Report on WP4 activities

S. Gennai[1] & A. Pompili[2]

1 INFN Milano Bicocca

2 University of Bari

Main activities in 2024

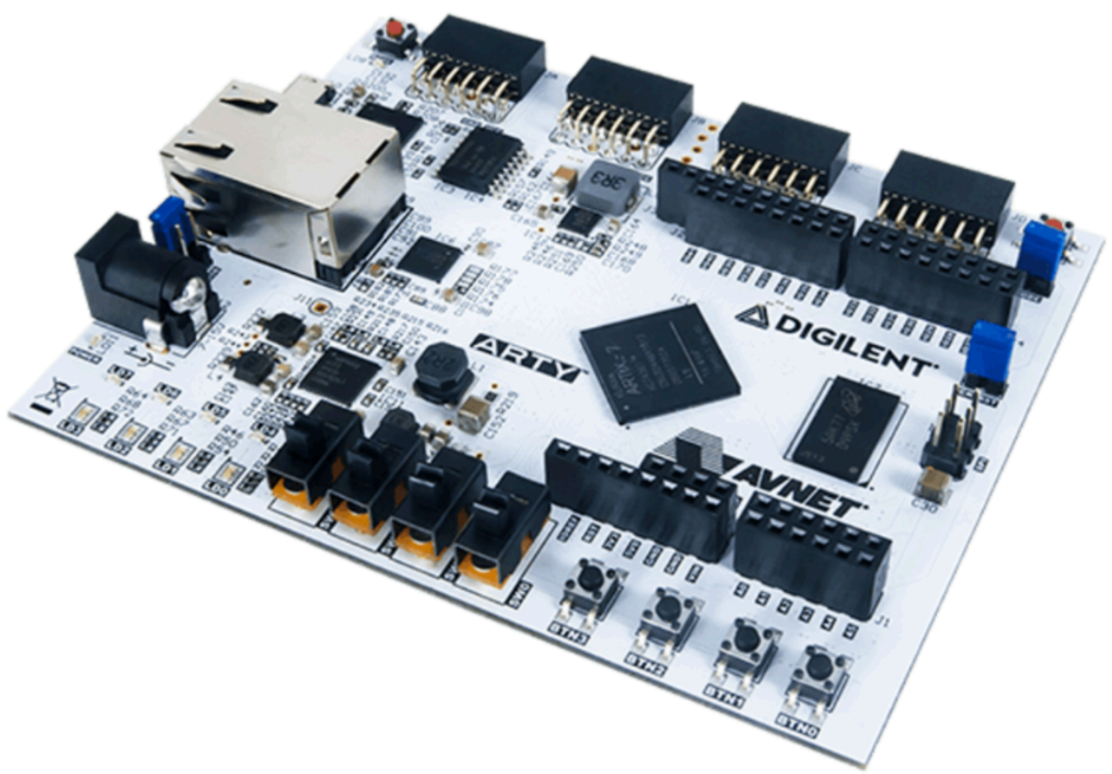
- Participated actively to two flagships for WP2:
  - FPGA
  - GPU
- Organized one FPGA course on VHDL
  - https://agenda.infn.it/event/39721/
- Organized a school of ML on GPU with Milano Bicocca and Bari university
  - CPU+GPU “offered” by CINECA
- Tried to organize an advance GPU course and a second FPGA one …
  - They will be made in 2025

FPGA course (4-6 of March 2024)

- Lecturers:
  - Andrea Triossi (University of Padova)
  - With the help of Mirko Mariotti
- Facilitators for the hands-on sessions:
  - Giulio Bianchini (University of Perugia)
- Course unit contents:
  - Introduction to FPGAs
  - FPGA Architecture
  - FPGA programming flow
  - VHDL language by examples
  - Introduction to the Vivado programming framework and the Arty A7 board
  - Combinational circuits on FPGA
  - Sequential circuits on FPGA
  - Arithmetic operations on FPGA
- Thanks to Mirko Mariotti for setting up remote FPGAs and efficient use of Vivado from Jupiter notebooks
  - CPU provided by Cineca

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

Introduction to FPGA programming



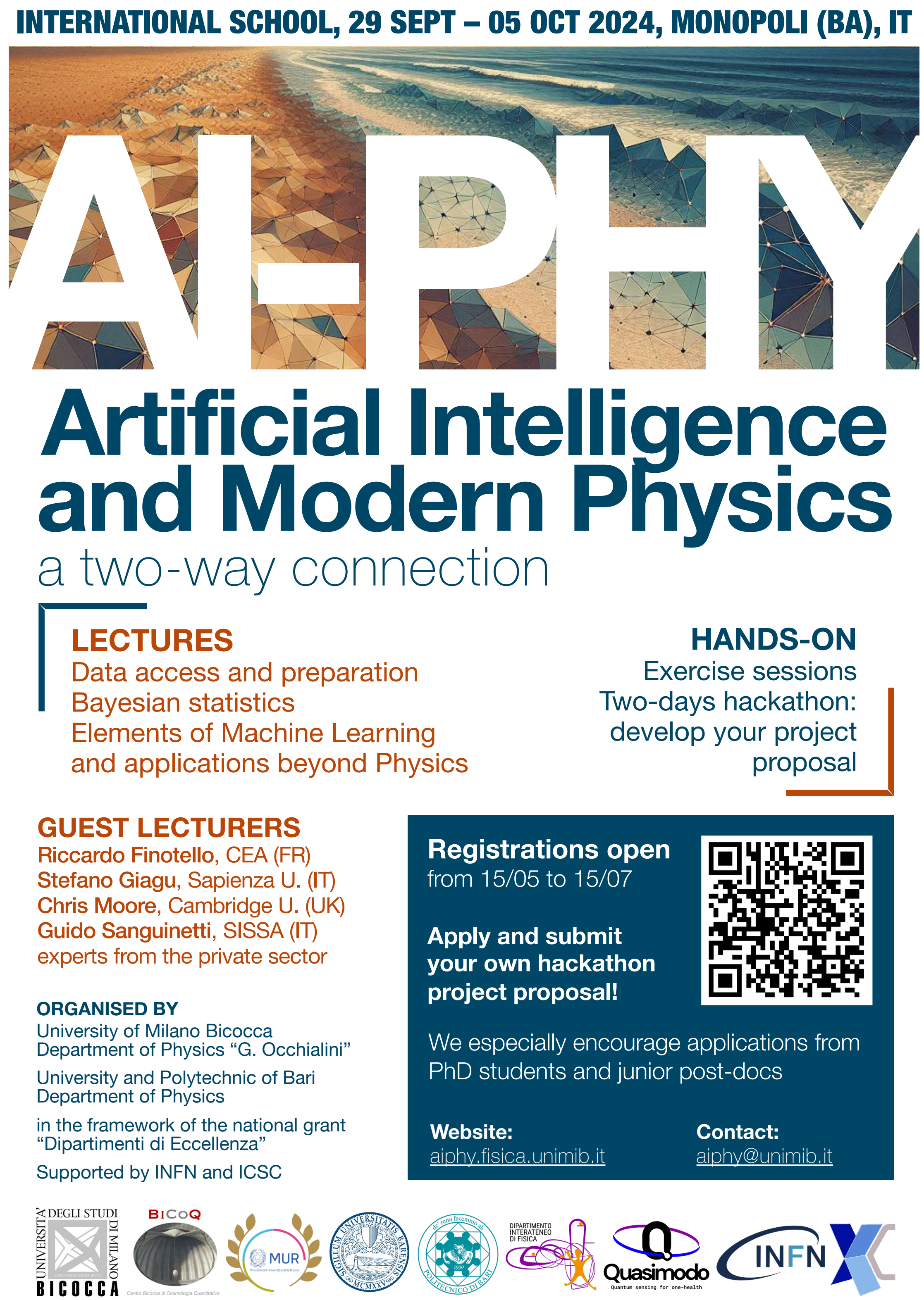
ML on GPU course

- Lecturers:

Andrea Beschi | McKinsey
Riccardo Finotello | CEA Paris-Saclay
Stefano Giagu | Sapienza University of Rome
Chris Moore | University of Cambridge
Guido Sanguinetti | International School for Advanced Studies (SISSA)

GPU provided by Cineca using SLURM access to backbone GPU with Jupyter notebooks

Some hiccups in the authentication procedures
But fast feedback form them



UC2.2.3 Development of ultra-fast algorithms running on FPGA

- Use cases, not all of them participate to the KPIs:
  - Trigger, DAQ and on-line processing
    - Development of algorithms based on neural networks and implementation on FPGAs, with application for trigger and anomaly detection at event level and object level for the Atlas experiment.
    - Development of a track reconstruction algorithm, at 30 MHz, on FPGA for LHC-b data acquisition.
      - Waiting for the Terabit FPGA clusters (ETA: end of 2024)
    - Development of digital trigger logic for a “missing energy” experiment with a positron beam at CERN (POKER/NA64)
      - Superseded by the OpenCall: AI-supported real-time data reduction algorithms for streaming readout systems.
    - Development of quantum-inspired Tree Tensor Networks for classification in Trigger on FPGA
    - Di-tau trigger development for the CMS Level-1 trigger system
      - PhD project, will be finalized at the end of the PhD term …
    - Scouting and processing of Level-1 trigger data using FPGA to run on-the-fly momentum object calibration with ML based algorithms
  - Developing FPGA tools
    - Development of a Customizable Framework for Multi-FPGA Accelerator Generation via architectures
    - Development and testing of RDMA over converged ethernet (ROCE) on FPGA for data transfer from detectors’ front-end to computing servers

UC2.2.3 Development of ultra-fast algorithms running on FPGA

- Status of KPI at the moment, not all the use cases participate to them

KPI ID	Description	Acceptance threshold	Status up to today
KPI2.2.3.1	Development of triggering algorithms, on-line analyses, data acquisition on FPGA	Submission of 1 paper to a peer-reviewed journal	1 paper already accepted
KPI2.2.3.2	Online scouting	Submission of 1 paper to a peer-reviewed journal	Abstract being submitted to ichep 2024 about scouting
KPI2.2.3.3	Development of tools to integrate several FPGAs together	Submission of 1 paper to a peer-reviewed journal	G. Bortolato et al 2024 JINST 19 C03038
KPI2.2.3.4	Organizing courses about FPGA programming on low and high level	At least two courses organized	1 course done at the end of 2023 1 VHDL course done in February 2024

UC2.2.4 Porting code to GPU platforms

- Algorithm porting activities
  - CMS pixel reconstruction on GPU in Alpaka (KPI2.2.4.2)
  - CMS strip reconstruction on GPU in Alpaka (KPI2.2.4.1 and KPI2.2.4.2)
    - CA-based Tracking extended from pixel to microstrip (CA=Cellular Automaton)
  - CMS Electron seeding on GPU (KPI2.2.4.1 and KPI2.2.4.2)
  - Multi-Objective (Particle Swarm Optimizer) Optimization tool for CMS GPU Algorithms Optimization (KPI 2.2.4.X)
    - for parameter tuning of the GPU pixel track reconstruction (efficiency vs fake rate)
  - Validation infrastructure
    - validation of the usage of Alpaka by CMS from 2024 datataking (KPI 2.2.4.3)
    - pixel tracks performance as metrics
    - tested also on pledged resources with nVidia GPUs but testing also on AMD GPU resources at CNAF
- Data Compression
  - ALICE data compression for asynchronous reconstruction (KPI2.2.4.2)
    - TPC data compression via entropy encoding (Asymmetric Numeral Systems encoders family)
    - TPC track-model decoding implemented on CPU and being offloaded to GPU
    - preliminary performance studies on AMD GPUs

UC2.2.4 Porting code to GPU platforms

- Status of KPIs

KPI ID	Description	Work Ongoing/Done (Threshold)
KPI2.2.4.1	At least XX offline algorithm ported to GPU (most probably an LHC algorithm)	2 (1)
KPI2.2.4.2	At least YY online algorithm ported to GPU (most probably an LHC algorithm)	2 (1)
KPI2.2.4.3	Preparation of a test infrastructure able to test codes on heterogeneous systems. At least ZZ architectures to be supported (eventually, AMD, nVIDIA, CPU)	3 (3), available resources for AMD, NVIDIA and CPUs
KPI2.2.4.4	Organize at least KK events to introduce students and collaborators to heterogeneous computing and train them to the usage of portability tools (joint with WP4).	1 (3)

Encountered Problems

- It is difficult to organize advanced courses
  - Most of lecturers have already prepared introductory courses or specialized slides on what they are doing
- If we want to organize something "different" from what offered we need to have the possibility to pay the lecturer for the extra work.
  - This is not so easy as access to the CN funding has revealed not to be so easy …

FPGA programming school

- A first attempt to build a kind of legacy for the WP4 activities
  - Using CN funding
- Originally thought to be structured in 3 years (or more)
  - Something similar to the TDAQ school
- We found quite a lot of problems to access fundings, though ….
  - So the project has been drastically resized

Foreseen advanced FPGA course

- Use of simulatorsi (GHDL/Gtkwave/Modelsim)
- Design flow Vivado + Vitis
- Scripting language for FPGA
- Advanced design flow
- Timing analysis
- Compilazione parziale
- Debug through ILA/VIO
- creating new IP with block design
- Tools at work: FPGA interconnection
  - Aurora transceivers
  - loopback mode
- Fine tuning (timing, analisi delle risorse)
- Upload bitstream

By P. Vinci et al.

Foreseen advanced GPU course

- Profiling tools
- More on portability tools
- Optimization for multi GPU nodes vs 1 node <-> 1 GPU
- Use of a GPU farm for ML training
  - Joint with WP5
- Hyper-parameter tuning while optimizing analysis FOMs

Conclusions

- Scientific developments ongoing and satisfactory (within a wide variety of use cases)
  - some activities ongoing beyond the scope of the KPIs;
  - some Innovation Grants and Open calls adding activities/use cases
- intensive collaboration with WP2 but less participation/ interaction from other WPs
  - good synergy with WP5
- We are committed to find funding to pay lecturers and tutors of the foreseen courses
  - especially now that we aim to deal with advanced contents