MECHANICAL ENGINEERING | PHYSICS | PRESERVATION OF THE ARCHITECTURAL HERITAGE | STRUCTURAL, SEISMIC AND GEOTECHNICAL ENGINEERING | URBAN PLANNING, DESIGN AND POLICY | AEROSPACE ENGINEERING | ARCHITECTURE, BUILT ENVIRONMENT AND CONSTRUCTION ENGINEERING | ARCHITECTURAL, URBAN AND INTERIOR DESIGN | BIOENGINEERING | DATA ANALYTICS AND DECISION SCIENCES | DESIGN | ELECTRICAL ENGINEERING | ENERGY AND NUCLEAR SCIENCE AND TECHNOLOGY | ENVIRONMENTAL AND INFRASTRUCTURE ENGINEERING | INDUSTRIAL CHEMISTRY AND CHEMICAL ENGINEERING | INFORMATION TECHNOLOGY | MANAGEMENT ENGINEERING | MATERIALS ENGINEERING | MATHEMATICAL MODELS AND METHODS IN ENGINEERING

PhD Yearbook | 2023

DOCTORAL PROGRAM IN INFORMATION TECHNOLOGY

Chair:
Prof. Luigi Piroddi

Introduction
The Ph.D. programme in Information Technology (Ph.D. IT) covers research topics in four scientific areas, associated to different facets of the field of Information and Communication Technology, namely Computer Science and Engineering, Electronics, Systems and Control, and Telecommunications. This broad variety of research topics perfectly captures the core mission of the corresponding sections of the Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB).
New research topics and cross-areas research fields are also covered, such as machine learning, big data, intelligent data analysis, Industry 4.0, internet of things, bioinformatics, quantum computing, ecology, environmental modelling, operations research, and transportation systems.
The Ph.D. IT programme is the largest at the Politecnico di Milano in terms of number of students. Every year, more than 90 new students join the programme, for an overall number of around 350.

Topics
Research at DEIB in the field of Information Technology is supported by 35 laboratories, and is organized in 4 main areas.

- Computer Science and Engineering (Vice-Chair: Prof. Francesco Amigoni):
  Information systems, Database management, Information design for the web, Methods and applications for interactive multimedia, Embedded systems design and design methodologies, Dependable systems, Cybersecurity, Autonomous robotics, Artificial intelligence, Computer vision and image analysis, Machine learning, Dependable evolvable pervasive software engineering, Compiler technology, Natural language processing and accessibility.
- Electronics (Vice- Chair: Prof. Angelo Geraci):
  Circuits and systems, Single photon detectors and applications, Radiation detectors and low noise electronics, Electronic circuit design, Electron devices.
- Systems and Control (Vice-Chair: Prof. Lorenzo Fagiano):
  Control theory and its applications, Robotics and industrial automation, Dynamics of complex systems, Planning and management of environmental systems, Operations research and discrete optimization.
- Telecommunications (Vice-Chair: Carlo Riva):
  Networking, Applied electromagnetics, Optical communications, Quantum communications, Wireless and space communications, Remote sensing, Signal processing for multimedia and telecommunications.

Industrial collaborations
Due to its intrinsic technological nature, the Ph.D. IT programme features many industrial collaborations. More than 50% of the Ph.D. candidates are funded by companies or by international research projects involving industrial partners. Indeed, the Ph.D. School envisions the collaboration between university and companies as the ideal ground to convert invention and scientific research into technological innovation. Nevertheless, alongside applied research projects in collaboration with industrial partners, the programme is also able to preserve a strong characterization in fundamental research.
To monitor the activities and development of the Ph.D. programme, the Faculty Board cooperates with an industrial Advisory Board, composed by members of public and private companies, working in management, production, and applied research. The two boards jointly meet once a year to identify and suggest new emerging research areas and to foster the visibility of the Ph.D. IT programme in the industrial world.

Educational aspects
The teaching organization and the course subjects reflect the scientific interests of DEIB faculties. The curricula include a wide choice of courses (about 20 per year), and more than 30 courses for basic soft and hard skills offered by the Ph.D. School of the Politecnico di Milano.
Access to external courses and summer schools is also encouraged. The challenge is to promote interdisciplinary research while offering advanced help to students to make the best choices according to the regulatory scheme of the programme.
Students must undergo a yearly evaluation of the progress in their research and course work.

Internationalization
Every year, several courses are delivered by visiting professors from prestigious foreign universities. Moreover, the Ph.D. IT programme encourages joint curricula with foreign institutions. The programme has several Double Degree and Joint Degree agreements with institutions from countries in all continents. Every year the programme receives more than 150 applications from foreign countries and about 15% of our selected Ph.D. candidates have applied from outside Italy.

Conclusions
The core mission of the Ph.D. IT programme is to offer an excellent doctoral level curriculum, through high-quality courses, a truly interdisciplinary advanced education, cutting-edge research, and international and industrial collaborations.

PHD BOARD OF PROFESSORS	
Prof. Francesco Amigoni - Vice Chair Computer Science and Engineering	Prof. Ivan Rech
Prof. Cesare Alippi	Prof. Alessandro Sottorcornola Spinelli
Prof. Luciano Baresi	Prof. Lorenzo Fagiano - Vice Chair Systems and Control
Prof. Cinzia Cappiello	Prof. Fabio Dercole
Prof. Nicola Gatti	Prof. Lorenzo Mari
Prof. Davide Martinenghi	Prof. Luigi Piroddi - Chair of the Doctoral Programme
Prof. Simone Garatti	Prof. Andrea Zanchettin
Prof. Maristella Matera	Prof. Carlo Riva - Vice Chair Telecommunications
Prof. Raffaela Mirandola	Prof. Matteo Cesana
Prof. Cristina Silvano	Prof. Paolo Martelli
Prof. Stefano Zanero	Prof. Andrea Monti Guarnieri
Prof. Angelo Geraci - Vice Chair Electronics	Prof. Massimo Tornatore
Prof. Giuseppe Bertuccio	Prof. Giancarlo Ferrigno - Representative from the Bioengineering Area
Prof. Giorgio Ferrari	

PHD ADVISORY BOARD	
Giorgio Ancona	Atos
Matteo Bogana	Cleafy
Mario Caironi	IIT
Paolo Cederle	Everis
Cristina Cremonesi	The European Ambrosetti
Riccardo De Gaudenzi	European Space Agency
Giuseppe Desoli	STMicroelectronics
Alessandro Ferretti	Tre-Altamira
Giuseppe Fogliazza	MCE Srl
Bruno Garavelli	Xnext s.r.l.
Maurizio Griva	Reply SpA
Sabino Illuzzi	Prospera
Renato Lombardi	Huawei Technologies
Renato Marchi	KPMG
Giorgio Parladori	PoliHub (ex Research Program Director, SM Optics srl)
Francesco Prelz	INFN
Enrico Ragaini	ABB S.p.A.
Paolo Giuseppe Ravazzani	CNR
Dario Regazzoni	Amazon Web Services (AWS) Italy
Carlo Sandroni	RSE S.p.A.
Massimo Valla	TIM
Luisa Venturini	Vodafone Italy
Stefano Verzura	Huawei Technologies
Roberto Villa	IBM Italy

Prizes and awards

In 2022 the following awards have been obtained by Ph.D. candidates:

- Autonomous Challenge @ CES, 1st Place – **Stefano Carnier, Luca Franceschetti, Sara Furioli, Alex Gimondi, Alberto Lucchini, Gianluca Papa, Filippo Parravicini, Solomon Pizzocaro, Stefano Raddrizzani, Giorgio Riva**
- 24th International Symposium on Formal Methods, Best Presentation Award – **Livia Lestingi**
- IEEE INERTIAL 2022, First Runner Up Student Paper Award – **Andrea Buffoli, Sarah Solbiati**
- SEAMS 2022 – 17th Symposium on Software Engineering for Adaptive and Self-Managing Systems, Best Paper Award – **Davide Yi Xian Hu, Luca Terracciano**
- Dimitris N. Chorafas Foundation Award – **Fabio Garzetti, Marco Manzoni**
- "Prof. Florian Daniel" PhD Thesis Award – **Micol Spitale, Emanuele Vitale**
- 17th Conference on Ph.D Research in Microelectronics and Electronics (PRIME), Gold Leaf Award – **Mauro Leoncini**
- 2022 Fabrizio Flacco Award – **Davide Bazzi**
- Springer Award – **Luca Buonanno, Michele Chiari, Luca Comanducci, Simone Disabato, Fabio Garzetti, Federica Guerrini, Marco Manzoni, Costanza Messeri, Francesco Picetti, Giorgia Ramponi, Marco Tangi, Michele Zanella**

ON THE EFFECTIVENESS OF NEIGHBORHOOD-BASED MODELS IN RECOMMENDER SYSTEMS

Cesare Bernardis - Supervisor: Prof. Paolo Cremonesi

In the history of Recommender Systems (RS), Neighborhood-based (NB) models played a crucial role in laying the foundation for the tackling of the recommendation problem. NB approaches are based on the perception that users with similar tastes tend to interact with akin items: a similarity measure, based on a heuristic or another more powerful techniques, is used to assess the commonalities between users and items, and to predict future user preferences.
Despite the simplicity of the concept that supports them, NB approaches turned out to be particularly effective since the early stages of the RS era.
More recently, thanks to the high relevance in the industrial environment of the recommendation problem, the interest on the research in the RS field has grown wider. This mainly led to the study of gradually more sophisticated and powerful models for the prediction of user preferences, and the consequent generation of recommendation lists. NB models lost their attractiveness in favor of more innovative solutions: Machine Learning (ML) first and Deep Learning afterwards took the main stage in the RS literature, promising to outperform the accuracy of classical methods as happened in various other application areas like computer vision, or natural language processing.
In the last few years, researchers noticed that the apparent continuous improvements to the state-of-the-art in Information Retrieval was often related to the usage of weak baselines and poor evaluation procedures. Consequently, the community started re-examining more classical algorithms for the recommendation task, and it was noticed that NB models still represent a valid alternative to more recent approaches for top-N recommendation in many scenarios.
NB techniques present complementary properties with those of modern and complex approaches. While ML and DL techniques can usually boast superior accuracy thanks to their enhanced prediction capabilities, NB models show a number of diverse advantages that go beyond the pure accuracy performance. (i) NB approaches are usually simpler models that do not require resource-wise expensive training phases and have a fast inference, granting, at the same time, a strong generalization potential. (ii) The recommendations produced by NB methods are easily justifiable, and the explainability of the outcome of a model is a fundamental characteristic of its success. (iii) NB algorithms result in stable models that base their single predictions on a considerable amount of information, and they are usually not subject to any source of randomness. From the user point of view, stability is a symptom of a trustworthy system. (iv) NB models are very useful and effective in transferring knowledge, as their performance in cross-domain recommendation demonstrates.
The potential of DL, and of ML in general, is undebatable, but it is hard to be exploited, and the pure accuracy is not the only relevant aspect of a model. The aim of this thesis is to show that it is possible to merge NB and ML to get the best from both worlds, exploiting their specific qualities. We want to show that the strengths of NB can be helpful in fulfilling several weaknesses of ML, in a field of application, like RS, that entails many challenges (highly sparse datasets affected by a number of different biases, high complexity of the recommendation task, etc.).
On the other hand, ML can boost the accuracy performance of NB models, aiding them to overtake their limits.
Only a few attempts to hybridize NB and ML models have been made in the past, and most of them focus on the common top-N recommendation problem. In this direction, we want to progress the current state-of-the-art with a threefold contribution on diverse aspects of recommendation.
The first aspect is confidence estimation in item-based models. We focus on the estimation of user-level confidence in item-based recommenders, which allows predicting the accuracy of a recommendation model.
A good accuracy predictor is useful for improving the quality of a recommender system under different points of view, including accuracy itself, but also the explainability of the recommendations, and user trust in the system. We highlight the analogy between the preference prediction formula of item-based models and the left eigenvector problem. Exploiting this property, we propose an ML technique to compute a novel estimator of user-level confidence for item-based models called Eigenvalue Confidence Index. Compared to the state-of-the-art, the correlation reported by the ECI is more consistent across datasets and algorithms, and it is often more than 2 times stronger. We also show that the ECI can be adopted to provide recommendations with maximum confidence in expectation, improving the accuracy of item-based techniques up to 20%.
The second contribution regards the study of the stability in Matrix Factorization (MF) models, one of the most famous ML instances for RS. MF is subject to several sources of randomness that heavily impact its optimization procedure. We study one of them in particular, i.e., the initial values of the model parameters, and we show that different initial values lead the model to find different solutions at convergence, resulting in a specific form of instability of its outcome. Since, as we show, stability and accuracy are positively correlated, a high degree of instability determines a degradation of the accuracy performance of the model.
We propose a new framework called Nearest Neighbors Matrix Factorization (NNMF) that generalizes MF to improve its stability. NNMF transfers neighborhood information extracted from NB methods into the embedding space learned by MF. We empirically prove that NNMF doubles the stability degree of common MF under different aspects, from the embedded representations of users and items to the recommendation lists generated, resulting also in higher accuracy especially on less popular items, where the stability of MF is particularly low.
The third and last contribution involves Deep Learning models training through similarities for the item cold-start recommendation. A branch of recent research works has demonstrated that extracting collaborative information from item similarities for the item cold-start recommendation scenario is effective. However, the potential of this new finding has not been explored yet, as only few simple content feature weighting techniques have been proposed. Following-up on this intuition, we explore two new approaches to improve the performance of NB techniques in cold-start scenarios. First, we propose a new feature weighting technique applied to a graph-based model called HP3, which improves the accuracy of common feature weighting techniques while keeping the advantages of a simple model: fast training and low requirements of computational resources.
Second, we propose a more sophisticated model called Neural Feature Combiner (NFC). NFC takes advantage of the expressive power of Deep Learning to train a NB model that is able to combine multiple features applying nonlinearities to represent high level concepts. We prove that NFC outperforms the current state-of-the-art for item cold-start recommendation by 10 to 20%, highlighting its ability to exploit collaborative information more effectively than its competitors in several scenarios.
In conclusion, our thesis demonstrates that if NB and ML are properly merged, the final ensemble has more value than its single components: the strengths of NB models, even if they are not directly bonded to the prediction power, can enhance the models under other aspects that are strictly related to the quality of the recommendations; on the other hand, the predictive accuracy and the adaptability of ML can help to overtake the limits that simple NB techniques inevitably have.

RECONCILING DEEP LEARNING AND CONTROL THEORY: RECURRENT NEURAL NETWORKS FOR MODEL-BASED CONTROL DESIGN

Fabio Bonassi - Supervisor: Prof. Riccardo Scattolini
Co-Supervisor: Prof. Marcello Farina

In recent years, the fruitful exchange of ideas between the control systems community and the deep learning community has paved the way for increasingly powerful and sophisticated data-driven control strategies. Indeed, the wide availability of data, the development of increasingly complex neural network (NN) architectures, and the advances in efficient training strategies and open-source software platforms have fostered the integration of deep learning tools with traditional modeling and control methodologies.

These tools have gained special interest in the context of indirect data-driven control, where NN models can be used for black-box nonlinear system identification with the goal of synthesizing model-based control architectures, thus conjugating the modeling power of NNs with the vast literature on model-based control strategies, such as Model Predictive Control (MPC). For this reason, this paradigm has been successfully adopted by practitioners to solve many challenging engineering problems, such as chemical and pharmaceutical process control, industrial manufacturing plant management, buildings HVAC optimization, optimal microgrid energy management, and many others.

Despite the popularity in practical applications, however, limited research efforts have been devoted to building solid theoretical foundations for the safe, proficient use of NN models for system identification and control. Nonetheless, especially when it comes to Recurrent Neural Networks (RNNs), the safety and robustness of these models to input perturbations is all but guaranteed. The scientific community is thus challenged to reconcile the use of deep learning tools with control theory, to develop theoretically sound methods for learning safe and robust RNN models, and to synthesize model-based control laws with guaranteed closed-loop performances.

This dissertation aims at providing a system-theoretic approach to address these gaps. To this end, we consider the most common RNN architectures for black-box identification, i.e., Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs), and Neural Nonlinear AutoRegressive eXogenous (NNARX) models. The stability properties of these architectures are analyzed adopting notions commonly used for nonlinear systems, such as the Input-to-State Stability (ISS), the Input-to-State Practical Stability (ISPS), and the Incremental Input-to-State Stability (δISS). Sufficient conditions under which the RNN architectures enjoy such stability properties are derived, in the form of nonlinear nonconvex inequalities on the weights of the network. Notably, these conditions can be used either a-posteriori to certify that a given RNN is stable or enforced during the training procedure. In this context, we propose a training procedure based on Truncate Back-Propagation Through Time (TBPTT) paradigm, which yields RNNs that are provenly ISS, ISPS, and δISS. This allows to obtain RNN models that are safe, i.e., not subject to unexpected divergence of states, whose modeling performances are asymptotically independent of initialization, and that enjoy a degree of robustness to input perturbations.

The design of control laws based on these black-box RNN models is another open question in the control systems community, due to the lack of solid theoretical foundations. To address this gap, in this dissertation we demonstrate how the δISS property of the model allows the design of control laws with closed-loop performance guarantees.

The first control architecture proposed in the thesis is a standard state-feedback nonlinear MPC based on the trained RNN model, see Fig. 1. Being the model black-box, a state observer needs to be designed to reconstruct the RNN state from the plant's input-output data. For this scheme, with particular reference to GRU models and Leuenberger-like observer structures, we show that the model's δISS allows *(i)* to recast the observer design problem as a (convex) optimization problem for which a solution is guaranteed to exists, and *(ii)* to compute a minimum prediction horizon above which the nominal closed-loop stability is guaranteed.

An additional MPC architecture is also designed, with the aim of ensuring asymptotic zero-error output tracking of piecewise-constant reference signals.

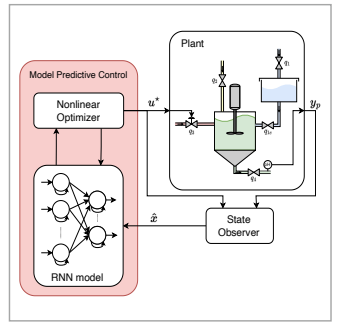

Fig. 1 - RNN-based nonlinear MPC with a state observer in the loop.

Such a strategy builds on the idea of including an integral action on the output tracking error, which allows to attain static performances provided that the closed-loop stability is preserved. The synthesis of such architecture consists in three steps: *(i)* a suitable choice of the integral action's gain, which makes the augmented system local asymptotically stable; *(ii)* the synthesis of a weak detector for the augmented system, i.e., a state observer with nominal exponential convergence guarantees; *(iii)* the design of a stabilizing nonlinear MPC for the augmented system. Notably, the δISS of the RNN model ensures that these three ingredients can be synthesized, yielding an offset-free control architecture with closed-loop stability guarantees.

Motivated by the computational cost of MPC laws, which could be prohibitive for real time control of systems with low sampling times or scarce online computational resources, a third control strategy, based on the Internal Model Control (IMC) approach, is considered. It is therefore proposed to learn an approximation of the inverse of the RNN model through a second RNN, which acts as a controller, and to close the control loop by feeding back the online modeling error. In this context, a controller training procedure, carried out using synthetically generated data sequences, has been proposed. The resulting control scheme allows arbitrary trajectories to be tracked at virtually zero online computational cost and, if both the RNN model and RNN controller are δISS, it allows to achieve performance guarantees. Although the results obtained support the potentialities of RNN-based control design, there are still open problems which are only partially addressed by this work. These include the verification of network safety, i.e., certifying that the output reachable set of the RNN model is included in some "safe set"; the problem of lifelong learning, i.e., how to tune the network based on the plant's operating conditions throughout its lifespan; the design of an MPC regulator that is robust to uncertainty; and finally, physics-based machine learning, i.e., the injection of qualitative physical knowledge of the plant, both in architecture design and network training, to obtain models that are more interpretable, physically consistent, and easy to train. These problems represent fertile ground for future research works, which could foster the use of RNNs for model-based control law design.

DATA DRIVEN AND SIGNAL PROCESSING TECHNIQUES FOR AUDIO FORENSICS

Clara Borrelli - Supervisor: Prof. Augusto Sarti

Digital media has established itself as the dominant communication strategy in nowadays society. In fact, it has been observed that re-posting of news containing video, images or audio is on average 11 times the re-postings of only text news. Social networks and search engines play as news aggregation platforms, often tailoring the recommendations on user's preferences. This affects the possibility to control the news cycle and the diffusion of information is lacking of inter-mediation of professional figures. Therefore, social media helps distributing multi-medial news but, at the same time, veridicity of the content is not guaranteed. In this new information environment, the spreading of falsified media has flourished, and its diffusion is boosted by the "echo chamber" effect. Often, fake news are crafted to damage the reputation of a public personality or institution, while gaining money through advertising, and they represent a serious threat to key areas of our society, like politics or economics. The creation of falsified media is facilitated by the availability of free video and voice editing software and the recent advances of AI techniques make it possible to create deep-fake in completely automatic fashion for all modalities.

In parallel, research in multimedia forensics has proposed several new methods to address the problem and detect falsified media. Impressive progresses have been done in audio, image and video analysis, detecting manipulation using single-modal input or exploiting multi-modal data. Nonetheless, fake media detectors are often challenged by the rapid evolution of attacks and anti-forensics methods. Common limitations are the lack of generalisation ability, necessary to address new synthesis and manipulation attacks, and the lack of robustness to different acquisition conditions or media compression and coding, operations that are frequently applied in social media sharing. For these reasons, recent trends of audio forensics focus on the extraction of high semantic information, like for instance emotion expressed in the speech, rather than analysing the signal at a lower level. The analysis of semantic inconsistencies can be aided with the usage of recent deep learning architectures, exploiting their flexibility and generalisation power. This approach can help understanding not only if the media has been forged, but also having a better insight of how the attack has been executed and what is the purpose of the attacker.

In the presented PhD thesis, we tackle three classic audio forensics topics applying novel methodologies and perspectives. The first problem we address is acoustic condition assessment. The main objective is to estimate from a single-channel audio signal an acoustic indicator able to express the characteristics of an acoustic environment. We do not focus simply on classic acoustic parameter extraction, but we aim at expressing at an higher semantic level the overall acoustic and noise properties of the recording location. This strategy finds several applications in the audio forensics investigation, since it allows to evaluate the authenticity of an audio recording by matching on different acoustic levels the hypothesised recording location of the audio evidence with the actual assessed one.

The second problem we face is synthetic speech detection and attribution for authenticity evaluation. In this case we decline the theme of authenticity verification looking directly at the source signal, i.e., at speech level. We develop different methods to identify and classify synthetic speech samples, taking into account the recent advances in speech synthesis. Regarding the detection problem, we propose two strategies. The first one envisages the use of low-level features, defined starting from the voice source-filter model. The second technique aims at exploiting more high semantic level features, exploiting recent NN architectures that allows to describe emotional and prosodic characteristics of voice. The two methods can be applied for the same task, i.e., voice authenticity verification, but they differ in the complexity and training data required

Finally we address the problem of integrity verification, taking advantage of the descriptors used for authenticity assessment. In particular, we focus on splicing operation detection and localisation. We propose two methods that start from different assumptions on the splicing operation. In the first one we assume that the splicing is a combination of two real recordings performed in two different environments, and we hence exploit reverberation time inconsistencies to detect and localise the splicing point. In the second scenario, we assume that the spliced file is a combination of synthetic and real speech. Therefore, we extract locally a descriptor of the audio signal strictly related to the origin of the speech signal, i.e., if it is real or fake. By looking at the behaviour over time of this representation we are able to spot partially synthetic audio files and locate the point in which the concatenation happened.

From a methodological standpoint, we choose a common framework for all methods. In most scenarios, we take advantage of ML and DL architectures, including both classic data-driven methods and more recent NN architectures. We exploit different NN architectures for extracting meaningful and compact embedding, related to different properties of the input. Usually, these networks take as input a simple time-frequency representation of the input and they are trained to learn a feature space related to a specific contextual attribute, which can range from the acoustic conditions of the environment to the emotional content of the speech. The fast development of new deep architectures and their increasing ability to model high semantic level concepts open up to new exciting solutions to audio forensics applications, preliminary investigated in this thesis. Obviously, deep-learning strategies require the availability of large training data corpora. Whenever this requirement is not met, the audio forensic analyst must rely on different tools, able to operate in limited training data scenario. In this case completely deterministic algorithms or systems based on handcrafted features and classic ML algorithms are preferable, even if less robust to signal-level changes. In the thesis we considered both scenarios, and we therefore propose solutions spanning different abstraction and semantic levels and requiring different resources.

SYNTHESIS OF FILTERS AND FILTERING ANTENNAS FOR MICRO AND MILLIMETER WAVES APPLICATIONS

Steven Kleber Caicedo Mejillones

Supervisors: Prof. Michele D'Amico, Prof. Matteo Oldoni

Filters and antennas are the closest building blocks to the air interface in modern wireless communications systems. Filters allow the transmission of signals in a desired frequency range and eliminate those that operate in the unwanted range. Antennas help radiate signals within their operating range. This thesis focuses on the development of new methods for the synthesis and design of these two building blocks and the integration between them, in other words, filtennas.

On the one hand, with the advent of different 5G solutions and the massive deployment of IoT, the need for highly selective filters also arises. With size as a constraint in many applications, the introduction of several transmission zeros at specific frequencies may be necessary. This conversely often leads to intricate topologies and therefore more complicated implementation. Extracted-pole and cascaded blocks are modular topologies that can overcome all these issues, allowing to design of even fully canonical filters, i.e., with as many transmission zeros as resonators. This thesis proposes novel, accurate and analytical methods for the synthesis of extracted-pole and cascade-block filters including resonating and non-resonating nodes. These methods are based on the well-known coupling matrix synthesis of filters. Then, suitable matrix operations are applied to the synthesized circuit to transform it into the target topology. Regarding the extracted-pole filters, these new methods allow to synthesize filters in a more accurate way compared to the state of the art at the time of publication of this work. Regarding the cascaded-block synthesis, these new methods allow to analytically synthesize filters that previously were only possible with optimization methods such as cascaded-doublets or filters with mixed topology (n-tuplets and extracted-poles) with the blocks arbitrarily placed along the circuit.

On the other hand, filters and antennas are devices that are usually connected between them. The interconnection between them may increase the filter-antenna footprint as a matching network may be needed. If that is the case, this also adds losses to the link budget. This is particularly critical in the millimeter wave frequency range. That is why the rest of the thesis focuses on the synthesis and design of filtering antennas addressing different use cases. Regarding this topic, this thesis first proposes printed circuit board filtering antenna solutions for the customer premises equipment of a fixed wireless access system, as well as for a phased array antenna that uses frequency division duplexing. A circularly polarized horn filtenna for space applications is also proposed (Figure 1). For all these works, a design procedure based on filter synthesis is presented. Synthesis-based designs work on the premise that an equivalent circuit approximates the real model reasonably well. The main advantage of this type of design is that the expected frequency response of the final full-wave prototype is approximated in advance. Therefore, the selection of the best solution that satisfies the given requirements can be done through fast but accurate circuit simulations. When the best solution is found, then the actual full-wave prototype is designed. In this work, this last step is conducted in a modular way, that is, the full wave prototype is designed by blocks according to the synthesized circuit, then all the blocks are assembled. These methodological procedures speed up the design process.

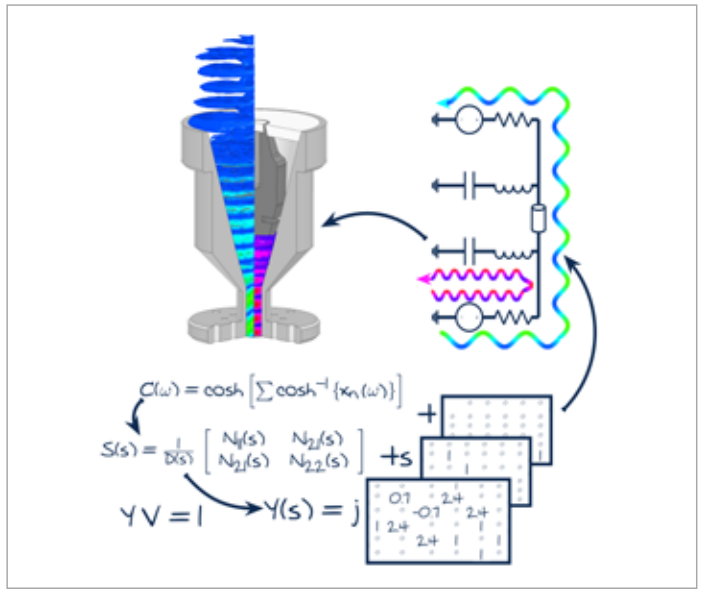

Fig. 1 - A schematic representation of a hierarchical cloud-enabled metro-area network spanning over four network levels

REDUCING THE GAP BETWEEN THEORY AND APPLICATIONS IN ALGORITHMIC BAYESIAN PERSUASION

Matteo Castiglioni - Supervisor: Prof. Nicola Gatti

This thesis focuses on the following question: is it possible to influence the behavior of self-interested agents through the strategic provision of information? This 'sweet talk' is ubiquitous among all sorts of economics and non-economics activities. In this thesis, we model these multi-agent systems as games between an informed sender and one or multiple receivers. We study the computational problem faced by an informed sender that wants to use his information advantage to influence rational receivers with the partial disclosure of information. In particular, the sender faces an information structure design problem that amounts to deciding 'who gets to know what'.

Bayesian persuasion provides a formal framework to model these settings as asymmetric-information games. In recent years, much attention has been given to Bayesian persuasion in the economics and artificial intelligence communities due also to the applicability of this framework to a large class of scenarios like online advertising, voting, traffic routing, recommendation systems, security, and product marketing. However, there is still a large gap between the theoretical study of information in games and its applications in real-world scenarios. This thesis contributes to close this gap along two directions. First, we study the persuasion problem in real-world scenarios, focusing on voting, routing, and auctions. While the Bayesian persuasion framework can be applied to all these settings, the algorithmic problem of designing optimal information disclosure polices introduces computational challenges related to the specific problem under study. Our goal is to settle the complexity of computing optimal sender's strategies, showing when an optimal strategy can be implemented efficiently. Then, we relax stringent assumptions that limit the applicability of the Bayesian persuasion framework in practice. In particular, the classical model assumes that the sender has perfect knowledge of the receiver's utility. We remove this assumption initiating the study of an online version of the persuasion problem. This is the first step in designing adaptive information disclosure policies that deal with the uncertainty intrinsic in all real-world applications.

In the first part of the thesis, we study persuasion in games with structure with a particular focus on voting scenarios. Information is the foundation of any democratic election, as it allows voters for better choices. In many settings, uninformed voters have to rely on inquiries of third party entities to make their decision. For example, in most trials, jurors are not given the possibility of choosing which tests to perform during the investigation or which questions are asked to witnesses. They have to rely on the prosecutor's investigation and questions. The same happens in elections, in which voters gather information from third-party sources. Hence, we pose the question: can a malicious actor influence the outcome of a voting process only by the provision of information to voters who update their beliefs rationally? We study majority voting, plurality voting and district-based elections, showing a sharp contrast in term of efficiency in manipulating elections and computational tractability between the case in which private signals are allowed and the more restrictive setting in which only public signals are allowed. In particular, we show that it is possible to compute an optimal private signaling scheme in polynomial time in all the elections that we considered, while the problem of approximating the optimal public signaling scheme is computationally intractable even for majority voting. Then, we explore how information can be used to reduce the social cost in multi-agents systems, focusing on routing games. In particular, we study Bayesian games where network vagaries are modeled via a (random) state of nature. We focus on the problem of computing optimal ex-ante persuasive signaling schemes, showing that symmetry is a crucial property for its solution. Indeed, we show that an optimal ex-ante persuasive signaling scheme can be computed in polynomial time when players are symmetric. Moreover, the problem becomes computationally intractable when players are asymmetric. Finally, we study persuasion in posted price auctions in which the seller tries to sell an item by proposing take-it-or-leave-it prices to buyers arriving sequentially. Each buyer has to choose between declining the offer or accepting it, thus ending the auction. We study Bayesian posted price auctions, where the buyers valuations for the item depend on a random state of nature, which is known to the seller only. Thus, the seller does not only have to decide price proposals for the buyers, but also how to partially disclose information about the state so as to maximize revenue. Our model finds application in several real-world scenarios. For instance, in an e-commerce platform, the state of nature may reflect the condition (or quality) of the item being sold and/or some of its features. As a first negative result, we prove that, in both public and private signaling, the problem of computing an optimal seller's strategy is computationally intractable. Then, we provide tight positive results by designing a polynomial-time approximation scheme for each setting.

In the second part of the thesis, we initiate the study of Bayesian persuasion with payoff uncertainty. First, we consider the setting with a single receiver and we deal with uncertainty about the receiver's type by framing the Bayesian persuasion problem in an online learning framework. We study a repeated Bayesian persuasion problem where, at each round, the receiver's type is adversarially chosen from a finite set of types. Our goal is the design of an online algorithm that recommends a signaling scheme at each round, guaranteeing an expected utility for the sender close to that of the best-in-hindsight signaling scheme. We rule out the possibility of designing a no-regret algorithm with polynomial per-round running time. Then, we provide two no-regret algorithms for the full and partial information model which require exponential per-round running time. We extend the online Bayesian persuasion framework to include multiple receivers. We focus on the case with no-externalities and binary actions. Moreover, to focus only on the receivers' coordination problem, we overcome the intractability of the single-receiver problem assuming that each receiver has a constant number of types. First, we show that designing no-regret algorithms is computationally intractable when the sender's utility function is supermodular or anonymous. Then, we focus on submodular sender's utility functions and we show that, in this case, it is possible to design a polynomial-time no-(1 - 1/e)-regret algorithm, which is tight.

The design of polynomial-time no-regret algorithms is impossible due to the hardness of the underline offline problems. Hence, the design of efficient algorithms for the offline problem is the bottleneck to the design of efficient online learning algorithms. In the last part of the thesis, we circumvent this issue by leveraging ideas from mechanism design. In particular, we introduce a type reporting step in which the receiver is asked to report her type to the sender. Surprisingly, we prove that, with a single receiver, the addition of this type reporting stage makes the sender's computational problem tractable. Then, we extend our framework to settings with multiple receivers. In such setting, we show that it is possible to find a sender-optimal solution in polynomial-time for supermodular and anonymous sender's utility functions. As for the case of submodular sender's utility functions, we provide a (1 - 1/e)-approximation, which is tight.

CONSTRAINT-AWARE PERFORMANCE AUTOTUNING IN LIVE PRODUCTION ENVIRONMENT

Stefano Cereda - Supervisor: Prof. Paolo Cremonesi

A modern IT system has hundreds of tunable configuration parameters that control its behaviour. Selecting the proper configuration is crucial to improving performance or reducing cost. However, manually finding well-performing configurations can be a daunting task since the parameters often behave in counter-intuitive ways and have inter-dependencies. Furthermore, a modern system sits on top of a complex IT stack that comprises several layers, like the Java Virtual Machine or the Operating System. Each layer has its tunable parameters, which affect the final behaviour of the system. To unlock the full performance potential of a system, we have to tune the entire IT stack jointly.

Unfortunately, we cannot run an extended search and find the optimal configuration which is the best one for our particular stack. Even if we had an infinite budget to run this search, we would still find a suboptimal solution as the optimal configuration depends upon the particular workload to which the system is exposed. We could even imagine running an extensive search to find the optimal configuration for each particular workload or at least a well-performing configuration for each workload. However, all this knowledge would become obsolete quickly, as new software versions are released, changing the effects of the parameters. Furthermore, new software releases also modify the available parameters, increasing the complexity of reusing old knowledge bases, which lack information about novel parameters.

As many IT systems are moving from a monolithic approach to a graph of thousands of microservices, the performance autotuning problem becomes even more complex. The entire system's performance is affected by all the microservices in complex ways, and even the slow-down of a single service can impact the user experience. Add to this that a system could be dependent on a service offered by another system that is not under our control, and this second system can potentially misbehave, slowing down the first system and making the tuning process harder. A proper autotuning system should thus be able to model and cope with these situations.

Finally, we must consider that the final goal of the performance analyst is not just to optimise performance, but to optimise performance while satisfying some constraints. For example, we might be interested in maximising throughput under a budget constraint or in minimising cost while meeting some service level agreements on the response time. The existence of constraints makes the optimisation harder, as we need to understand how the applied configuration affects the constrained metric, while testing configurations that should satisfy the constraints as much as possible. Depending on the tuning scenario, we might be interested in satisfying the constraints only for the current workload, or we might be interested in finding a configuration that satisfies the constraints on all the possible workloads. The same problem applies to the dependency on external services we explained above. Notice that there exists different families of constraint: we have constraints on the applied configurations (e.g., the heap size of the JVM must be smaller than the container memory limit), which we must satisfy in order to provide an applicable configuration, and constraints on the performance metrics resulting from evaluating an applied configuration. Among metric constraints, some are mandatory (e.g., the application has to start successfully), while others can be violated (e.g., an SLA on the response time), even if our goal is to guarantee that they are not violated.

This work hence considers different types of constraints, using different strategies to deal with them, considering constraint satisfaction the first goal of the tuning process. In conclusion, having an autotuner that respects SLA constraints would allow to deploy the autotuner directly in a production environment without resorting to a duplicate performance testing environment. Apart from the cost-saving, this would also allow obtaining a more reliable configuration, as it is tested on the real environment with the real workload, preventing any issue arising from human errors in the environment duplication.

The main goal of this work is to develop an autotuner that is:

- Generic: it works on a wide variety of target applications as it uses a black-box approach without making any assumptions about the underlying system. This avoids the burden of building and maintaining models of the various systems and allows to focus on the desired performance metrics.
- Holistic: in that it simultaneously targets various layers of the IT stack. This allows to explore the inter-dependencies of various parameters and achieve better performance.
- Contextual: the autotuner models the performance as a function of the tunable parameters and other external factors, called context. The context can be used to model the incoming workload, obtaining a workload-aware autotuner, but it can also be used to model external dependencies of the system, such as the response time of an external system that affects our target system but is not under our control.
- Safe: the autotuner is aware of Service Level Agreement (SLA) constraints and tries to suggest configurations that satisfy them. This capability works in conjunction with the Contextual part, as the autotuner can be configured to suggest a configuration that is safe for the current context (e.g., when we try to adapt the configuration to the incoming workload) or a configuration that is safe on all the previously observed contexts (e.g., when we want to take into account the possible slow-down of an external system).

The main research contributions of this work are:

- Introduction of evaluation framework: to compare different tuning approaches, some clearly defined evaluation criteria and metrics are necessary. As we are dealing with noisy performance measurements, it is crucial to discern whether an observed performance improvement is due to a change in the configuration or is just some random fluctuation. To this end, we introduce a set of normalised metrics that are interpretable and easy to understand.
- Context-aware autotuner: as mentioned above, the suggested configuration must be adapted to the context to which the system is exposed. We introduced an autotuner based on Contextual Gaussian Process bandits (CGP) able to deal with context variations, such as the workload.
- Context forecasting: when dealing with a variable context, it is crucial to be able to forecast the value of future contexts to modify the applied configuration proactively. Furthermore, testing a configuration when the context is unstable would lead to noisy measurements. Hence, we introduce a context forecasting module working in tight conjunction with the autotuner and scheduling when to run the experiments.
- SLA-constraints: in real scenarios, optimising a single performance metric is not the actual goal. Instead, it is essential to optimise a metric while meeting a set of Service Level Agreement (SLA) constraints. Violating these constraints, especially in a production environment, leads to economic penalties, which the autotuner must avoid at any cost. We thus modify the Bayesian Optimisation framework to deal with different kinds of constraints.
- Reaction matching characterisation: to speed up the tuning of an application, we can re-use the knowledge collected in the tuning of another application. To transfer knowledge in a helpful way, it is crucial to select relevant applications from an existing knowledge base of previous tuning sessions. To do so, we introduce the Reaction Matching characterisation, which can be used to characterise applications and thus find similarities, which we plan to use as a future extension to add a transfer learning module to the existing autotuner.

MODERN HIGH-LEVEL SYNTHESIS: IMPROVING PRODUCTIVITY WITH A MULTI-LEVEL APPROACH

Serena Curzel - Supervisor: Prof. Fabrizio Ferrandi

The exponential growth of data science and machine learning (ML), coupled with the diminishing performance returns of silicon at the end of Moore's law and Dennard scaling, is leading to widespread interest in domain-specific architectures and accelerators. Field Programmable Gate Arrays (FPGAs) and Application-Specific Integrated Circuits (ASICs) can provide the necessary hardware specialization with higher performance and energy efficiency than multi-core processors or Graphic Processing Units (GPUs). ASICs are the best solution in terms of performance, but they incur higher development costs; FPGAs are more accessible and can be quickly reconfigured, allowing to update accelerators according to the requirements of new applications or to try multiple configurations in a prototyping phase before committing to long and expensive ASIC manufacturing.

ASICs and FPGAs are designed and programmed through hardware description languages (HDLs) such as Verilog or VHDL, which require developers to identify critical kernels, build specialized functional units and memory components, and explicitly manage low-level concerns such as clock and reset signals or wiring delays. The distance between traditional software programming and HDLs creates significant productivity and time-to-market gaps and traditionally required manual coding from expert hardware developers.
The introduction of High-Level Synthesis (HLS) simplified this process, as HLS tools allow to automatically translate general-purpose software specifications, primarily written in C/C++, into an HDL description ready for logic synthesis and implementation. Thanks to HLS, developers can describe the kernels they want to accelerate at a high level of abstraction and obtain efficient designs without being experts in low-level circuit design.

Due to the mismatch between the requirements of hardware descriptions and the characteristics of general-purpose programming languages, HLS tools often require users to augment their input code through pragma annotations (i.e., compiler directives) and configuration options that guide the synthesis process, for example, towards a specific performance-area trade-off. Different combinations of pragmas and options result in accelerator designs with different latency, resource utilization, or power consumption. An exhaustive exploration of the design space does not require extensive changes to the input code, and it does not change the functional correctness of the algorithm, but it is still not a trivial process: the effect of combining multiple optimization directives can be unpredictable, and the HLS user needs a good understanding of their impact on the generated hardware.

Data scientists who develop and test algorithms in high-level, Python-based programming frameworks (e.g., TensorFlow or PyTorch) typically do not have any hardware design expertise: therefore, the abstraction gap that needs to be overcome is not anymore from C/C++ software to HDL (covered by mature commercial and academic HLS tools), but from Python to annotated C/C++ for HLS. The issue is exacerbated by the rapid evolution of data science and ML, as no accelerator can be general enough to support new methods efficiently, and a manual translation of each algorithm into HLS code is highly impractical.

This thesis proposes a multi-level, compiler-based approach to bridge the gap between high-level frameworks and HLS.
The key enabling technology is the Multi-Level Intermediate Representation (MLIR), a reusable and extensible infrastructure in the LLVM project for the development of domain-specific compilers. MLIR allows defining specialized intermediate representations (IRs) called dialects to implement analysis and transformation passes at different levels of abstraction, and it can interface with multiple software programming frameworks.
An MLIR-based approach is a "modern" solution to automate the design of hardware accelerators for high-level applications through HLS, as opposed to "classic" approaches that rely on hand-written template libraries.

A practical realization of the proposed approach is the SOftware Defined Architectures (SODA) Synthesizer, an open-source hardware compiler composed of an MLIR frontend and an HLS backend (Figure 1). SODA provides an end-to-end agile development path from high-level software frameworks to FPGA and ASIC accelerators, supports the design of complex systems, and allows to introduce and explore optimizations at many different levels of abstraction. The integration of an open-source tool in the backend allows to exploit years of HLS research and to introduce new features in the low-level hardware generation steps when necessary.

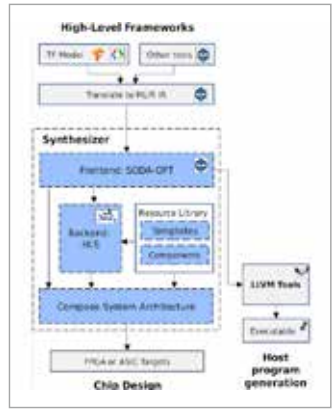

Fig. 1 - The SODA Synthesizer

The proposed design flow allows to implement and apply high-level optimizations before HLS, as compiler passes supported by dedicated MLIR abstractions; such an approach can improve productivity, performance, and portability of optimizations.
A new implementation of loop pipelining based on the MLIR affine dialect has been introduced as a case study, to test whether high-level compiler transformations can benefit HLS results without needing to modify the HLS tool itself. The proposed implementation can analyze dependencies between operations in the loop body and overlap the execution of iterations to increase parallelism; it can also forward results from one iteration to the other, support loops with variable bounds, and speculate execution of if-else blocks. All these transformations contribute to increasing the performance of the generated accelerators, and, since they do not introduce tool-specific annotations or code patterns, they also allow portability across different HLS tools.

Moreover, the SODA Synthesizer integrates a low-level synthesis methodology for the generation of coarse-grained, dynamically scheduled dataflow architectures with distributed control which are particularly suited to support streaming execution; analysis and transformation passes in the MLIR frontend support the low-level synthesis process and improve its results.

Finally, a multi-level compiler-based framework can adapt more easily to innovative input algorithms and hardware targets with respect to tools that generate code for HLS through a library of annotated C/C++ templates. For example, spiking neural networks are built of biologically-inspired integrate-and-fire neurons, and they are usually mapped on analog neuromorphic hardware; a new MLIR dialect has been introduced in the SODA Synthesizer to support the synthesis of SNN models into neuromorphic components. The dialect models concepts from the analog domain of spiking neurons through new types and operations that describe sequences of current spikes as lists of timestamps signaling their arrival.

SPAD-BASED INSTRUMENTATION FOR SINGLE-PHOTON TIMING AND COUNTING APPLICATIONS

Iris Cusini - Supervisor: Prof. Franco Zappa

The research presented in this dissertation thesis has been developed within the SPADlab, a research group at Politecnico di Milano mainly focused on Single-Photon Avalanche Diodes (SPADs), from the device design to in-field exploitation. SPADs have gained popularity over other single-photon detectors especially for the possibility to be fast gated (to time filter the incoming signal) and to precisely timestamp the detected photons (to measure their arrival time). The ability to detect single photons is a key feature in an increasing number of fields. Indeed, its scope is not limited to specific applications relying on single photons, such as quantum imaging, but extends to applications where a low signal is overwhelmed by background light, such as outdoor laser ranging, or in which faint excitation light is required not to damage a biological sample or harm a patient. After the implementation of the first SPADs in standard CMOS technology, researchers have been focusing on the design of large digital SPAD imagers. The front-end and processing electronics are following this path too, moving from off-chip data post-elaboration to progressively 'smarter' sensors including on-chip time-stamping and processing capabilities. During my research project, I have designed various electronic modules aiming at testing, characterizing, and exploiting several silicon multi-pixel SPAD chips with the main goal of acquiring optical signals through single-photon counting (SPC), time-correlated single-photon counting (TCSPC), and photon-coincidence detection. This thesis work introduces the characteristics and figures of merit of SPADs and SPAD arrays, with particular emphasis on their applications. After a general discussion about the FPGA-based modules for SPAD detectors and the methodologies and instrumentation applied to characterize the devices, six novel SPAD cameras are presented. These modules, conceived for portability and versatility, are stand-alone systems based on FPGA and USB 3.0 links to a laptop. They have shown remarkable improvements with respect to previous systems, in terms of programmability, stability, data transfer, power consumption, and heat dissipation. The performance of the chips, core of the modules, are reported and compared with the state-of-art. These novel SPAD cameras have triggered the interest of different international companies and research centers. In particular, the possibility to time-filter the incoming signal has been proven to be a highly desirable feature in Non-Line of Sight Light Detection and Ranging (NLOS LiDAR), and experiments at the 'Institute of Photonic Sciences' (ICFO), in Barcelona, have successfully proved the advantages of time-gated detectors in a novel Differential Interference Contrast (DIC) microscopy scheme.
Two of the modules have been designed in the framework of the 'Q-MIC' European Horizon 2020 FET project. Their design has been optimized for quantum microscopy: high-detection efficiency, low noise, and fast readout. One of the two imagers introduces a novel detection scheme based on an event-driven read-out that allows achieving 100% duty cycle and it is easily scalable to higher resolution chips. However, due to issues in the chip electronics, the maximum detectable photon coincidence rate that we could demonstrate is limited to 1 kpair/s, while the theoretical saturation level for the event-driven architecture is around 3 Mpair/s. This value should be reached with a second version of the chip, with bug fixing.
On the other hand, the second chip employs a more classical frame-based approach, with a row-skipping readout to optimize the frame rate. Measurements carried out by our partners at ICFO have proven the ability of this chip, when combined with a novel entangled photon source at 532 nm, to provide images of entangled photons roughly 5 orders of magnitude faster than what previously reported in literature.
The same camera has been also employed in quantum setups for weak measurements at 'Istituto Nazionale di Metrologia' (INRIM), in Turin, successfully proving the advantages of using these specific modules instead of general-purpose photon detectors.
Although SPADs are mostly exploited in photon-starved applications for their intrinsic high sensitivity, recently they have sparked interest also in high-rate scenarios since they can provide a wider dynamic range with respect to standard optical sensors.
Thus, the SPADlab group has developed four different pixels with short dead time and low afterpulsing probability. Indeed, the intrinsic dead time, which limits the SPAD maximum count rate, hence the dynamic range, is one of the main drawbacks of SPADs with respect to linear detectors. This work presents the characterization of these chips, with a specific focus on high photon rate scenarios, eventually proving short and stable dead time with negligible afterpulsing probability. The scope of these chips is to broaden SPADs exploitations into applications where at present their use is still limited, such as quantum cryptography and computing, and single-photon imaging, all of which require high detection rates. Moreover, in collaboration with 'Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS), Fondazione Don Carlo Gnocchi' and 'Istituto Auxologico Italiano', we have developed a wearable device based on one of the above-mentioned high-count rate pixels. This device employs photoplethysmography (PPG), a technique that is typically used for detecting blood volume changes in the microvascular bed of tissue by exploiting low-intensity light traveling through biological tissues, and it operates in transmission mode, which is considered in literature more robust to disturbances than reflection mode. Our aim is to monitor blood oxygenation in patients affected by obstructive sleep apnea while having concurrently a PPG signal not affected by variation in oxygenation. We have developed a first prototype based on a SPAD chip to prove SPAD suitability in contact PPG. The device has been tested by performing various protocols on volunteers and showed reliable continuous measurements. Moreover, in order to ascertain the pros and cons of using a SPAD instead of photodiodes (detector of choice in PPG devices), we have also developed a second prototype based on photodiodes. Validation tests on volunteers proved the suitability of SPADs in contact PPG and the possibility of concurrently extracting SpO2 percentage and PPG traces independent of oxygen variations. Compared to traditional photodiodes, SPADs have higher sensitivity, are more robust to electronic noise, do not need analog front-ends, and are more suitable for miniaturization. This may lead to innovative applications, such as implanted PPG systems for long-term monitoring of SpO2 and pulse velocity.

HIGH PERFORMANCE RESONANT SWITCHED CAPACITOR CONVERTER (RESCC) TOPOLOGY FOR HIGH CONVERSION RATIO DC-DC VOLTAGE CONVERSION

Alessandro Dago - Supervisor: Prof. Salvatore Levantino

In recent years, resonant switched-capacitor converters (ReSCC) proved to be a valid alternative to traditional buck converters, thanks to their key features such as very high power density, high efficiency, and capability of adjusting the output voltage through quasi-resonant operation.
However, since the components count increases with the conversion ratio, these advantages are usually limited to ReSCC topologies providing 2-to-1 nominal down-conversion. This thesis presents a novel DC-DC ReSCC topology derived from ladder structure with reduced number of components for nominal 4-to-1 voltage conversion. Two prototype converters have been implemented and derived from the proposed topology:
The first use case scenario is related to a 12V to point-of-load down converter. Thel 4-to-1 conversion ratio of the proposed topology allows to obtain a nominal output voltage of 3V, however, quasi-resonant regulation has been introduced to finely regulate the output voltage in the 2.5 to 3V range. The prototype has been partially integrated in 180nm BCD technology. Power MOSFETs (rated for 5V and 8V, which is lower than 12V input voltage thanks to the intrinsic voltage scaling of ReSCC converters), drivers, anti-cross conduction logic, zero-current detector and startup circuitry are all integrated in the ASIC. The 6 flying capacitors composing the converter are MLCC components, while the resonant inductors are implemented exploiting parasitic inductances of PCB vias. In Figure 1 it is reported the die micrograph and the board assembly. The limited dimensions of the design, which can provide up to 13W of output power, allows to achieve a state-of-the-art power density of 0.53W/mm2 at a resonant switching frequency of 1.4MHz.

Closed loop output voltage regulation is obtained through quasi-resonant regulation, which is a technique that allows to modify the converter conversion ratio (fixed in traditional SCC and ReSCC converters) by modulation of the resonant tanks current. A novel control scheme for 3-phase regulation has been proposed in this thesis, which is based on a double loop regulator. The first loop is used to directly regulate the output voltage, while the second one, slower, is used to synchronize the PWM control signals with the zero-current-switching of the converter.

Also, a light-load regulation mode has been proposed to achieve fine and efficient output voltage regulation even in no load conditions. Finally, quasi-resonant regulation has been also employed to perform a soft-start of the DC-DC converter.
The proposed control scheme has been implemented on a Xilinx Spartan-7 FPGA for testing purposes, and it feature a digital PWM generator with 52ps of resolution.

The efficiency curves and load transients responses of the implemented converter are reported in Figure 1. The measured peak efficiency is of 94.4%, however, as expected, it drops down when moving away from the nominal output voltage value. The dynamic performances are limited by the saturation of the voltage control loop, however a maximum deviation of 100mV for 1A load transient has been measured.

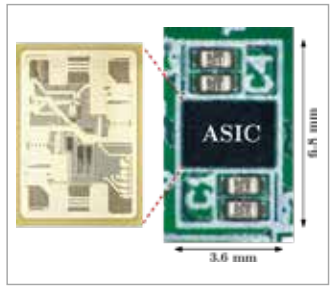

Fig. 1 - PCB assembly and die micrograph of the integrated ReSCC prototype.

The second use case scenario for the proposed topology is the down-conversion of 48V bus voltage towards the regulated high-bandwidth ASIC/CPU/GPU core voltage. This is usually obtained through a two steps conversion, performed by an unregulated intermediate bus converter (IBC) followed by a low voltage multi-phase voltage regulator. Here a high power discrete components unregulated IBC for the 48V to 3.4V down conversion is proposed. The nominal 4-to-1 conversion ratio of the proposed topology has been enlarged to 14-to-1 thanks to the use of a custom planar multi-tapped autotransformer. The hybrid converter obtained this way is characterized by zero-current and zero-voltage switching (ZCS and ZVS), which allows to obtain high conversion efficiency. The transformer copper losses at high current load (140A) have been minimized with the use of a 24 layers/3oz copper PCB stack. However, in order to limit the converter cost, the PCB uses a 12 layers stack, and only on the transformer area, two daughter boards of 6 layers each are soldered together with the main board. The ferromagnetic core is obtained assembling two standard commercial ER cores.
The converter prototype picture, along with the planar transformer assembly and the measured efficiency curve, are reported in Figure 2.

The implemented prototype is characterized by a height of 10.4mm over a total area of 32mm x 52 mm. The maximum output current that can provide to the load is 140A, at an output voltage of 3.14V, leading to a maximum output power of 440W. This way, a power density of 415W/mm3 is obtained at full load. A peak conversion efficiency of 96.3% has been measured at 30A current load, while the efficiency at full load operation, with forced air ventilation, is 91.4%.
Finally, the proposed topology has been validated by two prototypes designed for very different use case scenarios. In both cases optimal efficiency and power density performances have been obtained thanks to the limited number of components of the topology.

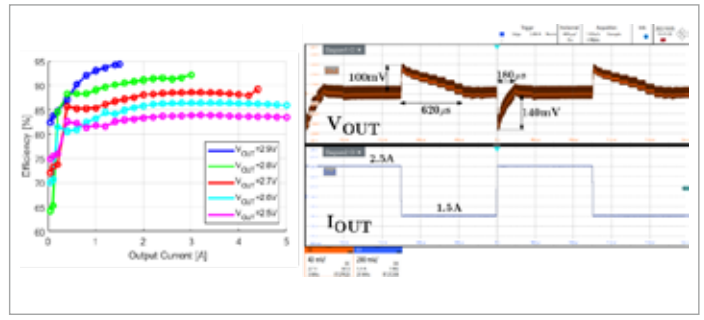

Fig. 2 - Conversion efficiency curves for different output voltages and load transient response.

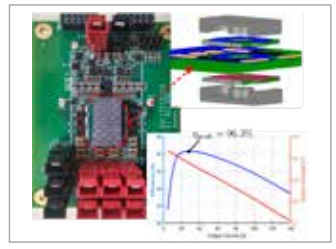

Fig. 3 - 440W prototype for 48-to-3.4V conversion featuring a planar multi-tapped autotransformer, with measured conversion efficiency.

ON HOW TO FACILITATE HARDWARE ACCELERATION OF MACHINE LEARNING FOR NON-EXPERTS IN HARDWARE DESIGN OVER THE EDGE, THE FOG, AND THE CLOUD

Andrea Damiani - Supervisor: Prof. Marco Domenico Santambrogio

Data Scientists cannot ignore the irruption of *Machine Learning (ML)* in their research field. Indeed, models that can learn based on data examples have been gaining increasing attention, fueled by the high Volume, Veracity, Velocity, and Variability of Big Data, because "More Data Beats a Cleverer Algorithm" in *ML*. Nevertheless, *ML* models have the terrible name of neither being explainable nor efficient. Fortunately, balancing these benefits and drawbacks in *ML* modeling has recently become possible, thanks to *Information Technology's (IT)* renaissance of specialization, with *Domain Specific Architectures (DSAs)* increasingly often replacing general-purpose software solutions based on *Central Processing Units (CPUs)* or *general-purpose Graphical Processing Units (gpGPUs)*. Unfortunately, designing, developing, programming, and deploying such *DSAs* on *Application-Specific Integrated Circuit (ASIC)* accelerators or, even worse, *Field Programmable Gate Arrays' (FPGAs) Programmable Logic (PL)* requires hardware design skills that rarely intersect with the area of expertise of Data Scientists and *ML* specialists.

The overarching research question, addressed by this Ph.D. research project, tackles exactly this trade-off and investigates "how to facilitate *ML* experts in the exploitation of the benefits of *DSAs* for hardware acceleration, without having them master hardware design and development." Three main research themes arise from this question: methods and tools for automatic translation of *ML* models into hardware accelerators; methods and tools for increasing the ease of access to the technological platforms enabling hardware acceleration; programming models that support *ML* experts to remain focused on their area of expertise, i.e., the Data, while exploiting hardware acceleration over distributed infrastructures. These themes are collectively targeted by the overall contribution of this research project, summarized in Figure 1.

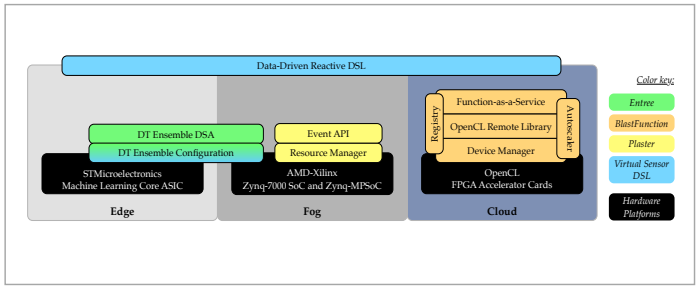

Fig. 1 - Overall contribution of the research project. Each publication is positioned according to the hardware platform it targets and its use in the Edge-Fog-Cloud stack.

The first theme is tackled by Entree, the first toolchain in the State of the Art for deploying large *Decision Tree (DT)* ensembles' inference (an explainable class of *ML* models) over embedded devices mounting an *FPGA*. It automatically converts the scikit-learn trained model to a hardware accelerator. Then, it supports the developer in fitting such accelerated models on embedded *FPGAs* even if they would statically exceed the available resources on the onboard *PL*. This optimization is achieved by employing a novel *DSA* (Figure 2) based on partial dynamic reconfiguration made available by recent advancements in heterogeneous *Systems-on-a-Chip (SoCs)*. Moreover, apart from the increased usability for non-experts in hardware design, Entree attains latency jitters up to hundreds of times lower than those obtained on embedded *CPUs*.

The second theme is addressed by BlastFunction and Plaster, two frameworks for distributing hardware-accelerated algorithms over the Cloud and the Fog, respectively. The former exploits a registry-based OpenCL extension for *FPGA* time-sharing over large Cloud infrastructures. The latter proposes a set of event-driven *Application Programming Interfaces (APIs)* for splitting data-intensive tasks (such as those related to *ML*) over multiple Fog nodes powered by *FPGAs*.

Once Edge, Fog, and Cloud devices have been enabled by the contributions listed so far, the third theme is targeted by the Virtual Sensor *Domain-Specific Language (DSL)* (Figure 3). This extension of the C++ programming language allows Data Scientists to create abstract sensors that measure high-level concepts instead of raw figures. This facilitation is obtained by only focusing at the language level on how the Virtual Sensor should source, aggregate, and process the necessary data, without any knowledge of the actual Edge-Fog-Cloud interface fulfilling it. Thanks to the Virtual Sensor *DSL*, a commercial toolchain for automatic code generation and workload distribution has been developed and released, completing the effort to ease access to hardware acceleration for non-experts in hardware design.

These contributions altogether give a robust and positive answer to the overarching research question that guided the project throughout, paving the way towards the extension of the approach to *ML* models other than *DT* ensembles and other applicative fields apart from the *Internet of Things (IoT)* targeted so far. Furthermore, the toolchains and frameworks presented in this research project are all designed with extensibility and contribution in mind, in an effort to build a solid foundation for bridging the gap between the two fields of expertise of *ML* and hardware acceleration. Finally, a distant goal of the project reaches the *Artificial Intelligence (AI)* world, bringing the same benefits of hardware-accelerating for *ML* to other pillar fields of this rising discipline, such as reasoning and knowledge representation.

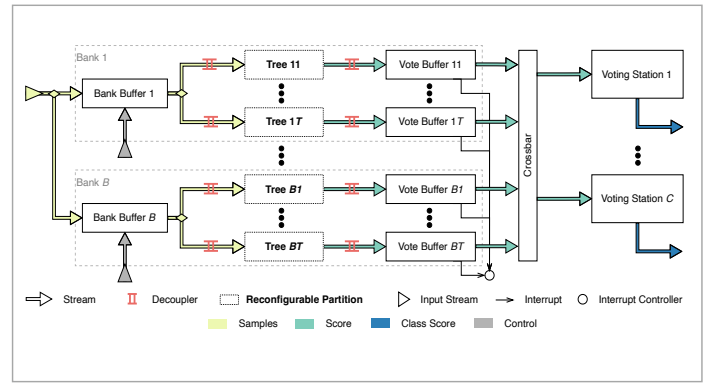

Fig. 2 - Entree template architecture. At its core, the Reconfigurable Partitions that host the Decision Trees to be evaluated, surrounded by the Static Shell providing the data path and control for the co-processor to fulfill a classification task.

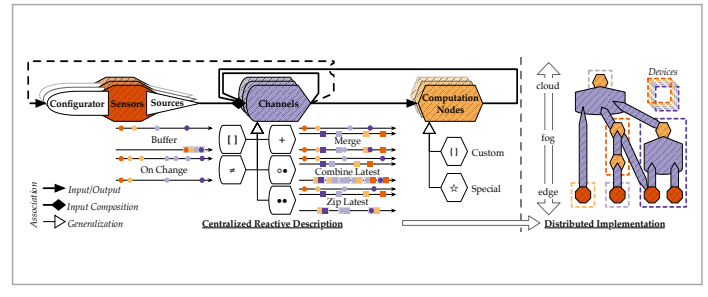

Fig. 3 - Entities in the Virtual Sensor DSL, their relations and distribution over Edge-Fog-Cloud infrastructures.

ON HOW TO OPTIMIZE MEDICAL IMAGE ANALYSIS: THE CHIMERA APPROACH

Eleonora D'Arnese - Supervisor: Prof. Marco Domenico Santambrogio

Medical image analysis has become the focus of a considerable number of research studies both in industry and academia, thanks to the increasing number of available imaging techniques and the quality of the obtained data. Moreover, the constant improvement in the acquisition machinery and digitalization process generates a massive amount of data that should be analyzed daily to extract relevant information for the diagnostic process. Indeed, data observation and analysis time are the de facto bottleneck of the clinical practice. Unfortunately, pure human analysis is unfeasible in a reasonable time; therefore, the introduction of automated workflow for the analysis of such data has become of great interest to speed up diagnosis and reduce the workload of physicians. Additionally, given the applicative scenario delivering accurate, fast, and efficient solutions is becoming central. In this scenario, two orthogonal paths can be considered and are followed by my dissertation: one pushes toward new algorithmic solutions, and the other revolves around identifying the most suitable hardware substrate for a given task.
In this context, this research thesis concentrates on optimizing widely employed and repetitive tasks in the medical image processing field. This dissertation focused on three solutions: two dealing with pre-processing, namely image registration and image segmentation, and a processing one that exploits the radiomic approach to identify and characterize cancer non-invasively. This thesis worked towards optimizing the proposed tasks by exploring one or both of the paths mentioned above. More in detail, this dissertation presents Chimera, which provides a comprehensive methodology and efficient solutions tailored to optimize the aforementioned most employed pre-processing and processing steps. Precisely, Chimera proposes open-source solutions exploiting different hardware accelerators. Additionally, they are extensible from other users and easy to adopt since they allow users with different levels of expertise to benefit from acceleration.

More into detail this thesis proposes a Hardware/Software toolchain for automating the generation of accelerated Image Registration (IR) pipelines. IR is a well-defined computation paradigm widely applied to align one or more images to a target. This paradigm builds upon three main blocks: the optimizer, the transformation, and the similarity metric. Although IR is widely employed, it is highly compute-intensive and represents many image processing pipelines' bottlenecks. Therefore, this thesis presents an open-source framework, shown in Figure 1, tailored to IR comprising of HW/SW highly-tunable registration components, that supports users with different expertise in building custom pipelines, and automating the design process.
Then, this dissertation proposes a framework to efficiently move the inference phase of different Deep Learning (DL) networks on Field Programmable Gate Array (FPGA)-powered devices to reduce power consumption while keeping accuracy and performance without knowledge of hardware development. Given the increasing quantities of medical imaging data, semantic segmentation and classification play a pivotal role for many down-stream applications such as surgery planning. In this context, DL is the way to go with a wide range of network architectures that are easily run on Graphics Processing Units (GPUs) thanks to a vast corpus of software libraries, helping the end users to benefit from their computational power. On the other hand, they are power-hungry devices, opening the introduction in the inference phase to more efficient devices. Based on these considerations, Chimera presents, NERONE, a framework that, given a pre-trained model and a dataset, allows the user to deploy the inference on FPGA transparently (Figure 2).
Finally, Chimera focuses on radiomics, which consists of mining massive arrays of quantitative features from routinely acquired digital medical images, and offers the possibility of calculating predictive biomarkers used as an in-vivo biopsy. Based on these considerations, this thesis describes a methodology and a tool, shown in Figure 3, that, starting from routinely acquired images, identifies and segments lung cancer masses and, thanks to radiomics, later on, characterizes them into primary cancer (and its subtypes) and metastases.

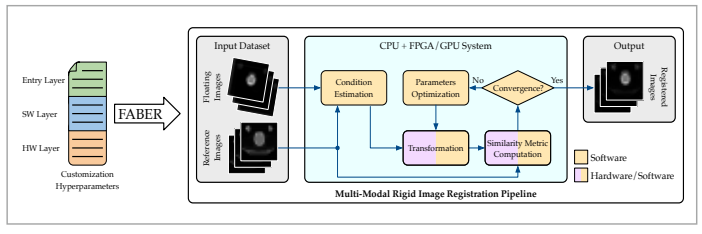

Fig. 1 - Overview of the entire Faber flow, from the user's input to the final Image Registration pipeline.

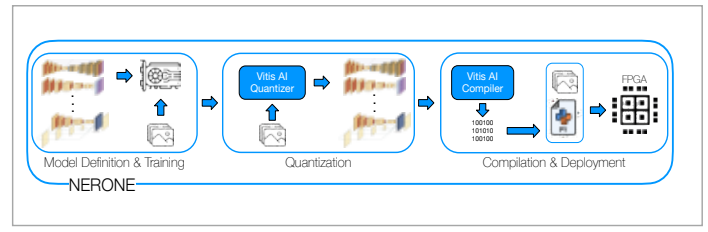

Fig. 2 - High-level view of the NERONE framework starting from GPU models training, models quantization, and compilation through Vitis AI compiler (Gpu icons created by Linector, Python file icons created by Flat Icons, and Gallery icons created by Freepik - Flaticon).

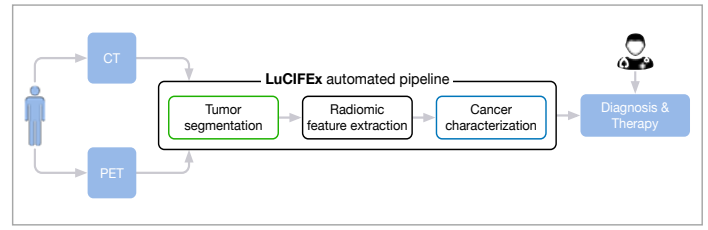

Fig. 3 -High level description of the radiomic-based tool that, starting from routinely acquired images, identifies, segments, and characterizes lung cancer masses.

BEAM-TYPE COMMUNICATIONS FOR THE 6G VEHICLE-TO-EVERYTHING SYSTEMS

Kai Dong - Supervisor: Prof. Umberto Spagnolini

In the coming Sixth Generation (6G) networks, advanced vehicle-to-everything (V2X) communication technologies will accelerate the evolution of intelligent transportation systems, with the goal of improving safety and traffic efficiency. The Society of Automotive Engineers (SAE) specified different automation levels, from level 0 (i.e., no automation) to level 5 (i.e., full automation) for autonomous driving. The higher the level of automation, the more critical Quality of Service (QoS) is required in the V2X systems. To support the advanced use cases of the evolving 6G V2X system (such as vehicle platooning, extended sensors, advanced driving and remote driving), beam-type Ultra-reliable Low-latency communication (URLLC) is required for the critical performance guarantees. The severe propagation loss of Electromagnetic (EM) signals operating in the Millimeter Wave (mmWave) and sub-THz frequency bands, on the other hand, limits the coverage range and makes beam-type communication links susceptible to being blocked. Although the Multiple Input Multiple Output (MIMO) system can compensate for this high path loss by pencil-like high-gain beamforming, there are still some challenges to be addressed in the 6G V2X systems, such as efficient Initial Access (IA), blockage modelling and mitigation, which are the main topics discussed in the thesis.

IA is mandatory for vehicles to discover and synchronize with the neighboring vehicles (or infrastructures) before information sharing and data transmission. Thus, an efficient IA scheme for the vehicles to select the best spatial beam pair among the candidate narrow beams to synchronize with the vehicular network is paramount of importance. Currently, a position-based beam sweeping scheme is widely utilized for the IA. However, the positioning information, e.g., from Global Positioning System (GPS), may be inaccurate or not always available in highly dynamic vehicular networks especially in the complex urban scenarios. Alternatively, Probabilistic Codebook (PCB) based beam selection scheme to select the best beam pair does not need the pre-knowledge of the position information but the spatial beam pointing directions. Inspired by the non-uniform distribution of the communication angles constrained by the road typologies in urban scenarios, a non-uniform quantization approach on the beam pointing angles for the PCB design is expected to achieve a performance enhancement compared to the uniform quantization scheme that specified in the current 3rd Generation Partnership Project (3GPP) standard. The simulation scenarios are extracted from OpenStreetMap, which is close to reality. Numerical simulation results assisted by professional simulation tools (e.g., Simulation of Urban MObility (SUMO), and Geometry-based, Efficient propagation Model for V2V communication (GEMV2)) consolidate our design benefits of a lower performance degradation in terms of spectral efficiency for the non-uniform PCB design compared to the uniform one defined in 3GPP.

Although the precise beamforming enabled by MIMO systems can compensate for the high propagation loss of the EM wave signals, the high sensitivity of the beam-type communication links to the blockages makes it challenging to guarantee the URLLC performance for some advanced use cases of 6G V2X systems. Thus, the prediction of the V2X link performance by considering the blockage impacts becomes of paramount importance. In this thesis, we proposed a novel vehicular blockage modelling approach for an arbitrary Vehicle-to-Vehicle (V2V) sidelink in a multi-lane highway scenario to derive an analytical model of Signal-to-Noise Ratio (SNR) distribution as well as the service probability constrained by a SNR threshold. In particular, the derived model considers multiple comprehensive impacting factors such as multiple vehicle blockers, random vehicles' height, and traffic density. Exhaustive simulation results show a good match between the derived analytical model and Monte Carlo simulations. This proposed analytical model provides an efficient tool for End-to-End (E2E) performance prediction, beam selection, and resource scheduling to mitigate the blockage effects.

In the case of the V2X communications links being blocked, candidate blockage mitigation solutions, e.g., advanced relaying technologies, are required for reliable link connectivity, which is the main focus of the remaining parts of this thesis. Specifically, the emerging relaying technologies, such as metasurface based and Amplify-and-Forward (AF) based, are promising for constructing a Smart Radio Environment (SRE) to meet diverse 6G applications. Firstly, benefiting from the new artificial metasurface technologies, the EM wave signals can be near-passively or fully-passively reflected towards the destination with a reliable Quality-of-Service (QoS). Unlike the conventional planar Reconfigurable Intelligent Surface (RIS), a novel Conformal Intelligent Reconfigurable Surface (C-IRS) is adopted to be deployed on vehicles' doors to assist the V2V sidelink communications in a multi-lane highway scenario. In particular, the proposed C-IRS is already set up with phase compensation for its curved shape so that EM wave signals can be reflected in a specular way (i.e., it is fully passive). Compared to the situation without C-IRS implementation, the significant blockage reduction can be as high as 20%. Moreover, if the phase of each element can be dynamically adjusted (i.e., Conformal Reconfigurable Intelligent Surface (C-RIS)), the blockage mitigation can be up to 70%. In the multi-lane highway scenario, the remarkable performance improvement in terms of the average SNR can be 10 – 20 dB for the C-IRS case with proper C-IRS selection. An average SNR gain of more than 30 dB is achieved with real-time phase adjustment at C-RIS. However, additional control signaling overhead is required, which will introduce an inevitable delay and computational cost. The proposed fully passive C-IRS has been demonstrated to achieve a remarkable improvement in performance and offers a new way to design a metasurface for blockage mitigation in 6G V2V communications.

In addition to the passive relaying scheme presented above, Smart Repeater (SR) is one key representative of active relaying technologies for the blockage mitigation that will be standardized in the upcoming 3GPP Release 18. The SR enables a smarter AF operation than the conventional Radio Frequency (RF) AF. In the thesis, we design a net-work controlled tri-sectoral Advanced SR (ASR) to support Vehicle-to-Infrastructure (V2I) communications, which is expected to achieve a doubly angular coverage (i.e., 240 deg) compared to the conventional SR, with a coverage of 120 deg. Specifically, the proposed ASR has three antenna arrays, one of which is towards Base Station (BS) and the other two towards the service areas (each with a Field-of-View (FoV) of 120 deg). Moreover, the multi-level multi-resolution codebook is designed to support multi-user communications. Numerical simulation results consolidate our tri-sectoral design benefits with the objective of maximizing the number of served vehicle User Equipments (UEs) constrained by per-UE rate and time-frequency resources, compared to the conventional SR. A more remarkable gain is achieved with a suitable trade-off between the number of served UEs and the time slots. Moreover, the benefits of the ASR design over the conventional SR in terms of cumulative spectral efficiency are also observed up to a factor of 2.

NEUROMORPHIC DEVICES BASED ON 2D MATERIALS AND THEIR APPLICATIONS IN COMPUTING

Matteo Farronato - Supervisor: Prof. Daniele Ielmini

In the last six decades, the semiconductor technology has experienced a fast improvement thanks to the scaling in device dimensions and the increase in the number of components per unit area. These two phenomena, namely the Dennard scaling and the Moore's law, drove the developing of ever more powerful computing systems, often based on the von Neuman architecture. However, in more recent years, the transistor dimensional downscaling in the 2D plane has stopped, encountering physical limits imposed by silicon-based materials. In addition, with the ubiquitous diffusion of mobile computing and Internet of Things (IoT), the amount of data produced and processed explodes, bringing out the critical issues of modern computing architectures. Indeed, von Neumann architecture-based digital processors are hindered by the performance gap between the central processing unit (CPU) and memory, which makes them generally inefficient in terms of both energy and latency, particularly in datacentric applications. To face these challenges, new computing paradigms categorized under the concept of in-memory computing, are gaining interest, since they suppress the memory bottleneck and have an improved energy efficiency. Neuromorphic computing, inspired by the functionality of human brain, is one of the main examples of architectures implementing this paradigm. The realization of such systems passes through the development of innovative memory devices, called emerging memories, which, thanks to their area scalability, low current, fast operation, and CMOS compatibility, are more appealing than standard charge-based memories. A key aspect of emerging memories is their operation principle, that often relies on the physics of the active material. For that reason, large effort is devoted to the study of materials with peculiar properties exploitable for the realization of efficient memory devices. Among all, 2D layered materials offer a unique physical structure and excellent electronic properties. After the demonstration of Graphene in 2004, a large number of 2D semiconductors have been discovered, studied and used for the realization of several electronic devices. This doctoral dissertation focused on the development of innovative memory devices based on 2D materials and their use in neuromorphic computing. First, devices have been fabricated in the clean room, following proper and reproducible nanofabrication steps. Particular attention has been devoted to set up processes involving 2D materials, which require dedicated deposition techniques. The work was made possible thanks to the availability of Polifab, the cleanroom facility of Politecnico di Milano. Two main devices, called memtransistors, have been developed. The first one, the ion-based memtransistor, is a three-terminal device with a transistor structure in back gate configuration. The channel of the device is composed by multilayer MoS_2, a bidimensional semiconductor mechanically exfoliated on top of the gate oxide. Source and drain terminals are made by silver with a distance $L_{channel} < 50$ nm. The device shows a considerable transistor characteristic with a large subthreshold swing. The application of a relative large drain to source voltage causes the formation of a metallic filament between the two terminals, thanks to the Ag ion migration. The behavior is volatile, meaning that the removal of the external electric field causes the spontaneous disruption of the filament in a time interval in the order of 100 ms. The two memory states can be independently controlled and tuned, which is a major advantage with respect to other similar devices. A memory chain like a NAND structure using three memtransistors was realized, demonstrating the possibility to independently program and read each device in the chain, opening the way for the future of memory technology based on 2D materials.

A second device with similar structure was also realized. In this case, the channel length is in the order of 100 nm, and the electrode material can be either Silver or Gold. This device, called electron-based memtransistor or charge trap memory (CTM), exploits the trapping and de-trapping of charges at the oxide/semiconductor interface to obtain a memory effect. This trapping causes the drift of the transistor threshold voltage that can be read as a change in the drain-source conductivity. Gradual potentiation of the device conductance is obtained with the application of gate/drain pulses. The highly linear synaptic curves obtained with the application of equal amplitude pulses make the device perfect for the implementation of AI hardware accelerators. Simulations of online training of a 2-layer, fully connected neural network (for digit image recognition) using CTM as synapses, show high test accuracy close to floating point, validating the excellent device characteristics. The spontaneous and non-linear depression of the device conductance, due to the relaxation of the threshold, can be further exploited for the implementation of neuromorphic computing. A pattern recognition system implemented by reservoir computing was demonstrated. A vector of CTMs composes the reservoir layer, which converts the spatiotemporal patterns coming from the image in a conductance vector. Thanks to the dynamic and non-linear response of the devices, the reservoir output state is unique for each input digit, and can be easily classified by a small feedforward neural network with just few weights. In addition, the weight matrix can be realized using crosspoint arrays of resistive memories, enabling a fully in-memory implementation of the proposed network. High test accuracy is obtained, considering also the variability of the reservoir response, making the device very attractive for in-memory computing applications. An insight of the future perspective of this work is finally given, with the characterization of a dual gate device realized starting from the CTM structure. These results open for the realization of more complex structures composed by several MoS_2-based devices in the same chip and the implementation of other neuromorphic functions.

In conclusion, the use of 2D-layered materials for the realization of emerging memories is still not very common, and several aspects need to be investigated, but all the results of this PhD dissertation pave the way towards future research works on emerging memories based on these innovative materials.

THE CAPACITY OF AMPLITUDE-CONSTRAINED VECTOR GAUSSIAN CHANNELS

Antonino Favano - Supervisor: Prof. Luca Barletta
Co-Supervisor: Dr. Marco Ferrari

Energy efficiency is a fundamental feature of modern wireless communication systems. The ever-growing requirements in terms of data rates, as well as the ubiquitous presence of wireless devices, have made the power consumption of wireless systems and the associated costs an extremely relevant concern. A green approach in the design of wireless communication systems is nowadays imperative to guarantee their sustainable growth.

One of the main components hindering the energy efficiency of wireless systems are radio frequency chains. To reduce their impact on power consumption, it can be useful to limit the peak power of the signals fed to their input. While the channel capacity of wireless systems subject to average power constraints has been investigated extensively, less is known about the capacity of channels subject to peak power constraints. Establishing a solid theoretic framework becomes essential to reliably assess the information capacity of such channels and, consequently, to efficiently exploit the resources available to wireless systems subject to peak power, or equivalently, peak amplitude constraints.

In this work, we investigate the capacity of nonfading and fading amplitude-constrained multiple-input multiple-output (MIMO) Gaussian channels. We consider three families of input constraints. The first, namely that of Total Amplitude (TA) constraints, limits the norm of the baseband representation of the input vector. The second family limits the absolute value of each complex component of the input. We refer to this family of constraints as Per-Antenna (PA) constraints. Finally, the third family is that of Antenna Subsets (AS) constraints, which limits the norm of the input subvectors resulting from any given partition of the input vector.

Nonfading Channels

In the case of nonfading channels under TA constraints, it is already known that the input distribution is uniform over its phase and that its support is composed of a finite number of concentric hyperspheres. In this work, we derive further insights on the structure of the capacity-achieving input distribution. The first contribution is the definition of a lower bound on the number of hyperspheres belonging to the support of the optimal input distribution, at any given signal-to-noise ratio (SNR). Furthermore, we prove that the capacity-achieving input distribution always includes the hypersphere of maximum radius allowed by the considered input constraint.

Another contribution is the definition of a numerical algorithm for the evaluation of an arbitrarily precise estimate of the optimal input distribution and of the associated channel capacity. For a given MIMO channel, numerically evaluating the capacity-achieving multi-dimensional input distribution is often a computationally unfeasible task. Nevertheless, in the case of nonfading amplitude-constrained channels, the spherical symmetry of the input distribution allows us to derive an equivalent channel model that depends only on the norm of the input vector and that, therefore, substantially simplifies the numerical estimation of the capacity-achieving input distribution. The algorithm iteratively refines the estimate of the input norm distribution by optimizing the position of the mass points, i.e., the radii of the hyperspheres in the multi-dimensional distribution, and the associated probabilities, uniformly distributed over the input phase. At each step, the radii are updated via a gradient ascent algorithm, while the probabilities are optimized via the Blahut-Arimoto algorithm. The convergence to the optimal input distribution is validated by verifying that the Karush-Kuhn-Tucker conditions for the maximization of the mutual information are satisfied.

The same iterative algorithm can be applied to other interesting case studies. Indeed, since both the PA and the AS constraints can be decomposed into independent TA constraints, the same algorithm is applicable to all the families of input constraints considered in this work. Moreover, we adapt and apply the derived algorithm also to the case of stochastically degraded wiretap channels under peak power constraints.

Finally, we provide an additional contribution in the specific case of PA constraints. We derive an approximated and discrete input distribution more suitable for real-world applications than that being theoretically optimal. Although being a suboptimal solution, designed specifically for more practical implementations, the proposed distribution is still close to being capacity-achieving and provides an information loss, from the true capacity, lower than 0.01 bit per channel use.

Fading Channels

In the case of amplitude-constrained fading channels, far less is known about the structure of the optimal input distribution and, therefore, it is also difficult to derive accurate estimates of the channel capacity. The most relevant results in the literature, available prior to our contributions, were upper and lower bounds on the channel capacity. In this work, we derive upper bounds that greatly improve upon the best previously available. Prior upper bounds would not converge to the best lower bound even asymptotically, as the SNR goes to infinity. We derive two novel upper bounds that are not affected by this drawback.

We refer to the first bound as Sphere Packing (SP) upper bound. The SP bound relies on the fact that the considered input constraints can be geometrically interpreted as convex regions belonging to a multi-dimensional Euclidean space, to which the input vector is bounded. The presence of fading induces a distortion of these constraint regions in the output signal space. The SP approach provides a volume-based upper bound that depends on geometric functionals of this distorted constraint region. These geometric functionals are called intrinsic volumes and can be defined, in general, for any convex constraint region. One critical issue is that the intrinsic volumes are known and can be efficiently evaluated just for a few selected cases, such as for TA constraints. To overcome this limitation, we define a variant of the SP upper bound, that can always be evaluated by upper-bounding the unknown intrinsic volumes. Furthermore, we define another variant based on a Gaussian maximum entropy argument, that improves the performance of the SP bound at low SNR. The main advantage of the SP bound is that it asymptotically converges to the best available lower bound and that it can be applied to all the considered families of constraints. The SP upper bound significantly improves upon the prior existing literature. Even at finite SNR, the SP bound is substantially closer to the lower bound than the previous best upper bound.

In the case of the PA and AS constraints, we derive a more targeted solution, namely the Quasi Parallel Channels (QPC) upper bound. While only applicable to the PA and AS constraints, the QPC bound is able to outperform the SP bound. The QPC upper bound exploits the inherent parallelism in the structure of the PA and AS constraint regions. Although channel fading distorts the parallelism of the constraint regions, we can still exploit the residual "quasi" parallelism to derive the QPC bound. Specifically, the upper bound is given by two terms. Roughly speaking, the first term ignores the effect introduced by the channel fading, while the second term acts as a distortion term and quantifies how much the channel deviates from being composed of truly parallel subchannels. Similarly to the SP bound, the performance of the QPC bound at low SNR is improved via a Gaussian maximum entropy argument. If compared to the SP bound or previous upper bounds, the main advantage of the QPC bound is that it is characterized by an even faster asymptotic convergence to the channel capacity.

VEHICULAR AUGMENTED REALITY SYSTEM FOR ADAS APPLICATIONS

Luca Franceschetti - Supervisor: Prof. Matteo Corno
Co-Supervisor: Prof. Sergio Matteo Savaresi

This PhD research project designed and developed a vehicular Augmented Reality system for ADAS (advanced driver assistance systems) applications. Augmented Reality (AR) is an interactive experience where real-world objects are "augmented" thanks to the integration of computer-generated virtual graphics, shown through the Head-Mounted Display (HMD). HMD is a wearable device, worn on the head of the user, that is capable of projecting virtual features and holograms in the field of view of the user by means of the semi-transparent lenses and holographic displays. In this way, real and virtual worlds are perfectly blended together, resulting in the complete immersiveness of the user wearing the device.
AR is commonly used only in small and static environments, like rooms or laboratories, where HMD built-in localization and mapping algorithms are precise and robust. These SLAM (simultaneous localization and mapping) algorithms build a map of an environment and at the same time use it to deduce HMD location through VIO (visual-inertial odometry). VIO combines camera and inertial measurements, that have complementary properties making them particularly suitable for fusion, in order to obtain robust and accurate localization and mapping. HMD (e.g. Hololens) uses SLAM and VIO algorithms for projecting stable AR features on top of the real-world mapped surfaces, in order to achieve a higher blending between virtual and real worlds.
HMD built-in algorithms do not work properly in wide and moving environments. During driving, the environment cannot be considered "static", because the car is travelling along the road and the background is expected to change in a short period of time. Also, cameras record static objects (the vehicle cockpit) and moving objects (the environment outside the vehicle), while IMU records lateral and longitudinal acceleration, returning conflicting information to the VIO algorithm.
This PhD work developed vehicular and ADAS applications of Augmented Reality, replacing the built-in HMD algorithm with some robust and specific solutions for driving contexts (car, bike, motorbike, tractor and off-road vehicle). ADAS help the driver with safety features to avoid collisions by offering technologies that alert the driver of potential problems. The integration of AR with ADAS could increase safety during driving even more: nowadays, ADAS warnings are shown to the driver through simple abstract symbols on the vehicle dashboard, without any specific information about the location of the danger. With AR is possible to show hazard warnings directly overlaid on the real danger, reducing the time used to look at the dashboard and the uncertainty of understanding the type of danger. AR is a well-studied topic in the driving and ADAS context, underlining a great interest in the use of AR while driving. Many studies confirm that AR fused with ADAS could bring a significant improvement in comfort and safety while driving.
While the potential of AR is clearly understood and acknowledged, the technological challenges are still abundant. To the best of the author's knowledge, no significant application of AR with HMD to real driving scenarios is present in the literature. The main reason is that, for the AR experience to be convincing, an HMD needs to address two aspects: stereoscopic holographic projection (usually done with waveguide technology) and accurate localization of the position and attitude of the user's head. Realism and immersiveness of AR experience strongly depend on the estimation of the position and orientation of the HMD (i.e. head of the user). An accurate pose estimation is the base on which the graphic engine renders the virtual objects so that they appear to occupy space in the real world as the user moves around. This task, also known as image registration, is paramount for a stable, precise and immersive experience. For this and other aforementioned reasons, this PhD project successfully developed algorithms of image registration and stereoscopic holographic projection for AR experience while driving or cycling.
Different achievements have been reached in this PhD research project. A bike/motorbike AR system, working in open cockpit context, has been developed, showing a holographic cyclist (Fig. 1), that competes with the real cyclist while wearing the HMD. Moreover, an AR system for closed-cockpit vehicles has been developed and further used for projecting holograms inside and outside the vehicle cockpit. A "braking points" use-case has been developed, that shows AR information placed outside the vehicle to inform the user where the braking manoeuvre should start (Fig. 2).
Since projecting AR information outside the vehicle requires high-precision estimation, a road detection algorithm has been implemented to estimate the 3D position of the road and to anchor the AR feature precisely on that. Also, a custom sensor-fusion algorithm for the vehicular AR system has been developed and used to compensate for the camera acquisition latency and the HMD projection latency by means of IMU measurement fused with fiducial marker pose estimation (Fig. 3). Furthermore, it has been analysed the robustness and reliability of the ArUco Markers pose estimation, which turned out highly sensitive to lighting conditions, causing misdetection and estimation errors during driving, since the cockpit cause shadowing and light reflection.
To overcome this limitation, it has been developed a new head pose estimation algorithm based on CNN and time-of-flight sensor, that is robust to different types of scene illumination. This new solution replaces markers as a fixed reference system inside the vehicle and uses all the cockpit features to estimate the driver's head pose.
This PhD research project led to the foundation of HMDrive s.r.l. in 2022, a start-up company whose main objective is the industrialization and consolidation of all the results and achievements discovered during this PhD years.

Fig. 1 - Holographic cyclist projected on the bike AR system.

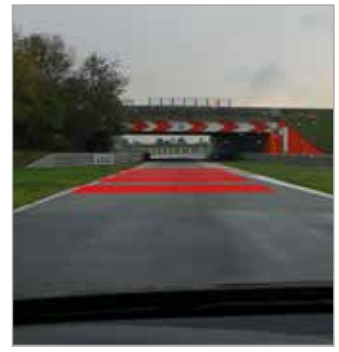

Fig. 2 - Braking point AR information projected outside the vehicle.

Fig. 3 - Sensor-fusion algorithm used to project AR obstacles in a moving vehicle application.

ADVANCED LEARNING METHODS FOR ANOMALY DETECTION IN MULTIVARIATE DATASTREAMS AND POINT CLOUDS

Luca Frittoli - Supervisor: Prof. Giacomo Boracchi

Anomaly detection is a challenging problem encountered in several domains, from quality control to cryptographic attacks. Several statistical and deep learning models have been proposed, each underpinning specific assumptions on the nature of the data to be analyzed. These models are typically configured on a training set of data generated in normal conditions, and then assess whether the testing data conforms to the model or not.

We present new solutions for anomaly-detection in two different settings. In the first part of the thesis, we assume that normal data are realizations of a random vector having a certain probability distribution. We focus on a change-detection problem, whose goal is to detect permanent changes in the data-generating process by analyzing a sequence of samples acquired over time, namely a datastream. Our most substantial contribution to the research on change detection is *QuantTree Exponentially Weighted Moving Average (QT-EWMA)*, an online and nonparametric change-detection algorithm for multivariate datastreams that can be configured to maintain the target Average Run Length (ARL_0), namely the expected time before a false alarm. In particular, we employ a QuantTree (QT) histogram to model the initial distribution from a training set and define a novel change-detection statistic based on the Exponentially Weighted Moving Average (EWMA) monitoring scheme. The properties of the statistics based on QuantTree histograms guarantee that QT-EWMA is nonparametric and allow us to compute thresholds that guarantee the ARL_0 independently on the data distribution. Our experiments on synthetic and real-world data confirm that QT-EWMA controls the ARL_0 more accurately than most competing methods, while achieving comparable or lower detection delays.

We extend our work to the concept-drift scenario, where the data samples are the object of a classification problem. Distribution changes, in this case referred to as concept drifts, require to update an underlying classifier. In the literature, concept-drift detection is addressed by monitoring either the overall data distribution (thus ignoring class labels) or the error rate of the classifier (thus ignoring distribution changes that have little impact on classification error). We propose *Class Distribution Monitoring (CDM)*, a novel concept-drift detection algorithm combining the information coming from the data distribution and the class labels. CDM uses multiple instances of QT-EWMA to detect changes in the class-conditional distributions of annotated datastreams. The main advantage of CDM is that it can detect drifts affecting a subset of classes and indicate which class triggered a detection, which might be crucial for diagnostics. Most remarkably, we show that CDM inherits from QT-EWMA the ability to control the ARL_0, which is rarely guaranteed by alternative solutions. Our experiments on synthetic and real-world data show that CDM controls the ARL_0 more accurately than alternative methods, while achieving lower detection delays in most cases. In particular, CDM achieves the best performance when the drift affects a small subset of classes and when the drift does not substantially increase the classification error, which are the most challenging scenarios.

As a new application of change detection, we address the problem of detecting errors in sequential cryptographic side-channel attacks. These attacks reconstruct a private key one bit at a time by using a distinguisher, namely a statistic involving side-channel data (e.g. the power consumption of a device) and some intermediate results of the target algorithm. We cast error detection as a change-detection problem since the distribution of the distinguisher changes after an error in the reconstruction of a key bit, and we propose to detect errors by monitoring the distinguisher sequence using a change-detection algorithm for univariate datastreams. Then, we propose an error-correction procedure based on a brute-force search over a small key window centered at the detected error, and a statistical test on the distinguisher values of each combination to select the correct one. Our experiments on synthetic and real-world side-channel data demonstrate that our procedure substantially improves the success rate of different sequential attacks against RSA-2048 decryption, outperforming existing techniques, which simply set a threshold on the distinguisher value. Our findings demonstrate that sequential attacks can be substantially strengthened and thus might be more dangerous than previously thought. For this reason, countermeasures such as blinding should be employed even when the low success rate of sequential attacks suggests that the cryptosystem is secure.

In the second part of the thesis, we address anomaly detection in point clouds. Point clouds are lists of the coordinates of points, e.g. describing the surface of an object, and are are gaining popularity since they provide a compact yet detailed representation for 3D data. Our goal is to assess whether individual point clouds belong to a certain normal class or not, a problem also known as one-class classification. Point clouds are high-dimensional data (a point cloud contains thousands of 3D coordinates) having complicated structures. The main challenge of handling point clouds is the lack of a grid structure: in fact, the points do not necessarily lie on a regular grid, so traditional Convolutional Neural Networks (CNNs) cannot be directly applied to this type of data. For this reason, several point-convolutional layers, namely convolutional layers designed to process point clouds, have been proposed in the literature.

We propose the *composite layer*, an original operator that extracts and compresses the spatial information from the coordinates of the points by a Radial Basis Function Network (RBFN) and then combines it with the features. Compared to the existing layers, our composite layer performs additional regularization by compressing the spatial information, and is substantially more flexible in terms of number of parameters and structure.
We use our composite layers to implement *CompositeNets*, neural networks achieving excellent classification performance. Most remarkably, we are among the first to address anomaly detection in point clouds by training our CompositeNet in a self-supervised fashion. Our solution achieves state-of-the-art performance in anomaly detection, outperforming the only existing solution and shallow baselines leveraging hand-crafted features.

Finally, we address a relevant anomaly-detection problem in an industrial scenario. In particular, we analyze Wafer Defect Maps (WDMs), namely lists containing the 2D coordinates of the defects in silicon wafers. In normal conditions, wafers contain few, randomly distributed defects, while defects forming patterns indicate problems in the production process. Since a few classes of defect patterns have been identified by production engineers, we cast anomaly detection as an open-set recognition problem, where the goal is to correctly classify WDMs belonging to the known classes and detect anomalous WDMs containing defect patterns that do not belong to any known class. Although the coordinates of a WDM lie on a grid, whose size is determined by the precision of the inspection machine, the grid is huge and prevents any CNN from processing WDMs at full resolution. For this reason, WDMs should be handled as 2D point clouds. To efficiently process WDMs, we train a Submanifold Sparse Convolutional Network (SSCN) on the known classes. Then, we detect anomalous patterns by applying an outlier detector based on a Gaussian Mixture Model (GMM) to the latent representation, namely the output of the penultimate layer of the SSCN. Our experiments on a dataset of WDMs acquired at STMicroelectronics show that our solution outperforms several methods from the literature, which we implemented on top of our SSCN for a fair comparison.

DESIGN AND IMPLEMENTATION OF A QC-MDPC CODE-BASED POST-QUANTUM KEM TARGETING FPGAs

Andrea Galimberti - Supervisor: Prof. William Fornaciari
Co-Supervisor: Prof. Davide Zoni

Public-key cryptography (PKC) allows sending encrypted messages over an insecure channel without sharing a secret key, and it has traditionally been a critical component of secure communication protocols such as TLS and SSH. Quantum computing is, however, expected to break the traditional PKC solutions in the upcoming decades, making it mandatory to design new security solutions that can also resist attacks carried out by quantum computers. Post-quantum cryptography (PQC) aims to design cryptoschemes that can be deployed on traditional computers and are based on problems that are computationally hard also for quantum computers, other than traditional ones, thus being able to resist both traditional and quantum attacks. US's National Institute of Standards and Technology (NIST) is currently undertaking a standardization process to define new standards for PQC, including key encapsulation mechanisms (KEMs) and digital signatures. The standard PQC solutions will have to satisfy not only security requirements but also performance ones. Providing effective hardware support for such cryptosystems is indeed one of the requirements NIST set within its standardization process, and it is particularly crucial to ensuring a wide adoption of post-quantum security solutions across embedded devices at the edge.
This thesis delivers a configurable FPGA-based hardware architecture to support BIKE, a post-quantum code-based KEM using quasi-cyclic moderate-density parity-check (QC-MDPC) codes. These codes are employed in a scheme similar to the well-studied Neiderreiter one, which dates back to the early 1980s. Compared to traditional Niederreiter schemes, whose underlying binary Goppa codes must have sizes in the order of millions of bits to provide quantum resistance, BIKE achieves a significantly smaller public key, in the order of tens of thousands of bits, through its usage of QC-MDPC codes.
The architecture proposed in this thesis aims to improve performance over the existing state-of-the-art software and hardware implementations of BIKE, and it is configurable through a set of architectural and code parameters that, through a single parametric design, allow using the resources available on FPGAs effectively, supporting different large QC-MDPC codes, and targeting the whole Xilinx Artix-7 FPGA family.
The hardware components implementing QC-MDPC bit-flipping decoding, binary polynomial inversion, and binary polynomial multiplication, i.e., the three most complex operations employed within the BIKE cryptoscheme, were first designed in a parametric way to exploit parallelism as desired according to the performance requirements and the area constraints given by the target platform. The QC-MDPC bit-flipping decoder implements the Black-Gray-Flip decoding algorithm, which requires the computation of two multiplications, performed respectively in the integer and binary domains, between a dense polynomial operand and a sparse one. The two dense-sparse multiplications are performed concurrently, in a pipelined fashion, and the number of the bits computed in parallel in both is configurable by the designer. The binary polynomial inversion component implements a Fermat-based inversion algorithm, which iterates binary polynomial multiplications and exponentiations.
The multiplications and exponentiations are carried out on dense-represented operands by two separate parametric components, and their computation is also scheduled in a pipelined fashion to maximize performance. The dense-dense binary polynomial multiplier used within the inversion component performs the multiplication between two large polynomial factors, with degree in the order of tens of thousands, by implementing a hybrid architecture that mixes the Karatsuba and Comba multiplication algorithms, where the number of partial products computed in parallel is configurable.

Two separate modules target the cryptographic functionality of the client and server nodes of the quantum-resistant key exchange, respectively. The client and server modules, whose architecture is depicted in Figure 1, make use of the binary polynomial arithmetic and QC-MDPC decoding components, as well as a SHAKE-based pseudorandom generator and an SHA-3 hashing core. The optimal parameterization of the configurable components, maximizing performance within the available FPGA resources, is identified by using a complexity-based heuristic that leverages the knowledge of such parametric components' time and space complexity to steer the design space exploration.
The performance was evaluated against state-of-the-art software and hardware implementations of BIKE. The proposed architecture's client- and server-side instances, targeting Artix-7 FPGAs and a 91MHz clock frequency, outperform by up to 1.91 and 1.83 times, respectively, the reference state-of-the-art software, which runs on a 4GHz desktop-class CPU supporting Intel AVX2 instructions. Moreover, compared to the best-performing state-of-the-art FPGA-based architecture, which also targets Artix-7 chips, the architecture described in this thesis provides a performance speedup of up to six times. Executing the whole KEM on the proposed architecture takes from 5.74ms to 0.61ms for AES-128 security, as shown in Figure 2, and from 19.35ms to 1.77ms for AES-192 security. Notably, both the client and server cores provide constant-time execution, thus avoiding information leakage that timing-based side-channel attacks could exploit.

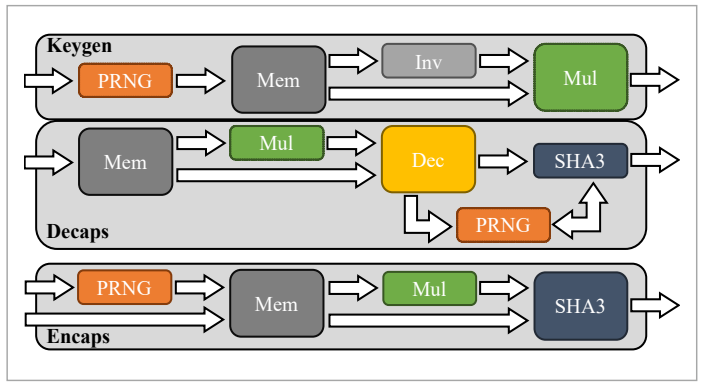

Fig. 1 - Top-level architecture of the BIKE client (Keygen and Decaps) and server (Encaps) cores.

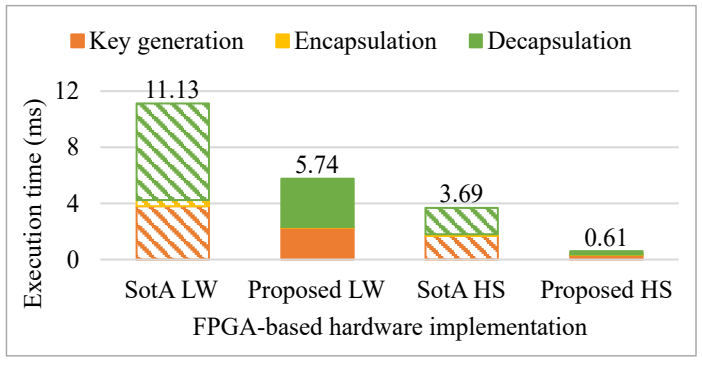

Fig. 2 - Execution times of BIKE with AES-128 security. Legend: SotA state-of-the-art, Proposed in this thesis, LW lightweight, HS high-speed.

ADDRESSING COLLABORATIVE MACHINE LEARNING CHALLENGES IN MEDICAL IMAGING

Edoardo Giacomello - Supervisor: Prof. Daniele Loiacono
Co-Supervisor: Prof. Luca Mainardi

The use of Deep Learning in medical diagnostics has shown remarkable results, often achieving performance levels comparable to human experts. However, Deep Learning relies heavily on large amounts of data for effective model development. In the medical field, data availability is a significant challenge due to privacy concerns. While anonymization techniques are commonly used to address this issue, they can be time-consuming and resource-intensive, and may not always be feasible. Additionally, generating and validating ground truth data for medical imaging tasks can be complex and labor-intensive.
To tackle this problem, a potential solution is to adopt a multi-centric approach, wherein the effort of data collection and model development is distributed among different entities, such as hospitals, without the need for centralized datasets. This approach could enable the sharing of data resources and expertise, while preserving patient privacy. By leveraging data from multiple sources, researchers can potentially overcome the limitations of data availability in medical imaging tasks and develop robust and generalizable Deep Learning models for practical clinical use. The proposed collaborative machine learning approach aims to address the challenges of data availability in medical imaging tasks by leveraging distributed learning and other machine learning techniques to exchange information among collaborating entities, instead of sharing raw data. This approach has several advantages, including addressing the privacy concerns associated with sharing patient data, reducing the need for data anonymization techniques, and distributing the effort required for data collection and annotation among multiple entities.
However, this approach also presents new challenges.
One challenge is the issue of different distributions in medical data, which can arise due to differences in population among different collaborating entities. Medical data collected from different hospitals or clinics may have variations in terms of patient demographics, imaging protocols, and disease prevalence, which can result in differences in data distributions. These distributional differences can affect the performance of machine learning models, as models trained on data from one entity may not generalize well to data from another entity due to the distribution shift.
Another challenge is the problem of missing data in multi-modality imaging datasets. Medical imaging datasets often include data from multiple imaging modalities, as in Magnetic Resonance Imaging (MRI), where physicians can prescribe one or more image *weighting* and possibly the use of a contrast agent (Fig. 1). However, not all patients may undergo all modalities of imaging, resulting in missing data for some modalities. This missing data can pose challenges for machine learning models that require complete data for training, as it may result in biased or incomplete learning. Addressing these challenges requires careful consideration and methodological approaches.
We proposed a framework for classifying distributed learning approaches and explored the integration of popular techniques like transfer learning in a collaborative setting (Fig. 2). Through empirical experiments, we demonstrated the effective utilization of various machine learning techniques, including ensembling methods, distributed learning algorithms, and transfer learning, to enable collaborative learning in the presence of data heterogeneity, model architecture variability, and diverse target labels. Moreover, we presented a novel approach based on adversarial networks to address segmentation challenges in datasets with missing modalities. Our research was based on real-world medical imaging tasks, including Brain Tumor Segmentation and Automated Chest X-Ray Diagnosis, serving as case studies.
Our study has yielded promising results, indicating that collaborative learning could be a feasible approach to address the challenges outlined above. Specifically, we found that ensembling methods can be effectively utilized in settings where multiple deep learning models with different architectures are already available. Our proposed methods based on entropy were particularly effective for classification tasks when no single model outperformed the others on every label. Additionally, we explored distributed learning as an optimal approach for designing collaborative systems when deep learning models are not readily available. We compared two recently introduced techniques, Federated Learning and Split Learning, in the context of data heterogeneity, and our results demonstrated that distributed learning can achieve performances close to a centralized model while maintaining good results with increased data heterogeneity and privacy requirements.
Furthermore, we investigated transfer learning as a collaborative tool. We leveraged embedding techniques to build different machine learning models based on feature extraction from multiple Convolutional Neural Networks, and we applied transfer learning to train models specific to smaller private datasets with a different set of labels. Our findings showed that this approach effectively enables training of machine learning models for medical imaging in settings with limited data and computational resources.
In addition, we explored the use of Adversarial Networks for transfer learning between segmentation models trained on different modalities, and generating missing modalities in medical datasets. Our results demonstrated the effectiveness of transfer learning in the context of adversarial networks, although the setting was more complex than standard neural networks. Moreover, we showed that a generative approach allows the use of machine learning models requiring multiple input modalities, with only a slight loss in segmentation performance. Overall, our study provides valuable insights into the potential of collaborative learning and the application of various machine learning techniques in addressing challenges related to data heterogeneity, model architecture variability, and limited data availability in medical imaging tasks. These findings have implications for advancing the field of collaborative machine learning and may contribute to improved healthcare outcomes.

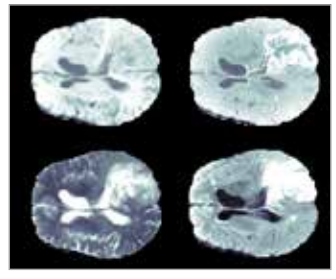

Fig. 1 - Different MRI modalities of the same brain affected by glioma: T1-Weighted (top left), T1-Weighted with Contrast (top right), T2-Weighted (bottom left), FLAIR (bottom right). Source: BraTS19 dataset.

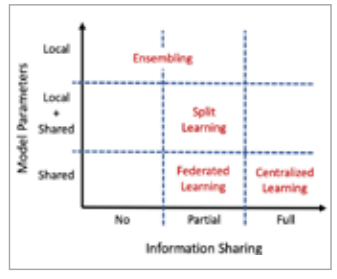

Fig. 2 - Overview of distributed learning paradigms, classified in our framework by degree of Information Sharing and Model Parameter displacement between the distributed models.

ADVANCED CONTROL TECHNIQUES FOR HETEROGENOUS AND DENSELY-INTEGRATED PHOTONIC CIRCUITS

Vittorio Grimaldi - Supervisor: Prof. Marco Sampietro

The potentialities of integrated photonic chip are enormous, ranging from the creation of short-range optical interconnection schemes, to the implementation of artificial intelligence and all-optical signal processors. However, integrated photonic systems have not reached the expected diffusion yet, hindered by the necessity of an electronic layer to be operated. Electronic control algorithms are crucial for mitigating the impact of fabrication mismatches and thermal and wavelength instabilities that prevent open-loop operations. This thesis explores effective methods for implementing control electronics and algorithms to stabilize the behavior of different photonic devices.

The proposed control scheme is based on the dithering technique that allows to extract the first derivative of the transfer function of the target device. With the use of an integral controller, a power-independent calibration-free control loop is designed and analyzed, allowing to stabilize any photonic device on the minima (or maxima) of its transfer function. Advanced uses of the dithering techniques are also discussed, showing the possibility to discriminate between the effects of distinct actuators with a single sensor by exploiting the orthogonality of different dithering frequencies and the frequency selectivity of the lock-in readout.

The concept of orthogonality helps in simplifying the complexity of the assembly. In particular, the control of a programmable photonic mesh without integrated sensors is presented. The results show how it is possible to use a single external power monitor to control 8 cascaded Mach-Zender interferometers, for which 16 different dithering signals were isolated from the same detector and hence the same electronic readout line. This approach comes at the expense of the maximum bandwidth of the loop, demonstrating how there is a substantial trade-off between the number of actuators controlled by a single sensor and the maximum speed of the control system.

The pilot tones technique was also introduced: by introducing a slow modulation of the optical intensity, it is possible to isolate and discriminate the effect of any individual carrier from the others. When the pilot tones technique is used in conjunction with the dithering one, intermodulated spectral components arise in the signal spectrum, each related to a single carrier but containing the information of the transfer function derivative. An experimental validation of this approach has been carried out, showing how a bi-diagonal mesh is able to separate and measure the effect of two orthogonal independently-labeled beams of lights.

If a working condition different from the maxima (or minima) of the device transfer function is targeted, the implemented loop based on the first-derivative signal would result in a control scheme that is not power-independent nor calibration-free. Hence, an extension of the dithering technique was investigated: by exploiting the non-linearities of photonic devices, it is possible to extract and use the second derivative of the transfer function. The performance of the second-derivative control loop was experimentally assessed on an integrated Silicon Photonics micro-ring modulator (MRM), showing how the locking point of the system is close to the theoretical optimum working condition of a modulator. The dithering technique and its second-derivative extension can be used together to control more complex and heterogeneous systems. For the purpose, a WDM-based multi-socket interconnect architecture is integrated in a single Silicon Photonics chip, where both resonant switches, locked on the maxima of their transfer function, and MRMs, locked on the maximum slope working condition, are simultaneously controlled.

The enormous flexibility of the platform was possible thanks to the use of a digital core for operating the system. To ensure maximum versatility, an FPGA was used for the application. The design had to be heavily optimized to reliably operate the system with commercially-available FPGAs. The basic digital signal processing chain for implementing the lock-in readout scheme was also expanded to cancel out the effect of spurious spectral components.

Finally, a plasmonic-assisted bolometric sensor was designed. Due to the plasmonic propagation, the free electrons interact with the lattice of the metal, dissipating energy and heating up the device, changing its resistance. FEM simulations were carried out to understand the effect of the device dimensions on the plasmonic propagation. Because of the presence of a higher-order mode, a strong beating effect was observed and experimentally demonstrated. The responsivity and the time response of the detector were also assessed, showing a sensitivity down to -20dBm and a bandwidth of 130kHz. The sensor was eventually used in a first simple control experiment on a micro-ring resonator, where a minimum chaser algorithm was programmed and operated in the custom UI.

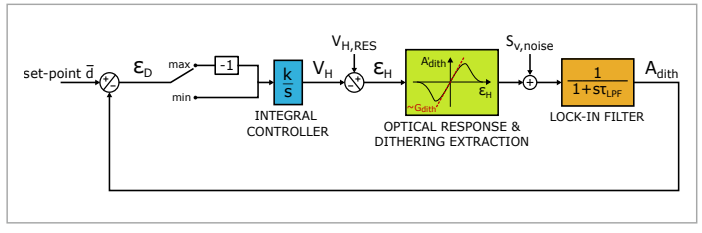

Fig. 1 - Block diagram of the proposed closed-loop architecture to lock the working condition of photonic devices, based on the dithering technique and an integral controller.

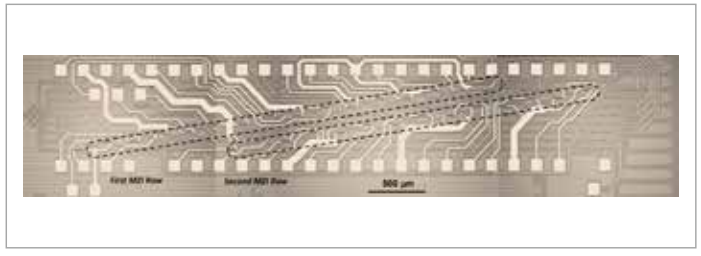

Fig. 2 - Microscope picture of the bi-diagonal mesh used in this work. The two diagonals are highlighted.

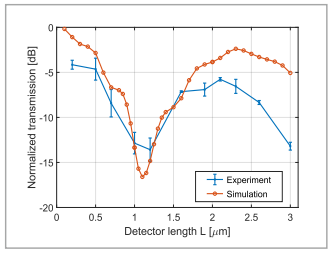

Fig. 3 - Experimental measurements of the transmission of the plasmonic-assisted detector vs. its length (in blue). The simulated curve is also reported (in orange) for comparison.

INNOVATIVE CROSS-LAYER OPTIMIZATION TECHNIQUES FOR THE DESIGN OF FILTERLESS AND WAVELENGTH-SWITCHED OPTICAL NETWORKS

Mëmëdhe Ibrahimi - Supervisor: Prof. Massimo Tornatore

Optical networks are playing a crucial role in the era of 5G-and-beyond communications to support applications requiring unprecedented capacity, reliability, low-latency and guaranteeing lightpath Quality-of-Transmission (QoT). To cope with such stringent requirements, network operators are driven to provide novel solutions while keeping network costs at bay. Recently adopted technologies, such as coherent transmission, allow to support high capacity demands and enable network designers to tune various configuration parameters to achieve better network performance, but at the expense of increased network design complexity. To reduce cost while enabling network expandability, new network architectures such as Filterless Optical Networks (FONs) are essential. In FONs, Reconfigurable Optical Add/Drop Multiplexers (ROADMs), based on costly Wavelength Selective Switches (WSSs), are replaced by broadcast-and-select switching architectures based on optical power passive splitters and combiners. We make use of FON network architecture and apply it in the context of horseshoe FON topologies, as they have shown to be a practical deployment approach, especially in metropolitan optical networks. We devise optimization approaches that minimize network costs by optimizing the deployment of optical equipment at network nodes and along the fiber while ensuring lightpaths' QoT. In particular, we optimize the deployment of Optical Amplifiers at network nodes, i.e., booster amplifiers and pre-amplifiers, and along the fiber, i.e., inline amplifiers, both in the context of traditional Wavelength Switched Optical Networks (WSON) based on ROADMs with WSSs and in the context of horseshoe FON. Additionally, we devise a cross-layer optimization approach for the placement of Optical Transport Network (OTN) traffic-grooming boards at the electrical layer while employing coherent and non-coherent transmission technologies, with the objective of minimizing overall network costs. Furthermore, we investigate the problem of Dedicated Path Protection (DPP) in the context of meshed FON and ensure link-disjoint primary and backup paths for each lightpath, by optimizing the deployment of additional equipment (transceivers and wavelength blockers) at network nodes and (colored passive filters) along the link.

To solve these problems, we have developed various optimization methods such as greedy heuristic approaches and meta-heuristics such as genetic algorithms, and Integer Linear Programming (ILP) models. The objective of these approaches is network cost minimization and depending on the context they are applied to, they are subject to different constraints such as lightpath QoT feasibility constraints, spectrum continuity and contiguity constraints, traffic-grooming constraints, FON-related constraints and capacity constraints.

In addition to the above-mentioned optimization techniques, we investigate the problem of estimating the QoT of unestablished lightpaths. Estimating lightpaths' QoT is a complex problem due to the nature of nonlinear impairments characterizing signal propagation in optical fiber and due to uncertainties of parameters used to describe various optical components/equipment. This problem has been generally tackled utilizing Machine Learning (ML)-classification approaches that estimate if a given lightpath configuration has satisfactory QoT. We address this problem utilizing ML-regression approaches as they allow to make more informed decisions about how conservative or aggressive a network operator can be when taking network planning choices, i.e., deploying a new lightpath. We propose several novel ML-regression approaches to estimate the distribution of a lightpaths' Signal-to-Noise Ratio (SNR).

The research carried during this thesis has been conducted in partnership with an industrial partner, SM-Optics, with whom we have jointly identified and investigated the research problems tackled during the thesis work. Additionally, we have conducted a joint research collaboration with two other universities and have investigated the use of Machine Learning - based approaches in network design.

SPAD ARRAYS AND SIPM FOR TIME-OF-FLIGHT LIDAR AND QUANTUM COMMUNICATION

Alfonso Incoronato - Supervisor: Prof. Franco Zappa

Light detection and ranging (LiDAR) exploits the "echo" light reflected back by an object illuminated by a light emitter. Nowadays, such technique is emerging in respect to others, due to the fine resolution and the possibility to reconstruct a 3D map of the scene. The 3D ranging spans from long-range automotive applications towards shorter scenes, such as in augmented reality and in smartphone cameras. The core of such a system is the optical sensor, needed to recover the light information and compute the Time-Of-Flight (TOF). The final system concerns the design of optics and lenses as in classical camera; however, the active illumination requires the management of lasers, scanning systems and many other components. Single Photon Avalanche Diodes (SPADs) have been investigated in many fields in which few photons are available; however, in the last decade, various companies focused the research on these sensors also for high laser power applications, in automotive scenarios. The reasons are mainly the possibility to obtain 2D and 3D spatially resolved imagers, and integrate into standard process all the electronics needed, obtaining a monolithic system-on-chip.

The main goal of this Ph.D. dissertation was to present the design and study of two different SPAD-sensors, for scanning single-spot LiDAR application with solar background rejection. Both sensors have been commissioned by a market leader customer and must fit into a specific scanning LiDAR system. The main idea behind the first microchip was the possibility to move toward a single-shot acquisition, using multiple SPADs within the single-pixel. The photon coincidence is exploited to track the incoming light and follow the detected peaks. Only the TOF related to the highest detected peak is provided, so it is possible to read out only one TOF per frame. The research consisted of validating such new method, that has been the subject of a journal paper and a patent application (number PCT/CN2022/075580). The second chip implements a more standard approach, known as Time-Correlated Single Photon Counting (TCSPC), exploited in many SPAD based setups. More TOFs can be computed, thanks to the multi-hit timing electronics, and a novel background rejection technique has been implemented: the background is observed for a defined interval, and a threshold is set accordingly. Every TOF computed after each threshold crossing, is associated to the number of triggered SPADs (i.e., number of photons), so a weighted histogram can be built off-chip. These new methods will be validated in the customer's final setup. The two microchips have been conceived, designed, and fabricated in a 160 nm BCD technology node. Such technology features SPADs with excellent performances, and is a mature node used since many years in the SPADlab at POLIMI. The first SPAD chip has been produced in 2021 and some preliminary validation and characterization have been performed after fabrication. A powerful 905 nm laser has been assembled to execute some preliminary long-range measurements.
The second chip has been entirely designed in 2021 and the tape out should have been in April 2022. Unfortunately, some manufacturing errors during fabrication required to start a second batch, and the tape out will be at the end of 2022.
The Ph.D. research also dealt with a secondary activity with completely different topic, concerning the design and characterization of a multi-channel SPAD chip to be integrated along with silicon photonics, as part of the "UNIQORN" Horizon 2020 FET project. This project had the objective of developing a Differential Phase Shift – QKD device. One of the main building blocks of the required system is a Quantum Random Number Generator (QRNG), which must also include a detector able to reveal the position of a photon randomly split through waveguides. Thus, I contributed to devise and develop a 32 × 1 linear SPAD array, in a 160 nm BCD technology, able to generate a raw random number by revealing the position on the array where a single photon impinges.
This chip has been extensively characterized and fully validated.

Fig. 1

GLOBAL PROTECTION FOR TRANSIENT ATTACKS

Niccolò Izzo - Supervisor: Prof. Luca Oddone Breveglieri

Computer security threats have a strong bond with information permanence, which is encoded in the physical state of a device. Attacks based on the trace left by the flow of information into a system in the form of its physical state are called transient. A secure device must store its state in a multitude of functionalities that have to be resilient to known and future attacks. In a mobile system, the security state of the device could switch between locked and unlocked, and the secure erasure of user data must be guaranteed during said transitions. Current DRAM-based main memories will be gradually replaced by Emerging Memories such as 3D XPoint, ReRAM, STT-RAM, Memristor or ULTRARAM, which are faster, more scalable and efficient than traditional NAND flash, even though their non-volatility is yet another potentially vulnerable state. Thus, a secure non-volatile storage architecture will have to employ well-known cryptographic building blocks to guarantee strong security properties on the stored data, such as confidentiality, integrity and replay protection, even when the device is turned off. Such properties must be guaranteed despite external physical threats, tampering with the bus signals, as well as internal threats, executing malicious code in a Virtual Machine on the same virtualized environment, or on the hypervisor itself. Another threat that originates from the variation of physical states are side-channel attacks. In fact, even the most efficient encryption architecture is rendered useless if a secret, e.g., a cryptographic key, is exposed through side-channel leakage, like power consumption, EM emission, or others. Masking techniques allow to implement effective software countermeasures, however their security proofs can be invalidated by hidden micro-architectural features. To restore the effectiveness of these countermeasures, a detailed model of the stateful elements of the data path has to be derived. Such model will allow the modification of the instruction scheduling of the sensitive code to implement side-channel countermeasures, e.g., masking, in a secure way.

HIGH PERFORMANCE ELECTRONIC SYSTEMS FOR MULTI - CHANNEL SINGLE PHOTON DETECTORS

Ivan Labanca - Supervisor: Prof. Ivan Rech

The analysis of optical signal can be crucial in many fields like medicine, biology, security and scanning systems. In this project, two different problems, related to single photon measurements, have been faced. The first is the study and implementation of a detector module, suitable for satellite LIDAR measurement system, and the second is the implementation of an innovative system to overcome the classic limits of the TCSPC technique.

1) The LIDAR technique measures the distance between the detector and the objects by illuminating the environment with a pulsed laser and measuring the time of flight using a suitable sensor. Satellite LIDAR systems allows the scan of the atmosphere by measuring the intensity as a function of time and the characteristics of the backscattering light returning to the satellite. This allows to obtain information on molecules, aerosols, clouds etc. along the path of the LASER pulse. The measurement of the light reflected by the first layers below the ocean level allows to obtain information also about the plankton under the surface. In the case of LIDAR measurements, the intensity of the signal is related to the reflectivity of the objects encountered along the path and to the distance from the laser that illuminates the scene. To maintain a good signal-to-noise ratio even in the parts with a lower signal (typically under the ocean surface), the laser pulse energy can not be too low also in the case of use of SPAD-based photodetector systems, useful for handling weak optical signals. Unfortunetely using a single channel SPAD, the dead time blinds the photodetector after a triggered avalanche, thus introducing information loss. To avoid these losses, an array approach is required: if some photodetectors are triggered, others can be ready to detect the back scattered photons. However, in case of particularly reflective objects or objects close to the measurement system, the intensity of reflected light can easily induce the saturation of the photodetector. Indeed, on the one hand it would no longer be possible to distinguish the variations of the optical signal that has to be measured; on the other hand, the saturation of the photodetectors hides the photons that arrive immediately after the intense optical signal. For instance, it would be hard to study what happens right after a water interface because the surface itself causes an intense bright flash, that can suddenly blind the photodetector. In order to avoid this issue, it is necessary to reduce the maximum number of photons impinging on the array. In this way, only a small percentage of SPAD is triggered just after the highest back scattered light pulse and the others are therefore ready to detect new incoming photons. Finally, the most important parameter is represented by the dynamic range of the system, that is intrinsically affected by the detector dark counts and dead time. This project responds to the requests for an ultra-wide dynamic range system with a peak signal rate up to 40 GigaPhotons/seconds and a temporal resolution of 10ns. The adopted solution involves the use of a SPAD sensor array. This allows you to take advantage of the high efficiency and fast recovery time of the array up to saturation of the entire SPAD array. The project has provided a first module to perform electro-optical characterization and irradiation tests, to study possible radiation damage in space. Subsequently, in order to satisfy the required dynamic range, a system based on an ARRAY of SPAD photodetectors, managed by a new active quenching circuit (AQC), was developed. In conclusion, the goal of this project is to design and develop a new Time-Resolved, Single-Photon sensing head able to overcome the current limits of Atmospheric Sensing via LIDAR satellite.

2) Time-Correlated Single-Photon Counting (TCSPC) is a technique performing the analysis of optical pulses with high timing precision. In a TCSPC measurement, a sample is excited by a periodic laser source and re-emitted photons are detected and recorded in a histogram depending on their arrival time within the laser period. After many measurements, the histogram shape represents the probability that a photon has arrived with a specific delay respect to the laser pulse, and this is the shape of the luminous signal. Conventional TCSPC modules can record only one photon per excitation period due to the dead time that characterizes the detectors. In this scenario, the pile-up represents the major limitation to speeding up the measurement in TCSPC experiments. In fact, to avoid distortion phenomena, up to now experiments have been conducted by limiting the rate of incident photons to below 5% of the excitation rate: the result is an inherently long acquisition time. However, by analyzing the impact of the dead time on the performance of the TCSPC technique, it is possible to find an optimal working point, with almost zero distortion, regardless of the shape and intensity of the light signal. The technique described shows how, as long as the dead time of the detector is exactly equal to the excitation period of the laser, it is possible to exceed the limit of 5% in the rate of incident photons, improving the speed of the TCSPC experiments and maintaining zero distortion. Compared to the limit imposed by the pile-up phenomenon, the improvement factor is almost equal to one order of magnitude, and the speed could further increase by extending this approach to multichannel solutions as well. However, the corresponding practical embodiment of the technique requires numerous expedients and specific characteristics of the detector, the quenching electronics and the conversion electronics. In my thesis we focused on these aspects, with the aim of creating the first prototype of a single-channel TCSPC system suitable for exploiting the technique. The acquisition system (FLASH system) consists of two blocks: the Detection Head and the Timing Conversion module. The Detection head is based on a custom technology SPAD, driven by an external Active Quenching Circuit (AQC) and a sensing circuit able to achieve a picosecond precision. The AQC dead time is finely tunable and this is necessary to match the excitation period. The Timing Conversion module implements a new Fast-TAC (F-TAC) structure based on sixteen TAC working in sequence to reach the high conversion rate required.

IMPROVED BEYOND 5G PHYSICAL LAYER DESIGN TO ENABLE V2X SYSTEMS AT HIGH FREQUENCIES

Francesco Linsalata - Supervisor: Prof. Maurizio Magarini

Beyond fifth generation (B5G) research has recently started and this thesis aims at contributing to it.
The advances in the automotive industry with the ever-increasing request for connected and autonomous vehicles (CAVs) are pushing for a new era of network systems. Vehicular communications, or vehicle-to-everything (V2X), are expected to be the main actors of the new network technology. The hard requirements of the enhanced V2X applications demand a radical transformation of the design of the network physical layer (PHY). Indeed, high frequency communications with large antenna arrays are the main candidate solutions to satisfy the high rate and low latency requirements.
However, the high mobility of vehicles and high frequencies introduce new challenges that are firstly presented, and then, addressed with novel solutions in this thesis.
The first problem that is considered regards the waveform design. The current 5G new radio (NR) waveform Orthogonal Frequency Division Multiplexing (OFDM) suffers inter carrier interference (ICI). Mobility introduces Doppler, and, thus, ICI. The high frequencies lead to high phase noise (PN). PN introduces further ICI. Thus, the combination of the two effects could be damaging for OFDM. Nevertheless, this aspect is not well-addressed in the current solutions. The first result of this thesis is a low complex joint PN and channel maximum likelihood estimation algorithm. The proposed solution outperforms the state-of-the-art approaches in terms of block error rate, requiring lower pilots' overhead to estimate the channel.
Fast and optimized initial access (IA) spatial synchronization methods for the vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) are among the main contributions of this work. A non-uniform distribution of the communication directions due to road topology constraints motivates the design of Probabilistic Codebook (PCB) techniques with prioritized beams during V2V IA. The proposed approach outperforms the 5G NR exhaustive space scanning technique in terms of training time and requires less signalling than the literature position-assisted methods.
Simulation results, supported by a set of experimental evidence, demonstrate that there is no winning method for IA in the V2I system. Even if the base station is equipped with radar or relies on the vehicle's onboard positioning systems. Numerical evaluations show that they provide complementary performance based on the position of the user in the served cell, moving the proposal to optimally combine radar and positioning information in a multi-technology integrated solution, which demonstrates to be optimal in terms of training time and gain losses.
Lastly, the thesis focuses on the problem of link and relay selection in V2X. The thesis proposes a novel proactive relaying strategy that exploits the cooperation between CAVs and environment information to predict the dynamic line of sight evolution, which is vital at high frequencies. The numerical results prove that the proposed approach can counteract the blockage and provide high network connectivity.

FIBER OPTIC CURRENT SENSORS FOR SYSTEM MONITORING

Andrea Madaschi - Supervisor: Prof. Pierpaolo Boffi

Nowadays, almost all devices, such as smart-phones, household appliances and industrial machines, require a source of electrical energy in order to work. Even in sectors where fossil energy sources were traditionally used, such as the automotive sector, a gradual transition to electricity sources is underway. A particular area of interest is represented by the electrical industry, where accurate analyses and measurements are required to evaluate performance and safety of power systems. To achieve this aim, it is often necessary to perform measures in very high current conditions and/or in presence of strong electromagnetic interference. Moreover in power systems where high voltages are used, even hundreds of kV, more and more stringent safety measures and guarantees of perfect insulation are required.
Traditional current sensors are struggling to keep up with all these requirements, mainly because they often require power to be supplied to provide current measurements and do not offer galvanic isolation. Providing complete isolation in high-voltage applications and at the same time making a reliable current measurement is an extremely critical, expensive and also unreliable task.
For these reasons, the study of alternative sensors that exploit optical fiber and the light propagating in it to obtain measurements of electric current is becoming increasingly important.
A fiber sensor has the advantage of having extremely small dimensions and weight, but also can be used in environments where strong electromagnetic interference are present without having to resort to particular precautions.
Being made of non-conductive material and practically immune to electromagnetic interference, fiber optic current sensors are a valid alternative to traditional current sensors, which even today, after decades of development, are not very reliable.
Thanks to the dielectric nature of the optical fiber, they can also easily meet the requirement of complete isolation, which is particularly critical in high-voltage applications.
In this thesis, two fiber optic current sensor configurations are presented: one based on a polarimetric approach and one based on a Michelson interferometer.
Both sensors measure the line integral of the magnetic field along one or more fiber turns, wrapped around the electrical conductor crossed by an electric current.
The use of the optical fiber arranged on a coil has the particular advantage of being sensitive almost only to the magnetic field produced by the wire that crosses it and of not being significantly affected by external electromagnetic field.
Current sensors based on optical fiber technology exploit the Faraday effect. Michael Faraday in 1845 discovered that when a linearly polarized beam travels in a magneto-optical medium, placed in a magnetic field aligned to the direction of propagation of the optical beam, its SOP rotates by an angle proportional to the intensity of the magnetic field and the length of the optical path exposed to the magnetic field.
This effect and more in general all the magneto-optic relations played a fundamental role in the history of electromagnetism, providing support to the electromagnetic theory of the light.
The first configuration presented is based on a polarimetric read-out and the most straightforward way to exploit this technique is represented in Figure 1.

Considering all components as ideals, the optical beam emitted by a generic light source, after being linearly polarized at 0°, propagates in a coil of optical fiber. In the center of the coil is placed an electrical wire which, generates a magnetic field with the components parallel to the coil and directly proportional to the electrical current. At the exit of the coil, the optical beam hits a second polarizer oriented at 45° with respect to the first one, then the transmitted light is detected by a photo-diode.
Analyzing the variation of intensity of the light, detected by the photo-diode, in presence of the magnetic field with respect to the case when there is no magnetic field, it is possible to derive the angle of rotation of the linear SOP of the light that is ruled by the following equation:

$$\Theta(t) = VNi(t)$$

where $i(t)$ is the current, N the number of turns of the coil and V is the Verdet constant of the fiber expressed in *rad/A*.
A simple configuration as the one reported in Figure 1 is not practical employable as a current sensor because it presents many flows as the sensitivity to external mechanical vibration, to the variation of the state of polarization of the light, to the intensity of the beam emitted by the light source and so on.
In the thesis all of these aspects are deeply analyzed and for most of them a solution is proposed.

The second configuration of FOS presented in the manuscript takes advantage of an innovative coherent configuration for demodulating the signals and relies on a modified version of the fiber optic Michelson interferometric architecture to detect the magnetic field produced by the electrical current.
The new coherent detection scheme is also used to retrieve in an efficient and passive way the phase information of the received optical signal. In order to address the high sensitivity to mechanical perturbation that affect classical in-fiber interferometers that use two different fibers or a single fiber in counter propagating mode to build the two arms of the interferometer, the new solution exploits a polarization diversity technique.
The solution leverages on the fact that two orthogonally polarized light beams don't interact with each other. This allows to collapse the two arms in a single fiber and makes the sensor almost immune to external reciprocal agents acting on the fiber sensor, such as mechanical noise.
An extensive laboratory analysis of the impact of the type of fiber implemented in the sensing head of the sensor has been also carried out. In particular a comparison between highly birefringence (Hi-Bi) spun fiber and rare earth terbium doped fiber has been performed, especially in term of sensitivity and stability of the measurement.
To conclude, in this thesis, a systematic analysis was carried out regarding the main factors that still prevent the use, in an uncontrolled environment, of a sensor based on fiber optic technology to reliably measure, the electric current using the Faraday effect and two novel implementations that aims to overcome those limiting factor have been presented.

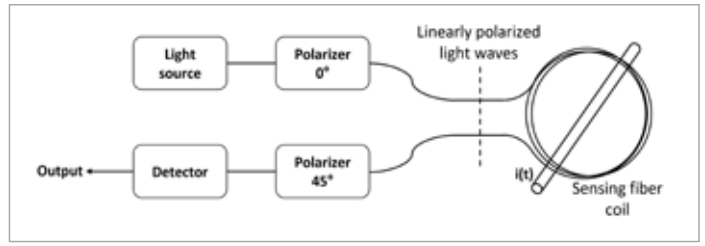

Fig. 1 - Schematic representation of a Fiber Optic current Sensor (FOS)

SINGLE-PHOTON AVALANCHE DIODE ARRAYS FOR QUANTUM-ENHANCED IMAGING AND SPECTROSCOPY

Francesca Madonini - Supervisor: Prof. Federica Villa

The second quantum revolution describes all new technologies enabled by quantum mechanics, not only to describe the physical world, but to address, control, and detect individual quantum systems. For example, observing the behavior of individual atoms or photons enables the achievement of phenomena such as quantum superposition and entanglement. The European Union has launched the Quantum Technologies Flagship initiative to push innovation in four main directions: quantum communication, to transmit data more securely; quantum simulation, to reproduce quantum behavior in well-controlled systems; quantum computation, to speed up computations; and quantum sensing and metrology, to improve measurements performance. Developing real-world applications is quite challenging, mainly because quantum systems require control to preserve their quantum properties from noise, decoherence and alterations caused by each observation. Just to mention the field of quantum measurements, the need to sense single quantum particles, rather than a macroscopic sample, requires detectors with single-photon sensitivity. Today, the most widely used are Single-Photon Avalanche Diodes (SPADs), which provide good overall performance together with the typical advantages of microelectronics, such as reliability, robustness, and compactness. They are also well suited to the development of large-format arrays for imaging applications.

Purpose of my Ph.D. research was the development of SPAD-based microchips and camera systems with high performance, targeted to: quantum-enhanced imaging, Raman spectroscopy; and quantum sensing.

Quantum imaging

Temporal photon correlation detection is a common ground for a variety of quantum imaging schemes. Thus, the development of high-fidelity photon coincidence sensors acquires key importance in the challenge of real-world quantum imaging platforms. In this context, I designed two geometrically identical SPAD imagers, but with definitely different electronics and on-chip processing: a so-called Event-Driven (ED) SPAD array and a Frame-Based (FB) SPAD array, both including 24 × 24 SPAD pixels with 50 µm pixel pitch, 10 µm SPAD diameter, and 60% equivalent fill-factor with microlenses mounted on-top.

The ED array features a TDC-free event-driven architecture that acknowledges an occurred photon coincidence and directly transfers the spatial coordinates (addresses) of the involved pixels in 330 ns, which does note scale with the pixel number and represents the lowest readout time among existing arrays for quantum imaging. The FB array implements a photon-timing approach with specific power-saving arrangements and fast readout to avoid overhead due to useless data. While reading the whole array takes 3 µs, by skipping the array rows in which none of the pixels detected photons, the readout time decreases to 240 ns in case of no photon detected over the entire area.

The post-silicon characterization of both array chips showed 43% peak PDP at 535 nm, 28 cps median DCR at room temperature, and $2.5 \cdot 10^{-4}$ maximum crosstalk between adjacent pixels.

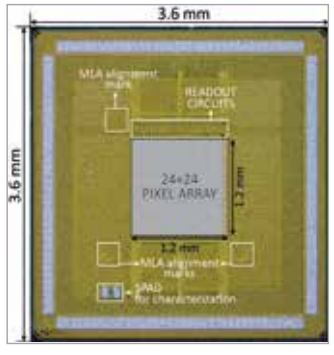

Fig. 1 - Chip layout of the Frame-Based 24 × 24 SPAD array.

Quantum imaging protocols carried out at ICFO (Barcelona, Spain) and INRiM (Turin, Italy) successfully validated the arrays' operation and proved their suitability for photon correlation detection, with both imagers ensuring minimum coincidence window around 2 ns. Future steps will include detailed design error fix, so to improve the overall performance for the future production of larger format arrays, such as with 48 × 48 or 96 × 96 pixels, since both the imagers consist of a modular core seamlessly scalable for further improvements in large-field and highly spatially resolved quantum microscopy.

Raman spectroscopy

The second valuable result of my Ph.D. research is the development of a test SPAD detector for ultrafast Raman-based protein sequencing. It consists in a 16 × 4 SPAD array operated in Time Gated Single Photon Counting (TG-SPC) mode, with both "hard-gating" and "soft-gating" time filtering modality (i.e., at SPAD level and at counting logic level, respectively), and readout time lower than single amino-acid translocation time (~1 µs) for single amino-acid Raman spectrum analysis. The SPAD array chip has been fabricated in a 40 nm CMOS technology, never employed before in our SPADlab at PoliMi. The chip has been characterized in depth, resulting in 44% peak PDE at 540 nm, afterpulsing below 0.1%, 3 ns minimum "hard-gating" window with 200 ps maximum rising/falling edges, 1 ns minimum "soft-gating" window, 23 ps SPAD timing jitter, and 200 cps median DCR.

Since the chip validation brought successful results, the imager is ready to be extended to a larger format, such as 128 × 4 or 256 × 4 pixels, with more advanced features tailored to ultrafast Raman spectroscopy measurements, e.g., on-chip gating generation (rather than externally provided signals, as currently) for <500 ps resolution in gate window width and position, and high-speed user-selectable readout for selective Raman bands investigation. Moreover, one of the advantages offered by the 40 nm CMOS node is the possibility to easily switch the planar design into a 3D stacked one, improving fill-factor performances by positioning the SPAD array in the top tier, while the processing electronics will be placed beneath it and specific hybrid bonding will be used in between them as connections. Eventually, the fabricated 16 × 4 SPAD array is ready to be exploited in experimental setups to detect ultrafast Raman, coupled to specific optics and protein translocation structures.

Quantum sensing

The last valuable result of my research is my contribution in the design of a novel detection system optimized for light extraction from single photon emitters that can be used as quantum sensors. Indeed, thanks to "Progetto Rocca" fellowship, I spent the last semester of my Ph.D. (March-August 2022) at Massachusetts Institute of Technology (MIT), Boston USA, hosted by Prof. Paola Cappellaro and her Quantum Engineering Group (QEG).

The work included the development of a new optical setup for quantum magnetometry, including a TG-SPC SPAD camera ("SPC3" camera commercialized by the POLIMI spin-off company MPD) and an ensemble of NV centers in diamond; the collection and processing of SPAD data through a new MATLAB interface; a benchmark of performance of the new system. First measurements have shown temperature and magnetic field gradients over the sample, not visible with previous non-spatially-resolved detectors. Currently, other experiments on NV charge transport are in progress. The goal is to lower the measurement time in observing dynamic behaviors, while providing µm-scale spatial resolution. The result of the collaboration was also to promote SPAD research at PoliMi to MIT staff member and to potentially find new SPAD cameras applications.

CHARACTERIZATION AND MODELING OF EMBEDDED PHASE CHANGE MEMORY DEVICES

Octavian Melnic - Supervisor: Prof. Daniele Ielmini

Since the dawn of the information technology age, non-volatile memories have been in increasing demand: the ever more capable computing machines need to store and access digital data in a fast and reliable way. The most successful memory technology up to now has been the flash memory, both for consumer-grade mass storage and for more specialized applications such as embedded memories in automotive or industrial microcontrollers (mC). Yet emerging memory technologies are still being developed as low-cost alternatives to flash, especially for the embedded applications. One of the most promising alternatives is Phase Change Memory (PCM), the object of study of this dissertation.
PCM cells have a chalcogenide alloy as the active material, and to store the digital data the material is made either crystalline (conductive) or amorphous (resistive) by a current pulse. Thus, the memory cell is a programmable resistor. The further integration in a memory array can use a standard selector device such as a MOSFET or BJT transistor. Beyond the conventional use of emerging memories for digital data storage, one appealing application is their use to store the synaptic weights in hardware-implemented neural networks (NNs). This dissertation also shows some use cases where PCM cells were successfully used in implementing such circuits with various architectures, and their performance experimentally demonstrated.

PCM reliability modeling

One advantage of PCM is the ability to be integrated in the back end of line in a standard CMOS process by adding very few process steps. This makes it a perfect candidate as embedded memory because it enables the addition of a small and low-cost embedded mC to many specialized integrated circuits, thus making these ICs smart and programmable. A very useful property of PCM is the ability to fine-tune the alloy composition in order to obtain certain electrical or reliability performances. The PCM studied in this dissertation uses a Ge-Sb-Te (GST) alloy heavily enriched with germanium. Such Ge-rich chalcogenide results in a very good data retention at high temperatures: a useful property for automotive-grade mCs and for the cases where the mC code must be programmed straight after production, before the high-temperature soldering on a circuit board. One of the downsides of PCM is the resistance drift over time. In well-studied alloys such as $Ge_2Sb_2Te_5$ the resistance drift impacts only the high-resistive state. The aforementioned germanium enrichment of the GST alloy comes at the cost of introducing resistance drift also to the low-resistive state. By drifting towards higher values this has the risk of flipping the read digital state, therefore it is a memory reliability issue. To better understand and predict the PCM resistance evolution over time, a drift model was developed in MATLAB. Experimental data was obtained from measurements on embedded PCM fabricated by STMicroelectronics in the Bipolar-CMOS-DMOS (BCD) 90-nm technology. The data was collected both from single analytical cells and from memory arrays.
PCM resistance drift is generally well-described by a power-law function of time, with the exponent n which describes the drift rate. But such simple description is valid at constant temperature because the value of n itself is a function of temperature. The model implemented a description of drift as the structural relaxation of defects in the amorphous material or in amorphous residues in the case of low-resistive states; such relaxation phenomenon leads to fewer and more distant trap sites, which in turn decrease the probability of carrier hopping events, thus increasing the measured resistance. The model considers a distributed activation energy for the annealing of various defects in a single cell. The internal state of the material is simulated as the partial and gradual annealing of defects from low to high activation energy, as a function of the temperature. The evolution of the material structural relaxation is then transformed into the activation energy for conduction, and the model also introduces stochastic variations at this step to account for the measured variability. The other phenomenon impacting PCM resistance, the crystallization of the amorphous material (or amorphous residues in the crystalline material), is also modeled in a similar way, with the time-to-crystallization being temperature-activated with a particular activation energy. A Monte Carlo implementation of the described model takes into account both the cell-to-cell variability of initial resistance, drift rate and crystallization time, and also the correlation between initial state and drift rate. Experiments at various constant temperatures were used to calibrate the model parameters and stochasticity. But the model is capable to run at a variable temperature profile, therefore further cross-temperature experiments were used to evaluate the model predictions, with good results. The capability to predict the impact of high-temperature soldering on the data integrity is especially useful in an industrial setting. More broadly, the model can be used to predict PCM reliability in various use case scenarios before on-field stress tests.

PCM in NNs

Artificial neural networks have been for a long time in the realm of computer science. In recent years their successful implementation in software encouraged the research of equivalent hardware architectures. The learning process of NNs implies the update of synaptic weights and the retention of learning implies such synaptic weights should be stored in a non-volatile memory. Many memory technologies were demonstrated suitable for such a task, and PCM are particularly suited for several reasons, some of which: they are two-terminal devices, have continuous range of programmable resistance, and they have gradual programming by applying sequential pulses. This dissertation proposes several neural network circuits with different architectures, where PCM cells are used both as synapses and as functional parts of a neuron.
The issue of resistance drift was tackled in a differential read scheme by comparing each synapse with a reference cell programmed in an intermediate state which drifts at the same rate as the synaptic weights. Another issue in the learning process in NN architectures is the catastrophic forgetting in the cases when the NN has to continually learn new classes to be classified. The proposed architecture combines supervised convolutional and fully-connected NNs, as well as unsupervised Spike-Timing-Dependent Plasticity (STDP) in order to mitigate this issue. Both the supervised and unsupervised parts of the network are implemented with PCM cells as synapses, and its functionality is tested for a simple image recognition and classification application.
In another architecture, the PCM variability is also leveraged inside a stochastic neuron circuit. A deterministic clock signal is fed to each neuron of a Recurrent Neural Network (RNN), and the PCM is used to gradually integrate the clock pulses in the Integrate-and-Fire circuit. The PCM cells from different neurons perform differently device-to-device and cycle-to-cycle, therefore resulting in a stochastic output from each neuron. The implemented RNN is aimed at solving constraint satisfaction problems, where this computational noise is fundamental for not getting stuck in a local solution but to converge to the global optimum result. In this hardware NN the synapses were implemented as couples of PCM cells, demonstrating the use of PCM in a single circuit both as hardware synapse and as integral part of the hardware neuron.
The proposed architectures are small-scale functional prototypes which pave the way to future research, which will have to tackle the integration of PCM in deeper, denser, and faster NNs.

OPTIMIZATION TECHNIQUES FOR VIRTUAL BASEBAND FUNCTION PLACEMENT FOR 5G RADIO ACCESS IN METRO-AREA NETWORKS

Ligia Maria Moreira Zorello - Supervisor: Prof. Guido Maier

Mobile operators face the challenge of deploying a flexible network to handle several emerging use cases. These applications, typically divided into enhanced Mobile BroadBand (eMBB), ultra Reliable Low Latency Communication (uRLLC), and massive Machine Type Communication (mMTC), present very specific requirements in terms of bandwidth, device density and latency. A new Radio Access Network (RAN) for 5G was proposed to provide more dynamic, low-cost and energy-efficient network management, resource allocation, and service provisioning compared to 4G. One of its main features is the deployment of functional splits. They define the separation of RAN baseband functions into different radio-network units: Radio Unit (RU), Distributed Unit (DU), Centralized Unit (CU). It enables decentralizing the RAN processing to improve network flexibility and potentially reduce costs.

Operators can take advantage of this functionality together with virtualization technologies to improve the network management and costs in terms of power consumption and capital expenditures. Nevertheless, this architecture brings new challenges to network operators. The optimal placement of these units is non-trivial: it depends on the application requirements, on the expected traffic volume, and on the costs and power derived by the placement. In addition, application requirements and expected traffic may vary in time due to users instantaneous needs. Therefore, performing hourly reconfigurations to comply with the traffic fluctuation may lead to service disruptions. Consequently, it is important to plan the baseband Virtual Network Functions (VNF) placement and service provisioning in advance to ensure efficient network reconfiguration.

This PhD dissertation optimizes the baseband function placement in mobile networks to minimize the overall network power consumption and guarantee quality of service. It proposes several optimization techniques to enable static and dynamic resource allocation, including mathematical programming, heuristic, black-box, machine learning, and auction. For this, this research integrated several mathematical models to define the requirements of functional splits in 5G RAN to be used in the optimization models produced a series of optimization algorithms that enable an efficient baseband function placement over network nodes both in static and dynamic scenarios.

We developed mixed integer linear programming models and heuristic algorithms to solve the baseband virtual network function problem in metro-area networks considering different scenarios. There is a significant effort in the literature to solve it. Nevertheless, aspects related to the costs and to the functional split constraints were not fully considered. The goal of the models proposed in this work is to minimize the overall system power consumption, including both network and node components. This optimization is subject to all split-related constraints and to the service being carried. We implemented the proposed algorithms to compare them to the current 4G distributed architecture as well as other solutions available in the literature. The results demonstrated that the partial centralization consumes less power than the current 4G distributed architecture and guarantees the compliance of all service and split requirements. Moreover, by carefully planning the baseband functions placement, it is possible to achieve more power-efficient and flexible results.

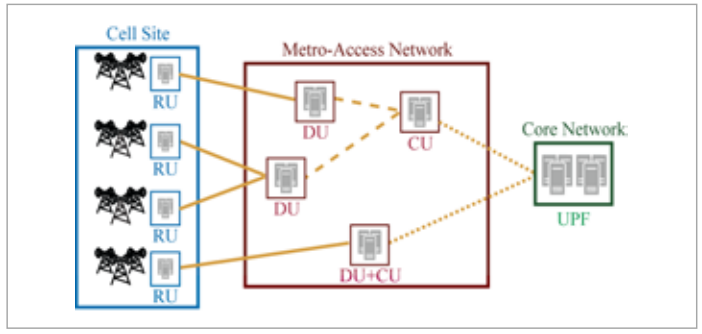

Fig. 1

Reconfiguring the network on a real-time basis is impractical, as it could lead to disruptions of service. In order to plan the baseband VNF placement in advance, the works available in the literature typically deploy machine learning algorithms, because they can provide accurate traffic forecast and improves the applicability of optimization frameworks in real-time scenarios. However, unpredictable events may create perturbations in these patterns. Therefore, the traditional techniques used to predict the traffic and then optimize the placement fail to provide solutions that ensures feasibility in real-time. Indeed, they do not guarantee that the real traffic could be carried by the anticipated placement. Instead, we proposed two different techniques to perform early and efficient planning of the baseband VNF placement. To accommodate unpredictable traffic variations, we need to reserve an extra buffer capacity in the nodes when computing the optimal placement. We proposed a multi-task algorithm that forecasts 1) the traffic mean used in the optimization objective function (cost estimation), and 2) the quantile value used in the constraints to ensure capacity. By forecasting the mean and quantile expected traffic, the deployment of the artificial capacity is no longer necessary. Furthermore, we propose a novel framework that uses black-box optimization to train a traffic prediction algorithm based on the optimization outcomes. The goal is to minimize a loss function related to power consumption and constraint violation to ensure that the predicted placement is feasible and that its consumption is close to optimal. The results of these works confirm that machine learning techniques cannot be blindly used into optimization models. However, by applying intelligent mechanisms that overestimate the demands, we can ensure the feasibility of the placement in real time.

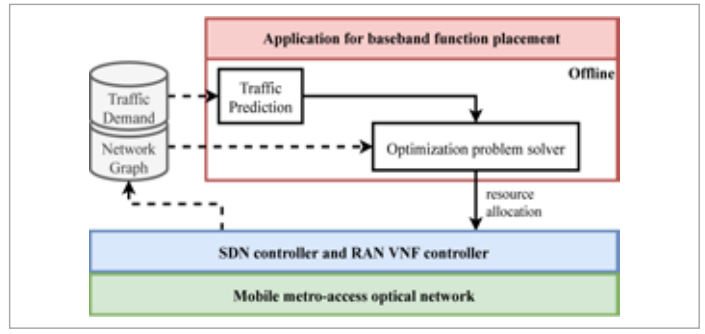

Fig. 2

MIXED SIGNAL GENERIC TESTING IN PHOTONIC INTEGRATION

Matteo Petrini - Supervisor: Prof. Andrea Melloni

The aim of this research is the investigation of novel testing approaches for Photonic Integrated Circuits (PICs). PIC testing (especially in mass production scenario) is one of the bottlenecks, that limits the diffusion and the use of these promising technologies. In fact, high-index-contrast PICs (such as those made of Silicon, like in this work) are very sensitive to the fabrication tolerances. For example, imperfections of 1 nm on the optical waveguide dimensions (which are 220 nm by 500 nm, for height and width, respectively) lead to a 100 GHz-shift (in frequency domain) of the device transfer function. The higher the complexity of the PIC the more detrimental are these effects.
The necessity of active control arises, for both testing and normal use. Before performing the optical testing (i.e., state if a PIC is compliant with specifications or not), a pre-tuning is, in fact, required.
Within this research, two novel architectures (tuneable WDM filter and controllable optical delay line) have been designed and validated, along with their tuning recipes, which are, as discussed, mandatory for the correct functionality of the optical circuit.
The first device is a fourth order Microring Resonator (MRR)-based filter [fig. 1(a)], complying quite challenging spectral features. It has a large passband (40 GHz), quite steep roll-off (30 dB, 50 GHz far from central wavelength) and infinite Free Spectral Range (FSR), thanks to the exploitations of MRRs having different (and non-proportional radii). It can be tuned along the C+L telecom band, in a non-traffic affecting way (i.e., hitless tuning), thanks to the actuators that control (individually the phases of the rings).
In the field of filter testing (not only in Photonics), all the solutions rely on the measurement of the spectral response of the device to be tested. But this is usually slow and indeed constitutes a bottleneck for the testing routines. At the same time, concerning the context of complex filters' tuning and locking, the literature is plenty of approaches and solutions. However, none of them is sufficiently fast and/or repeatable, to match the tight features required during the testing activities, in terms of throughput and robustness.
In this work, always avoiding any spectral measurement, we adopt a customized Power Spectral Density (PSD) and injecting it into the filter we shape its own transfer function onto the incoming signal. This arbitrary PSD is obtained by cascading flat spectrum source with a frequency-shaper (which is our golden reference, REF). This is coupled input of the device under test (DUT). By reading the output power by using a photodetector we are able to state how different is the DUT from the REF, from the spectral point of view. In other words, we demonstrate a correlation between the optical power coming out from the cascade, when the DUT satisfy these (sufficient) conditions: i) a mostly flat-top amplitude spectral response, with steep transitions and with a high rejection both in-band and out of band; ii) in addition, if the DUT is a tuneable-bandwidth device, the maximum bandwidth of the DUT has to be narrower or equal to the bandwidth of the REF.
This quantity is also used as a feedback signal to tune the DUT (the phases of the single MRRs are in fact independently tuneable exploiting thermo-optic actuators). By maximizing the power, according to a gradient descent algorithm, the similarity between REF and DUT is maximized, as much as DUT topology allows. Once tuning is complete, optical testing on the DUT can be executed.
Furthermore, by using this technique, the DUT spectrally replicates the REF. To obtain these results, custom electronics (composed of a commercial controller, a custom analog board and a custom probe card for electrical access of the PIC) to manage the whole testing/tuning loop has been designed and realized. This infrastructure is capable to test a filter in less than one second (lower-bounded by the time constant of the thermal actuators), which is fully compliant with massive silicon testing (an entire wafer would be tested in less than 8 hours).
Then, the achieved working points of the actuators are stored in a LookUp Table (LUT), to avoid another tuning stage during normal operation of the DUT. The obtained LUT shall be updated upon the users' needs (different wavelength channels to be filtered), strong temperature gradients or presence of high-power input signal (above 10 dBm). In fact, when silicon PICs are excited by high intensity signal a plethora of effects arises, having a significant impact on the phase and on the amplitude of the waveguides the PIC is composed of. These effects are even more detrimental if they affect devices made of resonant building blocks (due to the field enhancement).

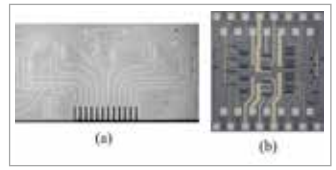

Fig. 1 - Microphotograph of tested devices: (a) optical filter and (b) true time delay line.

In particular, their spectral responses are heavily distorted. By monitoring a high-power probe signal propagating through the filter, we demonstrate that the gradient descent-based control loop can be successfully exploited to counteract the (phase) nonlinearities that are triggered, mainly by cooling down the thermo-optic actuators (since most of the observed nonlinearities induce a red-shift). When the convergence is reached, not only there is no more evidence of any nonlinear distortion (in frequency domain), but the new actuators' status can be stored in an updated version of the LUT.
Finally, we consider another complex (and novel) photonic PIC, a delay line [fig. 1(b)] capable to break the delay-bandwidth product constraint (nominally this figure of merit shall be around 1.75, while for delay lines in literature the best case is close to 0.25). This is composed by four nested Mach-Zehnder Interferometers (MZIs), arranged in such a way that the (N+1)-th MZI is the unbalance of the N-th MZI. On top of that, each single MZI is composed of a pair of tuneable couplers. Therefore, each MZI has three degrees of freedom (again controlled by thermo-optic actuators), the two coupling coefficients (supposed to be symmetric, to control the delay) and the optical unbalance (and thus the center of the passband). Hence, testing this PIC is quite challenging because both time and frequency behaviors must be validated. The direct measurements of the spectral response and of the group delay slow down the testing throughput. Anyhow, exploiting the electronics discussed before, a control loop to pre-calibrate the PIC is designed. In this case the feedback signals are provided by photodetectors placed at the output of every single tuneable coupler. In fact, it can be demonstrated that the splitting ratio (K) of each coupler has a correlation with the overall group delay.
The combination of novel testing recipes and custom hardware properly works, and we believe that this research provides some attractive solutions for these urgent needs. However, there is still room for improvement. First of all, bandwidth of the designed electronics is <100kHz (since sensors and actuators are quite slow) and for future developments this limitation has to be overcome. Then, the designed probe card must be suitably modified, to perform (simultaneously) both optical and electrical access of the DUTs (enabling for wafer level testing). More in general, concerning the broader topic of Photonic Testing, several steps should be performed to really spread Photonics in the mass-market, starting from the standardization of the chip layouts and from the uniformity of testing equipment and routines.

A STUDY ON DEEP LEARNING METHODOLOGIES APPLIED TO GEOPHYSICAL INVERSE PROBLEMS

Francesco Picetti - Supervisor: Prof. Stefano Tubaro

Introduction
Exploration Geophysics aims at estimating physical properties of the Earth subsurface from seismic data acquired close to the surface. For physical reasons, data are band-limited and corrupted by a great variety of noises, disturbances, and other phenomena.
Moreover, the acquisition campaigns result in massive datasets, limiting the algorithms to be computationally feasible.
Seismic data show a great variety of statistically relevant and independent patterns. To tackle the aforementioned challenge, I devise Deep Learning methods to solve several geophysical tasks by learning such patterns.

Processing - Seismic Interpolation through Deep Priors
In seismic exploration, imaging and processing algorithms greatly benefit from dense and regular grids. Unfortunately, the vast majority of acquisitions are coarse and degraded by a number of physical and economical factors.
To recover a functional grid, I approach the interpolation problem through Deep Priors, which are neural networks that precondition the inverse problem under the rationale that they might capture the innermost self-similarities of seismic data.
Figure 1 depicts the Deep Prior optimization scheme. The minimization variable is the interpolated data **m**, casted as the output of a neural network $f_\theta(\mathbf{z})$, **z** being a noise tensor that excites the generative modes of the network. The loss function compares the coarse acquired data **d** with a subsampled version of **m** through the sampling operator **S**. Notice that no ground truth is required. In fact, the minimization is performed on each set of **m** and **d**. Therefore, every time we change the data **d** or the sampling **S**, we must solve a new inverse problem.
In other words, the inverse problem is thus recast into the network parameters space, reshaping the objective functional and, possibly, reducing the null space. Therefore, the architecture design is critical for networks acting as preconditioners; every example required manual fine-tuning of the hyper-parameters.
Through an extensive experimental campaign, I demonstrated the Deep Priors to be a viable way to tackle 2D and 3D interpolation, at the cost of computational burden. Moreover, I pro posed an additional regularization term to handle highly aliased data based on the directional Laplacian.

Imaging - High quality images from coarse data
Here, I focus on a seismic imaging technique called Reverse Time Migration (RTM), that retrieves an image of the interfaces between rock layers of the subsurface from seismograms recorded at the surface. The quality of this approximation is degraded by several factors such as aliasing, limited aperture, noise, and non-uniform illumination, and complex overburden, which is more and more frequent in exploration areas. As a result, images obtained by RTM are contaminated by migration artifacts, uneven amplitudes, and limited bandwidth.
Instead of trying to study and designing an accurate and sophisticated imaging operator (e.g., regularized least-squares RTM, etc.), we can split the task into two different problems: a standard imaging operator followed by a post-processing operator.
In the thesis, I proposed such a operator to turn images migrated from a cheaper coarse acquisition into image migrated as if we had a finer acquisition geometry.
In particular, I train a U-Net in an adversarial scheme: the U-Net, called *generator*, is flanked by a a fellow CNN called *discriminator*. The latter is a binary classifier trained to discriminate whether its input is a pristine image or synthesized by the generator. Acting only during training, the discriminator can be seen as a regularizer driving the generator to output realistic images.
On both synthetic and real datasets, the results are pretty promising, showing that the proposed solution does not simply overfit the training set but can generalize. Unfortunately, due to training set limitations, a few details are lost on the test set, as depicted in Figure 2. However, the overall computational time dramatically reduces. For instance, the time needed to generate a single image with the dense geometry was around 40 minutes. In the proposed pipeline, the time required to create an output image of the network was about 2 minutes, almost entirely dedicated to migration with the coarse geometry to feed the network with.

Interpretation - Detecting landmines in Ground Penetrating Radar scans
Finally, I presented a humanitarian demining system that builds upon an auto-encoder anomaly detection scheme.
Specifically, such auto-encoder is trained to reconstruct mine-free GPR acquisitions; thus, while reconstructing mine-contaminated data, as depicted in Figure 3, it introduces some errors, clearly measurable in its hidden space. This error acts as an anomaly metric that, though thresholding, detects the presence of buried threats with a 90% true positive rate at a false positive rate of 0% even in cross-dataset scenario.
Furthermore, this light-weight architecture does not need a significant amount of training data; it can be retrained on a very small portion of safe soil, making the system a viable solution to fast-deployment in humanitarian scenarios.

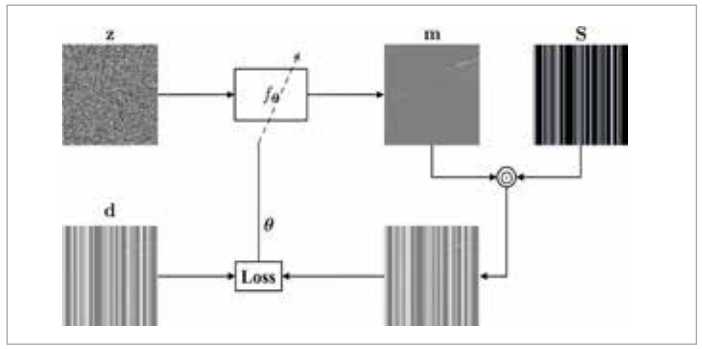

Fig. 1 - Deep Prior interpolation scheme.

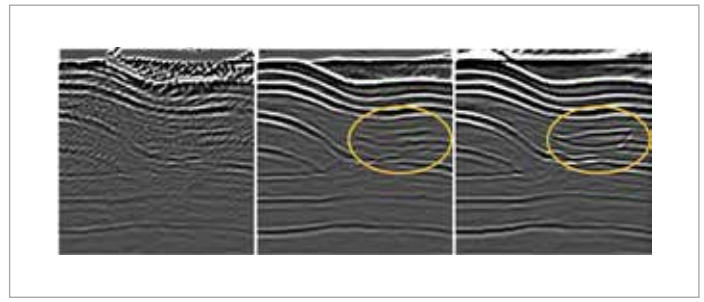

Fig. 2 - Test patches of the coarse-to-fine generative network: input coarse image (left), output (centre) and target (right). Overall, the imaging quality is preserved, in terms of kinematics and dynamics. In the yellow circles, we can spot a seismic feature (i.e., a channel) that is not recovered by the network.

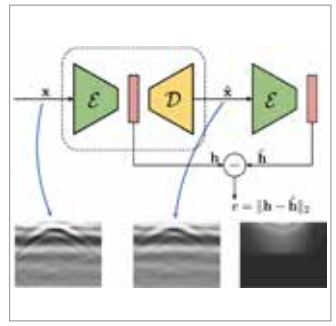

Fig. 3 - Anomaly detection scheme for landmine radar signatures.

ELECTRONIC BIO-RECONFIGURABLE IMPEDANCE PLATFORM FOR HIGH SENSITIVITY DETECTION OF TARGET ANALYTES

Paola Piedimonte - Supervisor: Prof. Marco Sampietro

Over the years, the spread of viral infections has posed a continuous threat to public health increasing the demand for more sensitive diagnostic methods and platforms. For the diagnosis of viral infections, it is possible to directly detect the whole virus or to determine the antibodies produced against virus proteins during and/or after the virus incubation period. The analytes such as viral nucleic acids (DNA and RNA), viral proteins, intact viral particles and antibodies can all be used for early diagnosis. In particular, combining the advances in electronic systems, microfabrication techniques and molecular bio-recognition, it is possible to realize extremely sensitive and compact platforms for a point-of care configuration.

Under this light, the main purpose of this work has been to revamp current diagnostic approaches for these epidemic diseases by using technological strategies, merging the most sophisticated biochemistry recognition mechanisms into an advanced electronic miniaturized platform whose detection sensitivity may reach levels down to the single target.

The developed biosensor system is based on the impedance variation between microelectrodes upon the capture of the target analyte, grafted over the biosensor surface through specific probe immobilization. A properly functionalized nanobead has been used to enhance the electronic signal since the dimensions and the structure of the target moiety would not allow the direct label-free detection.

The system proposed is composed by a biosensor integrated into a microfluidic path and electronically accessed to perform impedance detection by custom electronic board featuring high portability and multichannel operation. In this way, multiple sensing sites in parallel can be addressed, extremely important from a diagnostic point of view since they will allow performing multiplex analysis starting from a single clinical sample.
The core of the system is represented by gold interdigitated microelectrodes with an active area of 90x90 μm^2 and a comb dimension of 3 µm. The IDEs are designed in a differential configuration, reference and active sensor, to counteract all possible mismatches such as temperature fluctuations and variations in the ion content of the sample under test. The sensors were fabricated in the cleanroom facilities of the Politecnico di Milano (Polifab). The electrodes are realized on a borosilicate substrate with Cr/Au layer with a final thickness optimized for good conductivity and easy lift-off procedure. An SU8 protective layer ensures parasitic capacitance reduction in liquid environment and a final oxygen plasma guarantee a good superficial wettability.
In Figure 1, the experimental results shows that beads count by a truly differential sensors architecture operated in a lock-in scheme is very effective in monitoring specific IgG antibodies in human serum and buffer down to few single counts resolution, i.e. a LOD of 88 pg/mL. The sensitivity obtained by the system reaches and possibly outperforms other methods yet operating in a simple and clear protocol as demonstrated in the comparison of the proposed platform response to human serum positive to DENV with a commercial Dengue virus IgG kit. The technique therefore has the potential to become a valuable early diagnostic tool when levels of circulating antibodies are still low and might go undetected by conventional methods, in particular in regions where sensitive ELISA and PCR methods are not readily available. The precise temperature control channel of the electronic board and the on-chip temperature serpentine near to the IDEs structures open the possibility to control and actuate the temperature of the liquid over the chip (precision below 20 ppm), for example, in the denaturation/hybridization processes.

The system is perfectly suited to be easily configured for including novel probes able to detect concurrent viral strains or viral variants. Indeed, by simply modifying the preparation of the biosensor chip, the antibody and the probe, the differential impedance sensing concept can address a wide range of pathogens and diseases, making it very flexible to approach industrial interest, yet reaching clinically breakthrough results in the initial stages of infection.
As shown in **Figure 2,** the bio-reconfigurability and the multiplexing feature of the system has been successfully verified with oligonucleotides (DNA) detection down to pM target concentration. For this purpose, the commercial lock-in system has been replaced with a custom made electronics realized as an expansion board of the FPGA module XEM 7310.
A strength of the proposed system is the flexibility in addressing a wide range of targeted pathogens simply by updating the surface probe. Moreover, this proposed strategy offers the possibility of lab on chip development for the detection of infectious diseases with increased sensitivity down to the single target detection. The final system enables simple and effective bio-reconfigurability, leveraging advances in biomolecular recognition through appropriate selection of bio-probes, and allows extending the applicability of multiplex sensing to a broad range of needs.
For future work, the ongoing SARSCoV-2 pandemic represents a starting bench for the development and implementation of such a new biosensor. SARS-CoV-2 will be targeted by direct capture of the entire virus in solution, using DNA-labelled antibodies directed against the SARS-CoV-2 spike protein. This strategy will bring advantages in terms of reduced sample handling and processing (meaning less contamination and no loss of viral components) and no need for harsh chemicals nor for sample purification or amplification, resulting in a reduction of time and cost of the analysis.

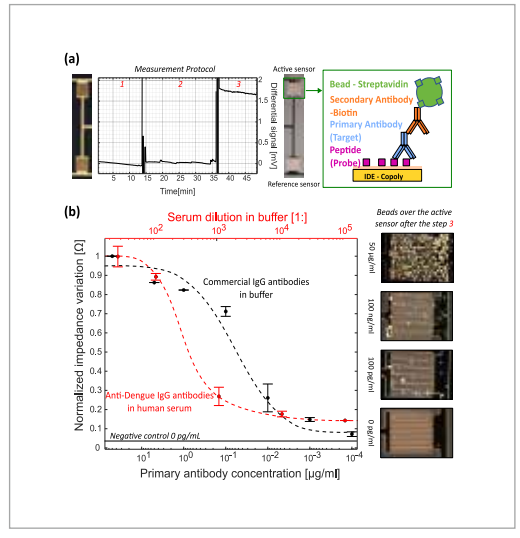

Fig. 1

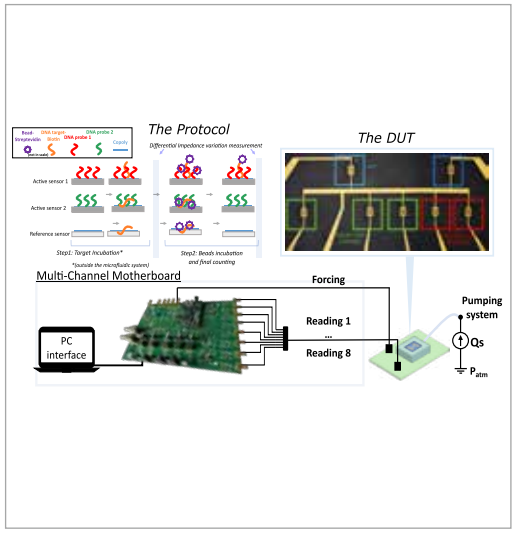

Fig. 2

TOWARD AUTONOMOUS VINEYARD OPERATIONS: DEVELOPMENT AND IN-FIELD TESTING OF A SELF-DRIVING TRACTOR

Solomon Pizzocaro - Supervisor: Prof. Matteo Corno

Automation has been used in farms for almost a century. Up to these days, the main automated operations are spraying, pruning, weeding, and soil monitoring. Some of these tasks still require the presence of the operator which fills the gap in perception and mobility. In recent years, thanks to the advancements in sensor technology and autonomous driving algorithms for road vehicles, more and more attention is being placed on self-driving tractors and complex crop management operations. Although there are various projects driven by these interests, fully autonomous navigation in orchards and many horticulture operations remain open problems. In the first place, it is clear the need for robust and tailored perception solutions, that can cope with the seasonal changes and the unstructured nature of the agricultural field. Furthermore, the uneven and soft ground makes the modelling and control of the robot's motion very challenging. This work aims to address these challenges with the analysis and development of the main modules for autonomous grape harvesting. In particular: localization; environment perception; navigation; and grape cutting point detection. The proposed modules were developed and validated for an unmanned tracked vehicle navigating in vineyards.

The proposed localization scheme fuses information coming from sensors available on the tractor that have a higher degree of dependability. A particular focus is given to the magnetometer sensor that is the core of the localization scheme. We show that an online accurate calibration of this sensor along with a sensor fusing algorithm (using IMU and track encoders) provides an accurate and robust absolute pose estimate during times of GNSS denial.

To solve the navigation problem, we propose a single controller that, given a path, drives the vehicle inside and outside the rows taking into account obstacles and crops vegetation. The proposed solution is framed on top of the ROS Navigation Stack, which makes it highly replicable for other applications. Path tracking is done on a local cost-map composed by two layers: a path cost-map (see Figure 1), that pools the robot towards and along the given path; an obstacles cost-map, which is built from the LiDAR data. To remove leaves, branches and other small objects from the obstacles cost-map, a row segmentation module is proposed. The row segmentation module is done with a custom implementation of the RANSAC algorithm with a two parallel planes model. We implemented the proposed modules for the navigation of a vineyard drone and tested in a real vineyard. The drone was able to track a path starting from the parking area, going trough all the rows and go back to the starting point. The RMS of the lateral error was 0.2 m, the RMS of the error from the middle of the rows was 0.15 m.

The last component proposed in this work is a computer vision software that recognizes grapes and compute the peduncle cutting point. After an analysis of the literature, we decided to implement the following work-flow: first, we train a YOLOv4 network to detect grapes and berries in images; then, we search for the peduncle in a Region Of Interest (ROI) on top of the grape bounding box; and lastly, we compute the cutting point on the peduncle. The proposed algorithm has an average cycle time of 12ms and a precision of 85% on a common laptop.

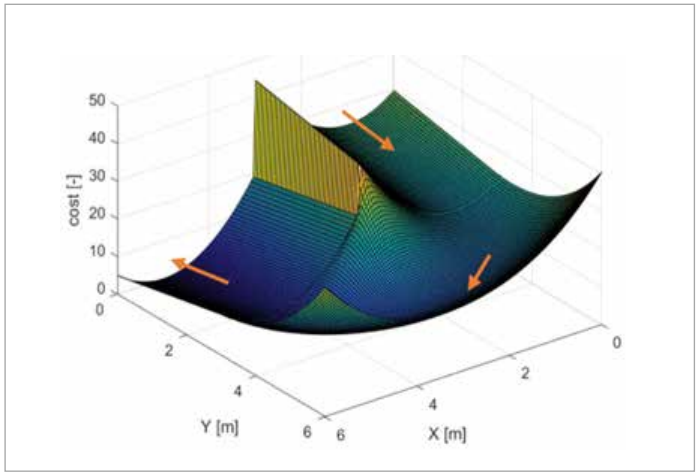

Fig. 1 - 3D visualization of the path cost. The orange arrows show the direction

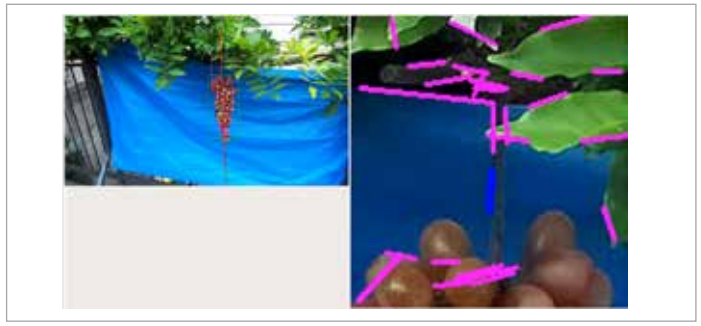

Fig. 2 - Grape, berries and cutting point on grape stem detection

DESIGN AND ANALYSIS OF ENERGY MANAGEMENT STRATEGIES FOR NON-STANDARD HYBRID ELECTRIC POWERTRAIN ARCHITECTURES

Stefano Radrizzani - Supervisor: Prof. Giulio Panzani

Environmental awareness and traffic congestion are at the basis of the transition to electrified vehicles, possibly light, small and shared. During this transition, hybrid electric vehicles (HEVs) are playing an important role, given their ability to combine features coming from complementary power sources and at least one uses electric power. The realization of an efficient integration of the multiple power source calls for energy management strategies (EMSs), i.e., control laws which optimally govern the energy flows in the system.

There exist different levels and configurations of hybrid electric powertrains. The internal combustion engine (ICE) itself, it can be made hybrid by electrifying the turbocharger shaft. Turbocharged engines compress the inlet air charge by recuperating energy from the exhaust gas, which is otherwise wasted. Thanks to the electrification of this process, the engine can be downsized and its efficiency can be improved. A higher level in the powertrain electrification can be reached by combining an internal combustion engine with one or multiple electric motors (EMs). Finally, a vehicle can be considered hybrid electric, if the electric energy storage is made with different technologies. For example, hybrid Li-ion battery packs merge complementary Li-ion cells technologies to manage the trade-off between power and energy density.

Generally, HEVs need an energy management strategy (EMS): a control law that optimizes the use of the multiple power sources. This optimization problem can be seen as a global optimization, in fact, the optimization horizon coincides with the entire life of the vehicle. Given the impossibility to know a-priori the complete use of the vehicle, many real-time solutions have been developed, here classified according to their level of optimality.
The simplest solutions are based on heuristic approaches, followed by the well-known Equivalent Consumption Minimization Strategy (ECMS) that reduces the global optimization problem to a local one to by minimizing an equivalent consumption at any time instant or by real-time solutions of the Pontryagin's Minimum Principle (PMP). The optimization problem can be also solved over a finite prediction horizon, applying Model Predictive Control (MPC) techniques.

In this scenario, where a high maturity on HEVs has been already reached, this thesis is oriented to non-standard hybrid electric powertrains, in order to provide novel contributions. In particular, we focus on a hybrid electric tractor, a human-powered series-parallel electric bicycle, and an electric racing car equipped with a hybrid Li-ion battery pack. With respect to traditional HEVs, these use cases introduce new challenges to face.

Hybrid electric tractor
When dealing with hybrid heavy-duty vehicles, such as agricultural tractors, it is not possible to consider them as a single category, when approaching the energy management problem. Indeed, each of them has unique features and driving-cycles. A tractor itself can be used in very different ways; indeed, it can move loads (transport scenario) or be involved in agricultural machining in field (working scenario). In this thesis, the focus is given to a hybrid parallel tractor in transport scenario. With respect to standard vehicles, the new challenge encountered during the design of an EMS for this prototype was induced by the presence of a built-in engine speed controller, typical of agricultural engines. Indeed, the speed controller regulates on its own the engine torque, so that it cannot be directly manipulated by the EMS. Therefore, the proposed EMS needs to be able to indirectly control the engine torque to improve the vehicle efficiency by controlling the electric motor only, in order to be plug & play with the built-in speed controller. The main contributions obtained while designing energy management strategies for this prototype can be summarized in the following main points: 1) the development of an MPC-based EMS that is plug & play with the built-in engine speed controller; 2) the reformulation of the ECMS into an efficiency framework, then extended to be integrated with the built-in speed controller; 3) the experimental evaluation in transport scenario of the ECMS-based approach, after its implementation on the tractor control unit.

Hybrid human-powered bike
The energy management design in hybrid human-powered bikes reveled to be very similar to the one proposed for the tractor, indeed the human power cannot be directly controlled. In this case, the energy management is designed to maximize the human efficiency. When sensors for oxygen and hearth rate are not available, the human efficiency can be associated to the comfort perceived by the rider. In this scenario, we proposed a comfort-oriented EMS, experimentally evaluated on a prototype. Particularly, the considered prototype is a series-parallel bike, where a free-wheel mechanism turns the vehicle from parallel to series by disengaging the chain and vice-versa. In human-powered vehicles, before designing an EMS, it is necessary to manage the human-vehicle interaction so to link the cyclist's behavior with the vehicle dynamics.
The main contributions on this topic are: 1) the extension of the virtual-chain approach for chain emulation in series bikes to series-parallel ones; 2) the extension of the virtual-chain to a virtual-bike framework, in order to impose also dynamical features on the systems; 3) the development and experimental evaluation of comfort-oriented EMS, based firstly on virtual-chain and secondly also on virtual-bike, in order to achieve better results.

Hybrid energy storage for electric racing cars
The last considered vehicle is a full-electric racing car, equipped with a hybrid energy storage, composed of different Li-ion cell technologies. In such a scenario, the goal is evaluating if hybrid solutions could provide an increase of performance, when the vehicle is pushed to the limits in order to minimize the race time. Toward to this aim, a co-design optimization problem is formulated to simultaneously optimize the energy management strategy and the battery design parameters, i.e., the coupling of different cell technologies and relative size. First of all, standard battery packs are considered in order to provide a validation of the proposed methodology; then, it is extended to hybrid battery packs to evaluate their potential benefits in race time minimization. Simulation results, considering Formula E as a case study, showed how the combination of high-energy Li-ion cells with high-power ones is able to significantly reduce the race time. Indeed, the hybrid battery pack provides a lighter solution (thanks to high-energy density cells) with high capability of regeneration (given by high-power ones), so to make the use of mechanical brakes practically unnecessary when considering the optimal sizing.

DEVELOPMENT OF CROSSPOINT MEMORY ARRAYS FOR NEUROMORPHIC COMPUTING

Saverio Ricci - Supervisor: Prof. Daniele Ielmini

With the advent of the Internet-Of-Things and with the ever-growing number of people gaining the possibility to purchase smartphones and tablets capable to store a large amount of photo, video, music and applications in a single portable device, the global amount of data has increased exponentially, which raises strong requirements in terms of energy efficiency and processing speed for data analysis. To satisfy these requirements, the computing performance of modern computers has increased steadily in the past few decades thanks to the scaling down of the transistor dimensions and the consequent higher density of information being stored in the same area, as predicted by Moore's law. The downscaling is now approaching its natural end mainly due to the increasing leakage of the complementary metal-oxide-semiconductor (CMOS) transistors due to their extreme miniaturization. If on one side we have reached a limit on data transport speed due to the transistors, on the other side we have to consider that there is an additional limit imposed by the fact that conventional computing systems are based on the von Neumann architecture, where memory and processing units are physically separated, which leads to an additional inevitable bottleneck due to the necessary data movement between the two separated units, which causes significant latency and energy consumption. This latency becomes significantly enormous when operation must be repeated thousand or millions of times, as happens to tensor products and matrix multiplications, where the operation between the elements of the matrices cannot be done in parallel but only one after the other, finally collecting all the results. Alternative computing approaches are becoming increasingly attractive to develop novel logics and neuromorphic computations in order to overcome Von Neumann bottleneck issues. Indeed, typical operations like image learning, pattern recognition and decision exhibit high computational cost for boolean CMOS processors, while, for human brain, they represent elementary resistive switching memories , also known as memristors, appear as one of the most promising technologies for in-memory computing, thanks to the CMOS-compatible fabrication process, the small area and the analog programming. Differently from conventional memories based on transistors, able to store binary values only, as 1 (transistor in pass mode) and 0 (transistor switched off), memristors can store information in the electrical properties, as the resistance (or conductance) for example, in an analog way. Moreover, by organizing these memories in a matrix configuration, also known as crosspoint architecture, in Figure 1, the matrix-vector multiplication is performed in one step only, carrying out all the single elements multiplications simultaneously exploiting the Kirchhoff's low. Because of the novelty, problems of reliability and integration with existing technologies affect the emerging memories and further studies are required to overcome the limits by optimizing the materials and their responses, the fabrication steps to be implemented in nowadays process flows and developing architecture designs and algorithms to exploit the innovative features and the strong parallelism of the physical multiplication. Merging novel in-memory, brain inspired and specialized architectures with nanoelectronics emerging memories is the goal of this Ph.D. dissertation. Different computing paradigms, namely neuromorphic computing and analog computing, have been explored and networks based on different emerging memories have been realized, demonstrating in hardware new computing concepts.

The first goal has been the development of a fabrication process flow for a suitable platform to be used as a starting point for the material analysis. The choice to host the memories into vias with controlled diameter allowed to avoid mismatches and to reduce the device-to-device variability. In most of the state-of-the-art memories, fabricated metallic lines separated by the oxide layer, the defects accumulated at the edges and at the corners of the line interconnections limiting the use of such memories due to the huge variability and low reliability. Pt/HfO2/Ti/Au non-volatile stack has been optimized by thinning the oxide layer till reaching the condition to be forming-free, where the forming initialization is suppressed (or comparable to a set operation) or even substituted by an initial reset, property which makes the devices suitable memories for a crossbar implementation. The tunability of the conductance was investigated both in quasi-static regime and in a pulsed one, revealing the devices can be linearly programmed between 1 µS to 500 µS by changing the compliance current or gradually reset by changing the reset amplitude. Crossbars based on forming-free RRAMs shows initial low resistive state values, in the order of 10 - 100 µS, which can be directly programmed without the need of an initialization. A suitable program and verify algorithm has been developed to automatically program the matrix considering all the possible sneak-paths, the IR drops and memory interferences. The feasibility of novel computing paradigms was demonstrated by implementing a fully memristive architecture to tackle the Iris dataset classification, mapping the covariance matrix in the RRAM crosspoint array and extracting the Principal Components by in-memory computing-based (IMC) power iteration and deflation technique. The results showed a clustering accuracy comparable with a 64-bit floating point (FP64) processor, with 98% overlap of the projected datasets, thus supporting IMC for high-efficiency, low-power hardware accelerators for machine learning applications. The use of a silver electrode, instead of the titanium one, gives to the devices a volatile behavior, meaning the programmed state is not stable and spontaneously comes back to its ground state. By studying the temporal dynamic using pulses, the devices showed a great tunability of the retention time by changing the pulse duration, ranging from 10 µs to tens of seconds. Moreover, by changing the pulse amplitude it is possible to change the switching probability of the device. The tunability of both the retention time and the switching probability was exploited to build a simple neuromorphic system able, with 5 only devices, to emulate the short-term memory mechanism that takes place in the human brain.

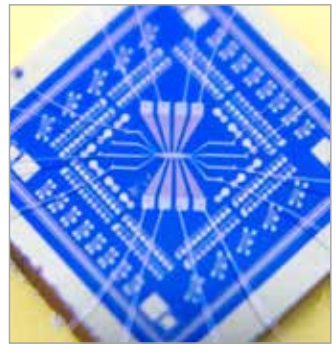

Fig. 1 - Bonded 8x8 RRAM-based crosspoint array

PRICING AND ADVERTISING STRATEGIES IN E-COMMERCE SCENARIOS

Giulia Romano - Supervisor: Prof. Nicola Gatti

In the last decade, *artificial intelligence* has been one of the main drivers of growth for digital markets. The use of AI tools in digital advertising has become increasingly common, opening up new opportunities that were previously unavailable. Some of the advantages over traditional advertising channels are the possibility of profiling a user from behavioral data, targeting ads in a precise way, running auction mechanisms to maximize specific objective functions associated with the revenue, and evaluating investment performance in real time. Although it is impossible to optimize these processes manually because of the vast amount of data provided by platforms and the numerous parameters that need to be set, algorithms and AI tools can efficiently perform such optimization.
In this thesis, we study new scenarios originating from recent innovations introduced by Web advertising and e-commerce platforms. Our particular focus is on the development of new economic mechanisms and learning algorithms that leverage techniques from the fields of algorithmic game theory, mechanism design, and online learning which can be applied to sell and advertise products on the Web. We study scenarios in which strategic agents, such as sellers, advertisers, and buyers, interact on Web platforms, and we analyze optimization problems faced by each party involved in the interaction. For instance, online marketplaces matching sellers/advertisers to buyers need to design mechanisms that incentivise agents to participate, while providing guarantees on their revenue. Taking the perspective of the online platform, we employ techniques and performance criteria from the mechanism design literature in order to design novel auction mechanisms and characterize their performance. Moreover, we study how to address problems faced by agents interacting on the platforms, such as sellers and advertisers. In particular, when an agent has to sell and/or advertise their products on the Web, they have to repeatedly interact with the mechanism operated by the platform. The structure of such interaction is distributed over time: agents are required to perform sequential actions, after which they observe a reward produced by the environment that also depends on their decisions. In this setting, online learning techniques are well suited to design *no-regret algorithms* which allow agents to learn effective strategies while addressing the exploration/exploitation dilemma. Inspired by novel real-world scenarios, we study non-standard learning processes in which, for instance, the feedback returned by the environment is affected by delays, or agents' actions are subject to time-varying constraints. These scenarios are common in practice when, for instance, agents are financially constrained by their budget or want to reach a target profitability in the form of a return-on-investments (ROI) constraint. This thesis expands upon classical models for online markets, incorporating novel e-commerce frameworks that have emerged as a result of the continuous expansion of web platforms. In doing so, it bridges the gap between theory and the latest real-world applications.
In the first part, we take the perspective of a seller who aims at selling their products or services on the Web. Most of the online economic transactions are carried out by posted-price mechanisms, in which sellers need to propose a *take-it-or-leave-it* price to each potential buyer. In this part of the thesis, we study the problem of setting prices over time in scenarios where a single item or multiple units of the same item have to be sold. We first analyze the scenario in which a single unit of a single item has to be sold within a finite period of time, when the value of the item is discounted over time according to an arbitrary continuous and non-increasing discount function. Our main result is a new posted-price mechanism, for which we provide guarantees in the form of bounds on the competitive ratio, that quantifies the worst-case difference in revenue between our mechanism and an optimal one that uses additional information about the user, typically unknown to the seller.
Then, we analyze the scenario in which multiple units of the same item have to be sold. The solution we propose is a new no-regret algorithm that can effectively address the problem at hand, and can also be applied to recommendation problems. Specifically, our algorithm is well-suited to situations where rewards received from the environment are distributed over a time horizon, thereby bridging the gap between non-delayed and delayed scenarios in the existing literature.
The second part of the thesis is centered around the problem of devising mechanisms for novel advertising scenarios. Our initial focus is on investigating a new type of ad auction that displays ads for similar products together with their respective prices. This can significantly influence user behavior and presents an opportunity for jointly optimizing ad allocation and pricing. To address this challenge, we propose several auction mechanisms differing in the payment rule and the level of information requested to the participants. Subsequently, we provide a study of their efficiency.
Another new problem that we study is advertising in the metaverse. Specifically, we initiate the study of a user model and algorithms to allocate ads optimally in the metaverse. Our model extends those currently adopted for search and mobile advertising. In particular, we assume that, during their experience, users will traverse several scenes during which they could be targeted with multiple ads, whose performance may depend on the specific scene in which they are displayed. Furthermore, the ads may be subject to externalities due to their sequential display. In this setting, we study the problem of computing an optimal allocation of ads. In particular, we assess the computational complexity of finding an optimal ad allocation for several model flavors and provide approximation algorithms with tight theoretical guarantees. Finally, we study advertising from the perspective of media agencies, whose recent proliferation has been driven by the increasing complexity of digital advertising. We extensively explore the effects of coordinating the bidding strategies of a group of advertisers who are participating in the same ad auction. Such coordination can lead to significant changes in the strategic interactions underlying the auction and may result in various forms of collusion, potentially increasing the revenue for the advertisers involved. We exploit the specific structure and features of the framework to provide approximate solutions for maximizing the revenue of the agency and the social welfare of the coordinated advertisers.
The third part of this thesis studies the problem faced by a constrained agent that has to learn effective bidding strategies. For example, advertisers need to optimize their revenue while adhering to limitations such as budget constraints or a minimum profitability threshold expressed as a ROI constraint. Our main contributions are new no-regret algorithms that can tackle general problems in which a decision maker has to take sequential actions subject to *uncertain* and *long-term time-varying* constraints. In particular, we propose a *best-of-both-worlds algorithm*, with no-regret guarantees both in the case in which rewards and constraints are selected according to an unknown stochastic model, and in the case in which they are selected at each round by an adversary. Our framework can be instantiated to handle *full-feedback* as well as *bandit-feedback* settings. Finally, we show how it can be applied to constrained bidding in repeated first-price and second-price auctions, since they are de facto standard in large Internet advertising platforms.

A SCALABLE, RECONFIGURABLE, AND ADAPTIVE FRAMEWORK FOR CHATBOTS IN EDUCATION

Donya Rooein - Supervisors: Prof. Barbara Pernici, Prof. Paolo Paolini

This Ph.D. thesis is an interdisciplinary research that examines the challenges of managing learning content using chatbot technology, which is typically developed by technology providers and delivered to educational institutions. This study has been completed by Donya Rooein under the guidance and supervision of Prof. Barbara Pernici and Prof. Paolo Paolini. The research focuses on designing and developing educational chatbot technology, emphasizing empowering non-technical actors. To address the challenges of managing learning content within chatbot technology, a configurable-driven approach with fully modular chatbot architecture techniques has been adopted. The study includes a set of case studies on different educational contents, with continuous validation of the chatbot's development through large-scale experiments with teachers and students at various educational levels.

The lack of provided controls in the design and development process represents limitations for using this technology on a large scale by focusing on empowering non-technical actors. This thesis emphasizes a configurable-driven approach including fully modular chatbot architecture techniques to face a set of identified requirements: 1) the management of learning content within the chatbot technology, 2) the strong separation between content and conversation design, 3) extensive customization for the content and conversation delivery, and 4) the empowerment of non-technical actors in education to be direct actors in the chatbot production process and maintenance.

These requirements are met through case studies on different educational contents with continuous validation of the chatbot's development through large-scale experiments with teachers and students at various educational levels. A new design of chatbots to separate content and conversation is investigated in the context of educational chatbots for tutoring tasks. To this end, I develop a scalable, configurable, and adaptive framework for building chatbots and analyzing content and conversation for the adaptive learning process. I investigate the problem of creating and maintaining chatbots in the properties of non-technical actors to empower their roles.

I propose a new modular architecture design with a configurable-driven methodology to bring more controls to the high-level actors and reduce the effort of IT support in the education domain. This thesis presents the design and implementation details of a framework to support the development of educational chatbots through the continuous validation of different use cases, with a discussion of the architecture and the development of three prototype contents. Finally, compared to traditional learning, chatbot-mediated learning is explored with experiments by teachers and students in different subjects from different education levels, from schools to higher education.

ON DATA-DRIVEN OPTIMIZATION IN THE DESIGN AND CONTROL OF AUTONOMOUS SYSTEMS

Lorenzo Sabug Jr. - Supervisor: Prof. Lorenzo Fagiano
Co-Supervisor: Prof. Fredy Ruiz

In numerous science and engineering contexts, there are optimization problems whose objective and/or constraint functions are not given explicitly, i.e., they are black-box. In other words, the only way they can be evaluated is through expensive simulations and/or experiments. Such problem settings arise in various applications where the systems under test contain different physical mechanisms (electro-mechanical, hydrodynamic, pneumatic, biological) interacting with each other, as well as with the environment. This dissertation addresses this concern by using the Set Membership (SM) framework for the first time to design a black-box optimization method. The resulting algorithm, called the Set Membership Global Optimization (SMGO), builds the SM-based models of the objective (and if required, also constraints) using a Lipschitz continuity assumption, with Fig. 1 showing increasing fidelity of the bounds with respect to the black-box function as the samples increase. These models are then used to strategically choose a sampling point from a set of candidate points. In choosing a candidate point for sampling, SMGO automatically trades off between exploitation of the current best feasible data point, and exploration around the search space to acquire more information about the shape of the pertinent hidden functions.

We have investigated the theoretical properties of SMGO, and we lay out sufficient conditions to guarantee convergence to a global optimum. Practically-motivated concerns regarding the treatment of noise and computational complexity are considered, and methods to address such issues are discussed, by means of iterative computation of SM-related bounds, as well as tweaks on the memory management. Furthermore, an extension to time-varying functions is discussed, and a novel approach for the forgetting-remembering trade-off is proposed by directly using the SM-based bounds.

The effectiveness of the proposed SMGO is verified and validated in the context of a synthetic benchmark test, statistically comparing it against other commonly-used optimization methods. Time-invariant black-box optimization problems are considered, as well as time-varying ones. The results of the tests demonstrated the competitiveness of the proposed SMGO with regards to the iteration-based performance. In addition, SMGO demonstrated less computational times than the state of the art.

Lastly, we have validated the use of SMGO in several case studies, in various levels of design difficulty. In an industrial test case, we used SMGO to experimentally tune a robotic gearshift mechanism, and we demonstrate its merits on the performance and convergence compared to other commonly-used tuning techniques. Furthermore, we tackle the difficult engineering problem of designing a combined passive-active attitude control for a small spacecraft (as in Fig. 2), considering non-trivial mission objectives and high-level constraints. A high-fidelity simulation environment is used to evaluate the individual settings combination for the passive and active attitude control, taking into account the non-trivial interactions between the different actuator types, and the space environment. Simulation results and statistical tests show that a simultaneous design using SMGO resulted in the best spacecraft design, in comparison with other commonly-used design approaches in the literature and engineering practice.

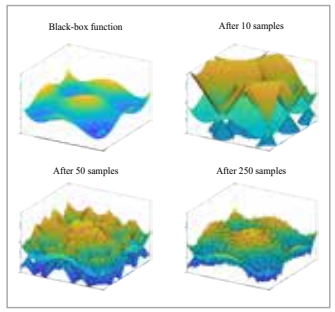

Fig. 1 - Evolution of Set Membership-based bounds with increasing samples

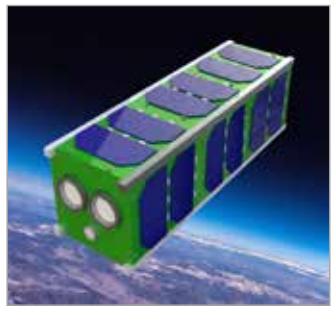

Fig. 2 - An artist's rendition of a 3U CubeSat for design of a passive- and active attitude control

MODEL PREDICTIVE CONTROL FOR CONSTRAINED NAVIGATION OF AUTONOMOUS VEHICLES

Danilo Saccani - Supervisor: Prof. Lorenzo Fagiano

Autonomous vehicles are a rapidly growing field, with numerous applications in transportation, logistics, and other industries. However, the safety of autonomous vehicles is paramount, and there are significant challenges involved in developing reliable navigation algorithms that can ensure safe operation in a variety of environments. This thesis addresses these challenges by proposing a theoretical framework for the constrained navigation of autonomous vehicles and demonstrating its potential in a variety of practical scenarios.
The proposed framework is designed to consider three important objectives: safety, exploitation, and exploration. Safety is considered in the form of constraint satisfaction and persistent obstacle avoidance, which are essential for ensuring that the vehicle does not collide with other objects in its environment. Exploitation refers to the vehicle's ability to make the most of its current knowledge of the environment and reduce the conservatism of a guaranteed collision-free approach. Finally, exploration refers to the vehicle's ability to discover the surrounding potential unknown environment while avoiding getting stuck in blocked areas.

To achieve these objectives, we exploited Model Predictive Control (MPC) schemes. MPC is an advanced control technique that utilizes a model of the system being controlled to predict the future behaviour of the system over a finite horizon. The control inputs are then optimized over the horizon subject to constraints, with the goal of minimizing a cost function that captures the system's performance objectives. The control inputs for the current time step are applied to the system, and the process is repeated at the next time step. MPC is particularly well-suited for constrained navigation of autonomous vehicles since they need to navigate in a way that is safe while accomplishing their autonomous task by avoiding obstacles. MPC can help achieve these goals by optimizing the vehicle's trajectory over a finite horizon while incorporating constraints on the vehicle's motion and the surrounding environment.
The main challenge of autonomous navigation problems is that the system evolves in a partially or totally unknown environment, which is detected by means of onboard sensors, such as LiDAR sensors, cameras or antennas. This information can be interpreted as a safe-set around the vehicle state, leading to time-varying state constraints. These kinds of constraints are rarely considered in the literature, but they represent an important tool to describe the perception of a possible unknown environment. To this aims, the first part of the thesis aims at presenting strategies to guarantees persistent constraints satisfaction despite the time-varying nature of the state constraints.

However, the safe set generated at each time step by exploiting sensor readings generally changes over time and can evolve more favourably to reach a given target. Thus, relying only on local sensor measurements can lead to too-conservative behaviour in terms of performance. To tackle this problem, we developed a novel MPC formulation, named multi-trajectory MPC (mt-MPC), which is particularly suitable for tracking problems with time-varying constraints and its peculiarity is to consider different future trajectories in the same Finite Horizon Control Problem (FHOCP). This approach allows one to partially decouple constraints satisfaction (safety) from cost function minimization (exploitation). This result is achieved considering two predicted state trajectories: one trajectory operates inside the time-varying feasible set guaranteeing the constraints satisfaction, the other one, that shares the first predicted state with the previous one, steers the system to the desired reference without considering state constraints.

Another challenge of autonomous navigation is exploring unknown environments. To address this challenge, the thesis proposes a novel exploration approach called Graph-Based Exploration And Mapping (G-BEAM). G-BEAM is a higher-level logic that provides the system with a temporary reference point that it can use to navigate towards a target or unexplored location.
At the heart of G-BEAM is a reachability graph, which serves as a map of the environment. Unlike traditional occupancy maps, which can be very large and require a lot of storage resources, the representation of the environment in G-BEAM's reachability graph requires less storage and can be directly exploited to compute the system's path towards a given target or unexplored location.
One of the key elements of G-BEAM is that the nodes of the graph are ranked according to the expected information gain that is realized when they are visited. This information gain is then used in the cost function of the navigation strategy, which is based on a receding horizon concept. Similarly, to a traditional MPC approach, G-BEAM computes the optimal future path over the graph in order to maximize the reward function describing the amount of information earned by visiting that node and then, the first node on the path is provided to the system in a receding horizon fashion. This means that the vehicle can explore its environment while avoiding getting stuck in blocked areas, which is essential for ensuring that the vehicle can continue to operate even in complex unknown environments.

The theoretical framework is then applied to a variety of practical scenarios, demonstrating its potential in real-world applications. The first example involved the use of the mt-MPC approach to successfully navigate a drone in an a-priori unknown environment. Specifically, we focused on navigating a commercially available drone equipped with a low-level flight controller and a planar LiDAR sensor, which was utilized to detect the surrounding environment, to its final destination in an unknown environment. The use of a low-level commercial flight controller allowed the system to be approximated with a linear model from the perspective of the mt-MPC allowing to execute the FHOCP on hardware with reduced computational power. By exploiting a dataset of autonomous flight, different sources of uncertainties are also taken into account in a set membership framework. The proposed formulation enabled the achievement of persistent constraint satisfaction throughout the navigation process in the unknown environment. The approach is experimentally validated out-of-the-lab on a prototype, together with the G-BEAM mapping strategy, demonstrating the feasibility of the proposed approach.
The framework is then extended to the case of multi-agent systems, which are becoming increasingly common in many industries. In one example, the navigation of a system of tethered multi-copters in a partially unknown environment is explored. The systems consist of multiple drones that are tethered to each other and to a ground station attachment point.
The framework is shown to be able to manage physical couplings between agents, which are described by the intersection of time-varying state constraints. Finally, the framework is applied to swarms of agents, where the multi-trajectory MPC approach is employed to enable automatic plug-and-play requests that cannot be rejected in a time-varying network topology of agents with limited communication capabilities.

Overall, this thesis provides a theoretical framework for the constrained navigation of autonomous vehicles that can consider the competing objectives of safety, exploitation, and exploration. The proposed framework has been demonstrated to be effective in a variety of practical scenarios, and it has the potential to make significant contributions to the field of autonomous navigation.

SCALABLE INTEGRATED ELECTRONICS FOR CLOSED-LOOP CONTROL OF LARGE RECONFIGURABLE PHOTONIC CIRCUITS

Fabio Toso - Supervisor: Prof. Giorgio Ferrari

Technological advances in integrated photonics have made the realization of complex photonic architectures possible. Several applications benefit from this progress, from the implementation of complex optical networks functionalities, to artificial intelligence, all-optical information processing and innovative light-based sensing. A key factor in enabling the successful operation of complex and dense photonic circuits is the implementation of a real-time electronic control layer, able to configure and stabilize the functionalities of optical devices against their intrinsic sensitivity to temperature fluctuations and fabrication tolerances. In this thesis, two complementary approaches were studied to address the lack of an integrated, scalable and low-power solution for the control and stabilization of large programmable photonic circuits: the miniaturization of the control electronics and the integration of electronic devices directly inside the photonic circuits, as a mean of reducing the required electrical input/output connections.

A multichannel ASIC controller was developed, capable of autonomously stabilizing and controlling up to four integrated Mach-Zehnder interferometers in parallel. Everything necessary to perform the control action was integrated inside a single chip, from sensor read-out to signal processing and driving of the actuators. In each channel, two integral controllers work in parallel to maximize or minimize the output optical power of a single interferometer. A dedicated low-noise lock-in input stage was designed to allow the ASIC to work with CLIPP detectors, which enable to monitor the state of the target photonic devices in a non-invasive way, with sensitivities down to -50dBm. A double dithering technique was used to discriminate the action of the two actuators included in a single interferometer through a single output detector.
For a simple and low power implementation in an integrated circuit, the technique is based on two square-wave dithering signals at the same frequency and in quadrature of phase. The implementation of two parallel switched-capacitor processing chains enabled to correctly extract the two signals and to perform the integral control action on the working point of the Mach-Zehnder with low noise and low offset. Finally, an automatic saturation detector and reset system was integrated to each channel, in order to avoid a controller stuck on a saturated condition and to enable a truly autonomous operation of the system.
The ASIC controller was successfully validated with a 4x1 binary-tree mesh of Mach-Zehnder interferometers.
The system was able to simultaneously control and stabilize the three interferometers of the photonic circuit, while compensating for input phase perturbations up to 50Hz, with residual output oscillations lower than 1dB. A re-configuration time of 10ms was measured in case of abrupt input variations. The proposed ASIC is the first fully integrated solution for an arbitrary and automatic control of two-input MZIs. Thanks to its limited are occupation ($2mm^2$) and power consumption (20mW) per controlled device, the ASIC has a clear advantage with respect to other discrete-component solutions found in the literature. The reduction of the area and power requirements of the electronics to the same order of magnitude of the controlled photonic devices is a key achievement in enabling the scaled control of large photonic circuits with hundreds of devices, as needed by the most advanced applications.

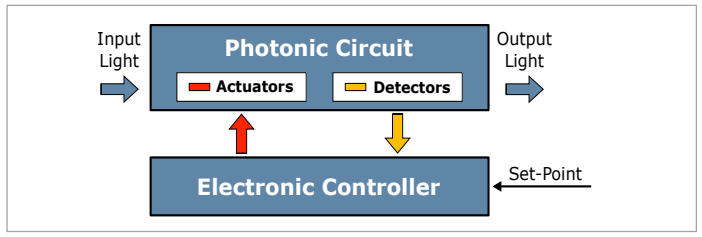

Fig. 1 - An electronic controller monitors the state of the circuit through light detectors and keeps the working point stable with on-chip actuators.

Time-multiplexing of input and output signals was explored as a possible solution to scaling the control electronics for very large photonic systems. In this thesis, it was demonstrated that a zero-change approach to the integration of electron devices in a standard photonic platform is possible. Both n-type and p-type MOSFET transistors were designed in a commercially-available active SOI silicon photonics platform. Side-wall MOSFETs, that leverage the cladding of waveguides as gate oxide (200nm) and the SOI thickness (220nm) as channel, were designed to overcome the lack of fabrication steps dedicated to the realization of standard electron devices. Although the performances of the fabricated components are limited if compared to the state-of-art CMOS technology, the possibility to integrate simple logic gates and analog switches inside a standard photonic circuit allowed to design an on-chip electronic multiplexer which enabled to time-multiplex the read-out of 16 sensors through just a single output connection. The device was used to control a 16-to-1 optical router in real-time, with full-mesh configuration times lower than 10ms and guaranteeing an average output extinction ratio of 9dB when routing a 10Gb/s on-off keying (OOK) modulated signal.

The possibility to time-multiplex the read-out of sensors in standard photonic technologies can greatly benefit applications that rely on electronics for control and stabilization and do not require the integration of high-performance electronic circuits. In fact, a significant improvement in the management of the input/output connections can be achieved, while preserving optimal photonic functionalities and without incurring in the higher cost of experimental technologies combining photonic components with a state-of-art CMOS process.

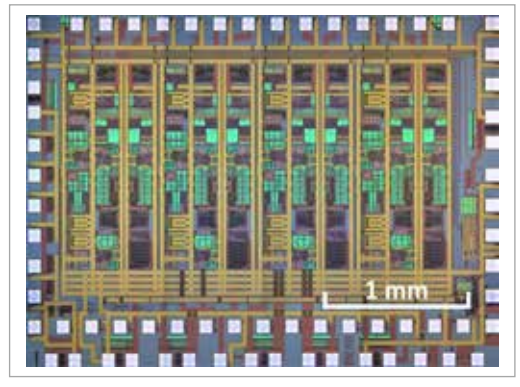

Fig. 2 - Photomicrograph of the fabricated ASIC.

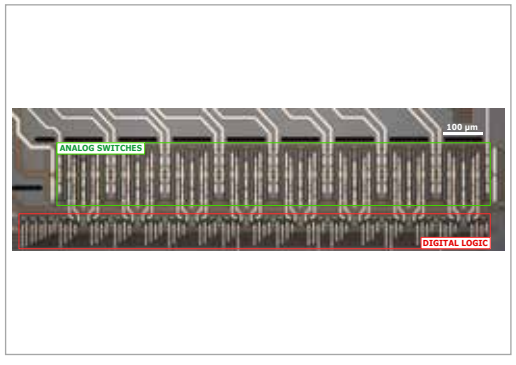

Fig. 3 - Fabricated electronic multiplexer in a standard Silicon Photonics technology.

HYPERSPECTRAL IMAGE ANALYSIS AND ADVANCED FEATURE ENGINEERING FOR OPTIMIZED CLASSIFICATION AND DATA ACQUISITION

Ava Vali – Supervisor: Prof. Sara Comai
Co-Supervisor: Prof. Matteo Matteucci

Hyperspectral imaging is a technique that combines spectroscopy with imaging capacities to capture cubic images in a non-invasive manner, by considering a wide range of wavelengths within the electromagnetic spectrum, that contain valuable diagnostic information for detecting objects and precisely distinguishing their constituent materials. Although such technology has been known for almost five decades, due to the hardware and computational limitations, its usage has been mostly exclusive to a few applications of its domain of origin, Remote Sensing. However, within the last decade, new advances have raised new opportunities, which recently led to a significant rise in the popularity of this technology among the research community. The increased computational capacity, enhancement of the acquisition quality, and the significant decrease in the size of hyperspectral imaging sensors realize new applications for a vast range of new domains, i.e., healthcare, pharmaceuticals, manufacturing, and food safety and quality control. Recent studies determine the great potential of machine learning in making these applications happen. Despite proven promising prospects, passing the technology to actual use cases is still complicated by some open critical challenges and requires further studies and adaptations. Due to the hyperspectral computational complexity, many studies propose deep learning as a powerful technique that allows skipping the complex feature introspection. Over the last decade, we have witnessed the domination of deep learning in image computations, typically in the supervised mode and in an end-to-end fashion. However, despite the similarities that hyperspectral images share with classic RGB images, their distinctive characteristics make such end-to-end deep learning techniques impractical for their real applications.

Hyperspectral images are high-dimensional by nature, which essentially leads to the curse-of-dimensionality phenomenon and causes several inconsistencies and computational inefficiencies for end-to-end deep learning approaches. High dimensionality also intensifies the problem of ground-truth scarcity, which is a critical problem when it comes to supervised deep learning solutions. In addition, the common challenges with any machine learning approach and the concerns regarding the forthcoming issue of reaching the physical limitations of Moore's Law are even more significant in the case of hyperspectral image analysis. In this thesis, we focus on these challenges and propose a scalable and holistic solution based on the classic machine learning pipeline structure that adopts advanced techniques and strategies to tackle these challenges.

Contrary to popular end-to-end approaches that ignore the potential of feature engineering, we start by highlighting its impact within the machine learning pipeline structure on the performance and efficiency of supervised hyperspectral-based classification tasks. We then present a strategy to better exploit these potentials of feature engineering by breaking it down into a sequence of distinct steps: i) feature transformation, ii) feature selection and iii) feature extraction. Each of these steps performs a set of tasks that enhance the quality, remove redundancies, or re-define the features to improve the classification and prediction performance. Accordingly, we revive the classic 4-stage machine learning pipeline structure -which consists of these feature engineering steps- and propose a dynamic and scalable strategy to adapt this classical structure to the hyperspectral-specific computation needs and empower the idea with advanced methods for automated optimization.

To accomplish the proposed strategy, we design and develop a prototype Automatic Machine Learning (AutoML) framework that generates and holistically optimizes several models by involving (or skipping) any possible combination of the feature engineering and other pipeline steps and eventually outputs the best-performing constituent model (schematically shown in Fig. 1). This framework also allows us to observe and determine the impact of feature engineering steps on the efficiency and performance of a given hyperspectral predictive task. More specifically, the framework performs an optimized model selection by providing transparent mid-process reports that let us conduct a comparative evaluation among different pipeline configurations and establish proper argumentative reasoning for this thesis's claim. Moreover, we deploy the framework considering several potential techniques as alternatives for the pipeline steps. In this case, the framework also indicates which techniques are most suitable for the given predictive task and its input data.

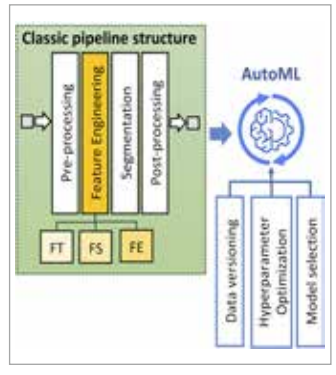

Fig. 1 - Schematic diagram demonstrating the proposed idea of adopting the classic pipeline structure to implement out prototype AutoML framework.

We define a holistic scheme for the framework to meet different optimization requirements, including data versioning, model selection, and hyperparameter tuning. These optimization procedures ensure the generalizability, reliability, robustness, and repeatability of the yielded model. We also present efficient implementations of the designed optimization scheme, which allows us to split the execution effort for optimized resource management and mitigate the risk of the system's failure. Moreover, we empirically investigate the possibility of automating feature engineering as a stand-alone procedure, independent of the forthcoming steps of the pipeline. More specifically, we investigate whether the feature engineering full-optimization can be conducted solely based on the input data, so its trained model can be used to enhance data for any potential following image segmentation technique. We discuss how stand-alone feature engineering is beneficial to optimize the whole process, from the acquisition of data to the optimized data analysis.

At last, we perform two experiments to evaluate the implemented framework. The experiments are chosen from two distinct families of hyperspectral-based applications to support this thesis argument. The first experiment is a well-known problem of remote sensing with the most-cited hyperspectral dataset in the literature that allows the reader to compare the results with the state-of-the-art. The second experiment is part of an exploratory study -a joint project of Politecnico di Milano University with a couple of industries- which investigates the potential of hyperspectral imaging technology in detecting contaminants as residuals of the production line on the chassis of the washing machine prior to the painting stage. Finally, we discuss the outcomes, supported by the mid-process reports, to establish the framework's effectiveness in both the hyperspectral problem scenarios and raise a debate about when, how, and where relying solely on the proposed set of stand-alone feature engineering steps can be most beneficial.

RESOURCE MANAGEMENT FOR MILLIMETER-WAVE ACCESS NETWORKS BASED ON ARTIFICIAL INTELLIGENCE

Bibo Zhang - Supervisor: Prof. Ilario Filippini

Millimeter wave (mmWave) communications have been envisioned in the fifth-generation (5G) standardization process as a promising direction, due to their attractive potential to provide a huge capacity extension to traditional sub-6 GHz technologies, thus meeting the demand of huge wireless access data rates. However, such high-frequency communications are susceptible to harsh propagation conditions such as high path losses and blockages that can be only partially alleviated by directional phased-array antennas. This makes mmWave networks coverage-limited, thus requiring the dense deployment of a number of base stations. Integrated Access and Backhaul (IAB) network, which is proposed by the 3rd Generation Partnership Project (3GPP) and generally organized in a multi-hop architecture, is gaining momentum as a cost-effective solution to the end of network densification. Self-backhauling is a peculiar aspect of this architecture, where both radio access and backhaul links share the same radio resources and interfaces. Therefore, a proper radio resource allocation is fundamental to efficiently operate this network. In particular, since the adopted medium access control (MAC) scheme is based on time-division multiple access (TDMA), routing paths and scheduling of directional transmissions along established links must be optimized.

Flow routing and directional link scheduling in mmWave IAB networks have become hot topics recently, which inherit from the classical problem of resource optimization in traditional wireless multi-hop networks. The studies relevant to both have appeared in a long-standing literature, mainly resorting to optimization techniques that assume always-available links and static users. Routing and scheduling in wireless multi-hop networks have been traditionally considered as hard problems due to interference constraints. The problems can become harder if uncertainties and dynamics (e.g., random obstacles, mobile users) are introduced into the networks. These dynamics can make the optimization techniques inappropriate to provide mmWave IAB networks with practical solutions that simultaneously deal with the harsh propagation environment, the strong impact of obstacles on link availability, and the users' mobility. Indeed, the optimal performance provided under ideal link conditions and static network layouts can be hindered by 1) the stochastic on-off behavior due to sudden obstacles and 2) the varying access topology caused by mobile users, which can destroy the advantages of a careful optimization. We could in principle re-optimize the network periodically or every time it undergoes a change, however, this will lead to huge computational costs and, most likely, it would not be practical. Therefore, flexible and adaptive solutions are required to schedule real-time operations so as to tackle the dynamics mentioned above.

Given the above context, we believe that Reinforcement Learning (RL) techniques can play an important role due to their intrinsic ability to adapt to the environment conditions. Indeed, an RL agent can be trained in the mmWave access networks and eventually discover a proper resource management strategy, even when the networks are dynamic and their replies are stochastic. RL agents can automatically capture relevant network statistics during the training and apply the obtained resource allocation strategy that provides the best long-term performance in front of any random instance of the dynamic network. Indeed, RL techniques have been successfully applied to play prototype games. However, to apply RL techniques, especially DRL techniques, to control complex systems such as wireless multi-hop networks, different aspects of the system need to be considered and the models are expected to be fine tuned. Although such RL applications on wireless networks have been largely investigated in recent years, mmWave IAB network, as an emerging network architecture solution proposed by 3GPP recently, leaves many special working points (e.g., heterogeneous devices, in-band backhaul, etc.) to be discussed and addressed. Therefore, directly applying the existing well-designed RL models to mmWave IAB networks studied in this thesis is not feasible, which requires a newly customized RL framework. Firstly, the state / observation space, action space and reward function need to be carefully designed according to the faced problem, while the neural networks need to be properly tuned. Secondly, how to facilitate RL agents' learning process also needs to be addressed. Thirdly, we need to consider the complexity of obtaining system data, which is fed into the learning process.

In this thesis, we aim to deal with the problem of flow routing and transmission scheduling to maximize the user throughput in mmWave IAB networks, taking into account both the stochastic link behaviors caused by obstacles and varying access topology due to mobile users. In particular, we investigate different scenarios, starting from the simplified and ideal cases and stepping forward to more realistic ones.

Firstly, we optimize the flow routing and link scheduling in static mmWave IAB networks where no dynamic factors are introduced. We mainly focus on coordinating interference among backhaul and access transmissions potentially activated in parallel so as to maximize the link capacities. Sets of compatible links (called link patterns) are generated, which can always be simultaneously activated, satisfying hardware and physical constraints including antenna patterns, half-duplex / full-duplex operation modes, radio frequency chain, power, etc. resorting to these generated patterns, the problem can be reduced from link scheduling to pattern scheduling, based on which, a quasi-optimal resource allocation scheme can be achieved.

Subsequently, we further study a simplified dynamic 5G Enhanced Mobile Broad-band (eMBB) scenario where users are assumed to be static and a few of the scheduled backhaul and access links are exposed to random blockages. In this first dynamic scenario, the link's availability is characterized by a slotted Bernoulli process that uses probability to indicate the obstruction intensity of a link. In the second scenario, we introduce a more refined temporally-correlated blockage model into the mmWave IAB scenario where we still assume users to be static. Specifically, we adopt a binary-state signal fading model to characterize the alternately blocked and available duration of each access link. Furthermore, in the third scenario, differently from the above scenarios, we consider mobile users, which are served by a star-topology backhaul network where all IAB-nodes are directly connected to the IAB-donor and operate in half-duplex mode. This simplified backhaul deployment, in addition to be one of the most common layouts envisioned for IAB mmWave netowrks, helps to focus on coping with uncertainties due to users' mobility rather than other more complicated aspects introduced by a multi-hop architecture. In the final scenario, the simplified star backhaul topology in the third scenario is extended to a tree topology where IAB-nodes are connected to the IAB-donor directly or via multi-hops. In addition, we consider 3D mobile obstacles, modeled as cylinders. Every time a 3D obstacle stands in the line-of-sight (LOS) path of an access link, a blockage occurs and the UE can't receive any bits from the IAB-node. At last, we explore the impact of IAB-nodes' duplex modes (half-duplex / full-duplex modes) on the performance of the downlink transmission. The adaptive RL-based resource allocation approaches are proposed accordingly for these dynamic scenarios. Numerical experiments have been extensively carried out and the results have shown the effectiveness of these approaches.