International Symposium on Grids & Clouds 2019 (ISGC 2019)

Contribution ID: **33** Type: **Oral Presentation**

The BondMachine toolkit: Enabling Machine Learning on FPGA

Friday, 5 April 2019 09:00 (30 minutes)

Future systems will be characterized by the presence of **many computing cores** in a single device, by **heterogeneous architectures** built to optimize power and “silicon” consumption as much as possible and by **re-configurable hardware** technologies. These concepts have been demonstrated, both in software programming and hardware evolution, by the multi-core, GPGPU, OpenCL and re-programmable logic devices which populate the spectrum from small devices up to large-scale data centers. A key to the success in the era of hybrid computing will be how coherently HW/SW systems will take all these components into account.

The BondMachine (BM) is an innovative prototype software ecosystem aimed at creating facilities where both hardware and software are co-designed, guaranteeing a **full exploitation of fabric capabilities** (both in terms of concurrency and heterogeneity) with the smallest possible power dissipation. The disruptive innovation of the BM is to provide **a new kind of computer architecture**, where hardware **dynamically** adapts to the specific computational problem, rather than being static and generic, as in standard CPUs synthesized in silicon.

In order to exploit the dynamic nature of the BM, its main goal is to create the described heterogeneous and flexible architectures on top of re-configurable technology devices (such as FPGAs). Moreover, the overall BM vision is based on the reduction of the number of hardware/software layers, which as a byproduct guarantees a simpler software development.This is precisely why the BM project has been thought as a complete re-configurable computing ecosystem, that starting from a high-level description creates both the hardware and the software that runs on it.

The BM uses Go as main language for the codesign. Its concurrency primitives are perfect to be mapped in the BM architecture and to allow writing concurrent applications on FPGA with a small overhead compared to the HDL code.
The flexibility of the BM makes it possible the implementation of any computing system, ranging from networks of small agents, like IoT (Internet of Things), to high performance devices for ML (Machine Learning) or real time data analysis, and even systems that mix all these different characteristics together.

In the context of the HEP domain, we are developing new BM components to **deploy complex AI systems on hardware**, providing a high-level mechanism to translate into silicon Deep Learning networks, created via standard Tensorflow and Keras toolkits.
For what regards deployment models, the BM provides several solutions, such as a standalone FPGA, accelerators coupled to workstations, as well as a BM as a Service running on hybrid Clouds.

In this talk we will provide a technical overview of the key aspects of the BondMachine toolkit, highlighting the advancements brought about by the porting of Go code in hardware. We will then show a cloud-based BM as a Service deployment. Finally, we will focus on Tensor Flow, and in this context we will show how we plan to benchmark the system with a ML tracking reconstruction from pp collision at the LHC.

Primary authors: Prof. BONACORSI, Daniele (University of Bologna); Dr SALOMONI, Davide (INFN-CNAF); Dr STORCHI, Loriano (Dipartimento di Farmacia, Universitá degli Studi G. D’Annunzio, Chieti); Mr MARIOTTI, Mirko (Department of Physics and Geology, University of Perugia); Dr SPIGA, daniele (INFN-PG); Dr BOCCALI, tommaso (INFN)

Presenter: Mr MARIOTTI, Mirko (Department of Physics and Geology, University of Perugia)

Session Classification: Supercomputing, High Throughput Computing, Accelerator Technologies, and their Integration

Track Classification: Supercomputing, High Throughput Computing, Accelerator Technologies, and their Integration