Flash Attention on FPGA

Aakarsh A
Department of Computer Science and Automation
Indian Institute of Science

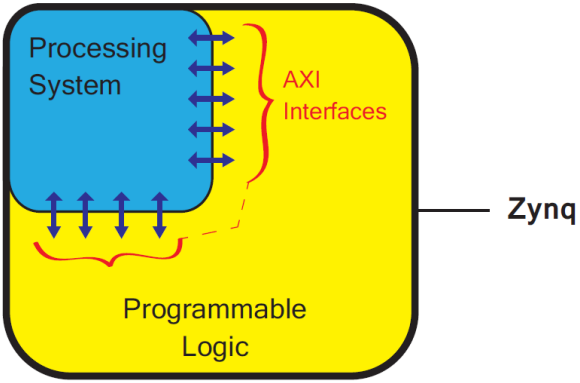
June 06, 2025

Overview

1. e^{-x} on Zedboard
2. Relation with AMD Alveo V80
3. Future Work

Implementation Result: 1

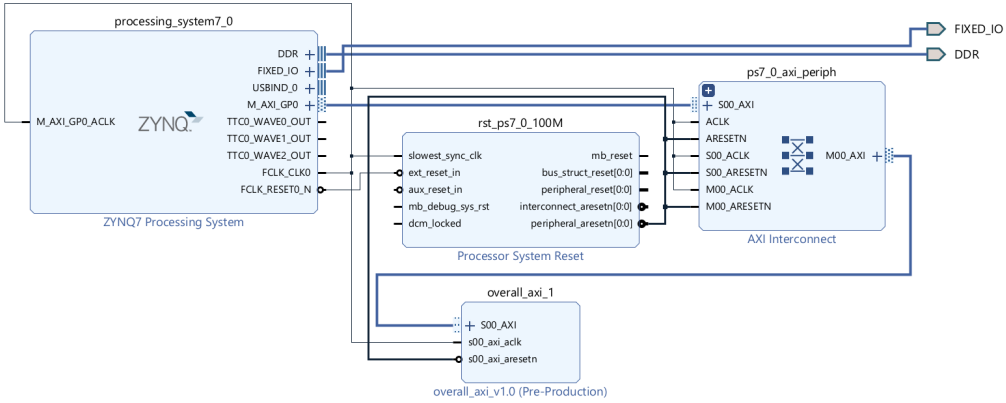
Zedboard: Zynq 7000 SoC + PL



The previous results of the implementation on Zedboard were just PL with GPIO parts, it does not give a true estimate of the operating frequency, as it did not involve PS.

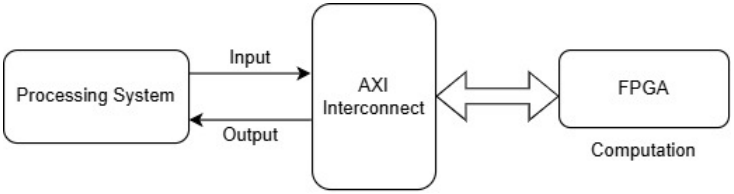
Implementation Result: 2

PS & PL Approach



Block Design integrating PS and PL via AXI Interconnect, and our designed RTL block is created as another AXI-4 Lite block to effectively interface with PS.

Implementation Result: 2



Implementation on Zedboard

Setup		Hold		Pulse Width	
Worst Negative Slack (WNS):	3.995 ns	Worst Hold Slack (WHS):	0.033 ns	Worst Pulse Width Slack (WPWS):	4.020 ns
Total Negative Slack (TNS):	0.000 ns	Total Hold Slack (THS):	0.000 ns	Total Pulse Width Negative Slack (TPWS):	0.000 ns
Number of Failing Endpoints:	0	Number of Failing Endpoints:	0	Number of Failing Endpoints:	0
Total Number of Endpoints:	1506	Total Number of Endpoints:	1506	Total Number of Endpoints:	669

All user specified timing constraints are met.

Timing Analysis for 3 CC Latency design. Based on WNS for 100 MHz, we can still push the operating frequency to 166 MHz.

Implmentation Result: 2

Setup		Hold		Pulse Width	
Worst Negative Slack (WNS):	-0.397 ns	Worst Hold Slack (WHS):	0.027 ns	Worst Pulse Width Slack (WPWS):	4.020 ns
Total Negative Slack (TNS):	-19.071 ns	Total Hold Slack (THS):	0.000 ns	Total Pulse Width Negative Slack (TPWS):	0.000 ns
Number of Failing Endpoints:	48	Number of Failing Endpoints:	0	Number of Failing Endpoints:	0
Total Number of Endpoints:	1680	Total Number of Endpoints:	1680	Total Number of Endpoints:	716

Timing constraints are not met.

Timing Analysis for 2 CC Latency design [2 Multiplications in the same CC]. Based on WNS for 100 MHz, Timing violations are observed, the maximum operating frequency is limited to 96.1 MHz.

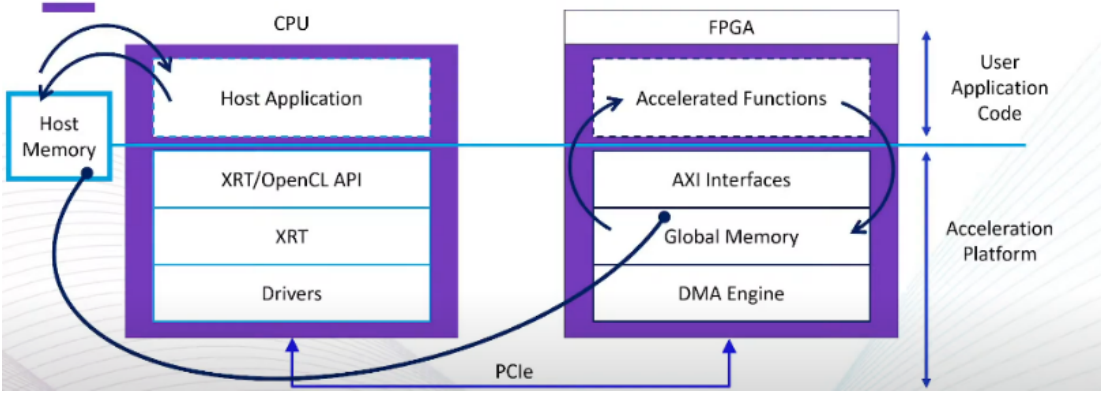
Xilinx SDK

```
xil_printf(ctrl1: "Writing x_val = %d, inp_valid = %d\r\n", x_val, inp_valid);
Xil_Out32(Addr: AXI_ADDRESS + 0x00, Value: x_val);
Xil_Out32(Addr: AXI_ADDRESS + 0x04, Value: inp_valid);

xil_printf(ctrl1: "Waiting for o_valid = 1...\r\n");
do {
    read_data = Xil_In32(Addr: AXI_ADDRESS + 0x0C);  // read entire 32-bit reg
    out_valid   = read_data & 0x1;                   // extract bit[0]
} while (out_valid == 0);

// At this point, o_valid == 1
xil_printf(ctrl1: "o_valid asserted by PL\r\n");

y = Xil_In32(Addr: AXI_ADDRESS + 0x08);

// while (1) {
// Might add extra software logic.
// }
xil_printf(ctrl1: "Read exp_value = 0x%08x (decimal %d)\r\n", (unsigned) y, (int) y);
```

Generate Bitstream, export Hardware, and set it as Platform hardware in Vitis SDK, and Code up the application/task as needed.

Relation with AMD Alveo V80

Integrating PS / PL Flow is similar



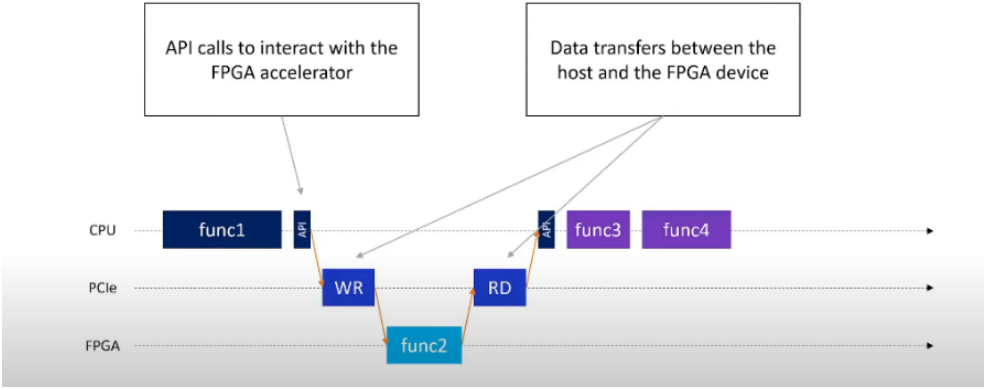
- Replace: ARM processor with x86, Replace AXI with PCIe.
- Each generated RTL block is wrapped and made an AXI-4 peripheral or AXI-MM peripheral and are called kernals.

FPGA Inference Flow in V80

1. The application on the host initiates data transfer using the **XRT API**.
2. XRT and the drivers move the data over PCIe and using the DMA engine, the memory goes directly to **DDR** (HBM or URAM or BRAM).
3. Using the AXI interface connected via **NoC**, the data reaches the **Kernals**, which are interconnected by a NoC as well.
4. The processed result is sent back to DDR through DMA and is stored directly in host memory over PCIe and is accessed via application when needed.
5. The whole process is done in Vivado till Kernel wrapping, after which it is done in Vitis.

.

Hardware Acceleration: A More Accurate View



Future Work

1. Verification of Parametrized Phase-2 on Zedboard [By Wednesday]
2. RTL implementation of Phase-3 with integration with Phase-2 [By Friday]
3. Verification of the above on Zedboard
4. Verification of Phase-2 on the Alveo V80 FPGA.

THE END

Feedback & Improvement Ideas?