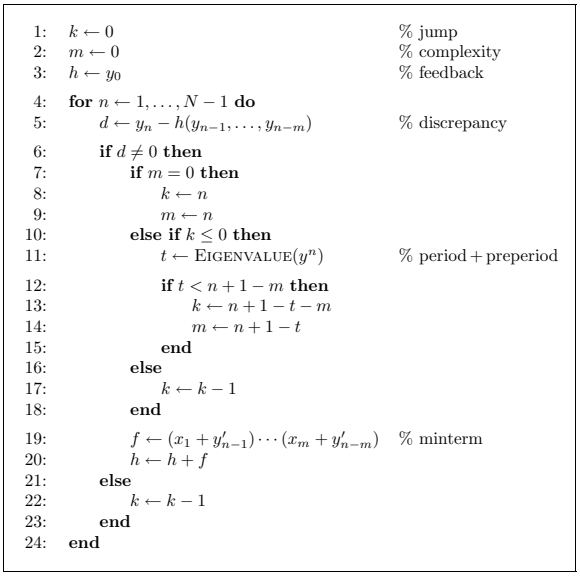


Fig. 3. A recursive algorithm for minimal FSR synthesis of binary sequence y^N

The next result exhibits that the eigenvalue profile uniquely determines the complexity profile.

Theorem 6 ([14]) *If two sequences have the same eigenvalue profile, then they necessarily have the same nonlinear complexity profile.*

The above are used to develop a recursive algorithm that computes the minimal FSR of any binary sequence, which comprises the generalization of the BMA to the nonlinear case. This algorithm is illustrated in Fig. 3 [14],[16]. The feedback function of the minimal FSR is given in the ESOP representation. In Lines 11, 12 of the algorithm, we examine whether a jump in the complexity occurs based on Theorem 5; if a jump occurs, then its value is computed in Line 13. Note that this step has linear computational complexity due to the existence of the *Knuth-Morris-Pratt* (KMP) algorithm for pattern matching. The computational complexity of the algorithm mainly rests with line 5 where the boolean function $h^{(n)}$ is evaluated. For each $n \leq N$ the ESOP representation of $h^{(n)}$ has less than n terms, each consisting of at most $c(y^n) \leq n$ variables. Since no term is present in the ESOP representation of $h^{(n)}$ having more than $c(y^N)$ variables, the computational complexity of the algorithm in the average case highly depends on the expected value of the nonlinear complexity of random binary sequences of given length N. It is known that for large N it holds $E(c(y^N)) \approx 2\log_2 N$, and the average computational complexity of the algorithm becomes $O(N^2 \log_2 N)$. Note that our algorithm has the same computational complexity with the one proposed in [4] for determining the minimal FSR of a given sequence. However, its recursive nature is an important advantage since it eliminates the need to know the entire sequence in advance.

Connections between nonlinear complexity and Lempel-Ziv compression ratio are also established. More precisely, the next result is proved.

Theorem 7 ([14]) *Let y be a binary sequence with period N, and let $c(y) = m$. If y^N denotes the first N terms of y and n is the largest integer such that $2^n \leq m < 2^{n+1}$, then*

$$\rho_{y^N} > \min\left\{ \frac{1}{m}\lceil \log_2(2m) \rceil, \frac{1}{2^n}(n+1) \right\}, \tag{14}$$

where ρ_{y^N} is the compression ratio of y^N according to the Lempel-Ziv compression algorithm of [27].

As it is proved in this thesis, the bound in (14) decreases as the complexity $c(y)$ of the periodic sequence y increases. Therefore, it is possible to design a construction for generating sequences of very high complexity, which however are highly compressible. Since truly random sequences do not present such behavior, we expect that compressibility is used in conjunction with nonlinear complexity to filter out sequences having this type of deficiency. A specific class of sequences achieving high nonlinear complexity and low compression ratio is identified in this thesis, the so-called s-optimal sequences, thus revealing the cryptographic value of compressibility.

5 Efficient computations of best quadratic approximations of Boolean functions

This section presents new efficient formulas for determining best quadratic approximations of boolean functions. The main result is the following.

Theorem 8 ([9]) *Let $f \in \mathbb{B}_n$ be a cubic function, where there exists variable x_j such that $f = (q + l_0) \parallel_j (q + q_j + l_1)$ (q, q_j are quadratic). Then, the best quadratic approximations of f have one of the following forms*

i. $\xi_f^0 = (q + l_0) \parallel_j (q + l_1 + \lambda_{q_j})$;
ii. $\xi_f^1 = (q + q_j + l_0 + \lambda_{q_j}) \parallel_j (q + q_j + l_1)$.

Corollary 1. *The second order nonlinearity of any cubic function $f \in \mathbb{B}_n$ of the above form is equal to $\mathcal{NQ}_f = 2^{n-2} - 2^{n-2-h_{q_j}}$, for some $1 \leq h_{q_j} \leq \lfloor (n-1)/2 \rfloor$.*

The above is proved by means of special properties that characterize the Walsh transform of quadratic boolean functions. The importance of Theorem 8 rests with the fact that it enables direct computation of all the best quadratic approximations of a particular subset of cubic Boolean functions on n variables, which have a variable being present in all cubic terms, by determining the best affine approximations of quadratic Boolean functions on $n-1$ variables; direct formulas for determining these best affine approximations are also proved in this thesis, which exploit the representation of quadratic functions according to Dickson's theorem [18], without using the Walsh transform. Cubic functions of the form described in Theorem 8 have been recently proposed for contemporary stream ciphers, thus revealing the cryptographic importance of the above result.

Best quadratic approximations of functions with degree 4 are also proved.

Theorem 9 ([9]) *Let $f \in \mathbb{B}_n$ be a Boolean function of degree 4, and let $f = f_0 \parallel_j f_1$, for some $1 \leq j \leq n$, such that f_0 is cubic function of the form described in Theorem 8 and $f_1 = q + l$ is a quadratic function, where q, l are its quadratic and linear part respectively. If $\mathcal{NL}_{f_0+q} \leq 2^{n-2} - 2^{n-4}$, then all functions*

$$g = (q + \lambda_{f_0+q}) \parallel_j f_1 \tag{15}$$

are best quadratic approximations of f and $\mathcal{NQ}_f = \mathcal{NL}_{f_0+q}$. Otherwise, it holds $\mathcal{NQ}_f > 2^{n-2} - 2^{n-4}$.

It is evident from the above that constructions of Boolean functions based on the concatenation of low-degree functions with fewer number of variables are susceptible to successful best quadratic approximation attacks if the sub-functions are not properly chosen, and in particular if the resulting Boolean function has low second order nonlinearity. Since many constructions of bent functions or correlation-immune functions (both admitting important cryptographic properties) present this structure, our results determine new design principles that need to be considered in constructions of boolean functions, so as to guarantee resistance in low order approximation attacks [10].

6 Conclusions

This thesis studies cryptographic features of sequences and Boolean functions, by using signal processing techniques. This leads to new methods for constructing cryptographic primitives achieving good cryptographic properties. Research is progress focuses on generalizing the results of Section 5 in a wider class of Boolean functions, while the security of several contemporary stream ciphers with respect to low order approximations is also currently studied. Moreover, the connection between several other cryptographic criteria of sequences, apart from nonlinear and Lempel-Ziv complexity, remains an interesting open problem.

References

1. Carlet, C.: Recursive lower bounds on the nonlinearity profile of Boolean functions and their applications. Cryptology ePrint Archive, Report 2006/459 (2006) http://eprint.iacr.org.
2. Courtois, N., Meier, W.: Algebraic attacks on stream ciphers with linear feedback. Advances in Cryptology - Eurocrypt '03 (Lecture Notes in Computer Science, Springer-Verlag) **2656** (2003) 345–359.
3. Erdmann, D., Murphy, S.: An approximate distribution for the maximum order complexity. Des. Codes and Cryptography **10** (1997) 325–339.
4. Jansen, C. J., Boekee, D. E.: The shortest feedback shift register that can generate a given sequence. Proc. Advances in Cryptology-CRYPTO '89 (1990) 90–99.
5. Kalouptsidis, N.: Signal Processing Systems. Telecommunications and Signal Processing Series, John Wiley & Sons (1996)
6. Kalouptsidis, N. and Limniotis, K.: Nonlinear span, minimal realizations of sequences over finite fields and De Brujin generators. Proc. Int. Symp. Inf. Theory and Appl., (2004) 794–799.

7. Key, E. L.: An analysis of the structure and complexity of nonlinear binary sequence generators. IEEE Trans. Inform. Theory **22** (1976) 732–736.
8. Kolokotronis, N., Kalouptsidis, N.: On the linear complexity of nonlinearly filtered PN-sequences. IEEE Trans. Inform. Theory **49** (2003) 3047–3059.
9. Kolokotronis, N., Limniotis, K. and Kalouptsidis, N.: Best affine approximations of boolean functions and applications to low order approximations. Proc. IEEE Int. Symp. Inf. Theory (2007) 1836–1840.
10. Kolokotronis, N., Limniotis, K. and Kalouptsidis, N.: Efficient computation of the best quadratic approximations of cubic boolean functions. 11th IMA International Conference on Cryptography and Coding, (Lecture Notes in Computer Science, Springer-Verlag) **4887** (2007) 73–91.
11. Kurosawa, K., Iwata, T., Yoshiwara, T.: New covering radius of Reed-Muller codes for t-resilient functions. IEEE Transactions on Information Theory **50** (2004) 468–475.
12. Lempel, A., Ziv, J.: On the complexity of finite sequences. IEEE Trans. Inform. Theory **22** (1976) 75–81.
13. Lidl, R., Niederreiter, H.: Finite Fields. vol. 20 of Encyclopedia of Mathematics and its Applications. Cambridge University Press 2nd ed. (1996)
14. Limniotis, K., Kolokotronis, N., and Kalouptsidis, N.: On the nonlinear complexity and Lempel–Ziv complexity of finite length sequences. IEEE Trans. Inform. Theory **53** (2007) 4293–4302.
15. Limniotis, K., Kolokotronis, N., and Kalouptsidis, N.: New results on the linear complexity of binary sequences. IEEE Int. Symp. Inf. Theory, (2006) 2003–2007.
16. Limniotis, K., Kolokotronis, N., and Kalouptsidis, N.: Nonlinear complexity of binary sequences and connections with Lempel–Ziv compression. Sequences and Their Applications, Berlin, Germany: Springer-Verlag, **4086**, (2006) 168–179.
17. Limniotis, K., Kolokotronis, N., and Kalouptsidis, N.: On the linear complexity of sequences obtained by state-space generators. IEEE Trans. Inform. Theory **54** (2008) 1786–1793.
18. MacWilliams, F. J., Sloane, N. J. A.: The Theory of Error Correcting Codes. Amsterdam, The Netherlands: North-Holland (1977).
19. Massey, J. L.: Shift register synthesis and BCH decoding. IEEE Trans. Inform. Theory **15** (1969) 122–127.
20. Massey, J. L., Serconek, S.: Linear complexity of periodic sequences: a general theory. in Proc. Advances in Cryptology - CRYPTO '96, Lecture Notes in Computer Science 358–371.
21. Matsui, M.: Linear cryptanalysis method for DES cipher. Advances in Cryptology - Eurocrypt '93 (Lecture Notes in Computer Science, Springer-Verlag) **765** (1993) 386–397.
22. Menezes, A. J., van Oorschot, P.C., Vanstone, S.A.: Handbook of Applied Cryptography. CRC Press (1996).
23. Niederreiter, H.: Some computable complexity measures for binary sequences. C. Ding, T. Helleseth, and H. Niederreiter, eds., in: Sequences and Their Applications, Discrete Mathematics and Theoretical Computer Science, Springer-Verlag (1999) 67–78.
24. Rizomiliotis, P., Kalouptsidis, N.: Results on the nonlinear span of binary sequences. IEEE Trans. Inform. Theory **51** (2005) 1555–1563.
25. Rizomiliotis, P., Kolokotronis, N., Kalouptsidis, N.: On the quadratic span of binary sequences. IEEE Trans. Inform. Theory **51** (2005) 1840–1848.
26. Rueppel, R. A.: Analysis and design of stream ciphers. Berlin, Germany: Springer-Verlang (1986).
27. Ziv, J., Lempel, A.: Compression of individual sequences via variable-rate coding. IEEE Trans. Inform. Theory **24** (1978) 530–536.

High density Integrated Optoelectronic Circuits for High Speed Photonic Microsystems

K. Minoglou*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications

k.minoglou@imel.demorkitos.gr

Abstract. The study of high density integrated optoelectronic circuits involves the development of hybrid integration technologies and the generation of models for the optoelectronic devices. To meet these goals, in the beginning a methodology for the heterogeneous integration of epitaxial GaAs wafers with fully processed standard bipolar complementary metal-oxide-semiconductor Si wafers is presented. The complete low-temperature wafer bonding process flow, based on SOG/SIO_2, including procedures for the Si wafer planarization and GaAs substrate removal, has been developed and evaluated. The implementation of an in-plane optical link, consisting of an edge-emitting laser diode, a waveguide and a photodiode, is demonstrated. Further investigation on heterogeneous integration is achieved by presenting a second methodology. The integration of complete optoelectronic dies, consisting of optical sources and detectors connected by waveguides for the employment of a photonic layer above CMOS integrated circuits has been proposed. Photonic dies are integrated to CMOS circuits through a novel metallic bonding technique that utilizes a thin multilayer structure of the Au-20Sn eutectic alloy along with a starting layer of a rare earth element (Gd). Its main advantage is the accomplishment of mechanical bonding and electrical connectivity of the heterogeneous devices in a single step. The study of photonic microsystems demands also the modeling of specific OE devices. Under this scope, an efficient model scheme that combines the non-linear behavior of the input parasitics with the intrinsic fundamental device rate equations of the Vertical Cavity Surface Emitting Lasers (VCSELs) is proposed. A systematic methodology for the model parameter extraction from dc and ac, electrical and optical measurements, is also presented and simulation results are compared with the experimental measurements. Extraction and simulation procedures are implemented in commercial integrated circuit design tools and they are proved to be very fast while they preserve adequate accuracy.

Keywords: heterogeneous integration, SOG/SiO_2, photonic link, metallic bonding, Au-80Sn, VCSELs, circuit model, parameter extraction, rate equations.

* Dissertation Advisor: Dimitris Syvridis, Professor

1 Hybrid Integration-Technologies

High-density optical interconnections require the integration of III–V optoelectronic (OE) devices along with Si integrated circuits (ICs). Currently, hybrid integration, based on flip-chip bonding, is the most mature technology for the combination of III–V optoelectronics with Si ICs [1]-[2]. The main drawback of the hybrid approach is the demanding fabrication sequence which results in increased manufacturing cost. A different approach for the integration of Si ICs with III–V OE devices is the fabrication of the OE devices in layers grown by heteroepitaxy on the Si wafer. However, the heteroepitaxial approach suffers from poor III–V material quality. Furthermore, process incompatibilities and cross contamination problems, between the complementary metaloxide-semiconductor (CMOS) and III–V technologies, cannot be easily eliminated.

In order to overcome the limitations of the hybrid and heteroepitaxial integration, we have developed a process having the wafer-scale characteristics of heteroepitaxial integration and at the same time being compatible with commercial bipolar CMOS (BiCMOS) technology. Our first approach is a low-temperature (LT) process, employing SiO_2 and spin-on glass (SOG) layers for the planarization and bonding of epitaxial GaAs wafers onto fully processed BiCMOS wafers. The second approach for hybrid integration that we propose is based on the use of a metallic alloy. This metallic bonding methodology is introduced for bonding OE dies above CMOS circuitry and it is based on the use of a proposed multilayer structure of Au-20Sn eutectic alloy over a thin film of Gd, as bonding agent.

1.1. Bonding with SOG/SiO_2

The basic process flow for the wafer-scale integration scheme starts with the fabrication of the ICs on a 4 in. Si wafer using a commercially available BiCMOS technology and the OE device layers are grown on a 3 in. GaAs wafer. The OE device layers are grown with the inverse sequence of the regular device structure, after the epitaxy of a thin AlAs etch-stop layer. Then, the fully processed Si wafer is covered by successive SiO_2 layers, deposited at low temperature (LT) using plasma enhanced chemical vapour deposition (PECVD), and the surface is planarized by chemical mechanical polishing (CMP). Next, the planarized and polished SiO_2 top layer is covered by SOG, to eliminate any remaining surface imperfection and to act as a bonding agent. Subsequently, a proper baking procedure is followed to remove the volatile elements from the SOG before the wafer bonding. The Si wafer is then bonded at room temperature face to face with the epitaxial GaAs wafer. The two wafers are centered and their major flats are mechanically aligned. The bonding is further strengthened by an annealing step performed at 200°C (the temperature must be kept below 250°C to avoid debonding due to the difference in the thermal coefficients of expansion (TCE) of GaAs and Si). The backside of the entire GaAs substrate is then removed by an appropriate thinning process so that only the epitaxial III–V structure remains bonded onto the top surface of the BiCMOS wafer. After this stage of the process flow the

temperature can rise up to 400 °C because the constraint of the different thermal coefficients of expansion becomes more tolerant due to the preceding thinning of the GaAs wafer, which can now sustain larger deformations. The remaining III–V film is thereafter processed to form OE devices using conventional III–V processing technology. The mask alignment of the OE devices to be fabricated on the bonded GaAs film can be performed either by infrared backside alignment or by alignment marks placed on the uncovered annulus of the 4 in. silicon wafer. The final steps of the process are: (a) the opening of via holes through the III–V film and the insulating intermediate layers and (b) the fabrication of electrical interconnections between the Si ICs and the OE devices using either wire bonding or on-wafer metalizations.

In order to evaluate the critical steps of the process flow, we have undertaken a series of experiments focusing on different aspects of the process. Both unprocessed and fully processed BiCMOS 4 in. Si wafers were bonded with 3 in. epitaxial GaAs wafers. The processed Si wafers were fabricated with the commercially available 0.8 mm BiCMOS technology of Austria Mikro Systems AG (AMS™). The epitaxial GaAs wafers consisted of various GaAs/AlGaAs heterostructures [3] grown by molecular beam epitaxy (MBE). The investigation of the LT SiO_2 /SOG based wafer bonding process has been initiated with the examination of the surface morphology of the commercial fully processed BiCMOS wafers. Atomic force microscopy and profilometry showed that the two-level metal interconnection stripes presented unevenness of 0.6–1.3 mm. Furthermore, irregular edges of 2 mm height existed in the IC testing pads opened within the capping nitride layer. This surface unevenness was reduced to about 90 nm by the planarization procedure. For that purpose, multiple low-temperature oxide layers, with 2mm total thickness, were deposited by PECVD and then polished using fumed silica as polishing media. Further planarization was accomplished by deposition of a SOG layer so as to achieve the bondability condition (microroughness under 1 nm). The average interface energy of the bonded surfaces was measured using the crack opening method by inserting a blade between the bonded wafers. The interface energy, before annealing, was 0.46 J/m^2 and reached the value of 1.4 J/m^2 after annealing at 200°C. The backside thinning of the bonded GaAs substrates has been investigated using two different methods. In one approach, CMP was used to thin the 400-mm-thick GaAs substrates to approximately 30 mm while in the second approach the GaAs substrates were thinned to an average thickness of 80 mm, using a fast wet etching process (using H_3PO_4:H_2O_2:H_2O). Laser diodes (LDs) have been fabricated on films bonded onto fully processed BiCMOS wafers and compared to similar devices processed on GaAs substrates.

In order to demonstrate the feasibility of the entire process, a complete in-plane OE link [4] was designed and fabricated. The link consists of a BiCMOS laser diode driver (LDD), an edge-emitting LD, a planar waveguide (WG), a photodiode (PD) and a BiCMOS transimpendance amplifier (TIA) [5]. The output of the laser driver was designed to be at a distance of 500 mm from the input of the TIA. In the space of 500 mm are placed the LD, the WG and the PD so as to form a complete OE link. The microphotograph of the integrated OE link is shown in Fig. 1.

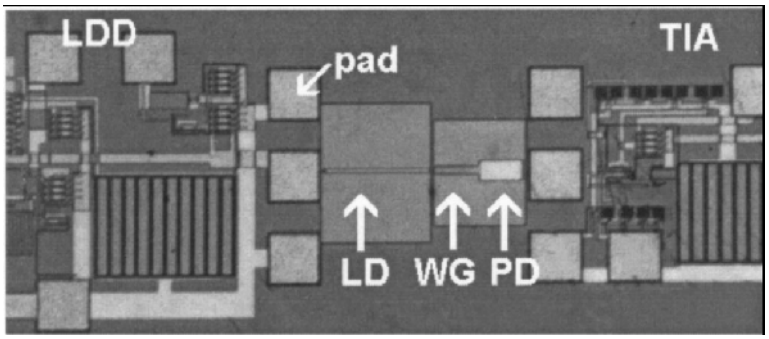


Fig. 1. Nomarski optical microscopy image showing the fabricated LD-WG-PD GaAs optical link.

The entire OE link was also successfully tested, unpackaged, for its functionality under low duty-cycle pulsed signal operation (0.4ms pulse width and 10KHz repetition rate). High frequency characterization has not yet been attempted because the self-heating effects on the unpackaged OE link can be deleterious for the OE devices. Furthermore, the influence of the wafer bonding and the subsequent GaAs process on the Si ICs has been examined using on-wafer *S*-parameter measurements up to 8GHz. Discrete devices as well as Si ICs were characterized before and after the wafer bonding process and no apparent performance degradation was observed

1.2. Metallic Bonding

Environmental friendly processes restrict solders used in packaging to the Pb-free group. On the other hand, successful flip-chip assembly is favored by eutectic alloys. Since die attachment is realized by solder fusion, a low melting point is desirable for CMOS and OE devices thermal reliability. Various candidates, such as Sn-Ag [6-7], In-Ag [8] or Au-Sn [9-10] alloys fulfill the former requirements. The reasons for selecting 80/20 weight percent (w.t.%) Au-Sn alloy among them are its physical and thermal properties. As a hard solder it remains in elastic deformation and the thermal expansion coefficient (TEC) mismatch becomes important only when bonding large chips. Moreover, its high thermal conductivity provides adequate power dissipation. It is compatible with III-V device metallization and its advantage over other hard solders is the comparatively low melting temperature (280^{o}C), as verified by the binary phase diagram of the Au-Sn alloy [11].

In order to control accurately the alloy composition and accomplish adequate intermixing during fusion, a multilayer structure of the alloy is adopted. In addition to that, rare-earth participation in the solder is exploited by the incorporation of Gd, since rare-earth elements improve adhesion to passivating surfaces, such as SiO_2. The following step is to integrate the optical link over CMOS circuitry lying in wafers with a passivated surface. Therefore, sample bonding via metallization on bare Si as well as on SiO_2 is investigated. Consequently, an initial thin film of 50nm vacuum-evaporated Gd, which corresponds to 3.5w.t.% of the solder, can precede alternating Sn and Au

layers of appropriately selected thickness preserving the Au-20Sn eutectic alloy composition. The Sn, Au layers, also deposited by evaporation in vacuum to reduce oxidation, ranged between 250nm and 750nm in total thickness. Au has been used as a cap layer to prevent alloy surface oxidation. Either full SOI-wafer to CMOS-wafer bonding or SOI-die to CMOS-wafer bonding can be followed.

Alloy patterning capability by standard lithography processes has been verified by fabricating pads of varying thickness and diameter on 4-inch Si wafers. The ability of precise patterning over both planar and structured substrates has been examined. Since contact pads provide point-to-point electrical connectivity to optoelectronic/CMOS devices, their pitch is critical in case of alloy spreading during fusion, which could cause short circuits and therefore device malfunction. To confirm the degree of spreading, samples containing pad arrays were flipped over SiO_2 substrate, while they were annealed at 330^{o}C and pressed by 2.5MPa. An alloy spreading less than 1μm has been observed. This fact enables dense CMOS/optoelectronic device integration.

A variety of bonding experiments have been conducted in order to determine the conditions to be followed for successful bonding results. All bonding experiments took place at an annealing temperature of 330^{o}C in ambient forming gas (95%Ar-5%H_2) for 40min. The annealing temperature of 330^{o}C is within the allowed post-processing thermal range for CMOS device functionality [12-13]. On the other hand, Sn is expected to segregate to the surface of the Au-rich alloy, due to its lower surface free energy compared to that of Au. This creates a Sn-rich surface layer, which is oxidized even in a very small amount of O_2, i.e. in ambient 95%Ar-5%H_2 gas, despite the retarding influence of H_2 in oxidation. The surface oxide formation, which is a major obstacle for the bonding process, is broken up with pressure application during annealing.

The samples used in bonding experiments have a surface ranging from 0.25cm^2 to 1.5cm^2. Samples have been fabricated by direct deposition of the Gd-(Au-20Sn) alloy on substrates such as Si, SiO_2 and BCB on Si. Additional samples were fabricated by depositing the Gd-(Au-20Sn) alloy on Al or Ti/Cu layers over Si and SiO_2 substrates to examine the possibility of bonding above standard CMOS metallization schemes.

Table I summarizes the parameters for five experiments using pairs of dies with different topologies. The total alloy thickness is 800nm, consisting of 50nm Gd followed by 750nm of Sn/Au alternating layers. In the cases (c)-(c) and (d)-(d) 16 cross-section squares are formed with an area of either (100x100)μm^2 or (50x50)μm^2, respectively. In Table I, S_{total} stands for the total die area, S_{bond} refers to the contact area between flipped samples and R is the overlap ratio of S_{bond} over S_{total}. P is the pressure applied on the bonding surface (actually flipped dies accept the same force in all cases, while the pressure increases due to decrease of the contacting area) and V_{bond} stands for the total alloy volume participating in bonding. Die shear strength tests for 0.25cm^2 samples have shown a 25g/mm^2 shear stress to be sufficient for detaching the die off the host substrate, revealing edge effects prohibiting the achievement of adequate bonding strength. However, for 1.5cm^2 samples with bonding area and overlap ratio larger than 2mm^2 and 1.1% respectively, shear strength has exceeded 2.5kg, complying with MIL-STD-883G, Method 2019.7 [14].

1. DIE PAIR	1. (a)-(a)		1. (a)-(c)	1. (a)-(b)		1. (c)-(c)	1. (d)-(d)
2. S_{total} (mm^2)	2. 25	2. 150	2. 150	2. 25	2. 150	1. 150	1. 150
3. S_{bond} (mm^2)	3. 25	3. 150	3. 4	3. 1.1	3. 2	2. 0.16	2. 0.04
4. R (%)	4. 100	4. 100	4. 2.67	4. 4.4	4. 1.3	3. 0.11	3. 0.03
5. P (MPa)	5. 0.2	5. 0.03	5. 1.2	5. 4.5	5. 2.5	4. 30.6	4. 122
6. V_{bond} 7. ($10^{-3}mm^3$)	6. 50	6. 240	6. 6.4	6. 2.2	6. 3.2	5. 0.26	5. 0.06

TABLE I. Parameters of bonding experiments with the 800nm thick Gd-(Au-20Sn) alloy

Moreover, since the proposed methodology is oriented to wafer scale integration, bonding experiments of multiple dies over specific regions of an entire 4-inch wafer have also taken place (Fig. 2). In this case, coarse alignment using alignment marks on the wafer surface and accurate die dicing has been successfully accomplished. For the fine alignment of dies to wafer pads with 10μm accuracy, either a passive or an active alignment methodology should be used [15]. An alignment technique using convex and concave features on both surfaces is under development.

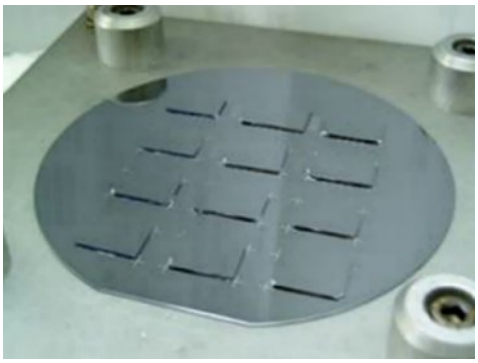


Fig. 2. Four-inch Si wafer with multiple dies bonded on specific regions

The metallic bonding technique possesses the advantage of achieving adhesion and electrical connectivity in a single step. The electrical properties of the Au-20Sn alloy are derived through a series of electrical measurements before and after it is annealed. Electrical tests are performed on mm-long alloy lines, which vary from 50μm 100μm in width. Carrier confinement across such lines ensures consistent results out of I-V testing.

Typical I-V characteristics for a 4mm long and 50μm wide line before annealing as well as for a 4mm long and 100μm wide line after annealing at 330°C are shown in Fig. 3. These I-V curves translate into a sheet resistance of 150mOhm/sq and 1.1Ohm/sq before and after the annealing respectively.

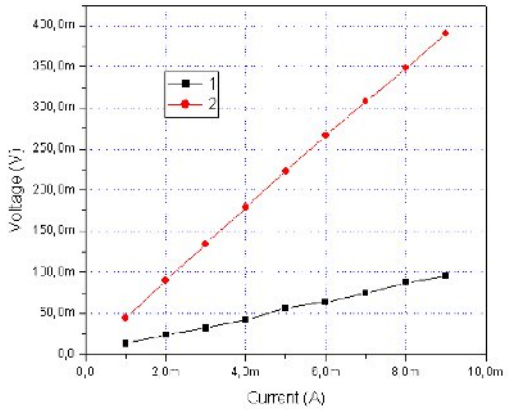


Fig. 3. I-V curves for alloy lines 4mm long and (1) 50 μm wide before annealing, (2) 100μm wide after annealing

2 OE Device Modeling

In recent years the Vertical Cavity Surface Emitting Lasers (VCSELs) have emerged and threaten to supplant the standard laser technology in a variety of applications, such as short haul high speed networks and Micro-Opto-Electro-Mechanical Systems (MOEMS). Two-dimensional VCSEL arrays are key components for achieving the highest aggregate bandwidths in tomorrow's parallel optical transceivers. Fabrication of 2D VCSEL array with 4 x 8 elements [16] as well as successful integration of 2x10 arrays of VCSELs with gigabit-per-second CMOS circuits have been reported [17].Thus, motivated by the fact that the ability to model VCSELs is critical to the design and analysis of optoelectronic microsystems, we propose a new model scheme, based on equivalent circuit optimization methodology, which will combine the non-linear behavior of the input parasitics with the intrinsic fundamental device rate equations. The complexity of the model requires a systematic methodology for the extraction of the parameters. The performed dc measurements at different temperatures as well as the ac s-parameter measurements provide a combination of electrical and optical data sufficient to define all the related parameter values. The results are validated through comparisons with simulation and measured data taken from commercially available VCSEL devices. The complete extraction procedure which has been developed proved to be very fast while it preserves adequate accuracy.

2.1. VCSEL model

The proposed circuit model for a packaged VCSEL is illustrated in Fig. 4. L_0, R_0 and C_0, model the connection to the measurement equipment and inductance L_p with capacitance C_p represent the parasitics of the package-leads as well as the wire- bonds of the package. The intrinsic VCSEL is modeled by a series resistance R_a in shunt with a non-linear capacitance C_j and the combination of a non-linear temperature-dependent current-controlled voltage-source *Einp* with a series resistance R_{int} and an ideal diode D_{vcsel}. Intrinsic voltage drop *Einp* and the intrinsic capacitance of the VCSEL are modeled according to the semi-empirical equation and to the junction diode's equation presented in [18]. The internal device temperature *T,* the carrier density and the photon density, which is equivalent to the output optical power *L*, are dynamically calculated using the respective rate equations. Moreover non-linear gain and transparency number and temperature dependent leakage current are included in the model.

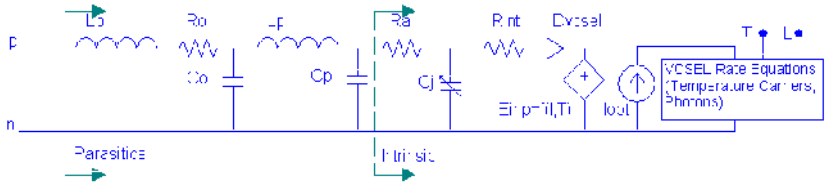


Fig.4. The proposed circuit model for the VCSEL

Since a circuit simulator is a differential equation solver, the rate equations can be solved with such a tool by mapping the dynamic quantities (i.e. the electron and photon populations) into node voltages, which are dynamically calculated. Working towards this direction, we have implemented all rate equations based on the analysis of Mena [19] in OPSIM™ of Anacad®. As an example in fig.5 is presented the equivalent circuit (with the expressions for the circuit elements) that corresponds to the following photon density equation:

$$\frac{dS}{dt} = -\frac{S}{\tau p} + \frac{\beta}{\tau n} No + \frac{Go \cdot zn\left(\gamma_o No - \gamma_1 N1 - \gamma_o Nt \cdot zn\right) S}{1 + \varepsilon S}$$

Where τp, τn, β, *No*, *Go*, γ_o, γ_1 and ε are model parameters and *zn* is an arbitrary constant used for convergence purposes.

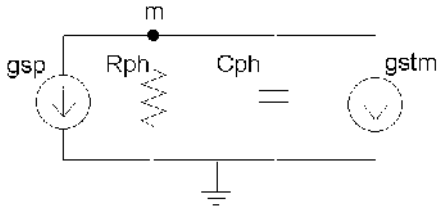


Fig.5. Equivalent circuit for the photon rate equation

$$g_{sp} = \frac{\tau p \cdot \beta \cdot k_f \cdot No}{\tau n \cdot \left(V(m) + \delta_m\right)}$$

$$g_{stm} = \tau p \cdot zn \cdot G_o \cdot \frac{\left(\gamma_o No - \gamma_1 N1 - \gamma_o N_t zn\right) \cdot \left(V(m) + \delta_m\right)}{1 + \varepsilon \dfrac{\left(V(m) + \delta_m\right)^2}{k_f}} - \delta_m$$

$$C_{ph} = 2 \cdot \tau p$$

$$R_{ph} = 1 \qquad and \qquad S = \frac{\left(V(m) + \delta_m\right)^2}{k_f}$$

2.2. Parameter extraction procedure

Due to the large number of the model parameters (16 for the input circuit and 25 for the rate equations) a three-step parameter extraction methodology is proposed to estimate them by dividing them into distinct groups. The parameter estimation is achieved by using I-L-V dc characteristics measured at four ambient temperatures and S_{11} and optical signal ac responses for various bias currents. The optimization algorithm is taken from the OPSIM™ tool, a modified Levenberg-Marquardt method, which efficiency and robustness have been proven by years of usage. In the first step the dc dependent parameters of the input circuit (such as $R_{in,}$ R_{int} and R_a) are estimated using as input to the optimization tool the dc current-light (I-L) characteristics and as targets the measured I-Vs. In the second step, using the previously calculated values, the optimization target is changed to S_{11} vector measurements and the remaining parameters of the input circuit, which influence its ac behavior (such as L_o, R_o, C_o, L_p, and C_p) are estimated. In the above procedures the rate equation that affects the results is only the thermal one, which is used to determine the internal device temperature. Thus, in the third step, the parameters of the carrier and photon rate equations as well as gain, transparency number and leakage current parameters are estimated using as optimization targets the dc I-L characteristics and the ac optical response. In Table II it is presented the flowchart of the generic algorithm used for the complete parameter extraction procedure. It is clearly illustrated the crucial dependence of each step on the results of the previously completed ones. Also Table II summarizes the grouping of the parameters depending on which extraction step procedure is applied. As it is shown in Figures 6 and 7 satisfactory agreement between measured values and simulation results for a commercially available VCSEL is achieved using the proposed model and extraction methodology. Uniqueness of the extracted parameters is not a constraint since the only limitation of the method is the convergence between measured values and simulation results.

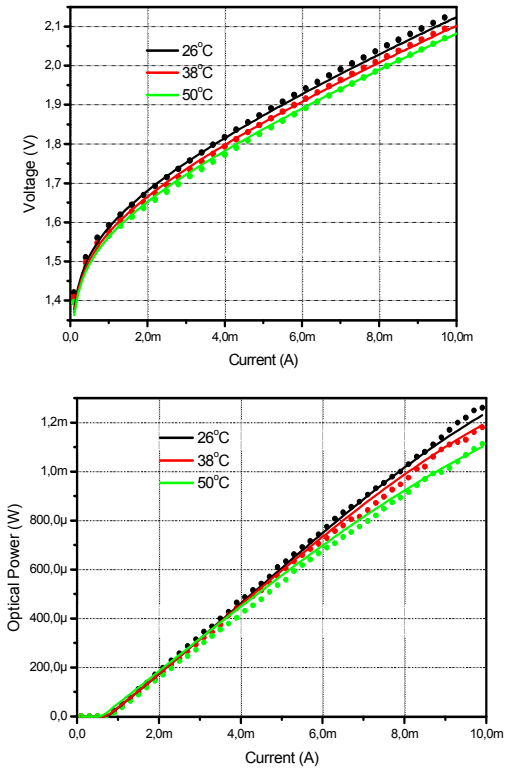


Fig.6 Measured values (dots) and simulated (continuous line) (a) I-V and (b) I-L characteristics at 26ºC, 38ºC and 50ºC.

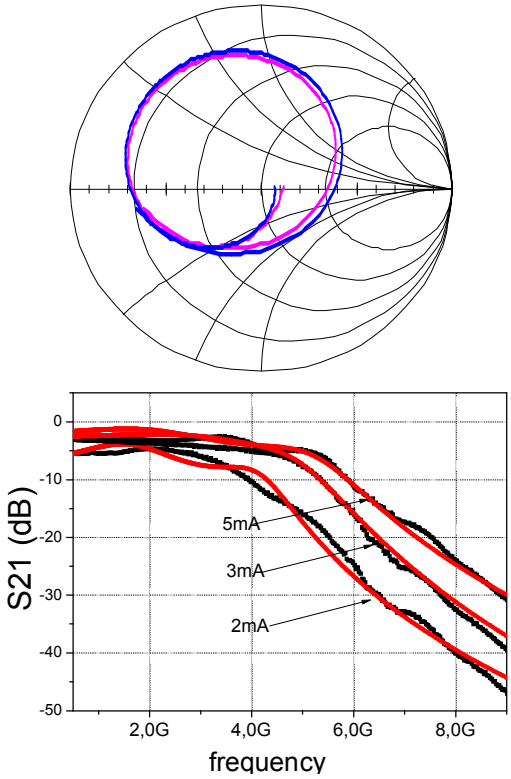


Fig.7 Measured (dots) and simulated (continuous line) S_{11} parameter at a bias of 3mA and S_{21} at 3 different biases

2.3. Driving VCSELs

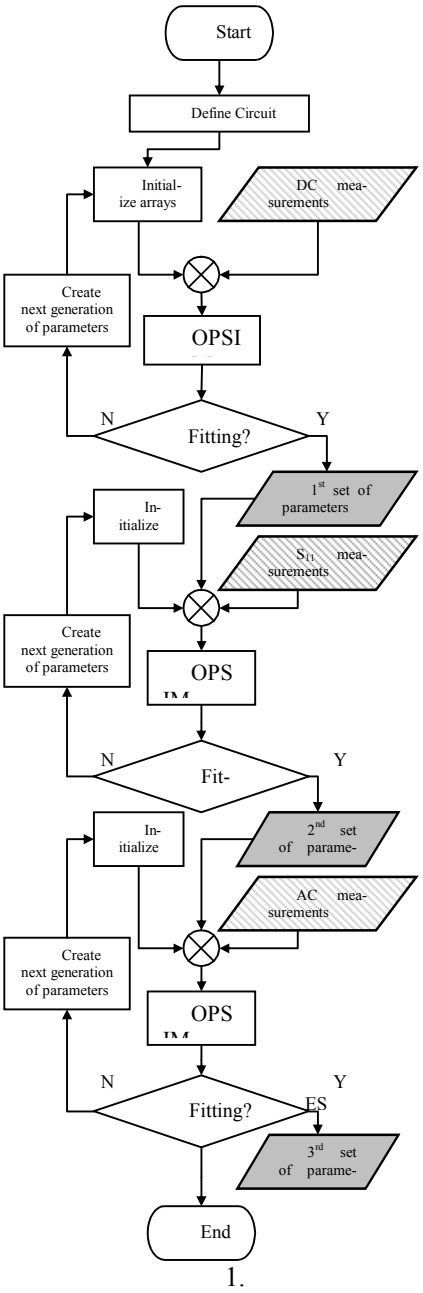
The design of a laser diode driver (LDD) is a challenging task since clear eye diagrams at high speeds require abrupt signal edges and low jitter. Traditionally, the LDDs were designed to drive Edge-Emitting Lasers (EELs) and were based on differential pair topologies acting as current-pulse sources for the laser diodes. This choice of the circuit topology is straightforward for the EELs exhibiting low series resistance very often combined with considerable series parasitic inductance. At first glance, the design task of LDDs for VCSELs seems relaxed due to the lower threshold and modulation currents required by these devices. However, high-speed VCSELs present significantly larger series resistance in combination with relatively high shunt capacitance than EELs leading to distinct requirements and consequently different design

approaches for the LDD. The particular aspects of the VCSEL drivers become crucial especially at the multi-Gb/s data rates because the packaging parasitics further aggravates the shunt capacitance effect. A comparison of the small signal and the transient response of high-speed VCSELs using ideal as well as realistic current and voltage drivers is performed. A non-linear VCSEL equivalent model is implemented and the associated parameters are extracted from dc and ac measurements of a commercially available packaged device. Simulations, using the extracted realistic VCSEL models, show that the employment of voltage drivers results in an improvement of 74% of the 3-dB cut-off frequency of the optical current signal. Moreover, an 80% reduction of the rise and fall time and a 66% reduction of the signal delay are observed.

3 Conclusions

We have proposed two different approaches for hybrid integration of OE devices on Si wafers. The reported results on the methodology using SOG/SiO_2 have clearly demonstrated the feasibility of the monolithic-like process flow for the integration of GaAs OE devices on fully processed Si ICs. Further improvement of the planarization of the BiCMOS wafer will greatly upgrade the value of the entire process. Also, a metallic bonding technique permitting dense integration of photonic structures on CMOS wafers has been proposed. A multilayer metallization structure with well-controlled alloy composition has been developed and precise as well as uniform lithographic patterns of the bonding alloy have been reproduced at a 4-inch scale. The ability to achieve bonding of various passivatiing surfaces and standard interconnect metallizations has been shown.

Investigating OE devices, a compact and efficient model for the VCSEL that models by means of equivalent circuits the fundamental device rate equations, the thermal effects, the non-linear gain and transparency number functions and the input parasitics elements has been presented. The parameter extraction is based on standard dc and ac measurements and it is achieved by a three-step procedure, which divides model parameters into distinct groups.



1.

	Symbol	Description
1st step	R_{in}	*internal resistance at temperature T_0*
	T_0	*effective reference*
	n_f	*thermal-voltage sensitivity factor*
	B	*saturation current at temperature T_0*
	C	*material related constant*
	n	*temperature exponent of the saturation current*
	R_a	*access resistance*
	R_{int}	*series resistance*
2nd step	C_{jo}	*zero bias junction capacitance*
	Φ_o	*built-in junction voltage*
	m_p	*grading coefficient*
	C_p, L_p	*parasitics of the package*
	C_0, L_0, R_0	*parasitics of the connector*
3rd step	Il_0,a_0,a_1,a_2,a_3	*Leakage parameters*
	Nt_0	*Temperature independent transparency*
	cn_0,cn_1,cn_2	*Transparency number fitting constants*
	G_0	*Temperature independent gain constant*
	$ag_0,ag_1,ag_2,$ bg_0,bg_1,bg_2	*Gain fitting constants*
	T_{therm}	*Thermal time constant*
	a	*Ratio W_m/W*
	β	*Spontaneous emission coupling coefficient*
	h_1	*Parameter modeling diffusive effects*
	ε	*Gain saturation factor of mode 0 due to mode*
	k_f	*Output-power coupling coefficient of*
	t_p	*Photon lifetime*
	t_n	*Carrier lifetime*
	n_i	*Current injection efficiency*

1.

TABLE II Flowchart of the 3-step parameter extraction procedure and model parameters

References

1. A. V. Krishnamoorty, K. W. Goosen, L. M. F. Chirovsky, R. G. Rozier, P. Chandramani, W. S. Hobson, S. P. Hui, J. Lopata, J. A. Walker, and L. A. D'Asaro, IEEE Photonics Technol. Lett. **12**, 1073 (2000).
2. D. A. B. Miller, SPIE Critical Reviews of Optical Engineering, Heterogeneous Integration: Systems on a Chip **CR70**, 80 (1998).
3. A. Georgakilas, M. Alexe, G. Deligeorgis, D. Cengher, E. Aperathitis, M. Androulidaki, S. Gallis, Z. Hatzopoulos, and G. Halkias, in *CAS 2001 Proceedings* (IEEE, Piscataway, NJ, 2001), p. 239.
4. W. J. Grande, J. E. Johnson, and C. L. Tang, Appl. Phys. Lett. **57**, 2537 (1990).
5. G. Halkias, N. Haralabidis, E. D. Kyriakis-Bitzaros, and S. Katsafouros, in Proc. of IEEE International Symposium on Circuits and Systems 2000 (2000), p. V417.
6. Shangguan, D., Achyuta, A. and Green, W., "Application of Lead-Free Eutectic Sn-Ag Solder in No-Clean Thick Film Elecronic Modules", *IEEE transactions on Components, Packaging, and Manufacturing Technology-Part B*, Vol. 17, No. 4 (1994), pp. 603-611.
7. Bigas, M. and Cabruja, "Characterisation of electroplated Sn/Ag solder bumps", *Microelectronics Journal*, Vol. 37 (2006), pp. 308-316.
8. Chen, Y.-C., So, W.-W. and Lee, C.-C., "A Fluxless Bonding Technology Using Indium-Silver Multilayer Composites", *IEEE transactions on Components, Packaging, and Manufacturing Technology-Part A*, Vol. 20, No. 1 (1997), pp. 46-51.
9. Matijasevic, G. S., Lee, C. C. and Wang, C. Y., "Au-Sn alloy phase diagram and properties related to its use as bonding medium", *Thin Solid Films*, Vol. 223 (1993), pp. 276-287.
10. Buene, L., "Characterization of evaporated gold-tin films", *Thin Solid Films*, Vol. 43 (1977), pp. 285-294.
11. Massalski, T. B., Binary Alloy Phase Diagrams, 2nd ed. Metals Park, ASM (1992), Vol. 3, pp. 2863.
12. Takeuchi, H., Wung, A., Sun, X., Howe, R. T. and King, T.-J., "Thermal Budget Limits of Quarter-Micrometer Foundry CMOS for Post-Processing MEMS Devices", *IEEE transactions on Electron Devices*, Vol. 52, No. 9 (2005), pp. 2081-2086.
13. Sedky, S., Witrouw, A., Bender, H. and Baert, K., "Experimental Determination of the Maximum Annealing Temperature for Standard CMOS Wafers", *IEEE transactions on Electron Devices*, Vol. 48, No. 4 (2001), pp. 377-385.
14. Department of Defence, United States of America, Test Method Standard Microciruits, (2006), MIL-STD-883G, METHOD 2019.7, Die Shear Strength, pp. 1-6.
15. Sasaki, J., Itoh, M., Tamanuki, T., Hatakeyama, H., Kitamura, S., Shimoda, T. and Kato, T., "Multiple-Chip Precise Self-Aligned Assembly for Hybrid Integrated

Optical Modules Using Au-Sn Solder Bumps", *IEEE transactions on Advanced Packaging*, Vol. 24, No. 4 (2001), pp. 569-575.

16.R. King, D. Wiedenmann, P. Schnitzer, R. Jäger, R. Michalzik and K. J. Ebeling, "Single-Mode and Multimode 2D VCSEL Arrays for Parallel Optical Interconnects", IEEE Int. Semiconductor Laser Conference,Nara,Japan, 103 (1998).

17. A.V.Krishnamoorthy, et. al., "Vertical-Cavity Surface-Emitting Lasers Flip-Chip Bonded to Gigabit-per-Second CMOS Circuits", Photon. Tech. Lett., **11**(1), 128 (1999).

18. .K.Minoglou,, E.D. Kyriakis-Bitzaros, D. Syvridis, G. Halkias., "A compact non linear equivalent circuit model and parameter extraction method for packaged high-speed VCSELs", J. Lightwave Technology, **22**(12), 2823 (2004).

19. Mena, P.V., Morikuni, J.J., Kang, S.M., Harton, A.V., and Wyatt, K.W, "A comprehensive circuit-level model of Vertical-Cavity Surface-Emitting Lasers", J. Lightwave Technology, **17**(12), 2612 (1999)

Cube-Lifecycle Management and Applications

Konstantinos Morfonios*

National and Kapodistrian University of Athens, Department of Informatics and Telecommunications, University Campus, 15784 Athens, Greece
kmorfo@di.uoa.gr

Abstract. A common operation involved with the majority of algorithms relevant to On-Line Analytical Processing is aggregation, which can be extremely time-consuming if applied over large datasets. To overcome this drawback, scientists have proposed the precomputation and materialization of a large volume of aggregated data into a structure called data cube. Nevertheless, the construction and usage of the data cube itself has been found very demanding in terms of computational and storage resources. In the thesis summarized here [1] (hereafter called the thesis), we study this problem in depth and propose comprehensive suites of scalable algorithms that perform efficient cube construction, storage, query answering, incremental updating, indexing, and caching. Our extensive experimental evaluation indicates that our solutions are viable even when applied over very large datasets with arbitrary hierarchies. Some key points in our work include the introduction of a novel storage scheme for cubes that is based on the use of row-id references, a new external partitioning algorithm, an efficient construction plan, and the use of on-demand approaches for processes that are too expensive to perform in a given window of time. A unique property of all our algorithms is their compatibility with the relational model, which makes them easy to incorporate into existing relational servers. Finally, as an application in data mining, we study the usage of aggregate queries for solving the problems of feature selection and classification and propose a disk-based, lazy, and accurate solution, which exhibits great potentials in a broad range of applications.

1 Introduction

Modern data analysis "mines" knowledge from data stored in database systems discovering trends useful for decision making. To achieve such knowledge discovery, analysts pose complex queries that extensively use aggregation in order to group together "similarly behaving tuples". The response time of such queries over extremely large fact tables in modern data warehouses can be prohibitive. This inspired Gray et al. [2] to propose the implementation of the data cube. Implementation of the data cube is one of the most important, albeit computationally expensive, processes in On-Line Analytical Processing (OLAP). It

* Dissertation Advisor:Yannis Ioannidis, Professor

involves the computation and storage of the results of aggregate queries grouping on all possible dimension-attribute combinations over a fact table in a data warehouse. Such precomputation and materialization of (parts of) the cube is critical for improving the response time of OLAP queries and of operators such as roll-up, drill-down, slice-and-dice, and pivot, which use aggregation extensively [2]. Materializing the entire cube is ideal for fast access to aggregated data but may pose considerable costs in computation and maintenance time, as well as in storage space.

In order to overcome this problem and balance the tradeoff between query-response times and cube-resource requirements, implementation of the complete data cube has been studied using various data structures to construct and store the cube. In general, the data-cube implementation algorithms that have been proposed in the literature can be partitioned into four main categories, depending on the format they use in order to compute and store a data cube. On the one hand, Relational-OLAP (ROLAP) [2–5] and Multidimensional-OLAP (MOLAP) [6] methods use materialized views and multidimensional arrays, respectively, focusing mainly on the efficient sharing of computational costs (like sorting or hashing) during cube construction. On the other hand, Graph-Based approaches [7, 8] exploit specialized graphs (that usually take the form of tree-like data structures) in order to compute and store cubes more efficiently. Finally, Approximation-Based methods [9, 10] exploit various in-memory representations (like histograms), borrowed mainly from statistics. In this thesis, we focus on ROLAP methods and do not further study methods that belong to the other categories for the following reasons:

- MOLAP methods are poor performers when data is sparse, which is the case in most real-life applications. Although challenged by some, this has been observed by many researchers [3, 4].
- Graph-Based methods appear to have superior performance for cube construction and storage, but are currently not supported by any widely used product. Hence, they require nontrivial implementation effort, mainly due to the use of specialized data structures and algorithms.
- Approximation-Based methods generate and store approximate results, which are much more difficult to manage at run time compared to precise results generated by ROLAP methods.

As we show in this thesis, existing ROLAP methods that implement data cubes focus mainly on construction and storage of flat cubes (i.e., cubes constructed over flat datasets). The lifecycle of a data cube, however, does not involve (off-line) construction and storage only, but also query answering and incremental maintenance. Moreover, real-world datasets are not always "flat" but are usually organized in hierarchies, whose nature introduces several complications into all phases of the cube lifecycle, making existing techniques essentially inapplicable in a significant number of real-world applications. To overcome these problems, in this thesis, we develop comprehensive ROLAP solutions that address efficiently all functionality in the lifecycle of a cube and can be implemented easily over existing relational servers. They are families of algorithms

developed around novel, purely-ROLAP construction methods that provide fast computation of a fully-materialized cube in compressed form, are incrementally updateable, and exhibit fast query-response times that can be improved by low-cost indexing and caching, even when dealing with very large datasets with arbitrary hierarchies. The efficiency of our methods is demonstrated through comprehensive experiments on both synthetic and real-world datasets, whose results have shown great promise for the performance and scalability potential of the proposed techniques, with respect to both the size and dimensionality of the fact table.

The rest of this paper is organized as follows: In Section 2, we provide a detailed description of the data-cube implementation problem, focusing mainly on ROLAP techniques, and define some basic terminology that we use throughout this paper. Then, in Section 3, we summarize the main contributions of the thesis. Finally, we conclude in Section 4 and describe the directions of our future work.

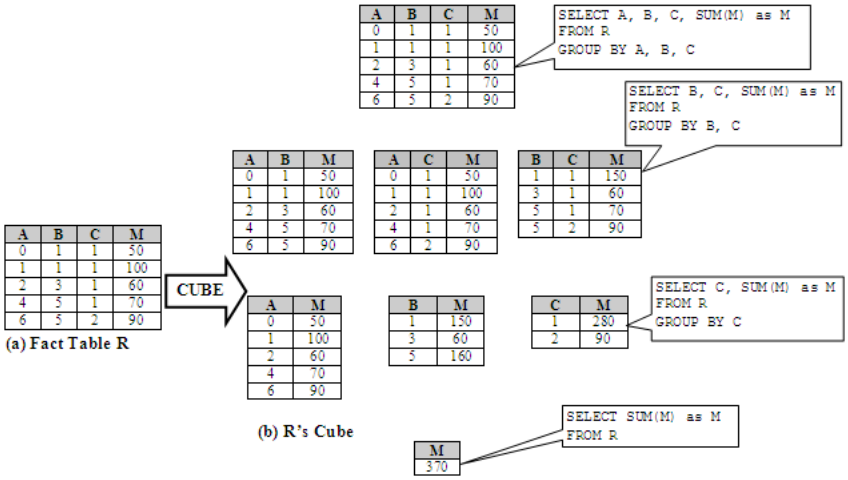


Fig. 1. Fact table R_1 and its data cube

2 Problem Description

Consider a fact table R_1 (Figure 1a) consisting of three dimensions (A, B, C) and one measure M. Figure 1b illustrates the corresponding cube. Each view that belongs to the cube (also called cube node hereafter) materializes a specific group-by query as shown in Figure 1b. Clearly, assuming that the data in the fact table is flat, i.e., that it is not organized in hierarchies, if D is the number of dimensions of a fact table, the number of all cube nodes is 2^D. (The situation becomes even more challenging in the presence of hierarchies.) The factor 2^D

implies that the cube size is exponentially larger with respect to D than the size of the original data (in the worst case). In typical applications, this is in the order of gigabytes, so development of efficient data-cube implementation algorithms is extremely critical.

A common representation of the data cube that captures the computational dependencies among different group-by queries is the cube lattice [11]. Let "grouping attributes" be the fields of a table that participate in the group-by part of a query. All group-bys computed for the data-cube construction can be partially ordered in a lattice structure [11], which is a directed acyclic graph (DAG), where each node represents a group-by query on the fact table and is connected via a directed edge with every other node that contains exactly one fewer grouping attribute. The source of an edge is called its parent node and the destination of the edge is called its child node. The node whose grouping attributes consist of all dimensions (highest) is called the root of the lattice. The node whose grouping-attribute set is empty (lowest) is called the ALL node. For example, Figure 2 presents the lattice that corresponds to the fact table R_1 (Figure 1a). Nodes in higher lattice levels contain more grouping attributes and are thus more detailed (hold data of finer granularity) than nodes in lower levels, which are more specialized (hold data of coarser granularity). Exploiting the "1-1" mapping between group-by queries and lattice nodes, we call node size the size of the result of the corresponding query.

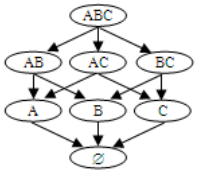


Fig. 2. Example of a cube lattice

In most cases, computation of the data cube is an extremely time-consuming process due to the size of the original fact table and the exponential number of cube nodes with respect to the number of dimensions. Hence, it is common in practice to select a proper subset of the data cube, pre-compute it off-line, and store it for further use. We call this process "data-cube implementation". Over the last years, there has been intense interest in efficient data-cube implementation and a variety of methods has been proposed for this task.

For ROLAP approaches, data-cube implementation consists of data-cube computation and data-cube selection. These, however, are not necessarily two separate procedures applied sequentially. A-priori knowledge of the strategy used for selection may affect the computation procedure and vice-versa. Hence, several methods have been proposed that combine fast computation and efficient selection in one step; we call them integrated methods.

3 Thesis Contribution

Taking into account that existing ROLAP methods exhibit weaknesses in one or more issues related to the lifecycle of data cubes, in this thesis, we propose novel techniques that deal with the problem in a comprehensive fashion, providing efficient solutions for the entire cube lifecycle, including construction and storage, indexing, caching, query answering, and incremental maintenance. We study techniques that are applicable not only on flat data cubes but on datasets that include hierarchies as well. The results of a rather extensive experimental evaluation on both real-world and synthetic datasets are very positive, giving strong indications on the power of our methods and of ROLAP overall.

In more detail, we start with a study of related work and provide a qualitative examination of ROLAP methods. We go beyond a simple review of existing algorithms and contribute to the following issues [12]:

- **Comprehensive Review:** We offer a comprehensive review of existing cubing methods that belong to the ROLAP framework, highlighting interesting concepts from various publications and studying their advantages and disadvantages. To the best of our knowledge, this is the first overview of its kind.
- **Problem Parameterization:** We carefully analyze existing ROLAP cubing methods and identify six orthogonal parameters/dimensions that affect their performance, namely "Traversal of Cube Lattice", "Partitioning of Original Fact Table", "In-Memory Processing Algorithm", "Storage Granularity", "Selection Criterion", and "Lattice Reference". The resulting 6-dimensional parameter space essentially models the data-cube implementation problem.
- **Method Classification:** We place existing techniques at the appropriate points within this parameter space and identify several clusters that these form. These clusters have various interesting properties, whose study leads to the identification of particularly effective values for the space parameters and indicates the potential for devising new, better-performing algorithms overall.

Based on the findings of the aforementioned thread of our work, we investigate new methods and propose a suite of algorithms that exhibit efficiency in all phases of the lifecycle of flat cubes. In particular, we focus on the following issues [13]:

- **Cube Construction:** We incorporate redundancy reduction into the dominant ROLAP cube construction methods and devise three new cubing algorithms, which exhibit considerable reduction in the size of the cube to be stored as well as some minor reduction in its construction time. We study their behavior under large and multi-dimensional datasets and show that the best among them, namely TRS-BUC, is the first ROLAP method that scales well up to at least 25 dimensions.
- **Query Answering:** Under a comprehensive query model, which is broader than the models used in the past for the same purpose, we evaluate novel

algorithms for answering queries on top of the cubes produced by the new methods and demonstrate that the cube resulting from TRS-BUC exhibits significantly better query execution performance compared to all earlier techniques, including those considered as "champions" with respect to construction time and storage space.
- **Incremental Maintenance:** We introduce novel incremental update algorithms and demonstrate that those based on TRS-BUC exhibit again significantly better performance compared to their counterparts. Moreover, they produce a cube identical to the one that would be produced by full reconstruction, i.e., TRS-BUC preserves its compact format unaffected, regardless of the frequency of updates.
- **Indexing and Caching:** We propose indexing and caching schemes and study their effect on both query answering and incremental updating of ROLAP cubes. Our experiments show that TRS-BUC is the only known method that can benefit significantly from such techniques, consuming inexpensive additional resources.

Moving on to hierarchical datasets, we show that the nature of hierarchies introduces several complications in all phases of the cube lifecycle that render existing ROLAP techniques (including those built around TRS-BUC) impractical. To overcome this drawback, we propose an extension of TRS-BUC, called CURE [14], and revisit all its surrounding algorithms related to cube usage. CURE contributes a novel lattice traversal scheme, an optimized data partitioning method, and a suite of relational storage schemes for all forms of redundancy. The last two are useful to "flat" datasets as well, but they are mostly necessary in the presence of hierarchies. In more detail:

- **Lattice Traversal with Dimension Hierarchies:** To the best of our knowledge, CURE is essentially the first comprehensive ROLAP solution capable of constructing a complete cube not only at the leaf level of each dimension hierarchy, but also at all higher levels, precomputing group-by queries at all granularities. To achieve this, CURE uses an efficient way of traversing an extended lattice that includes dimension hierarchy levels (first proposed elsewhere [11]), which enables great cost sharing of sorting operations through pipelining.
- **External Partitioning:** We propose an efficient algorithm for partitioning fact tables that store hierarchical data of any size into memory-fitting segments, while computing a very small subset of the cube using inexpensive additional resources. Exploiting this early-computed data, CURE accelerates the construction of the final cube significantly, making it feasible even when the original fact table is extremely large. Existing techniques partition data according to values in a single dimension and require that segments of tuples with the same value in this dimension fit in memory. As shown in this thesis, however, this is not always possible in cases that include hierarchies, due to small domain sizes at coarse granularities.
- **Efficient Storage:** Unlike previous ROLAP methods that rely only on avoiding redundant-tuple storage for cube size reduction, we further study

alternative schemes for storing nonredundant data efficiently as well. To the best of our knowledge, CURE is the only ROLAP method that condenses the cube both by rejecting all kinds of redundancy and by further exploiting appropriate data representations.
- **Query Answering:** We develop a straightforward algorithm for answering arbitrary queries using data materialized in an (unindexed) CURE cube and show that its practicability is limited in real-world applications that typically involve selective queries over large datasets. To overcome this, we investigate the effect of indexing on CURE cubes and propose an efficient extension of the original algorithm that is based on low-cost indexes. We show that indexing the entire cube, which is potentially very expensive in hierarchical cubes, is not necessary; indexing only the original fact table is enough, primarily because of the particular storage format of CURE cubes.
- **Query Optimization:** We examine customized query optimization policies to identify when using an index is beneficial and to indicate which index combination to use for a given query, based on cost estimations.
- **Incremental Maintenance:** We study different approaches for the incremental maintenance of CURE cubes and conclude that common eager tactics that refresh a cube periodically during a dedicated window of time are prohibitively expensive, due to the storage format of CURE and the nature of hierarchies. Alternatively, we propose a novel lazy method that only performs some lightweight operations during the update window and updates data on-line, when necessary, during query processing. Interestingly, the additional query cost is marginal making the lazy approach the method of choice. To the best of our knowledge, this is the first time a lazy method has been applied to a cube-related method. Finally, we propose a hybrid combination of the eager and the lazy method, which is very promising under certain conditions.

Having found efficient algorithms for materializing and using a large number of aggregate views, we study an application of aggregation on data-mining techniques, including classification and feature selection. In this thread of our work, we propose LOCUS, a lazy classifier implemented exclusively with aggregate queries expressed in standard SQL. To the best of our knowledge, LOCUS is the first disk-based lazy classifier with all the following properties [15]:

- **Classification Accuracy:** It exhibits good classification accuracy, which improves as training sets become larger. This can be justified theoretically based on its convergence to the optimal Bayes classifier, which minimizes the classification error probability. The same is also verified experimentally in comparison to Decision Trees, a very popular and accurate existing classifier.
- **Disk-Based Implementation:** It is database-friendly and, to the best of our knowledge, the only lazy method that uses a small and constant number of highly selective range queries in order to classify unknown objects. Such queries actually need to access a very small part of the underlying database and have been well studied in the database literature. They can be expressed

in standard SQL and existing query optimizers guarantee their fast response times with the use of traditional indexes, e.g., B^+-Trees.
- **Feature Selection:** Its classification accuracy can be efficiently used as a promising criterion for feature selection as well, in the sense that features are selected based on the accuracy that the classifier achieves when applied on them.
- **Parallelization:** It can be efficiently parallelized with essentially unlimited scalability.

4 Conclusions and Future Work

In this thesis, we presented a comprehensive study of efficient (mostly ROLAP) algorithms related to data cubes. We started with a thorough review of existing algorithms for efficient data-cube implementation in a ROLAP environment and identified six orthogonal parameters: "Traversal of Cube Lattice", "Partitioning of Original Fact Table", "In-Memory Processing Algorithm", "Storage Granularity", "Selection Criterion", and "Lattice Reference". We placed the existing algorithms at the appropriate points within the problem space based on their properties. We observed that the algorithms form clusters, whose study led to the identification of particularly effective values for the space parameters.

Based on the findings of the first phase of our research, we incorporated redundancy reduction into the best existing pure ROLAP methods for cube implementation and proposed TRS-BUC and a suite of novel algorithms built around it that deal with all aspects of cube usage, including efficient construction, storage, query answering, incremental updating, indexing, and caching. To the best of our knowledge, this has been essentially the first such comprehensive approach to the problem in the ROLAP context, treating all the above aspects in an independent fashion.

Furthermore, we studied ROLAP cubing in the presence of hierarchies and presented CURE, a novel ROLAP cubing method that addresses all challenges imposed by the nature of hierarchies and constructs complete data cubes over very large datasets with arbitrary hierarchies. CURE introduced an efficient execution plan suitable for hierarchical cube construction and revisited external-partitioning and size-reduction methods, which are complicated due to the existence of hierarchies. The effectiveness of CURE has been demonstrated through experiments on both real-world and synthetic datasets (including the APB-1 benchmark in its highest density), which have given very promising results with respect to the potential of CURE overall. Moreover, we developed efficient algorithms for query processing and incremental updating over CURE cubes in the presence of hierarchies, including some lazy policies that were never applied on cubes before. Interestingly, our solutions are ROLAP compatible, matching the design goals of CURE.

Finally, we applied ideas borrowed from data cubing on data mining and proposed LOCUS, an accurate and efficient disk-based lazy classifier that is data-scalable and can be implemented using aggregate queries expressed in standard

SQL. We showed that in most cases LOCUS exhibits high classification accuracy, which improves as training sets become larger, based on its convergence to the optimal Bayes. Furthermore, we used its classification accuracy as an efficient and reliable criterion for feature selection and proposed parallelizing both the classification and feature-selection processes, essentially achieving unlimited scalability. Overall, the results are very promising with respect to the potential of LOCUS as the basis for feature selection and classification, especially over large or inherently complex datasets.

In the future, we plan to compare CURE directly with Dwarf [8] and QC-Tree [7], prominent cubing methods that use specialized graph-like data structures. We expect this comparison to reveal the fundamental strengths and weaknesses of the two underlying techniques. Furthermore, we are interested in applying a CURE-like algorithm for the construction of skyline cubes.

Finally, with respect to data mining, we plan to further explore our feature selection method in the direction of identifying multiple reliable nodes, to implement a parallel version of LOCUS, and to study its applicability in regression problems (possibly replacing the count aggregate function with average).

Although cubing is already more than 10 years old, it remains an exciting problem with several fundamental aspects as well as applications still remaining in the dark and waiting to be investigated.

References

1. Morfonios, K.: Cube-lifecycle management and applications. (Ph. D. Thesis. 2007)
2. Gray, J., Bosworth, A., Layman, A., Pirahesh, H.: Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-total. In: Proc. of International Conference on Data Engineering (ICDE). (1996) 152–159
3. Beyer, K.S., Ramakrishnan, R.: Bottom-up computation of sparse and iceberg cubes. In: Proc. of ACM Special Interest Group on Management of Data (SIGMOD). (1999) 359–370
4. Ross, K.A., Srivastava, D.: Fast computation of sparse datacubes. In: Proc. of Very Large Data Bases (VLDB). (1997) 116–125
5. Wang, W., Lu, H., Feng, J., Yu, J.X.: Condensed cube: An efficient approach to reducing data cube size. In: Proc. of International Conference on Data Engineering (ICDE). (2002) 155–165
6. Zhao, Y., Deshpande, P., Naughton, J.F.: An array-based algorithm for simultaneous multidimensional aggregates. In: Proc. of ACM Special Interest Group on Management of Data (SIGMOD). (1997) 159–170
7. Lakshmanan, L.V.S., Pei, J., Zhao, Y.: Qc-trees: An efficient summary structure for semantic olap. In: Proc. of ACM Special Interest Group on Management of Data (SIGMOD). (2003) 64–75
8. Sismanis, Y., Deligiannakis, A., Roussopoulos, N., Kotidis, Y.: Dwarf: shrinking the petacube. In: Proc. of ACM Special Interest Group on Management of Data (SIGMOD). (2002) 464–475
9. Gunopulos, D., Kollios, G., Tsotras, V.J., Domeniconi, C.: Approximating multi-dimensional aggregate range queries over real attributes. In: Proc. of ACM Special Interest Group on Management of Data (SIGMOD). (2000) 463–474

10. Vitter, J.S., Wang, M.: Approximate computation of multidimensional aggregates of sparse data using wavelets. In: Proc. of ACM Special Interest Group on Management of Data (SIGMOD). (1999) 193–204
11. Harinarayan, V., Rajaraman, A., Ullman, J.D.: Implementing data cubes efficiently. In: Proc. of ACM Special Interest Group on Management of Data (SIGMOD). (1996) 205–216
12. Morfonios, K., Konakas, S., Ioannidis, Y., Kotsis, N.: Rolap implementations of the data cube. (submitted)
13. Morfonios, K., Ioannidis, Y.E.: Supporting the data cube lifecycle: The power of rolap. (to appear). In: VLDBJ. (2006)
14. Morfonios, K., Ioannidis, Y.E.: Cure for cubes: Cubing using a rolap engine. In: Proc. of Very Large Data Bases (VLDB). (2006) 379–390
15. Morfonios, K., Ioannidis, Y.E.: Locus: Lazy optimal classification of unlimited scalability. In: HDMS. (2006) 80–89

Propagation limitations in all-optical networks due to nonlinear effects

Ioannis Neokosmidis*

Department of Informatics and Telecommunications, University of Athens
Panepistimiopolis Ilissia, Athens, Greece, GR-15784

i.neokosmidis@di.uoa.gr

Abstract. The subject of this PhD thesis is the study of the impact of the nonlinearity in the performance of an optical telecom system. In particular the FWM phenomenon which is detrimental in WDM systems is analyzed theoretically and numerically, while various methods are proposed for its compensation. Finally the stability of nonlinear soliton pulses in CROWs is proposed in order to realize compact optical delay lines.

Keywords: Chirp, Four-Wave Mixing, modulation formats**,** nonlinear optics, photonic crystals, solitons, wavelength-division multiplexing (WDM).

Introduction

In this thesis, propagation limitations in all-optical networks due to nonlinear effects and especially Four-Wave Mixing (FWM) [1]-[3] are studied. In the first part of this thesis an accurate analysis of the statistical nature of the FWM noise was carried out using Monte Carlo (MC) and Multicanonical MC (MCMC) simulations. Such an analysis is of great importance in network design and modeling since FWM is an important source of noise in dense wavelength-division multiplexing (DWDM) networks that employ nonzero-dispersion fibers (NZDF). In these systems, FWM causes the dominant nonlinear distortion, especially if high input powers, narrow channel spacing and dispersion compensation are used. The MCMC method was also applied to study the impact of IP traffic burstiness on the performance of an IP over MPλS-based DWDM network limited by FWM and In-band crosstalk.

In the second part of this thesis two new methods based on a hybrid amplitude-/frequency-shift keying (ASK/FSK) modulation and pulse prechirping are proposed for the suppression of the FWM effect. The next part addresses the issue of the choice of optical modulation format in a multispan WDM system using G.655 fiber. Various modulation formats are assessed by numerical simulation in the presence of fiber nonlinearities with FWM being the dominant effect. Finally, in the last part of this thesis, the propagation of optical slow light solitons in Coupled

* Dissertation Advisor: Thomas Sphicopoulos, Professor

Resonator Optical Waveguides (CROWs) is proposed as an alternative optical delay line design method for reducing dispersion effects.

Statistical nature of the FWM noise

In this section, the MCMC method was applied for the study of the statistical behavior of the FWM noise in a WDM network. At the receiver, the photocurrent is proportional to the optical power and hence to $|E|^2$ where $E=E^{(m)}$ or $E=E^{(s)}$. In practical applications, it can be assumed that $\Delta\beta>>a$ and $\exp(-aL)<<1$, where a, L and $\Delta\beta$ are the optical losses, the length of the optical link and the phase match factor [3] of the FWM process respectively. In this case the photocurrent at the detector is written as:

$$S^{(m)} = k\left|E^{(m)}\right|^2 \approx kP_n e^{-aL} + 2k\delta\sqrt{P_n e^{-aL}}\, I_m \tag{1a}$$

$$S^{(s)} = k\left|E^{(s)}\right|^2 \approx k\delta^2 I_s \tag{1b}$$

where k is the receiver responsivity, P_{n} is the input peak power of the channel n and

$$\delta = \frac{\gamma c}{2\pi\lambda^2 D\Delta f^2} P_{in}^{3/2} e^{-aL/2} \tag{2a}$$

$$I_m = \frac{1}{3}\sum_{pqr} B_p B_q B_r \frac{d_{pqr}}{|p-n||q-n|}\cos\left(\theta_{pqr} - \theta_n\right) \tag{2b}$$

$$I_s = \left(\frac{1}{3}\sum_{\substack{pqr \\ r\neq n}} B_p B_q B_r \frac{d_{pqr}}{|p-n||q-n|}\cos\theta_{pqr}\right)^2 + \left(\frac{1}{3}\sum_{\substack{pqr \\ r\neq n}} B_p B_q B_r \frac{d_{pqr}}{|p-n||q-n|}\sin\theta_{pqr}\right)^2 \tag{2c}$$

where γ is the nonlinear coefficient, D is the fiber chromatic dispersion coefficient, λ is the wavelength of the signal, c is the speed of light in vacuum, Δf is the channel spacing, d_{pqr} is the degeneracy factor ($d_{\mathrm{pqr}}=3$ when $p=q$, $d_{\mathrm{pqr}}=6$ when $p\neq q$) and θ_{n} is the input phase in the mark state, respectively, of the given channel n.

Equations (1a) and (1b) provide an expression for the photocurrent in the mark and space state in terms of two new variables I_{m} and I_{s} given by (2b) and (2c) respectively. It is interesting to note that for a given number of channels, these new variables depend only on the bits and the phases of the optical signals. Once the pdf of I_{m} and I_{s} is determined, the pdf of $S^{(\mathrm{m})}$ and $S^{(\mathrm{s})}$ can also be determined, using the theorem of transformation of random variables.

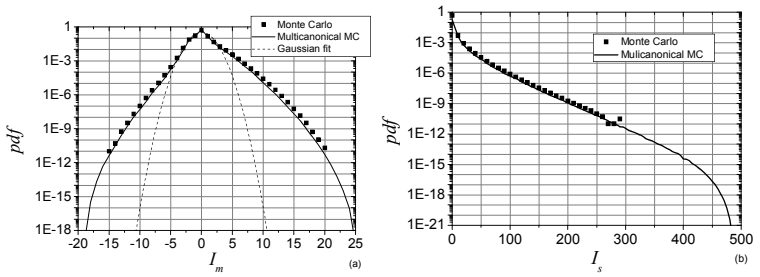


Fig. 1. The PDFs of a) I_m and b) I_s for N=16 calculated using the MCMC (solid lines) and conventional MC (dots) methods. Also shown with dotted lines in (a) is a Gaussian distribution with the same standard deviation as I_m.

The MCMC method [4] provides a simple, accurate and efficient tool for the computation of the PDF of the FWM noise (figure 1) and overcomes the limitations of the conventional MC method. The optical phases of all channels are assumed to be uniformly distributed within [0, 2π], due to phase noise, and the data bits are assumed to be in the mark and space state with equal probability, $P(B_i=0 \text{ or } 1)=1/2$. It is shown that the pdf in the mark state is not symmetric as assumed in previous studies [1], [2].

The MCMC method is also far more accurate than the Gaussian model (figure 1a) because it takes into account the correlation of the FWM noise components. Using the MCMC method it was also shown that the PDF converges quickly to its asymptotic form as the number of channels N increases (figure 2a). Indeed for N>32, the change in the PDF is hardly distinguishable even in a log-scale.

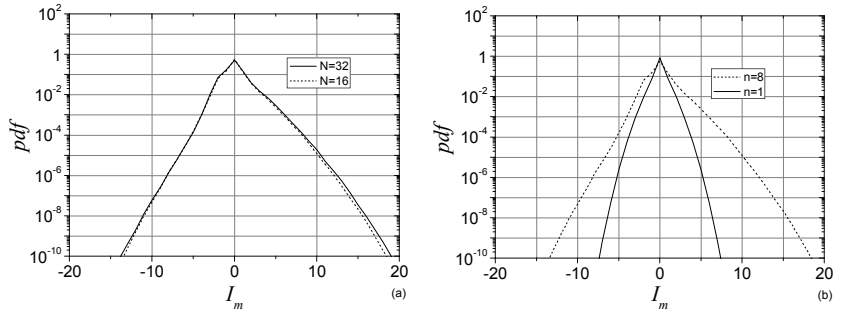


Fig. 2. a) The PDFs of I_m for the central channel in the case where N=16 (dashed lines) and N=32 (solid lines) and b) The PDF of I_m for N=16., n=8 (dashed lines) and N=16, n=1 (solid lines).

These results can prove useful for a system designer who wants to estimate the implications of the FWM-induced distortion in a WDM network. In particular, the error probability P_e can be computed from the pdfs S_m and S_s using numerical integration (figure 3).

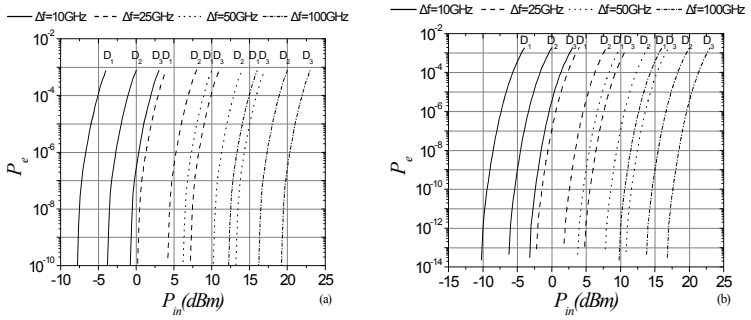


Fig. 3. BER as a function of the input peak power P_{in} in the mark state, for a) N=8 and b) N=16. The values of the chromatic dispersion used are $D_1 = 2\,ps/nm/Km$ and $D_2 = 5\,ps/nm/Km$

Impact of IP traffic Burstiness

The performance of Wavelength Division Multiplexing (WDM) optical networks can be severely degraded due to the existence of signal dependent noises such as the Four-Wave Mixing noise whose statistical behavior will depend on the statistics of the signal. To assess the implication of such noises, the bits are usually assumed to be in the mark and space state with equal probability. This approach however, does not take into account the effect of traffic burstiness in packet switched networks. In this section, the influence of traffic burstiness in the performance of the network is numerically analyzed in the case of an IP over WDM network. The network employs the MultiProtocol lambda Switching (MPλS) scheme [5] (figure 4a).

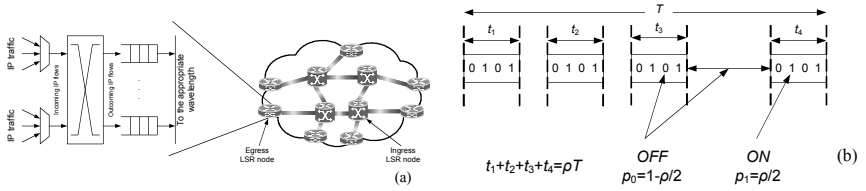


Fig. 4. a) The Label Edge Router of an IP over MPλS-based DWDM network and b) "1" and "0" generated under bursty traffic

At the ingress nodes, the packets are forwarded according to their wavelength. In order to take into account the traffic burstiness, each wavelength is modeled as an M/G/1 system. In this case the probabilities of "1" and "0" are p_0=1-ρ/2 and p_1=ρ/2 [6] where ρ is the traffic load (fig. 4b). The dependence of p_i on ρ is a first indication of the influence of traffic burstiness in the system performance: the decision variable in a FWM limited system depends on the bit statistics and hence on ρ.

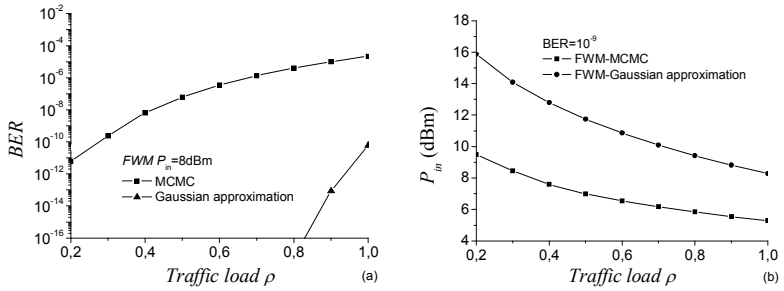
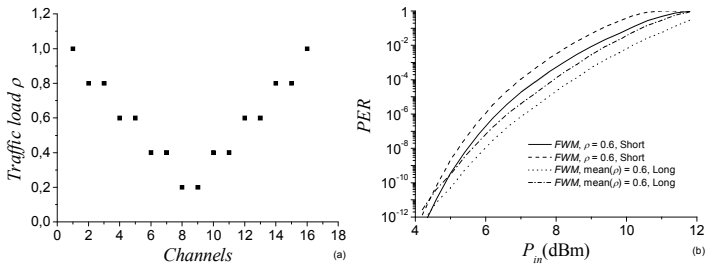


Fig. 5. Comparison between the MCMC method and the Gaussian approximation. a) BER vs traffic load ρ for P_{in}=8dBm and SXR=12dB and b) P_{in} and SXR for obtaining BER=10^{-9}.

The numerical analysis is carried out using the MCMC method. It is deduced that the network performance is intimately related to the traffic load (figure 5a). The Gaussian approximation based on the Central Limit Theorem cannot lead to accurate results in the presence of the FWM noise since the decision variable D can not be written as a sum of independent random variables. To further understand the implications of the error in the Gaussian approximation, the required values of the input power corresponding to a BER equal to 10^{-9}, are plotted in fig. 5b for various ρ.

It was previously shown that the FWM degradation was more severe for the central channels than the edge channels of a WDM system. It is then obvious that one way to reduce the FWM noise and increase the transmission power is to redistribute the traffic so that the edge channels carry heavier traffic than the central ones. Figure 6 illustrates that careful traffic engineering can improve the system performance in terms of the BER by at least one order of magnitude. These results imply that, when analyzing the performance of the network, traffic burstiness is an important issue that must be taken into account.



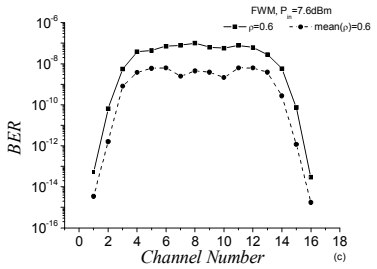


Fig. 6. a) Traffic load distribution along the channels, b) PER as a function of the input peak power P_{in} when all the channels are equally loaded with ρ=0.6 (solid line) and unequally loaded with mean ρ equal to 0.6 (dashed line) and c) BER values for all channels at P_{in}=7.6dBm.

New Techniques for the Suppression of the FWM-Induced Distortion in NZD Fiber WDM Systems

In order to reduce FWM-induced distortion two new techniques, the hybrid ASK/FSK modulation and the use of pre-chirped pulses, are proposed. In a WDM system with equally spaced channels, the central frequency of the products will coincide with some of the central frequencies of the channels. In order to reduce the number of FWM products that coincide with the WDM channels, one solution is to modulate the WDM signals using a special kind of FSK modulation. In the context of this special scheme, the WDM channels are divided into pairs and on each pair the channels follow the same FSK modulation.

Another solution for the reduction of the effect of the FWM induced distortion is optical pre-chirping. Since the efficiency of the FWM products are inversely proportional to the phase mismatch, it follows that reducing the phase coherence may reduce the power of the FWM noise. One way to reduce this coherence is through pulse pre-chirping. There are several methods to produce a pre-chirped signal such as cascading intensity and phase modulators or using dispersion-compensating devices like chirped fiber gratings and DCFs. In this work, the latter technique was chosen due to its ease of implementation.

It is shown that both techniques can greatly improve the *Q*-factor in a 10Gb/s WDM system (figure 7a). This happens even for very high input powers (~10dBm), where the degradation of the conventional WDM system is prohibitively high.

The proposed methods are also applied and tested in higher bit rates - 40Gbps (figure 7b). It is deduced, that although the hybrid ASK/FSK modulation technique marginally improves the system performance, the optical pre-chirp technique can still be used to greatly increase the maximum allowable input power of the system.

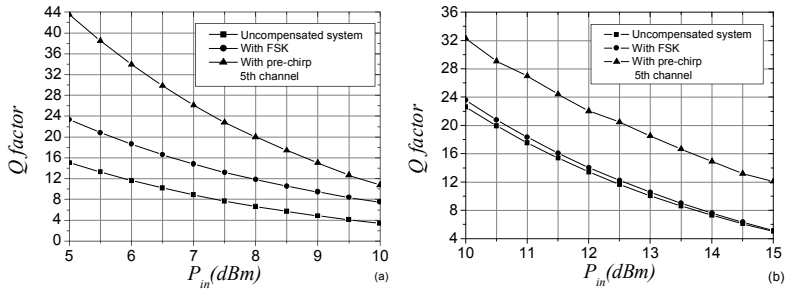
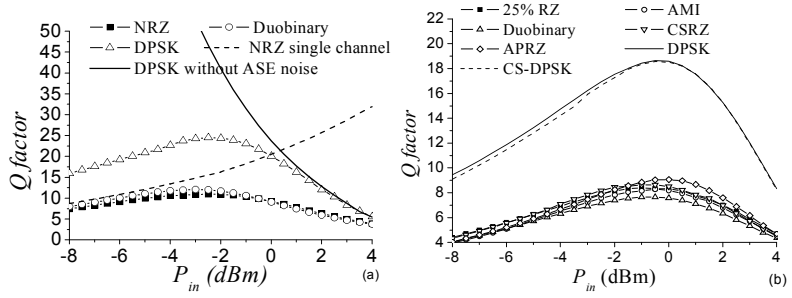


Fig. 7. Q factor of the central channel as a function of the input power P_{in} for an a) 10Gbps system with 8 channels and 50GHz channel spacing and b) 40Gbps system with 8 channels and 200GHz channel spacing

Non-linearity Tolerance of Optical Modulation Formats in NZD Fibers

This section addresses the issue of the choice of optical modulation format in a multi-span WDM system using G.655 fiber. Various modulation formats [7] are assessed by numerical simulation in the presence of fiber non-linearities with Four Wave Mixing (FWM) being the dominant effect. FWM noise has complex statistical behavior and thus the impact of the modulation format cannot be readily understood. Phase modulation formats are also degraded by the Gordon-Mollenauer (G-M) effect, which is due to the conversion of amplitude fluctuations created by the Amplified Spontaneous Emission (ASE) noise to phase fluctuations through Kerr non-linearity. Hence, one must resort to numerical simulations to accurately take into account the impact of fiber non-linearity. It is shown that the various amplitude modulation alternatives result in more or less the same performance. Phase modulation schemes such as DPSK [8], [9] drastically increase the system performance leading to an increase of the Q-factor by almost 3dB and of the optimum power by 1dB (figure 8).



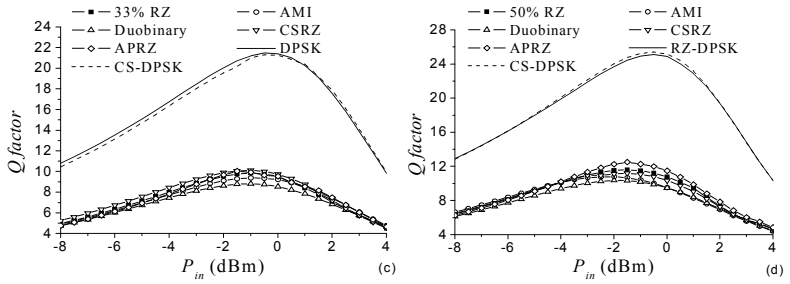


Fig. 8. *Q* factor as a function of input peak power for a) NRZ pulses, b) 25% RZ pulses, c) 33% RZ pulses and d) 50% RZ pulses

Optical Delay Lines Based on Coupled Resonator Optical Waveguide Soliton Propagation

Soliton Pulses are investigated as a means to mitigate dispersion-induced broadening and further reducing the group velocity in CROWs. High index contrast devices based on photonic wires or photonic crystals may enable further miniaturization and increased optical chip functionality. One of the main challenges in the field, is the realization of compact all-optical delay lines. Slow light structures such as coupled resonator optical waveguides (CROWs) [10] could possibly provide an attractive means of achieving large optical delays on a chip scale. As light slows down, the nonlinear properties are enhanced and this could also find applications in optical signal processing. Second and higher order dispersion however currently limits the performance of such devices [11].

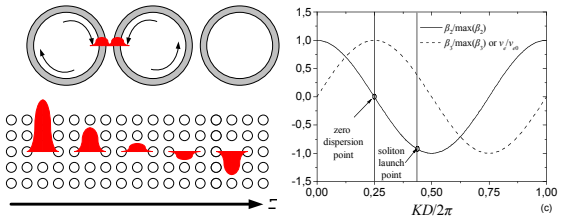


Fig. 9. Linear and nonlinear (soliton) propagation in CROWs

Figure 9 illustrates the group velocity second and third order dispersion coefficients v_g, β_2 and of a microring or photonic crystal defect CROW, where K is the propagation constant and D is the distance between the centers of two consecutive resonators. At $KD=\pi/2$, one obtains $\beta_2=0$ and hence second order dispersion does not cause pulse broadening. This is the point usually considered when the CROW is operated in the linear regime. Unfortunately however, the group velocity is maximum there implying small achievable delays and the signal can still be distorted by third

order dispersion. Although residual third order dispersion can in principle be compensated, the device length can still be long.
On the other hand, one may think of launching the pulse near $KD=\pi$, where the group velocity is minimum and $\beta_3=0$, while the second order dispersion can be compensated using the Kerr-induced, Self Phase Modulation (SPM) using sech soliton pulses. At this regime, the residual soliton broadening is only due to the fourth order dispersion β_4 and is much smaller than that experienced near the zero-dispersion point. These arguments point out that soliton pulses propagating at the nonlinear regime would experience much lower pulse broadening than linear pulses launched at the zero dispersion point. To ascertain whether CROW soliton propagation could indeed have a practical bearing on optical delay line design, one must evaluate the required pulse peak power and estimate the amount of optical losses that can be tolerated in the system. To describe optical propagation in CROWs, one uses the propagation equation:

$$j\left(\frac{\partial\Phi}{\partial t}+v_e\frac{\partial\Phi}{\partial z}+\frac{a}{2}\Phi\right)+\sum_{l\geq 2}j^{m(l)}\frac{\beta_l}{l!}\frac{\partial^l\Phi}{\partial z^l}+\gamma|\Phi|^2\Phi=0 \tag{3}$$

where Φ is the optical pulse enveloppe, v_e is the group velocity, a is the optical loss coefficient, β_l is the l^{th} order dispersion coefficient and γ is the SPM coefficient of the CROW, while $m(l)=\text{mod}(l,2)$. The values of these coefficients in terms of the CROW modal fields can be found in [12]. The initial soliton pulse is given by $\Phi_0\text{sech}(z/z_0)$, where z_0 is the soliton spatial width, related to the soliton duration t_0 by $t_0=z_0/v_e$ while Φ_0 is determined by $\Phi_0=|\beta_2|/\gamma z_0^2$.
To calculate the power, one uses the well known Poynting theorem formula,

$$P(z)=\langle\Pi(z)\rangle_t\cong\left\langle\int(\mathbf{E}\times\mathbf{H})\cdot\hat{z}dS\right\rangle_t \tag{4}$$

where **E** and **H** are the electric and magnetic field. The fields inside the CROW which can be calculated from the modes of the cavities and the envelope $\Phi(z,t)$ based on the tight binding approximation as outlined in [12]. The modes of the cavities can be calculated using the Plane Wave Expansion (PWE) method.

To examine the performance of the CROW delay line in the presence of higher order dispersion, one can numerically solve (4) including higher order dispersion terms (up to l=6). Figure 10a, depicts the broadening factor for the case of 2 rod spacing. As shown in figure 10a, in the case of two rod spacing between the cavities, the broadening exceeds 30% for A Slow Down factor $S=c/v_e=800$ in which case, $R=v_e/v_{e0}=0.01$. Note that a broadening of 30% roughly corresponds to a power penalty of 1dB. This implies that the nonlinear CROW can achieve about 100 times smaller delays than a linear CROW operated at resonance ($v_e=v_{e0}$). Note that $S=c/v_e=800$ implies that a 10ns delay (required to store 100 bits at 10Gb/s can be accommodated in just 3.8mm of propagation length.

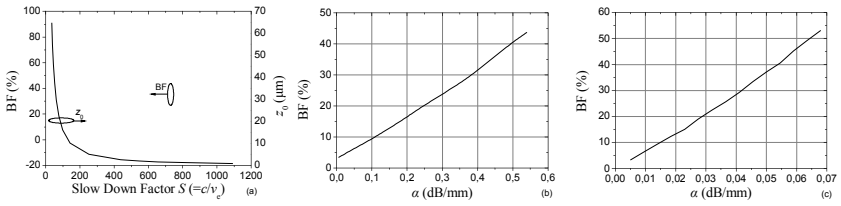


Fig. 10. a) Variation of *BF* and z_0 with respect to *S*, b-c) Variation of *BF* with respect to *a*

The influence of optical loss is depicted in figure 10b and 10c where the broadening factor is plotted as a function of the loss coefficient for the case of a 1ns and 10ns delay respectively. In both cases the spacing between two successive cavities is two rods and *R*=0.1.The figure suggests that soliton pulses are sensitive to loss. For example, in the case of 10ns delay, increasing the loss from a=0.03 dB/mm to *a*=0.04dB/mm leads to a 10% increase of the pulse width. A 30% broadening is obtained for *a*=0.042dB/mm and this indicates that the required loss value is very low, even compared to the present state-of-the-art loss values for ring resonator CROWs. These considerations demonstrate the importance of reducing the optical losses and towards this end, various distributed amplification schemes can be used such as Raman amplification or quantum wells. Efficient PC slab designs can also lead to optical loss reduction. For smaller delays, the required losses are somewhat relaxed and a 30% broadening is obtained for *a*=0.383dB/mm in the case of a 1ns specified delay.

Conclusions

In this thesis, the MCMC method has been applied for a study of the statistical behavior of FWM noise in a WDM network. The obtained results were used in estimating the performance of a system assuming continuous or bursty traffic models. The MCMC method was proved far more accurate than the Gaussian model because it takes into account the correlation of the FWM noise components. It was also illustrated that careful traffic engineering can improve the system performance in terms of the BER by at least one order of magnitude.

Two techniques—hybrid ASK/FSK modulation and prechirping the optical pulses—were proposed to suppress the FWM-induced distortion which can pose important limitations on the input power of a WDM system. It was proved that both techniques greatly improve the performance of a WDM system.

A performance comparison of modulation formats for a multispan WDM system with G.655 fibers was presented. It was shown that advanced amplitude formats do not result in increased performance while phase modulation schemes (DPSK and CSDPSK) are more advantageous.

Finally, the performance of a soliton-based CROW delay line where nonlinearity is used to compensate for the second order dispersion-induced pulse broadening was analyzed. A propagation equation describing the soliton propagation under the influence of higher order linear effects was given. Design equations were provided in

order to calculate the peak power of the soliton and to choose the soliton width in order to eliminate soliton attraction and collision. The importance of higher order nonlinear effects was also investigated. It was shown that if the optical losses are kept low, the soliton-based CROW delay line can achieve nanosecond delay at a propagation length of a few millimeters due to the high slow down factors obtained. The soliton-based delay line could, therefore, provide a path towards realizing integrated optical buffers.

References

1. Eiselt, "Limits on WDM Systems Due to Four-Wave Mixing: A Statistical Approach," *Journal of Lightwave Technology*, Vol. 17, pp. 2261-2267 (1999).
2 K. Inoue, K. Nakanishi, K. Oda and H. Toba, "Crosstalk and Power Penalty Due to Fiber Four-Wave Mixing in Multichannel Transmissions," J. Light. Techn. 12, 1423-1439 (1994).
3. S. Song, C. T. Allen, K. R. Demarest and R. Hui, "Intensity-Dependent Phase-Matching Effects on Four-Wave Mixing in Optical Fibers," J. Light. Techn. 17, 2285-2290 (1999).
4. R. Hozhonner and C. R. Menyuk, "*Use of multicanonical Monte Carlo simulations to obtain accurate bit error rates in optical communication systems*," Opt. Lett, 28, 1894-1896 (2003)
5. D. Awduche, Y. Rekhter, "Multiprotocol lambda switching: combining MPLS traffic engineering control with optical crossconnects," *IEEE Com. Magazine* 39, 111–116, (2001).
6 R. Cooper, *Introduction to Queueing Theory*. New York: North Holland 1981
7. P. J. Winzer, R.-J. Essiambre, "Advanced optical modulation formats," in Proc. of ECOC, Th2.6.1 (2003)
8 H. Kim and A. H. Gnauck, "Experimental investigation of the performance limitation of DPSK systems due to nonlinear phase noise," *Photonics Technol Lett*, vol. 15, No.2, pp. 320-322, February 2003.
9. C. C. Hiew *et al*., "BER Estimation of Optical WDM RZ-DPSK Systems Through the Differential Phase *Q*," *Photonics Technol. Lett*, vol. 16, No. 12, pp. 2619-2621, Dec. 2004.
10. A. Yariv, Y. Xu, R.K. Lee and A. Scherer, "Coupled-resonator opticalwaveguide:a proposal and analysis," *Optics Letters*, Vol. 24, pp. 711-713 (1999).
11. J. B. Khurgin, "Expanding the bandwidth of slow-light photonic devices based on coupled resonators," *Optics Letters*, Vol. 30, pp. 513-515 (2005).
12. D. N. Christodoulides and N. K. Efremidis, "Discrete temporal solitons along a chain of nonlinear coupled microcavities embedded in photonic crystals," *Optics Letters*, Vol. 27, pp. 568-570 (2002)

Dynamic Context Aware Service Provision in Beyond 3G Mobile Networks

Spyros Panagiotakis*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
spanag@di.uoa.gr

Abstract. The evolution of mobile communication systems to 3G and beyond introduces requirements for flexible, customized, and ubiquitous multimedia service provision to mobile users. One must be able to know at any given time the network status, the user location, the profiles of the various entities (users, terminals, network equipment, services) involved and the policies that are employed within the system. Namely, the system must be able to cope with a large amount of context information. Present paper focuses on location and context awareness in mobile service provisioning and proposes a flexible and innovative model for user profiling. The innovation is based on the enrichment of common user profiling architectures to include location and other contextual attributes, so that enhanced adaptability and personalization can be achieved. For each location and context instance an associated User Profile instance is created and hence, service provisioning is adapted to the User Profile instance that better apply to the current context. The generic model, the structure and the content of this location- and context-sensitive User Profile, along with some related implementation issues, are discussed.

1 Introduction

The challenge with mobile, distributed computing is that exploits the user's dynamic environment with a new category of applications that are aware of the context in which they run. As context we define the combination of information relevant to the nearest environment of a user, such as the user location, the serving network, his terminal device, etc. Context-aware applications present information and services to a user, as well as automatically execute services and commands, sensoring context and its changes. Changes of the contextual environment, are modelled as events, and are communicated to the application for real-time service re-adaptation.

A context-aware service takes into account the current context of the user and based on this information it adapts its behaviour to the respective user's needs including personal preferences and environment's capabilities [1][2]. The contextual informa-

* Dissertation Advisor: Lazaros Merakos, Professor

tion can be encoded in various related profiles such as, the user preferences profile and the terminal, ambient, network, and service profiles. The combination of all these profiles constitutes the User Profile [3][4].

The issue of adapting service provision and providing personalised services based on user preferences is what 3GPP introduced in Virtual Home Environment (VHE) [5]. VHE is a concept for Personal Service Environment (PSE) portability across network boundaries and between terminals. Primary aim of VHE is to consistently present with the same personalised features, User Interface customisation and services in whatever network and whatever terminal (within the capabilities of the terminal and the network), wherever the user may be located. VHE is enabled by user profiles since they encode parameters that are essential to the user, such as the users' preferences for communication and service presentation on the terminal.

In particular, the User Preferences profile encodes desirable service provision features that are particular to an individual user. User preferences can be categorised into service-independent, which apply to all services that are accessed by the user and service-specific, which pertain to a particular application.

Since profiling information is exchanged among different administrative boundaries, to assure interoperability the XML [6] and the other XML-like languages (e.g., WSDL [7], SMIL [8], OWL [9]) are the best candidates for describing profiles. Moreover, profiles should be laconic so that they are transmitted efficiently. In situated–aware architectures user profiles must be dynamically composed since their constituting segments may be distributed. Contextual profiles influence greatly a service's deployment and execution, since context-aware services should adapt to context and related updates.

To enable third parties to develop context-aware services above mobile communication systems, various efforts have been undertaken by standardization working groups and fora towards the introduction of open, network-independent interfaces enabling context retrieval [10]-[13]. These interfaces provide applications with transparent access to network functionality (e.g., call control, location information, messaging, profiles retrieval), thus offering third party application developers the opportunity to create advanced, network-independent and context-full services with standard software engineering tools and general-purpose programming languages.

Let us now illustrate how all these technologies above should ideally work assuming a hypothetical, but realistic, service provision scenario. We assume that a mobile user is receiving real-time video news from an application provider using a large-screen WLAN-enabled laptop while at home. As the mobile user is boarding a vehicle to his office and changes from the large-screen terminal to a portable 3G phone, a context information server in the service support centre notifies the application of the location and the terminal and radio access update. Application successfully recognizes, by the user's profile or after interacting with the user, the user wish to change content (media) from video to text and a media converter executes content conversion.

The above is only one example of requirements for the flexible, customized, context-aware and ubiquitous multimedia service provision to mobile users that the evolution of mobile communication systems to 3G and beyond introduces. Definitely, in a system that aims to provide flexible and context aware service provision and adaptation, the knowledge of the system status as well as the various entities' states and events are significant factors. One must be able to know at any given time the network status, the user location, the profiles of the various entities (users, terminals, network equipment, services) involved and the policies that are employed within the system. Namely, the system must be able to cope with a large amount of context information.

However, in order for this scenario to be optimally and operationally deployed in real life, several technological requirements should be fulfilled by involved infrastructures such as:

- A sensoring infrastructure effectively monitors the context of the user (location, terminal, Radio Access Technology (RAT)) and notifies the application whenever the applicable context changes.
- The preferences of the user with respect to the application provision should be, at least, location-dependant, so the application retrieves the applicable application customisation set whenever the context changes.
- This implies that the favourite geographical areas of the user should have, at first, been defined and then, that for each of these areas the corresponding sets with the user preferences for application customisation have been defined and corresponded to them.
- The application scenario is defined in a standardised document, which describes all the alternative provisioning and media adaptation aspects it can support.

All these requirements further burden service provisioning. However, taking into account that development and deployment for end-user applications should be kept as simple as it can, as well as that the additional intelligence required by these applications for the step towards context awareness should be minimal, it is induced that the certain additional intelligence required should be transferred from applications to networks or middleware applications [2]. Furthermore, most of the requirements above concern some provisioning attributes and components common to most context aware applications. Hence, it would be beneficial, in terms of minimising complexity and potential useless interactions, if some common services and components were offered to open access, so these can be reused and shared among interested parties or applications. For example, an infrastructure provided for context monitoring and events notification can be easily shared by several applications. The same states also for some user profiling issues, since the information related to the favourite geographical zones of a user or its generic preferences for services consumption from within them could be shared among authorised applications of the specific user. The latter issues enhance our position for transferring the intelligence required for context awareness from applications to a middleware layer. This aspect is illustrated in Fig. 1. Adopting a middleware layer with open access to authorised applications and assigning to it the required intelligence for context awareness, simplifies and facilitates

adaptation to context as well as development, composition and deployment of context aware applications, while minimises redundant interactions between applications and underlying network infrastructure.

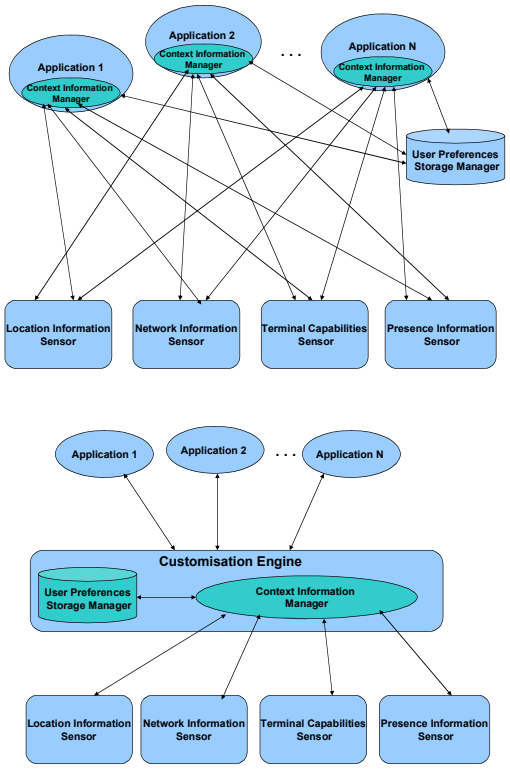


Fig. 1. Context awareness with and without mediation

Present paper focuses on location and context awareness in service provisioning and proposes an open and shared middleware framework that enables efficient management of location and context information. Furthermore, a more flexible and innovative model for user profiling is introduced that facilitates adaptation to context. The generic model, the structure and the content of a context-sensitive User Profile, along with some related implementation issues, are discussed.

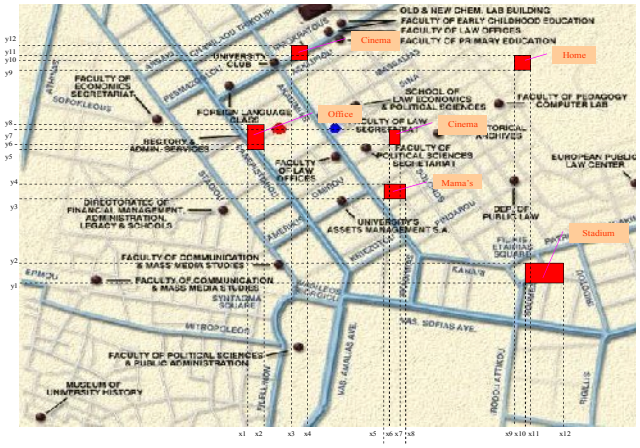


Fig. 2. User-defined Home Zones of a subscriber

2 Location- and context- sensitive user profiling

The User Profile is a key means to provide subscribers with truly location aware and customised services. Additionally to the service deployment and execution mechanisms, which take the user location into consideration, the management of user profiles can be also location sensitive. Location awareness in User Profiling is based on the concept of user-defined Home Zones. Each Home Zone comprises a geographical area into which a user wishes to experience personalised and customised service provisioning (e.g., the Home, the office, the car) [14]. Ideally a Home Zone should be as wide as the user wishes, so that truly customisation can be achieved. For example, a user may wish to experience differentiate service provision in each room of his house or office. In such a case each room should be considered as distinct Home Zone for that user. However, limitations in the location measurement accuracy induced currently by various positioning technologies, forbid Location Based Services (LBS) to distinguish among very narrow Home Zones. Due to this fact, certain distance among the defined Home Zones of a user, depending on the accuracy of the location measurements, should exist, so the position detection system can follow the moves of subscribers from one Home Zone to another. In the near future, when the location estimation technologies mature further, location based services shall provide users with more accurate positioning and Home Zone definition. Fig. 2 Illustrates the example user-defined Home Zones of a subscriber, along with the associated coordinates bounding each one.

In the proposed framework the location-sensitive user profiles are managed by the User Profile Manager (UPM) assisted by the available location and context sensors of the underlying infrastructure. Location sensor will be identified hereafter as the Location Manager. In this context, personalisation and customisation during service provision is achieved by discriminating the user preferences according to the location (e.g., Home, work) of the user and by maintaining different User Preferences Profiles (or Service Customisation Sets) for each instance (e.g., Home-, work-dependent profiles). Other contextual parameters that can be also taken into account are the serving RAT, the type of the Mobile Equipment (ME), the time of the day, or the presence status [15] of the user (e.g., lazy- or work- or mood- specific). These attributes constitute the user context. For each context an associated User Preferences Profile is created and hence, service provision is adapted to the User Preferences Profile instance that better apply to the current user context.

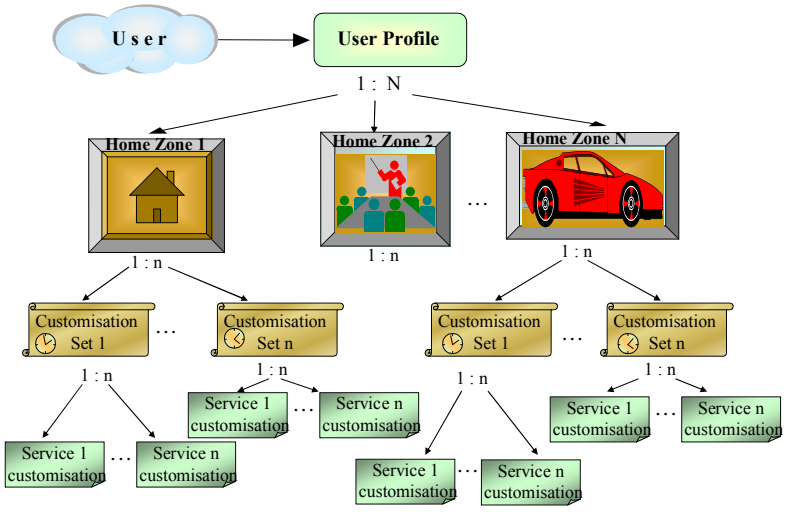


Fig. 3. Home Zone-sensitive user profiling

Fig. 3 depicts the Home Zone-sensitive user profiling approach. Each user is associated with a single User Profile that contains general information for the subscriber such as his general user preferences, the available terminal capabilities, the subscribed network capabilities, etc [3]-[4]. That User Profile consists of several, Home Zone-dependent, Profile instances that inherit from the User Profile and contain various, Home Zone-dependent, personal attributes such as the QoS values related to home use

or travelling. A user may have one or more instances (or Service Customisation Sets) of a specific Home Zone profile; each individually customised by the user and associated, for example, with a specific time or period of the day or other context attributes. Each Service Customisation Set includes the applicable user preferences, with respect to the User Interface preferences, the Browser appearance, the preferred memory usage etc., and the service subscription Profiles, with the preferred settings for the subscribed services (e.g., for pricing) and associated privacy policies. The user experiences service offering according to the current active Service Customisation Set.

From that point of view each Home Zone-sensitive User Profile can be considered as a tree with the subscriber's identity (e.g., his IMSI, e-mail or SIP identity) at the root, the Home Zone and context attributes as nodes and the service customisation sets as leaves. Storing locally such a tree-like User Profile for each subscriber, finding his current Home Zone and context and crossing that tree from top to down, the UPM can retrieve the most up-to-date user profiling data each time it is required. Although in Fig. 3 only the Home Zone and time attributes are taken into account in the user preferences selection and service differentiation, further contextual parameters could be used, such as the type of ME and the serving RAT of the subscriber, for better elaboration and specification.

In that context, each subscriber upon registration with the proposed platform specifies his applicable Home Zones describing each one of them with addresses or street names. It is then up to the Location Manager of the platform to map the Home Zones specified by the user to the appropriate geographical coordinates or network areas (e.g. cell Ids, Location or Routing Areas), making use of a data base with the appropriate spatial information. Thus, whenever a user enters the platform and accesses an application, the UPM retrieves the current location and the required context attributes of the user in order to identify the Home Zone he is currently situated in and to retrieve the user preferences that apply to this specific Home Zone and context. This has as result only the profile that better applies to the current Zone and status of the user to be taken into account, customising the whole service provisioning accordingly. It is implicit that the user is always prompted to confirm the User Profile instance selection or alternatively to choose the one he desires.

```
<Profile112>
        <ProfType> General Preferences </ProfType>
        <ProfileID> GP112x </ProfileID>
        <ProfileURI> http://erotocretos.di.uoa.gr/profiles/profileGP112x </ProfileURI>
        <RefContextZone>
                <HomeZone> HZ1</HomeZone>
                <MTType> MT1</MTType>
                <RATType>RAT2</RATType>
        </RefContextZone>
</Profile112>
```

Fig. 4. Specification of a new user preferences set

Then, each time the subscriber specifies a new user preferences set, it is associated with the current Home Zone and context of the subscriber or the Home Zone and context indicated by the subscriber. For each new Home Zone or context value (e.g., new ME, or new RAT) occurs, a new node in the tree is inserted. Equally, for each new customisation profile a new leave is added to the tree. Definitely, a user preferences customisation set can be associated with more than one Home Zones and contexts, if the user wishes. For each active subscriber the UPM stores locally a data structure that represents the tree of his location sensitive User Profile. The data itself is not included. Instead a reference to the repository that stores each profile is kept, along with a unique data reference generated upon data creation and storage. The Home Zone and the context attributes are used by UPM to identify the path to the active profile data. To this end, a pointer that crosses the User Profile tree and points to the active customisation profiles is created for each user. The pointer is updated each time a change on active Home Zone and context occurs. Fig. 4 illustrates the specification of a new user preferences set for a user and its correlation with specific Home Zone and context attributes (defining the path to the associated profiling data in the User Profile tree). This data structure is inserted in the User Profile tree of the specific subscriber constituting a new leave in the User Profile structure. After insertion, this data structure will, also, be the result of search of UPM in this User Profile tree for the "General Preferences" profile of the subscriber that corresponds to the designated Home Zone and context. Retrieving this data, the UPM can, then, access the corresponding profile repository (specified by the referenced profile URI) to retrieve profile GP112x.

Hence, whenever a user accesses an application, the application requests by the UPM the applicable preferences set of the user. The UPM, then, retrieves the user location in order to identify the current Home Zone of the user, along with the context attributes required for profile identification (e.g., the ME type and the type of the serving RAT) to identify the current context and the applicable profile instance of the user and, finally, to retrieve by the corresponding repository the user preferences set that applies to the specific context. It is implicit that the user is always prompted to confirm the User Profile instance selection or alternatively to choose the one he desires. The UPM stores locally the current Home Zone and context of each user for later use and faster searching in the User Profile tree.

With location and context aware user profiling only the profile that best applies to the current Home Zone and status of the user is taken into account, customising the service offering accordingly. During service access, the underlying Location Manager (and the other context sensors) tracks the location (and the other context attributes) of the user and upon entrance in or exit from a Home Zone an appropriate notification is sent to the UPM notifying it for the new Home Zone the user entered. The UPM, then, retrieves the applicable user preferences in the new Home Zone of the user in order to announce them to the executing applications by the user and enable them to tailor their service provision to the user accordingly. Fig. 5 illustrates a User Profile instance transition induced by a Home Zone change. As the user moves from Home Zone 1 to Home Zone 3, it leads to a change of his current context from Context Zone 1 (Home Zone 1, Mobile Equipment 1, Radio Access 2) to Context Zone 2 (Home Zone 3,

Mobile Equipment 1, Radio Access 2), which in turn leads to a different user preferences set (from User Preferences Set N to User Preferences Set 2). Such a change of the active User Profile instance triggers the UPM to generate and propagate an associated alert event, so that the applications and modules registered for receiving such alerts are properly informed.

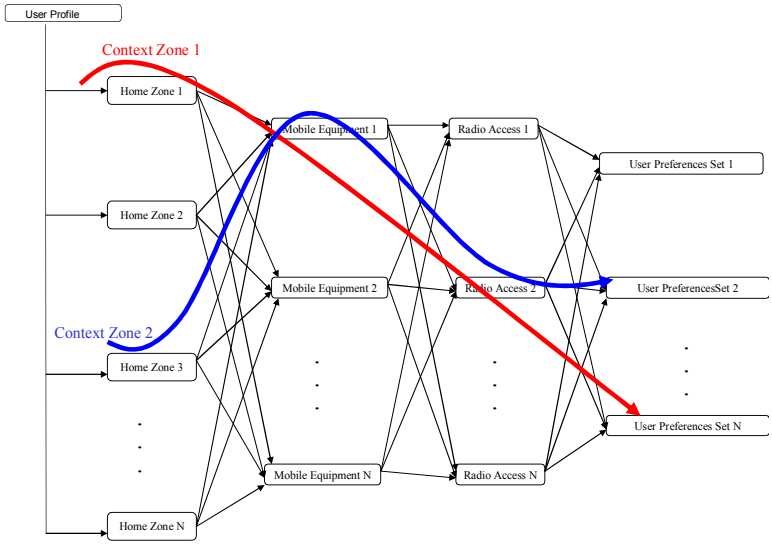


Fig. 5. User Profile instance transition induced by a Home Zone change

In Fig. 5, apart from the Home Zones, the ME and the serving RAT type context attributes are also used for further profile elaboration. Taking into account that within a single Home Zone a subscriber can switch from one type of mobile equipment to another (e.g., from a UMTS mobile handset to a PDA or laptop) and access different radio environments (e.g., from GPRS or UMTS networks to WLAN/WiFi or Bluetooth), further classification of user profiles within a Home Zone can be achieved. This is why we consider that the current Home Zone, current ME and current serving RAT attributes of a subscriber are the three key context attributes that uniquely identify the current context of a user. Hence, we define each triplet of type {current Home Zone, current ME, current RAT} as a user specific *Context Zone* that can be used for identifying the user status and customising the service provisioning accordingly.

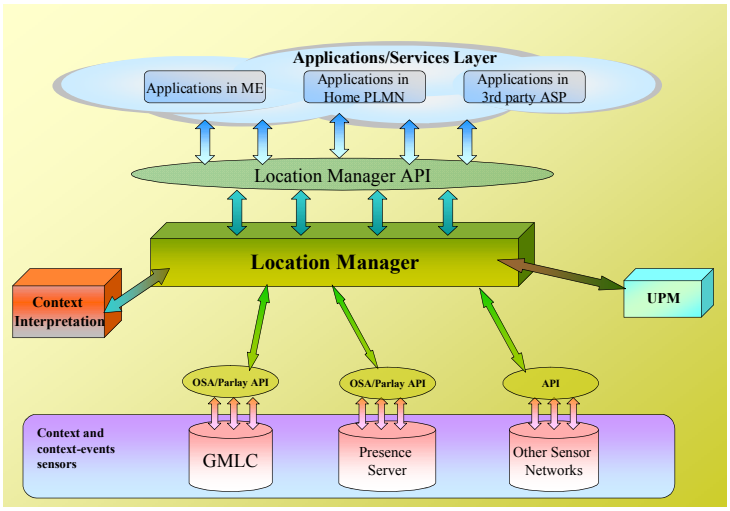


Fig. 6. Environment of the Location Manager

3 Location Management

The Location Manager constitutes an independent module responsible for retrieving, managing and exploiting the information related to the location and mobility of the subscribers. It interacts with the location and presence information's sources of the underlying network infrastructure (e.g. the 3GPP GMLC [16],[11],[12] or the Presence Server [15]) to track the location, presence and mobility of the subscribers [17]-[20]. To translate the retrieved raw location information into a recognizable and usable format instead of the geographical coordinates or network areas that the underlying Location Sensors or Server (e.g., the GMLC) provides, the Location Manager interacts with the appropriate Interpretation Component of the platform. Then, location, presence and mobility data and events along with the preferences of the corresponding subscriber, taken from the user profile, are processed to provision the user in the new location he entered. Fig. 6 illustrates the environment of the Location Manager. The users' location can be used, for example, to determine, based on the user's preferences, the reconfiguration policies that are propagated from the Location Manager to the underlying network infrastructure [17]-[20]. Combining the location information with the user preferences, the Location Manager is able to provide end users and any authorised entity internal (e.g., the Application/Service logic component) and external (e.g., the third-party Application/Service Providers (ASP)) to the platform with new advanced location aware services. Furthermore, it enriches the service provi-

sioning approach of our platform with location information features, enabling better customisation and personalisation of the whole service offering.

Primary goal of the Location Manager is to enable easier development of LBS, hence it does not focus on the development of a specific application or service (location-based or not). Instead, our primary goal is to build a generic framework for location-sensitive service development and deployment that can accommodate any service or application, gathering the necessary informational resources and building blocks for facilitating development and deployment. To this end, we provide applications with an execution environment and some reusable and customisable location sensitive building blocks for structuring their functionality, while we hide from application developers the complexity and the physical interactions required for retrieving and adapting the location information. For transparency and independence the functionality provided by the Location Manager is accessed by internal modules of the framework as well as authorised ASPs through an open API provided to authorised entities [19]. We have designed the methods and built the services that we expect a location-based application/service to need frequently. In particular, this open interface includes methods that enable:

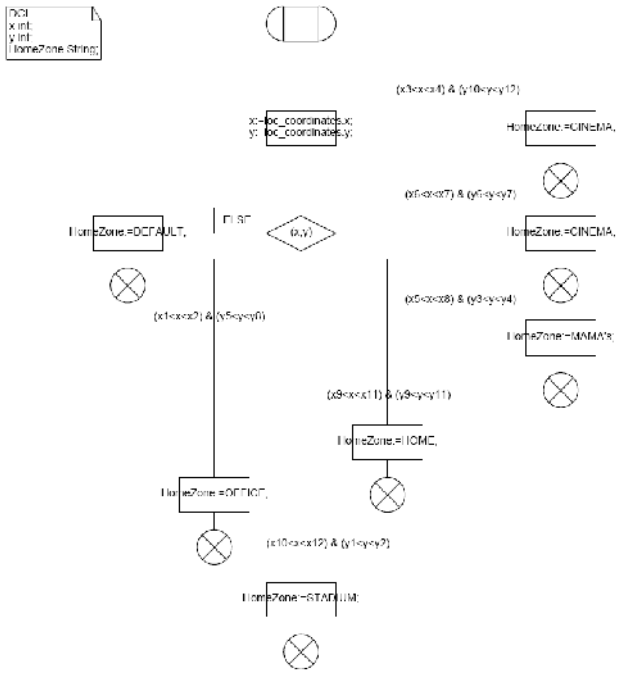


Fig. 7. Translation of geographical coordinates to applicable Home Zone

- Retrieval of the location of the specified user. Location retrieval can be immediate (in case that the current location of the user is requested) or deferred (in case that the location of the user when a specific event takes place is requested) [16], [17]-[20]. Making use of appropriate spatial data bases or GIS systems, the Location Manager is able to map the current raw user location taken by the underlying Location Sensors (expressed in geographical coordinates or network areas) to the requested higher-level format (e.g., street address or predefined geographical Zones (Home Zones)). Hence, the retrieved location information can be in a recognizable and usable format. Fig. 7 illustrates a simple algorithm that can be used for the translation of the geographical coordinates (x, y) retrieved from the underlying location sensors to the corresponding Home Zone of the user having defined the Home Zones of Fig. 2. It is based on simple comparison of retrieved coordinates against the coordinates bounding each established Home Zone for the user.
- Creation/Modification/Deletion of Location-sensitive Policy Classes. Such a policy class can be, for example, the registration of the Home Zones of a user with the Location Manager, so the Location Manager monitors the user mobility within them.
- Activation/De-activation of the Location-sensitive policies.
- Creation/Modification/Deletion of Policy Events. Such a policy event can be, for example, the crossing (entrance or exit) of registered Home Zones by their owner.
- Registration/Deregistration for receiving Location-sensitive event notifications. With this method the registered applications can receive notifications from the Location Manager whenever a registered event occurs (e.g., whenever a specific user crosses his registered Home Zones).
- Handling of event notification arising from the underlying network.
- Submission of notifications to the registered end-users and applications for the available policies, restrictions, updates, tariffs, reconfigurations and other events that are associated with the current location of the user or induced due to the location updates that occur.

4 User Profile Management

The architecture we propose for User Profile management enhances the 3GPP GUP architecture [3]-[4] with the concepts of Home Zones-based Profiles. Hence, it adopts the distribution and information model of the GUP incorporating, additionally, in its logic provision for enhanced context sensitivity. The enhanced GUP server proposed here is called User Profile Manager (UPM). The UPM interacts with the Location and Presence Manager, as well as the context sensors of the proposed framework to retrieve the location, presence and other contextual information needed to compute Home Zones and provide location sensitive User Profiling.

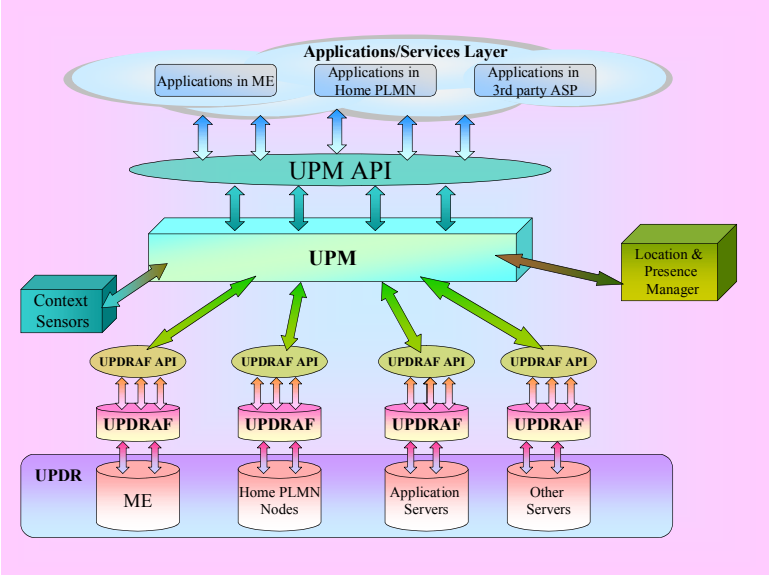


Fig. 8. Environment of the UPM

The UPM is responsible for managing the User Profiling information distributed in several data repositories across the network and disseminating the user-specific information to the requesting applications and services. It mediates between application/services and User Profile Data Repositories (UPDR), hiding from applications the underlying infrastructure and facilitating the interaction with the profiling sources. The applications that may request access to the User Profile data can vary from applications in the Mobile Equipment (ME) to applications in the Home PLMN or third party Application/Service Providers (ASPs). The UPM enables authorised applications to insert/delete or modify user profiling data in UPDRs, retrieve user profiling data upon request and receive profiling dependent event notifications, each time a registered events occur. Fig. 8 illustrates the environment of the UPM.

To hide the implementation details of the profiling architecture from applications and services and assure service transparency, the interaction between internal modules or services of the profiling architecture, as well as authorised third party applications, and the UPM is accomplished through an open Application Programming Interface (API) provided by UPM to authorized entities [19].

Profiling data stored in the various User Profiling components are identified by the identity of the associated subscriber/user, the corresponding Home Zone and context attributes of the User and the profile type. The specified methods provide:

- Creation/deletion/update of user profile data. These procedures are always related to a single subscriber, Home Zone and context, which are identified in the request. If Home Zone or context attributes are missing the current Home Zone and context of the identified subscriber is assumed.

- Retrieval of the whole user profile data or some specific components. The queried data are identified by subscriber identity, Home Zone and context attributes and the data reference. In case that Home Zone and context are not passed in the request the UPM retrieves the current Home Zone and context of the targeting user and returns to the requesting application the requested profile data that corresponds to the current Home Zone and context of the user.
- Listing of the existing profile items in the various User Profiling Data Repositories that are associated with the specified Home Zone and context of the targeting user.
- Creation/Modification/Deletion of profile-dependent policy events.
- Registration/Deregistration for receiving profile-dependent event notifications.
- Submission of event-driven notifications to the registered applications whenever some of the registered events occur. The synchronisation of the profile data kept by an application can be performed by last three methods (creation of a policy event (e.g. monitoring of the active preferences profile of a specific user for the application XYZ), registration for receiving notifications related to this event and receipt of associated notifications whenever an update of the active preferences profile is required (e.g., due to the user transition from Home Zone 1 to Home Zone 2)).
- Notifications to the users for the profile-dependent policies, restrictions, updates associated with the current Home Zone and context of the user.

The user profiling information is distributed and stored in various UPDRs. Each UPDR stores the primary master copy of one or several profile components. Possible candidates for UPDR are the ME, the HSS/HLR, and various application and management servers in Home PLMN or 3rd party Application/Service Providers. Synchronization between profiling data in UPDRs and UPM is required.

Access to UPDRs is accomplished through the associated User Profile Data Repositories Access Functions (UPDRAF). Each UPDRAF can be viewed as the front end to the underlying repository that realizes the harmonized access interface. It hides the implementation details of the UPDR from the rest UPM infrastructure. The UPDRAF performs protocol and data transformation where needed. The protocol between the UPDRAF and the UPDR is implementation dependent and not standardised. The UPDRAF can take also part in the authorization of access to UPDR. Through UPDRAFs the UPM can insert/delete/modify the underlying profiling data, read them, and receive synchronisation notifications whenever some change on profiling data occurs.

Example Interactions of UPM

A. Retrieving the user preferences

Fig. 9 illustrates the required interactions between the components of the UPM environment whenever an application accesses the UPM to retrieve the current user preferences of a subscriber. Interactions include:

1. An authorised application requests to retrieve some User Profile components of a specific subscriber. The application does not include the Context Zone parameter in the request.
2. The UPM authenticates the application and checks its authorisation to receive the requested data.
3. Since Context Zone is not provided, the UPM presumes that the profile data requested are those associated with the current Context Zone of the targeting user. Hence, the UPM contacts the Location Manager of the architecture to retrieve the current Home Zone of the user.
4. The UPM retrieves the type of the Mobile Equipment from the appropriate context sensor.

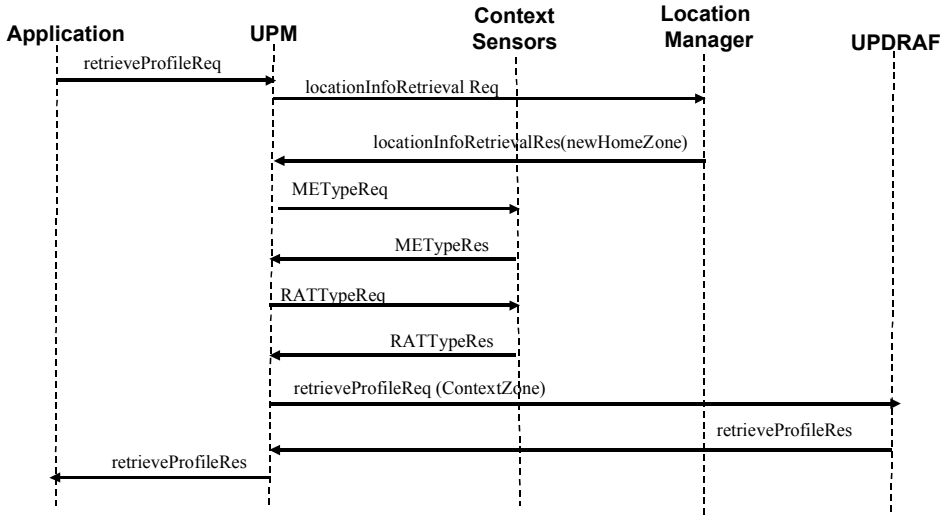


Fig. 9. Required interactions for retrieving the user preferences

5. The UPM retrieves the type of the Radio Access Technology serving the user from the appropriate context sensor.
6. The UPM identifies the current Context Zone of the user and based on that it crosses the Context Zone-dependent structure of the User Profile to identify the storage location of the requested components.
7. The UPM accesses the API of the identified UPDR to request the specified data. The UPM includes Context Zone and data reference in the request.
8. The UPDRAF searches the stored profiling information to retrieve the requested data and responds to the UPM.

9. The UPM responds to the application with the requested profiling data. Since the data requested by the application might are stored in several UPDRs, it is likely the UPM to have to interact with all involved repositories to request the data. In that case, the UPM should properly combine the returned data before responding to the application.

B. Profiling dependent event notifications

The example below presents how an authorised application remains informed and updated about changes on the active user preferences of the targeting subscriber, induced by changes on the current Context Zone of the user.

Fig. **10** illustrates the required interactions between the various components of the profiling architecture:

1. The application accesses the API provided by UPM to create a profiling dependent policy event related to a specific subscriber. The specific event intends to constitute the application aware of the changes on the active user preferences profile of the targeting subscriber.
2. The UPM authenticates the application and checks its authorisation for the requested operation.
3. The application registers itself to receiving notifications related to the aforementioned event.
4. Sometime, the UPM receives an event from the Location Manager of the architecture indicating that the specified subscriber has entered a new Home Zone. It is assumed, here, that registration of the UPM with the API of the Location Manager, according to the procedures described in section 0 for receiving such notifications from the Location Manager, has preceded.
5. The UPM contacts the Location Manager of the architecture to retrieve the current Home Zone of the user. This step can be skipped if the location notification received includes the new Home Zone of the subscriber.
6. Since the UPM has not received notifications from the associated context sensors for changes on the Mobile Equipment or the serving RAT of the subscriber, it assumes that only the geographical location of the user has changed. The UPM identifies the new Context Zone of the user and based on that crosses the Context Zone-dependent structure of his User Profile to identify the available profiles of the user in the new Context Zone.

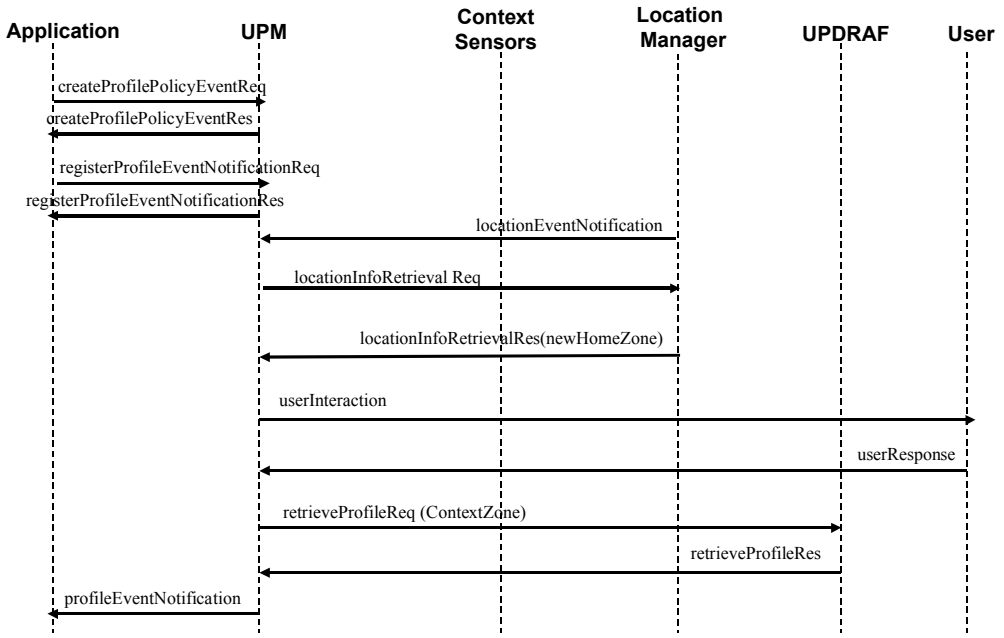


Fig. 10. Required Interactions for enabling profiling dependent event notifications

7. The UPM, optionally, interacts with the user to inform him about his available service customisation profiles in the new Context Zone.
8. The user selects and activates the desired profile.
9. The UPM updates the User Profiling pointer to point to the newly activated profile of the user, and locates the UPDR that stores the new profile components. Then, the UPM accesses the API of the identified UPDR to request the specified data. The UPM includes Context Zone and data reference in the request.
10. The UPDRAF searches the stored profiling information to retrieve the requested data and responds to the UPM.

The UPM notifies the requested application for the change occurred. The notification to the application can be a simple announcement of the change occurred on the user preferences, unless the application has requested to also receiving the new profiling data along with the notification (illustrated in

Fig. **10**). In the former, the application should contact again the UPM to receive the new data

5 Conclusions

Present paper focused on location and context awareness in mobile service provisioning and proposed a framework that enables efficient management of contextual profiles. Furthermore, a flexible and innovative model for user profiling was introduced. Innovation is based on the enrichment of common user profiling architectures to include location and other contextual attributes, so that enhanced adaptability and personalization can be achieved. For each location and context instance an associated User Preferences instance is created and hence, service provisioning is adapted to the User Profile instance that better apply to the current location and context.

Comparing the framework of location and context sensitive user profiles with the traditional, non-context sensitive, profiling architectures in mobile communications (e.g. the 3GPP GUP), someone can easily note that the former offers really challenging advantages to applications enabling their efficient adaptation to location and context, minimising, in parallel, the additional intelligence required to this end. The price to be paid for these advantages is mainly in terms of additional delay, which is due to the interactions required between UPM and location and context sensors. First results from evaluation of our framework show that in the worst scenario of its functionality, the UPM requires about 35 % more time than GUP to retrieve a requested profile and respond to an application (scenario of Fig. 9). However, taking into account that profile retrieval in the UPM approach takes place in two steps; the heavy one of Fig. 9, which is performed only once during an application execution scenario, and the lighter one of

Fig. **10**, which is repeatedly performed after the first profile retrieval (which requires almost the same time as a GUP-like profile retrieval), we conclude that the additional time expense that incurs the adoption of UPM greatly varies depending on the number of times that the scenario of

Fig. **10** is executed. Thus, in the execution scenario of a mobile application during the provision of which to a mobile subscriber the UPM is required to propagate ten times to the application the applicable user preferences, the additional time expense falls to 15 % of the respective GUP time. The latter result shows that for applications with high mobile characteristics, where the location and the context of the subscribers change often, the UPM is the ideal solution for enabling adaptation to context. Similar, for applications targeting to static end-users, the UPM solution has little to offer.

References

[1] Spyros Panagiotakis, Athanassia Alonistioti, "Context-Aware Composition of Mobile Services", IEEE IT Professional, July 2006, Volume 8, Number 4, pp. 38-43

[2] Spyridon Panagiotakis, Maria Koutsopoulou, Athanassia Alonistioti, Nikos Houssos, Vangelis Gazis, "A Middleware Framework for Reconfigurable Mobile Networks", International Journal of Mobile Communications (IJMC), March 2004.

[3] 3GPP TS 22.240: “Service requirement for the 3GPP Generic User Profile (GUP); Stage 1”.
[4] 3GPP TS 23.240: “3GPP Generic User Profile-architecture (GUP); Stage 2”.
[5] 3GPP TS 23.127: “Service aspects; The Virtual Home Environment”.
[6] Extensible Markup Language (XML) home page, http://www.w3c.org/XML.
[7] Web Services Description Language (WSDL) 1.1, http://www.w3.org/TR/wsdl
[8] Synchronized Multimedia Integration Language (SMIL) 2.0 Specification, http://www.w3.org/TR/smil20/
[9] Web Ontology Language (OWL), http://www.w3.org/2004/OWL/
[10] User Agent Profile (UAProf) Specification, http://www.wapforum.org/what/technical.html
[11] Parlay Group, “Parlay API Spec”, available from URL http://www.parlay.org/specs/index.asp
[12] 3GPP TS 29.198: “Open Service Access (OSA); Application Programming Interface (API); Part 1-12”.
[13] J. Keijzer, D. Tait, R.Goedman, “JAIN: A new approach to services in communication networks”, IEEE Communications Magazine, January 2000.
[14] 3GPP TS 22.071: “Location Services (LCS); Service description, Stage 1”.
[15] 3GPP TS 23.141: “Presence Service, Architecture and functional description”.
[16] 3GPP TS 23.271: “Functional stage 2 description of LCS”
[17] Spyridon Panagiotakis, Athanassia Alonistioti, "Intelligent service mediation for supporting advanced location and mobility aware service provisioning in reconfigurable mobile networks", IEEE Wireless Communications Magazine, October 2002.
[18] Spyridon Panagiotakis, Athanassia Alonistioti, Lazaros Merakos, "An advanced location information management scheme for supporting flexible service provisioning in reconfigurable mobile networks", IEEE Communications Magazine, February 2003.
[19] Athanassia Alonistioti, Spyridon Panagiotakis, Maria Koutsopoulou, Vangelis Gazis, Nikos Houssos, “Open APIs for Flexible Service Provision and Reconfiguration Management”, chapter contribution to the book entitled: “Software Defined Radio (4th volume): Architectures, Systems and Functions”, published by John Wiley & Sons Ltd. in May 2003.
[20] Spyros Panagiotakis, Nancy Alonistioti, “Location-based Service Differentiation”, contribution to the book entitled: “The Handbook of Mobile Middleware”, edited by Paolo Bellavista, Antonio Corradi, published by Auerbach Publications (Taylor & Francis Group) in September 2006, chapter 30, pp. 787-818.

Contribution in the Analysis and Coding of Three-Dimensional Image Sets

Nicholas P. Sgouros

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Panepistimiopolis, Ilissia
Athens 15784, Greece

nsg@di.uoa.gr

Abstract. Three-Dimensional (3D) visualization systems are used in many specialized applications where 3D observation is required. Technological achievements in the areas of integrated optics, sensors and network infrastructures guarantee that these systems will be used soon in a large number of different applications. This dissertation summarizes the author's research on the major aspects of a specific 3D visualization method, called Integral Photography (IP). This work describes the implementation of a prototype device capable of producing high resolution, near field, IP images of real 3D objects based on a flatbed scanner. In addition, this work also presents a novel automated method for calibration of the sensors and the optics used in an IP acquisition device without prior knowledge of the system's characteristics. This is accomplished by post-processing of the acquired IP images using image analysis and pattern recognition methods. Finally two encoding schemes are developed in order to properly exploit the inherent redundancy of an IP image and achieve high performance in a rate distortion sense.

Keywords: Three-dimensional Imaging, Integral Photography, Three-dimensional Image Analysis, Three-dimensional Image coding, Autostereoscopic Applications

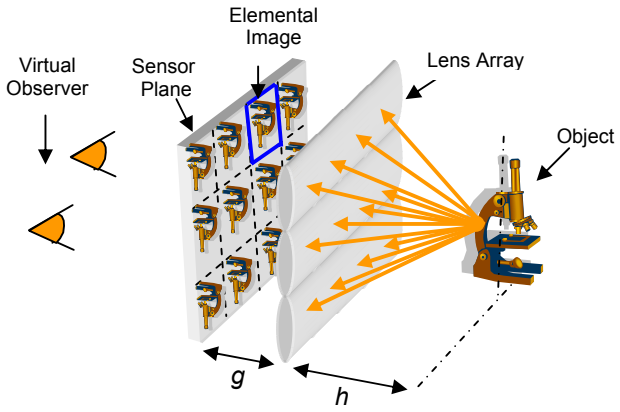
1 Introduction

In general, three-dimensional (3D) viewing methods can be divided in stereoscopic and autostereoscopic. Stereoscopic methods require additional viewing aids to provide 3D viewing, like glasses or helmets, while autostereoscopic methods do not require additional viewing aids as all optical components needed in order to reproduce correct 3D objects and scenes are integrated in the display device [1-2].

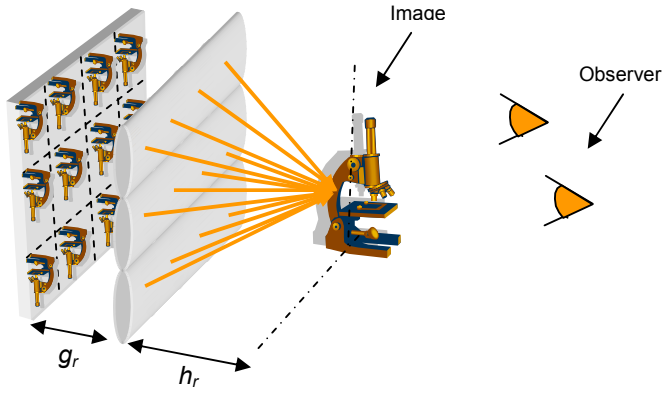
Integral Photography (IP) is an autostereoscopic technique for producing high realistic 3D images, initially devised by G.Lippman [3] back in 1908. Specifically, this technique provides natural viewing with full colour support, enhanced detail and adequate depth level. Moreover, multiple simultaneous viewers are supported, while most of the currently existent IP setups for capturing and display provide full parallax

Dissertation Advisor: Manolis Sangriotis, Assoc Professor

on both horizontal and vertical directions [4,5]. The basic principle for the capturing and reproduction methodology is depicted in Fig.1, (a) and (b) respectively.



(a)



(b)

Fig. 1. IP (a) Acquisition and (b) Display Device functionality principles.

An IP image is a two-dimensional (2D) lattice of small images, usually called elemental images, which are formed behind each lens of the lens array of Fig. 1a. The IP image is captured using a Charge Coupled Detector (CCD) placed directly behind the lens array located at the sensor plane in the same figure. When the resultant IP image is displayed using a Liquid Crystal Display (LCD) through a second lens array, as shown in Fig.1b, the initial 3D scene is reproduced.

Some of the key problems of an IP imaging system involve calibration, image restoration, compression, depth extraction, object recognition and others. Most of the standard image processing techniques can be applied in resolving these issues. However, certain adaptations should be made in order to take advantage of the inherent characteristics of an IP image structure and produce optimal results. Under this framework there are a number of attempts made, trying to tackle problems such as IP image compression [6,7], but there are still major processing issues, like the ones mentioned, that remain an open field for research and development.

This paper is organized in four sections. In section one, a novel architecture for high quality IP image acquisition is proposed which uses a Lens Array and a commercial flatbed scanner. General calibration of the acquisition device along with filtering for noise removal and IP image analysis issues are resolved in section two. Finally in section three of this work two novel coding schemes are presented for IP and generalized multiview architectures. Section four, concludes this paper and presents future work issues and enhancements.

2 IP Still image Acquisition Device

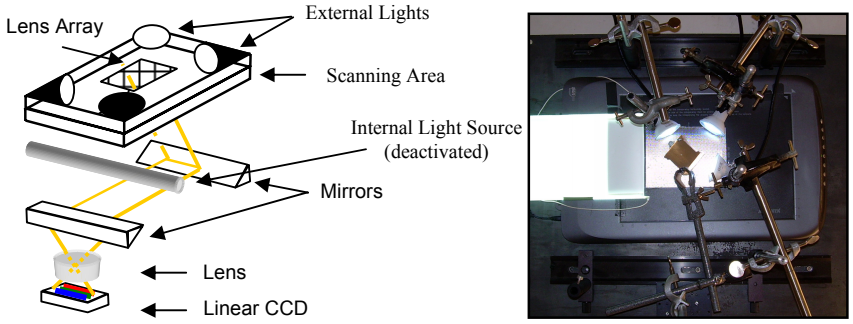
There are currently numerous IP acquisition setups based on digital sensors like area CCDs' [8-10]. The small size of the CCD sensor used in standard digital cameras imposes a number of limitations to IP acquisition setups. As a result, the IP images produced have a small number of lenses in an effort to increase lateral image resolution. On the contrary, an increase to the number of lenses in order to enhance depth resolution reduces lateral resolution. Taking into account both previous limitations a compromise is usually made based on the camera characteristics and available lens arrays. In this paper an IP acquisition device is described which is based on the functionality of a flatbed scanner and a large area lens Array. This device delivers high-resolution IP images with only a small fraction of the cost of solutions that are based on area CCD sensors.

2.1 Acquisition device

The developed prototype consists of a flatbed scanner with an effective scanning area of 21x30 cm and 0.1cm pitch for the rectangular lenses that assemble the lens array. The schematic of the device along with the functional prototype developed in this work are depicted in Fig. 2 (a) and (b) respectively.

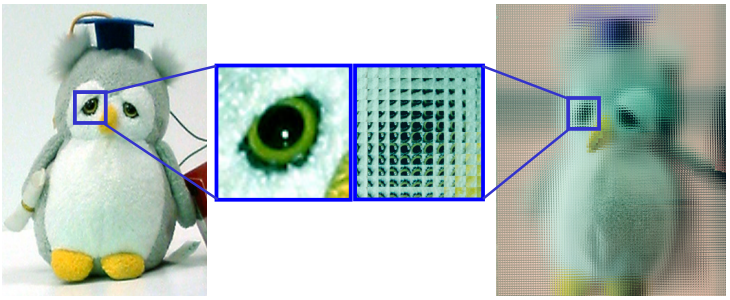
The optical resolution of the flatbed scanner used in the proposed setup is 3200dpi x 3200 dpi. When used in conjunction with a lens array of 0.1cm lens pitch high-quality IP images were produced, where each elemental image contained over 100 pixels in each dimension. The internal light source of the scanner was deactivated and external lighting was used in an effort to eliminate back-scattering occurring due to the incident light from the internal light source on the lens array. The only drawback of this device is the time needed for scanning the IP image of a 3D object, which limits its use to static objects. However, the resultant images unveil the potential of IP as a forthcoming standard for 3D acquisition and display. A high quality IP image,

along with the real world 3D object are depicted in Fig. 3 (a),(d). A magnification of a small part of the IP image in Fig. 3 (c) shows the rectangular structure of the lens array and the different views of the part of the object in Fig. 3 (b) in neighboring elemental images.



(a) (b)

Fig. 2. Acquisition Device: (a) Operation Principle (b) Experimental prototype.



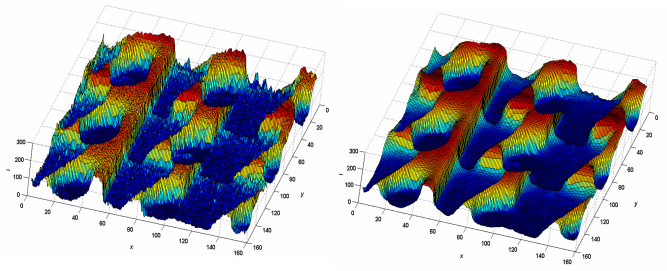
(a) (b) (c) (d)

Fig. 3. (a) Real world 3D object. (b) Magnified part of a portion of the object. (c) Corresponding IP image of the magnified part of the object in (b). IP image of the object in (a).

3 IP image Analysis

In this section, we develop a post acquisition filtering stage for noise reduction and propose a scene invariant method that automatically detects small rotational misalignments in IP image acquisition setups. In addition the technique described manages to measure the exact size and position of each elemental image without requiring prior knowledge on the system characteristics. In this way we can perform post acquisition processing in IP images or assist in the initial calibration of an IP image acquisition setup.

A Contrast-Limited Adaptive Histogram Equalization – CLAHE method is used in order to compensate for bad lighting conditions and a non linear 2D median filtering is applied in order to reduce excessive shot noise form the IP image. The results of the de-noising operation are depicted in Fig. 4.



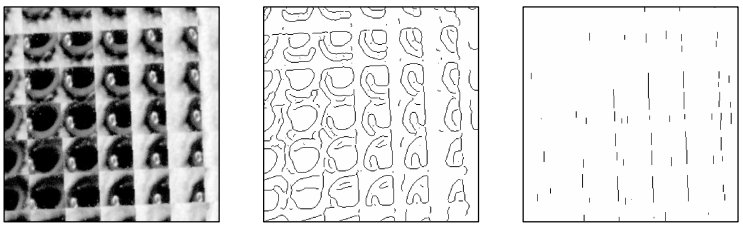
(a) (b)

Fig. 4. (a) Intensity values for a part of an IP image, (b) Intensity values for the same image after non linear 2D median filtering.

The rest of this section focuses on the accurate measurement of the skew angle of the Lattice Lines Structure (LLS) that reflects the rotational misalignment of the lens array used in the acquisition stage, in regard to the CCD sensor. It also accurately detects the lattice lines positions, which usually deviate from the equidistant case due to the non-integer values of the ratio of the lens array pitch to the CCD pitch. For this purpose, a non-linear filtering technique is used to enhance the boundaries of the elemental images along with the Hough Transform (HT) [11] in order to produce an accurate estimate of the skew angle. A lattice matching method is developed, based on the techniques used to solve the longest common subsequence (LCS) problem [12], in order to match a detected sequence of lines positions to a theoretical lattice model. The experimental data consist of a series of uncalibrated IP images acquired using the setup proposed in section two and computer generated IP images, using the method proposed in [13].

3.1 Skew Detection

A color IP image is initially converted to gray-scale as shown in Fig. 5a and the Canny edge detector [14] is deployed next in order to detect strong lines in the IP image and produce the edge image depicted in Fig. 5b. The image-wide lattice lines in the edge image are further enhanced using a one-dimensional (1D) median filter. On the contrary, shot noise and lines running to other directions are effectively attenuated due to their small size in regard to the median filter window. The filter is applied row-wise or column-wise and has the effect of attenuating everything except line segments that are almost horizontal or vertical respectively. In order to reduce the overall computational complexity of the algorithm, only one set of lattice lines, either horizontal or vertical, is considered in the skew detection process. As there is always a large number of elemental images in an IP image, there is also an adequate number of correctly detected lines running in a specific direction that is enough to produce an accurate estimation of the skew angle. The results of a vertical 1D median filtering are depicted in Fig. 5c.



(a) (b) (c)

Fig. 5. A portion of a physically acquired IP image. (a) Initial gray-scale image, (b) edge detection result, (c) edge image after vertical 1D median filtering.

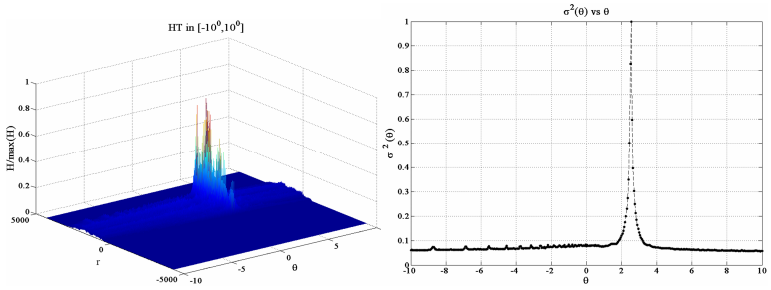
In the skew angle detection process the HT is used, which is a robust and extensively used technique in image processing applications. The HT succeeds in identifying only image-wide lines, which are more likely to correspond to true lines of the LLS that define the boundaries of the elemental images. In our framework we used a sampling rate of 0.05^0 which provides a high level of accuracy, and assumed that the skew angle of the IP image remains under $\pm 10^0$ in order to reduce the computational complexity of the algorithm.

The detection of the skew angle (θ_s) in the HT space is based on the fact that θ_s corresponds to a column in the HT accumulator array, **H**, having a large number of strong peaks, as shown in Fig. 6a. On the contrary the rest of the columns in **H** are relatively homogenous having a small number of weak peaks. In order to determine θ_s the statistical variance of each column in **H** is calculated and the results are shown in Fig. 6b.

3.2 LLS detection stage

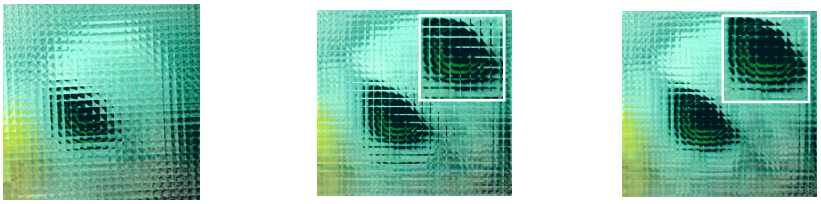
In order to derive the lattice lines positions that form the LLS the Canny edge detector and the 1D median filter are reapplied on the deskewed version of the initial gray-scale IP image. Since the image is deskewed, the lattice lines are properly oriented in horizontal and vertical directions. In order to enhance these image-wide lines the 1D median filter is applied row-wise and column-wise and the horizontal and vertical projection profiles [15] are calculated.

Peaks that are candidates for lattice lines in these two directions are identified using a global threshold at 20% of the maximum value of the corresponding profile. For each of the candidate peaks, a 1D version of the peak detection algorithm presented in [15] is used. The mean size of each elemental image is derived by calculating the average distance of all peaks detected in the previous stage in each direction. Finally a matching procedure utilizing the Smith-Waterman algorithm [16] follows that matches the detected LLS to a modeled LLS based on the mean size of each elemental image. The results of this procedure are shown in Fig. 7 as an application to pseudoscopic elimination [17].



(a) (b)

Fig. 6. (a) The column in the HT space corresponding to the skew angle θ_s has a large number of strong peaks, (b) the corresponding variance $\sigma^2(\theta)$ vs. θ.



(a) (b) (c)

Fig. 7. (a) Uncalibrated portion of an IP image. Pseudoscopy elimination in the deskewed IP image, using (b) constant and (c) variable, pitch size.

4 IP and Multiview Image Coding

IP images are regarded as an omnidirectional multiview image set. In order to properly exploit both intra-pixel and inter-elemental image redundancy which is inherent in IP images two encoders are proposed. The first encoder uses disparity estimation between adjacent elemental images operating on the basic principles of the MPEG-2 compression standard [18]. The encoder outperforms previously proposed 2D coding schemes like JPEG for IP images containing medium and large sized elemental images. However, as disparity estimation schemes become impractical with small sized elemental images an adaptive 3D Discrete Cosine Transform (3D-DCT) is used that effectively exploits the intra-pixel and inter-elemental image redundancy by properly grouping neighboring elemental images.

4.1 Disparity encoder

The perfect alignment of subsequent lenses in the array allows a unidirectional search to be performed at the disparity estimation step, which has a significantly lower complexity compared to a multidirectional search to estimate a disparity vector. This feature also allows the use of a full pixel search, which further improves the matching results. The best matching for a predicted block is decided by the use of a mean square error (MSE) estimator.

Disparity vectors produced by the disparity estimation process are finally DPCM coded, followed by Huffman encoding that further reduces the extra bits needed for the disparity vectors. Finally, standard intra-frame and residual frame coding were used as described in MPEG-2 for coding reference and residual elemental images respectively. The results of the comparison with baseline JPEG are depicted in Fig. 8.

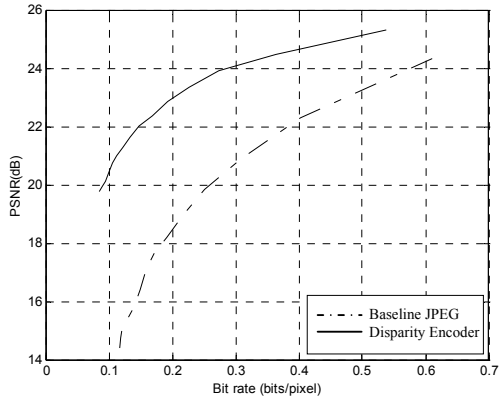


Fig. 8. Results for the Disparity Encoder vs Baseline JPEG.

4.2 Adaptive 3D-DCT encoder

The 2D lattice of elemental images is initially transformed to a 1D series of elemental images, using the a spiral curve as described in [7]. Consecutive elemental images are grouped together to form volumes on which the 3D-DCT will be applied. In our approach, the 3D-DCT is applied on groups of eight elemental images as a compromise between good quality and computational efficiency.

The encoder is assembled of a 3D-DCT unit, the quantizer and the entropy coder (EC). Two additional units are added to this standard setup used for detecting scene parts that belong to the background and determining the quantization values and the scan order of the 3D-DCT coefficients. The encoder layout is depicted in Fig 1.

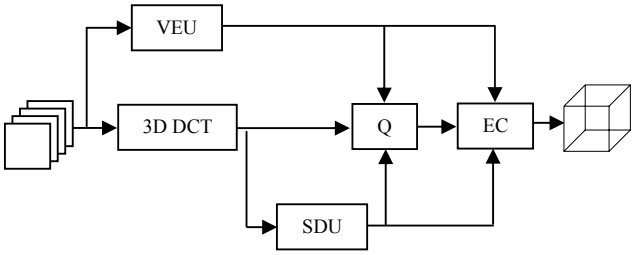


Fig. 9. Adaptive 3D-DCT Encoder.

The encoder uses a local variability strategy (VEU) on the initial data volume in order to point out parts of the image set that belong to the distant background. A global standard deviation strategy (SDU) is also used to locate the dominant 3D-DCT coefficients directivity and determine upon quantization values and scan order of the quantized coefficients. The performance results of the adaptive 3D-DCT scheme compared to baseline JPEG are depicted in Fig. 10.

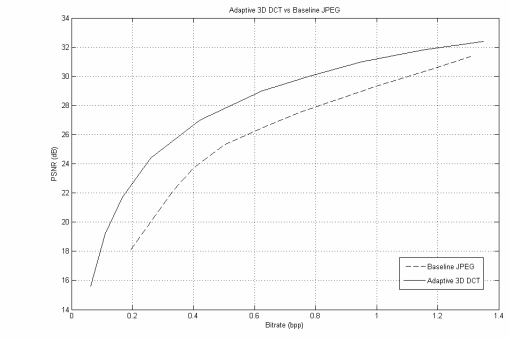


Fig. 10. Adaptive 3D-DCT vs Baseline JPEG.

4 Conclusions and Future Work

In this paper a holistic approach is presented for IP image acquisition, analysis, processing and encoding. The results showed that proper extensions of classic image analysis and processing algorithms achieve high accuracy and efficiency when applied on IP images. Future work includes 3D object reconstruction from IP images and development of algorithms and techniques for IP video acquisition and processing.

References

1. S.Pastoor, M.Wöpking, 3-D displays: A review of current technologies, Displays 17 (2), pp 100-110, 1997.
2. M.Halle, Autostereoscopic displays and computer graphics, Computer Graphics 31(2), pp. 58-62, 1997.
3. G. Lippmann, La Photographie integrale, C.R. Acad. Sci. 146, pp. 446-455, 1908.
4. J. S. Jang, B. Javidi, Two-step integral imaging for orthoscopic three-dimensional imaging with improved viewing resolution, Opt. Eng. 41(10), pp. 2568-2571, 2002.
5. J. S. Jang, B. Javidi, Formation of orthoscopic three dimensional real images in direct pickup one-step integral imaging, Opt. Eng. 42(7), pp. 1869-1870, 2003.
6. N.Sgouros, A.Andreou, M.Sangriotis, P.Papageorgas, D.Maroulis, N.Theofanous, Compression of IP images for autostereoscopic 3D imaging applications, In: Proc. IEEE ISPA03, pp. 223-227, 2003.
7. S.Yeom, A.Stern, B.Javidi, Compression of 3D color integral images, Opt. Express 12(8), pp. 1632-1642, 2004.
8. M. Levoy, "Light Fields and Computational Imaging," IEEE Computer, vol. 39(8), pp. 46-55, 2006.
9. J. Jang and B. Javidi, "Time-Multiplexed Integral Imaging For 3D Sensing and Display," Optics & Photonics News, vol. 15, pp. 36-43, 2004.
10. M. Martínez-Corral, B. Javidi, R. Martínez-Cuenca and G. Saavedra, "Integral Imaging with Improved Depth of Field by Use of Amplitude-Modulated Microlens Arrays," Applied Optics, vol. 43, pp. 5806-5813, 2004.
11. R.C.Gonzalez and R.E.Woods, *Digital image processing, second ed.*(Prentice Hall, NJ, 2002).
12. T.H.Cormen, C.E.Leiserson, R.L.Rivest, Introduction to algorithms, (MIT Press, MA, 2000).
13. S. S. Athineos, N. P. Sgouros et.al, "Photorealistic Integral Photography using a Ray Traced Model of the Capturing Optics," IS&T/SPIE Journal of Electronic Imaging, vol. 15(04), no. 043007, 2006.
14. J. F. Canny, "A Computational Approach for Edge Detection," Trans. Pat. Anal. Mach. Intell. 8, pp. 679-698, 1986.
15. R.C.Gonzalez, R.E.Woods and S.L Eddins, Digital image processing, using MATLAB (Prentice Hall, NJ, 2004).
16. T. Smith and M. Waterman, "Identification of common molecular subsequences," Journal of Molecular Biology, vol. 147, pp. 195-197, 1981.
17. M. Martinez-Corral, B. Javidi, R. Martínez-Cuenca, and G. Saavedra, "Formation of real, orthoscopic integral images by smart pixel mapping," Opt. Express 13, pp. 9175-9180 2005.
18. K. Rao, J. Hwang, Techniques & standards for image-video & audio coding, Prentice Hall,1996.

Study of All-Optical Wavelength Conversion and Regeneration Subsystems for use in Wavelength Division Multiplexing (WDM) Telecommunication Networks.

Hercules Simos *

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications

simos@di.uoa.gr

Abstract. In this thesis, we study all-optical processing techniques based on non-linear semiconductor optical amplifiers for the functionalities of second and third generation optical networks. In particular, devices for all-optical wavelength conversion with regenerative properties have been investigated. The configurations under investigation are based on the non-linear four-wave mixing process which occurs when fields with proper spectral and power arrangement are coupled into the active medium. The performance characteristics of the devices are determined by the properties of the optical amplifier, as well as by the operating conditions like the gain of the optical amplifier, the power levels and the frequency detuning of the pump and the information signal.

Keywords: All-optical wavelength conversion, four wave mixing, regeneration, semiconductor optical amplifier, amplified spontaneous emission noise.

1 Introduction

All-optical wavelength conversion is considered as a key function for the future WDM lightwave systems. Wavelength conversion addresses a number of key issues in WDM networks including transparency, interoperability, and network capacity. Wavelength conversion may be the first obstacle in realizing a transparent WDM network. High bit rate and efficient conversion has been demonstrated using techniques like cross-gain modulation (XGM), cross phase modulation (XPM) and four wave mixing (FWM) in a variety of passive or active media [1]. FWM in traveling wave semiconductor optical amplifiers (TW-SOAs) is probably the most favored and studied technique for λ-conversion, since it offers transparency to the modulation format and to the applied data rate up to tens of Gbps [1].

On the other hand, in high bit rate optical communication systems and networks, data signals suffer from degradation during their propagation, due to noise, pulse

* Dissertation Advisor: Dimitris Syvridis, Professor

distortion and crosstalk. The accumulation of amplified spontaneous emission (ASE) noise generated by amplifiers degrades the signal to noise ratio (SNR). Chromatic dispersion affects the pulse shape, while different types of crosstalk due to the propagation through optical cross connects, wavelength converters and filters, degrade the signal quality. All-optical regenerators are critical components for the restoration of these signal impairments, avoiding the optoelectronic conversion limitations. Regeneration can be either 2R for signal reshaping, or 3R for both signal reshaping and retiming. Several regenerators have been proposed so far, based on XPM, cross-gain modulation XGM in passive or active media. Although the four-wave mixing process is (FWM) is exhaustively investigated for λ-conversion applications, only recently, FWM in DSF has been employed as a 2R all-optical regenerator [2], [3].

In this thesis, we study all-optical processing techniques based on non-linear semiconductor optical amplifiers for the functionalities of second and third generation optical networks. In particular, devices for all-optical wavelength conversion with regenerative properties, based on FWM in a SOA, have been investigated in detail. In the first part of this thesis, the noise properties of the converted signal generated by the four-wave mixing based wavelength conversion process were investigated. The investigation of the noise characteristics was carried out by a theoretical analysis and the corresponding experimental confirmation, including the separate study of noise induced by the four wave mixing process itself and the amplified spontaneous emission noise from the amplifier. The study of the noise properties at the new wavelength was performed for the intensity as well as the phase noise of the converted signal. The useful conclusions from this study were used for the investigation of the regenerative properties of the mixing process.

The second part of this thesis starts with the experimental investigation and the confirmation of the regenerative properties of the four wave mixing process in a semiconductor optical amplifier. The regeneration and simultaneous wavelength conversion of 2.5 Gbps optical signals is experimentally demonstrated. Furthermore, the second part includes the design and optimization by numerical simulation of a wavelength conversion system with regenerative properties, for use in nodes of wavelength division multiplexing optical networks. In this thesis, an alternative configuration for the pump and data signal is proposed for the first time, in order to obtain non-linear response from the optical amplifier. Based on this approach the new signal exhibits improved noise characteristics and lower bit error rate. The numerical investigation showed successful regenerative operation at 40 Gbps.

2 Numerical modeling of FWM in SOAs

The FWM process in a SOA can be generally described as follows. Pump (A_1) and signal (A_2) waves at optical frequencies ω_1 and ω_2 respectively, with the same state of polarization, are injected into the SOA from the same facet. The beating of the two co-propagating input waves inside the SOA generates refractive index and gain gratings at the frequency $\Omega = \omega_2 - \omega_1$. The input waves are scattered by these gratings which in turn leads to the creation of product waves, the conjugate (A_3) and the

satellite (A_4) at optical frequencies ω_3 and ω_4 respectively, with $\omega_3 = \omega_1 - \Omega = 2\omega_1 - \omega_2$ and $\omega_4 = \omega_2 + \Omega = 2\omega_2 - \omega_1$. The numerical analysis is based on the position dependent gain rate equation and the field propagation equations. The propagation and the nonlinear interaction between the input waves and the FWM products are described by the coupled - mode equations. The expression for the first FWM signal is given below [4]:

$$\frac{\partial A_3}{\partial z} = \frac{1}{2}[g(1 - i\alpha_{CDP}) - a_{loss}]A_3 - \frac{1}{2}\left(\eta_{3,1}|A_1|^2 + \eta_{3,2}|A_2|^2 + \eta_{3,4}|A_4|^2\right)A_3 - \frac{1}{2}(\eta_{1,4} + \eta_{2,4})A_1 A_2 A_4^* - \frac{1}{2}\eta_{1,2}A_1^2 A_2^* + A_{SE,3} \quad (1)$$

The variation of the time dependent gain $g = g(z, t)$, for a subsection is given by [4]:

$$\frac{\partial g}{\partial t} = \frac{g_s - g}{\tau_s} - g \cdot \frac{P_{tot}}{P_{sat}\tau_s} \quad (2)$$

$A_i = A_i(z, t)$, $i = 1, 2, 3, 4$, is the slowly varying envelope of the fields; $A_{SE,i} = A_{SE,i}(z,t)$, represents the spontaneous emission noise generated and added to the waves in each subsection; it is modeled as a white Gaussian distributed process [4]; The coefficients $\eta_{i,j}$, $i, j = 1, 2, 3, 4$, $i \neq j$, represent the non-linear interactions among the mixing waves, which they are related to the inter-band and intra-band carrier dynamics: the carrier density pulsations (CDP), the carrier heating (CH) and the spectral hole burning (SHB). The expression for each mechanism, as well as all other parameters can be found in [4], [5]. The four output fields $A_i = A_i(z = L, t)$, $i = 1, 2, 3, 4$, are calculated by numerically integrating the equations (1) and (2) over the SOA length. Integration is accomplished by sampling the length of the SOA at subsections of length dz and the time-dependent fields at samples of duration dt. The simulations are carried out employing either continuous waves (CW) or Non Return to Zero (NRZ) pulsed inputs.

3 Noise properties of FWM in a SOA

The major drawbacks of the FWM process in SOAs, are the efficiency degradation at high detuning values and the optical signal-to-noise ratio (OSNR) degradation due to the amplified spontaneous emission (ASE). In general the spontaneous emission perturbs both the amplitude and phase of the converted signal, resulting to its intensity and phase noise degradation.

3.1 RIN performance of wavelength converters based on FWM in SOAs

The relative intensity noise (RIN) properties of the FWM converted signal have not been investigated up to now, neither theoretically nor experimentally. Such an

investigation is necessary since RIN affects the performance of lightwave systems using intensity modulation. In this section the numerical and experimental study on the RIN performance of the FWM in SOAs is presented. The influence of the power as well as the intensity noise properties of the input signals, was investigated. Although in most cases a degradation of the converted signal is observed, it will be shown that improvement in the RIN performance is possible under certain operating conditions [5]. This property reveals the regenerative capability of the FWM process.

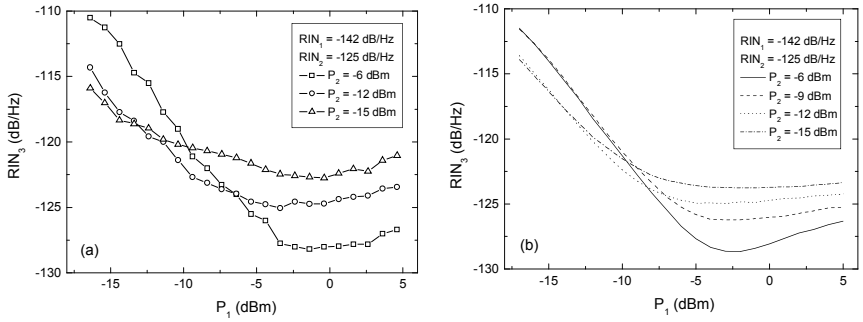


Fig. 1. RIN of the wavelength converted signal against input pump power with $RIN_1 = RIN_2 =$ -142 dB/Hz: (a) experimental and (b) numerical results.

The intensity noise of the converted signal against input power is shown in fig. 1. Figure 2 depicts the dependence of the output signal RIN (RIN_3) on the input pump power (P_1) for different levels of the input power of the signal (P_2). Figure 1a corresponds to experimental data and figure 1b to calculated results. In both cases the RIN of the input signals was RIN_1 = -142 dB/Hz and RIN_2 = -125 dB/Hz. The dependence of the output RIN on the pump power (P_1) is not monotonous, but it exhibits a saturation regime at high pump power levels. In addition, the output RIN depends inversely proportional on the input signal power (P_2), at high levels of P_1. The latter does not hold for low power levels of A_2 (P_2). In this case, the roles of the two input waves change as A_2 becomes the pump and A_1 the signal. The explanation for the behavior of fig. 1 can be given through the power transfer functions of the FWM process. As P_1 increases, the conversion efficiency gradually saturates reaching the regime where further increase of P_1 results in constant or even decreasing P_3. This regime appears at higher P_1 values as P_2 increases. At the same time the SOA gain has reached already the saturation, resulting in a more or less constant level of ASE noise. This means that for the P_1 range where the conversion efficiency is not saturated, while the ASE noise is saturated, the output RIN decreases with P_1, as observed in fig. 1. Further increase of P_1, results in decrease of the converted signal power and therefore increase of the RIN_3, in agreement with the experimental and numerical findings of fig. 1.

An interesting finding from fig. 1, which could lead to useful conclusions from the application point of view, is related to the fact that the output RIN_3 is lower relative to the input signal noise RIN_2. This occurs in the saturation regime (high pump power levels) when the signal power is also high (e.g. in fig. 1, curves corresponding to P_2 = -6 dBm, P_2 = -9 dBm). In order to clarify the physical mechanism behind this

behavior, calculations of RIN_3 against P_1 have been carried out, assuming a noise-free SOA. It was found that when the input noise is low (RIN_1 = RIN_2 = -142 dB/Hz) the output noise RIN_3 in the saturation regime is close to the input RIN_2. In the case where the input signal noise is higher (RIN_1 = -142 dB/Hz, RIN_2 = -125 dB/Hz), RIN_3 is reduced with respect to the input RIN_2. For the pump power level where the process is unsaturated, the combination of the low output power P_3 with the high noise from the signal, leads to output RIN_3 levels close to the high input value RIN_2.

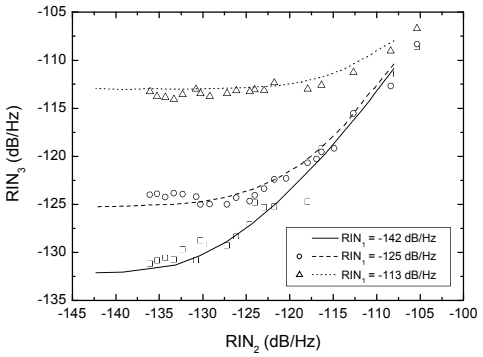


Fig. 2. Output noise RIN_3 versus input signal noise RIN_2, for different levels of the input pump noise RIN_1. Lines correspond to numerical calculations and symbols represent experimental data. P_1 = 0 dBm, P_2 = -6 dBm

In order to confirm the input noise influence, the dependence of RIN_3 on RIN_2 has been investigated both numerically and experimentally (fig. 2), for different levels of RIN_1. The power levels of P_1 and P_2 have been properly chosen in order to ensure highly efficient conversion. A general observation that can be made is that the upper limit of RIN_3 is set by the signal noise (RIN_2), while the lower limit is set by the pump noise (RIN_1), except of the case of very low RIN_1 (-142 dB/Hz) where the lower limit is set by the ASE noise. The limit on the lower values is no longer set by the pump noise, when the input value of RIN_1 is very low.

It is obvious from the above discussion, that the overall RIN performance of the converter is a combination of two discrete processes. The first is the ASE noise and its dependence on the input power. The second is the transfer of the intensity noise from the input waves, to the output through the FWM process. ASE dominates in the low pump power regime, where the FWM efficiency is poor, whereas at the saturation regime, RIN carried by the input signal predominates over ASE.

3.2 Phase noise performance of wavelength converters based on FWM in SOAs

The result of this part was a detailed theoretical and experimental study of the phase noise characteristics of FWM in a SOA based wavelength converters. The theoretical study for the spectral linewidth of the conjugate, the new four-wave mixing component used for conversion, predicted that the new wave has a spectral

linewidth strongly dependent on the pump linewidth (x4) plus the probe linewidth. This effect was confirmed by experimental results for several operating conditions (different amplifiers, input linewidth). The results showed that the linewidth of the converted signal is independent of the amplified spontaneous emission noise induced from the amplifier, and that the linewidth enhancement observed, is caused by the non linear dynamics of the four wave mixing process. The numerical investigation confirmed the analytical theory for in high pumping conditions, however low pump power induces high levels of phase noise to the conjugate due to strong influence of ASE noise.

3 Experimental investigation of the regenerative properties of FWM in a SOA

In this part, the possibility to achieve extinction ratio (ER) enhancement and noise suppression in SOA based FWM λ-converters, is investigated for the first time in detail, assuming NRZ data format carried by one of the optical inputs according to the typical wavelength conversion operating scheme. We show that in most cases the SOA saturation dominates resulting even in ER degradation of the output relative to the input. Nevertheless there is an operating regime where significant ER enhancement can be achieved together with noise suppression while the obtained output remains NRZ modulated. The experimental setup for the measurements under dynamic operation is shown in fig. 3.

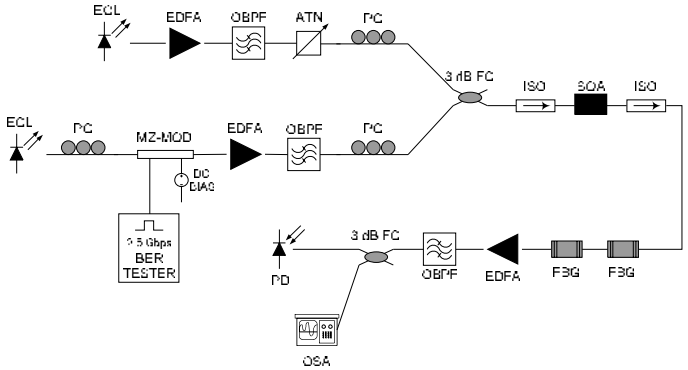


Fig. 3 The experimental setup used for the dynamic measurements of the wavelength conversion based on FWM in a SOA. EDFA: erbium-doped fiber amplifier, BPF: band-pass filter, PC: polarization controller, MOD: modulator, SOA: semiconductor optical amplifier, OI: optical isolator, FBG: fiber Bragg grating.

In order to characterize the regenerative properties of the FWM process, the static transfer function of the process was experimentally determined by measuring the output power of the conjugate product as a function of the pump wave input optical power while keeping the signal wave optical power, at a constant value. The results, for three different optical power values of the signal (16 μW, 35 μW and 50 μW) are

shown in fig. 4a. The results show a well understood behavior discussed already in the past [6]. The conjugate optical power assuming operation in the unsaturated regime, is given by $P_3 = G_1^2 G_2 P_1^2 P_2 R(\Omega)$ where P_1 and P_2 are the pump and signal optical power respectively, G_1, G_2 are the corresponding small signal gain and $R(\Omega)$ depends on the detuning between signal and pump. Only at small signal conditions and with totally unsaturated SOA the gain remains constant and therefore the optical power of the conjugate depends on the square of the pump power, resulting in extinction ratio improvement of the FWM product. Such conditions are practically infeasible as it shown in fig. 4a. Even for very low signal power of 16 μW and accordingly low pump power levels, the conjugate power increases for a while as the pump power reaches to a value of approximately 80 μW. Further increase of the pump results in decreasing the conjugate power due to the gain saturation of the SOA.

For the conditions of the transfer function corresponding to $P_1 = 50$ μW, dynamic measurements were carried out; the corresponding results (fig. 4b) show enhancement of the output extinction ratio of 2-3 dB. Bit error rate measurements showed a 1-3 dB improvement (negative sensitivity penalty) of the regenerated output versus the input signal and for BER values in the range from 10^{-11} to 10^{-7} [7]. Fig. 4b (inset) shows eye diagrams of the signal at the input (lower) and the regenerated output (upper).

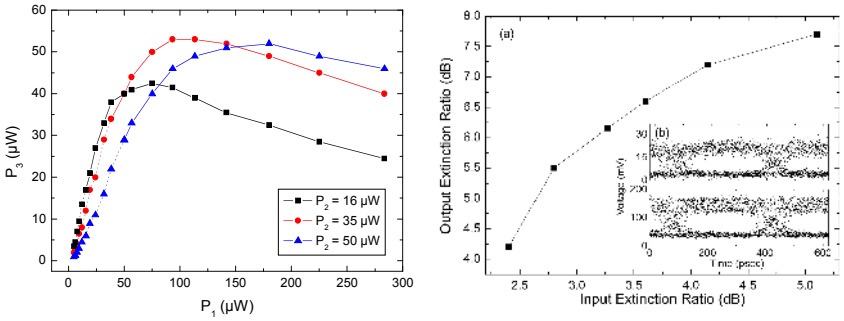
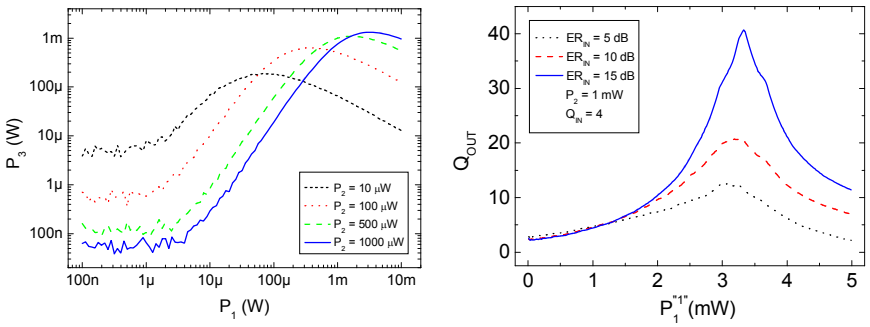


Fig. 4 (a) Experimental transfer functions of FWM in a SOA (b) experimental dynamic measurements of the output extinction ratio versus the input ER.

4 Design and numerical evaluation of a wavelength converter with 2R regenerative properties based on FWM in a SOA

Up to now, none of the conventional FWM configurations have exhibited regenerative properties. The experimental investigation of the previous section showed that if the data stream is applied on the pump wave instead of the signal, and under certain conditions for the power levels of the modulated wave relative to the CW one, regenerative characteristics appear at the λ-converted wave. However, the nature of the FWM based regeneration process, which is related to the SOA's gain saturation, sets an upper limit at the operation speed of the whole process due to the limited response of the SOA's carrier dynamics. Major objectives of the present

section is to investigate the above speed limitations and to identify the optimum SOA's structural characteristics as well as FWM operational conditions at which regenerative behavior is expected. In order to increase the maximum achievable bit rate for this type of regenerator, a fiber Bragg grating (FBG) is employed at the output of the SOA, to act as an optical discriminator increasing this way significantly the upper limit of the bit rate operation of the regenerator. The study is structured as follows: the numerical model is solved numerically for the case of CW inputs, in order to calculate a set of static transfer functions of the regenerator corresponding to different FWM operating conditions. Using these static transfer functions, the operating conditions which correspond to the best regenerative performance, are identified. Based on the optimized operating conditions, the above FWM model is solved in the time domain with NRZ modulated input data up to 40 Gbps, and the corresponding output ER and Q-factor are calculated.



(a) (b)

Fig. 5 (a) Static $P_3(P_1)$ transfer functions of the FWM process in a SOA. (b) The expected Q-factor at the output, calculated using the transfer function with P_2 = 1000 μW shown in (a).

4.1 Static results

The optical power P_3, of the product A_3 is plotted against the power P_1 of the input wave A_1 (fig. 5a), for different power levels P_2 of the input wave A_2, according to the FWM configuration mentioned above. Obviously from the group of curves of figure 5a, the increase of P_2 results in a much more satisfactory non-linear transfer function like the one of the ideal regenerator. The $P_3(P_1)$ behavior has been observed and discussed in the past [6]. Using the transfer function corresponding to P_2 = 1000 μW of figure 5a, we calculated the output Q-factor and extinction (ER) for a constant input Q factor of 4 and for different input ER values. The output Q-factor is plotted in fig. 5b. It is obvious that, the best performance in terms of Q-factor is achieved when the input power level of A_1 is in the saturation regime of the transfer function. On the contrary, maximum ER enhancement is achieved at lower input power levels, corresponding to the linear regime of the transfer function.

Similar calculations have been performed to study the wavelength detuning between the input signals and the SOA length influence to the regenerator's static

performance. It is shown [4] that low values of wavelength detuning and long SOAs device are preferable for our application where high conversion efficiency is critical.

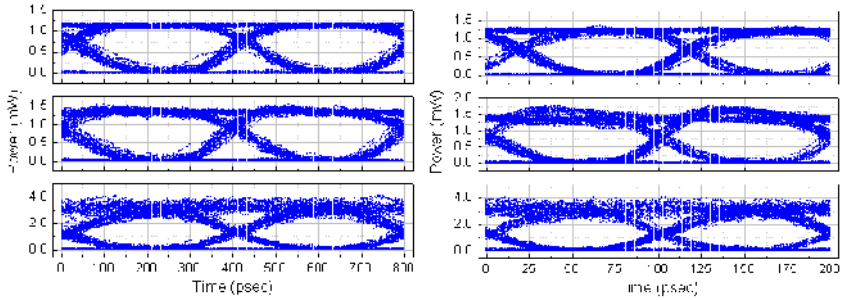


Fig. 6. Calculated eye diagrams at the input, at the SOA output and at the FBG output (from bottom to top) for operation at (a) 2.5 Gbps and (b) 10 Gbps.

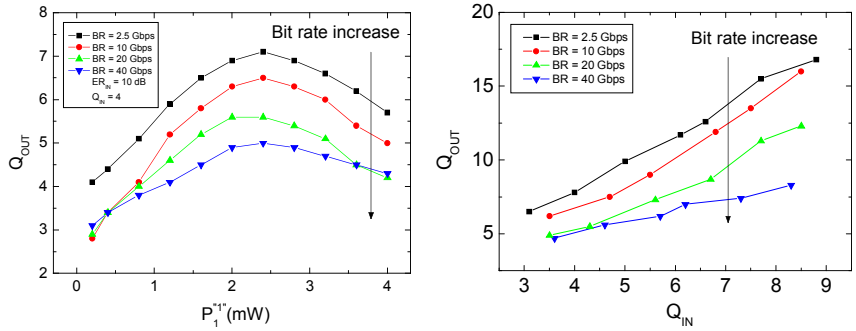


Fig. 7. Calculated Q-factor at the input, as a function (a) of the input power of the logic symbol "1" and (b) of the input Q-factor, for different bit rates.

4.2 Dynamic results

As mentioned earlier, the dynamic behavior of the FWM based regenerator is expected to be much different from the static one. This is because, at the dynamic operation of the system, several speed limitations are imposed by the SOA carrier dynamics, which do not appear at a typical FWM scheme. Due to the strong variation of P_1 and the fact that a part of the regeneration process is based on the SOA gain saturation, it can be easily shown both theoretically and experimentally, that the regenerator's performance at high data rates is limited by the carrier dynamics of the SOA. Typical eye diagrams at 2.5 and at 10 Gbps are shown at figure 6a and 6b, respectively. In the case of 2.5 Gbps very satisfactory results are obtained in both ER improvement and noise suppression even without the use of the grating filter. In the case of 10 Gbps, the noise suppression achieved by the FWM process is degraded by the pattern dependent response of the SOA (fig. 6b, middle), and finally restored by the FBG (fig. 6c, up). The FBG acts as a differentiator and restores the integration action of the SOA. Therefore, the overall regeneration performance at high bit rates is

a result of the combined action of the FWM originated noise suppression and the SOA speed enhancement through the FBG filter.

In fig. 7a we have plotted the output Q-factor as a function of the power level of the pump wave A_1, and in 7b the output Q versus the input Q, at several bit rates. The optimum performance for all rates occurs when the power of logic "1" is located at the saturated regime, we have maximum noise suppression for "1" power levels. Moving toward higher bit rates (10, 20, and 40 Gbps), a degradation mainly to the Q-factor performance, is observed. The degradation of the regeneration process is mainly caused by the speed limitations of the SOAs carrier dynamics as aforementioned. At 10 Gbps, there are remarkable regenerative characteristics. While at 20 Gbps, the performance is further decreased, and finally, at 40 Gbps, the regenerative behavior of the system is observed only at high values of the input ER. Finally the output Q shows linear dependence on the input Q for all rates.

5 Conclusions

Detailed investigation of the possibility to achieve all optical 2R regeneration using FWM in SOAs has been carried out. The alternative configuration of FWM proposed in this thesis is mainly determined by the operating conditions, mainly by the adjustment of the input power levels of the modulated signal. The FWM static transfer functions approximate the step-like characteristic of an ideal regenerator. The noise properties of the converted signal were investigated. The useful conclusions from this study were used for the investigation of the regenerative properties of FWM process. The expected regeneration properties are confirmed experimentally under static and dynamic operation at 2.5 Gbps optical signals. Furthermore, the design and optimization by numerical simulation of a wavelength conversion system with regenerative properties showed successful regenerative operation up to 40 Gbps.

References

1. S. J. B. Yoo, "Wavelength conversion technologies for WDM network applications," J. Lightw. Technol., vol. 14, pp. 955–966 (1996)
2. E. Ciaramella, S. Trillo, "All-Optical Signal Reshaping via Four-Wave Mixing in Optical Fibers", IEEE Photon. Technol. Lett., vol. 12, pp. 849-851 (2000)
3. A. Bogris and D. Syvridis, "Regenerative properties of a pump-modulated four wave mixing scheme in dispersion shifted fibers", J. Lightw. Technol., vol. 21, pp. 1892-1902, (2003)
4. H. Simos, A. Bogris, and D. Syvridis, "Investigation of a 2R all-optical regenerator based on four-wave mixing in a semiconductor optical amplifier", J. Lightw. Technol., vol. 22, pp. 595–604 (2004)
5. H. Simos, I. Stamataki and D. Syvridis, "Relative intensity noise performance of wavelength converters based on four-wave mixing in semiconductor optical amplifiers", IEEE J. of Quantum Electron., vol. 43, pp. 370-377, (2007).
6. A. D'Ottavi, E. Iannone, A. Mecozzi, S. Scotti, P. Spano, R. Dall'Ara, J. Eckner, G. Guekos, "Efficiency and noise performance of wavelength converters based on FWM in semiconductor optical amplifiers", IEEE Photon. Technol. Lett., vol. 17, pp. 357-359 (1995)

7. H. Simos, A. Argyris, D. Kanakidis, E. Roditi, A. Ikiades, and D. Syvridis, “Regenerative properties of wavelength converters based on FWM in a semiconductor optical amplifier”, IEEE Photon. Technol. Lett., vol. 15, pages 566–568 (2003)