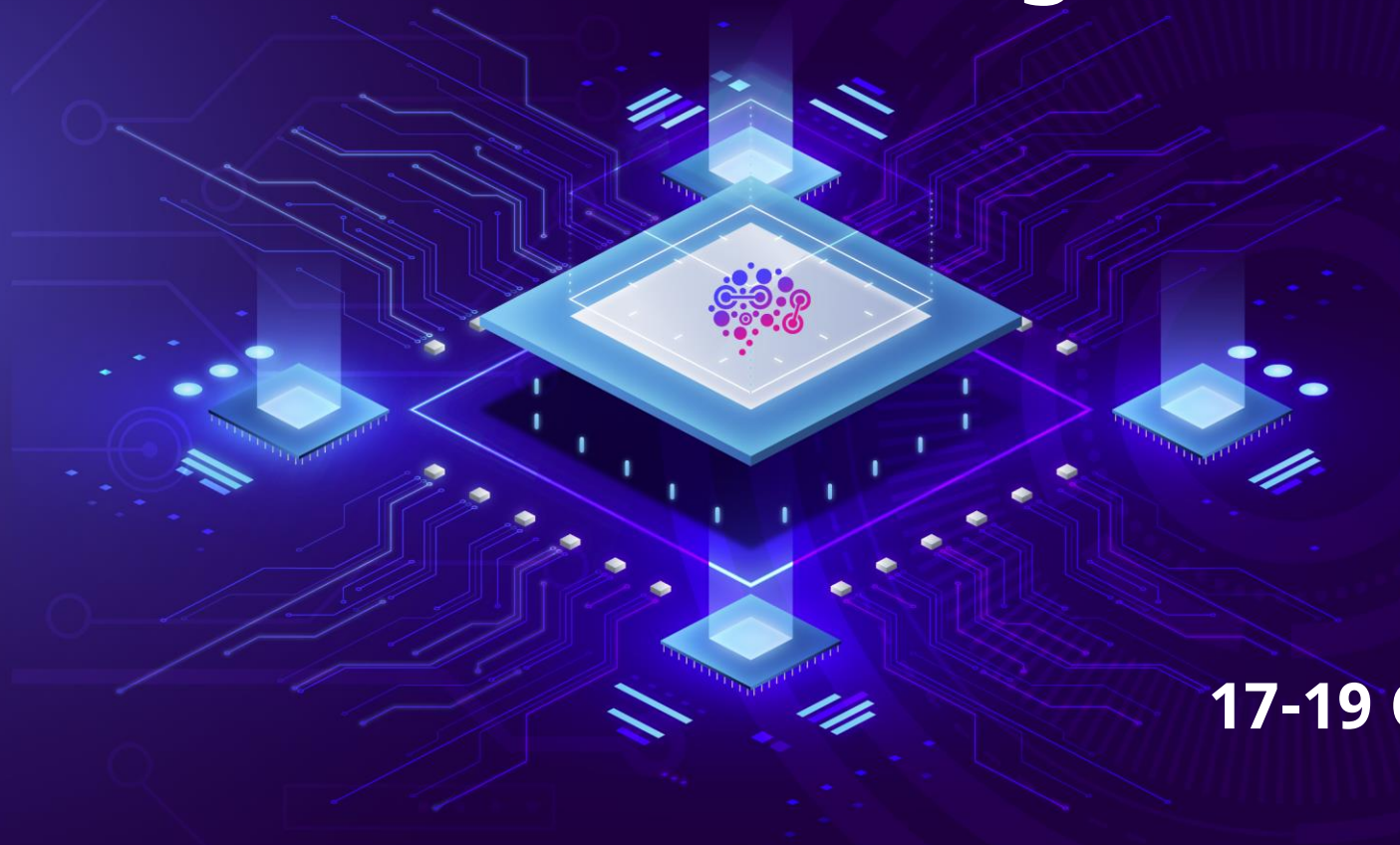
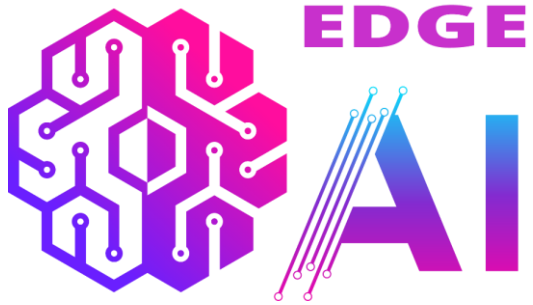
European Conference on EDGE AI Technologies and Applications - EEAI



Advancing emerging edge AI technologies and driving next generation intelligent applications.

17-19 October 2023, Athens, Greece

European Conference on EDGE AI Technologies and Applications - EEAI



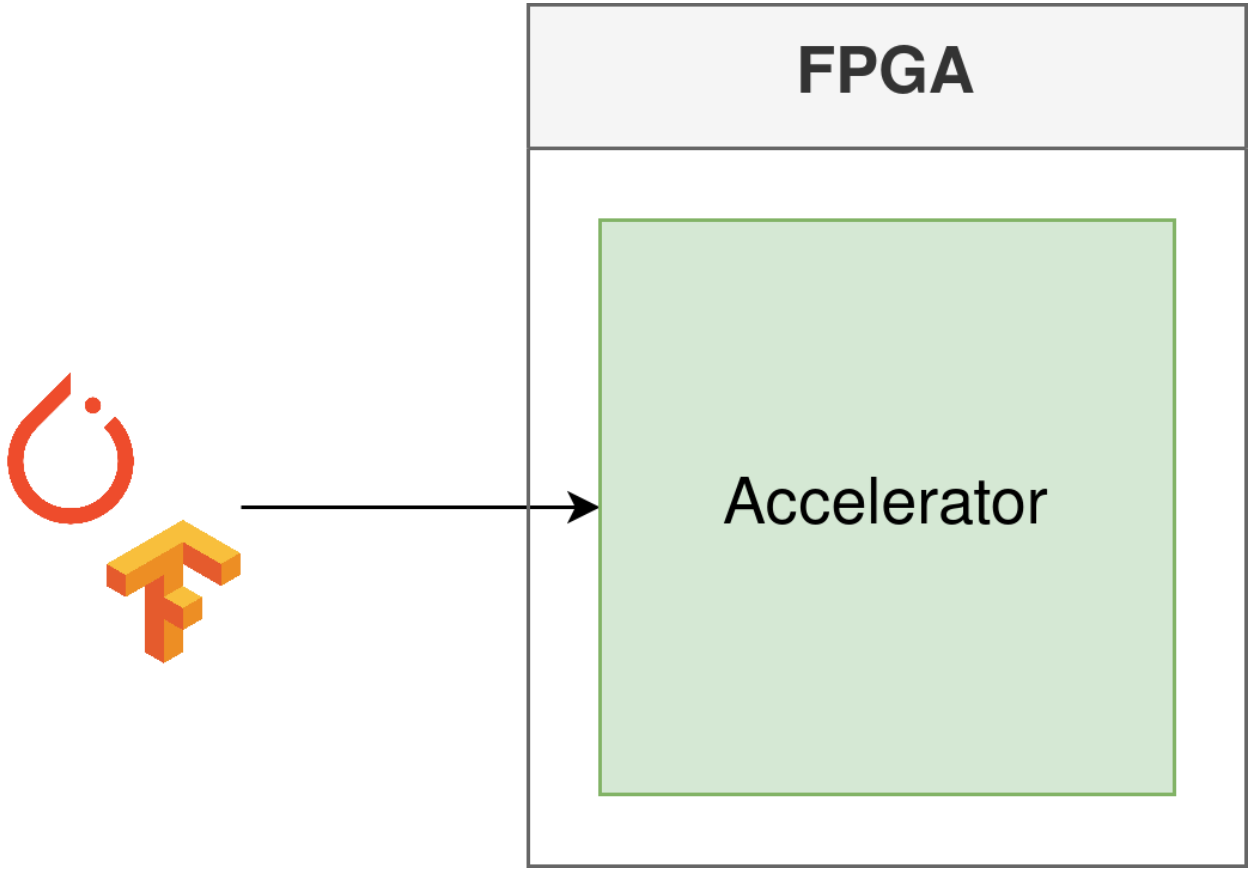
Efficient mapping of neural networks to FPGA

Thierry Delafontaine, ZHAW, Switzerland

17-19 October 2023 Athens, Greece

Copyright © 2023

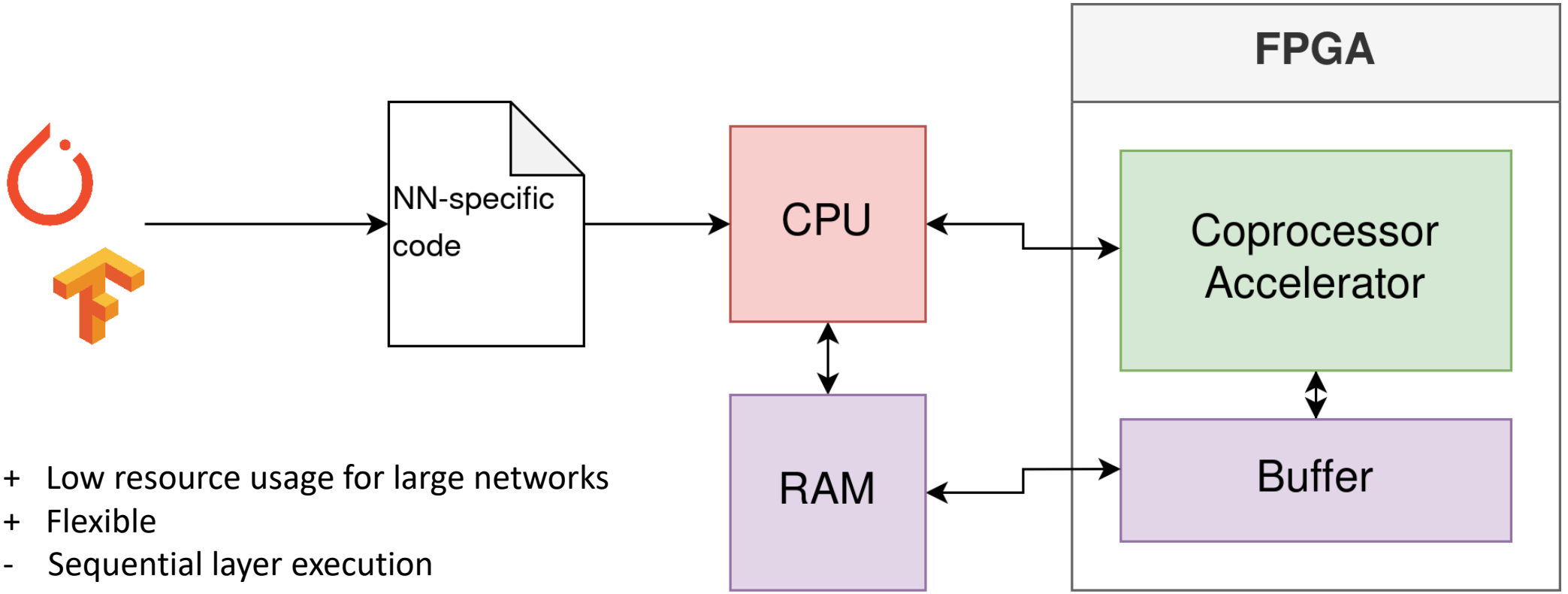
Introduction



Copyright © 2023

State-of-the-art

Coprocessor-style accelerator



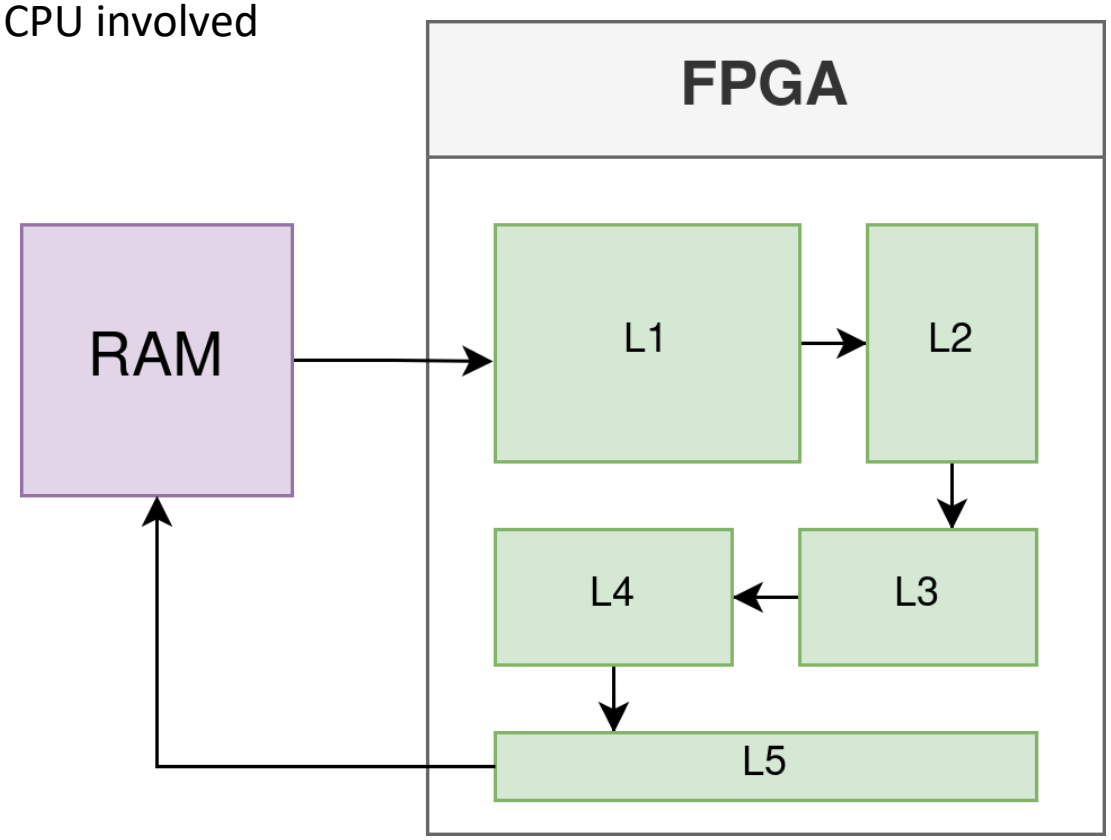
\+ Low resource usage for large networks
\+ Flexible
\- Sequential layer execution

Copyright © 2023

Methodology

Dataflow-style accelerator

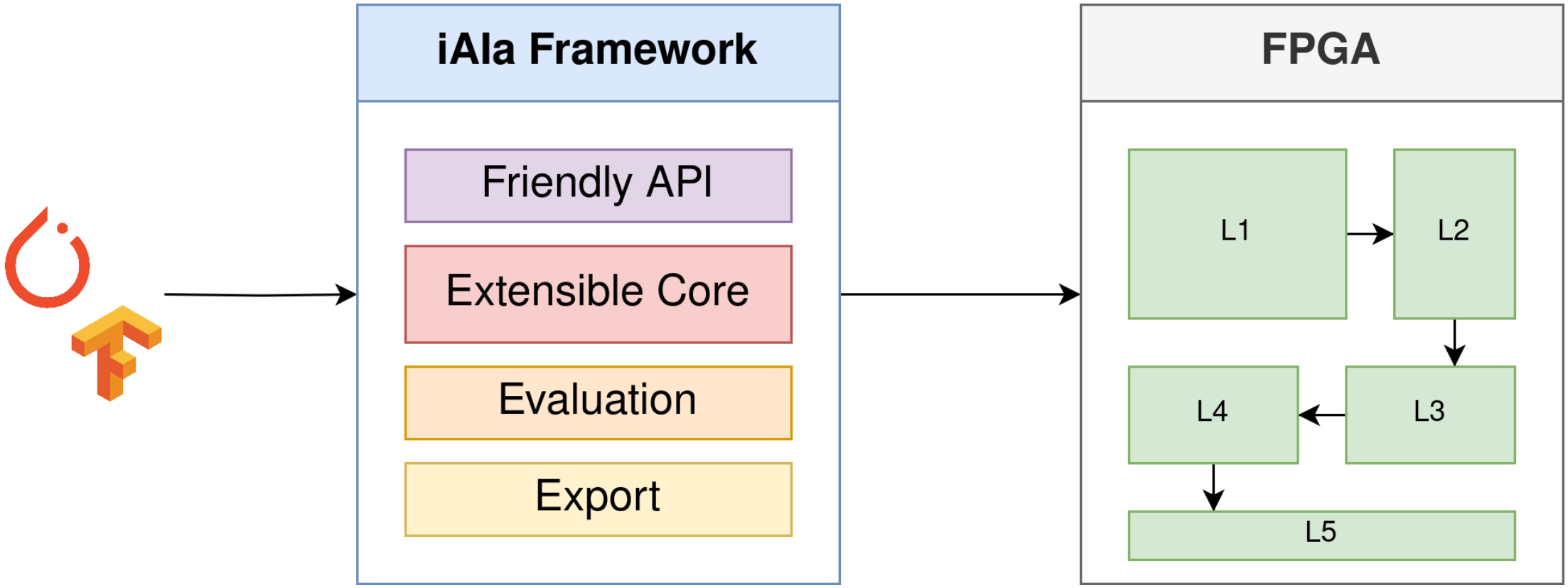
No CPU involved



+ Pipelined layer execution
- Specific
- High resource usage

Copyright © 2023

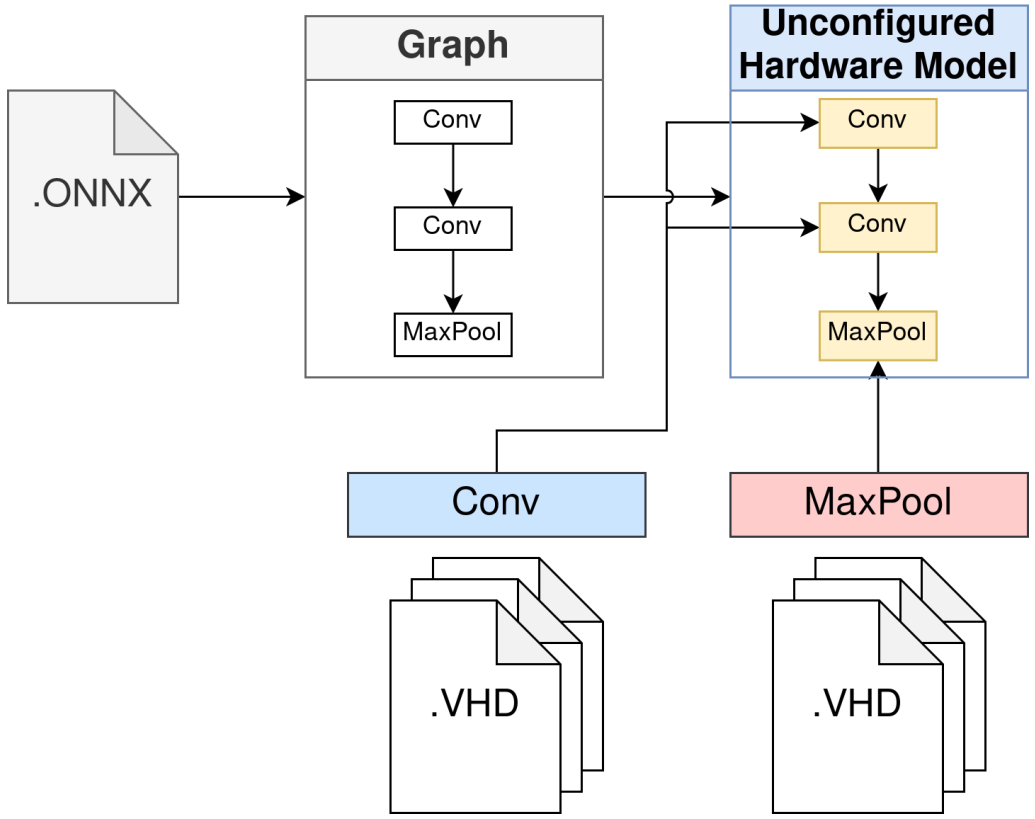
Our Goal



Copyright © 2023

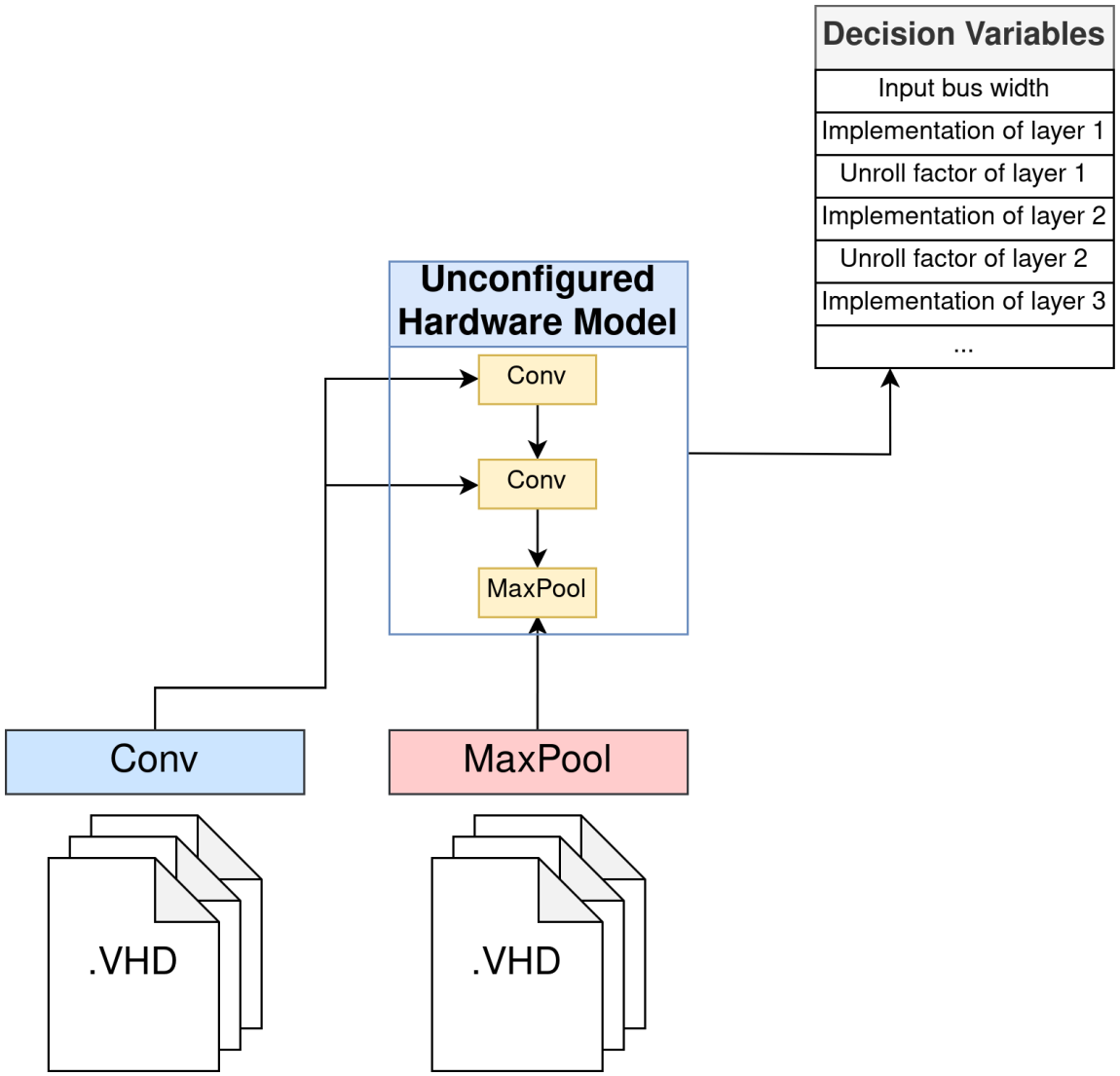
Concept

Open Neural Network Exchange Format



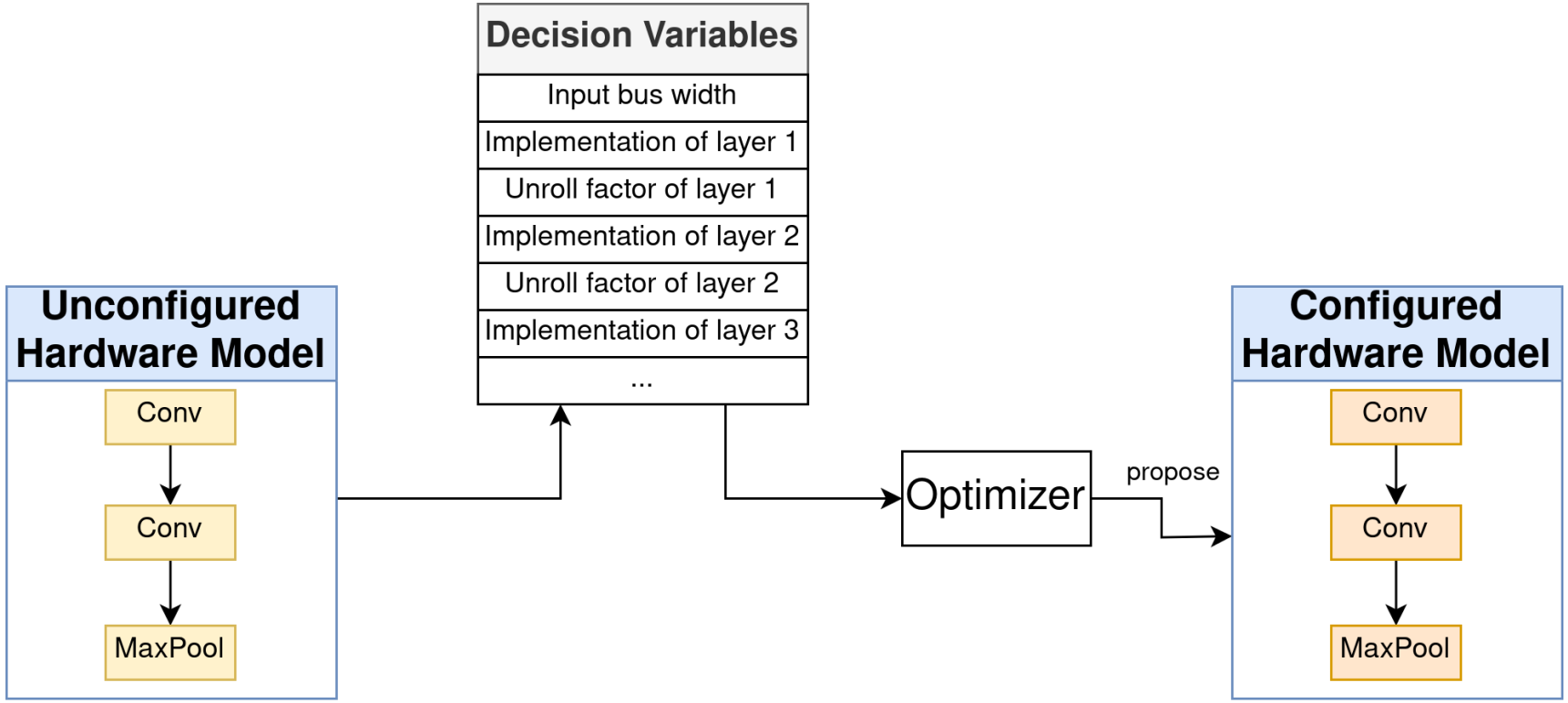
Copyright © 2023

Concept



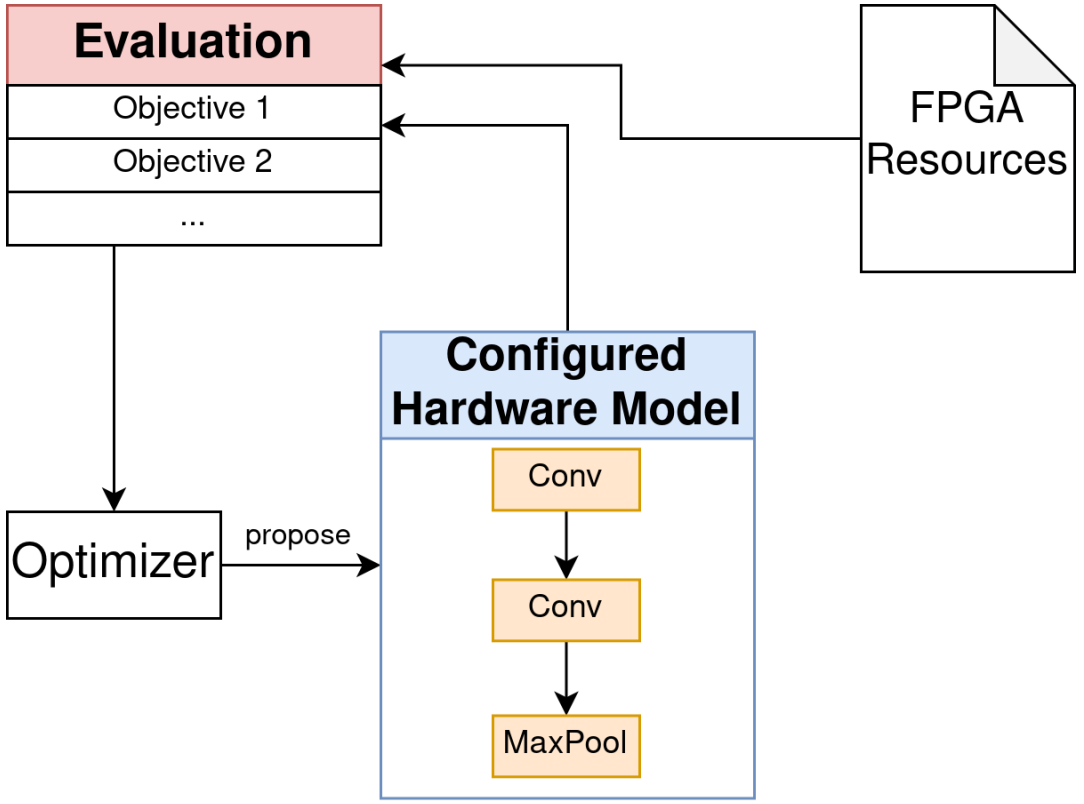
Copyright © 2023

Concept



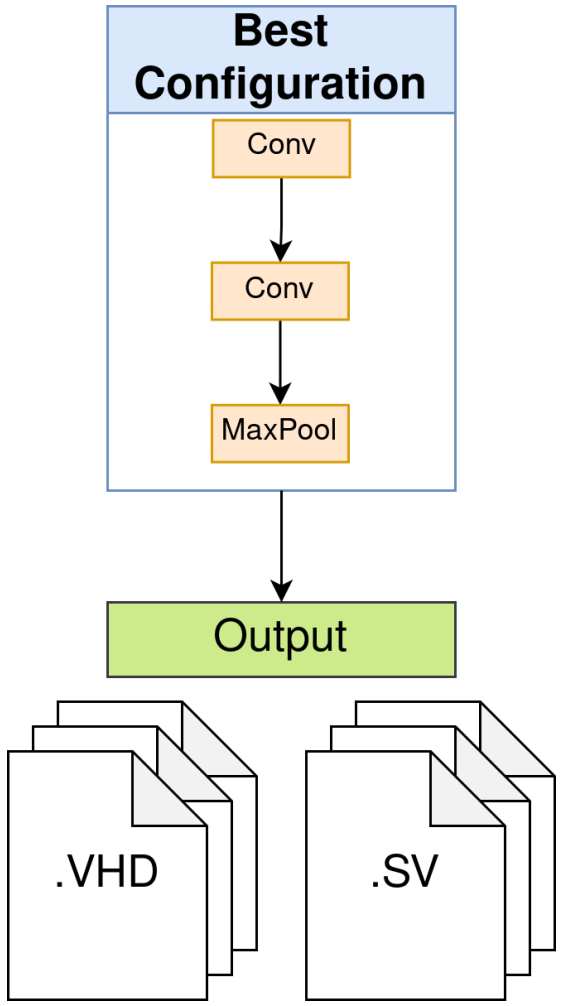
Copyright © 2023

Concept



Copyright © 2023

Concept



Copyright © 2023

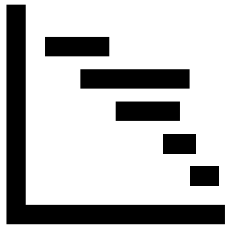
Experimental Results

```
module conv1 #(
    parameter string Implementation,
    parameter int UnrollFactor,
    parameter int XTDataWidth,
    parameter int XTIDWidth,
    parameter int XTDestWidth,
    parameter int XTUserWidth,
    parameter string XType,
    ...
) (
    input logic clk_i,
    input logic rst_ni,
    input logic x_axis_tvalid_i,
    output logic x_axis_tready_o,
    input logic[XTDataWidth-1:0] x_axis_tdata_i,
    input logic[XTDataWidth/8-1:0] x_axis_tstrb_i,
    input logic[XTDataWidth/8-1:0] x_axis_tkeep_i,
    input logic x_axis_tlast_i,
    input logic[XTIDWidth-1:0] x_axis_tid_i,
    input logic[XTDestWidth-1:0] x_axis_tdest_i,
    input logic[XTUserWidth-1:0] x_axis_tuser_i,
    ...
);

  localparam string AUTO_PAD = "NOTSET";
  localparam int[0:0] DILATIONS = '{1};
  localparam int GROUP = 1;
  localparam int[0:0] KERNEL_SHAPE = '{3};
  localparam int[0:0] PADS = '{0};
  localparam int[0:0] STRIDES = '{1};
  ...
endmodule
```

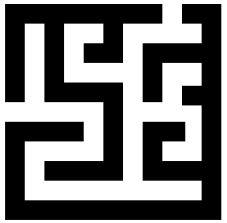
```
library ieee;
use ieee.std_logic_1164.all;
use ieee.numeric_std.all;

library ops;
use ops.all;

entity conv1 is
  generic (
    Implementation : string(0 to 64);
    UnrollFactor : integer;
    XTDataWidth : integer;
    XTIDWidth : integer;
    XTDestWidth : integer;
    XTUserWidth : integer;
    XType : string(0 to 64);
    ...
  );
  port (
    clk_i : in std_logic;
    rst_ni : in std_logic;
    x_axis_tvalid_i : in std_logic;
    x_axis_tready_o : out std_logic;
    x_axis_tdata_i : in std_logic_vector(XTDataWidth-1 downto 0);
    x_axis_tstrb_i : in std_logic_vector(XTDataWidth/8-1 downto 0);
    x_axis_tkeep_i : in std_logic_vector(XTDataWidth/8-1 downto 0);
    x_axis_tlast_i : in std_logic;
    x_axis_tid_i : in std_logic_vector(XTIDWidth-1 downto 0);
    x_axis_tdest_i : in std_logic_vector(XTDestWidth-1 downto 0);
    x_axis_tuser_i : in std_logic_vector(XTUserWidth-1 downto 0);
    ...
  );
end entity;

architecture rtl of conv1 is
  constant AUTO_PAD : string(0 to 64) := "NOTSET";
  constant DILATIONS : array (0 to 0) of integer := (1);
  constant GROUP : integer := 1;
  constant KERNEL_SHAPE : array (0 to 0) of integer := (3);
  constant PADS : array (0 to 0) of integer := (0);
  constant STRIDES : array (0 to 0) of integer := (1);
begin
  ...
end architecture;
```

Copyright © 2023

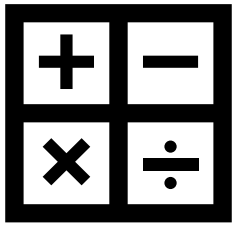
Future Steps



Process scheduling



Memory management

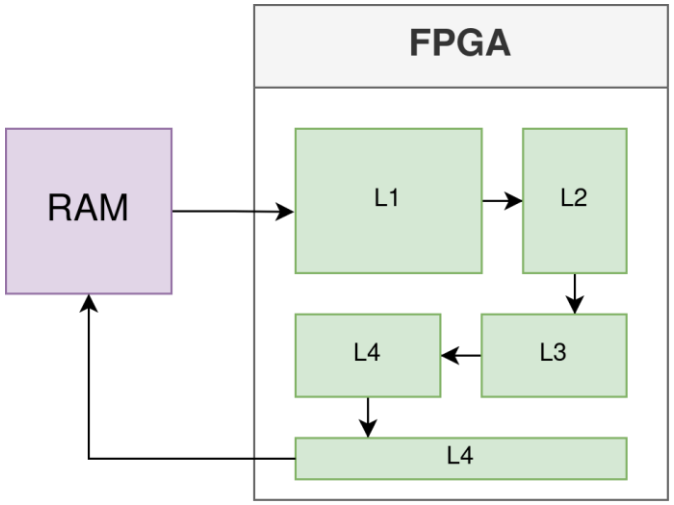


Operator support

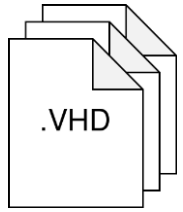
Copyright © 2023

Conclusion

In development



Dataflow-style accelerator



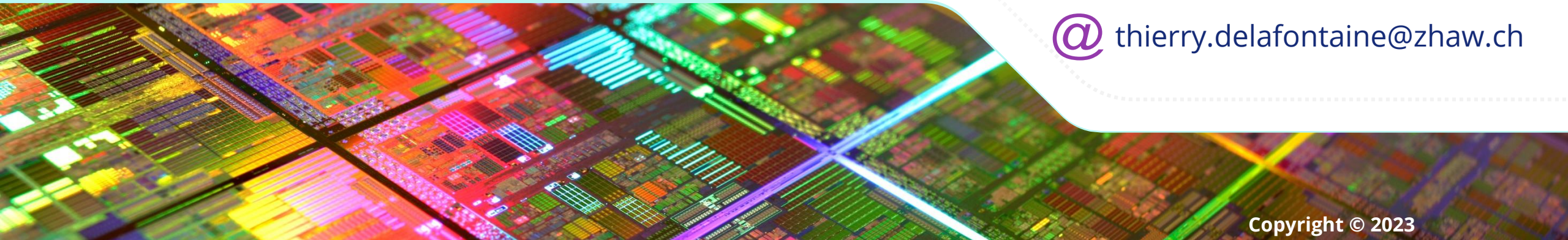
RTL templates

Copyright © 2023

EDGE AI

Thank You

For your attention



@ thierry.delafontaine@zhaw.ch

Copyright © 2023