

University of Calabria, DIMES
High Level Synthesis of Digital Systems
2023-2024
Prof.ssa PERRI
Prof. FRUSTACI
Sparse Matrix Vector Multiplication Analysis

Giorgio Ubbriaco
247284
bbrgrg00h11d086x@studenti.unical.it

June 2024

Index

| | | |
|----------|--|-----------|
| 1 | Introduction | 6 |
| 1.1 | Sparse Matrix | 6 |
| 1.2 | Compressed Row Storage (CRS) | 6 |
| 2 | Tasks to be performed | 7 |
| 3 | Definitions | 8 |
| 4 | C Simulations | 9 |
| 5 | Solutions | 11 |
| 5.1 | Solution 1 | 12 |
| 5.2 | Solution 2 | 15 |
| 5.3 | Solution 3 | 17 |
| 5.4 | Solution 4 | 19 |
| 5.5 | Solution 5 | 22 |
| 5.6 | Solution 6 | 25 |
| 5.7 | Solution 7 | 29 |
| 5.8 | Solution 8 | 33 |
| 5.9 | Solution 9 | 36 |
| 5.10 | Solution 10 | 44 |
| 5.11 | Solution 11 | 49 |
| 6 | Conclusions | 53 |

Listings

| | |
|--|----|
| definitions/definitions.h | 8 |
| c_simulations/smvmtb.cpp | 9 |
| c_simulations/smvmtb_output.cpp | 10 |
| c_simulations/stdm.cpp | 10 |
| solutions/s1/s1.cpp | 12 |
| solutions/s1/s1tripcount.cpp | 13 |
| solutions/s2/s2.cpp | 15 |
| solutions/s3/s3.cpp | 17 |
| solutions/s4/s4.cpp | 19 |
| solutions/s5/s5.cpp | 22 |
| solutions/s6/s6all3.cpp | 25 |
| solutions/s6/s6columnIndex.cpp | 26 |
| solutions/s6/s6values.cpp | 26 |
| solutions/s6/s6x.cpp | 26 |
| solutions/s7/s7.cpp | 29 |
| solutions/s7/s7columnIndex.cpp | 29 |
| solutions/s7/s7values.cpp | 30 |
| solutions/s7/s7x.cpp | 31 |
| solutions/s8/s8all3.cpp | 33 |
| solutions/s8/s8columnIndex.cpp | 33 |
| solutions/s8/s8values.cpp | 33 |
| solutions/s8/s8x.cpp | 34 |
| solutions/s9/s9.cpp | 36 |
| solutions/s9/s9columnIndex.cpp | 36 |
| solutions/s9/s9values.cpp | 37 |
| solutions/s9/s9x.cpp | 38 |
| solutions/s9/s9rowPtr4.cpp | 38 |
| solutions/s9/s9y4.cpp | 39 |
| solutions/s9/s9pipelineLoop1.cpp | 40 |
| solutions/s9/s9pipelineLoop1rowPtr.cpp | 41 |
| solutions/s10/s10all3.cpp | 44 |
| solutions/s10/s10columnIndex.cpp | 44 |
| solutions/s10/s10values.cpp | 44 |
| solutions/s10/s10x.cpp | 45 |
| solutions/s10/headermodified.h | 45 |
| solutions/s10/s10all3modified.cpp | 45 |
| solutions/s10/s10columnIndexmodified.cpp | 46 |
| solutions/s10/s10valuesmodified.cpp | 46 |
| solutions/s10/s10xmodified.cpp | 46 |
| solutions/s11/definitions.h | 49 |
| solutions/s11/s11all3.cpp | 49 |
| solutions/s11/s11columnIndex.cpp | 49 |
| solutions/s11/s11values.cpp | 50 |
| solutions/s11/s11x.cpp | 50 |
| solutions/s11/headermodified.h | 50 |

List of Figures

| | | |
|---|-----------------------------------|----|
| 1 | HLS Solution 1 Analysis | 13 |
| 2 | HLS Loop Pipelining | 15 |
| 3 | HLS Loop Unrolling | 19 |
| 4 | HLS Solution 4 Analysis | 20 |

| | | |
|---|---|----|
| 5 | HLS Solution 5 Analysis | 23 |
| 6 | HLS Array Partitioning | 25 |
| 7 | HLS Solution 10 Analysis | 47 |
| 8 | HLS Conclusions C/RTL Cosimulation Plot | 53 |
| 9 | Utilization Export RTL Plot | 56 |

List of Tables

| | | |
|----|--|----|
| 1 | SMVM Solutions To Be Performed | 7 |
| 2 | Detailed SMVM Solutions To Be Performed | 11 |
| 3 | HLS Solution 1 without Trip Count Timing Summary (ns) | 12 |
| 4 | HLS Solution 1 without Trip Count Latency Summary (clock cycles) | 12 |
| 5 | HLS Solution 1 without Trip Count Latency Loops Summary | 12 |
| 6 | HLS Solution 1 with Trip Count Timing Summary (ns) | 13 |
| 7 | HLS Solution 1 with Trip Count Latency Summary (clock cycles) | 13 |
| 8 | HLS Solution 1 Latency with Trip Count Loops Summary | 13 |
| 9 | HLS Solution 1 with Trip Count Utilization Estimates Summary | 14 |
| 10 | HLS Solution 1 with Trip Count C/RTL Cosimulation Summary | 14 |
| 11 | HLS Solution 1 with Trip Count Export RTL Resource Usage | 14 |
| 12 | HLS Solution 1 with Trip Count Export RTL Final Timing | 14 |
| 13 | HLS Solution 2 Timing Summary (ns) | 15 |
| 14 | HLS Solution 2 Latency Summary (clock cycles) | 15 |
| 15 | HLS Solution 2 Latency Loops Summary | 15 |
| 16 | HLS Solution 2 Utilization Estimates Summary | 16 |
| 17 | HLS Solution 1 with Trip Count C/RTL Cosimulation Summary | 16 |
| 18 | HLS Solution 2t Export RTL Resource Usage | 16 |
| 19 | HLS Solution 2 Export RTL Final Timing | 16 |
| 20 | HLS Solution 3 Timing Summary (ns) | 17 |
| 21 | HLS Solution 3 Latency Summary (clock cycles) | 17 |
| 22 | HLS Solution 3 Latency Loops Summary | 17 |
| 23 | HLS Solution 3 Utilization Estimates Summary | 18 |
| 24 | HLS Solution 3 C/RTL Cosimulation Summary | 18 |
| 25 | HLS Solution 3 Export RTL Resource Usage | 18 |
| 26 | HLS Solution 3 Export RTL Final Timing | 18 |
| 27 | HLS Solution 4 Timing Summary (ns) | 19 |
| 28 | HLS Solution 4 Latency Summary (clock cycles) | 19 |
| 29 | HLS Solution 4 Latency Loops Summary | 20 |
| 30 | HLS Solution 4 Utilization Estimates Summary | 20 |
| 31 | HLS Solution 4 C/RTL Cosimulation Summary | 21 |
| 32 | HLS Solution 4 Export RTL Resource Usage | 21 |
| 33 | HLS Solution 4 Export RTL Final Timing | 21 |
| 34 | HLS Solution 5 Timing Summary (ns) | 22 |
| 35 | HLS Solution 5 Latency Summary (clock cycles) | 22 |
| 36 | HLS Solution 5 Latency Loops Summary | 22 |
| 37 | HLS Solution 5 Utilization Estimates Summary | 23 |
| 38 | HLS Solution 5 C/RTL Cosimulation Summary | 23 |
| 39 | HLS Solution 5 Export RTL Resource Usage | 24 |
| 40 | HLS Solution 5 Export RTL Final Timing | 24 |
| 41 | HLS Solution 6 Timing Summary (ns) | 27 |
| 42 | HLS Solution 6 Latency Summary (clock cycles) | 27 |
| 43 | HLS Solution 6 Latency Loops Summary | 27 |
| 44 | HLS Solution 6 Utilization Estimates [#] | 27 |
| 45 | HLS Solution 6 C/RTL Cosimulation Report | 28 |
| 46 | HLS Solution 6 Export RTL Report | 28 |

| | | |
|----|---|----|
| 47 | HLS Solution 7 Timing Summary (ns) | 29 |
| 48 | HLS Solution 7 Latency Summary (clock cycles) | 29 |
| 49 | HLS Solution 7 Latency Loops Summary | 29 |
| 50 | HLS Solution 7 with columnIndex partitioning Latency Loops Summary | 30 |
| 51 | HLS Solution 7 with columnIndex and values partitioning Latency Loops Summary | 30 |
| 52 | HLS Solution 7 with columnIndex, values and x partitioning Timing Summary (ns) | 31 |
| 53 | HLS Solution 7 with columnIndex, values and x partitioning Latency Summary (clock cycles) | 31 |
| 54 | HLS Solution 7 with columnIndex, values and x partitioning Latency Loops Summary | 31 |
| 55 | HLS Solution 7 with columnIndex, values and x partitioning Utilization Estimates Summary | 32 |
| 56 | HLS Solution 7 with columnIndex, values and x partitioning C/RTL Cosimulation Summary | 32 |
| 57 | HLS Solution 7 with columnIndex, values and x partitioning Export RTL Resource Usage | 32 |
| 58 | HLS Solution 7 with columnIndex, values and x partitioning Export RTL Final Timing | 32 |
| 59 | HLS Solution 8 Timing Summary (ns) | 34 |
| 60 | HLS Solution 8 Latency Summary (clock cycles) | 34 |
| 61 | HLS Solution 8 Latency Loops Summary | 34 |
| 62 | HLS Solution 8 Utilization Estimates [#] | 35 |
| 63 | HLS Solution 8 C/RTL Cosimulation Report | 35 |
| 64 | HLS Solution 8 Export RTL Report | 35 |
| 65 | HLS Solution 9 Timing Summary (ns) | 36 |
| 66 | HLS Solution 9 Latency Summary (clock cycles) | 36 |
| 67 | HLS Solution 9 Latency Loops Summary | 36 |
| 68 | HLS Solution 9 with columnIndex partitioning Latency Loops Summary | 37 |
| 69 | HLS Solution 9 with columnIndex and values partitioning Latency Loops Summary | 37 |
| 70 | HLS Solution 9 with columnIndex, values and x partitioning Latency Loops Summary | 38 |
| 71 | HLS Solution 9 with columnIndex, values, x and rowPtr partitioning Latency Loops Summary | 39 |
| 72 | HLS Solution 9 with columnIndex, values, x, rowPtr and y partitioning Latency Loops Summary | 40 |
| 73 | HLS Solution 9 with columnIndex, values and x partitioning and loop1 pipelined Latency Loops Summary | 40 |
| 74 | HLS Solution 9 with columnIndex, values and x partitioning and loop1 pipelined Utilization Estimates Summary | 41 |
| 75 | HLS Solution 9 with columnIndex, values and x partitioning and loop1 pipelined C/RTL Cosimulation Summary | 41 |
| 76 | HLS Solution 9 with columnIndex, values and x partitioning and loop1 pipelined Export RTL Resource Usage | 41 |
| 77 | HLS Solution 9 with columnIndex, values and x partitioning and loop1 pipelined Export RTL Final Timing | 41 |
| 78 | HLS Solution 9 with columnIndex, values, x, rowPtr and y partitioning and loop1 pipelined Latency Loops Summary | 42 |
| 79 | HLS Solution 9 with columnIndex, values, x, rowPtr and y partitioning and loop1 pipelined Utilization Estimates Summary | 42 |
| 80 | HLS Solution 9 with columnIndex, values, x, rowPtr and y partitioning and loop1 pipelined C/RTL Cosimulation Summary | 42 |
| 81 | HLS Solution 9 with columnIndex, values, x, rowPtr and y partitioning and loop1 pipelined Export RTL Resource Usage | 43 |
| 82 | HLS Solution 9 with columnIndex, values, x, rowPtr and y partitioning and loop1 pipelined Export RTL Final Timing | 43 |
| 83 | HLS Solution 10 Timing Summary (ns) | 46 |
| 84 | HLS Solution 10 Latency Summary (clock cycles) | 47 |
| 85 | HLS Solution 10 Latency Loops Summary | 47 |
| 86 | HLS Solution 10 Utilization Estimates [#] | 47 |
| 87 | HLS Solution 10 C/RTL Cosimulation Report | 48 |
| 88 | HLS Solution 10 Export RTL Report | 48 |
| 89 | HLS Solution 11 Timing Summary (ns) | 51 |
| 90 | HLS Solution 11 Latency Summary (clock cycles) | 51 |

| | | |
|----|---|----|
| 91 | HLS Solution 11 Latency Loops Summary | 51 |
| 92 | HLS Solution 11 Utilization Estimates [#] | 51 |
| 93 | HLS Solution 11 C/RTL Cosimulation Report | 52 |
| 94 | HLS Solution 11 Export RTL Report | 52 |
| 95 | HLS Conclusions C/RTL Cosimulation Report | 54 |
| 96 | HLS Conclusions Export RTL Report | 55 |

1 Introduction

1.1 Sparse Matrix

Nell'analisi numerica, una **matrice sparsa** è una matrice in cui la maggior parte degli elementi presenta un valore nullo. Non esiste una definizione rigorosa della proporzione di elementi a valore nullo affinché una matrice possa essere considerata sparsa. Al contrario, se la maggior parte degli elementi è non nulla, allora la matrice è considerata densa.

Una matrice è tipicamente memorizzata come un array bidimensionale. Ogni voce della matrice rappresenta un elemento $a_{i,j}$ e vi si accede tramite l'indice di riga i e l'indice di colonna j . Per una matrice $m \times n$, la quantità di memoria necessaria per memorizzare la matrice in questo formato è proporzionale a $m \times n$ (senza considerare che è necessario memorizzare anche le dimensioni relative alla matrice).

Nel caso di una matrice sparsa, è possibile ridurre notevolmente i requisiti di memoria memorizzando solo le voci non nulle. A seconda del numero e della distribuzione delle voci non nulle, è possibile utilizzare diverse strutture di dati che consentono di ottenere enormi risparmi di memoria rispetto all'approccio di base. Il compromesso è che l'accesso ai singoli elementi diventa più complesso e sono necessarie strutture aggiuntive per poter recuperare la matrice originale senza ambiguità.

I formati possono essere divisi in due gruppi:

- Quelli che supportano una modifica efficiente, come DOK (Dictionary of Keys), LIL (List of Lists) o COO (Coordinate List), utilizzati solitamente per la costruzione della matrice.
- Quelli che supportano l'accesso e le operazioni matriciali efficienti, come CRS (Compressed Row Storage) o CCS (Compressed Column Storage).

1.2 Compressed Row Storage (CRS)

Il formato **Compressed Row Storage (CRS)** permette la rappresentazione di una matrice tramite tre array unidimensionali consentendo un accesso veloce alle righe e una moltiplicazione matrice-vettore efficiente. In particolare, i tre array utilizzati sono i seguenti:

- **values**
È un array contenente tutti gli elementi della matrice non nulli.
- **rowPtr**
È un array contenente gli indici, relativi all'array **values**, corrispondenti ai primi elementi non nulli di ogni riga.
- **columnIndex**
È un array contenente gli indici di colonna degli elementi non nulli.

2 Tasks to be performed

Prendendo come riferimento il formato CRS per il calcolo del prodotto tra una matrice sparsa ed un vettore e considerando il tool di sintesi ad alto livello per sistemi digitali, fornito da Xilinx[®], analizzare le soluzioni proposte nella seguente tabella utilizzando le direttive proprietarie citate e caratterizzando in termini di latenza e utilizzazione delle risorse.

| Solution | Loop1 | Loop2 |
|-----------------|--------------|------------------------------|
| 1 | - | - |
| 2 | - | Pipeline |
| 3 | Pipeline | - |
| 4 | Unroll=2 | - |
| 5 | - | Pipeline, Unroll=2 |
| 6 | - | Pipeline, Unroll=2, Cyclic=2 |
| 7 | - | Pipeline, Unroll=4 |
| 8 | - | Pipeline, Unroll=2, Cyclic=4 |
| 9 | - | Pipeline, Unroll=8 |
| 10 | - | Pipeline, Unroll=2, Cyclic=8 |
| 11 | - | Pipeline, Unroll=2, Block=8 |

Table 1: SMVM Solutions To Be Performed

3 Definitions

Qui di seguito vengono riportate le definizioni e le intestazioni dei metodi corrispondenti alle soluzioni implementate per la moltiplicazione tra una matrice sparsa e un vettore. In particolare, ogni definizione presenta la documentazione associata. Inoltre, è stata prevista l'implementazione per la moltiplicazione standard così da poter verificare i risultati ottenuti tramite formato CRS.

```
1 #ifndef DEFINITIONS_H
2
3
4 /**
5  * Square Matrix Size.
6  */
7 const static int size = 4;
8
9 /**
10 * Number of Non-Zero Elements.
11 */
12 const static int nnz = 9;
13
14 /**
15 * Number of Rows.
16 */
17 const static int rows = 4;
18
19 /**
20 * Data Type.
21 */
22 typedef int DTYPE;
23
24 /**
25 * Matrix Vector Standard Multiplication Design.
26 * @param matrix[size][size] Input matrix
27 * @param y Multiplication Result
28 * @param x Input Vector
29 */
30 void std_multiplication(DTYPE matrix[size][size], DTYPE *y, DTYPE *x);
31
32 /**
33 * Sparse Matrix Vector Multiplication Design (CRS format).
34 * @param rowPtr[rows+1] Indexes First Elements
35 * @param columnIndex[nnz] Indexes Non Zero Elements
36 * @param values[nnz] Input Values
37 * @param y[size] Multiplication Result
38 * @param x[size] Input Vector
39 */
40 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
    x[size]);
41
42
43 #endif
```


4 C Simulations

Qui di seguito viene riportato il file testbench per la C Simulation in HLS e il corrispondente output ottenuto. In particolare, qui di seguito verrà riportato il caso in cui venga scelta la prima configurazione con matrice di dimensione 4×4 . Le altre configurazioni, relative ad alcune solution implementate, verranno presentate nei paragrafi successivi.

```
1 #include "definitions.h"
2 #include <iostream>
3 using namespace std;
4
5 void std_multiplication(DTYPE matrix[size][size], DTYPE *y, DTYPE *x);
6 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
    x[size]);
7
8
9 int main() {
10     int fail = 0;
11
12     DTYPE matrix[size][size] = {
13         {3,4,0,0},
14         {0,5,9,0},
15         {2,0,3,1},
16         {0,4,0,6}
17     };
18     DTYPE x[size] = {1, 1, 1, 1};
19     DTYPE values[] = {3, 4, 5, 9, 2, 3, 1, 4, 6};
20     int columnIndex[] = {0, 1, 1, 2, 0, 2, 3, 1, 3};
21     int rowPtr[] = {0,2,4,7,9};
22
23     /*
24     DTYPE matrix[size][size] = {
25         {3, 4, 0, 0, 0, 0, 0, 0, 0},
26         {0, 5, 9, 0, 0, 0, 0, 0, 0},
27         {2, 0, 3, 1, 0, 0, 0, 0, 0},
28         {0, 4, 0, 6, 0, 0, 0, 0, 0},
29         {0, 0, 0, 0, 0, 0, 0, 0, 0},
30         {0, 0, 0, 0, 0, 0, 0, 0, 0},
31         {0, 0, 0, 0, 0, 0, 0, 0, 0},
32         {0, 0, 0, 0, 0, 0, 0, 0, 0}
33     };
34     DTYPE x[size] = {1, 1, 1, 1, 1, 1, 1, 1, 1};
35     DTYPE values[] = {3, 4, 5, 9, 2, 3, 1, 4, 6};
36     int columnIndex[] = {0, 1, 1, 2, 0, 2, 3, 1, 3};
37     int rowPtr[] = {0, 2, 4, 7, 9, 9, 9, 9, 9};
38     */
39     /*
40     DTYPE matrix[size][size] = {
41         {3, 4, 0, 0, 0, 0, 0, 0, 0},
42         {0, 5, 9, 0, 0, 0, 0, 0, 0},
43         {2, 0, 3, 1, 0, 0, 0, 0, 0},
44         {0, 4, 0, 6, 0, 0, 0, 0, 0},
45         {1, 0, 0, 0, 0, 0, 0, 0, 0},
46         {1, 1, 0, 0, 0, 0, 0, 0, 0},
47         {1, 1, 0, 0, 0, 0, 0, 0, 0},
48         {1, 1, 0, 0, 0, 0, 0, 0, 0}
49     };
50     DTYPE x[size] = {1, 1, 1, 1, 1, 1, 1, 1, 1};
51     DTYPE values[] = {3, 4, 5, 9, 2, 3, 1, 4, 6, 1, 1, 1, 1, 1, 1, 1};
52     int columnIndex[] = {0, 1, 1, 2, 0, 2, 3, 1, 3, 0, 0, 1, 0, 1, 0, 1};
53     int rowPtr[] = {0, 2, 4, 7, 9, 10, 12, 14, 16};
54     */
55
56     DTYPE ystd[size];
57     std_multiplication(matrix, ystd, x);
58     DTYPE y[size];
59     smvm(rowPtr, columnIndex, values, y, x);
```

```

60     cout << endl;
61     for(int i=0; i<size; ++i) {
62         cout << "ystd=" << ystd[i] << ", ";
63         cout << "y=" << y[i] << endl;
64         if(ystd[i] != y[i] )
65             fail = 1;
66         if(fail == 1)
67             cout << "i=" << i << " failed." << endl;
68         else
69             cout << "i=" << i << " passed." << endl;
70     }
71     cout << endl;
72
73     return fail;
74 }

```

```

1  INFO: [SIM 211-2] ***** CSIM start *****
2  INFO: [SIM 211-4] CSIM will launch GCC as the compiler.
3  Compiling ../.././smvmTB.cpp in debug mode
4  Compiling ../.././smvm.cpp in debug mode
5  Compiling ../.././stdm.cpp in debug mode
6  Generating csim.exe
7
8  ystd=7, y=7
9  i=0 passed.
10 ystd=14, y=14
11 i=1 passed.
12 ystd=6, y=6
13 i=2 passed.
14 ystd=10, y=10
15 i=3 passed.
16
17 INFO: [SIM 211-1] CSim done with 0 errors.
18 INFO: [SIM 211-3] ***** CSIM finish *****
19 Finished C simulation.

```

Nello specifico, qui di seguito, viene riportata l'implementazione della moltiplicazione standard utilizzata per verificare i risultati sopra allegati.

```

1  #include "definitions.h"
2
3  void std_multiplication(DTYPE matrix[size][size], DTYPE *y, DTYPE *x) {
4      for (int i = 0; i < size; i++) {
5          DTYPE ytmp = 0;
6          for (int j = 0; j < size; j++)
7              ytmp += matrix[i][j] * x[j];
8          y[i] = ytmp;
9      }
10 }

```

5 Solutions

Di seguito verranno illustrate e analizzate le soluzioni previste nella tabella sotto allegata. In particolare, nelle implementazioni dove è previsto l'utilizzo della direttiva di partitioning sono stati considerati tre array (columnIndex, values, x) a cui corrispondono quattro solution differenti. Nello specifico, è stata prevista una soluzione in cui viene effettuato il partitioning di tutte e tre gli array contemporaneamente e le rimanenti tre implementazioni in cui, per ognuna di essa, è stato previsto il partizionamento singolo di uno dei tre array appena citati. Tali implementazioni sono riportate qui di seguito.

| Solution | Loop1 | Loop2 |
|----------|----------|---|
| 1 | - | - |
| 2 | - | Pipeline |
| 3 | Pipeline | - |
| 4 | Unroll=2 | - |
| 5 | - | Pipeline, Unroll=2 |
| 6 | - | Pipeline, Unroll=2, Cyclic=2 |
| | - | • Pipeline, Unroll=2, Cyclic=2 (columnIndex) |
| | - | • Pipeline, Unroll=2, Cyclic=2 (values) |
| | - | • Pipeline, Unroll=2, Cyclic=2 (x) |
| | - | • Pipeline, Unroll=2, Cyclic=2 (columnIndex, values, x) |
| 7 | - | Pipeline, Unroll=4 |
| 8 | - | Pipeline, Unroll=2, Cyclic=4 |
| | - | • Pipeline, Unroll=2, Cyclic=4 (columnIndex) |
| | - | • Pipeline, Unroll=2, Cyclic=4 (values) |
| | - | • Pipeline, Unroll=2, Cyclic=4 (x) |
| | - | • Pipeline, Unroll=2, Cyclic=4 (columnIndex, values, x) |
| 9 | - | Pipeline, Unroll=8 |
| 10 | - | Pipeline, Unroll=2, Cyclic=8 |
| | - | • Pipeline, Unroll=2, Cyclic=8 (columnIndex) |
| | - | • Pipeline, Unroll=2, Cyclic=8 (values) |
| | - | • Pipeline, Unroll=2, Cyclic=8 (x) |
| | - | • Pipeline, Unroll=2, Cyclic=8 (columnIndex, values, x) |
| 11 | - | Pipeline, Unroll=2, Block=8 |
| | - | • Pipeline, Unroll=2, Block=8 (columnIndex) |
| | - | • Pipeline, Unroll=2, Block=8 (values) |
| | - | • Pipeline, Unroll=2, Block=8 (x) |
| | - | • Pipeline, Unroll=2, Block=8 (columnIndex, values, x) |

Table 2: Detailed SMVM Solutions To Be Performed

5.1 Solution 1

Qui, di seguito, viene riportata l'architettura relativa alla prima solution.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       ytmp += values[k] * x[columnIndex[k]];
8     }
9     y[i] = ytmp;
10  }
11 }

```

Si può notare come venga utilizzata una variabile temporanea *ytmp* poiché essa viene utilizzata per calcolare l'uscita corrispondente. In particolare, il risultato calcolato ad ogni iterazione *k* viene sommato a quello della precedente iterazione. Pertanto, essendo che l'uscita deve essere solo assegnata e non letta per ogni iterazione, si utilizza una variabile temporanea per calcolare il risultato. Solo alla fine delle iterazioni si potrà assegnare il risultato all'uscita corrispondente.

Effettuando la sintesi è possibile evidenziare il seguente report:

| Clock | Target | Estimated | Uncertainty |
|--------|--------|-----------|-------------|
| ap_clk | 10.00 | 8.510 | 1.25 |

Table 3: HLS Solution 1 without Trip Count Timing Summary (ns)

| Latency | | Interval | |
|---------|-----|----------|-----|
| min | max | min | max |
| ? | ? | ? | ? |

Table 4: HLS Solution 1 without Trip Count Latency Summary (clock cycles)

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | ? | ? | ? | - | - | 4 |
| + loop2 | ? | ? | 5 | - | - | ? |

Table 5: HLS Solution 1 without Trip Count Latency Loops Summary

Si può notare come la latenza associata a questa architettura risulta essere "?", cioè non definita. In particolare tale non definizione è dovuta al loop2 del quale non è definito il trip count associato essendo il numero di iterazioni corrispondente non noto a priori. Nello specifico, il ciclo 2 dipende dai valori presenti all'interno dell'array *rowPtr* che sono incogniti poiché dipendono dai valori in input all'architettura. Viceversa, la latenza per ogni iterazione (IL), essendo che dipende dalla tipologia di operazioni, risulta essere definita. Per quanto riguarda, invece, il loop1, esso presenta un'iteration latency non definita poiché dipendente direttamente dal loop2 di cui non si è a conoscenza della latenza totale come spiegato precedentemente. Pertanto, la latenza totale del loop1 e, di conseguenza, la latenza totale associata all'architettura risulta essere non nota a priori. Quindi, per poter risolvere si specifica all'interno dell'implementazione la direttiva *trip_count*. In particolare, si possono specificare tre valori all'interno di tale pragma: min, max e avg. Tali valori fanno riferimento rispettivamente al numero minimo, massimo e medio di iterazioni del loop di riferimento. In particolare, tenendo conto che in generale *rowPtr[i]* indica il numero di elementi non nulli al di sopra della *i*-esima riga all'interno della matrice, il loop2 è come se scandisse riga per riga la matrice. All'iterazione *i* del loop1 considera le iterazioni possibili all'interno del loop2 comprese tra il valore *k=rowPtr[i]* (corrispondente al numero di elementi non nulli presenti al di sopra della riga *i*) e il valore *k=rowPtr[i+1]* (corrispondente al numero di elementi non nulli presenti al di sopra della riga *i+1*). Quindi, la differenza tra *rowPtr[i+1]* e *rowPtr[i]* rappresenta il numero di elementi presenti all'interno della riga *i*. Pertanto, tale direttiva permette al tool di analizzare come la latenza del loop contribuisce alla latenza totale dell'architettura così permettendo al progettista di effettuare ulteriori ottimizzazioni al design.

Pertanto, si allega l'architettura risultante.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       ytmp += values[k] * x[columnIndex[k]];
9     }
10    y[i] = ytmp;
11  }
12 }

```

Effettuando la sintesi è possibile evidenziare il seguente report:

| Clock | Target | Estimated | Uncertainty |
|--------|--------|-----------|-------------|
| ap_clk | 10.00 | 8.510 | 1.25 |

Table 6: HLS Solution 1 with Trip Count Timing Summary (ns)

| Latency | | Interval | |
|---------|-----|----------|-----|
| min | max | min | max |
| 13 | 93 | 13 | 93 |

Table 7: HLS Solution 1 with Trip Count Latency Summary (clock cycles)

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 12 | 92 | 3~23 | - | - | 4 |
| + loop2 | 0 | 20 | 5 | - | - | 0~4 |

Table 8: HLS Solution 1 Latency with Trip Count Loops Summary

Si può notare come, dopo aver applicato la direttiva di trip_count, i valori di latenza risultano essere definiti numericamente. In particolare, il loop2 presenta un numero di iterazioni compresa tra 0 e 4, cioè rispettivamente il valore minimo e massimo specificati nel pragma di trip_count. Bisogna ricordare che tale direttiva non ha impatto sull'architettura ma ha solo impatto sui cicli di latenza.

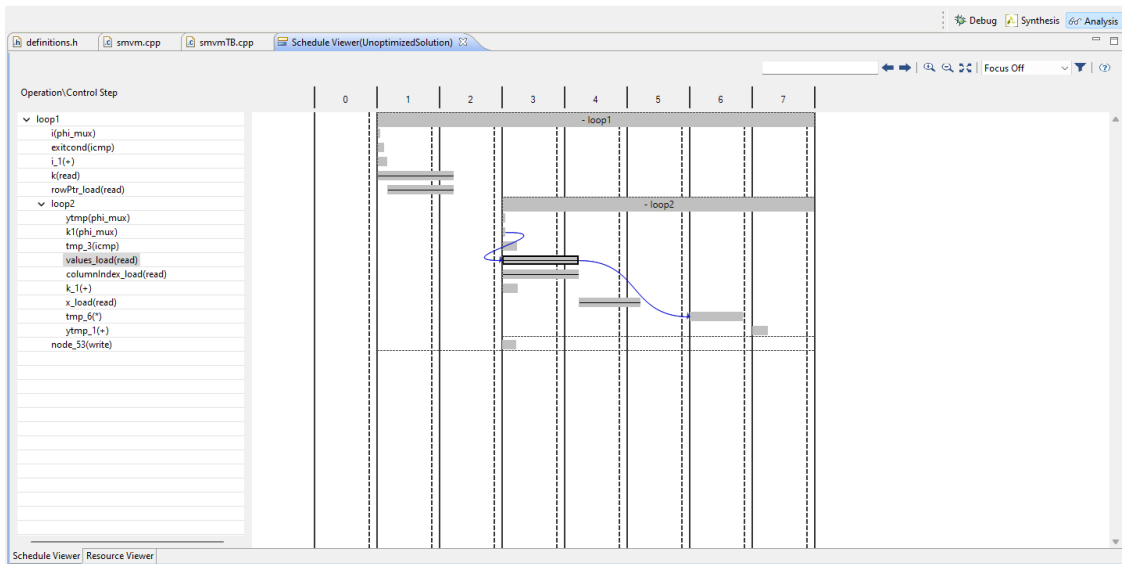


Figure 1: HLS Solution 1 Analysis

Qui di seguito, viene allegato l'utilizzazione delle risorse stimata dal processo di sintesi.

| Name | BRAM_18K | DSP48E | FF | LUT |
|------------------------|-----------------|---------------|-----------|------------|
| DSP | - | - | - | - |
| Expression | - | 3 | 0 | 137 |
| FIFO | - | - | - | - |
| Instance | - | - | - | - |
| Memory | 0 | - | - | - |
| Multiplexer | - | - | - | 71 |
| Register | - | - | 241 | - |
| Total | 0 | 3 | 241 | 208 |
| Available | 280 | 220 | 106400 | 53200 |
| Utilization (%) | 0 | 1 | ~0 | ~0 |

Table 9: HLS Solution 1 with Trip Count Utilization Estimates Summary

Successivamente effettuando la C/RTL Cosimulation e l'Export RTL è possibile evidenziare i seguenti report.

| RTL | Status | Latency | | | Interval | | |
|------|--------|----------------|-----|-----|-----------------|-----|-----|
| | | min | avg | max | min | avg | max |
| VHDL | Pass | 58 | 58 | 58 | NA | NA | NA |

Table 10: HLS Solution 1 with Trip Count C/RTL Cosimulation Summary

| Resource | VHDL |
|-----------------|-------------|
| SLICE | 48 |
| LUT | 93 |
| FF | 161 |
| DSP | 3 |
| BRAM | 0 |
| SRL | 0 |

Table 11: HLS Solution 1 with Trip Count Export RTL Resource Usage

| Timing | VHDL |
|---------------------------------|-------------|
| CP required | 10.000 |
| CP achieved post-synthesis | 5.745 |
| CP achieved post-implementation | 5.692 |

Table 12: HLS Solution 1 with Trip Count Export RTL Final Timing

5.2 Solution 2

Qui, di seguito, viene riportata l'architettura relativa alla seconda solution.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       #pragma HLS pipeline
9       ytmp += values[k] * x[columnIndex[k]];
10    }
11    y[i] = ytmp;
12  }
13 }

```

In particolare, nella soluzione hardware in questione, rispetto alla solution 1, è stato aggiunto la direttiva di pipeline all'interno del loop2. Pertanto, ci si dovrebbe aspettare una minore latenza totale dal momento che il pipelining permette di scindere le operazioni complesse in più operazioni semplici. In questo modo si può far lavorare l'architettura con dati temporalmente differenti.

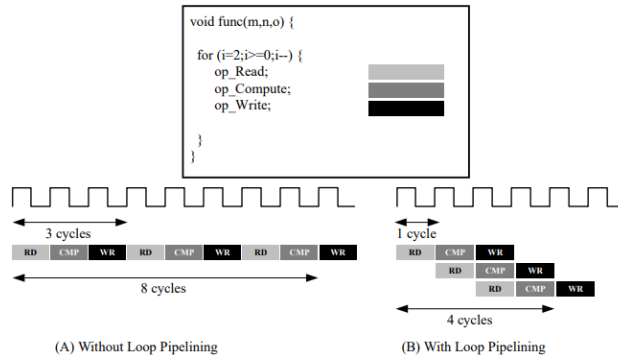


Figure 2: HLS Loop Pipelining

Effettuando la sintesi è possibile evidenziare il seguente report:

| Clock | Target | Estimated | Uncertainty |
|--------|--------|-----------|-------------|
| ap_clk | 10.00 | 8.510 | 1.25 |

Table 13: HLS Solution 2 Timing Summary (ns)

| Latency | | Interval | |
|---------|-----|----------|-----|
| min | max | min | max |
| 17 | 45 | 17 | 45 |

Table 14: HLS Solution 2 Latency Summary (clock cycles)

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 16 | 44 | 4~11 | - | - | 4 |
| + loop2 | 0 | 7 | 5 | 1 | 1 | 0~4 |

Table 15: HLS Solution 2 Latency Loops Summary

Si può notare, rispetto alla solution precedente, come in questo caso venga specificato un valore numerico di Initiation Interval (II). In particolare, l'II.achieved risulta essere il medesimo di quello target, cioè uguale

a 1. Teoricamente, come in questo caso, si dovrebbe ottenere $II_{target}=II_{achieved}$. Se così non fosse allora il tool non è riuscito a raggiungere l'obiettivo prefissato e si dovrebbero attuare modifiche all'architettura o eventualmente prevedere l'utilizzo di ulteriori direttive.

Qui di seguito, viene allegato l'utilizzazione delle risorse stimata dal processo di sintesi.

| Name | BRAM_18K | DSP48E | FF | LUT |
|------------------------|----------|--------|--------|-------|
| DSP | - | - | - | - |
| Expression | - | 3 | 0 | 141 |
| FIFO | - | - | - | - |
| Instance | - | - | - | - |
| Memory | 0 | - | - | - |
| Multiplexer | - | - | - | 78 |
| Register | - | - | 340 | 32 |
| Total | 0 | 3 | 340 | 251 |
| Available | 280 | 220 | 106400 | 53200 |
| Utilization (%) | 0 | 1 | ~0 | ~0 |

Table 16: HLS Solution 2 Utilization Estimates Summary

Successivamente effettuando la C/RTL Cosimulation e l'Export RTL è possibile evidenziare i seguenti report.

| RTL | Status | Latency | | | Interval | | |
|------|--------|---------|-----|-----|----------|-----|-----|
| | | min | avg | max | min | avg | max |
| VHDL | Pass | 38 | 38 | 38 | NA | NA | NA |

Table 17: HLS Solution 1 with Trip Count C/RTL Cosimulation Summary

In particolare, si può notare come, in seguito all'introduzione della direttiva di pipeline, il numero di risorse risulta essere cambiato. Nello specifico, l'utilizzazione delle LUT è aumentata di circa il 24% mentre quella dei FF è diminuita di circa il 14%.

| Resource | VHDL |
|----------|------|
| SLICE | 38 |
| LUT | 115 |
| FF | 139 |
| DSP | 3 |
| BRAM | 0 |
| SRL | 0 |

Table 18: HLS Solution 2t Export RTL Resource Usage

| Timing | VHDL |
|---------------------------------|--------|
| CP required | 10.000 |
| CP achieved post-synthesis | 5.745 |
| CP achieved post-implementation | 5.718 |

Table 19: HLS Solution 2 Export RTL Final Timing

5.3 Solution 3

Qui, di seguito, viene riportata l'architettura relativa alla terza solution.

```

1 #include "definitions.h"
2
3 #include "definitions.h"
4
5 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
6   loop1: for (int i=0; i<rows; i++) {
7     #pragma HLS pipeline
8     DTYPE ytmp = 0;
9     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
10      #pragma HLS loop_tripcount min=0 max=4 avg=2
11      ytmp += values[k] * x[columnIndex[k]];
12    }
13    y[i] = ytmp;
14  }
15 }

```

In particolare, nella soluzione hardware in questione, rispetto alla solution 1, è stato aggiunta la direttiva di pipeline all'interno del loop1.

Effettuando la sintesi è possibile evidenziare il seguente report:

| Clock | Target | Estimated | Uncertainty |
|--------|--------|-----------|-------------|
| ap_clk | 10.00 | 8.510 | 1.25 |

Table 20: HLS Solution 3 Timing Summary (ns)

| Latency | | Interval | |
|---------|-----|----------|-----|
| min | max | min | max |
| 13 | 93 | 13 | 93 |

Table 21: HLS Solution 3 Latency Summary (clock cycles)

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 12 | 92 | 3~23 | - | - | 4 |
| + loop2 | 0 | 20 | 5 | - | - | 0~4 |

Table 22: HLS Solution 3 Latency Loops Summary

Si può notare come, in questo caso l'Initiation Interval sia non specificato nel loop2 dal momento che la direttiva introdotta nella solution 2 è stata eliminata per la soluzione hardware in questione. Molto più importante è che, considerando la direttiva di pipeline all'interno del loop1, in corrispondenza dell'Initiation Interval di tale ciclo non è definito alcun valore numerico. Tanto è vero che, analizzando i log della sintesi presenti nella console è possibile identificare i seguenti warning.

WARNING: [XFORM 203-503] Cannot unroll loop 'loop2' (smvmProject/svm.cpp:21) in function 'smvm' completely: variable loop bound.

WARNING: [SCHED 204-65] Unable to satisfy pipeline directive: Loop contains subloop(s) not being unrolled or flattened.

In particolare, il tool non è riuscito a soddisfare la richiesta di pipeline a causa dei bound non noti e, pertanto, non riuscendo ad effettuare l'automatic unrolling del loop2. Infatti, si può notare come i valori di latenza siano i medesimi di quelli della solution 1.

Qui di seguito, viene allegato l'utilizzazione delle risorse stimata dal processo di sintesi. Anche in questo caso il numero di risorse è il medesimo di quello ottenuto in corrispondenza della solution 1.

Successivamente effettuando la C/RTL Cosimulation e l'Export RTL è possibile evidenziare i seguenti report. Anche in questo caso, sia il report del C/RTL Cosimulation sia quello dell'Export RTL risultano essere i medesimi di quelli della solution 1.

| Name | BRAM_18K | DSP48E | FF | LUT |
|------------------------|----------|--------|--------|-------|
| DSP | - | - | - | - |
| Expression | - | 3 | 0 | 137 |
| FIFO | - | - | - | - |
| Instance | - | - | - | - |
| Memory | 0 | - | - | - |
| Multiplexer | - | - | - | 71 |
| Register | - | - | 241 | - |
| Total | 0 | 3 | 241 | 208 |
| Available | 280 | 220 | 106400 | 53200 |
| Utilization (%) | 0 | 1 | ~0 | ~0 |

Table 23: HLS Solution 3 Utilization Estimates Summary

| RTL | Status | Latency | | | Interval | | |
|------|--------|---------|-----|-----|----------|-----|-----|
| | | min | avg | max | min | avg | max |
| VHDL | Pass | 58 | 58 | 58 | NA | NA | NA |

Table 24: HLS Solution 3 C/RTL Cosimulation Summary

| Resource | VHDL |
|----------|------|
| SLICE | 48 |
| LUT | 94 |
| FF | 161 |
| DSP | 3 |
| BRAM | 0 |
| SRL | 0 |

Table 25: HLS Solution 3 Export RTL Resource Usage

| Timing | VHDL |
|---------------------------------|--------|
| CP required | 10.000 |
| CP achieved post-synthesis | 5.745 |
| CP achieved post-implementation | 5.692 |

Table 26: HLS Solution 3 Export RTL Final Timing

5.4 Solution 4

Qui, di seguito, viene riportata l'architettura relativa alla quarta solution.

```

1 #include "definitions.h"
2
3 #include "definitions.h"
4
5 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
6   loop1: for (int i=0; i<rows; i++) {
7     #pragma HLS unroll factor=2
8     DTYPE ytmp = 0;
9     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
10      #pragma HLS loop_tripcount min=0 max=4 avg=2
11      ytmp += values[k] * x[columnIndex[k]];
12    }
13    y[i] = ytmp;
14  }
15 }

```

In particolare, nella soluzione hardware in questione, rispetto alla solution 1, è stata aggiunta la direttiva di unrolling con fattore pari a 2 all'interno del loop1.

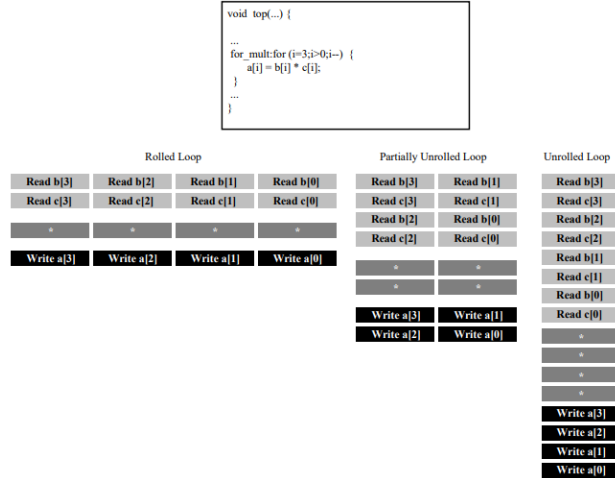


Figure 3: HLS Loop Unrolling

Effettuando la sintesi è possibile evidenziare il seguente report:

| Clock | Target | Estimated | Uncertainty |
|--------|--------|-----------|-------------|
| ap_clk | 10.00 | 8.510 | 1.25 |

Table 27: HLS Solution 4 Timing Summary (ns)

| Latency | | Interval | |
|---------|-----|----------|-----|
| min | max | min | max |
| 11 | 91 | 11 | 91 |

Table 28: HLS Solution 4 Latency Summary (clock cycles)

In particolare, si può notare come il valore di trip count del loop1 risulta essere dimezzato rispetto alla solution 1. Questo è dovuto all'attuazione della direttiva di unrolling di fattore 2 sul loop1 e così permettendo l'esecuzione in parallelo di due iterazioni.

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 10 | 90 | 5~45 | - | - | 2 |
| + loop2 | 0 | 20 | 5 | - | - | 0~4 |
| + loop2 | 0 | 20 | 5 | - | - | 0~4 |

Table 29: HLS Solution 4 Latency Loops Summary

Infatti, lo si può meglio notare tramite l'interfaccia analysis. In particolare, si può evidenziare come il loop1 venga parallelizzato. Nello specifico, vengono previsti due loop2 uno dopo l'altro facendo così aumentare la latenza per ogni iterazione del loop1. Quindi, il valore del trip count associato al ciclo 1 viene dimezzato mentre l'Iteration Latency associata allo stesso loop aumenta dal momento che vengono gestiti due loop2.

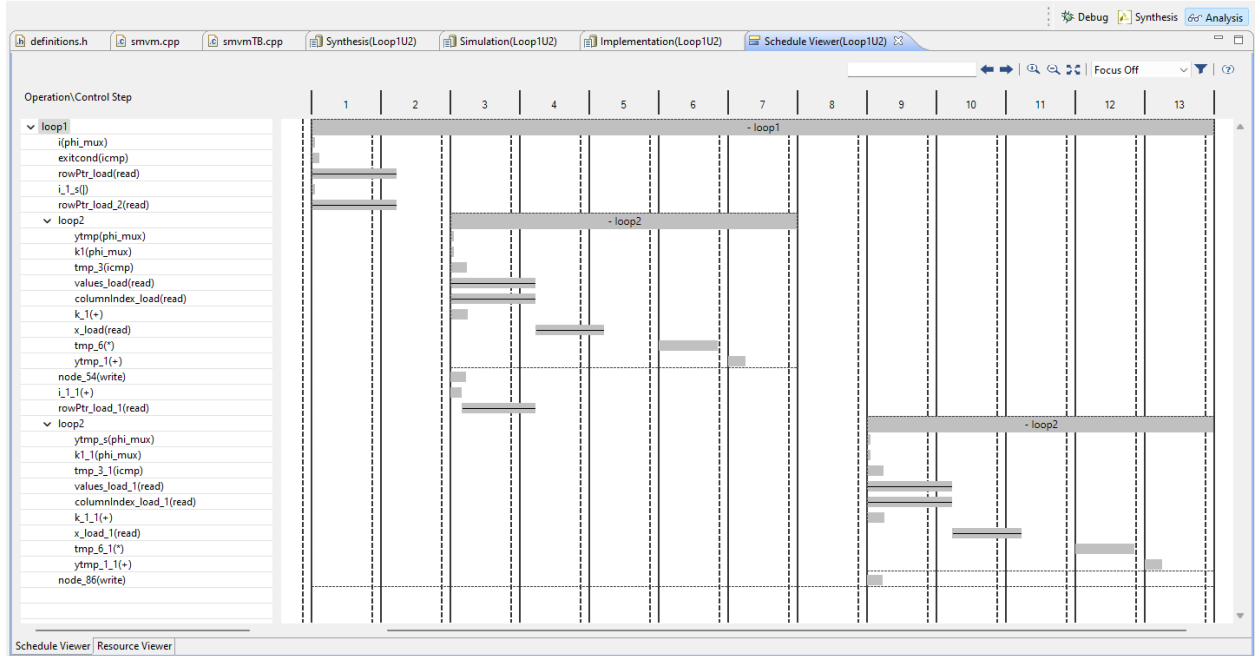


Figure 4: HLS Solution 4 Analysis

| Name | BRAM_18K | DSP48E | FF | LUT |
|------------------------|------------|------------|---------------|--------------|
| DSP | - | - | - | - |
| Expression | - | 3 | 0 | 235 |
| FIFO | - | - | - | - |
| Instance | - | - | - | - |
| Memory | 0 | - | - | - |
| Multiplexer | - | - | - | 197 |
| Register | - | - | 376 | - |
| Total | 0 | 3 | 376 | 432 |
| Available | 280 | 220 | 106400 | 53200 |
| Utilization (%) | 0 | 1 | ~0 | ~0 |

Table 30: HLS Solution 4 Utilization Estimates Summary

Successivamente effettuando la C/RTL Cosimulation e l'Export RTL è possibile evidenziare i seguenti report.

| RTL | Status | Latency | | | Interval | | |
|------|--------|---------|-----|-----|----------|-----|-----|
| | | min | avg | max | min | avg | max |
| VHDL | Pass | 56 | 56 | 56 | NA | NA | NA |

Table 31: HLS Solution 4 C/RTL Cosimulation Summary

In particolare, rispetto alla solution 1 si ha un aumento di circa il 73% delle slice, del 100% delle LUT e di circa l'81% dei FF dal momento che è stato introdotto un parallelismo all'interno dell'architettura.

| Resource | VHDL |
|----------|------|
| SLICE | 83 |
| LUT | 186 |
| FF | 292 |
| DSP | 3 |
| BRAM | 0 |
| SRL | 0 |

Table 32: HLS Solution 4 Export RTL Resource Usage

| Timing | VHDL |
|---------------------------------|--------|
| CP required | 10.000 |
| CP achieved post-synthesis | 5.745 |
| CP achieved post-implementation | 5.692 |

Table 33: HLS Solution 4 Export RTL Final Timing

5.5 Solution 5

Qui, di seguito, viene riportata l'architettura relativa alla quinta solution.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       #pragma HLS pipeline
9       #pragma HLS unroll factor=2
10      ytmp += values[k] * x[columnIndex[k]];
11    }
12    y[i] = ytmp;
13  }
14 }

```

In particolare, nella soluzione hardware in questione, rispetto alla solution 2 dove era presenta solo la direttiva di pipelining nel loop2, è stata aggiunta la direttiva di unrolling con fattore pari a 2 all'interno dello stesso ciclo.

Effettuando la sintesi è possibile evidenziare il seguente report:

| Clock | Target | Estimated | Uncertainty |
|--------|--------|-----------|-------------|
| ap_clk | 10.00 | 8.510 | 1.25 |

Table 34: HLS Solution 5 Timing Summary (ns)

| Latency | | Interval | |
|---------|-----|----------|-----|
| min | max | min | max |
| 33 | 41 | 33 | 41 |

Table 35: HLS Solution 5 Latency Summary (clock cycles)

In particolare, si può notare come, rispetto alla solution 2 dove il trip count relativo al loop2 era pari a $0 \sim 4$, in questo caso il trip count associato al loop2 risulta essere dimezzato dal momento che è stato previsto un unrolling di fattore pari a 2 all'interno del ciclo in questione. Inoltre, dal momento che è stato introdotto una direttiva di pipeline all'interno del loop2, si può evidenziare come l'Initiation Interval raggiunto risulta essere il medesimo di quello target.

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 32 | 40 | 8~10 | - | - | 4 |
| + loop2 | 4 | 6 | 5 | 1 | 1 | 0~2 |

Table 36: HLS Solution 5 Latency Loops Summary

Si può notare come, all'interno del loop2, vengono effettuate in parallelo due letture relative alle variabili columnIndex, due relative alle variabili *values* e due relative alle variabili x.

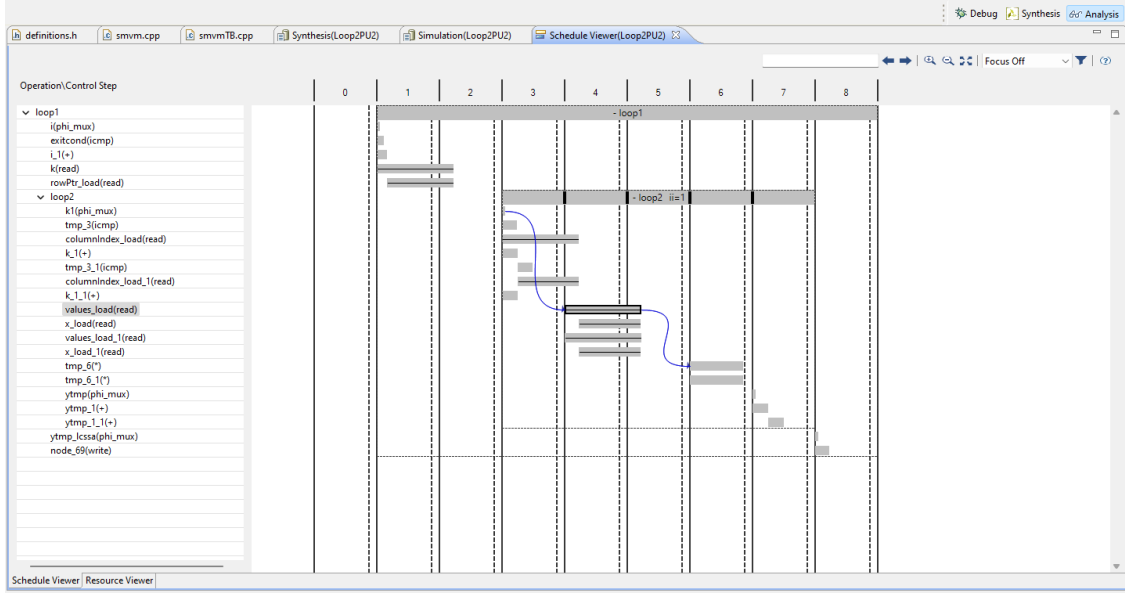


Figure 5: HLS Solution 5 Analysis

| Name | BRAM_18K | DSP48E | FF | LUT |
|------------------------|------------|------------|---------------|--------------|
| DSP | - | - | - | - |
| Expression | - | 6 | 0 | 257 |
| FIFO | - | - | - | - |
| Instance | - | - | - | - |
| Memory | 0 | - | - | - |
| Multiplexer | - | - | - | 78 |
| Register | - | - | 597 | 64 |
| Total | 0 | 6 | 597 | 399 |
| Available | 280 | 220 | 106400 | 53200 |
| Utilization (%) | 0 | 2 | ~0 | ~0 |

Table 37: HLS Solution 5 Utilization Estimates Summary

| RTL | Status | Latency | | | Interval | | |
|------|--------|---------|-----|-----|----------|-----|-----|
| | | min | avg | max | min | avg | max |
| VHDL | Pass | 37 | 37 | 37 | NA | NA | NA |

Table 38: HLS Solution 5 C/RTL Cosimulation Summary

Si può notare, rispetto alla soluzione hardware 2, un aumento dell'utilizzazione delle risorse di circa l'82% per quanto riguarda le slice, del 60% per quanto riguarda le LUT e di circa il 41% per quanto riguarda i FF. Inoltre, si può evidenziare come il numero dei DSP sia raddoppiato proprio in seguito all'introduzione della direttiva di unrolling di fattore pari a 2 all'interno del loop2.

| Resource | VHDL |
|-----------------|-------------|
| SLICE | 69 |
| LUT | 184 |
| FF | 196 |
| DSP | 6 |
| BRAM | 0 |
| SRL | 0 |

Table 39: HLS Solution 5 Export RTL Resource Usage

| Timing | VHDL |
|---------------------------------|-------------|
| CP required | 10.000 |
| CP achieved post-synthesis | 7.927 |
| CP achieved post-implementation | 7.465 |

Table 40: HLS Solution 5 Export RTL Final Timing

5.6 Solution 6

Qui, di seguito, vengono riportate le architetture relative alla sesta solution. In particolare, come già precedentemente citato, tale solution prevede l'utilizzo della direttiva di partitioning.

Il partizionamento serve per risolvere un problema tipicamente causato dagli array. Gli array sono implementati come BRAM, solitamente progettate per un dual-port massimo. Questo può limitare il throughput di un algoritmo ad alta intensità di read/write. La larghezza di banda può essere migliorata dividendo l'array (una singola BRAM) in array più piccoli (più BRAM), aumentando di fatto il numero di porte. Gli array vengono partizionati utilizzando la direttiva ARRAY_PARTITION. Vivado HLS offre tre tipi di partizionamento degli array. I tre tipi di partizionamento sono:

- **block**

L'array originale viene suddiviso in blocchi di uguali dimensioni di elementi consecutivi dell'array originale.

- **cyclic**

L'array originale viene suddiviso in blocchi di uguali dimensioni che interlacciano gli elementi dell'array originale.

- **complete**

L'operazione predefinita consiste nel dividere l'array nei suoi singoli elementi. Ciò corrisponde alla risoluzione di una memoria in registri.

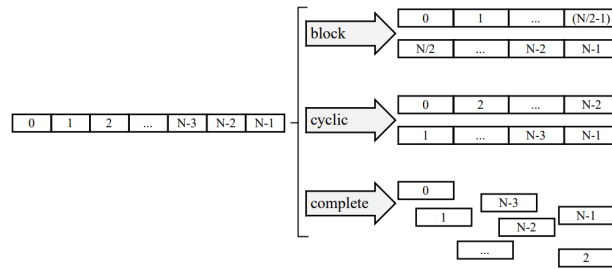


Figure 6: HLS Array Partitioning

Nella soluzione hardware in questione verrà utilizzata la direttiva di partizionamento di tipologia cyclic e nello specifico, verranno analizzate le seguenti implementazioni relative al loop2:

- Pipeline, Unroll=2, Cyclic=2 (columnIndex, values, x)
- Pipeline, Unroll=2, Cyclic=2 (columnIndex)
- Pipeline, Unroll=2, Cyclic=2 (values)
- Pipeline, Unroll=2, Cyclic=2 (x)

In particolare, è possibile evidenziare nel dettaglio le differenti soluzioni hardware nei seguenti allegati.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       #pragma HLS pipeline
9       #pragma HLS unroll factor=2
10      #pragma HLS array_partition variable=columnIndex cyclic factor=2
11      #pragma HLS array_partition variable=values cyclic factor=2
12      #pragma HLS array_partition variable=x cyclic factor=2

```

```

13     ytmp += values[k] * x[columnIndex[k]];
14 }
15 y[i] = ytmp;
16 }
17 }

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4     loop1: for (int i=0; i<rows; i++) {
5         DTYPE ytmp = 0;
6         loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7             #pragma HLS loop_tripcount min=0 max=4 avg=2
8             #pragma HLS pipeline
9             #pragma HLS unroll factor=2
10            #pragma HLS array_partition variable=columnIndex cyclic factor=2
11            ytmp += values[k] * x[columnIndex[k]];
12        }
13        y[i] = ytmp;
14    }
15 }

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4     loop1: for (int i=0; i<rows; i++) {
5         DTYPE ytmp = 0;
6         loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7             #pragma HLS loop_tripcount min=0 max=4 avg=2
8             #pragma HLS pipeline
9             #pragma HLS unroll factor=2
10            #pragma HLS array_partition variable=values cyclic factor=2
11            ytmp += values[k] * x[columnIndex[k]];
12        }
13        y[i] = ytmp;
14    }
15 }

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4     loop1: for (int i=0; i<rows; i++) {
5         DTYPE ytmp = 0;
6         loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7             #pragma HLS loop_tripcount min=0 max=4 avg=2
8             #pragma HLS pipeline
9             #pragma HLS unroll factor=2
10            #pragma HLS array_partition variable=x cyclic factor=2
11            ytmp += values[k] * x[columnIndex[k]];
12        }
13        y[i] = ytmp;
14    }
15 }

```

Effettuando la sintesi è possibile evidenziare il seguente report:

| Solution | Clock | Target | Estimated | Uncertainty |
|------------------------|--------------|---------------|------------------|--------------------|
| columnIndex, values, x | ap_clk | 10.00 | 8.510 | 1.25 |
| columnIndex | ap_clk | 10.00 | 8.510 | 1.25 |
| values | ap_clk | 10.00 | 8.510 | 1.25 |
| x | ap_clk | 10.00 | 8.510 | 1.25 |

Table 41: HLS Solution 6 Timing Summary (ns)

| Solution | Latency | | Interval | |
|------------------------|----------------|-----|-----------------|-----|
| | min | max | min | max |
| columnIndex, values, x | 33 | 41 | 33 | 41 |
| columnIndex | 33 | 41 | 33 | 41 |
| values | 33 | 41 | 33 | 41 |
| x | 33 | 41 | 33 | 41 |

Table 42: HLS Solution 6 Latency Summary (clock cycles)

Si può notare come, in corrispondenza di tutte e quattro le soluzioni hardware proposte in questa sezione, i valori di Iteration Latency, trip count, Initiation Interval relativa al loop2 e latenza totale del loop1 e loop2 risultano essere i medesimi. In particolare, il valore di trip count del loop2 risulta essere dimezzato, come quello della solution5, rispetto, ad esempio, alla solution2 dal momento che è presente la direttiva di unrolling di fattore pari a 2.

| Solution | Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|------------------------|------------------|----------------|-----|--------------------------|----------------------------|--------|-------------------|
| | | min | max | | achieved | target | |
| columnIndex, values, x | - loop1 | 32 | 40 | 8~10 | - | - | 4 |
| | + loop2 | 4 | 6 | 5 | 1 | 1 | 0~2 |
| columnIndex | - loop1 | 32 | 40 | 8~10 | - | - | 4 |
| | + loop2 | 4 | 6 | 5 | 1 | 1 | 0~2 |
| values | - loop1 | 32 | 40 | 8~10 | - | - | 4 |
| | + loop2 | 4 | 6 | 5 | 1 | 1 | 0~2 |
| x | - loop1 | 32 | 40 | 8~10 | - | - | 4 |
| | + loop2 | 4 | 6 | 5 | 1 | 1 | 0~2 |

Table 43: HLS Solution 6 Latency Loops Summary

| Solution | BRAM_18K | DSP48E | FF | LUT |
|------------------------|-----------------|---------------|-----------|------------|
| columnIndex, values, x | 0 | 6 | 535 | 623 |
| columnIndex | 0 | 6 | 533 | 472 |
| values | 0 | 6 | 533 | 472 |
| x | 0 | 6 | 599 | 463 |

Table 44: HLS Solution 6 Utilization Estimates [#]

| Solution | RTL | Status | Latency | | | Interval | | |
|------------------------|------|--------|---------|-----|-----|----------|-----|-----|
| | | | min | avg | max | min | avg | max |
| columnIndex, values, x | VHDL | Pass | 37 | 37 | 37 | NA | NA | NA |
| columnIndex | VHDL | Pass | 37 | 37 | 37 | NA | NA | NA |
| values | VHDL | Pass | 37 | 37 | 37 | NA | NA | NA |
| x | VHDL | Pass | 37 | 37 | 37 | NA | NA | NA |

Table 45: HLS Solution 6 C/RTL Cosimulation Report

Si può notare come, in corrispondenza della soluzione basata su partitioning dell'array columnIndex e dell'array values si ha la medesima utilizzazione dei FF. Molto probabilmente questo risultato è legato al fatto che entrambi gli array presentano medesima dimensione, cioè pari a nnz. Inoltre, la minore utilizzazione delle risorse si ha in corrispondenza dell'array x a cui corrisponde, infatti, la dimensione minore tra i tre array considerati. In particolare, la soluzione in cui viene considerato il partizionamento dei tre array contemporaneamente potrebbe essere considerata come la solution che richiede più risorse dal momento che presenta il maggior numero di slice utilizzate.

| Solution | SLICE | LUT | FF | DSP | BRAM | CP required | CP achieved post- synthesis | CP achieved post- implementation |
|------------------------|-------|-----|-----|-----|------|----------------|--------------------------------------|---|
| columnIndex, values, x | 113 | 316 | 198 | 6 | 0 | 10 | 7.927 | 7.799 |
| columnIndex | 84 | 259 | 224 | 6 | 0 | 10 | 7.472 | 7.843 |
| values | 99 | 327 | 224 | 6 | 0 | 10 | 7.502 | 8.184 |
| x | 99 | 250 | 198 | 6 | 0 | 10 | 6.541 | 6.931 |

Table 46: HLS Solution 6 Export RTL Report

5.7 Solution 7

Qui, di seguito, viene riportata l'architettura relativa alla settima solution.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       #pragma HLS pipeline
9       #pragma HLS unroll factor=4
10      ytmp += values[k] * x[columnIndex[k]];
11    }
12    y[i] = ytmp;
13  }
14 }

```

In particolare, rispetto alla soluzione hardware 5 dove era stato considerato un parallelismo di fattore pari a 2, in questa solution è stato considerato un unrolling di fattore pari a 4. In particolare, ciò che ci si aspetta è un aumento delle risorse ed eventuali problematiche relative al timing dal momento che il tool deve gestire all'interno del loop2 più accessi in memoria paralleli.

Effettuando la sintesi è possibile evidenziare il seguente log nella console:

WARNING: [SCHD 204-69] Unable to schedule 'load' operation ('columnIndex_load_2', smvmProject/s-mvm.cpp:34) on array 'columnIndex' due to limited memory ports. Please consider using a memory core with more ports or partitioning the array 'columnIndex'.

Tale log sta a significare che non riesce a schedulare correttamente, dal punto di vista degli accessi in memoria, la load operation relativa all'array columnIndex dato dal numero limitato di porte relative alla memoria. In particolare, analizzando il report relativo alla sintesi, si può notare come l'Initiation Interval, associato al loop2, raggiunto risulta essere maggiore di quello target.

| Clock | Target | Estimated | Uncertainty |
|--------|--------|-----------|-------------|
| ap_clk | 10.00 | 8.510 | 1.25 |

Table 47: HLS Solution 7 Timing Summary (ns)

| Latency | | Interval | |
|---------|-----|----------|-----|
| min | max | min | max |
| 37 | 45 | 37 | 45 |

Table 48: HLS Solution 7 Latency Summary (clock cycles)

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 36 | 44 | 9~11 | - | - | 4 |
| + loop2 | 5 | 7 | 6 | 2 | 1 | 0~1 |

Table 49: HLS Solution 7 Latency Loops Summary

Pertanto, si potrebbe aggiungere una direttiva di partizionamento relativa all'array menzionato all'interno del log, cioè columnIndex.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       #pragma HLS pipeline

```

```

9      #pragma HLS unroll factor=4
10     #pragma HLS array_partition variable=columnIndex complete
11     ytmp += values[k] * x[columnIndex[k]];
12 }
13 y[i] = ytmp;
14 }
15 }

```

Effettuando nuovamente la sintesi, si ottiene il seguente log nella console e i seguenti valori di latenza. **WARNING: [SCHED 204-69] Unable to schedule 'load' operation ('values_load_2', smvmProject/svm.cpp:34) on array 'values' due to limited memory ports. Please consider using a memory core with more ports or partitioning the array 'values'.**

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 32 | 40 | 8~10 | - | - | 4 |
| + loop2 | 4 | 6 | 5 | 2 | 1 | 0~1 |

Table 50: HLS Solution 7 with columnIndex partitioning Latency Loops Summary

Si può notare come in questo caso il warning sia relativo all'array values. In particolare, la tipologia di warning è la medesima facendo presupporre che il tool non riesca a schedulare correttamente, secondo le direttive imposte dall'architettura, gli accessi in parallelo all'array values. Infatti, anche in questo caso il valore di Initiation Interval raggiunto risulta essere ancora maggiore di quello di quello target. Pertanto, si potrebbe aggiungere una direttiva di partizionamento relativa all'array menzionato all'interno del log, cioè values.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       #pragma HLS pipeline
9       #pragma HLS unroll factor=4
10      #pragma HLS array_partition variable=columnIndex complete
11      #pragma HLS array_partition variable=values complete
12      ytmp += values[k] * x[columnIndex[k]];
13    }
14    y[i] = ytmp;
15  }
16 }

```

Effettuando nuovamente la sintesi, si ottiene il seguente log nella console e i seguenti valori di latenza. **WARNING: [SCHED 204-69] Unable to schedule 'load' operation ('x_load_2', smvmProject/svm.cpp:34) on array 'x' due to limited memory ports. Please consider using a memory core with more ports or partitioning the array 'x'.**

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 32 | 40 | 8~10 | - | - | 4 |
| + loop2 | 4 | 6 | 5 | 2 | 1 | 0~1 |

Table 51: HLS Solution 7 with columnIndex and values partitioning Latency Loops Summary

Si può notare come in questo caso il warning sia relativo all'array x. In particolare, la tipologia di warning è la medesima della precedente. Anche in questo caso il valore di Initiation Interval raggiunto risulta essere ancora maggiore di quello di quello target.

Pertanto, si potrebbe aggiungere una direttiva di partizionamento relativa all'array menzionato all'interno del log, cioè x.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4     loop1: for (int i=0; i<rows; i++) {
5         DTYPE ytmp = 0;
6         loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7             #pragma HLS loop_tripcount min=0 max=4 avg=2
8             #pragma HLS pipeline
9             #pragma HLS unroll factor=4
10            #pragma HLS array_partition variable=columnIndex complete
11            #pragma HLS array_partition variable=values complete
12            #pragma HLS array_partition variable=x complete
13            ytmp += values[k] * x[columnIndex[k]];
14        }
15        y[i] = ytmp;
16    }
17 }

```

Effettuando nuovamente la sintesi, si ottiene il seguente log nella console e il seguente report.

WARNING: [SCH204-21] Estimated clock period (10.208ns) exceeds the target (target clock period: 10ns, clock uncertainty: 1.25ns, effective delay budget: 8.75ns). WARNING: [SCH204-21] The critical path in module 'smvm' consists of the following:

'add' operation ('ytmp_1_3', smvmProject/smvm.cpp:34) [139] (2.55 ns)
 'phi' operation ('ytmp', smvmProject/smvm.cpp:34) with incoming values : ('ytmp_1_3', smvmProject/smvm.cpp:34) [68] (0 ns)
 'add' operation ('ytmp_1', smvmProject/smvm.cpp:34) [105] (2.55 ns)
 'add' operation ('ytmp_1_1', smvmProject/smvm.cpp:34) [117] (2.55 ns)
 'add' operation ('ytmp_1_2', smvmProject/smvm.cpp:34) [128] (2.55 ns)

| Clock | Target | Estimated | Uncertainty |
|--------|--------|-----------|-------------|
| ap_clk | 10.00 | 10.208 | 1.25 |

Table 52: HLS Solution 7 with columnIndex, values and x partitioning Timing Summary (ns)

| Latency | | Interval | |
|---------|-----|----------|-----|
| min | max | min | max |
| 29 | 33 | 29 | 33 |

Table 53: HLS Solution 7 with columnIndex, values and x partitioning Latency Summary (clock cycles)

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 28 | 32 | 7~8 | - | - | 4 |
| + loop2 | 3 | 4 | 4 | 1 | 1 | 0~1 |

Table 54: HLS Solution 7 with columnIndex, values and x partitioning Latency Loops Summary

Si può evidenziare come, in questo caso, il valore di Initiation Interval raggiunto sia il medesimo di quello target, cioè pari a 1. Inoltre, quello che si può notare è che all'interno della console viene visualizzato un warning indicante un periodo di clock stimato maggiore di quello target. In particolare, viene stimato un timing per ogni operazione in maniera dettagliata: 2.55ns per *add operation ytmp_1_3*, 0ns per *phi operation ytmp*, 2.55ns per *add operation ytmp_1*, 2.55ns per *add operation ytmp_1_1* e 2.55ns per *add operation ytmp_1_2*. Pertanto, calcolando la somma di tutti questi timing stimati si ottiene un periodo di clock stimato pari a 10.2ns. Nello specifico, il periodo di clock rimanente, cioè 0.008ns, evidentemente corrisponde al valore di timing relativo a *phi operation ytmp* che viene approssimato all'interno del report a 0ns. Inoltre, si può notare come l'utilizzazione delle risorse sia notevolmente aumentata. In particolare, l'utilizzazione

delle risorse, rispetto alle risorse disponibili della scheda, risultano essere pari al 5% per i DSP e al 2% per le LUT.

| Name | BRAM_18K | DSP48E | FF | LUT |
|------------------------|-----------------|---------------|-----------|------------|
| DSP | - | - | - | - |
| Expression | - | 12 | 0 | 534 |
| FIFO | - | - | - | - |
| Instance | - | - | - | 252 |
| Memory | 0 | - | - | - |
| Multiplexer | - | - | - | 99 |
| Register | - | - | 854 | 128 |
| Total | 0 | 12 | 854 | 1013 |
| Available | 280 | 220 | 106400 | 53200 |
| Utilization (%) | 0 | 5 | ~0 | 1 |

Table 55: HLS Solution 7 with columnIndex, values and x partitioning Utilization Estimates Summary

Si procede con successivi passi così da verificare se tale problematica, riguardo il periodo di clock stimato superiore a quello target, possa essere risolta dal tool, tramite ulteriori ottimizzazioni, durante la fase di Export RTL.

| RTL | Status | Latency | | | Interval | | |
|------|--------|----------------|-----|-----|-----------------|-----|-----|
| | | min | avg | max | min | avg | max |
| VHDL | Pass | 29 | 29 | 29 | NA | NA | NA |

Table 56: HLS Solution 7 with columnIndex, values and x partitioning C/RTL Cosimulation Summary

Si può notare come il tool sia riuscito a risolvere la problematica riguardante il periodo di clock stimato superiore a quello target. Infatti, si evidenzia come quello raggiunto post-implementation risulta essere pari a $8.110ns$. Bisogna notare, però, che l'utilizzazione delle risorse risulta essere notevolmente alta dal momento che il tool è riuscito ad attuare i partizionamenti di fattore pari a 4 su tutti e tre gli array precedentemente citati.

| Resource | VHDL |
|-----------------|-------------|
| SLICE | 314 |
| LUT | 915 |
| FF | 297 |
| DSP | 12 |
| BRAM | 0 |
| SRL | 0 |

Table 57: HLS Solution 7 with columnIndex, values and x partitioning Export RTL Resource Usage

| Timing | VHDL |
|---------------------------------|-------------|
| CP required | 10.000 |
| CP achieved post-synthesis | 7.449 |
| CP achieved post-implementation | 8.110 |

Table 58: HLS Solution 7 with columnIndex, values and x partitioning Export RTL Final Timing

5.8 Solution 8

Nella soluzione hardware in questione verrà utilizzata la direttiva di pipeline, la direttiva di unrolling di fattore pari a 2 e la direttiva di partizionamento di tipologia cyclic. Nello specifico, verranno analizzate le seguenti implementazioni relative al loop2:

- Pipeline, Unroll=2, Cyclic=4 (columnIndex, values, x)
- Pipeline, Unroll=2, Cyclic=4 (columnIndex)
- Pipeline, Unroll=2, Cyclic=4 (values)
- Pipeline, Unroll=2, Cyclic=4 (x)

In particolare, è possibile evidenziare nel dettaglio le differenti soluzioni hardware nei seguenti allegati.

```
1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       #pragma HLS pipeline
9       #pragma HLS unroll factor=2
10      #pragma HLS array_partition variable=columnIndex cyclic factor=4
11      #pragma HLS array_partition variable=values cyclic factor=4
12      #pragma HLS array_partition variable=x cyclic factor=4
13      ytmp += values[k] * x[columnIndex[k]];
14    }
15    y[i] = ytmp;
16  }
17 }
```

```
1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       #pragma HLS pipeline
9       #pragma HLS unroll factor=2
10      #pragma HLS array_partition variable=columnIndex cyclic factor=4
11      ytmp += values[k] * x[columnIndex[k]];
12    }
13    y[i] = ytmp;
14  }
15 }
```

```
1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       #pragma HLS pipeline
9       #pragma HLS unroll factor=2
10      #pragma HLS array_partition variable=values cyclic factor=4
11      ytmp += values[k] * x[columnIndex[k]];
12    }
13    y[i] = ytmp;
14  }
15 }
```

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       #pragma HLS pipeline
9       #pragma HLS unroll factor=2
10      #pragma HLS array_partition variable=x cyclic factor=4
11      ytmp += values[k] * x[columnIndex[k]];
12    }
13    y[i] = ytmp;
14  }
15 }

```

Effettuando la sintesi è possibile evidenziare il seguente report:

| Solution | Clock | Target | Estimated | Uncertainty |
|------------------------|--------|--------|-----------|-------------|
| columnIndex, values, x | ap_clk | 10.00 | 8.510 | 1.25 |
| columnIndex | ap_clk | 10.00 | 8.510 | 1.25 |
| values | ap_clk | 10.00 | 8.510 | 1.25 |
| x | ap_clk | 10.00 | 8.510 | 1.25 |

Table 59: HLS Solution 8 Timing Summary (ns)

| Solution | Latency | | Interval | |
|------------------------|---------|-----|----------|-----|
| | min | max | min | max |
| columnIndex, values, x | 29 | 37 | 29 | 37 |
| columnIndex | 33 | 41 | 33 | 41 |
| values | 33 | 41 | 33 | 41 |
| x | 29 | 37 | 29 | 37 |

Table 60: HLS Solution 8 Latency Summary (clock cycles)

Si può notare come i valori di latenza totale, delle latenze totali dei loop corrispondenti e di Iteration Latency risultano essere differenti tra le varie solution. In particolare, la soluzione che implementa il partizionamento dei tre array (columnIndex, values e x) presenta valori di latenza uguali a quelli della soluzione che implementa il partizionamento dell'array x. Bisogna notare, però, che il valore del trip count risulta essere il medesimo per ogni solution. In particolare, essendo previsto un unrolling di fattore pari a 2 all'interno del loop2, il corrispondente valore risulta essere dimezzato rispetto alla solution 1 in cui non è presente alcun pragma di parallelismo.

| Solution | Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|------------------------|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | | min | max | | achieved | target | |
| columnIndex, values, x | - loop1 | 28 | 36 | 7~9 | - | - | 4 |
| | + loop2 | 3 | 5 | 4 | 1 | 1 | 0~2 |
| columnIndex | - loop1 | 32 | 40 | 8~10 | - | - | 4 |
| | + loop2 | 4 | 6 | 5 | 1 | 1 | 0~2 |
| values | - loop1 | 32 | 40 | 8~10 | - | - | 4 |
| | + loop2 | 4 | 6 | 5 | 1 | 1 | 0~2 |
| x | - loop1 | 28 | 36 | 7~9 | - | - | 4 |
| | + loop2 | 3 | 5 | 4 | 1 | 1 | 0~2 |

Table 61: HLS Solution 8 Latency Loops Summary

| Solution | BRAM_18K | DSP48E | FF | LUT |
|------------------------|-----------------|---------------|-----------|------------|
| columnIndex, values, x | 0 | 6 | 630 | 628 |
| columnIndex | 0 | 6 | 535 | 560 |
| values | 0 | 6 | 634 | 581 |
| x | 0 | 6 | 596 | 441 |

Table 62: HLS Solution 8 Utilization Estimates [#]

| Solution | RTL | Status | Latency | | | Interval | | |
|------------------------|------------|---------------|----------------|-----|-----|-----------------|-----|-----|
| | | | min | avg | max | min | avg | max |
| columnIndex, values, x | VHDL | Pass | 33 | 33 | 33 | NA | NA | NA |
| columnIndex | VHDL | Pass | 37 | 37 | 37 | NA | NA | NA |
| values | VHDL | Pass | 37 | 37 | 37 | NA | NA | NA |
| x | VHDL | Pass | 33 | 33 | 33 | NA | NA | NA |

Table 63: HLS Solution 8 C/RTL Cosimulation Report

Si può notare come, in corrispondenza della soluzione hardware basata sul partizionamento dei tre array, si ha la maggiore utilizzazione di slice, LUT e FF. Inoltre, si può evidenziare come le soluzioni basate rispettivamente sul partitioning di columnIndex e values risultano avere pressoché la medesima utilizzazione dal momento che le dimensioni di tali strutture dati è la medesima. In particolare, rispetto alla soluzione hardware similare, dove era stato utilizzato un fattore di partitioning pari a 2 di tipologia cyclic (solution 6), si può notare, in corrispondenza del partizionamento dei tre array, un incremento dei FF di circa il 59%, un incremento delle LUT e delle slice di circa il 9%.

| Solution | SLICE | LUT | FF | DSP | BRAM | CP required | CP achieved post- synthesis | CP achieved post- implementation |
|------------------------|--------------|------------|-----------|------------|-------------|------------------------|--|---|
| columnIndex, values, x | 123 | 343 | 315 | 6 | 0 | 10 | 6.540 | 6.571 |
| columnIndex | 85 | 261 | 226 | 6 | 0 | 10 | 7.496 | 7.654 |
| values | 92 | 274 | 196 | 6 | 0 | 10 | 7.927 | 7.780 |
| x | 104 | 248 | 313 | 6 | 0 | 10 | 6.540 | 6.844 |

Table 64: HLS Solution 8 Export RTL Report

5.9 Solution 9

Qui, di seguito, viene riportata l'architettura relativa alla solution 9.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4     loop1: for (int i=0; i<rows; i++) {
5         DTYPE ytmp = 0;
6         loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7             #pragma HLS loop_tripcount min=0 max=4 avg=2
8             #pragma HLS pipeline
9             #pragma HLS unroll factor=8
10            ytmp += values[k] * x[columnIndex[k]];
11        }
12        y[i] = ytmp;
13    }
14 }

```

In particolare, rispetto alla soluzione hardware 5 e 7, dove rispettivamente era stato considerato un parallelismo di fattore pari a 2 e un parallelismo di fattore pari a 4, in questa solution è stato considerato un unrolling di fattore pari a 8. In particolare, ciò che ci si aspetta è un aumento delle risorse ed eventuali problematiche relative al timing, similari a quelle riscontrate nella solution 7, dal momento che il tool deve gestire all'interno del loop2 più accessi in memoria paralleli.

Effettuando la sintesi è possibile evidenziare il seguente log nella console:

WARNING: [SCHD 204-69] Unable to schedule 'load' operation ('columnIndex_load_6', smvmProject/s-mvm.cpp:34) on array 'columnIndex' due to limited memory ports. Please consider using a memory core with more ports or partitioning the array 'columnIndex'.

Tale log sta a significare che non riesce a schedulare correttamente, dal punto di vista degli accessi in memoria, la load operation relativa all'array columnIndex dato dal numero limitato di porte relative alla memoria. In particolare, analizzando il report relativo alla sintesi, si può notare come l'Initiation Interval, associato al loop2, raggiunto risulta essere maggiore di quello target.

| Clock | Target | Estimated | Uncertainty |
|--------|--------|-----------|-------------|
| ap_clk | 10.00 | 8.510 | 1.25 |

Table 65: HLS Solution 9 Timing Summary (ns)

| Latency | | Interval | |
|---------|-----|----------|-----|
| min | max | min | max |
| 45 | 61 | 45 | 61 |

Table 66: HLS Solution 9 Latency Summary (clock cycles)

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 44 | 60 | 11~15 | - | - | 4 |
| + loop2 | 7 | 11 | 8 | 4 | 1 | 0~1 |

Table 67: HLS Solution 9 Latency Loops Summary

Pertanto, si potrebbe aggiungere una direttiva di partizionamento con fattore pari a 8 relativa all'array menzionato all'interno del log, cioè columnIndex.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4     loop1: for (int i=0; i<rows; i++) {
5         DTYPE ytmp = 0;
6         loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7             #pragma HLS loop_tripcount min=0 max=4 avg=2

```

```

8      #pragma HLS pipeline
9      #pragma HLS unroll factor=8
10     #pragma HLS array_partition variable=columnIndex complete
11     ytmp += values[k] * x[columnIndex[k]];
12 }
13 y[i] = ytmp;
14 }
15 }

```

Effettuando nuovamente la sintesi, si ottiene il seguente log nella console e i seguenti valori di latenza.
WARNING: [SCH204-69] Unable to schedule 'load' operation ('values_load_6', smvmProject/smvm.cpp:34) on array 'values' due to limited memory ports. Please consider using a memory core with more ports or partitioning the array 'values'.

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 40 | 56 | 10~14 | - | - | 4 |
| + loop2 | 6 | 10 | 7 | 4 | 1 | 0~1 |

Table 68: HLS Solution 9 with columnIndex partitioning Latency Loops Summary

Si può notare come in questo caso il warning sia relativo all'array values. In particolare, la tipologia di warning è la medesima facendo presupporre che il tool non riesca a schedulare correttamente, secondo le direttive imposte dall'architettura, gli accessi in parallelo all'array values. Infatti, il valore di Initiation Interval raggiunto risulta essere ancora maggiore di quello di quello target. Pertanto, si potrebbe aggiungere una direttiva di partizionamento relativa all'array menzionato all'interno del log, cioè values.

```

1  #include "definitions.h"
2
3  void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
4  x[size]) {
5      loop1: for (int i=0; i<rows; i++) {
6          DTYPE ytmp = 0;
7          loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
8              #pragma HLS loop_tripcount min=0 max=4 avg=2
9              #pragma HLS pipeline
10             #pragma HLS unroll factor=8
11             #pragma HLS array_partition variable=columnIndex complete
12             #pragma HLS array_partition variable=values complete
13             ytmp += values[k] * x[columnIndex[k]];
14         }
15         y[i] = ytmp;
16     }
17 }

```

Effettuando nuovamente la sintesi, si ottiene il seguente log nella console e i seguenti valori di latenza.
WARNING: [SCH204-69] Unable to schedule 'load' operation ('x_load_6', smvmProject/smvm.cpp:34) on array 'x' due to limited memory ports. Please consider using a memory core with more ports or partitioning the array 'x'.

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 40 | 56 | 10~14 | - | - | 4 |
| + loop2 | 6 | 10 | 7 | 4 | 1 | 0~1 |

Table 69: HLS Solution 9 with columnIndex and values partitioning Latency Loops Summary

Si può notare come in questo caso il warning sia relativo all'array x. In particolare, la tipologia di warning è la medesima della precedente. Anche in questo caso il valore di Initiation Interval raggiunto risulta essere

ancora maggiore di quello di quello target.

Pertanto, si potrebbe aggiungere una direttiva di partizionamento relativa all'array menzionato all'interno del log, cioè x.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       #pragma HLS pipeline
9       #pragma HLS unroll factor=8
10      #pragma HLS array_partition variable=columnIndex complete
11      #pragma HLS array_partition variable=values complete
12      #pragma HLS array_partition variable=x complete
13      ytmp += values[k] * x[columnIndex[k]];
14    }
15    y[i] = ytmp;
16  }
17 }

```

Effettuando nuovamente la sintesi, si ottiene il seguente log nella console e il seguente report.

WARNING: [SCHED 204-68] The II Violation in module 'smvm': Unable to enforce a carried dependence constraint (II = 1, distance = 1, offset = 0) between 'add' operation ('ytmp_1.7', smvmProject/svm.cpp:34) and 'add' operation ('ytmp_1', smvmProject/svm.cpp:34).

WARNING: [SCHED 204-68] The II Violation in module 'smvm': Unable to enforce a carried dependence constraint (II = 1, distance = 1, offset = 1) between 'add' operation ('ytmp_1.5', smvmProject/svm.cpp:34) and 'add' operation ('ytmp_1.2', smvmProject/svm.cpp:34).

WARNING: [SCHED 204-68] The II Violation in module 'smvm': Unable to enforce a carried dependence constraint (II = 2, distance = 1, offset = 1) between 'add' operation ('ytmp_1.6', smvmProject/svm.cpp:34) and 'add' operation ('ytmp_1.2', smvmProject/svm.cpp:34).

In particolare, tali warning suggeriscono problematiche relative alla variabile ytmp. Nello specifico, ytmp è una variabile temporanea (di appoggio) di tipo DTYPE, cioè int, e pertanto il partizionamento su tale variabile non avrebbe senso poiché quest'ultimo funziona solo con gli array. Bisogna notare però che il problema non è relativo a ytmp ma è associato a y, cioè l'output, che è un vettore. Inoltre, y fa riferimento al loop1 e questo suggerisce che gli eventuali problemi sono relativi al ciclo1 ed, essendo che il loop1 presenta al suo interno il loop2, allora tali problemi si ripercuotono anche sul loop2. In aggiunta, si evidenzia come in questo caso l'Initiation Interval raggiunto risulta essere decrementato da 4 a 3 ma comunque ancora maggiore di quello target.

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 32 | 44 | 8~11 | - | - | 4 |
| + loop2 | 4 | 7 | 5 | 3 | 1 | 0~1 |

Table 70: HLS Solution 9 with columnIndex, values and x partitioning Latency Loops Summary

Un'idea potrebbe quella di effettuare un partizionamento anche sull'array rowPtr.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       #pragma HLS pipeline

```

```

9      #pragma HLS unroll factor=8
10     #pragma HLS array_partition variable=columnIndex complete
11     #pragma HLS array_partition variable=values complete
12     #pragma HLS array_partition variable=x complete
13     #pragma HLS array_partition variable=rowPtr complete
14     ytmp += values[k] * x[columnIndex[k]];
15 }
16 y[i] = ytmp;
17 }
18 }

```

Effettuando nuovamente la sintesi, si ottiene il seguente log nella console e il seguente report.

WARNING: [SCHED 204-68] The II Violation in module 'smvm': Unable to enforce a carried dependence constraint (II = 1, distance = 1, offset = 0) between 'add' operation ('ytmp_1.7', smvmProject/smvm.cpp:34) and 'add' operation ('ytmp_1', smvmProject/smvm.cpp:34).

WARNING: [SCHED 204-68] The II Violation in module 'smvm': Unable to enforce a carried dependence constraint (II = 1, distance = 1, offset = 1) between 'add' operation ('ytmp_1.5', smvmProject/smvm.cpp:34) and 'add' operation ('ytmp_1.2', smvmProject/smvm.cpp:34).

WARNING: [SCHED 204-68] The II Violation in module 'smvm': Unable to enforce a carried dependence constraint (II = 2, distance = 1, offset = 1) between 'add' operation ('ytmp_1.6', smvmProject/smvm.cpp:34) and 'add' operation ('ytmp_1.2', smvmProject/smvm.cpp:34).

In particolare, si riscontrano i medesimi warning allegati precedentemente.

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 28 | 40 | 7~10 | - | - | 4 |
| + loop2 | 4 | 7 | 5 | 3 | 1 | 0~1 |

Table 71: HLS Solution 9 with columnIndex, values, x and rowPtr partitioning Latency Loops Summary

Pertanto, a tale proposito si potrebbe effettuare un partizionamento anche su y.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4     loop1: for (int i=0; i<rows; i++) {
5         DTYPE ytmp = 0;
6         loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7             #pragma HLS loop_tripcount min=0 max=4 avg=2
8             #pragma HLS pipeline
9             #pragma HLS unroll factor=8
10            #pragma HLS array_partition variable=columnIndex complete
11            #pragma HLS array_partition variable=values complete
12            #pragma HLS array_partition variable=x complete
13            #pragma HLS array_partition variable=rowPtr complete
14            #pragma HLS array_partition variable=y complete
15            ytmp += values[k] * x[columnIndex[k]];
16        }
17        y[i] = ytmp;
18    }
19 }

```

Effettuando nuovamente la sintesi, si ottiene il seguente log nella console e il seguente report.

WARNING: [SCHED 204-68] The II Violation in module 'smvm': Unable to enforce a carried dependence constraint (II = 1, distance = 1, offset = 0) between 'add' operation ('ytmp_1.7', smvmProject/smvm.cpp:35) and 'add' operation ('ytmp_1', smvmProject/smvm.cpp:35).

WARNING: [SCHED 204-68] The II Violation in module 'smvm': Unable to enforce a carried dependence constraint (II = 1, distance = 1, offset = 1) between 'add' operation ('ytmp_1.5', smvmProject/smvm.cpp:35) and 'add' operation ('ytmp_1.2', smvmProject/smvm.cpp:35).

WARNING: [SCHED 204-68] The II Violation in module 'smvm': Unable to enforce a carried dependence constraint (II = 2, distance = 1, offset = 1) between 'add' operation ('ytmp_1.6', smvmProject/smvm.cpp:35)

and 'add' operation ('ytmp_1_2', smvmProject/smvm.cpp:35).

Nello specifico, si riscontrano nuovamente i medesimi warning allegati precedentemente.

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 28 | 40 | 7~10 | - | - | 4 |
| + loop2 | 4 | 7 | 5 | 3 | 1 | 0~1 |

Table 72: HLS Solution 9 with columnIndex, values, x, rowPtr and y partitioning Latency Loops Summary

Pertanto, dal momento che i warning sono ancora presenti e che il valore di Initiation Interval non tende a diminuire, si potrebbe provare effettuando anche il pipelining del loop1. In particolare, dal momento che i warning sono i medesimi, la solution verrà considerata senza la direttiva di partitioning su rowPtr e y.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     #pragma HLS pipeline
6     DTYPE ytmp = 0;
7     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
8       #pragma HLS loop_tripcount min=0 max=4 avg=2
9       #pragma HLS pipeline
10      #pragma HLS unroll factor=8
11      #pragma HLS array_partition variable=columnIndex complete
12      #pragma HLS array_partition variable=values complete
13      #pragma HLS array_partition variable=x complete
14      ytmp += values[k] * x[columnIndex[k]];
15    }
16    y[i] = ytmp;
17  }
18 }

```

Effettuando nuovamente la sintesi, si ottiene il seguente log nella console e il seguente report.

WARNING: [XFORM 203-503] Ignored partial unroll directive for loop 'loop2' (smvmProject/smvm.cpp:21) because its parent loop or function is pipelined.

WARNING: [XFORM 203-503] Cannot unroll loop 'loop2' (smvmProject/smvm.cpp:21) in function 'smvm' completely: variable loop bound.

WARNING: [SCHED 204-65] Unable to satisfy pipeline directive: Loop contains subloop(s) not being unrolled or flattened.

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 16 | 36 | 4~9 | - | - | 4 |
| + loop2 | 0 | 5 | 3 | 1 | 1 | 0~4 |

Table 73: HLS Solution 9 with columnIndex, values and x partitioning and loop1 pipelined Latency Loops Summary

| Name | BRAM_18K | DSP48E | FF | LUT |
|------------------------|----------|--------|--------|-------|
| DSP | - | - | - | - |
| Expression | - | 3 | 0 | 141 |
| FIFO | - | - | - | - |
| Instance | - | - | - | 63 |
| Memory | 0 | - | - | - |
| Multiplexer | - | - | - | 78 |
| Register | - | - | 211 | - |
| Total | 0 | 3 | 211 | 282 |
| Available | 280 | 220 | 106400 | 53200 |
| Utilization (%) | 0 | 1 | ~0 | ~0 |

Table 74: HLS Solution 9 with columnIndex, values and x partitioning and loop1 pipelined Utilization Estimates Summary

Si può notare come i warning relativi a ytmp non sono più presenti all'interno della console e che, inoltre, l'Initiation Interval, relativo al loop2, presenta un valore raggiunto uguale a quello target, cioè pari a 1. In aggiunta, si possono evidenziare ulteriori warning all'interno della console. In particolare, il tool non è riuscito a soddisfare la richiesta di pipeline a causa dei bound non noti e, pertanto, non riuscendo nemmeno a soddisfare la direttiva di unrolling all'interno del loop2. Infatti, si può notare come i valori di Initiation Interval (achieved e target) del loop1 risultano essere non definiti, mentre il trip count relativo al loop2 risulta essere il medesimo di quello della solution2 dove non era presente alcun pragma di parallelismo nel ciclo 2.

Pertanto, considerando l'implementazione finale ottenuta in corrispondenza della solution in questione, si effettua la C/RTL Cosimulation e l'Export RTL e si analizzano, di conseguenza, i report corrispondenti.

| RTL | Status | Latency | | | Interval | | |
|------|--------|---------|-----|-----|----------|-----|-----|
| | | min | avg | max | min | avg | max |
| VHDL | Pass | 30 | 30 | 30 | NA | NA | NA |

Table 75: HLS Solution 9 with columnIndex, values and x partitioning and loop1 pipelined C/RTL Cosimulation Summary

| Resource | VHDL |
|----------|------|
| SLICE | 75 |
| LUT | 248 |
| FF | 131 |
| DSP | 3 |
| BRAM | 0 |
| SRL | 0 |

Table 76: HLS Solution 9 with columnIndex, values and x partitioning and loop1 pipelined Export RTL Resource Usage

| Timing | VHDL |
|---------------------------------|--------|
| CP required | 10.000 |
| CP achieved post-synthesis | 5.745 |
| CP achieved post-implementation | 6.120 |

Table 77: HLS Solution 9 with columnIndex, values and x partitioning and loop1 pipelined Export RTL Final Timing

Inoltre, si potrebbe provare a considerare il pipeline all'interno del loop1 e il partitioning anche per gli array rowPtr e y così da verificare se è possibile ottenere ulteriori ottimizzazioni.

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4     loop1: for (int i=0; i<rows; i++) {
5         #pragma HLS pipeline
6         DTYPE ytmp = 0;

```

```

7   loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
8       #pragma HLS loop_tripcount min=0 max=4 avg=2
9       #pragma HLS pipeline
10      #pragma HLS unroll factor=8
11      #pragma HLS array_partition variable=columnIndex complete
12      #pragma HLS array_partition variable=values complete
13      #pragma HLS array_partition variable=x complete
14      #pragma HLS array_partition variable=rowPtr complete
15      #pragma HLS array_partition variable=y complete
16      ytmp += values[k] * x[columnIndex[k]];
17  }
18  y[i] = ytmp;
19  }
20  }

```

Effettuando nuovamente la sintesi, si ottiene il seguente log nella console e il seguente report.

WARNING: [XFORM 203-503] Ignored partial unroll directive for loop 'loop2' (smvmProject/smvm.cpp:21) because its parent loop or function is pipelined.

WARNING: [XFORM 203-503] Cannot unroll loop 'loop2' (smvmProject/smvm.cpp:21) in function 'smvm' completely: variable loop bound.

WARNING: [SCHD 204-65] Unable to satisfy pipeline directive: Loop contains subloop(s) not being unrolled or flattened.

| Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | min | max | | achieved | target | |
| - loop1 | 12 | 32 | 3~8 | - | - | 4 |
| + loop2 | 0 | 5 | 3 | 1 | 1 | 0~4 |

Table 78: HLS Solution 9 with columnIndex, values, x, rowPtr and y partitioning and loop1 pipelined Latency Loops Summary

| Name | BRAM_18K | DSP48E | FF | LUT |
|------------------------|----------|--------|--------|-------|
| DSP | - | - | - | - |
| Expression | - | 3 | 0 | 143 |
| FIFO | - | - | - | - |
| Instance | - | - | - | 84 |
| Memory | 0 | - | - | - |
| Multiplexer | - | - | - | 99 |
| Register | - | - | 175 | - |
| Total | 0 | 3 | 175 | 326 |
| Available | 280 | 220 | 106400 | 53200 |
| Utilization (%) | 0 | 1 | ~0 | ~0 |

Table 79: HLS Solution 9 with columnIndex, values, x, rowPtr and y partitioning and loop1 pipelined Utilization Estimates Summary

Pertanto, si effettua la C/RTL Cosimulation e l'Export RTL e si analizzano, di conseguenza, i report corrispondenti.

| RTL | Status | Latency | | | Interval | | |
|------|--------|---------|-----|-----|----------|-----|-----|
| | | min | avg | max | min | avg | max |
| VHDL | Pass | 26 | 26 | 26 | NA | NA | NA |

Table 80: HLS Solution 9 with columnIndex, values, x, rowPtr and y partitioning and loop1 pipelined C/RTL Cosimulation Summary

| Resource | VHDL |
|----------|------|
| SLICE | 101 |
| LUT | 345 |
| FF | 96 |
| DSP | 3 |
| BRAM | 0 |
| SRL | 0 |

Table 81: HLS Solution 9 with columnIndex, values, x, rowPtr and y partitioning and loop1 pipelined Export RTL Resource Usage

| Timing | VHDL |
|---------------------------------|--------|
| CP required | 10.000 |
| CP achieved post-synthesis | 7.603 |
| CP achieved post-implementation | 7.530 |

Table 82: HLS Solution 9 with columnIndex, values, x, rowPtr and y partitioning and loop1 pipelined Export RTL Final Timing

Si può notare come la latenza ottenuta risulta essere minore rispetto a quella ottenuta considerando soltanto i partitioning degli array columnIndex, values e x. Inoltre, l'utilizzazione delle risorse risulta essere leggermente maggiore. In particolare, si ha un aumento di circa il 35% delle slice, un aumento di circa il 39% delle LUT e una diminuzione di circa il 27% dei FF. Ovviamente il numero di DSP rimane il medesimo a causa dei warning precedentemente citati. Pertanto, dal momento che l'utilizzazione delle risorse è leggermente aumentata e la latenza è diminuita, come risultato della solution 9 verrà considerata quest'ultima soluzione hardware dove è stato considerato anche il partitioning degli array rowPtr e y.

5.10 Solution 10

Nella soluzione hardware in questione verrà utilizzata la direttiva di pipeline, di unrolling di fattore pari a 2 e la direttiva di partizionamento di tipologia cyclic. Nello specifico, verranno analizzate le seguenti implementazioni relative al loop2:

- Pipeline, Unroll=2, Cyclic=8 (columnIndex, values, x)
- Pipeline, Unroll=2, Cyclic=8 (columnIndex)
- Pipeline, Unroll=2, Cyclic=8 (values)
- Pipeline, Unroll=2, Cyclic=8 (x)

In particolare, è possibile evidenziare nel dettaglio le differenti soluzioni hardware nei seguenti allegati.

```
1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       #pragma HLS pipeline
9       #pragma HLS unroll factor=2
10      #pragma HLS array_partition variable=columnIndex cyclic factor=8
11      #pragma HLS array_partition variable=values cyclic factor=8
12      #pragma HLS array_partition variable=x cyclic factor=8
13      ytmp += values[k] * x[columnIndex[k]];
14    }
15    y[i] = ytmp;
16  }
17 }
```

```
1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       #pragma HLS pipeline
9       #pragma HLS unroll factor=2
10      #pragma HLS array_partition variable=columnIndex cyclic factor=8
11      ytmp += values[k] * x[columnIndex[k]];
12    }
13    y[i] = ytmp;
14  }
15 }
```

```
1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=4 avg=2
8       #pragma HLS pipeline
9       #pragma HLS unroll factor=2
10      #pragma HLS array_partition variable=values cyclic factor=8
11      ytmp += values[k] * x[columnIndex[k]];
12    }
13    y[i] = ytmp;
14  }
15 }
```

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4     loop1: for (int i=0; i<rows; i++) {
5         DTYPE ytmp = 0;
6         loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7             #pragma HLS loop_tripcount min=0 max=4 avg=2
8             #pragma HLS pipeline
9             #pragma HLS unroll factor=2
10            #pragma HLS array_partition variable=x cyclic factor=8
11            ytmp += values[k] * x[columnIndex[k]];
12        }
13        y[i] = ytmp;
14    }
15 }

```

Effettuando la sintesi si ottiene il seguente log.

ERROR: [XFORM 203-103] Cannot partition array 'x' (smvmProject/svm.cpp:11): incorrect partition factor 8.

In particolare, la console sta segnalando che effettivamente non riesce a partizionare l'array x dal momento che la dimensione dell'array risulta essere non compatibile con il fattore di partitioning dichiarato. Infatti, la dimensione di x inizialmente dichiarata in definitions.h è pari a 4 mentre il fattore di partizionamento che si sta utilizzando è pari a 8. A questo proposito si potrebbe modificare il valore di dimensionamento relativo a x all'interno dell'header. In particolare è necessario modificare sia il valore di size sia il valore di rows. Inoltre, bisogna anche modificare i parametri dichiarati all'interno della direttiva trip count dal momento che il numero di righe e, quindi, di iterazioni risulta essere differente.

```

1 #ifndef DEFINITIONS_H
2
3
4 /**
5  * Square Matrix Size.
6  */
7 const static int size = 8;
8
9 /**
10 * Number of Non-Zero Elements.
11 */
12 const static int nnz = 9;
13
14 /**
15 * Number of Rows.
16 */
17 const static int rows = 8;
18
19 ...
20
21 #endif

```

```

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4     loop1: for (int i=0; i<rows; i++) {
5         DTYPE ytmp = 0;
6         loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7             #pragma HLS loop_tripcount min=0 max=8 avg=4
8             #pragma HLS pipeline
9             #pragma HLS unroll factor=2
10            #pragma HLS array_partition variable=columnIndex cyclic factor=8
11            #pragma HLS array_partition variable=values cyclic factor=8
12            #pragma HLS array_partition variable=x cyclic factor=8
13            ytmp += values[k] * x[columnIndex[k]];
14        }
15        y[i] = ytmp;

```

```

16 }
17 }

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=8 avg=4
8       #pragma HLS pipeline
9       #pragma HLS unroll factor=2
10      #pragma HLS array_partition variable=columnIndex cyclic factor=8
11      ytmp += values[k] * x[columnIndex[k]];
12    }
13    y[i] = ytmp;
14  }
15 }

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=8 avg=4
8       #pragma HLS pipeline
9       #pragma HLS unroll factor=2
10      #pragma HLS array_partition variable=values cyclic factor=8
11      ytmp += values[k] * x[columnIndex[k]];
12    }
13    y[i] = ytmp;
14  }
15 }

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4   loop1: for (int i=0; i<rows; i++) {
5     DTYPE ytmp = 0;
6     loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7       #pragma HLS loop_tripcount min=0 max=8 avg=4
8       #pragma HLS pipeline
9       #pragma HLS unroll factor=2
10      #pragma HLS array_partition variable=x cyclic factor=8
11      ytmp += values[k] * x[columnIndex[k]];
12    }
13    y[i] = ytmp;
14  }
15 }

```

Pertanto, effettuando la sintesi si ottiene il seguente report.

| Solution | Clock | Target | Estimated | Uncertainty |
|------------------------|--------|--------|-----------|-------------|
| columnIndex, values, x | ap_clk | 10.00 | 8.510 | 1.25 |
| columnIndex | ap_clk | 10.00 | 8.510 | 1.25 |
| values | ap_clk | 10.00 | 8.510 | 1.25 |
| x | ap_clk | 10.00 | 8.510 | 1.25 |

Table 83: HLS Solution 10 Timing Summary (ns)

| Solution | Latency | | Interval | |
|------------------------|---------|-----|----------|-----|
| | min | max | min | max |
| columnIndex, values, x | 57 | 89 | 57 | 89 |
| columnIndex | 65 | 97 | 65 | 97 |
| values | 65 | 97 | 65 | 97 |
| x | 57 | 89 | 57 | 89 |

Table 84: HLS Solution 10 Latency Summary (clock cycles)

| Solution | Loop Name | Latency | | Iteration Latency | Initiation achieved | Interval target | Trip Count |
|------------------------|-----------|---------|-----|-------------------|---------------------|-----------------|------------|
| | | min | max | | | | |
| columnIndex, values, x | - loop1 | 56 | 88 | 7~11 | - | - | 8 |
| | + loop2 | 3 | 7 | 4 | 1 | 1 | 0~4 |
| columnIndex | - loop1 | 64 | 96 | 8~12 | - | - | 8 |
| | + loop2 | 4 | 8 | 5 | 1 | 1 | 0~4 |
| values | - loop1 | 64 | 96 | 8~12 | - | - | 8 |
| | + loop2 | 4 | 8 | 5 | 1 | 1 | 0~4 |
| x | - loop1 | 56 | 88 | 7~11 | - | - | 8 |
| | + loop2 | 3 | 7 | 4 | 1 | 1 | 0~4 |

Table 85: HLS Solution 10 Latency Loops Summary

| Solution | BRAM_18K | DSP48E | FF | LUT |
|------------------------|----------|--------|-----|-----|
| columnIndex, values, x | 0 | 6 | 762 | 900 |
| columnIndex | 0 | 6 | 539 | 752 |
| values | 0 | 6 | 640 | 773 |
| x | 0 | 6 | 727 | 492 |

Table 86: HLS Solution 10 Utilization Estimates [#]

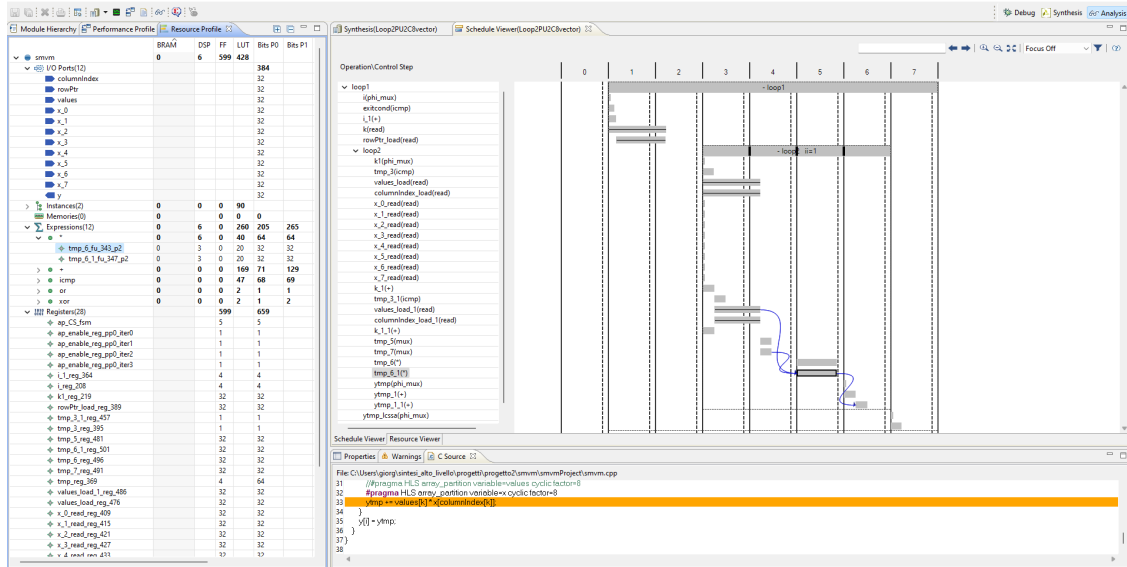


Figure 7: HLS Solution 10 Analysis

In particolare, dall'interfaccia Analysis sotto allegata (corrispondente al partitioning dell'array x), si può notare come l'unrolling di fattore 2 sia stato applicato correttamente dal momento che vengono effettuate 2 operazioni di moltiplicazione e 2 operazioni di somma in un'unica iterazione del loop2. Nello specifico, si può evidenziare come i prodotti richiedano l'utilizzo di 3 DSP ognuno tale da giustificare l'utilizzazione di tali risorse pari a 6 come riportato nel report di sintesi.

| Solution | RTL | Status | Latency | | | Interval | | |
|------------------------|------|--------|---------|-----|-----|----------|-----|-----|
| | | | min | avg | max | min | avg | max |
| columnIndex, values, x | VHDL | Pass | 61 | 61 | 61 | NA | NA | NA |
| columnIndex | VHDL | Pass | 69 | 69 | 69 | NA | NA | NA |
| values | VHDL | Pass | 69 | 69 | 69 | NA | NA | NA |
| x | VHDL | Pass | 61 | 61 | 61 | NA | NA | NA |

Table 87: HLS Solution 10 C/RTL Cosimulation Report

Si può notare come l'utilizzazione delle risorse sia aumentata rispetto alle soluzioni 6 (loop2 Pipeline, Unroll=2, Cyclic=2) e 8 (loop2 Pipeline, Unroll=2, Cyclic=4). In particolare, considerando le soluzioni hardware corrispondenti al partizionamento di tutti e tre gli array, si evidenzia rispettivamente un aumento di circa l'81% e di circa il 66% in corrispondenza delle slice rispetto alla soluzione 6 e 8, un aumento di circa il 77% e di circa il 63% in corrispondenza delle LUT, un aumento di circa il 127% e di circa il 43%. Per quanto riguarda le soluzioni hardware corrispondenti al partizionamento di columnIndex, rispetto alla soluzione 6 e 8, si registra rispettivamente un aumento di un'unità e una diminuzione di 2 slice, un aumento di nove unità e di 7 unità delle LUT, un aumento di 6 unità e di 4 unità dei FF. Per quanto riguarda le soluzioni corrispondenti al partitioning di values, rispetto alla soluzione 6 e 8, si registra rispettivamente un aumento di circa il 18% e di circa il 27% delle slice, un aumento di 15 unità e di circa il 25% delle LUT, una diminuzione di 25 unità e un aumento di 3 unità dei FF. Per quanto riguarda le soluzioni hardware corrispondenti al partizionamento di x, rispetto alla soluzione 6 e 8, si registra rispettivamente un aumento di circa il 44% e di circa il 38% delle slice, un aumento di circa il 24% e di circa il 25% delle LUT, un aumento di circa il 124% e di circa il 42% dei FF.

| Solution | SLICE | LUT | FF | DSP | BRAM | CP required | CP achieved post- synthesis | CP achieved post- implementation |
|------------------------|-------|-----|-----|-----|------|----------------|--------------------------------------|---|
| columnIndex, values, x | 204 | 558 | 449 | 6 | 0 | 10 | 6.540 | 6.840 |
| columnIndex | 83 | 268 | 230 | 6 | 0 | 10 | 7.496 | 8.115 |
| values | 117 | 342 | 199 | 6 | 0 | 10 | 7.927 | 7.603 |
| x | 143 | 311 | 444 | 6 | 0 | 10 | 6.540 | 6.790 |

Table 88: HLS Solution 10 Export RTL Report

Ovviamente questo aumento di utilizzazione delle risorse è dovuto all'incremento del fattore di partizionamento adottato nelle soluzioni hardware. Infatti, si evidenzia che il maggiore incremento delle risorse si ha rispetto alla soluzione 6 dove era previsto un fattore pari a 2 rispetto alla soluzione in questione dove il fattore di partitioning previsto è stato di 8.

5.11 Solution 11

Considerando il codice utilizzato nella precedente solution, nella soluzione hardware in questione verrà utilizzata la direttiva di pipeline, di unrolling di fattore pari a 2 e la direttiva di partizionamento di tipologia block. Nello specifico, verranno analizzate le seguenti implementazioni relative al loop2:

- Pipeline, Unroll=2, Block=8 (columnIndex, values, x)
- Pipeline, Unroll=2, Block=8 (columnIndex)
- Pipeline, Unroll=2, Block=8 (values)
- Pipeline, Unroll=2, Block=8 (x)

In particolare, è possibile evidenziare nel dettaglio le differenti soluzioni hardware nei seguenti allegati.

```
1 #ifndef DEFINITIONS_H
2
3
4 /**
5  * Square Matrix Size.
6  */
7 const static int size = 8;
8
9 /**
10 * Number of Non-Zero Elements.
11 */
12 const static int nnz = 9;
13
14 /**
15 * Number of Rows.
16 */
17 const static int rows = 8;
18
19 ...
20
21 #endif
```

```
1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
4 x[size]) {
5     loop1: for (int i=0; i<rows; i++) {
6         DTYPE ytmp = 0;
7         loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
8             #pragma HLS loop_tripcount min=0 max=8 avg=4
9             #pragma HLS pipeline
10            #pragma HLS unroll factor=2
11            #pragma HLS array_partition variable=columnIndex block factor=8
12            #pragma HLS array_partition variable=values block factor=8
13            #pragma HLS array_partition variable=x block factor=8
14            ytmp += values[k] * x[columnIndex[k]];
15        }
16        y[i] = ytmp;
17    }
18 }
```

```
1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
4 x[size]) {
5     loop1: for (int i=0; i<rows; i++) {
6         DTYPE ytmp = 0;
7         loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
8             #pragma HLS loop_tripcount min=0 max=8 avg=4
9             #pragma HLS pipeline
10            #pragma HLS unroll factor=2
```

```

10     #pragma HLS array_partition variable=columnIndex block factor=8
11     ytmp += values[k] * x[columnIndex[k]];
12 }
13 y[i] = ytmp;
14 }
15 }

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4     loop1: for (int i=0; i<rows; i++) {
5         DTYPE ytmp = 0;
6         loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7             #pragma HLS loop_tripcount min=0 max=8 avg=4
8             #pragma HLS pipeline
9             #pragma HLS unroll factor=2
10            #pragma HLS array_partition variable=values block factor=8
11            ytmp += values[k] * x[columnIndex[k]];
12        }
13        y[i] = ytmp;
14    }
15 }

1 #include "definitions.h"
2
3 void smvm(int rowPtr[rows+1], int columnIndex[nnz], DTYPE values[nnz], DTYPE y[size], DTYPE
  x[size]) {
4     loop1: for (int i=0; i<rows; i++) {
5         DTYPE ytmp = 0;
6         loop2: for (int k=rowPtr[i]; k<rowPtr[i+1]; k++) {
7             #pragma HLS loop_tripcount min=0 max=8 avg=4
8             #pragma HLS pipeline
9             #pragma HLS unroll factor=2
10            #pragma HLS array_partition variable=x block factor=8
11            ytmp += values[k] * x[columnIndex[k]];
12        }
13        y[i] = ytmp;
14    }
15 }

```

Effettuando la sintesi si ottiene il seguente log.

WARNING: [XFORM 203-105] Cannot partition array 'columnIndex' (smvmProject/smvm.cpp:11): indivisible factor 8 on dimension 1, which has 9 elements.

WARNING: [XFORM 203-105] Cannot partition array 'values' (smvmProject/smvm.cpp:11): indivisible factor 8 on dimension 1, which has 9 elements.

In particolare, la console sta segnalando che il tool non riesce a partizionare correttamente, tramite la tipologia block e il fattore 8 dichiarato, gli array columnIndex e values dal momento che presentano un numero di elementi che risulta essere non multiplo del fattore di array partitioning dichiarato. Pertanto, a tale proposito si modifica, di conseguenza, la variabile nnz, cioè il numero di elementi non nulli all'interno della matrice, nell'header. Nello specifico è stato scelto un valore pari a 16. Pertanto, ciò che ci si aspetta è che il tool divida tali array, columnIndex e values, in 8 sub-array aventi ognuno 2 elementi. Invece, per quanto riguarda l'array x, avente dimensione pari a 8, ci si aspetta che il tool lo divida in 8 sub-array da un elemento ognuno.

```

1 #ifndef DEFINITIONS_H
2
3
4 /**
5  * Square Matrix Size.
6  */
7 const static int size = 8;
8
9 /**
10 * Number of Non-Zero Elements.
11 */

```

```

12 const static int nnz = 16;
13
14 /**
15  * Number of Rows.
16  */
17 const static int rows = 8;
18
19 ...
20
21 #endif

```

Pertanto, effettuando la sintesi si ottiene il seguente report.

| Solution | Clock | Target | Estimated | Uncertainty |
|------------------------|--------|--------|-----------|-------------|
| columnIndex, values, x | ap_clk | 10.00 | 8.510 | 1.25 |
| columnIndex | ap_clk | 10.00 | 8.510 | 1.25 |
| values | ap_clk | 10.00 | 8.510 | 1.25 |
| x | ap_clk | 10.00 | 8.510 | 1.25 |

Table 89: HLS Solution 11 Timing Summary (ns)

| Solution | Latency | | Interval | |
|------------------------|---------|-----|----------|-----|
| | min | max | min | max |
| columnIndex, values, x | 57 | 89 | 57 | 89 |
| columnIndex | 65 | 97 | 65 | 97 |
| values | 65 | 97 | 65 | 97 |
| x | 57 | 89 | 57 | 89 |

Table 90: HLS Solution 11 Latency Summary (clock cycles)

| Solution | Loop Name | Latency | | Iteration Latency | Initiation Interval | | Trip Count |
|------------------------|-----------|---------|-----|-------------------|---------------------|--------|------------|
| | | min | max | | achieved | target | |
| columnIndex, values, x | - loop1 | 56 | 88 | 7~11 | - | - | 8 |
| | + loop2 | 3 | 7 | 4 | 1 | 1 | 0~4 |
| columnIndex | - loop1 | 64 | 96 | 8~12 | - | - | 8 |
| | + loop2 | 4 | 8 | 5 | 1 | 1 | 0~4 |
| values | - loop1 | 64 | 96 | 8~12 | - | - | 8 |
| | + loop2 | 4 | 8 | 5 | 1 | 1 | 0~4 |
| x | - loop1 | 56 | 88 | 7~11 | - | - | 8 |
| | + loop2 | 3 | 7 | 4 | 1 | 1 | 0~4 |

Table 91: HLS Solution 11 Latency Loops Summary

| Solution | BRAM_18K | DSP48E | FF | LUT |
|------------------------|----------|--------|-----|-----|
| columnIndex, values, x | 0 | 6 | 789 | 674 |
| columnIndex | 0 | 6 | 536 | 503 |
| values | 0 | 6 | 597 | 503 |
| x | 0 | 6 | 727 | 492 |

Table 92: HLS Solution 11 Utilization Estimates [#]

| Solution | RTL | Status | Latency | | | Interval | | |
|------------------------|------|--------|---------|-----|-----|----------|-----|-----|
| | | | min | avg | max | min | avg | max |
| columnIndex, values, x | VHDL | Pass | 64 | 64 | 64 | NA | NA | NA |
| columnIndex | VHDL | Pass | 72 | 72 | 72 | NA | NA | NA |
| values | VHDL | Pass | 72 | 72 | 72 | NA | NA | NA |
| x | VHDL | Pass | 64 | 64 | 64 | NA | NA | NA |

Table 93: HLS Solution 11 C/RTL Cosimulation Report

Si può notare come, rispetto alla solution 10 dove era stato applicato un partitioning dello stesso fattore ma di tipologia cyclic, l'utilizzazione delle risorse in corrispondenza del solo partizionamento dell'array x non cambia dal momento che è stata solo modificata il valore della variabile nnz, cioè la dimensione degli array columnIndex e values.

| Solution | SLICE | LUT | FF | DSP | BRAM | CP required | CP achieved post- synthesis | CP achieved post- implementation |
|------------------------|-------|-----|-----|-----|------|----------------|--------------------------------------|---|
| columnIndex, values, x | 179 | 454 | 450 | 6 | 0 | 10 | 6.541 | 6.677 |
| columnIndex | 90 | 267 | 227 | 6 | 0 | 10 | 7.489 | 7.664 |
| values | 122 | 391 | 232 | 6 | 0 | 10 | 7.506 | 7.768 |
| x | 143 | 311 | 444 | 6 | 0 | 10 | 6.540 | 6.790 |

Table 94: HLS Solution 11 Export RTL Report

6 Conclusions

Analizzando l'insieme delle soluzioni hardware proposte, è possibile effettuare alcune considerazioni finali riguardo la latenza e l'utilizzazione delle risorse corrispondente ad ogni solution. Bisogna specificare che nel plot sotto allegato, le solution che fanno riferimento al partizionamento degli array columnIndex, values e x saranno definite, per semplicità, nella seguente maniera:

- (columnIndex, values, x) \rightarrow a
- columnIndex \rightarrow b
- values \rightarrow c
- x \rightarrow d

Ad esempio, la solution 10 basata sul partitioning di columnIndex, values e x verrà identificata all'interno del plot come "10a" mentre la solution 8 basata sul partitioning di values verrà identificata all'interno del plot come "8c".

In particolare, è possibile notare come la soluzione 1 iniziale, dove non è stata considerata alcuna direttiva, prevede un valore pari a 58. Tale valore è possibile ottimizzarlo considerando, ad esempio, un approccio pipeline sul loop2. Infatti, in corrispondenza della soluzione hardware 2 è stato previsto un pragma di pipeline all'interno del ciclo 2 permettendo all'architettura di lavorare su dati temporalmente differenti e, quindi, permettendo una diminuzione della latenza totale. Invece, per quanto riguarda la soluzione 3, dal momento che il tool non è riuscito a sintetizzare la direttiva di pipelining nel loop1, la latenza associata risulta essere la medesima di quella ottenuta nella soluzione 1. Invece, in corrispondenza della soluzione 4, dove si è considerato un parallelismo di fattore pari a 2, cioè che ci aspetta è una diminuzione della latenza. Purtroppo questo non accade dal momento che il tool effettua una sintesi tale per cui le due istanze di loop2, previste in parallelo all'interno del loop1, vengono eseguite in maniera sequenziale. Pertanto, il valore del trip count associato al loop1 viene dimezzato mentre il valore di latenza associato all'architettura rimane quasi il medesimo. Per quanto riguarda, invece, la soluzione hardware 5 in corrispondenza della quale vengono considerate le direttive di pipeline e unrolling di fattore pari a 2 nel loop2, si ha sia un dimezzamento del valore di trip count sia una diminuzione della latenza totale.

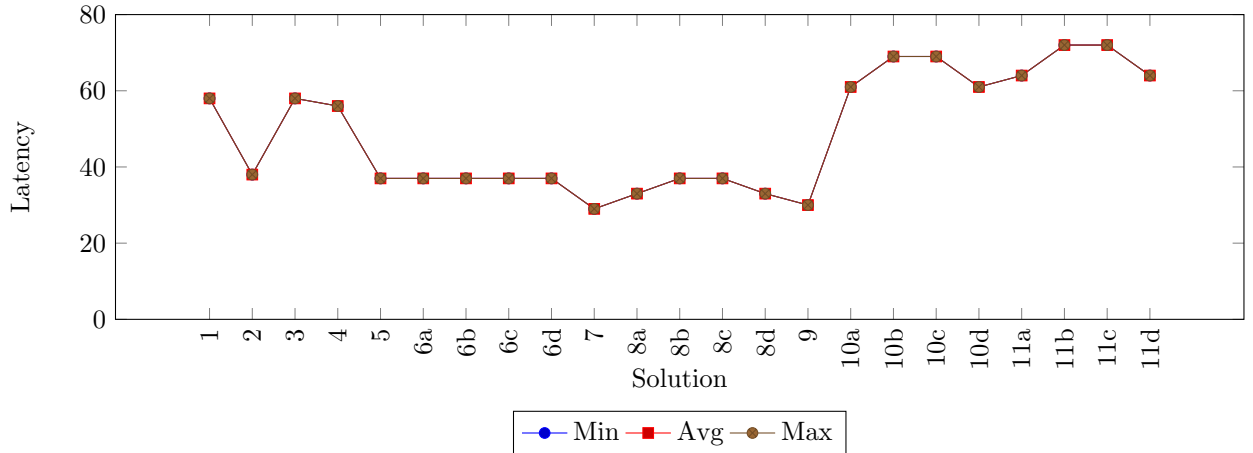


Figure 8: HLS Conclusions C/RTL Cosimulation Plot

Per quanto riguarda, la soluzione hardware 7, dove è stata prevista una direttiva di parallelismo di fattore pari a 4 all'interno del loop2, si può notare come la latenza associata risulta essere pari a 29 attestandosi come la seconda latenza minore tra quelle proposte. Tanto è vero che, in corrispondenza della soluzione hardware 9, si ottiene il valore di latenza minore tra quelle implementate, cioè pari a 26. Tale risultato è ottenuto in seguito all'applicazione della direttiva di pipeline nel loop1 e dei pragma di partizionamento agli array rowPtr e y dal momento che la direttiva di unrolling di fattore pari a 8 è stata ignorata dal tool a

causa dei warning citati nella sezione corrispondente. Tanto è vero che il tool, tramite log all'interno della console, segnala l'impossibilità di applicare l'unrolling all'interno del loop2 e ottenendo, di conseguenza, un valore di trip count uguale a quello ottenuto in corrispondenza della solution 1.

Per quanto riguarda la solution 6, dove è stato considerato all'interno del loop2 sia un parallelismo di fattore pari a 2 sia le varie configurazioni partitioning dei tre array, si può notare come la latenza corrispondente alle 4 configurazioni considerate risulta essere la medesima, cioè pari a 37. Inoltre, tale valore risulta essere il medesimo di quello ottenuto in corrispondenza della soluzione hardware 5 dove è stato considerato all'interno del loop2 la direttiva di unrolling di fattore 2 senza considerare quelle relative al partitioning. Pertanto, sarebbe opportuno, nella successiva analisi conclusiva dell'utilizzazione delle risorse, comprendere se effettivamente ha senso considerare i partitioning ulteriori. Inoltre, si può notare come, in corrispondenza, delle configurazioni proposte nella solution 8, si ottiene la medesima latenza tranne per le configurazioni dove si considera il partizionamento di fattore pari a 4 di tutti e tre gli array contemporaneamente e dove si considera soltanto il partitioning di fattore pari a 4 dell'array x. Infatti, in corrispondenza di tali configurazioni si ottiene una latenza pari a 33. Pertanto, analizzando successivamente l'utilizzazione delle risorse, nel caso in cui tale utilizzazione risulti essere maggiore rispetto alla solution 5, si potrebbe individuare un trade-off tra risorse e latenza.

| Solution | RTL | Status | Latency | | | Interval | | |
|--------------------------|------|--------|---------|-----|-----|----------|-----|-----|
| | | | min | avg | max | min | avg | max |
| 1 | VHDL | Pass | 58 | 58 | 58 | NA | NA | NA |
| 2 | VHDL | Pass | 38 | 38 | 38 | NA | NA | NA |
| 3 | VHDL | Pass | 58 | 58 | 58 | NA | NA | NA |
| 4 | VHDL | Pass | 56 | 56 | 56 | NA | NA | NA |
| 5 | VHDL | Pass | 37 | 37 | 37 | NA | NA | NA |
| 6 | | | | | | | | |
| • columnIndex, values, x | VHDL | Pass | 37 | 37 | 37 | NA | NA | NA |
| • columnIndex | VHDL | Pass | 37 | 37 | 37 | NA | NA | NA |
| • values | VHDL | Pass | 37 | 37 | 37 | NA | NA | NA |
| • x | VHDL | Pass | 37 | 37 | 37 | NA | NA | NA |
| 7 | VHDL | Pass | 29 | 29 | 29 | NA | NA | NA |
| 8 | | | | | | | | |
| • columnIndex, values, x | VHDL | Pass | 33 | 33 | 33 | NA | NA | NA |
| • columnIndex | VHDL | Pass | 37 | 37 | 37 | NA | NA | NA |
| • values | VHDL | Pass | 37 | 37 | 37 | NA | NA | NA |
| • x | VHDL | Pass | 33 | 33 | 33 | NA | NA | NA |
| 9 | VHDL | Pass | 26 | 26 | 26 | NA | NA | NA |
| 10 | | | | | | | | |
| • columnIndex, values, x | VHDL | Pass | 61 | 61 | 61 | NA | NA | NA |
| • columnIndex | VHDL | Pass | 69 | 69 | 69 | NA | NA | NA |
| • values | VHDL | Pass | 69 | 69 | 69 | NA | NA | NA |
| • x | VHDL | Pass | 61 | 61 | 61 | NA | NA | NA |
| 11 | | | | | | | | |
| • columnIndex, values, x | VHDL | Pass | 64 | 64 | 64 | NA | NA | NA |
| • columnIndex | VHDL | Pass | 72 | 72 | 72 | NA | NA | NA |
| • values | VHDL | Pass | 72 | 72 | 72 | NA | NA | NA |
| • x | VHDL | Pass | 64 | 64 | 64 | NA | NA | NA |

Table 95: HLS Conclusions C/RTL Cosimulation Report

Inoltre, si può notare come all'aumentare del fattore di partitioning, si ha un notevole aumento della latenza associata all'architettura. Infatti, in corrispondenza della solution 10, dove è stato considerato un unrolling di fattore pari a 2 e un partitioning di fattore pari a 8, si può notare come la latenza sia pari a 61 per le configurazioni basate su partizionamento dei tre array e partizionamento di x mentre pari a 69 per le rimanenti. Analogamente, in corrispondenza delle configurazioni associate alla solution 11, si può notare

come la latenza sia pari a 64 in corrispondenza delle configurazioni basate su partizionamento dei tre array e partizionamento di x mentre pari a 72 per le rimanenti.

| Solution | SLICE | LUT | FF | DSP | BRAM |
|--------------------------|--------------|------------|-----------|------------|-------------|
| 1 | 48 | 93 | 161 | 3 | 0 |
| 2 | 38 | 115 | 139 | 3 | 0 |
| 3 | 48 | 94 | 161 | 3 | 0 |
| 4 | 83 | 186 | 292 | 3 | 0 |
| 5 | 69 | 184 | 196 | 6 | 0 |
| 6 | | | | | |
| • columnIndex, values, x | 113 | 316 | 198 | 6 | 0 |
| • columnIndex | 84 | 259 | 224 | 6 | 0 |
| • values | 99 | 327 | 224 | 6 | 0 |
| • x | 99 | 250 | 198 | 6 | 0 |
| 7 | 314 | 915 | 297 | 12 | 0 |
| 8 | | | | | |
| • columnIndex, values, x | 123 | 343 | 315 | 6 | 0 |
| • columnIndex | 85 | 261 | 226 | 6 | 0 |
| • values | 92 | 274 | 196 | 6 | 0 |
| • x | 104 | 248 | 313 | 6 | 0 |
| 9 | 101 | 345 | 96 | 3 | 0 |
| 10 | | | | | |
| • columnIndex, values, x | 204 | 558 | 449 | 6 | 0 |
| • columnIndex | 83 | 268 | 230 | 6 | 0 |
| • values | 117 | 342 | 199 | 6 | 0 |
| • x | 143 | 311 | 444 | 6 | 0 |
| 11 | | | | | |
| • columnIndex, values, x | 179 | 454 | 450 | 6 | 0 |
| • columnIndex | 90 | 267 | 227 | 6 | 0 |
| • values | 122 | 391 | 232 | 6 | 0 |
| • x | 143 | 311 | 444 | 6 | 0 |

Table 96: HLS Conclusions Export RTL Report

In particolare, è possibile notare come la solution 3 prevede la medesima utilizzazione della solution 1. Infatti, come già precedentemente esposto nella sezione corrispondente, tale solution corrisponde alla solution 1 dal momento che la direttiva di pipelining prevista nel loop1 non viene interpretata dal tool a causa dei bound non noti. Pertanto, il tool ignora tale pragma sintetizzando il circuito come se tale direttiva non fosse presente all'interno del loop1.

Inoltre, è possibile evidenziare come nella soluzione hardware 5 è prevista l'utilizzazione di 6 DSP, rispetto alle solution precedenti ad essa in cui ne erano previste 3. Questo incremento è dovuto al fatto che nella solution 5 è previsto un unrolling di fattore pari a 2. Pertanto, se nelle soluzioni base, cioè dove non era presente alcun parallelismo all'interno del loop2, era prevista un'utilizzazione di 3 DSP, allora nella soluzione 5 ne saranno previste il doppio. Analogamente, nella soluzione hardware 7 è previsto un unrolling di fattore pari a 4. Pertanto, come si può evidenziare dalla tabella sopra allegata, si ha un notevole incremento in corrispondenza delle LUT e, inoltre, l'utilizzazione delle DSP risulta essere pari a 12. Invece, per quanto riguarda la soluzione hardware 9, dove è previsto un unrolling di fattore pari a 8, l'utilizzazione di tali risorse risulta essere leggermente maggiore delle solution precedentemente citate dal momento che sono stati considerati due ulteriori partizionamenti (per gli array rowPtr e y). In particolare, dopo aver previsto ulteriori direttive di partizionamento all'interno del loop2, il tool ha evidenziato un'ulteriore problematica dovuta a ytmp che si è cercato di risolvere introducendo un pragma di pipeline all'interno del loop1. Tale ulteriore direttiva ha evidenziato i problemi anche citati nella soluzione 3 dove non è stato possibile attuare il pipelining del loop1 a causa dei bound non noti. Per tale motivo, anche nella solution 9 sono stati riscontrati i medesimi warning che hanno portato il tool a sintetizzare l'architettura senza considerare l'unrolling di

fattore pari a 8 nel loop2 e il pipelining del loop1. Nel caso in cui, il tool fosse riuscito a sintetizzare tale solution prevedendo un parallelismo di fattore pari a 8, l'utilizzazione delle DSP, secondo i risultati delle solution basate su unrolling precedentemente citate, sarebbe stata pari a 24.

Inoltre, si può notare come, in corrispondenza della solution hardware 10, la quale prevede quattro solution in base al partizionamento effettuato, si ha un incremento dell'utilizzazione delle risorse rispetto alla solution 8 dove l'approccio risulta essere il medesimo tranne che per il fattore di partitioning utilizzato che nel caso delle soluzioni 10 risulta essere il doppio. In particolare, per le soluzioni 10 è stato previsto un valore di rows e size pari al doppio di quelli previsti per le soluzioni 8. Pertanto, il numero di risorse previste risulta essere maggiore. Tanto è vero che, per la solution 10 (columnIndex, values, x) si ha un incremento di FF di circa il 43%, un incremento di LUT di circa il 63% e un incremento di slice di circa il 66%. Per quanto riguarda, invece, il numero di DSP utilizzati, questo risulta essere il medesimo tra le soluzioni 10 e le soluzioni 8 dal momento che in entrambe è previsto un fattore di unrolling pari a 2. Riguardo, invece, le soluzioni 11 nelle quali è stato previsto un partizionamento di tipologia block e di fattore pari a 8, l'utilizzazione delle risorse risulta essere quasi la medesima delle soluzioni 10. In particolare, per l'attuazione di tale direttiva è stata prevista, come precedentemente citato nell'apposita sezione, la variazione del valore della variabile nnz da 8 a 16. Infatti, tale valore nelle soluzioni 10 risulta essere pari a 8 mentre nelle soluzioni 11 risulta essere pari a 16. Tanto è vero che, le variazioni di utilizzazione di risorse tra le due categorie di solution si ha principalmente in corrispondenza dei partizionamenti degli array che prevedono un dimensionamento basato sulla variabile nnz (columnIndex e values). Infatti, in corrispondenza del partizionamento dell'array x, sia nella solution 10 che nella solution 11, l'utilizzazione delle risorse risulta essere la medesima.

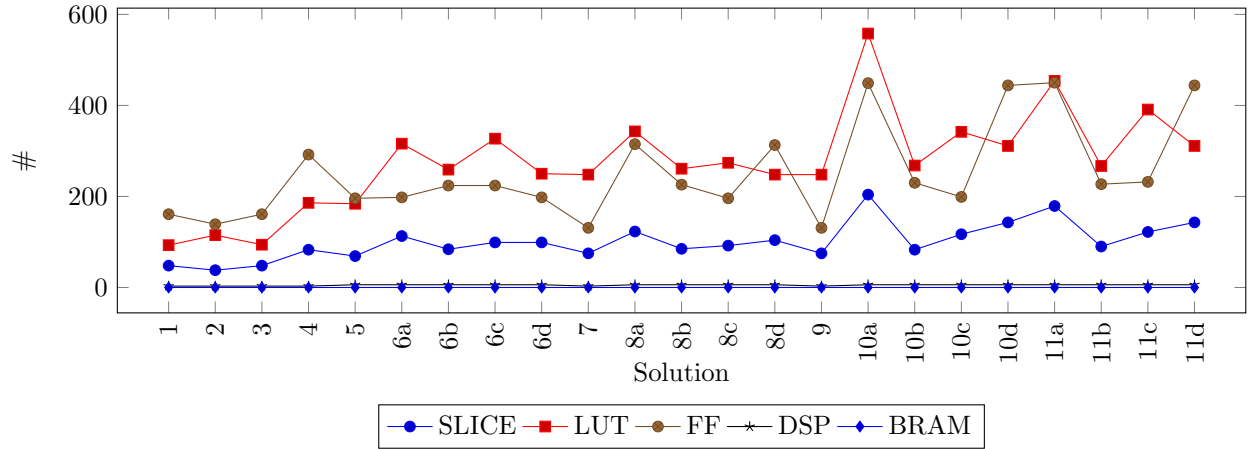


Figure 9: Utilization Export RTL Plot

In particolare, si può notare come, in corrispondenza delle solution 6 e 8 si hanno dei valori di latenza uguali e, in alcune configurazioni, anche minori rispetto alla solution iniziale come precedentemente analizzato. Nello specifico, tali solution presentano un numero di risorse che, confrontato con quello della solution iniziale, risulta essere maggiore. Pertanto, come precedentemente citato, si può notare come da una parte si ha una soluzione con un numero di risorse basso ed una latenza pari a 58 mentre dall'altra parte si hanno delle solution con latenza pari a 37 (in alcune configurazioni anche pari a 33) ma con utilizzazione delle risorse notevolmente maggiore (ad esempio il numero di DSP utilizzato risulta essere il doppio dovuto all'utilizzo della direttiva di unrolling di fattore pari a 2).

Nello specifico, la minor utilizzazione di risorse si ha in corrispondenza della solution 2 (anche se il numero di LUT risulta essere leggermente maggiore rispetto al numero di LUT minimo che si ha in corrispondenza della solution 1) dove è stata considerata la direttiva di pipeline all'interno del loop2. Inoltre, si può notare come la latenza associata a tale architettura risulta essere pari a 38. In aggiunta, si può evidenziare come la solution hardware 9 possa essere considerata anch'essa come possibile candidata ad essere soluzione ottimale. Infatti, si può notare come l'utilizzazione delle risorse associata risulta essere leggermente maggiore e la latenza inferiore rispetto alla solution 2. In particolare, se la solution 2 permette di ottenere l'utilizzazione delle risorse minima tra le soluzioni implementate con una latenza pari a 37, la solution hardware 9, invece,

permette di ottenere la latenza minima tra le architetture considerate con un'utilizzazione leggermente maggiore. Pertanto, dal momento che la differenza di utilizzazione delle risorse tra le due solution appena citate non è notevolmente alta, la soluzione hardware 9 risulta essere il miglior trade-off. Ovviamente, in questo caso, il trip count associato a tale solution risulta essere il medesimo della soluzione iniziale dal momento che la direttiva di unrolling di fattore pari a 8 prevista all'interno del loop2 non è stata attuata dal tool a causa dei problemi precedentemente citati.