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 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 | 2021

DOCTORAL PROGRAM IN INFORMATION TECHNOLOGY

Chair:
Prof. Barbara Pernici

Introduction

The PhD program in Information Technology (ITPhD) covers research topics in four scientific areas:
Computer Science and Engineering, Electronics, Systems and Control, and Telecommunications.
This broad variety of research topics is matched together by the common affinity to the ICT area and perfectly captures the core mission of the corresponding sections of the Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB). New research topics and some 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 PhD program in IT is the largest in Politecnico in terms of number of students. Every year, about 70 new students join the program, for an overall number of students around 220. Students must undergo a yearly evaluation of the progress in their research and course work.

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. Cristina Silvano): Information systems, Database management, Information design for the web, Methods and applications for interactive multimedia, Embedded systems design and design methodologies, Dependable systems, 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. Luigi Piroddi): Control systems, Robotics and industrial automation, Optical measurements and laser instrumentation, Dynamics of complex systems, Planning and management of environmental systems, Operations research and discrete optimization. Telecommunications (Vice-Chair: Matteo Cesana): Networking, Applied electromagnetics, Optical communications, Wireless and space communications, Remote sensing, Signal processing for multimedia and telecommunications.

Industrial collaborations

Due to its intrinsic technological nature, the PhD program features many industrial collaborations. More than 50% of the PhD candidates are funded by companies or by international research projects involving industrial partners. In the PhD School vision, the collaboration between university and companies is the ideal ground where to turn invention and scientific research into technological innovation. This collaboration also contributes to create a common terrain of friendly culture, to size research risk, and to preserve strong basic research. To monitor the activities and development of the PhD program, the PhD board cooperates with an industrial referee board, composed by members of public and private companies, working in management, production, and applied research. The board meets once a year to identify and suggest new emerging research areas and to foster the visibility of the PhD program 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 Polimi PhD School.
Access to external courses and summer schools is also encouraged. The challenge is to promote interdisciplinary research while offering advanced help students to make their best choices according to the regulation scheme of the programme.

Internationalization

Every year, several courses are delivered by visiting professors from prestigious foreign universities.
Moreover, the PhD program encourages joint curricula with foreign institutions. We have several Double Degree and Joint Degree agreements with countries in all continents. Every year we receive more than 150 applications from foreign countries and about 15% of our selected PhD candidates has applied from outside Italy.

Conclusions

The core mission of our PhD Program is to offer an excellent PhD curriculum, through high quality courses, a truly interdisciplinary advanced education, cutting-edge research, international and industrial collaborations.

PHD BOARD OF PROFESSORS

Area 1: Computer Science and Engineering	Area 2: Electronics	Area 3: Systems and Control	Area 4: Telecommunications
Cristina Silvano - Vice Chair Computer Science and Engineering	Angelo Geraci - Vice Chair Electronics	Luigi Piroddi - Vice Chair Systems and Control	Matteo Cesana - Vice Chair Telecommunications
Cesare Alippi	Giuseppe Bertuccio	Luca Bascetta	Paolo Martelli
Francesco Amigoni	Giorgio Ferrari	Andrea Castelletti	Andrea Monti Guarnieri
Florian Daniel	Ivan Rech	Fabio Dercole	Carlo Riva
Nicola Gatti	Alessandro Sottocornola Spinelli	Lorenzo Fagiano	Massimo Tornatore
Raffaela Mirandola		Simone Garatti	
Barbara Pernici			
Matteo Pradella		**Representative from the Bioengineering Area**	
Letizia Tanca		Giancarlo Ferrigno	

PHD ADVISORY BOARD

Giorgio Ancona	Atos
Matteo Bogana	Cleafy
Mario Caironi	IIT
Paolo Cederle	Everis
Luigi Cicchese	Concept Reply
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.
Sabino Illuzzi	Prospera
Renato Lombardi	Huawei Technologies
Renato Marchi	KPMG
David Oliva Uribe	Head Industrial Doctoral School, EIT Digital
Giorgio Parladori	SM Optics srl
Francesco Prelz	INFN
Enrico Ragaini	ABB S.p.A.
Paolo Giuseppe Ravazzani	CNR
Fabrizio Renzi	IBM
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 2020 the following awards have been obtained by PhD candidates:

- Dimitris N. Chorafas Foundation Award - **Alberto Marchesi**, **Alessio La Bella**.
- 2019 IEEE AP/ED/MTT North Italy Chapter Thesis Award: “Best Master Thesis Electron Devices Society” – **Matteo Farronato**
- Photonics West 2020 Young Investigator Award – **Giulia Acconcia**
- 2020 CAEN Best Young Speaker Award at 5th Topical Workshop on Modern Aspects in Nuclear Structure – **Luca Buonanno**
- ISSIP-IBM-CBA-JST-CISCO Student Paper Award – **Micol Spitale**, **Fabio Catania**, **Pietro Crovari**

DYNAMIC MANAGEMENT OF VIRTUAL NETWORK FUNCTIONS IN OPTICAL METRO NETWORK ARCHITECTURES FOR 5GCOMMUNICATIONS

Leila Askari - Supervisor: Massimo Tornatore

New services emerging in metro networks force network operators to seek new approaches for provisioning of the resources required by these services in the network. The traditional approach, that leverages the single-purpose hardware devices, imposes high Capital Expenditure (CapEx) and Operational Expenditure (OpEx) on network operators and will not provide an agile way to provision services. In other words, the hardware devices used as middle-boxes only provide a single functionality, therefore, to provide multiple functionalities in the network, several middleboxes are needed, resulting in high CapEx and OpEx, especially in terms of energy cost. Furthermore, these devices are typically expensive and have limited life cycle.

Network Function Virtualization as a new paradigm shift has emerged to enable network operators reduce the CapEx and the OpEx by replacing hardware devices by software instances that can run on generic purpose servers and switches. These software instance that can run on any generic server are referred to as Virtual Network Functions (VNFs). However, the decision about where to place these VNFs in the network and how to concatenate them in the specific order to provision a Service Chain (SC) is not trivial.

Service chaining of VNFs has attracted lot of attention in the context of metro networks , as metro is the network segment where future low-latency 5G services will reside. Therefore, in this thesis, we mainly focus on the problem of VNF placement in optical metro networks and develop novel algorithms to perform VNF placement for dynamic service chaining. We consider the problem of service chain provisioning under different scenarios:

- *Latency-aware service chaining* for which we propose an algorithm that based on the latency requirements of each SC decides about performing traffic grooming and bandwidth assignment of SCs. Our algorithm performs service chaining in a multi-layer optical network that is based on Optical Transport Network (OTN) over Wavelength Division Multiplexing (WDM) as shown in Fig 1.
- *Reprovisioning* for which we consider the situation in which the placement of VNFs can be changed in the network to achieve higher performance in terms of number of SCs provisioned in the network.
- *Protection* for which we propose algorithms to provide different protection strategies for dynamic service chaining.
- *Filterless* for which we propose an algorithm to perform dynamic service chaining in filterless optical networks.

To conclude this thesis, we further perform techno-economic analysis of different VNF placement strategies to figure out the best trade-off between bandwidth and computational cost in metro networks to guarantee a specific service availability defined in the Service Level Agreement of the users.

DEVELOPMENT & CHARACTERIZATION OF SILICON DRIFT DETECTOR ARRAYS & INTEGRATED READOUT ELECTRONICS FOR X-RAY DETECTION APPLICATIONS

Ata Baniahmad - Supervisor: Carlo Ettore Fiorini

The aim of this doctoral activity has been the modelling, design and characterization of X-ray detection systems based on silicon drift detectors (SDDs) and a low-noise integrated multichannel readout front-end for solid-state detectors signal processing in the field of nuclear electronics.

This PhD research activity is concluded on two main subjects: First part is dedicated to SIDDHARTA experiment. This experiment is an upgrade of the first one held in 2009 in Frascati (Rome) and both are meant to provide a deeper knowledge of the strong-interaction inside atoms, which is not completely defined due to the absence of experimental data. To gain new information, the experiment needs to measure as accurate as possible the shift and the amplitude of the K-alpha emission line of the Kaonic Hydrogen, an exotic atom. The first installation of the experiment provided useful data since it was the first-ever measurement of the Kaonic Deuterium X-rays, but still had some limitations in resolution mostly due to background events, so it was necessary an upgrade which will be running in 2020. Both the detectors and readout electronics have been improved to match the background noise requirements, so they have to be fully characterized before using them in final experiment.

There are two main topics covered in this research activity. First part is dedicated to installation of Silicon Drift detectors (SDD), development of the acquisition chain, and post-processing of the acquired data to analyze the functionality of the detectors and readout electronics. The second part is dedicated to characterization of SDDs and dealing with unexpected behaviors encountered during the characterization procedure. Beside that some developments in the low-noise integrated multichannel readout front-end electronics have been done.

Second part of this research activity is dedicated to HTRS ASIC. HTRS is a front-end ASIC for the readout of multi-elements Silicon Drift Detectors specifically designed for high count-rate X-ray spectroscopy applications. In particular, the focus is on the maximization of the event throughput and, at the same time, keeping a high energy resolution. HTRS is going to be used in eXTP (enhanced X-ray Timing and Polarimetry) mission, a scientific space mission designed to study the state of matter under extreme conditions of density, gravity and magnetism. Primary goals are the determination of the equation of state of matter at supra-nuclear density, the measurement of QED effects in the radiation emerging from highly magnetized stars, and the study of matter dynamics in the strong-field regime of gravity. The eXTP mission will revolutionize these areas of fundamental research by high precision X-ray measurements of NSs (Neutron Star) across the magnetic field scale and BHs (Black Hole) across the mass scale. In addition to investigating fundamental physics, eXTP will be a very powerful observatory for astrophysics, which will provide observations of unprecedented quality on a variety of galactic and extragalactic objects.

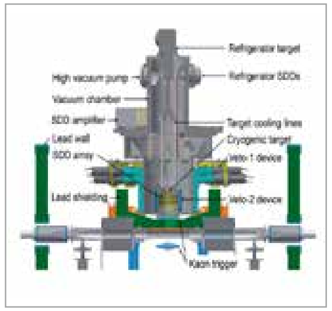

Fig. 1 - The SIDDHARTA-2 setup. The target is surrounded by the Silicon Drift Detectors..

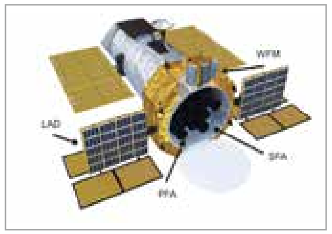

Fig. 2 - eXTP satellite, consist of four main parts including spectroscopic focusing array, the large area detector, the polarimetry focusing array and wide field monitor

Covering and solving the issues occurred during the HTRS ASIC installation and redesigning the ASIC carrier board and test setup have been done during this research activity.

MODEL, INTEGRATE, SEARCH... REPEAT: A SOUND APPROACH TO BUILDING INTEGRATED REPOSITORIES OF GENOMIC DATA

Anna Bernasconi - Supervisor: Stefano Ceri

Co-supervisor: Alessandro Campi

The integration of genomic data and of their describing metadata is, at the same time, an important, difficult, and well-recognized challenge. It is important because a wealth of public data repositories is available to drive biological and clinical research; combining information from various heterogeneous and widely dispersed sources is paramount to a number of biological discoveries. It is difficult because the domain is complex and there is no agreement among the various data formats, data models, and metadata definitions, which refer to different vocabularies and ontologies. It is well-recognized in the bioinformatics community because, in the common practice, repositories are accessed one-by-one, learning their specific metadata definitions as result of long and tedious efforts, and such practice is error-prone; moreover, downloaded datasets need considerable efforts prior to insertion in analysis pipelines.
Within the context of the European project data-driven Genomic Computing (GeCo), which supports genomic research by proposing bioinformatics abstractions and tools, this PhD thesis focuses on the data integration problem, sharing the motivations and methodologies of the project and addressing one of its objectives.
We have thoroughly analyzed the players involved in the genomic data context and proposed a conceptual model of metadata (the Genomic Conceptual Model, see Fig. 1) to represent in a general way the most common information attributes that describe genomic samples and experiments in the available sources. The model describes a typical genomic region data file by different perspectives (biology, technology, management and extraction) and sets the basis to query the underlying data sources for locating relevant experimental datasets. We then describe META-BASE (see Fig. 2), our architecture for integrating datasets, retrieved from a variety of genomic data sources, based upon a structured transformation process; we present a number of innovative techniques for data extraction, cleaning, normalization and enrichment and we show a general, open and extensible pipeline that can easily incorporate any number of new sources. The resulting repository – already integrating several important sources such as the Encyclopedia of DNA Elements, The Cancer Genome Atlas, the Roadmap Epigenomics, and the 1000 Genomes Project – is exposed by means of user interfaces to respond to biological researchers' needs. We provide both a graph-based endpoint for expert users, who need to explore the semantic structure of metadata, and GenoSurf (http://www.gmql.eu/genosurf/), a user-friendly search system providing access to the consolidated repository of metadata attributes, enriched by a multi-ontology knowledge base, locating relevant genomic datasets, which can then be analyzed with off-the-shelf bioinformatics tools.
This interface was evaluated by running an extended empirical study with both Biology and Computer Science knowledgeable participants, collecting many relevant insights on the practices of different user profiles and on their understanding of procedures for extracting relevant datasets for their research.
The models, frameworks and tools that are described in this thesis are already included in follow-up projects; they can be exploited to provide biologists and clinicians with a complete data extraction/analysis environment, equipped by a 'marketplace' of ready-to-use best practices. The process may be guided by a conversational interface, which breaks down the technological barriers that are currently slowing down the practical adoption of our systems. Our commitment is to continue the inclusion of relevant data sources for bioinformatics tertiary analysis, improving our process from a data quality and interoperability point of view.
Inspired by our work on genomic data integration, during the outbreak of the COVID-19 pandemic we searched for effective ways to help mitigate its effects with our contribution; we were able to successfully re-apply the *model-build-search* paradigm used for human genomics. Even if the domain of viral genomics is completely new, it presents many analogies with our previous challenges. Here we model viral nucleotide sequences as strings of letters, with corresponding sub-sequences – the genes – that encode for amino acid proteins. To highlight differences with previously considered data, we have devised the Viral Conceptual Model to account for their technological, biological and organizational aspects, in addition to computed annotations and variants on both nucleotides and amino acid sequences. We then integrate sequences with their metadata from a variety of different sources and propose a powerful search interface (ViruSurf, http://www.gmql.eu/virusurf/) able to quickly extract sequences based on their combined variants, to compare different conditions, and to build interesting populations for downstream analysis.
When applied to SARS-CoV-2, the virus responsible for COVID-19, complex conceptual queries upon our system are able to replicate the search results of recent articles, hence demonstrating considerable potential in supporting virology research.
This work has been realized in the last nine months, during the spread of the SARS-CoV-2 pandemic; after setting the first milestones we are now moving forward, considering the next challenges of this new domain with growing interest. These include the development of a requirements elicitation technique for emergency times, the extension of ViruSurf to other types of data (e.g., relevant for vaccine design), and the provision of visual and statistical support to the integrated data.
The results on this thesis are part of a broad vision: availability of conceptual models, related databases and search systems for both humans and viruses' genomics will provide important opportunities for genomic and clinical research, especially if virus data will be connected to its host, the human being, who is the provider of genomic and phenotype information.

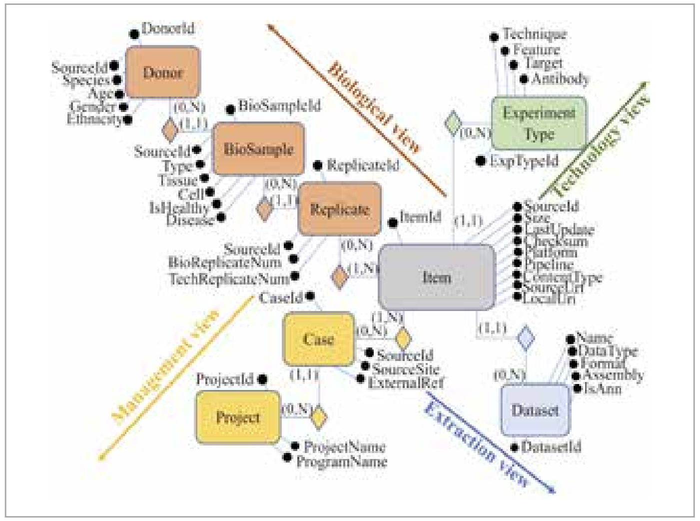

Fig. 1 - The Genomic Conceptual Model for describing a bioinformatics tertiary analysis genomic region data file (Item), using the four perspectives of biology, technology, management, and extraction.

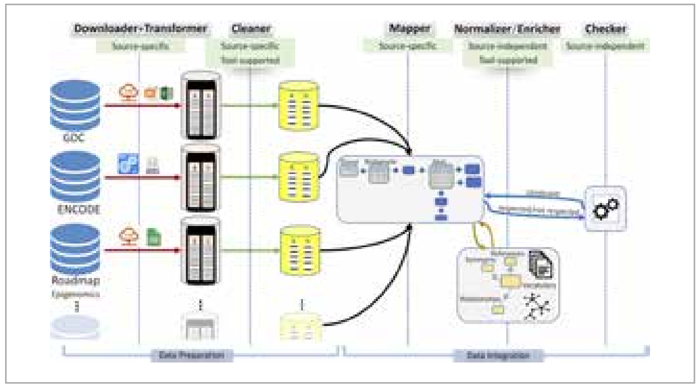

Fig. 2 - The META-BASE data and metadata integration architecture for building a repository of genomic datasets.

TEMPORAL NETWORK MODELS FOR THE SPREAD OF INFECTIOUS DISEASES, WITH A SPECIAL FOCUS ON LIVESTOCK FARM SYSTEMS

Alba Bernini - Supervisor: Renato Casagrandi

Co-supervisor: Luca Bolzoni

Epidemics in livestock systems represent a major concern for many reasons: they may compromise animal welfare, reduce productivity, induce large costs for their control and eradication, and have negative social impacts. The ongoing shift towards diets highly demanding for animal products will make coping against infectious diseases in domestic animals an even more important issue for policy makers and populations.
Epidemiological models represent a powerful tool to both investigate disease spread processes and support better informed decisions aiming at either reducing the risks of outbreaks or, at least, limiting their impacts.
This being the context, the goal of this Thesis was two-fold: on the one hand, it aimed at exploring the complex interplay between pathogen transmission routes in farm systems; on the other hand, it investigated the trade-offs emerging when models are used as Decision Support System for the definition of risk-based biosecurity plans.
Farm systems are characterized by complex contact structures featuring several potential pathogen transmission routes. Animal exchanges are considered the major pathway for between-farm transmission of many livestock infectious diseases. Yet, vehicles and operators visiting several farms during routine activities can also contribute to disease spread. Indeed, if contaminated, they can act as mechanical vectors of fomites, generating indirect contacts between visited farms. While data on animal exchanges is available in national databases in most EU-countries, information about the daily itineraries of trucks and operators is rare because difficult to obtain. Thus, some unavoidable approximations have been frequently introduced in the description of indirect contacts in epidemic models. In this Thesis, the importance of accurate information on on-farm visits by visitors and trucks potentially carrying fomites has been highlighted in different contexts. The analyses focused on the potential spread of a disease in a dairy cattle system in the Emilia Romagna Region, where a comprehensive data collection campaign was conducted on calf transportations occurred in a 3-month period. Two network models were developed to study between-farm contacts generated by either animal exchanges (direct contacts) and connections operated by trucks moving calves (indirect contacts). Nodes in the networks represented farms, considered as epidemiological units, while links described the potentially infectious contacts. The developed network models were either temporal, and thus could be interpreted as a sequence of snapshots (i.e., static networks) representing contacts occurred in a single day, and multiplex, because within each snapshot links were grouped into two layers (one for direct and the other for indirect links) (Fig.1). Difference between the two models were only in the description of indirect contacts, while other characteristics were in common. In particular, the complete model used the full knowledge of the daily trucks' itineraries, while the partially informed one used only a subset of such information.
The disease spread was described through stochastic epidemiological models. Results revealed the importance of an accurate description of indirect contacts for a deep understanding of epidemic spread processes and for the identification of farms playing a crucial role in shaping disease spread patterns. Sensitivity analyses corroborated these findings also in the case of resistant pathogens and of different disease dynamics (Fig.2). Taken together, these results suggest that indirect contact routes of transmission should be properly incorporated in epidemiological models to achieve a better understanding of transmission processes and to detect farms playing a key role in the case of epidemics. This information is crucial for the design of effective surveillance and control plans.
The potential functional role played by farms in case of epidemic spread was further analysed with the specific goal of identifying the most critical farms, where the implementation of biosecurity measures may be more effective in protecting the entire farm system. The goal was to investigate trade-offs emerging when epidemiological models are used to support the design of disease management interventions against possible future epidemics. Different scenarios of implementation of on-farm biosecurity measures (i.e., preventive measures reducing the risk of disease introduction) were investigated, varying both the number of farms implementing biosecurity and criterion to select them. In order to optimize the allocation of resources for biosecurity implementation, the performance of each scenario was assessed with respect to different objectives, such as the reduction of the probability of either frequent or extreme outbreaks in the farm system. Trade-offs between objectives did emerge, suggesting that the identification of key farms is not trivial, and that the definition of management objectives is a crucial phase to define risk-based biosecurity plan. The definition of the best strategies to protect the system against epidemics is not easy and only the discussion between modelers and decision-makers can help in finding proper solutions.

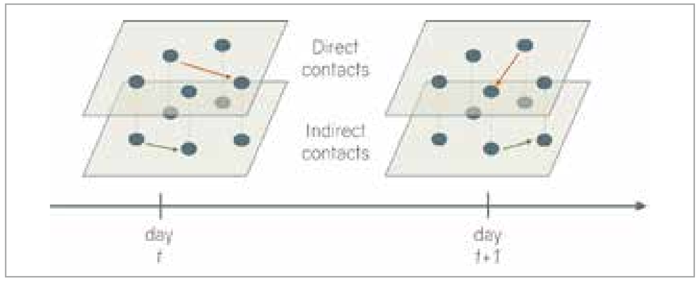

Fig. 1 - Simplistic yet illustrative example of a temporal multiplex network. Each day is represented as a snapshot, i.e., a static network, with two layers, grouping direct and indirect contacts occurred in that day. Nodes in the two layers are the same and represent farms.

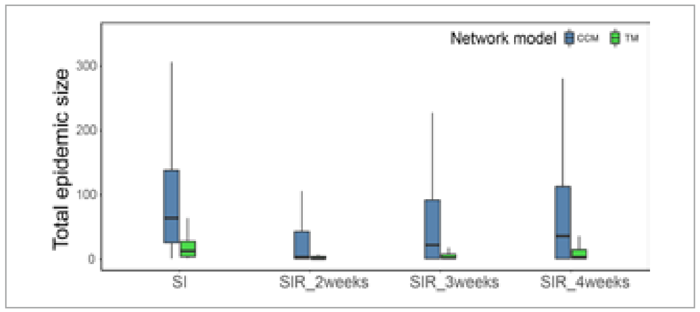

Fig. 2 - The boxplots summarize the distributions of total epidemic size (measured for each simulation as the number of infected farms at the end) obtained when the disease spread process was simulated on the partially informed model (blue) and on the complete one (green) using different epidemic model: an SI and an SIR with different infectious period (from 2 to 4 weeks). The modelling approach adopted for the description of indirect contacts can highly affect the simulation outcomes and the partially informed model systematically leads to higher values of total epidemic size.

HARDWARE DESIGN ANDIMPLEMENTATION OF MEMRISTIVE-BASED LEARNING SYSTEMS FOR EFFICIENT NEUROCOMPUTING

Stefano Bianchi - Supervisor: Daniele Ielmini

Neuromorphic engineering aims to reproduce brain-like reasoning in a silicon chip. Computers able to learn by sensory excitement from the external world, to infer abstract concepts and to make decisions, are spurring a new technological revolution reshaping all aspects of our life and society. The development of artificial intelligent hardware systems must overcome the current architecture used for standard computers, which relies on physically separated processing unit and memory: data are continually sent to the processing unit from the memory, elaborated in the processor and then sent back again to the memory. Such repetitive transmission of information is the main difference with respect to the biological computation, where the knowledge is elaborated in the same place in which it is stored, hence "in-memory computing". In order to introduce a significant improvement in the hardware design, the scientific research has recently focused on the memristive devices such as the phase change memory (PCM) and the resistive random-access memory (RRAM). The main advantage of memristor devices is related to the 3D stacking capability in the back end of the line, which zeroes the transfer of information and enables "in situ" computation. For this reason, as well as for the great area efficiency (size of only few square-nm), the memristive devices are the best candidates to foster the next technological era.
This doctoral dissertation describes some novel approaches to improve the intelligent computation by both bio-inspired and artificial standpoints. Concerning the bio-inspired learning algorithms, the STDP (Spike-Timing-Dependent Plasticity) is one the most plausible paradigms accepted for the description of the learning activity in the human brain. For this reason, STDP is here verified by simulations and experimental measurements in extended networks with resistive random-access memory (RRAM) devices used as synapses. Great attention is also given to the prediction of the behaviour of the bio-inspired networks via accurate analytical models. Furthermore, the Verilog-A modelling of the memristive devices is also investigated in order to introduce a complete framework for the simulation of memristor-based circuits in computer-based drafting tools. On the other hand, the pure algorithmic approach to artificial intelligence has led to the definition of deep neural networks based on training algorithms, such as backpropagation, able to perform complex tasks. These techniques, which rely on a large number of arithmetical operations (matrix-vector multiplications, MVMs), demonstrated high reliability and efficiency in fully connected and convolutional neural networks for object recognition, natural language processing and playing games. However, despite of the significant efficiency in specific tasks, the artificial neural networks lack the sufficient plasticity for the adaptation to continually evolving situations.
To both rely on the computational accuracy of convolutional networks and on the bio-inspired plasticity and resilience of STDP, this doctoral dissertation proposes a new kind of artificial neural network. The new architecture is capable of learning and classifying new input objects without catastrophically forgetting previously learnt information, thus achieving lifelong learning. The efficacy of the neural network is highlighted by PCM-based experimental demonstrations of continual learning for the MNIST and CIFAR10 datasets, with particular attention on the cohesion of stability and plasticity enabled by the multilevel programming of the PCM devices. In order to introduce a further computational novelty in terms of performances and resilience in the neuromorphic engineering, this doctoral dissertation also proposes a new PCM-based homeostatic neuron. At each fire event, an internal PCM device is partially crystallized, thus modulating the internal threshold of the neuron. This assures improved pattern specialization, significant reduction of power consumption, robustness against external perturbations and self-control of each neuronal activity as a function of the spiking rate. The homeostatic neuron is also shown to enable multipattern learning from Fashion-MNIST dataset via unsupervised asynchronous STDP. The PCM-based neuron is also useful in the framework of reinforcement learning, where the interaction with the surrounding environment contributes to the evolution of the network dynamics. In particular, concerning bio-inspired recurrent neural networks (RNNs), it is here demonstrated that the self-adaptation driven by PCM devices can control the internal states of the neurons in relation to the past experience of the network, thus enabling the fulfilment of complex decision making tasks. In order to validate the concept of reinforcement learning with memristive devices in large scale, an extended hardware system is also presented. In particular, the hardware is capable of reproducing bio-plausible cognitive functions with a SiOx RRAMbased architecture mastered by a digital system on chip (SoC). The computation of the system carries out significant results by merging Hebbian learning and homeostatic plasticity for improved efficiency and stability in bio-inspired RNNs. The hardware is tested for two main tasks: (i) the autonomous exploration of a continually evolving maze (i.e. a maze whose walls continually move in time); (ii) the Mars rover navigation, which concerns the exploration of Martian landscapes by using the NASA database of images. The hardware self-optimizes its policy starting from stochastic trials and plastically modifies its synaptic connections to reach the optimum escape path by progressive experience of penalties and rewards. Such studies introduce a novel approach to real-life problems and propose a significant boost of the technologies for improved computational capabilities. However, in order to compute the benefits of the memristor-based computation, an integrated design procedure in computerbased drafting tools must be also analysed. The integrated design of memristor-based neural networks, along with the definition of robust circuital architectures, promises significant breakthroughs in the future thanks to the enormous advantages in terms of compactness, efficiency, and speed. In particular, this doctoral dissertation also proposes the high-level analysis of an integrated fully connected artificial neural network designed using the software kit STMicroelectronics BCD 90 nm with embedded PCM cells. The memristor-based computation enables the use of only one-clock computation per single MVM, which is by far more advantageous with respect to the thousands of digital multiply and accumulate operations (MACs), each requiring at least one clock pulse, currently needed by standard Von Neumann computers. Furthermore, significant improvements in terms of hardware architecture are also proposed since the design is robust against device non-idealities such as the conductance drift of the PCM devices. This doctoral dissertation offers a wide range of analysis in various fields, ranging from pure device and network modelling to systems and integrated designs.
It offers several insights over bio-inspired computation and neural networks using memristive devices, deepening theoretical and experimental solutions to improve the state-of-the-art of artificial intelligence. The novelties here introduced, as well as the description of the experimental setups, are key elements to introduce new paradigms in the standard elaboration procedures, thus highlighting the fundamental steps for boosting the next technological era of computation.

SENSOR-ASSISTED COOPERATIVE LOCALIZATION AND COMMUNICATION IN MULTI-AGENT NETWORKS

Mattia Brambilla - Supervisor: Monica Nicoli

This doctoral dissertation, authored by Mattia Brambilla and supervised by Prof. Monica Nicoli, presents research advances on cooperative localization and communication. These two macro trends are investigated in multi-agent networks, where time-varying agents of unknown absolute location are asked to fulfill positioning and information sharing tasks.

The research on localization aims to develop an integrated solution where cooperative self-localization of agents is combined with multitarget detection, localization and tracking. These two problems are jointly resolved in a holistic approach where the graph theory is used to describe the relationships among agents, targets, and observations. Targets are not only considered as unknown objects to be just localized, but their statistical properties represent valuable information to be conversely used to refine the agents' positioning coming from self-definition of the multi-agent network. The operating framework is based on belief propagation, a massage passing algorithm working on the factor graph describing all relationships among agents, measurements and external variables. The final result is the development of a generalized solution that is flexible enough to accommodate heterogeneous measurements of diverse agent types, ensuring a wide range of applications. Both a centralized and a fully distributed implementations are considered, allowing for a customization of the processing algorithm to get the desired functionalities according to the needs of each specific application. In this regard, we considered the maritime and vehicular environments (as can be inferred from Fig 1), but internet of things in industrial and surveillance contexts are other major applications of interest. In maritime context, a centralized implementation is adopted as it better adhere to operative conditions of deploying multiple agents which communicate to one single processing unit. In vehicular settings, instead, a cooperative distributed implementation has to be preferred where vehicles need to exchange local information to their neighbors in proximity, creating fast Vehicle-to-Vehicle communication links without the support of any infrastructure, as allowed by 5G standard.

The research on communication is focused on vehicular applications, where the use of Vehicle-to-Everything (V2X) links to extend vehicle's sensing capability by establishing connections with nearby road users represents a remarkable upgrade for a safer mobility. Such paradigm is here reverted: this thesis proposes the use of vehicles' sensors to enhance communication performances. The specific research topic consists in developing sensor-assisted beam alignment techniques, where narrow directional V2X beams are steered according to the information extracted from perception sensors. Attentions are given not only to the specific technical aspects of the methodology and technologies, namely Millimeter-Wave (mmWave) and Free-Space Optics (FSO), but also a system-level architecture is proposed. Inter-vehicle cooperation and intra-vehicle sensor data fusion are combined in a unified system targeted to guarantee reliable beam-based V2X links in mobility scenarios.

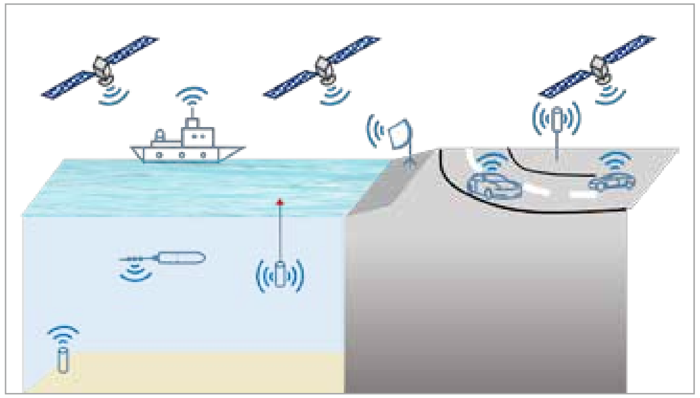

Fig. 1 - Representation of a multi-agent network in maritime and vehicular scenarios.

Both macro-researches deal with a tight integration of heterogeneous sensors with communication in dynamic multi-agent systems. The volatility of interconnections among agents due to mobility and instability of links mandatorily calls for flexible and adaptive techniques, capable of profitably fuse diverse types of information. The outcomes of this thesis demonstrate how a statistical approach is capable of handling realistic problems and developing versatile solutions.

ON THE MANAGEMENT OF POWER AND PERFORMANCE TRADE-OFFS IN DISTRIBUTED CLOUD-NATIVE INFRASTRUCTURES

Rolando Brondolin - Supervisor: Marco D. Santambrogio

Cloud computing is now the de-facto standard for the deployment of complex and scalable applications and systems at scale. In the last few years, cloud computing applications shifted from the monolithic architecture to a more flexible microservice-based one with the so-called cloud-native ecosystem. This shift allowed to separate concerns among different development teams, increased the scalability of the cloud applications and allowed to develop, test, and deploy each functionality almost independently from the rest of the system. Cloud-native applications fostered even more the growth of cloud computing and, for this reason, cloud providers have to manage an unprecedented amount of applications for a huge amount of users. This trend poses new challenges in the management of data-centers. In particular, the expected energy usage of data-centers will reach 8% of the whole energy consumption of the world by 2030. Moreover, power consumption represents 20% of the Total Cost of Ownership (TCO) of a data-center. If we consider that the CPU is currently the most power-hungry component of a server, there is the need to optimize how cloud applications are executed within cloud infrastructures to keep the cloud-computing growth sustainable.

It is clear that there exists a trade-off between performance and power consumption of cloud-native applications executed onto cloud infrastructures. Although a simple power-capping approach might be tempting, the effects that this activity might have on the performance of cloud-native applications are devastating. Increasing the latency of the responses of the Google search service from 100ms to 400ms reduces the number of searches per user from -0.2% to -0.6%. Such reduction persists for some time even if the latency is restored to the original value. For these reasons, any system that wants to reduce the power consumption of cloud-native applications should keep into account performances as the main aspect to protect.

Within this context, the goal of this thesis work is to design and develop power management techniques able to sustain the performances requested by cloud-native applications and workloads while reducing as much as possible the power consumption such applications generate. Given the complexity of cloud-native applications and environments, an automated approach is required. In particular, we exploited the Observe Decide Act (ODA) control loop, which is an autonomic methodology that enables the applications to adapt their behavior depending on their state and on the observations of the surrounding environment. Such methodology poses many challenges that we addressed throughout the development of this thesis work: (1) how we can measure the behavior of cloud-native applications and the environment in terms of performance and power consumption; (2) how we can define performance throughout cloud-native applications and how we can define meaningful performance targets; (3) how we can effectively and precisely reduce power consumption while preserving performances of the running workloads.

The first challenge deals with observability and with the ability to achieve self-awareness. Unlike many other previous techniques, we decided to tackle this challenge with a black-box approach. This means that we observe and obtain knowledge of the application components without instrumenting the workloads. Although this approach poses some limitations in the kind of information we can collect, it provides huge flexibility as any future application component can be observed and monitored out-of-the-box without modifications.

The second challenge deals with the correct definition of what performance is for a cloud-native application and encompasses both the Observe and the Decide phases of the ODA control loop. In particular, we experimented with two different performance metrics: CPU usage and average latency. On the one hand, CPU usage is typically defined by application developers as a performance constraint within Kubernetes [2]. However, setting CPU requirements correctly for each component is hard, thus there is room to optimize how the allocated CPU is used by the application components. On the other hand, latency is a performance constraint understandable both from developers as well as final users, as it defines the quality of the user experience. Moreover, latency allows using queuing theory to guarantee performance during execution.

Finally, the third challenge deals with how to enforce performance constraints to save power. Such challenge involves both the Decide and the Act phase of the ODA control loop. Within this thesis, we experimented with heuristic control based mainly on basic queuing theory and with fast and accurate power actuation, represented by Running Average Power Limit (RAPL). Within this context, RAPL is our source of information for power consumption as well as for power actuation.

To address the challenges described so far, this thesis work defines the following contributions. We designed and implemented DEEP-mon, which is a dynamic and energy efficient power monitor able to attribute power consumption and performance metrics to each container running on a given host. DEEP-mon leverages kernel features like extended Berkeley Packet Filter (eBPF) to avoid user-code instrumentation. The proposed approach is accurate and increases the system power consumption of 0.90%, representing a low overhead that is the current state of the art in the field.

We designed and implemented a monitoring methodology able to capture network activity and its performance without instrumenting the user code and we integrated it within DEEP-mon. The resulting tool monitors for each container its CPU usage, power consumption, instruction retired, cycles, cache references, cache misses, network bandwidth, average network latency, and network latency percentiles (from 50th to 99th). Results show an overall measurement error that is below 5% for almost all metrics with a lower overhead w.r.t. similar approaches in the state of the art.

We developed HyPPO, a Hybrid Performance-aware Power capping Orchestrator. HyPPO implements an ODA control loop that builds upon the data we collect with DEEP-mon to maintain a given level of performance while reducing power consumption. In particular, we manage CPU usage, which is commonly set as a constraint by application developers to each container in a Kubernetes cluster. HyPPO acts by applying a hardware power cap with RAPL and controlling the increase in CPU usage up to the level specified by the application developers. This approach allows reducing by 25% on average the power consumption with an average SLA violation of 5%.

We developed PRESTO, a latency-aware power capping system for cloud-native microservices. PRESTO implements an ODA control loop that exploits the power and performance data collected with DEEP-mon to maintain an average latency requirement while saving power. The latency requirement is split across all the microservices of the cloud-native application through queuing theory and a graph analysis that allows to precisely attribute service times to be enforced to each microservice. PRESTO leverages the same RAPL actuator of HyPPO and reduces the power consumption of 37.13% on average with a control error of 12.5% and of 1.5ms on average.

Finally, we explored how to improve energy efficiency of microservice-based applications by introducing heterogeneous architectures, merging together the elasticity of cloud-native applications and the performance and energy-efficiency of Field Programmable Gate Arrays (FPGAs). The results are: (1) an improvement in the FPGAs time utilization while maintaining the performance of the applications and (2) power savings w.r.t. a pure software system implementing the same functions. This work represents an interesting initial study and paves the way for more extensive research work on how to accelerate microservices and cloud-native workloads.

COMPUTATIONAL METHODS FOR DATA-DRIVEN PREDICTIONS AND UNDERSTANDING OF BIOLOGICAL INTERACTIONS

Gaia Ceddia - Supervisor: Marco Masseroli

Systems biology is a well-known approach aimed at integrating and comprehensively analyzing the growing amount of biological data. It leverages mathematical models such as complex networks to organize experimental data from complex systems and easily incorporate them. The use of complex network modelling permits a different perspective on biology since it avoids the study of isolated components (e.g., genes and proteins), enabling a comprehensive view of biological processes. For example, it has been proved that multiple disruptions in gene networks cause more damage than the removal of a particular gene. Thus, complex networks representations and approaches give interesting insights into the overall structure and dynamics of the cell organization. Moreover, various molecular networks can be studied, e.g., protein interaction, gene regulation or metabolic networks; however, the main goal is to merge these different networks into one coherent framework modelling an entire organism. While research is still far away from this, network-based studies like this Thesis may contribute to this aim by shedding light on how to use and interpret the enormous amount of data available.
This Thesis deals with the use of biological networks in order to obtain relevant results for each presented application. Particularly, I delivered three network-based implementations, each tackling one key concept of complex biological systems: 1) network inference, the discovery of well-connected structures of nodes (e.g., genes, proteins and drugs) where the study of isolated elements does not explain the functionalities carried by the entire system; 2) link prediction, detecting connections in a network leveraging the overall system architecture and available heterogeneous data; 3) network feature extraction, identifying special patterns of connectivity sharing similar functions within a complex system. All these three features have already been studied in system biology; however, my contribution to this field lies in the applied methods, integration techniques and exciting results. This Thesis aims to the successful use of complex networks to represent different phenomena such as transcription factor interactions, drug-target associations, drug-related annotations and gene communities, the network-related predictions retrieving protein networks or drug repurposing possibilities, and the biologically relevant module extraction to find novel cancer biomarkers. In order to achieve these results, I would ensure the reproducibility of my works by making them publicly available for testing and usage. Altogether, this Thesis gives optimal examples of how the central characteristics of complex networks may lead to biological discoveries in different applications of systems biology.
More in detail, for the network inference application, this work illustrates an unexplored possibility for predicting transcription factor (TF) interaction networks. I applied an association rule method, a well-known yet simple approach, that lead to network representations of ChIP-Seq data coming from diverse sources. This study reports a comprehensive and forthright pipeline for inferring interaction networks using a novel statistical index called the Importance Index, constituting the predicted weight of the links in the TF network. The specific goals of this application are: to develop a general approach for analyzing any set of ChIP-Seq data, for finding TF co-factors of a given TF and its network; to implement such a method into an R/Bioconductor package (called TFARM) available; to report relevant biological results obtained through complex network representations. Moreover, I proved that inferred networks contain well-known TF interactions using experiments on ENCODE ChIP-seq data. The hallmark of this work is to provide researchers with the ability to identify and understand TF networks managing ChIP-Seq data and allowing user-selected analyses.
Considering the growth of publicly available databases and the emergence of integration approaches, I wanted to address another intriguing challenge, which is to assess benefits and discoveries related to the combined use of heterogeneous data. For this reason, my purpose explored the advantages of using an integrated Non-Negative Matrix Tri-Factorization approach for drug repositioning. Thus, I both combined multiple data sources and explored the link prediction problem in complex networks by means of a factorization approach. To find new indications for already approved drugs, I modelled annotation, association and interaction data into one multipartite network where the drug repositioning question becomes a prediction problem, and the principal goal is to infer drug-indication, drug-target and drug-disease links. Not only link predictions computationally benefitted from additional data, but in particular novel biological results of the drug repositioning application confirmed the genuine advantages of modelling such data into complex networks. Indeed, a knowledge-based evaluation demonstrated that top-scored predictions are biologically relevant.
Finally, to have a complete analysis of complex networks, this Thesis deals with the research issue of finding particular patterns of connectivity, also called communities. Thus, the network feature extraction problem is addressed by selecting highly connected groups of nodes in the studied network. These clusters and their organization control the entire graph architecture, i.e., perturbations to these modules causes a more notable biological malfunction than a single-node elimination. To achieve this result, I employed an innovative integration of multiple gene co-expression networks through the Similarity Network Fusion (SNF) approach. Since the integration approach for drug repurposing lead to exciting discoveries, I decided to tackle the network feature extraction problem by combining various networks carrying miscellaneous information. The goal of this work is to simultaneously fuse networks representing different similarity metrics and select those nodes characterizing the structure of the fused graph. Indeed, I used the SNF method for its feature of contributing to the emergence of modules with shared similarities across networks. The analysis had as main objective to provide functionally and biologically relevant cancer biomarkers using an innovative network feature extraction approach. The final goal of extracting communities of cancer gene biomarkers is accomplished through the implementation of a complete pipeline.
All these case studies proved that integrating data sources using complex network approaches enables a wide range of significant discoveries, whether the goal is graph representations, link predictions or feature extractions. This work highlights fascinating features of complex networks applied to systems biology and motivates more efforts in network-based research.

ANALYSIS AND CONTROL DESIGN OF ADVANCED SUSPENSION SYSTEMS

Stefano Dattilo - Supervisor: Sergio Matteo Savaresi

Co-supervisor: Matteo Corno, Giulio Panzani

This dissertation deals with the technological analysis and control strategies design for advanced suspension systems for automotive applications. The suspension is typically composed by two elements: the elastic element and the dissipative element. The aim of the *elastic element* is to support the weight of suspended mass and to keep it in its nominal position, and it is usually a coil spring or a gas-spring. The aim of the *dissipative element* is to dissipate the energy introduced by road disturbances, and it is usually composed by a piston-cylinder system. The cylinder is filled with oil, and the viscous friction between oil and piston dissipate energy during suspension movements.

During the last decades, the continuous drop of electronic component cost has led to the development of electronically controllable suspensions systems.

This type of suspension can modulate the generated force, and through the development of control algorithms it is possible to overcome the passive suspension limitations in terms of comfort and safety.

In this dissertation we will analyze and explore the potentialities of three advanced suspension system. The **Multichamber suspension**, which is a suspension able to modulate both the dissipative element and the elastic element at high frequency. The **Proactive suspension**, which is a semi-active suspension, with short term active capability and energy recovery feature. It can generate active force up to 10 Hz, but its capability to recover energy reduces its energy request. The **Hydro-pneumatic suspension,** which is a load-leveling suspension controlled with on-off actuators. It is largely used in agricultural applications, where the energy demand is not an issue.

Those types of suspensions are already available on the automotive market and they represent the suspension state-of-the-art technology, but, at the best of the author's knowledge, their potentialities are under-explored. The objective of this thesis is to provide a complete technological analysis and propose control strategies able to exploit the capabilities of each type of suspension mentioned above.

Multichamber suspension

The innovative aspect of the Multichamber suspension (Fig. 1) is represented by the controllable air-spring. The elastic element is an air-spring equipped with an auxiliary chamber. The main chamber and the auxiliary chamber are connected through a controllable on-off valve. Controlling the valve, it is possible to modulate the suspension elastic force. Intuitively, the more is the available volume, the softer is the suspension stiffness; it means that the valve-state influences the suspension stiffness. At the best of the author's knowledge, this type of technology is widely unexplored.

A realistic model of the multichamber suspension is proposed and validated through test-bench experiments on a prototype. Once a reliable model of the system is obtained, the comfort-oriented potentiality of this type of suspension are studied in simulation environment. Firstly, an optimization problem is solved to define a comfort-benchmark of this technology, then a model-based comfort-oriented control law is proposed, which can reduce the vehicle vertical oscillation by 8% with respect to a classic semi-active suspension.

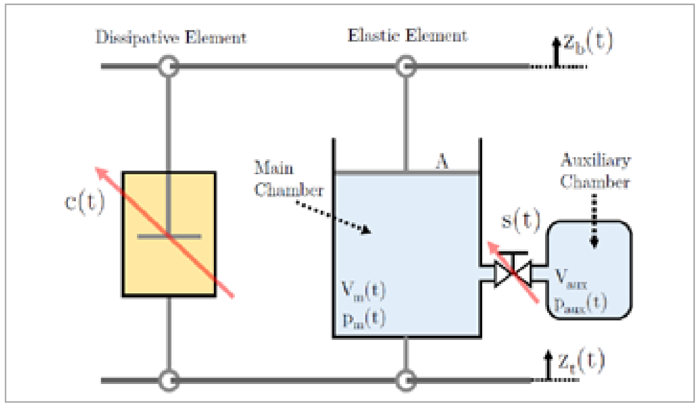

Fig. 1 - Multichamber suspension scheme

Proactive suspension

The suspension structure is like a classic damper (Fig 2). The innovative aspect is the presence of a motor-pump system (called ActiValve) able to move oil from one chamber to another generating an active force. Notice that, based on the working condition, the motor-pump system can be used as a generator, it means that this type of technology can also recover energy. Firstly, a model of this type of actuator is proposed and validated through test-bench experiments on a prototype. Then different centralized comfort-oriented control strategies are proposed. The objective is to control all four proactive suspensions, in a synchronized way, to reduce vehicle vertical, roll, and pitch accelerations. The results show that this type of suspension can reduce vehicle acceleration by at least 60% with respect to the passive configuration. Eventually, a single-corner safety-oriented control is designed. Considering a full brake maneuver, the designed control can reduce the braking time by 7% with respect to passive configuration.

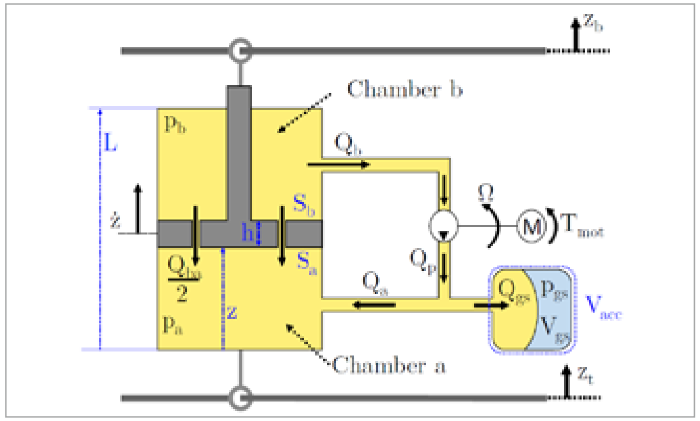

Fig. 2 - Proactive suspension scheme

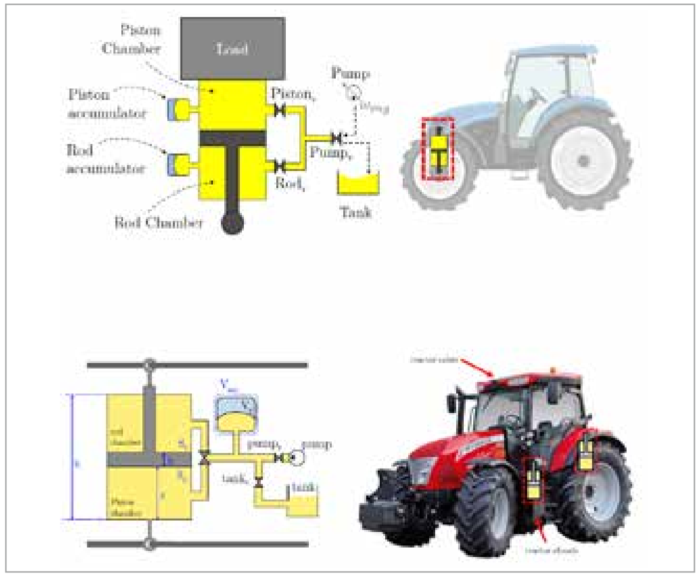

Fig. 3 - Hydro-pneumatic suspensions schemes

Hydro-pneumatic suspension

Two different hydro-pneumatic suspension applications are considered. The first one is a Front-Axle suspension for small size tractors, the second one is a fully suspended cabin for large size tractors. Both type of suspensions are connected to an external hydraulic circuit, which can be actuated in order to insert or extract oil from the cylinder changing its equilibrium positions. This external circuit cannot be continuously actuated, but it presents only ON-OFF actuation, and it should be controlled to compensate the system equilibrium alteration due to slope or mass variations. Considering the front-axle application, the control problem is a multi-variable load-leveling problem; indeed, the considered suspension allows to control both piston position and chamber pressure. Firstly, a control-oriented model of the system is developed, and then load-leveling control strategy is proposed. Eventually the designed control strategy is validated through experimental results. Considering the tractor-cabin application, the objective is to control the four hydro-pneumatic suspensions (under the four cabin corners) to control the cabin degrees of freedom: heave, roll, and pitch. Therefore, also this application represents a multi-variable load-leveling problem. Firstly, a control-oriented model of the system is developed. Even if the application is completely different, the obtained model is of the same form of the front-axle one. For this reason, we could apply the same control strategy designed for the front-axle application.

HYBRID SYSTEMS: COMPUTATIONAL APPROACHES TO VERIFICATION AND CONTROL

Riccardo Desimini - Supervisor: Maria Prandini

The goal of this thesis is introducing computational approaches to solve verification and control design problems for hybrid systems, a class of dynamical systems that has attracted a lot of attention due to its significant impact on various application domains. Indeed, hybrid systems provide a comprehensive modeling framework that allows to tackle problems involving complex systems characterized by intertwined physical and logical components.
We focus, in particular, on discrete-time PieceWise Affine (PWA) dynamical systems, a class of nonlinear dynamical models characterized by a finite collection of affine dynamics, each one associated to an element of a polyhedral partition of the state space. PWA systems are attractive from a computational viewpoint due to their simple description in terms of affine dynamics on domains described via linear inequalities. Also, despite their simplicity, they have powerful modelling capabilities: they naturally arise as models of certain classes of systems with a phased behavior, and they can be used to approximate smooth nonlinear systems with an arbitrarily set accuracy level, to the purpose of formal verification and control design. PWA systems are equivalent to Mixed Logical Dynamical (MLD) systems, a class of hybrid models that can represent various kinds of processes characterized by intertwined physical and logical components (e.g., finite state machines interacting with dynamical systems and systems with mixed discrete/continuous inputs and states), and are described via a collection of linear equalities and inequalities involving both real and integer variables, which makes mixed integer linear and quadratic programming a natural framework for their analysis and design.
Given a PWA system, we aim at determining if it can satisfy some property related to its state evolution and operate in a safe and/or efficient way. We consider the case when the PWA system evolution is affected by input variables representing either disturbances or control inputs: if some of them are control inputs, then we are solving a control problem; if no control input is present, then, we are addressing a verification problem.
In the latter case, the developed computational procedures return a response to the verification problem that includes a counterexample when the desired property is violated. This can be useful for the redesign of the original system, possibly including a controller.
The choice of adopting a discrete-time framework is a natural consequence of the control-oriented perspective of this work. Indeed, controllers are implemented through digital devices: the control input to a system is typically applied using a zero-order hold and is derived from state/output measurements obtained at sampling times.
The developed toolkit of computational methods for verification and control of PWA systems is inspired by existing approaches in the literature on reachability analysis, reach set representation and reduction, invariant set computation, model abstraction, and further extend and combine them in an original way within a comprehensive framework, including also the parameter varying case, thus enhancing the applicability of the model-based approach to PWA systems verification and control.
The thesis is structured in three parts on modelling, verification, and control, which are described next.
Part I - Modelling
In Chapter 1 we present the adopted discrete-time PWA modelling framework and present a procedure to obtain an equivalent MLD system starting from a PWA model.
Part II - Verification
In Chapter 2 we describe the proposed set-based methodology to address formal verification of discrete-time PWA systems. The applied approach combines and further extends existing techniques from the literature on reachability analysis and invariant set computation. We demonstrate the features of our verification approach by means of numerical simulations on some test cases. A comparison with an alternative approach based on MLD reformulation and mixed integer linear programming is included, where we show the better scaling properties of our method.
In Chapter 3 we present an extension of the method of Chapter 2 to the class of PWA parameter-varying systems. This extension rests on the introduction and combination of suitable PWA abstractions for which reachability analysis can be performed as in Chapter 2. We illustrate the extended method by revisiting a numerical example from Chapter 2.
In Chapter 4 we address safety verification of discrete-time smooth nonlinear systems by means of an abstraction and refinement approach. We adopt a conformant PWA abstraction to compute a polytopic robust control invariant set that is included within the assigned safe region and includes the set of initial conditions. We also devise a local guided refinement strategy of the PWA mode partition so as to reduce the computational load with respect to a global uniform refinement.
Part III - Control
In Chapter 5 we consider finite horizon control of PWA systems affected by an additive disturbance. We design a static state-feedback controller driving the system from an initial region to a target region, while also constraining the state trajectories to an assigned tube of polytopes. By resorting to a zonotopic control law parametrization, we develop a reachability-based control design procedure that is easy to be implemented online and that can cope with tighter constraints with respect to alternative tube-based approaches in the literature.
In Chapter 6 we address the problem of designing an explicit model predictive controller for constrained stochastic linear systems. The aim is to minimize an expected quadratic cost subject to robust state-input constraints. We present a novel design methodology resting on an affine parametrization of the control law that allows to reformulate the control problem as a convex quadratic program. The resulting explicit controller is sub-optimal because of the imposed parametrization of the control law but has a reduced complexity compared to alternative solutions obtained through min-max approaches where the worst case cost is minimized.

EXPLORING AND CHALLENGING THE LIMITS OF LANGUAGE MODELS AND THEIR APPLICATION TO HUMAN-GENERATED TEXTUAL CONTENT

Marco Di Giovanni - Supervisor: Marco Brambilla

In the field of Natural Language Processing (NLP), Language Models (LMs) have always been one of the most useful and used approaches to process textual data. During the last decades, researchers developed algorithms to embed texts into numerical vectors to perform Machine Learning tasks such as word-, sentence-, pair-of-sentences-classification (e.g., Machine Translation, Part-of-Speech Tagging, Named-Entity Recognition, Sentiment Analysis). Due to the intrinsic sparse nature of texts, the multitude of modern languages and their differences, and the scarce availability of big labelled datasets, the development of a model that completely understands natural language is a difficult task, not solved yet.
At first defined by simple unigram models, through the years LMs have been improved until the recent release of BERT, a pre-trained LM that reached state-of-the-art results in many heterogeneous tasks. BERT is a deep Transformer-based model whose architecture is a collection of self-attention layers with residual connections. The model is pre-trained, in an unsupervised way, on huge corpora of multilingual texts, on two tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP). It can be later fine-tuned on supervised tasks, adding on the top of its architecture an appropriate output layer. This transfer learning approach allows the model to obtain higher accuracies on small labelled datasets thanks to the encoded information from the pre-training step.

In this thesis, I apply LMs to textual content publicly shared on social media, intending to find meaningful representations of users that encode syntactic and semantic features. Once embeddings are defined, I compute similarities between users to perform higher-level tasks. I investigate knowledge extraction, in the form of “emerging” users and “entity-relation-entity” triplets, community detection and characterization of users, classification of users’ political inclinations and detection of controversial topics discussed in social networks, concluding with an evolutionary adversarial attack approach to check the robustness of the selected models.

The thesis starts with a collection of works about knowledge extraction from social networks. The first study is a tentative to detect emerging knowledge from social media, to automatically enrich knowledge bases. The proposed pipeline, given as input a set of Twitter users (seeds) belonging to a predefined community (e.g., fashion designers) usually selected by an expert in the field, outputs a set of candidate Twitter users, sorted by similarity to the seeds. The precision of the pipeline is calculated checking whether the best candidates belong to the same community of the seeds. This approach can be iterated using manually verified candidates as new seeds. Experiments on eight heterogeneous datasets show that the iterative approach finds an increasing number of entities and that its precision is approximately constant during the process. It detects candidates from the whole world in just a few iterations and a high fraction of them can be considered “emerging” since they do not have a dedicated Wikipedia page yet.

It is not easy to define how similar two Twitter users are. User similarity can be reduced to textual similarity by neglecting the social connections of users (follow and friend relationships) and their online behavior (comments and likes to posts), even if those features could potentially be useful to profile them.
A two-steps approach is designed to calculate the similarity between users: an initial embedding of the textual content into a vector, followed by the computation of distances among vectors.
An accurate selection of types of words, used to create the embeddings, is performed by collecting their POS tags. Nouns, proper nouns and verbs are selected as specific syntactic types to generate BoW embeddings.
Semantic alternatives are obtained searching for instances and their types with DBpedia to create other BoW embeddings. Finally, another approach to embed texts semantically is performed applying topic detection algorithms, such as Latent Dirichlet Allocation (LDA), so to obtain feature vectors representing the topics discussed by users.
The computation of similarities is analyzed empirically selecting the best metric to calculate distances between vectors.
Many experiments on different communities (ranging from Chess Players and Australian Writers to Italian Politicians belonging to different parties and their follower, or players of different sports) allows a safe selection of the best pipeline.

The thesis continues with the description of a study about the detection of controversies in online Twitter discussions.
The designed approach compares in accuracy to the state-of-the-art graph-based approaches, where the retweet graph is used to classify whether a topic is controversial or not. The hypothesis is that, even if the structure of the retweet graphs reflects the sides of controversies, just the shared textual content should too. When users have different opinions about a topic, what they write and share about it should reflect the difference and LMs should be able to detect this difference.
This content-based approach starts with the selection of a topic as a set of keywords and the collection of the related tweets. The retweet graph is built, and Louvain algorithm detects two clusters. Textual contents of Tweets from each cluster are embedded with FastText or BERT, supervisedly trained, and the embeddings are used to calculate a controversy score. Non-controversial topics get higher values than controversial ones, thus a threshold score can be selected for classification purposes.
The approach achieves state-of-the-art results both in terms of ROC AUC score and computational speed, on a dataset of 30 online multilingual topics, half controversial and half not.

Finally, I perform a study on the limits of modern language models. Nowadays, it is relatively easy to store huge deep models, with hundreds of billion parameters, and train them using enormous corpora of unsupervised texts collected from the web.
Since these models obtain superhuman performances in benchmark tasks such as GLUE and SuperGLUE, works about interpretability and analysis of weak points recently attracted the attention of the research community, being able to light a bit these huge black boxes. We aim to understand how models learn and store information, and how to improve them.
EFSG is a model- and task-agnostic algorithm developed to generate adversarial texts that fool language models. Fooling a model means finding weak points, and defense strategies are needed to obtain improved models.
I design an evolutionary approach: sentences are individuals, and their fitness is the score from LMs. Through mutations and crossover, the algorithm generates always new and fitter sentences during the epochs. An ablation study is performed to find the best hyper-parameters of the algorithm, while transferability checks quantify if weak spots are general or model-dependent. Adversarial training is proved to be a good defense strategy, but improved models are not completely immune to the attacks.

Due to the recent COVID-19 outbreak, I also performed a brief analysis of public Facebook posts, focusing on their correlation with misinformation, such as the relationship between 5G technology and the virus, the relationship between migrant flows and the virus and rumors that the virus was created in a laboratory in China. Results of a linguistic analysis suggest that users are mainly polarized to a single misinformation topic, but their lexicons, different from users not sharing misinformation news, largely overlap. Finally, I apply BERT model to detect positive and negative sentiment of posts and I map sentiment peaks with events related to the COVID-19 emergency.

ECO-EFFICIENT SMART MOBILITY: A FOCUS ON THE HUMAN FACTOR

Alessandra Duz - Supervisor: Matteo Corno

The increase of congestion, traffic and air pollution, which follows global population growth and urbanization, drives worldwide efforts to change current mobility models. The main target of this quest is the transition towards smart mobility, which rely on information and communication technologies (ICTs) to seek for accessible and sustainable ways of transportation. The development of eco-efficient solutions is one of the most significant trends that currently characterize smart mobility. In fact, as the unmindful usage of energy sources shows its effects on climate change and global pollution, the necessity of minimizing the mobility energy expense and environmental impact is evident. Smart mobility proposes solutions to minimize traveling energetic expense through the active control of energy sources and vehicle dynamics. The shifting towards this type of transportation systems determines a change in the role of the human subject as a driver and consequently calls for a redefinition of the methods to deal with the human perception of ride quality. This dissertation presents three eco-efficient smart mobility solutions, which are developed with a focus on the human factor. Each solution investigates a different type of engagement of the person, more specifically I considered: autonomous driving systems, power-assisted cycles, and supervised mobility.

Autonomous driving systems have many advantages in terms of safety and reduction of traffic and congestion. Moreover, fully autonomous vehicles allow the person to make a better use of the traveling time, as the control of the motion does not require the passenger engagement. Through the planning of the vehicle motion, it is possible to manage the traveling consumption and the inertial forces acting on the passenger, the vehicle speed profile in particular has a high impact on both aspects. For this reason, I propose a computationally efficient global speed planner for autonomous vehicles, which optimises a tuneable trade-off between consumption and comfort aspects, within space-based speed constraints and with an arrival time requirement. The optimization relies on the dynamic programming algorithm, which is highly suitable for non-linear problems. The main innovative contribution is the inclusion of the passenger's comfort as planning criteria, this is defined through a frequency-based evaluation of the triaxial vehicle accelerations and included as weighted cost in the optimization cost function, together with the vehicle consumption. Moreover, I propose guidelines to reduce the computational load of the algorithm though progressive reduction of the system discretization. The speed planner is validated though the analysis of the optimal speed profiles and of the computational times. The dissertation also offers a deepening on the comfort topic through the identification of models of the head's dynamic response to longitudinal vehicle dynamics, which affects the arise of motion sickness symptoms. Experimental driving tests shows a compliant passenger pitch response to the longitudinal vehicle acceleration, which is proven by to cause an increase of motion sickness incidence. To the best of the author knowledge this research provides a first attempt to experimental identify the passenger pitch response to the longitudinal vehicle dynamics.

Human powered vehicles, such as bicycles and kick scooters, have a positive impact on urban congestion and pollution, and foster a healthy life style. They are however not viable for everyone as they require a certain level of physical fitness. Powered assisted vehicles overcome this limitation by decreasing the required user effort through the usage of an additional power source, which is typically electric. Energy management for those applications needs to tackle the limited energy storage capacity of this type of vehicles, that can easily lead to energy depletion. A charge sustaining energy management strategy for the development of full hybrid electric bike (HEB) was proposed in - *M. Corno, D Berretta, P. Spagnol, and S.M. Savaresi. Design, control, and validation of a charge-sustaining parallel hybrid bicycle. IEEE Transactions on Control Systems Technology, PP(99):1–1, 2015* -. The basic idea of HEB is to exploit the gaps in the cyclist's pedalling efficiency to guarantee a zero net battery energy consumption and a reduction of the cyclist's pedalling effort. The dissertation proposes the adaptation of the HEB energy management strategy to a bike-sharing free-floating application. Shared mobility enhances efficient usage of energy sources and road occupancy. In other words, it overcomes the Single Occupancy Vehicle (SOV) model, which is costly and increments traffic and congestion through the sharing of vehicles (e.g. sharing services) and the sharing of the ride with other persons (e.g. vehicle pooling). Although indirectly, the integration of those solutions in the commuting habits also promotes public transport, by covering the so called ``last mile", which is the portion of the route that is not reachable through public infrastructure. The integration of HEBs in bike-sharing free-floating service requires additional electronics that considerably change the electric power absorption profiles with respect to private use. In particular, the absorption while the bicycle is not being used cannot be neglected. Moreover, the average trip is considerably shorter than in private use and the mission profiles have a larger variability than in private use. These factors call for a complete redesign of the bicycle energy management system. This doctoral research designs a novel energy management system for a HEB for bike sharing applications. In particular, at the bicycle-level, I derive a stochastic control-oriented model of the energy dynamics of the rider and bicycle system, I define a method to quantitatively profile, from the energy point of view, the users of the sharing system, and I develop an adaptive control strategy for the tracking of an energetic power reference. The development of the bicycle-level energy management explicitly considers both the reduction of user's cycling effort and the congeniality of the human-motor interaction, which is gained through a slow regulation of the control action.

At the fleet level, the dissertation proposes two possible ride pricing strategies, aiming at the economic sustainability of the free-floating bike-sharing service. The system validation considers the analysis of the energy tracking during the ride, the cyclist's effort reduction and the economic profitability of the overall service. To the best of the authors' knowledge, no work in the open scientific literature addresses the energy management of a sharing service of a fleet of HEB and its associated costs.

Wearable devices allow continuous monitoring of the human status without limiting normal daily activity, and collect physical measurements of the person that can be processed and shared with the surrounding environment. The doctoral research focuses on the evaluation of the running effort, in order to provide the athlete with information on their optimal run gait. The energy-oriented running style evaluation is a well-known problem, which however typically requires cumbersome instrumentation. I propose as novel contribution the usage of two wearable devices, which evaluate the heart beat rate and respiratory parameters respectively, for the identification process. The analysis considers an experimental campaign of data acquisition for a trained athlete. By processing the two sensors' data I obtain distinct evaluations of the running effort as function of running cadence and trunk flexion. From the data analysis I identify a polynomial fitting for the description of the effort dependence on cadence. The cadence configuration with minimum effort is successfully matching for both sensors' effort evaluations. I could not draw exhaustive conclusion on the effort dependence on trunk flexion based on the available data.

AN ASSESSMENT OF REPRODUCIBILITY AND METHODOLOGICAL ISSUES IN NEURAL RECOMMENDER SYSTEMS RESEARCH

Maurizio Ferrari Dacrema - Supervisor: Paolo Cremonesi

In the era of exponential information growth many online services provide the users with personalized recommendations, which are known to yield substantial benefits in terms of user satisfaction. The development of ever improving algorithms to generate personalized ranked item lists is therefore a crucial goal for the field of recommender systems. During the last three decades various families of methods have been proposed, from simple nearest-neighbour based on heuristics to machine learning and graph-based models. In recent years, the research community has been captivated by neural approaches based on deep learning, which have become dominant in the literature.

Motivation

The community has developed widely accepted best practices for the evaluation of newly proposed models. Typically, a newly proposed technique is compared in a selected experimental setting against other state-of-the-art methods on at least a dataset. The ample freedom to design the evaluation protocol can make a comparison of the measurements reported in different papers opaque, due to the presence of many differences in the evaluation procedure, such as training-test split, pre-processing and recommendation list length. Identifying what represents the state-of-the-art for a given scenario can be challenging.

Indications that this methodology hides a number of issues have been published in several research fields and stem from the limited reproducibility of published results, the arbitrariness of the experimental protocol and the often poor optimization of reported baselines. All these issues cause what we referred to as *phantom progress.* In order to assess the extent of those issues for the recommender systems field and the level of progress made by new complex models, we attempted to reproduce recent results in the area of neural recommendation approaches based on collaborative filtering. The analysis included several high-level scientific conferences and identified 24 relevant articles published between 2015 and 2018.

Limited reproducibility of published research

The first step was to reproduce the published results based on the source code and data provided by the authors and according to the description in the paper. On average only 50% of articles could be reproduced. This result is in line with other studies in other fields of artificial intelligence and machine learning. For many published papers, the conclusion does not hold to an independent evaluation, which violates a cornerstone of the scientific method.

Poor competitiveness against simpler methods

As the second part of the study, we compared the reproducible complex methods against a set of many older and simpler baseline models, from simple nearest-neighbours to linear machine learning and graph-based models. To ensure a fair comparison the hyperparameters of all baseline models were thoroughly optimized. The results indicate that out of 12 reproducible articles, only 1 shows to be consistently competitive against simple baseline methods. All other models, despite their much higher complexity and computational cost, sometimes thousands of times higher even on high-level GPUs, fail to consistently outp
erform the simple baselines.

Methodological issues

We discuss this surprising result and trace it back to several common issues in today's research practice, which, despite the many papers that are published on the topic, have apparently led the recommender system field, for the task considered in our analysis, to a certain level of stagnation.
The issues can be grouped in poor reproducibility, arbitrariness of experimental configuration, poor optimization, selection and propagation of weak baselines, technical issues in the evaluation and insufficient evidence to support claims.
Regarding the reproducibility, considering that in this thesis we relied on the original materials provided by the authors, the main reasons for the limited number of reproducible papers is due to the lack of said materials. Furthermore, contacting the authors for assistance almost never led to solving the reproducibility issues for a paper.
The heterogeneity of the evaluation procedures is too an issue. While it is expected that different scenarios will require different evaluation protocols, it is common for algorithms that target the same scenario to be evaluated in all sorts of different ways. Usually, no motivation is provided for the specific evaluation protocol, which is therefore rarely grounded on a clearly identified real setup. The 12 reproducible articles overall relied on 18 different datasets, 7 training-test splits and various data pre-processing steps. We could not find 3 evaluated in the same way.
The optimization of baselines is too a widespread problem. There is a tendency to rely on the hyperparameters reported in the article that proposed that complex baseline model, regardless for how different the evaluation protocol and the data may have been. Usually, even the most effective algorithm does not produce state-of-the-art results if it is poorly optimized and any conclusion drawn from that weaker result is mostly meaningless.
The combination of those problems creates a cascade effect that propagates weak baselines as new articles are published. It has become common to consider state-of-the-art complex neural methods and some of the algorithms we analyse have become indeed common baselines, despite them not being competitive against simpler models. This means that simpler and better performing models are being replaced in the state-of-the-art with a multitude of more complex yet weaker models. The achieved progress is therefore sometimes non-existent as these complex models do not outperform previous methods, the state-of-the-art stagnates.
We also encountered a number of other issues. Surprisingly, half of the published implementations caused information leakage using test data in the training process and a similar number provided data splits which were inconsistent with the description provided in the paper. We also encountered errors in the implementation of evaluation metrics. Lastly, some articles made incorrect claims due to the lack of sufficient supporting evidence, regarding, in particular, the use of convolutional networks applied to the outer product of embeddings obtained from known low dimensionality approximation models. The claim was made by assuming an analogy between embeddings and images and by showing that the proposed model outperformed some baselines. Upon further analysis we could demonstrate that embeddings do not share the properties of images, and therefore the use of convolutions is theoretically unsound, moreover the convolutional algorithms do not have the claimed modelling capacity and are not able to outperform simple and well-known baselines.

Ways forward

The thesis discusses the highlighted issues in detail providing suggestions and best practices, calling for improved scientific rigour in the experimental evaluation of published research.

STUDYING USER ACCEPTANCE OF AUTONOMOUS DRIVING IN THE WILD

Alessandro Gabrielli - Supervisor: Matteo Matteucci

In the past two decades, autonomous cars have become a reality for many car manufacturers and software companies, and, nowadays, there are many ongoing tests on autonomous cars in our streets. Indeed, the car's automation technology is progressing fast and has already reached an excellent safety level.
As autonomous cars will reduce pollution, reduce road traffic, and increase road safety when commercially available, it becomes crucial to have a fast and wide acceptance of this technology. Understanding how a driver wants to be driven becomes a key factor. With the higher car automation levels (level 3-5 of SAE), the human role changes from active driver to passengers. The lack of controls perceived as passengers can lead to a higher stress level and discomfort if the car does not behave as expected. If passengers do not feel relaxed and comfortable in autonomous driving conditions, they probably will not use it, and so, the acceptance and usage of this technology will be delayed. Therefore, to guarantee the success and the acceptance of this technology, it becomes essential to focus on the passenger's comfort, as a psycho-physiological aspect and not as an ergonomic issue, where there is little available knowledge, and it is also important to understand how to increase the comfort and reduce the stress from the physiological point of view.
This research is a part of the I.DRIVE Lab (Interaction between Driver, Road Infrastructure, Vehicle and Environment) that aims to develop inter-disciplinary competencies for the analysis and modeling of behavioral aspects related to the interaction between driver, vehicle, infrastructure, and the environment in conventional or autonomous vehicles, where the driver becomes, de facto, a passenger (so the term driver refers to both driver and passenger). The laboratory is composed of a fixed structural component, based on a Virtual Reality simulator and a mobile component, based on an instrumented vehicle that are the core of this research.
This PhD research aimed to develop a platform that allows studying, in a real environment, the interaction between driver and car, focusing on the driver stress and the driving factor that impact on driver stress, to increase the knowledge in this field and help the acceptance of the car automation technology in the future.
The majority of previous research studies that have studied driver stress and workload have been conducted in driving simulators. Driving simulators have many benefits, like doing experiments in a controlled environment, studying dangerous maneuvers or dangerous situations, and repeating the same experiment with different participants. Still, the findings in a simulator may not be easily generalizable to real-life. Indeed, in most simulators, it is impossible to feel the body's acceleration as in a real car, and the fact that a crash will not cause real harm makes a driver drive faster and less careful.
Since there are only a small number of on-road driver's stress studies available, our research focuses on real environments in which the results on driver's stress and what influences the driver's stress may be more accurate. However, the data are more challenging to measure and analyze. To achieve our goals, we have used an instrumented Tazzari Zero car. In the Tazzari we installed cameras, an IMU, a GPS and a LIDAR. From these sensors, we obtain some information such as speed, acceleration, and jerk of the car in the three dimensions, number and category of obstacles detected, their position with respect to the car and, for the obstacles in front of the car, their time headway with respect to the car.
Moreover, most previous research in this field uses only questionnaires or interviews to understand how stressful a drive is. Indeed, measuring the passenger/driver discomfort in real-time is not an easy task since the passenger/driver, in our case, is not in a controlled environment, but it is in a highly dynamic environment. There are few examples of real-time discomfort acquisition in literature, such as using a handset control.

However, it is possible to use a handset only in an autonomous, not in a manual drive, and, in our study, we want to be free to acquire the stress level in both autonomous and manual drive. For these reasons, the real comfort of a passenger/driver is usually acquired after driving with questionnaires or interviews. Questionnaires and interviews are not a direct stress measure since they are subjective and are retrospective information on the passenger's stress, and, for this reason, they may have strong biases. In the platform developed, in order to acquire a stress metric, physiological sensors are used. In particular, in our car, we have installed sensors to monitor the physiological signals of the driver/passengers, such as the ECG, GSR and respiration. From these sensors we are able to extract stress indexes such as the heart rate, the ratio between low frequency and high frequency of the heart rate variability, the respiration rate and the tonic component and phasic component of the skin conductance.
Moreover, the physiological sensors are an objective, direct and continuous stress measure. This research has used questionnaires and interviews only to compare and validate the acquired data analysis results.
With the platform developed in this research, we increased the knowledge of passenger comfort in a real instrumented car from the physiological perspective overcoming previous research limits of using simulators and questionnaires. To validate the entire data acquisition and processing pipeline and ensure that all data acquired is useful for future and better analysis, we had done a correlation analysis on data collected during an acquisition campaign with 6 participants from 24 to 32 years old.
For correlation analysis, we used the K-nn algorithm. This algorithm allows studying the correlation between a time series and event sequences. Time series were retrieved from the stress indexes while the event sequences were extracted from the driving features. Furthermore, this algorithm provides information on the delay relationship between the time series and the event sequences and if the event sequences cause an increase or a decrease in the time series values. We tested the algorithm on the data acquired from our platform.
Results demonstrates that the phasic component extracted from the skin conductance and the respiration rate are the only stress indexes showing a high correlation with the driving features. Furthermore, results show that the driving factors that impact the driver's stress the most are the longitudinal jerk, the angular velocity on x-axis, and the linear acceleration on y-axis.

DISTRIBUTED PROCESSING AND OPTIMIZATION METHODS FOR QUERYING BIG GENOMIC DATASETS

Andrea Gulino - Supervisor: Stefano Ceri

In 1953, at the Cavendish Laboratory (University of Cambridge), Francis Crick and James Watson discovered the molecular structure of deoxyribonucleic acid (DNA), marking a milestone in the history of science and giving rise to modern molecular biology.
Almost 70 years later, DNA research plays a crucial role in many disciplines. In medicine, understanding DNA functions, unveiled the mechanisms underlying many diseases, e.g. cancer, improving our ability of early-diagnosis and treatment. Vaccines, like those that are currently being tested against SARS-CoV-2, depend on the knowledge we have on the virus genome.
DNA sequencing is the process of determining the order in which the four nucleotides (A,C,T,G) appear in a DNA molecule. The advent of rapid and less costly DNA sequencing methods, such as Next Generation Sequencing (NGS) techniques, has significantly accelerated biological and medical research and discovery. Biologists and bioinformaticians have been working over the years to build and maintain public data repositories of sequences and tools for managing and processing those data.

The ERC winning Data-Driven Genomic Computing (GeCo) project, developed at Politecnico di Milano, addressed several open problems in what is called tertiary analysis, mainly focused on making sense out of genomic data. By proposing the Genomic Data Model (GDM), it introduced a formalization for describing and encoding collections of genomic DNA region sets; with the GenoMetric Query Language (GMQL) it provided a rich set of operations that allow processing thousands of region sets at a time.
GMQL queries are processed on the Genomic Data Management System (GDMS), a high-throughput cloud-based system implemented on top of the cloud computing framework Apache Spark. The implementation of complex GMQL operations, such as JOIN and MAP, which are based on the relative distance between regions, leverages genome binning, a partitioning technique that enhances parallelism.

The thesis presents some of the research work done within the Genomic Computing group. The contribution of the thesis is threefold and relates to three areas:

- System design: it introduces new and re-engineered cloud-based systems for supporting genomic data processing and analysis.
- Performance modeling: it proposed methods for optimizing such systems and modelling their performance.
- Distributed algorithms: it describes efficient distributed algorithms for performing complex operations on genomic regions and mutations.

The first contribution concerns performance modeling and focuses on the optimization of the most critical GMQL operations. The binning algorithms used within the GDMS make the implementations of GMQL's most critical operators highly parallel. Each operation, instead of being computed along the whole chromosome, is computed within each bin. High parallelism is achieved since every bin can be processed in parallel independently from the others. In particular, the thesis solves the problem of determining the optimal bin size, using complex analytical models that consider the input data's characteristics, the operation parameters, and the execution environment. The proposed models are validated through a rich set of experiments that show the model's accuracy and sensitivity to the variation of query parameters, using both synthetic and real datasets. Several examples show that the execution time doubles with a relatively small error in the choice of bin size, and a big error may even lead to cases where execution cannot be completed, as resources are exhausted. We also discuss the optimization of sequences of operations, showing that binning chaining or reuse can introduce significant savings.

The second contribution concerns the re-engineering of the GDMS architecture for supporting federated GMQL queries. Federated GMQL is an extension of GMQL that supports the sharing of genomic datasets and computational capabilities across a so-called federation. This solution was designed to answer the increasing need of laboratories and health organizations for sharing NGS data resources, and easily accessing and processing comprehensively shared genomic data.
A GMQL federation is made up of several GMQL instances (cloud computing sites running GMQL) that interact with each other to share data and jointly process GMQL queries. The extension required a significant re-engineering of the GDMS architecture, which now comprises a centralized -- and possibly replicated -- component, called *Name Server,* used by instances to advertise the availability of shared data and computational resources. The GMQL language was instead extended with simple data sharing instructions that allow referencing external datasets and set an execution policy, prescribing which instance of the federation should be in charge of processing a given instruction of the query. The thesis describes the main novelties introduced with Federated GMQL, focusing on the architectural aspects that enabled the extension of the language. Specifically, the Name Server and the GDMS extensions that support the authenticated communication between instances, including substantial changes to the GMQL Repository, the Execution Manager and the REST API. Those components are then part of a novel architecture that allows the cooperation of multiple geographically dispersed Apache Spark instances (clusters) for the execution of a (conceptually) individual application.

The third contribution considers a broader set of Apache Spark applications and focuses on performance prediction. Specifically, the thesis targets workflow applications and proposes a modular performance estimation approach that relies on Machine Learning (ML) and can be extended to any data-driven workflow (or DAG-based) application implemented on top of Spark (e.g. workflows built with Azure Databricks or Apache Airflow). Although the proposed approach does not require any specific knowledge on the workflow application on which it is applied, we show the advantage of pushing into the model context-specific knowledge and how ML and analytical models can co-exist within a performance estimation pipeline. The proposed approach is applicable to ach is queries, which are mapped into operation DAGs and implemented on Spark; in particular, the operation and query time predictions are used to improve the performance of classes of execution plans for federated queries.

The last contribution of the thesis concerns the work done in the context of mutation analysis. Specifically, it describes MutViz, a complex visualization tool for the analysis of somatic mutations in arbitrary user-provided regions of the DNA. Born as a simple visual interface, MutViz has quickly become a complex tool for identifying patterns of somatic single-nucleotide mutations observed in human cancers. Specifically, the tool matches the user-provided region sets against a database containing more than 65M mutations associated with 35 different tumor types. The computation is handled, depending on the type of analysis, by querying a PostgreSQL database, enabled for parallelism, or by distributed algorithms, specifically designed for this tool. Despite being completely independent on GMQL, the system architecture and the distributed algorithms implemented in MutViz leverage the experience gathered working on the Genomic Data Management System. An ongoing work re-adapts Mutviz to support the visual analysis of SARS-CoV-2 variants.

The solutions proposed in the thesis cover a large spectrum of optimization problems in the area of distributed systems, and their application to the domain of genomics has produced several efficient methods and systems dedicated to tertiary data analysis. In turn, tertiary data analysis is gaining more and more relevance, also thanks to the availability of huge collections of well-curated human genomic datasets and to the increased relevance of their application for understanding many open problems in personalized health care.

READOUT ELECTRONICS FOR HIGH-RATE HIGH-RESOLUTION ENERGY-DISPERSIVE X-RAY DETECTION SYSTEM

Idham Hafizh - Supervisor: Carlo Ettore Fiorini

In the past few years, detection module for X-ray spectroscopy based on Silicon Drift Detector (SDD) with a concentric anode, coupled with CMOS monolithic Charge Sensitive Amplifier (CSA), has shown to be very competitive in terms of performances and ease of fabrication. CMOS-based CSA is characterized by high transconductance, when coupled to SDD that has relatively small anode capacitance, it results in excellent noise performance and thus, good energy resolution. The combination is also known to be compatible with high-rate operation, provided several conditions such as small detector unit dimension, low-temperature operation, and fast detector pulse processing system. High-rate operation is essential to some applications, for example, in energy-dispersive detector in synchrotron beamlines performing experiments such as XRF, XAS, and micro XRF. The requirement is driven by the ever-increasing flux of x-ray incident beams in upgraded synchrotron facilities, leading to higher fluorescence flux available for the detector, potentially improving the quality of the acquired data and/or reduce significantly experiment time. The first part of the thesis focuses on this question: can we develop an X-ray spectrometer, adapted for experiments in synchrotron beamline, that can provide good energy resolution and high throughput capability, with immunity to electrical disturbances typically present in the beamlines? The conclusions for these questions have been achieved through the milestones in the ARDESIA project. ARDESIA (Array of Detectors for Synchrotron Radiation Applications), is an SDD-based, 4-channel X-ray spectrometer, optimized for synchrotron experiments such as X-ray fluorescence (XRF) and X-ray absorption fine structure (XAFS). Its detection module, has been established previously, is able to achieve high-count rate (>1Mcps/channel) and high-resolution (≈125eV of FHWM Mn-Kα line at optimum shaping time, <200eV at short shaping times) X-ray fluorescence detection. During the course of my PhD study, the complete instrument, including internal and auxiliary electronics, has been assembled and tested successfully in both laboratory and synchrotron beamline settings. In DAΦNE-Light DXR1 beamline, satisfactory results of XRF and XANES acquisition have been achieved in soft X-ray regime. In LISA BM-08 ESRF beamline, measurements including long-duration XAFS measurements and XAFS measurement on trace elements confirm the qualification and performance of the instrument, in terms of energy resolution, throughput capability, immunity against external disturbances, and stability over time. High-rate spectroscopy can only be realized with SDD-CSA based detection system, only if the following stages provide fast pulse processing. In a system with a limited number of detection channels (<10 channels), Digital Pulse Processor (DPP) is the most commonly adopted pulse processing solution, thanks to its satisfactory spectroscopic performance, high throughput, and robustness of digital system. On the other hand, when high-density multichannel detection is needed, as in future synchrotron-based experiments, or when the power and/or area constraints apply, Analog Pulse Processor (APP), implemented in an integrated circuit, is still an attractive solution, since it can provide compact and cost-effective solution compared to DPPs. ASIC-based APPs, if exist, with relatively high-throughput capability, appear to be a desirable solution to equip a compact instrument with high-density multichannel readout electronics (50-100 channels) within a reasonable cost and power budget. The second part of the thesis raises these questions: is it conceptually feasible to create an analog pulse processor with throughput capability above 1Mcps output count rate with satisfactory energy resolution and spectrum quality? If yes, can we implement it in ASIC and demonstrate experimentally such performance? In my PhD study, the questions are first tackled by theoretical study of a fast APP solution. Systematic assessment of various shaping amplifiers has been conducted, considering series noise, ballistic deficit immunity, and pile up immunity. Estimation and validation on its ballistic deficit effect to additional spectrum broadening are also explored. The study is concluded with analysis of analog-based pile up rejection (PUR) strategies and maximum throughput estimation taking into account several parameters on analog processing channel. The assessment provides some relevant insights to define the specification of TERA (Throughput Enhanced Readout ASIC), an APP implemented as an ASIC with a purpose to demonstrate experimentally the results obtained in the study.
The ASIC development has been started from its predecessor, SFERA (SDD Front-End Readout ASIC), designed in the same 0.35µm technology node. Each pulse processing channel includes a 7th-order semi-Gaussian shaping amplifier with controllable shaping times and full scale ranges, followed by a peak stretcher and a switched-capacitor analog memory. Each channel is also equipped with a dedicated peak detector and a novel pile up rejection (PUR) logic. Each pair of channels can be optionally digitized by a 12-bit on-chip SAR ADC. The architecture enables to achieve high throughput and satisfactory energy resolution. In fact, in ^{55}Fe spectroscopy measurements, when using the shortest shaping pulse width of 200 ns, 2.5 times shorter than the smallest shaping time implemented in SFERA, a FWHM Mn-Kα line of 159.4 eV was obtained at low rates (10 kcps). This energy resolution can be achieved thanks to the minimization of the effect of the ballistic deficit, achieved by optimizing the SDD detector in terms of optimum biasing, low operating temperature (-37°C), and small size (1 mm diameter). At 1.61 Mcps/channel input rate, an output count rate of 1.09 Mcps/channel and a resolution of 205.1 eV were achieved with a 4mm-diameter SDD and optimum PUR settings. Such high-rate performances are, to our knowledge, the best ones reported so far for an APP, based on a spectroscopy-grade ASIC, and very close to those achievable by a standard DPP. Therefore, TERA can represent an attractive, compact, and scalable pulse processing solution for high-rate multichannel energy-dispersive X-ray detection systems.

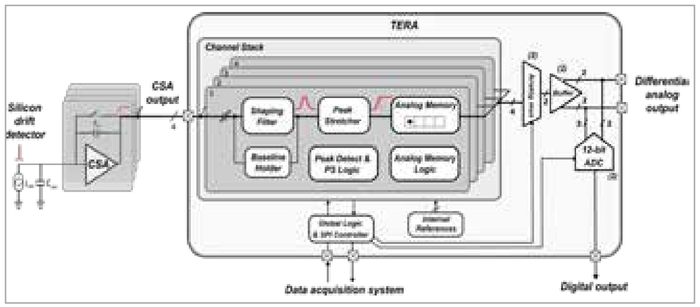

Fig. 1

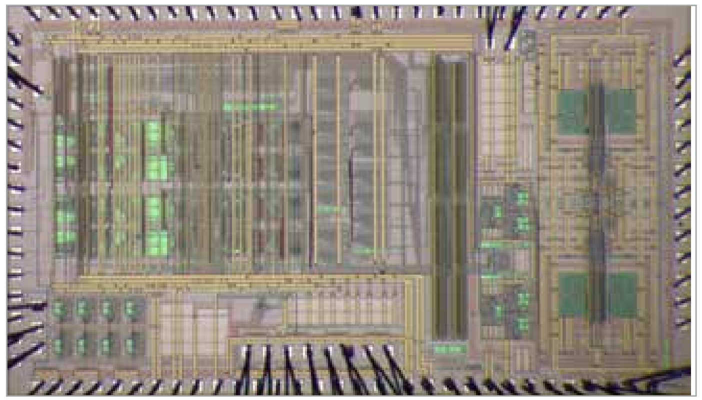

Fig. 2

MULTI-CORE FREQUENCY SYNTHESIZERS FOR MM-WAVE COMMUNICATIONS

Saleh Karman - Supervisor: Salvatore Levantino

This thesis discusses the analysis, design and implementation of three mm-Wave frequency synthesizers. Initially, a brief introduction to phase-locked loops (PLLs) is given, introducing the phase domain model of digital and analog synthesizers, the fundamental metrics that quantify their performances (phase-noise, integral jitter, EVM) and the link between them. Two applications are considered as targets for the designs presented in the thesis: the IEEE802.11ad standard and E-band mm-Wave wireless backhaul. An overview of PLL architectures found in literature and commercial products is briefly discussed, their pros and cons are qualitatively stated to motivate the adoption of a two-step synthesis for the presented architectures.
Chapter two presents a digital fractional-N PLL operating at 30GHz. Aim of this design is to compensate the existing FoM gap between mm-Wave PLLs and other sub-6GHz RF implementations. To this end i) a digital-subsampling architecture is adopted ii) the programmable divider on the feedback path is substituted by a novel low-power injection locking divider iii) a novel digital-to-time-converter (DTC) range reduction technique is proposed to further reduce power avoiding the need for an RF multiplexer. The PLL, followed by a multiplier by two, covers with margin three channels of the IEEE802.11ad standard granting an EVM of -21.5dB (197fs-rms) and -25.5dB (135fs-rms) with and without a receiver tracking filter respectively. The achieved EVM allows the employment of QPSK and 16QAM modulations, with ~5dB margin, in the two cases. The frequency synthesizer is implemented in a 65nm CMOS technology, occupies an area of 0.55mm^2 and consumes 35mW resulting in a FoM = -238.6dB.
The same chapter deals with the topic of quantization noise reduction in digital synthesizers based on lead-lag (or bang-bang) phase-detectors (BBPD). Lead-lag PDs are known for they minimalistic design and power consumption, unfortunately, due to their very nature they introduce quantization noise that raises the PLL's in-band noise level. To reduce the latter contribution on the phase-spectrum, multi-bit phase-detectors have been recently proposed in literature, however, the impact of this solution on the PLL's jitter performances has not been addressed in literature. The presented study quantifies the trade-off between the number of threshold and jitter improvement. Closed form expression for the output jitter as a function of the number of thresholds are given to determine the best compromise between complexity and noise performances.
Chapters 3 and 4 describe the implementation of two PLLs targeted for operation in a mm-Wave wireless backhaul transceiver. The test-chips have been realized within the European project TARANTO. To satisfy the stringent noise requirements demanded by E-band carriers and exploit the possibility to realize wide tracking loop bandwidth (up to 700kHz) multi-core techniques are explored to lower out-of-band noise (also in-band noise in Ch. 4) below the limitations set by the individual blocks (e.g. oscillator, DTC/reference buffer) trading power for improved noise performances.
Chapter three presents a digital fractional-N phase-locked loop implementing a dual core digitally controlled oscillator (DCO) operating at 12.3GHz. To reduce the out-of-band noise, two class-B double-transconductor oscillators magnetically coupled by means of their tail filters are embedded in a digital PLL. The reason for this choice was to avoid placing in parallel to the main tanks lossy and noisy components, moreover, this topology creates a virtual ground at the power supply nodes which makes the tail impedance independent from the supply network parasitics simplifying the coupling-network design. The presented PLL, followed by a by six frequency multiplier covers the lower E-band frequency range with margins. The achieved jitter is 264fs-rms while the spot noise of the DCO at 1MHz offset from a 12.5GHz carrier is -117dBc/Hz.

The design of the previous test-chip, from a practical standpoint, proved very difficult and brought out some issues related to this approach i) routing of the cumbersome digital interconnections was very challenging ii) the DCO buffers, placed outside the dual core DCO structure for better biasing, received the signals through long interconnections which impaired the tanks' quality factor iii) the floorplanning proved to be very complex and the optimization of the overall system penalized that of the individual cores. These considerations motivated the research of an alternative approach to lower the out-of-band phase-noise, hence, architectures based on "multi-core PLLs" are proposed as a promising alternative to multi-core oscillator based ones.
Chapter 4 is divided in two parts. The first one discusses multi-core PLLs based on coupled PLLs (CPLLs, i.e. PLLs coupled by means of an additional phase-detector) and introduces a novel topology of CPLLs, which is able to enforce the coupling mechanism also on the components located on the reference path. A detailed analysis of the noise transfer-functions is presented, and the conditions under which noise improvement is guaranteed are clarified and verified against simulations.
The second part, instead, discusses multi-core PLLs with output signal combination. By summing the outputs of two PLLs, 3dB phase-noise reduction are obtained. The proposed architecture is superior to the previous multi-core-oscillators-based approach for different reasons i) each oscillator core can be designed and optimized individually ii) each oscillator can be easily integrated within its PLL core with no need for lengthy and cumbersome interconnections iii) concurrent multi-mode oscillations, theoretically possible in multi-core oscillators, are avoided iv) there is no need to place switches and other lossy components in parallel to the oscillators tank, lowering the overall quality factor and increasing power consumption v) power scalability is preserved.
To validate this architecture, a digital fractional-N dual-core PLL operating at 20GHz is presented. This synthesizer features a i) a dual-core power-combining architecture ii) an offset calibration loop, which aligns the output waveforms to ensure constructive sum of the outputs iii) a class-B DCO with efficient second harmonic extraction. The 20GHz PLL, followed by a by four frequency multiplier, covers the upper and part of the lower E-band frequency range. The EVM calculated for an 80GHz carrier (two active cores) is -21.2dB (174.2fs-rms) and -40.4dB (19fs-rms), respectively, without and with a carrier tracking loop. Guaranteeing, with ~7dB margin, the employment of QPSK and 256QAM modulations. The PLL, implemented in a 55nm BiCMOS technology occupies a core area of 0.65mm^2. The power consumption is 18mW and 42.6mW respectively with one and two active cores. The PLL FoM is -241dB and -238.9dB in the two cases.
Based on the achieved results, the dual-core PLL matched the performances of the best quad-core oscillator found in literature, consuming the same power and with minor area overhead, confirming the superiority of the proposed approach.

SET MEMBERSHIP IDENTIFICATION AND FILTERING OF LINEAR SYSTEMS WITH GUARANTEED ACCURACY

Marco Lauricella - Supervisor: Lorenzo Mario Fagiano

Robust system identification addresses the problem of identifying models of unknown systems, and of providing a measure of the resulting model uncertainty, which is of great interest for several applications concerned with robust estimation and control, like robust Model Predictive Control, resource scheduling, smart grid robust optimal operations, energy usage and production planning. An interesting robust identification framework that relies only on deterministic hypotheses goes under the name of Set Membership identification. In the Set Membership identification theory, the uncertainty is described by additive disturbance signals which are only known to be bounded, where the bound is often unknown. Thus, all information about the system is provided by the available measurements and could then be summarized by sets of system state, or sets of model parameters or variables estimate, that are consisted both with the available data, and the constraints on the estimated bound on the identification error. The appeal of Set Membership lies in its directness, simplicity, and in its need for few assumptions, compared to probabilistic approaches.
In this doctoral dissertation, we address the problems of model identification and output filtering, for the case of linear time-invariant (LTI) systems, resorting to the Set Membership framework. Our goal is to derive tight uncertainty intervals for the unknown system output, and to provide an accurate filtering algorithm, where the accuracy is measured by means of guaranteed bounds. Moreover, we aim at developing an identification algorithm to obtain one-step-ahead models of LTI systems with guaranteed worst-case error bounds in simulation, enforcing at the same time a stability property on the resulting models. Here, we present new theoretical results pertaining to Set Membership identification, for the considered case of LTI systems with bounded additive measurement noise, where, contrary to most existing works, the bound is a priori unknown.

Noise bound estimation. In the Set Membership literature, the noise bound is often assumed to be known a priori, which can be a limiting assumption, or it is considered as a tuning variable, where its choice plays a direct role on the resulting accuracy bounds, which could be conservative, or could be falsified by future data. However, in many relevant applications, the noise bound is not known, and a guess on its amplitude can be difficult to verify, and improper initial assumptions on the disturbance bound could have strong effects on the desired estimate, leading to inconsistent estimators. In this thesis, new theoretical results pertaining to Set Membership identification are presented, allowing one to obtain a convergent estimate of the noise bound from data, for the case of LTI systems affected by bounded additive measurement noise, where the bound is unknown. The presented theoretical results are derived by addressing the problem of the identification of models of LTI systems having guaranteed simulation accuracy. To achieve these results, we resort to multistep prediction models and we analyze the properties of the corresponding multistep prediction error bound. The main conceptual innovation is that we analyze not only each value of the error bound separately, but also its course as a function of the prediction horizon p. We show that, under mild assumptions, the multistep prediction error bound is a convergent quantity, and its behavior over the prediction horizon, as well as the value to which it converges, are directly linked to the noise bound amplitude, and to some system properties, in particular its order, and its dominant decay rate.

Model stability. In LTI system identification theory, bounded input bounded output (BIBO) stability is perhaps the most common assumption. However, in the literature only few contributions address the problem of how to make use of this a priori information on the system stability to enforce the convergence of the model to be identified, since formal inclusion of stability constraints makes the estimation problem difficult to be solved. Here, we propose a novel approach to enforce model stability directly in the identification problem. To do so, we introduce new results allowing one to estimate the worst-case simulation error bounds for any simulation horizon, up to infinity, and to obtain a condition ensuring its convergence, while analyzing its converging behavior. We derive a clear link between the obtained infinite-horizon bound, and the estimated noise bounds, model order, system decay rate, and prediction horizon used in the model identification routine. Then, we provide a refined version of the classical Feasible Parameter Set (FPS) formulation, and we take advantage of this new concept of FPS to enforce a converging behavior on the parameters to be identified, by means of polynomial constraints included in the identification procedure. We finally prove that the models derived with the proposed procedure are guaranteed to be asymptotically stable, a property that is non-trivial to enforce during the identification phase.

Output filtering. The developed theoretical results are then employed to address the problem of output filtering, allowing one to improve over existing results in terms of filtering accuracy, measured by tight guaranteed bounds on the unknown output uncertainty region. The problem of process variables filtering has been widely studied, and the majority of the related works mainly focuses on the case of stochastic disturbances with known probabilistic properties. Since such probabilistic assumptions may be sometimes not realistic, or difficult to validate in real applications, optimality properties and the boundedness of the estimation error are in general not guaranteed. Resorting to the theoretical results derived in this thesis, we propose a novel filtering algorithm based on the combined use of independent multistep prediction models, that achieves tight accuracy bounds. The main idea is to define the true system output uncertainty interval as the intersection of the local accuracy regions of all the p-steps ahead predictions of the current output sample, obtained using independent multistep predictors. Then, the filtered output of the unknown system is given by the center of the resulting uncertainty interval, which is locally optimal in a worst-case error sense, attaining the minimal achievable accuracy bounds under the Set Membership framework.

Validation of the proposed approaches. Finally, we test the proposed prediction and filtering procedures both in a numerical example, where the true quantities are known and the developed methods can be evaluated in full, and in a real-world experimental application, pertaining to the roll rate dynamics of an autonomous glider, where their performance are compared to those of standard identification approaches, and of a robust Kalman filter based on the Riccati equation approach. Moreover, an interesting case study pertaining to the problem of forecasting non-residential building energy consumption is proposed to conclude this thesis, where, by introducing an artificial input signal able to capture the seasonal dynamics of the load time series, we are able to use a linear one-step-ahead model structure, which achieves a forecasting accuracy comparable to that of nonlinear algorithms like Neural Networks, while retaining the capability to provide guaranteed bounds on the accuracy, resulting by the application of the Set Membership identification approach presented in the first part of this thesis.

MECHANISMS AND ALGORITHMS FOR DYNAMIC RESOURCE SHARING IN NETWORK SLICING

Alessandro Lieto - Supervisor: Antonio Capone

The digital transformation taking place in industries and organizations is generating a large number of new use cases for advanced communication technologies. The huge economic interests of the many industrial sectors involved in this transformation meets the need of the telecommunications players to enlarge their business relationships and welcome new players in their ecosystem. *Network slicing* will be the key technology to offer efficient and reliable communication services to this wide heterogeneity of applications and use cases.
Despite the end-to-end nature of slices, in this thesis we focus on the management of radio network resources, which will be the real bottleneck of the network chain, given the scarce nature of wireless resources.
In particular, we address a specific network slicing scenario, where slice tenants retain the control and the management of their slices, e.g., by dynamic quality of service management (increase/reduce throughput and/or latency), traffic shaping or admission control policies. Indeed, we strongly believe there is a high dependency between the aforementioned configurations and the performance of slices, and, therefore, how they impact tenants' business models and overall profits. However, such flexibility unavoidably affects the network management and requires dynamic orchestration capabilities in order to handle those dynamics. Although we do not exclude cases where service level agreements (SLAs) strictly enforce limitations on the online configurable parameters for slices, we consider such flexibility in our work, i.e., that slice tenants may actively intervene in the management of their slices by real-time tuning a set of slice-parameters and requirements. In our vision, it will allow network operators to enable new business models, based on dynamic pricing, and encourage multiple entities, from different vertical sectors, to accept those solutions, giving them the full control of their actions, depending on their specific applications and business cases.

We define the conceptual solution of *dynamic slicing* and analyze the trade-off between resource sharing and degree of *isolation* for network slices. We first analyze slice-aware radio resource scheduling solutions based on customized utility functions. We propose a dynamic sharing optimization solution, where slices dynamically share the radio resources as in a common pool and the scheduler allocates resources to the users of each slice such as to maximize the utility functions of each slice. We enable customization and differentiation of slice utility functions, by translating their service requirements and expected quality in a measurable quantity that defines the level of satisfaction of the tenants for the quality of service provided for their slices. Furthermore, we compare our dynamic slicing solution with an optimal static slicing solution that enforces physical slice isolation, by optimally reserving portion of the spectrum to maximize the utility functions of the slices. We show that dynamic slicing outperforms static slicing even in case of critical applications with stringent service level agreement targets and define a metric to quantify these gains.

However, in case of congestion, it might not be possible to satisfy all the requirements at best. By considering the dynamic resource requirements and the conflicting and heterogeneous quality of service of the slices, we do not expect slice tenants to accept any decision taken by a central entity about the possible downgrading of their slices, when network conditions cannot guarantee the requirements for all of them. In this case, the diverse business interests of the tenants must be taken into account and be modeled as multiple real decision makers that compete on network resources and their achievable performance. We model this competition through the definition of a dynamic marketplace, where the infrastructure provider charges the tenants according to resource supply and demand. In the marketplace, the tenants take strategic decisions about when and where purchasing resources, depending on the trade-off between the quality of service they want to provide to their users and the current price charged by the provider. We develop an automated negotiation mechanism, where tenants locally (over cell or areas) trade resources in the medium-short term. We develop our scheme through game theory, analyze the properties of the game in terms of convergence to a Nash equilibrium and discuss the quality of those equilibria. In addition, we propose a heuristic algorithm that guarantees fast convergence to a single Nash equilibrium for any number of slices in the system and compare those equilibria with the one obtained with state-of-the-art algorithms.

Furthermore, we integrate the dynamic renegotiation schemes enabled by the slicing marketplace with slice-aware schedulers that implement slicing policies in terms of average resource share guarantees. We introduce a Slicing Management Framework that processes the high-level requirements of the slices defined by tenants at time of slice configuration and instantiation and map them into specific network domain parameters, adapting the resource demand based on their performance requirements and economic interests. We implement this framework on a 3GPP-calibrated system level simulator and define the interfaces for the exchange of information needed to run it in a real system.

Finally, we validate the framework on emulated per-slice data traffic traces. We test the capabilities of our approach in a realistic scenario with traffic fluctuations over hours and days to offer a general overview of system practicability. We also discuss the implementation of an online platform where tenants can monitor real-time their key performance indicators, both in terms of affordable performance and prices, and infrastructure providers to assess their pricing strategies and the ability to satisfy tenants' requirements.
This capability opens new scenarios where artificial intelligence can play a significant role to automate decisions and add values for all the stakeholders based on the knowledge of system behavior.

We leave space for future extensions of this work, where the implementation of Machine Learning and Artificial Intelligence can help the automation processes described in this thesis. We also highlight that the trend of dynamic renegotiation will become even more significant in next generation wireless systems, as nowadays it is of crucial importance in cloud computing systems and other vertical markets, like energy and transportation systems.

ENHANCING HUMAN-ROBOT COLLABORATION FOR FLEXIBLE MANUFACTURING IN INDUSTRY 4.0

Riccardo Maderna - Supervisor: Paolo Rocco

The present-day global economy is rapidly transitioning from mass production to mass customization, which is characterised by high-mix low-volume production and frequent changes of products and processes. The rigidity of the long-established paradigm for industrial automation and robotics does not fit the Industry 4.0 scenario. Instead, advances in human-robot collaboration are playing a key role in providing companies with an adaptable and powerful tool capable of enhancing efficiency and flexibility of manufacturing processes. Collaborative robotics removes the physical separation and allows robots and humans to share the same environment and work closely with each other. Robots can relieve workers from fatiguing and alienating tasks, whereas humans leverage their superior cognitive and manipulative skills to perform operations that are difficult to automate or that require high-level decision-making.
The thesis provides methods and tools to control the production of a flexible manufacturing cell, where multiple humans and robots cooperate to assemble different products. Flexible manufacturing is characterised by time-varying mix and frequent production changes, which reflect on modifications in the workspace layout and the operations to be performed. In this scenario, it is crucial to reduce set-up times as much as possible, favouring plug-and-play and easily reconfigurable approaches. Also, versatile scheduling algorithms are needed to organise the complex workflow and fully exploit the available flexibility, ensuring the optimal use of resources and the smart management of unexpected events. This is especially true when human behaviour is considered, which is difficult to fully define a priori and changes over time, e.g. due to training and fatigue. Understanding and predicting the human activity allows for better coordination of the agents and fosters correct task allocation and scheduling.
The first part of the thesis discusses the problem of monitoring the current human activity. Human tasks are complex operations that can be accomplished following several sequences of low-level actions, which compose different variants of the same task. Also, the possibility of execution errors and small pauses is considered. Two strategies have been proposed. The first one learns a model of the human task from demonstrations relying on the automatic segmentation of motion trajectories. Then, motion segments train a classifier to identify the variant being performed by the operator at run-time. The identification of low-level actions enhances collaboration in two ways. First, the human worker is notified in case of errors, so that he/she can immediately implement corrective actions. Second, early classification allows for a prompter prediction of the future evolution of the task and, consequently, better planning of the robot actions.
The second strategy has the primary objective of estimating the expected duration of the ongoing human activity. The proposed method is based on a modified version of the Dynamic Time Warping algorithm and does not require any training phase. Instead, it learns online from previous repetitions of the same activity and automatically recognises previously unseen variants, which are added to the activity model. Experimental results show that the algorithm provides accurate prediction also for long and complex operations and that the presence of many variants of the task does not degrade performance.
The second part of the thesis describes the overall control system architecture (Fig. 1) composed of a Digital Twin of the process, a dynamic scheduler, and a human monitoring unit. The proposed strategy endows the system with strong adaptation and learning capabilities, which are crucial to quickly react to changes in the workspace layout and the production. The Digital Twin, based on Petri Nets, tracks the status of the process in real-time based on information coming from the human monitoring unit and the robot controllers. The model is very general and able to describe a wide variety of cell configurations. Also, it can be generated automatically from an intuitive definition of the assembly tasks and easily updated to keep up with changes in the production.
The dynamic scheduling algorithm determines the optimal instructions for the humans and the robots working in the cell. The future evolution of the process is simulated starting from the current state of the Digital Twin. The optimal plan is selected to favour productivity and tracking of the target mix. Experiments carried out on a complex assembly task (Fig. 2) showed that the scheduler can optimally plan the task of the agents also in case of a time-varying production mix and the occurrence of faults. Also, a novel visuo-haptic interface to give instructions to human operators is introduced, which can be used in complex human-robot collaboration scenarios and allows expert users to improve performance.
Finally, a similar scheduling principle is applied to a collaborative kitting task to enhance both ergonomics and productivity. For this purpose, a method to associate an ergonomic measure to each picking action is also presented. The scheduler is formalised as a Mixed-Integer Linear Programming optimization and ensures the coordination of the agents to prevent collisions.

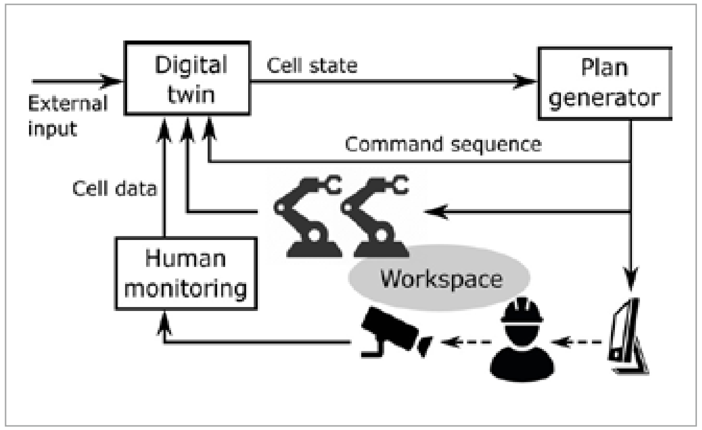

Fig. 1 - Control system architecture.

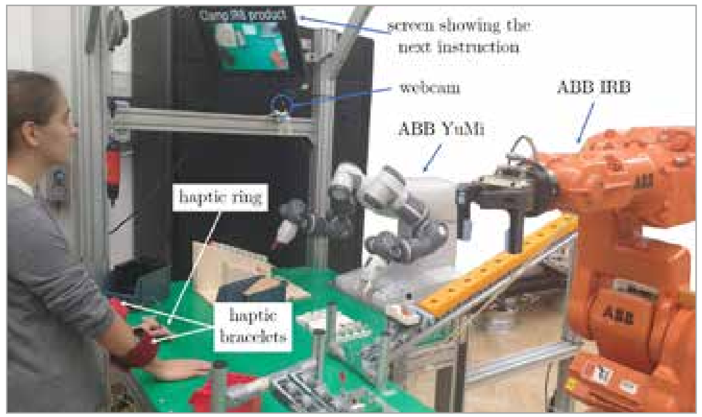

Fig. 2 - Experimental setup.

FROM THEORETICAL TO REAL WORLD CRYPTOGRAPHY: TOWARDS PRACTICAL PRIVACY-PRESERVING OUTSOURCED COMPUTATION AND ACCURATE PARSING OF DIGITAL CERTIFICATES

Nicholas Mainardi - Supervisor: Gerardo Pelosi

Cryptographic primitives are fundamental building blocks for the security of a system. Nonetheless, the security guarantees of cryptographic primitives do not depend only on their theoretical soundness, but also on their proper adoption in real-world applications. This process is often challenging, since there are many different issues that may compromise the security guarantees of cryptographic components, such as implementation flaws, side-channel attacks and misuses of the primitive. In this work, we focus on two relevant challenges that arise in the adoption of two cryptographic primitives in real-world applications: the unpractical performance overhead exhibited by cryptographic solutions for privacy-preserving outsourced computation; the security vulnerabilities stemming from the improper parsing of digital certificates.

Privacy-preserving outsourced computation techniques allow to offload a computation to an untrusted server while retaining the confidentiality of the data involved in the computation. Fully Homomorphic Encryption (FHE) is one of the most suitable techniques to perform privacy-preserving outsourced computation, as it allows arbitrary computation directly on encrypted data;

nonetheless, its adoption in real-world applications is currently hindered by its prohibitive

performance overhead. In this work, we aim at evaluating several strategies to reduce such unpractical performance overhead exhibited by FHE schemes. First, we investigate the security guarantees of existing noise-free FHE schemes, which are appealing due to their higher efficiency than common noisy FHE schemes. Our investigations allowed us to design two novel attack techniques against FHE schemes, which amplify the impact of existing vulnerabilities in the target FHE scheme by relying on its homomorphic capabilities, in turn allowing to fully subvert the security guarantees of the scheme even in a ciphertext-only scenario. Our attacks completely break the only two existing noise-free FHE schemes that were claimed to be secure in a ciphertext-only scenario, allowing the adversary to efficiently recover their secret key. Our attacks show the difficulty of designing a secure noise-free FHE scheme, thus leaving the existence of a secure noise-free FHE scheme as an interesting open problem.

Given the unavailability of secure and efficient noise-free FHE schemes, we evaluate two alternative strategies to reduce the performance overhead of privacy-preserving outsourced computation: employing Partial Homomorphic Encryption (PHE) schemes, which restrict the set of computations that can be performed on ciphertexts to gain some efficiency over FHE ones; relying on trusted hardware, such as the Intel SGX technology, which allows to securely run a computation over a machine whose software is completely compromised by a privileged adversary. We show the effectiveness of these two approaches by applying them to the design of two Privacy-Preserving Substring Search (PPSS) protocols, which allow to outsource the look-up of strings in outsourced documents while retaining the search and access pattern privacy of the queries as well as the confidentiality of the outsourced documents. The first protocol exhibit an extremely low communication cost and it enables queries from multiple simultaneous users, without requiring any synchronization among them; in addition, it is the first PPSS protocol that allows to perform privacy-preserving pattern matching queries, that is queries for strings containing wildcard characters. The second PPSS protocol, based on SGX, achieves an optimal communication cost for substring search queries and it is secure even against a malicious adversary who can arbitrarily tamper with the computation and/or the outsourced data; nonetheless, neither multiple simultaneous queries from different users, nor pattern matching queries are available in this solution. In conclusion, both our protocols show an extremely low communication cost and a practical response time on real-world use cases, highlighting the effectiveness of the two proposed approaches in reducing the performance overhead for privacy-preserving outsourced computation.

Digital certificates are widely employed in secure communication protocols to ensure the authenticity of the binding between a public key and its owner. Despite their crucial role, existing parsers for these certificates still exhibits a significant lack of accuracy, which has already been exploited to conduct several powerful attacks against the protocols relying on these certificates. For X.509 digital certificates, which are widely employed in the ubiquitous TLS protocol for secure communication, automatically generating a parser from a grammar specification has already turned out to be really effective in improving the parsing accuracy; nonetheless, the formal grammar for X.509 digital certificates is extremely complex, and thus hardly usable and maintainable in real-world implementations. To overcome this issue, in this work we propose a novel format for X.509 digital certificates; our format, while retaining the same expressiveness of existing X.509 certificates, can be described by a simple regular grammar, in turn allowing the automatic generation of a parser exhibiting optimal time and space complexities. Digital certificates in our format have a fixed preamble that allows legacy implementations to immediately recognize them as X.509 certificates with an unknown version, hereby avoiding possible unpredictable errors due to the attempt at parsing them as certificates in the current X.509 format. This feature enables a gradual roll-in of our new certificates without disrupting legacy implementations, which is a crucial factor to foster the adoption of our format in the X.509 infrastructure, as it removes the need to simultaneously replace all the existing certificates with new ones and to immediately upgrade all the parsing implementations.

Given the high accuracy showed by automatically generated parsers for X.509 digital certificates, in this work we also take a first step towards the automatic generation of a parser for certificates and messages in the OpenPGP protocol, which is the distributed and peer-to-peer solution opposed to the centralized and hierarchical approach employed in the public key infrastructure relying on X.509 certificates. In particular, we analyze the format of OpenPGP certificates and messages, showing that such format can be described by a deterministic context-free grammar (DCF), from which an efficient parser can be automatically derived. Nonetheless, we show that such grammar requires a prohibitively high number of rules that makes it unusable in practice to automatically generate a parser. To overcome this difficulty, we also discuss possible modifications to the OpenPGP format that allows to describe it with a simple DCF grammar from which an automatically generated parser can be practically derived. Furthermore, we outline several attacks that rely on different flaws that we identify in our analysis of the OpenPGP format, assessing their effectiveness against existing implementations of the OpenPGP protocol. The experimental evaluation reveals that fortunately such implementations are not vulnerable to our attacks, mostly because of wise choices of the developers in dealing with poorly or ill specified portions of the OpenPGP standard, hereby showing that the developers had a clear understanding of the security impact of their choices on their implementation of the protocol. Nonetheless, this pleasant outcome does not imply that our attacks are harmless in general; thus, we also propose some simple modifications to the OpenPGP format specification that allows to avoid the identified attacks.

FLASH MEMORY TECHNOLOGIES: EVOLUTION TOWARDS 3-D ARCHITECTURES AND APPLICATION TO NEUROMORPHIC COMPUTING

Gerardo Malavena - Supervisor: Christian Monzio Compagnoni

Since their first introduction, Flash memory technologies have been the object of an uninterrupted scaling process that allowed to increase their bit-storage density and become the most successful solution in the non-volatile memory market. However, shrinking the single cell size up to decananometer dimensions has faced some fundamental issues related not only to the manufacturing process itself, but also to inherent limitations of the physical mechanisms involved in the device operation that have undermined the memory array reliability. In this regard, as they target distinct applications, NAND and NOR Flash technologies have been following different scaling roadmaps over the years.

In particular, as NAND Flash technology aims to provide a low cost solution to store a large amount of data, high integration density and operating data throughput are undeniable features to pursue during the technology development. For this reason, the improvement of NAND Flash memory arrays has been the real driving force behind the efforts devoted to push the technology scaling to its physical limits, and NAND Flash memory cells have reached the smallest feasible feature size equal to 14 nm in the middle of 2010's decade. From then on, the conventional scaling approach has been replaced by an equivalent one, consisting in stacking many memory cells in the direction perpendicular to the plane of the wafer, thus breaking the trade-off between the dimension of each memory cell and the array storage density. Although the resulting three-dimensional (3-D) NAND Flash memories determined a general improvement in terms of reliability, some new issues have emerged due their novel architecture. One of them is absence of a body contact, preventing to directly access the string channel to raise its potential during the erase operation, similarly to what is done in planar technologies. To this purpose, gate-induced-drain-leakage (GIDL) occurring at the source-line and bit-line sides is exploited to inject a hole current towards the center of the string, thus raising its potential and triggering the emission of electrons from or the injection of holes into the storage layer.

On the other hand, NOR Flash arrays target code storage applications, therefore fast random access operation at the byte level and strong raw array reliability represent two mandatory requirements to be met. For this reason, differently from NAND Flash, the minimum feature size of NOR Flash technology has never been scaled beyond the 40 nm technology node. Despite this, in the last few years NOR Flash memory arrays attracted renewed interest for the implementation of hardware neural networks, which represent a promising solution to outclass (in terms of speed, power efficiency and integration density) conventional CMOS systems based on the Von-Neumann architecture in problems dealing with unstructured data, such as image recognition and classification. Hardware neural networks are computing systems, inspired to biological neural networks, made of arrays of computational units (neurons) interacting through connections (synapses) of different strength (synaptic weight). In hardware neural networks NOR Flash memory arrays are operated as artificial synaptic arrays connecting layers of adjacent neurons; each floating-gate (FG) memory cell in the array behaves like an artificial synapse receiving a voltage input at its control-gate and producing an excitatory post-synaptic current at its drain depending on cell threshold voltage. The threshold voltage value of the cells in the array is set during a learning phase to reproduce suitable synaptic weights, allowing the network to specialize its behaviour to perform a well defined task.

In this framework, the present thesis aims on one hand to investigate the GIDL-assisted erase operation in 3-D NAND Flash memory arrays. In particular, a compact model able to describe both the string dynamics and the threshold voltage transient during erase is developed and challenged for different string geometries and working conditions. On the other hand, a novel operational scheme allowing to employ mainstream NOR Flash memory arrays in neuromorphic systems is suggested, and its feasibility is successfully demonstrated by means of the implementation of a prototype hardware neural network. Besides, the impact on the performance of NOR Flash-based hardware neural networks of reliability issues typical of Flash memory cells, such as program noise and random telegraph noise, is explored in detail.

TCAD MODELING OF CURRENT TRANSPORT AND MAIN RELIABILITY ISSUES OF POLYSILICON-CHANNEL 3-D NAND FLASH STRINGS

Aurelio Mannara - Supervisor: Christian Monzio Compagnoni

NAND Flash technology represents today one of the leading solutions for highly performing non-volatile memories. Thanks to the incessant technology development since its first introduction, NAND Flash memories went through a steady decrease of their characteristic cell feature size (F), with a consequent increase of their storage density.

Although the scaling of planar NAND Flash arrays allowed to reach feature size values close to 15 nm around 2015, reliability issues arose by getting close to physical limits. To solve this, a complete architectural change was needed. Therefore, 3-D NAND Flash memory arrays were developed, in which the third dimension is exploited by stacking multiple memory layers. This allowed to relax the cell size scaling, while still increasing the storage density. Thanks to the increased cell size and to the gate-all-around structure of the cells, the impact of reliability issues of planar NAND arrays was reduced. Though, this new architecture is not free of issues. One of the main drawbacks coming with these new structures is given by the channel being polycrystalline, due to the involved manufacturing processes, for which silicon is deposited in cylindrical cavities by low pressure chemical vapor deposition (LPCVD).

The polycrystalline nature of the silicon channel in 3-D NAND Flash arrays results in the presence of multiple grain boundaries, where dangling bonds bring trap states, resulting in the generation of potential barriers and a localized reduction of the free electrons density. On top of causing an increase of the string channel resistivity, the random nature of grain profiles introduces a new variability source. As a matter of fact, different memory cells can display a different behavior and storage properties, depending on the specific grain configuration. Moreover, the transport properties at grain boundaries introduce peculiar temperature dependence of the device behavior. These issues must be taken into account in order to properly design the array and its control circuits. Due to the complexity of such devices and the involved reliability issues, care must be taken in order to avoid issues during operation.

In this framework, this thesis aims to describe current transport and main reliability issues in vertical-channel 3-D NAND Flash strings by means of Technology Computer Aided Design (TCAD) modeling, accounting for the polycrystalline nature of the channel. First of all, a custom tool able to generate virtual polycrystalline structures was developed and integrated in an environment compatible with a commercial TCAD suite in order to simulate such structures. Thanks to the flexibility of the tool, different partitioning algorithms were implemented and different interface properties could be applied at the grain boundaries, allowing to define different transport models in the inter-grain and intra-grain regions.

After defining the TCAD model, a comparison of different modeling approaches typically adopted to describe current transport in polysilicon-channel MOSFETs by means of TCAD simulations was performed. In particular, results under the effective medium approximation were compared to ones obtained from simulations where the polycrystalline nature of the channel was kept into account. Simulations were then performed to address the impact of introducing a silicon dioxide filler of different radii in the polysilicon channel, therefore switching to a Macaroni-like structure, effectively removing some of the defects on the grain boundaries.

Moreover, the current variability caused by the haphazardness of the polysilicon grain configuration was investigated on a three-gates string when simulating the read operation on a single cell in a 3-D NAND string, comparing the results coming from two different models to describe transport across the polysilicon channel. Then, the TCAD model was calibrated against experimental data coming from real devices and it was used to analyze the statistical dispersion of the temperature activation of the string current during a read operation on a single cell caused by the haphazardness of grain configurations.

Finally, the calibrated TCAD model was employed to analyze Random Telegraph Noise (RTN) fluctuations in 3-D NAND strings. In particular, the peculiar temperature sensitivity of RTN fluctuations in 3-D arrays was attributed to the strong constraints coming from grain boundary defects and it was validated by means of Monte Carlo experiments.

LOW-NOISE LOW-POWER INTEGRATED CIRCUITS FOR HIGH RESOLUTION X AND GAMMA RAY SEMICONDUCTOR DETECTORS

Filippo Mele - Supervisor: Giuseppe Bertuccio

The increasing demand for performance improvements in radiation detectors, driven by cutting-edge research in nuclear physics, astrophysics and medical imaging systems, is causing not only a proliferation in the variety of radiation sensors, but also a growing necessity of tailored solutions for front-end readout circuits, signal elaboration, storage and digital transmission. Within this work two main projects in the field of nuclear microelectronics for semiconductor radiation detectors have been presented, using complementary design flow approaches. In the first project, the study and development of a crucial building block in radiation detection systems has been carried out: the Charge Sensitive Amplifier (CSA). The CSA operates the fundamental conversion from the charge signal generated by the detector to the voltage signal processed by the following filtering stages; due to the specificity of its purpose, its design is inevitably linked to a target application, and is hardly found as an off-the-shelf component. However, through the definition of some flexible design strategies, a new preamplifier optimized for sub-microsecond processing times, named SIRIO-6, was realized in CMOS technology showing successful results in different applications based on low-capacitance semiconductor radiation detectors. An original study of the pulsed-reset discharge transient, which addresses the increasing demand in high-rate/low-deadtime applications was described, as well as an accurate modeling of the stray capacitance in compact CSA assemblies; a novel patented solution, currently granted in Italy and under international application process, has been disclosed for the optimization of parasitic capacitances in ultra-low noise CSAs. Within a controlled laboratory environment, the SIRIO preamplifier demonstrated an excellent performance with moderately cooled Silicon Drift Detectors (SDDs), obtaining an optimum noise of 3.4 electrons r.m.s. (corresponding to 122.7 eV FWHM on the ^{55}Fe 5.9 keV line) at t_{peak} = 5.6 µs, and 8.1 electrons r.m.s. (138.4 eV FWHM on the 5.9 keV line) at t_{peak} = 0.1 µs. At room temperature, the SIRIO preamplifier coupled with a CdTe pixel detector recorded a state-of-the-art resolution of 463 eV FWHM on the ^{241}Am 59.5 keV line at t_{peak} = 1 µs, which improves by a factor of two the best resolution performance present in literature at the time of writing for comparable experimental conditions. The SIRIO preamplifier was also integrated and characterized on two radiation detection instruments based on SDDs and realized entirely within the ReDSoX (INFN) project. The first one, shown in Fig. 1, is a Low-Energy X-Ray Fluorescence (LEXRF) detection system based on multi-element SDD arrays (12'300 mm^2 of total area) for the TwinMic beamline of the ELETTRA synchrotron (Trieste, Italy), which allows a wide solid angle coverage, and a spectroscopic resolution of 125 eV FWHM on the 5.9 keV line, keeping an excellent capability of detecting lightweight elements down to the C-Kα line (277 eV). The second detector system, realized for X-ray Absorption Fine Structure and X-Ray Fluorescence (XAFS/XRF) experiments, steers the design focus on high count-rate performances (up to 15.5 Mcps) with a best energy resolution below 150 eV FWHM. The resulting state-of-the-art detector, completely realized in Italy, was successfully shipped and inaugurated at the SESAME synchrotron, in Allan (Jordan) at the end of 2019. A detail of the SIRIO/SDD mounting on the XAFS/XRF system is shown in Fig.2.

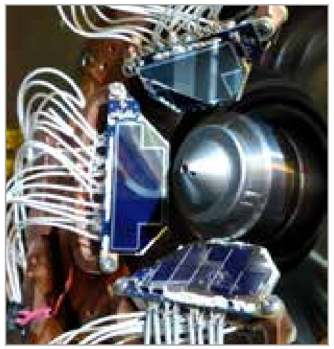

Fig. 1

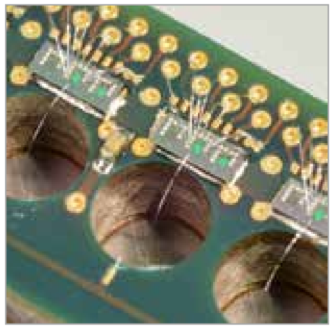

Fig. 2

The second main project of this work is an Application Specific Integrated Circuit chipset designed for the THESEUS space mission of the European Space Agency. Within this project, the high-level of customization required by space application instruments is combined with the complexity of a large scale multi-channel architecture which, in its flight version, requires a total of 25'600 front-end and 12'800 back-end channels. The ASIC constellation, named ORION, is thus a solution tailored to the requirements of the THESEUS mission, and in particular to the X and Gamma Imaging Spectrometer (XGIS), which is based on the detection module reported in Fig. 3. The ORION chipset features a dual dynamic range architecture for separate X-ray (2 – 30 keV) and Gamma-ray (30 – 20'000 keV) processing based on a double detection mechanism which combines the usage of SDDs and Cesium-Iodide scintillators. With an on-chip photon discrimination, filtering, analog-to-digital conversion and time-stamping capability, the ORION ASIC, realized in collaboration with University of Pavia, can interact directly with an on-board FPGA via SPI protocol.

Fig. 3

Among the future studies and applications connected to the first part of the presented work, it seems of particular interest the further investigation of fast discharge solutions for pulsed-reset CSAs, which might address the future interest of high-rate applications in synchrotron radiation facilities, and eventually the study of a noise-scalable CSA version for sub-milliwatt power consumption in multi-channel or vacuum operated applications. On the other hand, the result achieved using the CdTe pixel in combination with the SIRIO preamplifier, opened the path for new high resolution and high count-rate applications of such of detectors, which demonstrated to be well suitable for spectroscopy at room temperature in an energy range that cannot be covered with Silicon Drift Detectors. Concerning the ORION chipset, a multi-channel version of the ASIC is programmed for the beginning of 2021, which will be used in a module demonstrator to be presented to the ESA evaluation committee before the final selection and commissioning of the THESEUS space mission.

IMPLANTABLE AND INGESTIBLE DEVICES: NEW OPPORTUNITIES FOR ELECTRONICS AND HUMAN BODY INTERACTION

Filippo Melloni - Supervisor: Dario Andrea Nicola Natali

The area of Biomedical Engineering dealing with the interaction between electronic technologies and human body is increasingly gaining relevance and has long since revolutionized the process of doing therapy and diagnosis. The design of medical devices able to operate in an intense interaction with organs and tissues is particularly attractive. They are sophisticated tools capable to monitor electrical and chemical biological signals, replace parts of the body and restore compromised physiological functions. These advancements were achieved thanks to an interdisciplinary approach bridging electronics, chemistry, material science, mechanical engineering and medicine. In this complex knowledge blend, organic materials and organic electronics are increasingly proving to play a relevant role. Bioinert and biodegradable organic materials with favourable mechanical and chemical properties are key elements for a successful integration within the body tissues. While the ability of organic electronics to enable a mixed ionic and electronic charge transport is a win strategy for a proficient operation within physiologic medium. In this dissertation, advancements in this context are reported, approaching this wide field from different point of views: from the proposal for an inert system for chronic implantation, to the study of materials and devices for transient operation in the gastro intestinal tract.

In the first part, the document reports new ideas for a comfortable monitoring of the human bladder volume, aiming to an improvement in the life quality of patients affected by different urinary tract dysfunctions. The “proof-of-concept” device proposed for this task overcomes part of the limitations imposed by the semi-invasive technologies currently used as gold standard, and of the implantable solutions recently proposed in literature for the same purpose. Its working principle is based on a capacitive linear encoder integrated with a passive wireless radio frequency resonator, which can be remotely interrogated. The sliding encoding mechanism allows a wide sensing range without stringent requirements on materials properties and overall device stability, validating this encoding approach as a worthwhile alternative to stretchable devices proposed for the same therapeutic function. Figure 1 shows a photograph of the device and its working principle.

Later, I focused on the study of new materials and devices to be introduced in the novel context of edible electronics, with the goal to develop a harmless and digestible technology, as an alternative to the currently used ingestible medical devices. Appropriate attention is posed in the selection of valuable materials for the design of active electronic components, electing only edible approved materials, or proceeding with a cytotoxicity assay on promising candidates, in order to provide a preliminary degree of biocompatibility. In particular, the electrical properties of novel organic semiconductors and an edible solid electrolyte were investigated. Effort was spent in understanding the charges accumulation mechanism at the semiconductor electrolyte interface and thus detecting the semiconductor permeability to ions. As a result, low-power and air-stable transistors were fabricated, embedding only edible or at least non-cytotoxic materials. The characterization overtime shows a device shelf-life over 100 days without evident degradation. Figure 2 shows the architecture of the devices fabricated and a photograph of a fully printed transistor prototype. These results make significative steps toward the fabrication of fully edible electronic devices.

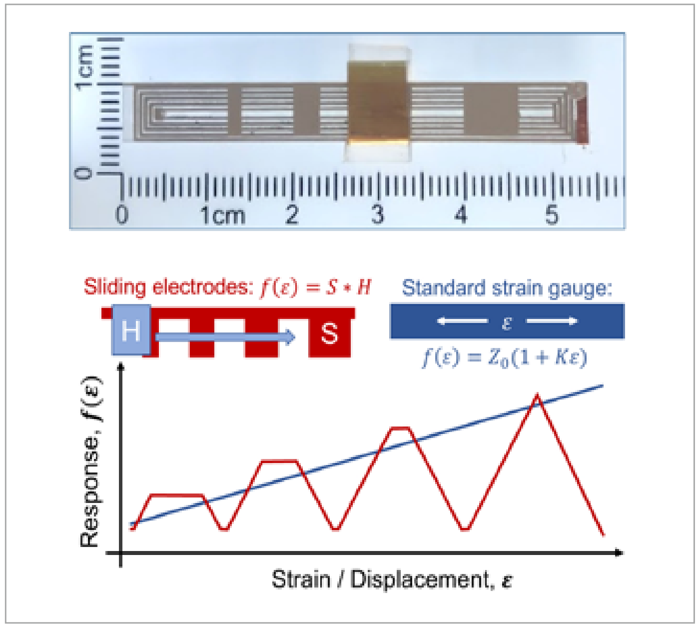

Fig. 1 - A photograph of the proposed proof-of-concept device and its readout compared to the one of a standard strain gauge.

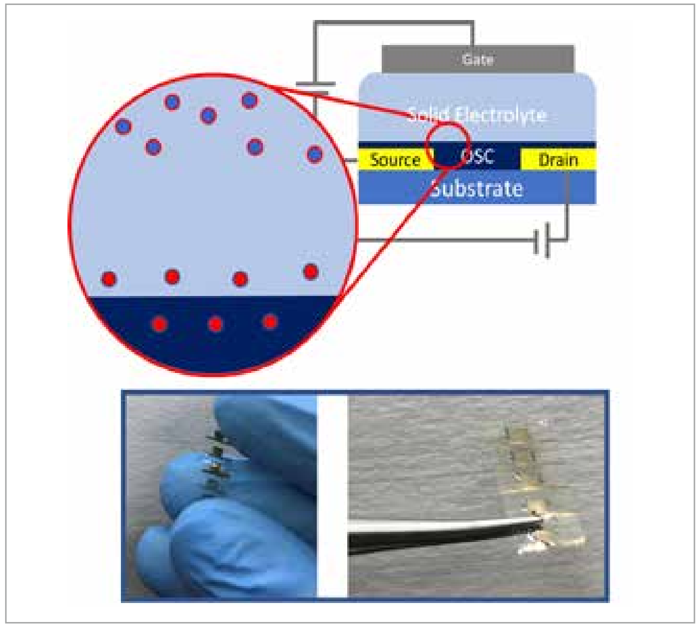

Fig. 2 - On the left the top-gate/bottom-contact architecture of the solid electrolyte transistor fabricated, and embedding a semiconductor permeable to ions. On the right, a photograph of preliminary experiment toward a fully printed and biodegradable device.

EXPLOITING ENVIRONMENT CONFIGURABILITY IN REINFORCEMENT LEARNING

Alberto Maria Metelli - Supervisor: Marcello Restelli

When we think of the process of learning for human beings, we realize that interaction with the surrounding environment plays a crucial role. Human beings acquire abilities in different ways, but all of them involve a certain degree of interaction with either the external environment or other agents (biological or artificial). A baby, an example of a biological agent, learns how to walk in a trial and error fashion. They try the first movement and then they likely fall down, so they try another one and, sooner or later, they manage to stay upright. No teacher is, in principle, needed in this process, as the effects of the movement are associated with a feedback signal (falling down or staying upright) that tells the baby whether it was profitable or not. This feedback triggers an adjustment in the behavior and, hopefully, over multiple trials, leads to the realization of the ultimate goal of walking. This kind of interaction is at the basis of Reinforcement Learning (RL).

In an RL problem, an artificial agent interacts with an environment by sensing perceptions and performing actions. The perception defines the ability to measure the state of the environment. The actuation possibilities, instead, are concerning the ability to perform actions on the environment. Whenever an action is played, it produces an evolution of the environment state and the agent is provided with a feedback signal, the immediate reward. The agent acts with the purpose of finding the proper actions, so as to maximize some utility function. In RL, such a utility function is defined as a notion of long-term reward, i.e., the cumulative (possibly discounted) sum of the immediate rewards collected during the agent's experience.

Environment Configurability

Besides the remarkable success demonstrated in recent years, RL appears to be deeply rooted in the definition of the environment as an immutable entity out of any control. In the traditional model, the agent can indirectly control the environment by means of the performed actions, but cannot directly change the environment dynamics. This is certainly true in a large number of applications, although we can identify a huge number of examples in which a "partial control" on the environment can be exercised. For instance, a human car driver has at their disposal a number of possible vehicle configurations they can act on (e.g., seasonal tires, stability, vehicle attitude, engine model, automatic speed control, and parking aid system) to improve the driving style or quicken the process of learning a good driving policy. Another example is the interaction between a student and an automatic teaching system: the teaching model can be tailored to improve the student's learning experience (e.g., increasing or decreasing the difficulties of the questions or the speed at which the concepts are presented). It is worth noting that the active entity in this configuration process might be the agent itself or an external supervisor (or configurator) guiding the learning process. Another example is product placement in a supermarket. A supervisor can dynamically adapt where to place the products in order to maximize customer satisfaction. Differently from the previous examples, it might be possible that the configurator (e.g., the supermarket staff) has a goal that is different from that of the agent (e.g., the customer).

In all these scenarios, whenever altering some portions of the environment or some environmental parameters is allowed, we speak of environment configurability (Fig. 1). Environment configuration arises in several real-world scenarios, with different objectives, involving different levels of cooperation and competition between agents and configurators. Therefore, we believe, the nature of this kind of interaction deserves additional study. The dissertation is subdivided into three parts.

Part I: Modeling Environment Configurability

The first part of the dissertation is devoted to the formalization of environment configurability. We propose a novel extension of the traditional Markov Decision Process (MDP), named Configurable Markov Decision Process (Conf-MDP), in order to properly represent the configurability possibilities of the environment. Then, we devise a taxonomy for the Conf-MDPs, according to the properties of the interaction between the agent and the configurator. Specifically, we identify different settings based on whether the agent is aware of the configurator presence and whether their objectives coincide. This latter distinction reveals two wide settings that characterize the interaction between the agent and the supervisor: the cooperative and the non-cooperative setting.

Part II: Learning in Cooperative Configurable Markov Decision Processes

In the second part of the dissertation, we focus on the cooperative Conf-MDP setting and we study the learning problem consisting of finding an agent's policy together with an environment configuration so as to maximize the long-term reward. We start with the simpler setting in which the environment is characterized by a finite state-action space. Then, we move to the continuous environment case and we devise an approach, able to work with no knowledge of the environment model.

Part III: Applications of Configurable Markov Decision Processes

The last part of the dissertation is dedicated to the study of two applicative scenarios in which environment configuration can play a relevant role. The first application we examine is policy space identification, i.e., the problem of identifying the space of policies that an agent can access during the learning process. Knowing the agent's policy space can be particularly convenient when environment configuration is possible. Indeed, agents optimizing the same objective but having access to different policy spaces might benefit from different environment configurations. The second application we explore is the control frequency adaptation. RL problems are typically formulated as discrete-times problems, but often derive from the time discretization of continuous-time ones. Thus, the control frequency is a relevant design choice and can be considered an environmental parameter that can be configured.

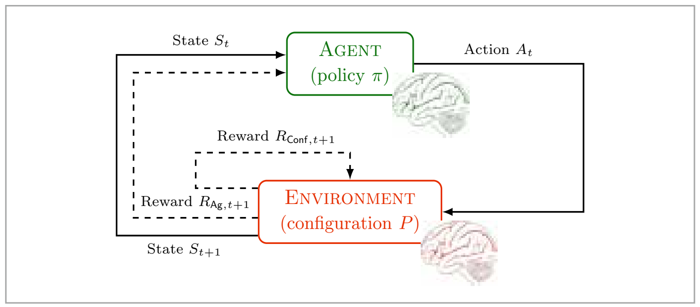

Fig. 1 - Graphical representation of the interaction between an agent and an environment in a Conf-MDP.

A MIXED-SIGNAL INTEGRATED CIRCUIT BASED ON PHASE CHANGE MEMORY SYNAPSES FOR DEEP NEURAL ACCELERATORS

Irene Munoz - Supervisor: Daniele Ielmini

Nowadays, machine learning empowers several different consumer applications, where artificial intelligence is well supported by statistical models and mathematical algorithms that allow computer systems to accurately perform specific tasks. Deep neural networks (DNNs) have dramatically enhanced classification and recognition operations by exploiting a general-purpose learning procedure in a multi-layer architecture.
However, these architectures have several limitations. First, the training and inference of DNNs using standard digital systems is time expensive and power consuming. Secondly, trained DNN cannot adapt to a constantly changing environment. For instance, biological organisms steadily acquire and modulate knowledge with respect to the environment in which they live (lifelong learning), while DNNs are affected by catastrophic forgetting whenever new data is learnt. For these limitations, the scientific community and the industry are looking for novel methods to improve the performance and efficiency of DNNs.
Taking advantage from the latest advances related to innovative computing approaches such as in-memory computing, the training and testing of DNNs could be highly improved in terms of speed and power efficiency. In-memory computing rises as a very effective method for overcoming the limitations of typical von Neumann architectures since it massively parallelizes the operations and performs the calculations where the data are stored, avoiding the so called ``von Neumann bottleneck”.
In particular, in-memory computing requires the use of memory elements capable of storing data and performing calculations at the same time. New emerging non-volatile memories (NVM), such as phase change memory (PCM) or resistive switching RAM (RRAM) give solution to these requirements, as they have small size and show fast switching, multilevel capability, low-voltage operation, and time-dependent dynamics. In addition, they can be arranged in array architectures.
As the main operations in DNNs are related to dense matrix-matrix multiplications, trained weights can be mapped in NVMs crossbar arrays as conductance values, exploiting the Ohm’s and Kirchhoff’s laws for performing matrix-vector-multiplication (MVM). These new advances could outperform current GPUs and CPUs in terms of power consumption and speed since the use of MVM allows to perform multiply-and-accumulate (MAC) operations in just one step.
There are many circuital challenges to be faced related to this kind of computation, as the device non-idealities, the resolution of the readout voltages, the large analogue currents, or the trade-off between area/power consumption with respect the classification accuracy of input images.
This doctoral dissertation proposes several solutions to reduce the impact of some of the previous limitations. Therefore, the presented theoretical analysis and implementing results aim to the improvement of DNNs from a technical point of view, where two different research paths have been followed.
The first one regards the introduction of bioinspired methods into the general architecture of neural networks for improving the capability of DNNs in the recognition of unknown images. Spike-Timing-Dependent plasticity (STDP), brain-inspired homeostasis and neural redundancy are some of the elements that have been included in the network to stabilize the learning processes.
One of the main projects developed during the doctoral program has dealt with the digital development of a hybrid supervised - unsupervised neural network capable of performing lifelong learning. The supervised part is formed by a convolutional neural network that allows the extraction of generic features from a training dataset; the unsupervised section is constituted by a spiking winner-take-all (WTA) network that follows the STDP protocol. The inference results are validated for the correct classification of up to 5 non-trained classes of the MNIST and fashion-MNIST datasets and compared with PCM-based approaches. The digital implementation of the network has been made by using the Xilinx Zynq-7000 System-On-Chip (SoC).
This doctoral dissertation also proposes an insight over the state of the art of bio-inspired computation, explaining the design of a homeostatic neuron based on the gradual crystallization of a PCM device. The partial crystallization of the PCM memory allows to improve the learning accuracy of unsupervised STDP by specializing the output neurons to a particular pattern by modifying the internal neural threshold. This latter spiking-control system has been named spike-frequency adaptation (SFA).
Other research studies have introduced the design of a new hardware implementation based on SiOx RRAM that supplements the accuracy of convolutional neural networks with the flexibility of bio-inspired spike-timing dependent plasticity for performing lifelong learning. To enable the cohesion between the stable and the plastic section of the network, the bioinspired SFA has been exploited, as it enhances the efficiency and accuracy of the network.
The second research line have covered the hardware design of a mixed-signal integrated circuit based on PCM synapses for the development of deep neural accelerators. Following the in-memory computing hardware approach, DNN weights are mapped in NVM arrays. A generic 1-layer fully connected (FC) multi-layer perceptron (MLP) is proposed, where the training weights are mapped into 4-bit unsigned digital words of `0s’ and `1s’, taking advantage from the wide resistive window of PCM devices between the high resistive state (HRS) and low resistive state (LRS), respectively.
The circuit has been designed using the ST Microelectronics BJT-CMOS-DMOS (BCD) 90 nm design kit with an embedded 1T1R PCM cell, which has been organized forming 4 arrays of 5 Kb each. The memory cells are manufactured with an optimized Ge-rich chalcogenide alloy and are stacked over the CMOS circuitry in the back end of the line (BEOL). Based on experimental measurements of the PCM 1T1R cell, Monte Carlo simulations have been performed to validate the hardware architecture. The results have supported the choice of the hardware design and the methodology to follow for the mapping of weights.
The design faces several circuital challenges such as the implementation of the analogue-to-digital interface between the array and the input-output (IO) peripherals, and the signal processing for driving the PCM devices. The implemented ADC follows an 8-bits SAR architecture and optimized layout to limit the presence of parasitic capacitances (top-plate) that worsen the conversion efficiency of the ADC. The analogue frontend is formed by a column decoder, an amplification stage, a voltage-dependent current generator based on a band-gap circuit, and a rail-to-rail output buffer. The circuit performs intelligent tasks of recognition of handwritten digits (MNIST) at high bandwidth (500 kHz) and low power (~ 200 mW). All the MNIST inference activity is performed in less than 0.8 s (256 mega operations per second, MOPS), which is far less with respect to the state of the art of standard von Neumann processors.
Furthermore, the whole ASIC relies on a significant robustness with respect to the non-idealities of PCM devices, since the results are resilient both to drift and resistance variability, achieving almost the same software classification accuracy of the MNIST dataset (~85 %).
The presented research activity offers a complete insight about the current panorama of neuromorphic engineering, both from the bioinspired and technical points of view. In particular, this work highlights the main features, problems, and design requirements for efficiently implementing a hardware integrated DNN using PCM cells. The adopted solutions and the obtained results are extensively described by pointing out the advantages of analogue in-memory computing for the realization of arithmetic calculations.

COMPUTATIONAL INFERENCE OF DNA FOLDING PRINCIPLES: FROM DATA MANAGEMENT TO MACHINE LEARNING

Luca Nanni - Supervisor: Stefano Ceri

Co-supervisor: Colin Logie

DNA is the molecular basis of life and consists of approximately three billion base pairs, which would total about three meters if linearly untangled. To fit in the cell nucleus at the micrometer scale, DNA has, therefore, to fold itself several times. Understanding the mechanisms of genome folding is a major biological research problem.

In this thesis, I first present the design and implementation of the PyGMQL software for interactive and scalable data exploration for genomic data. PyGMQL is based on the big data genomic engine GMQL and presents itself as a easy-to-use Python library, interacting seamlessly with other data analysis packages.

I then apply my software to the study of chromatin conformation data. I discover a set of spatial rules based on the insulator protein CTCF and its orientation which correlate with genome topology, highlighting the existence of a “grammar of genome folding”.

I finally focus on the relationship between chromatin conformation and gene expression. I use graph representation learning to encode chromatin topological features of genes. The learnt gene embeddings are then used to predict if two genes are co-expressed or not, given independent gene expression data. The results indicate a correlation between chromatin topology and co-expression, shedding a new light on this debated topic and providing a novel computational framework for the study of co-expression networks.

CLOUD-EMPOWERED DSP LEVERAGING BIG DATA ANALYTICS IN C-RAN ARCHITECTURE

Armin Okic – Supervisor: Alessandro Enrico Cesare Redondi

The constant increase of demands from mobile users is putting hard pressure on mobile network operators to expand the network infrastructure to satisfy the requirements of their users and offer them reliable and affordable mobile services. Therefore, the deployment of new infrastructure and investments in new technologies is necessary and very costly, while the revenue is not proportional to the invested capital into the network. The main reason for this disproportionality is in the paradigm shift of how, when, and why mobile users are utilizing mobile devices. The great influence comes from content application vendors that are offering over-the-top services to mobile users, like Facebook, Google, or Amazon. These vendors are getting the most from mobile networks as their users are using the cellular networks to access their content, applications, and services, while, from the other side, those providers do not invest anything into mobile network infrastructure (or pay anything to mobile network operators). On top of this, adding versatile functionalities of modern smartphones, it is clear that the dominant utilization of mobile networks moves from conventional voice calls or SMS messaging towards mobile data and Internet connectivity. For these reasons, mobile network operators need to develop new approaches in managing mobile networks in general. For the emerging 5G technologies, it was already shown that the promising technologies for the network operators will rely on network virtualization and network slicing.

The development of Network Function Virtualization (NFV) technologies, as one of the pillars of emerging 5G and beyond networks, is spreading out to all segments of mobile networks, including both Core Network (CN) and Radio Access Network (RAN). Once virtualized, the network functions become scalable, flexible and could be computed in different parts of the network. This flexibility is supported by the development of powerful Cloud systems that are deployed closer and closer to the end-users, creating in such a way Edge cloud resources and the concept of Multi-access Edge Computing (MEC). From the other side, the generation of enormous amounts of network monitoring data offers an opportunity to improve and optimize any task related to network resource management by analyzing those datasets and extracting meaningful network knowledge.

In this thesis work, we focus on exploiting virtualization techniques within the Cloud-RAN (C-RAN) architecture in combination with outcomes of mobile network monitoring datasets analysis at different levels of the network. We define and estimate computational requirements and constraints of computing virtualized RAN (vRAN) functions in the Cloud, for which we create an experimental platform built out of virtual machines specifically designed for processing vRAN functions. Through the detailed analysis of the C-RAN architecture, we investigate several design solutions by mainly tackling the problem of centralizing Digital Signal Processing functions to be computed in the Cloud with General Purpose Processing (GPP) hardware. We mainly focus on highly processing-intensive DSP functions of the RAN network.

Latest advances in network traffic data analysis indicate that mobile users' behavioral patterns are predictable in space and time as a high correlation between current and previous usage of mobile network infrastructure is present. Those meaningful insights about spatial and temporal network utilization, together with the support of the NFV concept, motivate mobile network optimization to follow the actual requirements of mobile users.

Therefore, we perform an extensive big data analysis and exploit its outcomes to provide an optimization framework that allocates computational resources of virtualized RAN functions. The data investigations are carried out on realistic city-wide and highway area datasets in order to demonstrate how mobile network traffic varies over time and space. Moreover, for the purposes of providing a dynamic allocation of virtual resources, we compare different traffic forecasting algorithms, analyzing in detail the impact of differently sampled datasets and prediction horizons. More precisely, we introduce a unique approach for forecasting mobile traffic traces, providing detailed research on the influence of sampling frequency and machine learning approaches on future samples anticipation results. This approach is recognized in other research areas, but it was never before applied in mobile networks. The study focuses on the application-level dataset, which ensures that final outcomes can be used for making decisions in C-RAN optimization, which includes a grouping of Radio Remote Heads (RRUs) into Baseband Unit (BBU) pools per application or network slice. Furthermore, by showing that forecasting of high-frequency time series samples of mobile traffic can be obtained directly from low-frequency time series with high performances, we provide valuable insights for data storage optimization of network operators, which might be helpful in future works regarding the big data storage systems.

The knowledge gained from the big data analysis and forecasting algorithms is further used to improve the proposed optimization framework that allocates virtualized RAN resources in a dynamic and optimal manner over time by following the variations of actual network traffic loads.

We also develop a unique procedure for estimating computational loads of virtualized RAN functions by combining open-source software tools and realistic mobile traffic datasets. By running the proposed optimization algorithm, we can ensure a minimal number of allocated virtual instances for the controllable level of service interruptions caused by the presence of dynamic allocation of RAN resources. Furthermore, we analyzed the influence of different forecasting algorithms and forecasting horizons on the proposed optimization algorithm. To deal with intensive computational requirements of the proposed optimal solution and to be able to solve the allocation problem for a high number of instances and over higher observational times, we provide near-optimal and fast heuristic algorithms.

Our proposed solution of allocating virtual RAN resources by exploiting the data analysis provides higher savings in the mobile network compared to the conventional C-RAN approaches. Additionally, the framework can be potentially used by network operators to support the requirements of various mobile applications and network services requested by the mobile users as a part of 5G and beyond network use cases.

We conclude that this work paved the path towards simple estimation of computational requirements of RAN functions, connecting it with realistic network data. The results provided in this work are highlighting the importance of considering network utilization feedback and the presence of big data systems in mobile networks for the improvements of overall network performances. The paradigm of introducing machine learning into the optimization of mobile networks is more and more spread into each domain of telecommunications, while our work expands the analysis in the domain of RAN network and optimization of its performances.

SAFE POLICY OPTIMIZATION

Matteo Papini - Supervisor: Marcello Restelli

Motivation

Reinforcement Learning (RL) is the most promising approach to decision making for autonomous agents. Its fundamental principle is simple and general: let the agent interact with the unknown environment, provide it with a reward signal that informs it of the appropriateness of its actions, and let it autonomously learn how to maximize the rewards. This makes RL applicable to any sequential decision problem, at least in principle. Indeed, RL achieved marvelous results in games, where the environment is easy to simulate and collecting experience data is just a matter of computation.

Moving from games to real-world applications, new challenges arise. Take industrial robots as an example of agents that interact with humans or take decisions that affect physical entities. First, data collection can be a slow and expensive process. Simulated experience can alleviate this problem, but engineering a reliable simulation is simply not possible in many applications. This makes sample complexity, the amount of data required for learning, a fundamental problem for the successful application of RL to real-world problems. The second challenge is a universal dilemma of sequential decision making, which is even more relevant in real-life RL: exploration. Learning from scratch, an agent needs to perform some actions just to increase its knowledge of the world. This exploratory behavior may be suboptimal according to its current knowledge, but it unlocks the ability to learn even better behavior. In a simulated environment, we typically only care about the final solution, so exploratory behavior is always worth the effort. Instead, when the agent is learning in the real world, intermediate performance can really matter. The third big challenge is safety. Learning in the real world can have concrete consequences, ranging from monetary losses to irrecoverable damage, up to threatening the life of human operators. Hence, it is important to identify all potential hazards and reduce the risk as much as possible. Careful engineering of the agent's hardware and of the surrounding environment can help a lot by reducing the number of hazards. However, it is equally important to design learning algorithms that do not take unnecessary or unacceptable risks. Designing a truly safe RL agent requires to intervene at multiple levels: in defining the reward (to ensure it is aligned with our desires), in planning the interaction, and in designing the learning process itself. In this work, we take the so-called Seldonian approach to RL safety. This means that we do not focus on how safety concerns can be embedded in the problem definition, nor on how the agent can plan its interactions to reduce hazards, but on the risks introduced by the learning process itself. In other words, we focus on preventing the negative effects of epistemic uncertainty, the lack of knowledge of the environment and its workings which can lead the agent to learn unsafe behaviors.

Scope and Methodology

We address the safety challenge outlined above for a specific class of RL algorithms: Policy Optimization (PO). These methods are currently the most promising for real-world applications of RL, due to their ability to deal with continuous decision variables, their robustness to noise, and their general convergence properties. Compared to other RL methods, they also make it easier to exploit prior knowledge about the task at hand, which can help to enforce a minimal amount of safety. Indeed, PO underpins the main achievements of RL in robotics. We study the problem of safety of policy optimization algorithms, in the online-learning, the Seldonian sense outlined above, and the relationships between safety, sample complexity and exploration.

We take a mostly theoretical approach, studying the properties of existing PO algorithms and seeing how the latter can be modified to provide formal guarantees. Our original contributions pertain a specific class of PO algorithms known as Policy Gradient (PG) methods. These are based on stochastic gradient descent and are by far the most used by practitioners, especially due to their applicability to deep neural policies. Our theoretical work is complemented by some algorithmic contributions, sometimes heuristic, and by empirical evaluations on simulated continuous-control problems. The latter should be intended mostly as an empirical verification of the theoretical findings.

Contributions

Monotonic improvement guarantees: we study monotonic improvement properties of policy gradient algorithms. This is the problem of avoiding oscillations in agent's performance during the learning process, which may correspond to undesired behavior. We follow previous work to carefully tune the meta-parameters of the policy gradient algorithm to guarantee monotonic performance improvement. Our findings are applicable to a much broader class of policies without sacrificing rigorousness. We provide a policy gradient algorithm, called SafePG, with monotonic improvement guarantees.

Conservative off-policy optimization: we address a problem that lies in the intersection of safety and sample complexity: that of re-using historical data for policy optimization. This practice can help to reduce the need of costly interaction data, but can produce overconfident, hence unsafe behavior if performed in a naive way. We propose an algorithm called POIS which employs importance-weighted estimators to perform multiple policy updates with the same data. A thorough analysis of these estimators allows to include a penalization term in the objective of POIS which explicitly prevents overconfident learning. The fundamental intuition is to distrust data that were collected with very different policies from the one that is currently being optimized. A practical version of POIS is evaluated on benchmark continuous-control problems and is shown to be competitive with state-of-the-art PO algorithms.

Convergence guarantees: we study the total amount of interaction data that policy gradient algorithms require to converge to a locally optimal solution. Although not explicitly about safety, this is very relevant to any real-world application of policy optimization. Although the convergence properties of basic policy gradient algorithms such as REINFORCE are well known, we raise awareness on some issues that may have been overlooked by previous literature, such as the classes of parametric policies to which these guarantees actually apply and the problems introduced by adaptive policy stochasticity. Then, we use insights from the stochastic optimization literature to design SVRPG, a policy gradient algorithm that is provably more data efficient than REINFORCE.

Safe adaptive stochasticity: we study the relationship between safety and exploration. We focus on the specific problem of guaranteeing monotonic performance improvement when the agent also has control over the amount of stochasticity of the policy. The naïve approach would be to treat the latter as an additional policy parameter, as practitioners often do. We show that this would be both inefficient and risky. Hence, we design a heuristic policy gradient algorithm that explicitly takes into account the long-term advantages of exploratory behavior. Then, we provide an adaptive meta-parameter schedule that guarantees monotonic improvement with high probability.

REMOTE BIOMETRIC SIGNAL PROCESSING BASED ON DEEP LEARNING USING SPAD CAMERAS

Marco Brando Mario Paracchini - Supervisor: Marco Marcon

Being able to constantly check, in real time and without any contact, the health condition of a person could have a significant impact in many different situations. Possible applications include: fitness assessments, medical diagnosis and driver monitoring. The act of extracting biomedical information by just analyzing a video is called remote PhotoPlethysmoGraphy (rPPG) or imaging PhotoPlethysmoGraphy (iPPG). This is an evolution of contact PPG, a technique introduced in early 20^{th} century that is nowadays commercially and clinically implemented in order to monitor the health state. The basic concept of PPG is placing a light emitter and a light receiver in contact of the subject skin (typically a finger) and analyzing the light intensity variation in order to estimate information about the cardiac activity. This is possible since the light intensity fluctuations that could be observed with a PPG device are caused by the periodic passage of blood in the vessel underneath the skin which changes how the light is reflected and transmitted by the subject's skin. On the other hand, rPPG aims at conducting the same analysis in a remote way without making any physical contact with the subject. The benefit of using rPPG are numerous in many different situations and in particular this could have a significant impact in the automotive industry. The possibility of constantly check the driver's health state could be extremely useful especially if this could be achieved without distracting or disturbing the driver. A computational unit equipped on the car that is able to extract a rPPG signal and analyze it in real time in order to consistently monitor the driver's cardiac activity could be pivotal in many situations. For example, these data could be used to enable particular features of the vehicle, such as autonomous driving, that could take control of the vehicle itself and avoid accident by simply safely parking the car in case of detected driver sickness or altered emotional state. With the evolution of smart cars this could become an important on board safety feature. Moreover, the biometric data could also be exploitable for other purposes, for example by insurance company or live remote assistance. For all these motivations, many automotive companies are researching on rPPG and the task of developing an rPPG automotive system was part of project DEIS which was a H2020 project that ran from 2017 to 2020 which has the purpose to develop methods in order to assess the dependability of many Cyber-Physical Systems. The main goal of this research is to develop a rPPG system able to estimate numerous biomedical measurements in real time and in a dependable fashion. Moreover, this work explores the possibility of adopting a SPAD (i.e. Single-Photon Avalanche Diode) array camera instead of traditional RGB camera, as done in the majority of publications in the this field. SPAD cameras are capable to detect even a single photon, have extremely high frame rate and have proved their usefulness in a very large range of applications, such as 3D optical ranging (LIDAR), Positron Emission Tomography (PET) and many others. In rPPG applications SPAD's high precision could be useful in measuring accurately the fluctuations in the light intensity reflected by the skin produced by the blood flow. On the other hand, the main drawback of using a SPAD sensor is their low spatial resolution due to technical limitation. In order to overcome this problem and use as much spatial information as possible, an ad-hoc deep learning based method is proposed. This is one of the first work in which deep learning methods are used in this field, being the adoption of this kind of techniques very recent in rPPG, the first publications started in 2019. Moreover, all the few other rPPG methods based on deep learning completely substitute the classical signal processing techniques with data driven ones using end-to-end networks. On one hand, the use of an end-to-end deep learning model has proven to achieve state of the art results on many computer vision tasks such as image segmentation, object detection, and many others. On the other hand, this kind of methods requires a massive amount of training data in order to learn how to extract heart related information directly from video frames and no prior domain knowledge is incorporated. This makes the performance of these methods tightly linked to the training dataset and potentially unable to generalize in different setting conditions. Moreover, the complete substitution of classical signal processing techniques developed using solid theoretical backgrounds (signal filtering, Fourier transform, etc.) with data driven ones could lead to non-optimal solutions. For the best of our knowledge no prior work has been done in trying to combine traditional and deep learning based signal processing in this field. We propose a deep learning method able to efficiently extract the pulse signal from the SPAD frames by performing skin segmentation. This method makes use of separable depth wise convolution layers in order to reduce the number of trainable parameters with the multiple beneficial effects of speed up the test time execution, not requiring massive training dataset and reduce the overfitting probability. Moreover, this research proved effectively that reducing the learnable parameters in a deep learning model, thus reducing its representation capabilities, it is not directly correlated to a reduction of the method's accuracy. Furthermore, as a side outcome of this work, a novel face skin segmentation dataset has been developed and made publicly available for the research community. One other novelty of this work is the adoption of a SPAD camera for rPPG applications since in all previous works in this field the cameras used are traditional RGB cameras. A major contribution of this work is to validate the effectiveness of performing rPPG using SPAD cameras, in particular in low illumination conditions, coupled with a deep learning technique in order to compensate for the low spatial resolution of Single-Photon cameras. Adopting a SPAD camera could also be beneficial in the use of the propose rPPG system in uncontrolled environments in which there could be sudden light variations (for example, if this technology is used in order to monitor a driver, this could happen in tunnel or in presence of car light reflexes). In this kind of scenarios, the best strategy in order to remove this high frequency noise is oversampling and SPAD cameras are the best one in this field. During this research, multiple experiments have been conducted investigating the possibility of using SPAD cameras in rPPG applications and evaluate their performances in respect to RGB cameras. In particular, a group of subjects has been acquired in controlled conditions in order to: select the best wavelength for performing rPPG with SPAD camera and compare the rPPG estimations obtained using SPAD camera and the ones that could be obtained using traditional RGB cameras. Another important aspect tackled in this work is the dependability of rPPG based heart rate estimation. In particular, since the biomedical parameters estimation is performed exclusively using optical information two scenarios were evaluated in which the pulse signal could be masked by other signals, thus resulting in an incorrect estimation. The two scenarios studied are the presence of subject head periodic movements and background pulsating light. In both these situations the resulting skin reflected light is strongly periodic but it is not related to the heart activity. Two different methods were proposed in order to detect these situations. Finally, a complete rPPG system, which includes all the processing described above and makes use of a SPAD camera has been implemented on a single-board ARM computer in order to estimate biometric parameters, such as heart rate, in real time and in a dependable way.

SPACE-TIME PARAMETRIC APPROACH TO EXTENDED AUDIO REALITY (SP-EAR)

Mirco Pezzoli - Supervisor: Augusto Sarti

The term extended reality refers to all possible interactions between real and virtual (computed generated) elements and environments. The extended reality field is rapidly growing, primarily through augmented and virtual reality applications. The former allows users to bring digital elements into the real world. Conversely, the latter lets us experience and interact with an entirely virtual environment. While currently extended reality implementations primarily focus on the visual domain, we cannot underestimate the impact of auditory perception in order to provide a fully immersive experience. As a matter of fact, effective handling of the acoustic content is able to enrich the engagement of users.

We refer to Extended Audio Reality (EAR) as the subset of extended reality operations related to the audio domain. In this thesis, we propose a parametric approach to EAR conceived to provide an effective and intuitive framework for implementing EAR applications. The main challenges of EAR regard the processing of real sound fields and the rendering of virtual acoustic sources (VSs); hence, EAR requires the development of properly designed sound field representations.

As far as sound field representation is concerned, two main paradigms are present in the literature: parametric and non-parametric. The former describes the acoustic field assuming a signal model governed by few meaningful parameters, e.g., the source signal and location, while the latter relies on the solutions of the wave equation providing accurate results at the cost of higher complexity and lower model interpretability. Therefore, in the context of the EAR, parametric models represent an appealing approach. They provide a compressed and intuitive description of the sound field, promoting the integration of VSs through the parameters of the model and their manipulation thereof.

Here, we introduce a novel parametric model for sound field representation based on few parameters. This model allows both the navigation and manipulation of a recorded sound scene. The main feature of the proposed solution is represented by the modeling of the acoustic source directivity integrated among the parameters of the representation. The directivity is a function describing the spatial property of the source sound radiation. As a matter of fact, sound sources typically present a directional acoustic emission imposed by their physical characteristics. It follows that the source directivity information influences our acoustic scene perception. Therefore, the integration of the directivity is a fundamental aspect for providing a more natural and immersive EAR, enhancing the user experience. In order to analyze the sound field, we adopted spatially distributed acoustic sensors. This configuration allows us to evaluate the acoustic field from different observation points in order to estimate the parameters required by the proposed representation. Successively, we exploit the estimated parameters to provide a sound field reconstruction technique that enables the *six-degrees-of-freedom* interaction (virtual navigation) with the sound field.

Conveniently, the parameters adopted for describing the acoustic sources can be exploited for characterizing a VS. Therefore, we can seamlessly implement EAR within the same parametric representation. The addition of the source directivity into the model is appealing since it allows the accurate rendering of VSs, including their directional characteristics. Hence, we can further lead the real-virtual interaction by implementing VS replicas of actual acoustic sources. A VS replica mimics the source spatial sound radiation through the VS directivity parameters. For instance, the VS parameters can be estimated from measurements on the real source. Conversely, we can rely on fully simulated acoustic sources, e.g., employing Finite Element Method (FEM) simulations, from which the VS parameters are derived. Therefore, an accurate estimate, prediction, and analysis of the directivity of VSs are fundamental to obtain an effective EAR.

In this thesis, we studied the VS implementation through a case study. In particular, we focused on the VS implementation of violins. Whereas violins present a peculiar directional radiation characteristic, we need to carefully analyze and model their directivity in order to provide an accurate VS implementation.

Regarding the analysis of the violin directivity, we can outline different solutions according to their invasiveness. In the first place, one can perform measurements directly on played violin. During our collaboration with *Museo del Violino* settled in Cremona (Italy), we had the unique opportunity to measure, for the first time, a relevant number of valuable historical violins made by the renowned masters of the Cremonese school such as *Antonio Stradivari* and played by professional violinists. From the acquired data, we derived a compressed representation of the violin directivity pattern based on the spherical harmonics expansion. Besides the VS modeling, the adopted representation allowed us to study and characterize the directivity patterns of the instruments, giving insights of their directional behavior. Although the measurement of played instruments allows an analysis scenario closer to the actual listening conditions, it might not be applicable for particularly fragile instruments. Less invasive techniques, such as nearfield acoustic holography (NAH), can be employed when conventional measurements cannot be carried out. It is known that the acoustic radiation of vibrating objects, such as violins, is determined by their dynamical behavior. Hence, from the knowledge of the vibration velocity field, we can estimate the directivity of the source. NAH allows the contactless estimation of the velocity field of a vibrating source from acoustic pressure measured in its proximity. Here, we introduced a novel NAH technique based on deep learning. In particular, we proposed a convolutional neural network (CNN) with an autoencoder-inspired structure to estimate the velocity field of both rectangular and violin plates.

Alternatively, simulations allow us to predict the directivity of a source relying on the FEM simulation of its vibroacoustic behavior. This approach minimizes the invasiveness at the cost of reduced accuracy caused by inherent approximations of the simulated model. It follows that an effective violin simulation requires a 3D model of the instrument geometry and the mechanical parameters of the material. Unfortunately, we can typically only acquire the outer surface of existing instruments. Therefore, we developed a practical technique for reconstructing the 3D model of violin plates, starting from outer surface scans and sparse thickness measurements taken at reference points. Furthermore, as regards the estimation of the material mechanical parameters, we proposed the evaluation of the Young's modulus from the sound wave velocity of wood. As a matter of fact, the Young's modulus is a fundamental parameter for mechanical simulations. The developed technique estimates the sound wave velocity from responses of the wood to an impulsive excitation in a rake receiver fashion. Successively, from the knowledge of the sound wave velocity, the Young's modulus is indirectly derived.

Lastly, we propose an EAR proof of concept through which we showcase the benefit of the proposed parametric approach to EAR. We display an EAR scenario in which two VSs, a VS replica of a prestigious violin, and a simulated generic model of the instrument are virtually co-located in a real sound scene with the presence of actual sound sources. The results give a sneak peek of the power of EAR, showing that the proposed parametric approach is able to provide the blend between real and virtual sound elements. Hence, we envision that the proposed solutions will pave the way to the development of parametric EAR frameworks for extended reality applications.

METRO NETWORK CAPACITY INCREASE VIA INNOVATIVE PHOTONIC TECHNOLOGIES AND COMMUNICATION SYSTEMS

Mariangela Rapisarda - Supervisor: Pierpaolo Boffi

In the next years, future Metropolitan Area Networks (MANs) will have to support a large range of resource-hungry applications and cope with requirements in terms of reduction of costs, energy consumption, and footprint. Several solutions to increase the transmission rates have been proposed. EU H2020 PASSION project (Photonic technologies for progrAmmable transmission and switching modular systems based on Scalable Spectrum/space aggregation for future aglle high capacity metrO Networks) is developing a solution consisting in the transmission of up to 160 25-GHz spaced channels in the C-band in the range between 191.9 THz and 195.875 THz. A modular approach in the design of the transmitters and of the receivers, the Sliceable Bandwidth/Bitrate Variable Transceivers (S-BVTs), provides flexibility and adaptability to the network conditions. The source constituting each Submodule of the transmitting part of the S-BVT is the Vertical-Cavity Surface-Emitting Laser (VCSEL), which is directly modulated with Discrete Multitone (DMT) signals. Thanks to its multicarrier nature, DMT allows to optimize the signal bandwidth depending on the conditions of the channel, giving more flexibility to the transmission. The MAN network is organized in a hierarchical topology. The focus is put on the intermediate Hierarchical Levels (HLs), the HL4 and HL3 nodes. In order to further increase the transmission rate, Space Division Multiplexing (SDM) can be employed by exploiting a bundle of fibers or MultiCore Fibers (MCFs). Since MAN networks cover distances in the order of hundreds of km, the accumulated Chromatic Dispersion (CD) heavily impairs the performance. This is the reason why Coherent Detection (COH-D) is preferred to Direct Detection (DD): the signal field recovery allows to compensate for the CD by Digital Signal Processing (DSP). This thesis partially falls in the objectives of PASSION project. In particular, the focus of the work is to analyze the potentialities of S-BVTs in MANs organized in hierarchical levels in terms of transmitted capacity. For this purpose, a MATLAB® simulation tool has been developed to emulate the behaviour of the main MAN network components: the VCSEL and its direct modulation with DMT format, the network nodes, the Standard Single-Mode Fiber (SSMF), and the coherent receiver.

In order to validate the complete simulator, a set of experimental measurements has been arranged. They imply the crossing of 1, 3, and 5 25-GHz spacing Wavelength Selective Switches (WSSs) filters for OSNR values of 30 dB, 35 dB, and 40 dB, direct modulation of a 18-GHz VCSEL and SSB DMT modulation. The target 50 Gb/s can be achieved for up to 5 WSS filters and 30-dB OSNR. A reach of 260 km can be achieved for 65-km span length, and 735 km for 35-km span length. The interplay between optical filtering and chirp parameters has been investigated for the same OSNR values and number of 25-GHz WSS filters with three different sources: two directly-modulated VCSELs, a 20-GHz short-cavity VCSEL and an 8-GHz tunable VCSEL, and a CW 500-kHz followed by an Electro-absorption Modulated Laser (EML). The short-cavity source has resulted to be a worse device with respect to the tunable one in terms of chirp parameters. The comparison among the sources has been performed for both DSB and SSB DMT modulation, resulting in the tunable VCSEL outperforming the short-cavity laser. The target capacity of 50 Gb/s is again achieved for both the VCSELs at 30-dB OSNR for SSB DMT modulation. DSB is instead heavily impaired by the interplay between chirp and tight filtering, which causes an average reduction of the transmitted capacity around 45%. All these results have been obtained for a single-channel transmission. An evaluation of the interchannel crosstalk has been performed for three 25-GHz spaced channels modulated with a 16-GHz and a 20-GHz SSB DMT signals, and the transmitted capacities have been compared with the ones achieved with single-channel transmission. When the channels are modulated with the smaller bandwidth signal, they suffer less from the interchannel crosstalk than with the wider one. Moreover, filtering the signal three times appears to be the best solution, since with two filters the interchannel crosstalk has a more significant impact, while a higher number of filters suppress the useful components of the signal. The HL4 and HL3 nodes have been deployed to evaluate the minimum OSNR required by the network to achieve a given target capacity (25 Gb/s, 40 Gb/s, and 50 Gb/s). For this purpose, every possible combination of 11 or less nodes has been investigated in presence of only linear effects in the fiber propagation. The impact of HL3 nodes in the performance is more significant than the effect of HL4 nodes, due to the tight filtering. When many of them are cascaded, part of the useful signal is suppressed, leading to higher requirements in OSNR to achieve a given target capacity. The values of required OSNRs have been achieved with linear propagation, so, in order to be applied to a real network, a margin should be taken. With this evaluation in mind, some network topologies have been evaluated by measuring the transmitted capacity for an optimized launch optical power and setting up different EDFA distributions along the network. A network composed of 2 HL4 nodes, 6 HL3 nodes (with only 2 "active" nodes and the other 4 in express path configuration), and 3 HL2/1 nodes (2 "active" nodes and the other in express path configuration) for 172-km SSMF propagation, a capacity of 51.19 Gb/s can be achieved with a launch power of 1 dBm, considering the effect of nonlinear impairments, and by aggregating some short spans together. A trade-off between the span length and the number of EDFAs to put in-line has to be found: if many short spans are cascaded, the accumulated nonlinear impairments, along with the noise introduced by many amplifiers, are detrimental for the performance, while few long spans cause a significant signal attenuation.

The crossing of a HL4 node based on Semiconductor Optical Amplifiers (SOAs) as wavelength blockers has been evaluated by experimental results. The employed polarization-insensitive SOA has been characterized in terms of its CW gain and TF for several bias currents and input powers. Then, the SOA response to DMT signal has been investigated. Two different sources have been employed: a DFB laser emitting at 1558 nm and a 1535-nm VCSEL. They have been modulated externally and directly respectively with both a DSB and an SSB 20-GHz electrical signal. In direct modulation, the effect of SOA high-pass filtering is more pronounced than in external modulation, with a reduction of 10 dB for low frequency subcarriers. On the other side, with external modulation of the optical signal, the amplifier introduces a Vestigial Sideband (VSB) filtering on the optical spectrum, which is more marked for high saturation conditions (i.e., for high bias currents and input powers). The complete HL4 node crossing, including the presence of the SOA, has been evaluated for both single and three-channel transmission, for propagation lengths of 16 km (only single channel) and 116 km (both single and multichannel), and coherent detection. The introduction of the SOA in the HL4 node has caused a reduction of 6.5 Gb/s in the transmitted capacity, which, however, exceeds the target capacity of 50 Gb/s for 116-km SSMF propagation and three 25-GHz spaced channels, provided the exploitation of 3rd order Volterra equalizer to compensate for the fiber nonlinear impairments and the interchannel crosstalk.

This thesis has demonstrated that S-BVTs deployed in hierarchical MAN networks with several node architectures can guarantee at least 50 Gb/s per channel per polarization in almost all the analyzed scenarios, including real network topologies and also in presence of SOA-based nodes.

A COMPREHENSIVE MODEL FOR THE INVESTIGATION AND PREDICTION OF DEPOLARIZATION ALONG HIGH FREQUENCY EARTH-SPACE LINKS

Eric Regonesi - Supervisor: Lorenzo Luini

Satellite communication systems working in microwave bands are impacted by the phenomenon of depolarization, induced by rain and ice particles present in the troposphere. Current models available in the literature do not fully characterize the large variability observed in beacon based depolarization measurements. Such models rely on the prediction of the rain contribution to the depolarization in terms of Cross Polar Discrimination (XPD). The ice effect, when present, is considered as an additive term whose statistical distribution is inherited by the one of rain, on an equiprobability basis.
The development of improved models requires thorough theoretical considerations and valid supporting data. Unfortunately, the availability of reliable satellite beacon data is scarce. Furthermore, the processing of experimental depolarization data is a nontrivial operation, which requires careful inspection, consistent filtering and reasonable hypotheses application. This work exploits the Italsat satellite propagation campaign carried out in the 90s, by means of the data recorded at the Italian sites of Spino d'Adda and Pomezia at three different microwave bands (Ka, Q and V). The results accomplished comprise a statistical analysis of depolarization and quasi-physical parameters, a novel model of ice induced XPD and a new methodology to scale dual polarized data to different link conditions.
First, the data processing is presented: data are carefully polished by removing the system biases and errors, and filtered to counteract the noise. For each event, a microphysical model is built by means of a novel method for the selection of the raindrop size distribution and slant path parameters.
Second, the data analysis is performed. The activity is carried out both on a statistical and an event basis. All the data are used to calculate the XPD, whose statistical analysis is presented with a direct comparison with models available in literature. For dual polarized signals, a further stage is carried out. The quasi-physical parameters, i.e. anisotropy and canting angle, are extracted and statistically assessed. This passage, joint to the developed microphysical rain model, permits the separation of the contributions by rain and ice particles to the overall electric field transfer matrix. The analysis is also shown through time series of a selected event. The effectiveness of the selection criteria and the signature of ice are highlighted.
Finally, a physical modelling part is presented. The ice integrated content is used to propose an ice induced XPD distribution model. Furthermore, with the aid of the data extracted from the ECMWF database, an average model relating the ice XPD and the ice integrated content is proposed, for the site of Spino d'Adda. These models differ from the current literature models by relying on the physical parameter of ice content, rather than being derived empirically by rain related quantities. Last, a methodology is proposed and validated to scale the overall transfer matrix to different link conditions, by separate processing on the previously extracted rain and ice contributions. This approach differs from methodologies present in literature by not considering the scaling as a whole operation. Rather, the overall transfer matrix is reconstructed after both the single particle families scaling.

HIGH ENERGY RESOLUTION X AND γ - RAY SPECTROSCOPIC SYSTEM BASED ON CDTE DETECTOR AND ULTRA LOW-NOISE AND HIGH SPEED FRONT-END ELECTRONICS

Martina Sammartini - Supervisor: Giuseppe Bertuccio

Co-supervisor: Bruno Garavelli

X and γ-ray spectroscopic systems with imaging capabilities are nowadays an essential tool employed throughout a large variety of fields from medical diagnostics, to homeland security and more in general wherever non-destructive testing (NDT) is desirable or needed. The more advanced are the applications, the stricter are the requirements that the spectroscopic system must satisfy. This is the case of XSpectra, a new patented technology developed by the Italian company Xnext, that allows to perform highly accurate and real time material characterization, which is of interest especially for quality control in food industry. The peculiarity of XSpectra with respect to state of the art is that it employs a multi-energy approach and divides in 1024 energy windows the acquired photons, instead of the only maximum two windows of conventionally employed scanners, which allows to obtain much richer and detailed radiographies, where also low density contaminants are detectable.

To guarantee such results, XSpectra is required to discriminate accurately photons from 130 keV to well below 20 keV, moreover XSpectra needs to handle very high photon fluxes (>106 mm-2s-1) which are employed to gather enough statistics, since in production lines objects are moved typically up to 1 m/s, and lastly operation as close as possible to room temperature is preferable, since cooling systems occupy space, are expensive and need maintenance. XSpectra, like many other advanced spectroscopic systems, employs detectors in Cadmium Telluride (CdTe), a compound semiconductor that, owing to its advantageous physical properties, allows to satisfy most of the requirements.

The predominant part of my PhD's research is carried out within this framework and its main focus is the study and the characterization of CdTe pixel detectors and the spectroscopic systems that employ them, trying to overcome the limits emerged from literature in the last ten years, in particular as far as the spectroscopy at high energy resolution is concerned, especially at low photon energy (<20 keV). One of the goals was to evaluate both their ultimate performance in terms of energy resolution and its dependence from some parameters of interest, such as the signal processing time, the internal electric field and the spectra acquisition time. The objective was to determine the limits to viable operation and to highlight the presence of peculiar phenomena that may affect the overall spectroscopic performance, such as the operation at deep sub-microsecond signal processing time and the effect of ballistic deficit on spectra quality. To this purpose, a research prototype spectroscopic system, based on the same CdTe detectors employed by XSpectra, but paired to a dedicated fast and ultra-low noise front end (the charge sensitive preamplifier SIRIO, developed in our laboratory), has been designed and developed, thus allowing to study the behavior of the detector at the fullest of its capability. Both the electrical and spectroscopic characterizations have been driven by the requirements of the application, therefore room temperature operation and high rate acquisition have been carefully examined. Moreover, owing to the advantageous features of the dedicated electronics, it has been possible to perform a study on the detector signals.

More in detail, the presentation of all the research carried out on the prototype has been organized as follows.

After a brief introduction on compound semiconductor detectors, in the first Chapter is reported an overview on the physical properties of the CdTe material used for detectors. The advantages of employing CdTe detectors with Schottky contact are presented and also a focus on the polarization effect is provided.

In the second Chapter are shown the results of the electrical characterization of the detector which was aimed at defining the limits for the subsequent spectroscopy tests. Besides the acquisition of the I-V characteristics, the isolation between pixels and pixel and guard electrode has been evaluated, as well as the influence of neighbor pixels to the one under test. Since reverse current transients highlighted the presence of the polarization phenomena, the effect of repetitive bias removal has been studied, to understand if it is an acceptable solution to the problem.

The results of the spectroscopic characterization constitute the third Chapter. After a brief recap of the energy resolution results in literature, the spectrum with the highest energy resolution obtained during the tests at room temperature is shown, along with its analysis that allowed to estimate the Fano factor for CdTe at room temperature, a parameter still not present in literature. Spectra at various bias voltages and changing the irradiated side are studied, thus allowing to gather insight on the physics of the detector through the evaluation of the charge collection efficiency and the derivation of the mobility-lifetime products. Energy resolutions at temperatures higher than the +20 °C have been evaluated and, addressing the detrimental problem of polarization effect, the limits to continuous acquisition time have been experimentally found.

In the context of another important requirement for advanced application, the high flux irradiation of the detector, the performance in the sub-microsecond peaking time range has been studied demonstrating the capability of the system to comply with the task.

In the fourth Chapter, the results of the study on the signals from the detector are presented. The shape of the experimental signals has been fitted on the basis of a theoretical model, allowing to gain knowledge on the transport properties of electrons and holes and on how they are affected by the peculiar physics of the detector. A crucial parameter of the model is the slope of the electric field, that is known to be affected by the polarization effect. Starting from the parameters extracted from the fit of the experimental data, the behavior in time of the slope of the electric field has been studied allowing to derive its impact on the percentage of the photons acquired. The conclusions on the polarization effect obtained through signal analysis are compared with the results from spectroscopy and they are found in agreement.

A review of all the results found during this work and their commentary is presented in the conclusions, where also are pointed out future perspective of this research.

DEEP LEARNING IN MULTI-STEP FORECASTING OF CHAOTIC DYNAMICS

Matteo Sangiorgio - Supervisor: Giorgio Guariso

In the last few decades, many attempts to forecast the evolution of chaotic systems, and to discover how far into the future they can be predicted, have been done adopting a wide range of models. Some early attempts were performed in the 80s and 90s, but the topic became more and more debated in recent years, due to the development of lots of machine learning techniques in the field of time series prediction.
Forecasting chaotic dynamics one or a few steps ahead is usually an easy task, as demonstrated by the high performances obtained on many systems. The situation dramatically changes when considering a long-term horizon because infinitesimal errors lead to a completely different evolution of the system even when one knows the actual model of the chaotic system. The most widely used prediction tools are artificial neural networks, which can be divided into those that present a feed-forward structure and those that include recurrent neurons. The first are static approximators capable of reproducing the relation between input and output, in principle with arbitrary accuracy.
When adopting these models, the forecasting of a chaotic time series over a multi-step horizon is commonly done by recursively performing one-step-ahead predictions (FF-recursive). An alternative consists of training the model to directly compute multiple outputs (FF-multi-output), each representing the prediction at a future time step. Both the forecasting methods have a weakness. The recursive one is optimized only to predict one step in the future. Thus its performance is not guaranteed on mid-long-term, in particular, when considering chaotic dynamics. The multi-output predictor takes into account the whole forecasting horizon: each neuron in the output layer focuses on the forecast of the considered variable at a different time step. The issue is that we are not able to specify that the outputs are sequential: the model acts as if the outputs were independent variables, rather than the same variable sampled at subsequent steps.
To overcome these critical aspects, it is necessary to adopt a neural model that is able to deal with the temporal dynamics of a certain variable: the recurrent neural networks (RNNs). Recurrent neurons (LSTM cell) have been demonstrated to be efficient when used as basic blocks to build up sequence to sequence architectures, which represent the state-of-the-art approach in many sequence tasks (e.g., natural language processing). The RNNs are almost always trained using a technique known as “teacher forcing” (LSTM-TF). It consists of using the ground truth as the input for each time step, rather than the output predicted by the network at the previous step. This technique is necessary when considering natural language processing related tasks, and it is currently always adopted even in numerical time series prediction. Training with teacher forcing does not allow the network to correct small errors because, during the training phase, the prediction at a time step does not affect future predictions.
We thus proposed to adopt a recurrent architecture and to train it without teacher forcing (LSTM-no-TF). Coupling these two elements solves the drawbacks of FF-recursive, FF-multi-output, and LSTM-TF. First, this structure is trained to reproduce the entire set of output variables. Second, it explicitly takes into account that these outputs represent the same variable computed at consecutive time steps. Third, small prediction errors propagate along the predicted sequence during training, and thus the training process should be able to correct them.
We tested the capability of the neural predictors on four chaotic systems: the logistic and the Hénon maps, the prototypes of chaos in non-reversible and reversible systems, and two generalized Hénon maps, a low- and a high-dimensional case of hyperchaos. First, the predictors have been trained on noise-free data generated by chaotic oscillators, without taking advantage of any physical knowledge on the systems. LSTM -no-TF predictors are able to couple the strengths of all the benchmark competitors, and provide the best performances on all the considered chaotic attractors (Fig. 1). The results obtained are robust because the predictors rank in the same way in all the chaotic systems.

The absence of noise is an ideal condition that is rarely verified when considering practical applications. We thus introduce additive white Gaussian noise on the signals obtained simulating the deterministic systems. A sensitivity analysis considering different levels of noise have been performed. As expected, the performances are considerably worse than those obtained in the noise-free case due to the chaotic behavior of the considered systems, which exponentially amplify the noise on the initial condition. This analysis confirms the ranking already obtained in the noise-free case: LSTM -no-TF turns out to be the best-performing architecture. Another test takes into account a modified logistic map, with a slow-varying growth rate. Testing the predictors on a slow-fast system is interesting because the task requires to retain information about both the slow-varying context (long-term memory) and the fast dynamics of the logistic map. The recurrent structure of the LSTM nets provides better predictive accuracy than feed-forward ones due to the LSTMs dynamic nature: they have an internal memory, and the values of their gates change at each step.
At last, we consider two real-world applications: solar irradiance measured in Como, and the ozone concentration in Chiavenna, Northern Italy. Both the time series exhibit a chaotic behavior (positive largest Lyapunov exponent). In general, the results confirm that the LSTM without teacher forcing outperforms the competitors. However, the ranking seems to be more system-dependent than that obtained with the artificial systems. For instance, the LSTM with teacher forcing provides the worst performance on the solar irradiance dataset. Another interesting result is that the feed-forward multi-output net reaches comparable (though still worse) forecasting accuracy of LSTM without teacher forcing in both the time series.
Besides the forecasting accuracy, another essential characteristic of the forecasting models is their generalization capability, often mentioned as domain adaptation in the neural nets literature. It means the possibility of storing knowledge gained while solving one task and applying it to different, though similar, datasets. To test this feature, the neural networks developed to predict the solar irradiance at the Como station (source domain) have been used, without retraining, on other sites (target domains) representing quite different geographical settings. The predictors developed in this study have proved to be able to forecast solar radiation in the target domains with a minimal loss of precision.

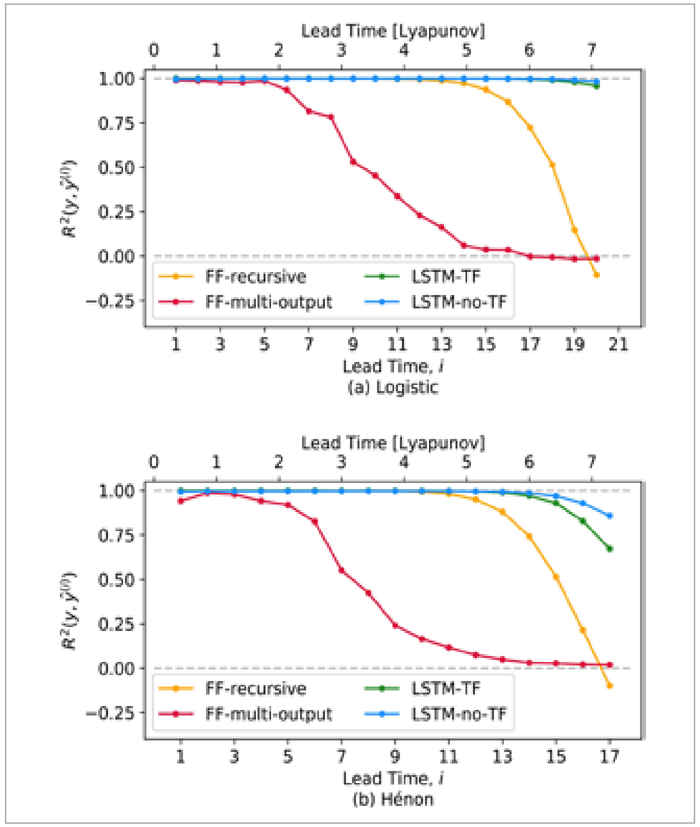

Fig. 1 - Forecasting accuracy of the neural predictors on the Logistic (a) and Hénon (b) maps.

MODELING, CONTROL AND AUTOMATIC CALIBRATION OF A SEMI-ACTIVE SUSPENSION SYSTEM FOR HIGH-PERFORMANCE CARS

Gianluca Savaia - Supervisor: Sergio Matteo Savaresi

The automotive industry is undergoing disruptive times. The traditional mobility patterns based on private ownership, fossil fuel and general-purpose (*i.e.*, big) dimensions are eventually doomed to be eradicated from our culture in favor of the revolution carried on by autonomous, shared, electric and tailor-made vehicles.
Aware of this new vision, most car manufacturers offer their customers the possibility to purchase mobility rather than a private vehicle because it grants more flexibility and fewer costs. This trend is bounded to escalate as autonomous driving becomes more reliable, to the point in time when nearly no one will own a car, preferring to pay a ride on a *robo-taxi* to reach their destination. This scenario will have an extreme impact on the automotive field since it has been shown that when sharing meets autonomous technology, the demand for vehicles will significantly drop. Today already, vehicles in the US spend 96% of their time in a parking lot, and an autonomous fleet in Shanghai could meet the mobility needs of the entire population with 1/3 of the current vehicles.
The only cars which are not expected to be overwhelmed by the autonomous and shared revolution are the supercars, which will survive as we know them today. There will always be a market for the Lamborghinis and the Ferraris because they are not really in the mobility business but rather in the entertainment business.
In the future, you will likely ride either a *robo-taxi* or a sports-car. Whatever the category, the car will be a high-performance vehicle: the sports-cars already are, and the robotaxis will have to be since they will need to drive more kilometers, faster and more comfortably.
In this thesis, we focused on comfort, which is mostly affected by the suspension system. Semi-active suspensions outperform passive systems because they can modulate the damping coefficient with respect to the driving scenario; the objective of electronic control consists in designing an algorithm that can conjugate comfort and vehicle stability, two factors which must be compromised in a passive suspension where the damping coefficient is constant.
Semi-active suspension systems are most often preferred over the more advanced active suspensions since the latter are based on electric motors, which require a significant amount of power to work at high bandwidth. In contrast, semi-active suspensions modulate the damping coefficient without introducing any active force.
Despite the dissipative constraint, a semi-active suspension can significantly improve comfort and handling simultaneously: softening the suspension guarantees good filtering of the road asperities, whereas increasing the damping coefficient improves the control of chassis movements.
Electronic suspensions give the possibility to shape the vehicle dynamics; designing the chassis' movements is equivalent to designing an emotion: the feeling of the driver along his ride.
In a sports-car, the electronic suspensions play a major role in shaping the vehicle dynamics. I like to think that my work's objective consists in shaping the emotion of the driver behind the wheel.

Main Contributions

The investigation starts, in the first part of this document, with the analysis of a magneto-rheological (MR) damper; this technology has been chosen because it is the most promising and effective way to control the damping of a semi-active suspension in a high-performance vehicle. The literature on this device is plentiful, but the author believes there exists a lack of contributions aimed at modeling the MR damper for control purposes.
For this reason, the proposed contribution consists of a control-oriented model for a MR damper based on a hammerstein-wiener scheme. The model considers all of the most significant phenomena impacting the damping force which can be exerted by the device: nonlinear characteristics, hysteretic behavior, magnetization dynamics and temperature effect.
In particular, to the author's knowledge, the investigation on the magnetization dynamics (*i.e.*, actuation) has never been examined before. Furthermore, the proposed model has been validated on a real damper mounted on a testbench equipped with a load cell which could measure the actual damping force.
The second part of the document is devoted to the design of control strategies for the improvement of comfort in the vertical dynamics.
A hierarchical architecture is proposed which aims at achieving the boundaries of performance for both low and high-frequency road disturbances.
In this architecture, a force tracking controller serves as an interface between the chassis motion and the MR damper; this layer decouples a force request coming from a higher level of abstraction to a low-level current setpoint to the damper.
Following this philosophy, the high-level logic only focuses on the chassis motion regardless of the underneath hardware, whereas the low-level logic consents to assign a desired shape to the damping characteristics other than the original one intrinsic of the MR damper.
The high-level control logic builds upon a variant of the skyhook logic; the main novelties proposed in this regard consists of: (i) a control strategy tailored to dampen the pitch motion of the chassis by feedbacking the measured pitch rate and distribute the reference damping accordingly to the geometry of the vehicle; (ii) an enhanced skyhook logic derived via machine learning techniques which improves comfort and stability of the chassis when subject to low-frequency excitation.
All the reported results have been validated on an actual sports-car equipped with MR dampers on a diverse set of road profiles exciting a broad range of frequencies. The validation has been performed by means of objective data measured by the on-board sensing units and subjective feedback of the test driver.
In the last part, we propose a novel framework for the automatic calibration of a semi-active suspension control software. The current paradigm consists of a calibration engineer tuning the parameters given the subjective feedback of an experienced test driver. Aiming at automating this process, we came up with a fully self-calibrating methodology where the vehicle performance is evaluated by the sensing unit (in place of the driver) and the control parameters are tuned via an algorithm based on Bayesian Optimization.
The core motivation is that a simulation environment, however sophisticated, cannot reproduce with high accuracy the dynamic phenomena occurring on the real vehicle; hence, the optimal calibration cannot be searched in simulation.
On the other hand, performing experiments on a real vehicle is economically expensive and time consuming; for this reason, it is important to design a methodology which guarantees to reach the global optimum with as few experiments as possible.
The complexity of the optimization problem is tackled in two phases: (i) in simulation, the control parameter set is reduced by identifying those parameters which affect the most the vehicle dynamics; (ii) on the actual vehicle, this subset of parameters are calibrated via repeated experiments.
The simulation phase is important to reduce the number of parameters to be optimized on the real vehicle, which translates to fewer experiments needed to reach the optimum. The experimental phase consists of repetitive experiments which are evaluated via a performance index computed on measured data; the proposed algorithm can automatically suggest a new calibration at each iteration, rapidly converging to the optimum.
The proposed methodology has been validated on a commercial vehicle equipped with a proprietary software for the control of semi-active suspensions. Eventually, the experiments have been performed on a proving ground with a test road containing a diverse set of road disturbances. The improvement on vehicle performance perceived in terms of comfort and vehicle stability between the initial and optimal calibration is significant.

CONTROL ORIENTED LEARNING IN VEHICLES

Dario Savaresi - Supervisor: Simone Formentin

Co-supervisor: Sergio Matteo Savaresi

The raising demand for more performing vehicles has led the research towards smart solutions, both from the mechanical and the control system design point of view. In many practical cases, however, the availability of all the measurements required for the control is not guaranteed. Estimation is therefore the solution to cope with such a limitation, in order to achieve new control objectives. Dealing with three categories of means of transport (heavy-duty, cars, and bicycles), the goal of the thesis is to propose some examples of learning and estimation in any of the most meaningful research lines in vehicles.

One of the most classical and important research topics in vehicles involves parameter estimation, because it has benefits in several fields, such as in emission reduction, fuel efficiency improvement, safety enhancement, and drivability. A second research line is about fault detection and recovery. Tightly linked to the fault detection and recovery field, particular interest raises when dealing with safety critical systems. Lastly, estimation plays a fundamental role also in signal reconstruction, acting as virtual sensors were deployed on the vehicle. In all these cases, estimation gives rise to new control objectives, setting the trend in vehicle enhancement.

The thesis proposes novel contributions in each one of the four aforementioned estimation research lines, to the purpose of show how learning and estimation algorithms can be adopted to improve vehicles' control strategies, thanks to a meaningful example for each major topic. In the parameter estimation field, a solution has been provided as far as the agricultural tractor mass estimation problem is concerned. Vehicle control logics can be tailored to each load condition, with many purposes: minimizing fuel consumption, estimating contact forces for ABS, calibrating controllers for electronic suspensions. Among the others, it is particularly of interest dealing with agricultural tractors, because, as examples of heavy-duty vehicles, their payload may significantly vary, even up to three or four times the unloaded vehicle configuration. Furthermore, for such a class of vehicles, mass estimation is exceptionally challenging, because, in the majority of the cases, two masses are linked together: the tractor and the trailer. This peculiarity, that reduces the possible estimation techniques, has to be explicitly taken into account in designing the estimation algorithm. Thanks to the preliminary wheel torque model identification, the tractor's mass is correctly retrieved. Related to the fault detection and recovery field, two algorithms for the semi-active damper state of health monitoring are proposed. These are the first two cases in the scientific literature in which this problem is faced also from a fault recovery perspective. Semi-active suspensions' behavior, in fact, may continuously drift away from the nominal one while driving. As such, fault detection and recovery algorithms are desirable to be implemented. Usually the damping force is always regulated acting on a current, that in turn is directly related to the damping ratio, through a current damping map. This map is usually provided in nominal conditions by the manufacturers or it can be experimentally obtained on a test bench. Throughout the life of the vehicle, however, this map may vary. As a consequence, the current imposed to the shock absorber results in a damping ratio that is different from the desired one. Among those causes directly related to the damper (thus disregarding possible faults in the control loop), it is worth mentioning: wear, aging, temperature, and component manufacturing dispersion. Concerning this latter case, even a brand-new damper is not characterized exactly by the maps provided in the data-sheets, because the manufacturing process is not perfectly repeatable and the components employed are not ideal. The phenomena here described represent a fault of the shock absorber itself and possible

compensation techniques may be adopted to mitigate any undesirable behavior. Thanks to the proposed algorithms, the damper current, under the nominal functioning hypothesis, can be accurately estimated. The deviation of such a current from the imposed one is an indicator for the suspension's state of health. In the safety critical application area, instead, with a focus on a Brake by Wire bicycle application, theoretical bounds for a DC Brushless motor to be drivable have been found. Then, they have been related to motor estimated parameters and working conditions, to prove that a fault recovery mode is often possible, quantifying also the performance losses through an additional efficiency, function of the rotor angle measurement error. The interest lies here in what to do if a fault in the Brushless DC Motor (BLDCM) that actuates the brakes happens: whether the motor is still drivable (and thus the braking can be correctly performed), or not. In today's applications, in fact, BLDCMs are widespread and popular with respect to their brushed counterparts, but they need a control system to drive the current, based on the actual rotor position. For a standard three-phase motor, such a measure is usually obtained by including a set of three Hall sensors. If a sensor fails, it is mandatory for the vehicle safety's sake to know if minimal functionalities are still ensured, possibly providing also an estimation for the performance losses. Finally, dealing with signal reconstruction, an indicator for the optimal gear shifting in bicycles has been obtained, providing an estimation set-up that is automatically able to learn and retrieve all the missing information from the system, potentially without the need of any external intervention. Thanks to this estimation, new frontiers in the bicycle comfort problem have been investigated. Bicycles have been increasing their usage and popularity since the last years and they are becoming the main mean of transport for short-range mobility. Beside the recreational usage, their growing number allows to mitigate the impact of atmospheric and acoustic pollution as well as traffic- and parking-related issues in big cities. To foster the use of bicycles, governments are thus connecting cities with new cycle routes. Inspired by the evolutionary trend that vehicles underwent in the last decades, the external comfort problem (i.e. the comfort with respect to the surrounding pedaling environment) has been recently taken into account by researchers. Two other major investigation topics are represented by the riding safety and the energy management for electric bicycles. On the other hand, however, the internal comfort problem should be taken into consideration, too, as the result of the interaction between the bicycle and the rider, that this time is an active agent in the (dis)comfort generation. In this case, comfort is primarily related to the amount of torque exerted on the pedals during gear shifting. In those moments, the chain is instantaneously re-engaged on the sprocket, with a sudden variation of the vehicle velocity and acceleration spikes. The problem of optimal gear shifting should be taken into account especially for non-professional cyclists, whose shifting rationale is seldom compliant with the optimal one. To account for this issue, an automatic gear shifter can be adopted. Such an actuator, however, needs an estimate for a shifting flag signal to perform properly, which is provided by the algorithm proposed. The comfort improvement, obtained through the estimated signal, is significant if compared to a standard pedaling condition.

MACHINE LEARNING-DRIVEN INTEGRATION, KNOWLEDGE EXTRACTION AND UNCERTAINTY MANAGEMENT FOR SCIENTIFIC DATA

Gabriele Scalia - Supervisor: Barbara Pernici

Managing scientific data, such as those found in the chemical and biomedical research, poses unique and challenging problems. Unique features characterize this data, including the impossibility of representing reality on a one-to-one scale, the imprecision in the observations and quality limitations introduced by technologies and models that continuously evolve.

This is an interdisciplinary research that, as a whole, investigates the management and the analysis of scientific data focusing on the challenges emerging in fields such as chemistry, genomics and biomedical research. For this work, we focus on machine learning techniques to face a set of identified requirements: 1) the management of uncertainty for complex data and models such as deep neural networks, 2) the estimation of system properties starting from imprecise, low-volume and evolving data, 3) the continuous validation of scientific models through large-scale comparisons with scientific data and 4) the unsupervised integration of multiple heterogeneous data sources related to different technologies to overcome individual technological limitations. Common to virtually all fields driven by experimental data, these requirements are faced through a set of case studies on different applications in chemistry, biology, and genomics, focusing on machine-learning driven solutions.
Uncertainty estimation and evaluation is investigated in the context of deep neural network (DNN)-based *molecular property prediction* (Figure 1). Advances in this field have recently led to the development of models which can efficiently learn properties based on the complex structures of molecules, with *graph convolutional neural networks* (GCNNs) reporting state-of-the-art performance for this task. One of the main challenges currently hindering the usage of these models in applications such as drug discovery is *uncertainty quantification*. Experimental data are always characterized by some level of intrinsic variability and imprecision. Models trained on those data inherit that uncertainty and are also affected by another kind of uncertainty that comes from insufficient training samples, often difficult to quantify, especially for complex models such as DNNs. In this work we investigate how both kinds of uncertainties can be modeled in DNNs, developing scalable approximate Bayesian GCNNs for molecular property prediction. Furthermore, we propose a framework to qualitatively/quantitatively evaluate resulting uncertainty estimates and to assess what they really capture about the model and the dataset. Extensive experiments are carried out on major public datasets. The analysis of the results leads to a better understanding of the role of uncertainty in this context and paves the way to its usage in applications.

We investigate the problem of estimating system properties starting from imprecise and low-volume data in the context of biological systems characterized by some kind of internal information exchange. Such exchange can be described from a communication point of view, allowing defining a *capacity* for the channel. We face the problem of estimating this capacity only using a set of inputs/outputs for the system (e.g., in-silico/in-vitro experiments).
We design a novel methodology that frames the estimation of the capacity as the optimization problem of finding an upper and a lower bound on the true value. These bounds are optimized starting from the data and using an *evolutionary iterative algorithm*. Being estimated

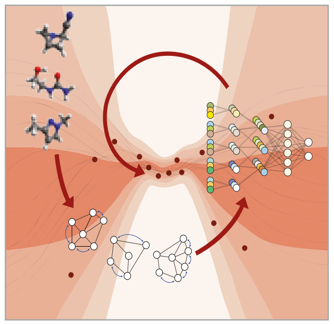

Fig. 1

from the data, the result is affected by the uncertainty and the volume of the available data. Therefore, particular emphasis has been placed on overcoming data scarcity and managing uncertainty. For this, and since gathering new experiments is usually costly and time consuming, the proposed methodology is equipped with a deep learning-based *data augmentation* module.
The proposed methodology is experimentally evaluated on a system composed of two prokaryotic cells.

Next, the design of a framework to support the development of *scientific models* through the *continuous validation* on integrated scientific experiments is presented, with the discussion of an architecture and the development of a prototype.
In many scientific settings, models are developed independently and externally with respect to

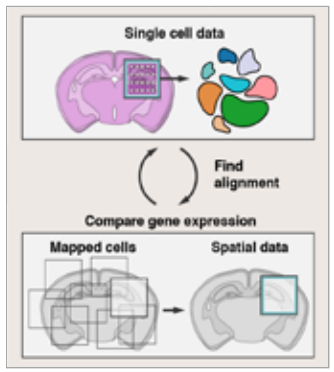

Fig. 2

the collected experimental data. Nonetheless, *experimental data* have a key role in the development/validation/refinement cycle. We discuss the design of a framework to manage scientific models along this cycle, taking advantage of large amounts of scientific data (experiments) extracted from the literature. This includes managing changes in models (which need to be validated against existing data), cross comparisons of models and experiments to find global patterns, and the automatic acquisition of new experiments (which triggers the validation of existing models). These activities are hindered by uncertainty in models and experiments, and by the need to automatically interpret and integrate available data. As a case study we consider chemical kinetics models, but the approach taken is general and domain-agnostic. Finally, the *unsupervised integration* of heterogeneous data sources with different features and characterized by a varying quality is explored, proposing a methodology to integrate datasets measured with complementary transcriptomics technologies (Figure 2). As a single technology for measuring spatially resolved whole transcriptomes of single cells does not currently exist, we show how this result can be achieved through integration of multiple, complementary technologies. Such outcome can enable integrated, multi-scale *biological atlases* at single cell resolution of organs, that is currently recognized as next frontier to understand cellular basis of health and disease.
To this end, we develop *Tangram*. Technically, Tangram is based on learning an *alignment function* that allows obtaining a *mapping*, through which complementary transcriptomics datasets are integrated.
Through a large set of experiments using various datasets measured with different technologies, we show how we can improve the resolution, the throughput, the quality and the available modalities of the starting datasets via integration. A final case study shows how we can learn a histological and anatomical integrated atlas of the somatomotor area of the healthy adult mouse brain starting from publicly available datasets.

DEMAND-DRIVEN TIMETABLING OPTIMIZATION FOR AUTOMATED METRO LINES

Tommaso Schettini - Supervisors: Federico Malucelli, Ola Jabali

With each passing year, the world is becoming increasingly urbanized. As of 2018, 55% of the world's population lives in urban areas, according to a report of the United Nations. With a world population of 7.63 billion people, this equates to 4.22 billion people who live in cities. As a result of this increased level of urbanization, major cities are becoming increasingly reliant on mass transit systems to satisfy their growing travel demand.

In particular, rail-based mass transit systems are becoming an increasingly vital part of the transit infrastructure, as they provide a reliable, high-speed, high-capacity, and sustainable means of transportation in congested, high-density areas. However, unlike road-based transit systems, rail-based systems are heavily restricted by their infrastructure and cannot be easily scaled up to meet increased demand. Thus, efficiently managing those systems has become a top priority to local authorities.

Traditionally, metro management is performed mainly at a tactical planning level, through a strictly hierarchical approach. Initially, each day of the year is categorized under a set of typical days, e.g., winter-workday, winter-festive, summer-weekend. This categorization aims at capturing the variations of demand observed throughout the year. Each day is then further divided into several time periods, such as morning peak and off-peak. For each time period, a service frequency is determined based on the historical demand to guarantee a certain level of service while minimizing costs. Lastly, based on these frequencies, the scheduling of trains and crew is established. This hierarchical approach, while functional, is rather inflexible. Consequently, the necessity for efficiency pushes towards the development of novel control paradigms to optimize the usage of the available resources and improve the service quality. To this end, two recent trends in metro systems are of particular interest.

The growing adoption of smart card-based automated fare collection (AFC) has lead to increased availability of passenger demand-related information. With these systems, it is possible to extract detailed journey information on a user-by-user basis and obtain high-fidelity Origin-Destination demand information for the network at a high temporal resolution. The current availability of demand data is in stark contrast to previously available methods for estimating demand, which required laborious and expensive surveys. As temporal Origin-Destination demand data becomes more readily available, more accurate demand predictions can be achieved. Thus, the prospect of demand-driven control paradigms becomes an attractive possibility to exploit the available data to their fullest.

With the advances of automated control systems, it is becoming more common for metro lines to be automated. From a control perspective, automated systems provide several advantages to transit systems and pose interesting new optimization challenges on how to fully harness their benefits. Indeed, in the timetabling of traditional lines with drivers, issues related to the crew scheduling can be particularly challenging and impose highly binding constraints on the scheduling of the trains. Automated transit systems remove the need for personnel scheduling and rostering, thus allowing for more freedom in their control strategies. Such freedom can be leveraged to employ more flexible train schedules, which can specifically be tailored to the passenger demand.

In this thesis, we develop novel control strategies for automated lines aimed at exploiting their benefits to achieve a fully demand-driven timetabling approach and improve the passengers' service quality. In particular, we deviate from established methods in two significant ways.

Firstly, we do not impose any structural constraints in the definition of the scheduling of the trains. Instead, we let the model organically determine the structure of the schedule to match the demand and achieve optimal service quality. To this end, we will also use a time-dependent description of the demand. Secondly, we allow the usage of short-turning in the scheduling of the trains. By using short-turning trains are not required to serve the line from terminal to terminal. Instead, the trains may reverse their direction before reaching the terminal of the line, effectively performing a short cycle. By doing so, we can shape the schedule to fit the distribution of demand on the line and offer increased frequencies to high-demand stations and lower frequencies at low-demand stations. In metro lines, short-turning is frequently utilized to handle service disruptions. However, its usage in metro lines is rarely systematic and is commonly achieved through the operators' experience.

We introduce an alternative demand-driven control paradigm, called Direct Timetabling. According to this paradigm, trains are controlled individually rather than imposing a predetermined structure on the entire train schedule. Additionally, short-turning decisions are jointly optimized with the schedule of the trains. This allows far more flexibility in the control decisions of the line, which is used to adapt to variations in passenger demand and yields a better service.

The flexibility granted by direct timetabling is matched with a considerable increase in the complexity of its associated optimization problem. Indeed, when controlling the trains individually and allowing short-turning decisions in the optimization, the spectrum of decisions associated with direct timetabling is far more complex than traditional approaches. Therefore, we will develop efficient exact algorithms based on cut generation for a general timetabling setting. We then present extensive computational results to evaluate the effectiveness of the cut generation algorithm and analyze the solutions obtained, comparing the optimized schedules to regular timetables, evaluating the benefits of the direct timetabling paradigm.

Then, we consider a special case of direct timetabling, which can be more efficiently solved under mild assumptions on the structure of the train schedule. Specifically, we assume the presence of a root station in the line, which cannot be skipped by short-turning, and we assume that idling is only allowed immediately after a short-turn, and for a maximum of a given number of time-steps. We present an alternative formulation for the problem and an efficient exact algorithm based on Benders decomposition. Through the computational experiments, we demonstrate the effectiveness of the developed algorithm and evaluate the impact of the assumptions made on the structure of the schedule.

Lastly, we develop a direct timetabling strategy to handle the demand associated with a large event (e.g., a football match or a concert), which is served by a single station of a metro line. Such events cause a significant increase in demand at a single station. Thus, operating the ordinary timetabling of a line may lead to prolonged passenger waiting times. We develop a formulation for the problem and an efficient iterated local search heuristic. Through the computational experiments, we evaluate the effectiveness developed heuristic and show the utility of the proposed paradigm over an ordinary timetable.

MULTICHANNEL CMOS INTEGRATED CIRCUITS FOR BROADBAND SPECTROSCOPY APPLICATIONS

Giuseppe Sciortino - Supervisor: Giorgio Ferrari

Advanced spectroscopy techniques are more and more gaining the interest of the research community because find outspreading applications in many different fields, like industrial quality control, tumor analysis and new materials engineering. The thesis contributes to this field by designing and operating Application Specific Integrated Circuits (ASICs) for fast, yet highly resolved, acquisition of spectra, obtained with two powerful optical techniques, namely broadband Stimulated Raman Scattering (SRS) microscopy and time-resolved Pump&Probe spectroscopy.
The first ASIC was developed in the framework of VIBRA (Very Fast Imaging by Broadband Coherent Raman), an ERC H2020 project coordinated by Prof. D. Polli of Politecnico di Milano. The challenge of the ERC project is to develop a Stimulated Raman Scattering (SRS) microscope, that in real time can distinguish different chemical species, with sub-micrometer spatial resolution. The microscope finds its main application in the histological field, allowing a fast, safe, contactless, and repeatable tumor analysis. To reach such a result, a multichannel ASIC implementing a pseudo-differential readout frontend has been designed and integrated in a complex electronic acquisitions system.
The chip is a 4-channel and lock-in based differential readout frontend, able to measure down to 10 ppm Stimulated Raman Scattering (SRS) signal over a wide range of input optical powers (50 µW - 600 µW per channel) and with a pixel dwell time of only 100 µs. The main noise contribution, given by the laser intensity power fluctuations, is rejected as a common mode signal, thanks to the pseudo-differential measuring approach, based on the acquisition of a Signal and a Reference optical beam. A novel nonlinear Automatic Gain Control (AGC) network compensates all the possible optical mismatches between the Reference and the Signal optical paths, that could lead to a relative compensation error as high as ±30%, thus ruin the laser noise rejection. Thanks to the auto-balancing network, the Refence and Signal beams are automatically equalized in less than 10 us, allowing the laser intensity noise suppression. The result is a shot noise limited measurement (SNL), leading to a pixel dwell time reduction of a factor up to 225, respect a simple differential acquisition stage. The chip has been designed and fabricated in AMS 0.35 µm technology (Fig 1a) and it was tested in the optical and electronic laboratory, for the functionality and performances validation. Eight identical ASICs, combined in a complex electronic system (Fig. 1b), have been so integrated in a broadband SRS microscope prototype (Fig. 1c). The instrument has been used to discern polymers with sub-micrometres resolution (Fig. 1d), acquiring up to 32 SRS wavelengths in parallel. The fast pixel dwell time (100 us), combined with the multispectral resolution capability, allows to perform chemical fingerprint with up to 1 frame per second (fps) speed, thus achieving the final goal of the VIBRA project.

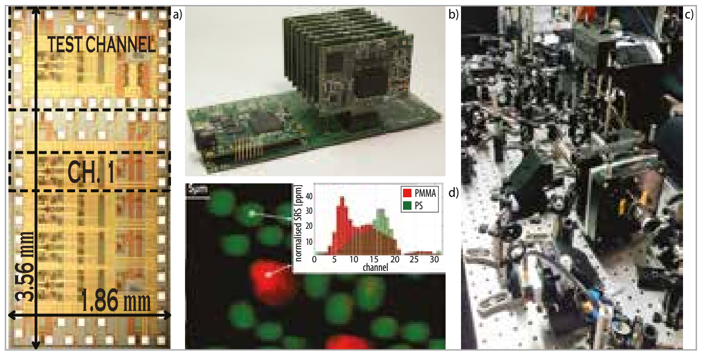

Fig. 1 - (a) Micrograph of the 4-channel ASIC designed for SRS spectroscopy in 0.35µm technology. (b) 32 channel PCB including 8 ASICs working in parallel and included in (c) the custom SRS microscope. (d) SRS image of few micrometers beads of polystyrene (PS) and polymethyl-methacrylate (PMMA).

The second framework of the thesis has been dedicated to the development of a ground-breaking spectrometer for ultra-fast Pump&Probe spectroscopy. The limitations of the current instruments, in terms of speed and resolution, will be overcame using fast repetition rate lasers (up to 1MHz) and a parallel acquisition and processing of many wavelengths (up to 40). These improvements are made possible thanks to a very flexible 20-channels custom ASIC, named ECLIPSE, specifically designed for Pump&Probe spectroscopy. The chip is switched capacitor technique based and can extract the useful optical information with few ppm resolutions. It was designed and fabricated in TSMC 0.18µm technology (Fig. 2a) and a PCB testing board has been designed too, including firmware and the Graphic User Interface software for data saving and visualisation. The platform permits a full characterisation of the device and can be used also as the main core of the future instrument.
From preliminary measurments performed in the electronic laboratory, the chip matches perfectly the project specifications (Fig. 2b), paving the way for the future next generations of Pump&Probe spectrometers.

I acknowledge support by the European Research Council Consolidator Grant VIBRA (ERC-2014-CoG No. 648615) - Very fast Imaging by Broadband coherent RAman.

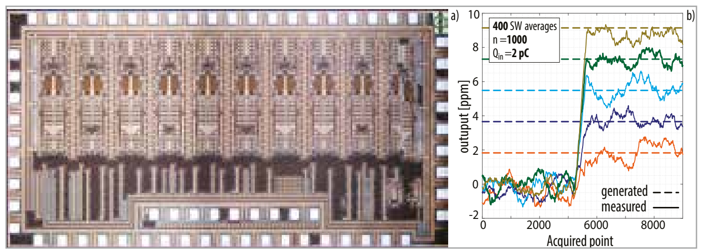

Fig. 2 - Fig. 2 (a) Micrograph of ECLIPSE, the custom ASIC specifically designed in 0.18µm technology for high resolved Pump&Probe spectrometers. (b) Preliminary validation of the maximum ASIC achievable resolution (few ppm).

OPTIMIZATION AND CONTROL OF SMART THERMAL-ENERGY GRIDS

Stefano Spinelli - Supervisor: Marcello Farina

Co-supervisor: Andrea Ballarino

This thesis deals with the development of novel algorithms and methodologies for the optimal management and control of thermal and electrical energy units operating in a networked configuration. The transformation of the energy and utility industry is characterized by a transition from centralized to distributed generation, that requires the introduction of new paradigms for energy management and control. Multiple and integrated energy vectors - i.e., electrical, thermal, etc. - must be considered together, holistically. On-site generation, on the other hand, complements the standard utility sources, enabling a symbiotic integration of central and localized energy production and distribution. This enhances the flexibility of the generation, as well as the complexity of the overall system. The smart thermal-energy grid is a large-scale networked system, where a set of common resources are shared by the producers and where the main objective is to sustain efficiently the time-varying demand of different forms required by a set of consumers, providing the optimal scheduling and the economic dispatch of the units. The integrated multi-utility configuration requires also a dynamic control of the operating point of each unit, considering the interaction among the subsystems and the fluctuation of the demands.
The work aims to foster the creation of a smart thermal-energy grid (smart-TEG), by providing supporting tools for the modeling of subsystems and their optimal control and coordination.
The configuration and the dimension of the problem intrinsically pose the main issue of its tractability with standard centralized approaches. Therefore, the distribution of control intelligence is the key point to reach a plant-wide dynamic optimal control. Hierarchical and distributed schemes are proposed in this thesis to address optimally the management and control issues of the smart-TEG.
The main types of generation units composing a smart-TEG are studied and modelled: different steam generation technologies, steam accumulation systems, as well as combined heat and power units (i.e., reciprocating engines, gas turbines and steam turbines) are introduced. For each unit, detailed first-principles models are discussed: these dynamical models – combined together – are used for the closed-loop validation, and additionally for the derivation of simplified models required for the various control schemes. Moreover, in the perspective of the unit commitment problem, a high-level model of each unit is derived, based on the discrete hybrid automata framework, which considers all the operating modes of the system and a static map in each mode.
Regarding the dynamic control of single units and coupled sub-systems, Model Predictive Control solutions are designed, discussed, and different approaches are investigated. For the fire-tube boiler, both a linear and a multi-rate MPC schemes have been developed; in addition, a nonlinear MPC has been proposed for its optimal start-up trajectory. A similar scheme is applied to the regulation of a water-tube steam generator integrated with a steam accumulator. In this work, the nonlinear MPC is implemented via a linear-parameter varying MPC, linearizing online the system along the optimal trajectory. The control scheme is specifically designed – e.g., in the start-up scenario – to reduce the prediction horizon, and therefore the computational burden, by the introduction of an artificial steady-state target as an additional decision variable.
The scheduling problem is also tackled with optimal receding-horizon model-based schemes: in this context, the presence of continuous and Boolean decision variables implies the solution of mixed-integer programs.
This research work is supported by an industrial case-study for the validation activities: the hierarchical architecture for unit commitment and dynamic control has been designed abstracting the main requirements from the actual case-study. The solutions are in general validated in a simulated environment, derived and identified from it, while the high-level scheme has also been tested on the field.
Furthermore, starting from this, two extended case-studies are conceived to investigate innovative control schemes. The first considers a network of similar generation units operating in a parallel configuration and in a scenario with limited shared resources: the steam generator of the industrial use-case is replicated to compose the ensemble case-study.
A hierarchical architecture for the unit scheduling and dynamic control of the ensemble is proposed, aiming to control it in a scalable and flexible way: an aggregated model of the ensemble is dynamically constructed based on the sharing factors, which represent the relative contribution of each unit to the network. These shares are adjusted in receding horizon by a top-tier optimization; this is integrated with a medium-layer robust MPC that drives the aggregated model of the ensemble towards the target. At the lower level, decentralized closed-loop systems are considered.
A second extension of the case-study regards the overall unit commitment of the smart-TEG, that is a network of heterogeneous generation units. It integrates different energy vectors, e.g., electricity and thermal energy. The main difference with respect to the electrical smart grids is related to the additional directionality constraints, induced by the steam distribution network. Consequently, a hierarchical configuration of the generation units must also be considered in the smart-TEG. It discriminates among central units, able to supply different production units on a wide area, and local generation units, which represent the on-site generation systems locally installed on production unit premises, that cannot support the demand of remote or of higher-level units.
The scheduling and economic dispatch of this network is a complex mixed-integer optimization problem that requires appropriate strategies to guarantee scalability, flexibility and to preserve the privacy of the different agents involved. This network can be indeed partialized into several agents, each representing an atomic subset of generation units with computational and control capabilities, as well as, decision autonomy. All of them must cooperate to supply a cumulative demand.
While a centralized approach can guarantee the best optimality result, the reliability and the flexibility of the system is compromised, as well as privacy. To overcome these issues hierarchical, decentralized and distributed optimization strategies are studied and proposed for the unit commitment problem.
Cutting-edge and state-of-the-art non-centralized schemes are implemented for the smart-TEG problem, requiring a deep study on the proper formulation of the unit commitment problem and subsystem models. Moreover, novel approaches for the performance improvement of these non-centralized schemes are proposed in this work.
The enhanced schemes, relying on the feasibility guarantees inherited from the standard approaches, permit to obtain in decentralized (and distributed) context an optimality level comparable with the centralized approach.
The work provides a wide analysis of control solutions for a smart thermal-energy grid to enable efficient utilization of multi-dimensional energy systems, developing control algorithms and tools at different levels.

EXPLOITING STRUCTURE FOR TRANSFER IN REINFORCEMENT LEARNING

Andrea Tirinzoni - Supervisor: Marcello Restelli

Recent advancements have allowed reinforcement learning algorithms to achieve outstanding results in a variety of complex sequential decision-making problems, from playing board and video games to the control of sophisticated robotic systems. However, current techniques are still very inefficient, in the sense that they require a huge amount of experience before learning near-optimal behavior. The main reason behind this limitation is that reinforcement learning agents, by design, learn new tasks from scratch, i.e., without any prior knowledge. This is in striking contrast to how humans learn. In fact, a human naturally reuses his/her lifetime of experience to quickly learn new tasks. Consider, for instance, buying a new car and driving it for the first time. It might take some time to perfectly master its control, yet anyone who has driven a car before is likely to decently drive the new one without any complication. This is because a human can generalize the skill of "car driving" beyond the specific car model. This feature of human learning, called knowledge transfer, is one of the natural directions to make artificial reinforcement learning more sample efficient. In particular, the goal of transfer in reinforcement learning is to build artificial agents that are able to reuse experience obtained while facing previous tasks to speed-up the learning process of new related problems.

The literature on transfer in reinforcement learning is vast.. While a variety of approaches have been proposed with promising results, many open problems, concerning both practical and theoretical aspects, still exist. These include: (i) building generalizable transfer methods, i.e., that are not tied to a specific learning algorithm or domain; (ii) building scalable methods, i.e., that can be applied to complex real-world problems; (iii) building provably-robust methods, i.e., those guaranteeing that knowledge transfer never negatively affects the learning process; (iv) understanding the fundamental theoretical limits of transfer, i.e., what can and cannot be achieved with knowledge reuse. Motivated by these open problems, in this thesis we offer a number of contributions to the field of knowledge transfer in reinforcement learning. The common pattern behind all the approaches we propose, and possibly behind the majority of transfer methods in the literature, is the concept of structured domain that we introduce. Informally, a structured domain is defined by a collection of tasks together with a process generating them. Tasks in the same domain possibly share some underlying structural properties, i.e., they have some similarities that enable knowledge transfer. An agent that faces multiple tasks from the same domain should be able to understand their hidden structure, so as to transfer this knowledge to improve the learning process of new problems. While framing everything in the context of structured domains, we study two complementary problems in the transfer literature: (P1) How can the agent reuse experience samples collected while facing a set of source tasks to speed-up the learning process of a new target task? (P2) How should the agent generate new experience in the target task (i.e., how should it explore the new environment) given the available prior knowledge? The thesis is divided in three parts, each addressing one of these two main problems from a different perspective.

The first part deals with problem (P1). In particular, we design methods to transfer experience samples (i.e., states, actions, and rewards collected from the interaction with different environments) across tasks. We start by formalizing and motivating the problem, while discussing its main challenges. We then propose algorithms for transferring experience samples in batch reinforcement learning and online policy search. For both settings, we follow the same pattern: we transfer all the available samples, while weighting their contribution to the learning process using importance sampling to compensate for the distribution mismatch between source and target domains. This is in contrast to existing works, which either transfer samples directly without accounting for distribution mismatch, make strong assumptions, or carry out an expensive sample selection process. Since computing the importance weights requires knowledge of the unknown source/target environment models, we propose a sound technique to estimate these quantities by directly reducing the mean square error of the resulting estimators. In both settings, we provide theoretical insights one the performance of our methods and report good empirical performance on continuous control tasks.

In the second part, we begin studying problem (P2) from a practical perspective. In particular, we use ideas from randomized value functions to design an algorithm that estimates a distribution over the optimal value functions of tasks in the underlying structured domain and uses it as a prior to enable exploration via posterior sampling in new tasks. Since computing the corresponding posterior given target samples is intractable in most cases of practical interest, we propose an efficient variational approximation. Notably, the generality of our design allows our algorithm to be combined with complex value-function approximators (such as neural networks) and posterior distribution classes. We theoretically study the finite-sample properties of this approximation and report good results on increasingly-complex continuous domains.

In the last part, we address (P2) from a theoretical perspective. We start by better formalizing the problem of transfer in structured domains. We use existing ideas to decompose the problem in two sub-problems: (1) learning structure from previous tasks and (2) exploiting structure to quickly learn new tasks. We focus on the second sub-problem for the case where the structure is exactly known and for the one where it is only approximate (e.g., learned from experience). First, we consider a multi-armed bandit problem with known structure in a general form and design an arm-elimination strategy that exploits the given structure to quickly discard sub-optimal arms. Our theoretical results, differently from those of previous works, clearly show the performance gain obtained by using prior knowledge while still certifying that our algorithm never performs worse than unstructured baselines. Next, we restrict our attention to structures where the rewards are linear functions of given features and unknown parameters. We design a computationally-efficient incremental algorithm that is asymptotically optimal (in a problem-dependent sense) for this specific structure, while providing finite-time guarantees on its performance. Finally, we consider the problem of best policy identification in Markov decision processes with approximate structure. Under the assumption that a generative model of the target task is accessible, we design an algorithm that actively demands samples from state-action pairs that yield high information for finding a near-optimal policy. We derive an upper bound to its sample complexity that certifies how the algorithm exploits the approximate structure while never resulting worse than baselines not using structure at all.

All together, these results advance our understanding of knowledge transfer, one of the key components towards the deployment of reinforcement learning agents to the real world.

DESIGN AND CONTROL OF ACTUATORS FOR MODERN VEHICLES

Davide Todeschini - Supervisor: Giulio Panzani

Co-supervisor: Sergio Matteo Savaresi

Modern mobility is characterized by an endless improving of vehicles performances and passenger comfort. The drive-by-wire paradigm of modern vehicles led to the replacement of traditional mechanical actuators with electronically controlled ones. In this thesis, some specific cases have been selected, where the peculiar nature of the actuator and its application calls for a deeper study. In particular, new tuning methods and actuator design procedures have been investigated and developed, to improve the efficiency and comfort of modern vehicles. The design and the control of such actuators reveal some challenges and opportunities for performance improvements.
The first, and most straightforward, example is the one where the electronic actuator becomes a "smart actuator", which serves the requests coming from higher lever vehicle controllers, regardless from their specific nature. For these kind of actuators, the challenge is to develop a general purpose control strategy, that maximizes the tracking performance of a generic reference signal. The Brake-By-Wire and the Clutch-By-Wire actuators perfectly fit this category: they replace the standard mechanical actuators with an electronically controlled one and their control goal is the fast and precise tracking of the reference variable (force, speed, position), regardless the higher level vehicle control strategy that generates them. For these actuators the development of a control strategy is an almost solved problem while one scientifically relevant challenge, that has been addressed in the thesis, is the automatic and data-driven calibration of such strategies.
In other situations, the integration and inter-dependency between the actuator and the vehicle control strategy is so deep that the design of the two cannot be addressed separately. As selected case study, the gyroscopic stabilization of a motorcycle has been addressed. In such application the actuator design and sizing are strongly interwoven with the vehicle stabilizing control strategy, calling for an genuine mechatronic design of both.
A last research opportunity that modern actuators can pose is related to the innovative actuator architecture that opens the way for different vehicle control strategy paradigms, compared to the standard ones. An example of these are electro-mechanical suspensions which in this research have been studied applied to a mountain-bike. This type of vehicle perfectly fits with the modern mobility transition, since from a pure mechanical vehicle it is becoming a very technological one on which different intelligent devices and actuators find their spot. In this case the energy-oriented vehicle control strategies partially conflict with the standard suspension comfort-oriented goal and thus require a deeper understanding and analysis.
Throughout the thesis, the challenges related with these four type of actuators has been addressed, proposing new design procedures and control methods.
Starting from the Brake-By-Wire, data-driven tuning methods have been applied to pressure controller in order to develop automatic tuning procedures that can be used both with a dedicated experiment, but can also adapt the control parameters during the actuator lifetime. The selected data-driven technique is the so called Virtual Reference Feedback Tuning (VRFT), for cascade controller architectures. In addition, the pressure oscillations problem, arising from the combination of long hydraulic braking lines and highly dynamic actuation requests, has been studied. A mathematical model that catch this phenomenon has been developed and an open-loop control strategy that attenuates the oscillations has been obtained. The proposed solution has been tested both in simulation and on the experimental setup, showing good performances in weakening the oscillations keeping the same actuation speed. Eventually, an automatic parameter calibration procedure for the proposed control strategy has been proposed and experimentally validated.
The Clutch-by-Wire actuator shares a similar mechatronic architecture with the BBW one. However, the specific nature of clutch mechanics is responsible for a highly nonlinear system behavior. A fixed-structure closed loop controller has been designed and its parameters tuned so to enhance the system robustness w.r.t the mentioned non-linearity. Due to the difficulty in getting a mathematical quantitative model of the system, multiple identifications with linear transfer functions have been carried out in different operating points. The whole set has been used for a robust tuning of the controller, based on the Scenario approach. This tuning method exploits a convex optimization problem that minimize the distance of actual closed-loop characteristic polynomial coefficients from the desired ones. Starting from that, some modifications to the algorithm have been performed, in order to achieve better tuning results. The idea behind these, is to keep the optimization problem convex, conducting at the same time a poles distances minimization through a sensitivity matrix. Moreover, since the sensitivity matrix is valid only in the neighborhood of desired poles, an iterative procedure has been proposed in order to move the farthest poles closer to desired one. The new tuning method has been experimentally validated and compared with a traditional tuning on the average system, showing more robust performances.
Regarding the gyroscopic stabilization problem, both the design and the control issues have been analyzed. In particular, an optimization-based mechatronic design procedure has been developed in order to obtain both actuator and control parameters. In addition to this, two different control algorithms that can deal with non-idealities of the experimental setup has been proposed with two different purposes. On one hand, an algorithm based on the sliding-mode theory has been developed; it can be easily tuned even if system parameters are not perfectly known but shows inferior performances with respect to the other. On the other hand, a state-feedback control algorithm has been proposed; it shows superior performances, but requires a more precise characterization of the system. Both the control algorithms have been experimentally tested on a real motorbike equipped with a prototype constructed with the indications obtained from the design procedure. Experimental tests validate both the design and the proposed control, showing good stabilization performances both with and without driver
Finally, the electro-mechanical suspension applied to a mountain bike has been studied, discussing its design and its energy-harvesting abilities. The EMS application to the bicycle vertical dynamics required the development of a new analytic model to describe the peculiar vehicle dynamics. The model takes its moves but differs from the quarter-car one, typically used in suspension control design and analysis. This model, together with a traditional suspension bench characterization has been used for design the electro-mechanical suspension. In particular, from a simulation analysis some indications for the suspension design oriented to energy-harvesting have been obtained. These indications, can guide the designer in order to select the best parameters of the suspension components that allow to harvest the maximum energy. Once the suspension has been designed, another simulation analysis has been conducted in order to study the energy-harvesting properties related with the applied damping. From these, it comes out that the maximum regenerated energy is obtained when the control algorithm is capable of emulating a low damping mechanical suspension. The experimental tests confirms such results. They have been conducted with a prototype based on the proposed design, experimentally proving the energy-harvesting properties of the system.

LOW-NOISE MIXED-SIGNAL ELECTRONICS FOR CLOSED-LOOP CONTROL OF COMPLEX PHOTONIC CIRCUITS

Francesco Zanetto - Supervisor: Marco Sampietro

During the last decade, an increasing research effort has been carried out to profit from the advantages of photonics not only in long-range telecommunications, where it has become the standard technology, but also at short distances, to implement board-to-board or even chip-to-chip interconnections. The performance scaling required by datacenter, automotive and telecommunication applications is in fact requiring very high data-rates achieved with reduced losses and power consumption, that are pushing copper-based connections close to their physical limitations even at short distance. In this context, Silicon Photonics emerged as a very promising technology, allowing to integrate many optical devices in a small silicon chip fabricated with the well-consolidated manufacturing processes of the microelectronics. However, the integration density made possible by Silicon Photonics revealed the difficulty of operating complex optical architectures in an open-loop way, due to their high sensitivity to fabrication parameters and temperature variations. Local light monitoring and active control of each photonic device thus emerged as strong requirements to correctly operate complex optical systems (Fig. 1).

In this thesis, the true potential of integrated photonics was unlocked thanks to the design of a low-noise mixed-signal electronic platform implementing feedback control of complex optical architectures. The system exploits the innovative ContactLess Integrated Photonic Probe (CLIPP), a detector that senses light in silicon waveguides in a non-invasive way by measuring their electrical conductance and can thus be integrated in many points of an optical circuit without penalties. The electrical behaviour of native waveguides was first investigated by measuring the resistance of their silicon core. The resulting value is orders of magnitude larger than what expected from the doping level and this was attributed for the first time to charges and electronic traps present at the interface between core and cladding, a phenomenon not fully investigated and established in Silicon Photonics. The knowledge acquired thanks to this characterization allowed to better understand the electrical behaviour of CLIPP detectors and propose a new sensor layout. A 10-fold improvement in the detector sensitivity was obtained with the new design, allowing to measure light variations down to -50 dBm. This unprecedented level of performance for a transparent detector, much better than what currently established in literature, demonstrates that fully transparent sensors can be a viable option to monitor the behaviour of photonic circuits. The obtained resolution would not have been possible without a proper design of the electronic readout frontend. Tailoring the electronics to the sensor is in fact mandatory obtain the best performance possible. In this case, a custom multichannel lock-in based ASIC (Fig. 2, 0.18 µm STMicroelectronics CMOS technology) was developed to read CLIPP detectors. The ASIC features an ultra-low-noise wide-bandwidth transimpedance amplifier with capacitive feedback, biased thanks to an innovative low-frequency network, followed by square-wave integrated demodulators to perform lock-in detection. The input referred noise of the ASIC well below 100 fA/sqrt(Hz) reflects in a resolution in the measurements of conductance below 1 pS, a level of performance required to detect very small light variations with a CLIPP detector. A mixed-signal FPGA-based motherboard (Fig. 3) was then developed to acquire the outputs of the ASIC and perform in the digital domain the most suitable control algorithms to stabilize the behaviour of integrated optical circuits. The platform reads the signal from 16 sensors and drives 16 actuators, thus enabling real-time control of complex photonic architectures. The digital core of the system was optimized to reduce the area occupation and process all the data in parallel without significant latency. Several control strategies were tested and the dithering technique in combination with a digital integral controller was identified as a very effective way to stabilize the behaviour of photonic circuits. The standard technique was studied and an extension was proposed to allow easy control of many devices in parallel, in particular in cascaded optical architectures. The designed system was finally tested in several experiments. First, a mesh of Mach-Zehnder interferometers, designed for mid-range free-space optical communications, was successfully tuned with a configuration time two orders of magnitude lower than in previous implementations. The capability of the electronic platform to compensate fluctuations of the mesh operating point was also successfully demonstrated. In a second experiment, the system was used to stabilize an optical architecture conceived for very high data-rate communications between processors inside datacenters. Thanks to the control action of the electronic platform, a 30 Gb/s optical link insensitive to temperature and wavelength fluctuations was successfully demonstrated. These two proofs of concept demonstrate that silicon photonics architectures, complemented by a proper electronic control layer, can be a realistic alternative to copper-based interconnections even in harsh and unstable environments and open the way to new sophisticated optical systems.

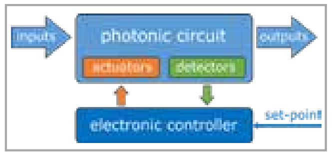

Fig. 1 - Control scheme of photonic systems. The state of the circuit is monitored with light sensors and stabilized with actuators, driven by an electronic controller.

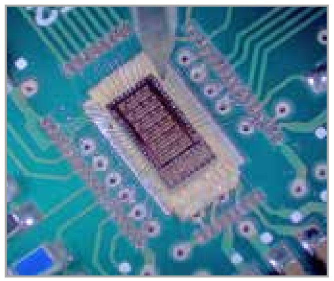

Fig. 2 - Microphotograph of the CMOS chip for CLIPP readout.

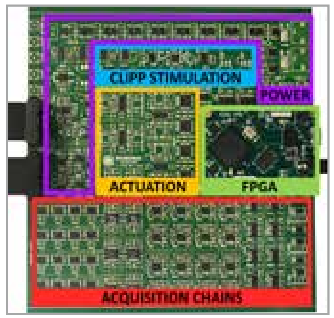

Fig. 3 - Photograph of the FPGA-based mixed-signal control motherboard.

FEATURE REPRESENTATION LEARNING IN COMPLEX WATER DECISION MAKING PROBLEMS

Marta Zaniolo - Supervisor: Andrea Castelletti

The success of a control policy highly relies on its feature representation, i.e., the information set it is conditioned upon. In real world control problems, defining an appropriate feature representation is a complex task, given the coexistence of multiple interacting processes whose relevance for the control task is often unclear.

In this thesis, we address the control problem of water resources systems, where a dam release policy is designed accounting for multiple water demands. This decision-making problem is challenged by the presence of non-linearities, multiple disturbances, possible alternative problem framings, and multiple conflicting objectives. Currently, the control rules of most water reservoirs are conditioned upon basic information systems, comprising reservoir storage and time index, however, the value of a more informative feature representation has been demonstrated unequivocally.

By conjunctively addressing the fields of water resources systems, feature selection, and multi-objective control, we propose methodological innovations in feature representation learning for complex water decision problems.

We capitalize on recent advances in monitoring and forecasting water availability to develop novel feature representation learning strategies to enhance water systems' resilience towards their crucial vulnerabilities, including droughts, critical phases in reservoir development (i.e., construction and filling), and multisectoral conflicts. Additionally, in multi-purpose systems, different control targets might be heterogeneous in their dynamics and vulnerabilities, and likely benefit from a tailored feature representation that varies across different objectives tradeoffs in terms of information and lag time.

We revise current literature on feature representation learning in the field of Reinforcement Learning and propose a taxonomy comprising *a priori*, *a posteriori*, and *online* approaches to feature representation learning (Figure 1), named after the alternative ways of interfacing the Feature Extraction (FE) with Policy Search (PS) steps.

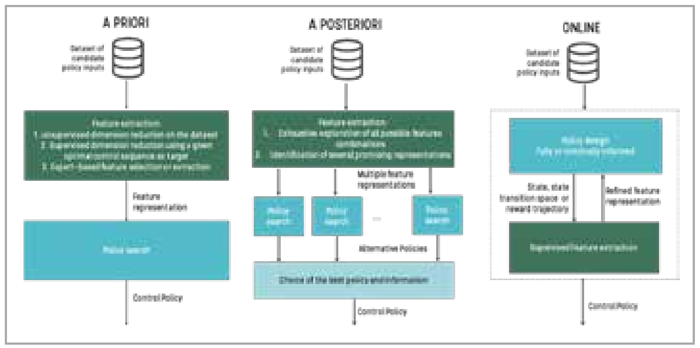

Fig. 1

In the *a priori* approach, the FE step is antecedent and independent from the PS step, and the dimensionality of the dataset of candidate features is reduced on the basis of intrinsic properties of the controlled system and does not depend on the policy search process. The *a posteriori* approach evaluates the suitability of a feature representation by assessing the performance of the policy conditioned upon it. In general, multiple policies are designed with alternative feature representations, and the desired representation is identified as the one producing the best performing policy. In the *online* approach, the FE step is simultaneous and codependent from PS, and an efficient policy representation is learned in conjunction with the policy.

For each approach, we propose novel methods targeting the control problem of multipurpose water systems, including:

- FRIDA, FRamework for Index-based Drought Analysis, that automatically designs an index representing a surrogate of the drought conditions of a basin, by selecting and combining relevant information about the water circulating in the system. FRIDA targets highly regulated water systems, where natural water availability is conditioned by the operation of water infrastructures, and traditional drought indicators fail in detecting critical events.
- A novel methodology that supports sustainable dam planning addressing the critical initial filling phase of large dams. The core novelty of this work consists in informing filling timing and operations by analyzing the climate oscillations that affect the region.
- CSI, Climate State Intelligence, a framework designed to capture the state of multiple climate signals from global datasets of oceanic temperatures to improve seasonal forecasts. The designed forecasts are employed to inform water system operations at the basin scale.
- NEMODPS, NeuroEvolutionary Multi-Objective Direct Policy Search, a novel direct policy search routine that conjunctively searches the policy functional class and its parameterization in a hyperspace containing policy architectures and coefficients. NEMODPS addresses multi-objective problems producing a tradeoff-dynamic architectural selection in one single run.
- AFS-NEMODPS, where NEMODPS is combined with an Automatic Feature Selection routine to conjunctively and progressively refine feature representation and policy design. AFS-NEMODPS targets multi-objective problems by tailoring feature representation and policy architecture to different objective tradeoffs.

Among the most valuable outcomes of this collection of works, we highlight that i) learning an appropriate feature representation is valuable not only in relation to the control policy performance improvement, but it also generates insights on the learned task by identifying relevant policy drivers and system vulnerabilities.
ii) The benefit of certain representations is especially noticeable in critical situations. For instance, the benefits of considering a drought index, or long-term hydrological forecasts are mainly noticeable in dry years.
iii) The value of hydrological forecasts in informing reservoir operations, a very common choice to enhance water system resilience, might be afflicted by modeling distorsions, and a less processed alternative could bring more benefit to the control policy.
iv) The concepts of feature representations can be valuable beyond strictly control problems to applications of planning and filling. Lastly, v) We demonstrate how guiding representation learning towards conflict mitigation determines a cross-sectoral information selection that tends to improve the performance in the weaker objective.

The common thread in these outcomes is that, in an operational context, the quest for an *optimal* feature representation might be an elusive concept. Instead of chasing optimality per se, we found several opportunities to mitigate recurrent water systems failures, that catalyze the largest concentrated impacts related to water management. As a consequence, our efforts were mostly concentrated on mitigating damages of drought emergencies, of critical phases in reservoir development (i.e., construction and filling), and social tensions deriving from conflicts between different users and their demands.