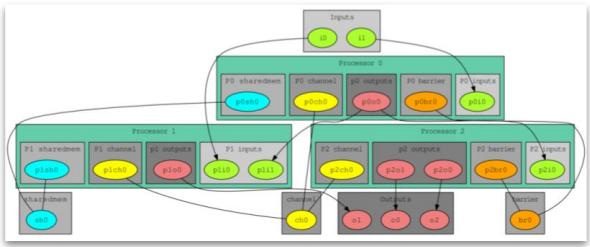
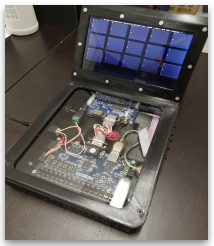
BondMachine

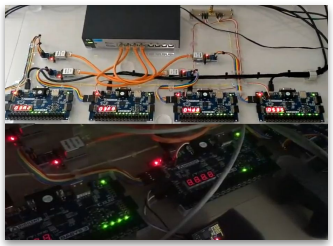
BondMachine: A framework to build dynamical computer architectures

The BondMachine is an **open source** (https://github.com/BondMachineHQ) 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







History and Major Highlight

- CCR
  - 2015 First ideas
  - 2016 Poster
  - 2017 Talk
  - 2022 Talk
  - 2023 Talk
- **InnovateFPGA 2018 Iron Award, Grand Final at Intel Campus (CA) USA**
- Invited lectures at FPGA workshops ICTP 2019 and 2022
- Golab 2018 talk and ISGC 2019 PoS
- **Article published on Parallel Computing, Elsevier 2022 DOI:10.22323/1.351.0020**

Firmware development

Customized to specific use case starting from a high-level language

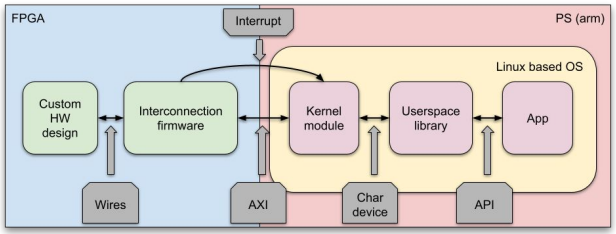
Main activities

- Fast machine learning inference with FPGA

  ? From a machine learning model trained with standard frameworks, synthesized in FPGA as a graph of heterogeneous and interconnected processors.

  Features
  - Optimized resource usage (LUTs, REGs, DSP ...)
  - Highly customizable
  - Available at a high level (Jupyter Notebooks, PYNQ)
  - Vendor independent (Xilinx-Amd, Altera-Intel)

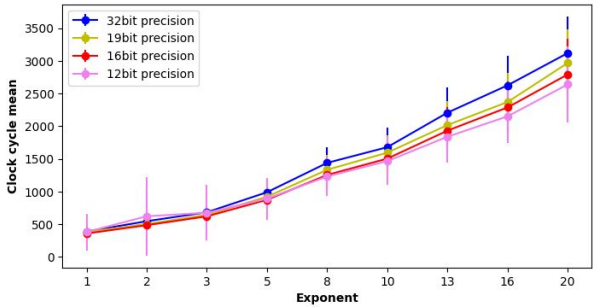
- Development of accelerated systems on hybrid processors (ARM & FPGA)

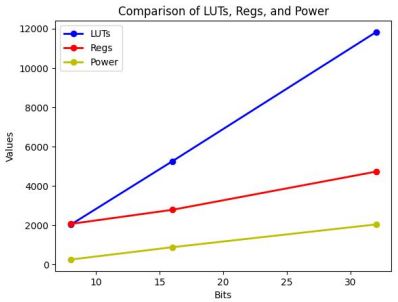


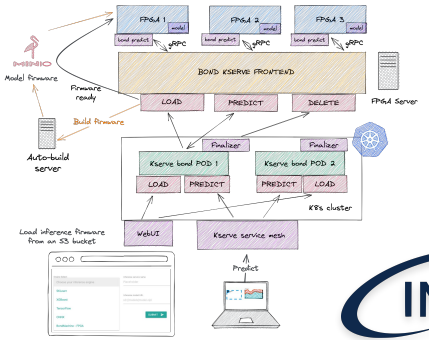
Exploiting FPGA clusters with these approach for benchmarking and supporting real scenarios

What are we doing

- **Model's compression**
  - Tests with different numerical precisions
    - We have integrated both static and dynamic numerical types to find the right tradeoff between resource optimization and accuracy degradation.
    - We are working on Quantization inside the BondMachine framework
  - Reduction of model complexity
    - using APIs provided by ML frameworks
    - using the high customization of BondMachine (pruning and collapsing of connecting processors)
- **Real energy measurements**
  - We want to analytically identify real energy consumption (we currently have estimates of energy consumption extrapolated from development software i.e. Vivado)
- **Bring the solution to cloud level**
  - We have integrated our system with cloud-native inference as-a-service solution
    - Implementing a KServe FPGA extension
    - We have validated an end-to-end workflow with a generic ML algorithm







What we want to do (and what we need)

- What we want to do
  - Continue the development of the core part of the BondMachine, optimizing the architecture both in computational terms and resource usage
  - Continue to develop techniques for compressing models
  - Multi FPGA vendor-independent accelerators (not only ML/DL, also accelerators for generic algorithm)
  - Better support for Intel FPGAs
  - Build automations to further abstract complexity
  - Improve cloud implementation in order to provide ML inference as a service system
- What we need
  - Intel FPGA
  - Cluster of Xilinx and Intel FPGA to run our multi FPGA system
  - Man power to develop the low level part (HDL programming)
  - Man power to develop the high level part (Python/Go programming)

ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing