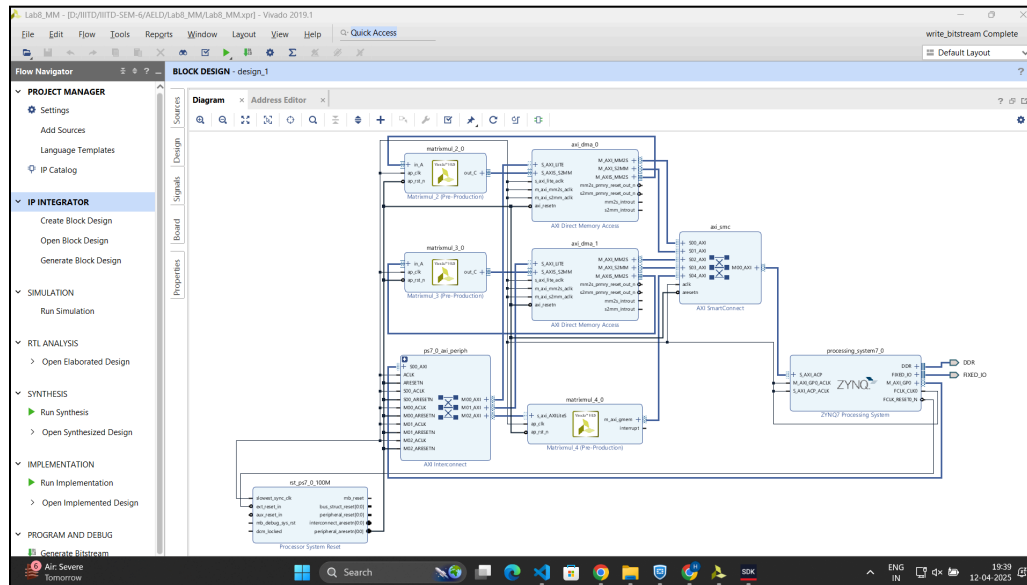


# Design of a Custom IP for 8x8 matrix multiplication

## Vivado Block Design



## Pragmas Used for Optimisations:

- **#pragma HLS PIPELINE:** Directive used in High-Level Synthesis (HLS) tools like Xilinx Vivado HLS to instruct the compiler to pipeline a specific section of code, typically a loop or function. Pipelining improves performance by allowing the overlapping execution of operations across multiple clock cycles. By Default, it tries to reduce the **initiation interval** of a loop to 1.
- **#pragma ARRAY\_PARTITION:** In hardware synthesis, arrays are typically mapped to memory structures like block RAMs (BRAMs). When multiple operations try to access the same memory in a single clock cycle, memory access conflicts arise, which can limit parallelism and pipelining.

Array partitioning removes this bottleneck by splitting the array into smaller, independently accessible memory blocks or registers, enabling parallel access to different parts of the array. This is essential for achieving high throughput in pipelined designs.

Example :

Matrix A[2][1] = [[1,2,3],[4,5,6]]

If we use **#pragma ARRAY\_PARTITION variable = A complete dim = 1**, then the array is split into two arrays, [1,2,3] and [4,5,6]. Now, each row can be mapped to a different BRAM or register block, allowing concurrent access to elements from different rows, which would not be possible with a single BRAM.

## Results

Method	Avg Execution Time (Micro-Seconds)
PS	26.074459
Pipelined MMUL in PL	31.892307
Pipelined + AP MMUL in PL	7.283692
Pipelined + AP + Memory Mapped MMUL in PL	6.415385

### Reasons :

The table shows average execution times for different methods of matrix multiplication (MMUL) on a processing system (PS) and programmable logic (PL). Here's an analysis of the reasons for the differences in execution times:

#### **PS (26.074459 $\mu$ s):**

Reason: The PS (Processing System) refers to a general-purpose processor (ARM Cortex-A9). It executes the matrix multiplication sequentially using software. While this method benefits from high clock speeds and optimized CPU instructions, it lacks parallelism, leading to a moderate execution time compared to PL-based methods.

#### **Pipelined MMUL in PL (31.892307 $\mu$ s):**

Reason: This method uses pipelining in the programmable logic (FPGA). Pipelining breaks down the matrix multiplication into stages, allowing concurrent processing of different parts of the computation. However, the execution time is higher than PS because the PL clock frequency is typically lower than the PS, and the overhead of setting up the pipeline (e.g., data transfer between PS and PL, pipeline stalls) may outweigh the benefits for the given matrix size or implementation.

#### **Pipelined + AP MMUL in PL (7.283692 $\mu$ s):**

Reason: Array Partitioning (AP) in the PL involves dividing arrays into smaller blocks that can be processed in parallel. This enables better resource utilisation on the FPGA, such as using multiple DSP slices or LUTs concurrently. Combined with pipelining, this parallelism significantly reduces execution time compared to both the PS and the basic pipelined method by allowing simultaneous computations on different array segments.

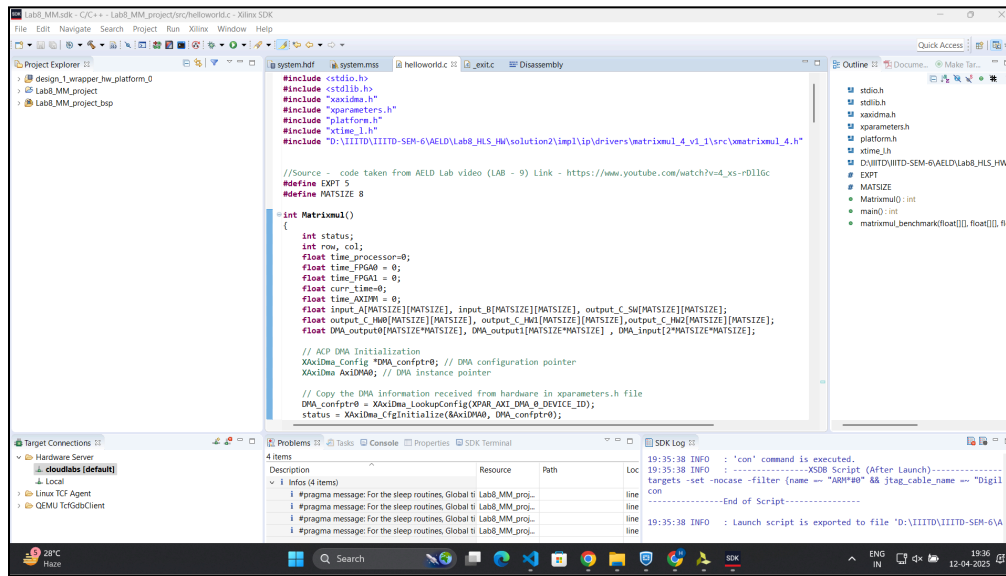
#### **Pipelined + AP + Memory Mapped MMUL in PL (6.415385 $\mu$ s):**

Reason: The addition of Memory Mapped maps the PL directly to memory (without using DMA). This reduces data transfer bottlenecks between the PS and PL, allowing faster access to matrix data. The slight improvement over the previous method (7.283692  $\mu$ s) is due to reduced latency in memory operations, making the overall process more efficient.

## Results:

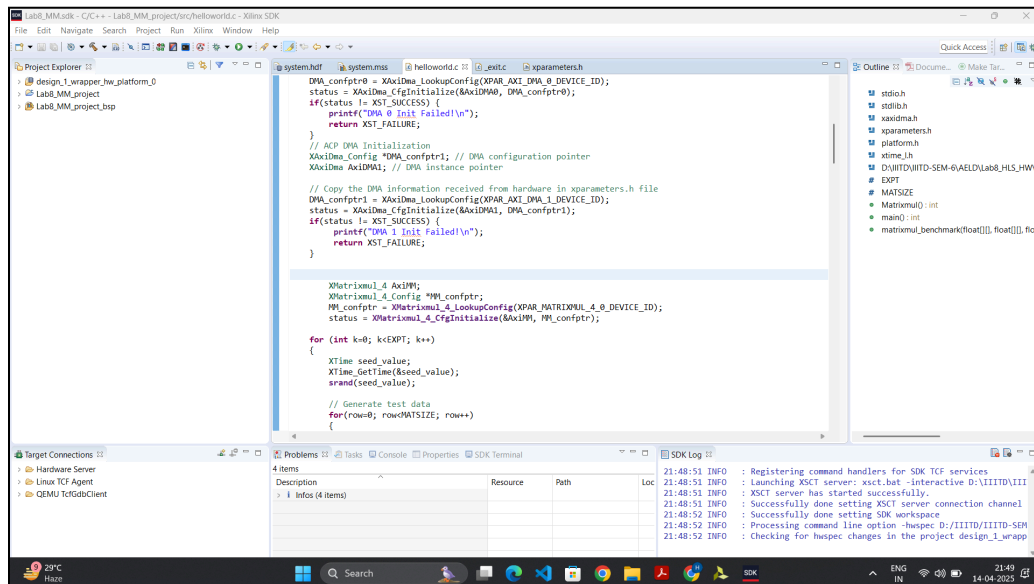
- Achieved 4 times speedup, reducing running time from 26.07 $\mu$ s to 6.41 $\mu$ s using memory-mapped interface.

## Xilinx SDK (Software Development Kit Code)



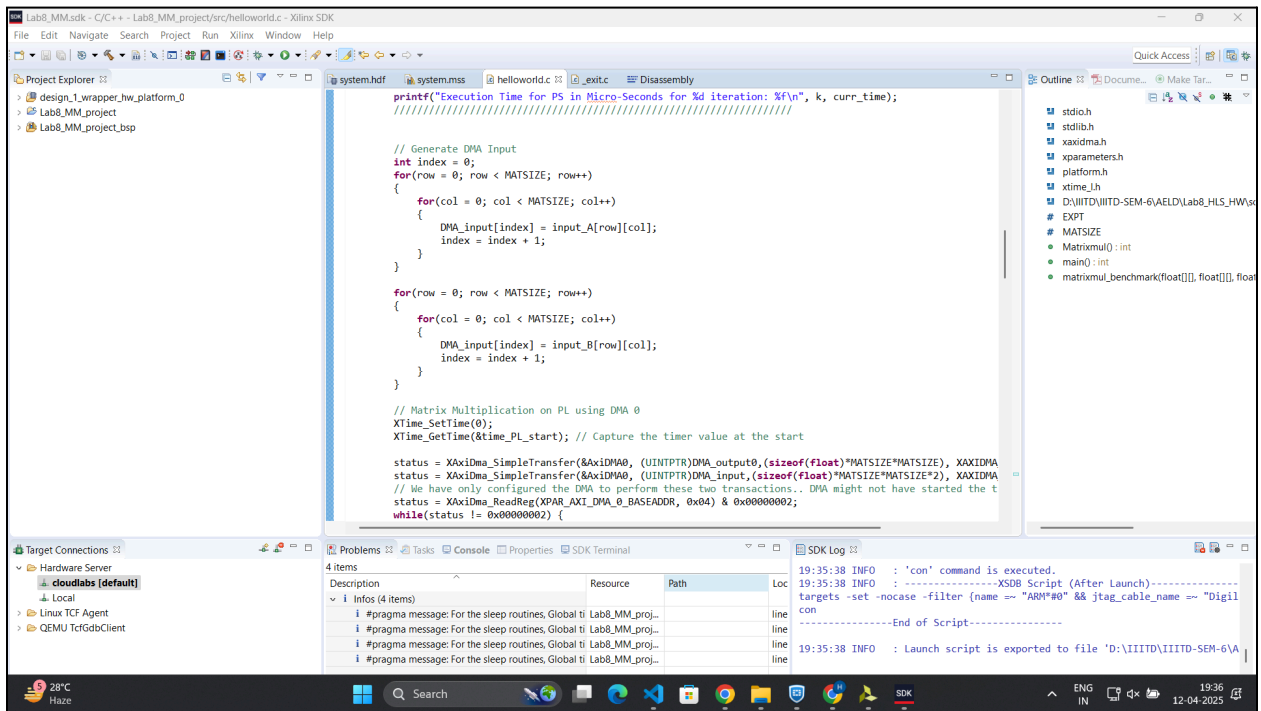
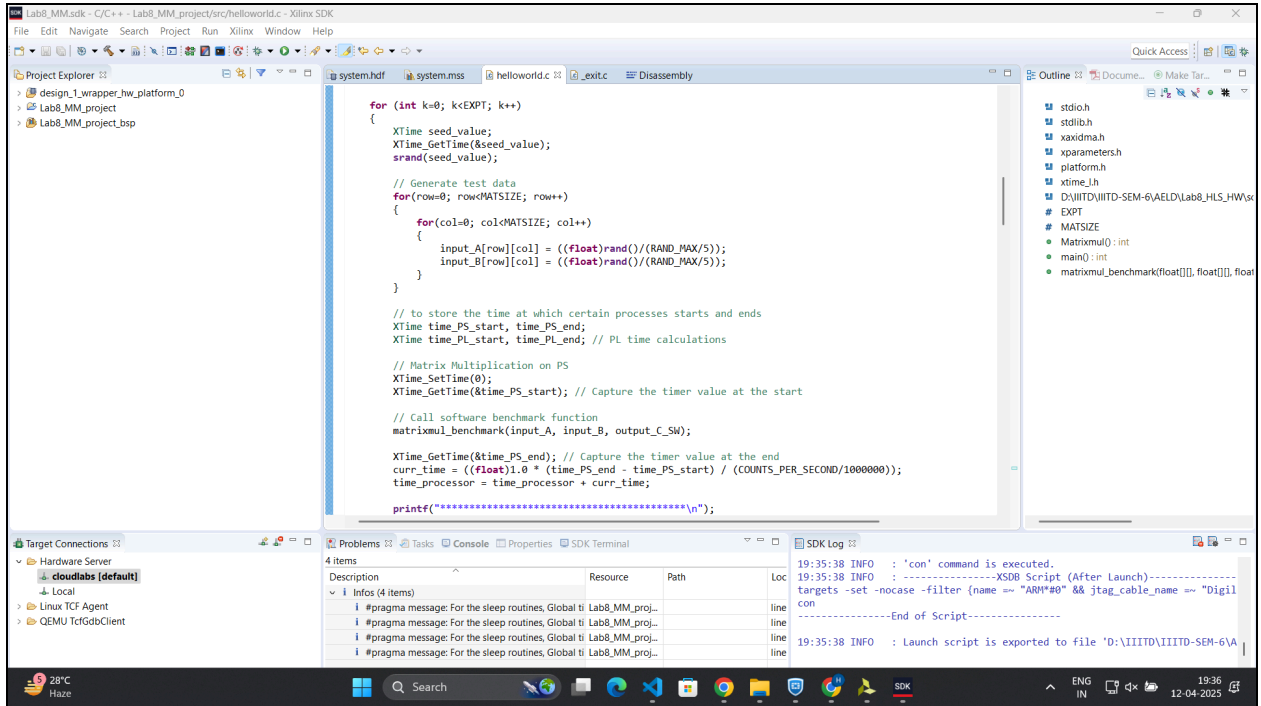
The screenshot shows the Xilinx SDK IDE with the following components:

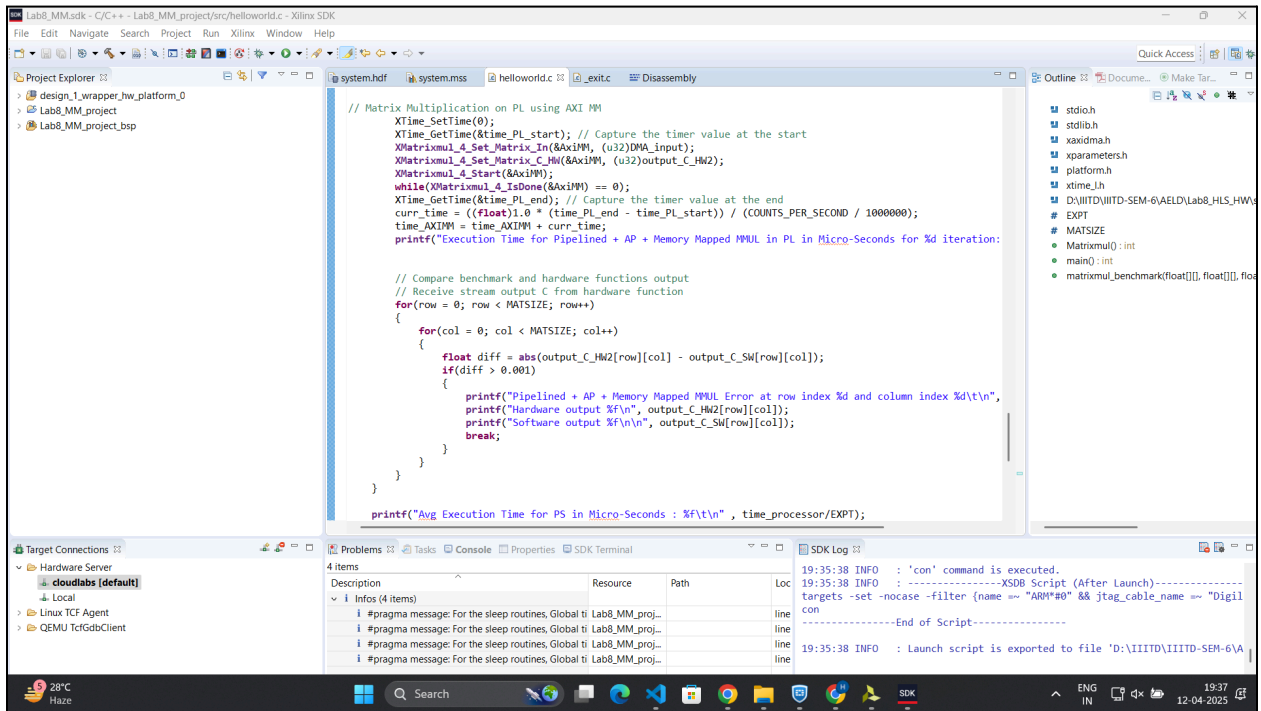
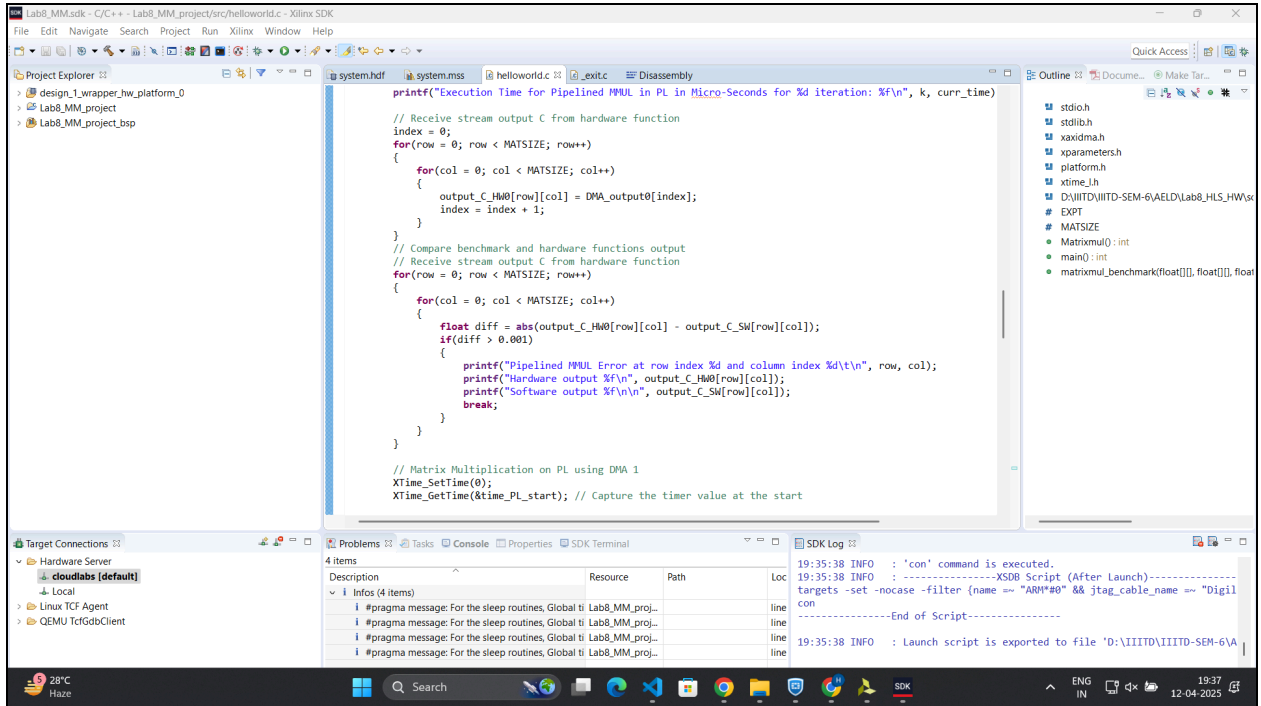
- Project Explorer:** Shows the project structure with files like `design_1_wrapper_hw_platform_0`, `Lab8_MM_project`, and `Lab8_MM_project_bsp`.
- Source Editor:** Displays the `helloworld.c` file. The code includes headers for `stdio.h`, `stdlib.h`, `xsaxidma.h`, `xparameters.h`, `platform.h`, `xtime.h`, and `xiitd\IIITD-SEM-6\AELD\Lab8_HLS_HW\src\matrixmul_4_v1_1\src\matrixmul_4.h`. It defines `EXPT` as 5 and `MATSIZE` as 8. The `int Matrixmul()` function is defined, which initializes DPA and AXIDMA, then performs a matrix multiplication benchmark.
- Outline:** Lists the files included in the project, including `stdio.h`, `stdlib.h`, `xsaxidma.h`, `xparameters.h`, `platform.h`, `xtime.h`, `xiitd\IIITD-SEM-6\AELD\Lab8_HLS_HW\src\matrixmul_4_v1_1\src\matrixmul_4.h`, `EXPT`, `MATSIZE`, `Matrixmul`, `main`, and `matrixmul_benchmark`.
- Problems:** Shows 4 items, including pragma messages for the sleep routines.
- SDK Log:** Shows the execution of the `con` command and the export of the launch script to `D:\IIITD\IIITD-SEM-6\AELD\Lab8_HLS_HW\src\matrixmul_4_v1_1\src\matrixmul_4.h`.

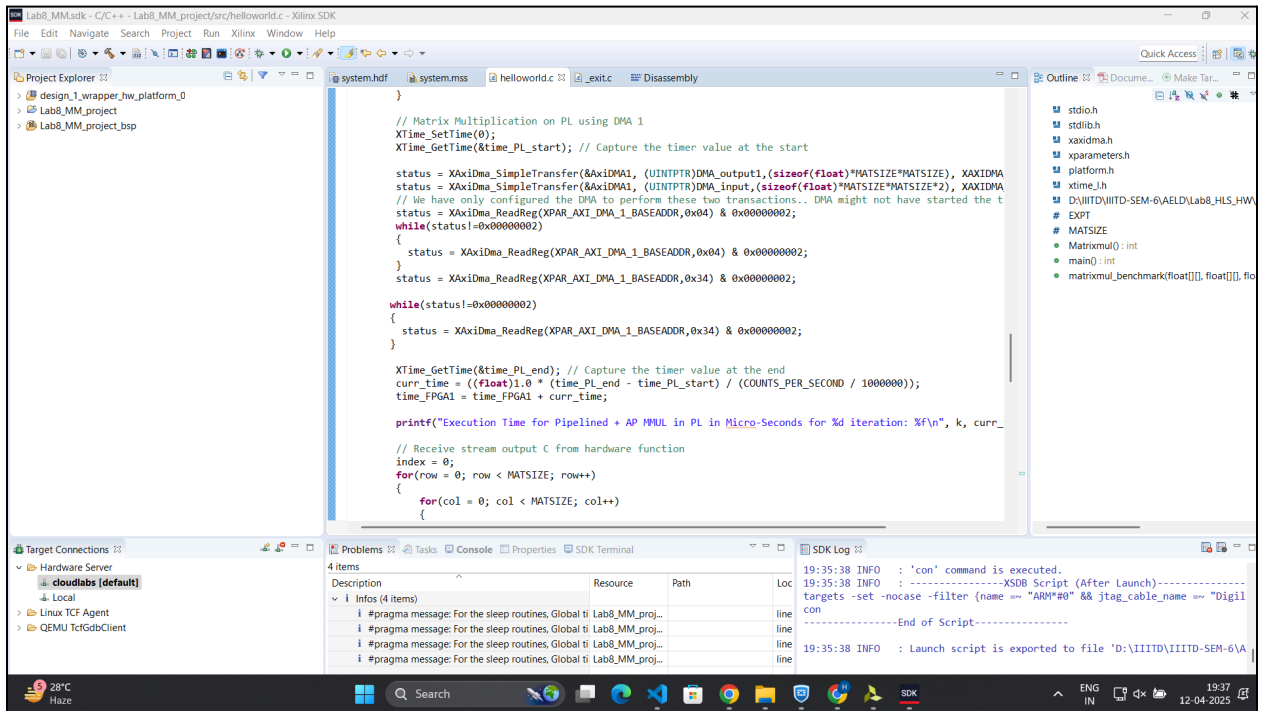
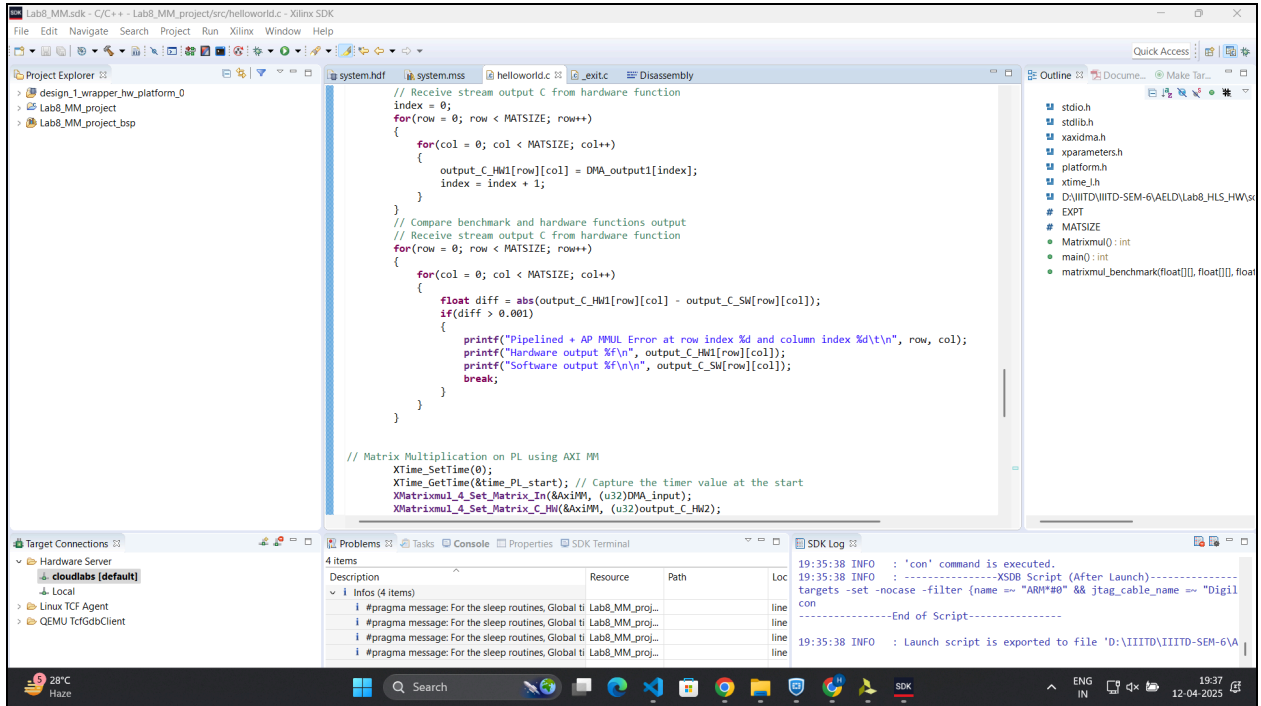


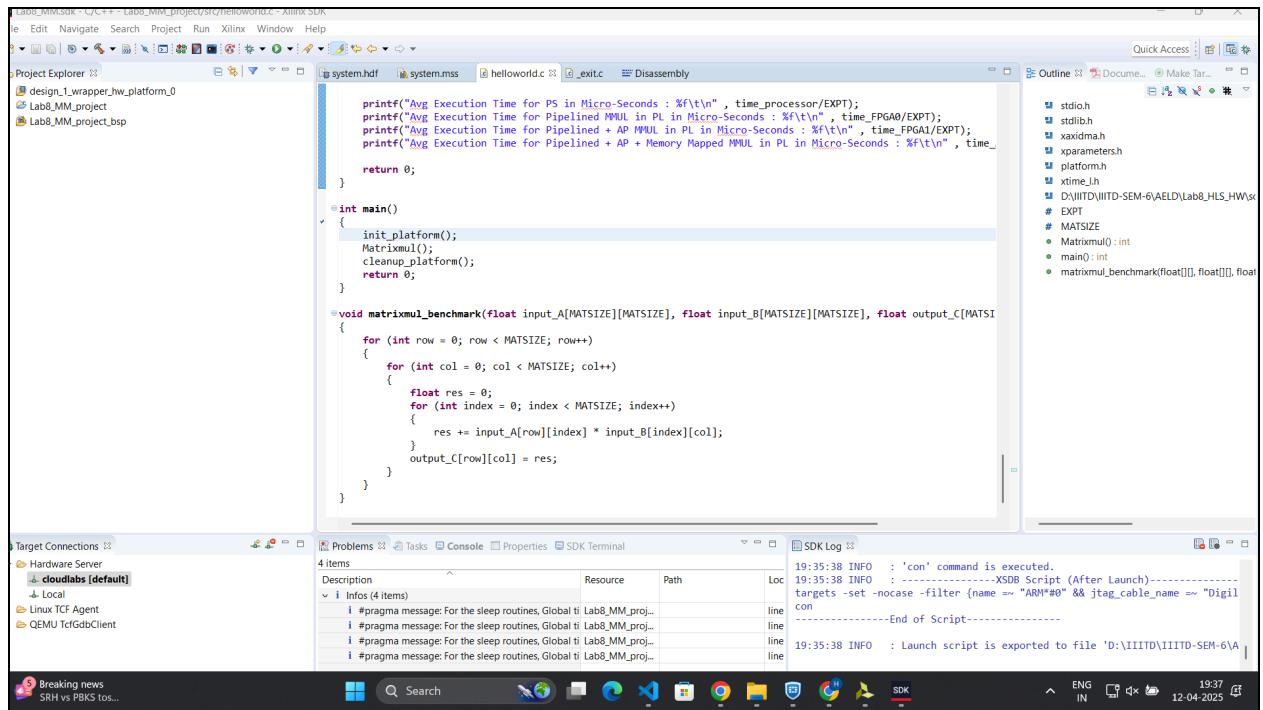
The screenshot shows the Xilinx SDK IDE with the following components:

- Project Explorer:** Shows the project structure with files like `design_1_wrapper_hw_platform_0`, `Lab8_MM_project`, and `Lab8_MM_project_bsp`.
- Source Editor:** Displays the `helloworld.c` file. The code includes headers for `stdio.h`, `stdlib.h`, `xsaxidma.h`, `xparameters.h`, `platform.h`, `xtime.h`, and `xiitd\IIITD-SEM-6\AELD\Lab8_HLS_HW\src\matrixmul_4_v1_1\src\matrixmul_4.h`. It defines `EXPT` as 5 and `MATSIZE` as 8. The `int Matrixmul()` function is defined, which initializes DPA and AXIDMA, then performs a matrix multiplication benchmark.
- Outline:** Lists the files included in the project, including `stdio.h`, `stdlib.h`, `xsaxidma.h`, `xparameters.h`, `platform.h`, `xtime.h`, `xiitd\IIITD-SEM-6\AELD\Lab8_HLS_HW\src\matrixmul_4_v1_1\src\matrixmul_4.h`, `EXPT`, `MATSIZE`, `Matrixmul`, `main`, and `matrixmul_benchmark`.
- Problems:** Shows 4 items, including pragma messages for the sleep routines.
- SDK Log:** Shows the execution of the `con` command and the export of the launch script to `D:\IIITD\IIITD-SEM-6\AELD\Lab8_HLS_HW\src\matrixmul_4_v1_1\src\matrixmul_4.h`.

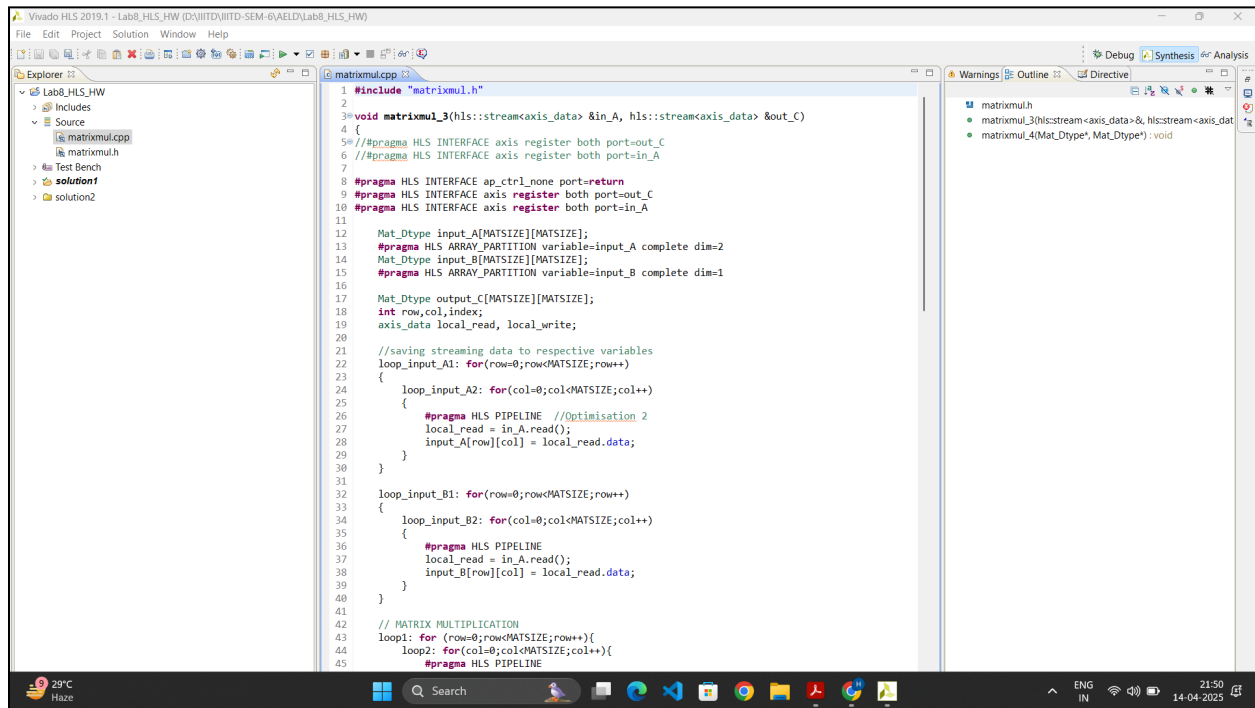




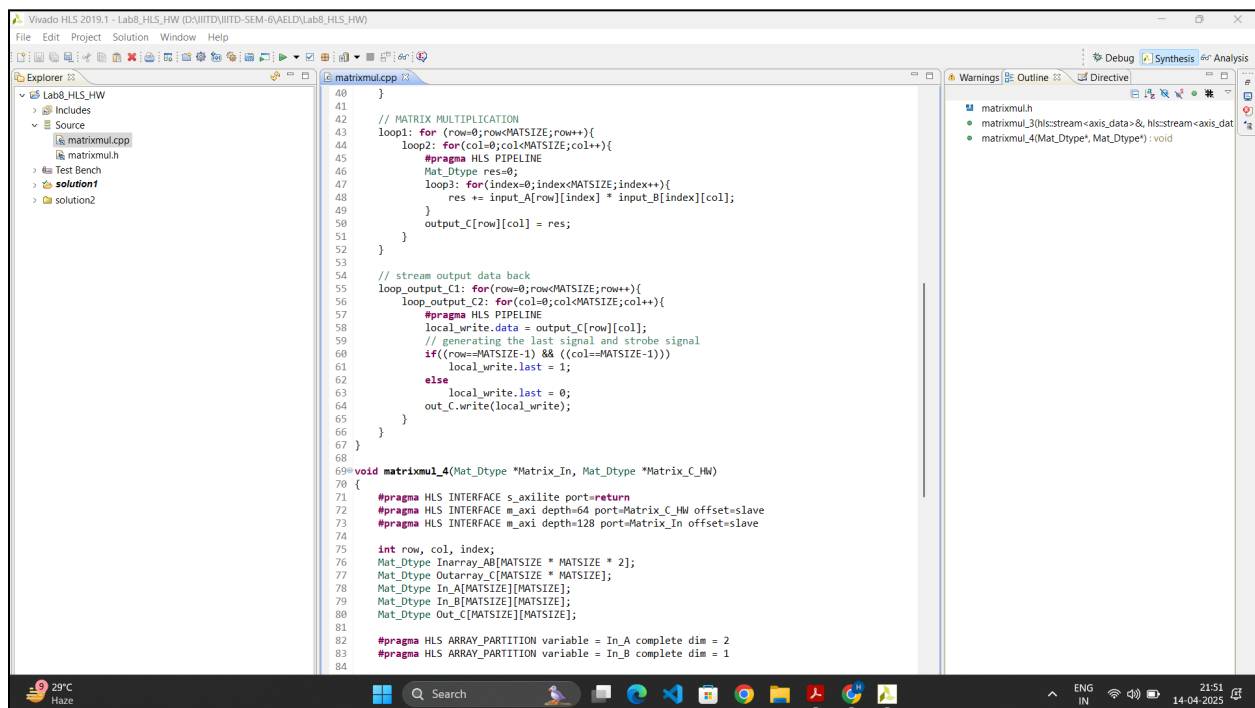




# Vivado HLS (high-level synthesis) C code

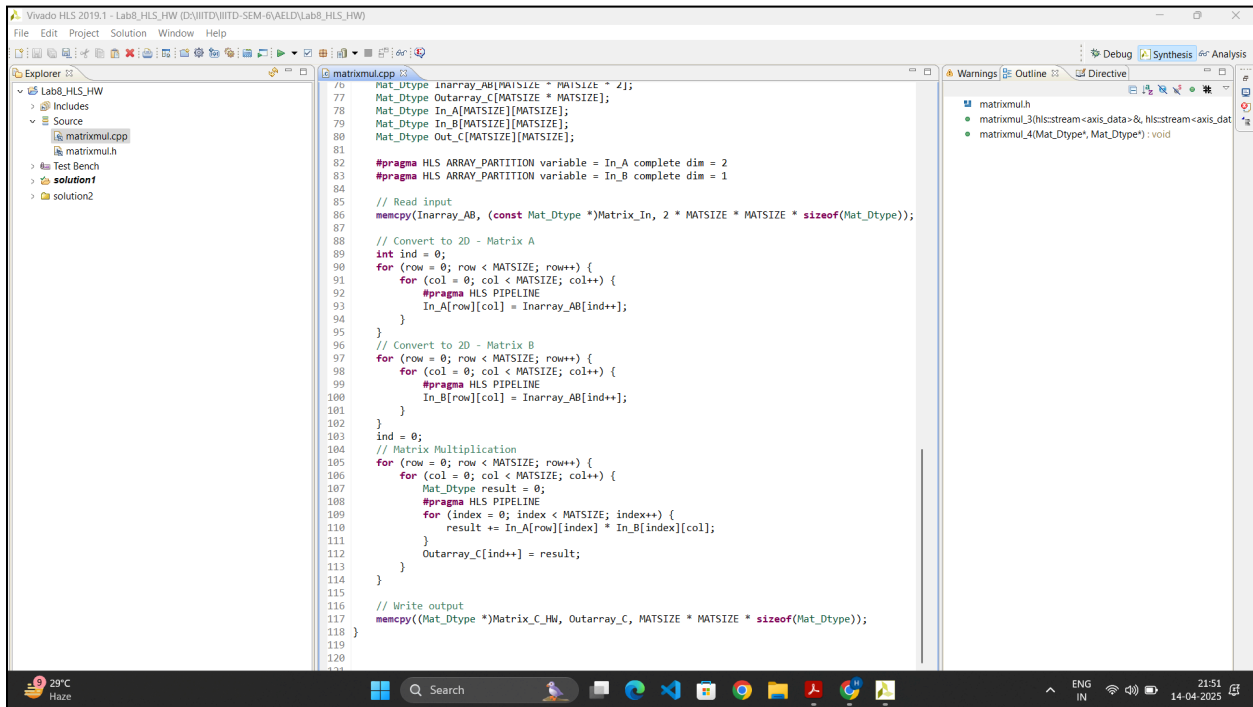
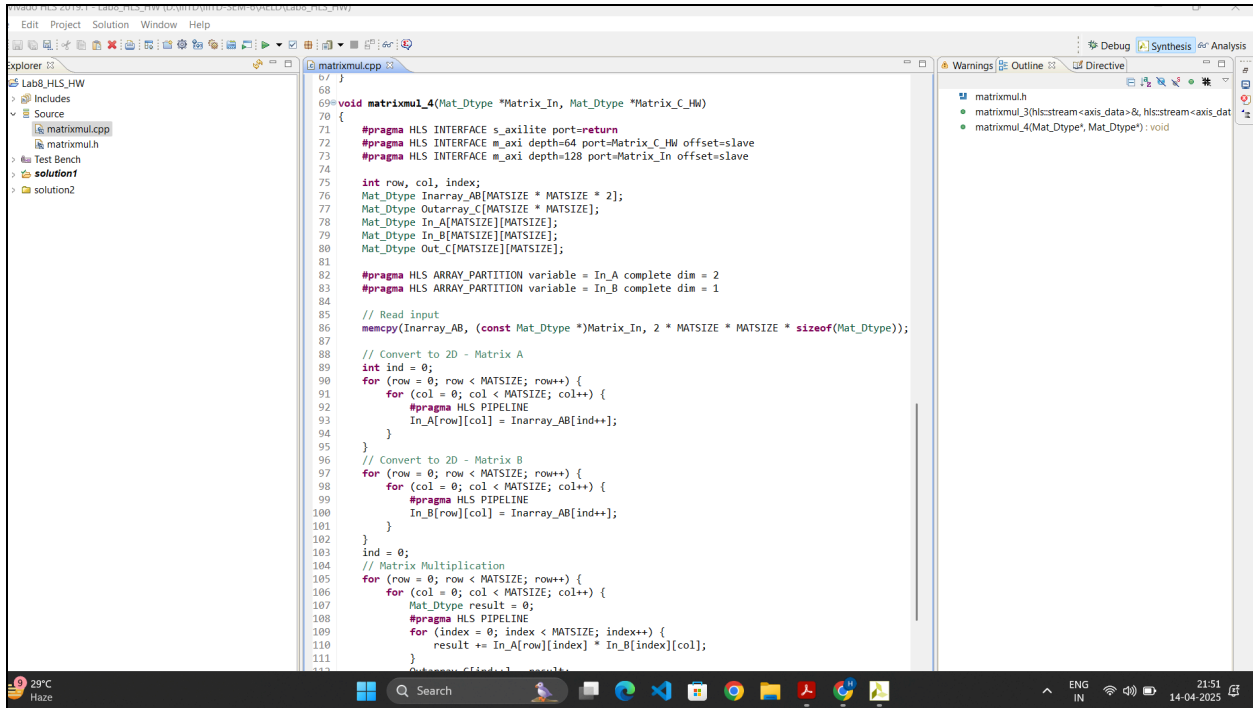


```
1 #include "matrixmul.h"
2
3 void matrixmul_3(hls::stream<axis_data> &in_A, hls::stream<axis_data> &out_C)
4 {
5     // #pragma HLS INTERFACE axis register both port=out_C
6     // #pragma HLS INTERFACE axis register both port=in_A
7
8     #pragma HLS INTERFACE ap_ctrl_none port=return
9     #pragma HLS INTERFACE axis register both port=out_C
10    #pragma HLS INTERFACE axis register both port=in_A
11
12    Mat_Dtype input_A[MATSIZE][MATSIZE];
13    #pragma HLS ARRAY_PARTITION variable=input_A complete dim=2
14    Mat_Dtype input_B[MATSIZE][MATSIZE];
15    #pragma HLS ARRAY_PARTITION variable=input_B complete dim=1
16
17    Mat_Dtype output_C[MATSIZE][MATSIZE];
18    int row,col,index;
19    axis_data local_read, local_write;
20
21    // saving streaming data to respective variables
22    loop_input_A1: for(row=0; row<MATSIZE; row++)
23    {
24        loop_input_A2: for(col=0; col<MATSIZE; col++)
25        {
26            #pragma HLS PIPELINE //Optimisation 2
27            local_read = in_A.read();
28            input_A[row][col] = local_read.data;
29        }
30    }
31
32    loop_input_B1: for(row=0; row<MATSIZE; row++)
33    {
34        loop_input_B2: for(col=0; col<MATSIZE; col++)
35        {
36            #pragma HLS PIPELINE
37            local_read = in_A.read();
38            input_B[row][col] = local_read.data;
39        }
40    }
41
42    // MATRIX MULTIPLICATION
43    loop1: for (row=0; row<MATSIZE; row++){
44        loop2: for(col=0; col<MATSIZE; col++){
45            #pragma HLS PIPELINE
46            Mat_Dtype res=0;
47            loop3: for(index=0; index<MATSIZE; index++){
48                res += input_A[row][index] * input_B[index][col];
49            }
50            output_C[row][col] = res;
51        }
52    }
53
54    // stream output data back
55    loop_output_C1: for(row=0; row<MATSIZE; row++){
56        loop_output_C2: for(col=0; col<MATSIZE; col++){
57            #pragma HLS PIPELINE
58            local_write.data = output_C[row][col];
59            // generating the last signal and strobe signal
60            if((row==MATSIZE-1) && (col==MATSIZE-1))
61                local_write.last = 1;
62            else
63                local_write.last = 0;
64            out_C.write(local_write);
65        }
66    }
67 }
68
69 void matrixmul_4(Mat_Dtype *Matrix_In, Mat_Dtype *Matrix_C_HW)
70 {
71     #pragma HLS INTERFACE s_axilite port=return
72     #pragma HLS INTERFACE m_axi depth=64 port=Matrix_C_HW offset=slave
73     #pragma HLS INTERFACE m_axi depth=128 port=Matrix_In offset=slave
74
75     int row, col, index;
76     Mat_Dtype Inarray_AB[MATSIZE * MATSIZE * 2];
77     Mat_Dtype Outarray_C[MATSIZE * MATSIZE];
78     Mat_Dtype In_A[MATSIZE][MATSIZE];
79     Mat_Dtype In_B[MATSIZE][MATSIZE];
80     Mat_Dtype Out_C[MATSIZE][MATSIZE];
81
82     #pragma HLS ARRAY_PARTITION variable = In_A complete dim = 2
83     #pragma HLS ARRAY_PARTITION variable = In_B complete dim = 1
84 }
```

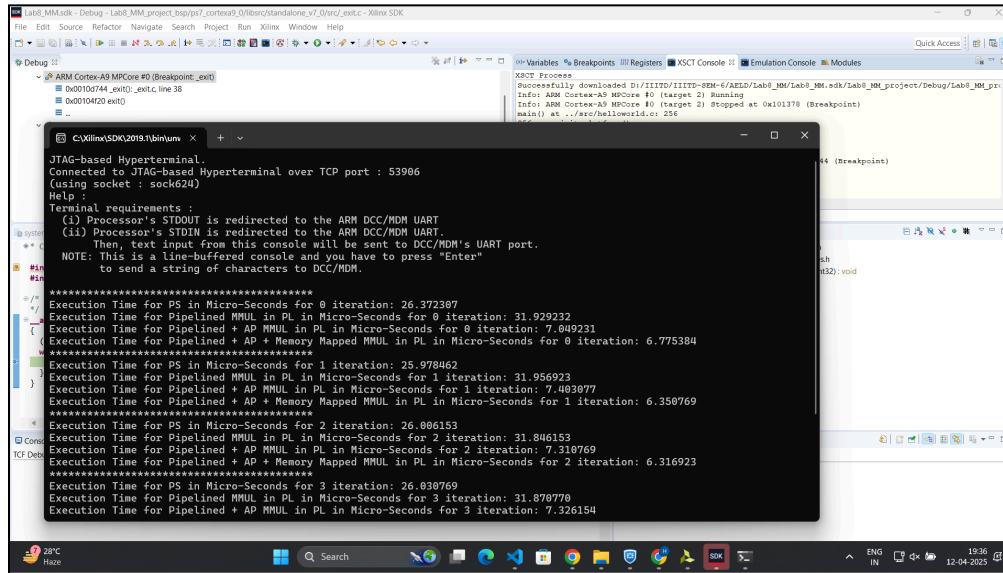


```
40 }
41
42 // MATRIX MULTIPLICATION
43 loop1: for (row=0; row<MATSIZE; row++){
44     loop2: for(col=0; col<MATSIZE; col++){
45         #pragma HLS PIPELINE
46         Mat_Dtype res=0;
47         loop3: for(index=0; index<MATSIZE; index++){
48             res += input_A[row][index] * input_B[index][col];
49         }
50         output_C[row][col] = res;
51     }
52 }
53
54 // stream output data back
55 loop_output_C1: for(row=0; row<MATSIZE; row++){
56     loop_output_C2: for(col=0; col<MATSIZE; col++){
57         #pragma HLS PIPELINE
58         local_write.data = output_C[row][col];
59         // generating the last signal and strobe signal
60         if((row==MATSIZE-1) && (col==MATSIZE-1))
61             local_write.last = 1;
62         else
63             local_write.last = 0;
64         out_C.write(local_write);
65     }
66 }
67 }
68
69 void matrixmul_4(Mat_Dtype *Matrix_In, Mat_Dtype *Matrix_C_HW)
70 {
71     #pragma HLS INTERFACE s_axilite port=return
72     #pragma HLS INTERFACE m_axi depth=64 port=Matrix_C_HW offset=slave
73     #pragma HLS INTERFACE m_axi depth=128 port=Matrix_In offset=slave
74
75     int row, col, index;
76     Mat_Dtype Inarray_AB[MATSIZE * MATSIZE * 2];
77     Mat_Dtype Outarray_C[MATSIZE * MATSIZE];
78     Mat_Dtype In_A[MATSIZE][MATSIZE];
79     Mat_Dtype In_B[MATSIZE][MATSIZE];
80     Mat_Dtype Out_C[MATSIZE][MATSIZE];
81
82     #pragma HLS ARRAY_PARTITION variable = In_A complete dim = 2
83     #pragma HLS ARRAY_PARTITION variable = In_B complete dim = 1
84 }
```





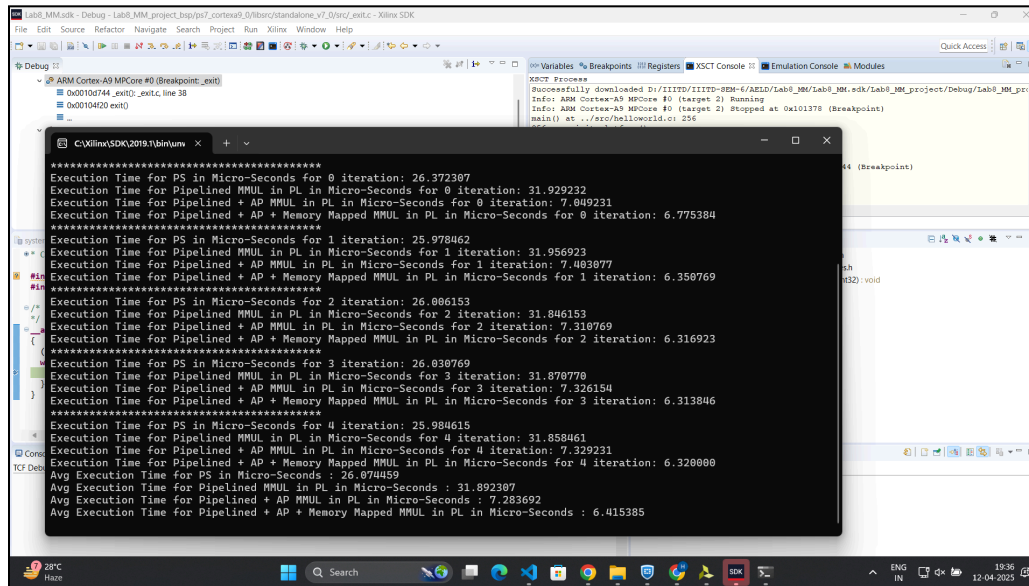
# Output [On JTAGTERMINAL]



The screenshot shows the JTAG Terminal window with the following output:

```
JTAG-based Hyperterminal.
Connected to JTAG-based Hyperterminal over TCP port : 53906
(using socket : sock624)
Help :
Terminal requirements :
(i) Processor's STDOUT is redirected to the ARM DCC/MDM UART.
(ii) Processor's STDIN is redirected to the ARM DCC/MDM UART.
Then, text input from this console will be sent to DCC/MDM's UART port.
NOTE: This is a line-buffered console and you have to press "Enter"
to send a string of characters to DCC/MDM.

*****
Execution Time for PS in Micro-Seconds for 0 iteration: 26.372307
Execution Time for Pipelined MMUL in PL in Micro-Seconds for 0 iteration: 31.929232
Execution Time for Pipelined + AP MMUL in PL in Micro-Seconds for 0 iteration: 7.049231
Execution Time for Pipelined + AP + Memory Mapped MMUL in PL in Micro-Seconds for 0 iteration: 6.775384
*****
Execution Time for PS in Micro-Seconds for 1 iteration: 25.978462
Execution Time for Pipelined MMUL in PL in Micro-Seconds for 1 iteration: 31.956923
Execution Time for Pipelined + AP MMUL in PL in Micro-Seconds for 1 iteration: 7.403077
Execution Time for Pipelined + AP + Memory Mapped MMUL in PL in Micro-Seconds for 1 iteration: 6.350769
*****
Execution Time for PS in Micro-Seconds for 2 iteration: 26.006153
Execution Time for Pipelined MMUL in PL in Micro-Seconds for 2 iteration: 31.846153
Execution Time for Pipelined + AP MMUL in PL in Micro-Seconds for 2 iteration: 7.310769
Execution Time for Pipelined + AP + Memory Mapped MMUL in PL in Micro-Seconds for 2 iteration: 6.316923
*****
Execution Time for PS in Micro-Seconds for 3 iteration: 26.030769
Execution Time for Pipelined MMUL in PL in Micro-Seconds for 3 iteration: 31.870770
Execution Time for Pipelined + AP MMUL in PL in Micro-Seconds for 3 iteration: 7.326154
```



The screenshot shows the JTAG Terminal window with the following output:

```
*****
Execution Time for PS in Micro-Seconds for 0 iteration: 26.372307
Execution Time for Pipelined MMUL in PL in Micro-Seconds for 0 iteration: 31.929232
Execution Time for Pipelined + AP MMUL in PL in Micro-Seconds for 0 iteration: 7.049231
Execution Time for Pipelined + AP + Memory Mapped MMUL in PL in Micro-Seconds for 0 iteration: 6.775384
*****
Execution Time for PS in Micro-Seconds for 1 iteration: 25.978462
Execution Time for Pipelined MMUL in PL in Micro-Seconds for 1 iteration: 31.956923
Execution Time for Pipelined + AP MMUL in PL in Micro-Seconds for 1 iteration: 7.403077
Execution Time for Pipelined + AP + Memory Mapped MMUL in PL in Micro-Seconds for 1 iteration: 6.350769
*****
Execution Time for PS in Micro-Seconds for 2 iteration: 26.006153
Execution Time for Pipelined MMUL in PL in Micro-Seconds for 2 iteration: 31.846153
Execution Time for Pipelined + AP MMUL in PL in Micro-Seconds for 2 iteration: 7.310769
Execution Time for Pipelined + AP + Memory Mapped MMUL in PL in Micro-Seconds for 2 iteration: 6.316923
*****
Execution Time for PS in Micro-Seconds for 3 iteration: 26.030769
Execution Time for Pipelined MMUL in PL in Micro-Seconds for 3 iteration: 31.870770
Execution Time for Pipelined + AP MMUL in PL in Micro-Seconds for 3 iteration: 7.326154
Execution Time for Pipelined + AP + Memory Mapped MMUL in PL in Micro-Seconds for 3 iteration: 6.313846
*****
Execution Time for PS in Micro-Seconds for 4 iteration: 25.984615
Execution Time for Pipelined MMUL in PL in Micro-Seconds for 4 iteration: 31.858461
Execution Time for Pipelined + AP MMUL in PL in Micro-Seconds for 4 iteration: 7.329231
Execution Time for Pipelined + AP + Memory Mapped MMUL in PL in Micro-Seconds for 4 iteration: 6.320000
Avg Execution Time for PS in Micro-Seconds : 26.074459
Avg Execution Time for Pipelined MMUL in PL in Micro-Seconds : 31.892307
Avg Execution Time for Pipelined + AP MMUL in PL in Micro-Seconds : 7.283692
Avg Execution Time for Pipelined + AP + Memory Mapped MMUL in PL in Micro-Seconds : 6.415385
```