This project has received funding from the Electronic Components and Systems for European Leadership Joint
Undertaking under grant agreement No 876925

ECSEL Joint Undertaking
Electronic Components and Systems for European Leadership

ANDANTE

AI for New Devices And Technologies at the Edge

D3.4 Building blocks specifications

Deliverable No.	D3.4	**Due Date**	*31-May-2021*
Type	Report	**Dissemination Level**	*Confidential*
Version	1.1	**Status**	Final
Description	Building blocks specifications		
Work Package	WP3 – AI building blocks, Methods and Tools.		

Abstract (Published Summary)

Abstract (Published Summary)

This document defines the neural network building blocks being developed in the context of the ANDANTE project. ANDANTE is targeting edge applications, where the data analysis occurs locally, at device level, to heavily reduce the traffic to the cloud and guarantee real-time responses, security and privacy. In such systems, particular characteristics like low power, low latency, accuracy, reliability, endurance, and long life must be optimized. For the different application uses cases, the associated neural networks require specific elements that must be specified, and its implementation should match the use case requirements in terms of functionality and performances. Furthermore, training algorithms targeting low-energy and low memory requirements will be developed to get efficient network models.

More specifically, this document addresses the following aspects:

- On-chip learning algorithms and efficient inference accelerators will be designed and implemented on FPGA platforms. For instance, memory efficient training algorithms for time series will be experimented using the SpiNNaker2 implemented on an FPGA.
- Signal processing modelling, like channel modelling and path loss estimation, are also investigated using a novel NN-based approach. Software layers for event driven SNN with training capabilities are developed together with a hardware/software partitioning and co-design.
- Distributed and hybrid networks will be studied and implemented to get the best trade-off between accuracy, scalability, and power.
- Tailored layers will be designed for spiking neural networks, event-based networks, and input data processing; they will enhance the overall performance of the system.
- Mixed precision data converters to realize MAC operations and crossbars implemented using FeFETs transistors will be designed for mixed-signal (deep) neural networks. They will be eventually fabricated in ASICs to guarantee reduced area and power.
- Analog neural networks will be implemented, as well, to address tinyML applications.
- From the infrastructure perspective, tools developed in ANDANTE will be used to better manage the design space and to optimize the energy efficiency of the neural networks.
- FPGA-based platforms will be used as a support to the development of low-power and possibly real-time neural network architectures together with neural network accelerators.

The content of this report is the starting point for the implementation of the proposed solutions that will happen within WP3 and WP4, and the realization of the demos related to the targeted applications defined in WP1, which will happen in the context of WP5.

Each individual contribution is part of an effort to build a common framework of development for neural networks for edge applications.