Scalable training on scalable infrastructures for programmable hardware

Dr. Marco Lorusso [1,2]

Prof. Daniele Bonacorsi [1,2] Dr. Riccardo Travaglini [2]
Prof. Davide Salomoni [2,4] Dr. Paolo Veronesi [2]
Dr. Mirko Mariotti [2,3] Dr. Giulio Bianchini [2,3]
Dr. Alessandro Costantini [4] Dr. Doina Cristina Duma [4]

[1]University of Bologna - Department of Physics and Astronomy

[2]INFN - National Institute for Nuclear Physics

[3]University of Perugia - Department of Physics and Geology

[4]INFN - CNAF Bologna

9th May 2022

Learning Machine Learning

- **Machine Learning** algorithms are **everywhere**;
  - This is reflected in the **number** of **workshops** and schools that are held every year;
    - More than **200** summer **schools** were organized in the last **5 years** according to a community redacted list [1];
  - Also popularization of **Massive Open Online Courses** (MOOC) on Artificial Intelligence, e.g:
    - **Coursera** with teachers from top universities ($\sim$ 1600 different courses on ML);
    - **ECMWF** course: *Machine Learning in Weather and Climate*

[1] https://github.com/sshkhr/awesome-mlss

And embedded and programmable logic?

- Increased **popularity** of IoT devices and **programmable SoC** (FPGA) in general;
- **However** lack of schools and workshops on fusing ML and FPGAs;
- Very **different** set of skills:

- Very high level programming
- Knowledge of libraries like TensorFlow and Pytorch
- Data Science

- Background of electronics
- Hardware Descriptive Languages (HDL)
- Lower level software programming

Even thought the field is gaining traction, **difficulties** in finding people who knows enough of **both** worlds to make progress.

Workshop on ML on FPGA in Bologna

- *Machine learning techniques with FPGA devices for particle physics experiments*, organized by INFN Bologna with CNAF technical support and funded by the INFN Training Program (2-4 November 2022);
- **First step** towards a greater focus on education in this field, first of a kind in Italy;
- **Leading lecturers** involved in the development of tools to make hardware more approachable at a higher level;
- Support from the **AMD/Xilinx University Program** (XUP);
- 20 in person participants.

Corso Nazionale di formazione
"Tecniche Di Machine Learning Con
Dispositivi FPGA
per Gli Esperimenti Di Fisica Delle Particelle"
INFN

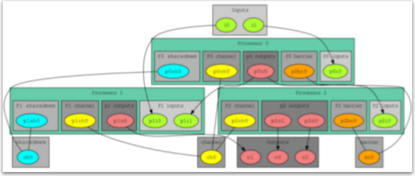
XILINX
| UNIVERSITY PROGRAM

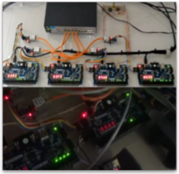
Topics of the workshop

- Introduction to efficient use of Machine Learning in HEP;
- Crash course on what FPGAs are;
- *HLS4ML* and how to translate Python to something implementable in hardware (see next slides);
- *Vitis-AI*, the AMD/Xilinx solution to Artifical Intelligence on programmable hardware;
- A new kind of computer architecture (multi-core and heterogeneous) which dynamically adapt to the specific computational problem rather than be static: the *BondMachine* (see next slides);
- How Quartus and Intel make ML on FPGA possible;
- (More than) half of the duration of the course spent on tutorials;

The BondMachine Toolkit

The BondMachine is an open source software ecosystem for the dynamical generation of computer architectures that can be synthesized on FPGA.



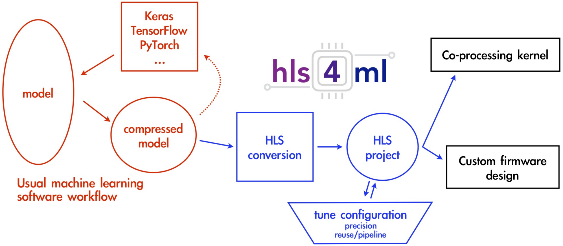


- High level programming language (Golang) for both the hardware and software
- Functional style programming
- Computational graph and Neural Networks
- Architecture generating compiler
- Fast machine learning inference with FPGA
- Development of accelerated systems on hybrid processors (ARM + FPGA)

https://github.com/BondMachineHQ

The hls4ml package

https://fastmachinelearning.org/hls4ml



- Developed by members of the HEP community to translate **ML algorithms** written in **Python** into **High Level Synthesis** code;
- HLS allows the **generation** of **hardware descriptive code** (HDL) from *behavioral descriptions* contained in C++ program;
- The translated Python objects can be injected in the automatic workflow of proprietary software like Vivado from Xilinx Inc.

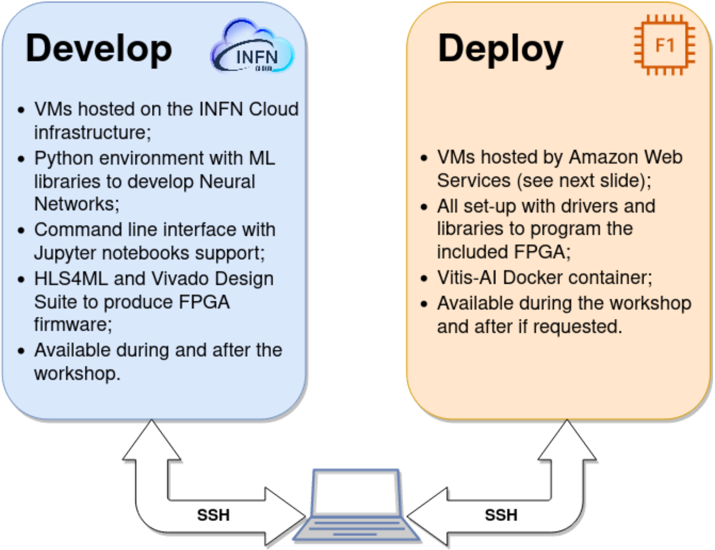
Offer a testing ground to the students

- Workshop as a chance to **start working** with this technology and new workflows;
- The need for **specific software** and libraries to **develop** both ML algorithms and FPGA firmware raise the need for a **dedicated development machine** available to the attendees;
- Wanted to work with **actual hardware** to test first-hand the firmware, BUT:
  - FPGAs are generally **not accessible** to more people at once for programming;

  ⇒ A board for each person

    ⇒ **Too expensive** and generally not feasible;
- Solution: **FPGAs in the Cloud!**

Cloud classrooms



Develop

- VMs hosted on the INFN Cloud infrastructure;
- Python environment with ML libraries to develop Neural Networks;
- Command line interface with Jupyter notebooks support;
- HLS4ML and Vivado Design Suite to produce FPGA firmware;
- Available during and after the workshop.

Deploy

- VMs hosted by Amazon Web Services (see next slide);
- All set-up with drivers and libraries to program the included FPGA;
- Vitis-AI Docker container;
- Available during the workshop and after if requested.

AWS F1 Instance

Cloud computing offers a **Pay-per-use** place to test FPGA firmware with relatively low cost.

- A specific kind of machines in the **AWS Cloud Computing** catalogue includes FPGAs;
- EC2 F1 instances use **FPGAs** to enable **delivery** of **custom hardware accelerations**;
- Packaged with **tools** to **develop**, simulate, debug, and **compile** a design.

Integrating local and cloud resources

- Workshop itself as a **test** for starting to understand if a **seamless integration** between **INFN Cloud** and a Cloud provider like **AWS** would be useful;
- A sketch of how it would work is the following:
  1. I **authenticate** myself on INFN Cloud with a kind of federated authentication;
  2. I select the type of **resource**(s) I want, e.g. a Xilinx U250 or U55C or an Intel Terasic;
  3. It may be that the desired FPGA resource is not available on INFN Cloud, but it is available on AWS, for example;
  4. In this case, the resource would be **instantiated on AWS transparently** and the user would be given the endpoint to connect to, without needing a different authentication or interface.

What we have learned - Next steps

- The workshop sparkled a **great interest** which pushed us organizers to **increase** the maximum number of participants ⇒ thanks to the Cloud solutions it was possible with **minimal efforts**;
- Nevertheless, it is still a **first attempt** for such an event, more work to be done e.g. time for **even more tutorial** and possibility to "come prepared" with **access** to the machines even **before** the workshop;
- New and more efficient **teaching techniques** could be tested, like *inverted learning*;
- Creation of a **VM template** with all tools for this kind of development, and publication of an **AMI** for deployment;
  - ⇒ Useful for both education purposes and research work;