FIDA: a framework to automatically integrate FPGA kernels within Data-Science applications

Luca Stornaiuolo
Dipartimento di Elettronica Informazione e Bioingegneria (DEIB)
luca.stornaiuolo@polimi.it

Alberto Parravicini, Donatella Sciuto and Marco Santambrogio
Context Definition

Huge amount of data that need to be processed to get aggregated information

- Machine Learning (e.g. Neural Network)
- Data Mining (e.g. K-means)
- Image Analysis (e.g. Vessel segmentation)
- Signal Processing (e.g. Autocorrelation)
Problem Definition

• Most computational pipelines rely on high-level languages like Python, R and MATLAB.

• These pipelines can be accelerated with OpenCL and FPGAs.

• Integrating OpenCL kernels with high-level languages is slow and error-prone.
Proposed Solution

• Integrating OpenCL kernels with high-level languages is slow and error-prone.

• **FIDA automatize this process!**
  From a **simple description of the kernel**, we create interfaces for Python, R and MATLAB.
...to this!

```json
{
    "kernel_name": "mmult",
    "board": ["xilinx_adm-pcie-7v3_1ddr_3_0"],
    "xclbin": ["kernel_7v3.xclbin"],
    "num_iterations": 3,
    "inputs": [
        {
          "type": "array",
          "name": "a",
          "length": 256,
          "class": "int",
          "position": 0
        },
        {
          "type": "array",
          "name": "b",
          "length": 256,
          "class": "int",
          "position": 1
        }
    ],
    "outputs": [
        {
          "type": "array",
          "name": "c",
          "length": 256,
          "class": "int",
          "position": 2
        }
    ]
}
```
Architecture

```python
var = python_interface.var_fpga(x, len(x))
```
Experimental Results

• We tested our architecture on different computational kernels.

• We profiled executions to find where bottlenecks are.

• We analyzed introduced execution time overheads.
Experimental Results

- **Execution profile:**

![Execution profile graph](image)
Experimental Results

- **Execution profile:**

- Most of the time is spent in interprocess communication: **shared memory** could drastically improve performance!
FIDA: a framework to automatically integrate FPGA kernels within Data-Science applications

https://necst.it/
https://www.slideshare.net/necstlab
https://www.facebook.com/groups/ReconfigurableArchitecturesWorkshop/

Luca Stornaiuolo
Dipartimento di Elettronica Informazione e Bioingegneria (DEIB)
luca.stornaiuolo@polimi.it

Alberto Parravicini, Donatella Sciuto and Marco Santambrogio