Call for applications – PhD position

Lyon Institute of Nanotechnology – INL
Ecole Centrale de Lyon
36 av. Guy de Collongue
F-69134 Ecully, FRANCE
http://inl.cnrs.fr/

Institut de Recherche en Informatique et Systèmes Aléatoires – IRISA
University of Rennes – ENSSAT
6 rue de Kerampont, BP 80518
F-22305 Lannion, FRANCE
http://www.irisa.fr

Optical Interconnect for efficient and flexible AI hardware accelerators

The increasing need to distribute AI applications from the cloud to edge devices is becoming a pressing concern for addressing data privacy, bandwidth limitations, power consumption reduction, and low latency requirements, especially for real-time, mission- and safety-critical applications. Consequently, there is an ongoing effort to design custom and embedded AI hardware architectures (AI-HW) that can support energy-intensive data movement, speed of computation, and large memory resources that AI requires to achieve their full potential. However, current AI-HW architectures are mainly based on GPU, TPU, or specialized designs, which are devoted to improving the energy/performance efficiency for a specific class of AI applications, such as Convolutional Neural Networks. Thus, they are not designed to provide the high flexibility and massive parallelism needed to support a wide range of AI algorithms, including dynamic networks, Recurrent NNs, Transformers, etc.

To address these limitations, the **AdaptING** project, part of the large-scale **PEPR IA** nationwide research initiative, proposes a new architectural paradigm called adaptive architecture, which aims to make HW adaptable to any given AI application and its constraints in terms of accuracy, energy, latency, and reliability. The adaptive architecture is designed to provide flexibility, efficiency, sustainability, and reliability for embedded AI. To achieve such objectives, the target architecture will be composed of **heterogenous components** (e.g., different HW accelerators for training and inference). Each component will be characterized by a given level of energy efficiency, latency, precision and trustworthiness.

The main goal of this PhD is to investigate how components must be interconnected in terms of technology (e.g. electronics, photonics), topology (e.g. NoC) and communication protocol to implement the adaptive AI hardware accelerator for a given use case. Indeed, one of the main bottlenecks in terms of energy and latency is the **need for transferring data between components** and **shared memories** (in the memory hierarchy). Scaling communication protocols to ensure coherence is a recurrent issue, since it requires coherence messages to be broadcast to all the shared memories, or to multicast these messages to an identified subset of the memories. Providing broadcast capabilities with negligible latency to enable fast chip-scale information distribution is a means to solve these issues. Networks-on-Chip (NoCs) have emerged as the communication backbone for such systems but existing NoCs with planar metal interconnects, are limited by high latency and power consumption of multi-hop wired links for long distance communication across the chip. While 3D integration technology allows the distance between computation nodes and memories to be reduced to improve latency, power consumption and bandwidth at the scale of a group of components (cluster), **it cannot be extended to inter-cluster communications**. The silicon photonics technology appears a suitable candidate to enable long links and will be investigated to evaluate the opportunity for efficient interconnects and for AI data traffic. Related to these opportunities, specific routing protocols will be proposed to reduce interconnect overhead on application execution time. In particular, broadcast transfers will be addressed as they represent a large part of traffic between ANN layers, or between memory and processing elements (for neuron weight transfers for example).

Background: The candidate should hold or be about to obtain an MSc in Electronics Engineering and Computer Science, and should have a very good background in computer architecture, as well as skills in networks-on-chip.

Application deadline: April 30th 2024

Starting dates: October 2024.

Environment: The PhD will be carried out in collaboration between IRISA and INL laboratories.

Send CV and statement of purpose (in English or French) to

- Ian O'Connor / INL - Lyon Institute of Nanotechnology - Ecole Centrale Lyon – email: ian.oconnor@ec-lyon.fr
- Alberto Bosio / INL - Lyon Institute of Nanotechnology - Ecole Centrale Lyon – email: alberto.bosio@ec-lyon.fr
- Daniel Chillet / IRISA – University of Rennes – ENSSAT – email: daniel.chillet@irisa.fr