Horizon 2020 Program

WIDESPREAD-03-2018-TWINNING

Innovation and excellence in massive-scale communications and information processing

INCOMING (Project No. 856967)

D3.3: Novel distributed algorithms, software and hardware implementation, and 5G-IT-Hub showcasing[1]

Abstract: This document presents a report on the WP3 research and development activities that resulted in significant fundamental results published in leading journals and conferences. The report also includes activities related to the development of testbed and experimentation capacities at the ICONIC centre and their applications in real-world scenarios.

Contractual Date of Delivery	30/06/2023
Actual Date of Delivery	30/06/2023
Deliverable Security Class	Public
Editors	Alexandre Graell i Amat (CHALMERS) Rastislav Struharik (FTN) Dejan Vukobratovic (FTN)
Contributors	AAU, FTN, DLR, CHALMERS
Quality Assurance	Milica Petkovic (FTN) Zivko Bojovic (FTN)

Document Revisions & Quality Assurance

UNIVERZITET U NOVOM SADU FAKULTET TEHNIČKIH NAUKA (FTN)	Coordinator	RS
AALBORG UNIVERSITET (AAU)	Participant	DK
CHALMERS TEKNISKA HOEGSKOLA AB (CHALMERS)	Participant	SE
DEUTSCHES ZENTRUM FUER LUFT - UND RAUMFAHRT EV (DLR)	Participant	DE

[1] The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 856967.

Internal Reviewers

1. Milica Petkovic (FTN)
2. Zivko Bojovic (FTN)

External Reviewer

1. Cedomir Stefanovic (AAU)

Revisions

Version	Date	By	Overview
v3.0	30/06/2023	Editor	Final version.
v2.0	25/06/2023	Editor and External Reviewer	Second draft.
v1.0	15/06/2023	Editor and Internal Reviewers	First draft.

Legal Notice

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AAU	Aalborg University
ADMM	Alternate Direction Method of Multipliers
AI	Artificial Intelligence
BP	Belief Propagation
CHALMERS	Chalmers University of Technology
CSI	Channel State Information
DL	Deep Learning
DLR	German Aerospace Centre
EU	European Union
FL	Federated Learning
FPGA	Field Programmable Gate Array
FTN	Faculty of Technical Sciences, University of Novi Sad
GNN	Graph Neural Networks
GPU	Graphical Processing Unit
ICT	Information and Communication Technologies
IoT	Internet of Things
ML	Machine Learning
MPI	Message Passing Interface
RF	Radio Frequency
RPi	Raspberry Pi
UNS	University of Novi Sad
VLC	Visible Light Communications

Executive Summary

Large-scale information processing and machine learning algorithms are one of the two main research and development topics of the INCOMING project. Massive-scale information processing covered in WP3 complements the research and development on massive-scale information acquisition carried out in WP2. Aiming excellence in large-scale information processing and machine learning algorithms, ICONIC researchers focused their efforts in two directions: i) development of fundamental theoretical results, and ii) development of a testbed capacity that will provide support for proof-of-concept demonstrations and create a pathway towards the innovative solutions. This deliverable provides an in-depth explanation of the main research results and testbed and demonstration activities at the ICONIC centre developed during the INCOMING project in the domain of large-scale information processing and machine learning algorithms.

1 Introduction

Large-scale distributed information processing algorithms are in the focus of WP3 that coordinates the research and development activities and knowledge transfer from the EU partners to the ICONIC centre researchers. Through well-balanced blend of staff exchanges, summer schools and expert trainings, WP3 will boost the research level of the ICONIC researchers through joint research efforts, while gradually shifting towards development of fast prototyping skills using hardware-based computing modules ranging from a simple and distributed stack of Raspberry Pi 4 (RPi4) platforms for parallel processing, via graphical processing units (GPU) to specialized field-programmable gate array (FPGA)-based implementations.

As one of the objectives of WP3, the INCOMING project identifies enhanced development and innovation capabilities of ICONIC researchers on hardware-based platforms, through focused development-oriented mini-projects aligned with the staff exchanges. *However, due to COVID-19 situation that affected the INCOMING project already during the M3 (March 2020) of the project realization, the decision is made to move activities of all WP3 tasks to online mode. The situation persisted until 2022 when staff exchange activities could finally start.* This has significantly affected joint development and prototyping activities that benefit from face-to-face meetings and day-to-day interaction. Nevertheless, the second half of the project, including the project extension, led to completion of the project activities according to the work plan.

For completeness, we start this report by presenting a list of research- and development-oriented mini-projects that are identified as promising in M15 of the project (March 2021). After the mid-term review meeting (June 2021), it was agreed that this list should be pruned to a shorter list of promising state-of-the-art distributed information processing and machine learning algorithms for development that should be pursued during the rest of the project duration. The report follows with description of the main research and development areas, where for each area we present both fundamental research results and related proof-of-concept demonstrators.

2 List of Initial Research and Development Mini Projects

During the first reporting period (RP1), 7 research mini-projects are set up by ICONIC researchers in order to initiate collaboration with EU partners in the domain of distributed information processing, machine learning and mobile edge computing. The list of research mini-projects, including the staff members involved, starting month and the status of activities during the mid-term project review (M15, March 2021), is presented in the table below.

Table 1. List of research mini-projects initiated in WP3 during RP1

MP No.	Mini-project title	Start	Team	Status (M15)
MP-R-3.1	Deep Learning for OFDM receiver design – the case study of IEEE 802.11ah standard	M3	Vukan Ninkovic (FTN) Dejan Vukobratovic (FTN)	Conference paper published (IEEE VTC Fall 2020). Journal version submitted.
MP-R-3.2	Gaussian Belief Propagation in Power Systems	M3	Mirsad Cosovic (FTN) Dejan Vukobratovic (FTN)	Work by ICONIC team. Journal paper accepted for publication.
MP-R-3.3	Deep-learning based design of unequal error protection codes	M9	Vukan Ninkovic (FTN) Dejan Vukobratovic (FTN) Christian Haeger (CHALMERS) Henk Wymeersch (CHALMERS) Alexandre Graell i Amat (CHALMERS)	Initial version submitted for journal publication. Work continues towards additional results.
MP-R-3.4	Graph Neural Networks for Smart Grid applications	M12	Mirsad Cosovic (FTN) Ognjen Kundacina (FTN) Dejan Vukobratovic (FTN)	Initial work done by FTN ICONIC team.
MP-R-3.5	Sparse coding for memristive memories	M12	Jovana Zoranovic (FTN) Stanisa Dautovic (FTN)	Initial work done by FTN ICONIC team.
MP-R-3.6	Hardware acceleration of sparse support vector machines for edge computing	M6	Vuk Vranjkovic (FTN) Rastislav Struharik (FTN)	Work by ICONIC team. Journal paper accepted for publication.
MP-R-3.7	Hardware acceleration of convolutional neural networks based on kernel clustering and resource-aware pruning	M6	Damjan Rakanovic (FTN) Vuk Vranjkovic (FTN) Rastislav Struharik (FTN)	Work by ICONIC team. Journal paper accepted for publication.
MP-R-3.8	Universal reconfigurable hardware accelerator for sparse machine learning predicting models	M6	Predrag Teodorovic (FTN) Vuk Vranjkovic (FTN) Rastislav Struharik (FTN)	Work by ICONIC team. Journal paper in preparation.
MP-R-3.9	Heterogeneous multi-core architecture for hardware acceleration of convolutional neural networks	M6	Rastislav Struharik (FTN) Vuk Vranjkovic (FTN) Predrag Teodorovic (FTN)	Initial work done by FTN ICONIC team.
MP-R-3.10	Hardware acceleration of 3D convolutional neural networks	M6	Vuk Vranjkovic (FTN) Rastislav Struharik (FTN) Nikola Kovacevic (FTN)	Initial work done by FTN ICONIC team.
MP-R-3.11	Hardware acceleration of capsule networks	M6	Vuk Vranjkovic (FTN) Rastislav Struharik (FTN) Djordje Miseljic (FTN)	Initial work done by FTN ICONIC team.

MP-R-3.12	Information-theoretic methods in neural information processing	M6	Gorana Mijatovic (FTN) Tatjana Loncar-Turukalo (FTN) Ivan Lazic (FTN)	Work by ICONIC team. Journal paper accepted for publication.

Note that one of the research mini-projects (MP-R-3.3) resulted in the first joint journal paper between ICONIC and CHALMERS which achieved the milestone MS8 (first joint journal paper accepted for publication).

During the first reporting period (RP1), 3 development mini-projects are set up with active participation of ICONIC researchers and guidance provided by EU partners in the domain of distributed information processing, machine learning and mobile edge computing. The list of development mini-projects, including the staff members involved, starting month and the current stage of the project, is presented in the table below.

Table 2. List of development mini-projects initiated in WP3 during RP1

MP No.	Mini-project title	Start	Team	Status (M15)
MP-D-3.1	Large-scale Belief Propagation experimentation testbed based on networked Raspberry Pi4 stack	M3	Mirsad Cosovic (FTN) Dragisa Miskovic (FTN) Dejan Vukobratovic (FTN)	Initial version of the solution successfully demonstrated. Work towards journal paper in progress.
MP-D-3.2	FPGA demonstrator for hardware accelerators of convolutional neural networks	M3	Predrag Teodorovic (FTN) Vuk Vranjkovic (FTN) Rastislav Struharik (FTN)	Work in progress.
MP-D-3.3	Deep Learning algorithms for packet detection and CFO estimation in IEEE 802.11ah standard	M3	Vukan Ninkovic (FTN) Dejan Vukobratovic (FTN)	Initial version of the solution successfully demonstrated.

In this report, we present a research and development work performed during the INCOMING project in the domain of selected distributed information processing, machine learning and edge computing algorithms. For each topic, we discuss the following objectives:

- **Fundamental research results:** Where applicable, we start review of each research and development topic by presenting theoretical results developed during the project that have been published in renowned (IEEE) conference and journal venues.
- **Information processing testbed development:** Where applicable, we describe the process of integration of hardware-based and hardware-accelerated testbeds that enable ICONIC researchers to gain experimental and development skills and quickly turn their ideas into the proof-of-concept prototypes.
- **Distributed information processing, machine learning and edge computing demos:** Within each of the identified testbeds, we identify, develop and describe specific demo use cases that will demonstrate the testbed capabilities and direct the development towards innovative information processing and machine learning solutions.
- **Testbed/Use case exploitation:** Finally, we present exploitable results generated during the project or planned beyond the project lifetime such as data sets, research and demo papers, EU and national funding project applications.

In the following sections, using the above objectives, we describe research and development large-scale information processing, machine learning and edge computing solutions pursued during the INCOMING project. For every solution, we describe relevant fundamental results developed during the project and the testbed developed to present specific demo use cases finalising with discussion on how such use case will be exploited during the project and after its lifetime.

3 Large-scale distributed information processing

Background: One of the central research topics of a group for information processing in ICONIC centre is distributed and decentralized information processing algorithms. More precisely, information processing group focuses on large-scale and distributed information processing and machine learning methods. This includes both scalable and secure algorithms and system architectures:

- *Large-scale distributed information processing algorithms*: 1) Distributed optimization methods, 2) Large-scale machine learning algorithms including deep learning, 3) Distributed probabilistic inference, 4) Privacy-aware learning methods.
- *Large-scale network architectures*: 1) Mobile Edge Computing (MEC) and Mobile Cloud Computing (MCC), 2) Software defined networking (SDN) and Network function virtualization (NFV), 3) Distributed storage, 4) Virtualized hardware architectures (virtual FPGAs/GPUs in the Edge/Cloud)

3.1 Fundamentals of Distributed Information Processing

ICONIC researchers are active in the domain of distributed state estimation algorithms in power systems using factor graphs and the Gaussian Belief Propagation (GBP) algorithm. During the INCOMING project, this research is both expanded and extended to the domain of Graph Neural Networks (GNN), including comparison between model-based and data-based methods.

Distributed State Estimation in Power Systems: The state estimation (SE) is a key functionality of the energy management system (EMS) whose aim is to provide a timely estimate of the system state variables (magnitude and angle of the voltage) at all the buses of the power system. Traditional EMS is centralised, where data collection and its processing is concentrated at a single node. 5G supports centralised EMS by deploying its functions as cloud-native applications within the central cloud. Recent trends see shifting the SE functionality from centralised to distributed system architecture. Distributed SE implies the absence of the central coordinator, where each local area communicates only with its neighbors. In terms of estimation accuracy, the distributed approach is equivalent to the hierarchical and centralized. For latency-critical scenarios, distributed SE over 5G networks with distributed information acquisition and processing represents the most promising approach. In the envisioned architecture, the distributed SE could be virtualized and deployed in numerous edge servers across the 5G network. In the following, we describe two main approached investigated by ICONIC researchers: GBP as a representative model-based approach, and GNNs as representative data-based approach.

Model-Based Distributed SE: The state-of-the-art model-based distributed SE algorithms exploit the matrix decomposition techniques applied over the Weighted Least Squares (WLS) method. In particular, the SE algorithms based on distributed optimization combined with the alternating direction method of multipliers (ADMM) have become popular in the literature. To decentralise an optimisation problem, ADMM decouples the objective function with consensus variables. The resulting algorithm can be interpreted as an iterative message-passing procedure, in which agents solve subproblems independently.

Another efficient iterative message-passing algorithm for distributed inference is GBP. Therein, the power system network with a given measurement configuration is mapped onto an equivalent factor graph containing the set of factor and variable nodes, as illustrated in Figure 1. Factor nodes are defined by the set of measurements, measurement error and measurement function. The variable nodes are determined by the set of state variables. When applied on factor graphs, the GBP algorithm calculates the marginal distributions of the system of random variables. GBP-based

SE was a principal research topic during INCOMING, building upon the results published before the INCOMING project [1], [2]. During the project, GBP is analysed in the contect of distributed observability analysis [3] and potential combination of distributed GBP and weighted least squares (WLS) methods [4].

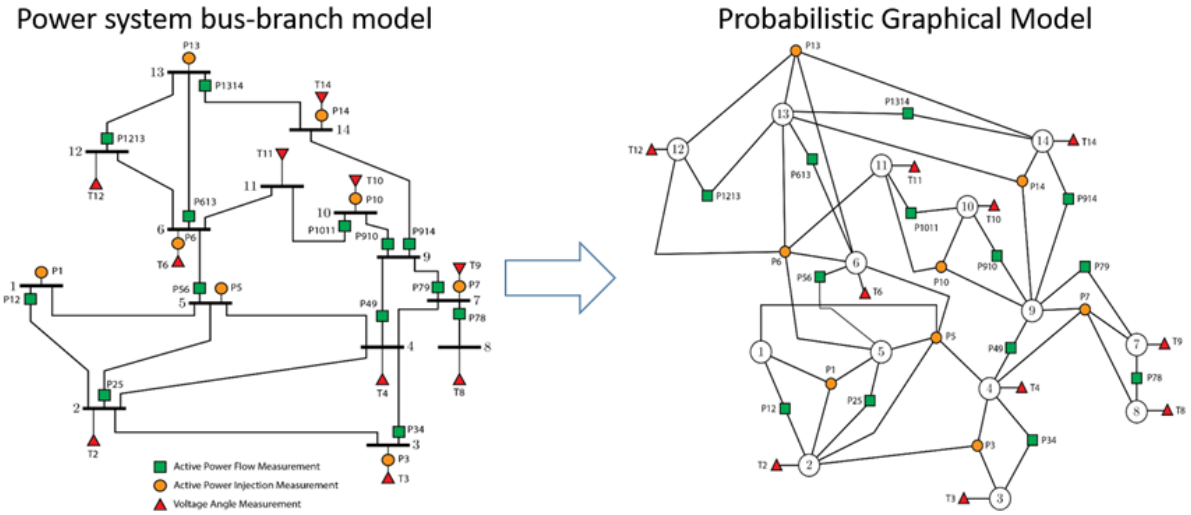


Figure 8. Conversion of a power system bus-branch model into a factor graph [2]

Data-Based Distributed SE: The growing collection of historical measurement data in juxtaposition with complex, often untractable problems has increased interest in developing data-driven SE methods. Data-driven SE, unlike model-based methods, can be designed to avoid using power system's parameters if they are highly uncertain. Deep learning-based approaches that are completely data-driven or hybrid, are typically designed using feed-forward or recurrent neural networks which must be trained on sets of samples with a fixed power system topology, and do not offer a possibility of distributed implementation.

Recent advancements in GNNs can be applied to solve the problems specific to applying deep learning in power systems. GNNs learn from the graph- structured data by recursively aggregating the neighboring node vector embeddings, and transforming them non-linearly into the new embedding space. The node embeddings are initialised by dataset inputs, followed by a predefined number of neighborhood aggregations, the GNN outputs the final node embeddings that can be used for classification or regression problems, as illustrated in Figure 2. Apart from not being restricted to training and test examples with fixed topologies, GNNs have fewer trainable parameters, lower memory requirements, and can easily incorporate connectivity information into the learning process. The centralised implementation of the trained GNN model's inference results in linear computational complexity with the number of nodes in the power system (assuming the constant node degree). Additionally, GNN-based SE can be computationally and geographically distributed across multiple processing units, with the requirement that all of the measurements in the k-hop neighbourhood are gathered and sent into the unit that predicts the state variables for each node. Our recent work on GNN-based in SE is presented in more details in [5], [6].

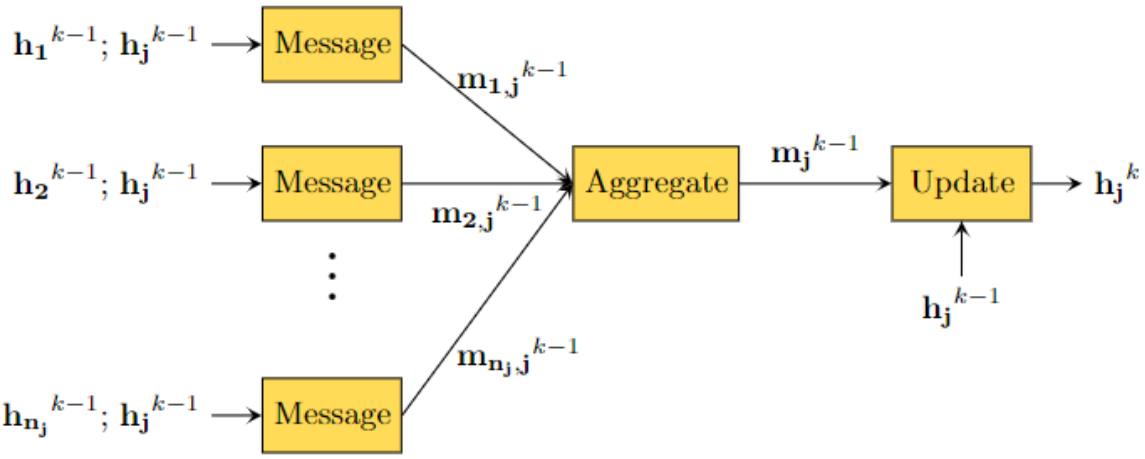


Figure 9. Message-passing GNN architecture for power system SE [6]

Distributed SE and 5G Edge Computing: ICONIC researchers started working on Smart Grid and 5G integration before the INCOMING project, where the main idea was on how to efficiently integrate distributed state estimation from a signal processing [7] and system architecture [8] perspective. This work is expanded during the INCOMING project in both directions. In terms of fundamental signal processing exploration, we focused on the performance and optimization of GBP applied over a large-scale factor graph divided into areas, where each area is running in parallel on different and geographically separated edge node, as illustrated in Figure 3. We proposed so called alternating GBP (AGBP) and presented its theoretical and numerical analysis which demonstrate it performs more efficiently than the GBP with synchronous scheduling [9].

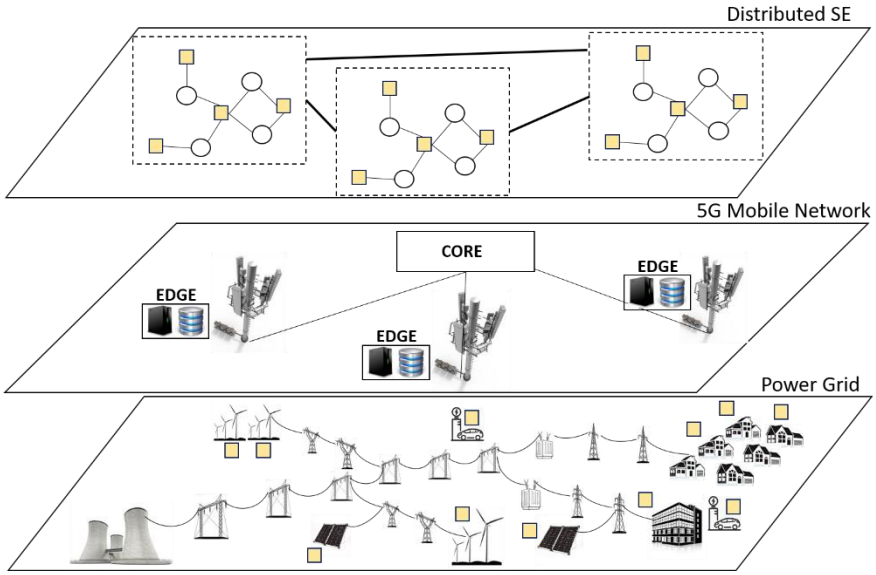


Figure 10. Distributed factor graph and GBP across 5G edge computing nodes [8]

Integration of distributed SE as a set of cloud-native centralised or distributed functions along with the sensory devices connected via 5G RAN that provide inputs for the SE process is investigated in [10]. 5G O-RAN architecture and virtualised cloud-native 5G core network service based architecture are investigated as an ideal support for integration of 5G and Smart Grid services. Figure 4 below illustrates concepts of 5G and SG integration.

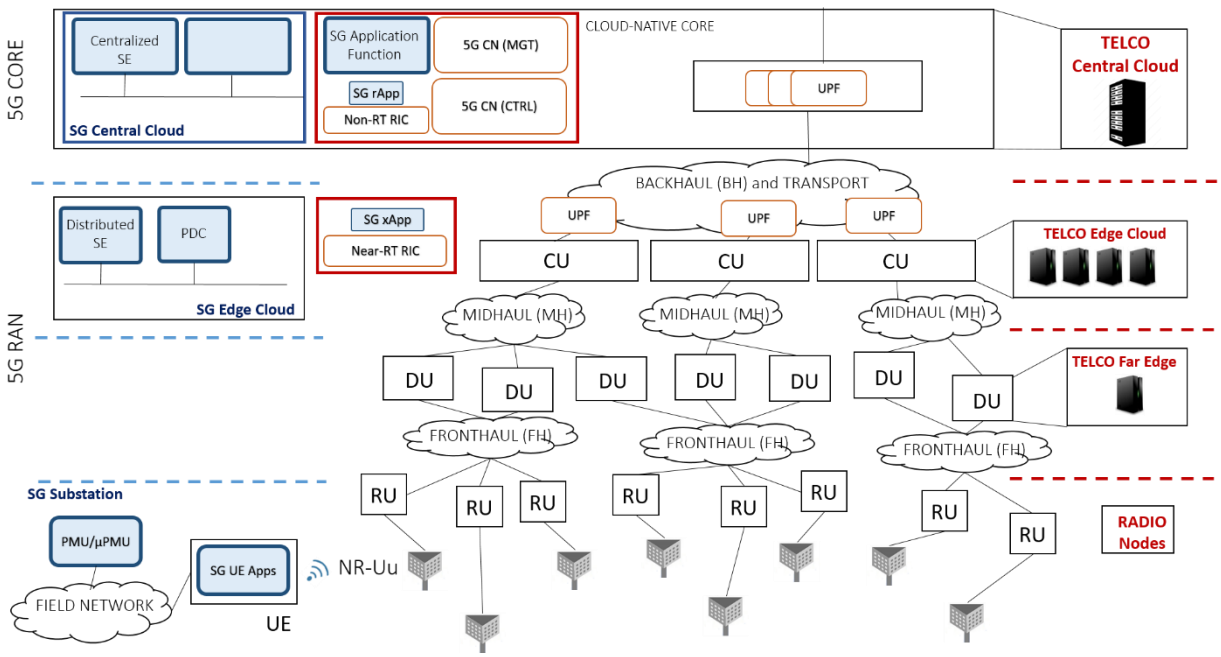


Figure 11. 5G and Smart Grid architecture integration [10]

The work on large-scale distributed algorithms and their applications in the context of 5G edge computing resources has been significantly extended during the project, leading to three top-tier journal publications (IEEE Transactions on Power Systems, IEEE Internet of Things Journal and Elsevier Sustainable Energy, Grids and Networks) and a number conference papers published at top-tier IEEE conferences such as IEEE Smart Grid Communications conference.

3.2 Testbed and Experimentation for Distributed Information Processing

- Distributed State Estimation in Power Systems: In Figure 5, we illustrate the setup for distributed state estimation (SE) testbed, where a power system is emulated on a server and measurements from the system are forwarded to distributed computing nodes in order to perform dynamic state estimation in a distributed fashion. We use factor graphs and BP algorithm as a principal algorithmic framework [1]-[4], however, our ongoing work also involves integration of data-based methods that use Graph Neural Networks (GNNs) as an SE solver evaluated on this testbed [5].
- Federated learning: Distributed machine learning methods based on federated learning (FL) are a part of several ongoing research investigations in ICONIC, ranging from Smart Building applications (illumination modelling), digit recognition for legacy electricity/water/gas metering applications, and wireless fingerprinting methods. RPi testbed represents suitable platform for emulation and testing of FL methods.

In order to design, implement and test various distributed algorithms in realistic and controllable testbed environment, we have developed a distributed computing testbed based on 40 quad-core Raspberry Pi4 (Rpi 4) devices networked via a gigabyte Ethernet switch. The testbed, illustrated in Figure 5, enable us to deploy various information processing and machine learning algorithms on individual RPi 4 nodes, that perform local computing tasks and exchange information between themselves, with the goal of solving a global computation, optimization or learning problem. The

testbed is based on message passing interface (MPI) library that is supported both for projects developed in C/C++ or Python. Communication between the computing nodes can be additionally controlled for packet drop rates and packet delays which enables testing of the algorithms in different conditions.

Hardware support: For the development of distributed information processing testbed, ICONIC possesses all necessary equipment, including:

- 40 Raspberry Pi 4 devices each with 8GB RAM memory
- Gigabit Ethernet switch for network connectivity

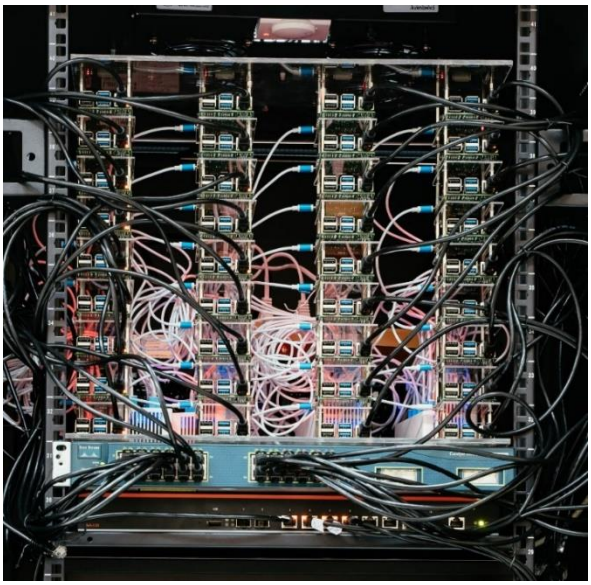


Figure 12. Stacked RPi 4 testbed (40 nodes)

Software support: All the software deployed and used in distributed information processing testbed is developed by ICONIC researchers. The underlying framework for distributed computation is message passing interface (MPI) library. MPI enables a framework for distributed computation and intermediate results exchange between the working nodes. It is suitable for development of various algorithms that rely on message-passing computation methods, such as distributed optimization methods (e.g., alternate direction methods of multipliers - ADMM), probabilistic inference methods (e.g., Belief Propagation algorithm), distributed learning methods (federated learning and Graph Neural Networks - GNNs) and many others. As such, it is generic enough to enable experimentation with a wide set of algorithmic tools, as illustrated in Figure 6.

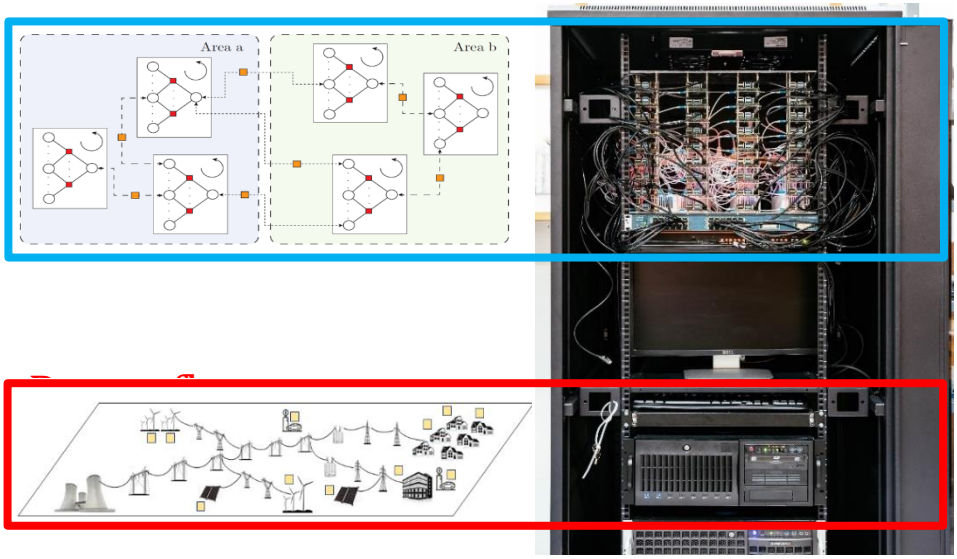


Figure 13. RPi 4 testbed for distributed power system state estimation

Final integrated testbed (M42): At its present state, the testbed represents a useful development tool for the design and development of distributed algorithms. The development of the final testbed included improvements along several directions. For example, we introduced additional software modules for control of inter-node communication delays and packet lossess. We developed a simple and flexible tool that integrates a desired communication link/network model into the

distributed computation framework. For example, modelling inter-node communication through mobile core network, or modelling communication between sensing unit and edge processing via 4G/5G radio interface is integrated to emulate realistic conditions of algorithm deployment as part of the mobile cellular edge/fog/cloud infrastructure.

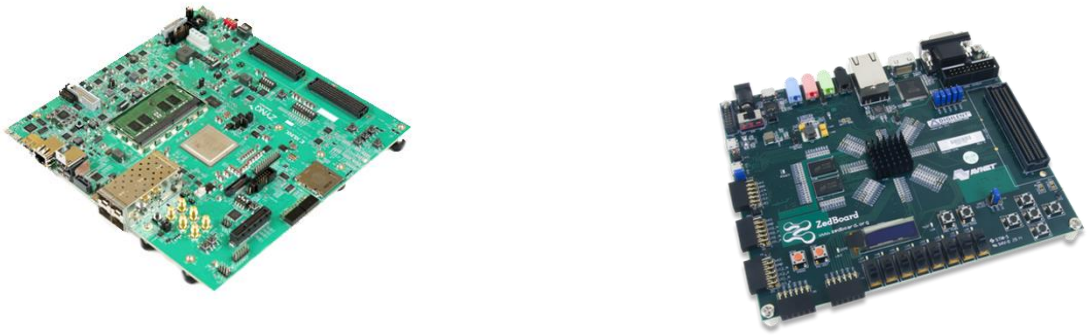
Exploitation and outreach directions: The distributed information testbed is and will be used for the design, development, testing and benchmarking various distributed information processing, optimization, inference and learning algorithms. The testbed will be used to support publishing demo and research papers at conferences and journals. Besides research and development, the testbed will be used in teaching, demonstration, and for training students.

4 FPGA-accelerated machine learning

Background: In recent years moving AI algorithms to the Edge became an increasing trend, both in academic research and industrial domains. Deploying AI algorithms, and deep learning algorithms in particular, on the Edge requires development of customized computing architectures that can meet severe power, compute and memory constraints, while still being able to reach required performance levels. Furthermore, given the volatile nature of AI algorithms which are still changing rapidly, developed solutions should be able to accommodate these changes. FPGA technology is an ideal platform for enabling rapid development of hardware accelerators of state-of-the-art AI/ML/DL algorithms. Part of ICONIC staff has a significant experience in designing such acceleration systems, having developed a number of different ML hardware acceleration IP cores. However, to be able to easily validate correct operation, measure performance in real-life settings of developed AI/ML/DL hardware accelerators, and act as a technology demonstration platform, a standardized hardware testbed is currently missing.

Hardware support: ICONIC possesses all necessary equipment required to develop an FPGA-based testbed for FPGA-accelerated machine learning, including:

- One Xilinx Zynq UltraScale+ MPSoC ZCU102 FPGA development board, based around Zynq UltraScale+ XCZU9EG FPGA device, enabling development of high-performance FPGA-based ML/DL hardware accelerators, shown in Figure 7a,
- 3 Xilinx ZedBoard development systems, based on the Zynq-7000 SoC XC7Z020 FPGA device, that are ideal for development of performance constrained, cost-optimized FPGA-based ML/DL hardware accelerators, shown in Figure 7b,
- A number of cameras, which will be used for image and video acquisition, to validate the correct operation of developed solutions,
- A monitor that will be used to visualize the testing process.



a) b)

Figure 14. Development boards used to build the FPGA-based testbed for accelerated Edge-based machine learning, a) ZCU102 development board, b) ZedBoard development board

Software support: For testbed that will be developed, we will also develop our own firmware, that will be executed on the embedded ARM processor cores, located within the selected FPGA devices that are present on two target FPGA development boards. This firmware will enable an easy way of integrating novel FPGA-based ML accelerators within the complete FPGA-based system. Firmware will also perform the image acquisition from the attached camera, or generation of artificial test data, as well as any post-processing of data generated by the ML accelerator IP core undergoing testing, and visualization of final results.

Final integrated testbed (M32):

Currently there is no one, standardized, testbed that can be used to validate correct operation and measure performance of developed ML hardware accelerator systems. Instead, for each ML accelerator a separate testbed has to be developed. This is not an efficient approach in terms of development time, and it can also lead to problems when comparing performance of different accelerator systems. An initial effort has been put into developing one, universal testbed, based on available FPGA development boards, that will offer a standardized way of integrating different ML accelerator IP cores into a fully-functional system, that could then be used to validate correct operation, measure performance, and also serve as a technology demonstrator.

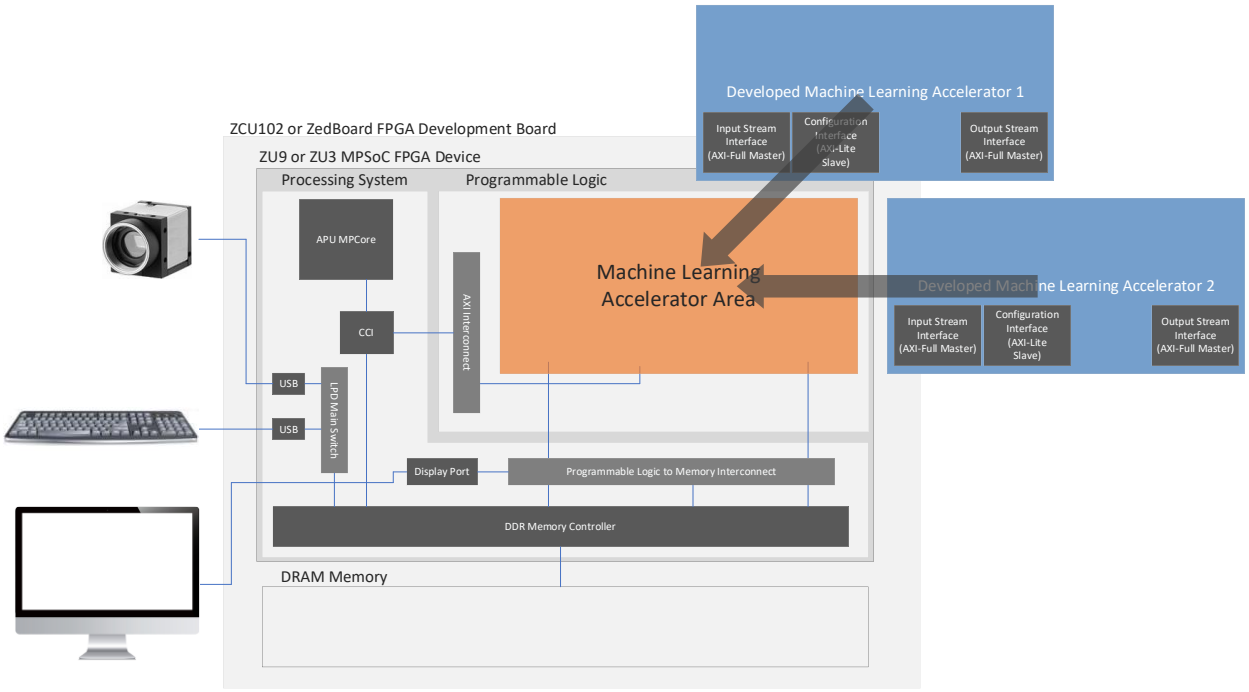


Figure 15. Overall structure of hardware-accelerated ML testbed

Structure of the final testbed, developed during this project, is shown in Figure 8. Testbed is based either on ZCU102 or ZedBoard FPGA development boards. ZCU102 development board is used to validate, measure performance and showcase more complex, high performance, ML accelerator systems. In contrast, ZedBoard development board is used to showcase more cost-optimized solutions. It is important to note that both development boards are used to implement one unified HDL design of the testbed, thus saving testbed development and maintenance time. HDL testbed design is based on the existing Processing System, containing an embedded ARM processor, which is available on Xilinx Zynq UltraScale+ MPSoC FPGA devices. The embedded ARM processor is used to configure, control and monitor the operation of the currently integrated ML accelerator IP core, through a firmware designed specifically for this purpose. This firmware is based on the

Embedded Linux operating system, which allows easy integration of various additional peripheral units into the testbed system, like camera, monitor and keyboard. Camera, using either an USB or Ethernet interface, provides video stream or individual images that are transferred to the integrated ML accelerator core for processing. Processed images, or video, are then visualized using an attached monitor device. Using a keyboard, operation of the testbed is controlled by a person. During the testing procedure, various performance data can be collected and later transferred to the PC computer for post-processing.

Developed ML hardware accelerator IP cores are implemented using a Programmable Logic zone of the Xilinx Zynq UltraScale+ MPSoC FPGA devices. To allow easy integration of different ML accelerator IP cores, all of them implement a standardized communication interface, comprising of a number of AXI-Full interfaces for data movement, and one AXI-Lite interface for control by the embedded ARM processor, as shown in Figure 8. For each of developed ML accelerators, a specific Linux device driver is developed, enabling easy integration in the Linux operating system and the main testbed control application.

This setup allows standardized, quick, and easy integration of different ML accelerators into one universal testbed system, thus significantly reducing development time required to prepare a hardware setup for validation and performance estimation of selected ML accelerator.

FPGA-accelerated ML solution for research, demonstration and innovation: After the ML hardware accelerator testbed system is developed, we use it to validate the correct operation and measure performance in real-life scenarios of several DL hardware accelerator IP cores that are being developed by the ICONIC staff members, as part of the INCOMING project. These include the following:

- Hardware accelerator IP core for sparsified convolutional neural networks,
- Hardware accelerator IP core of convolutional neural networks based on kernel clustering and resource-aware pruning,
- Heterogeneous multi-core IP architecture for hardware acceleration of convolutional neural networks,
- Universal reconfigurable hardware accelerator IP core for sparse machine learning predicting models.

The details of specific solutions of the new accelerated platform developed during the project is presented as a video tutorial at: https://youtu.be/KQpFTID6QKk or on the following ICONIC webpage link: https://iconic.ftn.uns.ac.rs/gemini-deep-learning-hardware-acceleration-platform/.

Exploitation and outreach directions: Developed ML accelerator testbed will mainly be used for validation and performance measurements of different developed ML hardware accelerator IP core, in real-life settings. Collected data will dominantly be used during the preparation of conference and journal publications. The testbed will also be used as a technology demonstration platform for interested industrial partners for possible future joint commercialization of selected promising ML hardware accelerator IP cores.

5 Machine Learning for Physical Layer Design

Background: End-to-end design of communication systems using deep autoencoders (AEs) is gaining attention due to its flexibility and excellent performance, and therefore, learning transmitters and receivers for a given channel model using deep AEs optimized for a specific loss function has been investigated in the recent literature [11]-[13]. Such AE-based encoders and decoders achieve close-to-optimal performance for some baseline communication scenarios (one-bit quantization channels [14], optical communications [15] and OFDM [16]) and have been demonstrated in a proof-of-concept real-world implementation [13].

These works consider AE-based encoders and decoders that provide equal error protection across the set of transmitted messages. However, due to the specific system requirements in many communication scenarios (transmission of control signals along data, multi-resolution source coding, and ultra-reliable and low-latency communication protocols), one is interested in the design of different types of AE-based error correction codes, such as unequal error protection (UEP) codes.

Besides single-user transmission, AE-based design is recently explored in multi-user setup [11], [13] as the recent trends show that, except for point-to-point communication systems, shifting the design of encoding and decoding procedures from conventional to machine learning (ML)-based methods is expanded towards multi-user non-orthogonal multiple access (NOMA) setup [17], [18], e.g., for NOMA constellations design. Moreover, except for reliable message recovery after receiving the entire encoded message (codeword), AEs can be used for code design in many practical scenarios (low-earth-orbit satellite communications), where the transmission process may be interrupted before receiving the complete codeword.

As we observe from the above examples, different types of AE-based code design can be integrated in different critical communication scenarios, and therefore, they have huge potential for practical implementation in modern communication systems (5G and beyond).

5.1 Fundamentals of Machine Learning Solutions for the Physical Layer

We consider the problem of communicating a message m from a set of messages over a noisy channel. Encoder and decoder can be represented as functions *f* and *g*, respectively, and the main goal is to optimize these functions and design (*f*, *g*) pair which will minimize average message error probability.

From a deep learning perspective, the above communication system can be represented as an AE [1]. An AE consists of a set of encoder layers representing the encoder mapping $x=f(m)$, normalization layer which ensures that codeword met power constraints, the noise layer modeling the channel *W* that transforms *x* into *y*, and a set of decoder layers representing the decoder mapping $g(y)=\hat{m}$, as shown in Figure 9. In this setup, functions (*f*, *g*) are jointly optimized in end-to-end training process by applying stochastic gradient descent (SGD) optimizer on minimization of the cross-entropy loss (used as a surrogate for minimizing the message error probability). This is baseline scenario which is further expanded within INCOMING partners [19], [20].

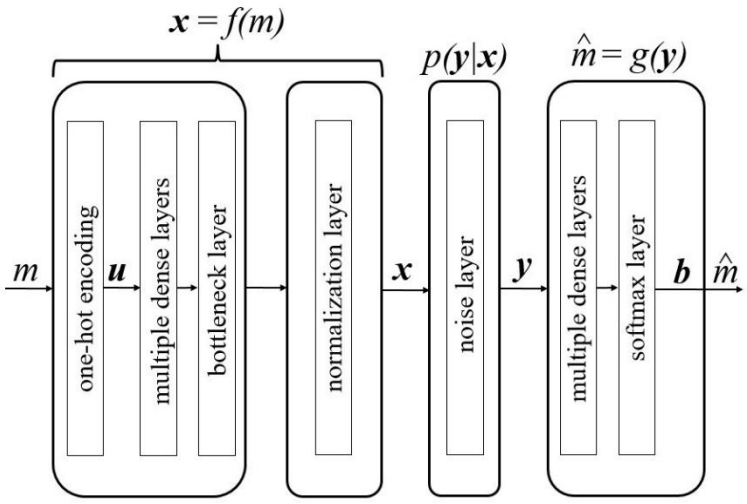


Figure 16. Communication system represented as a deep autoencoder

The first extension relates to the design of AE-based UEP codes [19], where a flexible and efficient method for design encoders and decoders for both message-wise and bit-wise UEP codes by training deep AEs is presented. The key idea of the proposed AE-based design is to define an appropriate compound loss function that comprises a weighted contribution of each importance class and generalizes the cross-entropy loss function to the UEP case. In such a way, by using an associated weight vector, the generalized loss function can be used to flexible trade off error probabilities corresponding to different importance classes and to explore the region of achievable error probabilities.

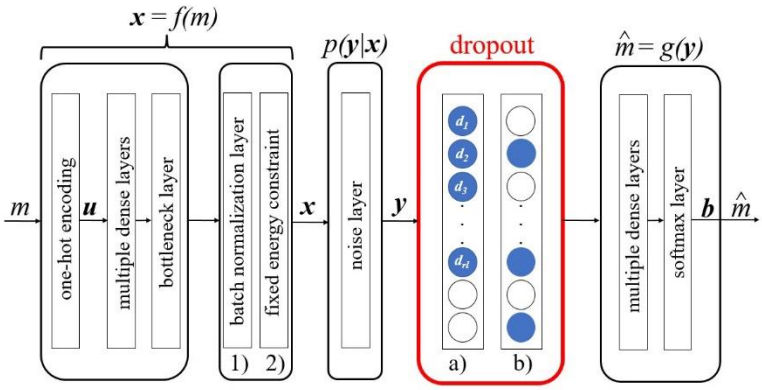


Figure 17. Communication system represented as a deep autoencoder with dropout block extension (red rectangle)

Further, conventional AE-based code design is extended to a class of AE-based codes that shares with rateless codes that receiving additional codeword symbols progressively improves the successful message decoding probability [20]. In more detail, by introducing a suitably defined randomized dropout strategy to match the AE-based code design to a different erasure channel of interests, we obtained a novel class of AE codes, that allow to trade off decoding delay and reliability. To define a generic randomized dropout strategy covering various erasure channel models, we introduce a channel dropout vector ***d*** associated to the channel noise layers, as shown in Figure 10.

AE-based approach is further used to design encoding and decoding solution for downlink NOMA [21], as conventional NOMA downlink communication system with successful interference cancelation (SIC) decoding can be implemented in a DNN fashion as an extension of the end-to-end AE-based scheme (Fig. 11, a), blue blocks), as illustrated in Fig. 11 (b) red blocks) [18]. Building upon the work in [18], we apply: 1) a weighted loss function to control error probability balance across different users, 2) SICNet architecture [17] to enhance deep AE-based decoding capability. Here, the transmitter sends jointly encoded L messages to L different users, i.e., an encoder is jointly optimized with the set of L decoders using an end-to-end AE-based training approach. In order to neglect any knowledge about the channel (which can experience extreme variations in time), weighted sum approach can be also used in the AE-based downlink NOMA by introducing weights associated to L different users, leading to a weighted total loss function (similar to the UEP case, but now we can flexibly trade off error probabilities corresponding to different users).

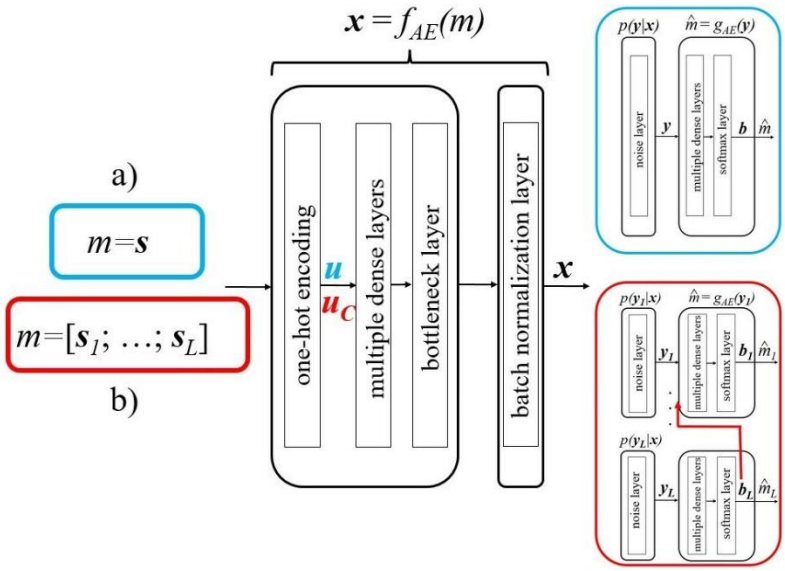


Figure 18. Communication system represented as a deep autoencoder: a) Conventional system - Blue blocks [1] b) NOMA downlink system with SICNet [7] - Red blocks; Black blocks are used by both a) and b))

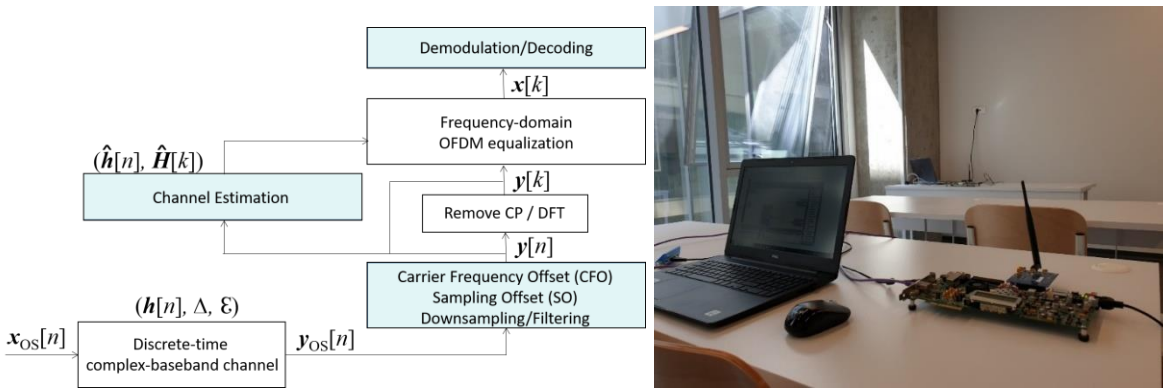


Figure 19. ML-based design of selected PHY blocks at the received: a) Receiver structure and blocks that have been implemented using deep learning, b) testbed implementation

Besides autoencoder-based communication system design, during INCOMING project, we also considered ML-based design of various PHY blocks. This work, done in collaboration with an

industrial partner, focused on the design on timing and carrier frequency offset estimation using deep learning for IEEE 802.11ah receivers, as illustrated in Figure 12. The results of this study is published in leading IEEE conferences and journals (IEEE Access [22] and IEEE VTC 2020).

5.2 Testbed and solutions for ML-aided PHY layer

Background: Integration of wireless communications and machine learning (ML) is gaining momentum in recent years. ML tools can be applied both to improve the design of communication systems. For example, as we detailed in the previous section, ML is shown to be efficient in learning new designs of wireless physical layer. On the other hand, ML can be used to extract features from received wireless signal in order to distinguish from different wireless transmitters (identification), recognize different transmission waveforms (modulation recognition), identify the transmitter position (localization), and many other known and yet unknown applications.

In ICONIC, we are focused on two topics. The first topic is applications of ML/AI methods in the design of various processing blocks at the PHY. The work in this domain is initiated as a collaboration with an industrial partner, where we worked on the design of physical layer at the receiver side using ML to replace conventional methods for IEEE 802.11ah (emerging Wi-Fi IoT standard) packet detection and carrier frequency offset estimation [22]. This work is followed by an extensive investigation of autoencoder-based communication system design in several communication scenarios described in the previous section that involved collaboration between project partners (UNS and CHALMERS) and further extended collaboration between ICONIC researchers and other leading research organisations in Europe, such as CTTC.

In the second topic, we are focused on ML for wireless fingerprinting. The goal is to collect data and extract features from wireless physical layer in order to learn some transmitter properties. We focused on extraction of channel state information (CSI) collected from short range (Wi-Fi) and long range (NB-IoT/LoRa) IoT transmitters embedded at various IoT platforms (RPi 4, ESP 32, custom-designed nodes). Such signals are logged and represented in the form of a data set on which various ML methods are applied. The goal of these studies are reliable device identification and localization in indoor environment using wireless fingerprints.

Hardware support: For this testbed, we use different types of IoT nodes:

- RPi 4 platforms with embedded IEEE 802.11n Wi-Fi modules
- ESP 32 platforms with IEEE 802.11n Wi-Fi and LoRa modules
- Custom based NB-IoT devices fabricated at ICONIC for testing and measurements
- 20 USRP software defined radio modules

Software support: Custom designed software for extraction of wireless CSI from various communication modules (Wi-Fi, NB-IoT). For Wi-Fi, extraction of channel gains of different OFDM carriers is possible and is usually used as CSI. Higher-level features such as received signal strength/power metrics (SNR, RSRP, RSSI) are also used, both in the case of Wi-Fi and NB-IoT and LoRa. Finally, using software defined radios, extraction of raw received baseband signal samples is also possible, providing the lowest level of physical layer signal as an input to ML methods in the form of received complex baseband signal samples.

Testbed integration: The testbed integration assumes placing a set of devices at different indoor locations (for example, a grid of locations in a given space) and collecting a data set extracted from a large set of transmitted data packets (e.g, thousands of packets per location point). From each packet transmission, various information including CSI is extracted and stored in a data base for further processing. During the project, several data gathering campaings have been performed and representative data sets will be shared with the research community in the final stages of the project (for example, we have deployed and created a data set for a set of 20 NB-IoT devices placed in a grid of locations of a single room, see Figure 12). Preliminary results using various ML algorithms demonstrate high distinguishability of signal received from different indoor location (although the receiver, i.e., the macro-cellular base station, is outdoor at the distance of about 300m). The testbed and results have been repeated using Wi-Fi based modules based on RPi 4 and ESP 32 platform. Furthermore, a platform for experimentation based on USRP SDR devices is also established to demonstrate signal reception and extraction of CSI for various wireless technologies (NB-IoT, Wi-Fi, 4G LTE).

Figure 20. Wireless fingerprinting (NB-IoT) setup

Final integrated testbed (M32): Final testbed setup has been integrated to provide fast and automated collection of CSI from devices deployed at different location and spatial resolution and for different use cases, either using commercial off-the-shelf (COTS) devices (RPi 4, ESP 32) with appropriate CSI extraction software, or using USRP SDR modules. For the latter, we used a recent framework for synchronisation and management of SDR lab nodes called Wiscanet (https://github.com/WISCA/wiscanet-source). In the present state, our lab enables fast creation of data sets suitable for training ML algorithms for various problems such as device identification, localisation, context and activity recoginition, indoor space occupancy detection, etc.

Integrated solution for ML research, demonstration and innovation: Regarding end-user scenarios and applications, our work on ML for wireless communications is focused on two principal technologies:

- Wireless fingerprinting for ML-based localisation, identification and context recognition: Channel state information based wireless fingerprints are powerful source of information that depends on both the transmitting device location as well as the environment in which it is deployed. For device identification and localisation, the goal is to extract the features that are dependant only on the device location and harware properties, while averaging out the effects of the environment. In the context recognition case, it is crucial to extract environmental features in order to recognize specific indoor ambient parameters such as the presence of people or objects, their number and location, etc. In this use case, we mainly focus on extraction of CSI from COTS devices for ease of deployment.
- ML for wireless PHY design: Design of PHY using ML ideas and tools are active research area in ICONIC. Our recent work included the design for OFDM-based receivers that use deep learning to optimize performance of specific receiver blocks (packet detection, timing and frequency offset estimation) in IEEE 802.11ah receivers. We have also contributed to the design of autoencoder-based error correction codes and constellation design that

provide for unequal error protection properties. In such work, we typically use SDR modules to demonstrate the proposed concepts as a real-world demo.

Exploitation and outreach directions: We will exploit the above work in the direction of advancing ICONIC research and demonstration capacities. ML based PHY design already resulted in two top-tier journal papers (IEEE Access and IEEE Communication Letters) and a number of conference papers (including top-tier IEEE venues such as IEEE VTC 2020, IEEE ICC 2023 and IEEE SPAWC 2023). ML based autoencoder code design resulted in a milestone result – joint publication between ICONIC and CHALMERS team that is published in IEEE Communication Letters. We have recently collaborated with AAU to prepare Marie Curie Doctoral Networks program that focuses on Smart Building services based on ML applied on wireless fingerprints (e.g., for indoor occupancy statistics and localization).

6 Summary

This report provided an overview of the selected outcomes of the research and development mini-projects developed during the INCOMING project. It presents testbed capacities evolved towards a set of promising use case demonstrators that provide a basis for various dissemination and outreach activities, such as publication of conference and journal papers, presentation of demos at various public events, potential spin-off of startup companies and tighter integration with regional industry in joint research and development projects. The report also underlines the most relevant theoretical findings and respective journal and conference publications developed during the project as part of WP3 activities.

7 References

[1] M. Cosovic, D. Vukobratovic: "Fast Real-Time DC State Estimation in Electric Power Systems Using Belief Propagation," IEEE SmartGridComm 2017 (Best Student Paper Award), Dresden, Germany, October 2017.

[2] M. Cosovic, D. Vukobratovic: "Distributed Gauss-Newton Method for AC State Estimation Using Belief Propagation," IEEE Transactions on Power Systems, Vol. 34, No. 1, pp. 648-658, January 2019.

[3] M. Cosovic, M. Delalic, D. Raca, D. Vukobratovic: "Observability Analysis for Large-Scale Power Systems Using Factor Graphs," IEEE Transactions on Power Systems, Vol. 36, No. 5, pp. 4791-4799, September 2021.

[4] M. Delalic, M. Cosovic, D. Raca, D. Vukobratovic: "Distributed Weighted Least Squares and Gaussian Belief Propagation: An Integrated Approach," IEEE Smart Grid Communications Conference SmartGridComm 2021, Aachen, Germany, October 2021.

[5] O. Kundacina, M. Cosovic, D. Vukobratovic: "State Estimation in Electric Power System Leveraging Graph Neural Networks," 17th Int'l Conference on Probabilistic Methods in Power Systems PMAPS 22, Manchester, UK, June 2022.

[6] O. Kundacina, M. Cosovic, D. Miskovic, D. Vukobratovic: "Graph Neural Networks on Factor Graphs for Robust, Fast, and Scalable Linear State Estimation with PMUs," Sustainable Energy, Grids and Networks (Elsevier), Vol. 34, No 101156, June 2023.

[7] M. Cosovic, D. Vukobratovic, V. Stankovic: "Linear State Estimation via 5G C-RAN Cellular Networks using Gaussian Belief Propagation," IEEE Wireless Communications and Networking Conference WCNC 2018, Barcelona, Spain, April 2018.

[8] M. Cosovic, A. Tsitsimelis, D. Vukobratovic, J. Matamoros, C. Anton Haro: "5G Mobile Cellular Networks: Enabling Distributed State Estimation for Smart Grid," IEEE Communication Magazine, Vol. 55, No. 10, pp. 62-69, October 2017.

[9] M. Cosovic, D. Miskovic, M. Delalic, D. Raca, D. Vukobratovic: "Distributed Inference over Linear Models using Alternating Gaussian Belief Propagation," to appear, IEEE Internet of Things Journal, 2023.

[10] O. Kundacina, M. Forcan, M. Cosovic, D. Raca, M. Dzeferagic, D. Miskovic, M. Maksimovic, D. Vukobratovic: "Near Real-Time Distributed State Estimation via AI/ML-Empowered 5G Networks," IEEE Smart Grid Communications Conference 2022, Singapore, October 2022.

[11] T. O'Shea and J. Hoydis, "An introduction to deep learning for the physical layer," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 4, pp. 563-575, Dec. 2017.

[12] T. O'Shea, T. Erpek, and T. C. Clancy, "Deep learning based MIMO communications," Jul. 2017., arXiv:1707.07980v1 [cs.IT]. [Online]. Available: https://arxiv.org/abs/1707.07980

[13] S. Dörner, S. Cammerer, J. Hoydis, and S. ten Brink, "Deep learning based communication over the air," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 132-143, Feb. 2018.

[14] E. Balevi and J. G. Andrews, "Autoencoder-Based Error Correction Coding for One-Bit Quantization," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3440-3451, June 2020.

[15] S. Li, C. Häger, N. Garcia and H. Wymeersch, "Achievable Information Rates for Nonlinear Fiber Communication via End-to-end Autoencoder Learning," in *Proc. 2018 Eur. Conf. on Opt. Commun. (ECOC)*, Roma, Italy, Sept. 23-27, 2018, pp. 1-3.

[16] A. Felix, S. Cammerer, S. Dörner, J. Hoydis and S. Ten Brink, "OFDM-Autoencoder for End-to-End Learning of Communications Systems," in *Proc. IEEE 19th Int. Workshop on Signal Process. Adv. in Wireless Commun. (SPAWC)*, Kalamata, Greece, June 25-28, 2018, pp. 1-5.

[17] T. Van Luong, N. Shlezinger, C. Xu, T. M. Hoang, Y. C. Eldar, and L. Hanzo, "Deep Learning Based Successive Interference Cancellation for the Non-Orthogonal Downlink," *IEEE Trans. Veh. Technol.*, vol. 71, no. 11, pp. 11876-11888, Nov. 2022.

[18] F. Alberge, "Constellation design with deep learning for downlink non-orthogonal multiple access," in *Proc. 2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Bologna, Italy, 2018, pp. 1-5.

[19] V. Ninkovic, D. Vukobratovic, C. Haeger, H. Wymeersch, and A. Graell i Amat, "Autoencoder-Based Unequal Error Protection Codes," *IEEE Commun. Lett.*, vol. 25, no. 11, pp.3575-3579, Nov. 2021.

[20] V. Ninkovic, D. Vukobratovic, C. Haeger, H. Wymeersch, and A. Graell i Amat, "Rateless Autoencoder Codes: Trading off Decoding Delay and Reliability," in *Proc. Int. Conf. on Commun (ICC2023)*, Rome, Italy, May 28- June 1, 2023, pp. 1-5.

[21] V. Ninkovic, D. Vukobratovic, A. Pastore, and C. Antón-Haro, "A Weighted Autoencoder-Based Approach to Downlink NOMA Constellation Design," in *Proc. IEEE 24th Int. Workshop on Signal Process. Adv. in Wireless Commun. (SPAWC2023)*, Shanghai, China, Sept. 25-28, 2023, pp. 1-5.

[22] V. Ninkovic, A. Valka, D. Dumic, D. Vukobratovic, "Deep Learning Based Packet Detection and Carrier Frequency Offset Estimation in IEEE 802.11ah," *IEEE Access*, 2021.