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

PhD Yearbook | 2025

DOCTORAL PROGRAM IN INFORMATION TECHNOLOGY

Chair:
Prof. Luigi Piroddi

Introduction

The Ph.D. programme in Information Technology (Ph.D. IT) covers research topics in four scientific areas, associated to different facets of the field of Information and Communication Technology, namely Computer Science and Engineering, Electronics, Systems and Control, and Telecommunications. This broad variety of research topics perfectly captures the core mission of the corresponding sections of the Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB).
New research topics and cross-areas research fields are also covered, such as machine learning, big data, intelligent data analysis, Industry 4.0, internet of things, bioinformatics, quantum computing, ecology, environmental modelling, operations research, and transportation systems.
The Ph.D. IT programme is the largest at the Politecnico di Milano in terms of number of students. Every year, more than 90 new students join the programme, for an overall number of around 400.

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

Industrial collaborations

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

Educational aspects

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

Internationalization

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

Conclusions

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

PHD BOARD OF PROFESSORS	
Amigoni Francesco	Gulinatti Angelo
Ardagna Danilo	Jabali Ola
Baresi Luciano	Loiacono Daniele
Bertuccio Giuseppe	Luini Lorenzo
Bestagini Paolo	Maier Guido Alberto
Bonfanti Andrea	Margara Alessandro
Boracchi Giacomo	Mari Lorenzo
Cappiello Cinzia, vice-coordinator of Computer Science and Engineering area	Martinenghi Davide
Carminati Marco	Matera Maristella
Corno Matteo	Monti-Guarnieri Andrea Virgilio
Dercole Fabio	Parolari Paola
Farina Marcello	Pelosi Gerardo
Ferrari Giorgio	Piroddi Luigi, Coordinator
Ferrigno Giancarlo	Rech Ivan, vice-coordinator of Electronics area
Filippini Ilario	Santambrogio Marco Domenico
Fornaciari William	Silvano Cristina
Garatti Simone, vice-coordinator of Systems and Control area	Tornatore Massimo, vice-coordinator of Telecommunications area
Gatti Nicola	Zanchettin Andrea Maria
Geraci Angelo	Zanero Stefano

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

Prizes and awards

In 2024 the following awards have been obtained by Ph.D. candidates:

- **Dimitris N. Chorafas Prize** – Tommaso Alfonsi, Riccardo Giampiccolo
- **Prof. Florian Daniel PhD Thesis Award** – Luca Barbieri, Edoardo Ramalli
- **Springer Award** – Tommaso Alfonsi, Francesco Buccoleri, Federico Cacciamani, Daniele Cattaneo, Federica Filippini, Piergiulio Mannocci, Francesco Tesolin, William D'amico, Matteo Italia, Lucrezia Manieri
- **International Conference on Process Mining, CPM Workshop on ML4PM – Leveraging Machine learning in Process Mining - Best workshop paper award** – Musa Salamov
- **IEEE International Conference on Quantum Software - Best Paper Award** – Simone Reale
- **European X-ray Spectrometry Conference (EXRS 2024) – Best Poster Award** – Beatrice Pedretti
- **2nd CINI Summer School on High Performance Computing and Emerging Technologies – Best Poster Award –** Davide Baroffio
- **31° Reconfigurable Architecture Workshop (RAW 2024) - Best Poster Award -** Beatrice Branchini, Marco Venere
- **IEEE Applied Power Electronics Conference and Exposition - Best Presentation Award –** Paolo Melillo, Simone Zaffin
- **Agricultural Technology Excellence Doctoral Prize 2024 – Best Thesis and Research –** Mirko Usuelli

CALIBRATION TECHNIQUES IN RECONFIGURABLE TIME-INTERLEAVED ADCS

Gabriele Bè - Supervisor: Salvatore Levantino

This Ph.D. dissertation presents the concept, modeling, and practical design of time-interleaved ADCs for wideband wireless communications. In particular, the target is the 5G NR FR2 standard, which exploits channels at mm-wave frequencies to avoid the crowded low-frequency sub-GHz spectrum. The standard defines single channels as wide as 400 MHz, with various modulation schemes up to 256-QAM. These specifications then translate to the ADC requirements for the receiver chain. The large channel bandwidth implies speeds of a few GS/s, and the high modulation order demands medium- or high-resolution ADCs. On top of that, the ADC must be flexible to support the standard features, have a small area to reduce the cost, and have a low power consumption to avoid draining the battery.

The architecture of choice for medium-resolution GS/s ADCs is the time-interleaving (TI) of lower-speed slices. In particular, this thesis focuses on SAR-based designs due to their simple architecture and digital-friendly nature. Time-interleaving consists of multiple identical sub-ADCs working in parallel on a different sample of the same input signal to increase the effective sampling rate. Even though the concept is simple, many practical aspects complicate their effective deployment, as shown in Fig. 1. First, the channel mismatches introduce spectral artifacts that limit the overall resolution and SFDR. As such, offset, gain, and timing skew calibrations are introduced to compensate for the interleaving mismatches and restore the original sub-ADC performance. In particular, the proposed timing skew estimation algorithm extends the state-of-the-art difference-based error identification techniques to a general interleaving factor, enabling TI ADCs with any number of parallel channels.

Two TI ADC prototypes are presented. The first targets multiple frequency bands, introducing a low-cost, flexible interleaving architecture. It consists of a 12-bit, 900-MS/s, SAR-based, 6x TI ADC. The interleaving factor is reconfigurable from one (single-core operation) to six, resulting in a sampling rate variable from 150 to 900 MS/s in steps of the core sampling rate of 150 MS/s via a digitally programmable phase generator embedded into each slice. The prototype integrates offset, gain, and timing skew calibration algorithms to track the environmental variations in the background. Implemented in a TSMC 1P8M 28-nm bulk CMOS process, the TI ADC achieves remarkable interleaving spurs in all the configurations, with a mismatch spur level better than -73 dBc. For the single-channel case, the SNDR and SFDR for a sub-ADC near-Nyquist input are 63.3 dB (10.2 ENOB) and better than 72dBc, respectively. For the 6x TI ADC configuration, instead, the low-frequency SNDR and SFDR are 55.9 dB (9 ENOB) and 65.6 dBc, respectively, while the high-frequency ones are 52 dB (8.35 ENOB) and 58.8 dBc. This performance degradation is mainly due to the crosstalk between the sub-ADCs via the reference distribution network and the limited linearity due to the kickback into the front end, exacerbated by the absence of an input buffer (Fig. 1). The resolution (SNDR) and the spectral purity (THD) are lower than the target, pushing for further investigations. In particular, the latter is critical for wireless applications to avoid strong, unwanted signals falling into the signal band.

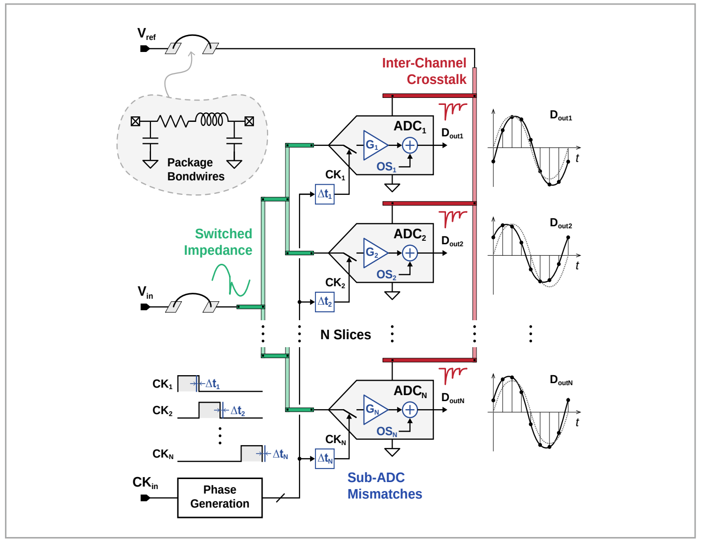

Fig.1 - Main factors degrading the performance of an N-channel TI ADC compared to its ADC slices: input switched impedance (green), inter-channel crosstalk (red), and sub-ADC mismatches (blue).

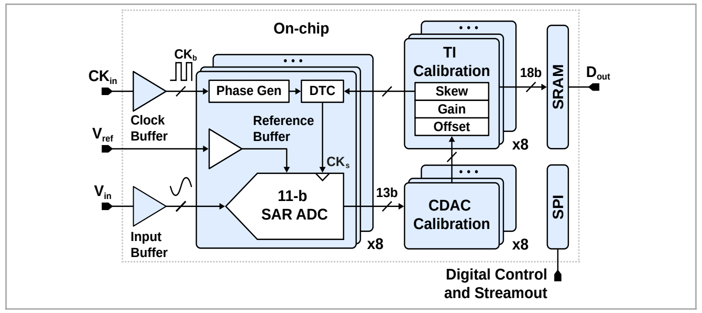

Fig.2 - Block diagram of the presented TI ADC comprising the analog section with input buffer, clock buffer, and SAR sub-ADCs, and the digital processing unit including sub-CDAC mismatch corrections, interleaving calibrations, and the auxiliary digital logic for controlling and testing the prototype.

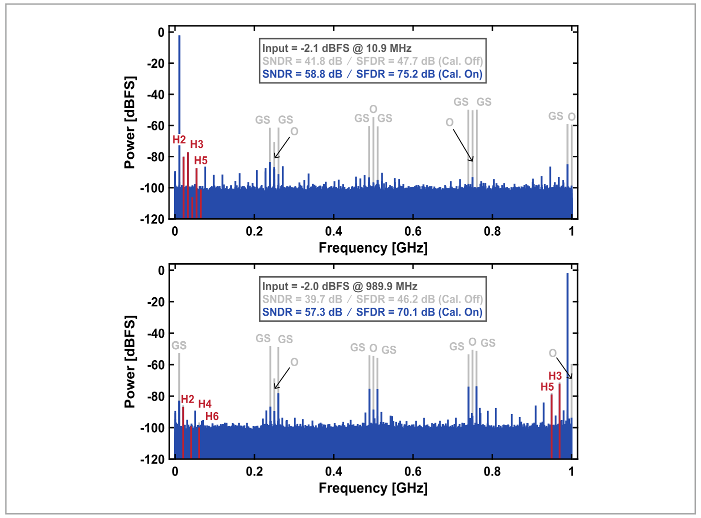

Fig.3 - Output spectra before (gray) and after (blue) the interleaving calibrations for a 10.9-MHz (top) and a 989.9-MHz input tone (bottom). The offset and gain/skew tones are indicated in gray by the letters O and GS, respectively, while the distortion terms are highlighted in red with their corresponding order.

The second prototype then started from the architecture and the interleaving calibrations of the first one, introducing an input buffer and a reference voltage distribution network embedding a reference buffer into each slice to limit the performance degradation of the TI ADC compared to the sub-ADCs. Figure 2 shows the overall architecture. It consists of an 11-bit, SAR-based TI ADC implemented in a TSMC 1P8M 28-nm bulk CMOS process, where eight 250-MS/s slices are time-interleaved for an aggregate sampling rate of 2 GS/s. The presented techniques lead to a TI ADC performance similar to the sub-ADCs. In particular, the average peak SNDR decreases by less than 1.4 and 1.76 dB in the amplitude and frequency sweeps, respectively. The low-frequency SNDR and SFDR are 58.8 dB (9.47 ENOB) and 75.2 dBc, respectively, while the high-frequency ones are 57.3 dB (9.23 ENOB) and 70.1 dBc (Fig. 3). Moreover, the SFDR exceeds 63.5 dB across the TI ADC 1-GHz bandwidth. With a total power consumption of 118.6 mW (including the input buffer, the sub-ADCs array, and the interleaving calibrations), the TI ADC achieves a Walden and a Schreier FoM of 99 fJ/conv-step and 156.6 dB, respectively. Compared to the state of the art, this prototype achieves the best resolution and linearity among 1-bit/cycle SAR-based TI ADCs, maintaining a comparable Schreier FoM. These results prove the effectiveness of the presented techniques for the time-interleaving of several sub-ADCs without significant performance degradation.

ENHANCING SITUATIONAL AWARENESS THROUGH REAL-TIME SOCIAL MEDIA DATA ANALYSIS

Carlo Alberto Bono - Supervisor: Barbara Pernici

Social media has revolutionized the way people experience and share reality, producing an unprecedented volume of real-time information. This dynamic and pervasive information flow makes it an invaluable data source for a timely understanding of events and trends. Among the others, social media platforms offer a rapid access to situational insights during rapidly changing events, such as emergency events. However, the task of extracting relevant and actionable information from noisy, unstructured, and unpredictable data streams remains a significant challenge. This research addresses these issues focusing on event detection, data filtering, multimodal enrichment, and adaptive learning. Tackling these challenges by developing interconnected methodologies ultimately enhances the ability to use social media as a reliable tool for gathering information, including crisis scenarios.

Regarding the research objectives, this work is structured around four key research questions: (1) How can systematic approaches support the extraction of relevant data from social media streams, especially during emergencies? (2) How can a language-independent, context-aware event detection system be designed for real-time crisis scenarios? (3) How can data filtering methods continuously adapt to evolving events, improving real-time situational awareness? (4) What role do multimedia attachments (images, videos, geotagged data) play in enriching crisis-related social media analysis? These questions have been explored through a multidisciplinary approach that integrates Machine Learning (ML), Natural Language Processing (NLP), Computer Vision (CV), Active Learning (AL), and multimodal data fusion in order to develop a scalable, adaptable framework for real-time social media analysis.

As effective analytics begin with data preparation, this research first focuses on a goal-driven, iterative methodology for configuring data preparation pipelines, ensuring real-time filtering of social media data. The approach optimizes efficiency, cost, and usability constraints, integrating human-in-the-loop validation to iteratively leverage context-awareness. Empirical validation on real-world emergencies (e.g., floods) demonstrates its robustness in isolating relevant data while minimizing noise, at the same time giving the designer control over which quality dimensions to maximize.

A critical limitation of existing social media-based event detection is dependence on static keyword lists and language-dependent approaches. In the context of multilingual, context-aware event detection, this research introduces a language-independent framework that can be used to configure dynamic event-specific lexicons. A downstream lightweight CNN model enables then the rapid understanding of crisis-related social media signals, outperforming traditional methods. Validation is conducted with empirical studies on diverse flood alerting scenarios.

Moreover, since emergency events are non-stationary and information needs shift rapidly, our research focused on how to integrate adaptive filtering. We propose an active learning-driven adaptation strategy that continuously refines filtering models using quality-driven data selection. By prioritizing high-value training samples – based on redundancy, singularity, and contextual relevance – this approach enhances classification accuracy while minimizing manual labeling, which is usually constrained by a limited budget. Experimental benchmarks on various events of interest show that this method improves performance over traditional uncertainty-based selection strategies in our scenario, making it particularly effective for resource- and data-scarce environments.

While most social media analysis focuses on text, multiple modalities such as images and videos often contain critical information. We propose methods to integrate visual data (including image and video attachments) and text, providing crisis responders with higher quality data. We demonstrate the ability of the method ability to improve rapid impact assessment using major disasters as case studies.

By bridging the gap between academic research and crisis management practice, the proposed research provides a foundation for next-generation social media-based situational awareness systems, ultimately enhancing the reliability and timeliness, and thus the effectiveness, of emergency response efforts. The proposed methodologies address fundamental challenges in crisis informatics and have a number of practical implications, with key advantages including scalability, geeralizability, adaptability and actionability. Building on these contributions, future directions intend to explore (1) How to leverage Large Language Models (LLMs) for zero-shot and few-shot crisis detection, (2) How to integrate graph-based methods for assessing relevance and leverage information propagation patterns, (3) How to address ethical concerns, including algorithmic bias and privacy risks in AI-driven disaster response.

NEXT GENERATION SINGLE-PHOTON AVALANCHE DIODES FOR HIGH-EFFICIENCY, NEAR-INFRARED PHOTON-TIMING

Andrea Bonzi - Supervisor: Angelo Gulinatti

Thanks to their remarkable performance, Single Photon Avalanche Diodes (SPADs) find nowadays application in an ever-growing number of fields, spanning from scientific research to consumer electronics. Among the many SPADs now available, silicon SPADs are the most widespread ones, because of their excellent performance in terms of Dark Count Rate (DCR), afterpulsing probability, and time resolution.
Despite the incredible steps forward of the last decade, already-existing detectors make not possible to envision a good combination of high detection efficiency and good timing resolution in the near-infrared spectrum (i.e. between 800 and 1000 nm of wavelength). The reason is an intrinsic trade-off that plagues the current SPAD structures: as the detector is illuminated from the top, increasing the detection efficiency requires a thicker depletion region which results in a larger time jitter, owing to the dispersion of the carriers' transit times. My doctoral work aims at developing SPADs that overcome the trade-offs between detection efficiency and timing jitter at near-infrared wavelengths.
First, with the goal of helping the growth and validation of new and more accurate SPAD models, I developed an extraction technique able to provide accurate doping profiles along the depletion region of the measured SPADs. Indeed, SPAD models, thanks to their ability in accurately forecasting detectors' metrics, play a key role in the evolution of SPAD technology. However, to verify the effectiveness of these models, a thorough and extensive validation against experimental data is required. To this aim, being most physical phenomena that determine detector performance strongly dependent on the electric field, accurate doping profiles are mandatory. Unfortunately, widely-adopted profiling techniques provide results which are not precise enough for SPAD modeling. Starting from an initial, approximate, doping profile provided by process simulations and/or Secondary Ion Mass Spectroscopy (SIMS) measurements, the technique I developed refines it up to an accuracy level high enough to correctly predict SPAD metrics.
To do this, I adopted an inverse scheme in which the doping profile is continuously corrected until a close matching is obtained between electrical simulations and experimental capacitance-vs-voltage (C-V) measurements.
To verify the method, I applied the technique to several real SPADs with different internal structures and different doping profiles and used the obtained results to compute their breakdown voltage. Fig. 1 reports the results for all the SPADs under test. Calculated outcomes in good accordance with experimental data were obtained, providing a convincing validation of the proposed method.
Second, with the aim to push farther the boundaries set by the trade-off between PDE and timing jitter, a side-illumination approach, in which the light propagates transversally, has been proposed (see Fig. 2). This way, the absorption efficiency can be improved by increasing the detector length rather than its thickness and the latter can be reduced to the typical values of the thin SPADs with a potential time jitter of about 30 ps FWHM. To do this, the SPAD will be embedded in a silicon waveguide and the photons, generated externally, will be coupled into it through an optical fiber. An upper electrode and a silicon contact region present on one side of the guide are necessary for biasing the structure. As a first step, an optical analysis has been conducted in order both to size the main parameters of the waveguide structure and to verify the robustness of the system. Afterwards, a simplified but realistic fabrication process was devised and simulated in order to investigate the electrical properties of the detector.
The simulations I performed confirmed that this solution is very promising and that non-idealities, such as the surface roughness of the guide or a small misalignment or tilt between the detector and the input optical fiber, have only a minimal impact on the detector performance.
Third, in order to overcome SPAD limitations within existing detector structures (thin and red-enhanced SPADs in particular), the following two technological solutions have been explored: back-illumination technology and resonant cavity integration. Indeed, despite the promising results predicted for the side-illumination approach, a waveguide-SPAD is not suitable for a number of applications. Simulation results allowed me to conclude that impressive improvements may be attained in terms of both detection efficiency and time response through the aforementioned solutions, either singularly or in combination.
Lastly, as there are still many applications for which red-enhanced SPADs may be suitable, the final part of my Ph.D. work was dedicated to this class of detectors. Modifications have been applied to the RE-SPADs' fabrication process in order to satisfy a series of requirements, such as a lower junction depth, an improved compactness, and a reduced thermal budget. With the goal of evaluating the effectiveness of all the modifications I applied, a simplified process flow with a reduced manufacturing time was also devised to produce test devices. In addition, to be able to extract the boron profiles from the wafers where the RE-SPADs will be manufactured, the potential and the feasibility of different Metal-Oxide-Semiconductor (MOS) structures have been investigated.

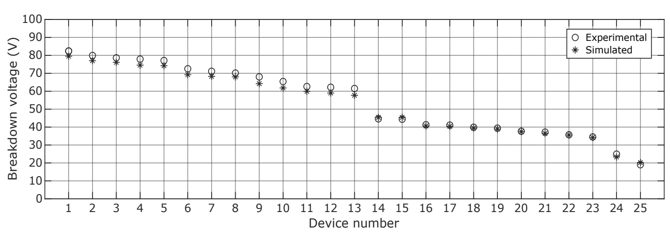

Fig.1 - Simulated breakdown voltages are reported along with the experimental ones for all the SPADs under test.

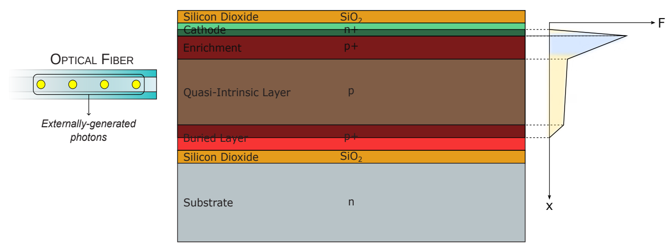

Fig.2 - The SPAD is embedded in a silicon waveguide and the photons, generated externally, are coupled into the detector through an optical fiber.

EXPLORING THE FUTURE OF NAVIGATION: HIGH TRL PIEZORESISTIVE GYROSCOPES

Andrea Buffoli - Supervisor: Giacomo Langfelder

MEMS gyroscopes experienced a rapid and vertical growth since their first conception in 1991, attracting different market areas ever since. Indeed, nowadays consumer, automotive and tactical applications widely exploit MEMS-based angular rate sensors thanks to their low SWaP-C. Their fast development have raised interest also for high-end applications, such as inertial navigation, which could greatly benefit from high-performance, lightweight and miniaturized gyroscopes to accurately retrieve the heading of an object.
The work developed in this PhD thesis discusses advancements towards navigation-grade performance for piezoresistive based M&NEMS gyroscopes, with the long term objective of providing a medium-to-high TRL, miniaturized and stand-alone prototype. The starting point is a laboratory-based setup exploiting off-the-shelf components and instrumentation, only, that showed near-navigation grade performance for yaw devices. First, a temperature characterization campaign showed competitive results in terms of ZRO and SF stability in temperature for near-navigation grade M&NEMS angular rate sensors, and a theoretical analysis on the major SF drift source has been discussed and experimentally validated. Then, the same performance both in terms of noise, stability and thermal behaviour have been confirmed by coupling the gyroscope to a custom low-noise ASIC designed specifically for this technology, including drive loop, AGC and open-loop sense chain. The IC and gyroscope coupling has been used as the analog core for the development of a FPGA-based stand-alone pre-industrial demonstrator, shown in fig. 1, that maintained (and even improved) performance, whilst miniaturizing and digitizing the system. This step paves the way for a full validation of the M&NEMS technology, complying with the industry and market requests for a medium/high TRL prototype, which shows the sensors performance with high fidelity and repeatability grade. Afterwards, further advancements in the integrated electronics design have been presented regarding: (i) the addition of demodulation and automatic quadrature compensation stages in a second version of the IC, that showed unexpected extra 1/f noise due to a non-optimized buffer in the signal output chain, and (ii) preliminary study and design for the integration of a high-resolution continuous-time ΣΔ ADC. Finally, the design of the new generation electromechanical structures has been reported, with particular focus on pitch/roll architectures. An exploratory fabrication run at higher MEMS thickness is indeed devoted to boost the in-plane sensitive device performance, for which an innovative mechanically equilibrated architecture has been presented, targeting both more efficient Coriolis-induced coupling and quadrature compensation.
Overall, this work demonstrated near-navigation grade performance on yaw M&NEMS gyroscopes, while at the same time extending the technology readiness level by reducing both dimensions and power consumption of the system. Indeed, a noise level of ARW = 0.006 °/√h, and BI = 0.016 dph have been experimentally measured, as shown in fig. 2, together with promising thermal behaviour showing ZRO and SF drifts lower than 1.3 dph/K and 850 ppm/K in the worst cases, respectively. These results well compare with some of the most advanced MEMS gyroscope products or prototypes present in the market, as shown in tab. 1. Tactical grade performance, with ARW = 0.018 °/√h and BI = 0.12 dph, have been also demonstrated on pitch/roll M&NEMS devices, ranking them among the best ever reported for in-plane sensitive devices. Furthermore, new designs at higher MEMS thickness target at obtaining performance very close to the one shown by the yaw counterpart.

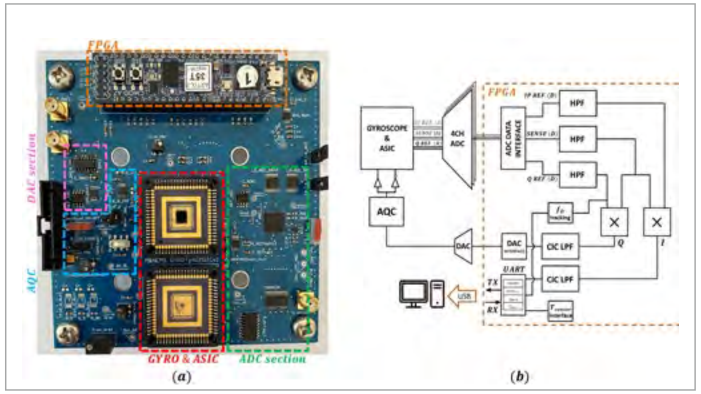

Fig.1 - (a) top image of the produced demonstrator, (b) high level schematic of the system.

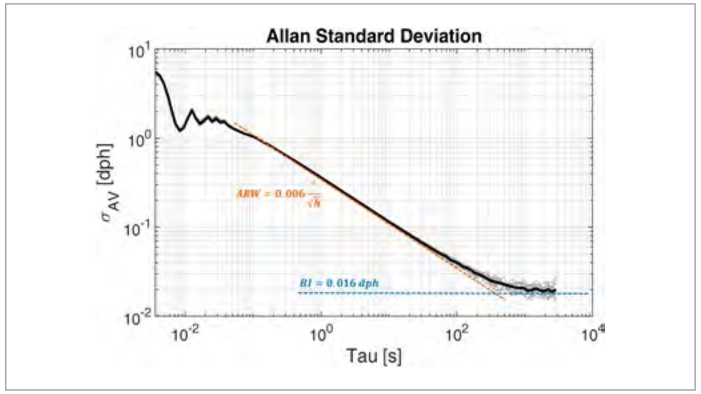

Fig.2 - Fifteen 8-hours long Allan standard deviations acquired in uncontrolled laboratory environment, showing good repeatability.

		ARW [°/√hr]	BI (25°C, 1000s) [°/hr]	T.O.T.O. [°/hr rms]	ZRO residual [°/hr rms]
Tronics GYPRO 3300 (product)		0.14	0.8	10	70
Safran STIM300 (product)		0.15	0.4	4	9
Honeywell HG7930 (prototype)		0.0035	0.01	±0.1	0.2
This work (prototype)		0.006	0.02 (uncalibrated)	1	6

Tab.1 - Comparison between this work and some of the most advanced MEMS gyroscope products/prototypes in the market/literature.

DEVELOPMENT OF A GOMMA-RAY DETECTION UNIT FOR A SPECT SYSTEM FOR REALT-TIME DOSE MONITORING IN BNCT

Anita Caracciolo - Supervisor: Carlo Ettore Fiorini

This doctoral research focuses on the development of a gamma-ray detector designed for use in a Single Photon Emission Computed Tomography (SPECT) configuration, specifically for quantifying and localizing the dose delivered to patients during Boron Neutron Capture Therapy (BNCT). BNCT is a highly targeted form of hadron therapy that exploits the interaction between boron-10, selectively accumulated in tumor cells via specific binding compounds, and thermal neutrons used to irradiate the patient. This interaction produces two high linear energy transfer (LET) particles that release their energy within the targeted tumor cells. Due to its precise targeting capability, BNCT is particularly effective for treating infiltrative or diffuse tumors and tumors located near critical structures. Its clinical adoption has been limited by several challenges, including the generation of sufficient clinical neutron fluxes, the development of selective and non-toxic boron compounds, and the absence of real-time dose monitoring systems. Recent advancements in compact accelerator-based neutron sources have revitalized interest in BNCT, with several new facilities now operational or under construction worldwide. As a result, there is an urgent need to further develop BNCT-related technologies.

The objective of this thesis was to develop a single unit of a BNCT- SPECT system, that aims to detect gamma rays at 478keV emitted by boron neutron capture reactions to enable real-time dose monitoring. The development of this system is particularly challenging due to the unique requirements of BNCT, including the need to detect low concentrations of boron (tens of ppm) over a huge background of mixed radiation produced by the high-intensity neutron flux necessary for treatment.

The development focused on an indirect conversion gamma-ray detector based on a $LaBr_3(Ce^{3+}+Sr^{2+})$ scintillator crystal, coupled with silicon photomultipliers (SiPMs) and read by custom electronics based on Application-Specific Integrated Circuit (ASIC) and Field-Programmable Gate Array (FPGA). The detector is combined with a pinhole collimator and an Artificial Neural Network (ANN) algorithm for two-dimensional position reconstruction of the gamma events. Over the course of this thesis, three successive versions of the detector were developed and are presented herein.

A series of experimental measurements at the TRIGA Mark II research nuclear reactor of Pavia University validated the feasibility of the BNCT-SPECT unit developed. The system successfully detected 1cm lateral displacements of a borated vial during neutron irradiation. These results strongly suggest that the BNCT-SPECT system proposed and developed in this research holds significant promise for clinical applications.

EXPLOITING HARDWARE AND SOFTWARE SPECIALIZATION FOR EFFICIENT REGULAR EXPRESSIONS PROCESSING

Filippo Carloni - Supervisor: Marco Domenico Santambrogio

Context and Background

The ever-increasing computational demand is straining current processing systems to fulfill the strict constraints of today's applications. This trend leads to a scenario where the computational systems can no longer sustain the incoming latency and throughput requirements in data analysis, potentially creating critical bottlenecks. Among the attractive but challenging modern computational kernels, pattern-matching based on Regular Expressions (REs) is among the most pervasive and intriguing. Specifically, REs are particularly useful due to their straightforward and powerful pattern representations beyond simple string matching, which allows the specification of complex behavior such as bounded and unbounded pattern repetitions, pattern alternation, character classes, and different execution strategies. Moreover, the processing simplicity of their equivalent-power Finite State Automata (FSA) recognizers combined with the advanced pattern expressiveness widely increases the utilization of REs in a plethora of application fields ranging from network and computer security to bioinformatics. For instance, they can be used to specify malicious packets, malware signatures, wildcards in database queries, and motifs in DNA.

Modern Challenges in Pattern Matching

Efficiently executing REs is an insidious task due to the high data dependency and irregular memory accesses of automata processing, which strongly limits the parallelization opportunities in performance-demanding contexts. Specifically, modern applications require low-latency, high-throughput, and energy-efficient processing in conjunction with the possibility of dynamically adapting the patterns to analyze at runtime, like in database queries. Applications such as Intrusion Detection Systems (IDSs) have an additional degree of complexity by requiring parallel RE analyses. These hard-to-fulfill requirements, combined with the intrinsic issues of RE efficient processing, demand different and advanced execution approaches to sustain the contemporary and future challenges.

Dissertation Approach and Contributions

This dissertation proposes to benefit from hardware and software domain specialization in a tightly coupled strategy to address the modern RE challenges.

RE benchmarking: Since the current set of RE execution engines comprehends heterogeneous execution platforms, before introducing novel approaches, it is essential to leverage previous efforts in the literature to characterize RE processing. This dissertation proposes a fair and standard evaluation methodology, called Yet Another REs Benchmarking framework (*YARB*), for measuring execution latency varying the data input size of publicly available and reusable RE execution engines. Thanks to its simplicity and extendibility, YARB creates the assessment baseline for the remainder of the evaluations considered here and for future RE engine characterization.

DSA benefits in embedded devices: This dissertation explores the use of an existing Domain Specific Architecture (DSA)-based approach in resource-constrained scenarios, like the Internet of Things (IoT) and Cyber-Physical Systems (CPS), as a motivating example of the potential benefits of domain specialization for executing resource-hungry and challenging tasks, such as REs used in Host-based IDS, when the computational power is limited. In particular, by using an existing DSA from the literature, it was possible to demonstrate that a DSA-based strategy can provide remarkable improvement when taking into account energy efficiency and deliver better performance than more powerful boards when executing more complex tasks, as in the case of multiple RE alternation.

Full-stack framework for efficient execution of REs: The proposed central approach focuses on an advanced hardware/software framework for efficient RE processing, targeting execution latency and energy efficiency as the primary evaluation metrics. This method relies on employing the *REs as a programming language* by compiling and optimizing the input REs into a set of instructions efficiently executed on the target DSA. After analyzing modern benchmarks (thanks to the help of YARB) to extract the basic set of RE features that a contemporary general-purpose engine must support, this dissertation proposes a novel Reduced Instruction Set Computer (RISC)-based Instruction Set Architecture (ISA) to support all of them in an advanced and efficient solution. Based on the presented optimized ISA, a tailored microarchitecture is introduced for efficiently executing the supported simple and advanced RE operators by leveraging a backtracking algorithm to restore from misprediction during the pattern analysis. Remarkably, the microarchitecture currently supports *greedy* and *lazy* strategies - the two most common RE execution strategies - which classical automata execution engines typically do not support but are included in the PCRE standard. In addition, this dissertation presents a domain-specific compiler to translate and optimize the input RE into an ISA-compliant binary, eventually executed by the proposed DSA. The introduced framework prototyped on an embedded FPGA can provide lower latency than advanced solutions like RE hardware accelerator on SmartNICs or GPU-based execution engines and similar performance compared to optimized software libraries (which strongly rely on advanced x86 features available in desktop and server-grade CPUs). Furthermore, the proposed framework can provide better energy efficiency than all evaluated approaches, thanks to its highly optimized computation and a better compiler/execution ratio, making it an attractive solution for applications requiring runtime RE adaptability.

Advanced compilation flow: Existing DSA-based strategies commonly rely on complex hardware designs to provide competitive performance, often overlooking the benefits of the software. This aspect leads to oversimplified compilation flows, which lack structured infrastructures to properly optimize the input REs. To overcome this limitation, this dissertation explores the use of MLIR multi-dialect abstractions to avoid *premature lowering* issues in a literature DSA-based RE compiler, enriching the overall efficiency by relying on compiler optimizations without requiring complex microarchitecture changes. In particular, the compiler has both an abstract dialect to optimize REs from an architecture-agnostic perspective and an architecture-oriented dialect for final DSA-specific optimizations. The described compilation strategy provides better performance and code structure than the monolithic approach without requiring overcomplicated hardware design.

Novel automata model: Since modern RE applications can require the analysis of large rulesets composed of hundreds to thousands of REs, this dissertation proposes to leverage similarities among REs within the same ruleset to improve the processing throughput and reduce the overall memory footprint. In particular, it introduces a multi-level compilation framework and a merging procedure to combine a group of REs that share common sub-patterns into a novel automaton model called *Multi-RE FSA* (*MFSA*). The MFSA extends the 5-tuple FSA model with the *activation function* and *RE set* concepts. In particular, the activation function allows the tracking of the *active* REs during the MFSA processing, removing the ones that are not allowed to continue matching because they do not *belong* to the next available transition. This way, it is possible to traverse one single MFSA while actually analyzing multiple REs parallelly, avoiding false positive matches potentially generated during the automata merging. A literature NFA-based processing engine is then extended to execute the MFSA model and evaluate its performance.

TECHNOLOGY AND APPLICATIONS OF COMPILER-BASED PRECISION TUNING

Daniele Cattaneo - Supervisor: Giovanni Agosta

Introduction

As the need for computing power constantly increases, there is also an increasing need to use it more efficiently. One way to achieve this is to exploit *approximate computing* techniques like *precision tuning*, whose goal is to reduce the precision of a program by changing its data types. At the moment, precision tuning has a limited appeal, because selecting lower precision data types and algorithms is non-trivial. The possibility of generalized automatic precision tuning, although appealing, is still an open research question.
To answer this question, we built a new extensible precision tuning framework, TAFFO, in order to explore the design space of precision tuning tools and build upon the current state-of-the-art. The architecture of TAFFO is based on LLVM, and it is implemented as five distinct compilation passes.
The Initialization pass scans the application in search of user-inserted annotations which provide contextual metadata, and the Value Range Analysis pass extrapolates the ranges of all variables in the program from the annotations. The ranges are then used to determine new data type assignments in the Data Type Allocation pass, the Conversion pass modifies the program to apply the new types, and the Error Analysis pass verifies a-posteriori if the error constraints are satisfied.

Optimization-based Data Type Allocation

The most crucial part of precision tuning is the Data Type Allocation (DTA). In TAFFO, we exploit a fully static analysis, which translates the program to an integer linear programming (ILP) optimization problem solvable by an off-the-shelf tool. This allows seamless support of floating point and fixed point data types, and of any other future number representation.
To optimize the accuracy of the transformed program, compare between different alternative representations using a new more general metric for the error, the IEBW, defined as the minimum number of fractional bits needed by a fixed point representation to represent all numbers in a range with the same precision as a given representation.
Additional parameters of the ILP model include target-hardware-specific characterization data, and the constraints reflect the LLVM-IR representation of the program. The data type assigned to any floating point register is modeled by mutually-exclusive binary variable for each supported data type. Fixed point formats are handled by separate integer variables defining the position of the point. More constraints are defined to model the peculiarities of LLVM-IR, such as type rules. The objective function takes into account the execution time of each operation, and the overall precision, expressed as the sum of the IEBW of all virtual registers. The terms related to execution time and cast are weighted separately to allow the user to fine-tune the required precision. The ILP-based DTA pass just described allowed us to obtain speedups up to 800% for benchmarks in the Polybench/C 4.2.1 suite on an embedded systems board with a 120 MHz ARM Cortex-M3processor.

Handling of Mathematical Functions

Many computational kernels can gain considerable speedups from application of mixed-precision techniques, but this speedup is often limited if the kernel makes heavy use of external library functions for performing operations such as trigonometric functions. Often there are no readily available alternative reduced-precision implementations of these, and higher-precision ones are used instead. This unnecessary precision has a measurable impact, especially in embedded systems. Ideally, precision tuning should rely on multiple implementations of each mathematical function, tailored to each supported data type. However, if the precision tuning tool supports configurable representations like fixed point ones, the amount of functions that should be provided grows substantially. In the context of programming languages, this kind of issue is solved using *template metaprogramming*. Our proposal, FIXM, is to adapt the *metaprogramming* approach to the automatic generation of the implementation of mathematical functions when exploiting the fixed point representation, rather than providing every possible function ahead-of-time. FIXM detects where the program uses one of the supported POSIX standard mathematical functions, and generates the specialized fixed point mathematical functions for each call site. The original function calls in the program are then replaced with a call to the newly generated mathematical function. If a suitable function had already been generated, the existing function is used. The process of generating the functions is performed by changing parameters in a set of template functions internally stored as LLVM-IR code. These parameters include the aforementioned function argument and return values. Our generator can support two kinds of mathematical library implementation: CORDIC-based and Look-Up-Table-based. These different approaches have specific strong points and drawbacks that complement each other, giving a more complete set of options to the user. With FIXM we achieve speedups up to 125% with negligible error, on a microcontroller platform with an ARM Cortex M4 CPU running at 168 MHz, in benchmarks from the FBench suite, which is specifically tailored for reduced-precision applications.

WCET Optimization

Finally, let us consider an application scenario where precision tuning may be applied: *critical* and *mixed critical* systems. In the development of such systems, one important aspect is to analyze the *Worst-Case Execution Time* (WCET) of each routine involved in the critical task. Knowing the WCET allows to schedule multiple tasks such that the CPU time allotted for each one of them will be appropriate. Henceforth, in many applications there is no single WCET target to hit, and the WCET is a metric to be optimized. This optimization can be performed through precision tuning. The basic idea is to adapt the Integer Linear Programming optimizer to make an estimation of the WCET. Indeed, the mathematical model already computes a parametric estimation of the relative execution time slowdown and of the round-off error, with respect to a fully floating-point-based implementation. These estimates are not useful in general to gauge the real error or execution time of the generated program, because the per-instruction coefficients employed are ratios rather than absolute errors or instruction timings. However, in principle, if we replace such coefficients with values corresponding to the actual error being inserted by a computation or its actual execution time
in clock cycles, the optimiser will gain visibility to a realistic estimation of these metrics, allowing
to select a given WCET to optimize for, and to estimate it.

Conclusion

We described novel precision tuning techniques that makes it more applicable to real-world applications: a new data type allocation approach based on Integer Linear Programming, and a metaprogramming-inspired approach to automatically generate lower-precision mathematical function implementations. Exploiting our methodologies, we achieved significant speedups with error figures that are overall under 1%.
As an application example, we describe how to exploit precision tuning to predict and optimize the Worst-Case Execution Time required by a software routine.
This work is just a step towards the ultimate goal of *plug-and-play precision tuning*, i.e. the capability of enabling precision tuning at the turn of a switch. Further improvements to TAFFO as a tool will therefore be necessary to reach that ideal. We expect the adoption of nextgeneration compiler frameworks such as MLIR to be crucial to allow using precision tuning in an even wider range of contexts.

MACHINE LEARNING METHODS FOR SIGNAL PROCESSING WITH APPLICATIONS TO THE WIRELESS PHYSICAL LAYER

Lorenzo Cazzella - Supervisor: Matteo Matteucci
Co-Supervisor: Dario Tagliaferri

This multidisciplinary thesis, at the intersection of machine learning, signal processing, and wireless communications, aims to study and develop adaptive signal processing techniques that exploit prior knowledge on the features of physical systems and signals to produce accurate models of physical phenomena. The need for specific machine learning algorithms able to deal with the noisy and partial nature of real-world signals has recently flowed into the Machine Learning for Signal Processing (MLSP) binomial. Whereas the observed signals can be noisy and their observability can be limited, the existence of an underlying system structure can be exploited to improve model design and performance.
Machine learning models have achieved remarkable performance in many tasks involving, e.g., natural language, images, videos, and audio signal processing. Nevertheless, a vastly unexplored research area involves the design of learning models to manage real-world signals and systems. These settings introduce specific challenges and opportunities: real-world signals are noisy, the number of available observations can be highly limited, low-complexity solutions are commonly necessary in critical conditions, and systems can present complex nonlinear and time-variant dynamics.
A setting featuring these challenging conditions is provided by the physical layer of wireless vehicle-to-everything (V2X) communications, which is considered as application scenario in this thesis. We focused on two tasks: (i) the design of low-rank signal denoising methods by means of clustering-based and deep learning-based subspace estimation, with application to high-frequency channel estimation in V2X communications, and (ii) the association of a radar target with the corresponding communication user in integrated sensing and communication (ISAC) systems.

Unsupervised low-rank signal denoising with robust subspace clustering

Our first contribution is a method based on robust clustering to denoise low rank-structured noisy signals. Many real-world signals, despite being high-dimensional, are known to lie on low-dimensional subspaces owing to the physical features of the systems by which they have been generated. The identification of these low-dimensional subspaces is instrumental for signal denoising: prior knowledge on the signal structure can help to filter out noisy components in a suitable subspace domain. Low-rank signals emerge in many real-world scenarios, and several common signal domains can be mapped to subspace representations and effectively processed accordingly.
When it becomes difficult to dispose of explicit labeling of the subspace structural features of a noisy signal, unsupervised techniques can play a significant role in retrieving signal patterns without explicit supervision. We tested the proposed method on high-frequency V2X channel estimation. Channel estimation techniques in the communication standard, like Least Squares (LS), are efficient but can produce extremely noisy estimates in high mobility scenarios. Since the communication channel at extremely high frequencies is known to be low-rank in the space-time (ST) domain, we propose a novel low-rank denoising method based on robust K-medoids clustering exploiting the subspace ST similarity of LS channel estimates to denoise them. This approach leverages the traffic patterns emerging from recurrent vehicle passages in typical urban scenarios. Numerical results showcase the distribution of the identified ST similarity clusters on the considered urban scenario and the benefits in terms of Normalized Mean Squared Error (NMSE) on channel estimation (achieving ~15 dB less) with respect to the standard LS solution.

Deep learning-based subspace estimation for low-rank signal denoising

As our second contribution, we propose a method for subspace estimation based on deep learning aimed at denoising noisy signals known to be low-rank. We focused on the design of a method that exploits the representational power of deep learning to achieve high estimation performance in noisy and low-data conditions. We devised a deep learning architecture that takes as input a single noisy estimate known to be low-rank and estimates the low-dimensional subspace on which it lies. The noisy estimate is then projected on the predicted subspace to filter out its noisy spectral components. As for the unsupervised learning methodology explored in the previous contribution, we applied this method to V2X channel estimation, where the noisy estimates are LS channel estimates. Numerical results show that the proposed method outperforms LS channel estimation in terms of NMSE by ~15 dB on average. Moreover, we investigated the generalization of the trained model to scenarios presenting diverse macroscopic space-time channel features, showing that the proposed method can be directly transferred to the other scenarios with minimal performance loss. Fine-tuning of the last two network layers has shown to further reduce the average NMSE by ~2 dB.

Learning to associate radar targets with communication users

The third contribution of this thesis is a deep learning method for the association of a vehicular radar target with the corresponding communication user in ISAC systems. In the latter, the communication infrastructure is provided with the capability to perceive the surrounding environment by means of radar sensing. Beamforming is a widespread signal processing technique used to achieve directional gains (beams) at the transmitter and receiver antenna arrays to counteract the high path loss emerging at high frequencies. We propose to perform the target-to-user (T2U) association in the beamforming space exploiting per-target beam prediction and the set of beams used at the communication infrastructure. We obtained joint beam prediction and multi-target detection by adopting the You Only Look Once (YOLO) model architecture and integrating it with a further head for beam prediction. T2U association is performed by solving a linear assignment problem in the beamforming space using the Hungarian algorithm. Then, we investigated the use of deep learning-based radar multi-target tracking to keep the target identity (and, therefore, its association with the corresponding communication user) over consecutive time steps. Simulation results over three scenarios with diverse mobility patterns show the effectiveness of the method in terms of beam prediction accuracy, radar detection and tracking performance, and probability of correct association.

Simulating accurate high-frequency wireless propagation environments

To provide a realistic environment for the evaluation of the proposed methods, we designed a dynamic multi-modal co-simulation framework integrating vehicular traffic, camera, LiDAR, V2X channel and mmWave radar channel simulators. The proposed co-simulation workflow can be exploited for the generation of massive synchronized multi-modal sensing and communication datasets, which can be employed to train and evaluate learning methods as well as to assess the performance of V2X, ISAC and multi-modal sensing techniques in realistic urban vehicular settings.
The methods developed in this thesis can have a direct impact on application scenarios that can benefit from adaptive low-rank signal denoising and cross-domain spatial-temporal data association. In the next generation wireless communication systems, machine learning-based techniques can provide robust performance under noisy conditions. Furthermore, the developed co-simulation environment can be a grounding framework for the realization of digital network twins of dynamic high-frequency urban environments. The availability of low-power embedded computing boards expressly designed towards machine learning applications can enable prototyping and testing the proposed techniques, while the availability of special purpose integrated circuits for machine learning can allow their deployment in real-world scenarios.

ADVANCED MODEL PREDICTIVE CONTROL OF WATER RESERVOIR SYSTEMS

Raffaele Giuseppe Cestari - Supervisor: Simone Formentin

In this thesis project, we developed advanced MPC (Model Predictive Control) techniques in water resources management. Controlling one or more water basins means managing multiple objectives simultaneously, typically in conflict with each other. The controller must satisfy the water demand of downstream agricultural districts and protect coastal cities from damage caused by flooding during the rainiest periods. The regulator is also responsible for preventing dry conditions of the water basin itself as they can cause an impediment to the normal functioning of river and lake transport, which can lead to indirect economic losses due to lower tourist attractiveness. Furthermore, if the basin corresponds to a hydroelectric power production plant, it is the controller's responsibility to maximize its yield. The combined management of all these control objectives highlights the complexity of the problem. However, it is necessary to add a further complicating factor in regulating this type of system: the disturbance entering the system represented by the inflow to the basin. This quantity is difficult to measure due to the countless sources from which it comes. The inflow depends on meteorological phenomena, large-scale climatic phenomena, and the management of any upstream basins, which may be independent of the management of the basin for which the controller must be designed. During the thesis project, we developed predictive control methodologies in this panorama that could include these delicate aspects in the design phase. Starting from an MPC control architecture with hourly resolution and assuming we have perfect knowledge of the inflows, we demonstrate the approach's effectiveness in the control of Lake Como by outperforming the historical regulation and being able to track the ideal performance of a DDP controller. Subsequently, still working on Lake Como, we removed the hypothesis of perfect knowledge of the inflows and concentrated on designing control architectures that could manage them.
Initially, we proposed a Scenario MPC architecture, in which we optimize the control action based on a set of possible system evolutions derived from scenarios identification on historical data. This solution provides robust control that statistically outperforms deterministic counterparts relying on a single prediction model. Subsequently, we treated the same problem by changing perspective: inserting the best possible model into the prediction model, which we demonstrate is not necessarily the best model identified on the data but is undoubtedly the best model to use in a closed loop. To find this model, we select within a chosen class the one that minimizes a cost given by the combination of the identification of historical data and the performance obtainable using that model on a backtesting window within the control scheme. We demonstrate that the methodology is particularly effective compared to traditional identification and control methods. We validate the performance on Lake Como based on six years of actual data, demonstrating its particular effectiveness in dry periods.
The thesis project continues by trying to answer the question: Is it possible to integrate an optimization of the cost function parameters from a multi-objective perspective into the MPC formulation? To answer this question, we designed a control architecture that automatically calibrates the combination of weights that constitute the cost function through Bayesian Optimization to maximize the quality of the Pareto frontier. At each iteration, we simulate the effect of the control action of a set of identical controllers except for the combination of weights of their cost function. This allows us to obtain a Pareto frontier in simulation, evaluated via multi-objective metrics and progressively improved according to the convergence of Bayesian Optimization. This solution can overcome the limitations of equivalent approaches in the literature since it guarantees convergence of the calibration algorithm. The case study of this methodology is the Red River system, a complex combination of interconnected water reservoirs and irrigation channels. The system's complexity requires using an internal model for the MPC, which can capture the input/output relations of the original system while maintaining the possibility of expressing the Jacobians explicitly to be included in the prediction model effectively. The architecture identified that respects these constraints is the Neural States Space. As far as we know, this represents a unique example in the literature for predictive control in water resources.
Finally, in this thesis, we discuss the further development of advanced predictive control in the context of water resources, applied to Lake Como, aiming to respond to a need: guaranteeing optimal control of the regulated basin even in the event of cyber threats, specifically Denial-of-Service attacks. This need arises from the awareness that these infrastructures represent a privileged target in an international cyber warfare context. Our proposed architecture consists of 2 MPC controllers tackling 2 different optimization problems at the same time. The first primary controller determines an optimal control action that balances system performance with the deviation from the control action calculated by the second controller. The second controller assumes that system control will be handed over to a hacker at the next time step for the remainder of the prediction horizon. Specifically, the aggressor is assumed to maximize the system's damage (i.e., in a cyber warfare context, its goal is sabotaging the targeted infrastructure). Consequently, the attacker's optimization problem is adversarial, seeking to achieve the opposite of the optimal control regulator solution at each time step. Recognizing that it is not trivial to correctly balance the relative weight of the safe component concerning the performance component within the main controller, we find that the best solution is to calibrate this weight online, called the resilience factor, identifying it by training a Hawkes process on historical series of attack occurrences, improving the estimate of the return time of the next attack and adjusting resilience accordingly. We validate the architecture in two attack scenarios. We outperform our counterparts in both cases based on the exclusive focus on performance or security. In summary, this doctoral thesis offers the scientific community an in-depth study of the problems of water resource management and how these problems have been addressed through the development of MPC architectures, with increasing complexity, treating the management of disturbances from a robust and from a control-oriented perspective, addressing the issue of multi-objective optimization and the critical role that regulated basins play as strategic infrastructures of national importance.

DATA-BASED CONTROL DESIGN FOR LINEAR AND RECURRENT NEURAL NETWORK MODELS WITH STABILITY GUARANTEES

William D'Amico - Supervisor: Marcello Farina

Co-Supervisor: Alessio La Bella

In control systems engineering, data-based techniques are gaining increasing attention due to the ever-growing availability of data in many large-scale applications and processes (e.g., chemical industry, metallurgy, machinery, electronics, electricity, transportation, and logistics). These techniques enable the design of high-performance controllers with moderate time and computational effort. They are mainly categorized into direct methods, where a controller is identified directly from data through optimization, and indirect methods, where a system model is first identified from data, based on which the controller is designed.

Among direct methods, Virtual Reference Feedback Tuning (VRFT) has gained wide popularity due to its simplicity and satisfactory performance. However, a major issue in VRFT, as well as in many direct data-driven methods, concerns the lack of explicit stability guarantees for the closed-loop system. VRFT only ensures that the control system approximates a given reference closed-loop model based on available data. If the controller identification is poor and the obtained regulator deviates significantly from the ideal one, the resulting feedback system may exhibit degraded performance or even instability. In the literature, only "a-posteriori" validation tests have been proposed to verify closed-loop stability in the linear setting.

The increasing availability of informative data has also facilitated the adoption of neural networks (NNs) for learning features from data. NNs have gained considerable interest in many engineering fields, including control applications. They can be divided into feedforward neural networks (FFNNs), which reproduce static systems, and recurrent neural networks (RNNs), which incorporate internal loops representing state variables, allowing them to model dynamical systems. RNNs have been successfully applied to real-world control problems in both indirect and direct frameworks. For example, a model predictive control approach based on an identified long short-term memory (LSTM) network has been used to control the cooling system of a large business center. Additionally, controllers based on echo state networks (ESNs) and LSTMs have been learned directly via VRFT to control an experimental electronic throttle body.

Despite the popularity and potential of RNNs, their theoretical properties have been only marginally analyzed in the literature. Since they are nonlinear dynamical systems, it is crucial to characterize conditions that guarantee the stability of their motions, especially when RNNs are integrated into control systems. In this context, incremental input-to-state stability (δISS) plays a key role. This property ensures that the dynamics of a δISS RNN is asymptotically independent of its initialization and imply other common stability properties, such as global asymptotic stability of equilibria. Interestingly, the δISS property can be directly enforced during the RNN training phase. However, related studies focus on open-loop RNNs and do not address the design of stabilizing RNN-based feedback controllers.

Regarding control systems, model predictive control strategies have been proposed for certain RNN architectures, ensuring closed-loop stability if the RNN-based model of the controlled system satisfies the δISS property. Additionally, local asymptotic stability guarantees have been explored for feedback systems employing event-triggered FFNN controllers.

Motivated by the lack of closed-loop stability guarantees by design in VRFT and the limited investigation of RNN stability properties, particularly in closed loop and with reference to δISS or regional stability properties, this dissertation addresses these open challenges.

A novel framework is introduced for applying data-driven methods to control systems described by both linear and RNN models. The primary objective is to develop robust methods for data-driven control design by combining direct and indirect data-driven algorithms to leverage the advantages of both. The VRFT method is used to enforce desired performance directly from data, whereas indirect methods are employed to ensure closed-loop stability by imposing constraints on uncertainty sets or RNN models identified from data. These lead to efficient, unifying optimization problems based on linear matrix inequalities (LMIs). The advantage of using VRFT despite the availability of uncertainty sets or RNN models is that VRFT provides a computationally lightweight and flexible framework for performance-oriented controller design. Analytic approaches for imposing performance specifications such as settling time or transient behavior are not readily available for RNNs and can be difficult to enforce even for uncertain linear systems. However, with suitable choices of the reference closed-loop model, a VRFT-based cost function allows for the direct imposition of such specifications from data.

Initially, the focus is on applying VRFT to linear time-invariant single-input single-output discrete-time systems affected by measurement noise. In this regard, a novel LMI reformulation of the VRFT method is developed to ensure robust closed-loop stability by design through a scenario-based polytopic set membership identification. Additionally, extensions for handling disturbance rejection, input saturations, and alternative ellipsoidal uncertainty sets are proposed.

The study then shifts to RNN models due to their exceptional modeling capabilities. A novel scenario-based method for accurately selecting the RNN model class in presence of noisy data is proposed, outperforming standard selection approaches in the literature.

Subsequently, the stability properties of general RNN classes, including ESNs and shallow neural nonlinear autoregressive exogenous (NNARX) models, are examined. A novel condition for δISS is introduced, which is proved to be less conservative than existing conditions in the literature. Additionally, alternative less conservative local stability conditions for the origin of RNN systems are formulated, along with procedures for maximizing the estimated subset of the basin of attraction.

Finally, based on identified RNN models or uncertainty sets for the system, the stability conditions are enforced in the form of LMIs for specific RNN-based control architectures, leading to controllers that guarantee δISS or local stability properties for the closed-loop system. The integration of these conditions into unifying LMI-based optimization problems is also explored, where performance metrics are enforced, possibly via VRFT.

Successful applications to simulated case studies, including realistic examples such as the pH neutralization process and a water-heating benchmark system, underscore the practical relevance of the proposed methodology.

INNOVATIVE TECHNOLOGICAL SOLUTIONS FOR MICROWAVE AND SUB-THZ PASSIVE DEVICES

Cristina D'Asta - Supervisor: Gian Guido Gentili

Co-Supervisor: Giuseppe Macchiarella

In modern wireless communication systems, new challenges are always arising. Technology is constantly advancing to follow the needs of the market, that is richer and richer of new communication standard and protocols operating in a variety of frequency ranges. For this reason, passive hardware such as filters and antennas, which are essential components in every communication system, must become more and more efficient, selective and low-cost. Waveguides are typically the first choice when a good performance is desired, thanks to their low losses and high power handling capabilities. However, their fabrication is traditionally very expensive, due to their bulky size below the *Ku* band, corresponding to 12-18 GHz, but also the serial nature of the metal milling process. Conversely, in the sub-THz range the high operational frequency calls for very precise manufacturing with a small feature size in the order of micrometers, making fabrication of components very challenging and expensive. This doctoral project focuses on exploring different approaches to size reduction below the *Ku*-band and enabling mass production of sub-THz passive devices at a reasonable cost.

Choosing the right fabrication techniques can also contribute to overall cost reduction of components, therefore, another important objective of this research is proposing new and innovative manufacturing techniques for passive devices manufacturing. Processes such as additive manufacturing and micromachining are presented, explained and applied to filter and antenna manufacturing at different frequencies. In particular, additive manufacturing of ceramic materials is proposed for applications below 10 GHz, instead, soft lithography, specifically replica molding of polymers, is applied to large-scale fabrication of sub-THz passive devices based on waveguide.

In the first part of the thesis, various filter synthesis techniques introducing transmission zeros are explained and applied to waveguide filters. Here miniaturization is achieved by improving the selectivity without increasing the number of resonators and therefore the size. The main results comprise:

1) A new design method for TE_{201} waveguide singlet applied to extracted-pole filters, based on the creation of a database of *ad-hoc* normalized charts to be reused at need for setting an initial combination of physical parameters, able to yield an equivalent circuit already very close to the desired one. Minor rapid adjustments allow to complete the design fast, as opposed to dimensioning approaches fully based on optimization, that, for such a sensitive structure, would rarely reach a good result without a suitable initialization. Keeping a database of charts could be the basis for the development of an automatized design procedure.
2) The first ever physical implementation, to our knowledge, of a path filter in waveguide technology with a transmission zero also in the direct path. The prototype filter, working in *Ku* band, has been designed by exploiting a new simulation method based on 2.5D finite element method, that solves the waveguide circuit in 2D and applies radial FEM across the height of the waveguide for the areas around the coaxial connectors and the screws. The excellent matching between the simulation and the measurement of the prototype, built at Polimi through CNC-milling, proved the accuracy of the employed simulation method and also the concept of path filter itself.

In the second part of the thesis, a step further in the miniaturization of filter is taken thanks to the usage of a high-permittivity material, specifically alumina, as part of the filter structure. A dielectric-filled waveguide filter is proposed and built through additive manufacturing, specifically lithography-based ceramic manufacturing. The linear size reduction is proportional to the square root of the relative electrical permittivity, which in this case is a factor of around 3. Moreover, by using additive manufacturing, we have the utmost geometrical flexibility and we also avoid the need for an *ad-hoc* mold, as it instead happens when traditional ceramic fabrication is employed. This results in a significant cost reduction, as molds are typically very expensive and can constitute even 70% of the total manufacturing cost of ceramic parts, but also in a straightforward implementation of any design change of the geometry without the need for further equipment besides the printer itself. We remark that additive manufacturing, especially in the field of ceramics, can be a true breakthrough for the initial prototyping of devices, but also for production in low volumes.

For our project, specifically, we studied the feasibility of a C-band ceramic-filled waveguide filter, partnering with the company Lithoz, which is one of the most prominent ceramic manufacturers exploiting 3D printing. We identified metallization as the main bottleneck of this technology, due to the extreme smoothness of the printed blocks. Various techniques were attempted, such as conductive silver polish, copper electroless plating and evaporation in the clean room, in the end none of them gave satisfactory results for good confinement of the fields. Nevertheless, with a simple morphological analysis of the printed blocks, we noticed that the process itself was very repeatable and also quite accurate, indeed a simulation of the real profile yielded fluctuations that can arguably be recovered with tunable elements in the structure.

The last project, carried out in collaboration with Huawei Technologies Italy and Università di Pisa, concerns the fabrication of a slotted waveguide antenna, working in G-band going from 110 GHz to 300 GHz, which will be soon standardized for future communications systems. The sub-THz frequency range truly poses serious challenges in manufacturing of high-performance passive components, as the waveguide size is too small for metal milling, which cannot be used due to physical limitations, bad accuracy and limited repeatability, as well as extremely high costs. Micromachining comprises a family of techniques coming from different application fields, like microelectronics and microfluidics, and has recently been proposed for application also in the microwave field, specifically for the sub-THz range. Silicon micromachining and thick SU-8 techniques have already been applied successfully to waveguide circuits manufacturing for this range, however their cost is still somewhat high, mainly due to the necessity of clean room equipment for all the process steps.

In this work, replica molding is proposed as a potential lower-cost alternative for large-scale production of geometrically complex sub-THz waveguide circuits. Our demonstrator device is a slotted waveguide antenna array spanning five different layers. We identified replica molding as a valid alternative to more established technologies for waveguide manufacturing at high frequency. Indeed, replica molding has been widely applied in microfluidics, but, to our knowledge, never fully employed in waveguide circuits manufacturing. It consists in creating a mold, made of photoresist through standard lithography, to be reused indefinitely to produce various replicas of the device in polymeric material. An electromagnetic measurement of the fabricated prototype has not been completed yet, but, from the morphological analysis performed on the prototypes, an excellent compatibility with the antenna sensitivity has been observed. In fact, replica molding can be a true enabler for mass production of waveguide-based components for sub-THz range, giving the proposed process flow a very high chance of future industrialization.

BATTERY STORAGE SYSTEM MANAGEMENT AND OPTIMIZATION IN HETEROGENEOUS ENVIRONMENTS

Massimiliano Maurizio De Benedetti

Supervisors: Luca Bascetta, Maria Prandini - Tutor: Simone Garatti

The proliferation of Distributed Energy Resources (DERs) and the electrification of transport systems with the introduction of Electric Vehicles (EVs) is posing several challenges related to their coordination, control, and optimal operation.
In particular, a key element is the Battery Energy Storage System (BESS) which needs to be controlled so as to fulfil the requirements of different DERs and EVs applications. Indeed, both Virtual Power Plant (VPP) and Electric Vehicle Routing (EVR) require the coordination of multiple units to solve an optimization problem where different contrasting objectives are traded-off.
Given the large size of the problem and the consequent high computational effort involved, the constraints on the time available to obtain a solution, and the fact that the information needed even just to formulate the problem is available only locally to the involved units, approaches resorting to a distributed architecture where computational capabilities of the units are exploited appear the natural solution to go. This in turn entails that BESSes have to be able to solve locally a part of the global optimization problem related to VPP and EVR.
This thesis addresses battery operation and aims at formulating an efficient algorithm for constrained sequential battery charge/discharge planning that is easy to integrate within a distributed control scheme so as to optimize the joint operation of multiple energy storage systems in diverse scenarios and applications.
The importance of energy storage systems in the broader context of decarbonization, industrial processes, and the energy transition is discussed in this thesis, showing their versatility spanning from stationary to electric mobility applications, and emphasizing their pivotal role in shaping a sustainable energy landscape.
In particular we establish the foundation for the developments of the optimization problems associated with energy storage systems in both stationary and electric mobility applications according to an unifying perspective, thus providing a theoretical framework for understanding the interconnections between various applications.
Then two different use cases, related to a stationary and a mobility related application, are described to showcase the versatility and effectiveness of the proposed Constrained Sequential Battery Planning (CSBP) algorithm.
The first use case focuses on the optimization of energy self-consumption in a residential setting, demonstrating promising results through validation on real data. The second use case addresses automated inspection of a photovoltaic plant using robots that can autonomously recharge their battery. %, with a market-based approach.
The adaptability of the CSBP algorithm to diverse applications is highlighted through its vertical customization for each use case. Also, simulation results obtained in the use cases indicate significant advantages of the CSBP algorithm over reference optimization strategies. Notably, a 40% reduction in energy costs for a residential customer showcases the algorithm's potential impact on the proliferation of Distributed Energy Resources (DERs).
The reliability introduced by the CSBP algorithm in battery charge/discharge planning is emphasized, with implications for robust DER aggregation and potential contributions to grid services markets.
In the context of renewable plant inspection, the algorithm demonstrates the equalization of mechanical depletion and reduction in battery degradation for Unmanned Ground Vehicles (UGVs) and Unmanned Aerial Vehicles (UAVs), suggesting the potential for full automation of such tasks.
Looking ahead, the thesis proposes intriguing avenues for future research. The evaluation of the CSBP algorithm in conjunction with Virtual Power Plant (VPP) optimization algorithms is suggested, along with the exploration of tailored forecasting methodologies for enhanced self-consumption maximization. In the electric mobility domain, the integration of the CSBP algorithm with heuristic approaches for solving the Electric Vehicle Routing Problem (EVRP) is proposed.
In conclusion, the CSBP algorithm emerges as a robust and versatile tool for optimizing energy storage systems in heterogeneous scenarios, with far-reaching implications for sustainable energy applications. The thesis not only contributes to the literature on optimization algorithms for the management of energy systems involving storage units but also presents tangible outcomes with potential real-world impact.

MACHINE LEARNING AS A NETWORK MANAGEMENT PRIMITIVE: FROM END-TO-END OPTIMIZATION TO ATOMIC NETWORK FUNCTIONS

Nicola Di Cicco - Supervisor: Massimo Tornatore

Machine Learning (ML) is rapidly becoming a central technology in future communications networks, with applications including (but not limited to) traffic engineering, traffic classification, traffic prediction, intelligent transport protocols, fault management, physical layer modeling, and the list goes on. The promise of ML for networks is an ambitious one: by leveraging the abundant amount of data produced by modern communication networks, we can automate a broad set of tasks that require heavy human intervention and specialized domain knowledge.

Unfortunately, despite the extensive research efforts concentrated in the past few years, we are still far from harnessing the full potential of ML for network management. Contemporary literature has extensively investigated algorithms and design principles for producing well-trained models from historical data and/or experiences, e.g., identifying which ML methodologies are the more effective for attacking specific network management problems, or which features are the most important for solving a certain task. These contributions, while fundamental, comprise only a first step towards a complete production pipeline. Deploying, using, and maintaining ML-based network management solutions poses nontrivial research and engineering challenges, orthogonal with respect to training a ML model, that are presently not yet thoroughly addressed. This hinders widespread adoption of ML by network operators, who prefer relying on classical, battle-tested network management solutions.

To take a step toward solving this fundamental problem, this Thesis identifies five core challenges in ML for network management: 1) Generalizability to data beyond training, 2) Adaptability to dynamic network environments, 3) Reliability, in terms of providing performance guarantees after model deployment, 4) Data efficiency, for minimizing the amount of labor required for training and updating models, and 5) Performance, i.e., ML-based solutions must tangibly improve over conventional methods to be worth considering. We address these challenges through multiple representative network management applications: online and offline network optimization with Reinforcement Learning, focusing on generalizability and performance; ML-based hardware fault classification in microwave networks, focusing on data efficiency and reliability; Continual in-network ML, focusing on data efficiency and adaptability; and intent-based networking with Large Language Models, focusing on performance and data-efficiency. We quantitatively validate the practical effectiveness of our proposed solutions through extensive comparisons against the state of the art, and by leveraging novel real-world datasets, which we make publicly available.

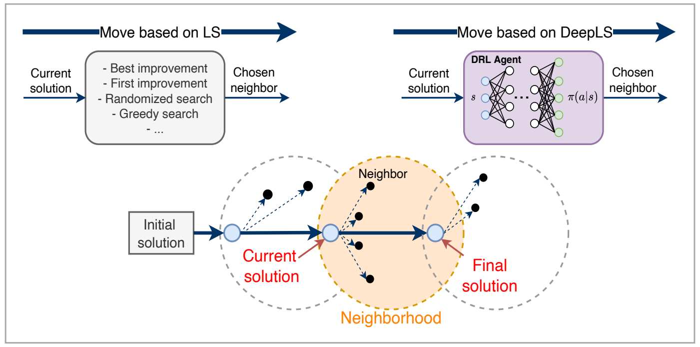

Fig. 1 - Local Search with Deep Reinforcement Learning (DeepLS). We augment a conventional algorithm for network optimization, Local Search, with lightweight intelligence. Our approach results in a generalizable algorithm with similar or better performance than the current state of the art.

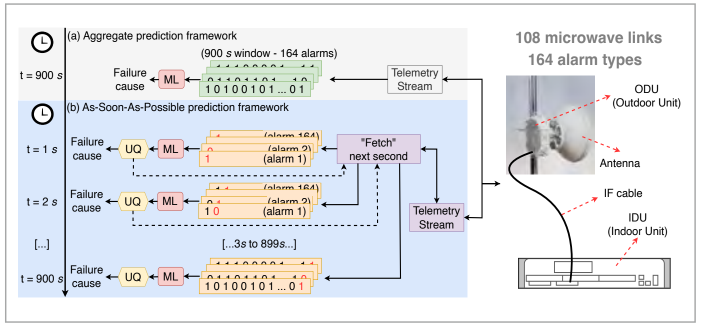

Fig.2 - As-Soon-As-Possible (ASAP) fault classification in microwave networks. We propose using Machine Learning to analyze a telemetry stream for predicting hardware faults. Our system leverages Uncertainty Quantification (UQ) to supply a prediction only when its certainty exceeds a user-provided safety threshold. In this way, we can integrate data-driven inference with fault management in a reliable way.

RESOURCE ALLOCATION AND SCHEDULING PROBLEMS IN COMPUTING CONTINUA FOR ARTIFICIAL INTELLIGENCE APPLICATIONS

Federica Filippini - Supervisor: Danilo Ardagna

The problem of optimizing the execution of Artificial Intelligence (AI) and Deep Learning (DL) applications in the Computing Continuum gained remarkable popularity in recent years, due to both the widespread adoption of AI in real-life scenarios and the challenging environment introduced by a distributed Edge-to-Cloud paradigm. We tackled the resource selection, scheduling and placement problem both from a design-time and runtime perspective, considering, on one hand, AI inference applications characterized by complex workflows with multiple heterogeneous components and, on the other hand, resource-demanding DL training jobs executed on public or private GPU-accelerated clusters.

Performance Models

Predicting execution times of computing tasks across different resources is essential for optimal placement, since it enables accurate estimates of their computational requirements and the expected costs related to the Computing Continuum resources usage. We addressed this challenge using white-box analytical methods based on queuing theory, guaranteeing fast computation and the possibility of considering components interference due to co-location, and black-box Machine Learning (ML) models trained on profiling data for prediction in unforeseen conditions. We developed the OSCAR-P auto-profiling framework to test applications on diverse hardware and build individual regression models leveraging the aMLLibrary auto-ML tool to predict the response time of each component-resource pair. Models achieve a Mean Absolute Percentage Error (MAPE) below 10% for interpolation and 20% for extrapolation. Moreover, we leveraged aMLLibrary to train models that estimate the execution time of Distributed Deep Learning training applications developed and executed on Ray clusters, achieving a MAPE between 3 and 15% for both interpolation and extrapolation.

Design-Time Optimization

Deploying AI applications in the Computing Continuum is complex, requiring careful Resource Selection and Component Placement (RS-CP) to minimize costs while meeting hardware, network, privacy, and QoS requirements. At design time, according to the expected value of input workload, application owners must select optimal resources in each computational layer, determine the best Deep Neural Network (DNN) component deployment across layers, and possibly leverage DNN partitioning to facilitate offloading, minimizing the placement costs under the aforementioned constraints. We proposed a Mixed-Integer Non-Linear Programming (MINLP) formulation for this problem and developed SPACE4AI-D, leveraging Random Search and heuristic methods as Local Search, Tabu Search, Simulated Annealing and Genetic Algorithms to determine good-quality solutions in a reasonable time. SPACE4AI-D yields 28–58% cost reductions against the state of the art, with a deviation between real costs in a prototype environment and predicted costs lower than 13%. To the best of our knowledge, our work represents one of the first attempts to effectively address resource contention, by adopting queuing theory to model components response times, while existing tools usually focus on single application instances. Similarly, considering multiple candidate DNN partitioning points and identifying the optimal deployment as part of the design-time optimization represents a novel contribution with respect to the existing literature.

Runtime Management for Inference Applications

In real-world applications, design-time decisions must be continuously adjusted due to workload fluctuations, requiring fast and efficient runtime reconfiguration to avoid resource saturation or underutilization. While design-time tools take minutes to optimize placement, runtime tools must adapt in seconds to support running applications. We adopted both heuristic methods (fast and effective, but requiring deep domain knowledge and sometimes restrictive theoretical assumptions) and Reinforcement Learning (RL) methods. These automatically adapt to variations in the environment conditions, but require a long time to learn effective policies and incur in severe constraints violations or overspending while exploring the possible actions. Our approach accelerates RL training by leveraging design-time knowledge, combining the strengths of both techniques for optimal resource management.

SPACE4AI-R uses Random Search and Stochastic Local Search to quickly optimize resource scaling, DNN partitioning, and component placement, achieving a 100x speed-up over SPACE4AI-D by handling pipelines with up to 15 AI components in less than 1.5s while reducing costs by 2–36% over static approaches and 10–40% over state-of-the-art scaling methods that aim at keeping the resources utilization within a predefined interval.

FIGARO is an RL-based approach that automatically scales resources to adapt the current deployment. It exploits imitation learning to develop an initial Deep Q-Learning-based policy that mimicks SPACE4AI-R, later refined via interactions with a simulated environment. The initial offline training reduces the cumulative number and entity of constraints violations by 2–4 orders of magnitude, and decreases by 25% the number of required training iterations.

Runtime Management for Training Applications

Training DL applications is challenging: jobs are resource-intensive even with GPU acceleration, and high-performance servers are characterized by considerable costs, both in private clusters and in the public Cloud. Optimizing Resource Selection and Job Scheduling (RS-JS) is essential for both Cloud users (to minimize execution costs) and Cloud Service Providers (CSPs) (to reduce energy consumption while meeting applications due dates).

Focusing on the Cloud users' perspective, we considered a multi-node cluster where servers can be provisioned on-demand and configured with heterogeneous VMs. Jobs can be executed concurrently, sharing all computational resources except GPUs. In this context, the RS problem consists in identifying, for each job, the optimal VM type and number of GPUs to minimize execution costs and due date violations; the JS problem involves determining which jobs to run if resources are not enough to execute all of them concurrently, and performing jobs-to-nodes assignments effectively partitioning the selected resources. We tackle these problems jointly and online, without prior knowledge of future job submissions. We developed a MILP formulation and a heuristic algorithm based on Randomized Greedy and Path Relinking, which sorts jobs according to the distance from the due date to avoid starvation and assigns to each job the cheapest configuration that allows to meet the due date, if possible (otherwise, the fastest one to limit the delay penalties). A best-fit jobs-to-nodes assignment limits idle resources. We achieve a 23–97% average cost reduction against the state of the art, handling up to 100 nodes and 450 jobs in less than 7 seconds with a deviation between real and predicted costs lower than 5%.

Under the CSPs perspective, we relaxed the assumption that jobs run for a predefined number of epochs and considered stochastic execution profiles to model termination criteria based on, e.g., the accuracy level. Guided by a solution model that identifies the optimal schedule for a single job, we proposed a MINLP formulation to minimize the expected energy costs and designed a STochastic Scheduler (STS) to select the best type and number of GPUs to execute jobs within the due date, incorporating GPU sharing. The resource profiles designed by STS start jobs execution with a low-power configuration, usually less expensive, and accelerate the training by progressively increasing the resources as the due date approaches. STS achieves 38–80% cost reductions on average against both state-of-the-art methods and an adapted version of our Randomized Greedy algorithm, while the adoption of GPU sharing guarantees additional advantages, further reducing the costs by 19–29% on average.

PERSUASIVE CONVERSATIONAL AGENTS TO FOSTER SUSTAINABLE BEHAVIOURS: DESIGN, EVALUATION, AND TECHNOLOGY

Mathyas Giudici - Supervisor: Franca Garzotto

Individual energy consumption habits are crucial for sustainability and mitigating climate change. Residential electricity consumption accounts for a significant portion of global energy use and will rise with increased electrification as a sustainability strategy.
Interactive technologies and Human-Computer Interaction research play a key role in offering technological solutions, raising awareness, and guiding people toward sustainable behaviors. Interactive computer systems that alter attitudes or behaviors without force are known as Persuasive Technology. These technologies are widely used in various domains (e.g., healthcare, business) and have proven effective.
In environmental sustainability, real-time energy consumption visualizations promote responsible behavior, with long-term exposure showing lasting benefits. Persuasive technologies aid in managing energy tariffs and workplace energy-saving behaviors. Additionally, environmental games and social robots foster sustainable habits.
Conversational Technologies, which allow users to interact with digital tools via natural language, are increasingly common. These technologies are integrated into phones (e.g., Google Assistant, Siri), cars, and customer service. Home automation further enables monitoring appliances, optimizing usage. The release of AI tools — Large Language Models (LLMs) — has revolutionized conversational technology applications, such as code development, creative writing, and text summarization, paving the way for integrating LLMs into smart home automation.
Conversational technology designed to persuade behavior change is called Persuasive Conversational Agents (PCA). PCAs are effective in promoting healthier habits like regular exercise, reducing snacking, or limiting alcohol consumption. However, their role in environmental sustainability is still novel and underexplored.
My thesis contributed to this emerging field by addressing the following research questions:
(R1) What is the current literature on Persuasive Conversational Agents for environmental sustainability?
(R2) How effective are Persuasive Conversational Agents in promoting sustainable electricity usage at home?
(R3) What design principles enhance the effectiveness of Persuasive Conversational Agents in home settings?
(R4) What opportunities and challenges exist in utilizing LLM-based Persuasive Conversational Agents for sustainable electricity use?
(R5) How can Persuasive Conversational Agents foster collective energy sustainability in communities?
To address RQ1, I conducted a systematic review of the literature on Persuasive Conversational Agents for environmental sustainability. This review examined design elements, persuasive techniques, sustainability concerns, and empirical assessment methods. The findings outlined a research agenda and a checklist to guide future studies.
For RQ2, we conducted an exploratory study to determine which persuasive techniques are most effective in conversational technology for sustainability. The study tested three persuasive strategies in chatbot interactions with 531 participants to assess their impact on energy-saving intentions and user experience. While no significant behavioral changes were observed, notable differences in user experience emerged.
Building on this, we examined the effectiveness of multiple persuasive strategies compared to a chatbot using a single strategy. A custom dialogue manager facilitated experimentation with 29 participants in a between-subjects study. Participants perceived a significant difference in parasocial interactions and dialogue quality, favoring multi-strategy persuasion, though no differences in usability or effectiveness were found. Qualitative feedback revealed specific user intents the chatbot failed to address.
To explore RQ3, we investigated how digital tools and conversational technologies engage individuals in sustainable behaviors at home. A two-hour focus group with seven participants interested in home automation and sustainability identified key factors for designing persuasive agents. This led to the refinement of CANDY, a framework for designing sustainability-focused Conversational Agents.
Additionally, we explored how multimodal Conversational Agents integrated into Smart Mirrors could encourage lower energy consumption. We developed *Leafy*, an agent employing gamification strategies for short- and long-term sustainability engagement. Furthermore, we built *Green*IFTTT, an application leveraging GPT-4 to create and maintain home automation routines. A lab evaluation with 13 participants assessed usability, engagement, and effectiveness, revealing that the tool was practical, entertaining, and easy to use.
During my Ph.D., the introduction of LLMs transformed conversational tools and human-AI interaction. To explore RQ4, we evaluated four LLMs in ecological sustainability through two assessments. First, we compiled reliable sources and analyzed how well model-generated content aligned with them. While ChatGPT performed best at the time, no statistical differences were found among models. Second, we tested ChatGPT's responses to 167 sustainability-related questions gathered from 75 people. Experts evaluated answers based on fluency, coherence, correctness, and logic, finding ChatGPT highly reliable for general sustainability queries.
We further investigated LLMs in smart home automation, specifically within the HomeAssistant framework. Our research highlighted GPT models' effectiveness in generating accurate automation routines compared to alternatives like Llama, Code-Llama, and Mistral. While LLMs successfully understood complex commands and created valid JSON outputs, challenges such as syntax errors persisted. Despite minimal quantitative differences between "green" and "non-green" prompts, qualitative feedback showed a positive shift toward sustainability. A subsequent user study with 56 participants demonstrated high usability and engagement compared to rule-based chatbots.
To address RQ5, we collaborated with the University College London's Interaction Centre on a project aimed at shifting digital technology from individual persuasion to connecting people with environmental policies in energy communities (groups sharing sustainable energy sources). We extended the project's platform with a conversational interface and conducted a field study with four groups of four people each. Interviews and Thematic Analysis provided insights into the role of conversational agents in community-based energy initiatives.
As an extension of this collaboration framework, we explored LLM-based multimodal, multi-user conversational interfaces for managing group communication and collaboration; a need emerged from our energy communities research. Given the complexities of designing multi-user conversational agents, we modeled our approach on a well-known psychology scenario — participants read documents and discuss a murder mystery. The experimental design and evaluation yielded unique empirical results, making a significant scientific contribution to Human-Computer Interaction research.

DATA-DRIVEN APPROACHES FOR MANAGING RESOURCES OF CLOUD, EDGE, AND HIGH-PERFORMANCE COMPUTING SYSTEMS

Bruno Guindani - Supervisor: Danilo Ardagna

Introduction

The presence and importance of software are tangibly increasing in modern society. This trend has become even more pronounced with the recent popularity of Artificial Intelligence (AI) tools for everyday tasks. By now, humanity heavily relies on software for many crucial endeavors, ranging from health to transportation. For instance, with the outbreak of the COVID-19 pandemic, molecule simulation programs for drug discovery became essential to medical research to find cures for the coronavirus infection.

Due to this growing pervasiveness and pivotal importance, the complexity of the software used in several fields has skyrocketed over time. At the same time, the hardware needed to support such software increased in size, energy demand, performance, and complexity. Three significant examples of sophisticated architectures for modern software are cloud computing, edge computing, and High-Performance Computing (HPC).

Main challenges

An essential part of running complex software on machines with sophisticated hardware architectures is choosing the correct configuration regarding software settings and hardware resources available (CPU, memory, disk, network, etc.). However, a poor choice for such settings can lead to application under-performance or additional costs for the end users. The impact of incorrect configuration on the execution cost or the performance of a piece of software is potentially paramount. However, the complexity mentioned above of both software and hardware has consequences on the ability to perform the tuning of these applications or predict their performance from the input settings. In the past, Information and Communication Technology (ICT) systems were often represented by white-box models, a term that refers to models in which the internal structure and workings are visible and understandable from the outside. Popular examples of white-box representations were Petri nets and queueing networks. However, when applied to the most modern software and hardware systems, these models often rely on unrealistic or unverifiable assumptions, preventing the application of these kinds of explicit formulations.

Black-box data-driven tools are ideally suited to tackle such optimization and performance prediction issues. The literature defines black-box methods as techniques that attempt to infer input-output relationships of systems of which there is no available knowledge on their internal workings. They usually employ statistical models trained with data collected from the system under scrutiny.

Bayesian Optimization (BO) is one of the leading state-of-the-art techniques for black-box optimization. Most commonly, it starts by choosing and evaluating a handful of starting points and then using them to fit a Gaussian Process (GP) model. The fitted model estimates the function value at each point and the uncertainty around the prediction. BO represents a perfect candidate for finding the optimal configuration of complex ICT systems.

When using BO as a decision technique to minimize a certain cost metric for the purpose of software-hardware optimization, it cannot directly provide an estimate of another performance metric of the optimized application at the same time. Indeed, the latter is a regression problem on a metric that may or may not be dependent on the minimized cost. Some approaches from the literature employ independent models on both, but this is an unrealistic assumption for many practical scenarios. This is a relevant issue for both researchers and companies. Indeed, the capacity to reliably anticipate the execution time and performance metrics of software is paramount in numerous software-related domains. In industrial contexts, where Service Level Agreements (SLA) with clients are established, the apt allocation of resources based on anticipated workload assumes critical importance.

This necessity also extends to tasks such as job scheduling and cluster capacity planning, where performance modeling facilitates the effective management of data center infrastructures in real-time operations. In HPC systems, additionally, accurate and reliable thermal prediction models would enable more efficient utilization, such as direct integration with HPC schedulers.

Machine Learning (ML) is another category of powerful black-box tools that can provide the required performance predictions with reliable accuracy. According to the literature, ML is the prominent category of black-box approaches for performance analysis. ML tools range from simple linear regression models to complex graph-structured neural networks, but at their core, they all represent different kinds of data-driven statistical techniques. They generally require amounts of data that depend on the type of ML model under consideration, and they often need careful parameter tuning of their own to reach their highest potential in terms of prediction quality. Under these circumstances, they can often estimate software execution time or other performance metrics with minimal error.

Main contributions

In this dissertation, we present several data-driven techniques for the optimization and performance modeling of computer systems such as cloud and HPC centers. Our goal is twofold: we first study the best tool to predict the performance of a large spectrum of applications, then apply such models to the optimization of these applications.

One of our main contributions is the MALIBOO (MAchine Learning In Bayesian OptimizatiOn) technique and its parallel extensions. They are BO-based algorithms to optimize the execution of applications and jobs from applications running in cloud computing and HPC settings, and they integrate ML techniques to ensure that the chosen application configurations respect given constraints, whether on time, accuracy of results, or any other black-box metric. The novelty of the approach lies in the synergy and interaction between BO and ML to conduct constrained black-box optimization in a wide variety of computing environments, joining the prediction capabilities of ML with the iteration-efficient decision-making of BO.

Another significant contribution in the performance prediction field, which is the first step of the aforementioned two-part process, is the development of aMLLibrary and its applications. aMLLibrary is an open-source, high-level Python package that allows the parallel building of multiple ML regression models. It is focused on performance modeling and supports several methods for feature engineering/selection and hyperparameter tuning. The library also constitutes the last component of the OSCAR-P pipeline.

OSCAR-P is an auto-profiling tool designed to automatically test serverless application workflows running on multiple hardware and node combinations in cloud and edge environments. This tool collects relevant profiling information on the execution time of the individual application components. These data are later used by aMLLibrary to train ML-based performance models and are the final artifact produced by the OSCAR-P tool.

In this dissertation, we also explore several approaches to temperature prediction for HPC systems as a performance prediction case study of particular interest. Specifically, we focus on both node-level and holistic models, ranging from a room-level Graph Neural Network (GNN) spatial model to simpler compute node-level models.

PHYSICS-INFORMED ONLINE LEARNING OF GRAY-BOX MODELS BY MOVING HORIZON ESTIMATION AND EFFICIENT NUMERICAL METHODS

Kristoffer Fink Løwenstein - Supervisor: Daniele Bernardini

Co-Supervisors: Lorenzo Fagiano, Alberto Bemporad

Advanced human-engineered systems will inevitably play a more dominant role in the future and are expected to operate with increased autonomy benefiting society as a whole. Major challenges lie ahead to ensure that such complex systems operate in a safe manner while satisfying strict performance requirements. A compelling idea is to use optimization-based methodologies such as Moving Horizon Estimation (MHE) and Model Predictive Control (MPC) for autonomous decision-making as they allow incorporating complex objectives and critical measures through a cost function and constraints on system states and inputs. A fundamental requirement for MHE and MPC is a reliable, but computationally light, model enabling online evaluation. Acquiring such a model and, especially, adapting it to temporal variations of the underlying system dynamics, surrounding environment, or part-to-part variations found in any system is a challenging endeavor. This issue is addressed by introducing parametric gray-box models (see Fig. 1): Relying on physics-based modeling facilitates the integration of function approximators such as Feed Forward Neural Networks (FFNN) or Recurrent Neural Networks (RNN) of rather limited size as the data-driven component in the gray-box models, and also benefits from increased model interpretability compared to pure black-box models. By deploying these gray-box models in an MHE scheme (see Fig. 2), a novel MHE framework for physics-informed learning is introduced. The learning is made safe through constraints consistent with physical laws seamlessly integrated into the optimization problem. The proposed MHE scheme is inherently suitable for online application and provides concurrent state estimation and model adaptation with uncertainty quantification. Thus, the proposed MHE framework can support any advanced decision-making process that relies on accurate state and parameter estimates, being autonomous or human-assisted, extending beyond just MPC. Furthermore, an offline training algorithm capitalizing on the proposed MHE scheme tailored for gray-box models is presented, which provides a viable alternative to classical training algorithms for such a class of models. The utility of the methodology is demonstrated through two numerical experiments, showing promising results for estimation, prediction, closed-loop Nonlinear MPC (NMPC), and online model adaptation.

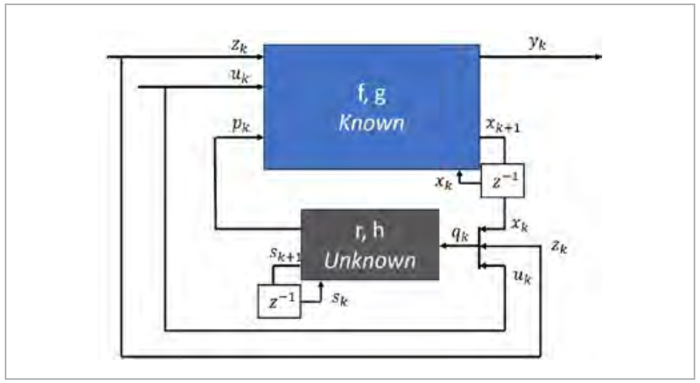

Fig. 1 - Gray-box model with a dynamical subsystem. The unknown part now has an internal state, therefore RNNs are a suitable modeling choice for the data-driven component.

For real-time implementation of MHE and MPC, the underlying optimization algorithms are of paramount importance. In linear MHE and MPC the problem boils down to a Quadratic Program (QP) while Sequential Quadratic Programming (SQP), a cherished solution procedure for Nonlinear MHE (NMHE) and NMPC, is based on solving a series of related QPs. To meet this demand, QPALM-OCP is introduced: QPALM-OCP is a Proximal Augmented Lagrangian Method (P-ALM) tailored for the multi-stage structure arising in MHE and MPC. The algorithm relies on a semi-smooth Newton method to solve the inner ALM problem allowing multiple active set changes between iterates and benefits from warm-starting. Due to specialized low-rank matrix modifications, the iterates remain cheap while providing fast execution times. Contrary to conventional Quadratic Programming solvers (QP solvers), QPALM-OCP comes with guarantees of R-linear convergence to a stationary point of nonconvex QPs, making it an ideal candidate for SQP methods, especially when using exact second-order information. A prototype implementation, despite being non-optimized, outperforms state-of-the-art general-purpose QP solvers and is slightly faster than state-of-the-art Optimal Control Problem (OCP)-specific solvers in numerical experiments (see Fig. 3), demonstrating great potential for further development.

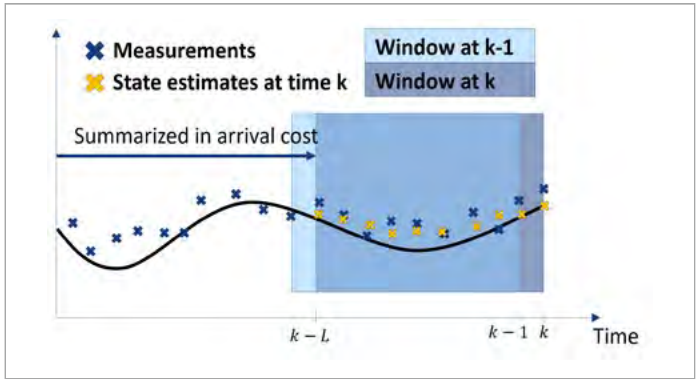

Fig. 2 - Illustration of the Moving Horizon Estimation principle. In MHE, the state and parameter estimates are found by optimizing over a fixed-sized window of past data. The estimation window is shifted in a receding horizon fashion while the so-called arrival cost tries to preserve the knowledge outside the estimation window.

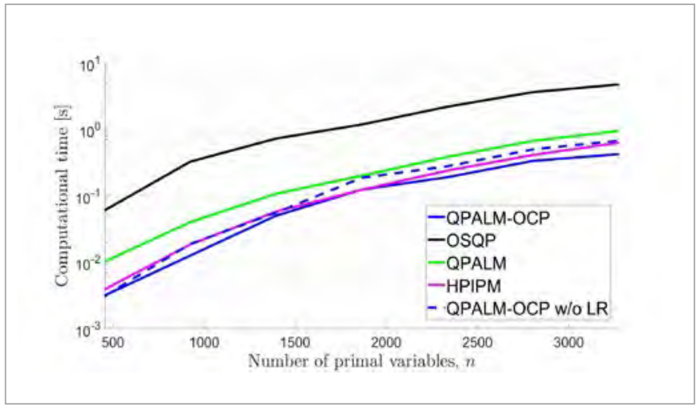

Fig. 3 - Comparison with state-of-the-art QP solvers for an increasing number of primal variables.

CONTROL OF LARGE-SCALE MLD SYSTEMS VIA MULTI-AGENT REFORMULATION AND DECENTRALIZED OPTIMIZATION

Lucrezia Manieri - Supervisor: Maria Prandini

This thesis addresses the optimal control of large-scale engineering systems that can be found in a variety of domains and, in particular, in the energy sector. We consider systems involving physical and logical components, described by linear equations and inequalities involving both discrete and continuous state and input variables. If also operational constraints and specifications are expressed in terms of linear inequalities, then one can adopt the Mixed Logical Dynamical (MLD) system modeling framework and translate the optimal control problem into a Mixed Integer Program (MIP), which is linear (hence a MILP) if the performance index is also linear.

Since the complexity of a MIP is intrinsically combinatorial and grows exponentially with the number of discrete decision variables, optimal control of an MLD system becomes computationally intensive or even prohibitive when its size increases, thus calling for appropriate resolution strategies. Additionally, if the MLD is affected by uncertainty only known from data, then one needs to devise solutions that are determined based on observed uncertainty instances but yet robust against unseen ones.

In this work, we propose an optimization framework that faces the combinatorial complexity arising in the optimal operation of large-scale MLD systems by taking advantage of the structure (if any) of the problem at hand to decompose it in smaller partially coupled sub-problems and recover computational tractability via decentralized solution-seeking schemes.

The decomposition strategy aims at disclosing the (hidden) partially separable structure of a linearly constrained optimization program via manipulation of its constraint matrix. Similarly to standard schemes, it translates the matrix permutation problem into a graph partitioning problem through a suitable graph representation of the sparsity pattern of the matrix at hand. The graph partitioning, however, is addressed via a novel strategy that attributes a probability to the arcs in the graph based on the matrix structure and identifies clusters of nodes based on the similarity between the evolutions of the probability distribution vectors obtained by initializing a random walk at each node. Fig.1 shows an example of constraint matrix and the permuted matrix obtained applying the proposed strategy.

For optimal MLD operation problems that can be reformulated as multi-agent constraint-coupled MILPs, we propose novel computationally efficient decentralized optimization schemes that distribute computation among the agents, with a coordinating unit enforcing the coupling constraints (cf. Fig.2). These schemes can also be applied to actual multi-agent systems. In such a setting, privacy of information may be a concern, and the fact that the proposed schemes do not require the agents to share with the central unit any information related to their individual operational limitation and contribution to the performance index can, then, be of interest.

Moving further, we extend the analysis to a wider class of non-convex constrained programs characterized by a single complicating constraint, i.e., such that its removal from the formulation simplifies the resolution of the problem. We propose an iterative bisection method for the dual problem that generates a sequence of feasible primal solutions with a cost that improves throughout iterations. Application to multi-agent problems with a scalar coupling constraint results in a decentralized resolution scheme where a central unit is in charge of updating the (scalar) dual variable while agents compute their local primal variables.

Finally, we consider the case when uncertainty affects the local constraints of a multi-agent constraint-coupled system. We derive probabilistic feasibility guarantees for a decentralized, not necessarily optimal, data-based solution of the uncertain problem while preserving the privacy of local information and improving over existing bounds in terms of sample complexity.

The optimization framework developed in this thesis is mainly motivated by challenges in the energy sector, as shown by the three main application examples adopted to assess the performance of the proposed methods, namely: optimal plug-in electric vehicles charging scheduling, Economic Dispatch Problems (EDPs) and provision of ancillary services in smart energy systems.

Our results have the potential for a practical impact on the energy sector as they can support the sustainable global electricity transition that is currently taking place.

In order to reach the net zero by 2050 target in the energy-related carbon dioxide emission, a synergistic decarbonization effort is needed in electricity generation and consumption, comprising, on the one hand, high penetration of renewable energy sources in the power grid and, on the other hand, the electrification of consumption with technologies and processes using fossil fuels like vehicles powered by internal combustion engines and gas boilers being replaced by electrically-powered equivalents like electric vehicles and heat pumps.

Since the intermittent and non-controllable nature of renewable energy production makes it challenging for grid operators to maintain a balance between supply and demand, new balancing mechanisms are being introduced, which are enabled by the proliferation of electrical appliances (e.g., electric air conditioners, heat pumps, electric vehicles) and the availability of smart metering, sensors, and actuation systems to operate them. These mechanisms include the implicit and explicit demand response in this thesis.

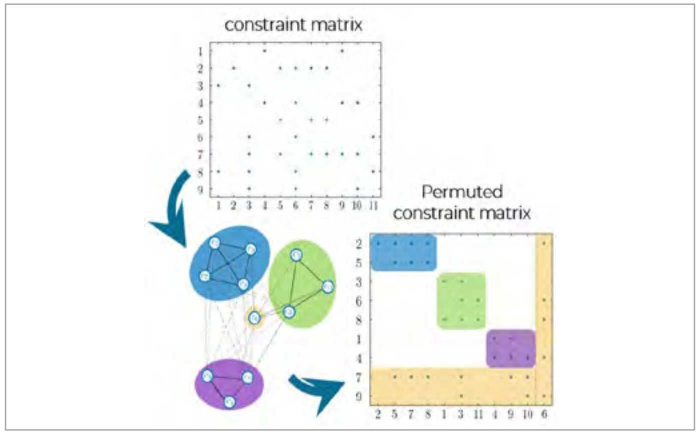

Fig. 1 - Decomposition strategy: original constraint matrix (upper-left corner), associated graph representation and partition and resulting permuted matrix.

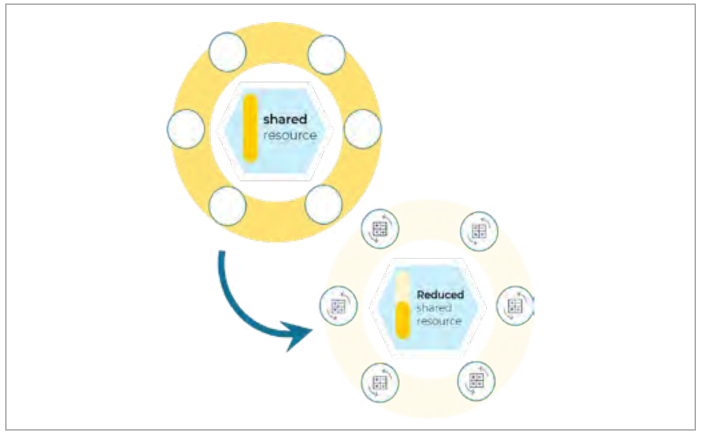

Fig. 2 - Visual representation of a multi-agent system with coupling constraints and the proposed resolution schemes that distributes computation among the agents.

ADVANCING REMOTE SENSING: DEEP LEARNING TECHNIQUES FOR InSAR DATA PROCESSING AND ANALYSIS

Gianluca Murdaca - Supervisor: Claudio Maria Prati

Synthetic Aperture Radar (SAR) represents an advanced microwave-based imaging system that operates on the principle of coherent signal processing. This system possesses a distinctive capacity for penetrating through cloud cover due to its utilization of microwave frequencies. Its active nature ensures continuous operation both during daylight and nighttime conditions. Leveraging these capabilities, the emergence of the Interferometric Synthetic Aperture Radar (InSAR) technique has revolutionized remote sensing, offering unparalleled accuracy in Earth surface monitoring. InSAR enables the precise measurement of ground movements, deformation, and topographical changes, leading to invaluable insights for various scientific and practical applications. However, despite its transformative potential, InSAR encounters several challenges due to the presence of noise and artefacts within the data, which can seriously compromise the result's accuracy and reliability. This thesis represents an in-depth investigation into InSAR data processing, employing the formidable capabilities of deep learning to forge novel methodologies. The exploration begins by addressing the phase filtering and coherence estimation task, a fundamental process crucial to exploiting the potential of the subsequent InSAR methodologies developed. In particular, we introduced an innovative model for noise reduction, pattern preservation, and artifact mitigation in filtered signals. The result's accuracy ensures the reliability of the subsequent phase unwrapping outcomes, thus enhancing the whole deformation or topographic analysis. A novel simulation strategy for generating SAR interferograms is also presented using parameters extracted from real SAR imagery, establishing the basis for a dataset that accurately mirrors real-world InSAR physical properties. Additionally, an assessment of the reliability in the filtered phase is computed, providing a comprehensive understanding of the quality linked to our phase estimations. After that, we addressed the intricate challenge of identifying phase discontinuities, a critical element in ensuring the precision of phase unwrapping. These abrupt transitions hold the key to accurate unwrapped phase values, preventing results from distortions. Within the complexities of real-world scenarios, where noise necessitates filtering and unwrapping depends on a noise-free phase, we introduced a comprehensive approach to estimating phase discontinuities starting from our filtered data. Finally, we addressed the task of change detection within open-pit mines. Recognizing the impact of mine alterations on safety, efficiency, and environmental impact, we leverage deep learning to detect changes using high-resolution SAR images accurately. In summation, this thesis delves deeply into the complexities of InSAR, harnessing deep learning methodologies to overcome challenges and unlock the technology's true potential. Our approaches leverage neural networks while intricately considering the underlying physics of the data. Indeed, we actively guide the learning process rather than passively feeding data and relying solely on deep learning for autonomous decision-making. Through the developed methods encompassing accurate phase filtering, coherence estimation, phase discontinuity identification, and change detection, this research redefines the boundaries of InSAR's capabilities, forging a path toward more profound insights and transformative applications across diverse disciplines.

ONLINE LEARNING METHODS FOR PRICING AND AVERTISING

Marco Mussi - Supervisor: Marcello Restelli

The exponential growth of online platforms and the increasing availability of data have driven companies to explore innovative methods for optimizing their pricing models and advertising strategies. Traditional approaches to these challenges, which often relied on static models and broad generalizations, have been progressively replaced by data-driven, adaptive methodologies powered by Machine Learning (ML). This research investigates the application of ML techniques, particularly Multi-Armed Bandits (MABs), in the domains of dynamic pricing and advertising budget optimization, with the ultimate objective of maximizing revenue.
Online learning algorithms must navigate the fundamental trade-off between exploration—discovering new actions with uncertain outcomes—and exploitation—leveraging known information to maximize a given reward function. This exploration-exploitation dilemma is central to this research and has broad implications beyond pricing and advertising, extending to fields such as healthcare, finance, and industrial process optimization. MABs offer computational simplicity, explainability, and strong theoretical guarantees, making them particularly attractive for the problems studied in this work. Given the need for efficient and interpretable solutions in pricing and advertising, MABs serve as the methodological foundation of this research. The study extensively explores various MAB models, demonstrating how they can be adapted and fine-tuned to meet the specific requirements of pricing and advertising scenarios. It also proposes several new MAB frameworks to address these settings efficiently, analyzing both their statistical and computational complexity.
The study is structured into three main parts, corresponding to the three areas in which contributions are made: dynamic pricing, advertising optimization, and joint pricing-advertising strategies.
Dynamic pricing is a critical task for businesses seeking to determine the optimal price for a product. The goal is to balance revenue generation with consumer demand. The study first introduces a practical MAB algorithm designed to optimize pricing strategies for e-commerce platforms featuring various products with differing market dynamics and data availability. Empirical evaluations on both synthetic and real-world data demonstrate the practical effectiveness of enforcing monotonic demand curves in scenarios with scarce data. Notably, the algorithm significantly improves revenue generation for long-tail products, which are typically priced less accurately, while its effect on popular products is less pronounced due to competitive pressures and pre-existing pricing expertise. A second contribution focuses on modeling temporal dependencies in pricing using autoregressive processes within a new MAB framework. The objective is to capture temporal dynamics while maintaining computational efficiency. This research introduces the concept of Autoregressive Bandits, proposing an optimistic algorithm to minimize regret while learning an action-dependent temporal structure. Theoretical performance guarantees on the statistical complexity of the algorithm are provided, and an experimental validation in realistic pricing scenarios underscores its efficacy.
Advertising budget optimization is another crucial aspect of selling products online. This part of the study focuses on budget allocation within Marketing Mix Models (MMMs), which present additional challenges due to interactions between different advertising campaigns. A theoretical framework called Dynamical Linear Bandits is introduced, modeling how rewards are influenced by both current actions and historical decisions. A regret lower bound is established, and an exploration-exploitation balancing algorithm is developed with provable performance guarantees. Synthetic experiments and real-world MMM simulations validate the algorithm's effectiveness.
A key insight from this work is the identification of budget optimization in MMMs as a relatively unexplored area, where linearity assumptions provide a useful yet simplifying perspective.
The final component of this research integrates pricing and advertising into a unified decision-making framework. A novel setting is introduced in which pricing and advertising decisions are jointly optimized using a factorized reward structure. The study presents the first general framework for addressing bandit problems with intermediate observations, a setting that generalizes joint pricing-advertising optimization. A theoretical analysis of the statistical complexity is provided, including both instance-dependent and worst-case lower bounds. Two regret-minimization algorithms are then proposed, each offering distinct advantages, with empirical evaluations confirming their superiority over state-of-the-art alternatives. The study highlights the inefficiencies that arise from treating pricing and advertising as independent tasks and demonstrates the benefits of their joint optimization.
This dissertation demonstrates the power of Multi-Armed Bandits in sequential decision-making problems relevant to e-commerce and digital marketing. By leveraging MABs, businesses can optimize pricing strategies and advertising expenditures in a data-driven, efficient, and interpretable manner, ultimately enhancing revenue generation in competitive online markets. The methodologies developed in this work pave the way for further advancements in online decision-making, offering new opportunities for businesses to leverage AI in optimizing their market strategies. With the rapid evolution of digital marketplaces and the increasing role of AI-driven automation, the findings of this research have broad implications for industries seeking to enhance their decision-making capabilities through intelligent and adaptive learning techniques.

OPPORTUNITIES AMIDST CHALLENGES: FREQUENCY-MODULATED MEMS ACCELEROMETERS FOR HIGH-END APPLICATIONS

Christian Padovani - Supervisor: Giacomo Langfelder

Frequency modulated MEMS (Micro-Electromechanical System) accelerometers have experienced an ever increasing performances in the last two decades in terms of stability and noise level. The thesis discusses the advantages and the critical points of two main working principles belonging to this typology: the mechanically induced frequency modulation and that electrostatically induced. The former has been designed with double-ended tuning fork (DETF) resonators, whose frequency is modulated by the axial stress induced by the displacement of a seismic mass. It is provided with a custom IC, with two oscillators and an SPI interface to provide the proper configuration, and a FPGA-based frequency readout and digital signal filtering. The Allan variance of the whole system is shown in Fig. 1. The results reach the state-of-the-art performances for this kind of device, particularly its dynamic range (164 dB) and time stability (> 10^4), as reported in Fig. 2, paving the way for the replacement of the bulky and expensive quartz accelerometers in high-end applications. On the other side, the architecture is based on a AM modulation of the frequency, also called Time-Switched Frequency Modulation (TSFM). The architecture is deeply investigated through four different devices: one exploits a different process technology (dual layer) to enhance the temperature stability and robustness against substrate bending, exploiting the possibility to shrink the dimensions of the anchor points and gather them at the center of the device. The measurements of ZGO (7 µg/°C) are compared with those of a standard implementation whose anchor points are placed at the corner of the masses. Another device has been simulated and fabricated with a disruptive application of shaped comb fingers to provide the tuning effect replacing the parallel plates to improve by an order of magnitude the frequency noise level. The main idea relies on the possibility to both increase the displacement on the masses otherwise limited by the parallel plates and at the same time keeping the link between the two modes involved, exploiting the quadratic nonlinearity of particularly shaped comb fingers. Fig. 3 shows the microscope picture of the full device and a SEM zoom of the shaped comb fingers. The last one has the same purpose but it is fabricated with a piezoresistive readout, replacing the capacitive one: it demonstrates promising results particularly in its out-of-plane sensitive implementation with a state-of-the-art dynamic range (132 dB).

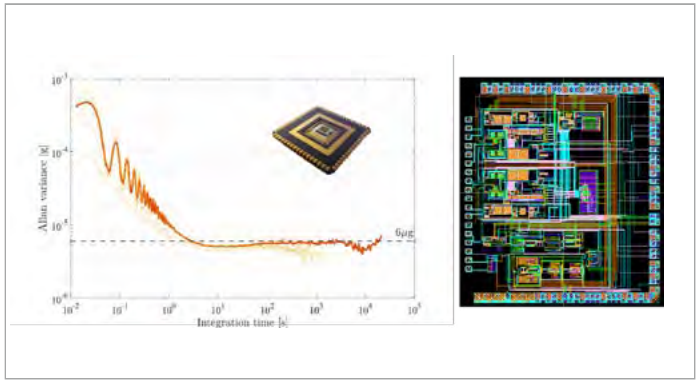

Fig. 1 - Allan variance performed on the output frequency of the IC glued to the MEMS and the carrier with a double-sided adhesive tape. Aside the IC layout used to perform all the measurements.

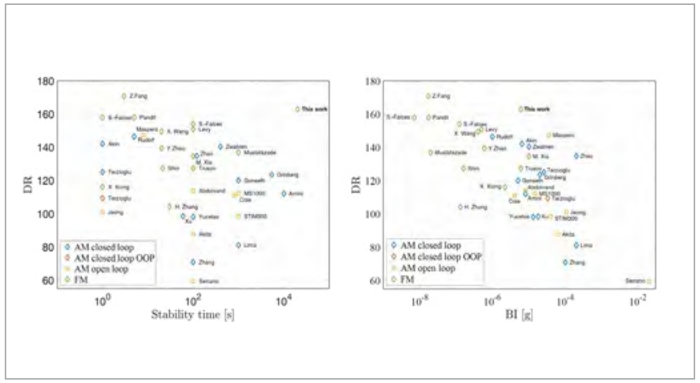

Fig. 2 - Dynamic Range (DR) against Bias Instability (BI) and stability time of this work compared with state-of-the-art accelerometers with different working principle reported in literature.

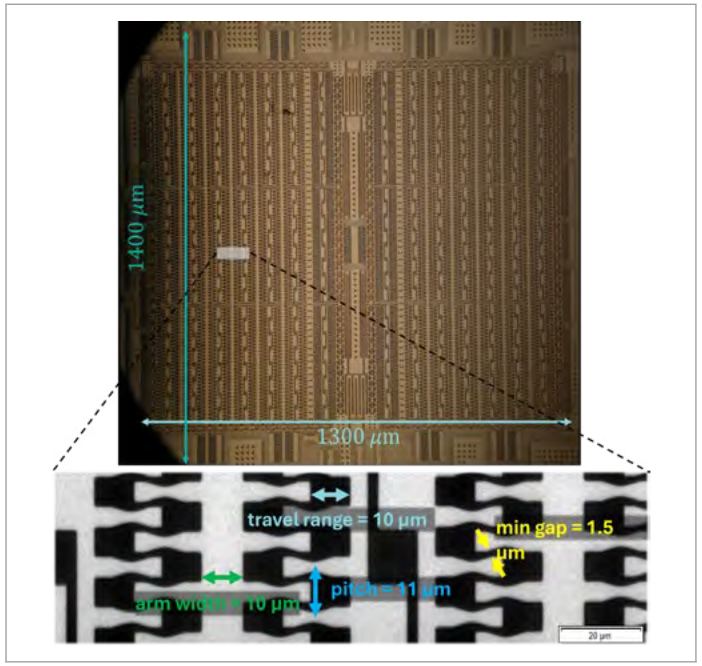

Fig. 3 - Microscope picture of the whole TSFM accelerometer based on shaped comb fingers with also a zoom on the shape profile with some relevant dimensions.

EXTENDED REALITY MODEL, A CONCEPTUAL MODEL FOR THE DESIGN OF EXTENDED REALITY EXPERIENCES

Alberto Patti - Supervisor: Franca Garzotto

Amid rapid technological evolution, Extended Reality (XR) technologies, including Virtual Reality, Augmented Reality, and Mixed Reality, have changed human interactions with digital environments and devices, blurring the lines between reality and the virtual realm in a seamless and coherent experience. However, as XR technology advances, creating successful experiences has become increasingly complex, requiring specialized expertise and often excluding stakeholders from the design process.
Recognizing this challenge, this research addresses the need for a comprehensive framework that facilitates the creation of XR experiences. The objective is to provide a structured language and vocabulary accessible to non-experts while offering sufficient depth for experts to use as a robust starting point for development.
The research began with a critical analysis of the existing models for XR experience design, conducted through a systematic literature review following the PRISMA method. This review examined 24 models across 33 papers, providing insights into their context, target audience, functionalities, and applications. Furthermore, the analysis highlighted the evolution and gaps in conceptual models within the XR field.
Subsequently, the Extended Reality Model (XRM) has been developed, with the aim of providing a conceptual framework that transcends specific technologies, offering a unified approach to XR experience design. It comprises three sub-models, each addressing distinct aspects of the XR experience: the Structural model, the Behavioural model and the Interaction model.
The Structural model describes the elements involved in the XR experience, distinguishing between components that can perform actions and react to events and those that just contribute to the environment.
The Behavioural model lists the actions available to the user during the XR experience by grouping them as User Action. Whenever the user acts within the XR environment, the corresponding User Action is triggered. The physical activities performed by the user are collectively termed Actions and are classified into Simple Action when it represents a singular user behaviour, Composite Action when it describes a combination or sequence of other Actions.
The Interaction model integrates all the elements articulated in the Structural and Behavioural models to shape the experiential flow. The sub-model utilizes the concept of states and transitions to articulate the sequential steps of the XR experience, specifying behavioural and property changes of the XR elements in each state.
The XRM has been validated through three case studies and an exploratory study to ensure the model's quality and expressiveness. The case study aimed to showcase projects that either contributed to the development or were designed with the XRM, demonstrating its impact in real-world scenarios. The exploratory study has been conducted to understand the model's usability and effectiveness by involving ten individuals knowledgeable in XR experience design and/or development.
The study was structured into sessions composed of two distinct phases: presentation and modelling. During the presentation phase, participants received an introduction to the model, including its application and examples. In the modelling phase, they individually described a hypothetical VR experience using the XRM, followed by a questionnaire.
The study results were generally positive. The XRM demonstrated acceptable usability, with an average SUS score of 69.5, though the NASA-TLX analysis indicated a "somewhat high" workload. Participants reported satisfaction with the models they generated, recognizing the tool's ability to clearly represent design concepts. However, feedback highlighted issues with the freedom provided by the XRM: while the model allows for clear and expressive representation, it also achieves verbose results.
This thesis has laid a foundation for a structured language in XR design, acknowledging both its strengths and limitations while setting the stage for future research.
By providing a structured vocabulary for XR designers and developers, the XRM enables high-level specifications for comprehensive XR experiences. While the results of the study were positive, continued refinement is necessary to address identified constraints and further enhance the model's usability and effectiveness.

IMPRESSION-AWARE RECOMMENDER SYSTEMS

Fernando Benjamín Pérez Maurera - Supervisor: Paolo Cremonesi

Impression-aware recommender systems (IARS) are innovative recommender systems that leverage a novel data type to learn users' preferences toward items in online systems. Such data type is called impression, a collection of items shown to a user on-screen and generated by an online system.

Traditionally, recommender systems leverage user's actions with the system, such as views or clicks, to learn user preferences. IARS take a comprehensive approach to achieve the same goal. Unlike actions, impressions represent the choices the recommender system provides to users at any time. IARS, then, join users' actions and displayed impressions to make accurate and serendipitous recommendations. Consequently, IARS, in conjunction with impressions, enable the design and development of novel and more refined recommendation techniques.

Nowadays, many open research directions fundamental for properly developing IARS still need to be answered. The Thesis addresses three directions: characterization, evaluation, and experimentation.

The first direction, characterization, addresses the need for sound mathematical foundations to define and describe IARS. In the Thesis, we design and propose a theoretical and mathematical framework to define impressions and IARS. Under this framework, we present a novel classification system for IARS comprised of three taxonomies and three design properties of recommendation models that learn from impressions. Additionally, we comprehensively review existing recommendation models proposed in previous research works.

The second direction, evaluation, addresses the need for standard evaluation methodologies and procedures to assess the recommendation quality of IARS. In the Thesis, we present, describe, and analyze a novel dataset with impressions suitable for evaluating IARS: ContentWise Impressions. We devise a standard evaluation framework for IARS, comprised of guidelines to process and clean datasets with impressions, to define recommendation tasks and goals, to define a set of baseline recommenders, and to search the optimal set of hyper-parameters.

The third direction, experimentation, studies IARS from a practical perspective, i.e., by performing experiments and evaluating their recommendation quality. In the Thesis, we assess whether the reviewed papers provide sufficient and necessary descriptions and tools to reproduce or replicate their results. Then, we describe two evaluation studies on IARS, one aiming at replicating the results achieved by the literature, while the other aiming to incorporate impressions into graph-based recommender systems in practical and simple approaches to improve recommendation quality. In both evaluation studies, we obtain positive results, where IARS obtain more accurate or diverse recommendations than the evaluated baselines in most scenarios.

DEVELOPMENT OF A PERCEPTION SYSTEM FOR AUTONOMOUS VEHICLES ON URBAN ROADS

Riccardo Pieroni - Supervisor: Sergio Matteo Savaresi

In the past decade, there has been significant progress in the development of autonomous vehicles (AVs), driven by extensive research and industry efforts. This is because autonomous vehicles promise transformative benefits, such as reduced vehicle ownership, increased utilization, and substantial safety and environmental improvements. By potentially replacing multiple conventional vehicles, AVs can alleviate parking demand and enhance road safety through reduced traffic conflicts and accidents. Furthermore, they contribute to environmental sustainability by optimizing traffic management and decreasing fuel consumption.
In this context, the development of perception systems for AVs, particularly for urban environments, is a major challenge in future mobility. These systems must perceive and interpret complex, dynamic scenes in real-time, integrating data from various sensors such as cameras, LiDARs, and radars. Each sensor provides different types of information: cameras offer high-resolution 2D images, LiDARs deliver 3D spatial data, and radars measure speed and movement. They also operate at different rates, adding complexity to real-time data processing and fusion. For instance, LiDAR typically updates at 10 Hz, while cameras can operate at 20-30 Hz. Effectively merging this data requires sophisticated algorithms to align and integrate different types of information, ensuring the vehicle has a consistent and accurate understanding of its environment. Specifically, a perception system for autonomous vehicles must identify and classify various objects such as vehicles, pedestrians, cyclists, traffic signals, traffic lights and other relevant objects on the road; it must continuously track the movement of detected objects over time to predict their positions and future movements; and it must also identify the drivable areas and lane markings on the road to determine where the vehicle can safely travel.
This research aims to improve the perception capabilities of autonomous vehicles by integrating sensor fusion and advanced deep learning techniques. The objective is to develop a robust perception framework that efficiently processes sensor data in real-time, ensuring safe and reliable decision-making in dynamic urban environments. In particular, this dissertation proposes advancements over the state of the art in various tasks, including the development of an advanced camera-LiDAR neural network for real-time 3D object detection, a versatile multi-object tracking framework that operates in both single-modality and multi-modality configurations, an innovative Bird's Eye View (BEV) semantic mapping algorithm, and a multi-task neural network for real-time segmentation of drivable areas and road lines.
Regarding 3D object detection and classification, the main contributions of this dissertation are the design and integration of mid-level and late fusion modules within a single architecture, which enhance the detection of small or distant objects while maintaining real-time performance. This method significantly improves the accuracy of LiDAR-based detectors for pedestrians and cyclists and achieves competitive 3D Average Precision compared to other state-of-the-art multimodal approaches. Moreover, our approach is optimized for real-time execution, achieving a processing rate of 22 Hz, which outperforms other multimodal methods that typically operate at 10 Hz. Regarding multi-object tracking, our approach outperforms other state-of-the-art solutions in pedestrian tracking and achieves comparable results for car tracking. The best results are obtained with sensor fusion, where LiDAR, camera, and radar data are combined for optimal tracking accuracy.
For BEV semantic mapping, the proposed method includes a novel mechanism that densifies the occupancy grid information by exploiting the spatial correlation of random variables within the lattice. This approach outperforms single-modality methods on the nuScenes dataset while operating at 25 fps.
Finally, the main contributions of this dissertation to the field of drivable area and road lines segmentation include the ability to differentiate drivable lanes in multi-lane scenarios and identify various types of road lines. Notably, our method can distinguish between *direct* and *alternative* lanes, a capability found in very few state-of-the-art methods. Additionally, it recognizes pedestrian crossings as a distinct line category, a feature absent in other leading approaches.

MACHINE LEARNING FOR PROBABILISTIC AND ATTRIBUTE-AWARE MACHINE LEARNING

Pietro Portolani - Supervisor: Matteo Matteucci

In the context of pervasive digitalisation, every process around us could be recorded and investigated for insights, predictions, better decision making and optimisation possibilities. Variability is always present in real-world processes and it is crucial to take it into consideration during the analysis and the development of solutions. Process attributes could explain part of the process variability, steering the process towards specific variants in presence of determined characteristics; it is therefore crucial to insert them into the process investigation. At the same time, a certain degree of stochasticity will always permeate a process being a source of variability. Process mining is a recently rising field that provides tools to extract knowledge about the process from data recorded by information systems. Process mining techniques are able to create process models, predict process behaviour as well as establishing the adherence of a model to specific data. Machine learning is another scientific field that has revolutionised many areas of our lives in recent times. It can automatically learn complex patterns linking an input to an output of a given system or detect similarities and relations inside data. In this thesis we investigate the use of machine learning techniques to deal with the process variability, taking into consideration process attributes and stochasticity inside the process mining analysis. Specifically, in the first part we study the uncertainty hidden in the forecast of neural networks in the next activity prediction task. Due to the hard explainability of neural networks, in the second part the thesis focuses on methods related to explicit process models. In particular, it proposes a novel approach to trace clustering and to decision discovery in a given model. The thesis also developed an extension to a state-of-the-art method to compose a stochastic process model that considers process attributes and activities previously executed during an instance execution. The developed approaches are then applied to an industrial case study to predict the most probable activities needed to produce a specific item in a highly flexible manufacturing process.

DATA-DRIVEN TECHNIQUES FOR SPEECH AND MULTIMODAL DEEPFAKE DETECTION

Davide Salvi - Supervisor: Paolo Bestagini

Nowadays, the creation and dissemination of multimedia content is becoming an increasingly simple operation. Thanks to the technological improvement of commonly used devices, every day we generate a large amount of data that is shared on the web and social media platforms. This opportunity comes from the resolution of some technical constraints that we had in the past, like storage, bandwidth, and computational power limitations. Individuals can now use their smartphone or any other similarly accessible device to produce and edit a large amount of high-quality media without any effort, be it photos, videos, or audio recordings. Then, this generated data can be stored on cloud platforms with unlimited capacity or instantly shared with anyone worldwide.
The created and shared content is also getting more and more complex, migrating in just a few years from simple images to high-quality videos. This transformation is evident simply by browsing online websites or scrolling through our social media feeds, where still images are replaced more and more often by high-resolution videos. "Instagram is no longer a photo-sharing app'", says Adam Mosseri, the head of Instagram, "but an app to be entertained", and videos undoubtedly offer the most captivating and immersive experiences for entertainment.
In addition to authentic multimedia content, i.e., captured through real acquisitions using cameras and microphones, a new type of media has started to circulate on the web: synthetic data. This includes any content that has been artificially generated or edited using signal-processing or deep learning techniques, such as Variational AutoEncoders, Generative Adversarial Networks or diffusion models, which are now commonly employed in various everyday applications. Examples of these models include ChatGPT for text generation, DALL-E for image forgery, Sora for video generation, and ElevenLabs for speech synthesis. Any type of multimedia content can be generated synthetically, achieving a level of realism that often makes it difficult to differentiate it from real data.
This progress opens the door to new, exciting possibilities and scenarios that were just imaginable a few years ago. For instance, it can lead to enormous progress in human-machine interaction, introduce new educational support systems, and improve technologies assisting individuals with disabilities or disorders. However, there are also negative implications associated with the diffusion of synthetic material, especially when this is produced with malicious intent.
An example in this regard are deepfakes, synthetic multimedia content generated through deep learning techniques that depict individuals in actions and behaviors that do not belong to them. Deepfake generation techniques allow, given an image, a short video, or an audio recording of a target person, to generate videos or audio signals in which the target appears to say or do things they have never uttered or done. This poses a significant threat, particularly when such content is generated without the consent of the depicted person and portrays unpleasant actions. In recent years, we have witnessed several cases where deepfakes have been used with malicious intent for deception, threats, or blackmail. Moreover, when the targets of these attacks are public figures or personalities with important roles, this content can be instrumental in spreading misinformation and propagating fake news.
To give an idea of the menace associated with this phenomenon, consider that the number of generated deepfakes doubles every six months, and only 7% of these are made for comedy and entertainment purposes. All the others are created with malicious intent to attack people's reputations by defamatory, derogatory, and fraud, and this material can spread on the internet without any control. Consequently, it is becoming paramount to automatically determine the authenticity of multimedia content to prevent dangerous situations. Also, given the rapid evolution and growing complexity of multimedia data circulating on the web, it is necessary to develop up-to-date detectors and analysis techniques capable of processing complex content such as high-quality videos and audio created with cutting-edge generative models.
In this thesis, we propose to tackle the deepfake detection problem using a multimodal approach, simultaneously analyzing different content modalities. For example, when examining a video of a person speaking, we aim to analyze both its visual sequence and speech content. By exploiting consistencies and inconsistencies between different modalities, we can increase the amount of helpful information available to solve specific tasks, such as deepfake detection.
Considering multiple modalities at the same time leads to the development of detectors with more robust and reliable predictions than monomodal models. This is because, in addition to investigating the artifacts present within the individual modalities, they can also analyze the coherence across them, which is often challenging to recreate for deepfake generators. For instance, consider a scenario with a highly realistic synthetic video sequence paired with equally realistic synthetic speech content. Given their realism, these signals would fool monomodal deepfake detectors. However, as multimodal systems can simultaneously analyze the two modalities, they can exploit the coherence between the two sequences and provide additional information to spot such synthetic content effectively. In general, the joint use of multiple data modalities opens the door to a new class of detectors that are intrinsically more robust to adversarial and post-processing.
Throughout the thesis, we aim to propose methods that combine signal processing and deep learning techniques. The goal is to exploit the potential of both approaches by overcoming their weaknesses. For example, using the information provided by model-based approaches, we can reduce the amount of data needed to train data-driven techniques, improving the results obtained. At the same time, such a hybrid strategy prevents the data-based models from learning certain content-based biases that are harmful to the final goal, improving the generalization capabilities of the developed systems. This is a crucial aspect in the forensic field, where we want to design systems that can face several different domains. An example of this type of application can be found in the semantic approaches we propose. These correspond to forensic detectors that base their predictions on the analysis of high-level features extracted from the input signal, which describe complex aspects such as emotions, speech prosody, or speaker-related information. Analyzing this type of feature allows our systems to not focus on individual artifacts left by specific generators but to investigate more complex aspects that are persistent across multiple datasets and scenarios. While model-specific artifacts change depending on the deepfake generator we are dealing with, analyzing more complex aspects allows us to better generalize our predictions across multiple contexts.

ANALOG FRONT-END CIRCUITS FOR HIGH-RESOLUTION WIDE-BAND ADCs

Lorenzo Scaletti - Supervisor: Luca Bertulessi

The relentless advancement of technology, characterized by rapid evolution, ever-increasing computing power, and miniaturization, is expected to persist well into the foreseeable future. This will lead to a proliferation of electronic devices, exponentially increasing the volume of data transmitted. The trend of integrating multiple interconnected devices into everyday appliances, spanning domains from home automation and smart vehicles to wearable technologies, Industry 4.0, and smart cities, is set to significantly increase the number of connected devices, including small, energy-efficient, battery-operated embedded systems. Furthermore, emerging data-intensive applications like 8K video streaming, online gaming, video conferencing, telecommuting, and virtual reality, will drive the demand for very high throughput wireless connections. In this context, wireless networks, including the currently expanding 5G networks and the forthcoming Wi-Fi 7 (802.11be) standard, take on a pivotal role, promising enhanced data rates and reduced latency without sacrificing the energy efficiency. Modern telecommunication standards mandate high-performance Analog-to-Digital Converters (ADCs) to enable effective communication between the wireless transceivers and the digital signal processing (DSP) units. Multi-Gigabit per second links, enabled by techniques like channel aggregation and high-order modulation schemes, dictate the requirement for wide-bandwidth (>1 GHz) and medium to high-resolution (>50 dB) data converters. However, designing high-resolution ADCs in the Giga-samples per second (GS/s) range poses formidable challenges. Time interleaving emerges as a viable technique to extend the bandwidth of energy-efficient, low-frequency ADCs. However, it introduces challenges such as inter-channel mismatches (i.e., offset and gain mismatches, and sampling clock skew) and interactions, which result in spectral artifacts that degrade the effective resolution compared to single-channel ADCs. This thesis revolves around the development of a time-interleaved data converter able to meet the stringent specifications. The simple and scalable Successive Approximation Register (SAR) topology, known for its commendable energy efficiency in modern technology nodes, emerges as an optimal candidate to implement the sub-ADCs of a time-interleaved converter array. The implementation of a SAR core, optimized for speed, was therefore the first step of the research endeavor. This involved the application of multiple techniques, including the introduction of a novel switching algorithm to drive the Capacitive Digital-to-Analog Converter (CDAC) and the incorporation of redundancy in the SAR conversion steps. This redundancy increases the robustness of the SAR core against settling errors, enabling higher conversion speeds compared to binary implementations. The SAR core was fabricated in a 28-nm bulk CMOS process as an integral component of a 6× time-interleaved converter. The presented sub-ADC is a 12-bit, 150-MS/s, 13-step redundant asynchronous SAR converter, achieving 63-dB Signal-to-Noise and Distortion Ratio (SNDR), which corresponds to 10.2 Effective Number Of Bits (ENOB), and 72-dB Spurious-Free Dynamic Range (SFDR), with an Equivalent Resolution Bandwidth (ERBW) larger than 100 MHz. To suppress the effects of time-interleave mismatches, the time-interleaved converter includes on-chip background calibrations. A novel skew calibration algorithm enables varying the number of active channels within the time-interleaved ADC, allowing adaptable signal bandwidth. This offers an interesting opportunity for optimizing the energy efficiency of multi-standard receivers. Despite the satisfying results, a degradation in the dynamic performance metrics was observed when transitioning from a single-core to a time-interleaved configuration. This degradation was primarily attributed to crosstalk between the channels through the shared reference distribution network, and to the absence of an on-chip input signal buffer. To address these impairments, a second prototype was designed, featuring a distributed reference buffer architecture that mitigates the channel interactions arising from the shared power supply and ground networks. Additionally, an input buffer with enhanced linearity was implemented to preserve the integrity of the input signal, mitigating the kickback of the input sampler of the ADC cores on the front-end network. The measurement results of this second prototype, a 2 GS/s 11-bit 8× -interleaved ADC, showcase an SNDR level consistent (within 1.4 dB) with the single-channel one across the entire 1 GHz input bandwidth, achieving 57.3 dB SNDR and 69.9 dB SFDR close to the Nyquist frequency.

AUTOMATIC CONTROL SOLUTIONS TO ENHANCE THE RELIABILITY OF AIRBORNE WIND ENERGY SYSTEMS

Mahmoud Hassan Mohammed Tareg - Supervisor: Lorenzo Fagiano

Airborne wind energy (AWE) is a technology to harvest wind energy and then convert it into electricity using crosswind flying kite connected to the ground through a tether. In the concept of a tethered crosswind flying kite, the aerodynamic force exerted on the kite produces traction force sufficient to keep the kite flying without using propulsion system; As long as wind is blowing and the kite will sustain enough flying velocity as a result a high aerodynamic traction force on the kite and a high tension on the tether will be exerted. To sustain the crosswind flying, the kite could be steered to follow circular shape flight trajectory.

One of the used methods to produce electricity out of the crosswind flying kite is the pumping cycle concept which considered in this study, Fig. 1 shows the main component of AWE system and the pumping cycle phases. During crosswind flying, the kite will fly in relatively higher speed than the wind speed, and then start to pull the tether, the tether is winded around a drum which is connected through a gearbox to a generator to produce electrical power, this phase of pumping cycle method is called traction phase. Followingly, the tether will continue to extend out from the drum as long as the wind speed remains above the cut-in threshold. When the tether nears its maximum extension, the kite will transition from crosswind flying to gliding back toward the drum to facilitate tether retraction, during which the ground unit acts as a motor and will consume fraction of the power harvested in the power generation phase; this power consumption phase is called the retraction phase. When the tether is reeled-in, the kite will exit gliding and transit to crosswind flying to start another pumping cycle. Fig. 1 shows the subsystems of the main pumping cycle Airborne Wind Energy System (AWES) and the operating principle. The kite unit can be of rigid or soft wings. The deployment of the rigid wing kite can be done by attaching a vertical take-off and landing (VTOL) subsystem to the kite.

A reasonable question is; why in particular this technology should be considered? In order to answer this question, it is good to recall the motivation behind sustainable energy needs: The Intergovernmental Panel on Climate Change (IPCC) of the United Nations has defined 1.5°C pathway to limit global warming to 1.5°C by 2050. The International Renewable Agency (IRENA) estimates that to complete the 1.5°C pathway it is necessary to add 8 Terawatt of wind energy by 2050, which requires an addition of 250 gigawatt every year from 2021 to 2050. The transition process is challenging; an aspect of the criticism of the possibility of this transition is related to the accessibility of the required minerals and metals. The estimate of the usage of steel material to build conventional wind turbines ranges from (107,000 to 132,000) tons per Gigawatt and that of aluminum ranges from (500 to 1,600) tons per Gigawatt for different wind turbines. When mapping these figures to the 1.5°C path requirement, adding 8 Terawatts to wind energy by 2050 will require from (856 to 1,056) millions of tons of steel and (4 to 12.8) millions of tons of aluminum. Most of this steel and aluminum is used to build wind turbine towers. However, AWES do not require towers to harvest wind energy; a tether made of polymer or conductor material is used to connect the flying apparatus to the ground and transfer the energy to the ground. Therefore, AWE is considered an attractive solution to reduce the levelized cost of energy (LCOE) for the wind energy sector. Moreover, AWE can be deployed offshore and onshore; Due to its lighter weight, ease of mobility and transportation, and less land and material use, AWE is considered to possess a great potential.

The AWE has not been used yet on a large commercial scale, given the current status of AWE technology, there are significant challenges to its successful implementation that must be addressed. These challenges include 1) ensuring fully autonomous operation even beyond standard operating conditions and 2) enhancing the system's robustness and failure tolerance. This research introduces strategies to enable autonomous operation beyond nominal operation condition, and to increase system reliability by applying fault tolerance control (FTC) for AWES. Specifically, this study introduces FTC algorithms tailored for Vertical Take-Off and Landing (VTOL) pumping cycle AWES, the method has been tested in a verified simulation environment, a control surface failure has been injected at different points during the production phase, and the proposed FTC algorithm effectively maintained system performance, see fig. 2. Additionally, a "standby" safety mode is proposed for operating AWES outside its nominal operational conditions; this mode notably enables AWES to function when wind speeds fall below the operational cut-in point see fig. 3, the method has been tested in verified simulator in wind speed from 0 to 12 m/s, the results indicate that the method successfully kept the kite aloft even at 0 m/s wind speed, and shows reliable performance in higher wind speed. Moreover, in this research the developed AWES simulator has been validated using experimental data, which is considered one of the contributions of this research given the limited availability of validated simulators in the AWE domain. This simulator is considered a crucial instrument for exploring AWES failure modes and assessing the corresponding mitigation strategies.

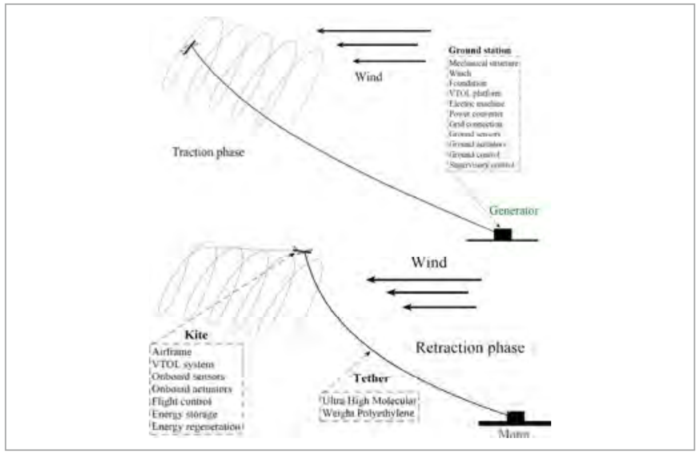

Fig. 1 - AWES main component showing pumping cycle concept.

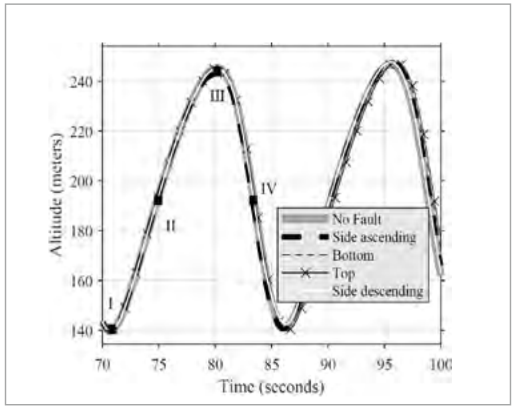

Fig. 2 - Kite altitude simulations showing the performance of the suggested FTC approach, when the fault is injected bottom top, and each side of the loop.

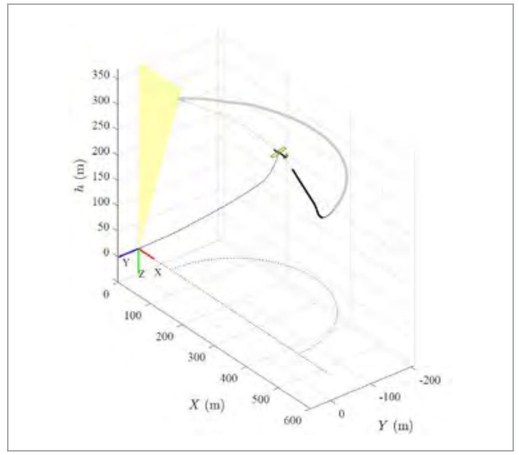

Fig. 3 - Kite flight trajectory in low wind condition during reverse pumping operation.

COOPERATIVE MACHINE LEARNING METHODS IN DISTRIBUTED SYSTEMS

Bernardo Camajori Tedeschini – Supervisor: Monica Nicoli

Cooperation between agents, particularly in dynamic and complex environments, is fundamental to achieving effective and efficient task execution. The necessity of such collaboration is pronounced in scenarios where tasks are too complicated or resource-intensive for individual agents to handle alone. This need spans various domains from medical fields and internet of things (IoT) to intelligent transportation systems and next-generation cellular networks (see Fig. 1).
According to the specific application domain, the focus can shift from learning the most accurate and/or reliable model to efficiently predicting the model outcome.
For example, in a medical network, the focus is on learning the most accurate and reliable models for tasks like diagnosis, ensuring data privacy and compliance with regulations. Whereas in vehicular networks, the emphasis is on efficiently predicting the model outcome in real-time for applications such as cooperative positioning.
Thus, we can distinguish between cooperative learning and inference based on their distinct roles and methodologies, each tailored to optimize different aspects of multi-agent systems (MAS).

Cooperative Learning Challenges. Cooperative learning involves centralized (C-ML) or decentralized (D-ML) machine learning approaches. In C-ML, a parameter server aggregates data from multiple agents to form a global model, emphasizing performance over scalability. This suits contexts with limited privacy concerns and robust data exchange, such as vehicular networks, where vehicles share motion and road data while roadside units aggregate information in real-time. A key challenge is accurate data association for cooperative perception and object tracking. Conversely, D-ML allows each device to learn independently, enhancing scalability and privacy. Medical settings use this approach to meet strict data protection rules, while IoT systems benefit from resilience and energy efficiency. Fully decentralized solutions handle non-IID data, node failures, and dynamic networks by reaching consensus without a central coordinator. Ultimately, cooperative learning must fully address non-stationarity in data and network topologies, ensuring models adapt to evolving conditions.

Cooperative Inference Challenges. Cooperative inference uses models learned cooperatively (via data exchange in C-ML or parameter sharing in D-ML) to deliver accurate, low-latency, reliable predictions for mission-critical tasks. Next-generation cellular networks (5G/6G) exemplify this, with stringent requirements on positioning (down to 20 cm), latency (10 ms), and reliability (99.999%). Traditional approaches struggle under complex conditions, prompting 3GPP's 5G-Advanced Release 18 to adopt ML-based positioning trained in digital twin environments. Relevant tasks include non-line-of-sight (NLoS) identification, positioning, and tracking, each relying on optimal data exchange and robust uncertainty quantification. Key challenges involve efficiently recognizing LoS/NLoS signals, combining outputs from base stations, and distinguishing aleatoric from epistemic uncertainties in real-time.

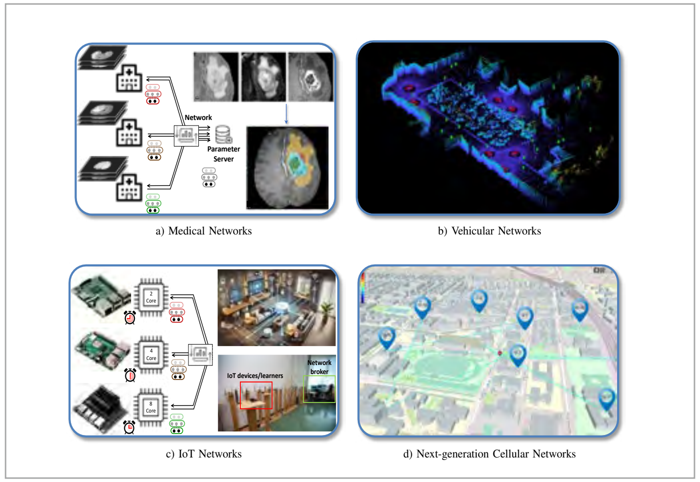

Fig. 1 - Application domain of multi-agent systems (MAS).

Contributions and Objectives. Driven by the increased demand for real-world cooperative applications, the thesis aims at proposing novel solutions to the main open problems in cooperative learning and inference to enhance robustness and adaptability under dynamic conditions. In particular, the thesis is divided into two main parts.
In **Part I**, we investigated cooperative learning in graph-aware centralized ML, privacy-preserving decentralized ML, and non-stationary frameworks. Graph-aware C-ML tackled data association and positioning in vehicular networks by leveraging logical graphs for complex, non-linear distributions and scalability. For sensitive data, D-ML transmits only model parameters or latent representations, mitigating privacy risks and enhancing communication efficiency. We built a robust platform for decentralized and fully-decentralized learning in medical and IoT networks, exploring asynchronous federated learning (FL), weighted consensus in non-IID settings, and decentralized split learning for resource-limited devices. Additionally, we addressed dynamic data distributions via multi-agent reinforcement learning (MARL). Notably, we developed an implicit cooperative positioning (ICP) approach in vehicular networks that exploits passive targets to refine agents' state estimates, thereby improving localization accuracy.

In **Part II**, we focused on cooperative inference for tasks like non-line-of-sight (NLoS) identification and positioning, both static and mobile, in next-generation cellular networks. Base stations (BSs) compress channel data into latent features for advanced sensing. We introduced an anomaly detection scheme to evaluate whether incoming samples belong to the line-of-sight distribution, thereby reliably identifying NLoS conditions. For static positioning, we designed a cooperative inference technique fusing latent representations among BSs in LoS scenarios, while each BS handles NLoS measurements independently. Addressing mobile positioning, we proposed a Bayesian neural network that produces real-time uncertainty estimates, which are then integrated into tracking filters to optimally fuse fingerprint-based likelihoods from multiple BSs, particularly in out-of-distribution environments.

Conclusions. In conclusion, the thesis presents a comprehensive exploration of cooperative learning and inference strategies within MAS, offering scalable and adaptive solutions across a diverse range of technological domains. By integrating advanced ML techniques with the implicit complexities of cooperative environments, we have developed robust models that significantly enhance both the precision and reliability of various application areas, ranging from vehicular and IoT networking to healthcare and cellular systems. These innovative approaches not only demonstrate the practical benefits of cooperative strategies but also highlight the potential for future advancements in data-driven technologies.

SINGLE-PHOTON AVALANCHE DIODES IN III-V COMPOUND SEMICONDUCTOR FOR NEAR-INFRARED APPLICATIONS

Fabio Telesca - Supervisor: Alberto Tosi

The arise of quantum theory in the first half of the 19th century was a cause for shock and awe in the scientific community all over the world. The newly theorized concepts disrupted the classical view of physics, but what initially seemed just an abstract and ungraspable theory slowly found mathematical proofs over the years and, perhaps most importantly, produced very real technological applications: for instance, concepts theorized in quantum physics led to the invention of the laser, extremely miniaturized transistors and sensors. In other words, quantum physics was a revolution of thought, that quickly led to the first quantum revolution of technology.
In the same way, today a second quantum revolution is ongoing, which involves actively utilizing quantum mechanics to manipulate and modify the quantum characteristics of the physical world for our specific needs, and has its own expanding field of science and engineering, known as "quantum technology." In this context, the exploitation of individual photons, including their properties and potential applications, holds paramount importance, as many applications necessitate the manipulation of one photon at a time. Consequently, single-photon detectors serve as essential components in virtually every optical quantum system. Some of the most promising quantum technological applications derived from the study of single photons include quantum key distribution, quantum imaging and microscopy, quantum computing, and the characterization of single-photon sources.
Single photon detection in the visible light range has already become a consumer technology, with silicon single photon detectors being easily integrated into CMOS processes. However, they have one major limitation: the maximum wavelength for which single photons can be detected is shorter than ~ 1.1 µm. Nevertheless, there are many cutting-edge applications requiring manipulation and sensing of single-photons in the near-infrared range (NIR), mostly at wavelengths of 1310 and 1550 nm. The 1550 nm wavelength is crucial for quantum communication based on optical fibers, or quantum imaging (such as ghost imaging, but also LiDAR) for eye safety, and extended-range biomedical applications (NIRS, DCS).This push from application requirements led to the development of single-photon detectors (SPDs) implemented with innovative technologies and materials capable of detecting NIR single-photons.
This dissertation focuses on advancing Indium Gallium Arsenide/Indium Phosphide Single-Photon Avalanche Diodes (InGaAs/InP SPADs), the key components of NIR single-photon detection. While superconducting nanowire detectors offer exceptional performance, their cryogenic requirements limit their practicality. SPADs, conversely, provide compact and portable solutions, which can be easily integrated in most complex systems.
My research work aimed to optimize the trade-off between the photon detection efficiency (PDE) and noise (referred to as dark count rate, DCR) of InGaAs/InP SPADs, enabling operation at higher temperatures and higher count rates. To achieve this, TCAD modeling and simulations of devices were studied and optimized, and combined with two production runs, to investigate different aspects of the devices. The typical structure of InGaAs/InP SPADs is shown in figure 1. I mainly worked on three primary application-oriented designs: one optimized for low noise, the second for high photon detection efficiency, and a last one for high count rates operation (with reduced afterpulsing probability).
The low-noise SPADs were designed with an optimized double zinc diffusion profile and charge layer thickness to finely shape the electric field inside the junction. It demonstrated DCR as low as 1 kcps and 4 kcps for 10 µm and 25 µm diameter devices, respectively, at 225 K and 5 V excess bias, while maintaining PDE of 33% at 1064 nm, 31% at 1310 nm, and 25% at 1550 nm. Timing jitter was measured in the order of 150 ps (FWHM). Conversely, the high-efficiency SPAD, designed with a thicker absorption layer and a specifically engineered double zinc diffusion profile, achieved a PDE of up to 50% at 1550 nm, with a DCR of 20 kcps and a timing jitter of approximately 70 ps (FWHM) at 225 K. Alternatively, operating at 37% PDE at 1550 nm it yielded a DCR of just 3 kcps and a timing jitter of approximately 100 ps (FWHM). When wire-bonded to a custom integrated circuit, all SPADs achieved afterpulsing probabilities of only a few percent at 225 K, with a 1 MHz gating frequency and few microseconds hold-off time, enabling photon count rates approaching 1 Mcps. Furthermore, the first fabrication phase incorporated controlled experiments to systematically evaluate the impact of variations in the double zinc diffusion profile on various key aspects of InGaAs/InP SPADs operation, such as afterpulsing probability and charge persistence.
In the subsequent generation, designs aimed at reducing afterpulsing and charge persistence were implemented by modifying the double zinc diffusion profiles and charge layer thickness to enable operation at lower excess bias and higher temperatures, based on combination of experimental evidence from the previous generation and simulation results. A new design, with a thinner multiplication region and remodeled double zinc diffusion profile, was fabricated. Preliminary testing on the thin-multiplication SPAD demonstrated a DCR of 4 kcps at 5 V excess bias, compared to 2 kcps for a standard control SPAD from the same generation, while both achieved comparable dark count rates of 1 kcps at 3 V excess bias. PDE was similar, with 25% and 18% at 3 V excess bias, respectively. However, the thin-multiplication SPAD exhibited a faster timing response of 140 ps at 5 V excess bias, and a reduced afterpulsing probability of 0.3% at 2 V excess bias, compared to 0.7% for the standard SPAD in the same conditions. These results indicate that a thin multiplication region design can improve afterpulsing and timing, without significantly affecting the PDE to DCR ratio when operated at lower excess bias. Finally, experiments for charge persistence reduction included the fabrication of one or two contacted guard rings around the active area, which could be bias independently, used to sink carriers before they reached the device center, and thus reduce the charge persistence effect.
This research served as a contribution to the progression of NIR single-photon detection, aiding the broader development of quantum technologies, and reinforced the ongoing need for optimization and innovation, acknowledging the collective work of the scientific community.

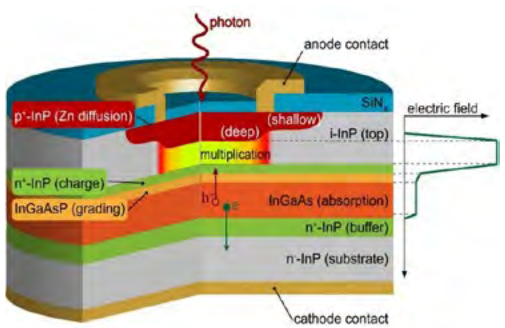

Fig. 1 - Typical cross-section of a front-illuminated planar InGaAs/InP SPAD and its electric field profile along the vertical direction.

PLL ARRAYS FOR PHASE NOISE REDUCTION AND BEAM STEERING

Francesco Tesolin - Supervisor: Salvatore Levantino

To meet the stringent requirements of next-generation communication standards, the demand for spectral purity in radio-frequency (RF) transceivers (TRX) has reached unprecedented levels. This challenge is exacerbated by the need for mass-volume production, which calls for the integration of complex baseband digital signal processing (DSP) units into the same die as the TRX RF chain, typically implemented in scaled CMOS nodes.

One of the most critical aspects of transceiver design is the local-oscillator (LO), typically realized as a phased-locked loop (PLL), which faces significant challenges due to the voltage-supply scaling and the limited analog performance of devices in modern technologies. In particular, the voltage-controlled oscillator (VCO), or the digitally-controlled oscillator (DCO) in digital PLLs, exhibits a fundamental upper bound in phase-noise performance that cannot be overcome even with increased power dissipation. Multi-core oscillators have emerged as a promising solution to restore the traditional trade-off between power and phase-noise by leveraging the combination of different oscillator elements. However, to meet the specifications dictated by the increasingly demanding communication standards, a relatively large number of oscillator cores should be used, making the design of a conventional PLL impractical due to the lossy coupling network between the cores and the large quantity of connections required [see Fig.1(a)].

This research has explored an alternative approach. Rather than employing multi-core oscillators within a single PLL loop, the outputs from an array of PLLs are combined together to improve phase-noise performance [see Fig.1(b)]. Since the phase noise from each DPLL core is uncorrelated, it sums in power at the system output, while the signal amplitude sums coherently. This produces an SNR improvement at the receiver proportional to N, where N is the number of cores used [see Fig.1(c)] and enables the full optimization of each PLL core while eliminating the challenges associated with the coupling networks and interconnections among oscillator cores. Moreover, adopting digital-PLLs (DPLLs) as core units allows achieving, at a spectral purity comparable to analog PLLs, a much smaller overall footprint with respect to the multi-core oscillator approach allowing to extend this architecture to a high number of elements.

This novel approach is applied to phased-array transceivers. Phased arrays utilize a number of elements to form highly focused antenna beams, enhancing antenna effective-isotropic radiated power (EIRP) and enabling spatial multiplexing. However, this comes at the cost of distributing high-frequency signals to each element and incorporating phase shifters at each node [see Fig. 2(a)], which are bulky and power-hungry components.

To address these challenges, the DPLL array architecture is adopted [see Fig. 2(b)]. Each DPLL in the array allows setting an arbitrary output phase shift by leveraging direct frequency modulation through an individual phase-control word (PCW) input [see Fig. 2(c)]. This replaces the conventional bulky phase-shifter array with a compact digital PLL array and removes the needs to distribute high-frequency signals to the elements, since the reference clock in the MHz range is distributed among the DPLL elements. Moreover, since each LO element operates independently, the over-the-air combination of the signals results in an SNR improvement proportional to the size of the array.

To validate the concept, a dual-element DPLL array has been implemented using 28nm TSMC technology [see Fig. 3(a)]. The measured phase-shifting accuracy between the elements is 2.1°/0.76° peak/rms over the -180° to +180° range [see Fig. 3(b)]. Additionally, the measured phase noise improves by -3dB from the single output to the combined output [see Fig. 3(c)], demonstrating the potential of the DPLL array to achieve low phase noise and beam steering.

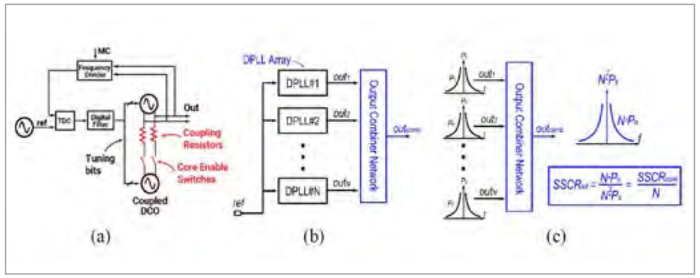

Fig. 1 - (a) PLL with embedded multi-core oscillators. (b) Digital PLL array concept, where the outputs of independent DPLL cores are combined together to generate the LO output.
(c) Theoretical improvement of the SSCR from a single DPLL output to the combined output.

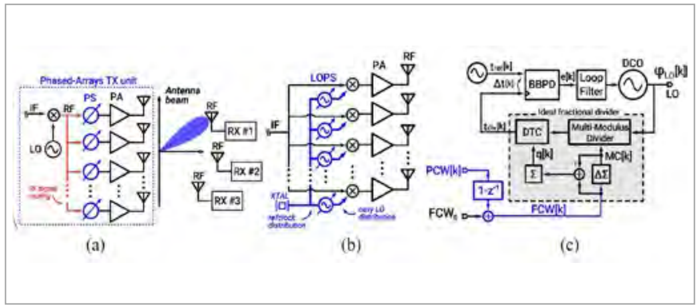

Fig. 2 - (a) Conventional phase array architecture with RF phase shifting. (b) Novel phase array architecture utilizing a digital PLL array. (c) Digital PLL architecture with a dedicated phase control input (PCW).

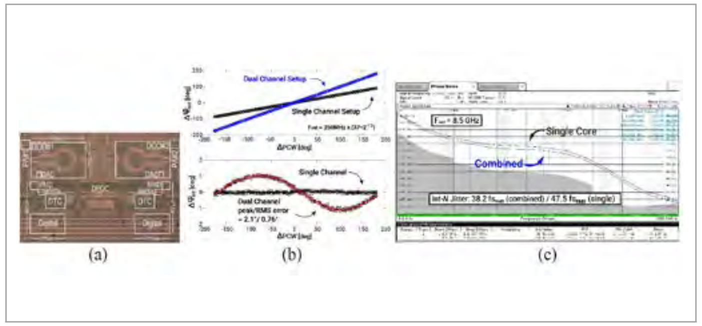

Fig. 3 - (a) Photo of the implemented dual-core prototype in 28nm CMOS. (b) Measured phase-shifting linearity. (c) Phase noise performance of single core and combined output.

GREEN, RESILIENT, AND SECURE NEXT-GENERATION OPTICAL NETWORKS

Qiaolun Zhang - Supervisor: Massimo Tornatore

1. Introduction

Future optical networks represent the essential backbone for various communication systems, such as long-haul, metro, and data center networks, etc. As future optical networks will have to handle a dramatic increase in volumes of highly sensitive data, it is crucial for future optical networks to energy-efficiently accommodate rising traffic demands as well as to be resilient and secure to prevent data loss or leakage to attackers. Specifically, the current communication systems already contribute approximately 2%-3% of global energy consumption, and the energy consumption of the future networks will increase with the ever-growing traffic demands. It is vital to develop sustainable solutions for future optical networks that reduce energy consumption. Furthermore, the future optical networks serve as an essential component for future networks such as Sixth Generation (6G) networks, which will serve various services such as autonomous driving and remote medicine. These services require developing proposals to ensure uninterrupted service during network failures (e.g., due to natural disasters). In the meantime, the data of these services contains various sensitivity information. With the development of quantum technologies, quantum computers pose new threats to future optical networks by comprising traditional cryptosystems. It is important to develop effective countermeasures against potential quantum attacks. In the thesis, I investigate the energy efficiency, resiliency, and security of future optical networks.

2. Solutions to Improve Energy Efficiency of Future Optical Networks

Achieving high energy efficiency for future optical networks is the first part of this thesis. Advances in novel optical-transmission technologies offer new possibilities for reducing the energy consumption of future optical networks. In this thesis, I investigate resource allocation algorithms to improve the energy efficiency of future optical networks using two emerging technologies: ZR/ZR+ pluggable optics and power profile monitoring (PPM) as follows.

- I performed an extensive power-consumption analysis of different IPoWDM network architectures, comparing the impacts of bypassing, regeneration, and grooming. Results show that transparent architectures are still more energy-efficient in both national-wide and continental-wide topology.
- I explored the optimal placement of PPM modules and compared their cost and power consumption to OTDR with an Integer Linear Programming (ILP) model and an optimized monitoring placement algorithm. Results show that, for a national-wide and a continental-wide topology, PPM should consume no more than 8% and 33% of transponder power, respectively. These findings can provide guidelines for the implementation and deployment of PPM.

3. Solutions to Improve Resiliency of Future Optical Networks

Guaranteeing high network resiliency for future optical networks is the second part of this thesis. This part focuses on guaranteeing resiliency for network virtualization in future optical networks. More specifically, this thesis focuses on designing proactive solutions to prepare the network before double-link failures and reactive solutions to recover the network after massive failures as follows.

- I proposed new techniques, SINC (SVNM with inter-VN capacity sharing) and SINC+ (SINC with spare slice sharing), to reduce required resources to guarantee network resiliency against double-link failures. To evaluate their performance, I developed ILP models for both SINC and SINC+, alongside a local search-based heuristic algorithm. Results show that SINC improves VN availability by up to 9.48% over SVNM without increasing total link resource consumption (TLRC). SINC+ ensures survivability against all double-link failures with up to 62.62% more TLRC compared to SVNM.
- I introduced the Progressive Slice Recovery (PSR) problem, which is crucial for rapid recovery in 5G networks with high-reliability demands. I classified recovery strategies with different service disruptions to the network and also considered ensuring PSR with guaranteed reachability among network nodes (network connectivity) or reachability of data centers (content connectivity). An ILP model and a scalable heuristic approach using deterministic rounding and Column Generation (CG) were developed to maximize the accumulative weighted number of recovered slices and virtual links (AWRSL). Results show that strategies allowing re-embedding improve AWRSL by up to 300%, and ensuring CC over NC boosts AWRSL by up to 8%.

4. Solutions to Improve Security of Future Optical Networks

Emerging threats posed by quantum computers put both current and future networks at risk. To secure optical networks from these quantum attacks, quantum technologies are being explored as a secure countermeasure. In this thesis, I focus on investigating novel resource allocation algorithms for Quantum Key Distribution (QKD) networks together with novel resource allocation algorithms for quantum networks, aiming to distribute sufficient keys to secure future optical networks as follows.

- I introduced a novel problem of Routing, Channel, Key-rate, and Time-slot Assignment (RCKTA) for QKD networks. I investigated four network settings, incorporating reducing the transceivers with optical bypass (OB) and improving secret key rate with trusted relay (TR). I formulated MILP models applicable to all the QKD network settings and proposed a scalable heuristic algorithm. Results show that the combination of OB and TR achieves the highest acceptance ratio, while OB has a higher acceptance ratio with limited QKD modules.
- I introduced the Link Configuration for Fidelity-Constrained Routing and Purification (LC-FCRP) problem in quantum networks, addressing the challenge of improving the quality of distributed entanglements. Through link configurations, I greatly improved the performance of the entanglement routing process. I formulated a simplified version of the LC-FCRP problem as a MILP model and proposed a scalable two-phase link configuration algorithm with a low optimality gap. Numerical results indicate that the proposed link configuration solution can improve the acceptance ratio achieved by existing entanglement routing algorithms by up to 87%.

5. Conclusions and Future Works

In this dissertation, I investigated resource allocation algorithms for green, resilient, and secure future optical networks. I began by discussing the state-of-the-art studies on energy efficiency, resiliency, and security for future optical networks and analyzed the corresponding key challenges. To address these challenges, I studied novel resource allocation algorithms for future optical networks. In particular, I focused on enhancing energy efficiency with novel optical-transmission approaches, improving network resiliency with proactive and reactive strategies, and enhancing network security with quantum technologies. The comprehensive results demonstrate the effectiveness of the proposed solutions for future optical networks.

Following the research problems of this thesis, some possible future works are listed below. a) Minimize power consumption of optical networks with ZR+ through network reconfiguration. b) Guarantee the resiliency of QKD and quantum networks. c) Enhancing security with optimized QKD and quantum network deployment.