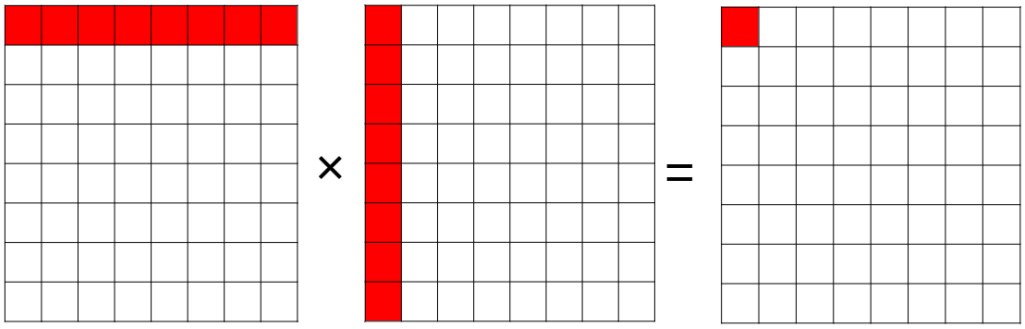
Report Lab B: Matrix Multiplication

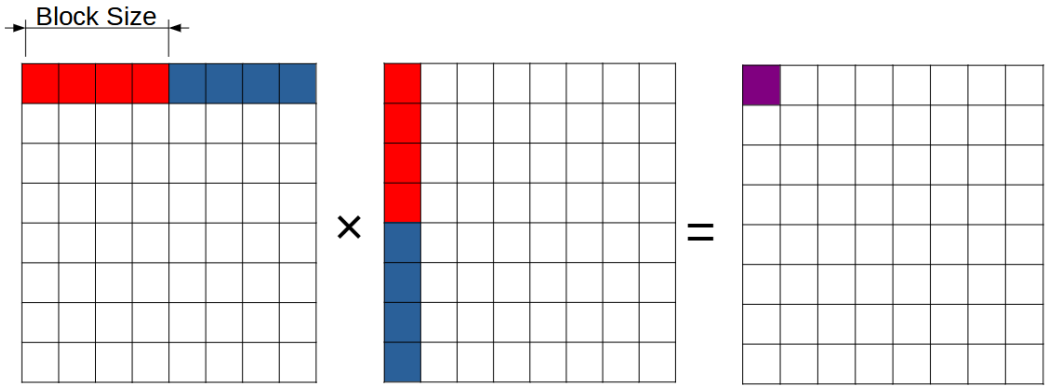
Introduction to Overall System

Matrix Multiplication (MM) of *N*N* matrices (*A*B=C*, and we set *N* to be power of 2) can be accelerated on hardware by unrolling (data parallelism) the *N^3* products and sums.
We use Vitis HLS and Vivado to implement our systems on FPGA. There are three different implementations for *N*N* integer matrices multiplication:

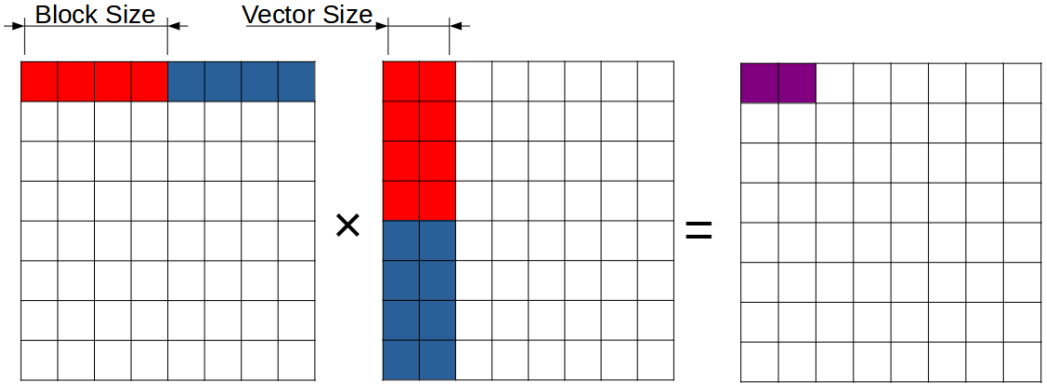
1. The first implementation is brute force implementation, and we completely partition rows of *A* and columns of *B* to perform data parallelism with unroll factor (*UrF*) *N*.



2. The second implementation is block matrix multiplication. The size of block (*BL*) should be equal to unroll factor in our implementation to ensure initiation interval (II) equal to 1. For each entry in *C*, we compute its sum of products in *N/BL* times. That is, we compute sum of *BL* products in one cycle.



3. The third implementation is block matrix multiplication, but use a vector of size *VT* to transmit data between host and kernel. Furthermore, since we can output *VT* entries of *C* in one cycle, we also perform data parallelism on it. That is, we have unroll factor *BL*VT*.



There are two phases in our implementation: first read in *A* and *B*, then do the computation and write out *C* at the same time.

Observed and Learned

By C-Synthesis report from Vitis HLS, we learn that the computation resource (DSP) is only related to unroll factor. (In our experiments, *# of DSPs=UrF*3*.)
The data-read-in phase takes time with order *N^2*. It is only reduced in the third implementation, with the order of time being *N^2/VT*.
The computation phase should take time with order *N^3* if there is no data parallelism.
The first implementation use unroll with factor *N*, which makes the time take is of order *N^2*. In this case the time took in computation is similar to the time took in data-read-in phase. However, the number of DSPs used linearly grows with *N*. From the utilization report in Vivado, we think this implementation can at most handle *N=64* case.
Block Matrix Multiplication method limits unroll factor and number of DSPs by block size. In second implementation, the block size should equal to unroll factor to ensure II being 1. We can see all implementation with II larger than 1 take significantly longer time.
In the third implementation, with limited resource, we can make trade of between vector size and block size to keep *UrF=BL*VT*, and we can see that a higher vector size can also reduce data transmitting time.
The conclusion is in the following table:

Implementation	*BL*	*VT*	*UrF*	Data transmit	Computation	# of DSPs
1st	-	-	*N*	*N^2*	*N^3/UrF*	*3*N*
2nd	+	-	*BL*	*N^2*	*N^3/UrF*	*3*BL*
3rd	+	+	*BL*VT*	*N^2/VT*	*N^3/UrF*	*3*BL*VT*

Experiments

For the first implementation we test *N=8, 16, 32* and we collect data from C-Synthesis reports. The raw data is in the figures below. Here we present a comparing table by setting N=8 case as baseline.

N	read data time	MM time	# of DSPs	# of FFs	# of LUTs
8	1	1	1	1	1
16	3.91	3.72	2	1.59	1.65
32	15.54	14.55	4	3.13	3.08

For the second implementation we set *N=128*, and test *BL=4, 8, 16, 32*. We only list result for cases with *BL=UrF* and set *BL=4* as baseline. If *BL≠UrF* we will get II violation, and latency increases dramatically. For cases with *BL≠UrF* go and see the raw data in figures below.

BL	read data time	MM time	# of DSPs	# of FFs	# of LUTs
4	1	6.6	1	1	1
8	1	3.4	2	1.69	1.48
16	1	1.8	4	3	2.39
32	1	1	8	5.68	4.27

For the third implementation we set *N=128*, and test *BL=16, VT=4* and *BL=VT=8*. And the result shows that their time spent for matrix multiplication is almost the same, which confirms that *UrF=BL*VT* and the time spent for data reading in *VT=4* is twice as large as *VT=8*.

The first and second implementations can be run on FPGA already, and there are Vivado utilization reports and FPGA implementation results in the following figures, and .bit and .hwh files in GitHub link.
The third implementation has not finished for FPGA, but there are Test Bench for Vitis HLS analysis, and we already check that it is correct in C-Simulation, C-Synthesis and Cosimulation.

GitHub link: https://github.com/b07901181/AAHLS-Lab-B-10

Figures

First Implementation: Brute Force

Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
matrix_mul				-	138	1.380E3	-	72	-	dataflow	0	24	2862	1876	0
read_data_A				-	66	660.000	-	66	-	no	0	0	10	314	0
read_data_B				-	66	660.000	-	66	-	no	0	0	10	330	0
brute_force				-	71	710.000	-	71	-	no	0	24	2276	820	0

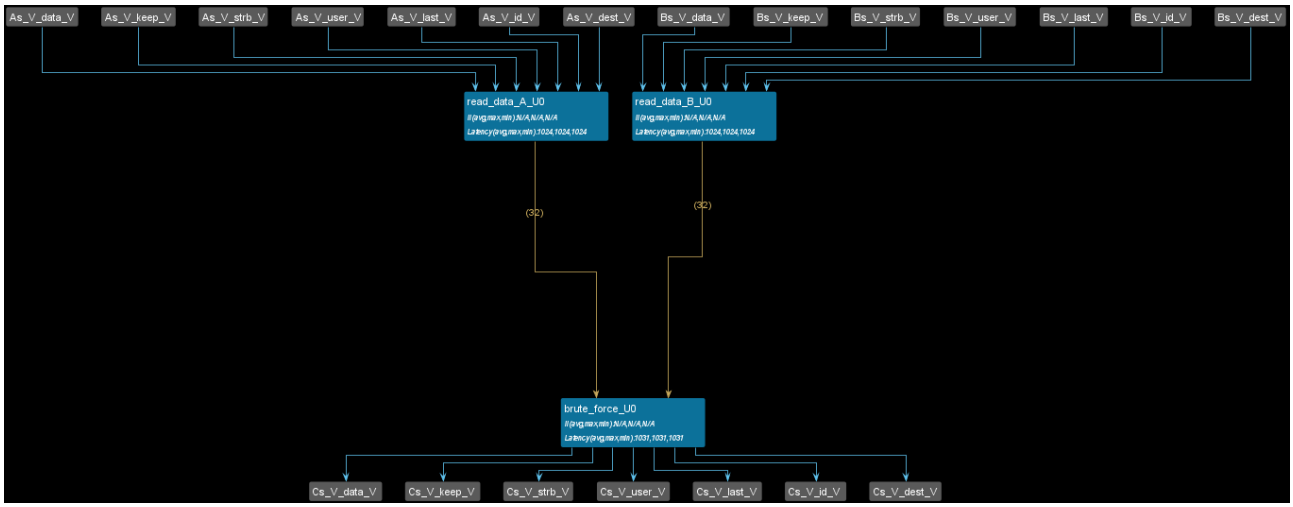
Synthesis report for first implementation with *N=8*

Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
matrix_mul				-	523	5.230E3	-	265	-	dataflow	32	48	4560	3091	0
read_data_A				-	258	2.580E3	-	258	-	no	0	0	12	538	0
read_data_B				-	258	2.580E3	-	258	-	no	0	0	12	570	0
brute_force				-	264	2.640E3	-	264	-	no	0	48	4466	1491	0

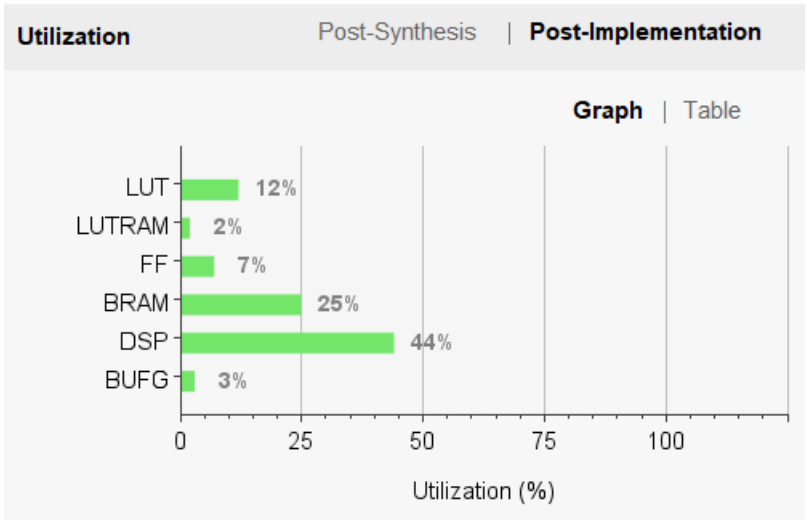
Synthesis report for first implementation with *N=16*

Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
matrix_mul				-	2060	2.060E4	-	1034	-	dataflow	64	96	8949	5786	0
read_data_A				-	1026	1.026E4	-	1026	-	no	0	0	14	985	0
read_data_B				-	1026	1.026E4	-	1026	-	no	0	0	14	1049	0
brute_force				-	1033	1.033E4	-	1033	-	no	0	96	8819	2844	0

Synthesis report for first implementation with *N=32*



Dataflow Viewer for first implementation with *N=32*



Vivado utilization report for first implementation with *N=32*

```
Kernel Time:
[0.0016689300537109375, 0.0013210773468017578, 0.0012938976287841797, 0.0018157958984375, 0.0014536380767822266, 0.0013206005096435547, 0.001277923583984375, 0.0012938976287841797, 0.0012695789337158203, 0.0012500286102294922]
============================
Python Time:
[0.0004584789276123047, 0.0004000663757324219, 0.00039196014404296875, 0.0003898143768310547, 0.0004017353057861328, 0.0004603862762451172, 0.0003936290740966797, 0.0003948211669921875, 0.0003921985626220703, 0.0003902912139892578]
============================
Correctness:
[True, True, True, True, True, True, True, True, True, True]
============================
============================
Exit process
```

Result for first implementation with *N=32* from FPGA

Second Implementation: Block Multiplication with 1-dim Partition and Unroll

Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
matrix_mul				-	557073	5.571E6	-	540687	-	dataflow	160	12	1670	1470	0
read_data_A				-	16386	1.640E5	-	16386	-	no	0	0	18	210	0
read_data_B				-	16386	1.640E5	-	16386	-	no	0	0	18	218	0
block_MM				-	540686	5.407E6	-	540686	-	no	32	12	1588	862	0

Synthesis report for second implementation with *N=128*, *BL=UrF=4*

Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
matrix_mul				-	294930	2.949E6	-	278544	-	dataflow	160	24	2821	2175	0
read_data_A				-	16386	1.640E5	-	16386	-	no	0	0	18	322	0
read_data_B				-	16386	1.640E5	-	16386	-	no	0	0	18	338	0
block_MM				-	278543	2.785E6	-	278543	-	no	32	24	2731	1231	0

Synthesis report for second implementation with *N=128*, *BL=UrF=8*

Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
matrix_mul				-	163859	1.639E6	-	147473	-	dataflow	160	48	5015	3515	0
read_data_A				-	16386	1.640E5	-	16386	-	no	0	0	18	546	0
read_data_B				-	16386	1.640E5	-	16386	-	no	0	0	18	578	0
block_MM				-	147472	1.475E6	-	147472	-	no	32	48	4909	1899	0

Synthesis report for second implementation with *N=128*, *BL=UrF=16*

Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
matrix_mul				-	98324	9.830E5	-	81938	-	dataflow	160	96	9487	6272	0
read_data_A				-	16386	1.640E5	-	16386	-	no	0	0	18	994	0
read_data_B				-	16386	1.640E5	-	16386	-	no	0	0	18	1058	0
block_MM				-	81937	8.190E5	-	81937	-	no	32	96	9349	3312	0

Synthesis report for second implementation with *N=128*, *BL=UrF=32*

Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
matrix_mul				-	2129933	2.130E7	-	2113547	-	dataflow	160	12	1464	5000	0
read_data_A				-	16386	1.640E5	-	16386	-	no	0	0	18	994	0
read_data_B				-	16386	1.640E5	-	16386	-	no	0	0	18	1058	0
block_MM	II Violation			-	2113546	2.114E7	-	2113546	-	no	32	12	1326	2040	0

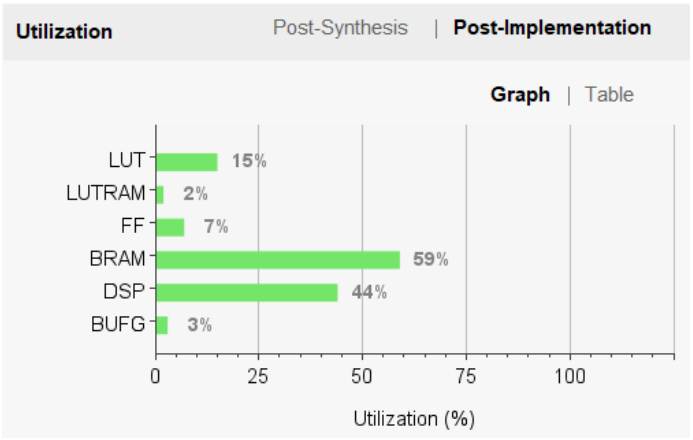
Synthesis report for second implementation with *N=128*, *BL=32*, *UrF=4*, Final II is 4

Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
matrix_mul				-	1081360	1.081E7	-	1064974	-	dataflow	160	24	2636	6333	0
read_data_A				-	16386	1.640E5	-	16386	-	no	0	0	18	994	0
read_data_B				-	16386	1.640E5	-	16386	-	no	0	0	18	1058	0
block_MM	II Violation			-	1064973	1.065E7	-	1064973	-	no	32	24	2498	3373	0

Synthesis report for second implementation with *N=128*, *BL=32*, *UrF=8*, Final II is 4

Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
matrix_mul				-	557073	5.571E6	-	540687	-	dataflow	160	48	4863	6201	0
read_data_A				-	16386	1.640E5	-	16386	-	no	0	0	18	994	0
read_data_B				-	16386	1.640E5	-	16386	-	no	0	0	18	1058	0
block_MM	II Violation			-	540686	5.407E6	-	540686	-	no	32	48	4725	3241	0

Synthesis report for second implementation with *N=128*, *BL=32*, *UrF=16*, Final II is 4



Vivado utilization report for second implementation with *N=128*, BL=UrF=32

```
Kernel Time:
[0.004660129547119141, 0.0043299198150634766, 0.004352569580078125, 0.004268169403076172, 0.004301309585571289, 0.0043454170227
05078, 0.004355907440185547, 0.004304170608520508, 0.004332304000854492, 0.0043604373931884766]
============================
Python Time:
[0.045684814453125, 0.05273175239562988, 0.052286624908447266, 0.052222251892089844, 0.05346512794494629, 0.05282020568847656,
0.05363011360168457, 0.056420326232910156, 0.05349159240722656, 0.052698612213134766]
============================
Correctness:
[True, True, True, True, True, True, True, True, True, True]
============================
============================
Exit process
```
Result for first implementation with *N=128*, BL=UrF=32, from FPGA

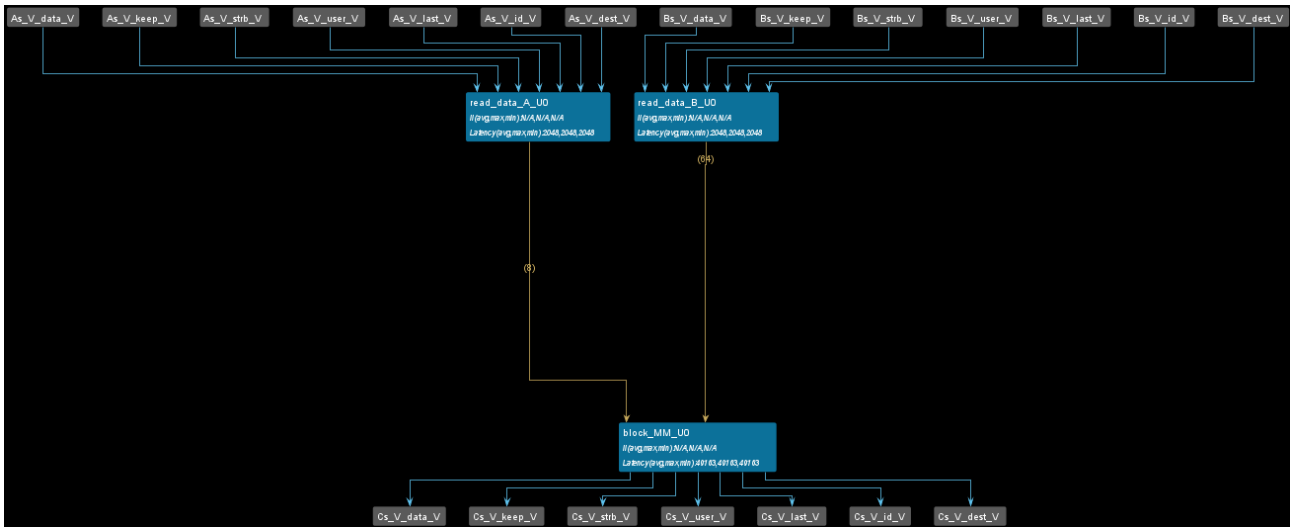
Third Implementation: Block Multiplication with Array Partition and Vectorization Transmission

Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
matrix_mul				-	53267	5.330E5	-	49169	-	dataflow	160	192	16857	9737	0
read_data_A				-	4098	4.098E4	-	4098	-	no	0	0	16	539	0
read_data_B				-	4098	4.098E4	-	4098	-	no	0	0	16	1915	0
block_MM				-	49168	4.920E5	-	49168	-	no	32	192	16707	6167	0

Synthesis report for third implementation with *N=128*, *BL=16*, *VT=4*

Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
matrix_mul				-	51218	5.120E5	-	49168	-	dataflow	160	192	17015	9774	0
read_data_A				-	2050	2.050E4	-	2050	-	no	0	0	15	313	0
read_data_B				-	2050	2.050E4	-	2050	-	no	0	0	15	1897	0
block_MM				-	49167	4.920E5	-	49167	-	no	32	192	16875	6552	0

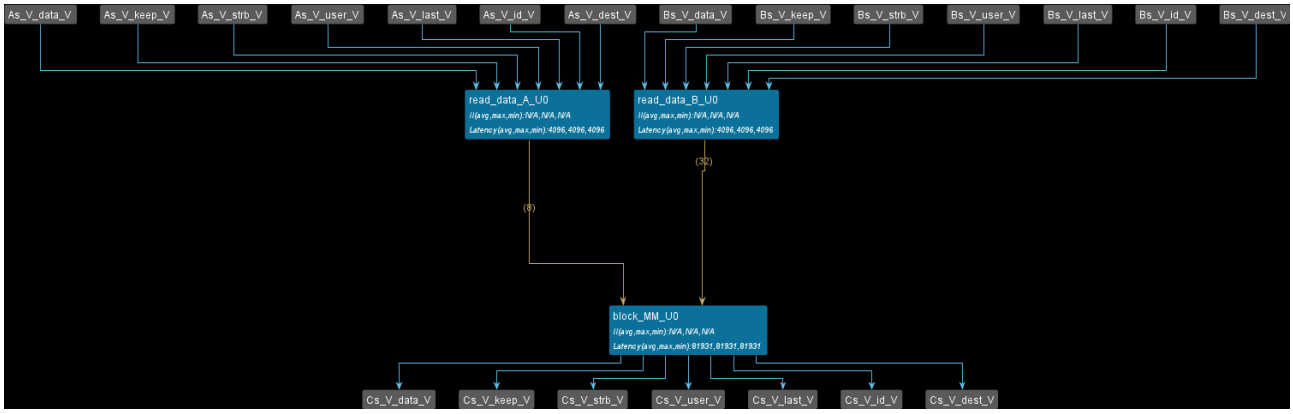
Synthesis report for third implementation with *N=128*, *BL=8*, *VT=8*



Dataflow Viewer for third implementation with *N=128*, *BL=8*, *VT=8*

Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
matrix_mul				-	86034	8.600E5	-	81936	-	dataflow	160	96	8855	5402	0
read_data_A				-	4098	4.098E4	-	4098	-	no	0	0	16	315	0
read_data_B				-	4098	4.098E4	-	4098	-	no	0	0	16	1003	0
block_MM				-	81935	8.190E5	-	81935	-	no	32	96	8745	3488	0

Synthesis report for third implementation with *N=128, BL=8, VT=4*



Dataflow Viewer for third implementation with *N=128, BL=8, VT=4*