In consideration of the determination of the Regione Piemonte – Direzione Coesione sociale No. 746 of May 30, 2019 which approved the following apprenticeship position for the PhD project proposal submitted by the Politecnico di Torino in the framework of a specific regional call for proposals (Apprendistato di Alta Formazione e Ricerca 2016-2018 - Avviso Pubblico per la realizzazione dei percorsi formativi di: Laurea triennale e magistrale, Diploma Accademico di primo e secondo livello, Master di primo e secondo livello Universitario, Dottorato di ricerca e Diploma accademico di formazione alla ricerca, Attività di ricerca approvato con Determinazione 537 del 3/8/2016 e successivo Aggiornamento dell'Avviso pubblico approvato con Determinazione n. 1486 del 17/12/2018):

PhD in Electrical, Electronics and Communications Engineering

Research project "Partitioning and Optimization Technologies for Edge Computing"

Politecnico di Torino – ADDFOR S.p.A.

Supervisors	Prof. Mario Roberto Casu – Politecnico di Torino mario.casu@polito.it Dr. Eng. Enrico Busto – ADDFOR S.p.A. enrico.busto@add-for.com
Contact	https://www.add-for.com/ http://www.det.polito.it/focus/research/research_groups/electronics/vlsilab_group
Context of the research activity	The recent development in the field of Deep Neural Networks (DNN) has made it possible to solve many complex problems with innovative solutions. Due to the wide applicability of DNNs to diverse problems, an extensive and widespread adoption of these technologies is expected in the near future. Many applications, however, will have to function autonomously, without the support of a continuous connection to the electrical and communication

	networks. Consider for example the huge amount of raw data that environment or industrial sensors would need to send to a centralized processing system for neural network inference. It is clear that to avoid the explosion of communication-related costs we need to relocate part of the processing power to the edge, whereas operations like incremental training, testing of new neural networks, storage of high-quality and high-valued data are bound to stay in the central processing system. The challenge is that edge devices are typically small embedded systems, with limited processing power and precision, small memory size, and severe constraints on low power consumption. To deploy DNNs on such devices require optimizing and tailoring the application on the specific embedded system, possibly with ad-hoc accelerators implemented in FPGAs or low-power GPUs. The right blend of competences is required to take on this challenge. Here comes the partnership between ADDFOR, with its core competences on the applications and neural networks, and the VLSI research group at the Politecnico di Torino, with its core competences on design and implementation of embedded systems, Systems-on-Chip, and specialized accelerators in FPGA. Notice that since this is a collaboration with ADDFOR S.pA., the selected PhD student will be hired by the company with a “high apprenticeship” pursuant to art. 45 of Legislative Decree 81/2015.
Objectives	The objective of this project is to study and to optimize algorithms based on DNNs on distributed and heterogeneous hardware architectures. More specifically, the activities will include studying, benchmarking and optimizing inference algorithms on commercial hardware platforms, such as NVIDIA Jetson Nano, Google edge TPU, AXIS ARTPEC 7, Intel® Movidius™ Myriad™ X, HiSilicon Kirin 9xx, Qualcomm Snapdragon 8xx, STMicroelectronics Orlando, without excluding systems based on FPGA technology such as Xilinx Zynq UltraScale +. The study should allow to understand and exploit the low-level structure of the processors (both general purpose processors and those specifically designed to support inference in neural networks), their memory architecture and, if dedicated hardware accelerators are used (e.g. GPU, FPGA), the communication between these and the processors.

	The goal is also to determine the most appropriate partitioning of a DNN software so as to be able to create systems that provide a secure remote update of data and networks and at the same time can transmit to the centralized systems the most significant data to allow a continuous updating of the system with Incremental Training techniques. These activities will contribute to developing specific skills for the development of security systems, energy systems and industrial production control.

Skills and competencies for the development of the activity	The candidate is required to have skills and competences such as those typically obtained with a Master of Science degree in areas like Electronic Engineering, Computer Engineering or Telecommunication Engineering. In addition, the candidate should possess the following competences: - C++ / Python programming for scientific computing; - Knowledge of the main Deep Learning Libraries (Tensorflow, Caffe, Pytorch); - Experience with GPU and FPGA hardware accelerators. Notice that the candidate must also be less than 30 years old at the moment of the hiring from the company.