MECHANICAL ENGINEERING | PHYSICS | PRESERVATION OF THE ARCHITECTURAL HERITAGE | STRUCTURAL, SEISMIC AND GEOTECHNICAL ENGINEERING | URBAN PLANNING, DESIGN AND POLICY | AEROSPACE ENGINEERING | ARCHITECTURAL COMPOSITION | ARCHITECTURE, BUILT ENVIRONMENT AND CONSTRUCTION ENGINEERING | ARCHITECTURAL, URBAN AND INTERIOR DESIGN | BIOENGINEERING | 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 | 2019

DOCTORAL PROGRAM IN INFORMATION TECHNOLOGY

Chair:
Prof. Barbara Pernici

Introduction
The PhD program in Information Technology (IT) started in year 2001, when two traditional PhD programs, in Automation- Computer Engineering and Electronics-Telecommunications, were merged. The new unified PhD program 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 Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB). However, following historical development of the Department, and new trends of modern society, some cross-areas research fields are also covered, such as robotics, machine learning, big data, intelligent data analysis, Industry 4.0, Internet of Things, 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 60 new students join the program, for an overall number of students of about 200. 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, Singlephoton 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, Information transmission and radio communications, 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. About 50% of scholarships 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 30 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 courses that spur innovative and cutting edge research. Therefore, particular attention is devoted to help students to make their best choices according to an internal regulation scheme.

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 20% 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		
BARBARA PERNICI	LORENZO FAGIANO	LUIGI PIRODDI
CESARE ALIPPI	GIANCARLO FERRIGNO	MATTEO PRADELLA
FRANCESCO AMIGONI	CARLO FIORINI	IVAN RECH
LUCA BASCETTA	SIMONE GARATTI	CARLO RIVA
GIUSEPPE BERTUCCIO	NICOLA GATTI	CRISTINA SILVANO
ANDREA CASTELLETTI	ANGELO GERACI	ALESSANDRO SOTTOCORNOLA SPINELLI
MATTEO CESANA	PAOLO MARTELLI	LETIZIA TANCA
FLORIAN DANIEL	RAFFAELA MIRANDOLA	MASSIMO TORNATORE
FABIO DERCOLE	ANDREA MONTI GUARNIERI	

PHD ADVISORY BOARD	
Mario Caironi	IIT
Luigi Cicchese	Concept Reply
Cristina Cremonesi	The European Ambrosetti
Massimo Crisci	European Space Agency
Alessandro Ferretti	Tre-Altamira
Giuseppe Fogliazza	MCE Srl
Bruno Garavelli	Xnext s.r.l.
Alessandro Grossi	Micron Semiconductor Italia Srl
Renato Lombardi	Huawei Technologies
Renato Marchi	Gruppo PAM
Giorgio Parladori	SM Optics srl
Enrico Ragaini	ABB S.p.A.
Piercarlo Ravasio	Prospera
Beatrice Rossi	STMicroelectronics
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 2018 the following awards have been obtained by PhD candidates:

ACM CHI2018 PhD Student Best Paper Award - 3rd place: Mirko Gelsomini
IEEE SSP 2018 Best Student Paper Award: Mattia Brambilla
IEEE Inertial 2018 Best Student Paper Award: Cristiano Rocco Marra
IEEE MEMS 2018 Outstanding Student Paper Award: Cristiano Rocco Marra
Dimitris N. Chorafas Foundation Award: Giuseppe De Nittis and Alessandro Falsone
2018 IEEE Nuclear Science Symposium Best Student Paper Award - 2nd place: Luca Buonanno
RecSys Challenge 2018, Runner-up: Maurizio Ferrari Dacrema

NEW HORIZONS FOR STREAM PROCESSING

Lorenzo Affetti - Supervisor: Prof. Gianpaolo Cugola

Stream processing has gained tremendous attention over the last years and many Stream Processors (SPs) have been designed and engineered to cope with huge volumes of data coming at high velocity. Streams could contain stock options, user clicks in web applications, customer purchases in an e-commerce application, positions of robots in a warehouse, or temperature measurements from sensors. The common requirement for streaming applications is to process unbounded streams of elements and continuously compute queries like “what is the top purchased product?”, or “what was the average temperature in the server room in the last second?” in order to take rapid compensating actions such as ordering a new stock of the top purchased product, or prevent fire in the server room. In order to continuously process huge amounts of elements and take real-time decisions, SPs exploit the computational power offered by multiple machines by distributing the computation and dividing data in shared-nothing partitions to avoid expensive data race management while processing. Stream processing is also a programming paradigm suited for designing novel event-driven applications with high throughput and low-latency requirements. Streams offer decoupling among the processing modules and, thus, enhance application modularity and composability. Indeed, SPs are playing a central role in the technology stacks of modern companies and they are covering more and more tasks that, in standard deployments, compete to other tools. The unification of multiple systems under a unique one reduces system integration complexity and facilitates application maintenance and modeling. Novel event-driven applications, indeed, embed a Database Management System (DBMS) in the state of computation of the SP, although with limitations such as multi-key transactions and consistent external querying. Moreover, their central role requires SPs to conform to a standardized execution semantics in order to improve their usability, interoperability, and interchangeability. This thesis takes a step towards SPs standardization through modeling the discrepancies between them, and a step towards their integration with DBMSs by extending their computational model to deal with transactional computation.

For SPs standardization, we use SECRET, a well recognized mathematical model to express their execution semantics, to model five distributed SPs that were developed after the introduction of SECRET itself and are today widely used in companies at the scale of Google, Twitter, and Netflix. We show that SECRET properly models a subset of the behavior of these systems and we shed light on the recent evolution of SPs by analyzing the elements that SECRET cannot fully capture.

In order to decrease system integration overhead and to overcome the limitations of the current approaches for DBMS over SP, we enhance the capabilities of the SP with DBMS’s ones by extending the SP computational model with transactional semantics: we develop a unified approach for multi-key transactions on the internal state of the SP, consistent external querying with respect to transactional operations on the state, and streaming data analysis. We implement TSpoon, a prototypal implementation of our extended model, as an extension to the open-source SP Apache Flink™. We evaluate our prototype using synthetic workloads in various configurations to understand which metrics mostly impact its performance. Eventually, we evaluate a real use-case scenario and compare the results with the ones obtained from VoltDB, a commercial in-memory database known for its excellent level of performance: TSpoon outperforms VoltDB in the execution of multi-key transactions and proves to be a promising future direction for the integration of DBMSs and SPs.

DISTRIBUTED SYNCHRONIZATION FOR DENSE WIRELESS NETWORK SYSTEMS

Maria Antonieta Alvarez Villanueva - Supervisor: Prof. Umberto Spagnolini

Distributed synchronization on dense wireless network constitutes a new regime for network synchronization, mainly due to the challenges involved on the inter-connectivity of large number of devices for management and coordination of the network. This inter-connectivity enables various services that are envisioned for Internet of Things (IoT), such as smart meters, smart traffic, public safety, medical metering, smart homes, etc. IoT paradigm embeds the inter-communication of massive number of devices that is challenging in terms of scalability, sustainability and improved efficiency. The inter-connectivity of massive number of devices is an open issue, and it is expected to be part of 5G ecosystem. Therefore, time and carrier frequency synchronization are critical aspects to be considered to guarantee the proper network communication. In this context, there are some characteristics to be taken into account for a proper synchronization solution in dense networks. On PHY-level context, the multiple connectivity for heterogeneous communication devices require scalability, and the constraint consumption of energy due to the capacity of the devices, requires a fast convergence and optimum processing. The proposed synchronization methodology considers the following statements: i) to allow scalability of the network; ii) to allow fast synchronization of the whole network; and iii) to mitigate power consumption.

The network scenario consists of a dense and not coordinated wireless connected nodes without any external agent as a reference, where the internode distance is small (compared to the bandwidth) to neglect any propagation delay. Each node is equipped with a local free-running reference that skews from the others, and local control is by changing timing offset (TO) and carrier frequency offset (CFO) on each node independently of the others. The periodic synchronization allocates specific signatures in data communication to exchange the synchronization state by every node so that the network iteratively reaches a global convergence. In this Thesis, a distributed synchronization algorithm based on consensus paradigms is proposed, that enables the network to reach asymptotically a global convergence based on the exchange of a common beacon (i.e., the same synchronization beacon is used by all nodes in the network) with features that enable a fast and accurate timing and carrier frequency synchronization. Each node broadcasts the same signature that superimpose (collide) with the others and compound signal of multiple collisions represents a reference signature that embeds the TO and CFO reference for the entire network. Once the network reaches a convergence, the frames are aligned giving the start of time-slot (Fig. 1). Contrary to conventional synchronization methods, the feature of the proposed distributed synchronization algorithm is that the collision of signatures does not impair the synchronization, rather it is used by the receiving node as ensemble reference to enable its synchronization.

The distributed synchronization algorithm is sequential as sketched in Fig. 1 where TO synchronization first reaches the steady state and this drives the convergence of the TO network. During frame alignment (TO acquisition, in Fig. 2), the CFO has a random behavior due to the large dispersion of TO not allowing a proper estimation of the CFO error. When TO is close enough to synchronization, the CFO correction takes place (CFO acquisition) by correcting the CFO impairments and allowing the fine synchronization of the network. The design of a unique synchronization signature (beacon) largely simplifies the setting and it is based on chirp-like sequences with good correlation properties (e.g. Zadoff-Chu sequences used on LTE), that allows the join estimation of TO and CFO errors. The accuracy of the correlator-based estimator is analyzed, and the impact of stochastic perturbations product of the oscillator's instability and estimator's error is studied to evaluate the convergence condition of the proposed distributed synchronization algorithm.

The proposed distributed synchronization algorithm is implemented on a hardware demonstrator, based on software-defined radios programmed in GNU radio, showing the algorithms ability to decouple the TO and CFO estimate and it is analyzed the convergence time of TO and CFO synchronization. Then, the optimization of the distributed synchronization is carried out based on i) an optimal duplexing strategy, and ii) an optimal synchronization protocol. In context of dense inter-connected networks with oscillators affected by drifts, two synchronization approaches are compared based on: collision-avoidance and collision of signals, to investigate the impact of the network scalability by comparing the convergence time and synchronization dispersion error. Finally, the impact of synchronization is evaluated for a resource management scheme (spectrum allocation) in Device-to-Device (D2D) communications. The selection of the resources to be added or released in the allocation is performed by minimizing the boundary extension of the time-frequency (TF) spectrum region, this criteria avoids fragmented region allocations with large boundary areas that could increases the cross-interference due to TF jitter.

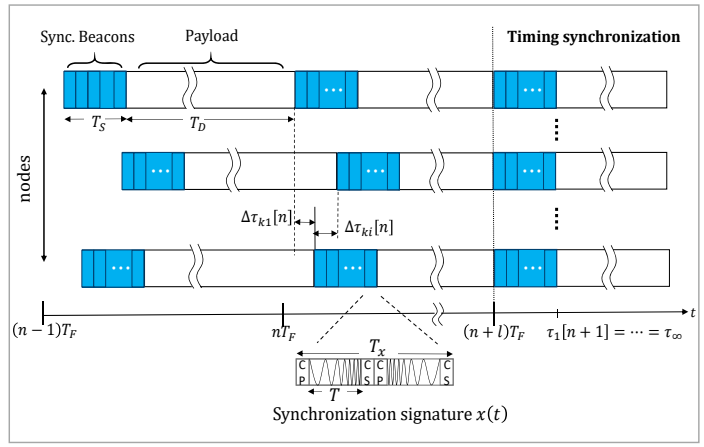

Fig. 1 - Sketch of TO synchronization evolution. All nodes transmit the same beacon periodically during and on consecutive frame-period . Beacons are allowed to collide among each other. At receiver node *k*, the relative TO error is the error from transmitted neighbor () with respect to the local reference. The network reaches a synchronization stage when all the frames are time-aligned given the start of the frame.

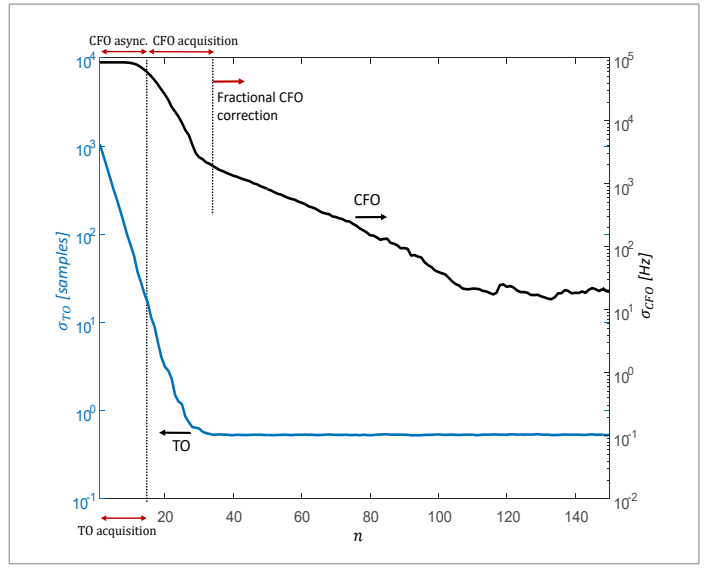

Fig. 2 - Root Mean-Square Deviation (MSD) of TO (left-axis) and CFO (right-axis) vs iteration for nodes mutually coupled and fully connected (all-to-all).

DESIGN OF ANALOG ASICS FOR X-RAY DETECTORS

Aidin Amirkhani - Supervisor: Prof. Carlo Ettore Fiorini

This research project is focused on the development of readout ASICs for two main applications. The first part is mainly focused on the ASIC development for the SIDDHARTA experiment. The SIDDHARTA experiment is designed to investigate strong nuclear interactions using exotic atoms in the field of nuclear physics. Silicon Drift Detectors (SDDs) used in this experiment are arranged in arrays of 2×4 elements with total area of 612 mm^2. At the final stage of SIDDHARTA experiment, 48 SDD arrays are needed to be utilized in a gantry structure to perform X-ray spectroscopy of exotic nuclei, like kaonic deuterium. Each single SDD unit in 2×4 formation of arrays is coupled to a charge sensitive preamplifier, namely CUBE, which is followed by shaping amplifier, and consequent analog and digital electronics that are all integrated on a custom developed multichannel chip called SFERA.

In the first chapter of the thesis, we introduce SIDDHARTA experiment in details. After giving an insight about the main objective of the SIDDHARTA experiment and its requirements, main characteristics and working principles of SDDs will be discussed with a focus on SDDs that will be employed in SIDDHARTA Experiment. The chapter then continues with recalling basic concepts of detector signal processing as they will be used frequently in the following chapters. Finally, the charge sensitive preamplifier that will be coupled to SDDs in SIDDHARTA experiment will be introduced and discussed briefly.

The second chapter introduces SFERA ASIC. SFERA (SDDs Front-End Readout ASIC), is a 16-channel ASIC that is composed of a 9^{th} order semi-gaussian shaping filter with selectable peaking times of 0.5us, 1us, 2us, 3us, 4us and 6us as well as peak stretcher, pile-up rejection logic and different readout modalities. A dedicated digital logic, manages the data transfer to the downstream DAQ system, synchronously providing it both the "firing channels" addresses and related stretched pulse-peaks in the same time order the events are detected. This chapter then continues with the description of the detector characterization procedure as well as DAQ system used to characterize and troubleshoot detectors that will be used in final SIDDHARTA experiment.

The third chapter focuses on the crosstalk and timing analyses made on SFERA and SIDDHARTA detection modules. In the SIDDHARTA experiment, in order to reject asynchronous background events, only X-rays detected within a time window opened by a kaon monitor trigger are selected. To test this detection operation, a beta source (^{90}Sr) was used to provide a trigger and simultaneously excite fluorescence X-ray lines on a multi-element target. This chapter aims to investigate the anomalous behaviour, observed in both time and energy spectra of the SDD acquisition chain, when the detectors were exposed to radiation dominated by charged particles.

The fourth chapter introduces the modifications made to SFERA ASIC to make it in-line with the experiment requirements. These modifications include but are not limited to increase of the ASIC time resolution as well as introducing new inhibit strategies that are vital for the final SIDDHARTA experiment.

The second part of my research, is focused on the development of a 128-channel low-power ASIC for the readout of silicon microstrip detectors with high energy resolution and counting rate efficiency for diffractometry applications. In chapter five, we introduce the microstrip detector read-out chain and we will continue our discussion on the choice of the best filter for our application. Then the chapter continues describing the criteria considered for the choice of the filter family and its order. After this choice, count-rate linearity of the filter will be studied and statistically discussed. Later, different implementation topologies will be discussed. After deciding the filter topology, we will demonstrate on how to tune the corresponding filter components. By implementing the shaper, gain and peaking time spread will be studied through Monte Carlo analyses and by realizing the layout of the shaper, post-layout results and effects of parasitic capacitances will be discussed. In the last part of the chapter, we will discuss on the overall performance of the shaper after fabrication.

AUTOMATIC SYSTEMS FOR UNSAFE LANE CHANGE DETECTION AND AVOIDANCE

Alessandro Amodio - Supervisor: Prof. Sergio Matteo Savaresi

Road Safety is currently recognized to be a major societal issue, with road crashes being among the major causes of death. In the past recent years, car manufacturers have been responding to this increasing need for safety by developing, and increasingly deploying on board of commercial vehicles electronic systems called Advanced Driver Assistance Systems (ADAS), designed to increase comfort and safety during everyday driving. In some cases, such systems only report information or alert the driver in dangerous situations, while in other cases they directly intervene on the vehicle dynamics to avoid potential hazards.

This thesis proposes a composite and integrated system that helps the driver avoid safety hazards due to unsafe lane change maneuvers, which are the most frequent scenarios that involve vehicle crashes. This task is addressed in two ways.

Prevention Drivers often underestimate his/her own level of drowsiness or impairment due to alcohol and start driving without being in healthy condition. In order to reduce potential risks in such scenarios, the goal of the system is to alert the driver in case unhealthy condition is detected, or a potential danger for a lane change maneuver is foreseen.

Intervention
In some cases, the driver loses control of the vehicle due to unhealthy condition, or intentionally initiates an unsafe lane change maneuver without noticing a potential danger due to inattentive driving. In such scenarios, the goal of the system is to have the vehicle actively intervene to correct the trajectory.

The Prevention task is performed by issuing alert signals to the driver in potentially hazardous scenarios; to do so, the proposed system addresses the following two main sub-tasks.

Before the drive starts, the system has the goal to evaluate driver's impairment condition due to alcohol abuse.

During the drive, the system has the goal to reveal potential danger for a lane change maneuver due to the presence of objects in the vehicle's surroundings.

The Intervention task requires the system to actively intervene to modify the vehicle's trajectory when an unsafe lane change maneuver is performed; in particular, the objective of the system is to activate and avoid the lane change maneuvers in the following cases.

Inattentive driving: the driver fails in noticing an upcoming danger in the adjacent lane and intentionally initiates a lane change maneuver in unsafe condition.

Drowsiness: the driver loses control of the vehicle due to occurrence of micro-sleep, which drifts towards the adjacent lane, resulting in an unintentional lane change maneuver.

The thesis' outline is given in the following.

Introduction
At first, the problem of road safety is outlined, then the Advanced Driver Assistance Systems are introduced, with a description of their main components, a classification and some examples.

The problem of the unsafe lane change is addressed by giving a definition of lane change maneuver and lane change crash, together with a distinction between different pre-crash scenarios; then, the main driver related factors are presented. To conclude, the thesis objectives and outline are presented, together with the main innovative contributions and results.

Experimental platforms and sensors
The main experimental platforms used for performing on-road tests are presented, which include a four-wheeled vehicle, a two-wheeled vehicle and a simulation environment.

Sensors are extremely important for the ADAS, since they provide information on the vehicle's surroundings they need to perform their functions: in the second part of the chapter, the main sensing technologies utilized in the work are described, which are radar, lidar and stereo cameras. In particular, for each sensing technology a technical background is given, in order to present the basic working principles; then, the results of comparative tests are presented in order to show the performance of each of the considered sensors and draw some conclusions on the main pros and cons of each technology.

Automatic System for Unsafe Lane Change Avoidance with Driver's State Monitoring
At first, the chapter presents some of the main related works that can be found in the scientific literature; then, the system structure is described.

The system is composed by two main blocks. The first is the Warning Module, which is in charge of addressing the Prevention task: after a brief outline, its two main components are presented. The second part is the Control Module, which is in charge of performing the Intervention task; unlike the other, it is not always active, but its activation is regulated by a Supervisor block. The last part of the chapter is dedicated to a third block, the Supervisor, which is in charge of controlling the activation of the Control Module based on the vehicle's position in lane, the driver's state and the state of the vehicle's surroundings.

Warning Module
This is the passive one among the two modules, in charge of giving alert signals to the driver in critical situations and to address the Prevention task.

The module is composed by two blocks; the first is the Driver's state monitoring block, which is in charge of revealing driver's drunkenness state before the drive start through dynamic analysis of the driver's Pupillary Light Response. A linear model is derived for the pupil diameter constriction due to light stimulus, from which a set is features is extracted; different classification techniques are then compared based on their capability on discriminating between sober and drunk subjects.

The second block is the Surrounding's state monitoring block, which is in charge of monitoring the vehicle's adjacent lane: in particular, the aim is to issue warning signals to alert the driver in case a potential danger to a lane change maneuver is detected. In order to do so, the block monitors the vehicle's Blind Spot and detects the danger due to far vehicles approaching at high speed.

Control Module
The Control Module addresses the Intervention task: this is the active module, which has actuation capability and is in charge of actively intervening on the vehicle's dynamics to avoid crashes with other vehicles.

At first, the main related works are described, and then the control objectives are presented: in particular, the goal is to control the vehicle's lateral position through a differential braking control action.

A detailed model of the system is then presented, to describe the vehicle dynamics from the control input (differential torque at the wheels) and the disturbance input (steering wheel angle) to the output (vehicle's lateral position). A linear model is identified from experiments at CarSim simulator and validated; the framework is considered to tune a PID controller that reaches optimal compromise between different control objectives. To conclude, experimental results are presented that show the effectiveness of the proposed strategy, validated against intentional and unintentional lane change maneuvers.

Conclusions
This chapter recaps the structure and the objectives of the work, summarizing the obtained results and the main contributions of the work.

Appendix
A new methodology for estimating the vehicle's lateral position in lane is presented: when active, the Control Module has the goal of closing a lateral position control loop, which requires an accurate estimation of the vehicle's lateral position.

In this chapter, a new methodology based on an array of magnetic sensors mounted on the road surface is proposed: the objective is to record the magnetic signature of a vehicle in transit above the equipment, and use it to estimate the vehicle model and its lateral transit position by comparing its magnetic signature with previously recorded ones of known vehicles, transited at known positions.

RESOURCE MANAGEMENT AND PLANNING IN CLOUD-ENABLED OPTICAL METRO-AREA NETWORKS

Omran Ayoub - Supervisor: Prof. Massimo Tornatore

The Internet is experiencing an exponential increase in terms of number of users, data traffic, connected devices and latency-stringent services. It is foreseen that the introduction of new technologies, like 5G, will boost this growth even further. To cope with this growth, future communication networks are required to provide unprecedented performance in terms of enhanced throughput, increased coverage, reduced latency and power consumption. Moreover, these networks must be continuously evolved, by resorting to novel technologies and architectural solutions, to meet such requirements.

A promising solution consists in enhancing nodes at the network edge, taking advantage of Network Function Virtualization and Cloud Computing, with cloud capabilities, i.e., with storage and computing capabilities, such that services can be pushed closer to users and terminated locally. In this view, there are some basic issues of this architectural solution that need to be investigated such as the energy and cost aspects. Moreover, as it is decisive to deploy and dimension edge-nodes guaranteeing minimum overall network resource occupation due to service distribution, strategic planning of the deployment of cloud-enabled edge nodes is needed to efficiently enhance the network performance.

In this thesis, we investigate the deployment of cloud-enabled edge nodes and propose novel strategies for improved network-resource management. We focus on optical cloud-enabled metro-area networks, as the one depicted in 1, which are currently evolving from a rigid ring-based aggregation infrastructure to a composite cloud-network ecosystem where novel 5G cloud-based services can be implemented and supported. Specifically, we investigate this architectural solution considering emerging network architectures such as Fixed/Mobile Convergent networks and Filterless Optical Networks. As for a use case, we consider Video-on-Demand (VoD) content delivery service for it being responsible for the elephant's share in global Internet traffic, and assume a case study where edge-nodes (e.g., caches) host and terminate VoD services.

In more details, the thesis first puts the energy efficiency of this architectural solution into question as the deployment of a significant number of cloud-enabled edge-nodes heavily impacts the overall network energy consumption. In particular, energy-efficient VoD content caching and distribution strategies that strikes the trade-off between the energy consumption due to data transport and energy consumption due to powering-on and operating the caches are proposed. Moreover, the thesis focuses on the cost-efficiency of this architectural solution and investigates cache deployments which guarantee a cost-efficient outcome. In addition, the thesis investigates edge-node deployments which guarantee the minimization of the overall network resource occupation while taking into account the network and service characteristics. Based on the obtained results, we provide a framework to identify the optimal cache deployment which minimizes the overall network resource occupation due to VoD delivery considering network and service characteristics. 2 shows an example of such a deployment for a specific case study of VoD content caching and distribution. The figure shows the dimension of the edge-nodes in terms of their storage capacity (GB) and processing power (vCPU), the number of VoD contents stored in each network level and the amount of traffic generated from each edge-node or mini data center to serve a number of users. Moreover, the thesis focus on moving (i.e., migrating) services hosted on Virtual Machines (VMs) between cloud-enabled edge-nodes and data centers. Specifically, the thesis investigates online VMs migration techniques which, on one hand, have several advantages such as allowing the migration of a service from one data center to another with minimal service interruption, while, on another hand, consume high amount of network resources. In this context, efficient routing and bandwidth assignment algorithms aimed at minimizing the overall network resource consumption due to online VMs migration are proposed.

Concluding, this thesis proposed strategies for energy and cost-efficient cloud-enabled edge-node deployments in optical metro-area networks for VoD content delivery service. Moreover, efficient strategies for resource management for VoD content distribution and online VM migration have been proposed.

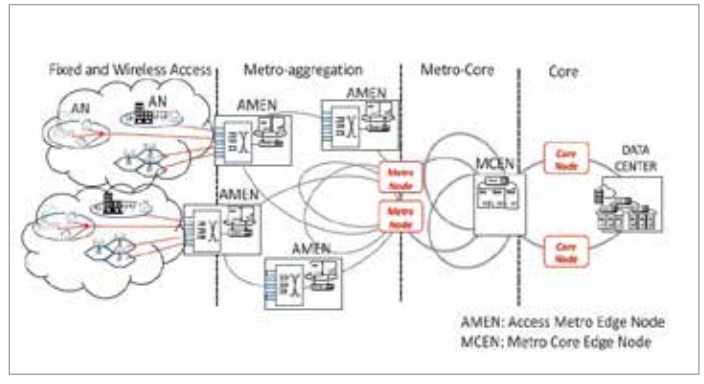

Fig. 1 - A schematic representation of a hierarchical cloud-enabled metro-area network spanning over four network levels

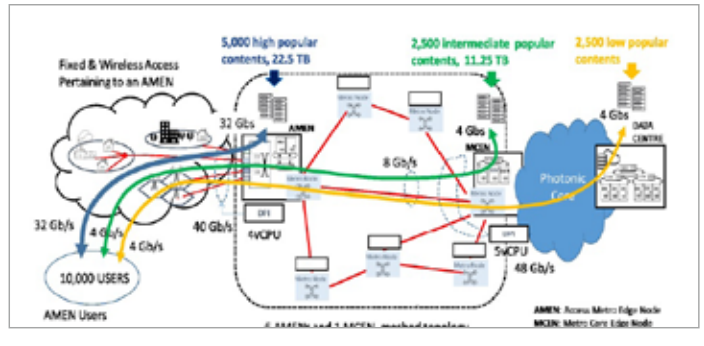

Fig. 2 - An example of a deployment of edge-nodes for a VoD content distribution case study.

ON THE CONTINUOUS AND REACTIVE ANALYSIS OF A VARIETY OF SPATIO-TEMPORAL DATA

Marco Balduini - Supervisor: Prof. Emanuele Della Valle

In recent years, an increasing number of situations call for reactive decisions making process to support solutions able to exploit heterogeneous streaming data. In this context, the urban environment results particularly relevant, because there is a dense network of interactions between people and urban spaces that produces a great amount of spatio-temporal fast evolving data. Moreover, in a modern city there is a multitude of stakeholders who are interested in reactive decisions for mobility management, tourism, etc. The growing usage of location-based social networks, and, in general, the diffusion of mobile devices improved the ability to create an accurate and up-to-date representation of reality (a.k.a. Digital footprint or Digital reflection or Digital twin). Five years ago, the state of the art was exploiting only a single data source (either social media or mobile phones). However, better decisions can result from the analyses of multiple data sources simultaneously. Multiple heterogeneous data sources, and their simultaneous usage, offer a more accurate digital reflection of the reality. In this context, we investigate the problem of how to create a holistic conceptual model to represent multiple heterogeneous spatio-temporal data and how to develop a streaming computational model to enable reactive decisions. The main outcomes of this research are: a) FraPPE conceptual model, b) RIVER streaming computational model and c) its implementations. FraPPE is a conceptual model, more precisely an ontology, to represent spatio-temporal data. In particular, it exploits digital image processing terms to bridge the gap between the data engineer perspective and visual data analysis perspective. FraPPE uses a digital image processing metaphor to enable visual analytics on spatio-temporal data, it proposes to capture the evolution of a physical world over time as a sequence of Frames. A Grid sits between the physical world and the Frames of the film. The physical world is decomposed in Cells (four in Figure 1), and each Frame is decomposed in Pixels. The Cells contains the Places where Events happen on different time. The Frame is the time-varying representation of the Grid, and the Events are captured in the Pixels. During my PhD, we first formalize the spatial and temporal concepts in FraPPE 1.0, and, then, we add concepts related to the provenance and the content in FraPPE 2.0. We check the adherence of both versions of FraPPE to the five Tom Gruber's principles, and demonstrate the validity of the conceptual model in the real world use cases that we illustrate below. RIVER is a streaming computational model. It is inspired by two principles: **(P1)** *everything is a data stream* - a variety-proof stream processing engine must indifferently ingest data with different velocities from any sources and of any size -, and **(P2)** *continuous ingestion* - the data in input is continuously captured by the system and, once arrived, it is marked with an increasing timestamp. The key innovation of RIVER is the way it tames variety and velocity simultaneously. Most of the stream processing engines in the state-of-the-art transform and adapt data at ingestion time. Contrariwise, RIVER is built around the idea of *Lazy Transformation*. So, a system that implements RIVER postpones data transformations until it can really benefit from

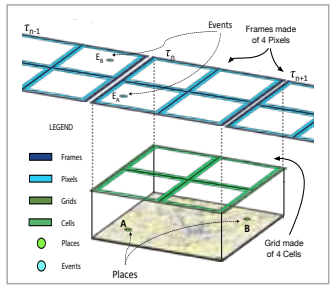

Fig. 1 - High-level overview of FraPPE.

them. Our hypothesis is that *Lazy Transformation* saves time and resources. RIVER relies on two main concepts: Generic Data Stream (S⟨T⟩) and Generic Time-Varying Collection (C⟨T⟩). Moreover, it proposes five different operators in order to ingest, process and emit data. Figure 2 depicts the five classes of RIVER operators and their interactions. The operators, defined as S2C⟨T⟩, C2C⟨T,T'⟩ and C2S⟨T⟩ allow to move from S⟨T⟩ to C⟨T⟩ and vice-versa, the *ingestion* (defined as IN⟨T⟩) and *emission* (defined as OUT⟨T⟩) operators allow to ingest and emit external data flow to/from a system implemented using RIVER. RIVER comes with the Pipeline Definition Language (PDL) -- our graphical language to abstract the operators' implementation complexity – that allows users to define computational plans, in the form of pipelines. In this thesis, we propose three different implementations of RIVER: Natron -- a singlethreaded vertically scalable implementation --, rvr@Spark and rvr@Hive -- two horizontally scalable implementations based on distributed technologies (Spark and Hive). In order to prove the validity of the *Lazy Transformation* approach, we first evaluate Natron against our legacy Streaming Linked Data engine that performs the data transformation at ingestion time. The result of this evaluation shows that Natron is cheaper -- it consumes less resources in terms of memory and CPU load -- and better approximates the correct answer under stress conditions. Moreover, we evaluate the cost effectiveness of Natron against rvr@Spark to prove that a distributed solution does not pay in all the situations. Indeed, in a mobile telco analysis, we observe that Natron is more cost-effective than rvr@Spark up to the scale of a nation. The results of those evaluations demonstrate the validity of the *Lazy Transformation* approach (assumed as a third principle **P3**) and confirm, in the stream processing engine field, that a distributed solution does not pay at all scale. Last but not least, in order to prove the feasibility and the effectiveness of FraPPE and RIVER in enabling reactive decision-making processes on heterogeneous streaming spatiotemporal data, we present five real world use cases in Milan and Como. For instance, Figure 3 illustrates a real case of visual analytics during the Milano Design Week with a grid of 6x3 cells overlaid to a city street map. Green circles visually represent the number of tweets posted in a time interval from each cell. The fill color opacity value is, instead, mapped to the number of mobile calls from each cell. During those case studies, we proved the guessability of our visual analytics interfaces discussing our data visualizations with different audiences (public users and stakeholders). Finally, we reflected on the limitations of FraPPE, RIVER and PDL to state the future directions of this research work. In particular, those reflections involve i) the future evaluation of FraPPE capabilities at Macro Level and the monitoring the OBDA field to improve FraPPE expressiveness in order to broadening the range of usage fields and fostering its adoption. ii) The future evaluations of RIVER against longer and more complex use cases and the definition of RIVER operators' cost model in order to enable automatic optimizations of the pipelines defined using PDL.

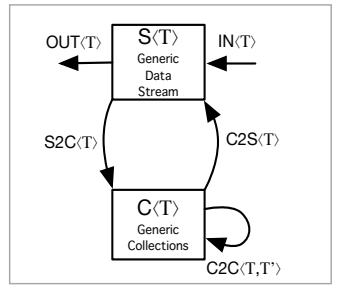

Fig. 2 - Overview of RIVER operators.

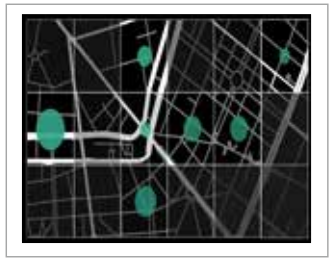

Fig. 3 - Visual interfaces created exploiting FraPPE and RIVER

WIRELESS PARTIAL-DUPLEX COMMUNICATIONS

Hamid Reza Barzegar - Supervisor: Prof. Luca Reggiani

Wireless systems cannot transmit and receive on the same frequency band at the same time (Full-Duplex wireless communications) due to strong Self-Interference (SI). The duplexing of transmission and reception must be done via either frequency division, i.e. (FDD), or time division, i.e. (TDD). In fact, Full-Duplex (FD) wireless communication means facing a high amount of SI that should be cancelled.
This research activity has been focused on the analysis and design of new schemes for wireless communication based on the reuse of the same channel for both communication directions (FD) in order to increase link performance in terms of transmission range or capacity. The proposed solutions are intended to be complementary to the implementation of efficient self-interference cancelers and they should be integrated into innovative resource allocation strategies.
Therefore, in order to overcome these issues, this research introduced a new scheme, named as Partial-Duplex (PD) approach: this solution consists of a communication link with the capability of supporting the connection in both directions at the same time in a portion of the bandwidth and with a frequency division for up-link and down-link in the rest of the band. The rationale behind this approach is to limit the level of SI finding a compromise between Half-Duplex (HD) and FD transmission and relaxing consequently the constraints on the echo canceller design in order to increase the distance range between transmitter and receiver. Equivalently, PD transmission aims to increase the overall bidirectional system rate w.r.t. an equivalent HD system, relaxing at the same time the high Self-Interference Cancelation (SIC) requirements that practical FD systems have to provide.
This hybrid transmission method between classical HD and FD is considered for point-to-point single career and OFDM links experiencing flat AWGN channels and also frequency selective fading.
In the first part of the study, the regions of SIC performance where PD systems outperform HD ones in terms of achievable spectral efficiency are analyzed by deriving the analytical distributions of the spectral efficiency gain regions in presence of random frequency-selective Rayleigh fading, for different strategies in the selection of FD sub-carriers in PD schemes; therefore, it is investigated the potential of this hybrid method, highlighting the role of the different parameters involved and the peculiarities of this flexible system design approach.
Nevertheless, we have wondered how to see this approach in the upcoming next generation of wireless networks 5G, where some common trends and promising technologies have been already identified. In addition to capacity gain and improvement of performance, which are expected from network densification, device-to-device communication and small cells, we can mention an increased spectrum sharing and integration as well as spectrum enhanced carrier aggregation and other advanced wireless communication technologies like massive MIMO and resource management based on machine learning.
Machine-type communication is one of the new use cases of 5G and anticipated to significantly increase both the number of connected devices to the network connections and traffic, respectively. In these cases, a proper use of wireless PD communication can significantly improve the overall throughput of the mobile network.
In this context, the second part of this study has been focused on the channel encoding for PD. One of the best candidates for efficient channel coding and error correction, is the class of Low-Density Parity-Check (LDPC) codes. We have identified that, in PD communication, the receiver faces a mixture of high and low bit (symbol) SNR in the same codeword, condition associated with a scheme in which part of the bits/sub-carriers is subject to FD interference (partial-duplex scheme). LDPC codes have shown, also w.r.t. turbo and polar codes, the best performance for PD, and we have studied how to optimize further their performance in this specific context.
Digital Video Broadcasting-Satellite-Second Generation (DVB-S2) standard was one of the standards including LDPC. These codes were implemented for long code-words and we proposed to create LDPC codewords with arbitrary length, derived from the original DVB-S2 and with the same parity check matrix structure.
Then, the application of this class of LDPC codes, derived from DVB-S2, to PD communication can be improved by a specific allocation techniques of the high and low SNR bits, suited to PD communications. The main innovative results of this activity are related to the fact that, in PD schemes, part of the band is transmitted in FD and the rest in HD and, consequently, some transmitted bits (in single carrier) and sub-carriers in presence of OFDM will be characterized by high SNR and the others by low SNR according to a pattern which is known, a-priori, by the system. Therefore, combining properly the patterns of these high and low SNR bits affects the coding performance of the system in a way that depends also on the parity check matrix structure of the code used in the transmission.
In addition, in order to validate further the results and according to upcoming 5G standardization, this channel encoding procedure has been applied also to the 5G encoding schemes recently considered by 3GPP. Mainly two types of channel coding are adopted in 5G, Polar and LDPC codes. Therefore, in this part of the study, we have investigated and compared our approach in the context of 3GPP standard for 5G and LTE, showing the performance of encoded PD. Results have turned out to be really promising for a specific class of LDPC, when, as in PD, we can exploit the a-priori knowledge of high and low SNR bits in the transmitted codeword.

ADVANCES IN WAVE DIGITAL MODELING OF LINEAR AND NONLINEAR SYSTEMS

Alberto Bernardini - Supervisor: Prof. Augusto Sarti

This doctoral dissertation, authored by Alberto Bernardini and supervised by Prof. Augusto Sarti, presents a contribution to the recent evolution of modeling and implementation techniques of linear and nonlinear physical systems in the Wave Digital (WD) domain. The overarching goal of WD methods is to build digital implementations of analog systems, which are able to emulate the behavior of their analog counterpart in an efficient and accurate fashion. Though such methods usually focus on the WD modeling of analog audio circuits; the methodologies addressed in this thesis are general enough as to be applicable to whatever physical system that can be described by an equivalent electric circuit, which includes any system that can be thought of as a port-wise interconnection of lumped physical elements. The possibility of describing systems through electrical equivalents has relevant implications not only in the field of numerical simulation of physical phenomena, but also in the field of digital signal processing, as it allows us to model different kinds of processing structures in a unified fashion and to easily manage the energetic properties of their input-output signals. However, digitally implementing nonlinear circuits in the Kirchhoff domain is not straightforward, because dual variables (currents and voltages) are related by implicit equations which make computability very hard. Spice-like software, based on the Modified Nodal Analysis (MNA) framework, is not always suitable for realizing efficient and interactive digital applications, mainly because it requires the use of iterative methods for solving multi-dimensional systems of equations. WD Filters (WDFs) are a very attractive alternative. During the seventies, Alfred Fettweis introduced WDFs as a special category of digital filters based on a lumped discretization of reference analog circuits. A WDF is created by port-wise consideration of a reference circuit, i.e., decomposition into one-port and multi-port circuit elements, a linear transformation of Kirchhoff variables to wave signals (incident and reflected waves) with the introduction of a free parameter per port, called reference port resistance, and a discretization of reactive elements via the bilinear transform. Linear circuit elements, such as resistors, real sources, capacitors and inductors, can be described through wave mappings without instantaneous reflections, as they can be all "adapted" exploiting the mentioned free parameter; in such a way that local delay-free loops (implicit relations between port variables) are eliminated. Series and parallel topological connections between the elements are implemented using scattering topological junctions called "adaptors", which impose special adaptation conditions to eliminate global delay-free loops and ensure computability. It follows that WDFs, as opposed to approaches based on the MNA, allow us to model separately the topology and the elements of the reference circuit. Moreover, WDFs are characterized by stability, accuracy, pseudo-passivity, modularity and low computational complexity, making many real-time interactive applications easy to be realized. Most classical WD structures can be implemented in an explicit fashion, using binary connection trees, whose leaves are linear one-ports, nodes are 3-port adaptors and the root may be a nonlinear element. However, WDFs are also characterized by important limitations. The first main weakness of state-of-the-art WDFs is that WD structures, characterized by explicit input-output relations, can contain only one nonlinear element, as nonlinear elements cannot be adapted. In fact, the presence of multiple nonlinear elements might affect computability, which characterizes classical linear WDFs, as delay-free loops arise. As a second main limitation of traditional WDFs, up to three years ago, there were no systematic methods for modeling connection networks which embed non-reciprocal linear multi-ports, such as nullors or controlled sources. Finally, very few studies were presented on the use of discretization methods alternative to the bilinear transform and potentially varying step-size in WD structures. This thesis presents various techniques to overcome the aforementioned limitations. After a review of the state of the art on WD modeling of lumped systems up to 2015, this thesis begins with redefining wave signals, first by offering a unified definition that accommodates the existing ones (those based on one free parameter per port), then by introducing a new class of waves with two free parameters per port, which leads to WD structures that exhibit a doubled number of degrees of freedom, called "biparametric" WDFs. The second part discusses how to implement nonlinear one-port and multi-port elements in the WD domain; finding a compromise between accuracy, efficiency and robustness. In particular, it shows how scalar nonlinearities can be represented in canonical piecewise-linear form in the WD domain; it presents a technique based on the Lambert function to make a class of exponential nonlinearities, e.g., diodes or Bipolar Junction Transistors, explicit in the WD domain; and it provides an in depth discussion on the modeling of nonlinear 3-terminal devices with various examples of applications on circuits containing transistors. The third part focuses on the description of connection networks as WD adaptors. In particular, this part discusses how to model arbitrary reciprocal and non-reciprocal junctions in WD structures based on various definitions of wave signals and how to compute waves reflected from the scattering junctions in an efficient fashion. The fourth part focuses on the cases in which we cannot do without iterative solvers. In particular, it introduces a novel relaxation method based on WD principles, developed for implementing circuits with multiple nonlinearities. The method is called Scattering Iterative Method (SIM). A proven theorem guarantees that SIM converges when applied to circuits with an arbitrary number of nonlinearities characterized by monotonic voltage-current characteristics. As an example of application of this method, it is shown that power curves of large Photovoltaic (PV) arrays (constituted of thousands of nonlinear PV units) with arbitrary topologies can be obtained far more efficiently using the proposed SIM than using MNA-based approaches, (SIM is at least 30 times faster). It is also shown that SIM is highly parallelizable and suitable for the implementation of time-varying circuits. Moreover, a strategy for deriving WD models of dynamic elements based on arbitrary linear multi-step discretization methods with potentially adaptive time-step size is here introduced and discussed. This strategy allows us to describe capacitors and inductors using time-varying Thevenin or Norton equivalents, and proves particularly useful in conjunction with SIM for implementing dynamic circuits with multiple nonlinearities. These results also motivated the simulation of some audio circuits (e.g., a dynamic ring modulator), which proved SIM to be a promising implementation method that is also employable for virtual analog applications, as it is characterized by high efficiency, parallelizability and inherent capability of handling time-varying elements. The combination of the new techniques for modeling arbitrary reciprocal and non-reciprocal connection networks with SIM is promising for the future development of a general purpose simulation program which might be more efficient than Spice-like software, as far as the time-domain analysis of arbitrary nonlinear circuits is concerned. As a further example of application of WD principles, the last part of the thesis discusses a novel approach for implementing a class of differential beamformers using WDFs.

TOOLS, SEMANTICS AND WORK-FLOWS FOR WEB AND MOBILE MODEL DRIVEN DEVELOPMENT

Carlo Bernaschina - Supervisor: Prof. Piero Fraternali

The proliferation of web enabled mobile devices in the last decade has presented great opportunities and challenges. In particular we have seen a shift from general purpose complex desktop applications to specialized web and mobile applications with particular focus on user experience, development time and cost. The industry addressed the challenge by introducing development tools specifically targeted at simplifying application development in these use-cases. Particular attention is devoted to Low Code/ No Code Development Tools, which are a valuable, and cost effective, alternative to classical development techniques. These tools are vendor specific and their selection is a critical step in the application development. Projects can be delayed or even fail due to vendor lock-in. In the academia Model Driven Engineering (MDE), and in particular Model Driven Development (MDD), has been adopted as a valuable approach able to address this challenge. Model Driven Engineering is the branch of Software Engineering that emphasizes the use of models, i.e., simplified descriptions of the application that capture its essential aspects at a certain level of abstraction, e.g., independently of the platform for which the application will be designed and of the technologies with which it will be implemented. Abstraction is the most important aspect of MDE and MDD. It enables developers to validate high level concepts and introduce details, which accommodate complexity, at later stages in the process. This distinction gives importance to a proper balancing between what is considered a high level concepts and an implementation detail. The goal of the research presented in this thesis is to propose tools, semantics and work-flows aimed at reducing the costs of Model Driven Development, especially in the field of web and mobile applications. We address the development and evolution of MDD tools, the semantics of modeling languages and the work-flow enabling the collaboration between developers and MDD experts. We specifically focus on the use-case of Web and Mobile applications development, given its unique set of requirements, i.e. details like styling of the User Interface (UI) or device sensors support are crucial. Differently from recent approaches, focused on the overloading of existing languages with newer and newer features, we focus on basic concepts, i.e., the flow of information in an application, and collaboration between developers and non MDD experts. In this thesis we explore the feasibility of reducing the friction between developers and the Model Driven Development approach, by proposing a set of tools, semantics and work-flow focused on the development of Web and Mobile applications via the high level definition and generation of the application front-end and the parallel manual introduction of requirements not captured by the high level definition. We present ALMOsT.js, an OpenSource agile model driven transformation framework for JavaScript, which enables non Model Driven Development experts to experiment with the MDD methodology and easily integrate the results in existing on-line tools. We define a Web and Mobile centric formal semantics for the Interaction Flow Modeling Language (IFML) (Figure 1). Instead of focusing on its extension, required to describe all the details involved in Web and Mobile development, we focus just on high level core concepts in order to give a unique and precise interpretation to each valid model.

DATA-DRIVEN AND HANDCRAFTED FEATURES FOR FORENSICS ANALYSIS AND SOURCE ATTRIBUTION

Luca Bondi - Supervisor: Prof. Stefano Tubaro

The communication power associated with visual content makes digital images a powerful and effective tool to deliver messages, spread ideas, and prove facts.
Smartphones, digital cameras, and camcorders are becoming more affordable every day, and thus constitute a rapid and convenient way of capturing and sharing photos quickly and inexpensively. The increasing diversity of brands, models, and devices, together with the ever-growing access to social network and picture sharing platforms poses a set of challenges, from the diffusion of illegal content to copyright infringement.
The wide availability and ease of use of image manipulation software makes the process of altering an image simple and fast. This could severely reduce the trustworthiness of digital images for users, legal courts, and police investigators. The fake news phenomenon is a well-known and widespread example of the malicious use of digital pictures and manipulation software. Modified images done with precision are used to create false proofs for made-up stories, exploiting the often-unquestionable trust with which readers take in visual content.
In this thesis we face several challenges related to the analysis of digital images. A first step in assessing image authenticity, and tracing an image back to its origins, consists in determining which device shot a specific picture. State-of-the-art techniques based on Photo Response Non-Uniformity (PRNU) prove to be very effective in determining the specific sensor that shot a picture. However, given the highly increasing number of devices, a full-range search over all the existing devices is impractical and time consuming. One of the ways to reduce the search space is to first find the camera model that took a picture, then test the image under analysis against the devices from the same camera model.
In this thesis we present the first data-driven method for camera model identification, showing how

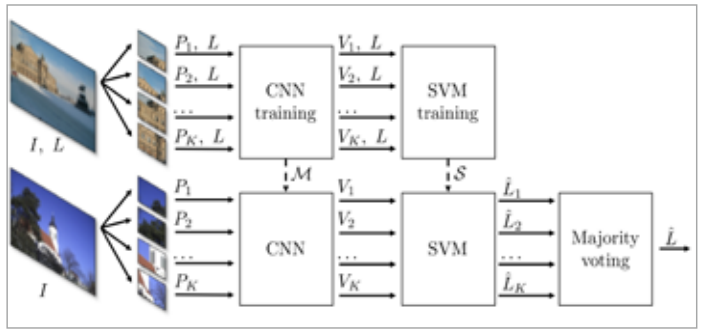

Fig. 1 - Proposed pipeline for camera model attribution

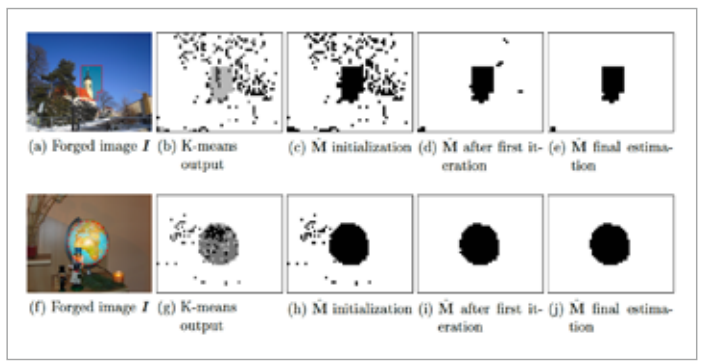

Fig.2 - Examples of tampering localization

modern deep-learning techniques based on Convolutional Neural Networks (CNN) can be adapted to multimedia forensics tasks. The pipeline of the proposed method is shown in Figure 1.
When it comes to a large-scale search of picture-device matches based on PRNU, at least two challenges arise: time and storage space constraints. To address such challenges, the forensics community explored a series of techniques to compress PRNU fingerprints and residuals. In order to reduce storage space requirements, while lowering the computational complexity, we introduce two techniques to address PRNU compression, by exploiting classical signal processing analysis and data reduction techniques. While determining the origin of a digital image is important to solve copyright infringement cases, digital images can be locally altered by adding, removing, or modifying objects with the goal of changing the semantics of the image. We present how to exploit the features learned with a CNN trained for camera model identification with the goal of detecting and localizing tampered regions within an image. Figure 2 shows two examples of the iterative procedure for tampering localization.

Under both device identification and camera model identification perspectives, we study a set of possible antiforensics attacks tailored at anonymizing an image to prevent the correct identification of its origin. This allows us to understand the limitations and weaknesses of the proposed camera model and device identification techniques. Finally, we leverage the knowledge and skills acquired in mixing together handcrafted signal processing and data-driven methods in two different forensics applications: Laser Printer Attribution and Single versus Double JPEG Detection. In both scenarios the key to tackle the forensics task at hand is fusing together a proper signal pre-processing technique with a carefully designed data-driven system.

FABRICATION AND CHARACTERIZATION OF RESISTIVE SWITCHING MEMORY DEVICES FOR HIGH-DENSITY STORAGE AND IN-MEMORY COMPUTING

Alessandro Bricalli - Supervisor: Prof. Daniele Ielmini

The von-Neumann architecture, which relies on the separation between logic and memory, is at the heart of modern computers. However, this architecture is showing major limitations especially in data-centric applications, which are becoming more and more common, mainly because of the wide performance gap existing between the memory and the central processing unit (CPU). In order to overcome this issue, fast and non-volatile memory technologies able to better sustain frequent data exchange with the CPU are being researched. Moreover, novel computing paradigms better suited for highly parallel data processing are being developed for enhanced performance and reduced power consumption in data-intensive tasks, such as image and voice recognition. Neuromorphic computing uses neurons and synapses as basic elements for computation, which eliminates the separation between memory and logic, similarly to the what happens in the human brain. Inmemory computing, instead, aims at performing logic operations directly inside the memory, thus eliminating the need to transfer data between the memory and the CPU. In this scenario, the study of novel devices such as resistive non-volatile memories (ReRAMs) and the development of highly engineered materials will play a key role in the implementation of new system for efficient computing. This work focused on the fabrication and characterization of memristive devices with vertical and planar structures and their use in novel systems for in-memory computing applications. The first part of the work consisted in the development of our own ReRAM technology. A ReRAM cell typically consists in a metal-insulator-metal (MIM) stack in which a conductive filament of defects can be created in the insulating layer, upon the application of a relatively low voltage between the two metallic electrodes. This consists in a soft breakdown event which is partially reversible, thus allowing to later program the device in at least two states. Those states can be distinguished by their respective resistive value, typically referred to as the low-resistance state (LRS) and the high-resistance state (HRS). One of the main issues with ReRAM is related to the variability of both HRS and LRS due to the stochastic nature of the switching event; this can be solved by increasing the ratio between the HRS and the LRS, namely the resistive window (RW), in order to compensate for the device variability. For this reason, we chose a material, i.e. silicon oxide (SiOx), that could in principle allow for a larger RW, thanks to its large band-gap and therefore better isolation properties of the HRS. The devices were obtained by evaporation of an ultra-thin SiOx layer and a titanium top electrode (TE) in sequence, on top of a carbon bottom electrode (BE). The films thickness and deposition rate were optimized to obtain the best electrical switching behavior, resulting in a very large RW at relatively low operating voltages and currents. Moreover, the HRS demonstrated to be highly tunable and stable, which is promising for multilevel cell operation and neuromorphic applications. The maximum endurance in pulsed regime proved to be in the order of more than 107 cycles, comparable to the state of the art and much better than common Flash technology. Finally, data retention at high temperatures was confirmed up to 260° C. Arrays of non-volatile emerging memories typically also require a selector in series to each memory element, i.e. a volatile switching device with a low subthreshold current, able to limit the leakage currents inside the array. In our work, a volatile switching device based on SiOx was also demonstrated by using a silver TE. The device demonstrated an on/off ratio of more than 107, which is interesting for use as a selector in crossbar arrays of resistive elements; however, the large spreading in the switching-off time limits the application of the device in high-performance memory application, while it may be interesting for neuromorphic computing. Afterwards, the development of the process for a full cross-point array of resistive memories was addressed. The geometry of the array was optimized to achieve a good uniformity of the oxide layer on the cell area for best switching uniformity. Both single memory elements and a full array were fabricated and tested. The devices showed a good RW of more than one order of magnitude, along with reasonable endurance, low variability and low voltage and current operation. In the near future, the implementation of a full array and the circuitry needed to drive it will allow the demonstration of true in-memory computing algorithms with ReRAM devices. Resistive switching can be exploited not only in common MIM structures, but also in planar devices. Particularly, we wanted to explore the possibilities offered by 2D TMDs such as MoS2, which are promising for the creation of low-dimensional, 3D stackable devices. TMD-based transistors have been widely investigated in literature and memristive effects mediated by 2D materials have attracted a lot of interest in recent years. We investigated the memristive behavior of such devices due to metal ion migration between the metallic source and drain contacts. By optimizing the lithographic technique to achieve a short channel, it was possible to demonstrate resistive switching phenomena in gated structures. This kind of device is very promising for use in neuromorphic applications such as the implementation of short-term potentiation and spike-rate dependent plasticity (SRDP) rules. Finally, the ReRAM devices described in the first part were used to demonstrate a preliminar implementation of brain-inspired logic gates and linear algebra operations using cross-point arrays. The logic gates exploit threshold switching in ReRAM devices to implement logic operations; using this approach, all the linearly separable logic functions can be implemented in just one step, while non-linearly separable functions such as XOR or 1-bit full adder can be implemented in multiple steps. Linear algebra operations, instead, can benefit from a cross-point structure in that Ohm's law in a cross-point array is the equivalent to a matrix-vector multiplication. By extending this concept, it was possible to obtain a circuit which allows to easily solve linear systems without iteration, which is a crucial problem in many common algorithms. In conclusion, this work presented a comprehensive study of emerging devices for the implementation of high-performance non-volatile memory arrays and in-memory computing systems.

LEARNING AND ADAPTATION TO DETECT CHANGES AND ANOMALIES IN HIGH-DIMENSIONAL DATA

Diego Carrera - Supervisor: Prof. Giacomo Boracchi

A primary concern in many applications is to detect whether a process departs from its normal conditions, since this could indicate a problem that has to be promptly alarmed and solved. Most often, changes have to be detected by monitoring a datastream generated by the process of interest. In this thesis, we investigate change and anomaly-detection problems from both a theoretical and a practical perspective. In particular, we formally study the intrinsic difficulty of monitoring a datastream when the data dimension increases and propose novel algorithms to perform monitoring in two real world scenarios: a quality inspection system monitoring the production of nanofibrous materials through the analysis of Scanning Electron Microscope (SEM) images, and ECG monitoring using wearable devices.

When monitoring this kind of datastreams we have to address three main challenges. At first, data feature complex structures and are characterized by a high dimension, and there does not exist any analytical model to describe them. Therefore, to perform successful monitoring, we have to resort to data-driven models, that are learned directly from data. However, we can acquired only data in normal conditions, since collecting data in alternative conditions is difficult, when not impossible. For example, in case of ECG monitoring it it is not possible to collect a large number of anomalous heartbeats of the same user, since these might be due to potentially dangerous arrhythmias. This limitation prevents the use of supervised models, such as classifiers to address this kind of problems, since we do not have training data for both normal and anomalous classes. The only viable solution is to resort to unsupervised learning techniques to learn a data-driven model that describes only normal data.

The second challenge to be addressed is the design of specific indicators and decision rules that allow to successfully detect whether data are generated in alternative conditions. Moreover, the amount of false detections has to be controlled and kept under a given tolerance level to make the detector effective in practical applications.

Finally, the third challenge to be addressed is domain adaptation since normal conditions might change over time. Thus, a model learned during the training phase, i.e., on data in the source domain, might not be able to describe new incoming data in the target domain. For example, in case of ECG monitoring, the heartbeats of a user get transformed when the heart rate increases, thus the morphology of heartbeats acquired during everyday activity is not the same of training ones, acquired in resting conditions.

We address all these challenges from two different perspectives. At first, we model data as realization of a random vector, i.e., we assume that data can be described by a smooth probability density function. In these settings we focus on the change-detection problem, where the goal is to detect permanent changes affecting the data-generating process. We investigate the intrinsic difficulty of performing change detection when the data dimension increases. In particular, we show the a popular change-detection algorithm that monitors the likelihood w.r.t. to a learned model suffers from the detectability loss, namely a decay in the change-detection performance when the data dimension increases as the change-magnitude is kept fixed. We analytically prove this result in case of Gaussian datastreams, and empirically in case of real world data. To experimental investigate the detectability loss, we develop CCM (Controlling Change Magnitude), a framework to manipulate real world datasets and introduce changes having a controlled magnitude.
Finally, we propose QuantTree, a novel algorithm to learn histograms for change-detection purposes. Peculiarity of the adaptive splitting scheme adopted by QuantTree is that it enables the non-parametric monitoring of datastreams. In particular, we theoretically prove that any statistic defined over such histograms does not depend on the data-generating distributions. This allows to control the false positive rate disregarding the data generating distribution. Moreover, to mitigate the effect of detectability loss, which affects also QuantTree, we propose an ensemble method that combines multiple histograms. Our experiments show that the non-deterministic nature of QuantTree increases the diversity of the computed histograms, improving the detection capability of the ensemble.

In the second part of the thesis we adopt a different modeling assumption: we assume that normal data can be well described by a dictionary yielding sparse representations, which is a model that have been successfully used in several signal and image processing applications. Moreover, we consider the anomaly detection problem, where the alternative conditions are not persistent, but affect only sporadic samples in the datastream.
We design an anomaly-detection algorithm that learns a dictionary yielding sparse representations of normal data and use these representations to extract low-dimensional indicators to assess whether new data conform or not to the learned dictionary, thus the normal conditions of the process. We propose two domain adaptation algorithms to adapt the anomaly detector, namely both the learned model and the decision rule, when the process generating normal data changes over time, as this happens in practical scenarios.
In particular, we customize our general anomaly-detection algorithm to perform online and long term monitoring of ECG signals directly on a wearable device. We have shown that dictionaries modeling normal heartbeats can be successfully adapted when the heart rate -- thus the heartbeat morphology -- changes. Moreover, while dictionaries used to detect anomalies have to be user-specific, they can be successfully adapted by user-independent transformations, learned from large and publicly available datasets. Thus, a few minutes of ECG signals acquired in resting conditions are enough to configure the device for long-term monitoring. Our algorithms have been implemented and successfully tested in a demo device performing online ECG monitoring.

We show that our anomaly-detection algorithm that can also successfully in a quality inspection system to detect defects in nanofibrous materials. Moreover, we make our algorithm scale-invariant by using a multiscale dictionary that aggregates atoms learned from synthetically resized normal images, and performing sparse coding by enforcing the group sparsity of the representations. This regularization term turns to be essential to achieve superior anomaly-detection performance.
Our experiments conducted on a large dataset of SEM images show that the proposed algorithm can effectively detect also tiny defects and demonstrate it can effectively handle changes in magnification level that typically occurs in industrial imaging applications.

Finally, we investigate the use of convolutional sparse representation as image model and in particular in white noise denoising. In particular, we show that these translation-invariant representations outperform the traditional sparse representations only when the image admits an extremely sparse representation, while the two approaches attain comparable performance in case of natural images. We explain this phenomenon by separately studying the bias and variance of these solutions, and by noting that the variance of the global solution increases very rapidly as the original signal becomes less and less sparse.

SPEECH ANALYSIS FOR AUTOMATIC PROSODY RECOGNITION

Sonia Cenceschi - Supervisor: Prof. Licia Sbattella

The thesis, developed in the context of the ARCSLab (arcslab.dei.polimi.it) research activities, presents a wide-ranging research work on prosody whose structure is shown in Figure 1. Prosody is defined as the group of audio paralinguistic and suprasegmental clues involved in the communicative and understanding process of human speech. According to the main Universals in language, each Speech Act expresses needs (e.g., talking about past or future events) that are similar for all humans, and are acoustically realized through linguistic and phonotactic language-related rules. At the same time, a spoken message can be uttered with a variable prosody because of countless factors as social context, emotions, intentions, rhetoric or spatial dislocation.

This work starts proposing a new descriptive model in order to analyse prosody complexity in a structured and orderly manner. The result is the **CALLIOPE model** (Combined and Assessed List of Latent Influences On Prosodic Expressivity), a conceptual multidimensional space defining all possible independent factors affecting the prosody of an Information Unit (IU), which is the smallest unit of analysis, intended as a sentence, or a portion of it, autonomous from the illocutionary point of view. An IU is composed by a vocal recording and its transcription.

A **psychoacoustic experiment** validates a subset of the model. It exploits the **SI-CALLIOPE corpus**, a series of IUs recorded in collaboration with actors of *Libro Parlato Onlus* and with actors and musicians of *Fondazione Sequeri Esagramma Onlus*.

The experiment investigates the influences of semantics, phonotaxis and intonation on the understanding processes of IUs, using three audio typologies: original sentences with real words, their pseudo-words versions and their pitch envelopes. Real and pseudo-words are in the corpus, while pitch envelopes are generated with the software Praat. The experiment asks listeners to recognize: *Exclamation, Statement, Vocative and Interrogative strucutures, Enumeratio, Aposiopesis*, and *Speech Loudness and Corrective Focus*. Results confirm that the understanding process requires a combination of linguistic and vocal cues, but also achieve interesting results for some tasks, for which prosody plays a major role or the linguistic meaning does not improve the recognition accuracy.

These conclusions are useful for the second part of the work, regarding the following two Automatic Prosody Recognition (APR) systems, both regarding the Standard Italian language.

PESI-net is a Neural Network able to classify an IU as a Question, Exclamation or Statement. It reaches an accuracy of about 68% with audio-only input, and 80% when considering both audio and textual information. When reduced to two classes (Question/Non-question), the network reaches 91%. Considering that the developing time has been too limited to allow for extensive optimizastions, results are really good.

PESI-net is composed by one Audio-based NN, one Text-based NN and a Master NN, which combines the results of the other two. Both the Audio and the Text-based NNs are based on a multi-layer, convolutional, bidirectional, Recurrent NN (RNN), based on Long Short-Term Memory (LSTM) nodes.

CoFOCUS-net, is based on bidirectional RNNs able to recognize the presence of the Corrective focus (only acoustic clues are considered for this task). Two different approaches to build the contrastive focus detection have been investigated: detection of focused syllables achieves an F-score of 0.693, while classification of the whole IU as focused or not achieved a really high F1-score of 0.826.

The SI-CALLIOPE corpus is too small for training PESI-net, while it is enough to generate the **CoFOCUS audio corpus** and train the CoFOCUS-net. PESI-net, instead, exploits the **ExInDe corpus**, based on a huge collection of EPUB3 audio books and TV shows.

The last part of this work is the **semi-automatic analysis** of the prosodic development in the EVA (Affective Vocal Education) recordings. EVA is an Esagramma Methodology® (www.esagramma.net) empowering expressiveness methodology for children with autism, intellectual and linguistic disabilities, activated by Fondazione Sequeri Esagramma (Milano, Italy) in October 2013. Analysing pre- and post- tests with semi-automatic methodologies we provide *objective* evidence of improvements (in contrast to the common *subjective* evaluation), laying the ground for further automatic recognition systems focused on prosodic skills.

The Polisocial (www.polisocial.polimi.it) LYV (Lend Your Voice) project is as an example of application of this thesis. LYV focuses on stimulating the improvement of prosodic skills of Italian young speakers with autism, intellectual and linguistic disabilities or Italian students with Specific Learning Difficulties (SpLD), as dyslexia, learning English as a second language, through the use of vocal, and technologically mediated, prosodic storytelling sessions. PESI-net is a central part of the LYV module related to the prosodic interaction shown in Figure 2. It plays the role of an assistant tool during the game, and integrated inside the editor gives the possibility to insert prosodic exercises into the story. The main contributions of this thesis are the definition of a new multidimensional conceptual model describing prosodic forms, two NN-based architectures for detection of structures and corrective focus, two new audio/textual corpuses composed by recited and read speech and used to feed the NNs, and a proposal for the semi-automatic analysis of some aspects of prosody and expressiveness. Results outlined in this thesis allowed to draw some conclusions on the interdisciplinary analysis of prosody through the combination of different perspectives to start a systematic and complete work on APR.

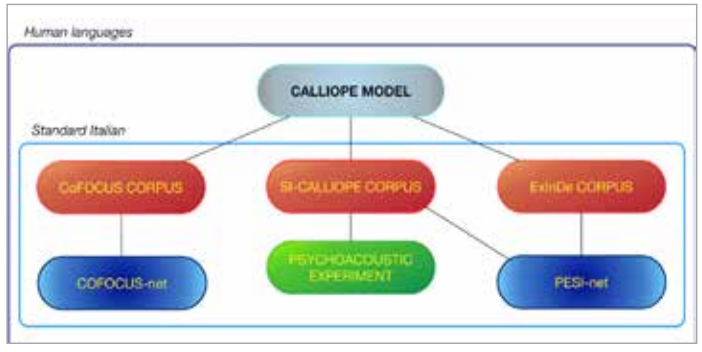

Fig. 1 - Structure of the thesis.

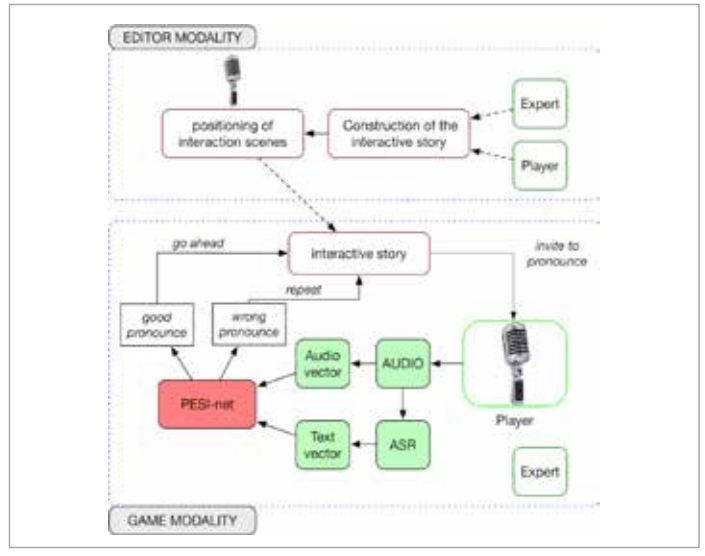

Fig.2 - The LYV prosodic interaction module.

DIGITALLY-INTENSIVE FREQUENCY MODULATORS FOR MM-WAVE FMCW RADARS

Dmytro Cherniak - Supervisor: Prof. Salvatore Levantino

The cost reduction is the cornerstone in nowadays radar systems. The radar technology, which was for a long time (since the invention in early 1930s) exclusive to military and defence, is emerging to wide spectrum of applications including automotive, industrial and consumer market. In the context of advance driver assistance systems and autonomous driving the radar technology has been employed for about two decades. In automotive applications the radar sensors are used for object detection as well as range, relative velocity and azimuth sensing, which in combination with machine learning enables vision of the vehicle. In comparison with other environmental perception sensors such as cameras and light detection and ranging (LIDAR), radar operates under foggy, dusty, snowy and badly lighted environment, which is essential for automotive application. Already in 1998 the high-end luxury vehicles have been equipped with an adaptive cruise control system which was based on a 76 GHz frequency-modulated contenious-wave (FMCW) long-range radar and two 24 GHz short range radars. Further, in 2014, the high-end vehicles already had up to six radar sensors which enabled 360° sensing in near range (up to 60 m) and far range (up to 200 m).

Initial automotive radar sensors were implemented based on discrete circuit elements which results in high cost realization. The next generation of the radar systems were implemented based on multiple Monolithic Microwave Integrated Circuits (MMIC) with the RF front-end typically realized in high-performance but still expensive GaAs technology. Further cost reduction and increased level of the integration was achieved with moving to SiGe Bipolar and BiCMOS technology. The BiCMOS technology allowed efficient integration of the digital part at the same chip with the RF front-end. Most of the nowadays radar systems are realized in BiCMOS technology, but the fast development of the automotive industry demands a single-chip radar solution in full CMOS technology which have already reached speed capability compatible with millimeter-wave (mm-Wave) circuit design. Moreover, there is also a need to bridge the gap between performance requirements for long-range and short-range radar solutions and CMOS would allow a single-chip FMCW radar covering all automotive application.

In recent years with introduction of Internet-of-Things concept, which exploits the human-to-machine and machine-to-machine interaction, the mm-Wave radar sensors found its application for object detection, motion sensing, imaging as well as complex gesture recognition. All of the above applications demand low-cost full CMOS implementation with high degree of reconfigurability in order to address different applications with a single-chip solution.

The performance of the complete FMCW radar sensor is, to the great extend, determined by the chirp synthesizers speed, linearity and noise parameters. All above-mentioned FMCW radar applications would demand different chirp configuration as well as different noise requirements. In general, enlarging the modulation bandwidth improves the range resolution and lowering the phase noise increases the signal-to-noise ratio (SNR). However, reducing the period of the chirp (enabling fast chirp synthesis) would be one of the key performance improvement. Fast chirp allows better separation of the targets in the frequency domain, increases the beat frequency beyond flicker corner which is essential for deep sub-micron technologies, it increases the maximum unambiguous velocity, improves velocity resolution as well as allows averaging which would enable additional noise reduction. Thus, improving the modulation speed and the phase noise of the FMCW modulator in combination with high degree of chirp reconfiguration would allow to address with a singe-chip solution all of the discussed above applications of the radar sensor.

The primary goal of this project is to develop a novel architecture of digital PLL-based FMCW modulator which is capable of fast and reconfigurable chirp generation whilst maintaining excellent phase noise performance. In order to reach this ultimate goal there is a number of original contributions that were developed along the project work. At first, a simple yet accurate FMCW system model has been derived in order to analyze the impact of the phase noise and chirp linearity on the performance of the FMCW radar system and to extract the specifications for the PLL-based modulator. There are several works in available literature that addresses the phase noise topic in FMCW systems, but they all lack to account for the impact of phase noise on the full range-Doppler map.

In PLL-based frequency modulators there is a trade-off between modulation speed and optimum phase noise which is to a certain extent defined by the closed-loop bandwidth of the PLL. Enlarging the closed-loop bandwidth allows to achieve faster modulation, however, the spot noise as well as the absolute jitter are typically compromised. Two modulation techniques known from communication applications, namely the two-point modulation and the digital pre-emphasis, are for the first time analyzed and compared in the context of fast chirp synthesis. It is demonstrated that though both modulation techniques are equivalent in terms of suppression of gain estimate errors, the two-point modulation has the advantage of reduced required phase detector range. Moreover, a new calibration method is proposed to match the transfer function of the digital pre-emphasis filter (DPF) with the PLL one. The calibration technique is based on the least-mean-square (LMS) algorithm and operates fully in background. The analysis of the proposed LMS-based algorithm is carried out as well.

The analysis of the two modulation techniques has demonstrated that they are both sensitive to DCO nonlinearity. Thus, the new concept of DCO nonlinearity correction by means of an adaptive piecewise linear digital pre-distortion is introduced. Two new digital pre-distortion schemes are compared to the conventional one, originally proposed to correct the nonlinearity of the digital-to-time converter (DTC). Additionally, an analysis of LMS-based background calibration algorithm for DCO nonlinearity correction is carried out.

Finally, the new architecture of a digital PLL-based FMCW modulator is proposed and implemented in a prototype fabricated in 65 nm CMOS process. The 23-GHz PLL-based modulator features the two-point modulation concept and one of the novel digital pre-distortion schemes to correct DCO nonlinearity. The measurement results demonstrate state-of-the-art chirp synthesis performance (1-µs fast chirp with 200-ns rest time) as well as lowest phase noise odBc/Hz among the digital PLL-based FMCW modulators reported in open literature.

In addition, the proposed modulation techniques are also applied for DTC nonlinearity correction to reduce the level fractional spurs in a digital PLL architecture. The 18-GHz PLL prototype which features a novel piecewise linear pre-distortion techniques is fabricated in a 28 nm CMOS process. The measurements demonstrate state-of-the-art spur level of -63 dBc which to the author's knowledge is the lowest fractional spur level among frequency synthesizers in 10 – 30 GHz range.

RADIO-RELAYING OVER OPTICAL FIBER FOR CLOUD RADIO ACCESS NETWORKS

Lorenzo Combi - Supervisor: Prof. Umberto Spagnolini

The focus of this Thesis is on analog signal transport and processing for future mobile radio access networks, and its content can be broadly divided into four macro-areas:

- proposal and experimental validation of optical pulse width modulation (PWM) for analog fronthaul;
- analysis of space-frequency multiplexing of fronthaul signals without channel equalization;
- hybrid beamforming for the mmWave radio access, with analog optical signal processing and transport;
- signaling to support analog optical signal processing at the remote antennas.

Analog optical PWM is suitable for analog fronthauling as it combines the best features of digital and analog transmission. Indeed, the 2-level waveform involves a relaxation of the linearity requirements, still allowing for an analog signal transport that avoids the bandwidth expansion of digital fronthauling. PWM is experimentally validated for different optical architectures. Besides the conventional optical network layout (that can be summarized in the cascade of a laser, fiber and photodiode), here an innovative architecture is proposed, based on reflection. In this reflective PWM the transmitter (at the remote antennas) is equipped with a modulated reflective semiconductor optical amplifier (RSOA) that receives, modulates and reflects back a continuous waveform (CW) from an optical source located at the receiver (at the centralized baseband unit, BBU). The modulated reflected signal which is the uplink fronthauling is polarization-separated from the CW seeding signal.
The experimental validation proves the effectiveness of reflective PWM in the transport of RF signals with 100-MHz bandwidth through up to 20km of fiber link.
Multilevel pulse width modulation is considered as a viable option to increase system capacity, and it allows to effectively transmit up to 16 aggregated 20-MHz LTE-like signals through a 7.5-km conventional PON link.

PWM analog fronthauling effectively avoids the bandwidth expansion related to digitization of fronthaul signals in state-of-the-art systems, but massive antenna arrays and larger signal bandwidth may call for additional capacity of the fronthaul link. Mode division multiplexing is considered in this Thesis to provide an additional multiplexing dimension and, to ensure the cost-effectiveness of the proposed solution, modal multiplexing is all-optical and passive, either based on photonic lanterns or on multiplane light converters, and paired with direct detection.
To cope with the arising intermodal interference, the degree of freedom offered by the mapping between radio resources and optical fronthauling resources is studied and it is shown that appropriate resource assignment can almost overcome the limitations introduced by intermodal interference.
In particular, the resources available on the two channels (wireless link antennas-users and wired fronthaul link) are defined in a space-frequency domain, the space being defined by the antenna array for the radio signal and by the propagation modes for the optical fronthaul.
System performance is evaluated considering beamforming at the centralized baseband unit to spatially separate uplink signals coming from users on the scene. The arising intermodal interference on the fronthaul link is shown to be successfully managed by an appropriate association between the space-frequency resource of the radio channel and those of the optical fronthaul.

Massive antenna arrays and larger signal bandwidth are key features of millimeter wave radio communication, and hybrid beamforming has been proved in literature to provide a valid mean to overcome the channel limitations (mainly the increased path loss) while keeping a reasonable hardware complexity and energy consumption. In this Thesis, the integration of hybrid beamforming and analog fronthauling is tackled. In particular, realistic analog signal processing techniques in the optical domain are considered in order to provide wideband analog processing via tunable delay lines. Tunability is shown to be essential in dealing with a fast-changing radio environment and with time division multiplexing (TDM) of users, but the speed of typical optical tunable elements in achieving a time-varying delay response is not enough. This causes a transient in system performance with degradation of the first transmitted symbol of the TDM frame. To counteract this effect, this Thesis proposes a technology-aware scheduling of time division multiplexed users, together with a digital precompensation of the transient in delay response.
A design algorithm for hybrid beamforming is proposed, based on the matching pursuit paradigm. It is numerically validated and tunability of typical optical components is addressed in the context of 5G NR specifications: on the one hand adaptive optical analog beamforming is shown to be mandatory to approach the performance of benchmark all-digital processing, and on the other hand the necessity to compensate for limited tunability speed of optical components is highlighted. Solutions based on digital precompensation and appropriate users scheduling are proposed and validated through numerical simulations.

Tuning the analog processing at the remote antennas requires a signaling channel parallel to the fronthauling. The proposed solution is based on joint A-RoF fronthaul transmission of the radio signals paired with a binary polarization shift keying (PolSK). This gives a low-cost, low-complexity parallel channel between the BBU and the RRH that can be used, e.g., to control the analog beamformer at the RRH and to feed back CSI from the radio channel.
Experimental validation provides a proof-of-concept, with parallel transmission of a LTE-like 20-MHz signal via intensity modulation and of a 1.5-Mbit/s stream via binary polarization shift keying. Results show a negligible degradation in the performance of the IM/DD link carrying the fronthaul signal and a high reliability of the digital signaling.

Last but not least, a precise pulse width modulation based on the concept of sigma-delta modulation is proposed and analyzed for its robustness against clock jitter noise. Results show a gain in signal to (jitter) noise ratio by at least 15 dB with respect to the open-loop pulse width modulation, and the convenience of low pass filtering for pulse width demodulation in the feedback path for achieving even better performance.

HIGH-SPEED, LOW-DISTORTION SOLUTIONS FOR TIME-CORRELATED SINGLE PHOTON COUNTING MEASUREMENTS

Alessandro Cominelli - Supervisor: Prof. Ivan Rech

Nowadays, Time-Correlated Single Photon Counting (TCSPC) represents a key measurement technique in many scientific and industrial applications demanding for the acquisition of extremely fast and faint luminous signals with picosecond resolution. In a typical TCSPC experiment, a sample is excited by means of a periodic laser source. Then, photons re-emitted by the sample are recorded to form a histogram, depending on their arrival times within the excitation period. In this way, after many periods the histogram represents a measurement of the average waveform of the luminous signal.
Unfortunately, TCSPC comes along with a major drawback, that is a relatively long acquisition time. In particular, two effects concur in limiting the maximum measurement speed of a TCSPC acquisition channel. First of all, a conventional TCSPC system can detect only one photon per excitation cycle. As a consequence, if more than one photon impinges on the detector during a period, the reconstructed waveform undergoes a distortion, which is known as classic pile-up. In order to avoid this issue, the intensity of the excitation source is typically adjusted to keep the average number of impinging photons in a period well below 1. It follows that a relatively high number of excitation cycles is required to accumulate a statistically relevant number of events in the histogram. The second limit to the measurement speed is related to a relatively long dead time of both detector and time-measurement electronics, which typically ranges in the order of 100 ns.
In the last decade, TCSPC acquisition systems have been subject to a fast trend towards the parallelization of many independent channels in order to speed up the measure. On one hand, some multichannel modules based on discrete components are already available in the market, featuring the best in-class performance in terms of resolution and linearity, but the high power dissipation and the volume occupied by a single channel have limited the degree of parallelism to only 4 or 8 channels so far. On the other hand, the exploitation of CMOS technology has permitted the integration of hundreds and even thousands of independent channels on the same chip, including detectors, represented by Single Photon Avalanche Diodes (SPADs), and the whole acquisition and conversion electronics.
Nevertheless, large arrays proposed so far with detectors and electronics integrated on the same chip suffer from a trade-off between number of channels and performance. In particular, the integration of both detectors and conversion electronics in the same pixel area has imposed tight constraints on power dissipation and area occupation of the electronics, limiting timing performance, both in terms of linearity and precision. Even worse, large multichannel systems are typically affected by a data-transfer bottleneck, which strongly limits the achievable measurement speed. In particular, the presence of a huge number of detectors can give rise to a considerably high data rate at the output of the system, which can easily reach 100 Gbit/s. Unfortunately, the real-time management of such a high rate demands for a huge bandwidth of the bus directed toward the external processor and for a considerable complexity of the system design. Instead, the maximum available transfer bandwidth is typically limited in the order of 10 Gbit/s. As a result, the efficient exploitation of a limited transfer bandwidth is, to date, one of the major challenges designers have to face to pursue the highest speed in TCSPC experiments.
Recently, different readout architectures have been proposed in literature to cope with a limited transfer bandwidth, trying to maximize its exploitation under typical operating conditions. Nevertheless, solutions proposed so far are affected by relatively low efficiency and the measurement speed still lies well below the limit imposed by the saturation of the transfer rate towards the elaboration unit.
The goal of this thesis work is to investigate novel approaches to speedup TCSPC measurements, avoiding at the same time any trade-off with performance.
First of all, I deeply investigated the problem of pile-up distortion, which currently represents the major limitation to measurement speed in a single TCSPC acquisition channel. In this context, I propose a novel solution to keep pile-up distortion below a negligible value, paving the way to a remarkable increase of the excitation power, well above the classic pile-up limit, thus leading, in turn, to a significant speedup of TCSPC experiments.
In particular, I theoretically demonstrated that negligible distortion (below 1%) is guaranteed if the dead time associated with the converter is kept below the dead time of the detector, and at the same time the detector dead time is matched to the duration of the excitation period. In this way, the speed of TCSPC experiments can be increased by a factor larger than 7.4, that is almost an order of magnitude, while providing negligible distortion regardless of the experimental conditions.
It is worth noting that the proposed technique allows a single acquisition channel to reach a remarkable measurement speed, which can be achieved, to date, only using eight independent TCSPC measurement channels operating in parallel. In this scenario, my solution requires a considerably lower complexity of the system design and, even better, it can be easily extended to a multichannel approach to further increase the measurement speed. Moreover, a practical use of my solution is already feasible exploiting recently-proposed electronics, that is time-measurement circuits with negligible dead time and a SPAD coupled to a fast active quenching circuit, featuring a short and finely tunable dead time.
In this work, I also present a novel readout architecture for large multichannel arrays, which has been conceived starting from a completely different perspective with respect to readout architectures proposed in the literature: a large detector array is shared with a limited set of high-performance time-measurement circuits, whose number is calculated starting from the maximum manageable data rate; then a smart routing logic has been designed to dynamically connect a large number of SPAD detectors to the external time-measurement electronics, in order to take full advantage of the available bus bandwidth. In addition, the proposed router-based architecture permits to exploit different technologies to design the various parts of the system, i.e. detectors, sensing electronics and time-measurement circuits, in order to optimize their performance.
The core of the router-based architecture is a selection logic, whose task is to select a subset of the detectors carrying a valid signal during each excitation cycle, to connect them to the external converters. It is evident that a certain elaboration time is required to carry out this operation, so a low-jitter delay line has been designed to be integrated along with each pixel of the array, in order to preserve the timing information related to a photon detection, until the logic elaboration has been carried out. The proposed delay line is able to provide a digitally-programmable delay up to 50 ns, while timing jitter is kept around 0.1% of the average delay, thus permitting excellent timing performance.
Then, I designed a novel routing algorithm exploiting digital gates distributed in a tree structure, aimed at the future realization of a 32x32 array. The proposed algorithm is able to dynamically connect the array to five shared conversion channels operating at 80 MHz, thus providing an overall throughput up to 10.4 Gbit/s, including 2 bytes for the timing information and 10 bits to address the selected pixels within the array. In addition, the designed logic has a double advantage: it permits to minimize at the same time the elaboration time and the number of interconnections crossing thesystem, which is a major issue in dense multichannel arrays.

ANALYSIS AND DESIGN OF ADVANCED ANTI-LOCK BRAKING SYSTEMS

Luca D'Avico - Supervisor: Prof. Sergio M. Savaresi

This thesis deals with the analysis and design of anti-lock braking systems. The research has been focused in particular on two vehicles such as bicycles and aircrafts that have been little explored so far from this point of view. In fact, such technology was born in the aeronautical field but never analysed from the analytical view point; furthermore, anti-skid for aircraft has always been based upon the very only wheel deceleration; in this work alternative solutions showing the features of slip-based approaches are applied to the aeronautical context; experimental results are reported for each one of the analysed solution. For what concerns bicycles, on the other hand, very little has been found on both scientific and industrial practice on this subject. The main goal of this work is to fill the gap between the wide popularity of bicycles and the lack of safety associated to this vehicle. The analytical description of the safety critical dynamics, design, implementation and experimental validation of the proposed solutions is reported in this thesis.

First of all, the analytical modelling of the most critical dynamics during braking manoeuvres are suitably modelled in order to be taken into account during the controller's design phase. The analysis of the wheel lockup has been thoroughly discussed in literature and it is reported for completeness. Then, the gear-walk in aircraft phenomenon is modelled and discussed in details; gear-walk can be described as an oscillatory motion of the landing gear itself in the longitudinal direction, taking place around a static vertical centre line. The nose-over(stoppie) motion of two-wheeled vehicles has been analysed; it consists of the rear wheel lift-off and constitutes one of the major safety concerns for bicycles.

The dedicated setup for this work is described in details focusing on the description of the dynamics of the internal control loops and the identification of the system dynamics. For the analysis and design of anti-skid for aircraft, a landing gear test rig has been implemented in a test facility simulating the landing manoeuvre. For bicycles, a mountain-bike has been equipped appropriately: an actuator has been placed in the front wheel braking system between the brake lever and the brake disk; the actuator position is controlled to obtain the desired front wheel deceleration.

Two different estimation algorithms for the detection of wheel lockup events have been proposed. The first one is dedicated to bicycles and it consists of a linear approach based upon data fusion of the wheel speed and the longitudinal acceleration. The second algorithm is dedicated to aircraft: a sliding mode observer (SMO) has been implemented taking as input the anti-skid current request and the measured wheel speed. Then, a detection algorithm for stoppie events for bicycles is proposed.

The design and experimental validation of anti-lock braking systems for both vehicles has been carried out. In particular, for bicycles three different deceleration controllers have been analysed: two switching control logics (Bang-Bang, Second Order Sliding Mode) and a modulating one (PI); the last one has proved to provide the best performances. For aircraft, a deceleration-based approach has been implemented to emulate the closed-loop behaviour of commercial anti-skids (5ph-ABS); furthermore, two different slip-based approaches has been proposed showing the improvements of performance at the cost of measurement/ estimation of the vehicle speed (SLIP, MSD). Experimental validation of these controllers is reported and analysed in details.

Considering both wheel lockup and stoppie event, suitable activation logics for the automatic selection of the reference wheel deceleration have been defined. In the stoppie control scheme, adaptivity has been introduced to take into account the shifting of the centre of gravity position once the rear wheel lift-off has occurred. Experimental results of these architectures are reported and discussed.

The tire wear is a problem of environmental and economic relevance in aircraft due to maintenance cost. For these reasons, the modelling of such phenomenon has been proposed and analysed from the point of view of the anti-skid intervention: slip -based approaches have been proved to provide same stopping distances drastically reducing the tire wear level with respect to a deceleration-based approach such as the one currently adopted on aircraft. Consistency with experimental results is shown as well.

MICROGRIDS ENERGY MANAGEMENT WITH A HIERARCHICAL DISTRIBUTED MODEL PREDICTIVE CONTROL APPROACH

Le Anh Dao - Supervisor: Prof. Luca Ferrarini

Starting from the first distribution system built in Manhattan and New Jersey in 1882 and the completion of the first long-distance transmission line which brought electricity from Niagara Falls to the city of Buffalo in 1896, the evolution of the power system has gone through a tremendous transformation in every factor from expansion, structure, efficiency to legislation and so on. In the last decade, with the ceaseless growth of distributed energy resources, especially renewable energy sources, the power system is facing a new scenario where many small and distributed generators connected to medium and low voltage grids, as opposed to a few large generators connected in high voltage. The main driver for this change comes from firstly to assure the security of energy supply. In addition, the second driver comes from the fact that fossil fuel price is volatile and increasing; also there are environmental concerns regarding their widespread usage, therefore the use of renewable energy resources which are often placed in the vicinity of end users is promoted. Indeed, the global renewable power capacity more than doubled during the last decade from approximately 1000 gigawatts in 2007 to 2195 gigawatts in 2017; in 2017 alone, 70% of the net addition of global power capacity was from renewable energy which marked the incredibly increasing trend of the renewable energy. This trend is expected to be maintained as many countries have put ambitious national targets for percentages of renewable energy and power for the near future. Notwithstanding the promising advantages it may bring, the penetration of Renewable Energy Sources (RESs) also puts many challenges to the power system due to its variability and thus the reliability of its supply together with the difficulty in predicting the output of renewable energy sources. These characteristics of renewable energy can cause significant problems of electricity supply, demand balance and power quality to the distribution network and the grid in general if large capacity renewable energy sources are connected. In this case, the power system becomes more intermittent and less predictable than in today's situation, while more difficult to control as the RES offers a much lower level in control with respect to traditional large-scale generators based on fossil sources.

To deal with foreseen situations of variation and intermittence in the power system, a clear direction is to improve the quality of RES production prediction. Another solution is to reconsider the paradigm of traditional "demand-driven supply" by changing to "supply-driven demand" or intermediate levels between them so that the demand side could play a more crucial role in supporting the power system between. Following this direction, Demand Response (DR) is generated from the late 1970s and introduced as a concept by authors in in 1980. In the future power system, DR is expected to play a key role in balancing power in the scenario that a major part of the electricity generation comes from intermittent renewable energy sources. Finally, another attractive solution for the intermittence of the power system is exploiting Energy Storage System (ESSs). Considering all these possible solutions together, the microgrid concept is introduced as the physical support for a system where ESSs, RESs and DR are all involved. The microgrid considered in this dissertation is sketched in Figure 1.

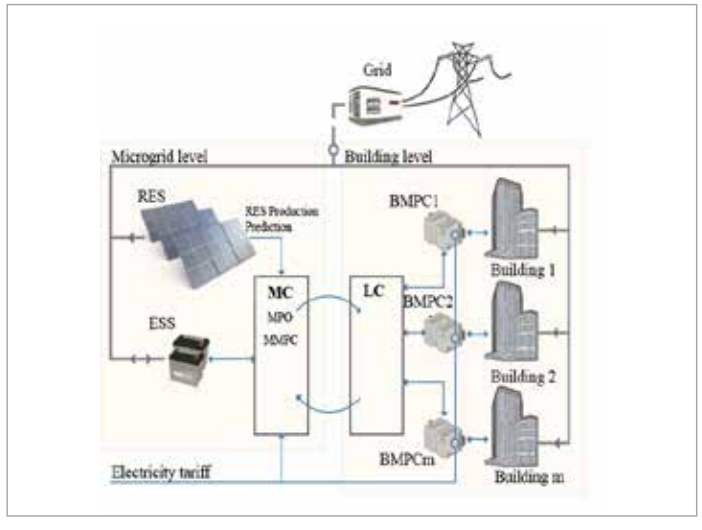

Fig. 1 - Microgrid setting and designed control system (Power flow: gray, bold line; Information, control signal: blue line)

On the described configuration of the microgrid, the dissertation aims at studying the impact of advanced control techniques, RES generated power prediction, and some other aspects on the optimal operation of a microgrid taking into account comfort for end-users and the economic terms for both end-users and the microgrid. Special attention on an innovative hierarchical distributed control architecture is designed for the integration of multiple end-users, photovoltaic RES and ESS in such a way that guarantees (or better, maximizes) privacy and minimizes the computational and communication burdens for the involved entities in the microgrid. Following the above discussion, our development of microgrid energy management system relies on two main pillars: (i) designing a comprehensive and innovative framework for microgrid energy management system and (ii) improving the predictability of RESs in microgrids to improve energy management.

Firstly, the overall goal and also the most important objective of the dissertation is to design a comprehensive and innovative framework for the microgrid energy management system which aims to maximize the overall benefit, while still accounting for possible requests to change the load profile coming from the grid and leaving every single building or user to balance between servicing those requests and satisfying his own comfort levels. The user involvement in the decision-making process is granted by a management and control solution exploiting an innovative distributed model predictive control approach with coordination.

In addition, to integrate the distributed Model Predictive Control (MPC) user-side with the microgrid control, and the microgrid control with microgrid scheduling which is also implemented under Model Predictive Control (MPC) framework, a hierarchical structure is proposed. Still, this structure's objective is also to assure the privacy for users and microgrid level by limiting exchanged information between them.

Secondly, as the development of RES production prediction is a crucial part of the predictive control framework which is employed in the microgrid control system, another objective of the dissertation is going to discuss a problem of performing Generated Power (GP) prediction. Specifically, the objective is to design a reliable prediction framework to improve the quality of available predictions. One promising direction is to employ ensemble method by combining different predictors together in such a way that could get the best characteristic in each of them so that the combination provides better prediction than any individual could.

In the end, the proposed overall approach is implemented and tested completely/partially in several experiments in the laboratory facility for distributed energy systems. Both results in simulation and experimental environments are expected to show the accuracy and the potential of the works, also in the perspective of implementation. On the other hand, the work on RES production is implemented on an embedded mini PC module Raspberry Pi 3 model B and the validation of the approach is performed by using a pilot PV plants and meteorological stations situated in North of Italy.

VIDEO RECOMMENDATION BY EXPLOITING THE MULTIMEDIA CONTENT

Yashar Deldjoo - Supervisor: Prof. Paolo Cremonesi

Video recordings are complex media types. For example, when we watch a movie, we can effortlessly register a lot of details conveyed to us (by the author) through different multimedia channels, in particular, the audio and visual channels. To date, most movie recommendation services base their recommendations on collaborative filtering (CF) that leverage user's consumption patterns and/or content-based filtering (CBF) models that use metadata (e.g. genre or cast). In most video-on-demand and streaming services, however, new movies and TV series are continuously added. CF models are unable to make predictions in such scenarios, since the newly added videos lack interactions -- a problem technically known as new item cold start (CS). Currently, the most common approach to this problem is to switch to a purely CBF method, usually by exploiting textual metadata. This approach is known to have lower accuracy than CF because it ignores useful collaborative information and relies on human-generated textual metadata, which are expensive to collect and often prone to errors. User-generated content, such as tags, can also be rare or absent in CS situations. Multimedia features, can provide the means to identify videos that 'look similar' or 'sound similar'. These discerning characteristics of heterogeneous feature sets meet users' differing information needs.

In the context of this PhD thesis, methods for automatically extracting video-related information from the multimedia content (i.e., audio and visual channels) have been elaborated, implemented, and analyzed. Novel techniques have been developed as well as existing ones refined in order to extract useful information from the video content and incorporate them in recommendation systems. Different video recommendation tasks are solved using the extracted multimedia information under recommendation models based on content-based filtering (CBF) models and the ones based on combination of CBF and collaborative filtering (CF). Finally, machine learning approaches have been proposed for cold video recommendation by training a CF model on warm items (items with interactions) and leveraging the learned model on the movie multimedia content features to recommend cold items (items without interactions)

As a branch of recommender systems, this thesis investigates a particular area in the design space of recommender system algorithm in which the generic recommender algorithm needs to be optimized in order to use a wealth of information encoded in the actual image and audio signals. The results and main findings of these assessments are reported via several offline studies or user-studies involving real users testing a prototype of developed movie recommender systems powered by multimedia content. The results show different scenarios in which movie recommender systems can benefit from multimedia content, outperforming the alternatives, most notably in new-item settings.

ON HOW TO EFFECTIVELY TARGET FPGAS FROM DOMAIN SPECIFIC TOOLS

Emanuele Del Sozzo - Supervisor: Prof. Marco D. Santambrogio

Heterogeneous System Architectures (HSAs) represent a promising solution to face the limitations of modern homogenous architectures, in terms of both performance and power efficiency. Thanks to the combination of hardware accelerators like GPUs, FPGAs, and dedicated ASICs, such systems are able to efficiently run performance demanding applications belonging to different domain on the most suitable device. In order to fully take advantage of HSAs, in the last years we have witnessed the rise of many and different solutions designed to simplify the development of applications for multiple target architectures. In this context, Domain Specific Languages (DSLs) represent one of the most interesting solutions. Indeed, current DSLs allow the user to quickly and easily develop portable designs for multiple architectures. Thanks to the restriction of the domain, DSL compilers are able to rapidly explore the design space and deeply optimize the resulting implementations. As a result, DSL applications often outperform hand-tuned libraries. On the other hand, Machine Learning (ML) frameworks represent another fascinating solution. Even though it has been around for decades, ML has been one of the major topics in research and engineering field over the last years. The reason for that is, on one hand, the availability of a huge amount of data to train ML algorithms, and, on the other, the possibility to efficient execute ML algorithms, like Convolutional Neural Networks (CNNs), on hardware. As a consequence, ML tools quickly evolved. Nowadays, frameworks like TensorFlow, Caffe, and Torch offer efficient solutions to easily both implement ML algorithms, without a significant expertise of the field, and target hardware accelerators.

Acceleration of Domain Specific Computations on FPGA

Although DSLs and ML frameworks are highly effective in assisting users towards the generation of efficient designs for CPUs and GPUs, they still lack a concrete support for FPGAs. Historically, hardware design for FPGAs has always been more complex with respect to the design for CPUs and GPUs. This limited the adoption of FPGAs in datacenters and HPC systems, in spite of the great design opportunities FPGAs can provide in such contexts (like arbitrary data precision and the possibility to create a custom architecture, basically a Domain Specific Architecture, tailored to the target application scenario). Similarly to what happened for CPUs and GPUs, over the last years FPGA toolchains have significantly improved and increased their features. For instance, High-Level Synthesis (HLS) tools facilitate the design on FPGA; indeed, they permit to hardware accelerate algorithms using languages like C/C++ and OpenCL, instead of languages like Verilog and VHDL. However, the whole FPGA design process remains complex and the integration with high-productivity tools and languages is still limited. In particular, on one hand, even though there exist some DSLs able to target FPGA, a common solution capable of supporting multiple DSLs, even the ones that do not have an FPGA backend, is still lacking. On the other, an official and fully-integrated FPGA support within industrial ML frameworks, like TensorFlow, is not available yet.

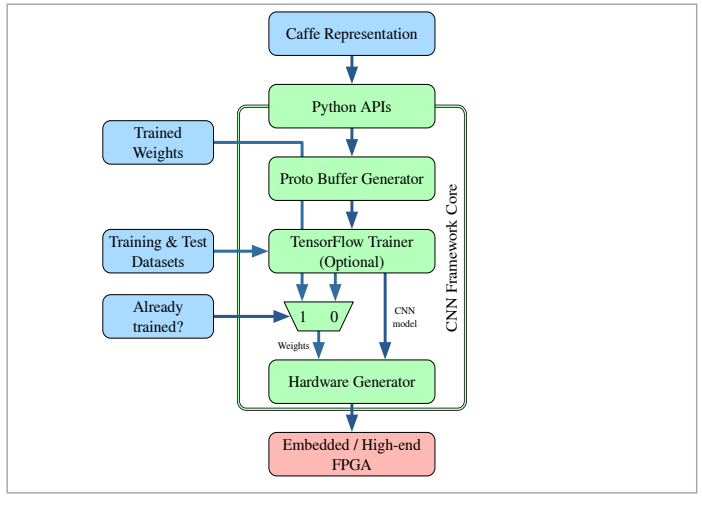

Fig. 1 - CNN framework overview.

Original Contributions

Given these motivations, this thesis focuses on the development of frameworks able to efficiently and easily target FPGAs from domain specific tools. In other words, the goal of this thesis is to demonstrate the efficiency of FPGAs when applied in a well-defined context, where the user can transparently take advantage of such devices and the frameworks manage all the complexity (i.e. the generation of an efficient hardware design). In particular, this work describes tools oriented to the hardware acceleration on FPGA of CNNs and DSLs. The purpose of such tools is to simplify the design of FPGA accelerators providing users with high-level abstractions to define the computation (and not its implementation). On the other hand, the proposed tools are able to build efficient hardware designs thanks to the restricted domain.

A Framework to Deploy CNNs on FPGA

The first tool is a framework for the fast-prototyping and deployment of CNN accelerators on FPGA (Figure 1). The goal of the framework is to bridge the gap between high-productivity ML frameworks, like TensorFlow and Caffe, and FPGA design process. The main features of the framework are:

- A novel framework written in Python, providing a set of modules that implement the toolchain for the design and the implementation of CNNs on FPGAs;
- A flexible internal representation based on Google Protocol Buffers that is compliant with a subset of the layer definitions of the Caffe deep learning framework, giving the possibility to provide existing models as input;
- The integration with TensorFlow for CNN training, providing the training set and the test set directly to the framework;
- A hardware library with customizable modules implementing the different type of layers of CNNs.

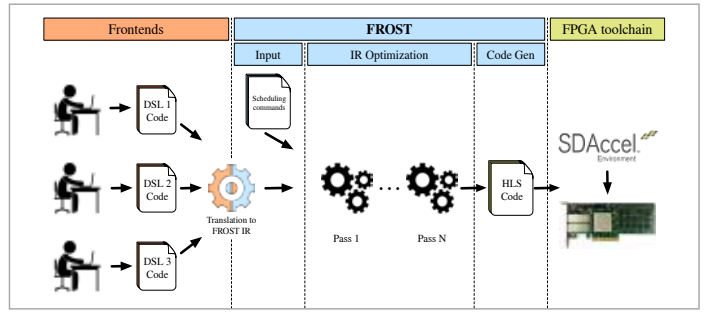

Fig. 2 - FROST workflow.

A Common Backend to Accelerate DSLs on FPGA

The second tool is FROST, a unified backend to efficiently hardware-accelerate DSLs on FPGAs (Figure 2). Starting from an algorithm described in one of the supported DSLs, FROST translates it into its Intermediate Representation (IR), performs a series of FPGA-oriented optimizations steps, and, finally, generates an optimized design for HLS tools. Here the main features of FROST:

- A common backend exposing an IR that DSLs can target in order to accelerate their computations on FPGA;
- Support for Halide, a state-of-the-art DSL for image processing, and Tiramisu, a code optimization framework for HPC systems;
- A high-level scheduling co-language the user can exploit to guide the optimizations to apply, specify the architecture to implement, and combine with the optimizations offered by the frontends;
- Generation of efficient hardware designs suitable for HLS tools.

ON THE EXPLOITATION OF UNCERTAINTY TO IMPROVE BELLMAN UPDATES AND EXPLORATION IN REINFORCEMENT LEARNING

Carlo D'Eramo - Supervisor: Prof. Marcello Restelli

Introduction

The recent exponential growth of Reinforcement Learning (RL) research has been made possible by the comparable significant improvement in computational power. Indeed, the advent of powerful and relatively affordable hardware, in particular Graphics Processing Units (GPUs), allowed researchers to extend the study of RL methodologies to highly-dimensional problems that were unpractical before, opening the line of research that is now commonly known under the name of Deep Reinforcement Learning (DRL). However, the groundbreaking results that DRL is achieving are obtained at the cost of a huge amount of samples needed for learning, along with very large learning times usually in the order of days. One of the reasons why this is happening, besides the outstanding significance of the results that fundamentally poses the problems of the efficiency of these methodologies in the background, relies on the fact that often experiments are run in simulations in which the sample efficiency problem is not such an issue as in real applications. Nevertheless, the issue of sample efficiency always constituted a matter of concern also in classic RL research where several works have been proposed to address the problem. It is historically well-known that this issue arises from the need of the agent to explore the environment it is moving in to improve its knowledge about it, and to exploit simultaneously the actions it considers to be the best to maximize its return, creating a trade-off known in RL as exploration-exploitation dilemma. The addressing of this trade-off is central and constitutes a measure of effectiveness of any algorithm available in literature.

The purpose of this thesis is to study the previously described problems proposing novel methodologies that explicitly consider the concept of uncertainty to speed up learning and improve its stability. Indeed, since a relevant goal of an RL agent is to reduce uncertainty about the environment in which it is moving, taking uncertainty explicitly into account can be intuitively an effective way of acting. This solution is not new in RL research, but there is still a lot of work that can be done in this direction and this thesis takes inspiration from the available literature on the subject extending it with novel significant improvements on the state of the art. In particular, the works included in this thesis can be grouped into two parts: one where uncertainty is used to improve the behavior of the Bellman equation and the other where it is used to improve exploration. The works belonging to the former group aim to address some of the problems of action-value estimation in the context of value-based RL, in particular in the estimate of the maximum operator involved in the famous Q-Learning algorithm. On the other hand, the works belonging to the latter group study different methodologies to improve exploration by studying the use of Thompson Sampling in RL or by introducing a variant of the Bellman equation that incorporates an optimistic estimate of the action-value function to improve exploration according to the principle of Optimism in the Face of Uncertainty (OFU).

Improving Bellman updates

The estimation of the Maximum Expected Value (MEV) of a set of random variables is required in several applications. For instance, in RL the optimal policy can be found by taking, in each state, the action that attains the maximum expected cumulative reward. The optimal value of an action in a state, in turn, depends on the MEV of the actions available in the reached states. Since the errors propagate through all state-action pairs, a bad MEV estimator negatively affects learning speed.

The most used approach to this estimation problem is the Maximum Estimator (ME) which simply takes the maximum estimated utility. This estimate is positively biased and, if used in iterative algorithms, can increase the approximation error step-by-step. More effective estimators have been proposed in recent years. The Double Estimator (DE) approximates the maximum by splitting the sample set into two disjoint sample sets. This approach has been proven to have a negative bias which, in some applications, allows to overcome the problem of ME.

In this thesis, we analyzed this problem and proposed the Weighted Estimator (WE) which approximates the MEV of a set of random variables by a sum of different values weighted by their probability of being the maximum. WE can have both negative and positive bias, but its bias always stays in the range between the ME and DE biases. Moreover, since all the previously mentioned approaches are limited to a finite set of random variables, in a subsequent work we extended the study to problems with an infinite set of random variables and proposed an extension of WE to address them.

Improving exploration

A desirable property of sampling strategies is to have theoretical evidence of their efficiency in terms of the number of collected samples. Among others, the principle of OFU has been well studied in the literature, where large evidence of its efficiency is given. OFU states that actions with statistically uncertain values must be favored (e.g., through an exploration bonus) in the action-selection process compared to the more certain ones in order to improve knowledge of the environment. This optimistic sampling strategy accelerates the learning of action values and exploits them efficiently as the uncertainty about the environment is reduced, and thus the effect of optimism lessens. This sampling strategy has been firstly used in the context of Multi-Armed Bandits (MABs) and is known as Thompson Sampling (TS). The idea of TS is to randomly choose an arm (i.e., an action) to be drawn according to its probability of being the optimal one. The use of TS in RL seems quite straightforward considering the actions of the Markov Decision Process (MDP) as the arms of the MABs; nevertheless, the presence of multiple states and dynamic transitions among them makes the estimation of uncertainty a challenging problem in this setting.

In this thesis we present a work that takes inspiration from the Bayesian RL framework in which a probability distribution is used to model the uncertainty over action value functions in each state. Our work is motivated by the fact that in the Bayesian approach the variance of the distribution can be used as a measure of the uncertainty of the action that allows to pursue OFU using TS. In this direction, we present several methodologies to efficiently compute the uncertainty of action values in online RL problems and to use the TS strategy showing its benefits w.r.t. other sampling strategies.

We also addressed the problem of optimism in exploration from another point of view, working on the proposal of the Optimistic Bellman Operator (OBE), a novel variant of Bellman Operator (BE) that results into an optimistic action-value estimate from an ensemble of action-value functions where the optimistic estimate is obtained from a maximum-entropy principle. The resulting action-value estimate favors the visit of unknown states over known ones; moreover, for the exploration bonus that OBE implicitly defines, we can prove that the bonus decreases consistently with the number of state visits resulting in a good balance of exploration and exploitation.

All the works presented in this thesis are described, theoretically studied, and eventually, empirically evaluated on several RL problems. The obtained results highlight the benefits that the explicit exploitation of uncertainty in RL algorithms can provide; indeed, we show how in a large set of problems that have been chosen in order to highlight particular aspects we were interested in, e.g. exploration capabilities, our methods prove to be more stable and faster to learn than others available in the literature.

DESIGN OF COLLABORATIVE ECO-DRIVE CONTROL ALGORITHMS FOR TRAIN NETWORKS

Hafsa Farooqi - Supervisor: Prof. Patrizio Colaneri

In today's age, green transportation remains one of the most important topics of research. This is due to the continuously increasing risk of global warming, of which transportation remains one of the major contributors. The main goal of green transportation is to promote vehicle technologies and driving styles which are energy efficient and environment friendly. Keeping this in mind, the European Union (EU) has set up certain restrictions/targets. In this thesis, the main focus is on railways, which is considered to be one of the most efficient means of transportation. For the railway sector, these targets are set together by the International Union of Railways (UIC) and Community of European Railway and Infrastructure Companies (CER). The short term target is to decrease the CO_2 emissions by 30 % over the period from 1990 to 2020, while the long term goal is to firstly further decrease these emissions by 50 % towards the end of 2030 and secondly to decrease the overall energy consumption of railways by 30 % towards the end of 2030 as compared to 1990. For this purpose, different methods have been proposed and implemented, which differ on the basis of certain factors. Some of these methods are operator based, which at times requires upgrade in the vehicle technology, while others are control based. For example, in order to reduce energy, an operator can deploy rolling stock that is more energy efficient by using more energy efficient engines or streamlining. On the other hand, the control based methods are mostly focused on development of techniques aimed at reducing energy consumption and emissions which are affected by the behavior of the driver, without necessarily upgrading the vehicle technology. In the literature, they are commonly referred to as Energy Efficient Train Control (EETC) or eco-driving strategies. Furthermore, in case of electrical trains, which can regenerate energy during braking, the control based techniques need to be further enhanced to be able to use this regenerated energy effectively.

Based on the above factors, this thesis has focused on two main research directions. The first research direction is associated with a single train control problem, where the control problem is to find the best driving strategy for the train to go from one stop to another, given an optimal timetable. EETC strategies can be either fully automated (ATO) or serve as an advisory system to the driver (DAS) for the purpose of assisting drivers in following an energy efficient driving style. For this purpose, three control strategies using Model Predictive Control (MPC) have been presented. In the first two strategies, shrinking horizon techniques have been combined with input parametrization approaches to reduce the computational burden of the control problem and to realize the nonlinear integer programming control problem which arises in the DAS scenario, while the third strategy is based on switching MPC with receding horizon. The second research direction falls under the paradigm of collaborative eco-drive control strategies, involving multiple trains belonging to a substation network. The main aim is to use the energy regenerated by the braking trains through collaboration among the trains connected and active in the network. In this case, three strategies to decide the collaborative law have been presented along with the extensions from the single train control strategies presented in the first part of the thesis. For the design of collaborative laws, techniques such as manual supervision, substation modeling and dissension based adaptive laws with concept similar to Markov chains have been used. All the strategies have been tested on the official simulation tool CITHEL of our industrial partner Alstom, and the obtained results in comparison with the existing techniques have proven to be more energy efficient.

DYNAMIC APPLICATION AUTOTUNING FOR SELF-AWARE APPROXIMATE COMPUTING

Davide Gadioli - Supervisor: Prof. Gianluca Palermo

With the end of Dennard scaling, the performance of modern systems is limited by the power consumed. This shifted the focus of system optimization toward energy efficiency in a wide range of scenarios, not only related to embedded systems but also related to high-performance computing (HPC).
Among all the possible directions that promise to improve the computation efficiency of a system, this thesis focuses on two approaches at the software layer. On one hand, when application developers write the source code, the best practice is to expose implementation parameters that alter the extra-functional properties of the application, such as execution time or power consumption. In literature, these parameters are also named \software-knobs, since a change on their value leads to a change in the extra-functional properties as well. On the other hand, several approaches aim at finding *good enough* results for the end-user, thus saving the unnecessary computation effort to further improve efficiency. In literature, this approach is named *approximate computing*. A large fraction of applications implicitly exposes software-knobs at algorithm-level to find accuracy-throughput tradeoffs. This happens especially in multimedia and whenever it is possible to use approximation techniques such as loop perforation or task skipping. Since approximate computing is able to significantly increase the application throughput by decreasing the result accuracy, several works in literature investigate the possibility to use also approximate hardware accelerators.
Among the implications of this trend, the application requirements are increasing in complexity. Due to the tradeoffs created by using software-knobs and approximate computing, the end-user might have complex requirements which involve extra-functional properties (EFPs) in conflict with each other, such as power consumption, throughput, and accuracy. In this context, the autonomic computing field investigates how to enhance the target system with a set of *self-** properties, such as *self-healing*, *self-optimization* or *self-protection*. This thesis focuses on the *self-optimization* property, where the target system shall automatically identify and seize optimization opportunities according to the system evolution.
Given the wide difference in performance between software-knobs configurations, a significant amount of research is spent on finding the ones that lead to optimal tradeoffs between EFPs of interest for end-user. However, finding a one-fits-all software-knobs configuration is complex if we consider the system evolution. The application requirements may change according to external events. For example, end-user might have different requirements according to whether the target platform is relying on batteries or not. Moreover, there might be changes in the underlying architecture. For example, a power capper might lower the core frequencies due to thermal reasons, or the available resources might change due to workload fluctuations. Furthermore, the EFPs might be heavily input dependent; therefore, a one-fits-all software-knobs configuration might lead to sub-optimal performance.

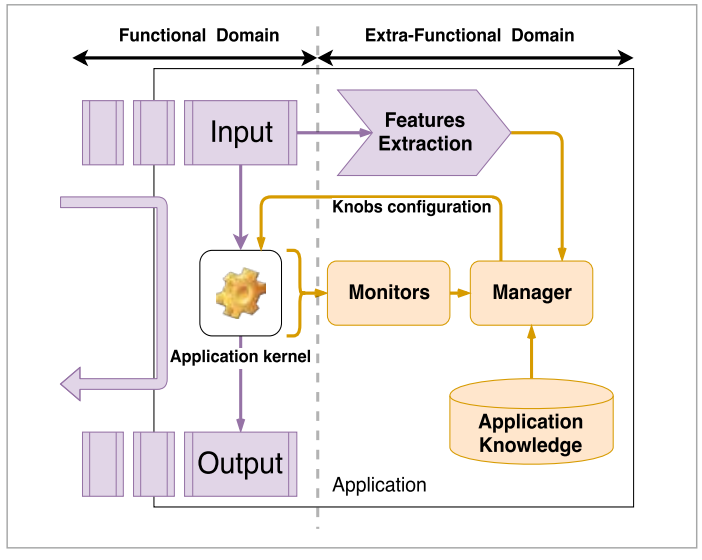

Fig. 1 - Overview of mARGOt. Purple elements represent application code, while orange elements represent mARGOt high-level components.

For these reasons, it is required an adaptation layer that tunes the software-knobs configuration at runtime for providing *self-optimization* capability. This is a known problem investigated in the literature using different approaches. However, how to provide to a target application the optimal software-knob configuration, according to end-user requirements and system evolution, is still an open question. The work carried out in this thesis aims at advancing the state-of-the-art toward this direction.
The main outcome of this thesis is a methodology to enhance a target application with an adaptation layer that exposes mechanisms to adapt in a reactive and proactive fashion. The methodology implementation, named mARGOt, is a C++ library that is linked to the target application and works at the function level. Fig. 1 depicts the overall architecture. mARGOt employs separation of concerns between functional and extra-functional requirement. End-user might define or change requirements at runtime, according to application phases. Moreover, by using feedback information from runtime monitors, it is possible to react to changes in the execution environment, providing to the application the most suitable software-knobs configuration. Furthermore, it leverages input features to identify and seize optimization opportunities according to the current input. The proposed adaptation layer has been evaluated in different scenarios, ranging from embedded to High-Performance Computing. Moreover, we deployed the proposed methodology in two real-world case studies. In the context of smart cities, we focused on a time-dependent probabilistic routing algorithm, by analyzing the relationship between end-user requirements, application software-knob and features of the input. Experimental results show how it is possible to drastically improve computation efficiency in the target use case. In the context of a drug discovery process, we focused on a geometrical docking miniapp, by analyzing the effect of approximation techniques for the extra-functional properties of interest. Experimental results show how it is possible to increase the application throughput by one order of magnitude, with an accuracy degradation less than 30%. By using mARGOt, end-user is able to harness this tradeoff to satisfy a time-to-solution constraint.
Additional contributions of this thesis are the methodology evaluation in different contexts. In particular, the orthogonality between resource managers and application autotuning has been evaluated, by applying different adaptation schemes in a dynamic workload with co-running applications. Our tests show that the average performance of using mARGOt as a lightweight resource manager is very close to the performance achieved with a combined approach based on a centralized resource manager. This approach is more portable and less intrusive from an application design point of view. However, it does not provide any guarantee of *fairness* nor *optimality* in resource allocation. Moreover, an approach has been proposed to combine the adaptation mechanisms of mARGOt, with source-to-source transformations of the LARA aspect-oriented language, and with insight provided by the COBAYN compiler autotuner, to provide to application developers a seamless online compiler and system runtime autotuning framework. The proposed approach is able to provide self-optimization capabilities to the target application, in terms of compiler options and OpenMP parameters, in a transparent way with respect to application developers.

INNOVATIVE APPROACHES TO THE LATERAL CONTROL PROBLEM IN CARS

Olga Galluppi - Supervisor: Prof. Sergio M. Savaresi

Cars represent the primary means of passenger transport in the world and road accidents are one of the mainspring of premature human death. Although, a decrease in the number of road deaths is observed throughout the last years. Technology advances in vehicle dynamics control are considered one of the reasons for this trend: the automotive industry is transforming, by becoming software intensive rather than mechanically intensive. Road vehicles are nowadays very complex systems, composed of several subsystem which are often interacting among each other. Although each module is responsible for a specific function, all of the subsystems influence the vehicle dynamic behaviour and, perhaps more importantly, the quality of the driving experience which is perceived by the driver and passengers, in terms of both performance and safety. The research studies of this Thesis describe innovative approaches to the lateral control problem in cars. The derived argumentations are human-oriented: they assort and embrace relevant perspectives on the lateral control question entailing sensibility over passengers safety, over the attainment of high-performance dynamical car behaviour, but also over the ecological burden. Throughout the discussion, multiple aspects which comprehend different layers of the problem are investigated and integrated. Placing in nowadays original core research on the subject, some of the studies are re-intepreted in an autonomous driving framework, though most of them mainly being conceived for the operation in all driving settings. Contemplated theorethical and applicative advances are endorsed by experimental and simulation studies. Concerned vehicle investigations include hybrid vehicles eco-routing, electrification of driving users, data-driven MIMO nonlinear control steering input solution, sideslip angle and longitudinal speed estimation, semi-active suspension control.

LOW-NOISE, LOW-POWER FRONT-END ASICS FOR HIGH-RESOLUTION X AND GAMMA RAY SPECTROSCOPY FOR RADIATION SEMICONDUCTOR DETECTORS

Massimo Gandola - Supervisor: Prof. Giuseppe Bertuccio

The aim of this thesis work is the study, design and development of CMOS ASICs as Front-End Electronics (FEE) for X and Gamma ray detectors for scientific applications.
During my PhD activities I worked on two different projects: *PixDD* and *HERMES*, sponsored by ASI (Agenzia Spaziale Italiana), for studying the Universe and deep space objects. All the projects foresee the developing of different ASICs starting from a common FEE architecture successfully implemented some years ago in *VEGA* ASIC for Silicon Drift Detector (SDD). Low-noise and low-power consumption are requirements for both the projects.
The PixDD project is developed within a collaboration among Politecnico di Milano, University of Pavia, National Institute of Astrophysics (INAF), National Institute of Nuclear Physics (INFN), Fondazione Bruno Kessler (FBK) and Karlsruhe Institute of Technology (KIT).
The goal of the *PixDD* project is to implement a system able to detect a low energy X-ray photon flux, coming from the outer space, in the range of 0.5keV up to 30 keV using a Pixelated Silicon Drift Detector for spectral-timing studies. The system is suitable for different space missions as, for example, the XTP missions (X-Ray Timing and Polarization) by Chinese Space Agency and LFA (Low-energy Focusing Array) by NASA. The detector realized by Fondazione Bruno Kessler (FBK), is shown in Figure 1, and it is a 4x4 matrix prototype where each 500 µm x 500 µm pixel is a drift detector. Low leakage current and small detector capacitance permits to reach a very high energy resolution in spectroscopy. The characterization of a pixel has been carried out with an ultra-low noise charge sensitive preamplifier called SIRIO obtaining a resolution of 127 eV at 5.9 keV of ^{55}Fe at 0 °C. The pixel leakage currents have been measured from 100 fA to 1 pA in the 0 °C and +20 °C temperature range. The final version of the detector will be a 16 x 8 matrix of 300 µm x 300 µm pixels. A multichannel ASIC, called *RIGEL,* has been developed for *PixDD* in 0.35 µm CMOS technology and it will be coupled through the bump-bonding techniques to the final version of detector.
In particular, we have developed the Read-out Pixel Cell (RPC) with the same dimensions of the pixel of the detector and containing all the analog and digital circuits for signal processing. The *RIGEL* ASIC including 16 x 8 RPCs, Wilkinson ADCs, Trigger management circuits, bias references and configuration registers has been assembled in collaboration with the University of Pavia.
A single RPC has been coupled to a pixel and studied in order to characterize the single-channel itself: 167 eV FWHM has been measured on the 5.9 keV line of ^{55}Fe at 1.45 µs of peaking time and 0 °C. A linearity error of ±2.7 % has been measured with a total power consumption less than 550 µW/Channel. A first prototype of the full system (*PixDD* detector and *RIGEL*) has been assembled at KIT (Figure 2) and it is currently under test.
The *HERMES* project is developed within a collaboration among National Institute of Astrophysics (INAF), Politecnico di Milano, FBK, Universities of Pavia, Cagliari, Udine, Ferrara, Napoli Federico II, Palermo, Tubingen, Nova-Gorica, Eotvos Budapest. The goal of *HERMES* is to implement a system able to detect and localise the Gamma Ray Bursts (GRB) and other high energy events coming from the deep space. One of the key feature of the *HERMES* project is the using of a constellation of nano-satellites (more than one hundred), called CubeSat 3U, instead of a dedicated satellite permitting to decrease the development period in a few years and the costs.
Since the photon energy range is from few keV up to almost 2 MeV, the "double detection" mechanism with a GAGG (Gadolinium Aluminium Gallium Garnet) scintillator coupled to two Silicon Drift Detectors has been used. The high energy photons (Gamma-ray) are acquired by the GAGG and converts in light and both the SDDs will be light up. On the other hand, X-ray photons interacts with only one of the two SDDs so it is possible to discriminate between X and Gamma events.
The detection unit payload is formed by 12 modules, each module is composed of 5 GAGG detectors and 10 SDDs. In total will be 60 GAGG and 120 SDDs.
The FEE developed for the *HERMES* project is called *LYRA* ASIC and it is organized in two ASICs, implemented in CMOS 0.35 µm Technology, called: FE-*LYRA* and BE-*LYRA* (Figure 3). FE-*LYRA* contains the preamplifier and the first shaping stage, BE-*LYRA* contains all the rest of the signal processing chain. The 120 FE-*LYRA* will be very close to the anode of detectors and the output will be elaborated by one of the four BE-*LYRA*. Each BE-*LYRA* has 32 channels and so each of them can manage 3 modules. The transmission between FE-*LYRA* and BE-*LYRA* is in current-mode to avoid spurious injection between all the output of the FE-*LYRA* that are close to each other.
The design specifications of the *LYRA* ASIC, confirmed by simulations, are:

- Energy range from 0.5 keV up to 120 keV
- Two selectable peaking time: 1.6 µs and 2.3 µs
- FWHM ≤ 180eV @ 6keV and T = -30 °C with detector leakage current less than 3 pA
- Maximum detector leakage current: 1 nA
- Linearity error within ± 1%
- Power Consumption ≤ 520 µW/ channel

LYRA ASIC has been manufactured and the test will be carried out in the middle of April 2019.

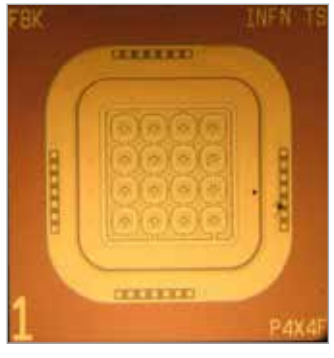

Fig. 1 - PixDD 4x4 prototype

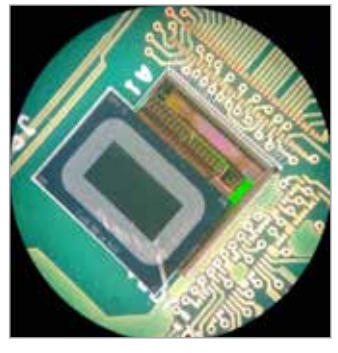

Fig. 2 - 16x8 PixDD detector coupled with the RIGEL ASIC

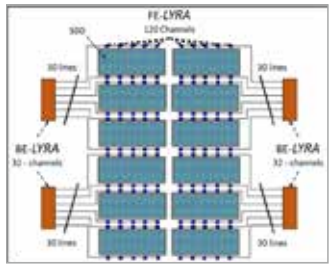

Fig. 3 - Schematic of the detection unit in the payload

PERFORMANCE MODELS, DESIGN AND RUN TIME MANAGEMENT OF BIG DATA APPLICATIONS

Eugenio Gianniti - Supervisor: Prof. Danilo Ardagna
Co-Supervisors: Dott. Michele Ciavotta, Dott. Marco Lattuada

Nowadays the big data paradigm is consolidating its central position in the industry, as well as in society at large. Lots of applications, across disparate domains, operate on huge amounts of data and offer great advantages both for business and research. As data intensive applications (DIAs) gain more and more importance over time, it is fundamental for developers and maintainers to have the support of tools that enhance their efforts since early design stages and until run time. The present dissertation takes this perspective and addresses some pivotal issues with a quantitative approach, particularly in terms of deadline guarantees to ensure quality of service (QoS).
IDC reports that big data used to concern highly experimental projects, yet its market is growing from $ 130 billion in 2016 to $ 203 billion in 2020, with a compound annual growth rate of 11.9 %. Big data applications offer many business opportunities that stretch across industries, especially to enhance performance, as in the case of recommendation systems. Furthermore, DIAs can also help governments in obtaining accurate predictions, for instance quality weather forecasts to prevent natural disasters and ease the development of appropriate policies to improve the population's life quality. In addition, big data systems are increasingly exerting a central force on society, thus requiring the development of intelligent systems providing QoS guarantees to their users.
Technically interesting scenarios, such as cloud deployments supporting a mix of heterogeneous applications, pose a series of challenges when it comes to predicting performance and exploiting this information for optimal design and management. Performance models, with their potential for what if analyses and informed design choices about DIAs, can be a major tool for both users and providers, yet they bring about a trade-off between accuracy and efficiency that may be tough to generally address. The picture is further complicated by the adoption of the cloud technology, which means that assessing operating costs in advance becomes harder, but also that the contention observed in data centers strongly affects big data applications' behavior. For all these reasons, ensuring QoS for novel DIAs is a difficult task that needs to be addressed in order to favor further development of the field.
Over this background, the present dissertation takes two main routes towards facing such challenges. At first we describe and discuss a number of performance models based on various formalisms and techniques. Among these, there are both basic models aimed at predicting specific metrics, like response time or throughput, and more specialized extensions that target the impact on big data systems of some design decisions, e.g., privacy preserving mechanisms or cloud pricing models. On top of this, the proposed models are variously positioned across the spectrum between efficiency and accuracy, thus enabling different trade-offs depending on the main requirements at hand. This is relevant in the second main part of this dissertation, where performance prediction is at the core of some formulations for capacity allocation and cluster management. In order to obtain optimal solutions to these problems, in one case at design time and in the other at run time, we adopt both mathematical programming and several performance models, according to the different constraints on solving times and accuracy.
More in detail, we propose performance models based on queueing networks (QNs), stochastic well formed nets (SWNs), and machine learning (ML). This variety is justified by the different uses of each methodology. ML provides algebraic formulas for execution times, which are perfectly fit to be added as constraints in our optimization problems' mathematical programming formulations, thus yielding initial solutions in closed form. Since ML can reliably provide accurate predictions only in regions properly explored during the training phase, the optimal solution is searched via a simulation/optimization procedure based on analytical models like QNs or SWNs, which in contrast are quite insensitive to the parameter range of evaluation, being devised from first principles. These kinds of models boast relative errors below 10% on average when predicting response times. Figure 1 shows the basic structure of a QN model for MapReduce, a framework that has been the main big data solution for years and remains a relevant starting point for the investigation of more recent frameworks boasting more complex programming models, such as Apache Spark. When considering convolutional neural networks (CNNs) and their typical use case with a deployment on general purpose GPUs, on the other hand, ML techniques are better suited for characterizing performance, given the daunting complexity of such software. In Figure 2 it is possible to observe the good accuracy obtained by applying linear regression to data regarding the well known CNN AlexNet.
In terms of optimization, first of all we consider the design time problem of capacity allocation in a cloud environment. The design space is explored via both ML and simulation techniques, so as to choose the best virtual machine type in the catalog offered by cloud providers and, subsequently, determine the minimum cost configuration that satisfies QoS constraints. We show also how this optimization approach was applied during the design phase of a tax fraud detection product developed by industrial partners, i.e., NETF Big Blu. Afterwards we also considered the run time issue of finding the minimum tardiness schedule for a set of jobs when the current workload exceeds predictions and the deployed capacity is not enough to ensure the agreed upon QoS. Thanks to the varied efficiency of performance models, it is possible to solve the design time problem in a matter of hours, whilst run time instances are solved within minutes, consistently with the different requirements.

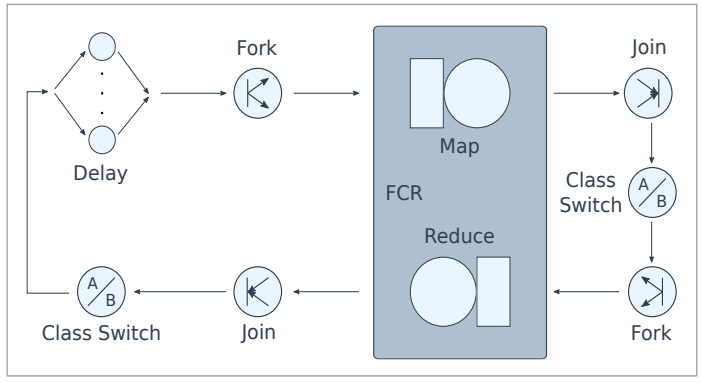

Fig. 1 - Queueing network model for MapReduce

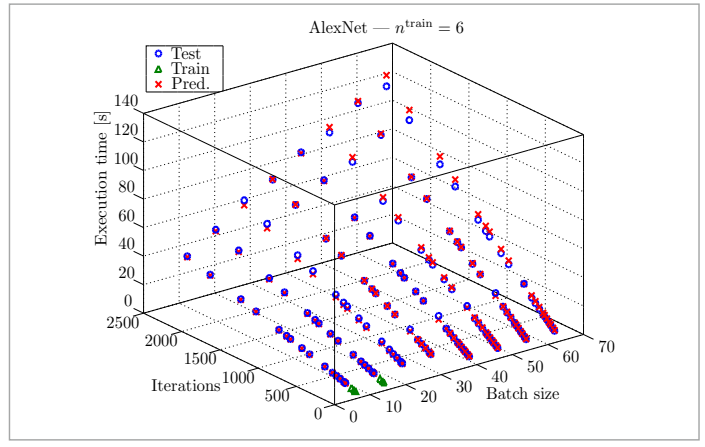

Fig. 2 - ML-based performance model for AlexNet

FREQUENCY SYNTHESIZERS BASED ON DIGITAL PLLS FOR CELLULAR RADIO APPLICATIONS

Luigi Grimaldi - Supervisor: Prof. Salvatore Levantino

Current and future mobile communication standards are targeting a new 10x increase in data rate in the following 10 years, as it is required from 5G standard for instance. To meet these expectations, new and more complex modulation schemes are being standardized, and wider bandwidths at higher carrier frequencies are being employed. Modern transceivers are asked to manage a continuously increasing data rate while keeping high spectral efficiency at a restrained power consumption. Nowadays techniques as the beamforming and phased-arrays proved to be effective solutions to meet the stringent targets imposed by the standard, in terms of SNR of the signal chain and power efficiency. In this context, frequency generation circuits, in both transmitter and receiver side, are asked to reach lower phase noise and spur levels at lower power consumption, while operating in the range of several tenths of gigahertz. At the same time, the implementation of such circuits in new and more scaled CMOS processes imposes a radical change in the design methodology. In the last few years, digitally-assisted analog design, applied not only to frequency synthesis, has been proven to be effective in improving performance in more scaled CMOS nodes. In this scenario, the Ph.D. activity has been devoted to study and implement digital phase-locked loops (DPLLs) for future mobile communications standards, to get a deeper understanding of the adaptive filtering techniques employed in such circuits and to extend their application with the aim of improving the noise and spur-versus power compromise. From this perspective, we focused the research activity on the study and implementation of digital techniques to improve the overall system performance to make it suitable for tight wireless communication standard requirements. The digital approach makes it possible to exploit calibration techniques, working in the background of the PLL operation, to correct for the analog impairments, providing all the advantages of low-power operation, low area occupation, repeatability and portability to more scaled technology nodes. The research carried out during the PhD demonstrated the effectiveness of digital PLL based on a single bit (bang-bang) phase detector when used as frequency synthesizer for modern transceiver architectures. We exploited digital calibration algorithms to achieve low-spurs and low-jitter frequency synthesis whereas we utilized a coarse phase detection based on a bang-bang phase detector, instead of a power-hungry multi-bit time-to-digital converter (TDC), to achieve low-power operation. In order to prove the effectiveness and the high performance achievable by the adoption of these two approaches, we realized two test-chips in 65-nm CMOS technology, the first one in the sub-6GHz band and the second one in the mm-Wave band (around 30GHz). The sub-6GHz prototype, which satisfies the tight spot noise specifications of GSM standard, features a new digital predistortion algorithm to linearize the digital-to-time converter characteristic (DTC) to improve spur level performance. Furthermore, we realized a new architecture capable of achieving fast lock over a wide frequency range, thus overcoming the limited bang-bang phase detector lock range. The mm-Wave PLL demonstrator benefits, in terms of both power and jitter, of the bang-bang phase detector used in a novel sub-sampling mode. We implemented a digital phase selection technique to reduce output jitter without affecting power consumption. Low-output jitter combined with a low power consumption makes this implementation the 1st 30 GHz digital PLL with the lowest Figure of Merit (FoM) for single loop PLLs above 24 GHz.

MODEL DRIVEN ENGINEERING FOR PRIVACY AWARE DATA INTENSIVE APPLICATION

Michele Guerriero - Supervisor: Prof. Elisabetta Di Nitto

The pervasiveness of the modern digital ecosystem is leading to dramatic changes in our society and in the way software systems are designed. Nowadays smartphones, wearables, sensors and, in general, smart devices sensing the environment, allow to continuously collect huge volumes of data. Having the ability to deal with and to make sense of this data can give to companies and institutions important advantages. This is known as the "Big Data" phenomenon and is pushing significant investments in order to re-design software solutions in a more data-centric way.

Researchers and practitioners are tackling the problem from various perspectives, ranging from the development of new technologies (storage solutions and application platforms for managing the execution of data-intensive applications) to the adoption of DevOps as a useful method to reduce the time to market and increase the possibility to feedback inputs from the operational environment back into the design time in order to speed up the innovation cycle. Nevertheless, bringing data-intensive applications (DIAs) into production is still very demanding for several reasons: (i) deploying DIAs is very time-consuming, as it requires the setup and fine tuning of multiple distributed and highly configurable systems on top of Cloud infrastructures; (ii) developing DIAs requires to become familiar with dataflow-based programming paradigms and with multiple non-trivial technologies; (iii) certain non-functional requirements, such as privacy and data protection, acquire primary importance and need to be carefully addressed.

This thesis provides three research contributions along this direction. First, we present a model-driven approach for the development and prototyping of DIAs across different target platforms. The approach adds support to the Unified Modeling Language (UML) for designing DIAs by proposing a novel UML profile. We show how the proposed profile can be used to automatically generate code for DIAs executing on different platforms (namely Apache Spark and Apache Flink) and we evaluate the approach through multiple case studies as well as a user study. Then, we complement our first result with a model-driven approach to simplify and automate the deployment of DIAs by means of the Infrastructure-as-Code paradigm. Also in this case we propose a novel UML profile for modeling the deployment of DIAs on top of Cloud infrastructures and we show how the profile foster the generation of Infrastructure-as-Code (relying on the OASIS TOSCA standard) to automatically deploy UML deployment models of DIAs. The approach has been adopted and evaluated by two industrial case studies. Finally, we present an approach for embedding privacy-awareness in DIAs to help simplifying the compliance with privacy requirements. This approach comprises (i) a novel language, based on a Metric Temporal Logic, for specifying privacy policies on DIAs, (ii) a framework that allows to automatically enforce policies at runtime by re-writing the structure of a DIA and to trade performance and privacy guarantees, (iii) a method to monitor policy violations by exploiting the formal semantics of the defined policies. We prototyped our approach on top of the Apache Flink platform and we performed several benchmarking experiments which show the general applicability of our solution.

Overall, by leveraging several results from Model-Driven Engineering, DevOps, formal methods and data protection techniques, this thesis proposes a novel methodology and a framework to design, develop and deploy modern DIAs, while addressing privacy and data-protection requirements.

ELECTRONICS BOOSTS PHOTONICS: DETECTOR AND ELECTRONIC DESIGN FOR NON-INVASIVE MONITORING AND CONTROL OF SILICON PHOTONIC SYSTEMS

Emanuele Guglielmi – Supervisor: Prof. Marco Sampietro

Electronics is an essential tool that can unlock the true potential of modern Silicon Photonic technologies, overcoming their limitations. My thesis contributes to the field of electronic-photonic integration, studying and improving the innovative CLIPP detector and developing the electronics to use it in novel scientific applications.

Silicon Photonic technologies can achieve outstanding datarates, low losses and power consumption, but require the closed-loop control of the optical devices, to compensate for their extreme sensitivity to fabrication tolerances and temperature fluctuations.

The large-scale implementation of feedback control systems is halted by the inadequate state-of-the-art photo-detectors, that introduce losses to monitor the working point of the photonic circuits.

The CLIPP, ContactLess Integrated Photonic Probe, is an innovative detector developed at Politecnico di Milano that overcomes these limitations by enabling non-invasive light monitoring in silicon photonic circuits, through an impedance measurement of the waveguide. My thesis explores the disruptive applications in Silicon Photonics that an innovative device like the CLIPP has unlocked.

The Detector

This work has deepened the knowledge of the CLIPP detector through experimental measurements, modelling and simulations. In particular, the key components of the electrical model of the device have been identified and linked to the geometrical parameters of the detector, highlighting the important role that the photonic circuit substrate plays in the operation of the device. The CLIPP miniaturization has also been explored, with focus on the optimization of the operating frequency of the detector.

The theoretical analysis concludes with practical layout guidelines for the integration of the CLIPP detector in dense photonic circuits. Particular attention should be devoted to the minimization of the direct coupling between the electrodes, main parasitic capacitance of the detector. CLIPP following these design guidelines have been fabricated within the context of the ICT-STREAMS project, obtaining parasitics as low as 1fF and never exceeding 5fF, an exceptional result compared to previous fabrications where the stray coupling was hundreds of fF, sometimes exceeding 1pF.

A new CLIPP architecture, called Embracing CLIPP, exploits deep n++ implantations at the same level of the waveguide (Fig. 1), available in Active Silicon Photonic technologies, to enhance the capacitive coupling with the waveguide. The device has been fabricated with IMEC technology and shows exceptional performance despite the high process variation. The Embracing CLIPP with n++ implantations at 400nm from the waveguide, achieved the record of the lowest detectable power measured so far with a CLIPP, -55dBm. At the same time, the non-invasive nature of the device is still unaltered.

The Electronics

The Electronics to read the CLIPP is based on a Lock-In Demodulation scheme to sense the resistive waveguide despite the capacitive nature of the detector.

The key aspects of the readout circuit have been discussed, highlighting the importance of using a TIA with capacitive feedback and minimizing the noise contributions of the input capacitance and the Lock-In filter. Input leakage currents needs to be handled with robust bias networks to avoid the saturation of the amplifier.

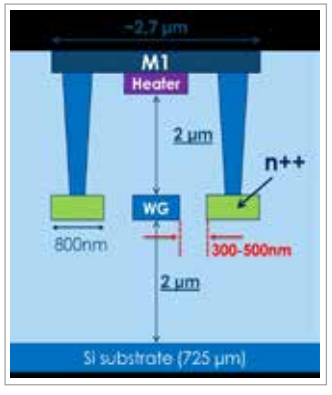

Fig. 1 - Cross section of the Embracing CLIPP electrode, showing deep n++ implantation to enhance the capacitive coupling to the waveguide.

The integration of the control electronic in an ASIC is essential for multichannel applications where a photonic circuit is monitored by tens of CLIPP detectors (Fig. 2). However, the n-type and p-type implantations of the CMOS technology can create parasitic photodiodes that aggravate the problem of leakage currents and make the system sensitive to the environmental light.

Pseudo-resistors made by active CMOS transistors are a compact integrated solution to bias TIA with capacitive feedback and to handle the leakage currents. After a comprehensive analysis of the state of the art, a new pseudo-resistor structure has been designed, realized and tested in AMS CMOS 0.35um technology. The device is tuneable through floating voltage generators, carefully designed to avoid the introduction of parasitic capacitances that would cause impair the frequency performance of the pseudo-resistor.

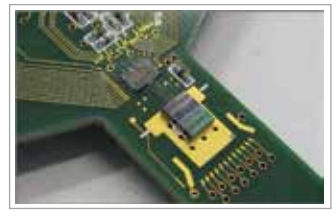

Fig. 2 - Readout electronics integrated in a CMOS chip (top left) directly connected to a Photonic Chip with CLIPP detectors.

The device can synthesize a tuneable resistance from 20MOhm to 20GOhm. Large signal non-linearities have been designed to achieve a dynamic reduction of the output offset when the AC signal is applied. For example, in presence of a 1nA parasitic current, and a 500mV sinusoidal signal, the output offset is reduced from 260mV to less than 50mV when the pseudo-resistor is set to 260MOhm. Furthermore, it is highly symmetrical to handle sinusoidal signals up to 10MHz.

The Applications

The CLIPP is a formidable tool to monitor the state of a photonic circuit and use the information to implement closed-loop control algorithms. The Pilot Tones and Dithering techniques offer the possibility to monitor specific signals or to extract the derivative of the optical components transfer function. Furthermore, they make the measurement insensitive to the parasitic effects and drifts, improving the robustness of advanced control algorithms.

CLIPPs have been used to control a very well-known 8×8 router architecture, showing light path tracking and automatic reconfiguration. Each switching element can be tuned sequentially in less than 500ms with very simple min/max chaser algorithms. Routing of multiple WDM signals through the matrix has been demonstrated with the pilot tones technique, that allows to track each signal independently with more than 50dB of isolation.

Thermal crosstalk effects between the switching elements have been counteracted with local feedback loops, achieving a stable optical crosstalk level of -30dB in both cooled and uncooled cases.

CLIPPs have also been applied to a Mode-Division Demultiplexer photonic integrated circuit, a novel application that requires non-invasive monitoring to avoid disruption of the orthogonality of the spatial modes. The employment of CLIPPs in strategic points of the circuit allows to reduce the tuning complexity from a 12 degrees of freedom global optimization on 4 outputs, down to simpler 2 degrees of freedom sequential optimizations on single outputs. Tuning of the chip has been achieved with a flexible FPGA-based multichannel platform allowing simultaneous monitoring and control of multiple CLIPPs and actuators of the circuit.

LIQDROID: A MIDDLEWARE FOR DIRECT INTERACTION BETWEEN MULTIPLE PROXIMAL ANDROID DEVICES

Anita Imani - Supervisor: Prof. Luciano Baresi

Nowadays, the speed of technology improvements in computing devices is very fast, and users benefit from them interchangeably to carry out various daily tasks. But despite the improvements that have occurred in the field of mobile technology and the connection protocols, the current situation as regards multiple- device interaction techniques is still far behind what it could be, and these computing devices are still mostly working in isolation. Existing multi-device interaction solutions enable the user to continue a task on another device, but their dependency on a specific set of devices and applications on these devices limits their usage and forces the user to act as a bridge between the devices. This entails the user having to perform some preliminary and time-consuming steps to configure the next device on which to resume the task.
This dissertation covers the motivation, design, and development of a novel paradigm to support multiple-device direct interaction by benefiting from the current potentials that exist in the Android OS to offer more advanced features. The proposed solution to support this novel paradigm is a middleware that enhances the creation of distributed Android applications and oversees their execution on a dynamically user-selected set of Android devices. This middleware, which is called LIQDROID, will create a bigger Android ecosystem between these devices that transforms the current pattern of single-user single-device to a fully cooperative environment.
Technically speaking, LIQDROID is an Android service that both augments each single Android machine and manages their cooperation. As well as it will provide a proper framework for the application developers to easily distribute their applications' components on the proximal devices and be relieved of the underlying complexities, and instead put their focus on designing and developing more innovative distributed applications.
In general, LIQDROID goes further than just letting applications installed on different proximal devices to interact with each other or synchronize their data or states while it can handle the distribution and execution of more complex devices' interactions. Instead of developing an OS which needs loads of time and efforts, we have benefitted from the existing mechanism inside the Android framework (i.e. intents, intent filters, content providers, and back stacks) to distribute its execution environment between the proximal devices to enhance the user's expectations.
The high-level architecture of LIQDROID is shown in Figure 1. LIQDROID will be installed on each Android device as a special-purpose service and is settled around the two layers as Connection layer and Interaction layer. LIQDROID will manage the integration and cooperation of the devices from the initialization phase up to the termination phase of a task. In the following we briefly explain the responsibilities of these two layers and their modules:

- The Connection layer sets the proper infrastructure for advertising, discovery and initializing the interaction between the proximal devices. The discovery in this layer happens at the device level which includes different kinds of Android devices (phones, cars, watches, televisions, etc.) while in the next layer it will be at the level of applications' components. For the privacy concern to establish the connection between the devices each user selected device will receive an authentication message and by accepting it, it becomes part of the user-defined ecosystem. The user is also able to create different groups of these connected devices to redirect the execution of a distributed task to his preferred group later.

The *Communication Manager* has been considered to enable LIQDROID benefit from the wide range of available communication protocols, like Bluetooth, WiFi, WiFi Direct, and others, that may support by each device to foster the idea of heterogeneous interactions to favour low-range protocols.

- After the ecosystem, has been created through directly applying the user's preferences, the Interaction layer will prepare the user selected device to become ready to accept the task as well as manage the task's execution during the devices' integration. To obey the Android framework rules, since, it uses intents to perform the interaction between different applications' components, LIQDROID supports the same behaviour, while the *Intent Manager* will enable the user to launch and interact with the components installed on the other proximal devices. It also provides the list of the capable applications' components available on connected devices to let the user choose the one that he prefers more.
- The Task Execution Manager preserves three different categories to handle the task's distribution such as: resume a task on another device(s) (Shifting), or integrate different devices to perform different parts of a single task jointly (Complementarity) or using several devices in parallel to execute a task and make them synched (Parallelization).
- As the user may change its place regularly, and different sets of devices may become accessible, it is mandatory to provide a data replication mechanism to support the data availability for the user at any time and in any place without having dependency to the physical storage of the device that has been initialized the task distribution or the one that have performed the task execution. This mechanism has been handled properly by the *Artifact Manager* module in LIQDROID to store and share the data between integrated devices. As LIQDROID is exploiting a peer to peer architecture, this mechanism also aims to enhance the accessibility to the shared data at the same time through several devices and supports the data versions inconsistency as well. During the task distribution, different kinds of events may happen such as receiving the low battery from one of the integrated devices or receiving a phone call which needs the direct intervention of LIQDROID to optimize the rest of the interaction. The *Event Manager* module is in charge of handling these events through available rules as well as receiving the user's preferences in case of need. Besides managing the executions of the activities, LIQDROID also allows the user to retrieve available services on the proximal devices, interact and control them through the *Service Manager*.

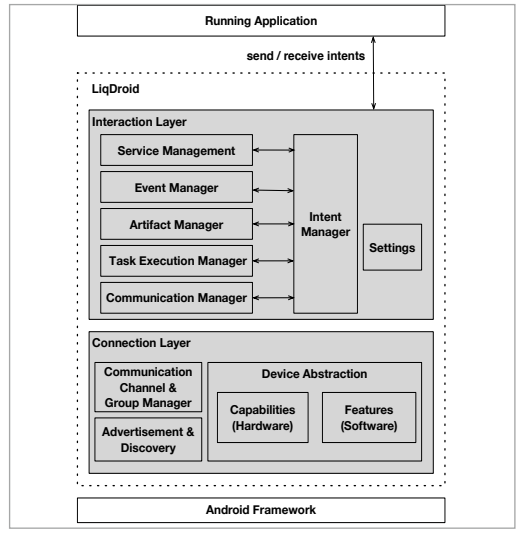

Fig. 1 - High-level Architecture of LIQDROID.

By benefiting from LIQDROID, users can better conceive and exploit the capabilities that are available in their proximal devices and go beyond the repetitive ways of performing different tasks in a single device. LIQDROID proposes a new solution for discarding the limitations that currently exist on the way of multitasking in the Android devices to execute multiple activities at the same time or binding to the services available on the other devices. It can support the currently available Android applications while its greater advantages will be achievable by developing LIQDROID-compatible applications.

SYNCHRONIZATION AND PERFORMANCE EVALUATION OF FUTURE WIRELESS CELLULAR SYSTEM BASED ON THE USE OF NEW MULTI-CARRIER TRANSMISSION TECHNIQUES

Atul Kumar - Supervisor: Prof. Maurizio Magarini

In recent years, the evolution of cellular technology and connectivity has led to a major revolution in the wireless communication industry, with differnt major application scenarios i.e. enhanced mobile broadband (eMBB), massive machine type communication (mMTC), ultra-reliable and low latency communications (URLLC), and vehicle-to-everything (eV2X). Current (4th generation) cellular transmission schemes, which are based on Long-Term Evolution (LTE) and LTE-Advanced (LTE-A), employ orthogonal frequency division multiplexing (OFDM) waveform in PHY design. As is well known, synchronization represents one of the most challenging issues and plays a major role in PHY design and its impact can also be higher in future wireless cellular system, where the use of non-proportional sub-carrier spacing will be required to accommodate the necessary bandwidth. Motivated by this, I studied and realized this Thesis work that concerns with "Synchronization and Performance Evaluation of FutureWireless Cellular System Based on the Use of New Multi-Carrier Transmission Techniques".
With this aim, I started studying about the two categories of multi-carrier waveforms: legacy OFDM waveforms and new waveforms for future wireless cellular systems. For the legacy OFDM waveforms, I studied mainly about the cyclic prefix (CP)-based OFDM system and also an OFDM system with an advanced transformation tool, known as discrete fractional Fourier transform (DFrFT). DFrFT-based OFDM system is analogous to the conventional OFDM one with the difference that DFT and inverse DFT (IDFT) are replaced by DFrFT and inverse DFrFT (IDFrFT), respectively. DFrFT is a generalization of the ordinary DFT with an DFrFT angle parameter (α). Moreover, I started analyzing the effect of synchronization issues, in particular, the problem of evaluating performance in the presence of symbol timing offset (STO) and carrier frequency offset (CFO). Here, we consider the chirp-based DFrFT method for both CFO and STO estimation. From the demonstrated results of MSE performance of CFO and STO estimations, it is realized that CFO and STO estimators with chirp-based DFrFT method performs better than other methods available in literature in case of the transmission over multipath fading channel.
However, in the presence of oscillator drifts and time-varying Doppler shifts, residual CFO and STO are still present in the received signal after the application of synchronization algorithms. In order to improve the robustness to residual CFO and STO, an OFDM system based on DFrFT is considered. Therefore, it would be better to analyse the analytical symbol error probability (SEP) performance of the DFrFT-based OFDM by considering residual CFO and STO together. This is demonstrated by quantifying its effect in an analytical expression of the term responsible for introducing inter carrier interference (ICI) and inter symbol interference (ISI), which is one of the novel contributions of this Thesis. The results of the SEP performance of the DFrFT-based OFDM system in the presence of residual synchronization errors, i.e., CFO and STO, demonstrate that the performance of DFrFT-based OFDM system depends on the αand, by properly selecting its optimal value the DFrFT-based OFDM system performs better than the one based on the DFT. Therefore, the calculation of optimal α is important for better performance of DFrFT-based OFDM system. In this Thesis, we also derived an analytical expression to calculate the optimal value of α. This is another novel contribution of this Thesis. Finally, we also considered more practical aspects related to the realization of a Software Defined Radio (SDR) system which is used to implement hardware co-simulation of multi-carrier transmission techniques. We have considered field-programmable gate array (FPGA)-in-the-Loop (FIL) co-simulation of receiver with equalization of DFrFT-based OFDM system transmission over a frequency selective Rayleigh fading channel in presence of CFO. My simulation results clearly demonstrate that the FPGA implementation of a DFrFT-based OFDM system in presence of CFO has the same performance as that obtained from Monte Carlo simulation. Also, the performance is validated with the fixed-point model of DFrFT-based OFDM. The approach described in the thesis constitutes an efficient way to convert the floatingpoint model into a fixed-point one to be run in an FPGA and then verifies its correctness through FIL co-simulation.
However, as it is well known, CP-based OFDM uses fixed set of waveform parameters, including sub-carrier spacing and length of CP that, are uniformly applied across the entire system bandwidth. Due to the lack of flexibility in supporting mixed services with different waveform parameters within one carrier, which is a key requirement in the PHY design of future cellular network. Additionally, high out-of-band (OOB) emission in frequency-domain is introduced by the time-domain rectangular pulse shaping filter. Also, OFDM signal with one CP per symbol may have a prohibitive low spectral efficiency and also increase the OFDM symbol duration. For this reason, new waveforms that are able to support variable and customizable pulse shaping filters in order to achieve a better trade-off between time-domain and frequency-domain localization is one of the research priorities. With this aim, generalized frequency division multiplexing (GFDM) is one of the proposed multi-carrier waveforms for future wireless cellular systems, which is based on the use of circular filtering at sub-carrier level. Compared to CP-based OFDM, the important advantages of GFDM consist in a reduction of OOB emission, achieved by means of filtering at sub-carrier level, in an increase of spectral efficiency, obtained through the introduction of tail biting, which makes the length of the CP independent from that of pulse shaping filter. Moreover, the flexible frame structure of GFDM is achieved, by changing the number of time slots and sub-carriers in a frame, covering both conventional OFDM and DFT-spread OFDM, which results in backward compatibility with LTE.
Motivated with the GFDM, we focus on the integration of GFDM in the timefrequency grid of LTE system and then analyse the impact of "Better than Nyquist" pulse shaping filters on OOB emission and SER. Moreover, we also consider the concept of wavelet for better time-frequency localization of pulse shaping filters by using the Meyer auxiliary function. After the impact of pulse shaping filter, for an efficient implementation of receiver in time-domain, a relationship between GFDM signal and discrete Gabor transform (DGT) is investigated for reducing the complexity. After that, we implement DGT-based GFDM when the synthesis function, i.e. pulse shaping filter, and the analysis function, i.e. receiving filter, satisfy the Wexler-Raz identity. Choosing functions that satisfy the Wexler-Raz condition allows for optimal symbol-by-symbol detection after DGT-based receiver in case of ideal channel. However, when transmission takes place over a frequency selective channel, symbol-by-symbol detection of sub-symbols is no longer optimal due to intersub-symbol interference (ISSI) generated by sub-symbols transmitted over the same sub-carrier, to improve the performance maximum likelihood detection (MLD) is implemented as a optimal detector for all the sub-symbols on the same sub-carrier, which is another novel contribution of this thesis.
Finally, we derive exact SEP expressions for GFDM waveform in the presence of CFO in AWGN channel and frequency selective Rayleigh fading channel. The analytical expressions of SEP are derived when matched filtering is implemented at the receiver for differnt modulation in case of AWGN channel and for BPSK only in case of frequency selective Rayleigh fading channel which is another contribution. Monte Carlo simulations are presented to demonstrate the exactness of the derived SEP expressions.

FORMAL VERIFICATION OF TIMED PROPERTIES FOR DATA-INTENSIVE APPLICATIONS

Francesco Marconi - Supervisor: Prof. Matteo Rossi
Co-Supervisor: Dr. Marcello Maria Bersani

In the past few years, cloud-based enterprise applications, leveraging the so-called data-intensive technologies, have emerged as pervasive solutions for modern computing systems.
Their adoption has been motivated by the growing need for systems that are able to collect, process, analyze and store huge quantities of data coming from various sources (social media, sensors, bank transactions, etc.) in a reasonable time.
Data-intensive applications (DIAs), taking advantage of those technologies, natively support horizontal scalability, and constitute a significant asset for the production of large-scale software systems.
However, the adoption of data-intensive technologies in the small and medium enterprises (SMEs), constituting the vast majority of the European industry, is still slow for a number of reasons, such as the steep learning curve of the technologies, the lack of experience and resources to keep up with such innovation.
The definition of methodologies and principles for good software design is, therefore, fundamental to support the development of DIAs.
Non-functional (quality) requirements are an aspect of software design that is typically overlooked at design time and turns out to be crucial in later stages of development. Usually expressed in terms of Service Level Agreements (SLAs), if they are not met, further refinements of the applications are needed, resulting in additional costs.
Design time quality analysis aims at detecting the presence of potential design flaws that could lead to later quality incidents, fostering the early detection of problems.
Different approaches exist for the quality analysis of parallel distributed applications.
Performance prediction is arguably the most common and is typically enacted on stochastic models by means of either analytical methods or simulation.
A different approach is pursued by formal verification: it performs an exhaustive check on the model of the system to assess whether certain properties are satisfied by the modeled system.
This thesis presents a unified model-driven approach for the formal analysis and verification of temporal properties for data-intensive applications.
It addresses the computation of the two main classes of DIAs, i. e., streaming and batch processing, by proposing two formal models based on metric temporal logic and analyzable through state-of-the-art solvers.
Both formalizations, albeit with relevant differences, capture the computational models of DIAs as DAGs, enriched with the most relevant quality aspects of the applications.
We first describe timed counter networks, a formal model capturing the computation of Storm-like streaming applications.
The model is devised by extending the CLTLoc metric temporal logic with positive discrete counters, and enables the analysis of properties concerning the growth of the queues of the nodes composing Storm topologies.
The model can be automatically analyzed (with some limited modifications) by means of the Zot satisfiability checking tool, but it presents some undecidability issues, therefore we propose an additional check to assess the soundness of the analysis results.
Then we present an analogous approach for the formalization of the deadline feasibility problem for Spark batch applications. The model is devised as well in terms of CLTLoc extended with counters, but describes a rather different scenario, in which application runs are finite and the goal of the analysis is to determine whether a specific deadline can be met by the designed application.
We also discuss a model reduction strategy that exploits the properties of partially ordered sets to partition the DAG underlying the formal model.
Next, we introduce D-VerT, a model-driven tool that allows designers to perform the formal analysis of streaming and batch applications by means of an automated toolchain, starting from suitably annotated UML diagrams.
D-VerT embeds the models and analysis devised for Storm and Spark applications, and it automatically translates the UML design diagrams to the corresponding instances of the formal models, which are then fed to a state-of-the-art satisfiability solver. In this way, users with limited expertise of formal methods can benefit from the analysis without directly dealing with the specific formalisms.
Experimental evaluation has been carried out over the two kinds of analyses performed by D-VerT.
The queue boundedness analysis of Storm applications was shown to capture design flaws leading to memory saturation in some of the nodes of Storm topologies, and provided some important hints to designers.
In the case of Spark applications, experiments show that the formal model reflects actual executions of the framework with good accuracy (error with respect to the actual execution time less than 10%). Moreover, the proposed optimization registered significant improvements in terms of time (up to 90%) and memory (up to 45%) needed to perform verification.
Future works include further refinements of the formal models to improve their accuracy and performance. For example, the optimization strategy for DAG-based computations could be improved and possibly embedded in the decision procedures implemented by the solvers. We also consider the investigation of new properties and different technologies. Moreover, we plan to extend the experimental analysis to new use cases to have a more extensive evaluation.
Finally, we are investigating the introduction of probabilities in our models to also carry out stochastic analysis.

ANALYSIS AND DEVELOPMENT OF ELECTROCHEMICAL MODEL-BASED STATE ESTIMATION ALGORITHMS FOR LI-ION BATTERIES

Stefano Marelli - Supervisor: Prof. Matteo Corno

Lithium ion (Li-ion) batteries are the most widely adopted technology for electric mobility and consumer electronics, thanks to their ability to store and deliver electric energy more efficiently and effectively than other chemistries. However, costs, performance limits and safety concerns are aspects that still require investments and research efforts. In Hybrid Electric Vehicles (HEVs) and Plug-in Hybrid Electric Vehicles (PHEVs), the battery pack is compact, because it is not the main or the only source of energy onboard, thus entailing less initial cost, but also higher powers to be stood by the battery (compared to its size). Thus, batteries need to be exploited to their limit; a conservative approach would be too costly both for companies and customers. Unfortunately, Li-ion batteries are chemically unstable systems, that require Battery Management Systems (BMSs) to be operated safely and efficiently. The BMS continuously monitors and controls the battery states, such as: temperature, current, voltage, amount of remaining energy, and battery degradation. Many of these states cannot be directly measured; one of the key functions of the BMS is therefore to provide an estimate of these states. The more accurate this estimate is, the closer the battery can be exploited to its fundamental limits, which allows for an efficient and cost-effective utilization.

Accurate state estimation and physical insights into cells behavior are enabled by electrochemical models. In the present thesis, two physics-based electrochemical and thermal models are implemented, namely a Single Particle and Thermal Model (SPTM) and a Pseudo 2-Dimensional and Thermal (P2DT) model. These models describe the dynamics of lithium concentrations and temperatures at different levels of detail (see Figure 1), respectively by means of Partial Differential Equations (PDEs) and Partial Differential Algebraic Equations (PDAEs).

Notice that inner concentration and temperature gradients arise during normal cell cycling, which stresses the importance of an estimator capable of capturing these gradients. The parameters of the SPTM are identified experimentally on a commercial $LiNiCoAlO_2$ 18650 cell. A dedicated laboratory setup, shown in Figure 2, is exploited to force specific input current profiles on the cell under test, while recording the output voltage and the ambient and cell surface temperatures. The model parameters are then fitted to input/output data, via a model-based procedure. The P2DT model is implemented by separating the Ordinary Differential Equations (ODEs) from the algebraic constraints, and proposing an efficient computational structure to solve the constrained ODEs. Also, the thermal dynamics are coupled with the electrochemical ones by means of a control-oriented numerical approach.

The SPTM is used to develop a sliding mode observer, to estimate lithium concentration inside the particles from the measured voltage. An analytic computation of the gain matrix allows to enforce mass conservation in the cell; an extremely efficient, yet robust observer is obtained. The State of Charge (SoC) estimation error converges to less than 2.5%. Also, the SPTM is used to design a backstepping observer, which is another computationally efficient and robust solution. This observer is developed directly on the PDEs, and encompasses a standard mathematical proof. Thanks to the inclusion of bulk thermal dynamics, this observer is validated with experimental tests, showing less than 2% SoC estimation error; importantly, local concentrations and bulk temperature estimates are provided by the observer. The P2DT model is used for the first time, without thermal dynamics, to develop an Unscented Kalman Filter (UKF). The observability issues are solved via a soft-constraint on total lithium mass, and the computational burden is reduced by more than a factor 3 with a parallel computation implementation of the UKF. This approach gives SoC estimation errors of less than 5% in realistic conditions with noisy voltage measurement, and local concentration errors of less than 3%. With the inclusion of distributed thermal dynamics, a novel Dual Unscented Kalman Filter (DUKF) is designed, with similar precautions. In this architecture, the overall observer is composed of two parts: the electrochemical UKF and the thermal UKF. Whilst the former part receives estimated bulk temperature as an additional input and estimates the local values of concentrations, the latter receives the estimated concentrations as additional inputs and estimates the local values of temperatures along the radius of a cylindrical cell. This structure not only allows to estimate the SoC with less than 1.5% error and the local concentration with less than 4% error, but also to estimate the temperature in any point of the cell with an error of only 0.2°C, under currents as large as 50C and noisy measured voltage (see Figure 3).

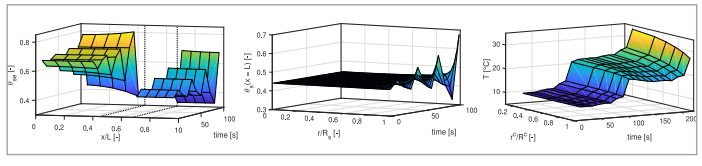

Fig. 1 - Descriptive capability of the P2DT model: (left) solid phase surface stoichiometry along cell film thickness in time; (center) solid phase stoichiometry along active material particle radius in time; (right) temperature along cylindrical cell radius in time.

Fig. 2 - Picture of the laboratory setup for the experimental identification tests, including an Arbin LBT system.

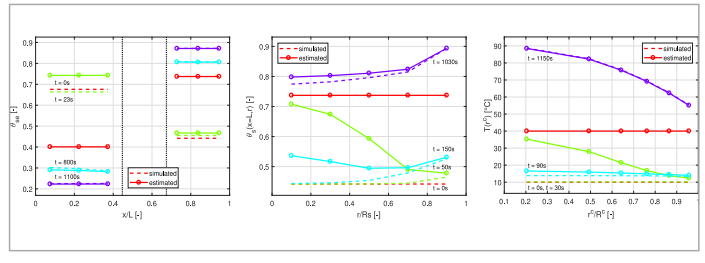

Fig. 3 - Estimation results of the DUKF, initialized with 50% SoC error and 30°C temperature error: (left) solid phase surface stoichiometry along cell film thickness; (center) solid phase stoichiometry along active material particle radius; (right) temperature along cylindrical cell radius.

TIME-SWITCHED FREQUENCY-MODULATION FOR LOW-OFFSET-DRIFT, WIDE RANGE, FULLY INTEGRATED 3-AXIS MEMS ACCELEROMETERS

Cristiano Rocco Marra - Supervisor: Prof. Giacomo Langfelder

Since a decade, the acceleration sensing unit mounted on board of the totality of consumer-grade devices relies on the same working principle: the AM capacitive MEMS accelerometer. Thus, this architecture has reached a mature and consolidated status, showing in a clear way its advantages and its major limitations.
Emerging next-generation applications (ranging from virtual/mixed reality to pedestrian inertial navigation in absence of GPS signal), are making urgent the availability of low-cost, low-power, ultra-high-accuracy MEMS inertial sensors. State-of-the-art AM accelerometers are mainly limited by their intrinsic offset-thermal-drift versus full-scale-range (FSR) trade-off and, unfortunately, both these parameters represent strong figures of merit in the aforementioned applications.
The aim of this thesis is to find an alternative methodology to measure acceleration with a micromechanical structure, bypassing the dead end presented by currently adopted solutions. Thus, a novel working principle for 3-axis frequency-modulated (FM) MEMS accelerometers is proposed.
Resonant accelerometers, in past years, were deeply investigated in literature, thanks to their intrinsic wide full-scale-range and to their immunity to electronics gain drift affecting on the other hand AM solutions. The weak point of this kind of solutions is the adoption of two distinct MEMS resonators with opposite acceleration sensitivity sign to perform differential frequency readout. This architecture is very sensitive to process non-uniformities and to the temperature dependence of resonant frequency in Silicon (TCf). In this thesis, the differential frequency readout is instead performed through the innovative time-switched approach: this methodology is based on a double sampling of the oscillation frequency of a single resonator, consecutively biased in two different configurations in time. The technique enables to avoid offset thermal drift contributions typical of differential resonant accelerometers based on two distinct resonators with unavoidable mismatch in the temperature coefficient of frequency.
The proposed approach can be described in an intuitive way, considering only one tuning fork resonators with a proper timing of the tuning stators. The single resonator is kept in oscillation, through a suitable electronic circuit exciting the anti-phase resonant mode of the two masses. In presence of an acceleration, the two masses experience an in-phase displacement that changes the gap of the two sets of parallel plates electrodes (named plates A and B), biased with two square-waves with a 180-degree phase shift. In this way, two different temporal phases can be defined: phase 1, when the A stators are biased at a voltage different from the rotor one and B stators are equipotential with the rotor and phase 2, in which the situation is specular. Let's consider an acceleration that, during phase 1, approaches the proof mass to the active A electrodes, resulting in a negative anti-phase frequency variation given by electrostatic softening effect. Conversely, in phase 2, only B electrodes are active, and the capacitive gap is increasing: the effect is a resonant frequency positive variation. Acquiring a frequency sample for each temporal phase and taking the difference, a single-resonator time-switched differential frequency readout is achieved. Again, the main advantage of this approach is the use of a single resonator instead of a couple of unavoidably mismatched resonating structures: TCf-related drifts are completely erased. Furthermore, changing the tuning waves DC point, the TCf-related drift can be amplified on purpose in order to counterbalance other known offset drift sources, theoretically trimming the total offset thermal drift to zero.
After a complete behavioral modeling of the conceived system, a compact tri-axis MEMS structure is designed and fabricated using the 30-m-thick STMicroelectronics-Thelma industrial process. Each axis of the sensor is fit into a 500 m x 500 m area, using a centrally anchored architecture in order to minimize substrate stress sensitivity and carefully choosing the resonant modes frequencies and quality factors.
The encouraging results obtained from the characterization of the sensor through a discrete-electronics board (sub 50 g/K at a FSR larger than 42 g) motivated the design of an integrated analog oscillator, in order to prove the possibility to keep the same key performance coupling the device with a low-power, small-footprint ASIC. To this aim, two different feedback oscillator loops are analyzed and fabricated, theoretically and experimentally identifying the optimum topology in terms of power-noise trade-off. Both the loops are realized on the same chip, with the 0.35-m AMS CMOS process, with an overall die area of 4.5 mm^2. The key difference between the two oscillators is the implementation of the phase shifter block (needed to satisfy the Barhausen oscillation criteria). In the first solution, it is implemented as an active integrator with unitary gain at the working frequency. The second approach relies on a phase-locked-loop (PLL) used as a 90-degree shifter. Noise at the comparator input, responsible for noise folding penalty, is different in the two cases, leading to different obtainable resolutions. The integrator performs a low-pass filtering action at the comparator input, reducing the noise bandwidth and thus the folding penalty. Conversely, the use of a PLL as a phase shifter forces the designer to place the comparator right after the front-end stage, in a point where the noise bandwidth, and thus the folding penalty, is large. A passive low-pass filter placed between the two stages can mitigate the issue, but the active integrator clearly performs a more efficient bandwidth limiting action. Furthermore, the PLL circuital complexity generally requires a larger power consumption: the optimal topology is thus represented by the loop based on the analog shifter.
Finally, the tri-axis MEMS-ASIC combo is coupled to an integrated frequency-to-digital converter, demonstrating the feasibility of a fully integrated, digital output, tri-axis FM time-switched MEMS accelerometer. The converter approach used in this thesis relies on a quantization in the time domain, introduced placing a D-flip-flop in a phase-locked-loop (PLL). The flip-flop provides the quantized information about the phase of the signal entering in the PLL (i.e. the integrated oscillator output). Differentiating two subsequent samples, with a period meter based on a counter and a differentiator, the measurement of the period is provided.
This additional differentiating operation introduces a further degree of quantization noise shaping, allowing a considerable decrease of the system clock frequency.
The designed system solves the trade-off between offset thermal drift and full-scale-range experienced by state-of-the-art capacitive AM accelerometers. The ratio between these two critical parameter results improved by more than an order of magnitude with respect to commercial best-in-class products, without any post-acquisition digital temperature compensation. At the same time, the other key parameters (as resolution, power consumption and bandwidth) remain in line with consumer-grade devices and, in particular, with next-generation applications as mixed-reality and pedestrian inertial navigation.

INTERFERENCE MITIGATION TECHNIQUES IN HYBRID WIRED-WIRELESS COMMUNICATIONS SYSTEMS FOR CLOUD RADIO ACCESS NETWORKS WITH ANALOG FRONTHAULING

Andrea Matera - Supervisor: Prof. Umberto Spagnolini

Centralized Radio Access Network (C-RAN) is an attractive solution to handle the huge number of user devices and antennas that are expected to populate next generation (5G and beyond) communication networks.
C-RAN is already adopted in current (4G) mobile networks, in which BaseBand Units (BBUs) and Remote Antenna Units (RAUs) exchange In-phase and Quadrature (I/Q) digitized signals over the so-called FrontHaul (FH) link. However, the expected increase in radio frequency bandwidth demanded by 5G calls into question the effectiveness of digital I/Q streaming, especially for RAUs equipped with multiple antennas.
In this regards, over the last years, several RAN functional split options have been proposed in order to relax the strict latency and bandwidth FH requirements. In contrast with such RAN functional splits, which are mostly based on digital FH, C-RAN based on analog fronthauling is emerging as a low-cost and bandwidth efficient RAN architecture. In analog C-RAN, the RAUs directly relay the radio signals to/from the BBU from/to the end-users, thus bypassing any bandwidth expansion due to digitization, reducing latency, and providing synchronization among multiple RAUs.

In more details, the focus of the thesis is on a particular C-RAN architecture characterized by analog FH links based on Local Area Network (LAN) copper cables, namely Analog Radio-over-Copper (A-RoC). In addition to the benefits provided by analog C-RAN, by leveraging the Power over Ethernet (PoE) technology, A-RoC allows for the powering of the RF equipment over the same copper cables, so that no additional power supplier is needed. A-RoC encompasses both the advantages of analog FH and the economical benefits of reusing the existing copper transport infrastructures, thus becoming the first candidate for extending indoor wireless coverage of next generation 5G networks into buildings and enterprises.
In this thesis, the A-RoC concept is extended to multiple-antennas RAUs and multiple twisted-pairs copper cables, e.g., LAN cables, thus leading to a more general FH architecture characterized by the cascade of a MIMO wired channel over a MIMO wireless channel. This architecture, referred to as Analog MIMO RoC (A-MIMO-RoC) and shown in Fig. 1, is at the center of this thesis.
The A-MIMO-RoC system design requires the optimization over the cascade of two different MIMO channels, i.e., the wireless channel and the cable channel, each of which with different characteristics and constraints. Hence, as a first step, advanced signal processing techniques have been investigated for these two channels, separately. In particular, on the one hand, the focus was on non-linear precoding techniques for next generation Digital Subscriber Line (DSL) systems, namely G.fast; on the other hand, interference mitigation techniques have been designed for wireless systems with focus on multi-operator scenarios, and for optical wireless systems, i.e., Visible Light Communications (VLC).
By merging the knowledge acquired from this preliminary step, the core of the thesis discusses the design and optimization of the Analog-MIMO-RoC architecture, which is complicated by the mutual interaction between the wireless and wired communication channels. In the proposed A-MIMO-RoC architecture, radio signals from multiple-antennas are opportunely mapped over the copper cables multiplexed both in space dimension, given by the multiple twisted-pairs bonded together into each LAN cable, and in frequency dimension, given by the bandwidth of each twisted-pair. It is shown in the thesis that such all-analog radio-cable resource mapping, referred to as Space-Frequency to Space-Frequency (SF2SF) multiplexing, enables the full exploitation of the transport bandwidth capability of copper cables.
The central chapters of this thesis provide extensive numerical analysis and simulation results showing the benefits of the proposed A-MIMO-RoC architecture for 5G indoor networks. The A-MIMO-RoC design and the SF2SF resource allocation problem have been investigated for both uplink and downlink channels, and considering both single- and multi-user settings. For the latter case, the aforementioned SF2SF multiplexing technique has been jointly designed with multi-user interference mitigation techniques, which have been inspired by vectoring techniques originally proposed for DSL systems. Furthermore, the proposed A-MIMO-RoC architecture has been extended to the context of heterogeneous 5G networks, in which different services with different requirements in terms of data-rate, latency, and reliability coexist in the same physical resources.
As a conclusive step, the real world implementation of the proposed architecture has been demonstrated by developing a first A-MIMO-RoC prototype, which was able to prove experimentally and for the first time the potentials of the proposed analog C-RAN architecture for FH indoor applications.
Concluding, starting from the native A-RoC idea, this thesis proposed A-MIMO-RoC: an analog C-RAN architecture characterized by the cascade of wired and wireless channels. Advanced precoding techniques and resource allocation strategies have been proposed for A-MIMO-RoC considering different communications scenarios: uplink and downlink channels, single and multi-user scenarios, but also heterogeneous 5G networks. The overall theoretical discussion has been finally supported by experimental results which make A-MIMO-RoC not only an interesting research topic providing numerous theoretical insights, but mainly a practical solution capable to cope with real-world problems that engineers and researchers are facing today in deploying next generation 5G networks. The fact is that RF signals hardly penetrate into larger buildings from the outside network and, in this direction, A-MIMO-RoC represents an important step towards the design of dedicated indoor wireless systems for all the buildings which nowadays still suffer from indoor coverage issues.

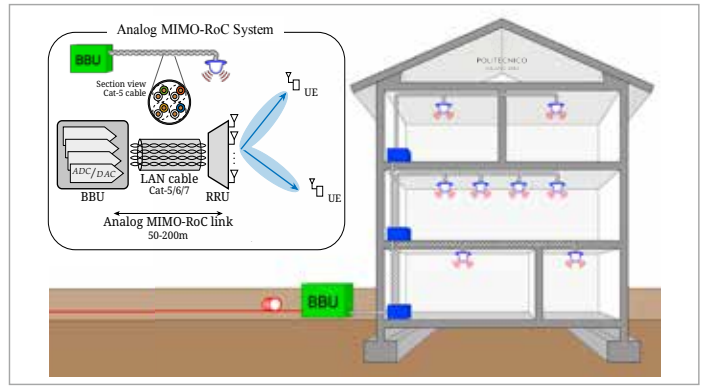

Fig. 1 - Analog MIMO RoC for enhanced indoor wireless coverage

SILICON OXYCARBIDE TECHNOLOGY FOR INTEGRATED PHOTONICS

Faisal Ahmed Memon - Supervisor: Prof. Andrea Melloni

Materials and their characteristics are the backbone of any technological field. The advancement of a technology demands novel materials with appealing properties. Integrated photonics is considered as emerging technology of 21st century. The fundamental block of integrated photonics technology is a waveguide that propagates photons. The waveguides can be used to design devices with enhanced functionality for telecommunication, optical signal processing, control and sensing. Unlike microelectronics, several material platforms in addition to silicon are being used in integrated photonics. The major passive waveguide platforms in integrated photonics are based on dielectrics. The well-established dielectric material platforms include doped-silica, silicon oxynitride and silicon nitride. These platforms have their own advantages and disadvantages and none of them can fulfill the needs of photonics applications such as CMOS compatibility, high index reconfigurability and low losses.

The primary theme of this doctoral dissertation was to develop a dielectric platform for integrated photonics that is CMOS compatible, widely index tunable, low loss, efficiently reconfigurable and can be integrated with typical dielectrics. To this aim, silicon oxycarbide (SiOC) a novel class of glass compounds has been exploited. Reactive RF magnetron sputtering was employed to deposit a system of silicon oxycarbide thin films over a wide composition range and large refractive index window from silica (n = 1.45) to amorphous silicon carbide (n = 3.2). The films properties were investigated in greater detail to assess its potential for micro-photonic device fabrication. Further to advance the development of the platform, medium-to-high contrast photonic waveguides and devices in silicon oxycarbide system were realized employing microfabrication process as shown in Fig. 1. The classical characteristics of the waveguides have been measured in the commercial telecom window to reveal its transparency and potential for photonic applications. Fig. 2 shows the propagation losses of SiOC photonic waveguides with refractive index n = 2.2 that have been estimated less than 2 dB/cm around wavelength λ = 1550 nm.

Silicon oxycarbides have been investigated for possible application in reconfigurable photonic integrated systems. The record high thermo-optic effect in silicon oxycarbides has been discovered that is one order of magnitude larger than typical dielectric platforms. As a further exploitation, integration of silicon oxycarbide with conventional dielectrics resulted in power efficient phase actuators that is a great achievement. The thermo-optic coefficient of SiOC and other optical materials is plotted against their inverse of direct energy gap in Fig. 3. The TOC increases with approaching energy gap and SiOC seems to follow the trend. The largest TOC is exhibited by germanium; however, it is not transparent in the telecom window of 1550 nm.

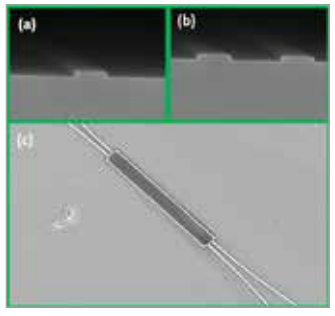

Fig. 1 - Photonic devices in silicon oxycarbide (a) optical waveguide (b) coupler (c) MMI

Within the scope of this work, we have been successful in developing a versatile platform with appealing characteristics of wide refractive index tunability providing low losses in the telecom wavelength range and efficient reconfigurability that was not possible with other typical dielectric platforms.

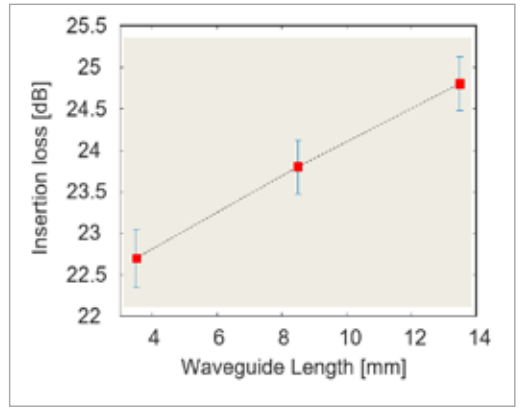

Fig. 2 - Plot showing insertion loss of SiOC (n = 2.2) optical waveguides as a function of waveguide length. The propagation losses less than 2 dB/cm were estimated.

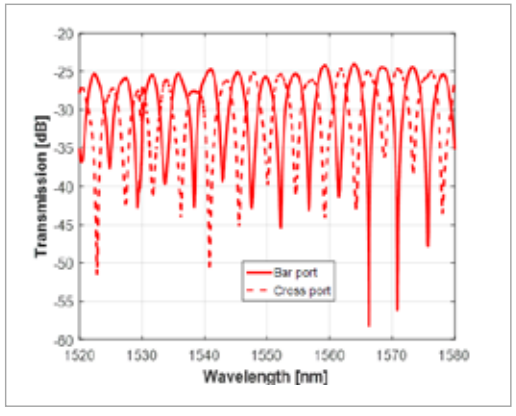

Fig. 3 - Transmission spectral response of MMI based MZI realized in SiOC (n = 2.2) over the telecom window between 1520 nm and 1580 nm

ARTIFACT-DRIVEN BUSINESS PROCESS MONITORING

Giovanni Meroni - Supervisor: Dr. Pierluigi Plebani

Traditionally, to monitor the execution of a business process, organizations rely on monitoring modules provided by Business Process Management Systems (BPMSs), which automate and keep track of the execution of processes. While the adoption of a BPMS to monitor a single-party, fully-automated business process is straightforward, the same cannot be said for multi-party processes heavily relying on manual activities.
In fact, a BPMS requires explicit notifications to determine when activities that are not under its direct control are executed. This requires organizations to federate their BPMSs, a complex task that has to be performed whenever a new organization participates in the process. Also, when activities are not automated, human operators are responsible for manually sending notifications to the BPMS, a task that disrupts the operators' work and, as such, is prone to be forgotten or postponed.
To continuously and autonomously monitor multi-party processes involving non-automated activities, a novel technique, named *artifact-driven process monitoring,* is presented. This technique exploits the Internet of Things (IoT) paradigm to make the physical objects participating in a process *smart*. Being equipped with sensors, a computing device, and a communication interface, such smart objects can then become self-aware of their own conditions and of the process they participate in, and exchange this information with the other smart objects and the involved organizations. This way, it is possible for the monitoring infrastructure to stay in close contact with the process, and to cross the boundaries of the organizations.
To be aware of the process to monitor, instead of using activity-centric process models which are usually adopted by BPMSs, smart objects rely on an extension of the Guard-Stage-Milestone (GSM) artifact-centric modeling language, named Extended-GSM (E-GSM). Normally, a BPMS expects the execution to *rigidly adhere* to the process model defined in advance. Therefore, whenever a deviation between the execution and the model is detected, a BPMS requires human intervention to resume process monitoring. E-GSM, on the other hand, treats the execution flow (i.e., dependencies among activities) in a descriptive rather than prescriptive way. Consequently, smart objects can detect violations during execution without interrupting the monitoring. Additionally, E-GSM can monitor if the physical objects evolve as expected while the process is executed. Finally, E-GSM provides constructs to determine, based on the conditions of the physical objects, when activities are started or ended.
To relieve process designers from learning the E-GSM notation, and to allow organizations to reuse preexisting process models, a method to instruct smart objects given BPMN collaboration diagrams, a widely adopted formalism to model business processes, is also presented. Firstly, a BPMN collaboration diagram is enriched with information on the physical objects participating in the process. Then, for each object, a BPMN process diagram, representing the activities interacting with that object, is extracted. After that, two E-GSM models are automatically derived from each BPMN process diagram: an E-GSM process diagram and an E-GSM lifecycle diagram. The former represents the process in terms of activities and their dependencies. The latter describes the expected evolution of the physical object in terms of finite states and transitions. Additionally, information on when each physical object starts and stops participating in the process is derived from the BPMN process diagram. This information is then used by the platform to determine the identity of the smart objects while the process is being executed.
In addition, an approach to determine the *monitorability* of a process, that is, to which extent smart objects are suited to monitor the process, is presented. To this aim, ontologies are adopted to formalize the capabilities of smart objects (e.g., which sensors are installed and which physical quantities they can measure). This way, given a process, its monitorability can be automatically quantified by querying the ontologies.
Finally, a prototype of an artifact-driven monitoring platform, named SMARTifact, is developed and tested against both historical and live sensor data. SMARTifact consists in four modules: On-board sensor gateway, responsible for collecting sensor data. Events Processor, responsible for inferring the conditions of the smart object based on sensor data. Events Router, responsible for forwarding information on the conditions of the smart object to the other smart objects participating in the process, and for receiving updates on their conditions. E-GSM Engine, responsible for identifying when activities are executed based on the conditions of the smart objects, for keeping track on how the process is being performed, and for detecting when the process deviates from the expected execution.

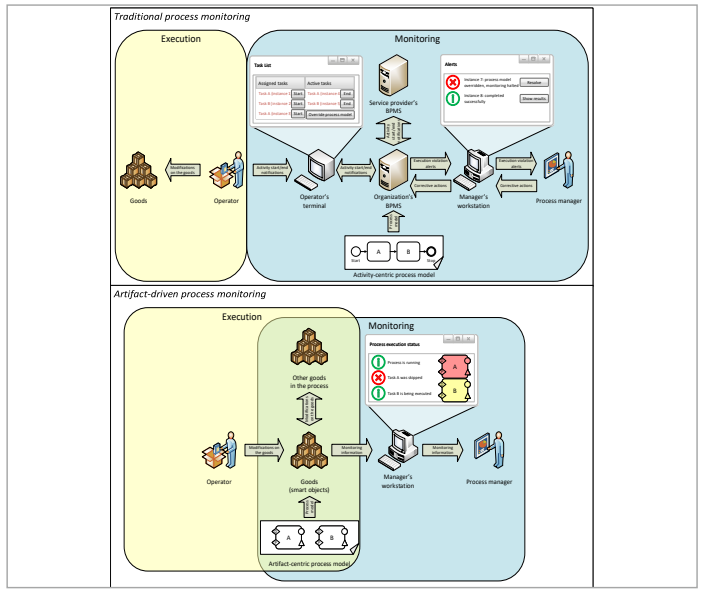

Fig. 1 - Traditional approach to monitor a business process (top), compared to artifact-driven process monitoring (bottom)

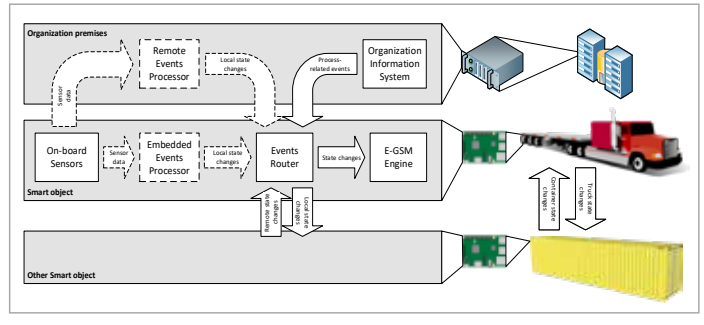

Fig. 2 - Reference architecture of SMARTifact, a prototype of an artifact-driven monitoring platform

MODELING AND SIMULATION OF SPIKING NEURAL NETWORKS WITH RESISTIVE SWITCHING SYNAPSES

Valerio Milo - Supervisor: Prof. Daniele Ielmini

During the last five decades, the microelectronics industry has been steadily evolving thanks to the Moore's law predicting an exponential increase of the number of transistors on the chip, and the increase of clock frequency at each technology generation. Currently, this scaling trend is coming to an end mainly due to the excessive power dissipation. In addition, performance gap between the central processing unit (CPU) and the off-chip working memory makes current digital processors based on conventional Von Neumann architecture inefficient in terms of energy and latency especially for the implementation of emerging data-centric applications such as big data analytics and machine learning tasks.
To face these challenges, emerging memory devices, also known as memristors, such as resistive random access memory (RRAM), phase change memory (PCM) and spin-transfer torque magnetic random access memory (STT-MRAM) have recently gained significant interest for their nonvolatility, scalability, low current operation and compatibility with complementary metal-oxide-semiconductor (CMOS) process.
Moreover, novel approaches aiming to radically subvert Von Neumann architecture blurring the distinction between computation and memory have also been subject of intensive research. Among these novel approaches, neuromorphic computing has rapidly attracted considerable attention for its ambitious objective to emulate the brain ability to carry out extremely complex cognitive functions such as learning, recognition, inference, and decision making with an unrivaled energy efficiency due to its event-driven information processing.
To achieve this goal, RRAM can play a key role enabling to replicate synaptic plasticity rules believed to be the origin of learning such as spike-timing dependent plasticity (STDP) and spike-rate dependent plasticity (SRDP) at device level thanks to its tunable resistance. Also, nanoscale size of RRAM devices offers the great opportunity to achieve high-density integration of resistive devices, thus paving the way for the hardware implementation of high-density spiking neural networks with resistive synapses capable of brain-inspired computing.

This doctoral dissertation covers modeling and simulation of spiking neural networks with hybrid CMOS/RRAM synapses capable of bio-realistic learning rules for implementation of brain-inspired cognitive tasks such as unsupervised learning of visual patterns and associative learning.
First, an overview of fundamental issues currently challenging the performance improvement of today's digital computing systems based on standard CMOS technology is provided. Moreover, the physical mechanisms and key characteristics of the main emerging non-volatile memory technologies, and the novel computing approaches proposed to overcome the current technology paradigm are extensively described.
Secondly, physics-based modeling of HfO_2 RRAM devices is described. A previous numerical model of HfO_2 RRAM providing a detailed understanding of the resistive switching mechanism at device scale is primarily reviewed. After, a previous analytical model of HfO_2 RRAM derived from numerical model is also reviewed. In addition, a stochastic model taking into account the statistical variability of set/reset processes in HfO_2 RRAM devices is described. Then, two hybrid CMOS/RRAM synapse circuits developed to replicate two fundamental bio-realistic learning rules such as STDP and SRDP are presented.
The implementation of STDP rule by a hybrid CMOS/RRAM synapse circuit with one-transistor/one-resistor (1T1R) structure is discussed by simulations and experiments. In addition, the implementation of SRDP rule by a hybrid CMOS/RRAM synapse circuit with 4-transistors/one-resistor (4T1R) structure is also discussed by simulations and experimental measurements. Moving from device to network level, the implementation of unsupervised learning and recognition of visual patterns by 2-layer feedforward spiking neural networks with 1T1R RRAM synapses capable of STDP is presented at simulation and experimental level. After discussing learning of a single pattern, on-line learning of sequential patterns and multiple patterns is also extensively addressed in simulation and experiments.
After that, the implementation of unsupervised learning of visual patterns by 2-layer feedforward spiking neural networks with 4T1R RRAM synapses capable of SRDP is presented. After discussing learning of a single pattern for variable initial weight configuration, on-line learning of sequential visual patterns is also investigated by simulations at network level.
Finally, a circuit implementation of a Hopfield recurrent spiking neural network with excitatory and inhibitory 1T1R RRAM synapses capable of STDP is presented. After discussing learning and recall of both a single attractor state and a sequence of two non-overlapping attractor states via simulations, RRAM-based Hopfield network is used to explore fundamental human brain primitives such as associative memory, pattern completion and error correction.

ON HOW TO DESIGN OPTIMIZED SPATIAL ARCHITECTURES: FROM ITERATIVE STENCILS TO CONVOLUTIONAL NEURAL NETWORKS

Giuseppe Natale – Supervisor: Prof. Marco Domenico Santambrogio

We are approaching the end of Moore's law and Dennard scaling, while demand for computing power increases constantly. Traditional processors are struggling to keep up with performance requirements, and both scientific research and industry are exploring reconfigurable architectures as an alternative. It is common knowledge that an increasing number of technology leaders, such as Microsoft, IBM, Intel, Google, Amazon to name a few, are currently exploring the employment of reconfigurable architectures as hardware accelerators. Fine grained inherent parallelism and low power consumption thanks to direct hardware execution are the key aspects that make reconfigurable architectures an attractive choice for High-Performance Computing (HPC).

This thesis is focused on a set of algorithms sharing a similar computational pattern, namely iterative stencils and Convolutional Neural Networks (CNNs).
The key computation for both algorithms consists in sliding a filter on the input data, computing new elements using a submatrix of the input. Iterative stencil algorithms are heavily employed in physics, numerical solvers and even finance, while CNNs are one of the recently developed deep learning algorithms, currently used in industry to perform image classification, analysis of video and even speech recognition and recommender systems.
While High-Level Synthesis (HLS) capabilities have improved dramatically over the recent years, the synthesis tools have yet to reach the level of sophistication required to properly optimize these algorithms, and extract sufficient parallelism or generate highly scalable solution, or automate the acceleration creation and integration in a way that resembles the development experience that is in place for CPUs and GPUs.
Indeed, the process of designing and deploying hardware accelerators for iterative stencils of CNNs is still a hard and complex task that requires expertise in FPGA programming and knowledge of hardware design tools. The complex dependencies that arise from the filtering conditions, the iterative nature of the algorithms, and the low operational intensity, make current HLS solutions inadequate to really optimize the resulting implementations. This thesis objective is to improve with respect to the proposed solutions targeting all the presented challenges.

For both the target algorithms, we designed optimized spatial architectures that are able to exploit different sources of parallelism of such algorithms, reduce the cost of data movements alleviating the burden on external memory and that can easily scale up to multi-FPGA systems. In particular, we propose spatial accelerators for iterative stencils and the features extraction stage of CNNs. The resulting designs, consisting of distributed architectures of independent elements communicating using read- and write-blocking FIFOs, can scale in size if enough resources are available, unfolding the algorithms computation in space. Such architectures can exploit different levels of parallelism offered by the target algorithms. For iterative stencils, we propose a dataflow accelerator template that is able to exploit both inter- and intra-timestep parallelizations, while keeping at minimum (and analytically optimal) the on-chip memory requirements. We thoroughly investigate the analytical properties of this accelerator, and to evaluate how the solution scales, we implement a multi-FPGA system to use as testing platform. Moreover, we propose a design automation flow that is able to generate the accelerator starting from a C/C++ input source.
For CNNs the resulting accelerator can exploit parallelism at different levels, pipelining the execution between layers, computing output feature maps concurrently, and processing input feature maps in parallel. To efficiently exploit DSPs, we introduce a time-sharing technique to compute each convolution with one DSP primitive, by which most of the data-path works at a submultiple of the clock frequency, resulting in the possibility to exploit higher levels of parallelism inside the accelerator, and achieving high energy efficiency and low off-chip bandwidth requirements. We also provide a simple performance model to assist the design process and allow to estimate the achievable performance with good accuracy.

The validation performed shows that for iterative stencils, the proposed solution is comparable with the state of the art for the single FPGA implementation, but we are able to perform substantially better on multi-FPGA, thanks to an approximately linear scaling in performance. Moreover, for deep CNNs our implementations for AlexNet and VGG-16 achieve a throughput of respectively 1.61 and 2.99 TOPS, where for VGG-16 we are the second-best implementation -- by a narrow margin -- available in the State of the Art, while for AlexNet we substantially outperform the previous work.

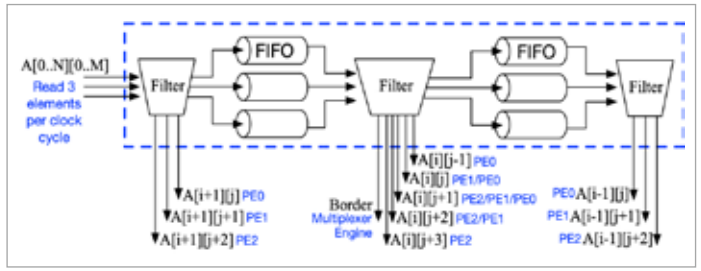

Fig. 1 - Memory subsystem of the accelerator template for iterative stencil algorithms

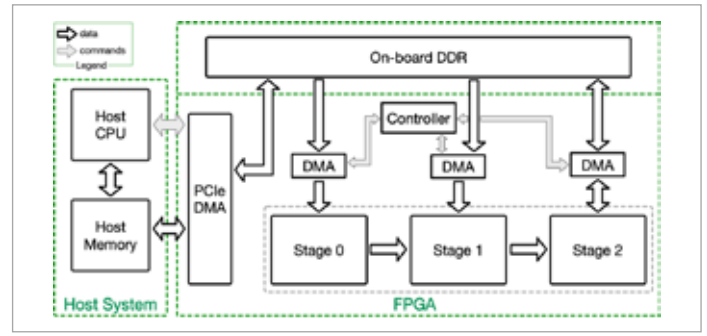

Fig. 2 - High level overview of the proposed hardware accelerator for CNNs

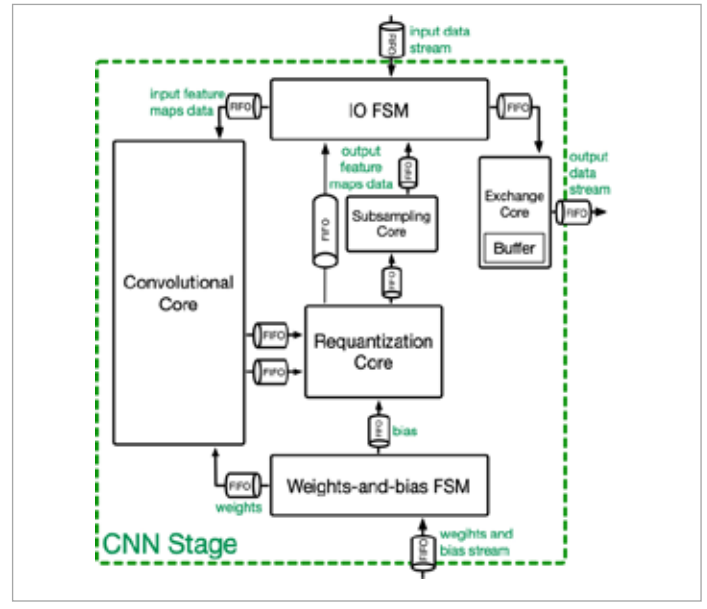

Fig. 3 - Microarchitecture of a single stage of the CNN accelerator

A GENERAL FRAMEWORK FOR SHARED CONTROL IN ROBOT TELEOPERATION WITH FORCE AND VISUAL FEEDBACK

Davide Nicolis - Supervisor: Prof. Paolo Rocco

Robot teleoperation has been long employed in a variety of applications where a human user is required to operate from a distance a robotic device, often a robot manipulator. This technology has seen its first applications in the '40s and '50s in nuclear material handling. In 1954 the electromechanical master-slave manipulator invented by Goertz laid the foundations of modern telerobotics and force reflecting devices, replacing the pure mechanical and hydraulic architectures of the time.
Currently, the topic of interaction between user and robotic devices has been receiving increasing attention from the research community and the industry.
The use of telerobotics is often motivated by the inaccessibility of the environment where the task must be performed, caused by hostile conditions. Noteworthy applications include minimally invasive surgeries, where a remotely controlled robot can minimize the procedure invasiveness and reduce tremors, space robotics for on-orbit servicing and planetary exploration, as well as search-and-rescue operations. The range of industries interested in this technology is growing to include sterile drug manufacturing and all those fields where access to a potentially hazardous environment is required while keeping a high degree of safety for personnel, or simply where a human himself would not be able to operate effectively.
The increasing attention for the application of these systems in such harsh conditions has also seen the interest of institutional organizations taking part in funding initiatives, such as the H2020 RoMaNS project for nuclear waste handling, or the WALK-MAN teleoperated humanoid robotic platform successfully tested in disaster scenarios.
As teleoperation applications and platforms grow more complex, the employed control framework should be able to relieve the user of some of the burden caused by operating such devices, establishing a sort of shared control. This work aims at proposing a comprehensive control framework for teleoperation systems, spanning from low level motion control aspects, interaction control, and system stability analysis when a human is in the loop, to higher level visual-aided control algorithms ensuring a simplified and intuitive use of the teleoperated platform, and a reduction of cognitive and physical fatigue for the operator.
At a local control level, robust control theory is employed to achieve a desired behavior of the teleoperated robots during interaction with the environment. Sliding mode control is used to robustly shape master and slave manipulators impedances irrespectively of uncertainties. This formulation has been addressed differently from previous literature, by directly specifying sliding manifolds in the robot operational space. The manipulator redundancy is solved via successive projections of sliding manifolds defined by lower priority tasks, into the null space of the higher priority ones. A hierarchical optimization outer layer considers control and motion constraints, allowing the inclusion of requirements better modeled via inequalities, such as torque or joint limits. The model predictive nature of the control also guarantees robust compensation of actuation delays and unmodeled filter dynamics (Fig. 1). To help and guide the operator, the specification of hard and soft virtual fixtures is tackled at this level, with virtual force feedback rendered through the analysis of the dual solution of the optimization. This enables the reduction of the user cognitive burden, by letting the robot autonomously control some of the tasks and providing kinesthetic information on the status of the remotely controlled device. A stability analysis of the overall control scheme in presence of variable communication delays during contact is performed preliminarily by relying on the Small Gain Theorem and then on passivity arguments, thanks to Llewellyn's absolute stability criterion. Therefore, clear tuning guidelines for master and slave robot impedance control parameters are obtained, guaranteeing stable bilateral control.
Furthermore, visual feedback by means of image-based visual servoing is integrated and experimentally validated on a teleoperated dual-arm platform, where one arm is teleoperated and the other completely autonomous and camera-equipped (Fig. 2). The proposed controller helps the user in navigating cluttered environments and keep a clear line of sight with its target by completely avoiding occlusions, reducing the operator workload required to complete a reaching task and delegating any camera reorientation task to the autonomous arm.
Finally, machine learning techniques in the form of Recurrent Neural Networks are employed to infer the user intention during collaborative tasks. Based on human motion limb studies and synthetic data, two neural networks are trained in order to predict the user motion and infer which goal the user wants to reach, given a set of possible targets. This information is the used to actively help the operator via the inclusion of an assistance control component, that has been experimentally verified to reduce user physical fatigue and autonomously complete the desired task (Fig. 3).

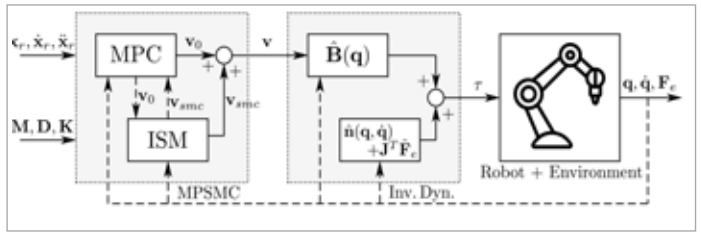

Fig. 1 - The proposed model predictive sliding mode control scheme for robust robot interaction control.

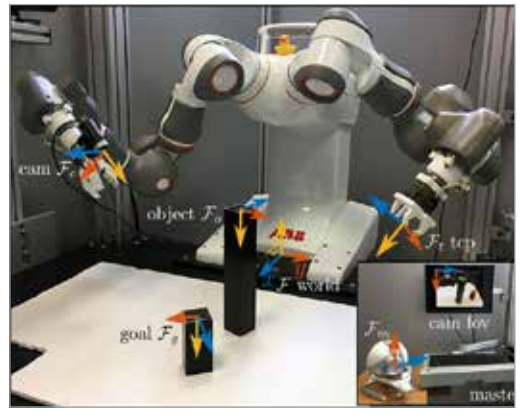

Fig. 2 - The experimental setup used for the proposed occlusion tolerant visual servoed dual-arm teleoperation.

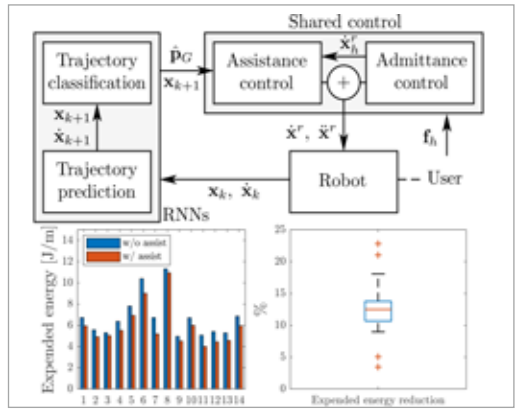

Fig. 3 - The shared control scheme employed for user assistance and the experimental results showing a systematic decrease in user fatigue.

PERFORMANCE AND RELIABILITY ISSUES OF NAND FLASH CELLS AT THE TRANSITION FROM PLANAR TO 3-D ARRAY ARCHITECTURES

Gianluca Nicosia - Supervisor: Prof. Chistian Monzio Compagnoni

The activities carried out during this Ph.D. aimed at shedding a light on the dominant reliability issues affecting the last planar technology node of NAND Flash arrays and at understanding the impact that the implementation of a polycrystalline silicon channel in three-dimensional (3-D) NAND Flash cells has on the performance of these devices.

Focusing, first, on planar arrays, the main source of threshold voltage (V_T) broadening at time 0 affecting deca-nanometer cells, i.e. program noise, was investigated for the first time with a single-electron resolution. The possibility of observing single-electron charging of the cell floating gate (FG) was exploited to directly study the statistical nature of the electron injection process. Moreover, the single-electron resolution allowed for the first time to extract previously physically unaccessible technological parameters like the spread of the control gate (CG) to FG capacitance (C_{pp}) among the cells in the array and electron leakage through the IPD stack on fully processed samples under real operating conditions. These results have been exploited to develop a Monte Carlo simulation tool of the programming transient that implements both C_{pp} variability and inter-poly dielectric (IPD) leakage. Simulation results showed that neither the spread of C_{pp} among the cells, nor the electron leakage from the FG to the CG impact on programming accuracy. Furthermore, a new physical picture for cycling-induced charge detrapping was presented, which explains experimental evidence that contradicts models already presented in the literature.

Through in-depth experimental activities, a detailed study of the dependence of string current (I_S) and V_T on temperature in 3-D NAND Flash arrays was carried out. Differently from what observed in planar devices implementing a monocrystalline silicon channel in which phonon scattering is the main source of mobility degradation at higher temperatures, I_S of 3-D NAND strings increases as temperature is increased not only in the subthreshold region, but also in the ON-state regime, with an activation energy that depends on gate bias. By reviewing the literature on polysilicon thin film transistors (TFTs), results have been explained in terms of thermionic transport across the polysilicon grain boundaries, which leads to an increase of the mobility of electrons through the channel as temperature is increased. This physical picture suggests that transport across the intergrain regions represents the bottleneck to channel conduction. Therefore, the random distribution of polysilicon grains could yield to a dispersion not only of the neutral V_T of the cells, but also, and more importantly, of their sensitivity to temperature. Technology computer-aided design (TCAD) simulations calibrated on the experimental results highlighted, for the first time, that the haphazardness in the configuration of the polysilicon grains in the string is responsible for a nonnegligible variability in the temperature-induced V_T shift of the memory cells.

Owing to the peculiar dependence of I_S on temperature in 3-D NAND Flash strings, also the temperature dependence of random telegraph noise (RTN) fluctuations was addressed. By studying the distribution of the V_T shift between two subsequent reads performed on many cells coming from a fresh array and fitting the data with a defect-centric model for RTN, it was found that the average fluctuation amplitude increases as temperature is reduced. Such a phenomenon represents a novelty as it has not been observed on planar arrays. RTN amplitude temperature dependence was further investigated by directly monitoring some RTN waveforms arising from single traps in cells out of both memory arrays or test elements. This experimental approach allowed to clearly observe, together with a change of capture and emission time constants, variability in the temperature dependence of the RTN fluctuation amplitude. Moreover, no correlation has been found between RTN time dynamics and the V_T shift introduced by the RTN trap. As the amplitude of RTN fluctuations is ruled by the degree of percolation in channel conduction, results have been explained in terms of increased nonuniformities in carriers flow in the string when temperature is reduced. As a matter of fact, first of all, lower temperature makes thermionic transport across the polysilicon grain boundaries more difficult. Moreover, the constant-current criterion for V_T extraction implies an increase of gate bias when reducing the temperature to compensate for the reduction of I_S. This yields to a stronger inversion of the intergrain regions and higher filling of trap states at the grain boundaries, making transport across them even more difficult. In this picture, RTN traps placed at or close to the grain boundaries increase their impact on cell V_T at lower temperatures, thus increasing their fluctuation amplitude. This physical picture has been confirmed by means of TCAD simulations implementing a random configuration of 3-D polysilicon grains and calibrated on the experimental results of I_S temperature dependence.

Finally, the impact of cycling on RTN and its temperature sensitivity has been investigated for the first time in 3-D NAND Flash arrays. Experimental results show that not only RTN in 3-D arrays is much lower than in their planar counterpart, but also its growth with program/erase cycling is much weaker. Moreover, it was shown that cycling has no impact on the RTN amplitude temperature sensitivity. As RTN in planar NAND is, together with program noise, one of the most serious reliability issues leading to distribution widening at time 0, this evidence is an additional proof of the reliability and performance improvements coming from the transition to 3-D integration.

Results obtained during this Ph.D. research represent an important step ahead towards understanding the reliability issues affecting state-of-the-art NAND Flash arrays, thus paving the way to prolong the historic increase in their storage density. Although single-electron charging during programming is no longer observable in 3-D NAND arrays owing to larger cell dimensions, the physical understanding of program noise and its modeling enabled by the analysis of the programming transients with single-electron resolution still holds true. Furthermore, the characterization and modeling results of I_S and RTN fluctuations in 3-D NAND strings provide important information in guiding the design of Flash arrays. They show, first of all, that read margins degradation when reading a page at a different temperature from the one at which it was programmed could be negatively affected by polysilicon-induced variability. Moreover, low temperatures represent a more critical operating condition in terms of time 0 V_T distributions for 3-D NAND arrays. As a matter of fact, they yield not only to an increased amplitude of RTN fluctuations, but also to a lower maximum sensing current as a result of higher string resistance. RTN amplitude temperature dependence also implies that extreme care must be taken in performing spectroscopic investigations of RTN traps which are based on the assumption, valid for monocrystalline devices, that the fluctuation amplitude is a signature of a specific defect and it is not affected by temperature.

RECONFIGURABLE PHOTONIC INTEGRATED CIRCUITS FOR HIGH CAPACITY OPTICAL NETWORKS

Douglas Oliveira Morais de Aguiar
Supervisor: Prof. Andrea Melloni

Day by day the use of Photonic Integrated Circuits (PICs) is growing and new applications are being explored in an accelerating rate. From biotechnology to aerospace, from telecommunications to agricultural monitoring, many are the markets in which Integrated Photonics Technology can provide key technological differentials. However, one of the main roadblocks is the initial cost that it is required to develop system prototypes and Proof of Concepts. One usually compares the current situation of the Integrated Photonics industry with the one from the electronics industry back in the '70s. At that time, there were few standard processes and few companies able to fabricate integrated electronic devices. The situation there changed when Multi Project Wafer (MPW) runs emerged. That helped companies and researchers to prototype their electronic Integrated Circuits (ICs) and demonstrate their work and devices with much-reduced costs. The basic idea of a MPW run is to allow companies and universities to share the expensive cost of the set of lithographic masks. This is a concept that only today is made possible in the Integrated Photonics industry, with the first generic foundry based PIC products being commercially available only in 2015.

In Fig. 1 it is depicted a typical development cycle in Integrated Photonics and in this thesis a complete cycle as the one shown is reported. The cycle starts with the study of the materials to be used, and it involves the analysis of diverse aspects including whether active or passive devices are needed, the expected optical propagation and coupling losses, if there is a need of polarization diversity and the required tuning efficiency of the devices. Then, the nanofabrication phase determines the feasibility of the devices by settling physical limits such as the minimum feature sizes, the achievable backscatter levels, and the efficiency of active devices. Next in order, the design of the components is carried out and is related to the application requirements such as system bandwidth, channel isolation and response time. Sub-sequentially follows the overall circuit design that should, above all, fit into a reasonable physical dimension, should also be controllable by an electronic system that does not consume much power and delivers the expected features. Everything should be planned to be used in external environments that are not necessarily as controllable as the laboratory environment. And finally, everything should be placed on an optical bench and tested to see if all the proposed specifications are met.

On another hand, given the escalation of demand for high-speed data interconnection, both between users and datacenters, high capacity optical networks need a boost in capacity, flexibility, and efficiency. After nearly 30 years of consistent capacity growth, optical telecommunication networks are approaching a fundamental capacity limit. Such networks need new technologies to solve the so-called "optical capacity crunch" problem. In few words, such problem states that networks are saturated, consume lots of power and occupy lots of space. Furthermore, network failures are frequent and compromise the availability of services.

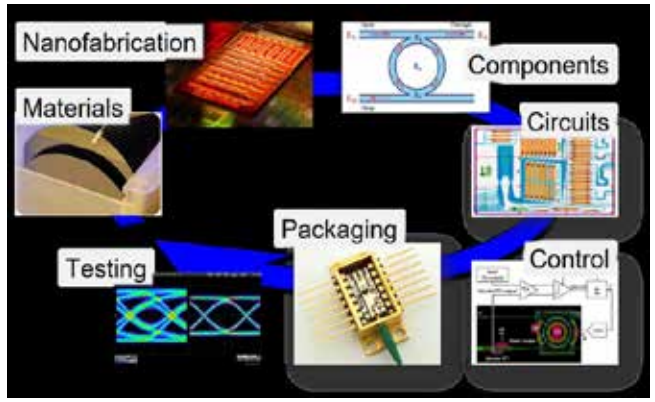

Fig. 1

To stand up for those problems, the network reconfigurability is a key feature in a saturated and power-hungry network operating scenario. Either to act in the case of equipment failures and fiber cuts or to support different traffic requirements and expand network capacity. To fulfill this demand the network elements must be flexible, compact and deliver high performance. The technology that is most widespread today, discrete photonics, does not address those requirements in a scalable and cost-efficient way. Integrated Photonics and the generic foundry approach are the state-of-the-art technology that permits such evolution.

This thesis has outcomes in diverse areas under the Integrated Photonics umbrella, and the most innovative ones are 1) **a telecom graded reconfigurable filter in Silicon Photonics (SiP).** A complete design, testing, and characterization of telecom graded optical filters in SiP are presented. A commercial foundry run was used to fabricate the device and the characterization and operation results of the device are presented throughout this thesis. 2) **The locking of high order Microring Resonator (MRR) filters in SiP using channel labels.** Using a novel technique to decouple the thermal crosstalk typical in SiP systems, locking to a data transmitting signal was demonstrated for the first time.

Fig. 2

3) **The automatic control of hitless SiP optical filter.** The dynamic control of the filter was done and a hitless reconfiguration of the hitless MRR filter was demonstrated by verifying that the Bit-Error-Rate (BER) of a neighbor channel was not affected by the reconfiguration, and 4) **an in-band Optical Signal to Noise Ratio (OSNR) measurement with channel labels.** A novel OSNR measurement technique is proposed and demonstrated making use of channel labels and lock-in demodulation.Most of the optical performance required by telecom networks were addressed in this work, however, one key aspect that is the polarization diversity scheme was identified to still be an open question.

The polarization extinction ratio obtained in commercial 2D grating couplers is not enough to cover the full C-band, and the same limitation is found on integrated polarization rotators. This imposes a limitation on the overall optical performance values that are achievable. The second most critical aspect that was observed in the design was the wavelength dependence of the directional couplers used in this work. In MRRs the optical bandwidth of the filters is determined by such couplers, thus the overall performance of the filter along the operation bandwidth is directly related to this variation, which cannot be made arbitrarily small in for small coupling coefficients.

A completely functional device (shown in Fig. 2) to operate hitlessly in an optical network was demonstrated, including the demonstration of the insensitivity of the neighbor channel to the reconfiguration of the filter. In addition, a novel technique to measure the in-band OSNR of an optical channel was proposed and demonstrated in a 10G system.

DATA ANALYSIS AND MODELS FOR LONG-DISTANCE BIRD MIGRATION: FROM CORRELATIONS ON RING RECOVERIES TO MACHINE LEARNING ON GEOLOCATOR MEASUREMENTS

Mattia Pancerasa - Supervisor: Prof. Renato Casagrandi
Co-Supervisor: Prof. Roberto Ambrosini

Climate is one of the fundamental shapers of ecosystems, thus its ongoing changes deeply influence the behavior, distribution and dynamics of plant and animal populations. Migratory birds are among the species most affected by this phenomenon, as they need to fine-tune their phenology according to the climatic conditions of their breeding and wintering areas. To investigate how and to what extent alterations of climate regimes may determine key changes in the movement ecology of migratory birds, a detailed knowledge of their staging sites, a trustable reconstruction of their migration routes and of the time schedules of their journeys is very necessary.
The classic methods for studying migration, such as bird ringing, can now be complemented by new technologies, such as GPS loggers of light level geolocators, that allow to record proxies of organisms' positions throughout their routes.
Focusing on a model species, the barn swallow (*Hirundo rustica*), in this work we first developed a method to investigate the occurrence of climatic connections between the African wintering and European breeding areas of this migratory passerine bird: surprisingly significant correlations between the average temperatures in the wintering and breeding locations of individuals emerged at the precise weeks of individuals' spring migration. The sign of the correlations found between the temperature series divides the set of individuals analyzed in two clusters that are extremely similar to a clustering produced from breeding and wintering geographical positions (cluster N: equatorial Africa; cluster S: southern Africa) (Fig. 1). Correlations have high significance only in the proximity of barn swallow wintering sites and if the temperature series refer to their migration period. This result could lead to hypothesize a possible selection of wintering locations based on the climatic connections present with individuals breeding areas.
In the second part of this thesis, we reconstructed from light level geolocator measures the migratory routes of 88 barn swallows (Fig. 2) belonging to three different populations (N: southern Switzerland; SW: Piedmont; SE: Lombardy). This analysis provides for the first time valuable ecological information on a large sample of small migratory birds. We verified the repeatability of the estimation method used (*FLightR R* package) showing how our results could be used in a large-scale analysis of the routes traveled by these birds.
The results obtained allowed us to identify four groups of individuals, as well as a possible effect of the year of migration on many indicators of the migration schedules obtained from the reconstructed routes. We have therefore analyzed every step of the migration of the reconstructed routes, highlighting the possible carry-over effects of one migratory stage on the following ones.
Finally, we have automated a time-consuming manual procedure of the pre-processing of geolocator data: the classification of twilight events based on the light curves associated with them. To achieve this goal, we have implemented three machine learning algorithms, thanks to the amount of data available as a result of the 88 routes reconstruction.
The classification performance of two of these methods was comparable with those of an expert human classifier. We tested the reliability of the results of the automatic classification, obtaining a reconstruction of the migratory routes very similar to that provided by the manual selection of the twilight events (Fig. 3).
This procedure could lead to the complete automation of the light level geolocator analysis process, thus allowing to quickly put in us large amount of data already collected on a wide array of species.
All the presented analyses are completely extensible to other species of migratory birds or animals. They provide a contribution to the definition of possible methodologies for investigating the climatic cues used by birds to calibrate their migration and for the reconstruction and the study of individual migratory routes from geolocator data.

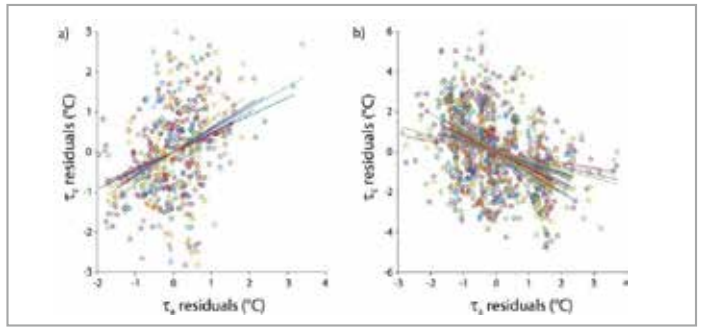

Fig. 1 - (a) Scatterplot of temperature anomalies in breeding vs wintering areas for individuals of cluster N whose migratory relevant temperature conditions in Africa explain more than 10% of variance of their European equivalent. Data referring to the same individual are denoted by a unique color: each circle represents the values of wintering and breeding temperatures anomalies for the focal individual in one of the 30-years of climatic reference for it. (b) As (a), but for cluster S.

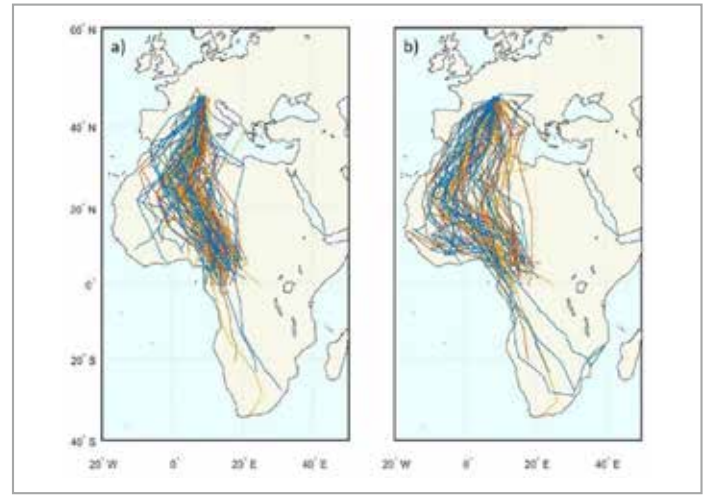

Fig. 2 - (a) Reconstruction of the 88 fall barn swallow routes estimated using FLightR. Each line connects mean positions of a single individual at each twilight. Color refers to population of origin (Blue: N, Orange: SW, Yellow: SE) (b) As (a) but for 66 spring barn swallow routes.

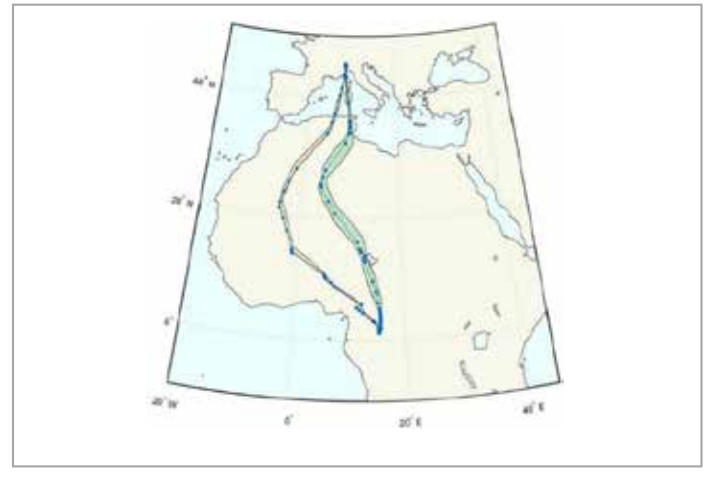

Fig. 3 - Migration paths estimated for one individual using FLightR software with twilight events classified by a Deep Neural Network. Filled polygons shows estimated route with manually classified twilight events (orange: fall migration, green: spring migration; boundaries are obtained as mean ± standard deviation of ten repetitions of FLightR).

DATA-DRIVEN METHODS FOR KNOWLEDGE DISCOVERY IN REGULOMICS

Stefano Perna

Supervisors: Prof. Stefano Ceri, Prof. Limsoon Wong

This thesis is focused on developing, testing and validating novel methods for transcription factor-transcription factor (TF-TF) interaction and coregulation prediction using predominantly data-driven models. The main result is the so-called TICA suite, which is composed of three algorithms and related web-applications: *TICA* (**T**ranscriptional **I**nteractions and **C**oregulation **A**nalyser), a novel algorithm that leverages genometric and positional information from ChIP-seq experiments targeting TF binding sites to infer interactions between two such TFs in human healthy and cancer cell lines; *NAUTICA* (**N**etwork **AU**gmented **T**ranscriptional **I**nteractions and **C**oregulation **A**nalyser), the first refinement of the TICA framework that can classify interacting TFs as either co-operating or competing based on PPI network analysis of shared interactors; and finally *ESTETICA* (**E**nrichment **S**ignal **TE**sting for **T**ranscriptional **I**nteraction and **C**oregulation **A**nalysis), a complimentary approach to TF-TF interaction classification that instead leverages on the signal enrichment values also provided by certain ChIP-seq experiments.

The common approach used by all the algorithms in the TICA suite is the following: first, given a set of many TFs in the same cell line, the null distribution is composed of TF pairs that do not interact; second, interacting TFs are distinguishable from the null case based on the relative positioning of their binding sites, i.e. the closer they are to the putative interactor's binding locations, the more likely is the interaction to be true. This translates into the use of the distance distribution tail size, a novel contribution to the field, together with more usual aggregators such as average, median, etc., to build null distributions for statistical inference.

TICA

The first model, TICA, is a novel methodology that employs genomic positional information of TF binding sites to predict physical interactions between TFs. The main advantages of TICA are three-fold: it leverages novel, parallel computing techniques to efficiently scan ChIP-seq point-sized binding site datasets and extract high-confidence binding sites and active transcription start sites; it does not require motif information for TF binding sites, bypassing incompleteness of selected motif databases and related accuracy issues; and it sports very high level of specificity even at the laxest levels of parameters, allowing investigators to screen out non-interacting TF-TF pairs with high levels of confidence before proceeding to wet lab confirmation experiments.

TICA leverages on GMQL, a novel language for the management, integration and querying of genomic information developed by the Genomic Computing (GeCo) research group at Politecnico of Milan. This language, which was created by pooling traditional distributed database techniques with computational genomics methods, supports a rich set of predicates describing distal properties of regions (e.g. being among the regions at minimal distance, possibly above a given threshold, from a given location). The development and testing process of TICA led to consequent modification and improvements of the GMQL language itself.

TICA has shown very good performance when validated with respect to curated protein complex and protein-protein interaction (PPI) network databases and outperforms competitors that require motif prediction in addition to binding site positions. TICA has shown to be as reliable if not better than similar interaction prediction algorithms that rely on motif information, while allowing for significantly higher output rates (ranging between 5000 to 22000 predictions on available cell lines). Moreover, TICA appears complementary to alternative TF-TF interaction prediction approaches (viz. TACO and CENTDIST), and combining their predictions greatly improves sensitivity at moderately reduced specificity.

NAUTICA

NAUTICA builds on the TICA framework by answering the following question: is it possible to distinguish the interacting TF pairs in co-operating and competing? One of the main limitations of TICA is that it cannot easily perform such distinction, as it can be shown that both co-operating and competing TF pairs tend to bind in close proximity to each other. One way to solve this problem is adding in an independent feature to the classifier, which is given by the number of shared interactors in a curated physical PPI network, such as the physical BioGRID subset. While co-operating TFs generally belong to the same regulatory module and thus are more likely to share many coregulators, competing TFs generally do not perform the same regulatory action when bound to the shared cognate partner and might not belong to the same module: the likelihood of them sharing many interaction partners should be much reduced.

NAUTICA classifications are confirmed by both literature investigation and protein complex databases, and the additional information extracted from BioGRID has been shown to improve TICA's predictions as well, allowing to relax its statistical thresholding on distance distribution tests and increase recall (the other main limitation of the framework). NAUTICA improves the TICA framework by leveraging protein-protein interaction network information (specifically, the number of shared interactors between two TFs in the network) for further classifying current TF-TF interaction predictions into co-operations and competitions. This classification is supported by both existing protein-complex databases and literature validation and improves the performance of TICA.

NAUTICA is a novel, effective tool for interaction classification that does not require motif prediction. To the best of our knowledge there is no method that performs wide-ranging TF interaction classification, so NAUTICA is a new contribution to the field. Several methods perform predictions on TF-TF cooperation, but those methods require TF binding motif predictions and/or knockdown experiments, making comparison difficult. Notably, the NAUTICA framework can take as input any TF-TF interaction prediction: it has been developed as an overlay for TICA, but it can easily be adapted to the usage in any pipeline which predicts and classifies TF interactions. As an example, one could use the results of CENTDIST or TACO.

ESTETICA

Finally, ESTETICA attempts to tackle the same problem as NAUTICA using the informative power of the signal enrichment found at each TF's binding sites. The higher the signal enrichment, the more copies of the protein are found at the site during the experiment and thus the stronger the binding at the target location. Co-operating TFs in general perform shared regulatory activity by recruiting each other to shared binding locations, and thus are expected to have higher joint value of signal enrichment when found in tight pairs. On the other hand, competing TFs fight for the same binding spots on the genome and/or on the trans-activating domains of the shared partners, generally in a mutually exclusive way. When one competitor is strongly binding a spot, it prevents the other from doing so. ESTETICA leverages on both the joint signal distribution and on this last postulate by separating the distribution in three equally sized sets (dubbed *HH*, *HL* and *LH*) and checking the spread of each set for significant difference.

MICROELECTRONICS AND INSTRUMENTATION FOR SINGLE-PHOTON IMAGING

Davide Portaluppi - Supervisor: Prof. Franco Zappa

In recent years, an increasing number of applications have emerged which can significantly benefit from the ability of detecting fast and faint light signals, thus requiring sensitivity down to the single-photon level and sub-nanosecond resolution in measuring the arrival time of each photon. These requirements are usually paired with the necessity of acquiring at high frame-rate, two- and three-dimensional movies of the scene under investigation. These applications range from industrial and automotive vision, such as object or obstacle recognition, road safety, distance-resolved ambient surveillance, to biomedical applications like fluorescence lifetime microscopy or time-resolved spectroscopy, study of physics of materials, and even to consumer applications such as gaming and gesture recognition.
This Ph.D. research aimed at developing an image sensor and camera system capable of acquiring such videos, employing the Time-of-Flight (TOF) technique to obtain distance information at the pixel level. The developed camera is based on SPAD (Single-Photon Avalanche Diode) detectors, which offer very high sensitivity, ruggedness, room-temperature operation, and can be fabricated through standard CMOS processes, allowing their monolithic integration with front-end and processing electronics.
In detail, this work initially focused towards designing a "smart pixel" structure capable of time-resolved detection of the incoming photons. This has been obtained by pairing the SPAD detector with a Time-to-Digital Converter (TDC) capable of a resolution of 75 ps and a full-scale range of 304 ns (equivalent to a 45 m maximum distance range and a 11 mm depth resolution). A typical drawback of "smart pixel" SPAD image sensors is the usually poor achievable fill-factor (that is, the ratio of optically sensitive area over the entire pixel area), due to the in-pixel integration of signal processing circuits, which take up silicon space. This work also focused towards mitigating this issue, by identifying the most space-consuming resources and sharing them among neighboring detectors. The final design employs a "macropixel" structure that is composed of four detectors with individual front-end electronics and photon counters (to measure light intensity), which share a single TDC without loss of spatial resolution by means of smart arbitration logic.
The design also targeted additional requirements, such as the ability of performing temporal selection of the incoming photons, in order to reject optical disturbances or other unwanted signals. To this end, the macropixel front-end provides a fast detector-enable (or "gating") functionality, capable of 250 ps transition time, which acts as a global shutter for the entire image sensor. Another accomplishment was to design a structure easily scalable toward a high number of pixels. Indeed, the developed image sensor only requires peripheral electronics to be located on two of the four edges, thus allowing an easy subdivision in quadrants; in-pixel management of data readout maximizes the pre-charge time of shared buses, allowing higher number of pixels and minimizing the silicon area required by bus drivers, at the same time providing the user with control over the type and amount of data to be transferred.

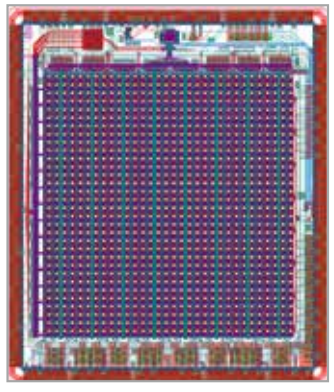

Fig. 1 - Layout of the designed SPAD image sensor.

The image sensor was fabricated in a 0.18 µm BCD (Bipolar-CMOS-DMOS) technology. Characterization shows remarkable noise performance, close to state-of-the-art for SPAD detectors, and state of art Photon Detection Efficiency (PDE) for "thin" SPAD devices, exceeding 60% PDE at 500 nm wavelength. The image sensor shows a single-shot timing accuracy of 60 ps rms (equivalent to about 9 mm depth accuracy), which can be increased through repeated distance measurements.
The image sensor was assembled on a custom designed chip carrier board, which provides user inputs for the global detector gating signal and synchronization with the active illumination source required for time-resolved photon detection. The carrier board also provides a threaded flange centered on the image sensor, intended for integration in optical setups or for mounting standard camera lenses through the use of a thread adaptor.

Fig. 2 - Photograph of the custom image sensor carrier. The sensor itself can be seen at the center of the black mounting flange.

A custom designed FPGA (Field-Programmable Gate Array) board is used to interface with the chip carrier and the image sensor itself. FPGA firmware takes care of the non-standard configuration and data interfaces and timings employed by the image sensor, as well as performing calibration and pre-processing operations on the sensor raw data. This way, the image sensor can be operated between 100'000 and 400'000 FPS (Frames Per Second) depending on the user choice of data to be output. Ultimately, the user is presented with a simple USB interface to connect the camera to a host PC.
Lastly, a user interface has been developed in LabView to control the camera system and easily perform acquisitions of 2D and 3D videos.

Fig. 3 - Photograph of the assembled prototype camera system.

CAOS: CAD AS AN ADAPTIVE OPEN-PLATFORM SERVICE FOR HIGH PERFORMANCE RECONFIGURABLE SYSTEMS

Marco Rabozzi - Supervisor: Prof. Marco D. Santambrogio

In current years we are assisting at a new era of computer architectures, in which the need for energy-efficiency is pushing towards hardware specialization and the adoption of heterogeneous systems. This trend is reflected in the High Performance Computing (HPC) domain that, in order to sustain the ever-increasing demand for performance and energy efficiency, started to embrace heterogeneity and to consider hardware accelerators such as Graphics Processing Units (GPUs), Field Programmable Gate Arrays (FPGAs) and dedicated Application-Specific Integrated Circuits (ASICs). Among the available solutions, FPGAs, thanks to their advancements, currently represent a very promising candidate, offering a compelling trade-off between efficiency and flexibility that is arguably the most beneficial. FPGA devices have also attained renewed interests in recent years as hardware accelerators within the cloud domain. Tangible examples of this are the Amazon EC2 F1 instances, which are compute instances equipped with Xilinx UltraScale+ FPGA boards. The possibility to access FPGAs as on demand resources is a key step towards the democratization of the technology and to expose it to a wide range of application domains.

Despite the potential benefits given by embracing reconfigurable hardware in both the HPC and cloud contexts, we notice that one of the main limiting factor to the widespread adoption of FPGAs is complexity in programmability as well as the effort required to port a pure software solution to an efficient hardware-software implementation targeting reconfigurable heterogeneous systems. During the past decade we have seen significant progress in High-Level Synthesis (HLS) tools which partially mitigate this issue by allowing to translate functions written in a high-level languages such as C/C++ to an hardware description language suitable for hardware synthesis. Nevertheless, current tools still require experienced users in order to achieve efficient implementations. In most cases indeed, the proposed workflows require the user to learn the usage of specific optimization directives, code rewriting techniques and, in other cases, to master domain specific languages. In addition to this, most of the available solutions focus on the acceleration of specific kernel functions and leave to the user the responsibility to explore hardware/software partitioning as well as to identify the most time-consuming functions which might benefit the most from hardware acceleration.

Furthermore, an additional degree of complexity is introduced when Partial Dynamic Reconfiguration (PDR) is employed for the target FPGA-based design. Indeed, PDR requires the designer to partition the design on different Reconfigurable Regions (RRs), define the floorplan of the RRs on the FPGA, and, finally, decide how to schedule the execution of the module on the RRs in order to minimize the reconfiguration overhead. All such decisions impact on the quality and performance of the final design, despite current vendors tools offer none or very small guidance on how to perform such tasks.

New methodologies for accelerating high-level applications through FPGA devices

A general method for translating high-level functions into FPGA-accelerated kernels has been proposed within CAOS. The core idea revolves on matching a software function with an architectural template, which is a characterization of the accelerator both in terms of its computational model and the communication with the off-chip memory. An architectural template constrains the architecture to be generated on the reconfigurable hardware and poses restrictions on the application code that can be accelerated, so that the number and types of optimizations available can be tailored for a specific type of hardware implementation. Within this context, we have integrated three architectural templates within CAOS to support algorithms with different characteristics: Master/Slave architectural template, Dataflow architectural template and Streaming architectural template.

CAD algorithms for partially reconfigurable FPGA designs

The CAOS backend has been conceived in order to support floorplanning, scheduling and mapping for PDR. In this thesis, we provide new approaches to automatically perform such steps in order to facilitate the integration of architectural templates that require the usage of PDR.

More specifically we present:

A novel floorplanning automation framework, integrated in the Xilinx toolchain, which is based on an explicit enumeration of the possible placements of each region. Moreover, we propose a genetic algorithm, enhanced with a local search strategy, to automate the floorplanning activity on the defined direct problem representation.

A new scheduling technique for partially-reconfigurable FPGA-based systems that allows to achieve high quality results in terms of overall application execution time. The proposed algorithm exploits the notion of resource efficient task implementations in order to reduce the overhead incurred by partial dynamic reconfiguration and increase the number of concurrent tasks that can be hosted on the reconfigurable logic as hardware accelerators.

An open-research platform to democratize high performance reconfigurable systems

The main objective of this thesis is the development of a platform able to guide the application developer in the implementation of efficient hardware-software solutions for high performance reconfigurable systems. The platform, dubbed as CAD as an Adaptive Open-platform Service (CAOS), targets both application developers and researches while its design has been conceived focusing on three key principles: usability, interactivity and modularity.

From a usability perspective, the platform supports application designers with low expertise on reconfigurable heterogeneous systems in quickly optimizing their code, analyzing the potential performance gain and deploying the resulting application on the target reconfigurable architecture. Nevertheless, the platform does not aim to perform the analysis and optimizations fully automatically, but instead interactively guides the users towards the design flow, providing suggestion and error reports at each stage of the process. Finally, CAOS is composed of a set of independent modules accessed by the CAOS flow manager that orchestrates the execution of the modules according to the current stage of the design flow. Each module is required to implement a set of well-defined Application Programming Interface (API) so that external researchers can easily integrate their implementations and compare them against the ones already offered by CAOS.

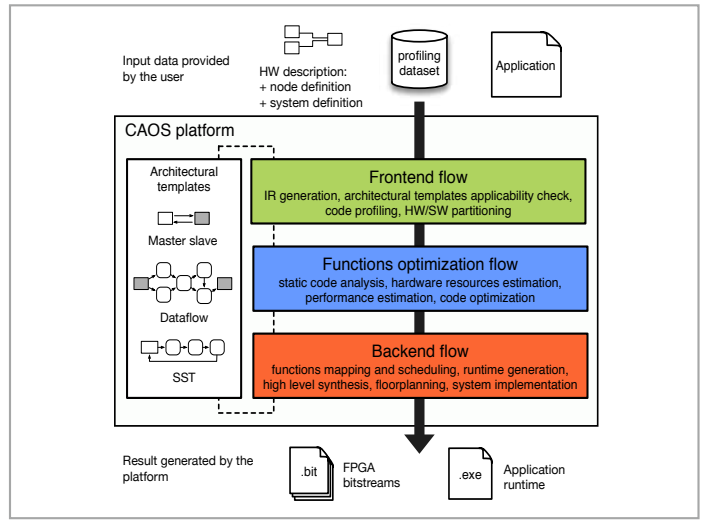

Fig. 1 - High-level overview of the CAOS platform

PERCEPTION AS BEHAVIOUR INDUCING MECHANISM: A REINFORCEMENT LEARNING PERSPECTIVE

Mirza Ramicic – Supervisor: Prof. Andrea Bonarini

Rapid advancement of machine learning makes it possible to consider large amounts of data to learn from. Learning agents may get data ranging on real intervals directly from the environment they interact with, in a process that is usually time-expensive. To improve learning and manage these data, approximated models and memory mechanisms are adopted. In most of the implementations of reinforcement learning facing this type of data, approximation is obtained by neural networks and the process of drawing information from data is mediated by a short-term memory that stores the previous experiences for additional re-learning, to speed-up the learning process, mimicking what is done by people. In this work, a multitude of techniques are presented, each of them concerned not just with data collected by the agent, but with how the feedback from its environment is encoded and managed by the learning mechanism through a mediating replay memory structure.
This opens up a possibility of implementing different replay memory architectures forming different modes of agent's artificial perception.
Given that we may have multiple optimal policies all of them achieving optimal behavior in different ways, through selecting different actions,it might be possible to influence the agent behavior by further narrowing the action selection, whilst still keeping it focused on the main goal of maximizing the reinforcement. This work exploits this possibility by modifying the agent's behavioral characteristics by changing the way the agent perceives the feedback it receives from its immediate environment; this creates an additional secondary drive that augments the main one given by the reinforcement function itself. The perception itself is not represented by the data collected by the agent but as the way the feedback from its immediate environment is encoded and managed by the learning mechanism. Influencing the way that the data is encoded we can modify the dynamics of the agent's perception and further influence its behavioral characteristics.
Environment is everything; we constantly remind ourselves of this fact when it comes to designing reinforcement learning problems where the ultimate goal is to enable an artificial agent to best adapt to its surrounding reality by learning to take better choices over time.
The perception modeling in reinforcement learning can be seen as a more than a rigid one-time feature design, but represented by a dynamic and state responsive mechanism that can by itself be capable of eliciting behavioral changes during an agent learning process. Because the agent actions depend solely on the information gained through its state representation, the dynamic processing of the information by a dynamic perceptual mechanism is bound to alter the action selection in a way that is responsive to the logic of the selection. This creates a secondary drive that complements the main meta-drive that is driven by the shear reinforcement. The learning process is still driven by the main drive of maximizing its reinforcement reward in order to reach an optimal behavior, but the way of getting to it is affected by the dynamic, subtle perception.
This is possible because the agent in any given time may have multiple options in terms of actions that are each optimal to the main reinforcement drive so the final choice of a specific action has the possibility of being driven by the perception layer instead.
The premise is that a simple difference of selecting information in the perception layer of the agent can lead to an emergence of complex and multi-layered behavioral patterns.
The process of a secondary drive driven by the perception mechanism allows for a further and more subtle modification of an agent's behavior without interfering with the primary reward maximization behavior. Similarly in humans we see different dispositions to a range of behaviors that alter they way they achieve their primary goals. These dispositions are known as emotional states and they are a product of the different physical characteristics of human brains altering the chemical compositions that govern our behavior. Our mental state, traits, mood, emotion that we are feeling affect the way we perceive reality as they alter the way the information is processed in our brains.
The experiments were performed using a specific mechanism called *cognitive filter* which differed in the selection of criteria that in term modified the sample probabilities and led to an emergence of a behavioral changing secondary drivers. Experiments in emergence of character were motivated by modeling some of the important human personality traits found in Five Factor Personality Model just by changing the perception of the agent using *cognitive filter* mechanism. First experiment used a sampling criterion of *information gain* in order to create a secondary behavioral drive that would represent the characteristics commonly associated with *Openness to Experience* personality spectrum. Second experimental setup used the same mechanism of *cognitive filter* in a multi-agent social setting and the sampling priorities were determined by the type of transition or whether it resulted in a social or exploration reinforcement. The sampling also simulated the difference in *cognitive bandwidth* that correlated with the main personality axis of Extraversion-Introversion.
The both experimental setups provided the results that have confirmed that the secondary drive effects are noticeable in the way the agents perform in the environment. The effect of a change in the agents perception induced complex behavioral patterns that significantly improved the adaptability to a variations in the environment.
Some of the traits adapted better to a variation that contained more negative reinforcement while others better adapted to the environment with scarce reinforcement that contained less amount of both positive and negative one. A social multi-agent sister where agents are able to communicate among each other bring us to the opportunity of being able to model agents that are more socially engaged and those that prefer to explore the environment without so much interaction with others. By tweaking the ratio between the two types we can achieve the intrinsic dynamics of a group social structure.
An evolutionary framework allowed us to model a perception that was best adapted to a specific environment and this showed us that the properly developed perception mechanism can greatly improve the efficiency of learning while producing complex secondary behavioral drives that complement the main one.
Some of the presented frameworks were focused on the general improvement of the reinforcement learning process rather than inducing behavioural characteristics. Experimental results using these frameworks show how can a *cognitive filter* mechanism with addition of a new criterion based on the entropy of the state improve the speed of convergence of the algorithm and another framework shows us how can an algorithm make use of the mechanism of *replay memory* in order to be able to learn complex actions that span across multiple learning steps.
The main contribution of this work is an underlying mechanism that is present across the chapters; the novel approach of modeling artificial perception in reinforcement learning by selectively filtering the agents raw cognitive stream. Experimental results show us that this seemingly simple mechanism is able to make a great difference to a reinforcement learning process by making it more dynamic and behaviorally complex. This opens up a new possibilities in the field that can be interesting to a wide range of researchers in behavioural sciences and artificial intelligence.

DEVELOPMENT AND ANALYSIS OF ALGORITHMS FOR INFORMATION-GATHERING IN AUTONOMOUS MOBILE ROBOTICS

Alessandro Riva - Supervisor: Prof. Francesco Amigoni

The gathering of information is a problem that implicitly appears in a large number of autonomous mobile robotics tasks. Some examples are exploration, coverage, and surveillance, in which the abstract concept of information assumes different meanings, e.g., representation of parts of an unknown environment or the presence of entities or objects.
In recent years, applications involving information gathering have received growing attention from both public institutions and private organizations.
The development of proper algorithmic solutions, however, is still particularly challenging, being most of the problems involved computationally hard.
In practice, several autonomous mobile robots adopt a layered architecture. Bottom levels include sensing and actuation, while the finding of a global plan to follow is demanded to higher levels.
When a task requires to gather information by means of one or more autonomous mobile robots, the high-level plan often consists of a navigation plan, i.e., a sequence of locations (or poses) to orderly reach, where information is gathered. A navigation plan is usually computed on a high-level representation of the environment, in which most of the geometrical features and the robot dynamics are neglected, and that is commonly based on grids or graphs. On this abstract model, the information-gathering tasks can be formulated as optimization problems, in which an objective has to be minimized (e.g., the completion time or the energy consumption) and some application-dependent constraints must be satisfied.
Unfortunately, most of the so-formulated problems are NP-hard and an effort has been made in the literature aimed at finding efficient sub-optimal algorithmic solutions. Although the approaches implemented can be broadly grouped in some standard classes (e.g., heuristics, greedy algorithms, linear programs relaxation), little is known about the quality of the navigation plans found. A typical tool to assess the performance of sub-optimal approaches is the analysis based on the approximation factor, that is, the ratio between the objective value of a sub-optimal solution and that of an optimal one.
In this thesis, we address two different topics related to the gathering of information by means of one or more autonomous mobile robots. The tasks addressed in these two contexts are here formalized as high-level optimization problems, for which offline algorithmic solutions are developed and analyzed. The first topic is that of coverage of a known environment using tools mounted on the robots. In their classic formulation, coverage models assume that the tools are operated by independent actions of the robots (e.g., the sensing of the surrounding environment). However, the concept of coverage can be extended to encompass a larger number of tasks, in which actions may require the joint operations of two or more robots to cover some features of the environment. A first notable example is that of the measurement of the signal strength among pairs of locations. In this broader coverage framework, the problems addressed and the results achieved are summarized as follows.
Classic coverage: a robot has to cover with a finite-range tool (e.g., a sensor or a brush) all the points of the free space of a given known environment. The underlying environment is perceived from a discrete set of points and each perception may require some time to be performed. In the literature, this problem is addressed mainly through practical solutions.
We revise some state-of-the-art approaches and we prove original approximation results for a broad class of algorithms, minimizing either the traveled distance or the completion time. For the most representative algorithm of this class, we show that the approximation factor cannot be improved. Then, we provide a novel polynomial-time algorithm, thanks to which we achieve the best-known approximation factor. Finally, we propose a practical metaheuristic (along with a polynomial-time guarantee) based on the Greedy Randomized Adaptive Search Procedure paradigm and capable of outperforming a state-of-the-art algorithm in all the considered instances.
Pairwise measurements: a team of robots have to jointly perform a given set of pairwise measurements in a given known environment. First, we prove that the minimization of either the completion time or the total traveled distance is NP-hard. Then, we provide theoretical bounds on the relation between these two objectives, showing that they cannot be always optimized at the same time. Our main contribution, however, is the development and the analysis of two approximation algorithms (one per objective), for which empirical results are also given and compared to those obtained by an exponential state-of-the-art algorithm. The approximation factors are proved to be quite tight and thus the room for future improvements is rather limited.
The second topic we consider is that of planning robot paths to reach given target locations. Differently from the typical path planning settings, we consider two scenarios in which communication constraints are imposed by practical application-dependent needs. These problems can arise in a number of information-gathering tasks, as robots deal with finite storage capabilities and it may be needed to flush data through a networked infrastructure, especially in long-term mission scenarios. Another critical aspect, emerging in multirobot tasks, is that of keeping robots connected in order to exchange information to enhance the decisions taken online. In this framework, the specific problems and the results achieved are the following.
Limited-buffer shortest paths: a robot, moving from a start to a goal location, is required to gather data along its path (e.g., a video feed in a monitoring scenario). The robot has at its disposal only a limited amount of memory and hence gathered data have to be periodically delivered to a base station exploiting the communication zones of a fixed network infrastructure. When the completion time of a path has to be minimized, we prove that the corresponding decision problem is NP-hard. We then propose a feasibility test which can tell whether a problem instance admits feasible solutions or not in polynomial time. We also devise a polynomial-time heuristic to obtain reasonable completion time and a polynomial-time algorithm to size the robot storage device. We carry out an experimental campaign to compare computational times and quality of results obtained by our heuristic versus those obtained by our original optimal algorithm.
Joint-connected paths: a team of robots moving in an environment must plan a set of start-goal joint paths which ensure global connectivity at each (discrete) time step, under some communication model. In the literature, this problem was suspected to be NP-complete, but we proved that it is actually PSPACE-complete, correcting a long-standing wrong belief. Then, practical algorithms, both deterministic and randomized, are devised and compared in an experimental campaign. Despite the hardness result, we show that many practical problem instances of realistic size can be solved within short time.
The findings contained in this thesis pave the way to a better comprehension of the criticalities inherent in tasks relative to information gathering, also highlighting the specific conditions in which existing algorithms break down (e.g., in the coverage task, when the covering-tool range enlarges, the algorithms perform worse and have weaker approximation guarantee). In order to deploy high-quality autonomous mobile robot systems, these critical issues need to be addressed and our algorithmic contribution provided an initial effort. In summary, the results achieved in this thesis contribute to further define the theoretical and algorithmic ground for the advancement of autonomous mobile robotics.

AN AUTONOMOUS NAVIGATION USE CASE FOR LAST MILE DELIVERY IN URBAN ENVIRONMENT

Stefano Sabatini - Supervisor: Prof. Sergio M. Savaresi

In the last three decades the field of autonomous mobile robot (AMR) navigation has been characterized by an extraordinary development. Different AMRs have been developed and commercialized in these years for a variety of applications. Most of the commercially available AMRs are designed either for indoor confined environments (such as industrial environments, hospital services, and buildings surveillance) or for outdoor applications in isolated and open areas like agriculture applications. The successful deployment of robots for these use cases has spurred great interest among industries in the research and development of mobile robots able to autonomously navigate far more complex environments like city centers. From the commercial point of view, the most attractive application of AMRs in urban environments is the autonomous parcel delivery. Many companies such as DHL, USPS and McKinsey have published trend reports in the last three years highlighting the big change that AMRs could bring to the delivery market, especially to solve the last-mile delivery problem. Infact, with the boom of e-commerce, people are purchasing more things online than in physical stores and they expect their goods to be delivered in a very short time to their homes. A system that automatically delivers parcels in city centers could make the last mile delivery (from the local warehouse to the final destination) not only more cost effective and efficient but it also may result in a more flexible solution (the final client could request the delivery at the exact time he is able to receive the goods). Even if it is a topic of great interest from the industrial and commercial point of view, AMR navigation in urban environments is still a big challenge from the technical point of view. Among many, three main critical requirements can be identified:
1) the need of robust localization solutions that are able to precisely provide the position of the AMR even in areas where Global Navigation Satellite Systems (GNSS) are not reliable (e.g. urban canyons or underpasses).
2) the need of extremely manoeuvrable robotic solutions in order to easily navigate even in most narrow sidewalks and a mechanical design that minimizes the possibility that the robot could get stuck in sidewalk irregularities.
3) in order to be a competitive solution, delivery AMRs should be cost effective. Up to now, reducing the cost of autonomous navigation sensors to a level that enables industrialization is an open point.

In the described framework, this dissertation presents the design, implementation and experimental evaluation of an AMR for parcel delivery (see Figure). The robot is named YAPE that stands for Your Autonomous Pony Express and has been designed by the author together with the homonym Italian start-up. The work addresses, at different levels, the aforementioned challenges that complex urban environments pose to autonomous navigation. From the mechanical design and motion control up to the mapping and localization modules, the main components of the system are designed having in mind the specific application of navigation on urban sidewalks. The chosen robot design is a self-balancing configuration that, although more difficult to control compared to common wheeled robots, carries evident advantages in terms of manoeuvrability and sidewalk roughnesses handling.

Fig. 1

The road-map for automating the delivery process envisages a gradual integration of autonomous driving functionalities. The navigation system presented here is targeted to solve a very specific use-case that constitute a realistic setting for a short term implementation of an autonomous delivery service. In fact, in a short-term horizon, AMRs will be remotely monitored with the capability to be remotely controlled by a human operator from an ad-hoc control station. When a delivery request is generated, the system will check if the request concerns an area that has never been explored by the robot. If this is the case, the delivery will be performed remotely controlling the robot from the control station. While the robot is remotely operated, all the sensors data are recorded and stored. These data are used in post-processing to build a map of the new area and extract relevant information (e.g. sidewalk width, sidewalk condition etc..). When a new delivery request is generated in the same area, the AMR is designed to autonomously perform the delivery leveraging on the information extracted in the previously remotely controlled operation. According to this paradigm, in this dissertation a target route is chosen as a delivery use-case. The route connects the Departments of Electronics at PoliMI and a supermarket in Milan. In the first place, the route is travelled remotely controlling the robot. Sensor measurements are acquired and a map of the urban environment along the route is constructed using state of the art Graph SLAM techniques. In a second moment, the robot is requested to travel the same route autonomously using a map-based navigation architecture (see Figure).

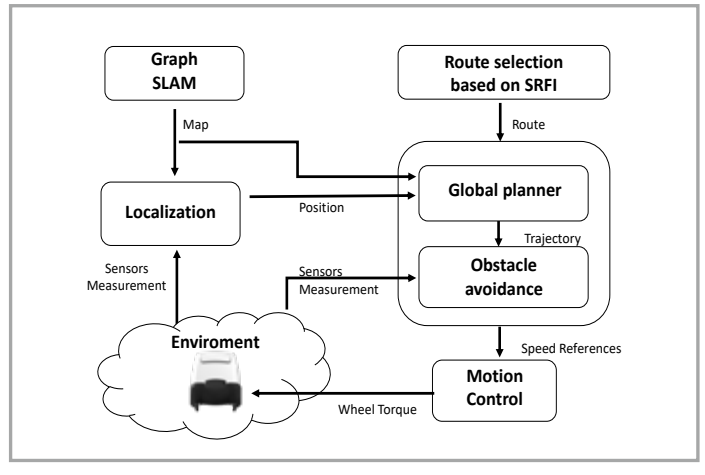

Fig. 2

A MODEL-CENTERED SOLUTION FOR TAMING THE HETEROGENEITY OF SMART DEVICES

Mersedeh Sadeghi - Supervisor: Prof. Luciano Baresi

Recent progress of Internet of Things (IoT) technologies has led to a competitive market and heterogeneity of communication protocols, large diversity of smart device types, and multitude of widely used Application Programming Interfaces (API) and standards. Even if, these advancements enhance the development of IoT domains including smart spaces, they intensify the interoperability challenge. In other words, the co-operability of smart devices is compromised by the numerous enabling technologies. Smart spaces are becoming more tightly bounded to the technology and standard proffered by their underlying frameworks.
Thus, the essential obligation to construct an effective ecosystem of smart devices is the interoperability --that is seamless cooperation and operation-- among the aforementioned enabling technologies, platform and standardization. Furthermore, the last link of the seamless interoperability chain is the integration of front-application to the smart environment. The ultimate goal of any smart space framework is to provide means for the end users to interact with surrounding smart devices which is achieved through visualization of capability and function of those devices. This requirement, hence, demonstrates a critical importance of a robust and preferably automated approach to creating GUI for a heterogeneous of smart objects. Finally, as a result of outspread of diverse smart devices with an affordable cost for end-users, the construction of a smart environment become more easy and omnipresent. In such environment, the intelligent and situation-aware systems pervasively collect data from smart objects to reason upon and respond to different needs of users. Inevitably, the smart devices themselves are a source of highly confidential information. In addition, they most often offer direct interaction with and manipulation of physical world. In this context, not only an unsupervised controlling of a device grounds adversely affects but even an unauthorized read-only access may disclose highly confidential data of user's lifestyle. Such high involvements of smart technologies with the fabric of everyday life of human meant to be utilitarian and highly advantageous but poses new security and privacy challenges.
This study started with a meticulous and critical analysis of the state of the art to identify the major complications of today IoT/WoT systems in different layers and aspects. The current literature tackles this issue within bottom layers of IoT stack, and, predominantly toward establishment of unified programming interface and communication means. Nevertheless, the harmonization of the interfaces to exchange data does not fully address the problem. In fact, a seamless integration and comprehensive interoperability do not accomplish if the exchanged data remain unstandardized and diversified. In this direction, although ontologies and shared vocabularies had promised the interoperability at the semantic level, yet, the upsurge of a competing domain-dependent and vendor-specific ontologies stirrers up the diversity trouble.
In the direction, to address above mention issues, we presented a unique model, called TDeX, to support the uniform description of diverse smart devices. The model stems from W3C standardization and extends their proposed TD (Thing Description), to support a uniform representation of diverse smart devices and manage the relevant contextual aspects together with the seeds for their visualization in the front-end application. TDeX thus does not only confine to a unified functional description of a device, but it covers other relevant concepts that constitutes the core building blocks of a ubiquitous computing. It thus paves the ground to make two more contributions: an innovative permission model for smart devices and an automated GUI generation for interaction with them.
In this thesis, we described a context-aware access control mechanism for smart devices which fulfills the emergent requirements of today pervasive environment. It is based on Attribute Based Access Control (ABAC) methodology; Our solution adapts this paradigm for IoT-based system and extends it to support a generic and extensible context model which makes it flexible and applicable for a wide spectrum of smart environment application domain. It has an application-oriented design supporting a generic and extensible context domain which makes it flexible and applicable for a wide range of IoT application domain. Moreover, the target users of access control system for smart spaces mostly includes non-expert people who desire to easily manage their own smart devices. The proposed framework offers an intuitive permission model and process for an owner to assert the required protection on his/her device. Finally, to have an effective access control model appropriate for advanced IoT device, the protection must be applied at device's function level. The granularity of proposed system affords an ultimate control over each and every aspect of a smart device.
Another pivotal challenge to envision a seamless pervasive application for IoT is the graphical user interfaces (GUI) generation for heterogeneous smart device. In this thesis, a model-based automated solution for the creation of (basic) GUIs for smart devices is proposed. The procedure is platform- and device-agnostic. Former enables the model to be rendered into many diverse concrete solutions (from web-based ones to the layouts of Android activities). Latter leads to a loosely couple visualization process which enables developers to generate GUI for multitude of heterogeneous smart devices on the fly without any prior application-device binding. Moreover, incorporation of active contexts of the use in this procedure resulted in, self-adaptive and context-aware GUIs, based on user permissions.
Finally, to encapsulate above mentioned solutions within a stand-alone and complete framework for management of smart space, this thesis also introduced a supporting middleware, M4HSD. It is A RESTful middleware that exploits the aforementioned model to abstract heterogeneous devices through standardized and homogeneous REST APIs and to support a secure and context-aware interaction with them. Special-purpose drivers and plug-ins harmonize peculiarities of different standards and data models into our unified one, TDeX. M4HSD thus envisions the syntactic interoperability in addition to network interoperability. Due to unified representation of heterogeneous devices achieved through TDeX, interactions with any device are generic and device agnostic. Hence, our framework decouples the front-end application from underlying devices and enables developer to write an application without any prior knowledge of functionality, type and API of specific devices.
Through this research study, we were following the incremental software development. It fragments the whole framework to smaller development phases and creates an evolutionary prototyping of each phase as we had been continued to build up the final completed software. Each phase of the project was critically tested, evaluated and enhanced accordingly. The feasibility, validity and effectiveness of the overall solution have been evaluated through several prototypes and well-designed case studies/ experiments.

OPTIMIZING DATA-INTENSIVE APPLICATIONS FOR MODERN HARDWARE PLATFORMS

Alberto Scolari - Supervisor: Prof. Marco D. Santambrogio

Data-intensive applications have become widespread in the years, especially in cloud-like environments. Among them, Data Analytics (DA) and Machine Learning (ML) applications are particularly important categories that deeply impacted business and science in the last decade, and are expected to have an even higher impact in the upcoming years.
In the latest years, we also saw hardware platforms evolving along different directions to overcome the limitations of Dennard's scaling and the end of Moore's law. While heterogeneity is coming into play for several applications, Central Processing Units (CPUs) have also evolved towards a growing number of cores, specialized Single Instruction, Multiple Data (SIMD) units, high memory hierarchies and, in general, a more complex and diverse set of features. On the other side, also data-intensive applications became more complex, to the extent that a current ML model may comprise tens of diverse operators to compute a single prediction value, while taking in input data of multiple types like text, number vectors and images. Oftentimes these applications are structured as "data pipelines" and go through many steps like input parsing, data pre-processing, analysis and possibly loading the output to some "data sink" at the end. Mastering this complexity to achieve the best implementation and deployment of an application in a given setting (hardware platform, software stack, co-located applications, etc.) is becoming a key issue to achieve best usage of the infrastructure and cost-effectiveness. This problem is especially hard with heterogeneous platforms, whose hardware features may not suit some parts of an application to be accelerated or may require a long redesign effort. Here, where the inevitable complexity of applications determines the diversity of operations, CPUs are still central, as they provide the flexibility and the maturity to efficiently run most of these workloads with sufficient performance even for today's needs, while being easier to program than other architectures. Moreover, their general-purpose design naturally fits the diversity of data-intensive applications.
This work explores the performance headroom that lies unused in modern CPUs with data-intensive applications. This headroom encompasses several dimensions, each one with specific problems and solutions. A first problem to solve is the performance isolation of co-located applications on the CPU, which has to take in account the sharing of the Last Level Cache (LLC). Here, we propose and evaluate a mechanism for partitioning the LLC that works on recent server-like CPUs and requires software-only modifications in the Operating System (OS), thus not impacting hardware nor applications. This solution proves to be effective with a diverse set of benchmarks and allows meeting Quality of Service (QoS) goals even in a contentious, hardly predictable environment. The second problem is the optimization of data-intensive applications, which can be composed of multiple, diverse computational kernels.
This work explores the limitations of current solutions, revolving around a black-box approach: application kernels are individually optimized and run, disregarding their characteristics and their sequence along the data path; indeed, a simple case study shows that even a manual, naïve solution can achieve noticeable speedups with respect to the current state of the art. Building on these findings, we generalize them into a white-box approach for applications optimization: while applications should be written as sequences of high-level operators, as current development frameworks already do, this sequence should also be exposed to the system where these applications run. By looking at this structure during the deployment of the application, the system can optimize it in an end-to-end fashion, tailoring the implementation to the specific sequence of operators, to the hardware characteristics and to the overall system characteristics, and running it with the most appropriate settings; in this way the application can make the best use of the CPU and provide higher QoS.
Such a re-thinking of current systems and frameworks towards a white-box also allows a cleaner support of heterogeneous accelerators. Indeed, the high-level description that we advocate allows the system to transparently map some operations to more specialized accelerators if need be. However, optimized solutions need to keep a sufficient degree of flexibility to cover the diversity of computational kernels.
As an example, we explore a case study around Regular Expression (RE) matching, which is a ubiquitous kernel in data-intensive applications with limited performance on CPUs, and we propose an architecture that enhances previous work in terms of performance and flexibility, making it a good candidate for the integration with existing frameworks. Overall, this work proposes several solutions for the main issues around modern CPUs and data-intensive applications, breaking some common abstractions and advocating for an appropriate description level of those applications. The solutions proposed here leverage this level of description that enables various optimizations, providing novel guidelines in order to make the best use of the architecture.
From this work, several research directions arise, especially around extending these abstractions and the related solutions to work with heterogeneous devices, whose usage for the masses calls for more automated optimization strategies and prediction models.

INCREASING CAPACITY OF WIRELESS NETWORKS THROUGH AERIAL BASE STATIONS

Navuday Sharma - Supervisor: Prof. Maurizio Magarini

In the past decade, due to immense high speed data and wider connectivity requirements, the cellular technologies have been continuously evolving leading to a major revolution in telecommunication industry. Currently, under IMT 2020, commonly known as 5th Generation of Cellular Technology, the targeted applications have been broadly classified as: enhanced mobile broadband (eMBB), massive machine type communication (mMTC), ultra-reliable and low latency communications (uRLLC), vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2X).

In order to address the increasing data requirements, particularly, during flash crowds such as concerts, rallies, festivals, sport events etc, where many people gathered around in an area, use data services such as video streaming, photo sharing, video calls etc, higher densification of terrestrial network architecture is gaining immense importance in the name of Ultra-Dense Networks (UDNs), as shown in Fig. 1. The data traffic increases manifold due to development of high resolution and big screen smart devices such as phones, tablets, laptops with 4K resolution, which eventually demands for higher data.

Recently, Unmanned Aerial Vehicles (UAVs), generally known as drones were started to be investigated to provide data service to the users in suburban areas. Such projects were initially started by Google and Facebook to develop solar powered drones providing internet services. Later, companies such as Qualcomm, China mobile and Nokia started to investigate on other aspects such as controlling the drones through LTE base stations, cooperative communication among the drone network etc. Successful handovers were reported by Qualcomm with zero link failure in autonomous drone control through LTE network. Similar project was conducted by Ericsson and China mobile with deployment of prototype in field trial. Also, with techniques such as swarm optimization and collision avoidance algorithms, the drone network could be deployed in the city environment as well, as reported by Nokia.

With such interest from industry and academia, research on UAVs acting as Aerial Base Stations (ABS) attained immediate attention. Therefore, this thesis mainly discusses about several technological aspects pertaining to such a system. Although, there are many research directions to the development of an ABS network, here we address certain directions. The work in this thesis, initially starts with the Air-to-Ground (A2G) Channel Modeling. There are many A2G channel models existing in the literature but since the work done here is to provide cellular network by UAVs, Low Altitude Aerial Platforms (LAPs) were preferred up to the altitude of 2000 m. The existing channel models mostly deal with commercial or military aircrafts, which fly at very high speeds near to subsonic or supersonic ranges. Such high speeds are not preferred for an ABS network to avoid frequent handovers. Also, some researchers have argued about the effect of Doppler shift at higher speeds which would finally lower the performance of the system. However, from the analysis and simulations provided in this thesis, effect of Doppler was not much observed with A2G channel and implementation of 5G waveforms. Due to unavailability of channel parameters for LAPs, measurements were performed using a radio propagation simulator for different environments: Suburban, Urban and Urban High Rise. These environments were generic since they were developed using ITU-R parameters for the simulator. Therefore, results from this simulator could be applied to practical environments with small degree of inaccuracy. Further, the work on cell coverage, capacity and interference analysis was conducted with simulation results obtained from the ray tracing simulator and verification of these results were done by performing analytical analysis and obtaining closed-form expressions. The graphs plotted using these expressions, matched with the graphs obtained from simulations with same simulation parameters. Later, in this thesis a new system was proposed as an optimal replacement to UDNs, to support the flash crowds. This system was termed as Ultra-Dense Cloud Drone Network (UDCDN), as seen in Fig 1. This system is advantageous as it offers reduction in Total Cost of Ownership (TCO) as compared to UDNs. Also, UDCDN is on-demand deployment system, i.e. it is deployed by the mobile operator only when required based on the data traffic information obtained from the cloud. Further, in the thesis, work has been done on implementing parameters for Ultra-Reliable Low Latency Communication (uRLLC) of 5G Physical Layer (PHY) on the ABS network to provide reliable and faster connectivity for ground users. Symbol Error Rate (SER) improvements were seen when uRLLC was implemented for A2G channel with Generalized Frequency Division Multiplexing (GFDM) modulation. Further, results were also provided by introducing Carrier Frequency Offset (CFO) in the system model.

Following the above work, other aspects were also considered in the thesis. Apart from primarily ABS network, a heterogeneous network (HetNet), consisting of multi-tier drone and terrestrial network, is seen as a more feasible options for the present cellular network. Therefore, further this thesis discusses about three major aspects which optimize the multi-tier ABS network: survivability, coverage and mobility laws to avoid issues related to inter-cell interference, frequent handovers, power deficiency etc. Later, work has also been done on interference alignment (IA) for maximizing the sum rate. However, IA demands for independent channels to provide better efficiency but at LAPs obtaining independent channels seems improbable. Therefore, such study serves better for a High Altitude Aerial Platform (HAP).

An optimal receiver separation distance was also defined for the system.

Another set of work has been performed in this thesis, which formed as a minor project apart from the major one reported throughout. The aim of this project was to develop an Internet of Thing (IoT) based Health and Usage Monitoring Systems (HUMS) for a helicopter. HUMS is an integrated recording and monitoring system that includes sensors, data acquisition technology and software algorithms (both on-board and ground-based) that are provided as a unit with the goals of reducing maintenance costs and improving safety and availability. For this system, the goal was to deploy various sensors all throughout the aircraft for monitoring different avionics components health and lifecycle. This data was to be send to a common gateway, known as a Transmission Data Concentrator (TDC) for data aggregations and processing. This processed data was sent to the cloud using cellular network when the aircraft is in the range of terrestrial network, otherwise using the satellite network.

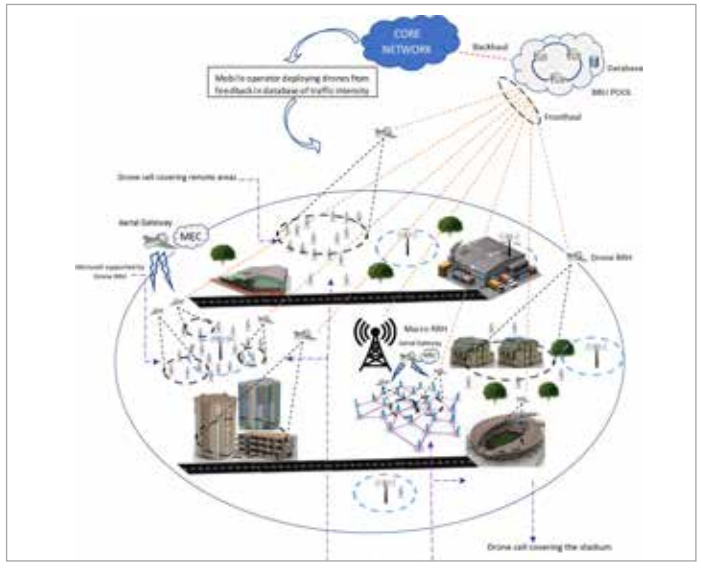

Fig. 1 - Ultra-Dense Cloud Drone Network system architecture

ENERGY EFFICIENCY AND SURVIVABILITY IN 5G CENTRALIZED ACCESS NETWORKS

Mohamed Shehata - Supervisor: Prof. Massimo Tornatore

The continuous demand for better wireless data services in terms of very high data rates (typically of Gbps order), extremely low latency, and significant improvement in users' perceived Quality-of-Service, has triggered the research on the fifth generation (5G) wireless systems that are expected to be deployed beyond 2020. Maintaining the current network architecture will lead to an unsustainable network-cost increase as well as to a dramatic expansion in the network power consumption. Hence, minimization of network cost and energy consumption have become a necessity for mobile network operators. In order to do so, the network infrastructure has to evolve from the old static architecture towards a more scalable, dynamic and agile one by resorting to novel technologies and architectural solutions to improve cost and energy efficiency.
In a traditional Distributed Radio Access Network (D-RAN), the Base Station (BS) comprises two modules, i) the Remote Radio Head (RRH) for transmission and reception of radio signals, Digital-to-Analog/Analog-to-Digital Conversion (DAC/ADC) of the baseband signals, frequency up/downconversion and power amplification, and ii) the Baseband Unit (BBU) performing the digital processing functions of layer 1, 2 and 3 (L1, L2, L3). As shown in Fig. 1(a) every BS hosts its "local BBU" and has a dedicated housing facility, which is not shared with other BSs. Hence, in D-RAN, power consumption as well as investment and maintenance costs increase linearly with the number of BSs. Given the rapid traffic growth envisioned for the next years, simply increasing BSs density in D-RAN does not represent a scalable solution.
A novel network architecture, called Centralized Radio Access Network (C-RAN), has been proposed as a more scalable alternative to D-RAN in terms of both power and cost efficiency. The main idea of C-RAN is that multiple BBUs are placed in a single physical location (BBU pool), which is connected to several RRHs through a high capacity "fronthaul" network, as shown in Fig.1(b). Thanks to this centralization, the baseband resources in the BBU pool can also be virtualized and shared among several BSs, and significant reduction in the overall computational resources can be achieved due to multiplexing gain. BBU centralization also allows to share maintenance costs and power consumption among several BSs and promotes the utilization of advanced interference-cancellation techniques such as the Coordinated Multipoint (CoMP).

In this thesis, we investigate the opportunities enabled by C-RAN. First we provide a quick survey on the C-RAN stat of art. Then, we model the computational savings (what we called multiplexing gain) enabled by C-RAN under four different functional splits. Furthermore, we show the cost savings arises from centralization. To estimate the power savings -resulting from reduction in the computational resources- for the various cases, we identify the main power consumption contributors in a BS and provide a power consumption model for the different RAN split options. The results show the cost and energy advantages of C-RAN with respect to classical RANs, due to "centralization" of BBUs into a few sites. Moreover, the results give insights on the latency and the transport network capacity on the BBU placement problem.

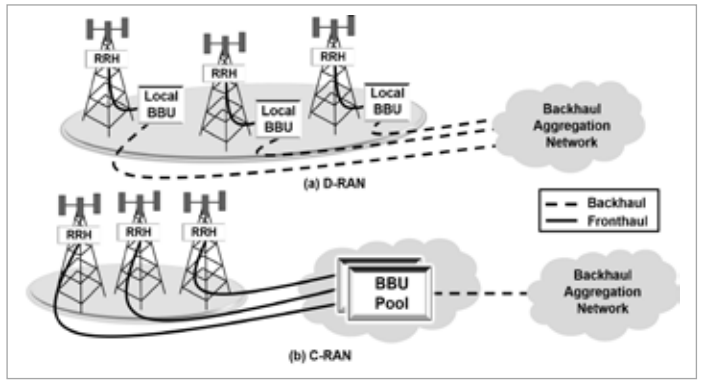

Fig. 1 - Distributed RAN VS Centralized RAN.

Following this centralization savings, we design a survivable C-RAN against BBU pool and link failures. We propose three different approaches for the survivable BBU pool placement problem and traffic routing in C-RAN deployment over a 5G optical aggregation network. We formalize different protection scenarios as Integer linear Programming (ILPs) problems. The ILPs objectives are to minimize the number of BBU pools, the number of used wavelengths and the baseband processing computational resources. Finally, we design survivable C-RAN based on shared path protection schemes with the objective to minimize the number of BBU pools and the number of used wavelengths.

VISIBLE LIGHT COMMUNICATIONS FOR NEXT GENERATION IN-FLIGHT SYSTEMS

Dario Tagliaferri - Supervisor: Prof. Carlo Capsoni

In recent years, aircraft passengers started complaining for the lack of a broadband Internet connectivity during flights. The provision of the required bandwidth on-board an aircraft is however challenging, for the limitations on the permissible R.F. interference, which is spuriously generated by user's electronic devices and has a detrimental effect on the avionic equipment. Despite some expensive solutions are nowadays in use, based on both cognitive Wi-Fi for on-board connectivity with users' equipment and satellite communications for delivering the aggregate bandwidth to the airplane, are currently not able to deliver enough bandwidth to the users. In order to avoid any R.F. interference inside the cabin, researchers started considering Optical Wireless Communications (OWC) as a viable solution to provide the required Internet connectivity to the passengers, either by deploying ad-hoc IR cellular-like systems or by exploiting the available reading lights to establish point-to-point Visible Light Communications (VLC) links with the users' terminals (Fig. 1). While the first option shares some of the drawbacks of the Wi-Fi-based systems, especially the extra costs and weight for deploying the system, the second solution reuses the same luminaries already present for illumination and is intrinsically not subject to the stringent eye-safety regulations owned by IR radiation. In parallel, a novel concept of airborne mesh networking was proposed to overcome the capacity limitation of satellite communications and provide the aircrafts with enough bandwidth to enable in-flight connectivity. The driver element is the availability of multi-Gb/s Free-Space Optics (FSO) links to connect the airplanes one to the other, relaying the signal received from ground, which allows to extend the coverage all over the world. Although airborne mesh networks are currently pioneering, few experimentations were given to demonstrate their feasibility and theyare expected to be implemented in the future.
The thesis core is the application of OWC inside the passengers' cabin of an Airbus A320. The study of a hybrid, bidirectional VLC (for the downlink)/IRC (for the uplink) link to the backbone network is addressed first, where the OWC channel is investigated with a simulative approach, characterizing the Channel Impulse Response (CIR) and the Signal-to-Interference-plus-Noise Ratio (SINR) for various configurations of sources and receivers. Constraints are on the complexity and on the integrability of the system in both the avionic environment and the users' devices, which translate into reduced size, weight, power consumption and overall costs. Consequently, the study is targeted to find novel solutions and results by considering only off-the-shelf devices. The thesis asserts that the VLC channel allows for high capacities with reduced levels of emitted light (compared to standard values) when considering common high-directional reading lamps. Enabling VLC in low-light conditions allows for a power consumption reduction still providing enough illumination when using a backlighted terminal. The limited passengers' mobility when seated and the consequent slow variability in time of the on-board VLC channel paves the way to the application of a Tomlinson-Harashima Precoding (THP) to operate at the transmitter, to remove the MLI among the passengers belonging either to the same seat group or to the same seat row. THP allows to retain the same modulation bandwidth of each LED (because no frequency reuse scheme is needed) while avoiding to add extra complexity at the receiving side. The thesis provides and discusses the optimum value of emission angle of the reading lamps as a trade-off of between the robustness against source-to-receiver misalignments (when users change the position of their terminals within the light cone) and the achievable performance, which are found to be comparable with the one pertaining to the latest release of 5G standardization. In addition, the thesis contains the design of a reduced complexity DCO-OFDM scheme to best suit the favourable on-board channel, where the only issue is represented by the Sampling Clock Offset (SCO) between transmitter and receiver. In the proposed OFDM system, the SCO is avoided without the use of any feedback loop, allowing for a noticeable simplification of the system. A novel wireless IRC system is also proposed, in order to enable the on-board uplink to request the desired multimedia content. Each passenger belonging to the same seat group is scheduled in a Time Division Multiple Access (TDMA) frame to avoid the mutual interference and connect with a single receiver placed between the reading lights. The design considers using low power and wide-angle transmitters to allow the transmitter-to-receiver alignment when the users reasonably change the position and orientation of their devices, and various types of receivers to choose the best performing one. In order to efficiently interconnect the reading lamps one to the other and with the external communication network (through the gateway), the thesis discusses three possibilities for the backbone network, respectively based on Power Line Communications (PLC), optical fibers and FSO, which is proposed for the first time. The study is addressed by taking advantage of the on-board avionic requirements, in terms of weight reduction, low power consumption and no R.F. emission, and the communication system requirements, which are set in order to be able to provide in-flight entertainment (enabling the download of the latest multimedia services). The thesis demonstrates the inadequate nature of the PLC solution for serving as a high-capacity backbone and compares the optical fiber- and FSO-based solutions in terms of achivable capacity, scalability and weight/energy reduction.
As last, the thesis work focuses on the design, realization and testing of a proof-of-concept VLC transceiver to serve both as a testing platform for an On-Off Keying (OOK) link and as a starting point for developing a more advanced system.

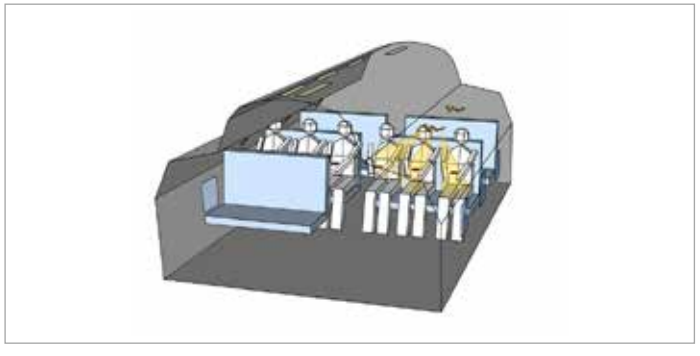

Fig. 1 - Sketch of a downlink VLC in a typical passengers' cabin.

BUILDING STRUCTURED HIERARCHICAL AGENTS

Davide Tateo - Supervisor: Prof. Andrea Bonarini
Co-Supervisor: Prof. Marcello Restelli

In modern real-world applications, autonomous agents are required to solve very complex tasks, using information taken from low-level sensors, in uncontrolled, dangerous, and unknown scenarios.
Among them, robotics systems share some common characteristics: most of these systems use continuous state and action variables that may need a fine grain precision.
They may exhibit different dynamics between different parts of the system, leading to a natural division based on different abstraction levels. Finally, some tasks are even difficult to formalize in the framework of Reinforcement Learning, making difficult to define a reward function, while some human (or non-human) experts may be able to provide behavioral demonstrations.
Based on these considerations, we propose two approaches to improve the applicability of Reinforcement Learning techniques in these scenarios: a new Hierarchical Reinforcement Learning framework based on the Control Theory framework, which is particularly well suited for robotics systems, and a family of Inverse Reinforcement Learning algorithms that are able to learn a suitable reward function for tasks (or subtasks) difficult to formalize as a reward function, particularly when the demonstrations come from a set of different suboptimal experts. Our proposals make it possible to easily design a complex hierarchical control structure and learn the policy both by interacting directly with the environment or providing demonstrations for some subtasks or for the whole system.

Most of the current Hierarchical Reinforcement Learning approaches are based on the concept of subtasks: the original task is decomposed in a set of smaller tasks.
The decomposition can be done using a hierarchical structure such as the task graph of the MAX-Q algorithm, or by creating structured policies as in the option or in the Hierarchy of Abstract Machines frameworks. These approaches work particularly well when dealing with MDPs with finite action and state space. However, they are really difficult to adapt to more complex scenarios, particularly when dealing with complex robotics systems. The main issues are that not all the systems can be modeled easily with a simple hierarchical structure, such as the one proposed by the aforementioned methods, particularly when we want to exploit particular aspects of the environment such as symmetry or translation-invariance. Furthermore, in most cases, it is not possible to write a generic hierarchical algorithm, but each task must be solved by a particularly hand-made algorithm, that exploits all the relevant characteristics of the environment and encodes the prior knowledge in the learning process.

One of the fundamental building blocks of control theory is the block diagram. Block diagrams can model complex systems composed by plants, sensors, controllers, actuators and signals. With block diagrams, it is easy to describe the control architecture of any kind of control system, from the simple control of an electrical motor to a complex industrial plant. The block diagram is based on blocks and connections. A block can represent a dynamical system or a function, and a connection represents the flow of the input and output signals from and towards other blocks.
Our approach is based on the same core idea, with some modifications in order to fully describe a Hierarchical Reinforcement Learning system. We believe that such representation is beneficial in general for a Hierarchical Reinforcement Learning system, but in particular for robotics applications, where the control system is often already structured (totally or partially) as a block diagram. We will call this approach Hierarchical Control Graph Learning.

Inverse Reinforcement Learning is a powerful tool that can be useful in practical applications in general, but in particular for hierarchical systems, when it is difficult to design a reward function for a sub-task, or when the reward function provided is not sufficiently informative e.g. when it is a step constant cost, or when the reward is sparse i.e. when the only reward value different than zero is either for task success or failure. In this scenarios, Inverse Reinforcement Learning can be an alternative method to shape the reward, as the reward function is optimal w.r.t. the desired behavior.
We focus on reward recovery instead of behavior cloning, as learning a reward function is more general: with a reward function, it is possible to learn the optimal policy even if the underlying environment dynamics changes w.r.t. the environment used by the experts. This is also useful when a sub-task needs to be solved in different parts of the state space, where the dynamics, and thus the experts' policy may change. Also, sometimes the imitator may not be able to exactly reproduce the behavior of a human demonstrator but can share the same goals, e.g. a humanoid robot.

To improve both the applicability and the performance of Inverse Reinforcement Learning methods, we focus on the multiple expert scenario. We assume that the experts share the same reward function, but their policies may be different due to suboptimal learning. We call this algorithm Single Objective-Multiple Expert Inverse Reinforcement Learning.

This work presents a set of techniques to face one of the most important issues in Reinforcement Learning: the design of hierarchical agents. Hierarchical agents are an extremely important approach to reduce the gap between Reinforcement Learning research and practical industrial applications.
The computation of (internal) reward signals is a key part of our framework: autonomous systems should be able to compute the intrinsic reward signals using just the information available from the measures of their sensors, and this is the main reason why the computation of intrinsic reward functions is tightly coupled with the interaction of the control system with the environment.
As the design of the reward function is one of the key elements of hierarchical agents, we focused on Inverse Reinforcement Learning, with the main objective of retrieving the reward function, instead of learning the expert policy from demonstrations.
With Inverse Reinforcement Learning, it is possible to automatically encode the expert knowledge about a task, or a subtask, into a reward function that specifies what is the objective the expert is trying to maximize. The exploitation of prior knowledge, both by manually inserting it in the system, or by inferring it automatically from the demonstration, is one of the key methods to face complex tasks, and the literature often lacks on this aspect as, at least in recent years, the focus has shifted towards completely automatic learning systems.
We believe that it is not reasonable, at the present time, to run the state of the art algorithms in most real-world scenarios: not only the amount of data required to learn is not compatible with real-world applications, but also the instability of the learning process, the possibility of undesired behavior in unexplored regions of the state space and the amount of preprocessing and implementation details to make the learning work, can have catastrophic consequences.

While in this work we do not provide a complete and definitive solution to any of the problems highlighted in the literature and above, we have put the focus on one of the possible ways towards the application of Reinforcement Learning in the real world: the structure of the agent.

STOCHASTIC METHODS FOR PERFORMANCE PREDICTION OF PHOTONIC INTEGRATED CIRCUITS

Abi Waqas - Supervisor: Prof. Andrea Ivano Melloni

Integrated photonic is the field where optical devices (such as couplers, interferometers, and so on) can be incorporated on a chip by the use of dielectric waveguides. Technology progress in integrated photonic has resulted in the implementation of complex photonic circuits combining many functions on a single chip, significant production volumes and reduced fabrication costs. While standard fabrication technologies are an essential condition for the commercial exploitation of photonic, they still have to face an unavoidable reality of uncertainties. As photonic devices are much longer when compare to wavelength, a slight variation in device geometry can cause a dramatic phase error, especially for devices based on interferometers such as Mach-Zehnder and micro-ring resonators. Each fabrication run is subject to several possible variations (waveguide width or height deviation, improper gap opening, change in material composition and surface roughness) that may eventually cause a fabrication deviations that reduce the yield at too low levels to be economically sustainable. This means that several realized devices, which are designed to be nominally the same, differ from one to another due to variations of fabrication process. As a result, the device response is no longer considered as deterministic but is more suitably interpreted as a stochastic process and the analysis of photonic circuit is incomplete without the inclusion of stochastic analysis due to the presence of unavoidable uncertainties. To obtain a design of a photonic circuit with high yield and reduce performance variation, it is essential to consider such uncertainties in process design kits and to isolate the most critical parameters of the circuits and to estimate and reduce the cost of post-fabrication correction of the process variability. Therefore, statistical data and efficient statistical tools to include this data to predict the statistical behaviour of the final circuit are becoming fundamental instruments in photonic.

In this doctoral dissertation, we focus on exploration and development of statistical tools to predict the performance of the integrated photonic circuit in the presence of unavoidable fabrication uncertainties. To address the problem of post-fabrication correction and to isolate the most critical parameters of the circuit, advanced sensitivity analysis methods were studied and applied. In the presence of stochastic uncertainties, sensitivity analysis method answers the questions such as which of the considered uncertain parameters is more important in determining the uncertainty in the circuit response. Or if it is possible to compensate the uncertainty in one of the considered uncertain parameter, which factor should we choose to reduce the most of the output variability. We have introduced variance-based sensitivity analysis method to investigate the behavior of a photonic circuit under fabrication uncertainties, with the aim to identify the most critical parameters affecting device performances. The comparison with other sensitivity analysis method namely elementary effect test (also known as Morris method) is also presented that is radically different in both the methodology and underlying assumptions to assess the credibility of the obtained results. We have demonstrated, for the first time, the possibility to use the results of these methods to estimate the post-fabrication correction (e.g. thermal tuning) cost and to reduce the power consumption for the mitigation of statistical variation of circuits' parameters and increase the yield.

In the case of traditional Monte Carlo analysis, thousands of repeated simulations are required to study the effect of stochastic uncertainties on photonic circuit. Even if we use more complex techniques based on spectral methods to represent uncertainty, we still normally need to run several simulations to sample the device uncertain response [see Fig. 1(a)]. In this work, we have subverted this approach by proposing the framework [see Fig. 1(b)] that can obtain a full description of the behaviour of a circuit under stochastic uncertainties with a single deterministic simulation, saving time and computational resources. To include information on the effect of fabrication uncertainties in each building block, we propose a Building-Block-based Generalized Polynomial Chaos method (BB-gPC) by properly exploiting stochastic collocation and Galerkin method to realize a completely novel class of device models for the preparation of stochastic process design kits. Using these proposed stochastic macro-models that inherently convey stochastic information, only a single deterministic simulation is required to compute the statistical features of any arbitrary photonic circuit, without the need of running a large number of time-consuming circuit simulations thereby dramatically improving simulation efficiency. The BB models, in the form of transmission or scattering matrices, are circuit independent and can be stored and replace the original deterministic macro-model of the building blocks in the process design kit. The new matrices can hence be combined according to the building blocks connections to derive with a single run of the deterministic circuit simulator the stochastic behaviour of any circuit, enabling an unprecedented simulation efficiency. The stochastic properties of the BB can reflect for example the foundry technological process and they are embedded in the PDK and should not be recalculated for every circuit. Also, the effect of spatial correlation on the yield of photonic devices is discussed and efficient method that includes the treatment of spatial correlation is presented.

The last part of the work is devoted to the modelling of the influence of temperature on the behaviour of optical devices. Thermal handling is also one of the fundamental issues in the effective exploitation of integrated photonic circuits. Unavoidable temperature gradients and fluctuations can significantly alter the behaviour of many devices. A method to take into accounts the wavelength, composition and temperature dependencies in the calculation of the refractive index and linear thermo-optic coefficient of $In_{1-x}Ga_xAs_yP_{1-y}$ alloys is presented and experimentally validated. The results provide a deeper understanding of the influence of the temperature on the behaviour of optical waveguides and devices, making possible an accurate and realistic modelling of integrated circuits.

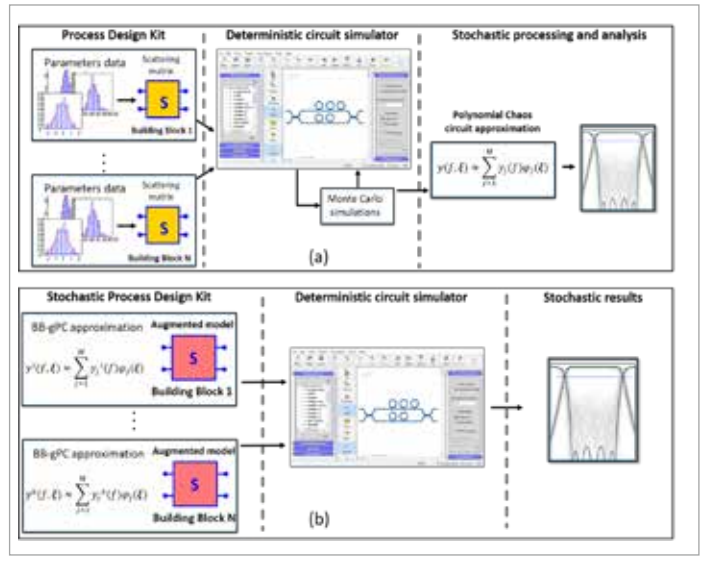

Fig. 1

ON RELEVANT QUERY ANSWERING OVER STREAMING AND DISTRIBUTED DATA

Shima Zahmatkesh - Supervisor: Prof. Emanuele Della Valle

Many modern applications require combining dynamic data streams with distributed data to continuously answer queries. Answering in a timely fashion, i.e., reactively, is one of the most important performance indicators for those applications. It is well known that remaining reactive can be challenging, because accessing the distributed data can be highly time consuming as well as rate-limited.
Consider the following example. In social content marketing, advertisement agencies may want to continuously detect influential Social Network users, when they are mentioned in micro-posts across Social Networks, in order to ask them to endorse their commercials. The number of followers may change in seconds, and the result of the query should be returned in a minute, otherwise, the competitors may reach the influencer sooner.
The Semantic Web community has recently started addressing the problem of evaluating queries over streaming and distributed data. They showed that RDF Stream Processing (RSP) engines provide an adequate framework for continuous query answering over the stream and distributed data. In this setting, distributed data is usually stored remotely or on the Web and accessible by using a SPARQL query over SPARQL endpoints. In order to access remote services, the query has to use federated SPARQL syntax, which is supported by different RSP query languages. However, remaining reactive can be challenging, especially when the distributed data is slowly evolving. High latency and rate limits in accessing the distributed data over the Web can put the applications at risk of losing reactiveness, i.e., the results of a query are no longer useful at the time they are returned.
State-of-the-art work addresses this problem by proposing an architectural approach that keeps a local replica of the distributed data. The local replica progressively becomes stale if not updated to reflect the changes in the remote distributed data. For this reason, recently, the RSP community investigated maintenance policies of the local replica that guarantee reactiveness while maximizing the freshness of the replica. The investigated maintenance policies focus on a class of queries that join a data stream with a distributed data source.
This thesis goes beyond the state of the art, focusing on finding the most relevant answers by continuously answering a query over streaming and distributed data, while considering the reactiveness constraints imposed by the users. The contributions of this study are various maintenance policies, which are tailored for two classes of queries: i) queries that have to filter data in the distributed dataset before joining it with streaming data, and ii) top-k queries where the scoring function involves data that appears both in the streaming and the distributed datasets.

We carried out our experimental evaluation over various Realistic and synthetic datasets. The experimental datasets are composed of streaming and background data. The streaming data is collected from 400 verified users of Twitter for three hours of tweets using the streaming API of Twitter. The background data is collected invoking the Twitter API, which returns the number of followers per user, every minute during the three hours we were recording the streaming data. As a result, for each user, the background data contain a time-series that records the number of followers.
For the class of continuous SPARQL queries that join the stream data with background data and the SERVICE clause contains a FILTER clause, Filter Update policy is proposed as maintenance policy. Intuitively, the Filter Update Policy focuses on a band around the filtering threshold for updating the data. The data in the band is likely to pass the filter condition and may affect future evaluations. Then, we introduced a group of policies as a combination of the Filter Update Policy with the state-of-the-art ones. The result of experiments showed that i) Filter Update Policy outperforms the state-of-the-art policies when the selectivity of filtering condition is above 60% of the total, and ii) the combined policies keep the replica even fresher than the Filter Update Policy.
We further investigated the combined approach, and the experimental evidence showed the difficulty of determining a priori the band around the filtering threshold to focus on. So, in the next step, relaxing the assumption in the combined policies, we proposed the rank aggregation approach, which let each policy to rank data items according to its criterion (i.e., to express its opinion), and then, aggregate them to take into account all opinions. The results of the experiments show that the proposed policies are comparable to the combined policies, but without requiring to determine a priori the band to focus on.
In the next step, focusing on the class of top-k queries, the contribution is an extended top-k query evaluation, which considers the join of streaming data with the distributed dataset. Various researches addressed the problem of top-k query evaluation in the streaming context, as the solutions proposed in the database community cannot be applied to streaming data. Introducing incremental query evaluation techniques, they try to avoid re-computing the top-k result from scratch at every evaluation, which is a major performance bottleneck in stream processing.
We extended the state-of-the-art approach for top-k query evaluation, considering distributed dataset with slowly evolving changes. The first proposed solution, Topk+N algorithm, works in data centers where the infrastructure is under control. We proposed Supre-MTK+N list, which is a data structure that keeps the top-k result in each evaluation. Then, we modify the state-of-the-art algorithm, we add the ability of handling the indistinct arrival of objects, and considering changed objects of distributed data as new arrivals.
This first solution may not work on the Web, where we have distributed data and we do not control the entire infrastructure. Therefore as a second solution, Considering the architectural approach presented in Semantic Web community as a guideline, we proposed AcquaTop framework, that keeps a local replica of the distributed dataset and updates a part of it based on a given update policy before every evaluation. When there are not enough refresh budgets to update all the stale elements, the result might have some errors. In order to approximate as much as possible the correct answer, we propose two maintenance policies (MTKN-F, and MTKN-T), which are specifically tailored to top-k query answering for updating the replica. MTKN-F policy maximizes the accuracy of the top-k result, i.e., it tries to get all the top-k answers in the result, but forsaking the order. \ MTKN-T policy, instead, maximize the relevance, i.e., minimizes the difference between the order of the answers in the approximate top-k result and the correct order, but forsaking accuracy of the less relevant results. The experimental evaluations empirically prove the ability of the proposed policies to guarantee reactiveness, while providing more accurate and relevant results than the state of the art.