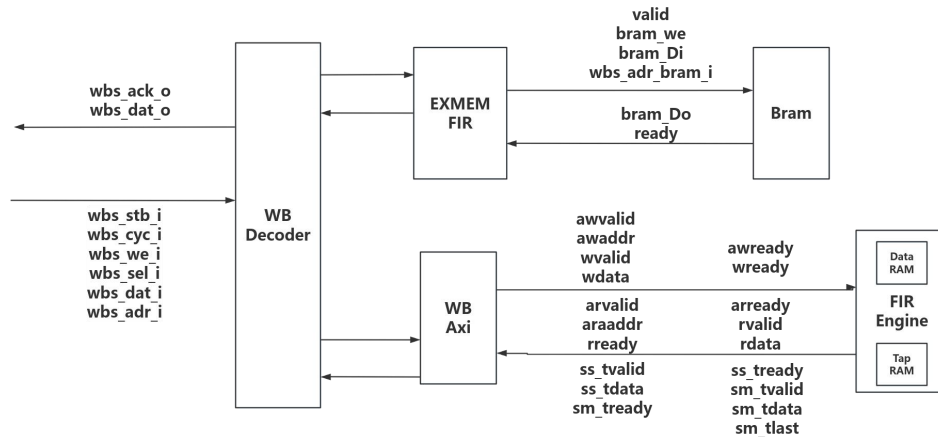


SOC Design

Lab4-2 Caravel FIR Report

21115793 ZHANG Mengmeng

- Design block diagram – datapath, control-path



The FIR accelerator adopts a co-processor model with tightly coupled firmware control.

Datapath:

Input Interface: AXI4-Stream for high-speed data ingestion.

Coefficient Storage: 12-bit addressable BRAM (Tap RAM) stores 11 filter coefficients.

Data Buffer: Circular buffer in BRAM (Data RAM) holds 11 historical input samples.

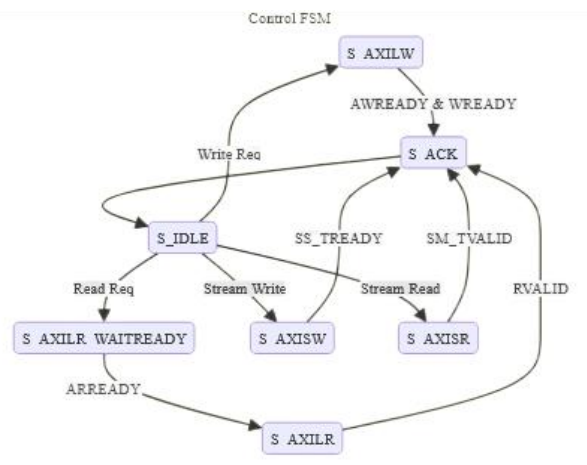
Compute Engine: Parallel MAC (Multiply-Accumulate) unit with 32-bit fixed-point arithmetic.

Output Interface: AXI4-Stream for result streaming.

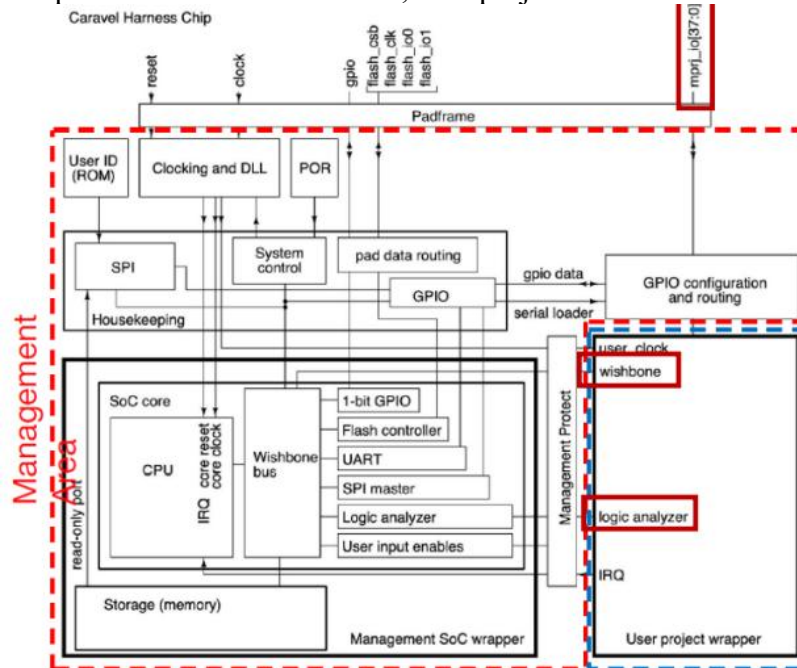
Control Path:

AXI4-Lite Control Registers: `ap_start`, `ap_done`, `ap_idle` for firmware coordination.

Wb State Machine:



- The interface protocol between firmware, user project and testbench



Base Address	Offset	Bit	Description
0x3000_0000	0x00	0	ap_start: set to 1 to start the FIR engine
		1	ap_done: assert when FIR engine processes and transfers all the data
		2	ap_idle: indicate whether FIR engine is actively processing data
		3	Reserved zero
		4	FIR engine is ready to accept input x[n]
		5	Output y[n] is ready to be read
	0x10 – 0x13	31:0	Data length
	0x40 – 0x7F	31:0	Tap coefficients
	0x80 – 0x83	31:0	Input x[n]
	0x84 – 0x87	31:0	Output y[n]
0x3800_0000			Execution memory that stores firmware code

Firmware \leftrightarrow Hardware:

AXI4-Lite:

Tap BRAM

Control registers: start/done flags

AXI4-Stream:

Input: ss tvalid/ss tready handshake for data streaming.

Output: sm tvalid/sm tready handshake with sm tlast marking frame end.

User Project \leftrightarrow Testbench:

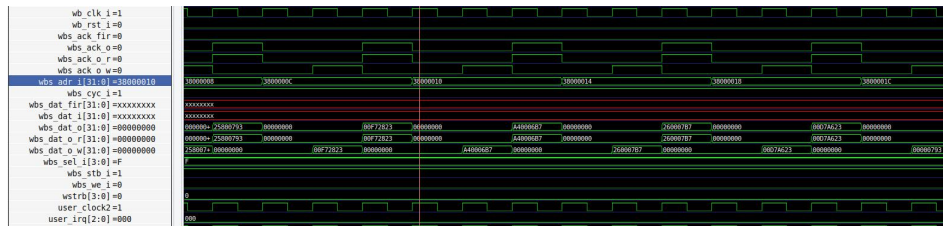
Data BRAM: Dual-port RAM for buffering input samples.

Tap BRAM: Stores filter coefficients (11-tap).

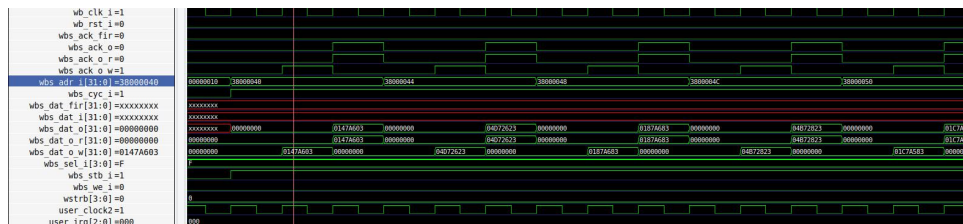
- Waveform and analysis of the hardware/software behavior.



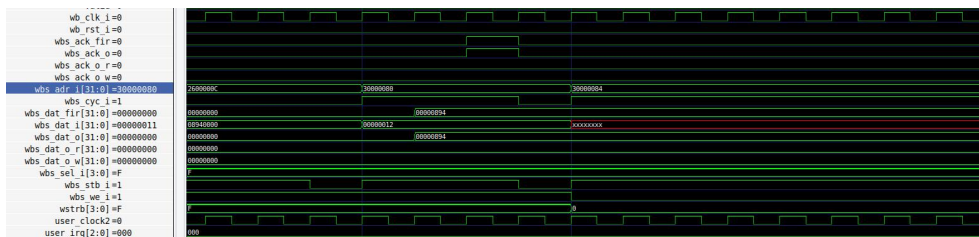
The captured interface timing demonstrates that during a Wishbone read operation at address 0x3000_0000, the assertion of the ack signal occurs after address validation, the hardware's readiness to accept X[n] input data streams.



Following the configuration of the test length via a Wishbone write transaction to address 0x3000_0010.



Firmware Send a write command to the wishbone, and write the address 3800_0040.



Firmware Send a write command to wishbone, address 3000_0080, indicating that X[n] is being transmitted.



Firmware Send a read command to wishbone, address 3000_0084, indicating that Y[n] is being delivered.

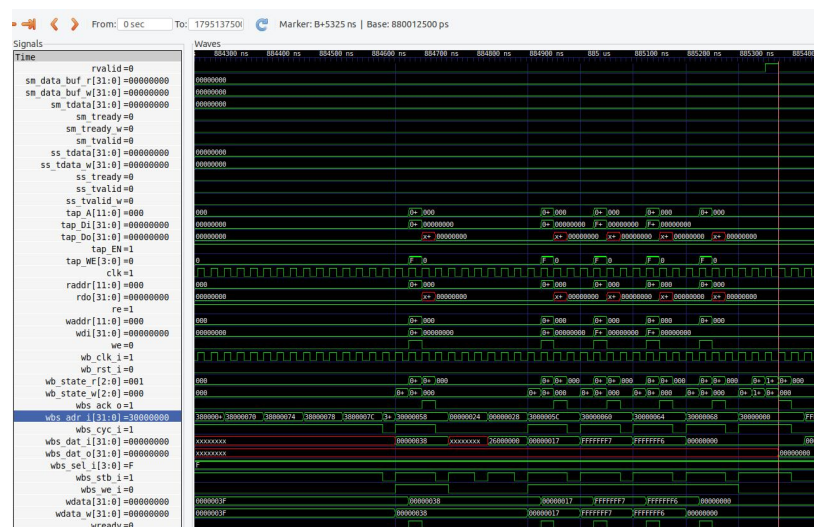
- What is the FIR engine theoretical throughput, i.e. data rate? Actually measured throughput?

FIR engine theoretical throughput: $32 \text{ bits} / 12 \text{ cycle} = 1.6 \text{ bits/ns}$

Actually measured throughput: $600 * 32 \text{ bits} / 104846 \text{ cycle} = 0.1824 \text{ bits/ns}$

- What is latency for firmware to feed data?

Latency for firmware: $5325 \text{ ns} = 213 \text{ cycle}$



- What techniques used to improve the throughput?

Dual BRAM Ports: Use one BRAM buffer for computation while preloading the next input, enables simultaneous coefficient/data access to reduces idle cycles between samples.

Firmware Optimization: Minimize loop overhead (e.g., unroll loops for bulk data transfers).

AXI Burst Transfers: Replace single-beat AXI transactions with burst mode to reduce handshake overhead.

- Does bram12 give better performance, in what way?

If the data RAM is replaced with BRAM12, it may reduce latency and improve throughput. The key rationale lies in utilizing the additional space provided by BRAM12 as an input buffer. This allows overlapping computation and data transfer phases, thereby optimizing resource utilization. BRAM12 reserves a dedicated buffer region to preload the next input sample while the FIR engine processes the current input. This eliminates idle cycles between consecutive computations by hiding data transfer latency.

While the FIR engine calculates the output for the first input, the buffer preloads the second input. The waiting time for output handshakes is reused for processing subsequent inputs.

- Can you suggest other method to improve the performance?

Parallel MAC Units: Process multiple taps concurrently.

Pipelined FSM: Overlap computation and data ingestion.

AXI DMA: Offload data transfer from firmware.

- Resource usage

Utilization report including FF, LUT and BRAM:

```

28 1. Slice Logic
29 -----
30
31 +-----+-----+-----+-----+-----+-----+
32 | Site Type | Used | Fixed | Prohibited | Available | Util% |
33 +-----+-----+-----+-----+-----+-----+
34 | Slice LUTs* | 439 | 0 | 0 | 53200 | 0.83 |
35 | LUT as Logic | 375 | 0 | 0 | 53200 | 0.70 |
36 | LUT as Memory | 64 | 0 | 0 | 17400 | 0.37 |
37 | LUT as Distributed RAM | 64 | 0 | 0 | 0 | 0.00 |
38 | LUT as Shift Register | 0 | 0 | 0 | 0 | 0.00 |
39 | Slice Registers | 348 | 0 | 0 | 106400 | 0.33 |
40 | Register as Flip Flop | 348 | 0 | 0 | 106400 | 0.33 |
41 | Register as Latch | 0 | 0 | 0 | 106400 | 0.00 |
42 | F7 Muxes | 0 | 0 | 0 | 26600 | 0.00 |
43 | F8 Muxes | 0 | 0 | 0 | 13300 | 0.00 |
44 +-----+-----+-----+-----+-----+-----+
45 * Warning! The Final LUT count, after physical optimizations and full implementation,
46
67 2. Memory
68 -----
69
70 +-----+-----+-----+-----+-----+-----+
71 | Site Type | Used | Fixed | Prohibited | Available | Util% |
72 +-----+-----+-----+-----+-----+-----+
73 | Block RAM Tile | 4 | 0 | 0 | 140 | 2.86 |
74 | RAMB36/FIFO* | 4 | 0 | 0 | 140 | 2.86 |
75 | RAMB36E1 only | 4 | 0 | 0 | 0 | 0.00 |
76 | RAMB18 | 0 | 0 | 0 | 280 | 0.00 |
77 +-----+-----+-----+-----+-----+-----+
81 3. DSP
82 -----
83
84 +-----+-----+-----+-----+-----+-----+
85 | Site Type | Used | Fixed | Prohibited | Available | Util% |
86 +-----+-----+-----+-----+-----+-----+
87 | DSPs | 3 | 0 | 0 | 220 | 1.36 |
88 | DSP48E1 only | 3 | 0 | 0 | 0 | 0.00 |
89 +-----+-----+-----+-----+-----+-----+

```

- Timing report

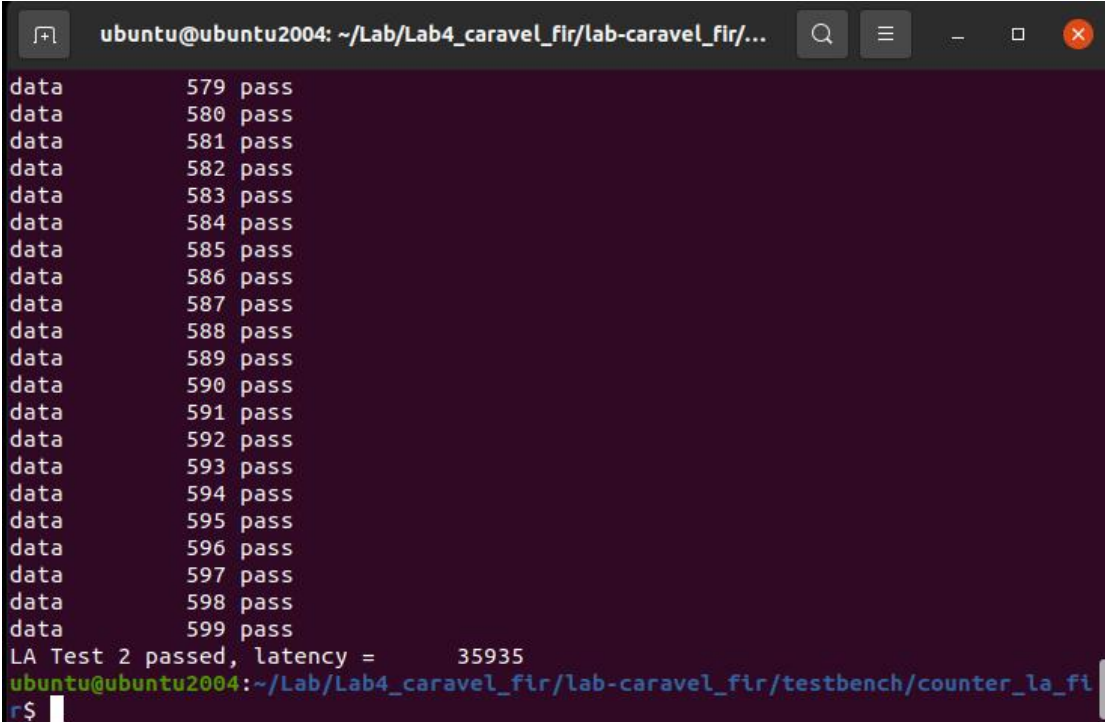
```

240 -----
241 | Clock Summary
242 | -----
243 -----
244
245 Clock      Waveform(ns)      Period(ns)      Frequency(MHz)
246 -----
247 wb_clk_i   {0.000 12.500}     25.000          40.000
248
249 -----
300 | Timing Details
301 | -----
302 -----
303 -----
304 -----
305 -----
306 From Clock: wb_clk_i
307 To Clock:   wb_clk_i
308
309 Setup :      0 Failing Endpoints, Worst Slack 14.622ns, Total Violation 0.000ns
310 Hold  :      0 Failing Endpoints, Worst Slack 0.137ns, Total Violation 0.000ns
311 PW   :      0 Failing Endpoints, Worst Slack 11.250ns, Total Violation 0.000ns
312 -----

```

Timing slack is MET, no timing violation

•Simulation log



A terminal window titled "ubuntu@ubuntu2004: ~/Lab/Lab4_caravel_fir/lab-caravel_fir/..." displays simulation results. The output consists of 21 lines, each showing "data" followed by a number and the word "pass". The numbers range from 579 to 599. Below these, a summary line states "LA Test 2 passed, latency = 35935". The prompt "ubuntu@ubuntu2004:~/Lab/Lab4_caravel_fir/lab-caravel_fir/testbench/counter_la_fi" is visible, followed by a new line starting with "r\$".

```
ubuntu@ubuntu2004: ~/Lab/Lab4_caravel_fir/lab-caravel_fir/...
data      579 pass
data      580 pass
data      581 pass
data      582 pass
data      583 pass
data      584 pass
data      585 pass
data      586 pass
data      587 pass
data      588 pass
data      589 pass
data      590 pass
data      591 pass
data      592 pass
data      593 pass
data      594 pass
data      595 pass
data      596 pass
data      597 pass
data      598 pass
data      599 pass
LA Test 2 passed, latency =      35935
ubuntu@ubuntu2004:~/Lab/Lab4_caravel_fir/lab-caravel_fir/testbench/counter_la_fi
r$
```