CryoAI – Prototyping cryogenic chips for machine learning at 22nm

Manuel Blanco **Valentin**

Fast Machine Learning for Science Workshop 2022

03-06 October 2022

On behalf of:

Fermilab - Chinar Syal, Farah Fahim, Giuseppe Di Guglielmo, Nhan Tran, Jules Muhizi

Northwestern U. - Seda Memik

Columbia U. - Joseph Zuckerman, Luca Carloni, Maico Cassel

1. CryoAI Overview & Motivation

- Prototype **SoC** for **machine learning applications**
  - **Autoencoder**: Early **anomaly detection**
  - Use case: Industrial failure prevention (audio)
  - **Quantized** training
- **Cryogenic** environment
- **New system-level design flow** based on:
  - HLS4ML (Siemens Catapult backend)
  - ESP
- Based on a **modular tile-base architecture**
  - LP **32bit RISC-V** microcontroller (Ibex)
  - **256kB SRAM** scratchpad
  - 18K-parameter **AI accelerator**
  - Auxiliary circuitry for IO handling
  - Tiles **interconnected** using **NoC**
- **GF 22nm process (22 FD SOI)**

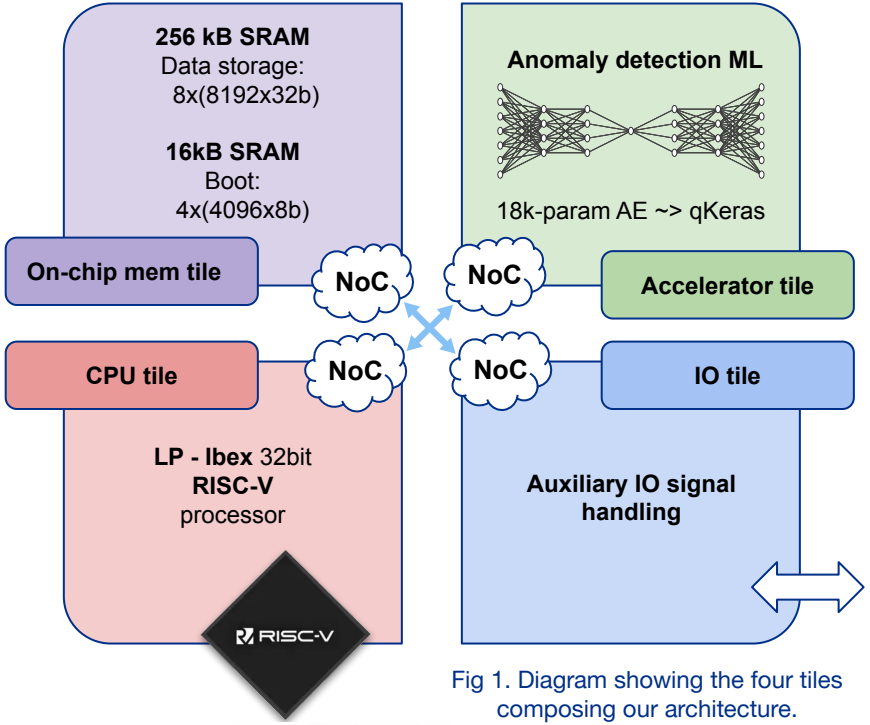
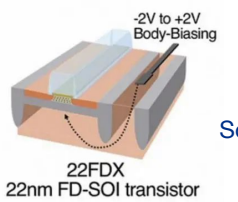


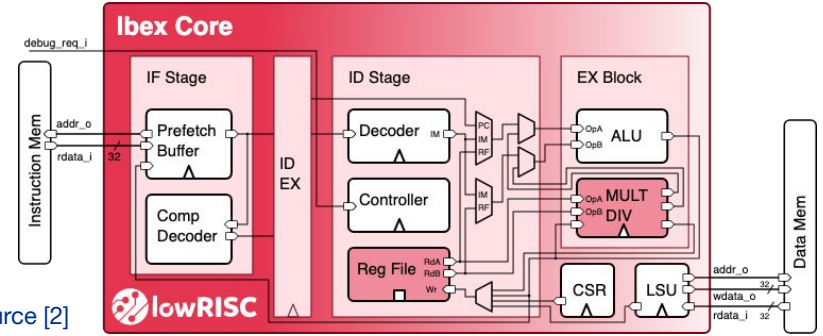
Fig 1. Diagram showing the four tiles composing our architecture.



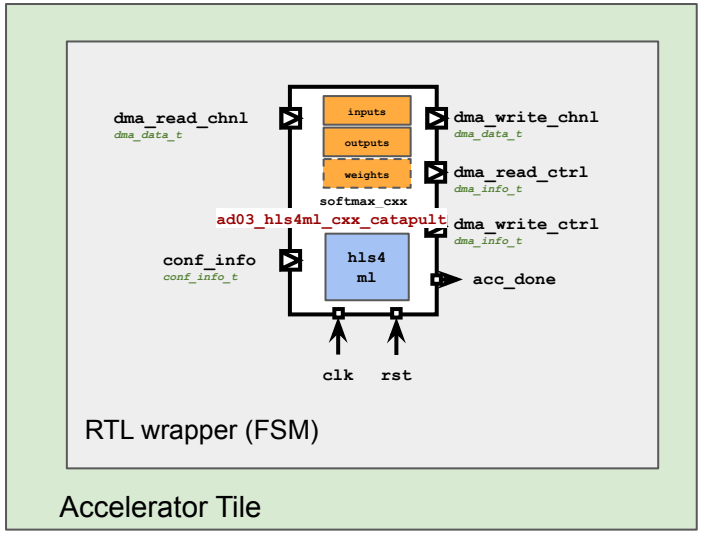
Source [1]

2. Modular infrastructure close-up

- **CPU tile**
  - **RISCV Ibex 32b** µcontroller [2]
    - LOW-RISC (formerly, µ-riscy)
    - RV32IMC instruction set
    - Highly **parametrizable**
    - Designed for minimal area
    - Very low power
    - Bare-metal
  - Program the accelerator
  - Constantly poll the accelerator to look for completion of inference.
  - Compute MSE for anomaly detection.
- **Accelerator tile**
  - **TinyMLPerf accelerator**
    - Modelled and trained → qKeras
    - Designed with hls4ml
    - Synthesized with Catapult HLS
  - Contains its own **DMA engine**
    - Lower memory access latency



Source [2]



2. Modular infrastructure close-up (II)

- **Memory tile (Scratchpad)**
  - Integrated on-chip SRAM
    - Fast exchange of data
    - High reliability
    - **Accessed by all other tiles by read/write requests**
  - **Ultra low-power** (0.5V)
  - **256 kB Data memory**:
    - 8 x (8192x32b) SRAM
    - CPU general ops mem
    - Accelerator data (Inputs, Outputs, Features, ...)
    - Stores main program
  - **16kB Boot memory**:
    - 4 x (4096x8b)
    - Stores simple bootloader

- **IO tile**
  - Brings data in/out chip
  - Handles voltage level conversion
  - Boot operation
  - Handles Wake/Reset CPU

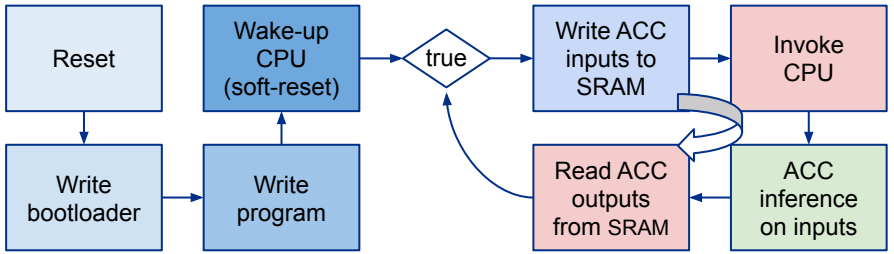


Fig 2. Diagram showing the normal operation of our chip for data inference

3. AI model & application

Neural network architecture:

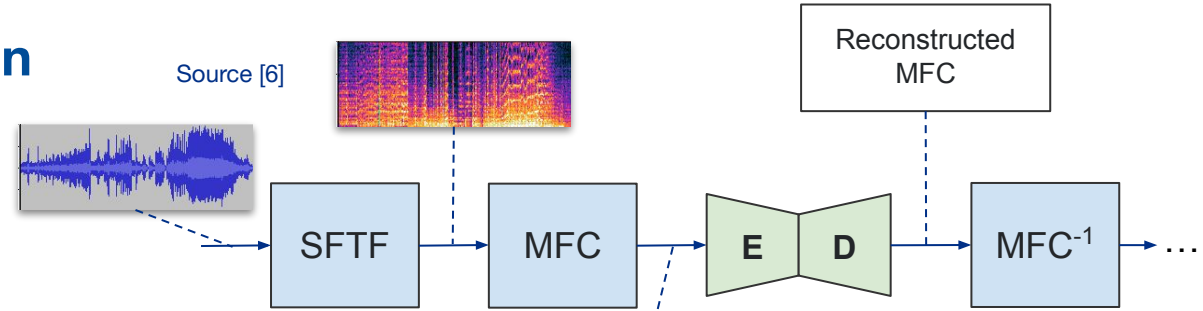
- **Autoencoder** (18K parameters)
- 7 dense layers
- Ref. model 270k parameters (AUC 85%)
- **Quantization** 10b W&B → **AUC 83%**

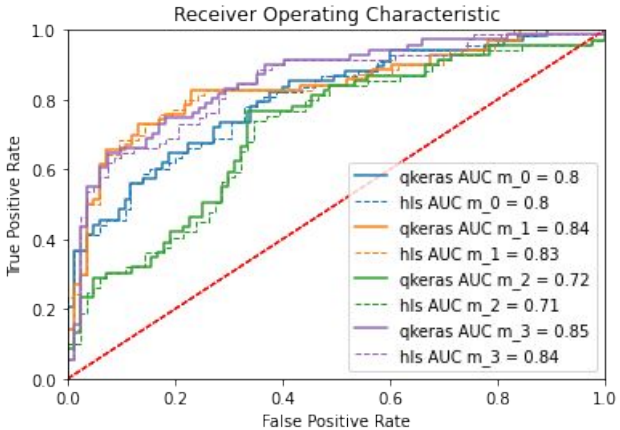
Based on **MLPerf Tiny** v0.5 [4]

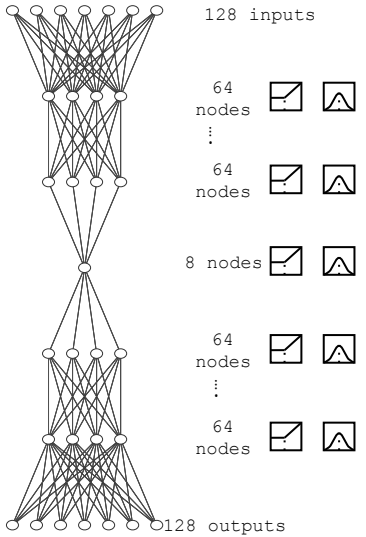
- Benchmark for ML embedded devices performance on-edge
- ToyADMOS dataset [5] (audio signals)
- **Pre-processing**:
  - Short-Time Fourier Transf. (STFT)
  - Mel-frequency Cepstrum of PWR

Anomaly detection:

- **Recreation of MFC features**
- Bad recreation → Anomaly



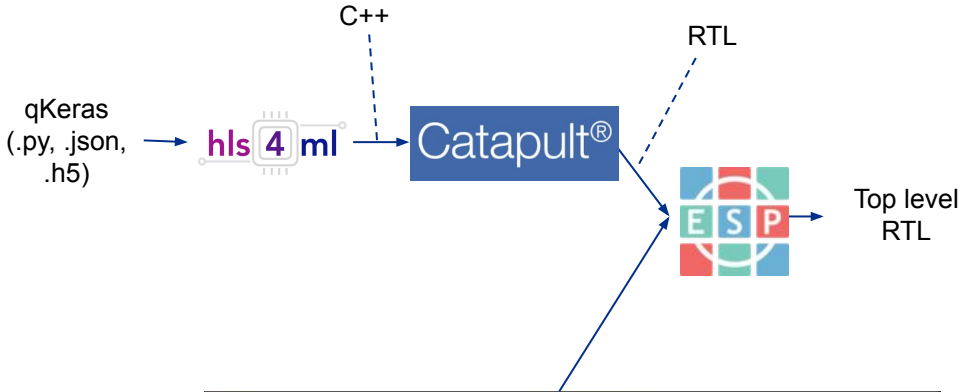


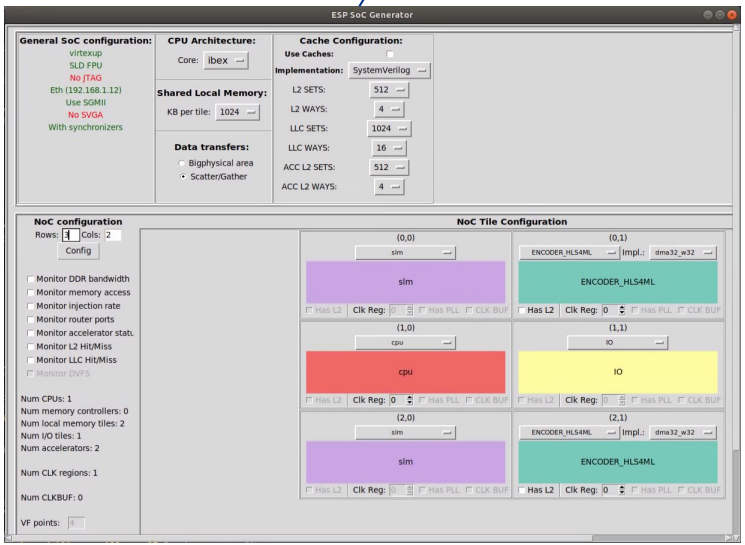


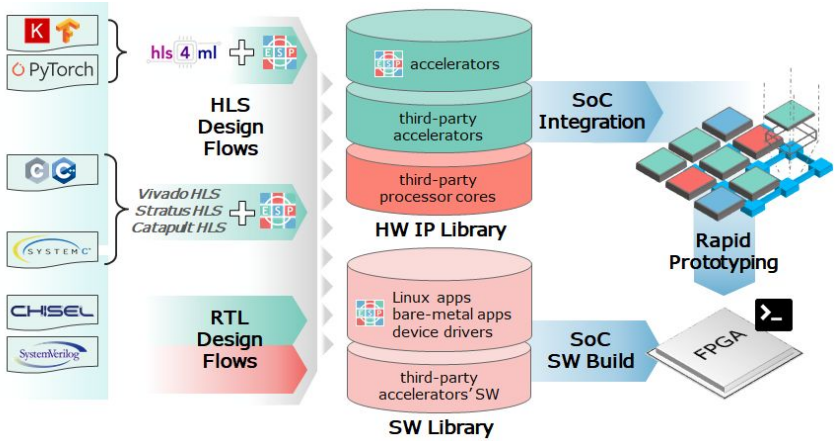
4. System level design flow

Our flow is based in **two main components**:

- **HLS4ML [8]** for **accelerator translation**:
  - qKeras ~> C++
- **ESP [9]** for **automatic tile RTL generation**
  - Ibex RISCV
  - Accelerator tile wrapper
  - Memories
  - IO





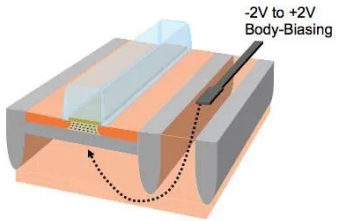


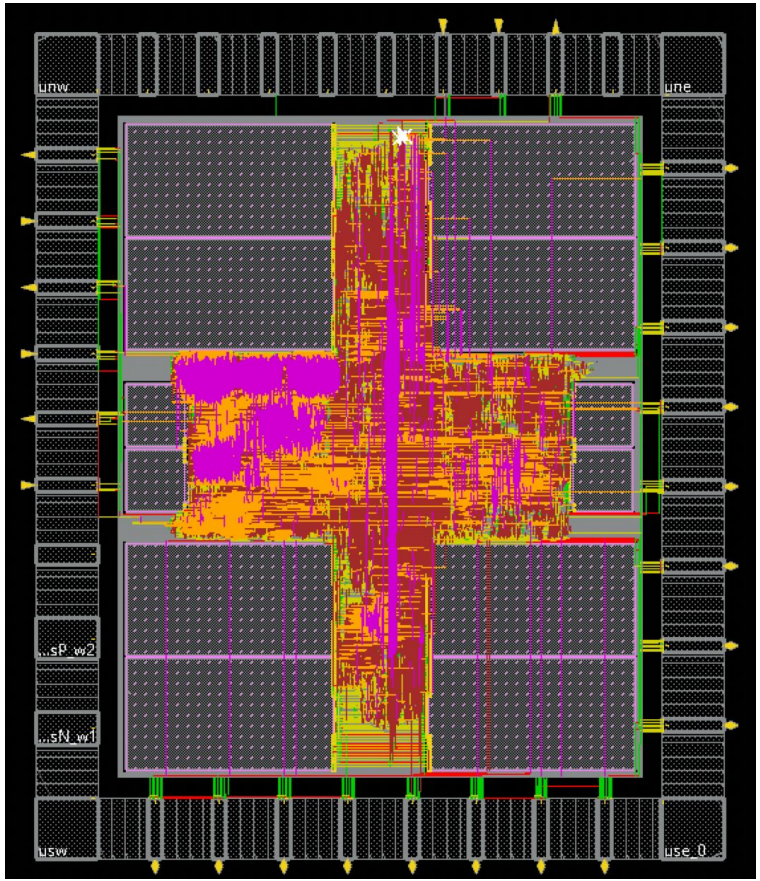
5. Digital implementation on 22nm

Digital flow:

- **GF 22nm process**
- Use of **SLVT devices** → Speed bottleneck (1GHz)
- **Forward body-biasing (FBB) required**:
  - Cryo temperatures → Increase in Vth
  - FBB → Compensate Vth
  - Operate devices at lower voltage
- Currently investigating:
  - How low can be bring the voltage down?

- Forward BB (FBB) enables low voltage operation down to 0.4v without speed loss
- Reverse BB (RBB) enables low leakage down to 1pA/micron
- Dynamic body biasing enables active tradeoff of performance vs. power
- Can be used to reduce variability across the die and/or die-to-die





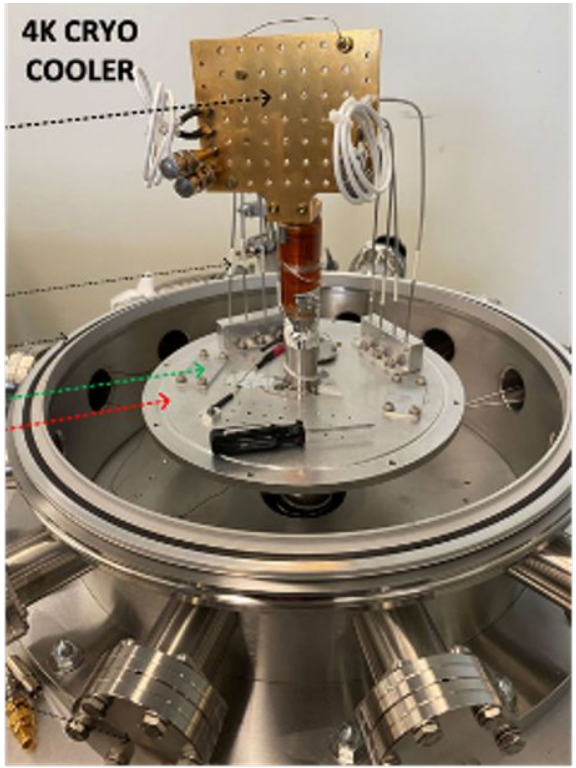
6. Current status

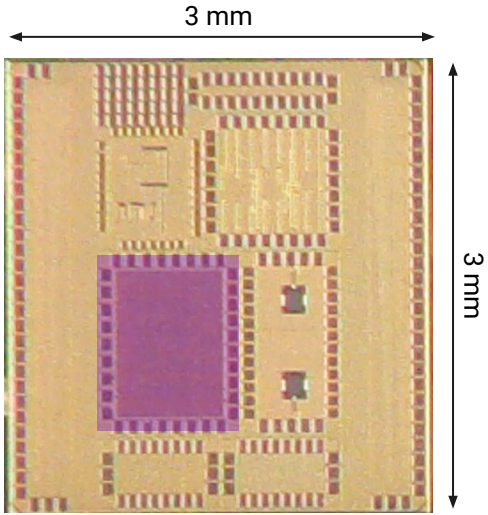
CryoAI chip specs. Version 1.0

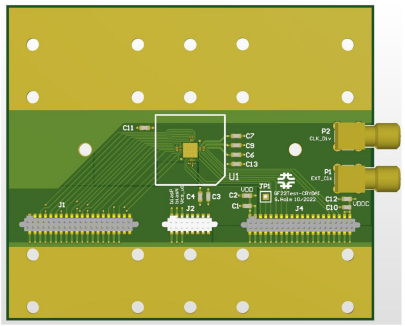
- Chip received in **Apr 2022**
- Final size **1x1.2mm**
  - ▉ Our chip (1x1.2mm)
- Integrated in a multi-project chip
  - Independently tested
- **PCB** designed for testing purposes

Cryo testing:

- Determine operability at cryo temps
  - 4K cryostat



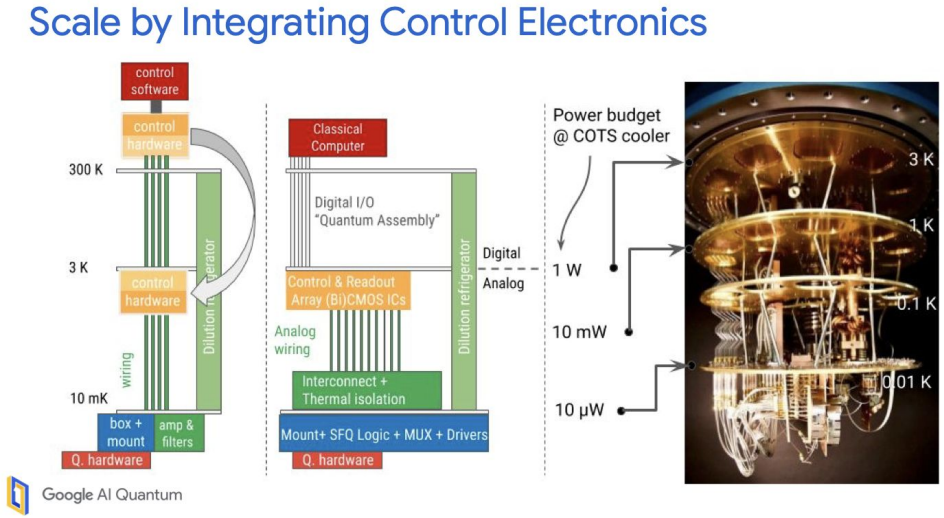




7. Conclusions

How does this fit into quantum control for quantum computing?

- **New design methodology**
  - ML applications
  - Cryogenic environments
  - Low-latency
  - Ultra-low-power
  - Integrated architecture
- Quantum applications can benefit from this flow.
- Flexible to be adapted to other ML applications.
- Alternative to mitigate the wiring problem



Authors / Partnerships / Collaborators

Chinar Syal (Fermi National Accelerator Lab. (US))
†Davide Giri (Columbia University)
Farah Fahim (Fermilab)
Giuseppe Di Guglielmo (Fermilab)
Joseph Zuckerman (Columbia University)
Luca Carloni (Columbia University)
Maico Cassel Dos Santos (Columbia University)
Manuel Valentin (Northwestern University)
Nhan Tran (Fermi National Accelerator Lab. (US))
Seda Memik (Northwestern University)
Jules Muhizi (Fermi National Accelerator Lab. (US))

In memoriam,
Davide Giri

Thank you!

Questions?

Manuel Blanco **Valentin**
manuelbv@fnal.gov
manuelvalentin2028@u.northwestern.edu

Sources

[1] 22 FD-SOI: https://monthly-pulse.com/2021/09/23/optimized-ip-for-gfs-22nm-fdx-technology/

[2] Ibex low-risc implementation: https://github.com/lowRISC/ibex

[3] Zero-Riscy Manual: https://www.pulp-platform.org/docs/user_manual.pdf

[4] MLPerf Tiny v0.5: https://mlcommons.org/en/news/mlperf-tiny-v05/

[5] ToyADMOS dataset: https://github.com/YumaKoizumi/ToyADMOS-dataset

[6] Spectrogram: https://manual.audacityteam.org/man/multi_view.html

[7] MFCC features: https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd

[8] HLS4ML: https://github.com/fastmachinelearning/hls4ml

[9] ESP: https://github.com/sld-columbia/esp