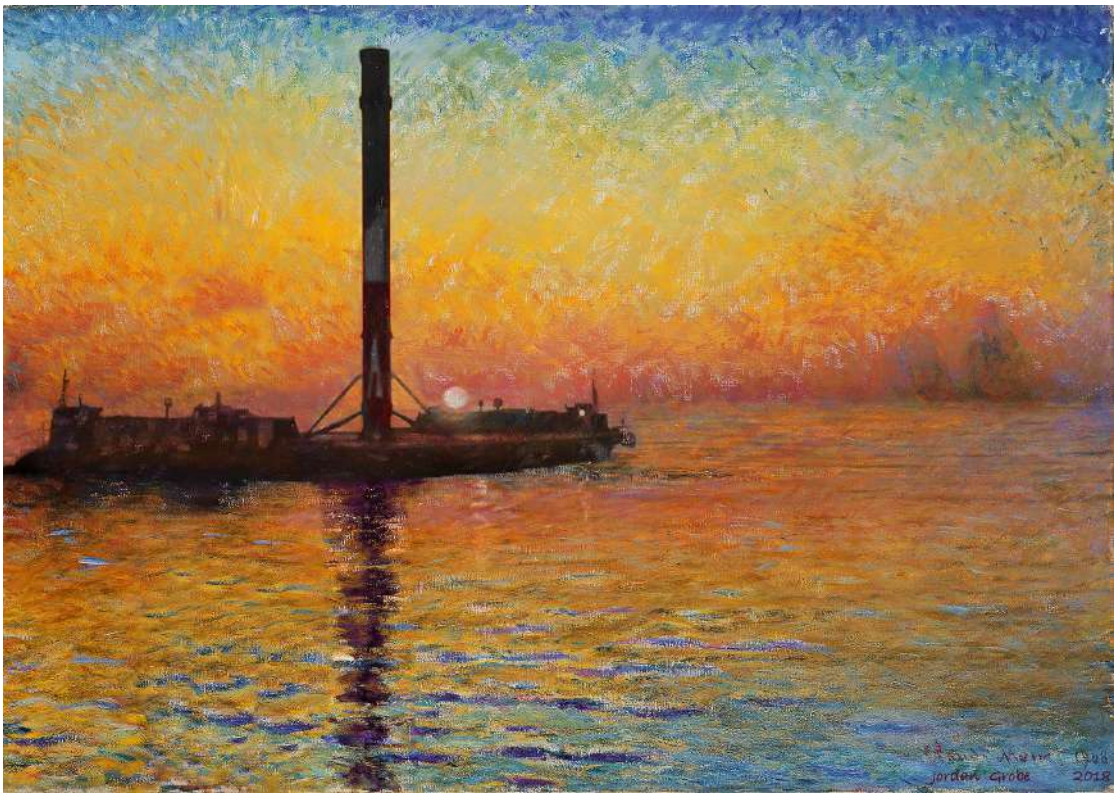
HELLENIC REPUBLIC
National and Kapodistrian University of Athens
Department of Informatics and Telecommunications

ABSTRACTS OF DOCTORAL DISSERTATIONS



Athens 2019
Volume 14

HELLENIC REPUBLIC
National and Kapodistrian University of Athens

Department of Informatics and Telecommunications

ABSTRACTS OF DOCTORAL DISSERTATIONS

The Committee of Research and Development Activities

M. Koubarakis

E.S. Manolakos

T. Theoharis

ISSN: 1791-7948

Volume 14

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

PREFACE

This volume includes extended abstracts of Doctoral Dissertations conducted in the Department of Informatics and Telecommunications, University of Athens, and completed from 1/2018 to 12/2018.

We publish this volume to demonstrate the breadth and quality of the original research performed by our Ph.D. students and faculty and to facilitate the dissemination of their innovative research results. We are happy to present the 14th yearly collection of this kind and expect this initiative to continue in the years to come. The submission of an extended abstract in English is required by all graduating doctoral students in our Department.

We would like to thank all graduates who contributed to this volume and hope that this was a positive experience for them. Finally, we would like to thank PhD candidate Nikos Bogdos for his help and attention to detail in putting together this volume.

The digital painting in the cover is called *Falcon 9 Booster at Dusk* (2018) by *Reddit user JGrobe.*

The DiT Dept. Committee on Research and Development Activities

M. Koubarakis
E.S. Manolakos (publication coordinator)
T. Theoharis

Athens, June 2019

Table of Contents

Context-based Resource Management and Slicing for SDN-enabled 5G Smart, Connected Environments

Sokratis N. Barmpounakis*[1]
National and Kapodistrian University of Athens, Greece
Department of Informatics and Telecommunications
sokbar@di.uoa.gr

Abstract. 5G mobile communication systems will address unprecedented demands in terms of system capacity, service latency and number of connected devices. Future 5G network ecosystems will comprise a plethora of 3GPP and non-3GGP Radio Access Technologies (RATs), such as Wi-Fi, 3/4/5G, etc. Deployment scenarios envision a multi-layer combination of macro, micro and femto cells where multi-mode end devices, supporting diverse applications, are served by different technologies. This thesis focuses on the radio resource management (RRM) from the perspective of the primary RAT and cell layer selection processes (i.e., cell (re)selection, handover, admission control); afterwards, it goes one step beyond, in order to link the RRM with Network Slicing approaches, as introduced in Software Defined Networking (SDN)-enabled environments, which creates smaller, virtual "portions" of the network, adapted and optimized for specific services/requirements. Towards this end, this thesis introduces a context-based radio resource management (RRM) framework, comprised of three distinct mechanisms: Two out of the three mechanisms exploit contextual information, while the third mechanism acts with an augmenting role to the former two, by pre-processing the contextual information required by such, context-based mechanisms and –thus- by limiting the signaling cost required for communicating this contextual information among network entities. Last but not least, a comprehensive analysis has taken place in relation to the architectural aspects, in the context of the forthcoming 5G network architecture and by mapping them with the latest 5G network components – as these were introduced in the latest 3GPP work-.

Keywords: 5G, Radio Resource Management, RAT selection, User Profiling, Handover, Context, SDN, Network Slicing.

1 Dissertation Summary

In the following years, the number of wireless and mobile devices is expected to increase considerably. Along with it, a huge increase of mobile traffic [1] will also take place. More specifically, the mobile traffic in 2016 was nearly 30 times the size of the entire global Internet in 2000. Almost half a billion mobile devices and connections were added in 2016, while at the same time, average smartphone usage grew 20% in

* Dissertation Advisor: Athanasia (Nancy) Alonistioti, Assistant Professor

the same year. In addition, the actual traffic volume per subscriber increases 25-40% per year, thus exceeding the expectations set by ITU [2]. The deployment of 5G cellular networks targets to support this vast number of devices, while at the same time existing 3GPP specifications will keep on supporting legacy cellular access networks (e.g., GSM, HSPA, LTE, LTE-A), as well as alternative radio access technologies (e.g., Wi-Fi). In this environment, the end users will have access to a diverse set of services (high definition video and audio, web browsing, games, etc.). It is worth pointing out in parallel that the high penetration of smartphones and tablets on the market [3] will enable end users to make use of all these services while on the move.

5G networks are expected to support billions of small end devices (e.g., sensors, actuators, etc.) as well as communicating vehicles [4] in the context of Machine Type Communication (MTC). The vision is that 5G networks will manage to materialize the Internet of Things (IoT) ecosystem. This realization will unveil new requirements to the network operators and telecom manufacturers. The aforementioned requirements should also be taken into consideration by the underlying network infrastructure. Overall, existing mechanisms used for the communication of end terminals are inadequate to support the future needs. Towards improving efficiency, well- established mechanisms have to be redesigned. One of the most important areas 5G networks have to improve is the mapping of smart devices and services to different RATs and layers (i.e., macro, micro, pico, femto cells). This mapping affects the Key Performance Indicators (KPIs) of a network in relation to the experienced grade of service (e.g., blocking probabilities, throughput, delay, jitter, etc.). The placement of a UE to a RAT or a cell layer, that is either in idle or a connected mode, is primarily realized via three vital RRM mechanisms, namely: a) Cell-(re)Selection (CS), b) Access/Admission Control (AC) and c) Handover (HO).

The contributions of this dissertation move towards the following major directions:

. a) COmpAsS, a user-oriented context-based scheme for RAT selection and traffic steering/switching, which processes context in real-time and produces a RAT suitability list to be used for handover management reasons,

. b) CEPE, a knowledge extraction engine based on data mining techniques towards user profiling and network policies formulation,

. c) CIP, a context information pre-processing scheme, which acts in an augmenting manner to the former two, in order to minimize the high context acquisition signaling overhead.

. d) A study on CEPE-COmpAsS interworking in an SDN-enabled, 5G architecture, capable of applying network slicing approaches.

Supplementary contributions, which enforce the research carried out in the aforementioned primary four directions and also comprise parts of the next steps to be made in the context of this research domain are:

- the architectural perspective of the proposed schemes, which takes into account the latest 3GPP standardization guidelines and attempts to provide a valid and viable solution towards the forthcoming 5G architecture,
- a study on network traffic engineering policies, which can exploit CEPE and user profiling methodology is included,

- an attempt to describe from a common point of view 3 primary RRM mechanisms, i.e. cell (re)selection, handover and call admission control via a comprehensive categorization of the existing approaches, both from the academic area, as well as from the industry, by incorporating the available patents as well, and
- a 5G use case application related to IoT and Precision Farming, which highlights specific requirements related to industrial applications, ultra-low delay requirements, etc.
-

The first major contribution of this thesis is COmpAsS, a context-aware, user-oriented RAT selection mechanism, which operates on the User Equipment (UE) side and ultimately produces a list of the most suitable RATs per active traffic flow/session, towards QoS optimization. One of the greatest advantages of the UE-based solution is the minimization signaling overhead over the air interface, as well as the computation load on the base stations. COmpAsS collects information related to the network status, such as the load of the base stations, the load of the backhaul link, the Reference Received Signal Quality (RSRQ), user mobility information, such as the velocity of the UE, as well as the specific QoS requirements of the type of traffic to be transmitted, in order to assess -in real-time- the most suitable RAT and/or cell layer, which should serve the UE's active sessions. COmpAsS mechanism adopts Fuzzy Logic (FL) as one the core logic modules, responsible for the perception of the network situation and, in combination with a set of pre-defined rules, calculates a list of the most suitable available access network options. Furthermore, we propose an evolution of 3GPP's Access Network Discovery and Selection (ANDSF) function, as one of the primary Evolved Packet Core (EPC) network functions collaborating with COmpAsS for the exchange of the required context information. The merits of COmpAsS are showcased via an extensive series of simulation scenarios, as part of 5G ultra dense networks (UDN) use cases. The results prove how the proposed mechanism optimises Key Performance Indicators (KPIs), when juxtaposed to a well-established LTE handover algorithm.

The second major contribution of the current thesis the Context Extraction and Profiling Engine (CEPE), a resource management framework, which collects diverse types of context information and performs data mining techniques in order to extract meaningful knowledge. The context information, which is aggregated, primarily relates to four categories: network operation data, user behavior information, terminal capabilities and application/service data. CEPE analyzes this information, extracts meaningful knowledge and performs user profiling in order to apply it for optimal resource planning, as well as prediction of resource requirements. CEPE collects information about users, services, terminals and network conditions and –based on offline processing– derives a knowledge model, which is subsequently used for the optimization of the primary RRM mechanisms, i.e. handover, cell selection and call admission control. From a methodological point of view, initially the KPIs that will be employed are identified in order to assess the efficiency of the mechanism. Next, the types of data that should be monitored are identified (network operation data, user behaviour information, etc.). Then, the extracted context information is translated into user profiles and is finally applied as input for enhanced cell (re)selection, handover or admission control. CEPE's operation is tightly connected to the scheme, which follows, CIP, and focuses

on the pre-processing of the vast amount of information, which is collected, towards minimizing the signaling overhead. The viability and validity of CEPE is demonstrated via an extensive set of simulation scenarios.

The third major contribution is CIP, a Context Information Pre-processing scheme, aiming to identify and discard redundant or unnecessary data before knowledge extraction. CIP could be considered as an integral part of the afore described profiling schemes, i.e. COmpAsS and CEPE. CIP comprises a framework that primarily relies upon data aggregation and pre-processing techniques. Context information processing and knowledge extraction is considered a great tool towards the optimisation of several network functions; nevertheless, the acquisition of the context is often a very costly process –in terms of signaling burden imposed on the network. The module comprises aggregating and compressing mobile network-related context information per unique identifier, such as the end device's International Mobile Subscriber Identity (IMSI), as well as techniques related to identifying and discarding user profile-redundant or unnecessary context data, before any transmission to CEPE. CIP gains are illustrated via a detailed analytical approach, guided by well-established 5G use case requirements.
The fourth major contribution of this thesis is a mapping of the proposed scheme in a Software Defined Networking-enabled 5G architecture, as proposed by the latest 3GPP standardization, capable of applying Network Slicing approaches for further optimizing the network resources distribution and sharing and addressing the challenging 5G use cases, such as massive IoT.

2 Results and Discussion

2.1 COmpAsS experimental evaluation

The performance of COmpAsS is demonstrated via a series of simulation scenarios. In this section, 3 rounds of experiments will be presented, along with the respective assumptions, topologies and outcomes. All simulations were carried out using the open-source NS-3 discrete-event network simulator (versions NS-3.19 and NS-3.23).

Experiment 1: Simplified Shopping Mall use case.

The first scenario presents a simplified version of a shopping mall test case; two rows of femto cells (assuming they are inside the mall's shops) and a pedestrian corridor in the middle, while macro cells (LTE eNBs) co-exist at a distance of around 1km which is a typical range for an urban – suburban location (**Fig. 1**). In our simulation scenarios, we included 2 eNBs and 5 HeNBs. For simplicity's sake, no WiFi APs were used in the scenario, as the large-small cell handling evaluation of our mechanism was achieved by using LTE macro and femto cells. It is furthermore assumed that inside the mall area, several UEs are either static or moving at pedestrian's speeds, i.e. 0 – 1.5 m/s. These UEs –being attached to the mall's HeNBs- contribute as well in the creation of the load that needs to be taken into account for selecting the appropriate RAT from UE.

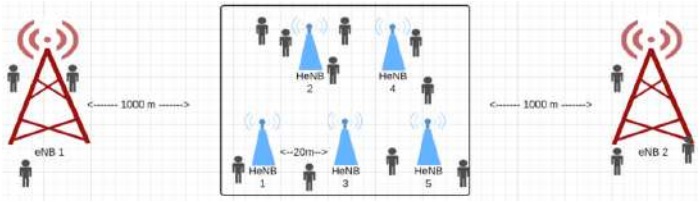


Fig. 1 Simplified Shopping mall use case

The simulations were made using three moving UEs at pedestrian speed: low, medium and high. In order to evaluate the performance of the algorithms, we increased the load of two out of the five overall HeNBs gradually, reaching from zero load to very high load. The background traffic in all rest (H)eNBs causes them to be in a medium load state during simulation time. By load, we are referring to both Load and Backhaul Load –as they were presented in the previous sub-sections-, which is calculated by the number of the rest static UEs, associated to each (H)eNB, as well as the number of bearers per UE. The KPIs, which are assessed, are the overall number of handovers that took place, the average throughput as well as the average experienced delay for the three moving UEs.

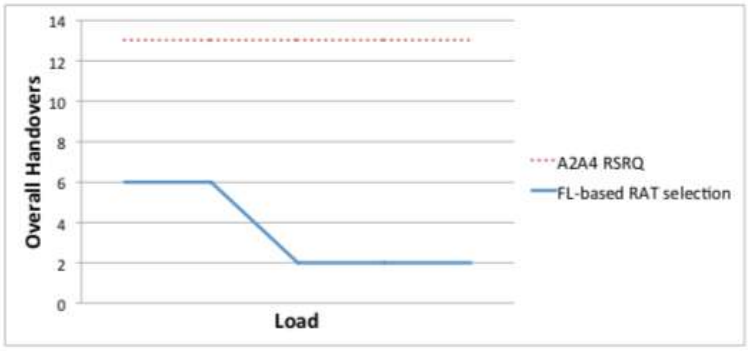


Fig. 2 Overall number of handovers

The above figure illustrates the performance of the two algorithms in terms of the overall number of handovers that took place from the 3 UEs. Noticeably, the A2A4 RSRQ algorithm's decisions are not influenced neither by the higher mobility of the UEs, nor by the increasing load of the HeNBs. On the contrary, the proposed mechanism tends to minimize handovers in the afore-mentioned cases. When the load is low, the number of handovers is reduced by 53.8% -due to reduced number of executed handovers of the high mobile UEs-, while when the load increases, the overall handovers are reduced by 84.6% since the suitability factor of candidate (H)eNB is low.

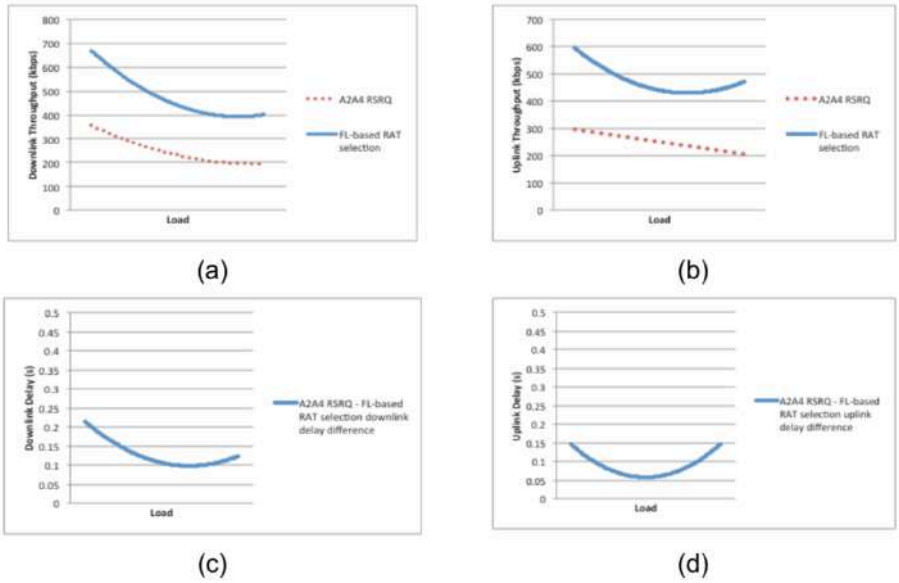


Fig. 3 (a) Downlink Throughput, (b) Uplink Throughput, (c) Downlink Delay, (d) Uplink Delay

Fig. 3 illustrates the performance of the algorithms with regard to the calculated average throughput both in the downlink and the uplink. The FL-based RAT selection outperforms the A2A4 RSRQ in terms of throughput; both in the downlink, as well as the uplink case by an offset of 300 kbps roughly as load increases. An interesting observation in the case of the proposed mechanism's performance is the increase of the throughput when the load of the HeNBs is extremely high; this advantage results from the fact that as load increases the suitability of the femto cells is constantly decreasing, tending to retain the UE from doing a handover to them. As a result, throughput and delay performance are directly related to the handover decisions presented earlier.

Similarly, the difference in the delay (measured as A2A4 RSRQ average packet delay minus the FL-based average packet delay) remains positive in all scenarios. Once more, as load increases radically, the FL-based mechanism tends to increase its performance since the UE keeps its connection to medium loaded eNB.

Experiment 2: Realistic Shopping Mall use case.

The 2nd experiment presents a realistic business case scenario of a shopping mall comprising 3 floors (ground floor, 1st and 2nd floor), and 20 shops per floor (Figure 31). The UEs are either static or moving, and are roaming around the shopping mall rooms (shops, cafes, etc.). Several HeNBs are deployed in the three floors. In addition, two macro cells (eNBs) exist outside the mall area in a distance of 200m to different directions. Due to the fact that COmpAsS handles Wi-Fi APs and HeNBs in a similar way, with regard to the pre-defined rules of the Fuzzy Inference Engine, for the sake of simplicity, in the simulations only macro and femto-cells are deployed.

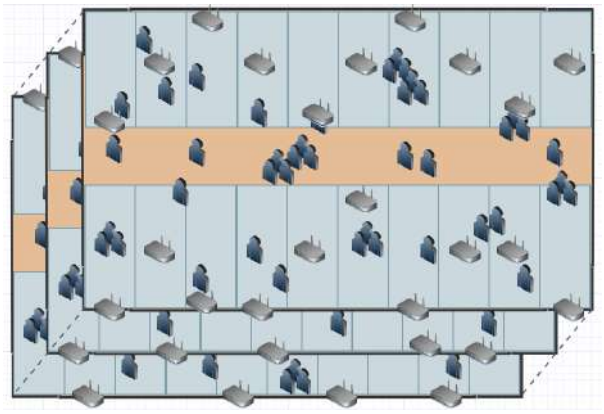
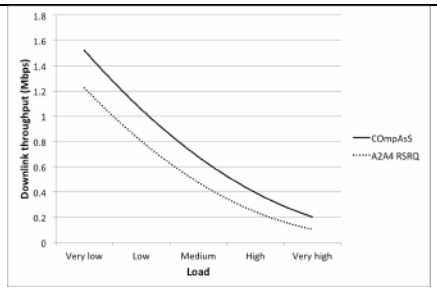


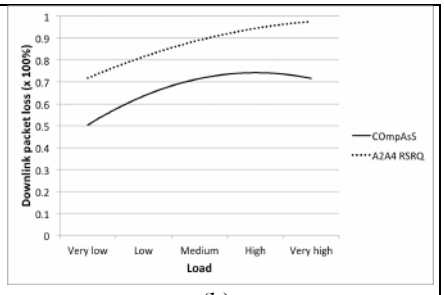
Fig. 4 Experiment 2: Shopping mall with 3 floors and 20 shops per floor

Besides the several UEs, which are roaming inside the mall area and creating respective traffic to the HeNBs, we use one "test UE", in which COmpAsS is deployed. Different simulations were carried out to test the UE at different velocities (low, medium, high), in each one of the scenarios in order to evaluate the proposed scheme for varying UE mobility, as mobility is one of the inputs, which are taken into consideration for the decision. The test UE is moving with linear velocity between the rows of the shops, on the 1st floor.

Indicatively, in the following figures we illustrate some of the measured KPIs, which resulted from the two mechanisms with regard to the number of overall handovers which took place during the simulation, the throughput of the test UE, the experienced delays, as well as the packet loss during the measurements. Variable load of the femtocells of the shopping mall was tested, calculated in relation to the overall associated users per base station and traffic that is generated. In particular, the load of the base stations varies from 10% up to 90% of their available resources (horizontal axis).



(a)



(b)

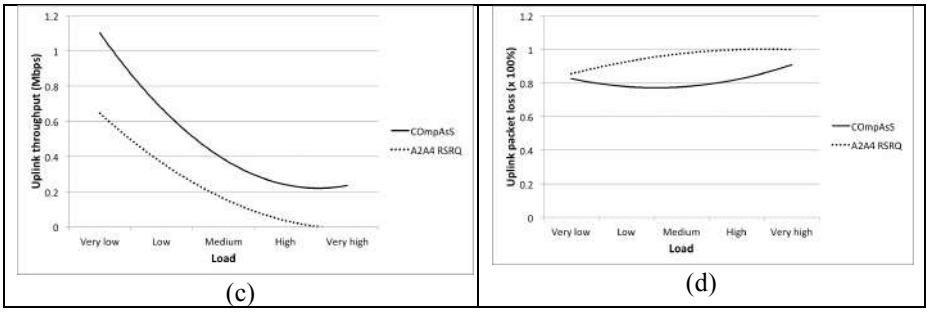


Fig. 5 Experiment 2 results

Experiment 3: Advanced Shopping Mall use case.

The final COmpAsS evaluation experiment is the most advanced one, among the three, which are included in this evaluation section. We evaluate this 3rd experiment on the basis of 4 different scenarios, which we describe in detail below. Overall, the network deployment allows seamless handling of services across different domains, e.g. mobile/fixed network operators, real estate/shop owners and application providers. The environment is similar to Experiment 2, presented earlier. Each floor's dimensions are 200x100m, containing 20 rooms/shops per floor, with several LTE Femto cell placed on each floors, depending on the scenario. Outside, two LTE eNBs are placed, 150m north and west of the mall respectively.

The proposed framework's algorithm uses two parameters, i.e. *Suitability Threshold and Hysteresis*. Different parameter values may alter radically COmpAsS's responsiveness and functionality, primarily in terms of triggering events frequency. Different network "states" (e.g., denser or scarcer deployments) would require different configurations of these two control parameters. Towards this fine-tuning process, hence, in the first two scenarios, we incorporate in our experimentation a range of values, both for Threshold and Hysteresis. Overall, the evaluation of COmpAsS moves along 4 axes-scenarios, each one of which focuses on a different varying parameter of the experiment's setup, in order to simulate -in the most realistic extent possible- all the radio conditions and network "states" that the proposed framework may encounter.

The following scenarios were evaluated:

- Per suitability threshold
- Per suitability margin
- Per deployment density (number of femto cells)
- Per network load (number of traffic bearers/UE)

For each one of the 4 scenarios, we ran 75 similar experiments in order to maximize the validity of our experimentation results. In order to define the number of runs per sub- scenario, as well as the experiment duration, we initially carried out some test scenarios; each one of the different runs incorporates a random generation of mobility patterns for the UEs, as well as slightly varying traffic models. We defined our confidence level at 95% in order to be able to demonstrate a satisfactory and statistically valid outcome. Indicatively, we provide the results for the 4th scenario (i.e. per network load):

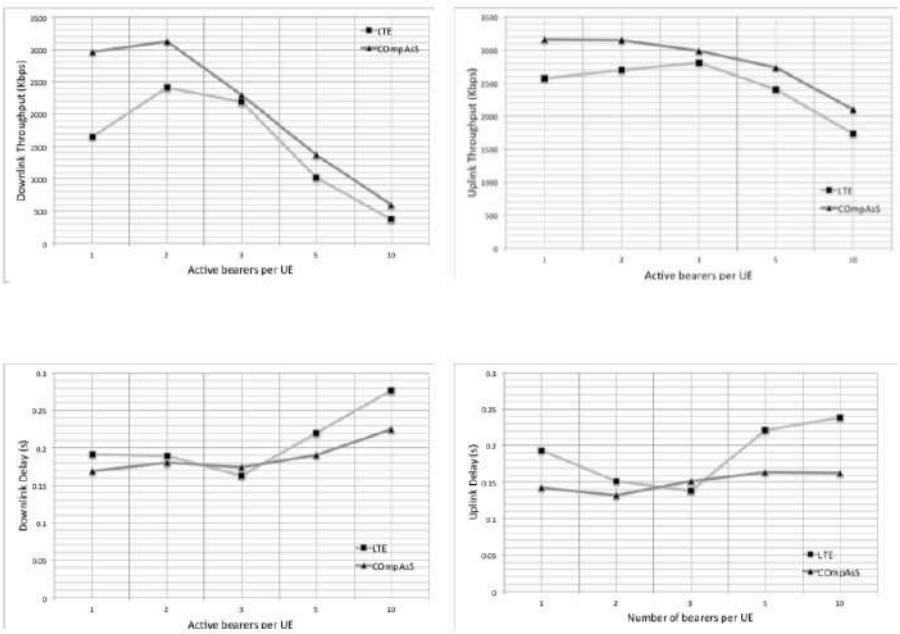


Fig. 6 Experiment 3 indicative results

2.2 CEPE experimental evaluation

The evaluation of CEPE followed the same simulation environment and principles with the earlier COmpAsS's evaluation steps. Indicatively, below the performance evaluation is presented in relation to the type of the Radio Access Technology, associated with the measured UE.

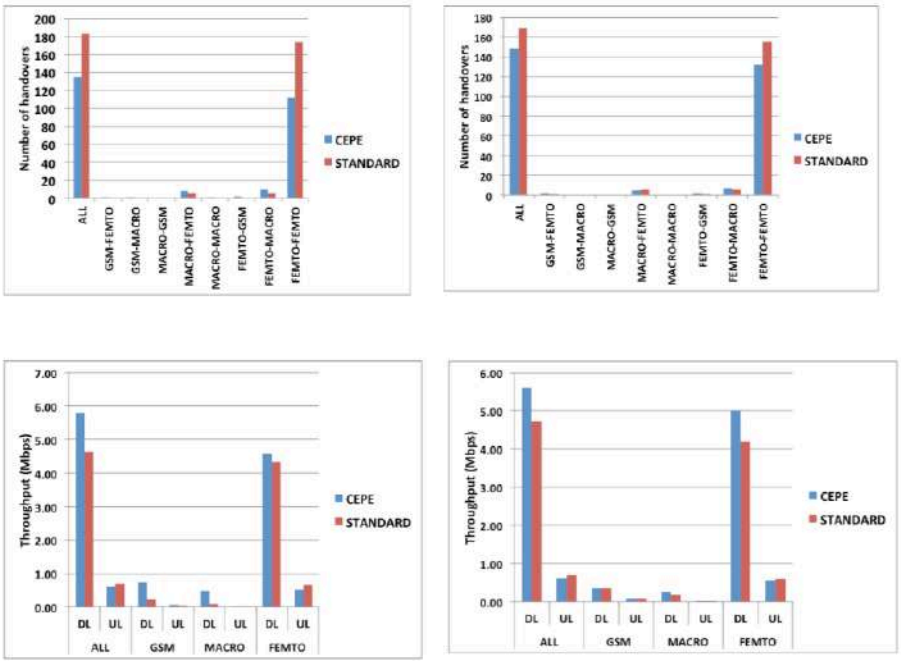


Fig. 7 CEPE evaluation results per type of RAT

3 Conclusions

Context awareness comes at a cost. The more information is acquired and processed, the higher the granularity of the context awareness, however the larger the burden, which is placed –both on the network, as well as the computing entities-; one of the crucial topics, which was analyzed in the context of this thesis was the signaling overhead evaluation for each one of the proposed mechanisms, as well as the type of the information acquired, which should be within the available information items and network entities, and in line with the latest standardization efforts towards 5G.

In the context of this thesis, the focus has been placed on three distinct context-aware mechanisms, which attempt to address the resource scarcity issues, which will be faced in the forthcoming ultra dense network deployments. All three mechanisms, focus on a different aspect of the context-aware approach: COmpAsS is a UE-based scheme, which attempts to select the most suitable radio access technology and point of access for the UE; CEPE, is an offline profiling engine, which is used to generate user and device profiles, and –based on these profiles- optimize the resource allocation- attempting to map in an optimal manner the mapping between the available resources and user/device profiles; CIP –the 3rd core mechanism- is a context information pre-processing and filtering scheme, which acts in a complementary manner to the aforementioned schemes, targeting to minimize the signaling and processing burden, which is posed by the diverse and numerous context information items –especially in dense 5G

deployments- with massive number of users, devices and coexisting access technologies.

This thesis provided a holistic study, which comprised comprehensive analyses from diverse aspects: architectural, algorithmic, experimental, analytical, etc. for all mechanisms. Besides the evaluation of the integrity and validity of each one of the three core schemes, a novel architecture is proposed, in line with the latest 3GPP standardization steps, comprising COmpAsS, CEPE and CIP as instances of the proposed novel network entities of the new 5G-EPC.

One of the crucial matters, on which focus was given, was the context acquisition process. In order to design COmpAsS's, as well as CEPE's system parameters, a comprehensive analysis on the network resources, respective interfaces and context information item types was made. In addition, an analytical approach was presented in the case of COmpAsS, which provided detailed insights on the information items, which are used, along with the signaling overhead required to aggregate them. To the best of our knowledge, there is no previous work, which attempts to quantify the signaling overhead of the proposed context-based mechanism, and juxtapose it with the gains measured in the network-related KPIs part.

The validity of each one of the mechanisms was showcased via an extensive set of experimental scenarios, carried out in line with the 5G Ultra Dense network scenarios and requirements, in realistic simulated topologies and with diverse access technologies and layers (macro cells, femto cells, Wi-Fi APs, etc.). The flexibility of the open source NS3 simulator, -which was used throughout all the experimentation-, enabled us to customize our environment according to very specific requirements, and –thus- achieve the realistic models we targeted.

This extensive demonstration proved a number of gains with regard to primary network KPIs, such as the maximization of the achieved throughput, the minimization of unnecessary handovers, as well the reduction of the latency measurements, particularly for delay-critical services, as described in the 5G verticals' requirements. The performance of the proposed schemes was juxtaposed to well established handover and RAT selection mechanisms, already deployed in 4G/LTE. This comparison highlighted numerous outcomes, both as far as the network and QoS KPIs are concerned (throughput, latency, packet loss, etc.), as well as the signaling overhead evaluation, since we compared with baseline –already deployed in the market- solutions, and not theoretical solutions found in the literature.

References

1. Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021, url: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile- white-paper-c11-520862.html, last accessed 2017/07/11.
2. ITU report, ICT Facts and Figures 2016, url: http://www.itu.int/en/ITU- D/Statistics/Documents/facts/ICTFactsFigures2016.pdf, last accessed 2017/07/11.

3. Henry Blodget, The Number Of Smartphones In Use Is About To Pass The Number Of PCs, https://www.businessinsider.com.au/number-of-smartphones-tablets-pcs-2013-12, last accessed 2017/07/11.
4. 5G Vision Brochure, 5G-PPP, The 5G Infrastructure Public Private Partnership: The next generation of communication networks and services.

Methodologies for Accelerated Analysis of the Reliability and the Energy Efficiency Levels of Modern Microprocessor Architectures

Emmanouil Kaliorakis[1]

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

Abstract. The evolution in computer architecture leads to increase in performance of modern microprocessors, which is also accompanied by decrease in products' reliability that is defined as their ability to avoid service failures that are more frequent and more severe than is acceptable. Thus, designers apply different techniques throughout microprocessors life-time in order to ensure the high reliability requirements of the delivered products.

This thesis proposes novel methods to guarantee the high reliability and energy efficiency requirements of modern microprocessors that can be applied during three phases of the processors' life-cycle: (a) MeRLiN methodology to accelerate (from 1 up to 3 orders of magnitude) the reliability assessments of hardware structures at the microarchitectural level against transient faults, (b) a methodology to accelerate the online permanent fault detection of many-core architectures providing up to 47.6X speedup using the Intel SCC chip as experimental vehicle, and (c) a comprehensive statistical analysis method based on linear regression with different feature selection methods to predict the safe voltage operation margins of the ARM-v8 cores of the enterprise X-Gene 2 micro-server.

Keywords: Reliability, transient faults, permanent faults, energy efficiency, statistical analysis

1 Introduction

The evolution in semiconductor technology and computer architecture give designers the opportunity to boost the performance of modern computing systems that are used in several domains of information and communication technology systems. Despite the changes in Moore's Law [1], computer architects and designers are still able to improve processor performance by using more aggressive and sophisticated techniques. However, the scaling in performance is also accompanied by increase in the vulnerability (or decrease in reliability) of microprocessors due to: (a) the strict deadlines that are required to minimize Time-to-Market (TTM) (minimizing also the time needed to test the circuits), (b) the modern device integration techniques that make processors more vulnerable to the radiation and also increase the occurrence of manufacturing defects,

[1] *Dissertation Advisor: Dimitris Gizopoulos, Professor.*

and (c) the increased design complexity that makes the testing process of the microprocessor products very difficult and unaffordable for the available TTM.

Specifically, the modern microprocessors face serious reliability issues during their entire life-cycle due to: (i) the errors that come from *transient faults* caused by cosmic rays, alpha particles and electromagnetic interference and are manifested as instantaneous flips of the values of real hardware bits, (ii) aging that leads to errors that appear at regular time intervals (*intermittent errors*) or exist indefinitely (*permanent errors*), and (iii) manufacturing defects that can either be manifested as permanent errors or lead to timing errors when the chips operate beyond their nominal voltage and frequency conditions. These manufacturing imperfections usually force computer architects to adopt pessimistic operation margins in terms of voltage in order to protect the chips, while sacrificing the energy efficiency of the delivered product.

2 Contributions during microprocessors life-cycle

Manufacturers use several validation techniques that are implemented throughout the processor life-cycle in order to protect the chips from the different types of malfunctions, which are very important to ensure the design requirements in terms of functionality, performance, power and reliability of the delivered products. The goal of this dissertation is to provide solutions to different validation challenges during the products' lifetime. The contributions of this thesis can be grouped in the two following categories (*Pre-Silicon* and *Post-Silicon Reliability Analysis* techniques) according to the time interval they can be used during the microprocessor life-cycle:

Pre-Silicon Reliability Analysis: A very important task during the early design phases is the reliability estimation of the hardware structures and the entire chips against transient faults. The reliability and performance requirements that are defined during the planning design phase can guide several design decisions in the next phases of the processor life-cycle, such as implementation of protection mechanisms or even determination of several microarchitectural features (size of hardware structures, policies, etc.) that can influence not only the vulnerability of a chip but also its performance. Statistical fault injection of transient faults (flips of real hardware bit values) on microarchitectural structures modeled in performance simulators is a state-of-the-art method to accurately measure the reliability, but suffers from low simulation throughput.

This thesis presents several contributions in the research field of *pre-silicon reliability analysis* phase of processors life-cycle. Firstly, in [2] we present a novel fully-automated versatile architecture-level fault injection framework (called MaFIN) that is built on top of a state-of-the-art x86-64 microprocessor simulator (called Marssx86), for thorough and fast characterization of a wide range of hardware components with respect to various fault models (transient, intermittent, permanent faults). Next, by using the same tool and focusing mainly on the transient faults we executed two reliability evaluation studies. In the first study, we evaluated the reliability and performance tradeoffs for major hardware components of an x86-64 microprocessor across several important parameters of their design (size, associativity, write policy, etc.) [3]. In the

second study [4], we used MaFIN in conjunction with a different tool (called GeFIN) that is also used for early reliability assessments at the microarchitecture level to evaluate in a differential way: (a) the reliability sensitivity of several microarchitecture structures for the same ISA (x86-64) implemented on two different simulators, and (b) the reliability of workloads and microarchitectures for two popular ISAs (ARM vs. x86-64). The conclusions of studies [3] and [4] can guide design decisions during the early design phase of the microprocessors concerning reliability and performance, while avoiding any costly redesign phase.

A major challenge of the early reliability assessments to soft errors at the microarchitecture level using statistical fault injection is that the campaigns that provide estimations of high statistical significance require excessively long experimental time. This thesis addresses this challenge by proposing two methodologies. Firstly, we propose to accelerate the individual fault injection runs by using several techniques that are implemented in the simulator and take place after the fault is actually injected in the hardware structure [5]. Secondly, to further accelerate the microarchitecture level fault injection campaigns we propose MeRLiN [6] that provides a final speedup of several orders of magnitude, while keeping the accuracy of the assessments unaffected even for large injection campaigns with very high statistical significance. The core of this methodology is the pruning of the initial fault list by grouping the faults in equivalent classes according to the instruction that finally accesses the faulty entry. Faults that belong to the same group are very likely to lead to the same fault effect; thus, fault injection is performed only in a few representatives from each group. MeRLiN methodology constitutes a major breakthrough in the field of accelerating the reliability estimations of hardware components at the microarchitecture level with negligible loss of accuracy.

Post-Silicon Reliability Analysis: Another important phase during the processor reliability life-cycle is the *Post-Silicon Reliability Analysis* that consists of the manufacturing testing and the in-field verification that take place during the fabrication process and after the release of the microprocessors to the market, respectively. Note that in contrast to *Pre-Silicon Reliability Analysis*, in this phase the validation targets implemented circuits and especially after their release to the market when the designers have no longer interaction with the design. The contributions of this thesis in this phase of the life-cycle cover two important research fields:

a) **Acceleration of permanent faults online detection in many-core architectures**: The extreme complexity of many-core processor architectures and the pressure for reduced time-to-market renders even the most comprehensive verification and testing process before and during mass production incomplete. A significant population of manufacturing faults escape in the field of operation and jeopardize correctness of the chip. Online functional testing is an attractive low-cost error detection solution, but it should be fast enough in order to not impact the system performance. This thesis faces this challenge by proposing an effective parallelization methodology [7] to accelerate online error detection for many-core architectures by exploiting the high-speed message passing on-chip network to accelerate the parallel execution of the test preparation phase of memory-intensive test programs.

To demonstrate the efficiency of the proposed methodology we used a 48-core real hardware chip, Intel's Single-chip Cloud Computer (SCC).

b) **Statistical analysis to predict the safe voltage margins in multicore CPUs for energy efficiency**: Reduction of the voltage operation margins of multicore chips is a major challenge for the designers to gain in terms of power. Unfortunately, this reduction leads to several reliability issues due to the manufacturing defects that make hardware cores of the same chip to present variations in their safe voltage and frequency operation limits. These variations that remain constant after the release of the chip to the market are classified as static variations. On top of that, transistor aging and dynamic variations in supply voltage and temperature, caused by different workload interactions can also affect the correct operation of a microprocessor. Thus, the designers choose to insert conservative guard-bands in the operating voltage (and frequency) to protect the chips from the effects of the static and the dynamic variations, despite the induced cost in terms of energy (and performance). The contribution of this thesis to this challenge is to propose a detailed statistical analysis methodology [8] [9] to accurately predict at the system level the safe voltage operation margins of the eight ARMv8 cores of the X-Gene 2 chip fabricated on 28nm technology. Our analysis uses as inputs the microprocessor's performance counters values of benchmarks that were collected in nominal voltage conditions execution and the results of the characterization phase when the chip operates in scaled voltage conditions.

In the next section, we present in more details the methodologies and the evaluation results of the major contributions of this dissertation.

3 Acceleration of reliability assessments against transient faults

The methodologies that were presented in this dissertation to accelerate the statistical fault injection campaigns that are used to assess the reliability of the hardware structures at the microarchitectural level can be summarized in two categories. In the first category, we accelerate the individual injection runs after the actual injection of the fault in the structure based on the faults lifetime (Section 3.1), while in the second category we further accelerate the fault injection campaigns of high statistical significance using MeRLiN methodology by pruning the faults of the initial fault list (Section 3.2).

3.1 Acceleration of injection campaigns based on the faults lifetime

In [5], we extended the baseline mode of an out-of-order cycle accurate full-system x86-64 fault injection framework (MaFIN) [2] with two extra modes of operation in order to speed up the statistical fault injection campaigns at the microarchitecture level. The common characteristic of the two proposed techniques of [5] is that they are implemented after the actual injection of the fault in the hardware structures during its lifetime. In the first mode, an injection experiment is forced to completion when the fault is overwritten before it is read and thus we classify it early and accurately as Masked. In the second mode, an injection experiment is forced to completion before the end of the application in two cases: (a) when the fault is overwritten before it is

read, or (b) when an x86 instruction reads the fault from the faulty entry and reaches the commit stage and before the actual termination of the benchmark. The second method provides a tradeoff between speedup and accuracy in order to deliver a fast but less accurate solution in the early reliability estimation problem.

For evaluation, we used MaFIN to carry out extensive fault injection campaigns of transient faults in six structures of the microprocessor that hold the majority of chip's area: L1 Data cache, L1 Instruction cache, L2 unified cache, Physical Integer Register File, LSQ (data field) and LSQ (address field). We used seven benchmarks from the MiBench suite [10], while we injected 2000 faults per campaign that corresponds to 2.88% error margin and 99% confidence level according to [11].

From the results of this study, we concluded that for the intra core structures (physical integer register file, address and data fields of LSQ), the best solution to speed up the statistical fault injection campaign is the second mode of operation with negligible loss of estimation accuracy, leading to a high speedup of 3.38X, 4.06X and 3.37X for the three structures respectively. Except for the second mode, the first mode could be also used for the same structures without any accuracy loss leading to a final speedup of 2.63X, 2.92X and 1.46X respectively.

On the other hand, the best choice for an architect to estimate the reliability of caches is the first mode operation. This conclusion comes from the fact that the inaccuracy of caches' reliability assessment using the second mode is not negligible (from 8.47 percentile units for L2 cache to 20.13 units for L1 Data cache) and the speedup is not as high as in the intra core structures (for instance only 1.06% increase of speedup for the L2 cache). Consequently, the first mode is the best choice for caches to speedup campaign (with 1.37X, 1.48X and 1.05X speedup for the L1 Data, L1 Instruction cache and L2 cache respectively) and to ensure the final estimation accuracy.

3.2 Acceleration of injection campaigns based on fault pruning (MeRLiN)

Exhaustive fault injection at the microarchitecture level using the entire statistically significant fault list (i.e. a list of all the flips for every bit of all hardware structures and for every program execution cycle) is infeasible. Thus, designers in the industry resort to statistical fault sampling to boost the throughput of massive fault injection campaigns, while maintaining a reasonably high accuracy of the final estimations. Despite the acceleration provided by the statistical fault sampling, the total simulation time that is needed to estimate the vulnerability of multiple hardware structures with different configuration parameters is still unaffordable and it leads to larger conservative design decisions. Consequently, the acceleration of the microarchitectural fault injection of array-based structures that occupy the majority of chip's area during the first steps of its design cycle is of major importance for the engineers that work in the industry.

MeRLiN methodology [6] accelerates the reliability estimation of hardware structures at the microarchitecture level from 1 up to 3 orders of magnitude without compromising the final accuracy of the assessment. MeRLiN consists of three main phases (see Fig. 1): *Preprocessing*, *Fault List Reduction* and *Fault Injection Campaign*. Next, we briefly describe these three phases:

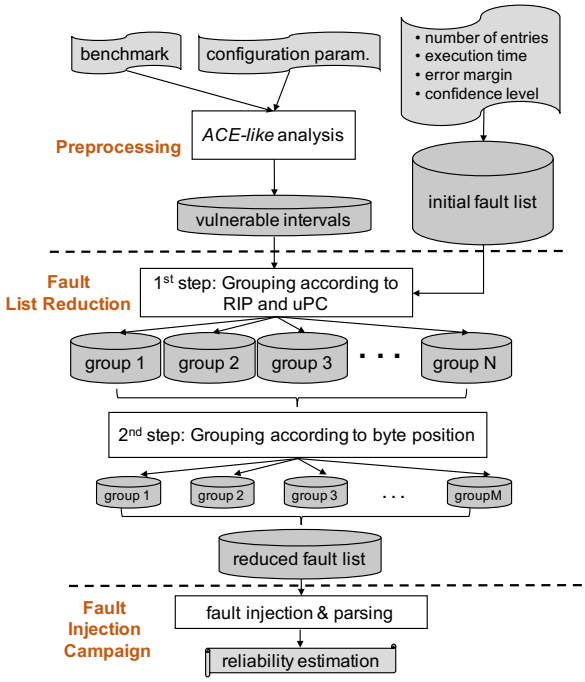


Fig. 1. Flowchart of MeRLiN.

Preprocessing: This phase takes place off-line and is responsible for two main tasks: the *ACE-like analysis* and the *Initial Fault List Creation*:

a) The *ACE-like analysis* is responsible to identify in a single pass all the *vulnerable intervals* of all hardware entries of the targeted structure (physical registers, store queue entries, cache words, etc.). For our analysis, a vulnerable interval of an entry starts with a write operation and ends with a committed read of the same entry or starts with a committed read and ends with another committed read of the same entry; all other intervals are non-vulnerable. This is different than classical ACE analysis [6] (see Fig. 2). All the faults that hit non-vulnerable intervals are excluded from the procedure of the actual injection and are classified as *Masked* as it is definite that they will not affect program execution. During the ACE-like analysis task, all the information concerning the vulnerable intervals is stored along with the information of the instruction pointer (*RIP*) and the microprogram counter (*uPC)* of the microarchitectural instruction that accesses the entry at the end of the vulnerable interval. This information is needed in the second phase of our algorithm (*Fault List Reduction*).

b) The *Initial Fault List Creation* is the second task of the first phase of MeRLiN in which the initial fault list of each injection campaign is created according to the typical statistical fault sampling formula of [11]. The initial fault list guarantees high statistical significance for all our results. The initial faults population is defined by: (1) the size (in bits) of the hardware structure, (2) the total execution (in cycles) of the benchmark that is already known from the *ACE-like*

task of the method, (3) the statistical confidence level and (4) the statistical error margin. The execution time required for the exhaustive fault injection campaigns using this initial population of faults at the microarchitecture level for all the combinations of benchmarks, hardware components and configuration parameters of each component is months or even years of execution, which is infeasible to take place especially in the early design phases of a chip. With MeRLiN we reduce this time by 1 to 3 orders of magnitude.

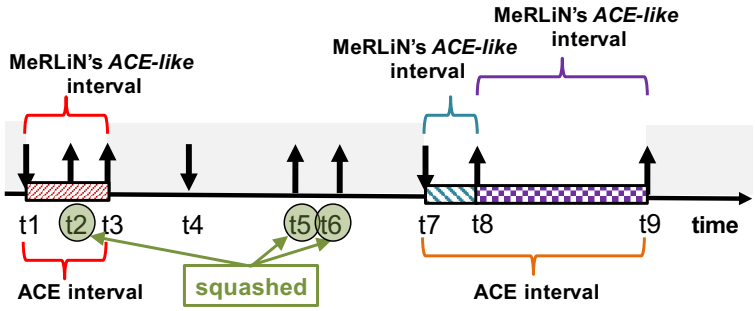


Fig. 2. ACE and *ACE-like* intervals definition example.

Fault List Reduction: This phase of MeRLiN consists of a two-step grouping algorithm:

a) In the *1st step of the group creation algorithm*, the remaining faults that hit ACE-like vulnerable intervals are stored in different subdirectories according to the *RIP* and the *uPC* of the instruction that reads the entry at the end of the interval (see Fig. 3). Each of the created groups consists of transient faults on the same or different entries of the hardware structure being analyzed, during the same or different ACE-like vulnerable intervals that are read by a micro-instruction with the same *RIP* and the same *uPC*. The classification of the faults in groups according to the *uPC* and the *RIP* of the microarchitectural instruction that accesses the faulty entry at the end of the vulnerable interval is necessary because different micro-instructions of the same instruction (x86 in our case study but generally applicable) can lead to different fault effects; thus, they should be classified separately in different groups (our experimental results validate this assumption).

b) In the *2nd step of the group creation algorithm*, to maximize MeRLiN's accuracy especially for groups with hundreds of faults, we select more than one fault for the actual fault injection runs in cases that faults hit a different byte of the entry. Moreover, faults in different bytes are selected from different dynamic instances of the same static instruction to increase time diversity (see Fig. 4). In this way, MeRLiN finally creates groups of equivalent faults at the byte granularity ensuring the accuracy of the final estimation. This can be further extended to separate faults hitting different nibbles or bits, but our experiments verify that this is not necessary and the byte granularity is sufficient.

Fault Injection Campaign: At the end, all the selected faults from all the created groups are stored in the *reduced fault list repository*. Only these group representative faults are actually injected using the microarchitecture level fault injector.

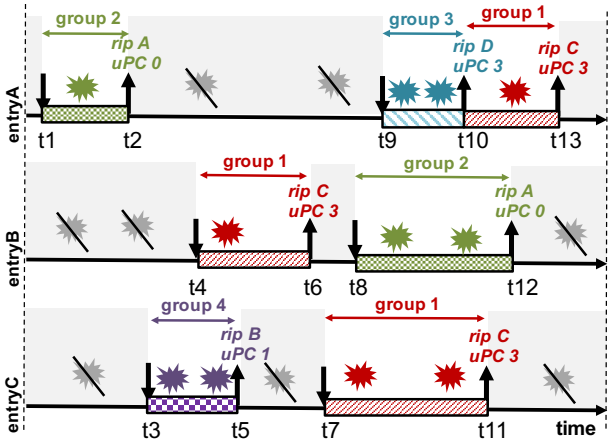


Fig. 3. 1st step example of the grouping algorithm.

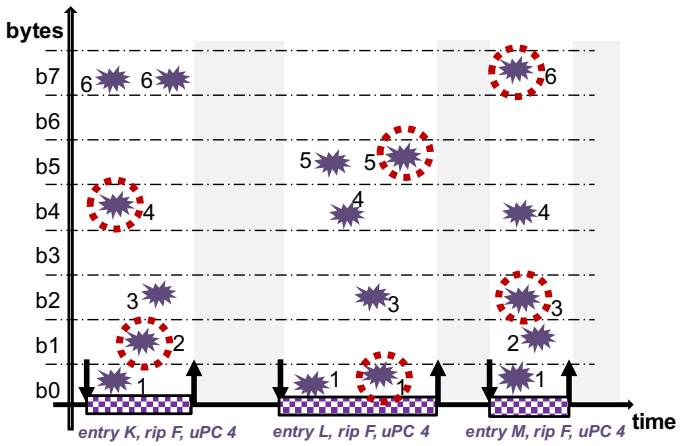


Fig. 4. 2nd step example of the grouping algorithm.

To evaluate the accuracy of the *Fault List Reduction* phase of MeRLiN we defined the *homogeneity metric* that expresses the effectiveness of our group creation algorithm to classify faults in groups that finally manifest the same fault effect. On average, the *homogeneity* of the created groups is very high (more than 91% for all our campaigns). In our study, we used the state-of-the-art microarchitectural injection tool (GeFIN) that is based on Gem5 simulator and extended it to implement and evaluate MeRLiN methodology on three data-related structures and one instruction-related structure of an x86-64 out-of-order processor model:

- The physical integer Register File for three sizes: 256, 128, 64 registers.
- The data field of the Store Queue of the Load/Store Queue for three sizes: 64 load and 64 store, 32 load and 32 store, and 16 load and 16 store entries. Gem5 does not implement data fields in the Load Queue.
- The data array of L1 Data cache for three sizes: 64KB, 32KB and 16KB.
- The destination register of the Issue Queue for two sizes: 32 and 60 queue entries.

For all our fault injection campaigns, we used 60,000 faults that correspond to 99.8% confidence level and 0.63% error margin. The key contributions of MeRLiN methodology are summarized below:

- It accelerates statistical microarchitecture level fault injection from 1 to 3 orders of magnitude. Our experiments with full runs of 10 MiBench benchmarks [10] show 93X, 225X, 68X and 28X speedup on average for different sizes of the register file, the store queue, the first level data cache and the issue queue, respectively. When applied to 10 SPEC CPU2006 benchmarks [12], MeRLiN reveals larger average speedups of 1644X, 2018X and 171X for the register file, the store queue and the first level data cache, respectively.
- It reports virtually the same reliability estimations as exhaustive (and infeasible) microarchitectural fault injection with extremely high statistical significance.
- It delivers fine-grained insights of the fault effects (Silent Data Corruptions – SDC, Detected Unrecoverable Errors – DUE, crashes, locks) unlike lifetime-analysis methods. This can be used to evaluate different protection schemes or to identify benchmarks more prone to SDCs or DUEs.

4 Acceleration of permanent faults online detection in many-core architectures

Functional testing techniques have gained increasing acceptance for microprocessor error detection during the last years. Functional online error detection approaches for many-core architectures are based on the application of the test programs during normal system operation and should adhere to the following requirements: (a) *reduced test program execution time*, (b) *small memory footprint*, (c) *test program replication* as all processor cores have to execute the same test program to detect faults and guarantee high fault coverage levels (this is different from traditional parallel programs).

In [7], we propose a functional online approach to accelerate the error detection of the Intel's Single-chip Cloud Computer (SCC) that contains 48 in-order Pentium cores. The SCC architecture consists of 24 tiles (two cores per tile) and 4 integrated DDR3 memory controllers supporting up to 64GB DRAM. Each processor SCC core has a 16KB L1 instruction cache, a 16KB L1 data cache and a 256KB L2 cache. Moreover, each tile has a 16KB message passing buffer (MPB) that bypasses L2 cache during communication.

For the experiments, we developed two test programs with different characteristics which represent typical test program formats used in functional online testing:

- *Load-Apply-Accumulate (LAA)* test program. It applies ATPG-generated test patterns stored in the off-chip DRAM. An LAA test program first reads two test vectors from the DRAM (assuming two-operand operations are being tested); it applies the target instruction (i.e. an arithmetic or logic instruction) and finally accumulates the results. We experiment with a loop-based LAA test program which applies a certain amount of test patterns (192KB or 384KB). LAA test program is memory-intensive and stresses the memory system of the SCC.
- *Linear-Feedback-Shift-Register* (LFSR) test program. This CPU-intensive test program applies pseudorandom patterns generated by a 32-bit LFSR. Similarly to LAA program, it first generates two pseudorandom test patterns, applies the target instruction and accumulates the results. LFSR test program generates either the same number of test data with LAA or 10 times more (e.g. for 384KB LAA, LFSR generates 3840KB test data).

The proposed method shown in [7] focuses on the efficient parallel execution of the test preparation phase. The test patterns are divided into 48 segments each one assigned to the private memory region of a core. The LAA test program is divided into two phases. First, all cores load in parallel the test patterns from their private memories, apply the tests and accumulate the responses. Subsequently, each core copies the corresponding test patterns from the local MPBs of the other 47 cores and applies/accumulates the tests. It is essential that in each cycle of the second phase each MPB serves the memory requests of only one core in order to limit the traffic congestion in the mesh and the routers. The rationale of the proposed method is that having every core to read test patterns from its private memory and distribute them to the other cores is the most efficient way to parallelize the test preparation phase of the LAA program. Our experimental results revealed up to 5.9X and 36.0X speedup when we applied our proposed method to 12 and 48 cores, respectively.

Regarding the LFSR test program, our experiments revealed that its test preparation phase cannot be parallelized in a more efficient way since the time each core requires to run the LFSR code to generate a certain number of test patterns is shorter than the time to copy these test patterns from the local MPB of an adjacent core. Thus, a second improvement in the parallelization of the entire online test program could be the parallel execution of memory-intensive test programs (e.g. LAA test) and CPU-intensive test programs (i.e. LFSR test) in the two cores of the same tile. This improvement increased the final speedup to 10.8X and 47.6X for the 12 and the 48 cores, respectively.

5 Statistical analysis to predict the safe voltage margins in multicore CPUs

Both static and dynamic variations lead microprocessor architects and designers to apply conservative guardbands (operating voltage and frequency settings) to avoid timing failures and guarantee correct operation, even in the worst-case conditions excited by unknown workloads. Predicting safe voltage operation regions of the microprocessor during manufacturing or after microprocessors' release to the market using as input the

performance counters provided by the system has recently gained the interest of the computer architecture community.

In [8] and [9], we implemented linear regression models with three different feature selection algorithms aiming to predict both the V_{min} and the Severity (a metric that indicates a region of chip's operation below the safe V_{min} and before the occurrence of any catastrophic for the system error) in the eight ARMv8-based cores of the X-Gene 2 chip. The inputs for our models came from the characterization phase of the chip in scaled voltage conditions and from the performance counters that were collected for each workload during its entire execution in nominal voltage conditions.

In our experiments, we ran all the benchmarks from the SPEC CPU2006 suite with all their inputs (40 programs in total). The evaluation of our models' accuracy targeting either the V_{min} or the Severity, was done by: (a) using the coefficient of determination (R^2) that assesses how well a model explains and predicts the future outcomes; also, it is indicative of the level of explained variability in the dataset. The larger the values of R^2, the better fit the model provides, while the best fit exists when R^2 is equal to 1, (b) using the Root Mean Square Error (*RMSE*) that represents the deviation between the predicted and the observed values (the smaller the *RMSE* the more accurate the model is), (c) comparing our models with the baseline (naïve) model, which is the average of the target values (V_{min} or Severity) of the training dataset.

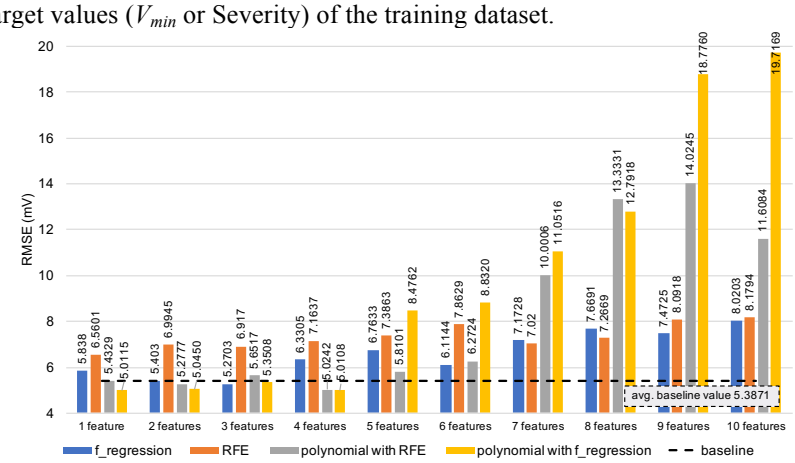


Fig. 5. Accuracy of predicting the V_{min} of the most sensitive core.

In general, our proposed method can lead to power gains from 11.87% up to 20.28% depending on the aggressiveness (Severity prediction is more aggressive than predicting V_{min}) of the prediction scheme. In this section and due to space limitations, we present in Fig. 5 the results for only one representative case of our analysis on predicting the V_{min} of the most sensitive core of the chip (core with the lowest V_{min} on average for all the experiments). The best accuracy (only 5mV inaccuracy) for this case was observed after using the polynomial transformation with *f_regression* selection and only 4 selected polynomial features leading to 11.87% power savings compared to the case of using the very pessimistic nominal voltage limit. Moreover, the R^2 that was measured for this prediction model is high (close to 0.75) indicating a good fit of the model.

6 Conclusions

In this dissertation, we propose several techniques to accelerate the analysis of the reliability and the energy efficiency levels of modern microprocessors that can be employed throughout the different phases of their life-cycle. For the *pre-silicon reliability analysis* phase, we proposed different methods to accelerate the statistical fault injection campaigns at the microarchitectural level either based on the faults lifetime after their injection that leads to an acceleration of up to 4.06X or by using fault pruning of the initial fault list provided by MeRLiN methodology that accelerates the reliability assessments even further (from 1 up to 3 orders of magnitude). Moreover, for the *post-silicon reliability analysis* phase, we proposed two methodologies. In the first, we proposed a parallelization approach to accelerate the online detection of permanent faults in many-core architectures (with up to 47.6X speedup), while in the second we proposed a statistical analysis approach to accurately predict (with only 5mV inaccuracy) the safe voltage operation margins of the ARMv8 cores of the X-Gene 2 chip.

References

1. G.Moore, “Cramming more components into integrated circuits”, In Electronics, April 1965.
2. N.Foutris, M.Kaliorakis, S.Tselonis, D.Gizopoulos, “Versatile architecture-level fault injection framework for reliability evaluation: a first report”, IEEE International On-Line Testing Symposium, 2014.
3. S.Tselonis, M.Kaliorakis, N.Foutris, G.Papadimitriou, D.Gizopoulos, “Microprocessor reliability- performance tradeoffs assessment at the microarchitecture level”, IEEE VLSI Test Symposium, 2016.
4. M.Kaliorakis, S.Tselonis, A.Chatzidimitriou, N.Foutris, D.Gizopoulos, “Differential fault injection on microarchitectural simulators”, IEEE International Symposium on Workload Characterization, 2015.
5. M.Kaliorakis, S.Tselonis, A.Chatzidimitriou, D.Gizopoulos, “Accelerated microarchitectural fault injection-based reliability assessment”, IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems, 2015.
6. M.Kaliorakis, D.Gizopoulos, R.Canal, A.Gonzalez, “MeRLiN: Exploiting dynamic instruction behavior for fast and accurate microarchitecture level reliability assessment”, ACM/IEEE International Symposium on Computer Architecture, 2017.
7. M.Kaliorakis, M.Psarakis, N.Foutris, D.Gizopoulos, “Accelerated online error detection in many-core microprocessor architectures”, IEEE VLSI Test Symposium, 2014.
8. G.Papadimitriou, M.Kaliorakis, A.Chatzidimitriou, D.Gizopoulos, P.Lawthers, S.Das, “Harnessing voltage margins for energy efficiency in multicore CPUs”, IEEE/ACM International Symposium on Microarchitecture, 2017.
9. M.Kaliorakis, A.Chatzidimitriou, G.Papadimitriou, D.Gizopoulos, “Statistical analysis of multicore CPUs operation in scaled voltage conditions”, IEEE Computer Architecture Letters, Jan. 2018.
10. M.R.Guthaus et al., “MiBench: A free, commercially representative embedded benchmark suite”, International Workshop on Workload Characterization, 2001.
11. R.Leveugle, A.Calvez, P.Maistri, P.Vanhauwaert, “Statistical fault injection: Quantified error and confidence”, ACM/IEEE Design, Automation & Test in Europe Conference, 2009.
12. Standard Performance Evaluation Corporation, https://www.spec.org [accessed 13/11/2017]

Managing Uncertainty and Vagueness in Semantic Web

Loukia Karanikola *

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

Abstract. Semantic Web has been designed for processing tasks without human intervention. In this context, the term machine processable information has been introduced. In most Semantic Web tasks, we come across information incompleteness issues, aka uncertainty and vagueness. For this reason, a method that represents uncertainty and vagueness under a common framework has to be defined. Semantic Web technologies are defined through a Semantic Web Stack and are based on a clear formal foundation. Therefore, any representation scheme should be aligned with these technologies and be formally defined. As the concept of ontologies is significant in the Semantic Web for representing knowledge, any framework is desirable to be built upon it. In our work, we have defined an approach for representing uncertainty and vagueness under a common framework. Uncertainty is represented through Dempster-Shafer model, whereas vagueness has been represented through Fuzzy Logic and Fuzzy Sets. For this reason, we have defined our theoretical framework, aimed at a combination of the classical crisp DL $\mathcal{ALC}$ with a Dempster-Shafer module. As a next step, we added fuzziness to this model. Throughout our work, we have implemented metaontologies in order to represent uncertain and vague concepts and, next, we have tested our methodology in real-world applications.

Keywords: Uncertainty · Vagueness · Dempster-Shafer Model · Description Logics · Semantic Web

1 Introduction

The Internet has paved the way for the evolution of alternative methods of communication. E-commerce, e-banking and online stores are some of them. Traditionally, computers were designed for performing numerical calculations. In addition, the content of Web information has been designed for human consumption, i.e. it is *human oriented*. The evolution of search engines gave a boost at the popularity of WWW, but at the same time made it necessary for the existence of a Web (or Web information) suitable for machines (or agents).

Towards this concept, Semantic Web was the vision of Tim Berners-Lee who stated: "*Machines become capable of analyzing all the data on the Web - the*

* Dissertation Advisor: Isambo Karali, Assistant Professor

content, links, and transactions between people and computers. "A Semantic Web", which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines, leaving humans to provide the inspiration and intuition. The "intelligent agents" people have touted for ages will finally materialize" [3].

The Semantic Web will contribute in the evolution of many web applications [1], such as Knowledge Management, Business-to-Computer, Electronic Commerce and Wikis.

Reliability, ambiguity or incompleteness issues are usual problems considering Web information, resulting in deficient knowledge. Any method that represents machine-oriented information should provide a well-defined description of imprecise knowledge [23].

Imprecise knowledge is usually divided into:

- *Uncertainty*
- *Vagueness*

Uncertainty refers to situations of information incompleteness whereas vagueness describes imprecise information, i.e concepts with not well-defined meaning. Generally, uncertainty and vagueness are considered two different notions and as such different theories have been defined for representing them. Probability theory, Dempster-Shafer theory and Possibility theory are some frameworks designed for uncertainty representation [6]. On the other hand, Fuzzy Logic and Fuzzy Sets [34] is the theory that lies behind vagueness representation. In many cases, we come across situations where both uncertainty and vagueness coexist. Thus, we need a common framework in order to represent uncertainty and vagueness concepts. Both notions can be defined as *imperfect information*.

Regarding Semantic Web, *ontologies* is the core concept for knowledge representation. Ontologies are represented through the Web Ontology Language (OWL) with OWL2 being the current version [33]. Description Logics (DLs) [2] have been employed extensively in Semantic Web, as they are the logics behind the most widely used version of OWL, OWL-DL. The necessity to capture uncertain and vague knowledge in Semantic Web has been employed in extensions of DLs, resulting in Probabilistic [21], Possibilistic [26] and Fuzzy extensions [29, 32]. These extensions capture the problem of uncertainty and vagueness separately and not as a common framework.

2 Objectives

2.1 Main Idea

The main goal of this dissertation is to define a framework for representing imperfect information, by extending crisp knowledge representation methods. By "imperfect", we refer either to uncertain or vague concepts. The general idea is to define a knowledge representation scheme, that allows for statements

with uncertainty and vagueness degree conditions. This representation assigns a truth degree in the interval $[0, 1]$ rather than a true/false value. Our framework is aligned with semantic web knowledge representation frameworks and it is defined based on these theories. Thus, our approach can be defined as a "*semantic web knowledge representation approach for representing uncertain and vague concepts*".

2.2 Thesis steps - Achievements

More precisely, throughout our dissertation we have proceeded through the following steps. For each step the reached achievements are also presented:

- Propose a definition of an "imperfect" Description Logic along with an "imperfect" Ontology, that captures both uncertain and vague concepts. Towards this concept the following sub-goals have been achieved:
  - Define ontologies that capture uncertain and vague concepts: An uncertainty ontology and an entailment method which is based on Dempster-Shafer model are described and implemented.
  - Define an extension of a crisp DL with Belief - Plausibility Degrees: We propose a framework that employs Dempster-Shafer theory in a Description Logic Knowledge Base environment. More precisely, we have defined a Dempster-Shafer DL Knowledge Base, in order to represent uncertainty in a Description Logics framework. In addition, a combination method of independent Dempster-Shafer DL Knowledge Bases has been proposed, based on Dempster's rule of Combination.
  - Define an extension of a fuzzy DL with Belief Degrees: Vague information has been emerged as a main issue in Semantic Web community. Vagueness is traditionally represented by Fuzzy Set theory. Besides vagueness, Semantic Web queries often have to deal with information incompleteness, aka uncertainty. This kind of information can be represented through Dempster-Shafer theory, that also enables distributed information fusion. Imperfect information, i.e uncertainty and vagueness, should be represented and manipulated under a common framework. We propose such a framework by defining a fuzzy Description Logic extended with Dempster-Shafer theory. Furthermore, we regard our method as a DL extension and we implemented it by a meta-ontology that captures Dempster-Shafer Fuzzy statements.
- Testing and evaluating our framework in real-world case studies: In order to test our methodology in real-world environments, we have tested two application areas, recommender systems and matchmaking environments. We have collected a set of data, detect uncertain and vague pieces of evidence and proceeded by employing suitable applications for manipulating them.

Consequently, for defining a unified framework for representing uncertainty and vagueness, we decided to combine the following theories:

- Fuzzy Logic
- Dempster-Shafer Theory
- Description Logics

3 Background and Related Work

At first, web data were designed taking into account human readers, with HTML being the most used language. The problem is that HTML does not provide for *metadata*, i.e. data about data. Metadata capture the semantic regarding Semantic Web data. Towards this concept, XML language have been employed.

In general, information processing within the Semantic Web is done by *agents*. As it is referred in [1], a semantic web agent "*will receive some tasks and preferences from a person, seek information from web sources, communicate with other agents, compare information about user requirements and preferences, select certain choices, and give answers to the users*". It seems that the role of an agent actually demands a decision making mechanism, which in turn presupposes a method for handling uncertainty and vagueness tasks. They are generally characterized as pieces of software that operate autonomously and proactively. In Semantic Web, an agent usually employs the following technologies:

- Metadata
- Ontologies
- Logic

3.1 Semantic Web Layers

Generally, the Semantic Web is regarded as a set of layers that form a stack, with each layer being built on top of another. At the bottom of the stack resides *XML*, which is a language that allows for structured web data with a user-defined vocabulary. Next, there is *RDF* and *RDF Schema*. *RDF* is a data model that is employed for writing simple statements about Web objects (resources). In addition, *RDF Schema* provides for organizing Web objects into hierarchies. Though tools for writing ontologies are provided, there is a need for more advanced ontology languages. Thus, the next level is the *ontology languages*, that allow for representations of more complex relationships, through a variety of dialects. Then, there is the *Logic* layer that provides with the means for writing declarative knowledge. Next, the *Proof* layer comes that is the deductive process, along with the representation of proofs and proof validation. Finally, the *Trust* layer considers digital signatures and in general knowledge based on recommendations by trusted agents.

3.2 Ontologies

As previously mentioned, there is a need for web information to be represented in a way that is understandable by machines. To achieve this, the Semantic Web incorporates a lot of technologies, which are described in what we call a *semantic web stack*. In addition, in [9], the semantic web architecture is regarded as "two-towers" rather than a stack . *Ontologies* and *rules* are the most significant among these technologies. Generally, an ontology "*is an explicit and formal specification of a conceptualization*" [1]. That means that it is a *conceptualization of a domain*

and provides a shared understanding of the domain. This term originates from philosophy and is the literal translation of the Greek word Οντολογία. As it is referred in [10] definitions for objects as well as types of objects should be provided. We can consider that an ontology consists of:

1. Types of entities that describe a specific domain
2. Properties of those entities

3.3 Description Logics

Description Logics is a family of *knowledge representation languages* and provide a way to "represent knowledge in a structured and formally well-understood way" [2]. They belong to a more general category called *description languages*. These languages allow the description of worlds providing constructors for building them [25, 2]. Generally, DLs support expressions that are built from atomic *concepts* and atomic *roles*. Each DL offers a specific level of expressiveness. DLs are a fragment First Order Logic (FOL), achieving lower complexity in expense of limited expressivity.

3.4 Uncertainty and Vagueness

Imperfect information includes uncertainty and vagueness concepts, which are described as follows:

- *Uncertainty*: It refers to situations when information incompleteness exist in order to decide about the truthness of a fact.
- *Vagueness*: It describes imprecise concepts, or concepts lacking clarity of definition

A good example of uncertainty and vagueness is given in [22], where the word "*degree*" is used to describe both uncertainty and vagueness measurements, but with different meaning. For example,

1. "*To some degree birds fly*" (uncertainty)
2. "*To some degree Jim is blond and young*" (vagueness)
3. "*Tomorrow, it will be a nice day*" (uncertainty and vagueness)

3.5 Fuzzy Logic and Fuzzy Sets

Fuzzy logic [34] is the logic of imprecision and approximate reasoning [36]. It is the framework for describing *vagueness*, by assigning truth values to linguistic variables [35] and aims at representing the human way of thinking. The general idea is that Fuzzy Sets' elements can belong to some degree to the set. More precisely, vagueness actually considers statements that are true to a certain degree, taken in the truth space [0,1]. In other words, statements are *graded*. Vagueness is associated with a set of vague concepts, e.g *low cost*. What is more is that vagueness is the result of ambiguity that describes information. For example a $100 ticket can be considered expensive for some people and low cost for others. The intuition behind the degree of membership is that the higher it is the more related is the object to the vague concept.

3.6 Dempster-Shafer Model and Dempster's rule of Combination

In the Semantic Web environment, usually, uncertainty comes as a result of ignorance, which in turn, is due to *incomplete information*. In other words, we talk about epistemic uncertainty. In those cases, the classical notion of probability cannot be considered suitable for the following reasons [7]:

1. Probability is not as good at representing ignorance.
2. An agent cannot always define probabilities for all sets of possible worlds.
3. In some cases, the computational effort demanded for probability definition, might be prohibitive.

Dempster-Shafer theory [28, 20] is considered a mathematical theory of evidence, that quantifies uncertainty in cases of ignorance and comes as a generalization of the Bayesian theory of subjective probability judgement. This theory is also known as *Theory of Belief Functions* or *Evidence Theory*. Bayesian theory quantifies judgements by assigning probabilities to the set of possible answers. Dempster-Shafer theory allows for deriving degrees of belief for a specific question based on probabilities for another related question.

4 Our Approach

As we have stated in the introduction, the Semantic Web vision introduces the concept of machine-processable information. In cases of imperfect information, i.e uncertainty and vagueness, the classical concept of ontology should be extended for capturing imperfect knowledge. Towards this concept, we aim at representing imperfect knowledge in an ontological environment.

In [18], an ontology for manipulating uncertainty, based on Dempster-Shafer theory, is described. The basic concepts of Dempster-Shafer model are represented through a Semantic Web ontology. Following, a set of entailment methods is combined through a method based on Dempster's rule of Combination.

In [19], an approach for representing uncertainty and vagueness is outlined. This approach considers vague knowledge represented through a fuzzy DL. In addition, an ontology is employed for representing information in a rule/event form, in order to perform reasoning. Both uncertainty and vagueness are represented by an *imperfection factor*. Big data processing have been also taken into account in this work.

In [15], an approach suitable for imperfect knowledge in a matchmaking case study is outlined. Matchmaking problems [13] can be considered as a case study of Semantic Web applications. In general, a matchmaking application considers a set of criteria, set by two parts. Towards this, we propose a matchmaking method of web data based on fuzzy criteria. Our method employs Dempster-Shafer theory and Dempster's rule of Combination in order to derive a combined constraint degree that represents the degree of matchmaking between the two parts (the seeker and the offer).

Following, we proposed a framework that employs Dempster-Shafer theory in a Description Logic Knowledge Base environment [16]. We name our model a *Dempster-Shafer DL Knowledge Base*.

As we have stated in the introduction, while developing Semantic Web applications, we often come across information incompleteness issues. As an example, let us consider a data source that contains information about *hotels*. We assume each hotel h to be assigned an interval cost per night rather than a crisp value, e.g:

$$h : [50 - 150]$$

In this case, if we want to make a reservation, we do not know exactly what the cost is but we know a lower-upper bound of the cost value. Moreover, consider the following query:

I'm looking for a hotel with cost no greater than 100

In a crisp logic framework, where each hotel has a unique value cost, the query could be answered with a yes/no statement. In our case, where we have to deal with interval value form, a yes/no statement cannot fully answer this query. The introduction of a *degree* notion seems to be more suitable to describe this kind of information.

In a Description Logics environment, if we consider a concept *DesiredHotel*, defined as:

$$DesiredHotel \equiv \ Hotel \sqcap \exists cost. \leq_{100}$$

then, the answer to our query is to decide whether a hotel individual is a member of the Class *DesiredHotel*.

Information incompleteness can be classified as an uncertainty problem. Dempster-Shafer theory, along with Dempster's rule of Combination [27], is a framework for dealing with information incompleteness, allowing integration of information from different independent sources. In our dissertation, we proposed an adaptation of Dempster-Shafer theory in a logic context.

More precisely, we define an extension of crisp Knowledge Bases with Dempster-Shafer modules. Dempster-Shafer Theory is more well-suited in modelling beliefs regarding the truthness of an event. Our method is an extension of the crisp DL $\mathcal{ALC}$. In our framework, we consider crisp DL axioms annotated with Dempster-Shafer belief and plausibility degree conditions.

As it is referred in the introduction, there is a need for representing uncertainty and vagueness through a common framework, especially in web-application areas. As a final step, we extended the theory of a fuzzy DL with a Dempster-Shafer framework. This framework is presented in [17, 14]. Our framework, denoted as *Dempster-Shafer Fuzzy Description Logic*, constitutes a generalization scheme of a crisp DL with fuzzy conditions along with a Dempster-Shafer module.

Taking into account the fuzzy DL interpretations introduced in [29], our framework considers any such interpretation as a *possible world*. The set of possible worlds is regarded as a frame of discernment. Thus, a basic probability assignment function is assigned on subsets of this set. This measure constitutes the uncertainty framework of our method.

A classical DL, assumes a universe $\mathcal{X}$ and subsets $\mathcal{A} \subseteq \mathcal{X}$, that constitute a DL *Concept*. Any element $x \in \mathcal{X}$ belongs to $\mathcal{A}$ or not, which is interpreted as a true/false value. The fuzzy extension assumes truthness interval on $[0, 1]$, where $\mathcal{A}$ is a *Fuzzy subset* and it is associated with a membership function $\mu_{\mathcal{A}}(x) : \mathcal{X} \rightarrow [0, 1]$. Any DL axiom, either crisp or fuzzy, has a truth value in a fuzzy interpretation $\mathcal{I}$. Our innovation, *Dempster-Shafer Fuzzy Description Logic* assigns probability masses into sets of fuzzy interpretations.

Let $\mathcal{W}$ a set of fuzzy DL interpretations. Let's denote a basic probability assignment function, m_{DS} on $2^{\mathcal{W}}$ as $m_{DS} : 2^{\mathcal{W}} \rightarrow [0, 1]$. Then, the extension of our method employs sets of fuzzy DL interpretations $\mathcal{I} \in \mathcal{W}$ in order to define Belief Degrees of fuzzy subsets of an interpretation domain $\Delta^{\mathcal{I}}$ (or $\Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$). This means that we assume a Fuzzy Description Logic and define Belief Degrees Conditions for axioms of this logic. In our case, we have considered the DL $\mathcal{ALC}$ and based on a fuzzy extension of it, we define our Dempster-Shafer Fuzzy DL. Since we extend fuzzy $\mathcal{ALC}$ based on Zadeh fuzzy logic, we also employ this logic in our framework.

For both our frameworks, i.e. the *Dempster-Shafer DL Knowledge Base* and the *Dempster-Shafer Fuzzy Description Logic*, we have examined decidability and complexity issues. In our approach, we adapt and extend the decidability procedure described in [29, 30] defined over fuzzy $\mathcal{ALC}$, in order to account for Dempster-Shafer Degree Conditions. This approach was first introduced in [4] in a propositional logic framework.

5 Conclusions and Future work

In our thesis, we defined an approach for representing uncertainty and vagueness under a common framework in a Semantic Web environment. In order to represent uncertainty we employed Dempster-Shafer model. Vagueness has been represented through Fuzzy Logic and Fuzzy Sets. At first, we examined our problem though an ontological point of view. Thus, we implemented suitable semantic web ontologies for capturing imperfect concepts. Following, for establishing our theoretical framework, we combined the classical crisp DL $\mathcal{ALC}$ with a Dempster-Shafer module. Next, we have proceeded by adding fuzziness in this model. Throughout our work, we formally defined the syntax and the semantics and examined decidability and complexity issues.

The main advantage of our method resides on the fact that we do not tackle uncertainty and vagueness as independent notions. This representation is in accordance with real-world applications, since very often uncertainty and vagueness coexist. The Dempster-Shafer model has been proven to be an ideal framework

for representing estimations, since it models a world in a way similar to human thinking, in cases of reasoning.

In addition, our theoretical framework has been built upon $\mathcal{ALC}$, a well-established DL. Our syntax has been defined as an extension of $\mathcal{ALC}$ syntax. Vagueness is represented through Zadeh's Fuzzy Logic, by considering membership degree conditions on crisp $\mathcal{ALC}$ axioms. In addition, we employ Dempster-Shafer theory for representing the uncertainty part. In order to employ this theory, we have defined belief degree conditions. The notion of *possible world* has an important role in defining the semantics of our framework. More precisely, we have regarded the set of possible worlds, i.e. fuzzy DL interpretations, as a frame of discernment and defined mass functions on subsets of this set. As a final step, we have considered the combination of statements from different Knowledge Bases, by employing our Combined Dempster-Shafer entailment, an entailment method based on Dempster's Rule of Combination.

The Dempster-Shafer framework was proven to be an ideal one for representing ignorance. Although it has many advantages, the complexity of the rule of Combination along with conflicts' modelling remains an issue to be tackled for representing real world case studies. As a future work, we will consider complexity and decidability issues more thoroughly, mostly aiming at Dempster's rule evaluation performance. In [27], other formulas for combining evidence are outlined. These formulas provide for lower complexity. Thus, the adaptation of these formulas in a DL environment can serve as a way to gain better complexity.

We shall also consider Big Data environments in a more thorough framework. Although we examine some Big Data issues through our dissertation, we do not consider some well known algorithms such as the one defined in [5]. As a future work, we will focus on the application of our model in a Big Data environment.

Another area of future work resides in the expressiveness level. As our dissertation has been defined upon DL $\mathcal{ALC}$, we may consider the extension of other DLs. Apart from fuzzy $\mathcal{ALC}$, other fuzzy extensions are described in [29], [22], [24], [32], [30],[31]. Moreover, although $\mathcal{ALC}$ is the basic DL, in cases of Semantic Web, a set of other DLs is usually employed, namely, $\mathcal{SROIQ}(D)$ [8], $\mathcal{SHOIN}$ [11] and $\mathcal{SHIF}$ [12]. So it will be useful to extend our framework to these DLs. In addition, for representing vagueness, we employed Zadeh's Fuzzy Logic. In future, we will consider other Fuzzy Logics as well.

In cases of strongly conflicting evidence, Dempster's Rule produces counter-intuitive examples. Towards this, other rules have been proposed [27]:

- The Discount and Combine method
- Yager's modified Dempster's Rule
- Inagaki's modified Dempster's Rule
- Zhang's Center Combination Rule

As a future work, we may consider the combination of evidence based on some of these rules.

In the area of applicability, case studies other than recommender systems and matchmaking environments can be examined. Some of them are:

– Semantic annotation
– Information extraction
– Ontology alignment
– Representation of background knowledge

These fields are described in [22] as some of the most representative ones of Semantic Web applications.

References

1. Grigoris Antoniou and Frank van Harmelen. *A Semantic Web Primer, 2nd Edition*. The MIT Press, 2008.
2. Franz Baader, Ian Horrocks, and Ulrike Sattler. Description Logics as Ontology Languages for the Semantic Web. In *Festschrift in honor of Jörg Siekmann, Lecture Notes in Artificial Intelligence*, pages 228–248. Springer-Verlag, 2003.
3. Tim Berners-Lee and Mark Fischetti. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*. Harper San Francisco, 1st edition, 1999.
4. Jianhua Chen and Sukhamay Kundu. A Sound and Complete Fuzzy Logic System Using Zadeh's Implication Operator. In *Foundations of Intelligent Systems*, pages 233–242, Berlin, Heidelberg, 1996. Springer Berlin Heidelberg.
5. Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, 51(1):107–113, January 2008.
6. Didier Dubois and Henri Prade. Possibility Theory, Probability Theory and Multiple-Valued Logics: A Clarification. *Annals of Mathematics and Artificial Intelligence*, 32(1):35–66, Aug 2001.
7. Joseph Y. Halpern. *Reasoning about uncertainty*. MIT Press, Cambridge, Mass., London, 2003.
8. Ian Horrocks, Oliver Kutz, and Ulrike Sattler. The Even More Irresistible $\mathcal{SROIQ}$. In *Proceedings of the Tenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'06, pages 57–67. AAAI Press, 2006.
9. Ian Horrocks, Bijan Parsia, Peter Patel-Schneider, and James Hendler. Semantic Web Architecture: Stack or Two Towers? In *Principles and Practice of Semantic Web Reasoning*, pages 37–41. 2005.
10. Ian Horrocks and Peter F. Patel-Schneider. *KR and Reasoning on the Semantic Web: OWL*, pages 365–398. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.
11. Ian Horrocks, Peter F. Patel-Schneider, and Frank van Harmelen. From $\mathcal{SHIQ}$ and RDF to OWL: The Making of a Web Ontology Language. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1(1):7 – 26, 2003.
12. Horrocks Ian and Sattler Uwe. A Description Logic with Transitive and Converse Roles Role Hierarchies and Qualifying Number Restrictions. Technical report, 1999.
13. Manish Joshi, Virendrakumar Bhavsar, and Harold Boley. Knowledge Representation in Matchmaking Applications. In *Advance Knowledge Based Systems, Model, Applications and Research*, pages 29–49. 2010.
14. Loukia Karanikola and Isambo Karali. Dempster-Shafer logical model for fuzzy Description Logics. In *SSCI 2016, Athens, Greece, December 6-9, 2016*, pages 1–8, 2016.

15. Loukia Karanikola and Isambo Karali. A Fuzzy Logic Approach for Reasoning Under Uncertainty and Vagueness - A Matchmaking Case Study. In *2016 2nd International Conference on Information Management (ICIM)*, pages 52–56, May 2016.
16. Loukia Karanikola and Isambo Karali. Semantic Web and Ignorance: Dempster-Shafer Description Logics. In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017.*, pages 68–73, 2017.
17. Loukia Karanikola and Isambo Karali. Towards a Dempster-Shafer Fuzzy Description Logic - Handling Imprecision in the Semantic Web. *IEEE Transactions on Fuzzy Systems*, Early Access(99):1–1, 2018.
18. Loukia Karanikola, Isambo Karali, and Sally I. McClean. Uncertainty Reasoning for the Semantic Web based on Dempster-Shafer model. In *4th International Conference on Information, Intelligence, Systems and Applications, IISA 2013, Piraeus, Greece, July 10-12, 2013*, pages 1–4, 2013.
19. Loukia Karanikola, Isambo Karali, and Sally I. McClean. Uncertainty Reasoning for the "big data" Semantic Web. In *Proceedings of the 15th IEEE International Conference on Information Reuse and Integration, IRI 2014, Redwood City, CA, USA, August 13-15, 2014*, pages 147–154, 2014.
20. Liping Liu and Ronald R. Yager. *Classic Works of the Dempster-Shafer Theory of Belief Functions: An Introduction*, pages 1–34. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.
21. Thomas Lukasiewicz. Expressive Probabilistic Description Logics. *Artificial Intelligence*, 172(6):852 – 883, 2008.
22. Thomas Lukasiewicz and Umberto Straccia. Managing Uncertainty and Vagueness in Description Logics for the Semantic Web. *Web Semantics*, 6(4):291–308, November 2008.
23. Miklos Nagy, Maria Vargas-Vera, and Enrico Motta. DSSim: Managing Uncertainty on the Semantic Web. Ontology Matching Conference'07, pages 160–169, Aachen, Germany, Germany, 2007. CEUR-WS.org.
24. Fernado Bobillo Ortega. *Managing Vagueness in Ontologies*. PhD thesis, Universidad de Granada, 2008.
25. Jeff Z. Pan. *Description Logics: Reasoning Support for the Semantic Web*. PhD thesis, School of Computer Science, The University of Manchester, 2004.
26. Guilin Qi, Jeff Z. Pan, and Qiu Ji. A Possibilistic Extension of Description Logics. In *Proceedings of DL'07, 2007*.
27. Kari Sentz and Scott Ferson. Combination of Evidence in Dempster-Shafer Theory. Technical report, 2002.
28. Glenn Shafer. Perspectives on the Theory and Practice of Belief Functions. *International Journal of Approximate Reasoning*, 4(5):323 – 362, 1990.
29. Umberto Straccia. A Fuzzy Description Logic. In *Proccedings of 15th National Conference on Artificial Intelligence*, pages 594–599, Madison,USA, 1998. AAAI-98.
30. Umberto Straccia. Reasoning Within Fuzzy Description Logics. *Journal of Artificial Intellignece Research*, 14(1):137–166, April 2001.
31. Umberto Straccia. Transforming Fuzzy Description Logics into Classical Description Logics. In *Logics in Artificial Intelligence*, pages 385–399, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.
32. Umberto Straccia. A Fuzzy Description Logic for the Semantic Web. In *Fuzzy Logic and the Semantic Web, Capturing Intelligence, chapter 4*, pages 167–181. Elsevier, 2005.

33. W3C OWL Working Group. OWL 2 Web Ontology Language Document Overview, October 2009.
34. Lofti A. Zadeh. Fuzzy Sets. *Information and Control*, 8(3):338 – 353, 1965.
35. Lofti A. Zadeh. The Concept of a Linguistic Variable and Its Application to Approximate Reasoning—i. *Information Sciences*, 8(3):199 – 249, 1975.
36. Lotfi A. Zadeh. Is There a Need for Fuzzy Logic? *Information Sciences*, 178(13):2751–2779, July 2008.

Adaptive epidemic dissemination in wireless ad hoc networks

Theofanis Kontos*

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
fanis1@di.uoa.gr

Abstract. The focus of this research is adaptive epidemic dissemination (AED) of information in wireless *ad hoc* networks. The main targets are the tradeoff between demands for broad information dissemination and reduced energy cost, the putsuit of an optimal solution to the problem and also the achievement of optimal performance in the aforementioned areas with high quality information.
Feedback-based AED schemes can be engineered that exploit context awareness in the form of channel state information (CSI) [1]
Besides such reactive and feedback-based schemes, others of a more proactive flavor can also be devised that use utility functions in order to adapt transmission characteristics according to predicted benefit [2]. With this approach a broad variety of benefit metrics can be exploited.
Introducing optimal stopping (OS) in the adaptation of transmission characteristics allows for visibility longer into the future and an optimal solution approach. The use of the one-stage-look-ahead, (*1sla*) OS rule displays improvement over non-adaptive schemes in terms of energy cost save while broad network infection remains as effective [3].
Optimization is further investigated in the broadcast scheduling problem, which is addressed with a transition to a *near-periodic* framework. According to this, scheduling is performed using the same toolcase of cross-layer utility functions and an optimal stopping mechanism. This problem is modeled as a *classical secretary problem* where the use of OS offers optimality. Compared against non-OS [2] and non-adaptive schemes this scheme wins in the energy cost save area and allows for successful network infection. In this case, the system converges to a state where dissemination cost is dramatically reduced while a large proportion of the residual, steady-state energy cost is due to CSI acquisition [4].
Furthermore, the possibility to improve information quality is investigated with information freshness originally assumed as the measure of quality. Adaptation of transmission characteristics based on information freshness is suggested aiming to lower the average age of the infecting information. In this manner, the idea of adapting transmission characteristics based on the actual payload of the infecting information is introduced.

* Dissertation Advisor: Efstathios Hadjiefthymiades, Professor

Tuning broadcasts down to a *polite gossip* during the route discovery phase in routing protocols for ad hoc networks is an additional attractive area. This problem is addressed within the framework of the AODV protocol. The tradeoff between energy cost save and fast routing information discovery emerges here. In an attempt to expand the technique to the actual routed data transmission (i.e. the user data), we move from broadcasting to a unicast landscape. Here, the tuning down of transmissions may have significant impact on information propagation. It is shown that the problem is best addressed using metrics reflecting a conceived quality of service and by extension quality of information.
The transmission scheduling problem enhanced by the additional information quality requirement can be formally expressed as an optimal stopping problem with known finite horizon. The quality of the information may be defined based on its properties (such as information freshness) or parameters of the protocol concerned with the dissemination, such as TTL in case of AODV. In this manner, information quality emerges as a third component besides energy cost and network infection, which participates the addressed tradeoff. Depending on the defined information quality the pursuit of a compromise between these three requirements is possible.

Keywords: Ad hoc networks · Adaptive epidemic dissemination · Cross-layer design · Optimal stopping.

1 The Problem Landscape

Epidemic dissemination (ED) [5], [6] aims to deliver broad infection of a network with some specified information while avoiding unconditional transmissions [7]. It is known that the latter cause problems, such as excessive energy cost and communication channel congestion without a corresponding benefit [8].

The performance of ED can be further improved with the introduction of adaptability of the transmission characteristics of the participating nodes. Hence adaptive epidemic dissemination (AED) is brought about [9], [10], [11], [12].

Previous research has delivered several AED schemes [13], [14], most of which are based on feedback mechanisms which exploit context awareness. Both context awareness and transmission characteristics adaptation concern several different network layers at the same time, thus rendering AED a technique of intrinsic cross-layer nature [15]. Adaptation decisions dictated by an AED scheme are taken in a decentralized, distributed manner rather than by a central and hierarchically superior node with full network awareness.

This research attempts to approach an optimal tradefoff among competing requirements in an AED environment and also deliver a strict formalization of this problem. The use of optimal stopping (OS) mechanisms is of central importance in our effort to reach optimality. The application of the technique in transmission scheduling is investigated and envisaged as future research field especially with an emphasis in ad hoc network routing protocols [16].

Let us consider a wireless ad hoc network where it is desired that a single information message be disseminated which is originally resident only in a small percentage of nodes. The latter, according to the established AED terminology are considered *infected*. Obviously, it is desired that the rest of the nodes (the *susceptible* ones) also become infected.

To this end every infected node periodically performs epidemic transmission of the infecting data, i.e. it *attempts* to transmit it: It performs a random experiment with success probability β. If the outcome is successful, then the node will broadcast in an attempt to infect its susceptible neighbors. Therefore, it has finite probability β (forwarding probability) to transmit within a predefined interval.

If the outcome of the experiment is successful and it indeed transmits, it does so using an adaptive modulation and coding (AMC) scheme. The AMC mode is described by the modulation type (e.g. QAM, QPSK, etc) and error-control coding. An important feature of the latter is the coding rate. A specific AMC mode μ shall be used for each transmission.

An infected node may for various reasons lose the infecting information, and then fall back to the susceptible state and be once again ready for new infection. This random process is modelled by a *cure probability*.

In AED models the following are important parameters:

- n: Network node count
- f: The *fan-out*, that is the nodes count targeted by a transmission.
- t: The time interval for which an infected node transmits. This can vary from 0 (*infect-and-die*) to ∞ (*infect-for-ever*).
- b: The node buffer capacity. If the infecting information consists of more than one messages, then this capacity introduces limitations and a mechanism to manage it is required.
- *partial view*: The network overview of an individual node. Essentially it is the count of nodes known to it.

2 Steps Towards the Solution

2.1 Epidemic Dissemination with Cross-Layer Context Awareness

The starting point of the present study is the feedback-based approach which is already familiar from previous research. The context awareness for each node consists of the following information:

- The signal-to-noise ratio (SNR). This knowledge constitutes channel state information (CSI) awareness.
- The proportion of received messages destroyed due to signal degradation while in transit in the channel (channel noise, multi-pathing, etc)
- The number of duplicates received. An infecting message received is considered a duplicate when received by an already infected node.

Exploiting such information awareness, each infected node adapts its AMC mode μ and forwarding probability β according to equations of the form shown in (1).

$$\begin{aligned} \beta(t+1) &= f(\beta(t), \text{context information}) \\ \mu(t+1) &= f(\mu(t), \text{context information}) \end{aligned} \tag{1}$$

Essentially the adaptation of β influences the outcome of the random experiment that the infected node performs when it is about to transmit. Decrementing the β means that the probability of successful outcome (=decision to transmit) is degraded, hence it is less possible for the node to transmit. Adaptation of the β, hence, implies decrease or increase of the number of transmissions and therefore the energy cost. Adaptation of the μ means decrease or increase of the probability of successful reception but also of the associated energy cost due to the respectively varying coding rates. The suitable simultaneous modification of μ takes care that, although reduced in number, transmissions are more likely to result in infections.

The following silent assumptions have been so far made:

- The fan-out equals the number of immediate (single-hop) neighbors of each node. It is essentially defined by the wireless transmitter range.
- The infect-for-ever epidemic model is followed, that is the infected nodes transmit indefinitely unless cured.
- The buffer size is adequate for the storage of an infecting message. This suffices to render the node infected. Saving two copies of the same message is meaningless and hence there is no need of storage space management.
- The node is aware only of the nodes it has received infecting information from

Simulations show that the adoption of the proposed scheme delivers considerably slower energy cost accumulation and also comparable to earlier heuristic AED schemes.

An interesting feature is the avoidance of the energy-intensive dialogues among nodes that cater for context information acquisition. This is a fully passive AED scheme. The overall novelty lies in the departure from heuristic schemes, the combined adoption of cross-layer context awareness, AMC with convolutional coding and immediate adaptation of the forwarding probability.

2.2 Epidemic Dissemination with Prediction: Utility Function with Simple Comparison (*beauty contest*)

A radical departure from the feedback-based approach is brought about by a new scheme that is based on benefit prediction.

It is from now on assumed that *the state of a node* at (discrete) time instance t is exactly described by the pair (β, μ). Therefore, the random experiment with success probability β described earlier is termed *transmitting from state* (β, μ)

Also, from now on, the forwarding probability is discretized to allow for a simplified model and calculations.

Moreover, the wireless channel is described as a finite state Markov channel (FSMC). Each state of the introduced FSMC is mapped to an SNR value range. Hence, the SNR value range corresponding to an AMC mode is mapped to an FSMC state.

Each time instance t the infected node evaluates which adaptation of its state is the optimal. The adaptation is achieved through performing an *action* out of a set of allowed ones presented in Table 1. The choice between increase and decrease of β and μ is based on context awareness and is presented in Table 2.

Table 1. Candidate actions

Action	Adaptation of β	Adaptation of μ
α_1	adapt	keep
α_2	keep	adapt
α_3	adapt	adapt
α_4	keep	keep

Table 2. Exploiting context awareness to decide decrease or increase of β and μ

Context	Actions
SNR increase	decrement β, increment μ
SNR decrease	increment β, decrement μ

The evaluation as to which action is the optimal is done based on the highest predicted immediate utility. This is calculated with the help of a suitable utility function.

This process of transitions can also be modelled with the use of a Markov chain. Each state of the Markov chain is mapped to a state (β, μ) of the node. Figure 1 depicts this approach.

The calculations also show that energy cost save and strong network infection are simultaneously achievable (figure 2).

Figure 2 displays the fact that infection is faster (shorter time-to-full-coverage *T2FC* and shorter time-to-90%-coverage *T29C*) and less energy intensive (lower energy-to-full-coverage *E2FC* and energyto-90%-coverage *E29C*) for adaptive schemes with various utility functions compared to the static approach.

A radical novelty of this scheme is the fact that it is proactive, based on prediction rather than pure feedback-based reactive. The choice of the utility function should depend on the needs of the individual problem and this flexibility is an additional virtue of the technique.

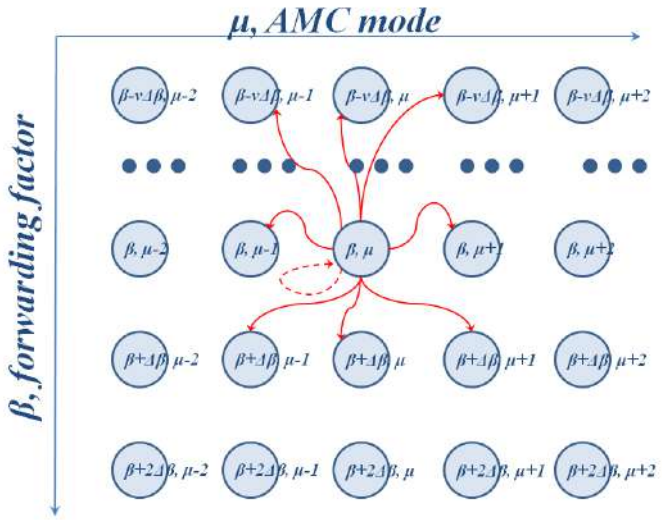


Fig. 1. Markov finite state machine describing the state transitions of a node.

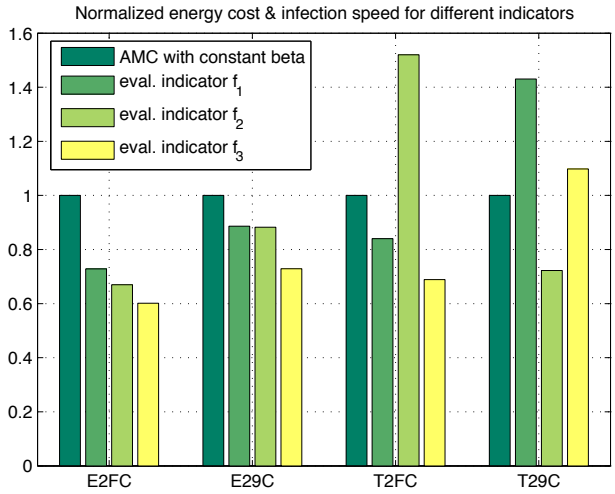


Fig. 2. Performance of a simple benefit prediction-based AED scheme.

2.3 Epidemic Dissemination with Optimization

The aforementioned technique may be further improved through an additional modification: *The most beneficial of the actions in Table 1 is actually adopted if and only if an optimal stopping (OS) condition is also simultaneously satisfied.*

Thus, OS is introduced as an additional term and the examined problem is addressed as an OS problem with infinite time horizon. The *one-stage-look-ahead* (1sla) rule has been chosen as the OS condition here to serve in a proof-of-concept setting. This is, however, optimal in problems with infinite temporal horizon only if the problem is monotonous with the specific utility function. With some assumptions it may be shown that the monotonicity condition is satisfied with a finite probability.

This "conservative" approach in tuning the transmission characteristics also delivers energy cost reduction with hardly any compromise in the epidemic infection.

2.4 Optimized Scheduling with Plesioperiodic Broadcasts

The acquired knowledge may be exploited for addressing transmission scheduling problems.

Many processes in modern wireless networks utilize temporal broadcasts for information dissemination. Examples are schemes like *directed diffusion* and some routing protocols such as AODV, DSR, LOADng, etc

Let us consider such a system in which the infected nodes periodically broadcast the information they possess -which is effectively the *infecting information.* Let the transmission period be ϵ. The following model is proposed instead of periodic broadcasts:

The temporal field is distributed in consecutive, non-overlapping time intervals of duration ϵ, which are termed *epochs.* Within each epoch the infected node solves the *classical secretary problem* with finite temporal horizon.

It starts from state (β, μ). At every time instance it evaluates the promised immediate benefit of the most beneficial of the actions of Table 1. When the OS condition is satisfied, as known from the secretary problem, this action is adopted and the node state is modified. Epidemic transmission takes place from this state and the node remains silent for the remainder of the current epoch.

Hence, a scheme of repeated epidemic broadcasts is adopted, which are performed within defined time limits but without the interval between two consecutive ones being constant.

In this manner, addressing the information dissemination problem transits from periodic to near-periodic or *plesioperiodic* as depicted in Figure 3.

The implementation of such a scheme assumes knowledge of the count of immediate neighboring nodes and the channel state information (CSI). This kind of context awareness is a quite common assumption in AED. Its advantages include the flexibility in choosing a utility function and also the energy cost reduction down to levels comparable with fully passive methods.

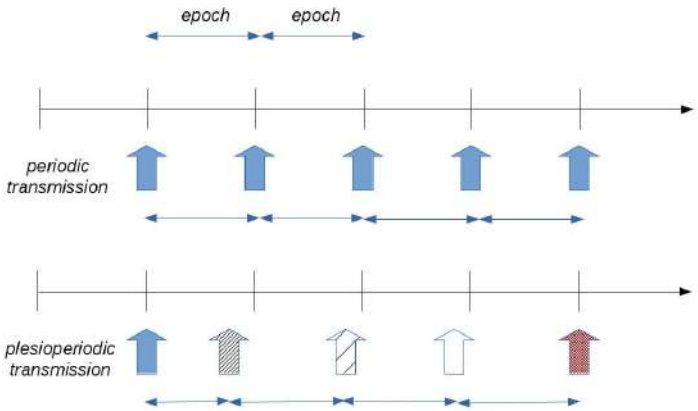


Fig. 3. Transition from strict periodicity to near-periodicity or *plesioperiodicity*.

The fields of application of such an improved scheduling method would include *route request* (RREQ) message dissemination in routing protocols. This constitutes an attempt to improve the performance of the dissemination of such messages using the described plesioperiodic approach.

Energy cost reduction compared to the unconditional message dissemination is expected. However, it is useful that other factors are taken into account. In the routing problem, a most interesting problem is that of encountering a node with the desired routing information. The latter is derived from the extent that the "epidemic", i.e. the information dissemination, finally assumes. Hence, the need for a tradefoff arises between energy cost save and routing information discovery.

Scheduling the infecting information transmissions is formulated as follows:

- At every time instance $t \in \mathbb{N}$ the state of an infected node is described by the pair $h = (\beta, \mu) \in B \times M$, where B and M the sets of possible values of β and μ respectively.
- The state may change through the adoption of an action α from a finite set of possible actions $\mathcal{A}$, such that $(\beta, \mu) \xrightarrow{\alpha \in \mathcal{A}} (\beta', \mu')$. It always holds that $(\beta, \mu) \in B \times M$ and $(\beta', \mu') \in B \times M$. That is the set $B \times M$ is closed under every action $\alpha \in \mathcal{A}$.
- Considering the time field divided up into *epochs*, the action α is adopted at time t_{ost}, that is one time *at maximum* within the current epoch. For the n-th ($n \in \mathbb{N}$) epoch of duration ϵ, same as the aforementioned period, this can be written as $t_{ost} \in E_n = \{n\epsilon, n\epsilon + 1, ..., (n+1)\epsilon - 1\}$, as we are dealing with a discretized time field. Therefore, in each epoch, that is $\forall n \in \mathbb{N}$, the optimal instance for state transition and broadcast from that new state is sought.
- Possible actions are evaluated using a utility function U. This is calculated $\forall t \in \mathbb{N}$, hence also for $\forall t \in E_n$.

For each epoch, i.e. $\forall n \in \mathbb{N}$, the pair $(\alpha, t_{ost}) \in \mathcal{A} \times E_n$ is termed the *optimal policy* for this epoch. That is:

The optimal policy is the decision as to which is the optimal moment within the epoch for the optimal action to be performed, which changes the node state.

As we saw, the node broadcasts from this new state. Therefore the problem is written formally as:

$$\boxed{\forall n \in \mathbb{N}, \text{ policy sought } (\alpha^*, t^*_{ost}) \in \mathcal{A} \times E_n : (\alpha^*, t^*_{ost}) = argmax_{(\alpha, t_{ost})} U_n(\alpha, t_{ost})}$$

The performance of the proposed technique is evaluated in an environment with mobile nodes where collisions in wireless channels and the CSI-acquisition energy cost are also considered. In figure 4 it is displayed in comparison to other adaptive and non-adaptive ones. It emerges as efficient as a fully passive benchmark and it also becomes apparent that a significant part of its residual energy cost is attributable to the context acquisition.

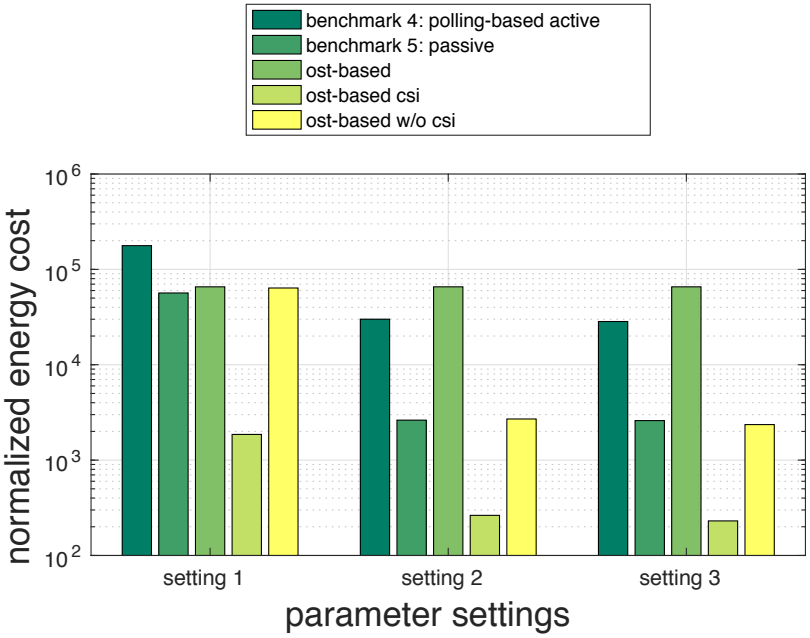


Fig. 4. Performance of an OS-based AED scheme. The discrete bundles of bars correspond to various problem parameter settings.

2.5 Applications in routing

Following the newly introduced formulation, the scheduling problem can be addressed under various conditions and requirements. In any case, of course, a suitable utility function is required for the evaluation of the benefit evaluation.

Context awareness can be broadened to include network density ρ changes for the decisions to increase or decrease β and μ. Table 1 then transits to Table 3. This, of course, is senseful mostly in a setting with mobile nodes.

Table 3. Context awareness in deciding to increment or decrement β and μ.

context information	SNR increase	SNR decrease
increment ρ	decrement β, increment μ	increment β, decrement μ
decrement ρ	decrement β, increment μ	increment β, decrement μ

Such a scheme can be utilized in the route request message (RREQ) broadcast scheduling problem.

An interesting aspect is the comparison of the behavior between random and scale-free networks. It is observable that in scale-free networks the same encouraging performance is observable. Naturally, the study of this type of networks is a whole new area by itself.

Context awareness can be further broadened through the inclusion of elements from the carried information itself (payload). Then the adaptation of the infected nodes state that was described earlier obtains a more information-centric character. For example, an application in the AODV routing protocol is amplifying the forwarding probability of a message when that is characterized by a still high TTL (time-to-live) value. In this manner, the disseminated information retains a young age.

The need for a tradeoff among three requirements arises:

- Energy cost reduction
- Broad information dissemination
- Retention or improvement of information quality

The techniques described have cross-layering in common. This property appears in our research in two aspects as both the context information and the adapted parameters are from various layers:

- The context information includes some of the following parameters: count of immediate neighbors, CSI between each node and its neighbors, proportion of malformed messages, number of duplicates and TTL. These are mostly associated with the physical layer and possibly the higher (application) layer.
- Adapted parameters include the forwarding probability and the AMC mode, hence the lower and network layers are involved.

3 Conclusions

Optimizing AED in demanding ad hoc network environments is an exciting problem. Exploiting the cross-layer nature of AED allows for pursuing a beneficial tradeoff among a number of competing requirements which may depend on the specific problem.

In this thesis we present a framework for flexible AED schemes that contributes in this direction as follows:

- Reaching optimal tradefoff among competing requirements
- Allowing for the inclusion of requirements associated with various layers depending on the specific problem and setting
- Addressing the transmission scheduling problem with the transition from periodicity to *plesioperiodicity*
- Showing that ad hoc network routing is a primary field of application for such schemes

In the schemes presented in this thesis broad infection of the network with information is achieved while the energy cost is significantly reduced. Retention or improvement of information quality is also investigated. Routing in ad hoc networks is shown to be a primary application field.

This approach achieves broad spread of high quality information while prolonging the network lifetime.

References

1. Theofanis Kontos, Evripidis Zaimidis, Christos Anagnostopoulos, Stathes Hadjiefthymiades, Evangelos Zervas, "An adaptive epidemic information dissemination scheme with cross-layer enhancements," in *Proc. IEEE Symposium on Computers and Communications (ISCC 2011:230-235)*, (Corfu, Greece), Jul. 2011.
2. Theofanis Kontos, Christos Anagnostopoulos, Stathes Hadjiefthymiades, "Wireless Channel State-Aware and Adaptive Epidemic Dissemination in Ad Hoc Networks," IJWIN 21(1): 58-67 (2014)
3. Theofanis Kontos, Christos Anagnostopoulos, Stathes Hadjiefthymiades, Evangelos Zervas, "Epidemic information dissemination controlled by wireless channel awareness," ISCC 2015: 721-726
4. Theofanis Kontos, Christos Anagnostopoulos, Stathes Hadjiefthymiades, Evangelos Zervas, "Adaptive epidemic dissemination as a finite-horizon optimal stopping problem," Wireless Networks, (), 1-18
5. "A Contribution to the Mathematical Theory of Epidemics," Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences. 115 (772): 700
6. Dietz K. "Transmission and control of arbovirus diseases," Epidemiology (eds Ludwig D, Cooke KL, editors.), pp. 104121 Philadelphia, PA: Society for Industrial and Applied Mathematics, 1974
7. Alan J. Demers, Daniel H. Greene, Carl Hauser, Wes Irish, John Larson, Scott Shenker, Howard E. Sturgis, Daniel C. Swinehart, Douglas B. Terry, "Epidemic Algorithms for Replicated Database Maintenance,"

8. S. Ni, Y. Tseng, Y. Chen, and J. Sheu, "The Broadcast Storm Problem in Mobile Ad Hoc Networks," *Wireless Networks*, vol. 8, Numbers 2-3, pp. 153-167, Dec. 1997.
9. Patrick T. Eugster, Rachid Guerraoui, Anne-Marie Kermarrec, Laurent Massoulie "Epidemic Information Dissemination in Distributed Systems," IEEE Computer 37(5): 60-67 (2004).
10. Wendi Rabiner Heinzelman, Joanna Kulik, Hari Balakrishnan "Adaptive Protocols for Information Dissemination in Wireless Sensor Networks," MobiCom 1999: 174-185.
11. Anagnostopoulos, Hadjiefthymiades, Zervas, "An analytical model for multi-epidemic information dissemination", *Journal of Parallel and Distributed Computing (JPDC), Elsevier*, vol. 71, issue 1, January 2011.
12. Christos Anagnostopoulos, Odysseas Sekkas, Stathes Hadjiefthymiades: "An adaptive epidemic information dissemination model for wireless sensor networks," Pervasive and Mobile Computing 8(5): 751-763 (2012)
13. Garbinato, Rochat, Tomassini, Vessaz, "Heterogeneous epidemic model for assessing data dissemination in opportunistic networks," *Future Generation Computer Systems*, vol. 26 issue 6, June 2010.
14. Lidia Yamamoto, "Epidemic Dissemination in Ad Hoc Networks", Extended Abstract, Dagstuhl Seminar 04411, Service Management and Self-Organization in IP-based Networks, Schloss Dagstuhl, Germany, October 2004.
15. Mihaela van der Schaar, Sai Shankar, "Cross-Layer Wireless Multimedia Transmission: Challenges, Principles, and New Paradigms," IEEE Wireless Commun. 12(4): 50-58 (2005).
16. Zygmunt J. Haas, Joseph Y. Halpern and Li Li, "Gossip-Based Ad Hoc Routing," IEEE INFOCOM, vol. 3, 2002, pp. 17071716

Distributed and Streaming Graph Processing Techniques

Panagiotis Liakos*

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

Abstract. Beneath most complex systems playing a vital role in our daily lives lie intricate networks. Such real-world networks are routinely represented using graphs. The volume of graph data produced in today's interlinked world allows for realizing numerous fascinating applications but also poses important challenges. Consider for example the friendship graph of a social networking site and the findings we can come up with when executing network algorithms, such as community detection, on this graph. However, the volume that real-world networks reach oftentimes makes even the execution of fundamental graph algorithms infeasible when following traditional techniques. In this short note we summarize our results on the study of two directions that allow for handling large scale networks, namely *distributed graph processing*, and *streaming graph algorithms*. In this context, we first provide contributions with regard to memory usage of distributed graph processing systems by extending the available structures of a contemporary such system with memory-optimized representations. Then, we focus on the task of community detection and propose i) a local algorithm that reveals the community structure of a vertex and easily facilitates distributed execution and ii) a streaming algorithm that greatly outperforms non-streaming state-of-the-art approaches with respect to both execution time and memory usage. In addition, we propose a streaming sampling technique that allows for capturing the interesting part of an unmanageable volume of data produced by social activity. Finally, we exploit the available data of a popular social networking site to empirically investigate a well-studied opinion formation model, using a distributed algorithm.

Keywords: Distributed graph processing · streaming graphs · graph compression · community detection · opinion formation.

1 Introduction

Real-life systems involving interacting objects are typically modeled as graphs and can often grow very large in size. A multitude of contemporary applications heavily involves such graph data and has driven to research directions that allow

* Dissertation Advisor: Alex Delis, Professor

for efficient handling of large scale networks. Two prominent such directions are distributed graph processing and streaming graph algorithms.

The tremendous growth of the Web graph has driven Google to introduce Pregel, a scalable platform with an API that allows for expressing arbitrary graph algorithms. Pregel is a distributed graph processing system that powers the computation of PageRank and has served as an inspiration to many systems that adopted its programming model. One such system is Apache Giraph which originated as the open-source counterpart of Pregel. Giraph is maintained by developers of Facebook that use it to analyze Facebook's social graph. Pregel-like systems follow a vertex-centric approach and address the task of in-memory batch processing of large scale graphs [26]. Communication details are abstracted away from the developers that implements algorithms for such systems. The latter offer APIs that allow for specifying computations with regard to what each vertex of the graph needs to compute whereas edges serve the purpose of transmitting results from one vertex to another. The input graph is loaded on start-up and the entire execution takes place in-memory. Consequently, the execution of a graph algorithm in a Pregel-like system depends on the available memory and will fail if the later is not sufficient enough to fit the graph.

The ever-increasing size of real-world networks has also motivated the design of algorithms that process massive graphs in the data stream model [42]. More specifically, the input of algorithms in this model is defined by a stream of data which usually comprises the edges of the graph. Therefore, graph stream algorithms are a perfect fit for problems dealing with networks that are formed as we attempt to analyze them, e.g., the network describing the activity taking place in a social networking site. However, many challenges arise in a streaming setting that need to be addressed when designing respective techniques. A streaming graph algorithm processes the stream in the order it arrives and each element of the stream must be processed immediately or stored as it will not become available again. In addition, the size of the stream and the speed in which its elements arrive do not allow for persisting the stream in its entirety. Therefore, processing cannot occur at a later stage.

In this dissertation we focused on both distributed and streaming graph processing techniques. We initially investigated the memory usage patterns that contemporary distributed graph processing systems adopt. We observed that graph compression techniques have not been considered in the design of the representations that distributed systems employ. Therefore, we built on compression techniques that assume centralized execution and provided numerous novel compact representations that are fitting for all Pregel-like systems. Our structures offer memory-optimization regardless of the algorithm that is to be executed, and enable the successful execution of algorithms in settings that state-of-the-art systems fail to terminate. We continued by studying a problem that has received considerable attention in the past, yet is still extremely relevant as previously proposed approaches fail to handle the massive volume of today's real-world graphs. In particular, we addressed the problem of community detection and our contribution was twofold as we proposed both a vertex-centric and

a streaming approach. We followed the trend of seed-set expansion methods in which small sets of nodes are expanded to communities. Our techniques offer impressive results with regards to all accuracy, memory usage and execution time. Next, we considered the stream of real-time activity of a social networking site and investigated ways of deriving the interesting part out of it based on network properties. More specifically, we used the interactions of the site's users to construct a network of authorities and assess whether each particular element of activity in the stream is interesting. This approach enables applications in numerous fields to exploit in real-time the enormous amount of information that is made available online everyday without being overwhelmed by the volume of the information. Finally, we investigated yet another field of study in the area of graph mining, namely opinion formation. We adopted a well-studied model and employed a distributed graph processing system to evaluate whether the predicted behavior of the users of a real social network according to this model matches the actual behavior these users.

Most of these results have appeared in [33–38]. What follows is a brief presentation of the topics and results of the dissertation, avoiding technical details.

2 Memory-optimized Distributed Graph Processing

The proliferation of web applications, the explosive growth of social networks, and the continually-expanding WWW-space have led to systems that routinely handle voluminous data modeled as graphs. *Facebook* has over 1 billion active users [15] and *Google* has long reported that it has indexed unique URLs whose number exceeds 1 trillion [2]. This ever-increasing requirement in terms of graph-vertices has led to the realization of a number of distributed graph-processing approaches and systems [1, 45, 40]. Their key objective is to efficiently handle large-scale graphs using predominantly commodity hardware [26]. Most of these approaches parallelize the execution of algorithms by dividing graphs into partitions [47] and assigning vertices to workers following the "*think like a vertex*" programming paradigm introduced with `Pregel` [41]. However, recent studies [26, 13] point out that the so-far proposed frameworks [1, 45, 40] fail to handle the unprecedented scale of real-world graphs as a result of ineffective, if not right out poor, memory usage [26]. Thereby, the space requirements of real-world graphs have become a major memory bottleneck.

Deploying space-efficient graph representations in a vertex-centric distributed environment to attain memory optimization is critical when dealing with web-scale graphs and remains a challenge. Related efforts have exclusively focused on providing a compact representation of a graph in a centralized machine environment [9, 5, 14, 39]. In such single-machine settings, we can exploit the fact that vertices tend to exhibit similarities. However, this is infeasible when graphs are partitioned on a vertex basis, as each vertex must be processed independently of other vertices. Furthermore, to achieve memory optimization, we need representations that allow for mining of the graph's elements *without decompres-*

sion; this decompression would unfortunately necessitate additional memory to accommodate the resulting unencoded representation.

A noteworthy step towards memory optimization was taken by *Facebook* when it adopted `Apache Giraph` [1] for its graph search service; the move yielded both improved performance and scalability [15]. However, *Facebook*'s improvements regarding memory optimization entirely focused on a more careful implementation for the representation of the out-edges of a vertex [15]; the redundancy due to properties exhibited in real-world graphs was not exploited.

We investigate approaches that help realize compact representations of out-edges in (weighted) graphs of web-scale while following the `Pregel` paradigm. The vertex placement policy that `Pregel`-like systems follow necessitates for storing the out-edges of each vertex independently. This policy preserves the *locality of reference* property, known to be exhibited in real-world graphs [8], and enables us to exploit in this work, patterns that arise among the out-edges of a *single* vertex. We cannot however utilize similarities among out-edges of different vertices, for we are unaware of the partition each vertex is placed into. Our first technique, termed `BVEdges`, applies all methods proposed in [9] that can effectively function with the vertex placement policy of `Pregel` in a distributed environment. `BVEdges` primarily focuses on identifying intervals of consecutive out-edges of a vertex and employs universal codings to efficiently represent them. To facilitate access without imposing the significant computing overheads of `BVEdges`, we propose `IntervalResidualEdges`, which holds the corresponding values of intervals in a non-encoded format. We facilitate support of weighted graphs with the use of a parallel array holding variable-byte encoded weights, termed `VariableByteArrayWeights`. Additionally, we propose `IndexedBitArrayEdges`, a novel technique that considers the out-edges of each vertex as a single row in the adjacency matrix of the graph and indexes only the areas holding edges using byte sized bit-arrays. Finally, we propose a fourth space-efficient tree-based data structure termed `RedBlackTreeEdges`, suitable for algorithms requiring mutations of out-edges.

Our experimental results with diverse datasets indicate significant improvements on space-efficiency for all our proposed techniques. We reduce memory requirements up–to 5 times in comparison with currently applied methods. This eases the task of scaling to *billions of vertices per machine* and so, it allows us to load much larger graphs than what has been feasible thus far. In settings where earlier approaches were also capable of executing graph algorithms, we achieve significant performance improvements in terms of time of up–to 41%. We attribute this to our introduced memory optimization as less time is spent for garbage collection. These findings establish our structures as the undisputed preferable option for web graphs, which offer compression-friendly orderings, or any other type of graph after the application of a reordering that favors its compressibility. Last but not least, we attain a significantly improved trade-off between space-efficiency and performance of algorithms requiring mutations through a representation that uses a tree structure and does not depend on node orderings.

3 Uncovering Local Hierarchical Link Communities at Scale

The neurons in our brains, the proteins in live cells, the powerplants of an electrical grid, and the users of an online social networking service, are all entities of *complex systems* that play a vital role in our daily lives. Networks are a powerful tool for modeling relations and interactions between the components of such complex systems. Respective real-world networks are often massive; yet they exhibit a high level of order and organization, which allows the study of common properties they exhibit, such as the power-law degree distribution and the small-world structure [46, 19]. Another important property that real-world networks exhibit is the presence of community structure [24]. At a high level, communities are groups of nodes that share a common functional property or context, e.g., two people that attended the same school, or two movies with the same actor. In several cases communities in a network are distinct; consider for example the fans of different basketball teams. However, it is often the case that communities overlap.

Effectively extracting the community structure of a node in a network has many useful applications, e.g., i) we can provide more informative and engaging social network feeds by better understanding the membership of an individual to various organizational groups, and ii) we can suggest common friends of an individual to connect because they share mutual interests. Early community detection approaches focused either on grouping the nodes of a network or on searching for links that should be removed to separate the clusters [20]. However, these approaches did not consider the fact that communities may overlap, and ultimately could not provide an accurate representation of a network's community structure. Algorithms that followed [4, 25, 48] allow for nodes to belong to several overlapping communities by employing techniques such as link clustering, matrix factorization, and personalized PageRank vectors. Still, these approaches are not applicable to the massive graphs of the *Big Data* era, as they focus on the *entire* graph structure and do not scale with regards to both execution time and memory consumption. Recent efforts have therefore shifted the focus from the global structure to a local view of the network [30–32]. More specifically, such approaches locally expand a set of target nodes in the community of interest, instead of uncovering the communities of the entire network.

Seed set expansion approaches employ techniques such as random walks to estimate the likelihood of a node to participate in the target community, and manage to scale to large networks [30–32]. These approaches consider that overlaps between communities are sparsely connected whereas the areas where communities overlap are denser than the actual communities. However, studies of real-world networks show that two nodes are more likely to be connected if they share multiple communities in common [49]. Hence, as the overlapping area is in fact denser than the actual communities, seed set expansion approaches are driven towards nodes that reside in the overlap. In addition to this, all scalable methods require *multiple seeds* to avoid detecting multiple overlapping communities as a single one. This constitutes a challenge, as it is usually the case that

we are interested in all communities of a single node, instead of seeking one community involving multiple predefined nodes. Finally, seed set expansion approaches are shown to perform well when detecting relatively large communities, whereas high quality communities are in fact small [49].

Here, we focus on the neighbors of a single node in the network, i.e., its egonet, and aim at extracting the –possibly overlapping– communities of this node. We build upon the ideas of *link clustering* [4, 18] and employ *similarity* measures that allow for effectively handling densely connected overlaps between communities. Our intuition is that when grouping pairs of links we should capture the *extent* to which a link belongs to multiple overlapping communities. To this end, we utilize a dispersion-based tie-strength measure that helps us quantify the participation of a link's adjacent nodes to more than one community. Our approach is both *efficient* and *scalable* as we focus on local parts of graphs comprising a target node and its neighbors. As we show through experimental evaluation, we produce a more accurate and intuitive representation of the community structure around a node for a number of real-world networks.

4 Community Detection via Seed Set Expansion on Graph Streams

Graph structures attract significant attention as they allow for representing entities of various domains as well as the relationships these entities entail. Real-world networks are commonly portrayed using graphs and are often massive. Despite their size, such networks exhibit a high level of order and organization, a property frequently referred to as community structure [24]. Nodes tend to organize into densely connected groups that exhibit weak ties with the rest of the graph. We refer to such groups as communities, whereas the task of identifying them is termed *community detection*.

Community detection is a fundamental problem in the study of networks and becomes more relevant with the prevalence of online social networking services such as Twitter and Facebook. Identifying the social communities of an individual enables us to perform recommendations for new connections. Moreover, by better understanding the membership of an individual to various organizational groups, we can provide more informative and engaging social network feeds. In addition to social networks, community detection is successfully applied to numerous other types of networks, such as biological or citation networks. In the former, we are particularly interested in inferring communities of interacting proteins, whereas in the latter we wish to uncover relationships between disciplines or the citation patterns of authors [20].

In the last two decades a plethora of community detection methods has been proposed [7, 16, 43, 44, 4, 25, 48]. However, these approaches are not applicable to the massive graphs of the Big Data era, as they focus on the *entire* graph structure and do not scale with regards to both execution time and memory consumption. Recent efforts manage to scale as far as execution time is concerned by focusing on the local structure and expanding exemplary seeds-sets

into communities [30, 32, 33]. Such a seed-set expansion setting can be applied to numerous real world applications, e.g., given a few researchers focusing on Big Data we can use a citation network to detect their colleagues in the same field. However, the space requirements of such algorithms rapidly become a concern due to the unprecedented size now reached by real-world graphs. The latter have become difficult to represent in-memory even in a distributed setting [37].

An increasingly popular approach for massive graph processing is to consider a *data stream model*, in which the stream comprises the edges of a graph [42]. This is a new direction in the field of community detection and to the best of our knowledge no prior approach has considered such a setting without imposing restrictions on the order in which edges are made available [27, 50]. We propose CoEuS, a novel community detection algorithm that is fully applicable on graph streams. CoEuS is initialized with seed-sets of nodes that define different communities. As edges arrive, we can process them but we cannot afford to keep them all in-memory. Therefore, CoEuS maintains rather limited information about the adjacent nodes of each edge and their participation in the communities in question. This information is kept using probabilistic data structures to further reduce the memory requirements of our algorithm. In addition to our original idea for community detection in graph streams, we propose two algorithms to enhance the effectiveness of CoEuS. The first one focuses on better quantifying the quality of each edge w.r.t. to a community. The second one is a novel clustering algorithm that allows for automatically determining the size of the resulting communities, in spite of the absence of the graph structure.

Our experimental results on various large scale real-world graphs show that CoEuS is extremely competitive with regard to *accuracy* against approaches that employ the entire graph structure and cannot operate on graph streams. More specifically, CoEuS can process with just a few MBs, graphs that prior approaches fail to handle on a machine with 16GB of RAM. Moreover, CoEuS is able to derive the communities in question inordinately faster. More importantly, CoEuS is able to return its resulting communities *on demand* at any time as we process the graph stream. This is particularly important, as even if we could afford to use space linear to the number of a graph's edges, no other approach is able to update communities as new edges arrive with no additional *significant* computational cost.

5 Adaptively Sampling Authoritative Content from Social Activity Streams

The tremendous scale of content generation in online social networks brings several challenges to applications such as content recommendation, opinion mining, sentiment analysis, or emerging news detection, all of which have an inherent need to mine this content in real time. As an example, the daily volume of new *tweets* posted by users of `Twitter` surpasses 500 million.[1] However, not

[1] http://www.internetlivestats.com/twitter-statistics/

all generated online social activity is useful or interesting to all applications. Using `Twitter` again as an example, more than 90% of its posts is actually conversational and of interest strictly limited to a handful of users, or spam [23]. Therefore, applications such as emerging news detection that operate on the entire stream, spend a lot of computational cycles as well as storage in processing posts that are not very useful.

One way to solve this problem is, instead of processing the social activity stream in its entirety, to take a sample of the activity and operate on the sample. Through sampling, our goal is to still capture the important and interesting parts of the activity stream, while reducing the amount of data that we would have to process. To this end, one obvious approach is to perform random sampling, i.e., randomly pick a subset of the activity stream and use that in the respective application. A more effective approach however, is to sample content published in the activity stream only from the users that are considered authoritative (or *authorities*).[2] By sampling the posts of authoritative users from the stream, we are reportedly [51] more likely to produce samples that are of *high-quality*, with limited conversational content and less spam.

The challenge in sampling high quality content from a social activity stream lies therefore in identifying authoritative users. Existing work deploys white-lists of users that are likely to produce authoritative content [22, 51] and samples their activity. Although such approaches have been shown to work well for certain applications, we will show experimentally that they are unable to cope with the dynamic nature of a social activity stream where, for example, new users emerge as authorities and old ones fade out. Other prior efforts on identifying authoritative users in social networks (not streams) have focused on computing a relative ranking of users based on network attributes [3, 11, 12, 28, 52]. We build on the findings of such approaches to identify authorities likely to produce useful content; our approach is different however, as we cannot presume that the complete structure of the social network is available, nor that we can afford to process the network offline.

We operate with the more practical assumption that we have incomplete access to the social network. In other words, we do not know which users exist in the network but we simply observe some partial activity from a social activity stream. Our goal is to produce high quality samples from such streams that will still be as useful as possible compared to being able to access the entirety of the social network and the activity within.

We propose RHEA,[3] an adaptive algorithm for sampling authoritative social activity content. RHEA forms a *network of authorities* as it processes a stream and includes in its sample only the content published by the top-K authorities in this network. Given a social activity stream with user interactions (e.g., answers in `Q&A` sites or mentions in the case of `Twitter`) we create a weighted graph used to quantify user authoritativeness. To deal with the potentially enormous

[2] We use terms *authoritative users* and *authorities* interchangeably.

[3] Rhea was the Titaness daughter of the earth goddess Gaia and the sky god Uranus. Her name stands for "she who flows".

amount of items that we encounter in the stream and limit memory blowup, we construct a highly compact, yet extremely efficient sketch-based novel data structure to maintain the authoritative users of the network. Our experimental results with half a billion posts from two popular social networks show significant improvements with regard to various binary and ranked retrieval measures over previous approaches. RHEA is able to sample significantly more *relevant* documents, with *higher precision* and remarkably more accurate *ranking* compared to sampling based on static white-lists of authoritative users. Our approach is generic and can be used with any online social activity stream, as long as we can observe indicators of authoritativeness in the stream.

6 On the Impact of Social Cost on Opinion Dynamics

An ever-increasing amount of social activity information is available today, due to the exponential growth of online social networks. The structure of a network and the way the interaction among its users impacts their behavior has received significant interest in the sociology literature for many years. The availability of such rich data now enables us to analyze user behavior and interpret sociological phenomena at a large scale. *Social influence* is one of the ways in which social ties may affect the actions of an individual, and understanding its role in the spread of information and opinion formation is a new and interesting research direction that is extremely important in social network analysis. The existence of social influence has been reported in psychological studies [29] as well as in the context of online social networks [10]. The latter usually allow users to endorse articles, photos or other items, thus expressing shortly their opinion about them. Each user has an internal opinion, but since she receives a feed informing her about her friends' endorsements, her expressed (or overall) opinion may well be influenced by her friends' opinions. This process may lead to a consensus.

The most notable example of studying consensus formation due to information transmission is the DeGroot model [17]. This model considers a network of individuals with an opinion which they update using the average opinion of their friends, eventually reaching a shared opinion. In [21] the notion of an individual's internal opinion is added, which, unlike her expressed opinion, is not altered due to social interaction. This model captures more accurately the fact that consensus is rarely reached in real word scenarios. The popularity of a specific article, for instance, may vary largely between different communities in a social network. This fact gives rise to the study of the lack of consensus, and the quantification of the social cost that is associated with disagreement [6]; the authors here consider a game where the utilities are the users' social costs and perform repeated averaging to get the Nash equilibrium. The resulting models of opinion dynamics in which consensus is not in general reached allow for testing against real-world datasets, and enable the verification of influence existence. Investigating game theoretic models of networks against real data is crucial in understanding whether the behavior they portray depicts an illustration that is close to the real picture.

We study the spreading of opinions in social networks, using a variation of the DeGroot model [21] and the corresponding game detailed in [6]. We perform an extensive analysis on a large sample of a popular social network and highlight its properties to indicate its appropriateness for the study of influence. The observations we make verify our intuitions regarding the source and presence of social influence. Furthermore, we initialize instances of games using real data and use repeated averaging to calculate their Nash equilibrium. We experimentally show that our model, when properly initialized, is able to mimic the original behavior of users and captures the social cost affecting their activity more accurately than a classification model utilizing the same information.

7 Conclusions

In this dissertation we studied two research directions that allow for handling large-scale graphs, i.e., *distributed graph processing* and *streaming graph algorithms*. Our focus was on improving contemporary distributed systems, introducing novel techniques for important graph processing problems, and employing scalable platforms to empirically study real-world networks. We proposed techniques to efficiently address challenges regarding: i) memory-optimized distributed graph processing, ii) large scale distributed and streaming community detection, iii) sampling authoritative content from streams of social activity, and iv) modeling the behavior of social network users. Our contribution in all above areas through extensive experimentation is shown to be significant.

References

1. Apache Giraph. http://giraph.apache.org/
2. We knew the web was big.... http://googleblog.blogspot.ca/2008/07/we-knew-web-was-big.html
3. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: WSDM 2008. pp. 183–194
4. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**(7307), 761–764 (2010)
5. Apostolico, A., Drovandi, G.: Graph compression by BFS. Algorithms **2**(3), 1031–1044 (2009)
6. Bindel, D., Kleinberg, J.M., Oren, S.: How bad is forming your own opinion? In: FOCS. pp. 57–66 (2011)
7. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), P10008 (2008)
8. Boldi, P., Rosa, M., Santini, M., Vigna, S.: Layered label propagation: a multiresolution coordinate-free ordering for compressing social networks. In: Proc. of the 20th Int. Conf. on World Wide Web. pp. 587–596 (2011)
9. Boldi, P., Vigna, S.: The webgraph framework I: compression techniques. In: Proc. of the 13th Int. Conf. on World Wide Web, May 17-20. pp. 595–602 (2004)

10. Bond, R.M., Fariss, C.J., Jones, J.J., Kramer, A.D., Marlow, C., Settle, J.E., Fowler, J.H.: A 61-million-person experiment in social influence and political mobilization. Nature **489**(7415), 295–298 (2012)
11. Bouguessa, M., Romdhane, L.B.: Identifying authorities in online communities. ACM Trans. Intell. Syst. Technol. **6**(3), 30:1–30:23 (2015)
12. Bozzon, A., Brambilla, M., Ceri, S., Silvestri, M., Vesci, G.: Choosing the right crowd: expert finding in social networks. In: EDBT '13. pp. 637–648
13. Cai, Z., Gao, Z.J., Luo, S., Perez, L.L., Vagena, Z., Jermaine, C.M.: A comparison of platforms for implementing and running very large scale machine learning algorithms. In: SIGMOD 2014, June 22-27. pp. 1371–1382 (2014)
14. Chierichetti, F., Kumar, R., Lattanzi, S., Mitzenmacher, M., Panconesi, A., Raghavan, P.: On compressing social networks. In: KDD 2009, June 28 - July 1. pp. 219–228 (2009)
15. Ching, A., Edunov, S., Kabiljo, M., Logothetis, D., Muthukrishnan, S.: One Trillion Edges: Graph Processing at Facebook-Scale. Proc. of the VLDB Endowment **8**(12), 1804–1815 (2015)
16. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. Physical review E **70**(6), 066111 (2004)
17. DeGroot, M.H.: Reaching a consensus. Journal of the ASA **69**(345), 118–121 (1974)
18. Evans, T., Lambiotte, R.: Line graphs, link partitions, and overlapping communities. Physical Review E **80**, 016105 (2009)
19. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. In: ACM SIGCOMM computer communication review. vol. 29, pp. 251–262. ACM (1999)
20. Fortunato, S.: Community detection in graphs. Physics Reports **486**(3), 75–174 (2010)
21. Friedkin, N.E., Johnsen, E.C.: Social influence and opinions. Journal of Mathematical Sociology **15**(3-4), 193–206 (1990)
22. Ghosh, S., Sharma, N.K., Benevenuto, F., Ganguly, N., Gummadi, P.K.: Cognos: crowdsourcing search for topic experts in microblogs. In: SIGIR '12. pp. 575–590
23. Ghosh, S., Zafar, M.B., Bhattacharya, P., Sharma, N.K., Ganguly, N., Gummadi, P.K.: On sampling the wisdom of crowds: random vs. expert sampling of the twitter stream. In: CIKM'13. pp. 1739–1744
24. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proc. of the National Academy of Sciences **99**(12), 7821–7826 (2002)
25. Gleich, D.F., Seshadhri, C.: Vertex neighborhoods, low conductance cuts, and good seeds for local community methods. In: KDD 2012. pp. 597–605 (2012)
26. Han, M., Daudjee, K., Ammar, K., Özsu, M.T., Wang, X., Jin, T.: An Experimental Comparison of Pregel-like Graph Processing Systems. Proc. of the VLDB Endowment **7**(12), 1047–1058 (2014)
27. Hollocou, A., Maudet, J., Bonald, T., Lelarge, M.: A linear streaming algorithm for community detection in very large networks. ArXiv e-prints (Mar 2017)
28. Jurczyk, P., Agichtein, E.: Discovering authorities in question answer communities by using link analysis. In: CIKM 2007. pp. 919–922
29. Kelman, H.C.: Compliance, identification, and internalization: Three processes of attitude change. Journal of conflict resolution pp. 51–60 (1958)
30. Kloster, K., Gleich, D.F.: Heat kernel based community detection. In: KDD '14. pp. 1386–1395 (2014)
31. Kloumann, I.M., Kleinberg, J.M.: Community membership identification from small seed sets. In: KDD '14, August 24 - 27. pp. 1366–1375 (2014)

32. Li, Y., He, K., Bindel, D., Hopcroft, J.E.: Uncovering the small community structure in large networks: A local spectral approach. In: WWW 2015 (2015)
33. Liakos, P., Ntoulas, A., Delis, A.: Scalable link community detection: A local dispersion-aware approach. In: 2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, December 5-8, 2016. pp. 716–725 (2016)
34. Liakos, P., Ntoulas, A., Delis, A.: COEUS: community detection via seed-set expansion on graph streams. In: 2017 IEEE International Conference on Big Data, BigData 2017, Boston, MA, USA, December 11-14, 2017. pp. 676–685 (2017)
35. Liakos, P., Ntoulas, A., Delis, A.: Rhea: Adaptively sampling authoritative content from social activity streams. In: 2017 IEEE International Conference on Big Data, BigData 2017, Boston, MA, USA, December 11-14, 2017. pp. 686–695 (2017)
36. Liakos, P., Papakonstantinopoulou, K.: On the impact of social cost in opinion dynamics. In: Proceedings of the 10th Int. Conf. on Web and Social Media, Cologne, Germany, May 17-20, 2016. pp. 631–634 (2016)
37. Liakos, P., Papakonstantinopoulou, K., Delis, A.: Memory-optimized distributed graph processing through novel compression techniques. In: Proc. of the 25th ACM Int. Conf, on Information and Knowledge Management, CIKM 2016, Indianapolis, IN, USA, October 24-28, 2016. pp. 2317–2322
38. Liakos, P., Papakonstantinopoulou, K., Delis, A.: Realizing memory-optimized distributed graph processing. IEEE Trans. Knowl. Data Eng. **30**(4), 743–756 (2018)
39. Liakos, P., Papakonstantinopoulou, K., Sioutis, M.: Pushing the Envelope in Graph Compression. In: Proc. of the 23rd ACM Int. Conf. on Information and Knowledge Management. pp. 1549–1558. Shanghai, China (2014)
40. Low, Y., Gonzalez, J., Kyrola, A., Bickson, D., Guestrin, C., Hellerstein, J.M.: Distributed GraphLab: A Framework for Machine Learning in the Cloud. Proc. of the VLDB Endowment **5**(8), 716–727 (2012)
41. Malewicz, G., Austern, M.H., Bik, A.J.C., Dehnert, J.C., Horn, I., Leiser, N., Czajkowski, G.: Pregel: A System for Large-Scale Graph Processing. In: SIGMOD 2010, June 6-10. pp. 135–146 (2010)
42. McGregor, A.: Graph stream algorithms: a survey. SIGMOD Record **43**(1), 9–20
43. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**(2), 026113 (Feb 2004)
44. Pons, P., Latapy, M.: Computing communities in large networks using random walks. In: Computer and Information Sciences-ISCIS 2005, pp. 284–293 (2005)
45. Salihoglu, S., Widom, J.: GPS: a graph processing system. In: SSDBM 2013, July 29 - 31. pp. 22:1–22:12 (2013)
46. de Sola Pool, I., Kochen, M.: Contacts and influence. Social networks **1**(1), 5–51 (1978)
47. Ugander, J., Backstrom, L.: Balanced Label Propagation for Partitioning Massive Graphs. In: WSDM 2013, February 4-8. pp. 507–516 (2013)
48. Yang, J., Leskovec, J.: Overlapping community detection at scale: a nonnegative matrix factorization approach. In: WSDM 2013. pp. 587–596 (2013)
49. Yang, J., Leskovec, J.: Structure and overlaps of ground-truth communities in networks. ACM Trans. on Intelligent Systems and Technology **5**(2), 26 (2014)
50. Yun, S., Lelarge, M., Proutière, A.: Streaming, memory limited algorithms for community detection. In: NIPS 2014, December 8-13. pp. 3167–3175 (2014)
51. Zafar, M.B., Bhattacharya, P., Ganguly, N., Ghosh, S., Gummadi, K.P.: On the wisdom of experts vs. crowds: Discovering trustworthy topical news in microblogs. In: CSCW 2016. pp. 437–450
52. Zhang, J., Ackerman, M.S., Adamic, L.A.: Expertise networks in online communities: structure and algorithms. In: WWW 2007. pp. 221–230

Quality of Experience Characterization and Provisioning in Mobile Cellular Networks

Eirini Liotou*

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

Abstract. Traditionally, mobile cellular networks have been designed with Quality of Service (QoS) criteria in mind. Quality of Experience (QoE) has, however, recently emerged as a concept, disrupting the design of future network generations. The emergence of the QoE concept has been a result of the inevitable strong transition that the mobile industry is experiencing from system-centric networks to more user-centric solutions. Motivated by this boost towards user-centricity, the objective of this dissertation is to explore the challenges and opportunities that arise in modern cellular networks when QoE is considered. In this direction, throughout this dissertation, QoE estimation models and metrics are explored and exploited in order to quantify QoE and improve existing network mechanisms. The core of this dissertation is the proposal of a QoE provisioning cycle that allows the control, monitoring (i.e., modeling) and management of QoE in a cellular network. In terms of modeling, QoE assessment methods and QoE-related performance indicators are described and classified, with an emphasis on parametric quality estimation. In terms of QoE management, novel QoE-aware mechanisms that demonstrate QoE improvements for the users are proposed, such as a radio scheduling algorithm that improves QoE by mitigating throughput fluctuations, and a context-aware HTTP Adaptive Streaming (HAS) mechanism that successfully mitigates stallings in bandwidth-challenging scenarios. Finally, a programmable QoE-SDN APP into the Software-Defined Networking (SDN) architecture is introduced, which enables network feedback exposure from Mobile Network Operators (MNOs) to Video Service Providers (VSPs), revealing QoE benefits for VSPs' customers and bandwidth savings for the MNOs.

Keywords: Quality of Experience (QoE), HTTP Adaptive Streaming (HAS), radio resource scheduling, Software-Defined Networking (SDN), mobile cellular networks.

* Dissertation Advisor: Lazaros Merakos, Professor

1 Dissertation Summary

1.1 Motivation and Scope

Over the last few years, there has been a tremendous increase in the network traffic generated by mobile users, a phenomenon which can be attributed to multiple factors. On the one hand, the emergence of smart phones and tablets along with the huge, recently emerged app market have changed the landscape in the telecommunications sector. In parallel, the charges even for intensive data usage are tolerable, as network operators offer very attractive subscription packets to attract customers. On the other hand, modern networks, such as the Long Term Evolution - Advanced (LTE-A) and emerging 5G networks, can offer very high bandwidth to their users, supporting a plethora of diverse, resource-hungry services, and further boosting the demand for data consumption. All these conditions make mobile users more and more demanding in terms of the quality they expect to achieve.

Recognizing this fact, there has lately been a momentum that pushes the epicenter of interest from the "network" to the "user". While network and service providers are trying to create or follow this "user-centric" trend, new terms have been coined that allow its more comprehensive description. The term "Quality of Experience" (QoE) is irrefutably the most dominant one, as it describes "*the overall acceptability of an application or service, as perceived subjectively by the end-user*". This means that older terms such as Quality of Service (QoS), traditionally used for years, are now considered only partial or incomplete. The reason behind that is that QoS can only record the technical characteristics of a service without giving a clear indication of the user's satisfaction when interacting with this service. In fact, the relationship between these two metrics (QoS and QoE) has been found to be non-linear.

The definition of QoE makes clear that it is a very broad and generic concept, and as such, it incorporates the complete end-to-end system effects (terminal, network, services, etc.) together with the human impressions of these effects. As vague as the concept of QoE may sound, reliable estimation methods have been developed with the assistance of subjective experiments with human evaluators. These experiments lead to reliable QoE assessment methods, which manage to automatically evaluate and rate the QoE of a user with respect to a specific application or service. This procedure is called "QoE modeling", and it is the most important first step towards QoE provisioning.

The awareness of an overall QoE score is very important for all involved stakeholders in the service communication chain. Once QoE is measured, this may be exploited in many aspects by network operators and service providers. First of all, the extraction of a QoE score of a service with respect to a user is the most attractive and absolute way to evaluate the performance of the offered services. Second, network problems such as bottlenecks or local failures may be identified by predefined QoE thresholds, and proactive or reactive actions may be triggered to correct them. A third important motive for QoE awareness is the possibility to incorporate QoE intelligence in the network mechanisms, and specifically in the network decision processes. This may lead to "QoE-driven" or otherwise called "QoE-aware" algorithms that can help the network function in a more efficient and effective way. For instance, QoE may become the

criterion or trigger mechanism of standard network algorithms (e.g., radio resource scheduling, mobility management, power control, etc.) replacing current QoS-based criteria, such as plain signal strength measurements. What is more, understanding and identifying the key factors that truly affect the user's experience creates the possibility to propose innovative algorithms that focus on targeted QoE performance indicators. Finally, QoE-awareness may drive a more resource-efficient network operation, by helping recognize moments and cases of operation when providing extra resources to the users would not improve their perceived QoE. In other words, "over-engineering" could be avoided.

QoE modeling and management in mobile cellular networks are fundamental components, part of a wider framework that enforces the end-to-end QoE provisioning. This framework also includes wider challenges such as the collection of appropriate input data that will lead to the awareness of QoE (i.e., QoE monitoring), the realistic implementation of such a framework in real networks, and the possible interaction between network providers and service providers, aiming at the holistic delivery of optimal QoE to the end-users, among others.

This dissertation focuses on exploring the challenges and opportunities that arise in modern mobile cellular networks in terms of QoE provisioning to end-users. Specifically, it aims to characterize and exploit QoE models and metrics in order to improve existing mechanisms in mobile cellular networks standardized by 3GPP (3rd Generation Partnership Project), but also towards the 5G horizon, such as radio resource allocation, Device-to-Device (D2D) communication setup, adaptive video streaming, etc.

1.2 Dissertation Contributions

In this dissertation, the reader will delve into details regarding the topic of QoE management in mobile cellular communication networks. The main contributions of the research conducted in this dissertation are the following:

- Proposal of a conceptual framework for achieving end-to-end QoE provisioning in mobile cellular networks. This framework is analyzed in terms of its design, its constituents and their interactions, as well as key implementation challenges, while its proof-of-concept in an LTE network is assessed.
- The identification and analysis of parametric QoE formulas and Key Performance Indicators (KPIs) that can be used for real-time QoE assessment of popular service types in communication networks (i.e., VoIP, online video, video streaming, web browsing, Skype, IPTV and file download services).
- A network management framework that exploits QoE awareness for controlling the operational mode of mobile users in LTE-A networks with D2D support. Simulation studies have revealed the twofold benefits of this mechanism, i.e., both for the users (increase in QoE) and the operators (increase in offered throughput).
- Proposal of a new radio scheduling logic, which takes into account the impact of throughput fluctuations on the QoE of interactive applications. By quantifying how traditional radio scheduling decisions influence the user-perceived QoE, a novel "consistent" resource allocation process is proposed, which further improves users' QoE by moderating these fluctuations.

- Analytical investigation of the video quality degradation problem as it is experienced by mobile users in vehicles, and proposal of a proactive context-aware HTTP Adaptive Streaming (HAS) strategy, which helps prevent stallings in light of bandwidth-challenging situations.
- Proposal of a Software-Defined Networking (SDN)-based architecture that promotes and enables a technologically feasible realization of a collaboration paradigm between Video Service Providers (VSPs) and Mobile Network Operators (MNOs). The potential of this architecture is highlighted through the proposal and evaluation of three use cases that are unlocked by this architecture, in the context of HAS. In this paradigm, feedback about the network throughput is provided to a VSP so that he can be in a stronger position to redefine encoding, caching, and per-user video segment selection.
- Identification of the essential attributes that can shape QoE-centric networks towards the 5G era, and introduction of the "experience package" concept. Experience packages can lead to a more personalized service provisioning to users, considering not only technical parameters, but also the user profile and the context of the communication.

Below, results concerning the most important contributions are presented.

2 Results and Discussion

2.1 A Conceptual Framework towards QoE Management in Mobile Cellular Networks

The first main study conducted in this dissertation provides insights on the issue of network-level QoE management, identifying the open issues and prerequisites towards acquiring QoE awareness and enabling QoE support in mobile cellular networks. A conceptual framework for achieving end-to-end QoE provisioning is proposed (Fig. 1), and described in detail in terms of its design, its constituents and their interactions, as well as the key implementation challenges. The main components of this framework are three building blocks, namely the QoE-Controller, QoE-Monitor and QoE-Manager. Apart from proposing and describing a high-level architecture for QoE management in mobile cellular networks, we use the LTE-A network as a case study to demonstrate the feasibility, performance issues and potential benefits of the proposed QoE management framework, using simulation.

Specifically, we describe how this high-level architecture may be customized and applied for the purposes of implementing a real-time QoE-aware admission controller in LTE-A. We study a scenario where the user density in an outdoors small cell is gradually increasing, representing for instance scenarios where this small cell is used to serve a stadium during a concert or a football game. We evaluate the proposed QoE management framework and compare it with the conventional case where no QoE management framework is present, and therefore, users are admitted based on their positions or on received signal strengths from the surrounding base stations. (Fig. 2). It is observed, that the QoE management framework surpasses the QoE achieved via

conventional admission control schemes, due to using the actual quality experienced by the users as the decisive criterion for admission.

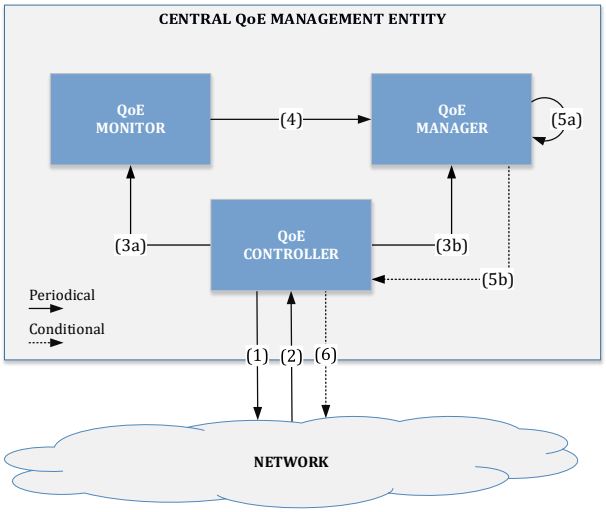


Fig. 1. The proposed QoE management framework.

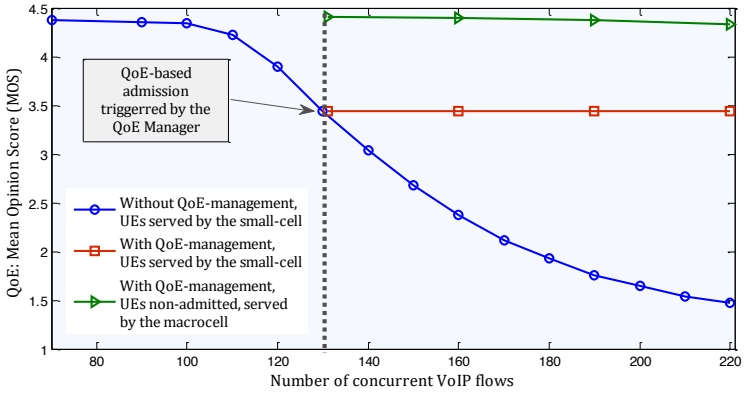


Fig. 2. QoE management framework evaluation for a QoE-driven admission control scenario.

2.2 Parametric QoE Estimation for Popular Services

As we are moving closer and closer to future network generations, the human factor is becoming the epicenter of attention and the driving force for the network design. Thus, the comprehension and, in extension, the control of the provisioned QoE to the users has become a necessity for network operators. Parametric QoE estimation models, i.e., formula-based QoE models, are a prerequisite for this purpose. They constitute the ideal tools towards live network quality monitoring and, hence, QoE management.

Nevertheless, despite the increased interest from academia and industry to push towards a QoE service provisioning model, a clear/comprehensive manual on the available parametric models and the critical QoE performance parameters per service type is currently missing. Identifying this gap, a second study conducted in this dissertation provides a thorough and handy "manual", currently absent from the literature, that identifies and describes appropriate parametric models for popular services nowadays, such as YouTube, Skype and IPTV, as well as summarizes standardized ones (Table 1).

Table 1. Parametric QoE estimation per service type.

Service Type	QoE Estimation Model	KPIs
File transfer	Data rate-based formula	Data rate, expected upper and lower data rate
Web browsing	Response time-based formula	Response time
Skype	Skype-specific formula	Frame rate, image quality, resolution
VoIP	ITU-T Rec. G.107, E-model	Packet loss ratio, delay, codec, coding rate
Video streaming	IPTV model	Data rate, frame rate
	YouTube (conventional) model	Number of stalling events, duration of stalling events, video duration
	YouTube with adaptive streaming model	Time on highest layer, amplitude, frequency of quality switches
Online video	ITU-T Rec. G.1070, E-model	Packet loss ratio for audio and video packets, relative delay between video and audio packets, data rate, frame rate, monitor size

2.3 QoE-Inspired Consistency in Radio-Scheduling

Radio scheduling is a well-studied problem that has challenged researchers throughout the last decades. However, recent findings that stem from the QoE domain come to give a new perspective to traditional radio scheduling approaches. In this study, we take advantage of recent subjective results regarding the impact of throughput fluctuations on the QoE of interactive applications and revisit well-known scheduling algorithms. By quantifying the impact of traditional radio schedulers on user-perceived QoE, we manage to draw new conclusions regarding the radio scheduling problem, such as the importance and impact of consistency of the resource allocation decisions on the users' QoE. As main result, fair algorithms inherently seem to be more consistent than greedy ones, providing less throughput fluctuations and, thus, better QoE. Based on this outcome, we propose a new scheduling approach, which further improves users' QoE by moderating throughput fluctuations. Such fluctuations' effect may be moderated, i.e., smoothed out, by introducing a new scheduling metric ($\boldsymbol{m_{i,k-fluct}}$) that tries to capture and mitigate the magnitude and occurrence of throughput fluctuations.

For evaluation purposes of the herein proposed scheduler, we compare the CDF of this scheduler with state of the art schedulers, for the case of 20 users uniformly

distributed in a cell. The results are presented in Fig. 3. We can observe that the proposed scheduler: a) is very fair, as shown by the steepness of the CDF, b) that the achieved minimum Mean Opinion Score (MOS) values are higher than for the other schedulers (CDF shifted to the right), while c) the larger MOS values are comparable to the other schedulers. This behavior is explained by the fact that the resource allocation procedure of the proposed scheduler is greedy in some sense. Trying to minimize the gap between the average throughput values and the potentially achieved data rates jointly for all the users, eventually this scheduler manages to first satisfy the low-throughput users. However, the low-throughput users do not necessarily take the "best" Resource Blocks, and therefore higher-throughput users are also served well.

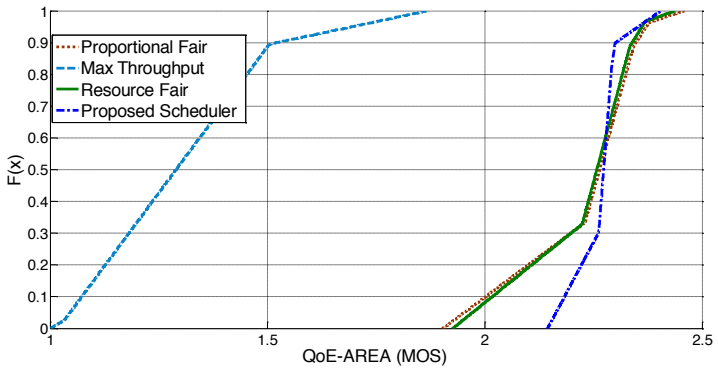


Fig. 3. MOS CDF for standard schedulers and the $m_{i,k-fluct}$ metric.

2.4 Enriching HTTP Adaptive Streaming (HAS) with Context Awareness

Video streaming has become an indispensable technology in people's lives, while its usage keeps constantly increasing. The variability, instability and unpredictability of network conditions poses one of the biggest challenges to video streaming. In this study, we analyze HAS, a technology that relieves these issues by adapting the video reproduction to the current network conditions. Particularly, we study how context awareness can be combined with the adaptive streaming logic to design a proactive client-based video streaming strategy. Our results show that such a context-aware strategy manages to successfully mitigate stallings in light of network connectivity problems, such as an outage. Moreover, we analyze the performance of this strategy by comparing it to the optimal case in terms of QoE-related KPIs for video streaming. The collected evaluation results encourage further research on how context-awareness can be exploited to further enhance video service provisioning by service providers.

In Fig. 5 we present: a) the conventional case, where no context awareness about an imminent outage event is available, and consequently, a standard HAS strategy is continuously executed, b) the case where context awareness about the starting point and duration of this outage event is available, which automatically leads to the selection of a proactive HAS strategy after a minimum required "advance time", and c) the optimal strategy in terms of maximizing the selected video quality layers and minimizing the

quality switches. Looking at Fig. 5 we can see that a stalling of around 80 sec is completely avoided when context awareness is deployed, or when optimal knowledge is assumed. The explanation behind the prevention of the stalling lies in Fig. 5: In the "without context" case higher HAS layers are selected as compared to the "with context" case. Having downloaded lower HAS layers in the "with context" case, the buffer of the client is fuller in terms of playtime than it would have been if higher HAS layers had been downloaded instead.

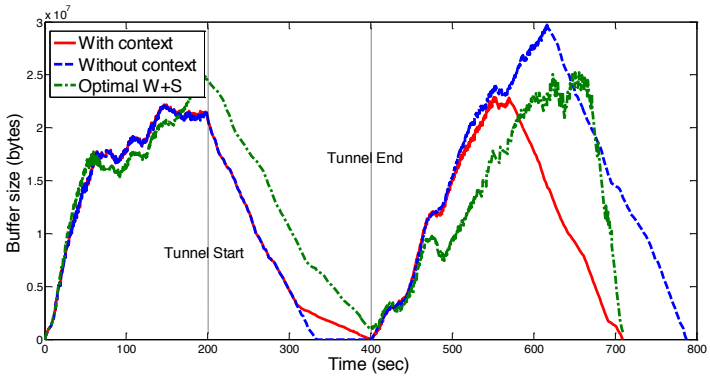


Fig. 4. Comparison of buffer size for different HAS strategies.

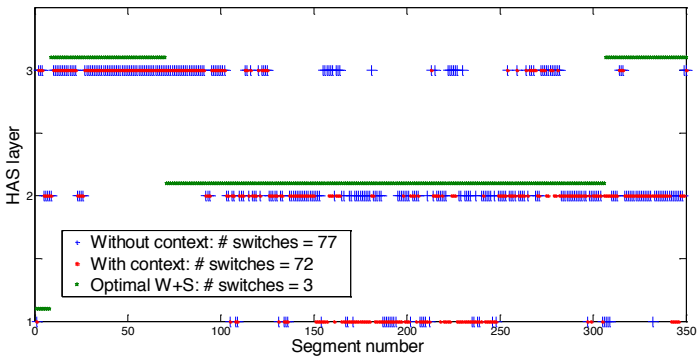


Fig. 5. Comparison of selected HAS layers for different HAS strategies.

2.5 QoE-SDN APP: A Rate-Guided QoE-Aware SDN-APP for HTTP Adaptive Video Streaming

While video streaming has dominated the Internet traffic, VSPs compete on how to assure the best QoE to their customers. HAS has become the de facto way that helps VSPs work-around potential network bottlenecks that inevitably cause stallings. However, HAS-alone cannot guarantee a seamless viewing experience, since this highly relies on the MNOs' infrastructure and evolving network conditions. SDN has brought new perspectives to this traditional paradigm where VSPs and MNOs are isolated,

allowing the latter to open their network for more flexible, service-oriented programmability. This study takes advantage of recent standardization trends in SDN and proposes a programmable QoE-SDN APP, enabling network exposure feedback from MNOs to VSPs towards network-aware video segment selection and caching, in the context of HAS. A number of use cases, enabled by the QoE-SDN APP, are designed to evaluate the proposed scheme, revealing QoE benefits for VSPs and bandwidth savings for MNOs.

The proposed QoE-SDN APP (Fig. 6) relies on the SDN architecture allowing the SDN controller to maintain a corresponding APP template. Such template offers VSPs the opportunity to program their QoE requirements and QoE assessment logic once subscribed. VSPs can then use the QoE-SDN APP to enhance their video segment encoding and distribution procedures by getting network feedback exposed by the MNOs.

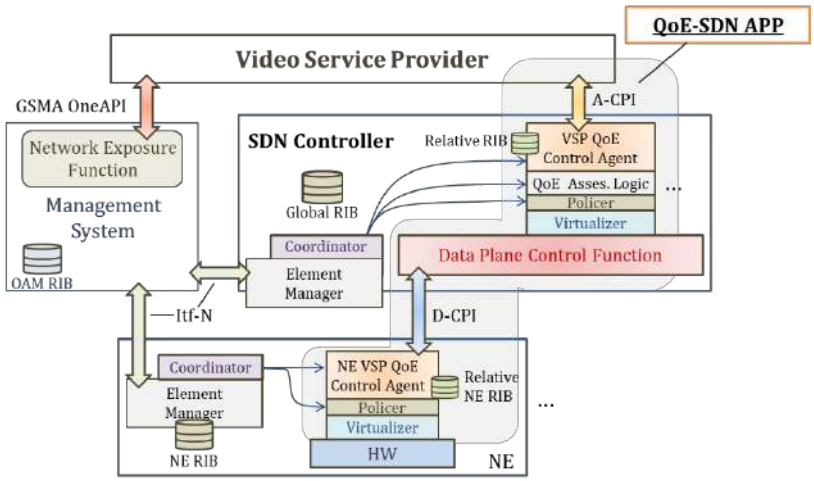


Fig. 6. QoE-SDN APP functions and architecture.

For our evaluation analysis, we adopted three use cases, considering first a HAS segment selection enforcement problem, then a segment encoding and placement (i.e., caching) problem, and finally a proactive segment selection and placement problem. Simulations were conducted comparing the aforementioned use cases with a standard, i.e., state of the art, version of HAS and with a conservative HAS variation that introduces minimum stalling events. In Fig. 7 we present the results with respect to the segment selection enforcement use case considering the aforementioned different HAS variations.

As shown in this figure, the experienced video bit rate in the system is higher for the standard case, followed by the rate-guided HAS (with the QoE-SDN APP) and the minimum stallings HAS. This is due to the fact that the standard HAS case allows users to select segments with a higher quality layer in contrast with the proposed rate-guided HAS, which takes a more conservative approach, guiding users to select segments with a lower quality. However, the proposed rate-guided HAS as well as the minimum stallings HAS allow more segments (i.e., more playtime) to be buffered, preparing the

video player better for imminent congestion and worse channel conditions. Therefore, such higher quality layer selection for standard HAS is the result of overestimated subjective bandwidth calculations that mislead users to request segments with a higher quality layer, and thus, eventually experience stalling events. This effect is illustrated in Fig. 7, where the QoE model for YouTube gives an estimation of the MOS as a function of the number and duration of stalling events, showing the benefits in terms of QoE for the proposed rate-guided HAS. Since stalling is the most important QoE shaping factor, such an improvement is highly desirable for the users and VSPs.

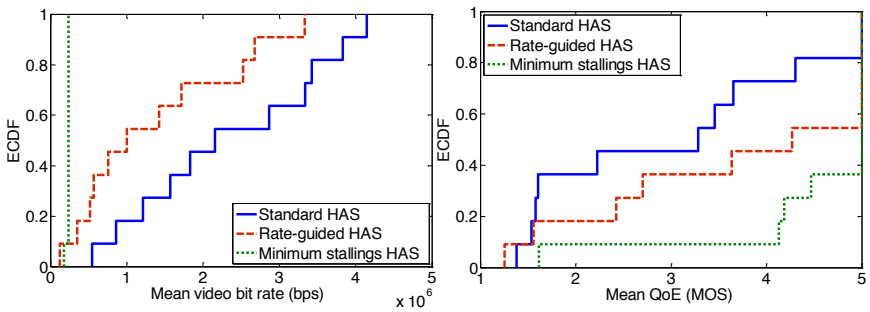


Fig. 7. ECDF of the mean video bit rate for all users (left), ECDF of MOS for all users (right).

3 Conclusions

The introduction of QoE intelligence and QoE-aware capabilities in mobile cellular networks changes the network management approach. Future network management implements a QoE management cycle, where a) QoE-related intelligence is gathered, b) QoE modeling and monitoring reveals the user satisfaction level or warns about imminent problems, and finally, c) a QoE control (i.e., management) procedure triggers proactive or reactive actions to appropriate network elements and functions.

This dissertation has dealt with the challenges arising from the need to integrate QoE intelligence in a mobile cellular network, which mainly concern the real-time evaluation of QoE, the improvement of existing network mechanisms, and the proposal of new QoE-inspired algorithms, stemming from the inherent characteristics of QoE and the non-linear impact of conventional QoS parameters on the user perception.

As a general comment, the research conducted in this dissertation has focused on the integration of QoE to research topics that are currently under intense research interest from academia and industry, such as D2D, HAS, radio scheduling and SDN. However, this is just a subset of potential solutions that may be proposed, when QoE intelligence is integrated into the real-time operation of a future network. Nevertheless, this dissertation provides valuable insights and useful findings in this direction, further encouraging research in the area of QoE characterization and provisioning in mobile cellular networks.

Overall, this thesis promotes the uniting of the domain of QoE with the domain of mobile communications, as well as the collaboration of mutual-interest between MNOs

(network layer) with service providers (application layer), presenting the high potential from such approaches for all involved stakeholders.

References

1. D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, "Enabling D2D communications in LTE networks," 24th International Symposium on Personal, Indoor and Mobile Radio Communications (IEEE PIMRC), London, United Kingdom, September 2013.
2. D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, "LTE-A access, core, and protocol architecture for D2D communication," Smart Device to Smart Device Communication, Springer International Publishing, Editors: S. Mumtaz and J. Rodriguez, ISBN: 978-3-319-04963-2, pp. 23-40, April 2014.
3. D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, "The need for QoE-driven interference management in femtocell-overlaid cellular networks," Mobile and Ubiquitous Systems: Computing, Networking, and Services, Springer International Publishing, Editors: I. Stojmenovic, Z. Cheng, and S. Guo, ISBN: 978-3-319-11569-6, vol. 131, pp. 588-601, September 2014.
4. E. Liotou, E. Papadomichelakis, N. Passas, and L. Merakos, "Quality of Experience-centric management in LTE-A mobile networks: The Device-to-Device communication paradigm," 6th International Workshop on Quality of Multimedia Experience (IEEE QoMEX), Singapore, September 2014.
5. E. Liotou, D. Tsolkas, N. Passas, and L. Merakos, "Ant Colony Optimization for resource sharing among D2D communications," 19th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (IEEE CAMAD), Athens, Greece, December 2014.
6. E. Liotou, G. Tseliou, K. Samdanis, D. Tsolkas, F. Adelantado, and C. Verikoukis, "An SDN QoE-Service for dynamically enhancing the performance of OTT applications," 7th International Workshop on Quality of Multimedia Experience (IEEE QoMEX), Costa Navarino, Greece, May 2015.
7. E. Liotou, H. Elshaer, R. Schatz, R. Irmer, M. Dohler, N. Passas, and L. Merakos, "Shaping QoE in the 5G ecosystem," 7th International Workshop on Quality of Multimedia Experience (IEEE QoMEX), Costa Navarino, Greece, May 2015.
8. E. Liotou, N. Passas, and L. Merakos, "Towards QoE provisioning in next generation cellular networks," IEEE Communications Society, Multimedia Communications Technical Committee E-Letter, vol. 10, no. 3, May 2015.
9. E. Liotou, D. Tsolkas, N. Passas, and L. Merakos, "Quality of Experience management in mobile cellular networks: Key issues and design challenges," IEEE Communications Magazine, Network & Service Management Series, vol. 53, no. 7, pp. 145-153, July 2015.
10. D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, "Addressing traffic demanding scenarios in cellular networks through QoE-based rate adaptation," 26th International Symposium on Personal, Indoor and Mobile Radio Communications (IEEE PIMRC), Hong Kong, China, August 2015.
11. D. C. Mocanu, J. Pokhrel, J. P. Garella, J. Seppänen, E. Liotou, and M. Narwaria, "No-reference video quality measurement: Added value of machine learning," Journal of Electronic Imaging, vol. 24, no. 6, December 2015.
12. E. Liotou, T. Hoßfeld, C. Moldovan, F. Metzger, D. Tsolkas, and N. Passas, "Enriching HTTP adaptive streaming with context awareness: A tunnel case study," International Conference of Communications (IEEE ICC), Kuala Lumpur, Malaysia, May 2016.

13. E. Liotou, D. Tsolkas, and N. Passas, “A roadmap on QoE metrics and models,” 23rd International Conference of Telecommunications (IEEE ICT), Thessaloniki, Greece, May 2016.
14. E. Liotou, D. Tsolkas, K. Samdanis, N. Passas, and L. Merakos, “Towards Quality of Experience management in the next generation of mobile networks,” 25th European Conference on Networks and Communications (EuCNC), Athens, Greece, June 2016.
15. F. Metzger, E. Liotou, C. Moldovan, and T. Hoßfeld, “TCP video streaming and mobile networks: Not a love story, but better with context,” Elsevier Computer Networks, Special Issue on “Traffic and Performance in the Big Data Era,” vol. 109, pp. 246-256, November 2016.
16. E. Liotou, R. Schatz, A. Sackl, P. Casas, D. Tsolkas, N. Passas, and L. Merakos, “The beauty of consistency in radio-scheduling decisions,” 59th Global Communications Conference (IEEE Globecom Wkshps) - International Workshop on Quality of Experience for Multimedia Communications (QoEMC), Washington, DC, USA, December 2016.
17. D. Tsolkas, E. Liotou, N. Passas, and L. Merakos, “A survey on parametric QoE estimation for popular services,” Elsevier Network and Computer Applications, vol. 77, pp. 1-17, January 2017.
18. E. Liotou, A. Sfikopoulos, P. Kaltzias, and V. Tsolkas, “An evaluation of buffer- and rate-based HTTP adaptive streaming strategies,” 22nd International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (IEEE CAMAD), Lund, Sweden, June 2017.
19. S. Tennina, I. Tunaru, G. Karopoulos, D. Xenakis, E. Liotou, and N. Passas, “Secure energy management in smart energy networks,” 22nd International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (IEEE CAMAD), Lund, Sweden, June 2017.
20. E. Liotou, A. Marotta, L. Pomante, and K. Ramantas, “A middleware architecture for QoE provisioning in mobile networks,” 22nd International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (IEEE CAMAD), Lund, Sweden, June 2017.
21. E. Liotou, N. Passas, and L. Merakos, “The emergence of experience packages in the 5G era,” IEEE 5G Tech Focus online journal, September 2017.
22. F. Metzger, T. Hoßfeld, L. Skorin-Kapov, Y. Haddad, E. Liotou, P. Pocta, H. Melvin, V. Siris, A. Zgank, and M. Jarschel, “Context monitoring for improved system performance and QoE,” Autonomous Control for a Reliable Internet of Services: Methods, Models, Approaches, Techniques, Algorithms and Tools, Springer International Publishing, Editors: I. Ganchev, R. van der Mei, and J. L. van den Berg, to appear.
23. R. Schatz, S. Schwarzmann, T. Zinner, O. Dobrijevic, E. Liotou, P. Pocta, S. Barakovic, J. Barakovic Husic, and L. Skorin-Kapov, “QoE Management for future networks,” Autonomous Control for a Reliable Internet of Services: Methods, Models, Approaches, Techniques, Algorithms and Tools, Springer International Publishing, Editors: I. Ganchev, R. van der Mei, and J. L. van den Berg, to appear.
24. E. Liotou, T. Hoßfeld, C. Moldovan, F. Metzger, D. Tsolkas, and N. Passas, “The value of context-awareness in bandwidth-challenging HTTP Adaptive Streaming scenarios,” Autonomous Control for a Reliable Internet of Services: Methods, Models, Approaches, Techniques, Algorithms and Tools, Springer International Publishing, Editors: I. Ganchev, R. van der Mei, and J. L. van den Berg, to appear.
25. E. Liotou, K. Samdanis, E. Pateromichelakis, N. Passas, and L. Merakos, “QoE-SDN APP: A rate-guided QoE-aware SDN-APP for HTTP adaptive video streaming,” IEEE Journal on Selected Areas in Communications, under review.

Methodologies for developing Academic/Research Skills in Computer Science and Telecommunications Departments

Irini M. Mamakou[1]

National and Kapodistrian University of Athens

Department of Informatics and Telecommunications

mamakou@unipi.gr

Abstract: Based on the demonstrated value of academic literacy and communication skills in academia and professional life, numerous experts, curricula designers and accreditation boards consider them integral in computer science curricula and value them as an important component in academic study. Although engineering education programs have kept pace with emerging disciplinary knowledge, research and technologies, they have been less successful in ensuring that their graduates acquire the skills, and attitudes desired by academia and the workplace.
More specifically, Computer Science and Engineering departments in universities in Greece focus almost exclusively on the "hard" skills, that is, the theoretical and technical subject areas, and ignore provision of support for the above-mentioned survival skills. An Action Research case study on engineering pedagogy was conducted at two departments of a university by means of a course designed to face the shortfalls for current generation of freshmen which is severe in Greek Universities and to fuse essential disciplinary needs, skills and knowledge in a module, by integrating it into the curriculum and contextualizing it to develop a novel course for Computer Science and Telecommunication undergraduate students; emphasis is laid on academic/professional skills, literacy and conventions specific to the scientific discourse community they have just entered through an integrated project-based approach to ensure active learning, engagement, enculturation, consolidation, collaboration and the teaching-research nexus. The purpose of this empirical investigation is to present the experimental findings of this integration, students' response, experience, challenges, and benefits, seeking to examine student understanding of academic conventions, and how well students transfer this learning to their projects. Data was collected through various research instruments and results reveal that students achieved substantial learning gains in academic skills and significant growth toward a more mature understanding of academic life, responsibility, ethics and integrity.

Keywords: Academic and Research Skills, Engineering Education, Project-based Learning

1 Introduction- Contribution/Statement of Purpose of this Dissertation

This dissertation captures and distils lessons learned through an empirical, experimental approach to developing skills and fulfilling key needs in engineering education.New requirements in engineering education concerning content and pedagogy, generates a growing concern about the imbalance between hard core and soft skills courses in many Engineering curricula, especially in Greece, as it deprives students from acquiring discipline specific conventions, enculturation and the teaching-research nexus, and graduates from work readiness and employability. It provides a proposal and empirical insight through the application of an eclectic methodology to address deficiencies and harmonise both instruction and content in Computer Science and Engineering Departments with European and International guidelines.
It is also a contribution to the exploration of the dialogue between research and practice in university education settings. The author speaks with involvement and from experience. An action research is applied to promote reflective practice, methodological innovation, understanding of academic and professional conventions, curriculum development/modification, institutional change and academic/professional development through skills acquisition.

[1]Dissertation Advisor: Konstantinos Halatsis, Professor Emeritus

The motivation and background for this research is to bridge the academic/professional competency gap between the ability of computer science/engineering undergraduates and the current and future needs of academic study and the computer science/engineering profession.
Despite the official guidelines for US and Europe concerning academic/research skills, few degree programmes provide explicit help for students who do not already possess such assumed skills [1]. The problem is more intense in Greece with traditional lecture-based instruction and traditional end-of-semester written assessment prevailing across the curriculum eliminating thus chances of making provision for generic study/research skills, let alone specific key skills, tailored to the students' academic and future professional needs.
In this framework, we have identified as a priority this imperative need and have been inspired to take up Action Research to design a course to face the shortfall for current generation of freshmen (which is severe in Greek Engineering Departments) by shifting towards academic, discipline specific literacy, professional skills and transferable competencies and integrating them in the curriculum. This novel contextualised course which fuses essential disciplinary skills and knowledge and exploits current pedagogy and a suitable methodology will help students survive throughout their studies and compete successfully in the global workplace afterwards.
Similar research efforts devoted to this end have been reported at an international level. Unfortunately, in Computer Science departments' curricula in Greece, dedicated, specially tailored "academic skills" modules are out of the question. This leads to deficiencies that impede academic advancement and integrity as well as student socialization and satisfaction and will be discussed extensively further down.
In response to voices of concern about attrition rates, the increased demands to develop academic and professional skills before leaving university setting [2] and to the alarm by "the plagiarism plague" [3] along with the multiple deficiencies detected in contextualised writing across the Greek student body, this action research has been put forth. The result is an interdisciplinary effort that has been taken in developing an integrated course for a one-semester course for Computer Science and Telecommunication freshmen student to address deficits. Identifying the discrepancy between 21st century needs and actual instruction provided, helps us raise awareness and recognition of the value of these skills. A review of best instructional/methodological practices recommended and applied as suitable at both the level of delivering CS core/content modules and the level of teaching academic skills and literacy in Computer Science departments worldwide has made us contemplate on strong and weak points, design and apply an innovative integrated2 approach which is the case study presented in this work. In this course, groups of students go through the various steps to carry out an original research project as a vehicle to acquire specific needed skills and learn professional writing conventions common to research papers and academic life in their scientific field.

2 Literature review

In this section, a consideration of what the targets of Computer Science and Engineering Education should be and an examination of the concepts, learning theories and instructional methodologies which have been developed and applied to cater the needs of this discipline are presented. For a course to fulfil its aims, we need to know how university students learn, to understand barriers to students' learning, and to develop modern, tailored, in and out-of-classroom techniques that promote content learning and academic soft skills acquisition among students in Computer Science and Telecommunications departments [4].

Reports from worldwide experience present that a shift to skills is increasingly recognized and appreciated in Higher Education and the business arena. Modern education does not only involve knowledge, as in traditional education, but it also entails what the students are able to do with this knowledge (and how) [4]. The measurement of competences or learning outcomes should be an important consideration in the evaluation of an educational method or a course/degree [5].

A review on international literature on skills particularly required in Computer Science reflects that numerous voices coming from faculty staff, curriculum designers and potential employers emphasize the inadequacy of support for undergraduate and post-graduate students concerning soft or behavioural skills such as analytical skills, communication skills, teamwork, academic writing, critical thinking, problem solving, dealing with written and spoken academic texts, understanding use of technical literature and other scholarly information sources, self-learning and suggest shift and adaptation [6][7][8][9][10][11][12][2][13], [14–16]. Professional communication is an important aspect in the development of effective CS and engineering graduates.

As Computer Science is a rapidly changing field, students should also be prepared to survive, compete and succeed in this changing environment by providing them curricula that address lifelong learning and include

professional practice as components of the undergraduate experience. Training students to integrate theory and practice is a key skill in CS education [17]. Inevitably, this has raised awareness to science faculty across the country to transform curricula and their instructional approaches they select for their undergraduate classrooms.

Enrollment and retention of students in computing departments/courses is another concern as enrollment numbers seem to be declining and students drop out of these programmes [18]. Successful academic performance is important to retention and it requires more than innate intelligence. Studies provide evidence that skills and intrapersonal factors are essential in meeting the challenges of demanding computing curricula [19].

Researchers have also recognised the need for cognitive skills for successful computing professionals [20–22] such as creativity, self-efficacy, high-emotional intelligence.

It therefore becomes evident that there is a worldwide concern/call to apply an educational philosophy that promotes the development of modes (or habits) of learning that are skills oriented and produces students with the right mix of technical and soft skills [23] to meet both national expectations and international exigencies.

Since going through the steps of project completion cultivates higher order thinking skills [24] and research is becoming a collective enterprise in industry and academia, Project-based instruction is undoubtedly a suitable and highly effective model for the CS context. Besides, it is quite probable that computer science graduates will find themselves working as members of a research team during their academic or professional life [7] therefore having engaged in research and team work previously will prove fruitful.

3 Research methods

For this study, various instruments were employed to gather data in two stages of this action research process which are described underneath.

3.1 Research methods- Diagnostic stage

To identify and define the problem a four-fold preliminary research was carried out. Most parts of it are the results of preliminary research performed through questionnaires while one is secondary research:

Instrument 1:

A questionnaire (Part A) which was completed by freshers to examine learning styles, habits, strategies, needs, factors that facilitate learning, preconceptions, misconceptions, existing skills, attitudes, anticipations. This questionnaire was anonymous to ensure that students feel free to express their deficiencies, problems, difficulties without fear of being looked down or considered inferior, in case of low-performance or low-achievement students.

Instrument 2: A pretest (Part B, followed by a post test at the end of the semester) which attempts to investigate prior knowledge useful for the level of education they are just entering and elicit prior acquaintance and experience of academic skills. This will be used as a control.

Instrument 3: A survey (secondary research) was carried out to discover the balance between hard and soft skills courses offered in the various Greek computer Science departments' curricula.

Instrument 4: discussion groups with faculty members to investigate needs that they have identified over their years in the departments through teaching and interacting with Computer Science and Telecommunications students.

3.2 Research methods- Main research (Therapeutic stage)

In order to evaluate and measure students' satisfaction and effectiveness and identify strengths, limitations and possible areas for improvement of this experimental intervention reported on here, various metrics have been gathered and considered for the analysis presented in this study, such as project evaluation, students' satisfaction, students' performance in the posttest.

For collection and reaching of the types of data mentioned above, the following 4 strategies (instruments) have been employed:

Instrument 1: posttest data, that is the second part of the questionnaires distributed in the preliminary diagnostic research which was once again filled in by the same students on completion of the course, after taking up the new content and mode of instruction, to measure students' improvement through comparing/contrasting the pretest with the posttest.

Instrument2: a criterion-referenced test: evaluation of students' projects submitted at the end of the semester as a culminating activity of all efforts made through this 13-week endeavour. Students' projects were marked according to rubrics.Projects were obviously not anonymous since students had to be informed about their grades/performance and provided with personalised feedback (at the level of groups for group work).

Instrument 3: students' self-report. This sort of data was collected through the completion of a questionnaire that was developed and administered at the end of the semester, targeted to investigate students' satisfaction of the experiential design, process, material, blended methodology, skills acquisition and perceptions of classroom environment. Anonymity was the only option for this student survey instrument in order to give participants the freedom to evaluate the instructor and her choice of methodology in a sincere, fair and objective way.

Instrument 4: discussion with students, field notes and observations

4 Preliminary/diagnostic pilot research

The first stage of the Action Research, namely the diagnosis or problem awareness [25] is described and analysed in this chapter. This stage has shed light to the instructional/learning methodology problems and students' deficiencies, helped us formulate the subsequent research questions and provided insight about the action that needs to be taken. The preliminary research pertinent to the main research, was carried out through a questionnaire and a pretest (in the form of questionnaires that have been distributed to students of the undergraduate programme of the department of Computing and the department of Telecommunications, University of Peloponnese), as well as surveys, in order to locate problems regarding academic/research skills within the student cohort of Greek computing education, select the appropriate paradigm and design the course accordingly. The areas that we have focused on in this initial diagnostic study are:

1. The level of exposure (if any) to research skills in formal instruction (while at school and while at university) and the level of students' linguistic skills in the English Language
2. The level of plagiarism owing to the absence of research skills training
3. Computing departments' curricula analysis aiming to reveal the proportion of soft skills.
4. Students' perception about academic skills and their lack of proficiency in this area

The data collected through the four research instrumentsin the framework of the pilot survey (diagnostic stage) and their analysis are presented and discussed. This survey reveals the needs of undergraduate students studying in a computer science department in a Greek university regarding academic/research skills and leads us to decide on the appropriate principles that should pertain our course design and implementation in our setting.

It becomes evident that HE Institutions in Greece do not run skills courses or workshops either specifically designed for the department's content embedded in the curriculum or outside the undergraduate programme (generic).

Therefore, computing and engineering students do not receive formal instruction and training in academic literacy and professional communication at any stage of either their formal compulsory education or in the course of their degree - not even in their native language.

Student pilot questionnaires to investigate their prior knowledge and perceptions on issues involving academic/research skills along with a survey in the curricula of CS/Engineering Departments in Greece and in the level of plagiarism that is thriving and discussions with faculty members verify the deficiencies and guide us to the formation of the suitable research questions and selection of the appropriate way to deal with the shortfall.

In this empirical research, through an innovative programme that is designed and applied, the following **research questions** are attempted to be addressed:

1. Student satisfaction

a. How were the students accepting the particular learning model which combined thoughtful integration of collaborative learning, online and face-to-face interactions as well as project-based learning?

b. Why are students willing to attend, engage and participate in the course? Do they identify its beneficial effects?

c. What is the students' opinion of the learning experience, the problem-based methodology/process, the challenges posed and its effectiveness?

2. Can this course support learning and skills development? What does it provide to students?

5 Action Planning- Theoretical Course Design Considerations

The action planning that has been developed based on the findings of the diagnostic stage. It involves the intervention that has been planned/designed and its specifics in terms of an eclectic methodology. The diagnostic procedureshave revealed that new forms of methodologies need to be explored and new models of participant behaviour need to be considered in an effort to understand and cope with the problems and deficiencies. This is the stage to select which theory to put into practice, thus a consideration of the concepts and theories which suitably underlie the instructional methodology and teaching/learning procedure to be used in the exploratory stage will be made.

This integrated approach which we have termed as "Research Project-Based Approach" encapsulates the most effective elements of four currently acceptable theoretical and methodological bases that have been reported in literature and applied in classroom settings. The four pillars that determine the specifics of our experiment are: Genre-based approach, Project-based learning, Content-based learning and the Revised Bloom's Taxonomy. These four theories provide the foundations for the design and application of the experiment that will be conducted and will enable the researcher to test his hypothesis by reaching valid conclusions.These approaches are the most suitable considering the challenges and constraints posed by the framework of the Greek Computer Science and Engineering departments and actual students' needs as detected through our diagnostic survey by questionnaires.

6 Overview of the Experimental Procedure- Application & Technics

In this chapter, the 4th stage of the Action research, namely the therapeutic stage in which the hypotheses are tested by a consciously directed intervention or experiment in situ, is described.

This Action stage or experimental manipulation analyses the intervention that has been planned and performed in the form of a novel venture at the department of Computer Science and the department of Telecommunications, University of Peloponnese, Greece. The purpose is to provide students with academic literacy support and skills (the language of delivery being English) in order to treat the deficiencies and cope with the problems and inconsistencies with computer science education essentials as advocated by specialists and diagnosed by our preliminary research.

This section describes the activities/deliverables completed by students while taking the course and carrying out their projects in a collaborative way exploring the importance of establishing student-to-student networks (collaboration) in which they become active and willing participants. This stage can be seen as a clear description of their own transformative pathway towards skills acquisition and computer science and telecommunications content knowledge and provides useful insight for instructors that wish to adopt this methodology or enrich their teaching portfolio with new learning tools [26]. Of course, this approach requires certain degree of flexibility and dynamism from all involved [27], as there are demanding areas entailed such as defining an angle of a broad topic to form challenging project titles, laying out the process and steps of how learners should go about developing a project and presentation, time and in-group management, avoiding plagiarism and so on.

The new course was introduced in the first semester of both the department of Computer Science and Technology and the department of Telecommunications Science and Technology of the University of Peloponnese, Greece. All students had to take this obligatory credit-bearing course on "Academic/Research Skills", however, attendance was optional, as is the case with most non-laboratory based courses in Greek Higher Education. Students are invited to attend a 3-hour session per week, which is of a hybrid form in the sense that half of it comprises a core session in the lecture hall and, the remaining, of a lab session in the computer laboratory, and to engage in self-study, self-exploration and team.

Aims

The aim of both modes of instruction, namely core classroom sessions and computer lab workshops, was to help students develop two types of skills:

a. Specialised skills, tailored to their field of study such as researching a CS/Engineering topic, retrieving appropriate, scholarly resources, understanding specialised vocabulary, writing up an academic project, developing oral communication skills

b. Transversal, cross-curricular competences, such as group work, autonomy, responsibility, accountability, self-study, collaboration

7 Results and discussion

7.1 Posttest- Instrument 1

As we can observe after comparing and contrasting students' answers in the initial and final test, we can conclude that a significant improvement has been performed. Students have started to abandon their High School habits and familiarise themselves with academic texts and the new genre along with the conventions and ethics that apply. The new course therefore has fulfilled its purposes as the benefits that students have reaped are evident.

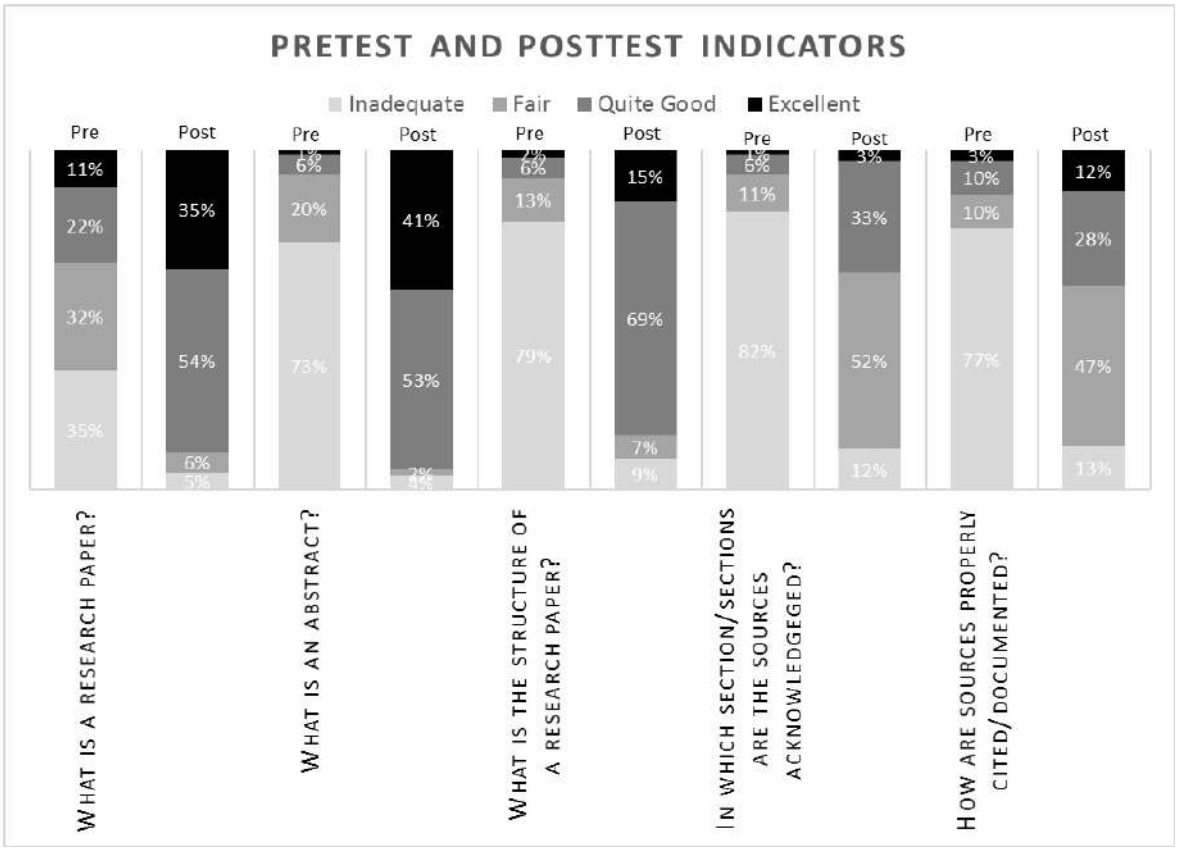


Fig. 1. Pretest and Posttest Performance Indicators

7.2 Criterion-referenced test (Project Evaluation)- Instrument 2

This study examined the results of the projects based on survey findings from projects of 72students of two departments. The results showed that the projects were successful in fulfilling the set requirements and in promoting students' academic engagement. The majority of students performing averagely in most criteria and above averagely in some other, is a significant predictor of academic progress. These findings suggest that study projects can potentially contribute to improving Computer Science higher education as they fulfil students' needs for competence, relatedness and autonomy and enhance students' academic engagement.

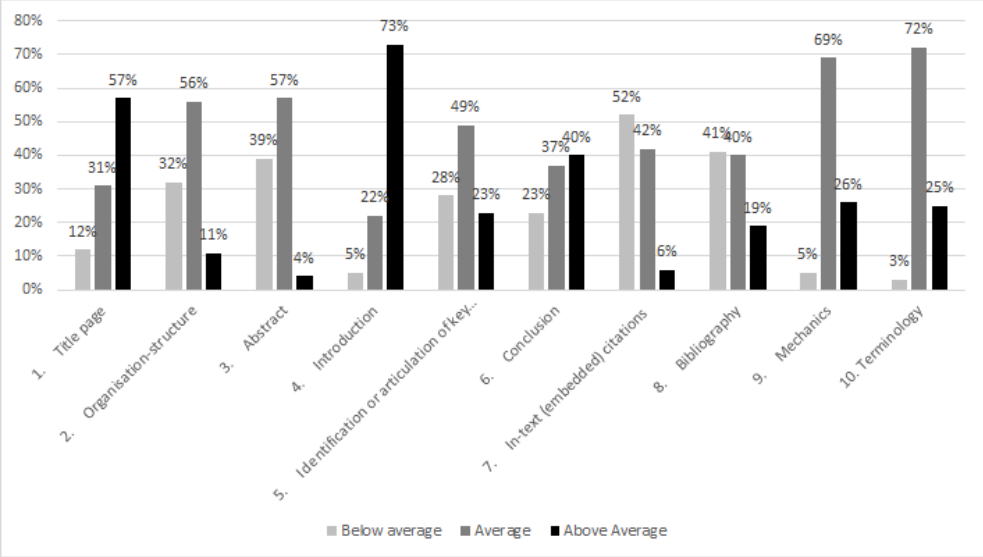


Fig. 2. Student performance towards the Academic/Research project requirements

7.3 **Student satisfaction- Course experience questionnaire - Instrument 3**

Finally, in the overall evaluation, this new learning methodology has been rated very highly by students. They appreciate the novelty-based motivation and challenges posed by this web-enhanced learning environment and would like to continue with this approach in other courses in the future. They liked the roles they assumed: a) as researchers brainstorming and specifying research questions, locating and evaluating resources and discovering information themselves b) as collaborators in the group allocating tasks, sharing information and the learning experience c) as authors producing an academic/research project about a state-of-the-art technology topic and d) as presenters performing in front of their peers for the first time.

Table 1: Aspects of the methodology that the students appreciated the most

Reason	Frequency (%)
1. The contextualized (genre-based) nature and Interdisciplinary mode	82%
2. Collaboration, group work	73%
3. The acquisition of knowledge and skills that will prove useful and transferable throughout their studies and professional life	57%
4. The fact that they were active learners and were invited to learn by doing	53%
5. Socialization with other students and smooth transition to the new academic environment	51%
6. The exposure to and exploitation of resources through the Internet and in the department library	50%
7. Innovation, originality	47%
8. The research, exploration and hands-on nature in the activities	45%
9. The fact that they learn academic/professional English discourse without realising it	43%
10. Having to study bibliography (internet sites, journal papers, books)	37%
11. The fact that they were encouraged to orally present their written project in front of audience (their peers) using IT (PowerPoint)	23%

7.4 **Summary of findings**

This content/project-based methodology can be considered an innovation in the curricula of Computer Science departments in Higher Education as it introduces a significant shift not only in the mode but also in the context of learning.

The empirical research findings, when combined with the practical advantages of integrating content and purpose-specific communication skills, provide persuasive arguments in favour of content & project-based instruction. The benefits for students are traceable in the instruments analysed about and are summarised in the following features:

1. project work focuses on content learning, in order to stimulate learners, enhance active engagement and responsibility and develop a sense of ownership in the process, by either selecting a suggested topic or an original one of their own, deriving from their discipline.
2. though the teacher plays a major role in offering support, guidance and feedback at critical moments in the process either during face-to-face instruction or on-line, this is a learner-centred approach because learners are engaged into active exploration, research, problem-solving, self-management and responsibility. Giving students freedom to immerse themselves in the project actively seeking information, relevant to the topic they are committed to examine, can lead to motivated and independent learners. The instructor inevitably relaxes control of the learners and assumes the role of guide or facilitator [28] as expected in the case of the interdisciplinary approach to integration [29].
3. eClass, and the Web in general, create a more vibrant environment for constructing/acquiring academic knowledge and soft skills, since they provide readily accessible variety of content resources. Sharing of information and documents, integration of ideas or information from various sources and collaboration are facilitated and encouraged. The "at- any-time" aspect of the Web is exploited since learners are free to access the platform and take up work at their own pace, anytime, anywhere convenient to them. Timetable difficulties usually prevent university students from attending on-site tuition. The lecture mode is not the unique way to acquire knowledge and cognitive abilities any more. Individualized needs are thus satisfied and certain categories of learners are facilitated.
4. a feedback loop is established between the instructor and the groups of learners (through drafting and redrafting projects and components of it). This feedback cycle may occur several times either in class or through the eClass platform before the final electronic submission of the project work encouraging students to stay current and alert.
5. the "one course for all" policy is abandoned. Uniform level education for everybody proved to be a wrong approach [30]. This project-based instruction applies customisation and user adaptation by making use of a combination of different backgrounds (High-school majors), different cognitive styles, different learning strategies, motivations, capacities, even hobbies and extra-curricular interest of students involved. It cannot be expected to receive final products of same level and uniform quality. Heterogeneity of the end-users -which is common in university classes - is not an obstacle since project-based learning meets the needs of learners with varying skill levels and learning styles [31] and is thus consistent with individualized learning and the socio-constructivist paradigm. After all, one of the important factors here is student engagement in the various steps of the process.
6. collaborative work over the completion of the project stimulates social interaction and contributes to the psychological adjustment of students to college life (enculturation), especially when the course is offered early in the beginning of their studies, as in the case presented here.
7. this approach culminates in an end product (e.g. a research essay and an oral presentation) that can be shared with others, giving the project a real purpose and situating authentic problem solving in an authentic physical and social context. The value of the project, however, lies not just in the final product but also in the process of working collaboratively towards the end point. Thus, research project work has both a process and a product orientation [32]. The atmosphere conducive for doing research is built, which is significant if we are to retain the bright minds.

To conclude, regarding the research questions in this thesis, we can say that students were very positive towards the innovative course as they achieve substantial learning gains in academic skills and professional communication and significant growth toward a more mature understanding of academic life, responsibility, ethics and integrity. The learning objectives were thus accomplished in order for students to reap the benefits and the education they receive be consistent with current requirements, building a rich understanding of themselves and their abilities and increasing their own self-confidence to perform the myriad tasks expected of them by potential employers.

8 Conclusions

As the results of the experiment show promising signs, this research is intended to inform practice, curricula decisions, instructional methodology and materials and contribute significant insights into Computer Science and Engineering education of both English and non-English speaking background computer scientists and engineers in Anglophone and non-Anglophone universities. The insights are relevant, significant, interesting, and have the potential to inspire and impact practice within the wider computer science and engineering education community. The methodology is informed by relevant theory and clearly demonstrates how the problem is approached and the design and application are developed accordingly.

References

1. Skinner, I., Mort, P.: Embedding Academic Literacy Support Within the Electrical Engineering Curriculum: A Case Study. IEEE Trans. Educ. 52, 547–554 (2009).
2. The Boyer Commission: Reinventing Undergraduate Education: Three Years After the Boyer Report, (2001).
3. Campbell, D.: The Plagiarism Plague. Natl. Crosstalk. 14, (2006).
4. Pérez, J., Vizcarro, C., García, J., Bermudez, A., Cobos, R.: Development of procedures to assess problem-solving competence in computing engineering. IEEE Trans. Educ. 60, (2017).
5. Goff, L., Potter, M.., Pierre, E.: Learning Outcomes Assessment. McMaster University Abeer SiddiquiMcMaster University (2015).
6. ABET: Accreditation Board of Engineering and Technology. (2003).
7. Mohan, A., Merle, D., Jackson, C., Lannin, J., Nair, S.S.: Professional Skills in the Engineering Curriculum. IEEE Trans. Educ. (2010).
8. Edgerton, R.: Education White Paper, (1997).
9. Felder, R.M., Brent, R.: Designing and teaching courses to satisfy the ABET engineering criteria. J. Eng. Educ. 92, 7–25 (2003).
10. Hissey, T.W.: Education and careers 2000. Enhanced skills for engineers. Proc. IEEE. 88, 1367–1370 (2000).
11. ACM/IEEE: ACM/IEEE Computer Society Curricula, (2004).
12. Committee on Science, E.: Careers in science and engineering : a student planning guide to grad school and beyond. National Academy Press (1996).
13. Reinventing Undergraduate Education: A Blueprint for America's Research Universities, (1998).
14. The Joint Task Force on Computing Curricula IEEE Computer Society Association for Computing Machinery: Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering. Comput. Curricula Ser. 664–75 (2004).
15. Joint Task Force on Computing Curricula IEEE Computer Society Association for Computing Machinery: Software Engineering 2014: Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering. Jt. Task Force Comput. Curricula IEEE Comput. Soc. Assoc. Comput. Mach. 134 (2014).
16. Jollands, M., Jolly, L., Molyneaux, T.: Project-based learning as a contributing factor to graduates' work readiness. Eur. J. Eng. Educ. 37, 143–154 (2012).
17. Computer Science Curricula 2013, Curriculum Guidelines for Undergraduate Degree Programs in Computer Science. (2013).
18. Matrisciano, A., Belfiore, N.P.: An investigation on Cognitive Styles and multiple intelligences model based learning preferences in a group of students in engineering. In: 2010 9th International Conference on Information Technology Based Higher Education and Training (ITHET). pp. 60–66 (2010).
19. Belanger, F., Lewis, T., Kasper, G.M., Smith, W.J., Harrington, K. V: Are Computing Students Different? An Analysis of Coping Strategies and Emotional Intelligence. IEEE Trans. Educ. 50, 188–196 (2007).
20. Denis, M.S.L., Trauth, E.M., Farwell, D.: Critical Skills and Knowledge Requirements of IS Professionals: A Joint Academic/Industry Investigation. MIS Q. 19, 313–340 (1995).
21. Michael A. Eierman, Hilbert K Schultz: Preparing MIS Students for the Future: A Curriculum. J. Educ. MIS. 3, 5–12 (1995).
22. Weiqi Li, Hanwen Zhang, Ping Li: Assessing the Knowledge Structure of Information Systems Learners in Experience-Based Learning. J. Inf. Syst. Educ. 5, 205–2012 (2004).

23. Chia, R.: The aim of management education: Reflections on Mintzberg's "Managers not MBAs" Organ. Stud. 26, 1090–2 (2005).

24. Johnson, E.B.: Contextual teaching and learning : what it is and why it's here to stay. Corwin Press (2002).

25. Lewin, K.: Group decision and social change. In: Readings in social psychology. pp. 197–211 (1947).

26. Mettetal, G.: Classroom Action Research as Problem-Based Learning. In: Energizing Teacher Education and Professional Development with Problem-Based Learning. pp. 108–120 (2001).

27. Santos, S. dos: PBL-SEE: An Authentic Assessment Model for PBL-Based Software Engineering Education. IEEE Trans. Educ. 60, 120–126 (2017).

28. Stoller, F., L., Sheppard, K.: Guidelines for the Integration of student projects into ESL classrooms. English Teach. Forum. (1995).

29. Drake, S.M., Burns, R.C.: Meeting standards through integrated curriculum. Association for Supervision and Curriculum Development, Alexandria, Va. : (2004).

30. Cristea, A.I., Okamoto, T., Cristea, P.: MyEnglishTeacher-an evolutionary Web-based, multi-agent environment for academic English teaching. In: Proceedings of the 2000 Congress on Evolutionary Computation. CEC00 (Cat. No.00TH8512). pp. 1345–1353. IEEE.

31. Project Based Learning | BIE.

32. Stoller, F.L.: Project Work: A Means to Promote Language Content. English Teach. Forum. 35, 2–19 (1997).

List of publications

1. Mamakou, I., M Grigoriadou and C. Halatsis, (2017) Developing Academic Research Skills in the Engineering/Computer Science Curriculum through Project–Based Learning under revisions in *European Journal of Engineering Education*.
2. Mamakou, I. & M. Grigoriadou (2010) An e-project-based approach to ESP learning in an ICT curriculum in higher education, *Themes in Science and Technology Education*, Vol 3 (No 1-2), p. 119-137
3. Mamakou, I. & M. Grigoriadou (2008), Project-Based Instruction for English in Higher Education in Marriott, R. & P. Torres (eds) *The Handbook of Research on E-Learning Methodologies for Language Acquisition*, IGI Global
4. Mamakou, I., C. Halatsis& M. Grigoriadou (2017, June) Modernising Instructional Methodologies in Computer and Telecomminication Engineering Education in *The Proceeding of the 27th EAEEIE (European Association for Education in Electrical and Information Engineering) Annual Conference*, Grenoble, available at IEEExplore
5. Mamakou, I., C. Halatsis& M. Grigoriadou (2017, June) Curricula Review in Computer and Telecommunication Engineering Education in Greece in *The Proceeding of the 27th EAEEIE (European Association for Education in Electrical and Information Engineering) Annual Conference,* Grenoble, available at IEEExplore
6. Μαμάκου, Ε. & Κ. Χαλάτσης (2010, Απρίλιος) Πρόταση Διδασκαλίας της Αγγλικής Γλώσσας στα τμήματα Πληροφορικής και Τηλεπικοινωνιών- Στόχοι, Μεθοδολογία και Περιεχόμενο. *5ο Πανελλήνιο Συνέδριο Διδακτική της Πληροφορικής*, Αθήνα, σελ. 216-223.
7. Mamakou, I. & M. Grigoriadou (2009, September) An E-Project-Based approach to ESP learning in an ICT Curriculum in Higher Education. In *5th International ICTATLL Workshop 2009*, Korinth, Greece

In-network processing based hardware acceleration for situational awareness

Md Fasiul Alam[1]

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

Abstract. In this research, an In-network processing (INP) based computational framework has been proposed that gradually refine the captured information while moving it upstream to the application. The refinement can go up to the point of knowledge extraction from the primitive or even fused data. It shows that the demanding tasks that previously were simply undertaken on the fixed infrastructure are now possible on the mobile end. It relies on a sophisticated in-network processing scheme that gradually refines the information captured by sensing elements to the level of application-exploitable knowledge. With this process, the scaling problem is tackled much more efficiently through a problem segmentation and exploitation of physics-based and human-based sources. The components executing on INP nodes are structured appropriately to delegate the demanding subtasks to some onboard accelerator. The core program executes on the central processing unit and exploits the particular characteristics of the side processor. Special system-on-chip (SoC) hardware for computational acceleration like central processor unit (CPU), digital signal processor (DSP) or field-programmable gate arrays (FPGAs) are desirable. These accelerated hardware supports software defined process on-chip memory that expedites all the advanced processing with a minimum energy overhead. In-network software defined processing (ISDP) capabilities has been considered that allows dynamic reconfiguration of the network topology. It is advantageous for the network to possess reliable and complete end-to-end network connectivity; however, even when the network is not fully connected, the system may act as conduits of information — either by connectivity gaps, or by distributing information from the network space.

Keywords: In-network Processing, Hardware acceleration, low power process, WSN, IoT, Situation awareness

[1]Desertion advisor: Stathes Hadjiefthymiades, Professor

1 Introduction

Due to the rapid development and spread of embedded computer technology over the last decade [1], sensor nodes are widely used in situational awareness and produce potentially abundant information for IoT applications. However, the disconnected, intermittent and limited communication environment that these nodes often operate in make the usage of internet-level communication solutions not applicable on the WSN level [2-4]. The direct consequence is the increase (already in place) of the amount of big data to be analyzed, which in the industrial field translates in the need to equip itself with platforms of data storage of enormous proportions [5]. The exponential increase of the data to be analyzed leads to the need for adopt new ways to provide answers immediate and reliable applications to high degree of criticality, which do not tolerate latencies in communication. A growing number of contemporary IoT applications require more from WSNs than simple data acquisition, (conditional) communication and collection to databases. For example, maintenance in nuclear applications, such as ISR systems, aim to improve the situation awareness of decision makers and expect preprocessed data that are already converted to human understandable form.

Data collection to central databases, as used in typical WSNs for monitoring, creates overhead, potential bottlenecks and offers limited resilience [6]. An additional aspect is that some of the potential WSN users, such as in-the-field extreme environment maintenance (i.e ATLAS, CERN), need situational information in a timely manner and would prefer to receive information tailored to their current information needs directly from the network to minimize delays and dependence on central infrastructure. In order to manage the large data flows and to minimize the bandwidth requirements, novel paradigms such as D2D and mist computing (an extension of fog computing) need to be exploited. Traditional data aggregation is a predecessor of these paradigms, but in its classical form is not enough for IoT applications, because the traditional flow of data from network edge (sensor nodes) to center (databases) remains.

In general, performing intelligence operations inside the network, such as eliminating irrelevant records and aggregating raw data, can reduce energy consumption and improve sensor network lifetime significantly [7]. This is referred to as sensor data processing, in which an intermediate proxy node is chosen to house the data transformation function to consolidate the sensor data streams from the data source nodes, before forwarding the processed stream to the sink. Sensor based IoT nodes sense, receive, process and transmit data to other nodes. In these functions, a node may exhibit different degrees of intelligence and sophistication. Some nodes involve little processing and transmit small quantities of information. Other are connected to sensors that generate large quantities of data (e.g. cameras), require large storage, high processing power and may transmit many data. Other types of nodes (i.e., knowledge discovery, consensus, reasoning or fusion) also exhibit varying needs. Consequently, node capabilities vary from highly restricted resources in terms of processing, storage, transmission bandwidth and energy to devices that compare to current smart phones in terms of resources. In-network processing is a technique employed in sensor database systems whereby the

data recorded is processed by the sensor nodes themselves. This is in contrast to the standard approach, which demands that data is routed to a so-called sink computer located outside the sensor network for processing. In-network processing is critical for IoT based sensor nodes because they are highly resource constrained, in particular in terms of battery power and this approach can extend their useful life quite considerably. Based on their capabilities, the nodes are assigned a specific role/type in order to achieve a certain task. Different nodes can cooperate in the execution of some workflow in order to achieve some goal/task also based on the semantics of the processed data. For instance, a set of nodes can participate in the execution of a workflow, which implements a high level fusion operator or, even, a distributed classification algorithm. The node can be envisaged as the basic INP units through which collaborative intelligence can be attained. Energy consumption is a crucial factor in a sensor network. INP is a useful technique to reduce the energy consumption significantly. Many different processing modules within the IoT infrastructure that can gradually refine the captured information while moving it upstream to the application. The refinement can go up to the point of knowledge extraction from the primitive or even fused IoT data. The components executing on such nodes (Fig. 1) are structured appropriately to delegate the demanding subtasks to some onboard accelerator (e.g., DSP, GPU). The core program executes on the central processing unit and exploits the particular characteristics of the side processor. Energy is the dominant constraint. Because the quantities of information to process are much larger and the processing algorithms are much heavier, nodes with higher processing capabilities are needed. Special hardware for computational acceleration like DSP or FPGAs is desirable.

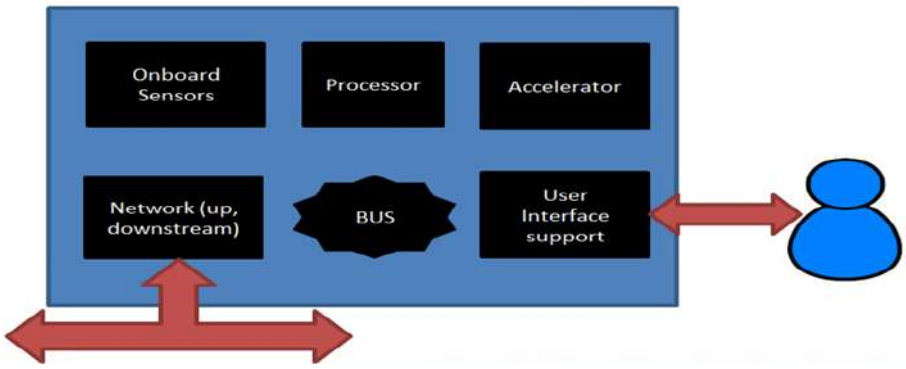


Fig. 1. INP Node

2 In-network Processing Architecture

In-network processing architecture usually involves information filtering/transformation nodes that rely on computationally demanding algorithms. Such nodes are based on accelerated, not conventional hardware that expedites all the advanced processing with a minimum energy overhead. This is imperative for the efficient operation of the

WSN as it addresses two important needs: performing demanding tasks at low energy cost, filtering the information to support energy efficiency and render the network sustainable over time. Different schemes for the processing of IoT generated data have been investigated in this chapter. The objective is to reduce the application processing needs in the discussed domains and drastically reduce the volume of data seen by the application. Currently, applications collect huge volumes of IoT data in their support databases for further processing in line with specific business logic. Many different processing modules within the IoT infrastructure have been orchestrated that gradually refine the captured information while moving it upstream to the application. The refinement can go up to the point of knowledge extraction from the primitive or even fused IoT data. With this scheme, the originally identified scaling problem is tackled much more efficiently through a problem segmentation and exploitation of in-network processing. The components executing on such nodes are structured appropriately so as to delegate the demanding subtasks to some onboard accelerator (e.g., DSP, GPU). The core program executes on the central processing unit and exploits the particular characteristics of the side processor. Energy is the dominant constraint. Because the quantities of information to process are much larger and the processing algorithms are much heavier, nodes with higher processing capabilities are needed. Special hardware for computational acceleration like digital signal processor (DSP) or field-programmable gate arrays (FPGAs) is desirable. As these nodes must be able to operate on batteries for relatively long periods devices that resemble modern smart phones may satisfy their requirements. IoT nodes sense, receive, process and transmit data to other nodes. In these functions, a node may exhibit different degrees of intelligence and sophistication. Some nodes involve little processing and transmit small quantities of information. Other are connected to sensors that generate large quantities of data (e.g. cameras), require large storage, high processing power and may transmit a lot of data. Other types of nodes (i.e., knowledge discovery, consensus, reasoning or fusion) also exhibit varying needs. Therefore, node capabilities vary from highly restricted resources in terms of processing, storage, transmission bandwidth and energy to devices that compare to current smart phones in terms of resources. Based on their capabilities, the nodes are assigned a specific role/type in order to achieve a certain task. Different nodes can cooperate in the execution of some workflow in order to achieve some goal/task also based on the semantics of the processed data. For instance, a set of nodes can participate in the execution of a workflow, which implements a high level fusion operator or, even, a distributed classification algorithm. The node can be envisaged as the basic INP units through which collaborative intelligence can be attained.

Nodes can be categorized into the following sub-types subject to their capability:

- Sensing (S) node,
- Fusion (F) node,
- Consensus (C) node,
- Knowledge extraction (K) node,
- Sink (M) node.

Nodes can participate in the execution of some workflow in order to achieve some task/goal. The set of nodes that participates in the execution of some workflow can be also viewed as a composite node with certain input, output, and capability. A composite node can be further aggregated with other composite and/or basic nodes in order to construct an even more complex node, forming tree like structures. Based on such constructors of nodes, the main network of nodes and an overall view of the architecture are depicted in Fig. 2. Fig. 2 shows the role hierarchy in the user (data) plane. The several types of node are explained and analyzed below.

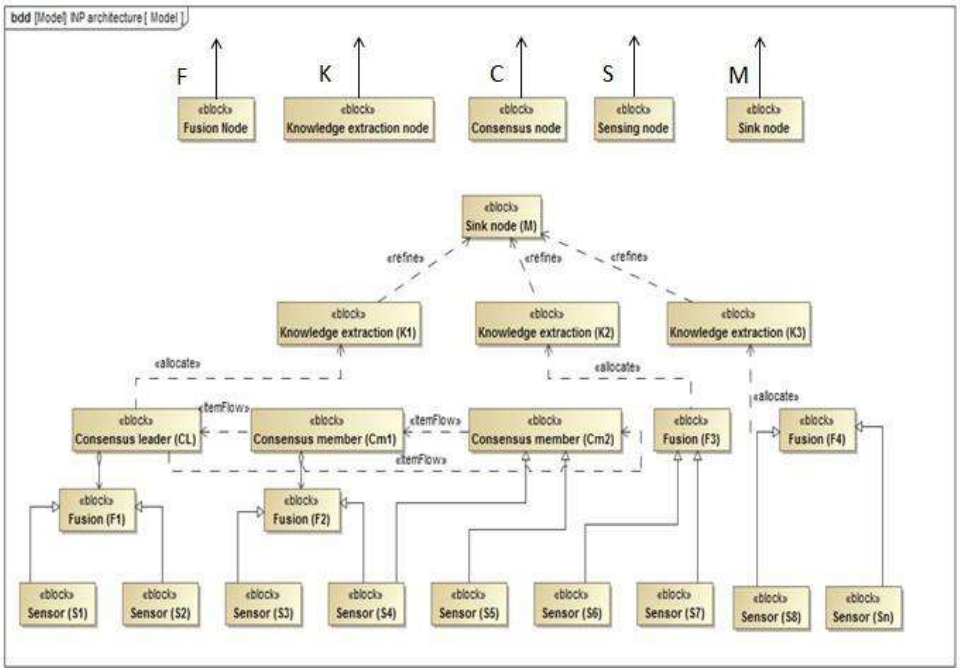
2.1 Sensing Node ('S' node)

The sensing node captures data from the environment and propagates this information to the network (either to 'F' or 'C' nodes). The processing capabilities of the sensing node can be limited since its main task is to sense and relay pieces of data to the upstream node. Additionally, several alternate routing schemes could be investigated as an add-on feature to optimize the overall network performance and avoid redundant retransmissions. The 'S' node, and each node that relays information, apart from processing, can adapt its relay/data dissemination mechanism. For instance, the 'S' node can adapt the key operational parameters of an epidemic-based information dissemination scheme (i.e., the forwarding probability and validity period). The sensing node can be embedded to various sensing devices, thus, a sensing device consists of numerous sensing 'S' nodes. The sensing device can, then, be abstracted as a composite 'S' node. For instance, numerous vision sensors can be implemented as an 'S' node with onboard cameras and DSP facilities. A vision sensor produces readings (not video) for image segments (tiles). An indicative example involves fire detection through a vision sensor that segments images into 16 x 16 tiles. Each tile is handled independently within the camera sensor network (and the originating node of course).

2.2 Fusion Node ('F' Node)

The fusion node collects data from heterogeneous sources ('S' nodes or other 'F' nodes) and performs fusion operations in order to deduce more accurate information. The heterogeneity of the received information is bridged by the data semantics profile of each 'S' and 'F' node. That is, the 'F' node is aware of the different type of pieces of data that is responsible to apply fusion operators. Obviously, conventional aggregator operators (e.g., statistical mean, min/max operators) can be applied on data of same type. However, the 'F' node can apply intelligent information fusion techniques (e.g., theory of evidence) on pieces of data of different types. The 'F' node is capable of fusing based on certain quality/validity metrics. The deduced information is provided as input either to 'K' nodes or to other 'F' nodes. In the latter case, the output of a first-level fusion process feeds a second-level fusion (multi-level fusion) in order to conclude to information with higher degree of confidence. That is, a composite 'F' node represents a two-level fusion scheme, which consists of a set of 'F' nodes that fuse data from diverse data types. For instance, consider an 'F' node which produces probability of a fire event by fusing the results of (a) 'F' nodes, which aggregate temperature values, (b) of 'F'

nodes, which aggregate humidity values, and (c) of ‘F’ nodes which produce probabil-



ity of smoke and fire detection. One of the most obvious examples is receiving the input of a number of ‘S’ nodes and fusing them using some unsupervised learning method, e.g. PCA, k-means, etc., for dimensionality reduction.

Fig. 2. INP Architectures (Roles, Connections)

2.3 Consensus Node (‘C’ node)

Consider a neighborhood of ‘S’ and ‘F’ nodes, which is spatially defined. The ‘S’ nodes monitor contextual parameters and the ‘F’ nodes aggregate the corresponding measurements to compute an estimate in a completely distributed way. However, in order to deliver the locally measured data to a common (composite) ‘F’ node, the ‘S’ and ‘F’ nodes exchange their data by performing pair-wise aggregator operations (e.g., average), thus, converging to a common value for such nodes; then consensus is reached. The local estimate of the neighborhood is recorded at each participant node and, thus, can be recovered from any ‘surviving’ node in the neighborhood. Such neighboring nodes process and store information locally, as typical ‘S’ and ‘F’ nodes, but they behave as a single unit, the so called ‘C’ node. The nodes of a consensus group (neighborhood) try to harmonize possible inconsistencies in the received information so the whole group to conclude to a shared and commonly accepted measurement and/or knowledge. In Fig. 2, CM nodes are just members (nodes) of a consensus group that

communicate to each other while node CL is the leader of the group that represents the whole group in the network (e.g., it exchanges information with external nodes).

2.4 Knowledge Extraction Node ('K' node)

The 'K' node performs machine learning and data mining operations to extract new knowledge, such as classification models, frequent patterns, novelty and outlier detection, from the different data sources generated from the network nodes. A 'K' node can use as input the output generated by different node types, 'F', 'C' and even 'K'. In addition a 'K' node can also coordinate the distributed execution of data mining and machine learning operators over the different nodes of the network, in order to reduce the communication load, for example either by performing local dimensionality reduction via the 'F' nodes, or by requesting the generation of local models, and subsequently have the parameters/outcome of these models communicated instead of the actual data.

2.5 Sink Node (M)

The main role of the sink is to conceal IoT heterogeneity in terms of sensors, actuators, networking, or middleware through an interconnection unit. The sink node (M) concentrates the data stemming from the underlying IoT devices and performs the operations needed before feeding the applications with the requested information. The "M" node can be considered as a mediator between the underlying IoT and the application layer. The applications are fed with information by the network and are agnostic to the operation of the underlying IoT. A middleware capable of supporting intelligent pervasive applications can materialize this layer.

3 Energy Efficient Parallel Processing

This section describes the acceleration of in-network operations by means of energy efficient hardware. The adoption of a scheme that relies heavily on INP renders the architecture highly efficient, long lasting and drastically reduces the data processing load that the (sensor-supported) application or middleware should sustain. Even though such IoT architectures with increased INP characteristics can be designed and operated over conventional IoT hardware (e.g., motes with conventional processing elements) the use of accelerated nodes would further increase efficiency, thus, boosting the benefits of INP. The approach to follow in such scenarios is to identify the merits of the accelerated hardware that is currently available for the IoT deployment and contrast these to the peculiarities of the algorithms that are foreseen in the INP architecture. In this chapter, Parallella platform [8], which facilitates the energy efficient parallelization of tasks within resource-constrained nodes, have been focused. Here the important aspect is parallelization. Hence, the idea is to identify all the roles/algorithms that feature internal operations/tasks with clear parallelization needs/capabilities (e.g., operations on matrices). Therefore, the structure on the INP building blocks that are assuming is the one shown in Fig. 3. The algorithm is fed with several sources of data that, generally, need to undergo some preprocessing phase. The general-purpose processor should perform this data-preprocessing phase together with coordination tasks, which is part

of the accelerated hardware. The non-parallelizable task may also orchestrate the delivery of data to the memories of the parallel processing element. Then, the parallelized tasks are invoked followed by the collection/consolidation of results and execution resumes at the general-purpose processor.

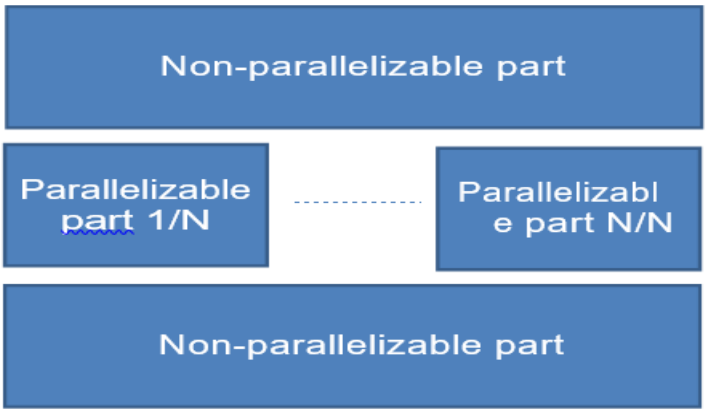


Fig. 3. Parallelized INP component

The algorithms that were introduced in INP integrated parts that can be efficiently handled in parallel. Typical examples are the PCA technique for dimensionality reduction and data compression, the multi-dimensional event detection and a data aggregation scheme that relies on FFT.

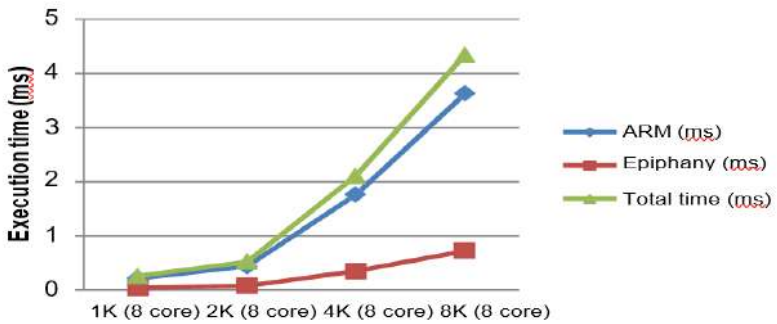


Fig. 4. FFT Turnaround times on Parallella (ARM Epiphany-ARM)

Fig. 4 shows the turnaround time (TAT) in the ARM-Epiphany combination for demanding FFT tasks. Specifically, the plot shows how TAT scales for different FFT sizes. FFT is invoked on the 4 and 8 cores processor in this benchmark but can be invoked in parallel on all 16 cores. Therefore, the Epiphany can be called to FFT 1024

samples (1K) for 16 streams in parallel. From the obtained results it is clear that times overlap and TAT minimally impacted as the number of streams increases from 1 to 16. Experimental results demonstrate up to 60% and 64% reduction in latency for GPU-to-GPU and CPU- to-GPU point to- point communications, respectively.

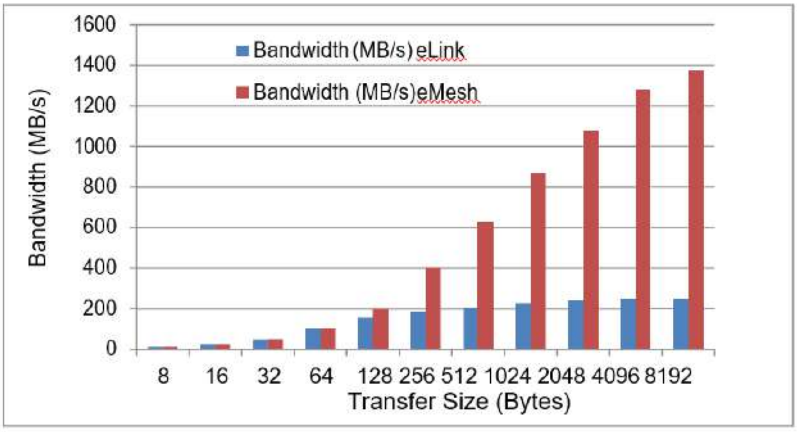


Fig. 5. Data transfer size over bandwidth

The results for the DMA and direct-write transfer size measurements are shown in Fig. 5. First the DMA transfer function provided by the SDK library was tested over the eMesh and eLink for different transfer sizes, capturing both the total transfer time and start-up time. Here we see a very large portion of the time spent sending data can be attributed to starting up the DMA engine (65.2% to 6.9%) in Fig. 6.

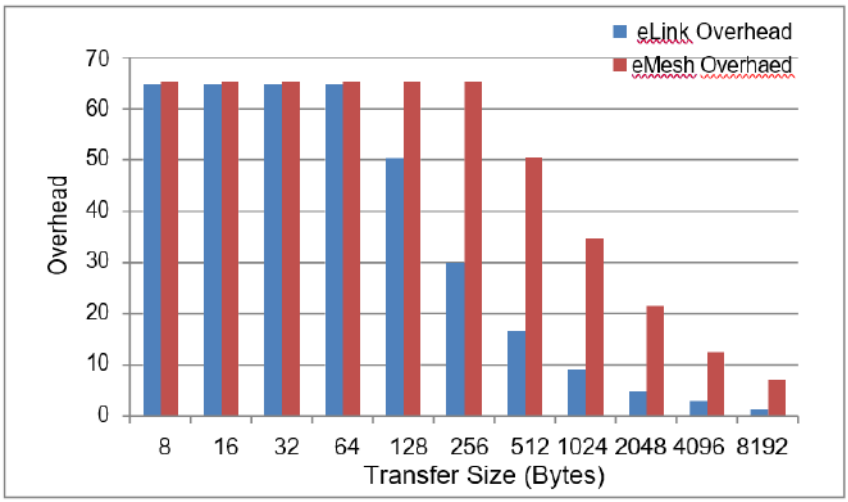


Fig. 6. DMA transfer size versus overhead

Several nodes within the IoT are implemented through accelerated hardware which

(a) Are assigned very specific roles in the data transformation process (raw data sampled somewhere are progressively turned into valuable knowledge)

(b) Have the capability of fulfilling such roles very efficiently and avoid unnecessary transmissions through the network paths thus rationalizing the use of scarce energy

(c) Have the capability to promptly react to changes in the network architecture and mitigate the "disturbances" to the overall data transformation plan through their reconfiguration capabilities.

The accelerated hardware that is considered is based on Software Define (SD) SoC elements that combine a conventional processing element like ARM processor with hardware programmability of FPGA [9-10]. The FPGA part of the node architecture is handling the core functionality that the assigned role prescribes for this particular node (e.g., event detection, spectral decomposition, fusion). This is installed/deployed in the FPGA according to the initial role assignment/planning that the network planners derive. Nodes should be equipped with the necessary components which are held inactive until external control stimulates the node. Within each node a supervisor/control process is typically executed on the processor part of the SoC (Fig. 7).

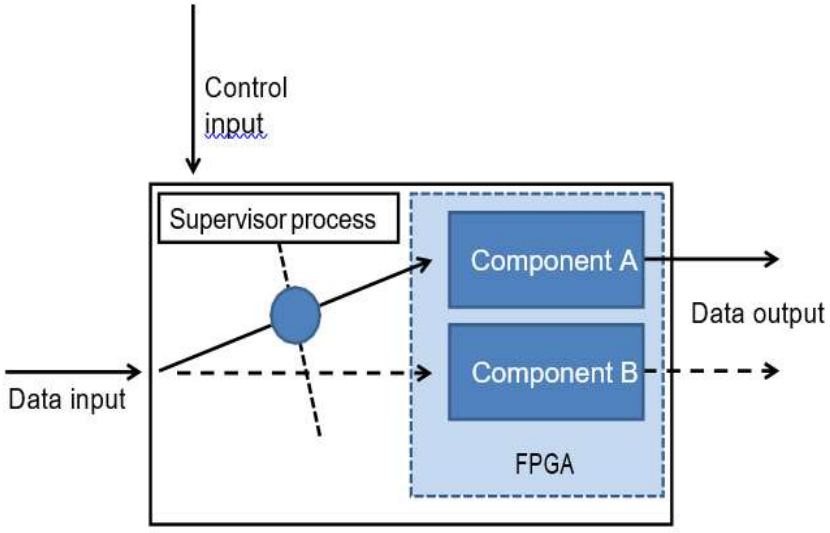


Fig. 7. Control process of IoT network

Table 1. Comparison of measurements

Packet size (Bytes)	32	128	512	2048	8192	32768
Number of packets	134217728	33554432	8388608	2097152	524288	131072
Time (s)	219	66	29	21	18.2	17.6
Speed (Megabytes/s)	19.61	65.08	148.10	204.52	235.99	244.03
PL clock @100 Maximum Burst size 16, total transfer size 4294967296 bytes						

Simulation is run for different packet sizes and for each packet size, performance is measured. The Table 1 summarizes the measurements. The measurement are carried out with 100 MHZ with burst size 16 and 256 respectively. These configurations absolutely depends on how we define ZYNQ HW design in Vivado (for example the Clock frequency, the burst size etc.). Four GBytes of data to DRAM have been transferred and from that speed of transfer is obtained. Packet size has been defined for each test. Total transfer size is divided by total packet in order to get number of packets. From that, performance is calculated (time and speed).

4 Conclusion

A generalized architecture of INP has been discussed in this thesis paper taking into account the role of each components interfaces to ensure efficient exchange of the information and optimization of the overall resources. In conventional network processing scenarios, problems frequently arise in a pipeline when certain data or pieces of information are not be readily accessible or available at the time they are needed by the pipeline or when computations take longer as a result of an exceptional condition. These problems contribute to an overall slowdown of the processing pipeline and lead to undesirable data transmission/processing stalls that markedly reduce performance. The INP system overcomes many of these issues and limitations by implementing discrete processing paths wherein each processing path is directed towards handling network traffic and data of a particular composition. The system architecture combines the energy efficiency with the flexibility and programmability of a system on a chip processor. As power consumption is the highest priority design constraint, the proposed system for WSNs/IoT uses two techniques to reduce power consumption. 1) Lightweight event handling in hardware: initial responsibility for handling incoming interrupts is given to a specialized processor, removing the software overhead that would be

required to provide event handling on a general-purpose processor. 2) Hardware acceleration for typical WSN/IoT tasks: modular hardware accelerators are included to complete regular application tasks such as data filtering.

References

1. Myers, Brad, Scott E. Hudson, and Randy Pausch. "Past, present, and future of user interface software tools." ACM Transactions on Computer-Human Interaction (TOCHI) 7.1 (2000): 3-28.
2. Carpenter, Mark Alan, Kathy Lockaby Khalifa, and David Bruce Lection. "Methods, system and computer program products for delayed message generation and encoding in an intermittently connected data communication system." U.S. Patent No. 5,859,973. 12 Jan. 1999.
3. Naeem, Tahir, and Kok-Keong Loo. "Common security issues and challenges in wireless sensor networks and IEEE 802.11 wireless mesh networks." 3; 1 (2009).
4. Fan, GaoJun, and ShiYao Jin. "Coverage problem in wireless sensor network: A survey." JNW 5.9 (2010): 1033-1040.
5. Minelli, Michael, Michele Chambers, and Ambiga Dhiraj. Big data, big analytics: emerging business intelligence and analytic trends for today's businesses. John Wiley & Sons, 2012.
6. Al-Karaki, Jamal N., and Ahmed E. Kamal. "Routing techniques in wireless sensor networks: a survey." IEEE wireless communications 11.6 (2004): 6-28.
7. Raghunathan, Vijay, et al. "Energy-aware wireless microsensor networks." IEEE Signal processing magazine 19.2 (2002): 40-50.
8. Adapteva, Epiphany Architecture Reference. http://adapteva.com/docs/epiphany_arch_ref.pdf, 2013
9. http://www.zedboard.org/
10. Prongnuch, Sethakarn, and Theerayod Wiangtong. "Heterogeneous Computing Platform for data processing." Intelligent Signal Processing and Communication Systems (ISPACS), 2016 International Symposium on. IEEE, 2016.

Survey of short area networks based on optical wired and wireless media

Nikolaos D. Raptis*

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

Abstract. In the present dissertation, the main target was to implement short area networks with range of tens of meters using optical media. The work was separated in two large sections. The first one concerned the implementation of indoor wired transmissions covering distances between 50 m and 100 m using Large Core Step-Index Plastic Optical Fibers (SI-POFs) made of polymethyl methacrylate (PMMA). The motivation was the necessity to deal with the rather large power losses and the low bandwidth of such fibers. The bandwidth of several SI-POF segments of different manufacturers increased by about 40% and the losses slightly decreased after passing them through a specific thermal treatment. The only limitation was the direct connection of the source with the SI-POF. The behavior of the SI-POFs remained improved to the same extent during an interval larger than a year. In the second part, diffused optical wireless transmissions at 265 nm were investigated. The target ranges were a few tens of meters. The bit rates that were applied were of kilobit per second order. The motivation was the necessity to deal with (i) the inability to deploy Non-Line-Of-Sight (NLOS) optical wireless links under several atmospheric conditions and (ii) the ambient noise that may have severe impact on such links. Both requirements were satisfied using the solar-blind part in the C-band of the Ultraviolet (UV-C) spectrum. It was proven theoretically and experimentally that the power losses of such links after applying transmissions at 265 nm are large but can be reduced significantly under harsh atmospheric conditions of artificial fog appearance. The lower losses under thick atmosphere conditions ensured the better performance of the low-rate signals that were transmitted, as well.

Keywords: thermal treatment of plastic optical fibers, bandwidth, bit error rate, pulse position modulation, solar-blind UV-C band, optical wireless communications, scattering, absorption.

1 Introduction

In this survey, two sections were covered that concerned the investigation of the potential deployment of links for short distances that are based on optical technologies. The first section concerns wired transmissions. Provided that data rates in the access

* Dissertation Advisor: Dimitris Syvridis, Professor.

network segment are expected to increase in the near future, the adoption of optical fibers becomes necessary. In an environment where optical cabling dominates, an issue that emerges is the type of fiber that will be selected for covering tens of meters in indoor areas. A cost-effective choice is the SI-POFs with 980 μm core diameter manufactured using PMMA. Despite their advantages, they have significant limitations that would prohibit their massive installation. The limitations are the increased power losses and the limited bandwidth. The question that was answered was how feasible is the direct intervention in the POF, in order its attributes to be modified. The results were a bandwidth increase of almost 40% and a slight decrease of the losses of the POFs that went through a specific heating procedure. For each 50 m POF segment that was thermally treated, a respective reference fiber of equal length from the same manufacturer was used for the comparisons. The profitable results appeared under the condition of direct connection of the laser with the POF, with the narrow optical beam launched focused in the center of the fiber core. However, under overfilled launching conditions, the thermally treated POFs gave similar results to those of the respective reference POFs in terms of bandwidth, whereas a slight decrease of the losses of the thermally treated fibers was measured again. During a longer than a year interval of occasional measurements, the improvement of the thermally treated fibers was preserved. The explanation of the phenomena was confirmed through simulations and the theoretical results were compatible with the respective experimental results.

In the second section, which concerns the optical wireless transmissions in an outdoor environment, the target was the coverage of short distances under the regime of diffuse transmissions with low data rates. The requirements that were defined were the absence of Line-Of-Sight (LOS) components in the link and the insignificant degradation of the link operation due to the ambient noise, especially by the sun. The NLOS regime imposes the existence of a mechanism that will enable the diffused transmissions. This mechanism is the scattering due to the molecules and the aerosols in the atmosphere. Wavelengths of conventional optical wireless communication systems do not scatter intensely, whereas the noise level of the sun that reaches the earth surface in such wavelengths is not negligible. A band that allows the realization of transmissions in a diffused way with a particularly low power level from the sun on the earth's surface lies between 200 nm and 280 nm and is called solar-blind UV-C band. Initially, the survey included the channel modeling of point-to-point links applying single and multi-scattering schemes at 265 nm. Moreover, transmissions of pulsed and multicarrier signals were considered and the operation of a simple network infrastructure was modeled, as well. In the network scenario, the access to the medium was regulated by a Code Division Multiple Access (CDMA) scheme. Apart from the transmission under clear atmosphere conditions, the issue of the operation under a thicker atmosphere regime was also investigated. According to the results, a significant decrease of the power losses and an amelioration of the bandwidth appeared. The deviation from the initial results obtained under clear sky conditions was a function of the atmosphere density and the size of the scattering centers' radii that were taken into account in the more complex version of the models that were regarded for some geometric configurations. In terms of the theoretical transmissions of signals, the propagation through a medium with larger density resulted in the reduction of the required

power levels for the operation of the links in specific performance boundaries and the limitation of the intersymbol interference in cases where the bandwidth was low under the clear atmosphere regime. Furthermore, in the networking scenario, when multiple nodes transmitted simultaneously according to a probability, the lower optical power levels required in order to achieve specific Bit Error Rate (BER) values under a thick atmosphere case compared to the case of clear atmosphere were verified. The simulation results indicated that NLOS links at 265 nm could be implemented and operate successfully. Therefore, the next step would be the experimental measurements. In this part, point-to-point links were deployed. In the optical part, the transmitter consisted of Light Emitting Diodes (LEDs) at 265 nm and at the receiving side an optical filter was followed by a Photo-Multiplier Tube (PMT). The first stage included the measurement of the power losses in clear atmosphere up to 20 m range, for a large set of transmitter and receiver elevation angles. When artificial fog appeared in the medium using a fog machine, a clear decrease of the losses was measured, which was different for several link distances and several elevation angle pairs when deploying NLOS configurations. For some links that were tested, the reduction of the losses was over 7 dB during a large number of successive samples of the received signals. The lowest values of the measured losses for some configurations were more than 10 dB lower than in clear atmosphere. The second stage of the experimental measurements included the performance evaluation of links in terms of BER for transmissions of a fourth order Pulse Position Modulated (4-PPM) signal [1] and a Flip-Orthogonal Frequency Division Multiplexed (Flip-OFDM) signal [2] for kbit/s rates. The BER for 10 kbit/s was measured for the largest number of elevation angle pairs that could be settled and for the same distances as in the losses measurements set. The optimum demodulation of 4-PPM was the reason of its superiority. Then, the fog machine was used for the evaluation of the impact of the medium thickening on the performance of the links under investigation. When artificial fog appeared, the expected decrease of BER at 10 kbit/s rate for 4-PPM and Flip-OFDM and at 4 kbit/s for Flip-OFDM was verified for several NLOS configurations. The amelioration was over 2 orders of magnitude for some link cases than the respective measurements under clear sky conditions. Finally, the impact of fog on LOS links was investigated, where the filter and the PMT were replaced by a proper lens and a pin photodiode. The appearance of fog had a devastating impact on performance, as a significant increase of the losses and deterioration of the BER appeared.

2 Optical Communication Systems Using SI-POFs

In the near future, the data rates will increase even in households. A cable that can support the increasing rates will be required in indoor environments for covering short ranges. A possible solution could be the Large Core PMMA SI-POFs, SI-POFs in short, due to their low cost and easy handling. Nevertheless, the SI-POFs suffer from rather large attenuation, whereas severe constraints are imposed on the transmission rate due to their increased modal dispersion. Intensive theoretical and experimental efforts have been made in order to mitigate the effect of limited bandwidth of SI-

POFs for distances between 50 and 100 m [3]-[9]. However, the bandwidth limitation is still present. It was examined if the power losses and the bandwidth could be improved by intervening directly in the channel. It was shown experimentally and verified theoretically that an increase in bandwidth and a slight decrease of the attenuation can be achieved by treating thermally an SI-POF following a specific procedure.

2.1 Theoretical Explanation of the Bandwidth Limitation of SI-POFs

Supposing that a narrow and well collimated beam is launched at the central part of the POF core, exciting only lower order modes, the origins of higher order modes activation are the inhomogeneities and impurities in the core material, which act as scattering centers for the incident beam, resulting in coupling of optical energy to higher order modes. The outcome is the lower bandwidth due to higher accumulated modal dispersion. If a mechanism existed that could "eliminate" as many as possible scattering centers along a POF, then the coupling length of the POF [10] would increase and the higher order modes would not be activated at short lengths. Additionally, the irregularities at the core-cladding boundaries may have critical effect on the accumulated modal dispersion, as well. The desired mechanism should deal with the two scattering "areas". The power coupling between modes and the influence of the non-ideal core-cladding boundaries can be visualized by the Gloge's time-dependent power flow equation given below [10] that outlines the light propagation along a POF

$$\frac{\partial p(\theta,z,t)}{\partial z} = -a(\theta)\times p(\theta,z,t) - \tau_{tr,rel}(\theta)\times\frac{\partial p(\theta,z,t)}{\partial t} + \frac{1}{\theta}\times\frac{\partial}{\partial\theta}\left[\theta\times D\times\frac{\partial p(\theta,z,t)}{\partial\theta}\right] \quad (1)$$

where z represents the length dimension, t is the time variable, θ is the propagation angle with respect to the fiber axis, $p(\theta, z, t)$ is the power distribution over angle, space and time, $a(\theta)$ is the mode dependent attenuation, $\tau_{tr,rel}(\theta)$ is the relative mode delay and D is the coupling coefficient.

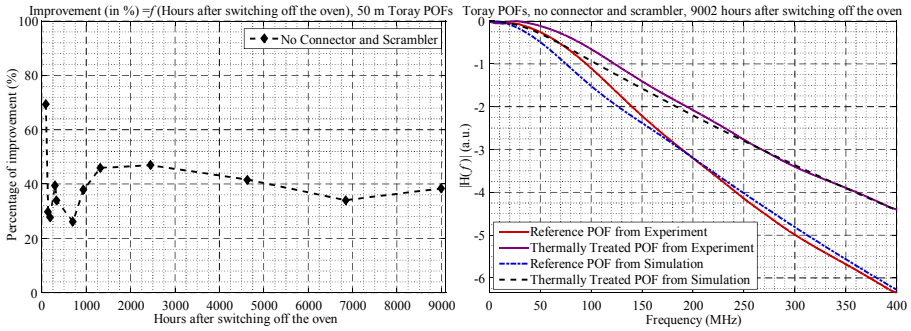
Three terms are added to give $a(\theta)$. One of them appears due to reflections of modes at the core-cladding interface and is proportional to $-ln(R_{refl})$, where R_{refl} is the reflection factor with $0 \le R_{refl} \le 1$ [4]. By decreasing R_{refl} and considering more irregular core-cladding boundaries, the reflected power into the POF is severely decreased during light propagation and the activated higher order modes will appear intensely attenuated at the end of the POF compared to a POF with higher R_{refl}, leading to bandwidth increase. Therefore, the R_{refl} decrease would have a favorable impact on bandwidth. The coupling coefficient D is related to all the scattering sources in the core [10]. If D is decreased, the core is considered more homogeneous, and assuming central launching of a focused laser beam, a lower number of higher order modes will appear at fiber end, resulting in larger bandwidth for a fiber with a length shorter than its coupling length. Therefore, a contradiction would appear between the improved core homogeneity and the improved irregularities at the core-cladding boundaries in terms of bandwidth of a properly manufactured fiber. We decided to proceed to a thermal curing of SI-POF specimens of 50 m and 100 m length, targeting at increasing their homogeneity. Such a procedure decreased D and increased slightly R_{refl}.

2.2 Heating Procedure and Experimental Results

Focusing on 50 m segments, two types of commercially available SI-POFs have been used: those having Numerical Aperture (NA) of 0.46, provided by Luceat and those having NA of 0.5, provided by Toray. For both types, the core and cladding diameters were 0.980 and 1 mm, respectively. For each POF type, a pair has been used in all the experiments; one serving as a reference and the other to be treated thermally. Both POFs of the same pair were wrapped in loose loops of equal diameters and were kept in the loop state before, during and after the heating procedure. It was experimentally confirmed that the POFs of the same pair had the same bandwidth and attenuation, before the procedure begun. For the thermal treatment, an oven with adjustable temperature was used. About 1.0 m of each end of the POFs to be heated was kept outside the oven. The heating process consisted of increasing the temperature of the oven by 5^{o}C every 20 minutes up to 75^{o}C. After 20 minutes at 75^{o}C, the temperature was reduced to 65^{o}C and kept at that level for 1 hour. Finally, the POFs were left to cool slowly, reaching the ambient temperature (25^{o}C) after approximately 1 hour. This process decreased the scattering centers and the ameliorated core homogeneity surpassed the impact of the improved core-cladding irregularities, resulting in larger bandwidth after the thermal treatment [11]. The requirement is the direct connection of the source with the POF. The source was a laser and the beam with a waist much smaller than the POF's diameter was launched in each POF sample aligned with the core. The disadvantages were a slight deformation of the thermally treated POFs and the inability to generate exactly the same improvement after repeating this procedure for different POF samples of the same length even from the same reel.

A set of bandwidth values obtained after occasional measurements in an interval of thousands hours for both SI-POF pairs gave a clear indication of the performance improvement. In the experimental setup, a Fabry-Perot Laser Diode at 653 nm was connected directly with the POF. The POF was directly connected to the photodiode, as well. After reception, the frequency response of the electrical signal was measured. Before each measurement, the whole system was calibrated using a 0.3 m piece of SI-POF in order to isolate exclusively the optical frequency response of the fiber. The evolution of the fiber bandwidth enhancement with time appears in Fig. 1(a) for the 50 m POF pair of Toray. The horizontal axis corresponds to the hours that have passed after the end of the treatment. The vertical axis corresponds to the bandwidth improvement of the treated POF relative to the reference one. A sequence of ten bandwidth measurements has been performed for both the thermally treated and the reference POFs and the difference of their mean bandwidths has been drawn as a percentage value. As it can be observed in Fig. 1(a), after some fluctuations during the first few hundreds of hours, the bandwidth of the thermally treated fiber stabilized at more than 30% higher values than those of the reference fiber. Even 9002 hours after the end of the treatment, the improvement was close to 38.40%. The mean bandwidth enhancement was close to 39.28% from all the measurements. Concerning the optical losses of the thermally treated POF, a decrease between 0.35 dB and 0.64 dB appeared during the respective measurements in the 9002 hours interval. A similar behavior appeared for the thermally treated POF of Luceat. More specifically, the mean

bandwidth increase of the treated POF compared to the one of the reference POF was around 42.78% and the losses were decreased in a range between 0.5 and 0.95 dB in an interval of 11900 hours after switching off the oven.



(a) (b)

Fig. 1. (a) Optical bandwidth improvement for the thermally treated Toray POF as a function of time for direct connection of the source with the fiber. (b) Experimental snapshots of the amplitude of the frequency response of the two Toray POFs, 9002 hours after switching off the oven and the respective frequency response amplitudes after numerical solution of (1).

Two snapshots of the magnitude of the frequency response of the thermally treated and the reference Toray POFs, captured 9002 hours after switching off the oven, are depicted in Fig. 1(b) as solid curves. The optical bandwidth at 3 dB of the snapshot of the treated POF is almost 267.47 MHz. About 189.03 MHz was the measured bandwidth of the reference POF. It was crucial to confirm if the impact of the reduction of D and the slight increase of R_{refl} achieved after the thermal treatment agreed with the experimental results. For this reason, (1) was solved numerically. The central launching of a Gaussian beam at 650 nm with $\sigma = 3^{o}$ (inside the POF) was considered. NA = 0.50 and core refractive index n_{core} = 1.492 were set for both POFs. $D = 3.18\times10^{-4}$ rad^2/m and R_{refl} = 0.99982 for the reference and $D = 1.51\times10^{-4}$ rad^2/m and R_{refl} = 0.99991 for the thermally treated POF were defined. In Fig. 1(b), the dashed-dotted and the dashed curves represent the amplitudes of the theoretical frequency response of the fibers that correspond to the reference and the treated POFs, respectively. These curves both follow quite well the respective experimental curves. The theoretically obtained optical bandwidths were 188.16 MHz and 267.52 MHz, for the reference and the treated POFs, respectively, verifying further the explanations of the improvement.

An issue that appeared was the behavior of each POF under overfilled launching conditions. With all modes present from the beginning of the span, the contention between the two mechanisms, i.e. the attenuation of higher order modes due to the irregularities at the core-cladding boundaries and the (re)activation of higher order modes due to the core impurities, is expected to result in no bandwidth improvement for the treated POFs. The experimental frequency responses of both Toray POFs verified the expectation when overfilled launching was applied 9002 hours after switching off the oven, as 96.72 MHz and 99.71 MHz were the bandwidths of the reference and

the thermally treated POF, respectively. Theoretical bandwidths close to the experimental values were estimated for both POFs. In terms of the losses, a decrease between 0.50 dB and 0.85 dB was obtained for the treated POF from all the measurement sets under this launching regime. A similar behavior to that of the Toray POFs appeared for the Luceat POFs under overfilled launching conditions, as well.

For direct connection of the laser with each POF, the bandwidth superiority of the heated POFs remained constant up to the moment of the last measurements. The bandwidth increase of the thermally treated POFs disappeared under overfilled launching conditions, though. A slight decrease of the losses appeared for the thermally treated POFs under both launching conditions. The assumptions that after the thermal treating the scattering centers in the POF core diminished (decrease of D) and the irregularities at the core-cladding boundaries are reduced (slight increase of R_{refl}), are realistic producing compatible theoretical results to the experimental ones.

3 Communication Systems Using Diffuse Light

Conventional communication systems may fail in environments with intentional jamming or interference, combined with dense atmosphere and/or complex configurations due to obstacles. A solution can be the solar-blind UV-C band, whose tempting features are the significantly reduced solar irradiance on earth's ground, the intense scattering [12] which offers the way to set up NLOS links and its combination with strong absorption which ensures covertness [13]. In the present section, apart from investigating theoretically and experimentally the losses and the BER performance of low rate short-range NLOS links at 265 nm under clear atmospheric conditions, a second crucial target was to examine how strong scattering in a denser medium affects the performance of such links.

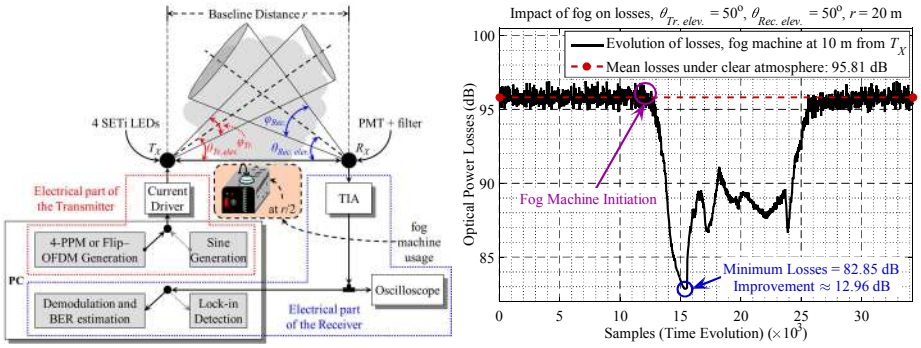
3.1 Theoretical estimations

The performance improvement of NLOS systems at 265 nm when the medium thickened was numerically demonstrated for short ranges and two scenarios; a point-to-point and a networking one. For the configurations and all the modulation formats that were investigated, the lower losses of the thicker medium imposed lower emitted power levels. For the same configurations, higher bandwidth values were estimated under thick atmospheric conditions [14], [15]. These results were evaluated using two theoretical channel models; a single-scattering channel model [15] and a Monte-Carlo approach that supports multiple scattering events per photon [16], which was more reliable in cases where higher order scattering events must be taken into account.

In a networking scenario where a few nodes transmitted to a central one at 265 nm establishing NLOS links with ranges up to 50 m in a sparse medium, the application of a CDMA protocol may cause significant limitations [15]. The enhanced bandwidth and the lower losses of the channels in a thick environment allowed the transmission of encoded signals without requiring extreme emitted power levels in order specific BER values to be achieved, making the connections to the central node feasible.

3.2 Experimental Setup for Measuring either the Losses or the BER

The setup could be used for measuring either the power losses or the BER, as shown in Fig. 2(a). The elevation angles of both the transmitter ($\theta_{Tr.\ elev.}$) and the receiver ($\theta_{Rec.\ elev.}$) and the range (r) could be adjusted. The transmitter consisted of 4 LEDs (by SETi) emitting at 265 nm. The receiver consisted of a PMT (R7154 by Hamamatsu) with an optical filter (transmittance = 0.17) in front of it. The LEDs' divergence angle ($\varphi_{Tr.}$) and the receiver's Field-of-View ($\varphi_{Rec.}$) were 10° and 30°, approximately. The PMT was followed by a Trans-Impedance Amplifier (TIA).



(a) (b)

Fig. 2. (a) Experimental setup. PC: Personal Computer. (b) Time evolution of the losses without and with fog in the medium, for $\theta_{Tr.\ elev.} = \theta_{Rec.\ elev.} = 50^{o}$, $r = 20$ m.

The LEDs intensity was directly modulated. The losses were estimated by producing a 1 KHz sine signal and applying a lock-in detection scheme [16]. The sources were biased to the same point that provided close to 1 mW mean optical power per LED. For the set of BER measurements, slightly less than 0.5 mW mean optical power was emitted by each LED. Non-linearity issues were avoided. 4-PPM and Flip-OFDM were examined [2], [17]. The same root mean square voltage was generated for both schemes for fair comparisons due to the AC coupled TIA output. For 4-PPM, maximum likelihood detection without threshold was applied after integrate-and-dump filtering in slot duration level. For each Flip-OFDM symbol, 256×2 subcarriers were used. A fog machine [17], [18] was used to increase the thickness of the medium artificially. It was set to $r/2$ in all link cases. It was examined if stronger scattering in a thicker medium would result in the reduction of the losses or BER of NLOS links.

3.3 Experimental Measurements of the Power Losses and the BER under Clear Sky and Foggy Conditions

The power losses were measured for a large combination of transmitter and receiver elevation angles and 5, 10 and 20 m ranges under clear atmosphere conditions. The losses were below 60 dB only when both LOS components existed in the links and r = 5 m. The losses increase with distance was intense. When r = 10 m and $\theta_{Tr.\ elev.}$, $\theta_{Rec.}$

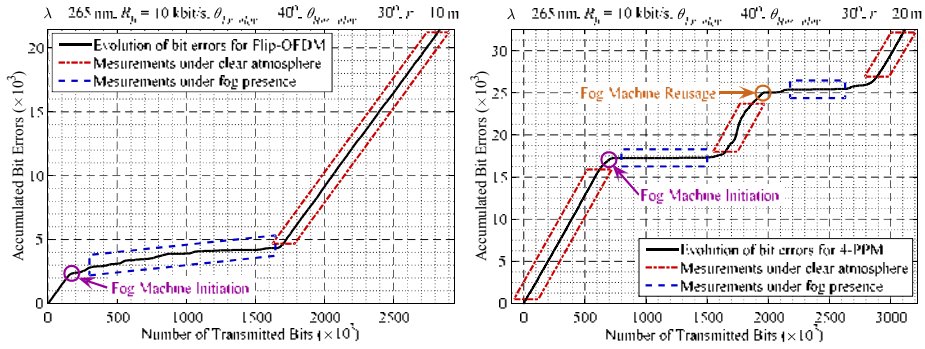
$_{elev.}$ were both over 40° in order NLOS links to be established, the losses were over 80 dB. For r = 20 m and $\theta_{Tr.\ elev.}$, $\theta_{Rec.\ elev.}$ over 40°, the losses were over 90 dB. Therefore, the NLOS links were deployed due to the dominance of scattering over absorption in a sparse medium at 265 nm. The losses were significantly high, though.

A wireless environment rich in scattering centers is expected to exhibit lower losses for short distances (r < 50 m) and transmissions at 265 nm, an expectation reinforced by the theoretical results in 3.1. The medium was thickened intentionally using the aforementioned fog machine for experimental verification. After the initial convergence of the lock-in stage, the operation of the fog machine started and during its operation, the impact of fog was observed until the machine was stopped, leaving the detector to converge again to the initial losses. This procedure was repeated for each examined configuration up to 20 m, reusing the machine in order to ensure the reappearance of the results. A representative scenario is depicted in Fig. 2(b). Time samples in thousands values are shown on the horizontal axis. The vertical axis represents the instant values of the losses. Each sample is obtained every 5 msec. After setting the elevation angles, the distance and the PMT gain to $\theta_{Tr.\ elev.} = \theta_{Rec.\ elev.} = 50^{o}$, r = 20 m and 4.51×10^{6}, respectively, the transmission of the modulated carrier with the sine signal started. The mean losses in clear atmosphere were estimated close to 95.81 dB for this set of measurements. After 12100 samples the operation of the fog machine started. The denser medium resulted in an immediate reduction of the losses. The minimum value of 82.85 dB was the lowest one than all the measured values for this configuration. The losses diminished by almost 13 dB meaning that the received optical power was 20 times higher than the level measured in clear atmosphere. Despite the instability of the medium, after the 18500th sample the losses remained lower than 90 dB for more than 5000 samples, i.e. almost 4 times higher received power than initially for more than 25 sec, before the rapid dissolution of fog. The losses of all the NLOS links at 265 nm that were examined exhibited definitely a decreasing behavior when fog appeared, even when an obstacle was placed at r/2 in some cases [16].

BER measurements were obtained for R_b = 10 kbit/s rate, whereas the ranges that were studied were 5, 10 and 20 m for many geometric configurations. For this rate, 30 subcarriers carried data Quadrature Phase Shift Keying symbols per Flip-OFDM symbol. Under clear sky conditions, the performance of 4-PPM was better compared to that of Flip-OFDM due to the optimal demodulation of the former. The superiority of 4-PPM was clear in intermediate elevation angles which ensured the NLOS operation of the links. However, for large elevation angles, the superiority of 4-PPM was rather mitigated due to the decreased received power levels and the deterioration of the Signal-to-Noise-Ratio. The limitations due to the higher losses at large elevation angles became more severe at r = 20 m. The transmissions were not bandwidth limited for the examined rates [19]. Consequently, for the tenuous medium and the emitted power levels applied here, 4-PPM outperformed Flip-OFDM, as 4-PPM fitted better to loss limited NLOS links at 265 nm up to r = 20 m, provided that the losses are not extreme due to the increased range and the large elevation angles.

The decreasing behavior of the losses under fog presence was exploited in order to increase the received optical power and decrease the BER in NLOS links for the distances and the bit rate examined so far. The impact of fog on BER is presented on

curves where the horizontal axis depicts the transmitted bits and the vertical axis depicts the respective accumulated bit errors [17], [18]. The slope of such curves is the BER. If any change in the medium affects the losses, such as the number of the scattering centers, then the slope will be modified locally. Therefore, the behavior of BER for a specific configuration due to the absence and then the presence of fog can be evaluated experimentally. In order to do so for Flip-OFDM, the following parameters were set: r = 10 m, $\theta_{Tr.\ elev.}$ = 40°, $\theta_{Rec.\ elev.}$ = 30° and the PMT gain close to 7×10^5. All measurements were carried out in real-time. The respective curve appears in Fig. 3(a). The values in both axes of both curves in Fig. 3 are in thousands bits. The fog machine was initiated after roughly 1.72×10^5 transmitted bits. The blue dashed parallelogram indicates an interval of 1529 accumulated bit errors for almost 1.345×10^6 received bits giving BER ≈ 1.14×10^{-3} locally. Then, the machine was stopped and the fog dissolved rapidly. For the rest of the curve covered by the red dashed-dotted parallelogram, the slope was 1.48×10^{-2} (16553 bit errors for 1.118×10^6 received bits).



(a) (b)

Fig. 3. Evolution of bit error accumulation for $\theta_{Tr.\ elev.}$ = 40°, $\theta_{Rec.\ elev.}$ = 30°, R_b = 10 kbit/s and (a) Flip-OFDM at r = 10 m and (b) 4-PPM at r = 20 m.

Keeping $\theta_{Tr.\ elev.}$ = 40°, $\theta_{Rec.\ elev.}$ = 30° and setting r = 20 m and the gain of the PMT close to 2.00×10^6, a 4-PPM signal was transmitted at the same rate. The impact of fog on BER is displayed in Fig. 3(b). Initially, under clear atmosphere, for 6×10^5 transmitted bits, 15393 errors occurred and BER ≈ 2.57×10^{-2} was measured. This part is highlighted by the first red dashed-dotted parallelogram. Just before the transmission of 7×10^5 bits, the operation of the machine initiated. During fog production which is pointed by the blue dashed parallelogram, only 20 errors appeared for a 7×10^5 transmitted bits, and the BER became 2.86×10^{-5}, that is 3 orders of magnitude lower than the initial BER. After fog dissolution, according to the second red dashed-dotted parallelogram, the BER was restored to 2.60×10^{-2}. Just after 1.95×10^6 bits, the machine was reused. For 4.5×10^5 received bits (noted by the second blue dashed parallelogram) the BER changed to 1.91×10^{-4}, almost one order of magnitude higher than the previous measurement with fog present, due to the more inhomogeneous coverage with fog than previously. The improvement is not canceled, though. The BER became 2.43×10^{-2} for the last part of the curve. It is evident that when fog appeared, the per-

formance was enhanced for both configurations that were presented and for many others that were investigated under the presence of fog for distances up to 20 m.

4 Final Conclusions

The main target of this work was separated in two branches. The first one was to cover distances of tens of meters in indoor environments using SI-POFs, while trying to deal with the disadvantages of the channel. The second one was the coverage of tens of meters distances in outdoor environments using the scattering of atmosphere as the mean to deploy optical wireless NLOS links. The wavelength range between 200 and 280 nm has some favorable attributes that can be exploited in order to deploy such links. A simultaneous task was the investigation of the impact of the atmosphere thickness on the performance of NLOS links.

In the wired transmissions, when SI-POFs from different manufacturers were treated thermally following a specific procedure, a mean 35-40% bandwidth increase appeared in more than a year interval compared to the bandwidth of the respective reference SI-POFs of the same manufacturers. The requirement was the direct connection of the laser with the POF. Otherwise, under overfilled launching conditions, both the thermally treated and the respective reference SI-POFs presented similar bandwidths. A slight decrease of the losses appeared in all treated POFs. An open issue is the optimization of the treating procedure is crucial in order to achieve the highest possible improvement. The systematic investigation of the impact of a similar thermal procedure on Graded-Index POFs (GI-POFs) could be an interesting issue, as well.

In the wireless part, the experimental results for transmissions at 265 nm showed the feasibility to establish functional outdoor wireless NLOS links in the solar-blind part of the UV-C band under clear atmosphere. The main attributes were the increased but not prohibitive power losses and the low bit rates that could be supported. A clear dominance of scattering over absorption was proven experimentally and validated by theoretical results when the atmosphere became thicker. This was verified for several NLOS links by measuring not only the losses but also the BER when artificial fog was inserted in the medium. Therefore, it was confirmed that the part of the UV-C band between 200 and 280 nm is a strong candidate in applications where the operation of short range and low bit rate NLOS optical wireless links is required under harsh atmospheric conditions and covertness. A further step can be the optimization of the emitted power levels in order to deploy as distant as possible NLOS links, in accepted limits in terms of eye safety, though. Other critical issues are the deployment of functional network clusters and the implementation of multiple access protocols that suit better to the nature of the power-limited NLOS channels. In terms of other applications, the underlying technology can be used as a supplementary choice to conventional technologies when the latter fail. For instance, combined with the 802.15.7 standard [20], the concept of intelligent transport can be enhanced. Finally, the immediate and intense change of the losses in NLOS links at 265 nm when the thickness of the medium altered is an attribute that can be exploited in local weather prediction systems as a supplementary alternative to other mature technologies.

References

1. He, Q., Sadler, B. M., Xu, Z.: Modulation and coding tradeoffs for non-line-of-sight ultraviolet communications. In Proceedings of SPIE, 7464, San Diego, CA, USA, (2009).
2. Fernando, N., Hong, Y., Viterbo, E.: Flip-OFDM for unipolar communication systems. IEEE Transactions on Communications 60(12), 3726–3733 (2012).
3. Lee, S. C. J. et al.: Discrete multitone modulation for maximizing transmission rate in step-index plastic optical fibers. Journal of Lightwave Technology 27(11), 1503–1513 (2009).
4. Breyer, F.: Multilevel transmission and equalization for polymer optical fiber systems. Ph.D. Dissertation, Dept. Tecnhische Universität München, München, Germany, 2010.
5. Grivas, E., Raptis, N., Syvridis, D.: An optical mode filtering technique for the improvement of the large core SI-POF link performance. Journal of Lightwave Technology 28(12), 1796–1801 (2010).
6. Raptis, N. et al.: Space-time block code based MIMO encoding for large core step index plastic optical fiber transmission systems, Optics Express 19(11), 10336–10350 (2011).
7. Raptis, N. et al.: Space-time block codes application in large core step-index plastic optical fibers. In Proceedings of POF 2011, pp. 31–36, Bilbao, Spain (2011).
8. Pikasis, E. et al.: A space-frequency block encoded OFDM scheme for short area POF networks. In Proceedings of POF 2011, pp. 537–541, Bilbao, Spain (2011).
9. Pikasis, E. et al.: Performance evaluation of CDMA-DMT for 1-mm SI-POF short-range transmission links, IEEE Photonics Technology Letters 24(22), 2042–2045 (2012).
10. Drljača, B., Djordjevich, A., Savović, S.: Frequency response in step-index plastic optical fibers obtained by numerical solution of the time-dependent power flow equation. Optics and Laser Technology 44(6), 1808–1812 (2012).
11. Raptis, N., Syvridis, D.: Bandwidth enhancement of step index plastic optical fibers through a thermal treatment. IEEE Photonics Technology Letters 25(16), 1642–1645 (2013).
12. Xu, Z., Sadler, B. M.: Ultraviolet communications: potential and state-of-the-art. IEEE Communications Magazine 48(5), 67–73 (2008).
13. Junge, D. M.: Non-line-of-sight electro-optic laser computations in the middle ultraviolet. M.S. thesis, Naval Postgraduate School, Monterey, CA, USA (1977).
14. Raptis, N., Roditi, E., Syvridis, D.: Power-spectrum requirements in ultraviolet optical wireless networks. In Proceedings of SPIE, 9354-2, San Francisco, CA, US (2015).
15. Raptis, N., Pikasis, E., Syvridis, D.: Performance evaluation of modulation and multiple access schemes in ultraviolet optical wireless connections for two atmosphere thickness cases. Journal of Optical Society of America A 33(8), 1628–1640 (2016).
16. Raptis, N., Pikasis, E., Syvridis, D.: Power losses in diffuse ultraviolet optical communications channels. Optics Letters 41(18), 4421–4424 (2016).
17. Raptis, N., Pikasis, E., Syvridis, D.: Performance evaluation of non-line-of-sight optical communication system operating in the solar-blind ultraviolet spectrum. In Proceedings of SPIE, 9991-3, Edinburgh, UK (2016).
18. Raptis, N. et al.: Experimental evaluation of modulation formats' performance in diffuse UV channels. IEEE Photonics Technology Letters 29(11), 887–900 (2017).
19. Chen, G., Xu, Z., Sadler, B. M.: Experimental demonstration of ultraviolet pulse broadening in short-range non-line-of-sight communication channels. Optics Express 18(10), 10500–10509 (2010).
20. Rajagopal, S., Roberts, R. D., Lim, S.-K.: IEEE 802.15.7 visible light communication: modulation schemes and dimming support. IEEE Communications Magazine 50(3), 72–82 (2012).

Semantics of Negation in Extensional Higher-Order Logic Programming*

Ioanna Symeonidou

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

Abstract. There exist two different extensional approaches to the semantics of positive higher-order logic programming, introduced by W. W. Wadge and M. Bezem respectively. We show that the two approaches coincide for a broad class of programs, but differ in general. Moreover, we adapt Bezems technique under the well-founded, stable model and infinite-valued semantics and show that only the latter succeeds in retaining extensionality in the general case. We analyse the reasons for the failure of the well-founded adaptation of Bezem's technique, arguing that a three-valued setting cannot distinguish between certain predicates that appear to have a different behaviour inside a program context, but which happen to be identical as three-valued relations. Finally, we define for the first time the notions of stratification and local stratification for higher-order logic programs with negation and prove that every stratified program has a distinguished extensional model, which can be equivalently obtained through the well-founded, stable model or infinite-valued semantics.

1 Intoduction

Out of the many extensions of traditional logic programming that have been proposed over the years, the transition to a higher-order setting has been a particularly intriguing and at the same time controversial potential course. The key characteristic of higher-order logic programming is that (roughly speaking) it allows predicates to be passed as parameters of other predicates.

Recent research [19, 2, 3, 5, 4] has investigated the possibility of providing *extensional* semantics to higher-order logic programming. Extensionality facilitates the use of standard set theory in order to reason about programs, at the price of a relatively restricted syntax. Actually this is a main difference between the extensional and the more traditional intensional approaches to higher-order logic programming such as the ones of [13, 9]: the latter languages have a richer syntax but they are not usually amenable to a standard set-theoretic semantics. For a more detailed discussion of extensionality and its importance for higher-order logic programming, the reader can consult the discussion in Section 2 of [15].

* Dissertation Advisor: Panos Rondogiannis, Professor

Despite the fact that only very few articles have been written regarding extensionality in higher-order logic programming, two main semantic approaches can be identified. The first one (called "Wadge's semantics" in the following) [19, 5] was originally proposed by W. W. Wadge [19] for positive programs and later refined and extended by Charalambidis et al. [5]. It has been developed using domain-theoretic tools and resembles the techniques for assigning denotational semantics to functional languages. The second approach (called "Bezem's semantics" in the following) [2, 3], was proposed by M. Bezem also for positive programs. This approach relies on the syntactic entities that exist in a program and is based on processing the ground instantiation of the program.

A natural question that arises is whether one can still obtain an extensional semantics if negation is added to programs. Wadge's semantics has been extended to apply to programs with negation in two ways. The extension proposed in [4] was built upon the infinite-valued semantics [18], a relatively recent proposal to the semantics of negation in logic programming, defined over a logic with an infinite number of truth values. Very recently, a second extension was proposed in [8], generalizing the well-founded semantics [11]. On the other hand, prior to the work reported in this dissertation, no similar extension of Bezem's semantics had been proposed.

In this dissertation we focus on Bezem's approach and attempt to evaluate it, first in comparison to Wadge's semantics as the sole existing alternative and second with respect to its potential to generalise to programs with negation. Our contributions can be summarized as follows:

- We show that for a very broad class of *positive programs* the approaches of [2, 3] and [19, 5] coincide with respect to ground atoms that involve symbols of the program. On the other hand, we argue that if we consider an extended language, which allows existentially quantified predicate variables in the bodies of program clauses, then the two approaches give different results in general. This result, published in [6, 7], will not be discussed in the present summary of the dissertation.
- We demonstrate that neither the well-founded [11] nor the stable model [12] adaptation of Bezem's technique leads to an extensional model in the general case. The result concerning the stable model semantics was published in [16], while the result concerning the well-founded semantics was stated in [15, 17].
- We study the reasons behind the failure of the well-founded adaptation of Bezem's technique and the more general question of the possible existence of an *alternative* extensional three-valued semantics for higher-order logic programs with negation. We indicate that in order to achieve such a semantics, one has to make some (arguably) non-standard assumptions regarding the behaviour of negation, for example as in the case of [8]. The argument was first presented in [15].
- We demonstrate that by combining the technique of [2, 3] with the infinite-valued semantics of [18], we obtain an extensional semantics for higher-order logic programs with negation. This result was published in [14, 16]. Note that

the infinite-valued semantics was the first approach to negation to enable the extension of the semantics of [19, 5] (see [4]).
– We define the notions of *stratification* and *local stratification* for higher-order logic programs with negation. These notions were first defined in [14]; note that similar notions have not yet been studied under the semantics of [19, 5, 4, 8]. We prove that the stable model, the well-founded and the infinite-valued adaptations of Bezem's technique give equivalent extensional models in the case of *stratified* programs. The extensionality of the well-founded model for stratified programs was shown in [15] and affirmed the importance and the well-behaved nature of stratified programs, which was, until now, only known for the first-order case.

The next two sections motivate in an intuitive way the main ideas behind extending Bezem's semantics in order to apply to higher-order programs with negation. The remaining sections develop the material in a more formal way.

2 An Intuitive Overview of the Proposed Approach

In this section we give an intuitive description of the semantic technique for positive higher-order logic programs proposed by Bezem [2, 3] and we outline how we use it when negation is added to programs. Given a positive program, the starting idea behind Bezem's approach is to take its "ground instantiation", in which we replace variables with well-typed terms constructed from syntactic entities that appear in the program. For example, consider the higher-order program (for the moment, we use ad-hoc Prolog-like syntax):

```
q(a).
q(b).
p(Q):-Q(a).
id(R)(X):-R(X).
```

In order to obtain the ground instantiation of this program, we consider each clause and replace each variable of the clause with a ground term that has the same type as the variable under consideration (the formal definition of this procedure will be given in Definition 9):

```
q(a).
q(b).
p(q):-q(a).
id(q)(a):-q(a).
id(q)(b):-q(b).
p(id(q)):-id(q)(a).
        ...
```

One can now treat the new program as an infinite propositional one (i.e., each ground atom can be seen as a propositional variable). This implies that we can

use the standard least fixed-point construction of classical logic programming in order to compute the set of atoms that should be taken as "true".

Bezem demonstrated that the least fixed-point semantics of the ground instantiation of every positive higher-order logic program of the language considered in [2, 3], is *extensional* in a sense that can be explained as follows. In our example, `q` and `id(q)` are equal since they are both true of exactly the constants `a` and `b`. Therefore, we expect that if `p(q)` is true then `p(id(q))` is also true, because `q` and `id(q)` should be considered as indistinguishable.

We use the same idea with programs that include negation: the ground instantiation of such a program can be seen as a (possibly infinite) propositional program with negation. Therefore, we can compute its semantics in any standard way that exists for obtaining the meaning of such programs and then proceed to examine whether the chosen model is extensional in the sense of Bezem [2, 3]. As we are going to see in the subsequent sections, when the infinite valued model of the ground instantiation of the program is chosen for this purpose, the semantics we obtain is indeed extensional, but the same does not hold for the well-founded model or the stable model(s).

3 Infinite-valued Semantics

In this section we discuss the motivation behind the infinite-valued semantics [18]. The key idea of this approach is that in order to give a logical semantics to negation-as-failure and to distinguish it from ordinary negation, one needs to extend the domain of truth values. For example, consider the program:

$$\begin{aligned} &\mathtt{p} \leftarrow \\ &\mathtt{r} \leftarrow {\sim}\mathtt{p} \\ &\mathtt{s} \leftarrow {\sim}\mathtt{q} \\ &\mathtt{t} \leftarrow {\sim}\mathtt{t} \end{aligned}$$

According to negation-as-failure, both `p` and `s` receive the value *true*. However, `p` seems "truer" than `s` because there is a clause which says so, whereas `s` is true only because we are never obliged to make `q` true. In a sense, `s` is true only by default. For this reason, it was proposed in [18] to introduce a "default" truth value T_1 just below the "real" true T_0, and (by symmetry) a weaker false value F_1 just above ("not as false as") the real false F_0. Then, negation-as-failure is a combination of ordinary negation with a weakening. Thus ${\sim}F_0 = T_1$ and ${\sim}T_0 = F_1$. Since negations can be iterated, the new truth domain has a sequence $\ldots, T_3, T_2, T_1$ of weaker and weaker truth values below T_0 but above a neutral value 0; and a mirror image sequence $F_1, F_2, F_3, \ldots$ above F_0 and below 0. Since our propositional programs are possibly countably infinite, we need a T_α and a F_α for every countable ordinal α. The intermediate truth value 0 is needed for certain atoms that have a "pathological" negative dependence on themselves (such as `t` in the above program). In conclusion, our truth domain $\mathbb{V}^\infty$ is shaped as follows:

$$F_0 < F_1 < \cdots < F_\omega < \cdots < F_\alpha < \cdots < 0 < \cdots < T_\alpha < \cdots < T_\omega < \cdots < T_1 < T_0$$

It is shown in [18] that every first-order logic program has a unique minimum infinite-valued model, under an ordering relation $\sqsubseteq$. For example, the minimum infinite-valued model of the program presented above may be described as the set $\{(\mathtt{p}, T_0), (\mathtt{q}, F_0), (\mathtt{r}, F_1), (\mathtt{s}, T_1), (\mathtt{t}, 0)\}$.

4 The Syntax of $\mathcal{H}$

In this section we define the syntax of our language $\mathcal{H}$. $\mathcal{H}$ uses a simple type system with two base types: o, the boolean domain, and ι, the domain of data objects. The composite types are partitioned into three classes: functional (assigned to function symbols), predicate (assigned to predicate symbols) and argument (assigned to parameters of predicates).

Definition 1. *A type can either be* functional*,* predicate*, or* argument*, denoted by σ, π and ρ respectively and defined as:*

$$\begin{aligned}\sigma &:= \iota \mid (\iota \to \sigma)\\ \pi &:= o \mid (\rho \to \pi)\\ \rho &:= \iota \mid \pi\end{aligned}$$

We will use τ to denote an arbitrary type. The binary operator $\to$ is right-associative. It can be easily seen that every predicate type π can be written in the form $\rho_1 \to \cdots \to \rho_n \to o$, $n \geq 0$ (for $n = 0$ we assume that $\pi = o$).

Definition 2. *The* alphabet *of $\mathcal{H}$ consists of the following:* predicate variables *of every predicate type π (denoted by capital letters such as $\mathsf{Q}, \mathsf{R}, \mathsf{S}, \ldots$);* individual variables *of type ι (denoted by capital letters such as $\mathsf{X}, \mathsf{Y}, \mathsf{Z}, \ldots$);* predicate constants *of every predicate type π (denoted by lowercase letters such as $\mathsf{p}, \mathsf{q}, \mathsf{r}, \ldots$);* individual constants *of type ι (denoted by lowercase letters such as $\mathsf{a}, \mathsf{b}, \mathsf{c}, \ldots$);* function symbols *of every functional type $\sigma \neq \iota$ (denoted by lowercase letters such as $\mathsf{f}, \mathsf{g}, \mathsf{h}, \ldots$); the* inverse implication *constant $\leftarrow$; the* negation *constant $\sim$; the comma; and the left and right parentheses.*

Arbitrary variables will usually be denoted by V and its subscripted versions.

Definition 3. *The set of* terms *of $\mathcal{H}$ is defined as follows: every predicate variable (resp., predicate constant) of type π is a term of type π; every individual variable (resp., individual constant) of type ι is a term of type ι; if f is an n-ary function symbol and $\mathsf{E}_1, \ldots, \mathsf{E}_n$ are terms of type ι then $(\mathsf{f}\ \mathsf{E}_1 \cdots \mathsf{E}_n)$ is a term of type ι; if E_1 is a term of type $\rho \to \pi$ and E_2 a term of type ρ then $(\mathsf{E}_1\ \mathsf{E}_2)$ is a term of type π.*

Definition 4. *The set of* expressions *of $\mathcal{H}$ is defined as follows: a term of type ρ is an expression of type ρ; if E is a term of type o then $(\sim\mathsf{E})$ is an expression of type o.*

We will omit parentheses when no confusion arises. To denote that an expression E has type ρ we will often write $\mathsf{E} : \rho$. We will write $vars(\mathsf{E})$ to denote the set

of all the variables in E. Expressions (respectively, terms) that have no variables will be referred to as *ground expressions* (respectively, *ground terms*). Terms of type o will be referred to as *atoms* and expressions of type o will be referred to as *literals*.

Definition 5. *A* clause *of $\mathcal{H}$ is a formula $\mathsf{p}\ \mathsf{E}_1 \cdots \mathsf{E}_n \leftarrow \mathsf{L}_1, \ldots, \mathsf{L}_m$, where p is a predicate constant of type $\rho_1 \to \cdots \to \rho_n \to o$, $\mathsf{E}_1, \ldots, \mathsf{E}_n$ are terms of types $\rho_1, \ldots, \rho_n$ respectively, so that all E_i with $\rho_i \neq \iota$ are distinct variables, and $\mathsf{L}_1, \ldots, \mathsf{L}_m$ are literals. The term $\mathsf{p}\ \mathsf{E}_1 \cdots \mathsf{E}_n$ is called the* head *of the clause and the conjunction $\mathsf{L}_1, \ldots, \mathsf{L}_m$ is its* body. *A* program P *of $\mathcal{H}$ is a finite set of clauses.*

Example 1. The program below defines the `subset` relation over unary predicates:

```
subset S1 S2 ← ∼(nonsubset S1 S2)
nonsubset S1 S2 ← (S1 X), ∼(S2 X)
```

The ground instantiation of a program is described by the following definitions:

Definition 6. *A* substitution *θ is a finite set of the form $\{\mathsf{V}_1/\mathsf{E}_1, \ldots, \mathsf{V}_n/\mathsf{E}_n\}$ where the V_i's are different variables and each E_i is a term having the same type as V_i. The domain $\{\mathsf{V}_1, \ldots, \mathsf{V}_n\}$ of θ is denoted by $dom(\theta)$. If all the terms $\mathsf{E}_1, \ldots, \mathsf{E}_n$ are ground, θ is called a* ground substitution.

Definition 7. *Let θ be a substitution and E be an expression. Then, $\mathsf{E}\theta$ is an expression obtained from E as follows: $\mathsf{E}\theta = \mathsf{E}$ if E is a predicate constant or individual constant; $\mathsf{V}\theta = \theta(\mathsf{V})$ if $\mathsf{V} \in dom(\theta)$, otherwise $\mathsf{V}\theta = \mathsf{V}$; $(\mathsf{f}\ \mathsf{E}_1 \cdots \mathsf{E}_n)\theta = (\mathsf{f}\ \mathsf{E}_1\theta \cdots \mathsf{E}_n\theta)$; $(\mathsf{E}_1\ \mathsf{E}_2)\theta = (\mathsf{E}_1\theta\ \mathsf{E}_2\theta)$; $(\sim\mathsf{E})\theta = (\sim\mathsf{E}\theta)$. If θ is a ground substitution with $vars(\mathsf{E}) \subseteq dom(\theta)$, then the ground expression $\mathsf{E}\theta$ is called a* ground instance *of E.*

Definition 8. *For a program P, we define the* Herbrand universe *for every argument type ρ, denoted by $U_{\mathsf{P},\rho}$ to be the set of all ground terms of type ρ that can be formed out of the individual constants, function symbols, and predicate constants in the program.*

Definition 9. *Let P be a program. A* ground instance of a clause $\mathsf{p}\ \mathsf{E}_1 \cdots \mathsf{E}_n \leftarrow \mathsf{L}_1, \ldots, \mathsf{L}_m$ *of P is a formula $(\mathsf{p}\ \mathsf{E}_1 \cdots \mathsf{E}_n)\theta \leftarrow \mathsf{L}_1\theta, \ldots, \mathsf{L}_m\theta$, where θ is a ground substitution whose domain is the set of all variables that appear in the clause, such that for every $\mathsf{V} \in dom(\theta)$ with $\mathsf{V} : \rho$, $\theta(\mathsf{V}) \in U_{\mathsf{P},\rho}$. The* ground instantiation of a program P, *denoted by $\mathsf{Gr}(\mathsf{P})$, is the (possibly infinite) set that contains all the ground instances of the clauses of P.*

5 The Semantics of $\mathcal{H}$

In [2, 3] M. Bezem developed a semantics for higher-order logic programs that is a generalization of the familiar Herbrand-model semantics of classical (first-order) logic programs. As such, the approach proposes that essentially predicates

are understood as mapping tuples of syntactic objects to truth values. Because of this, the following simplified definition of a higher-order interpretation is possible:

Definition 10. *A* (higher-order) Herbrand interpretation *I of a program* P *is a function which assigns to each ground atom of* $U_{\mathsf{P},o}$*, and to the negation thereof, an element in a specified domain of truth values.*

The truth domain used in [2, 3] is the traditional two-valued one, as only positive programs are studied. In this dissertation we also consider Herbrand interpretations with a three-valued truth domain, i.e. $\{false, 0, true\}$, as well as interpretations with an infinite-valued truth domain, i.e. $\mathbb{V}^{\infty}$.

The concept of "Herbrand model" of a higher-order program can be defined as in classical logic programming.

Definition 11. *Let* P *be a program and I be a Herbrand interpretation of* P*. We say I is a* model *of* P *if* $I(\mathsf{A}) \geq min\{I(\mathsf{L}_1), \ldots, I(\mathsf{L}_n)\}$ *holds for every ground instance* $\mathsf{A} \leftarrow \mathsf{L}_1, \ldots, \mathsf{L}_m$ *of a clause of* P*.*

Bezem's semantics is based on the observation that, given a positive higher-order program, the minimum model of its ground instantiation serves as a Herbrand interpretation for the program itself. We follow the same idea for programs with negation: we can use as an interpretation of a given higher-order program P, the model defined by any semantics that applies to its ground instantiation. It is trivial to see that any such model is also a Herbrand model of P.

In the following sections we focus on the well-founded [11], stable [12] and infinite-valued [18] models. We investigate if these models enjoy the extensionality property, formally defined by Bezem [2, 3] through relations $\cong_{I,\rho}$ over the set of ground expressions of a given type ρ and under a given interpretation I. These relations intuitively express extensional equality of type ρ, in the sense discussed in Section 2. The formal definition is as follows:

Definition 12. *Let I be a Herbrand interpretation for a given program* P*. For every argument type ρ we define the relations $\cong_{I,\rho}$ on $U_{\mathsf{P},\rho}$ as follows. Let* $\mathsf{E}, \mathsf{E}' \in U_{\mathsf{P},\rho}$*; then* $\mathsf{E} \cong_{I,\rho} \mathsf{E}'$ *if and only if: $\rho = \iota$ and* $\mathsf{E} = \mathsf{E}'$*; or $\rho = o$ and* $I(\mathsf{E}) = I(\mathsf{E}')$*; or* $\rho = \rho' \to \pi$ *and* $\mathsf{E}\,\mathsf{D} \cong_{I,\pi} \mathsf{E}'\,\mathsf{D}'$ *for all* $\mathsf{D}, \mathsf{D}' \in U_{\mathsf{P},\rho'}$*, such that* $\mathsf{D} \cong_{I,\rho'} \mathsf{D}'$*.*

Generally, such relations are symmetric and transitive [2, 3] (partial equivalences). Whether they are moreover reflexive (full equivalences), depends on the specific interpretation, which leads to the notion of *extensional interpretation*:

Definition 13. *Let* P *be a program and let I be a Herbrand interpretation of* P*. We say I is* extensional *if for all argument types ρ, $\cong_{I,\rho}$ is reflexive, i.e. for all* $\mathsf{E} \in U_{\mathsf{P},\rho}$*,* $\mathsf{E} \cong_{I,\rho} \mathsf{E}$*.*

6 Extensionality Study of Selected Models

In this section we demonstrate that the adaptation of Bezem's technique under the the two main approaches to negation known from first-order logic programming, i.e. the stable model semantics and the well-founded semantics, do not in

general preserve extensionality. On the other hand, we indicate that the infinite-valued semantics can lead to an extensional semantics for the programs of our higher-order language $\mathcal{H}$.

We begin by showing that the ground instantiation of a $\mathcal{H}$ program may not always have extensional stable models.

Example 2. Consider the program:

```
r Q ← ∼(s Q), ∼(r p)
s Q ← ∼(r Q), ∼(s q)
q a ←
p a ←
```

where, in the first two clauses, `Q` is of type $\iota \to o$. By examining the ground instantiation of the above program, one can see that it has the non-extensional stable model {`(p a)`, `(q a)`, `(s p)`, `(r q)`}. However, it has *no* extensional stable models: there are four extensional interpretations that are potential candidates, namely $M_1 = \{$`(p a)`, `(q a)`$\}$, $M_2 = \{$`(p a)`, `(q a)`, `(r p)`, `(r q)`$\}$, $M_3 = \{$`(p a)`, `(q a)`, `(s p)`, `(s q)`$\}$, and $M_4 = \{$`(p a)`, `(q a)`, `(s p)`, `(s q)`, `(r p)`, `(r q)`$\}$; but none of these models is a stable model of the program. □

The above examples seem to suggest that the extensional approach of [2, 3] is incompatible with the stable model semantics. Unfortunately, the same holds for the well-founded semantics.

Example 3. Consider the higher-order program P:

```
s Q ← Q (s Q)
p R ← R
q R ← ∼(w R)
w R ← ∼R
```

where the predicate variable `Q` is of type $o \to o$ and the predicate variable `R` is of type o. It is not hard to see that the predicates `p` : $o \to o$ and `q` : $o \to o$ represent the same relation, namely $\{(v, v) \mid v \in \{false, 0, true\}\}$.

Consider the predicate `s` : $(o \to o) \to o$. By taking the ground instances of the clauses involved, it is easy to see that the atom `(s p)`, under the well-founded semantics, will be assigned the value *false*. On the other hand, `(s q)` is assigned the value 0, under the well-founded semantics, since the ground instances of the relevant clauses form a circular definition involving negation. In other words, `p` and `q` are extensionally equal, but `(s p)` and `(s q)` have different truth values.

The above discussion is based on intuitive arguments, but it is not hard to formalize it and obtain the following lemma:

Lemma 1. *The well-founded model $\mathcal{M}_{\mathsf{P}}$ of the program of Example 3, is not extensional.*

In light of the above negative results, the next theorem suggests that the infinite-valued model adaptation of Bezem's technique is a more suitable candidate for capturing the extensional semantics of general $\mathcal{H}$ programs.

Theorem 1. *The infinite-valued model of every program of $\mathcal{H}$ is extensional.*

A question that arises is whether there exists a broad class of programs that are extensional under the well-founded semantics. The next section answers exactly this question.

7 Extensionality of Stratified Programs

In this section we argue that the well-founded model of a *stratified* higher-order program [16] enjoys the extensionality property. In the following definition, a predicate type π is understood to be *greater than* a second predicate type π', if π is of the form $\rho_1 \to \cdots \to \rho_n \to \pi'$, where $n \geq 1$.

Definition 14. *A program* P *is called* stratified *if and only if it is possible to decompose the set of all predicate constants that appear in* P *into a finite number* r *of disjoint sets (called* strata*)* $S_1, S_2, \ldots, S_r$*, such that for every clause* $\mathsf{H} \leftarrow \mathsf{A}_1, \ldots, \mathsf{A}_m, {\sim}\mathsf{B}_1, \ldots, {\sim}\mathsf{B}_n$ *in* P*, where the predicate constant of* H *is* p*, we have:*

1. *for every* $i \leq m$*, if* A_i *starts with a predicate constant* q*, then stratum*$(\mathsf{q}) \leq$ *stratum*(p)*;*
2. *for every* $i \leq m$*, if* A_i *starts with a predicate variable* Q*, then for all predicate constants* q *that appear in* P*, such that the type of* q *is greater than or equal to the type of* Q*, it holds stratum*$(\mathsf{q}) \leq$ *stratum*(p)*;*
3. *for every* $i \leq n$*, if* B_i *starts with a predicate constant* q*, then stratum*$(\mathsf{q}) <$ *stratum*(p)*;*
4. *for every* $i \leq n$*, if* B_i *starts with a predicate variable* Q*, then for all predicate constants* q *that appear in* P*, such that the type of* q *is greater than or equal to the type of* Q*, it holds stratum*$(\mathsf{q}) <$ *stratum*(p)*;*

where stratum$(\mathsf{r}) = i$ *if* r *belongs to* S_i*.*

Evidently, the stratification for classical logic programs [1] is a special case of the above definition.

Example 4. It is straightforward to see that the program:

```
p Q ← ∼(Q a)
q a ←
```

is stratified. However, it is easy to check that the program:

```
p Q ← ∼(Q a)
q a a ← p (q a)
```

is not stratified because if the term `(q a)` is substituted for `Q` we get a circularity through negation. The type of `q` is $\iota \to \iota \to o$ and it is greater than the type of `Q` which is $\iota \to o$.

As it turns out, stratified higher-order logic programs have an extensional two-valued well-founded model.

Theorem 2. *The well-founded model* $\mathcal{M}_{\mathsf{P}}$ *of a stratified program* P *is extensional and does not assign the value* 0*.*

8 The Restrictions of 3-Valued Approaches

In this section we re-examine the counterexample of Example 3 but now from a broader perspective. In particular, we indicate that in order to achieve an extensional three-valued semantics for higher-order logic programs with negation, one has to make some strong assumptions regarding the behaviour of negation in such programs.

Under the infinite-valued adaptation of Bezem's approach and also under the domain-theoretic infinite-valued approach of [4], the semantics of that program *is* extensional. The reason is that both of these approaches differentiate the meaning of `p` from the meaning of `q`. Under the truth domain in both approaches, i.e. $\mathbb{V}^{\infty}$, predicate `p` corresponds to the infinite-valued relation: $p = \{(v, v) \mid v \in \mathbb{V}^{\infty}\}$ while predicate `q` corresponds to the relation: $q = \{(F_{\alpha}, F_{\alpha+2}) \mid \alpha < \Omega\} \cup \{(0, 0)\} \cup \{(T_{\alpha}, T_{\alpha+2}) \mid \alpha < \Omega\}$ where Ω is the first uncountable ordinal. Obviously, the relations p and q are different as sets and therefore it is not a surprise that under both the infinite-valued adaptation of Bezem's semantics presented in this dissertation and the semantics of [4], the atoms `(s p)` and `(s q)` have different truth values.

Assume now that we want to devise an (alternative to the well-founded extension of Bezem's semantics presented in this dissertation) extensional three-valued semantics for $\mathcal{H}$ programs. Under such a semantics, it seems reasonable to assume that `p` and `q` would correspond to the same three-valued relation, namely $\{(v, v) \mid v \in \{\mathit{false}, 0, \mathit{true}\}\}$. Notice however that `p` and `q` are expected to have a different *operational* behaviour. In particular, given the program:

```
s Q ← Q (s Q)
p R ← R
```

we expect the atom `(s p)` to have the value *false* (due to the circularity that occurs when we try to evaluate it), while given the program:

```
s Q ← Q (s Q)
q R ← ∼(w R)
w R ← ∼R
```

we expect the atom `(s q)` to have the value 0 due to the circularity through negation. At first sight, the above discussion seems to suggest that a three-valued extensional semantics for all higher-order logic programs with negation is not possible.

However, the above discussion is based mainly on our experience regarding the behaviour of first-order logic programs with negation. One could advocate a semantics under which `(s q)` will also return the value *false*, arguing that the definition of `q` uses two negations which cancel each other. This cancellation of double negations is not an entirely new idea; for example, for certain extended propositional programs, the semantics based on approximation fixpoint theory has the same effect (see for example Denecker et al. [10][page 185,

Example 1]). We have recently developed such an extensional three-valued semantics for higher-order logic programs with negation, using an approach based on approximation fixpoint theory in [8].

9 Conclusions and Future Work

We have for the first time adapted Bezem's technique to define an extensional semantics for higher-order programs with negation, achieved through the infinite-valued approach [18]. On the other hand, we have shown that an adaptation of the technique under the well-founded or the stable model semantics does not in general lead to an extensional semantics. Finally, we have defined the notions of stratification and local stratification and proven that the class of stratified programs is a notable exception to the previous negative result.

It poses an interesting open question whether the class of *locally stratified higher-order logic programs* (see [16]) is well-behaved with respect to extensionality, or not. Another matter worth looking into is the relationships between the infinite-valued extension of Bezem's semantics presented in this dissertation and its domain theoretic counterpart. The most intriguing question, perhaps, is the comparative evaluation of the infinite-valued extensions of Bezem's semantics and the domain theoretic semantics against the three-valued domain theoretic semantics of [8].

References

1. Krzysztof R. Apt, Howard A. Blair, and Adrian Walker. Towards a theory of declarative knowledge. In Jack Minker, editor, *Foundations of Deductive Databases and Logic Programming*, pages 89–148. Morgan Kaufmann, 1988.
2. Marc Bezem. Extensionality of simply typed logic programs. In Danny De Schreye, editor, *Logic Programming: The 1999 International Conference, Las Cruces, New Mexico, USA, November 29 - December 4, 1999*, pages 395–410. MIT Press, 1999.
3. Marc Bezem. An improved extensionality criterion for higher-order logic programs. In Laurent Fribourg, editor, *Computer Science Logic, 15th International Workshop, CSL 2001. 10th Annual Conference of the EACSL, Paris, France, September 10-13, 2001, Proceedings*, volume 2142 of *Lecture Notes in Computer Science*, pages 203–216. Springer, 2001.
4. Angelos Charalambidis, Zoltán Ésik, and Panos Rondogiannis. Minimum model semantics for extensional higher-order logic programming with negation. *TPLP*, 14(4-5):725–737, 2014.
5. Angelos Charalambidis, Konstantinos Handjopoulos, Panos Rondogiannis, and William W. Wadge. Extensional higher-order logic programming. *ACM Trans. Comput. Log.*, 14(3):21, 2013.
6. Angelos Charalambidis, Panos Rondogiannis, and Ioanna Symeonidou. Equivalence of two fixed-point semantics for definitional higher-order logic programs. In Ralph Matthes and Matteo Mio, editors, *Proceedings 10th International Workshop on Fixed Points in Computer Science, (FICS 2015), Berlin, Germany, September 11-12, 2015*, volume 191 of *EPTCS*, pages 18–32, 2015.

7. Angelos Charalambidis, Panos Rondogiannis, and Ioanna Symeonidou. Equivalence of two fixed-point semantics for definitional higher-order logic programs (extended version of conference paper). *Theoretical Computer Science*, 668:27–42, 2017.
8. Angelos Charalambidis, Panos Rondogiannis, and Ioanna Symeonidou. Approximation fixpoint theory and the well-founded semantics of higher-order logic programs (*in press*). *Theory and Practice of Logic Programming*, arXiv:1804.08335, 2018, 2018. (To be presented at the 34th International Conference on Logic Programming (ICLP 2018), Oxford, UK, July 14 - 17 2018).
9. Weidong Chen, Michael Kifer, and David Scott Warren. Hilog: A foundation for higher-order logic programming. *The Journal of Logic Programming*, 15(3):187 – 230, 1993.
10. Marc Denecker, Maurice Bruynooghe, and Joost Vennekens. Approximation fixpoint theory and the semantics of logic and answers set programs. In Esra Erdem, Joohyung Lee, Yuliya Lierler, and David Pearce, editors, *Correct Reasoning - Essays on Logic-Based AI in Honour of Vladimir Lifschitz*, volume 7265 of *Lecture Notes in Computer Science*, pages 178–194. Springer, 2012.
11. Allen Van Gelder, Kenneth A. Ross, and John S. Schlipf. The well-founded semantics for general logic programs. *J. ACM*, 38(3):620–650, 1991.
12. Michael Gelfond and Vladimir Lifschitz. The stable model semantics for logic programming. In Robert A. Kowalski and Kenneth A. Bowen, editors, *Logic Programming, Proceedings of the Fifth International Conference and Symposium, Seattle, Washington, August 15-19, 1988 (2 Volumes)*, pages 1070–1080. MIT Press, 1988.
13. Dale Miller and Gopalan Nadathur. *Programming with Higher-Order Logic*. Cambridge University Press, New York, NY, USA, 1st edition, 2012.
14. Panos Rondogiannis and Ioanna Symeonidou. Extensional semantics for higher-order logic programs with negation. In Loizos Michael and Antonis C. Kakas, editors, *Logics in Artificial Intelligence - 15th European Conference (JELIA 2016), Larnaca, Cyprus, November 9-11, 2016, Proceedings*, volume 10021 of *Lecture Notes in Computer Science*, pages 447–462, 2016.
15. Panos Rondogiannis and Ioanna Symeonidou. The intricacies of three-valued extensional semantics for higher-order logic programs. *Theory and Practice of Logic Programming*, 17(5-6):974–991, 2017. (Presented at the 33rd International Conference on Logic Programming (ICLP 2017), Melbourne, Australia, August 28 - September 1 2017. **Best Paper Award**).
16. Panos Rondogiannis and Ioanna Symeonidou. Extensional Semantics for Higher-Order Logic Programs with Negation (Extended version of conference paper). *Logical Methods in Computer Science*, Volume 14, Issue 2, July 2018.
17. Panos Rondogiannis and Ioanna Symeonidou. The intricacies of three-valued extensional semantics for higher-order logic programs (*in press*). In *27th International Joint Conference on Artificial Intelligence (IJCAI 2018) ("Best Sister Conferences" track), Stockholm, Sweden, July 13-19, 2018, Proceedings*, 2018.
18. Panos Rondogiannis and William W. Wadge. Minimum model semantics for logic programs with negation-as-failure. *ACM Trans. Comput. Log.*, 6(2):441–467, 2005.
19. William W. Wadge. Higher-order horn logic programming. In Vijay A. Saraswat and Kazunori Ueda, editors, *Logic Programming, Proceedings of the 1991 International Symposium, San Diego, California, USA, Oct. 28 - Nov 1, 1991*, pages 289–303. MIT Press, 1991.