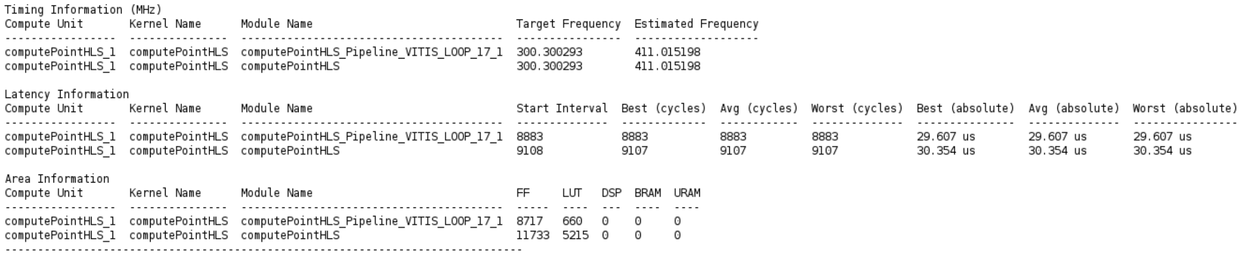
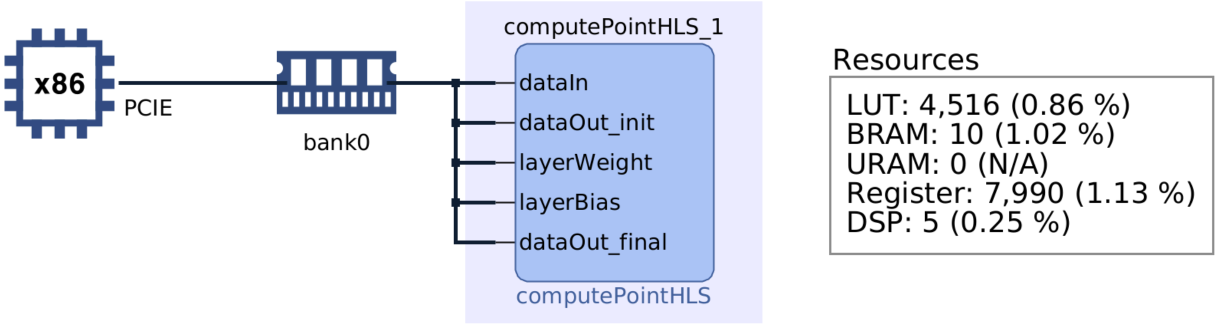
Single Convolution Layer Vitis Kernel Results:



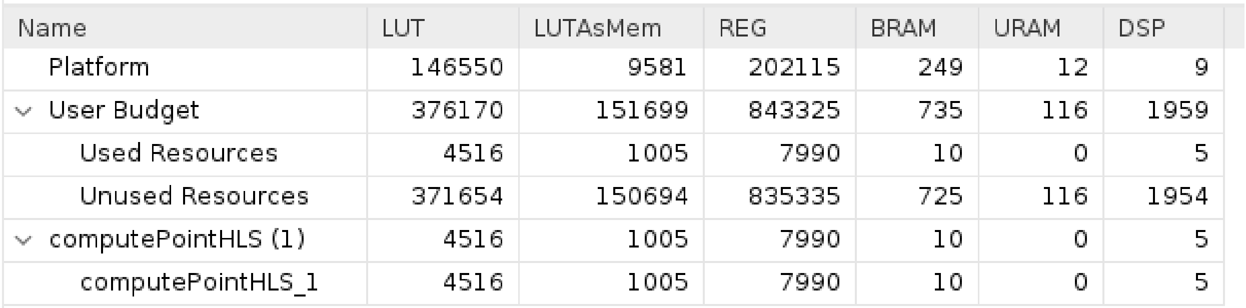
```
Timing Information (MHz)
Compute Unit       Kernel Name      Module Name                              Target Frequency  Estimated Frequency
----------------   --------------   --------------------------------------   ----------------  -------------------
computePointHLS_1  computePointHLS  computePointHLS_Pipeline_VITIS_LOOP_17_1  300.300293        411.015198
computePointHLS_1  computePointHLS  computePointHLS                           300.300293        411.015198

Latency Information
Compute Unit       Kernel Name      Module Name                              Start Interval  Best (cycles)  Avg (cycles)  Worst (cycles)  Best (absolute)  Avg (absolute)  Worst (absolute)
----------------   --------------   --------------------------------------   --------------  -------------  ------------  --------------  ---------------  --------------  ----------------
computePointHLS_1  computePointHLS  computePointHLS_Pipeline_VITIS_LOOP_17_1  8883            8883           8883          8883            29.607 us        29.607 us       29.607 us
computePointHLS_1  computePointHLS  computePointHLS                           9108            9107           9107          9107            30.354 us        30.354 us       30.354 us

Area Information
Compute Unit       Kernel Name      Module Name                              FF     LUT   DSP  BRAM  URAM
----------------   --------------   --------------------------------------   -----  ----  ---  ----  ----
computePointHLS_1  computePointHLS  computePointHLS_Pipeline_VITIS_LOOP_17_1  8717   660   0    0     0
computePointHLS_1  computePointHLS  computePointHLS                           11733  5215  0    0     0
----------------------------------------------------------------------------
```

1 Convolution estimated 30.354uS

To run full layer: 100352 * 30.354 = 3.046 seconds

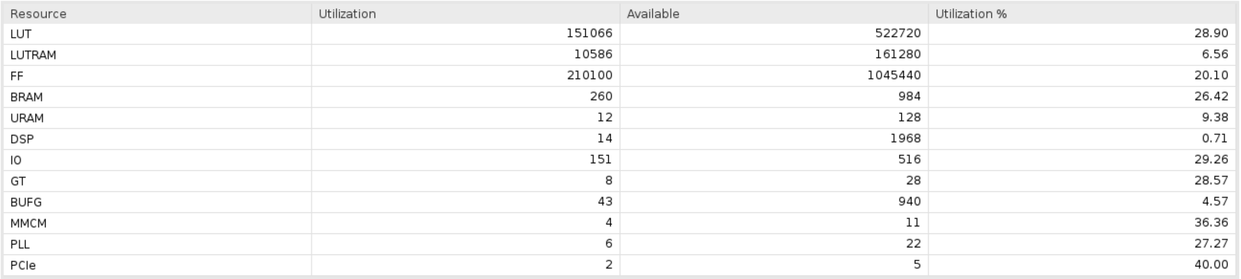


LUT: 4,516 (0.86 %)
BRAM: 10 (1.02 %)
URAM: 0 (N/A)
Register: 7,990 (1.13 %)
DSP: 5 (0.25 %)



Name	LUT	LUTAsMem	REG	BRAM	URAM	DSP
Platform	146550	9581	202115	249	12	9
User Budget	376170	151699	843325	735	116	1959
Used Resources	4516	1005	7990	10	0	5
Unused Resources	371654	150694	835335	725	116	1954
computePointHLS (1)	4516	1005	7990	10	0	5
computePointHLS_1	4516	1005	7990	10	0	5

Kernel Usage



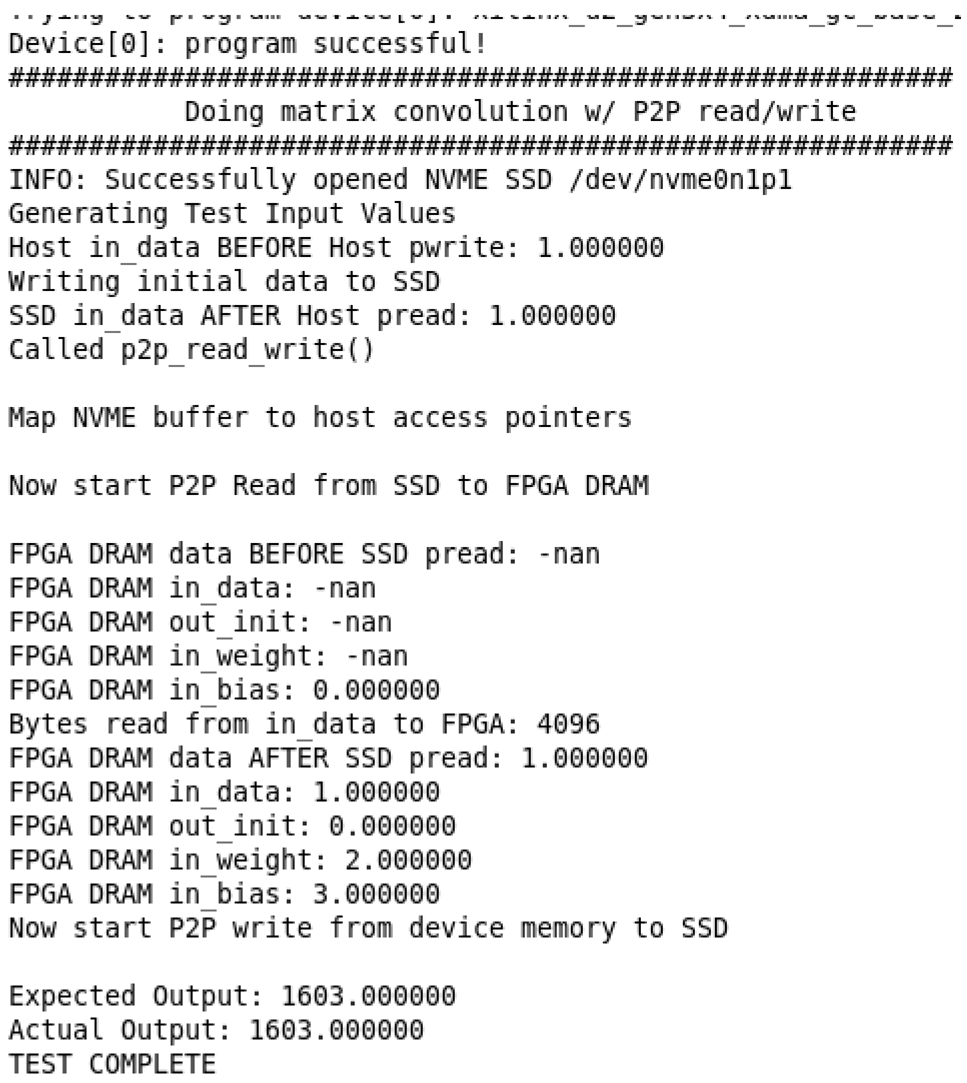
Resource	Utilization	Available	Utilization %
LUT	151066	522720	28.90
LUTRAM	10586	161280	6.56
FF	210100	1045440	20.10
BRAM	260	984	26.42
URAM	12	128	9.38
DSP	14	1968	0.71
IO	151	516	29.26
GT	8	28	28.57
BUFG	43	940	4.57
MMCM	4	11	36.36
PLL	6	22	27.27
PCIe	2	5	40.00

Device Utilization

Original Convolution Layer Kernel Call:

Data_Input: Array of 800, all value 1
Data_Weight: Array of 800, all value 2
Data_Bias: scalar 3
Data_Out_In: 0

Expected Output: = 800*1*2 + 3 = 1603



```
Trying to program device[0]: xilinx_u2_gen3x4_xdma_gc_base_
Device[0]: program successful!
###########################################################
           Doing matrix convolution w/ P2P read/write
###########################################################
INFO: Successfully opened NVME SSD /dev/nvme0n1p1
Generating Test Input Values
Host in_data BEFORE Host pwrite: 1.000000
Writing initial data to SSD
SSD in_data AFTER Host pread: 1.000000
Called p2p_read_write()

Map NVME buffer to host access pointers

Now start P2P Read from SSD to FPGA DRAM

FPGA DRAM data BEFORE SSD pread: -nan
FPGA DRAM in_data: -nan
FPGA DRAM out_init: -nan
FPGA DRAM in_weight: -nan
FPGA DRAM in_bias: 0.000000
Bytes read from in_data to FPGA: 4096
FPGA DRAM data AFTER SSD pread: 1.000000
FPGA DRAM in_data: 1.000000
FPGA DRAM out_init: 0.000000
FPGA DRAM in_weight: 2.000000
FPGA DRAM in_bias: 3.000000
Now start P2P write from device memory to SSD

Expected Output: 1603.000000
Actual Output: 1603.000000
TEST COMPLETE
```

Original Convolution Point Kernel Call Timing Results:

1 block = 4KB

Use seperate block for each array input. I think I can reduce this by combining blocks

Host to SSD Write (4 blocks): 221 us
SSD to FPGA writes (4 blocks): 185 us
Kernel run: 236 us
FPGA to SSD write (1 block): 89 us
SSD to Host write (1 block): 114 us

Original Convolution Layer Kernel Call Timing Results:

Repeat for full layer, for 100352 iterations.

Host to SSD Write (4 blocks): 221 us
SSD to FPGA writes (4 blocks): 185 us
Kernel run: 236 us
FPGA to SSD write (1 block): 89 us
SSD to Host write (1 block): 114 us

Full Layer: 32.417 seconds

Reducing write size for kernel result

Before this, used 4 blocks of 1024 entries, 4KB eachs for each input argument and 1 output argument.

Number of float values able to be written/read from pwrite/pread:

- 1024
- 512
- 256
- 128

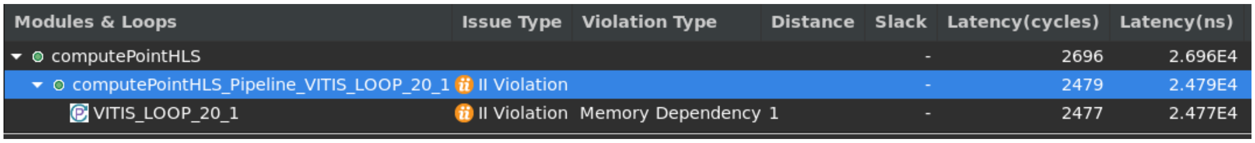
Got a WRITE FAIL when using 64 float values or less.

Changed output write array size to 128, reduced writeback time by 50us. NO CHANGES TO KERNEL

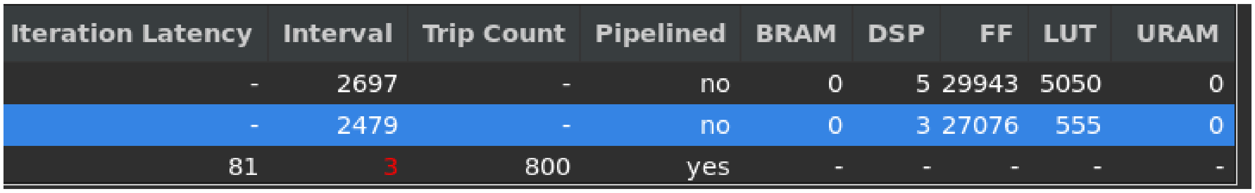
Optimizing Kernel Pipelining

Set pipeline stride with HLS pragma: #pragma HLS pipeline II=stride

Original Cycle Report:

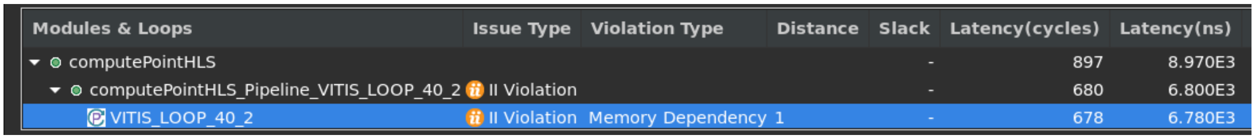


Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)
computePointHLS				-	2696	2.696E4
computePointHLS_Pipeline_VITIS_LOOP_20_1	II Violation			-	2479	2.479E4
VITIS_LOOP_20_1	II Violation	Memory Dependency	1	-	2477	2.477E4

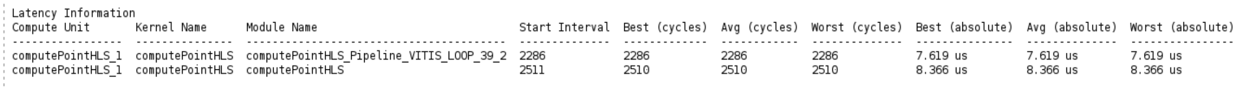


Iteration Latency	Interval	Trip Count	Pipelined	BRAM	DSP	FF	LUT	URAM
-	2697	-	no	0	5	29943	5050	0
-	2479	-	no	0	3	27076	555	0
81	3	800	yes	-	-	-	-	-

Cycle Report, Stride 4:

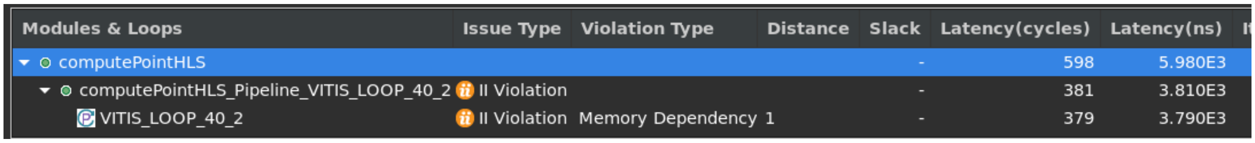


Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)
computePointHLS				-	897	8.970E3
computePointHLS_Pipeline_VITIS_LOOP_40_2	II Violation			-	680	6.800E3
VITIS_LOOP_40_2	II Violation	Memory Dependency	1	-	678	6.780E3



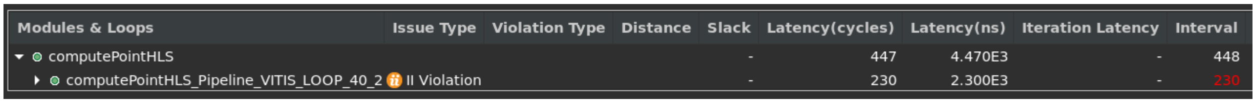
```
Latency Information
Compute Unit      Kernel Name      Module Name                               Start Interval  Best (cycles)  Avg (cycles)  Worst (cycles)  Best (absolute)  Avg (absolute)  Worst (absolute)
----------------  ---------------  ----------------------------------------  --------------  -------------  ------------  --------------  ---------------  --------------  ----------------
computePointHLS_1  computePointHLS  computePointHLS_Pipeline_VITIS_LOOP_39_2  2286            2286           2286          2286            7.619 us         7.619 us        7.619 us
computePointHLS_1  computePointHLS  computePointHLS                           2511            2510           2510          2510            8.366 us         8.366 us        8.366 us
```

Cycle Report, Stride 8:



Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)
computePointHLS				-	598	5.980E3
computePointHLS_Pipeline_VITIS_LOOP_40_2	II Violation			-	381	3.810E3
VITIS_LOOP_40_2	II Violation	Memory Dependency	1	-	379	3.790E3

Cycle Report, Stride 16 (lowest without bandwidth issue):



Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval
computePointHLS				-	447	4.470E3	-	448
computePointHLS_Pipeline_VITIS_LOOP_40_2	II Violation			-	230	2.300E3	-	230

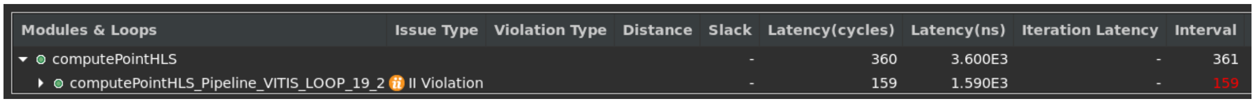
Once above stride 16, bus doesnt have enough ports

Note: Got same timing results for strides 4 and 16. Think I reached max improvement with this change

Setting separate m_axi ports for two read arrays

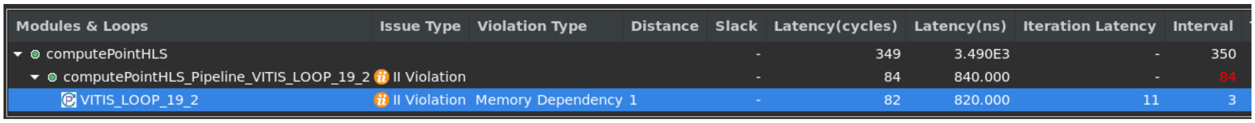
Reduced execution cycle time of kernel by half, can now read ports in parallel

Stride 16:



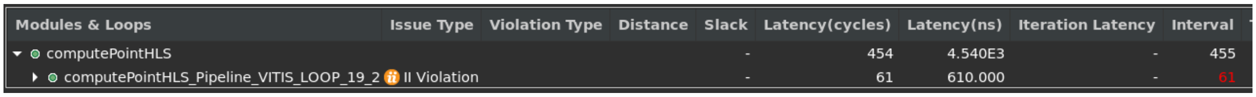
Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval
computePointHLS				-	360	3.600E3	-	361
computePointHLS_Pipeline_VITIS_LOOP_19_2	II Violation			-	159	1.590E3	-	159

Stride 32:



Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval
computePointHLS				-	349	3.490E3	-	350
computePointHLS_Pipeline_VITIS_LOOP_19_2	II Violation			-	84	840.000	-	84
VITIS_LOOP_19_2	II Violation	Memory Dependency	1	-	82	820.000	11	3

Stride 64:



Modules & Loops	Issue Type	Violation Type	Distance	Slack	Latency(cycles)	Latency(ns)	Iteration Latency	Interval
computePointHLS				-	454	4.540E3	-	455
computePointHLS_Pipeline_VITIS_LOOP_19_2	II Violation			-	61	610.000	-	61

Took longer, Stride 32 optimal

No improvement????

Multiple Iterations per kernel call

Less data reads/writes to FPGA

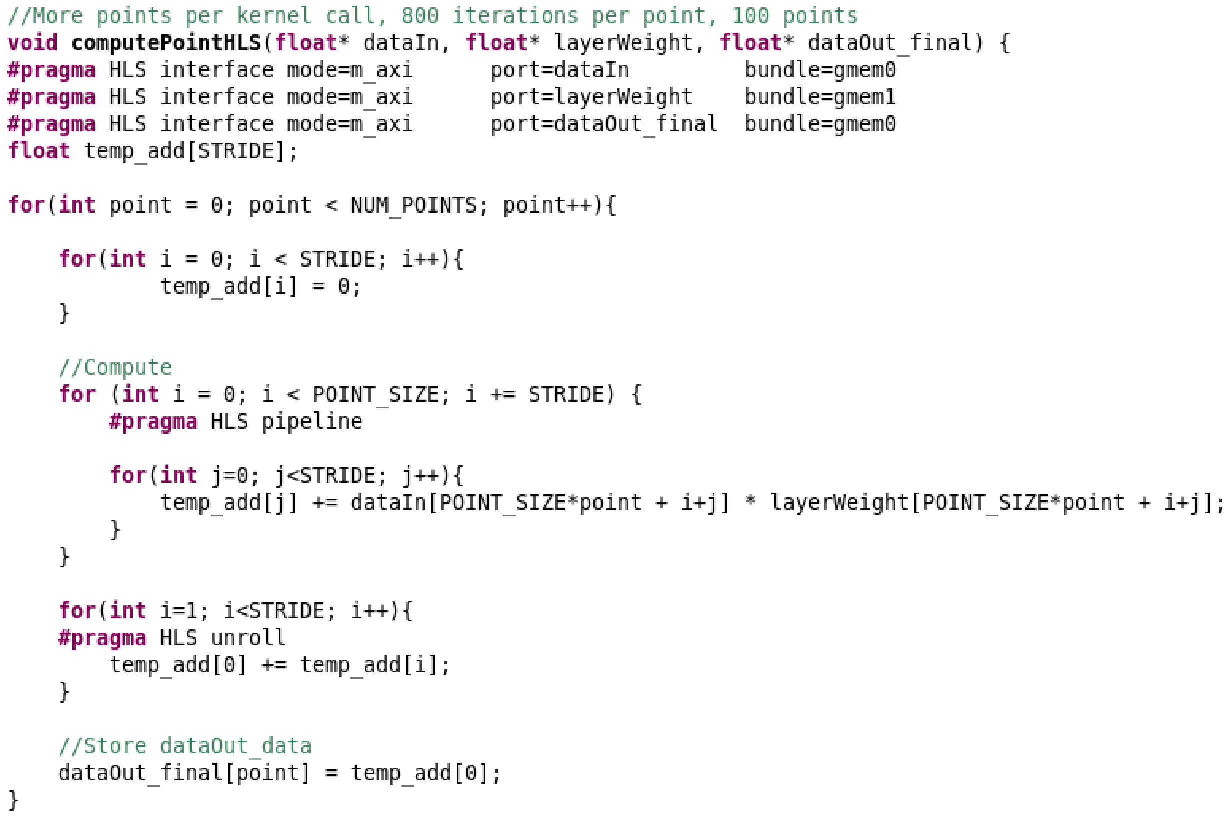
Did 40 points per call first, got down to 1.5 seconds

Tried 90 next

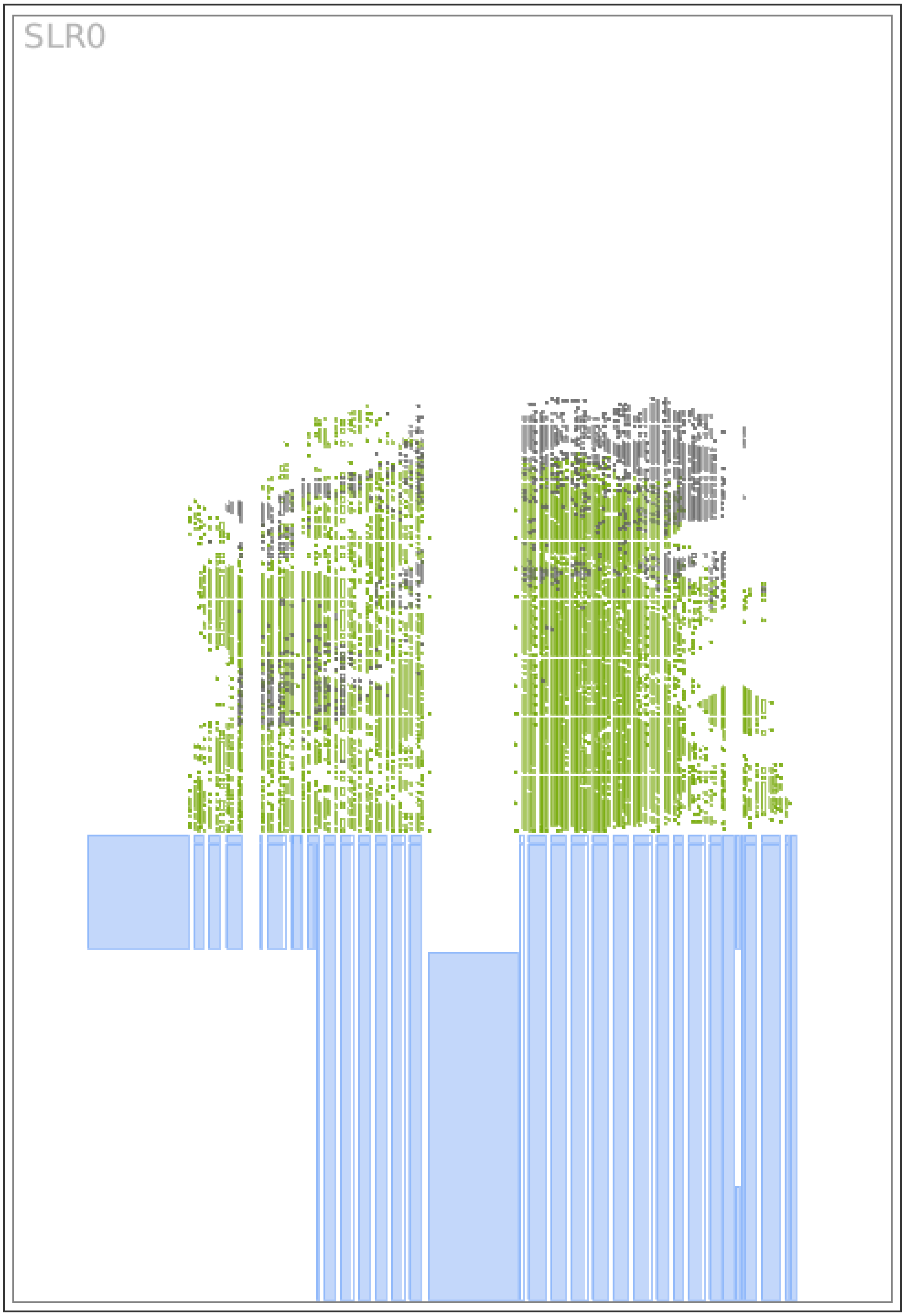
Then did 800

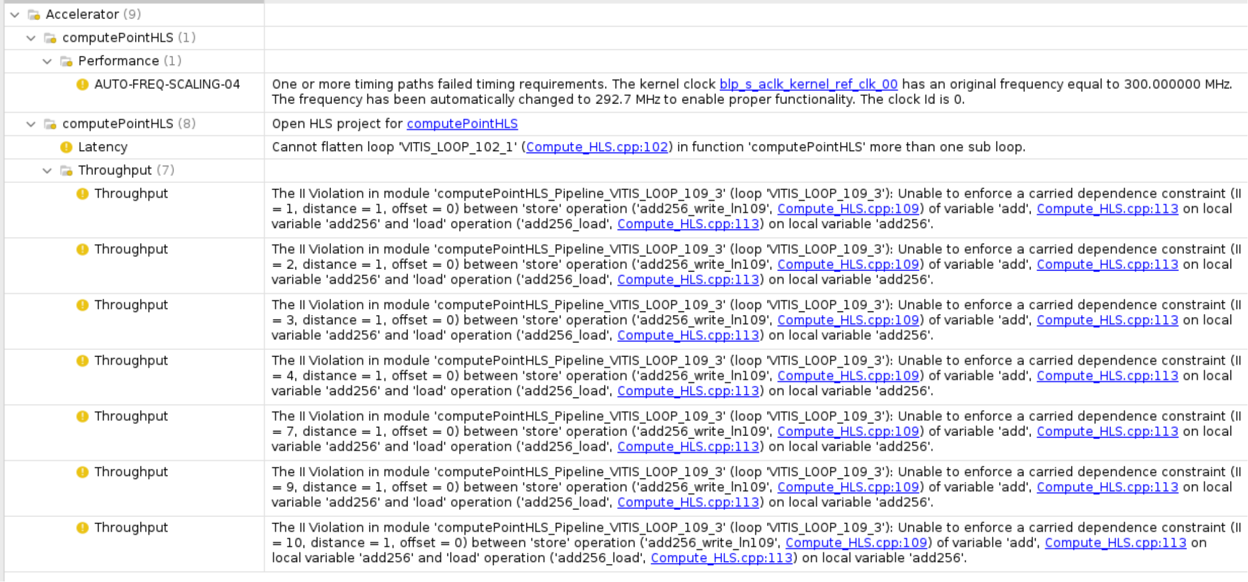
800 was largest size I could use, this achieved time lower than baseline!

Final Synthesis Results

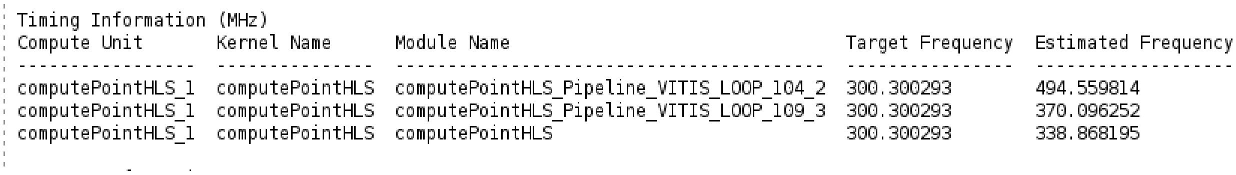


```
//More points per kernel call, 800 iterations per point, 100 points
void computePointHLS(float* dataIn, float* layerWeight, float* dataOut_final) {
#pragma HLS interface mode=m_axi      port=dataIn         bundle=gmem0
#pragma HLS interface mode=m_axi      port=layerWeight    bundle=gmem1
#pragma HLS interface mode=m_axi      port=dataOut_final  bundle=gmem0
float temp_add[STRIDE];

for(int point = 0; point < NUM_POINTS; point++){

    for(int i = 0; i < STRIDE; i++){
            temp_add[i] = 0;
    }

    //Compute
    for (int i = 0; i < POINT_SIZE; i += STRIDE) {
        #pragma HLS pipeline

        for(int j=0; j<STRIDE; j++){
            temp_add[j] += dataIn[POINT_SIZE*point + i+j] * layerWeight[POINT_SIZE*point + i+j];
        }
    }

    for(int i=1; i<STRIDE; i++){
    #pragma HLS unroll
        temp_add[0] += temp_add[i];
    }

    //Store dataOut_data
    dataOut_final[point] = temp_add[0];
}
```

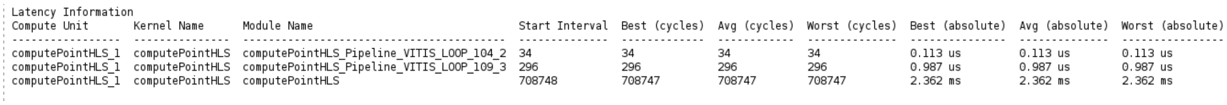

SLR0



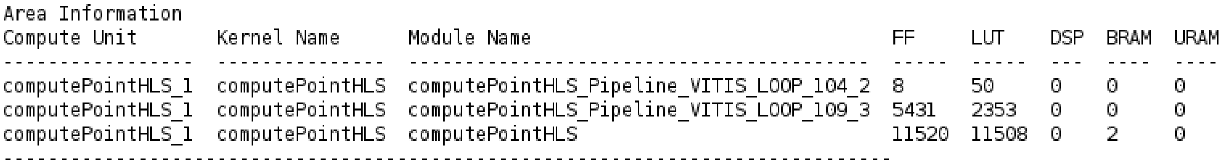
Accelerator (9)	
computePointHLS (1)	
Performance (1)	
AUTO-FREQ-SCALING-04	One or more timing paths failed timing requirements. The kernel clock blp_s_aclk_kernel_ref_clk_00 has an original frequency equal to 300.000000 MHz. The frequency has been automatically changed to 292.7 MHz to enable proper functionality. The clock Id is 0.
computePointHLS (8)	Open HLS project for computePointHLS
Latency	Cannot flatten loop 'VITIS_LOOP_102_1' (Compute_HLS.cpp:102) in function 'computePointHLS' more than one sub loop.
Throughput (7)	
Throughput	The II Violation in module 'computePointHLS_Pipeline_VITIS_LOOP_109_3' (loop 'VITIS_LOOP_109_3'): Unable to enforce a carried dependence constraint (II = 1, distance = 1, offset = 0) between 'store' operation ('add256_write_ln109', Compute_HLS.cpp:109) of variable 'add', Compute_HLS.cpp:113 on local variable 'add256' and 'load' operation ('add256_load', Compute_HLS.cpp:113) on local variable 'add256'.
Throughput	The II Violation in module 'computePointHLS_Pipeline_VITIS_LOOP_109_3' (loop 'VITIS_LOOP_109_3'): Unable to enforce a carried dependence constraint (II = 2, distance = 1, offset = 0) between 'store' operation ('add256_write_ln109', Compute_HLS.cpp:109) of variable 'add', Compute_HLS.cpp:113 on local variable 'add256' and 'load' operation ('add256_load', Compute_HLS.cpp:113) on local variable 'add256'.
Throughput	The II Violation in module 'computePointHLS_Pipeline_VITIS_LOOP_109_3' (loop 'VITIS_LOOP_109_3'): Unable to enforce a carried dependence constraint (II = 3, distance = 1, offset = 0) between 'store' operation ('add256_write_ln109', Compute_HLS.cpp:109) of variable 'add', Compute_HLS.cpp:113 on local variable 'add256' and 'load' operation ('add256_load', Compute_HLS.cpp:113) on local variable 'add256'.
Throughput	The II Violation in module 'computePointHLS_Pipeline_VITIS_LOOP_109_3' (loop 'VITIS_LOOP_109_3'): Unable to enforce a carried dependence constraint (II = 4, distance = 1, offset = 0) between 'store' operation ('add256_write_ln109', Compute_HLS.cpp:109) of variable 'add', Compute_HLS.cpp:113 on local variable 'add256' and 'load' operation ('add256_load', Compute_HLS.cpp:113) on local variable 'add256'.
Throughput	The II Violation in module 'computePointHLS_Pipeline_VITIS_LOOP_109_3' (loop 'VITIS_LOOP_109_3'): Unable to enforce a carried dependence constraint (II = 7, distance = 1, offset = 0) between 'store' operation ('add256_write_ln109', Compute_HLS.cpp:109) of variable 'add', Compute_HLS.cpp:113 on local variable 'add256' and 'load' operation ('add256_load', Compute_HLS.cpp:113) on local variable 'add256'.
Throughput	The II Violation in module 'computePointHLS_Pipeline_VITIS_LOOP_109_3' (loop 'VITIS_LOOP_109_3'): Unable to enforce a carried dependence constraint (II = 9, distance = 1, offset = 0) between 'store' operation ('add256_write_ln109', Compute_HLS.cpp:109) of variable 'add', Compute_HLS.cpp:113 on local variable 'add256' and 'load' operation ('add256_load', Compute_HLS.cpp:113) on local variable 'add256'.
Throughput	The II Violation in module 'computePointHLS_Pipeline_VITIS_LOOP_109_3' (loop 'VITIS_LOOP_109_3'): Unable to enforce a carried dependence constraint (II = 10, distance = 1, offset = 0) between 'store' operation ('add256_write_ln109', Compute_HLS.cpp:109) of variable 'add', Compute_HLS.cpp:113 on local variable 'add256' and 'load' operation ('add256_load', Compute_HLS.cpp:113) on local variable 'add256'.



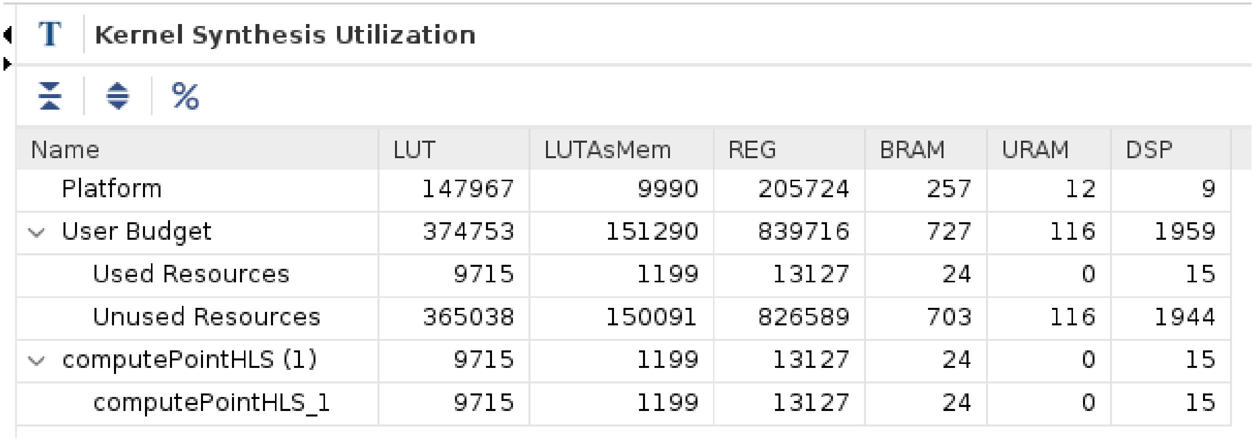
```
Timing Information (MHz)
Compute Unit       Kernel Name      Module Name                                Target Frequency  Estimated Frequency
-----------------  ---------------  -----------------------------------------  ----------------  -------------------
computePointHLS_1  computePointHLS  computePointHLS_Pipeline_VITIS_LOOP_104_2  300.300293        494.559814
computePointHLS_1  computePointHLS  computePointHLS_Pipeline_VITIS_LOOP_109_3  300.300293        370.096252
computePointHLS_1  computePointHLS  computePointHLS                            300.300293        338.868195
```



```
Latency Information
Compute Unit       Kernel Name      Module Name                                Start Interval  Best (cycles)  Avg (cycles)  Worst (cycles)  Best (absolute)  Avg (absolute)  Worst (absolute)
-----------------  ---------------  -----------------------------------------  --------------  -------------  ------------  --------------  ---------------  --------------  ----------------
computePointHLS_1  computePointHLS  computePointHLS_Pipeline_VITIS_LOOP_104_2  34              34             34            34              0.113 us         0.113 us        0.113 us
computePointHLS_1  computePointHLS  computePointHLS_Pipeline_VITIS_LOOP_109_3  296             296            296           296             0.987 us         0.987 us        0.987 us
computePointHLS_1  computePointHLS  computePointHLS                            708748          708747         708747        708747          2.362 ms         2.362 ms        2.362 ms
```

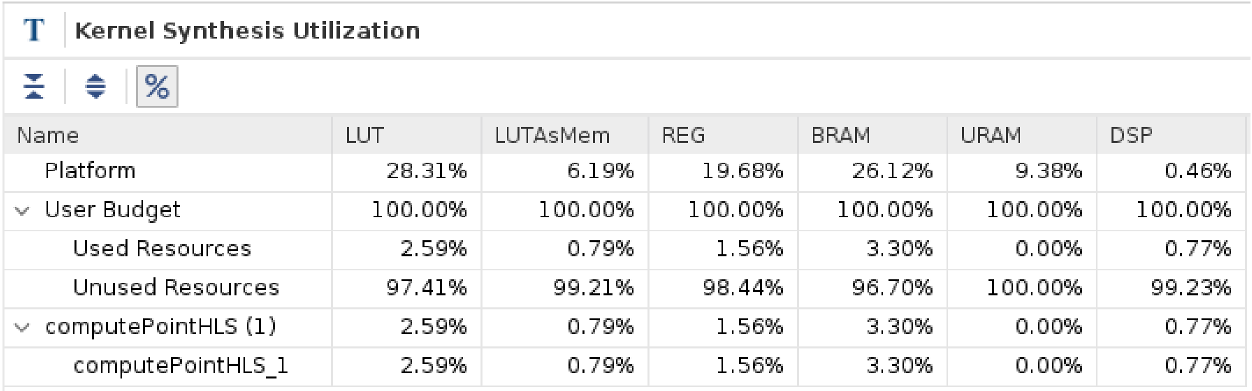


```
Area Information
Compute Unit       Kernel Name      Module Name                                FF     LUT    DSP  BRAM  URAM
-----------------  ---------------  -----------------------------------------  -----  -----  ---  ----  ----
computePointHLS_1  computePointHLS  computePointHLS_Pipeline_VITIS_LOOP_104_2  8      50     0    0     0
computePointHLS_1  computePointHLS  computePointHLS_Pipeline_VITIS_LOOP_109_3  5431   2353   0    0     0
computePointHLS_1  computePointHLS  computePointHLS                            11520  11508  0    2     0
-----------------------------------------------------------------------------
```



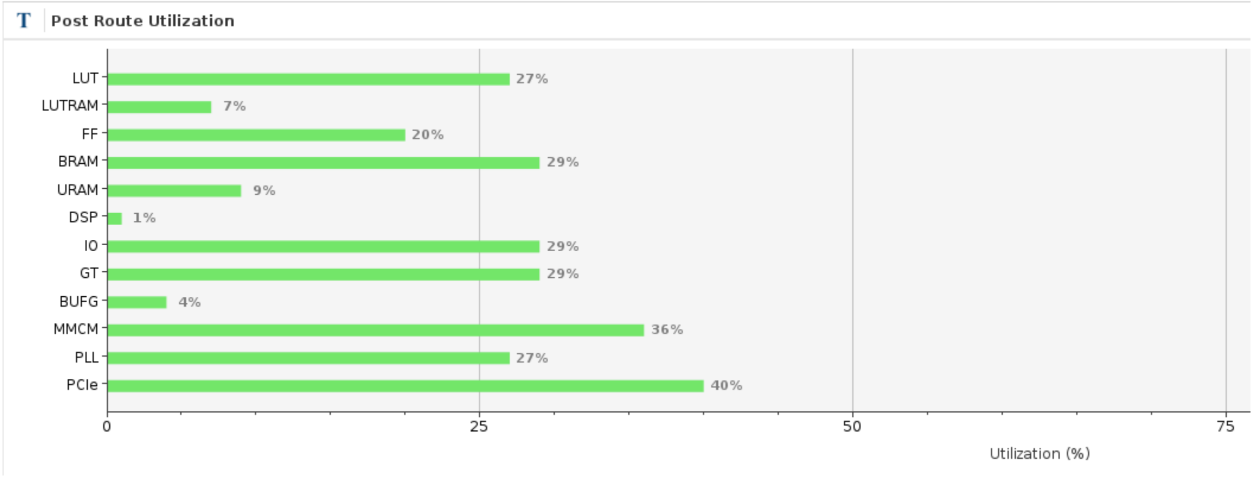
Kernel Synthesis Utilization

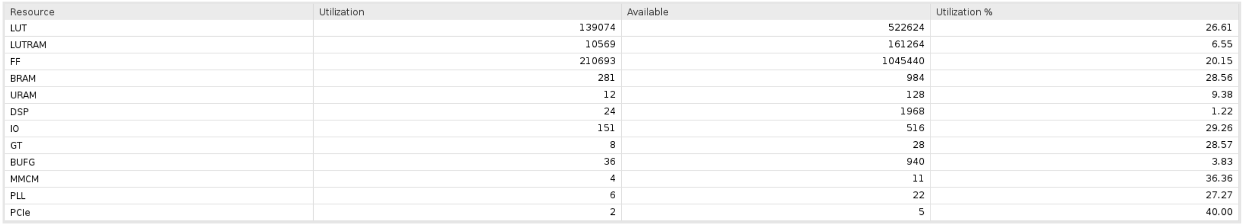
Name	LUT	LUTAsMem	REG	BRAM	URAM	DSP
Platform	147967	9990	205724	257	12	9
User Budget	374753	151290	839716	727	116	1959
Used Resources	9715	1199	13127	24	0	15
Unused Resources	365038	150091	826589	703	116	1944
computePointHLS (1)	9715	1199	13127	24	0	15
computePointHLS_1	9715	1199	13127	24	0	15



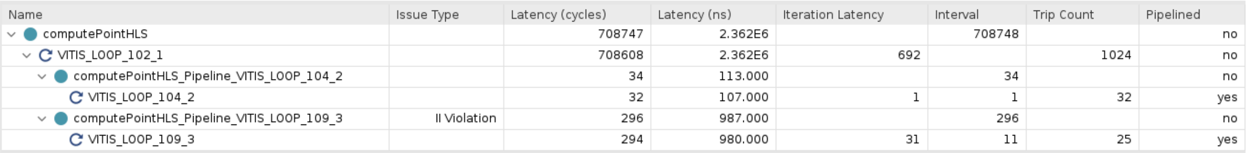
Kernel Synthesis Utilization

Name	LUT	LUTAsMem	REG	BRAM	URAM	DSP
Platform	28.31%	6.19%	19.68%	26.12%	9.38%	0.46%
User Budget	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Used Resources	2.59%	0.79%	1.56%	3.30%	0.00%	0.77%
Unused Resources	97.41%	99.21%	98.44%	96.70%	100.00%	99.23%
computePointHLS (1)	2.59%	0.79%	1.56%	3.30%	0.00%	0.77%
computePointHLS_1	2.59%	0.79%	1.56%	3.30%	0.00%	0.77%

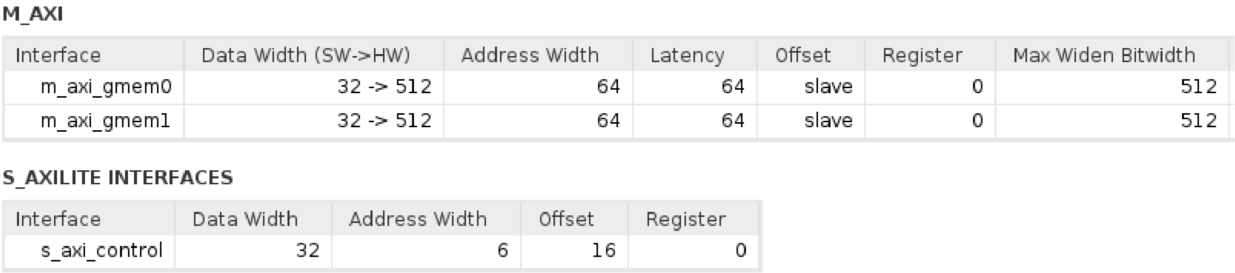




Resource	Utilization	Available	Utilization %
LUT	139074	522624	26.61
LUTRAM	10569	161264	6.55
FF	210693	1045440	20.15
BRAM	281	984	28.56
URAM	12	128	9.38
DSP	24	1968	1.22
IO	151	516	29.26
GT	8	28	28.57
BUFG	36	940	3.83
MMCM	4	11	36.36
PLL	6	22	27.27
PCIe	2	5	40.00



Name	Issue Type	Latency (cycles)	Latency (ns)	Iteration Latency	Interval	Trip Count	Pipelined
computePointHLS		708747	2.362E6		708748		no
VITIS_LOOP_102_1		708608	2.362E6	692		1024	no
computePointHLS_Pipeline_VITIS_LOOP_104_2		34	113.000		34		no
VITIS_LOOP_104_2		32	107.000	1	1	32	yes
computePointHLS_Pipeline_VITIS_LOOP_109_3	II Violation	296	987.000		296		no
VITIS_LOOP_109_3		294	980.000	31	11	25	yes

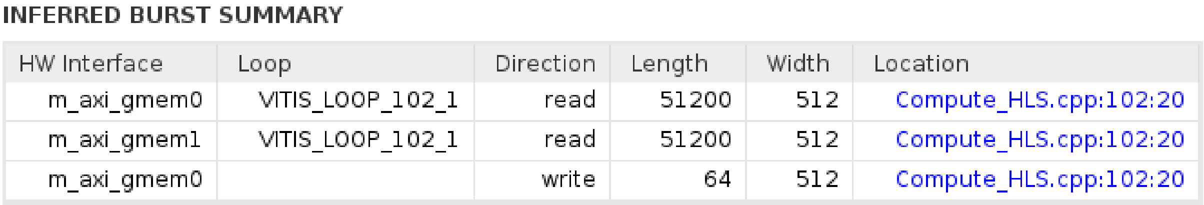


M_AXI

Interface	Data Width (SW->HW)	Address Width	Latency	Offset	Register	Max Widen Bitwidth
m_axi_gmem0	32 -> 512	64	64	slave	0	512
m_axi_gmem1	32 -> 512	64	64	slave	0	512

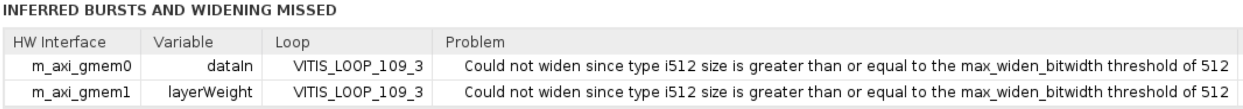
S_AXILITE INTERFACES

Interface	Data Width	Address Width	Offset	Register
s_axi_control	32	6	16	0



INFERRED BURST SUMMARY

HW Interface	Loop	Direction	Length	Width	Location
m_axi_gmem0	VITIS_LOOP_102_1	read	51200	512	Compute_HLS.cpp:102:20
m_axi_gmem1	VITIS_LOOP_102_1	read	51200	512	Compute_HLS.cpp:102:20
m_axi_gmem0		write	64	512	Compute_HLS.cpp:102:20



INFERRED BURSTS AND WIDENING MISSED

HW Interface	Variable	Loop	Problem
m_axi_gmem0	dataIn	VITIS_LOOP_109_3	Could not widen since type i512 size is greater than or equal to the max_widen_bitwidth threshold of 512
m_axi_gmem1	layerWeight	VITIS_LOOP_109_3	Could not widen since type i512 size is greater than or equal to the max_widen_bitwidth threshold of 512