Politecnico di Milano

PhD in
Information Technology

Research Area n. 1 - Computer Science and Engineering

Research Title: TINY-ML TOOLCHAIN FOR ULTRA-CONSTRAINED PROCESSORS

<table>
<thead>
<tr>
<th>Scholarships and Financial support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monthly net income of PhD scholarship (max 36 months)</td>
</tr>
<tr>
<td>Increase in the scholarship for stays abroad</td>
</tr>
<tr>
<td>Number of scholarships</td>
</tr>
<tr>
<td>Beginning of PhD</td>
</tr>
<tr>
<td>Deadline for application</td>
</tr>
</tbody>
</table>

Context of the research activity

Motivations and objectives of the research in this field: Advancements in ultra-low power processors and sensors as well as in artificial intelligence are paving the way towards increasingly autonomous intelligent Cyber-Physical Systems. A CPS environment at the deep edge is equipped with resource-constrained devices. In turn it embodies a sensor node which executes CPS edge applications exploiting sensing, computing, communication and actuation. Other than classic CPS deployments, where a centralized CPS cyber platform was the central point of data collection and processing, intelligent distributed CPS environments aim to rely (mostly) on computation capabilities implemented very close to sensors (or within same package) and actuators in order to decentralize intelligence and minimize cloud...
connectivity/latency issues and privacy concerns [1-3].

The Intelligent CPS advancements will allow the implementation of break-through use cases at unprecedented ultra-low power consumption. For example in manufacturing, a plant can be highly automated and shared by multiple tenants who utilize machinery from trusted third-party vendors. In this context, devices implementing tiny machine-learning applications can be used to monitor the manufacturing process and dealing with unknown behaviors or unknown defective products not foreseen at product qualification time.


Methods and techniques that will be developed and used to carry out the research

The objective of this PhD is to develop an innovative AI mapping toolchain capable of:

a) Architecting artificial neural networks at very low bit depth (i.e., down to 1 bit activation/weight data width vs. 32 bit floating point data width)

b) training resulting models on selected datasets representing use cases

c) optimizing their topologies with multi target objectives (e.g. memory, energy, computation).

The aim for implementing the software in a hardware is the ultra-low power DSP (code name STred – reduced energy DSP) processor unit designed by ST featuring very constrained memory and computational resources.

Potential commercial and industrial applications of this PhD is in the development of Distributed, Intelligent, Self-aware IoT Applications. For instance, this processor-software solution has applications in audio (e.g. ultrasound event detection) and vibration processing (e.g. anomaly detection) for predictive maintenance and within a power envelop comparable as order of magnitude to the energy consumption of the sensor alone (i.e. µW). Furthermore, interoperability with existing ST AI tools for automatic mapping and deployment of pre trained neural networks shall be in the scope of the activity.

Distributed AI is fast growing field requiring fast innovation to surf its wave. This can be enabled by offering neural network inference up to ultra-constrained devices within a coherent, productive toolchain and an end-to-end eco-system. Indeed, machine learning inference on the edge is an increasingly attractive prospect due to its potential for increasing energy efficiency, privacy, responsiveness, and autonomy of edge devices.

In this PhD is foreseen that development of the innovative
AI toolchain will support our development, optimization and mapping of Tiny Machine Learning applications on resource constraints processors on the edge. In particular the tools should: a) Architect artificial neural networks at very low bit depth (i.e., down to 1 bit activation/weight data width); b) training resulting models on selected datasets representing use cases, optimized for the deployment scenarios; c) optimizing their topologies with multi target objectives (e.g. memory, energy, computation). In all cases these tools and techniques should be designed to be interoperable and demonstrated with existing AI toolchain (e.g. ST and off the shelf).

**Educational objectives**

The main novelty of this PhD consists of application of **deep-edge** neural networks in intelligent CPS systems supported by **distributed AI** on CPS ultra-constrained processors in a sensor node. Such processors are **resource constrained** (typically, KB of memory, few MHz clock frequencies, µW power budget, dedicated binary/low bit depth instructions). The candidate will develop a strong background in the field.

The expected academic outcomes are:
- Techniques and tools for supporting the development, optimization and mapping of TinyML applications on resource-constrained-processors
- Knowledge and the answer to the key research questions of the PhD topic will be disseminated through several MSc level courses at POLIMI, by preparing ad Hoc seminars but also by supervising MSc students’ thesis or course projects.
- Publication of scientific papers on international journals and conferences related to architecture, compiler and exemplary ANN(s) are envisioned to ensure a significant scientific impact of the research.

Quantitative expected outcomes are at least 3 top-level conferences and 2 journal submissions.

The position is also included in the EIT Doctoral school program, a business development experience at the industrial partner premises is expected, together with innovation related training.

**Job opportunities**

The PhD candidate will address fundamental problems with a broad applicability in the field of embedded system design, edge computing and Machine learning. A PhD graduate with such a background can be very valuable in STMicroelectronics as well as many other large companies / SME. Post-Doc research opportunities are also available in academia.

**Composition of the research group**

The student will be part of the group of prof. Cristina Silvano at DEIB of Politecnico di Milano

http://home.deib.polimi.it/silvano/

Number of Full Professors 1
Number of Associate Professors 1
Number of Assistant Professors 0
Number of Post-Docs 2
Number of PhD students 2
Number of contracted researchers 0
| Names of the research directors | Prof. Cristina Silvano, Politecnico di Milano  
Dr. Danilo Pau, STMicroelectronics |
| E-mail address, phone number and web-page | cristina.silvano@polimi.it  
02 2399 3692  
http://home.deib.polimi.it/silvano/ |
| List of Universities, Companies, Agencies and/or National or International Institutions that are cooperating in the research | 1. Politecnico di Milano  
2. ST Microelectronics  
3. EIT Digital |

### Additional support

| Educational activities  (purchase of study books and material, funding for participation in courses, summer schools, workshops and conferences): financial aid per PhD student per year | 2\text{nd} year: max 1534,00 euro per student  
3\text{rd} year: max 1534,00 euro per student |
| Teaching assistanship: availability of funding in recognition of supporting teaching activities by the PhD student | There are various forms of financial aid for activities of support to the teaching practice. The PhD student is encouraged to take part in these activities, within the limits allowed by the regulations. |
| Computer availability: | 1\text{st} year: \textit{individual use}  
2\text{nd} year: \textit{individual use}  
3\text{rd} year: \textit{individual use} |
| Desk availability: | 1\text{st} year: \textit{individual use}  
2\text{nd} year: \textit{individual use}  
3\text{rd} year: \textit{individual use} |

### Additional information

The PhD will be co-supervised by the Danilo Pau IEEE fellow and Technical Director, System Research and Applications of STMicroelectronics. Part of the research activities will be done at the research Labs in STMicroelectronics sites Agrate and Castelletto (Milano), where the candidate will spend part of its time.