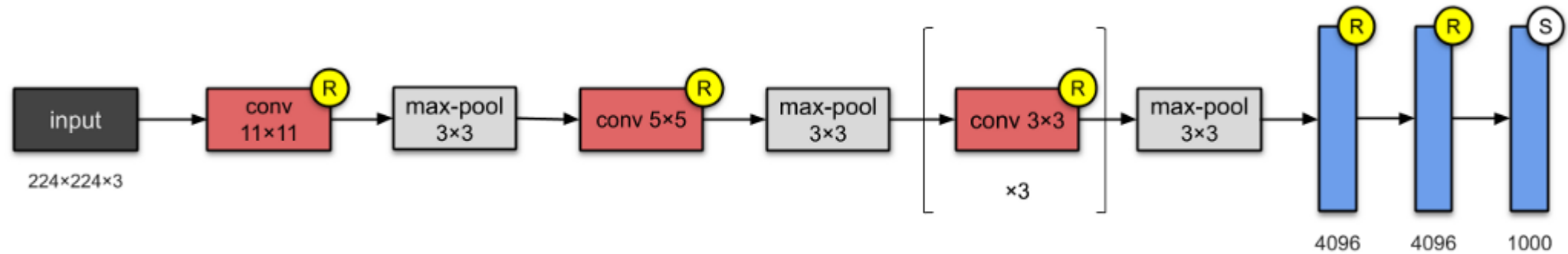


Acceleration of AlexNet for ImageNet Classification

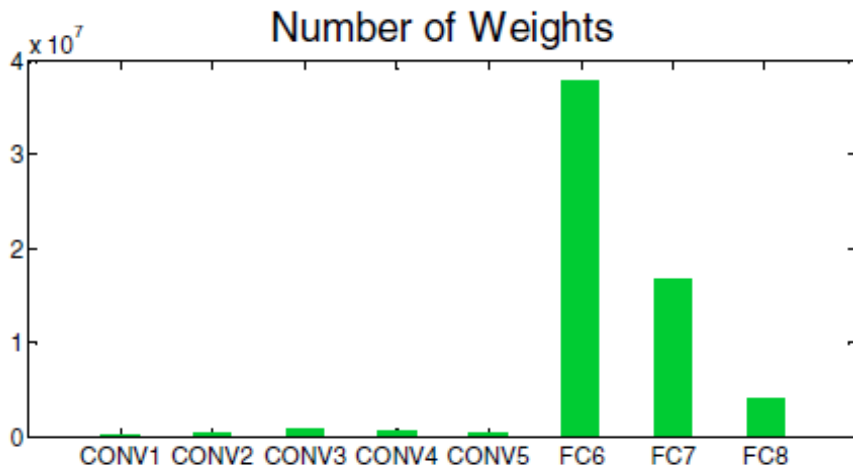
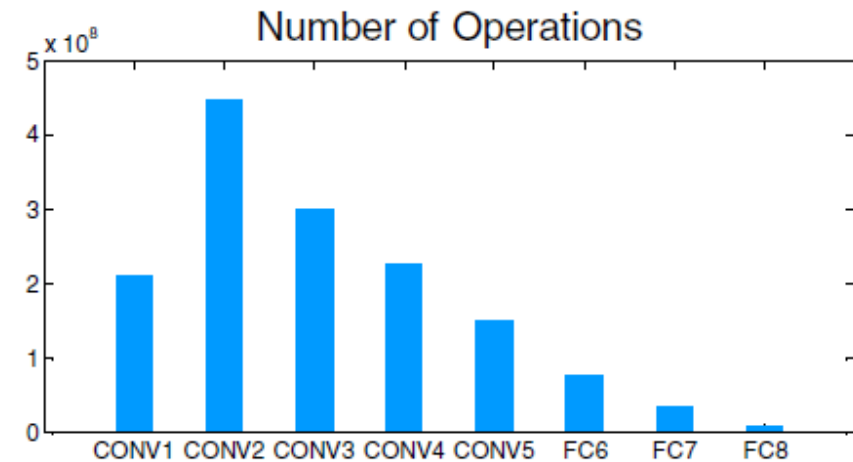


Dhruva Dilip Devasthale

Prabhleen Kaur Gill

Vandita Shetty

Problem statement



- AlexNet Architecture has 14 layers
- Convolution layers are computation intensive (large number of operations)
 - Higher parallelism degree
 - More resources used
- Fully connected layers are memory intensive (large data accesses from memory for weights)
 - Reduce memory bandwidth
- Limited by resources available on board, primary focus was to first bring down resource utilization followed by latency optimization

Workflow

Python Implementation – golden output and layer parameters

C++ Implementation – Individual Layers

C++ Implementation – Binding and logic verification

HLS – Unoptimized individual layers

Binding all unoptimized layers

HLS – Optimized individual layers

Binding all optimized layers

CONV1 deployment on FPGA board

Logic Implementation

Python Implementation

- Model - Tensorflow & Keras
- Dataset- CIFAR-10 (created by Alex Krizhevsky)
- Successfully trained model with 90% accuracy.
- Golden output, weights and bias for each layer generated were stored and used for C implementation

C++ Implementation

- Separate executable for each layer.
- Separate testbench for logic verification of each layer. (MSE (conv1) – 4.83×10^{-14})
- Weights and biases generated from Python were used.
- Wrapper testbench for binding all layers.
- Logic verified by accurate class prediction.

Optimization Techniques

Array Partitioning

Tiling

Unrolling and Pipelining

Data quantization – Accuracy, Speedup and Utilization

Loop Reordering

Efficient data reading and writing – Optimizing DRAM access

Dataflow – memory streaming

CONV1 (3X227X227)

Product family: zynqplus

Target device: xczu3eg-sbva484-1-e

			BRAM_18K	DSP	FF	LUT	LATENCY (ms)
AVAILABLE			432	360	141120	70560	
FXT <32, 8>	UNOPTIMIZED	TOTAL	798	30	4542	7331	8441
		UTILIZATION	184%	8%	3%	10%	
	OPTIMIZED	TOTAL	280	43	5699	11444	373 (22.63x)
		UTILIZATION	64%	11%	4%	16%	
FXT <16, 4> Underflow issue	UNOPTIMIZED	TOTAL	414	16	4012	7001	8441
		UTILIZATION	95%	4%	2%	9%	
	OPTIMIZED	TOTAL	159	21	5136	10308	1244 (6.78x)
		UTILIZATION	36%	5%	3%	14%	

- Output dimensions – 96x55x55
- OUT_BUF size – 32x11x11

			BRAM_18K	DSP	FF	LUT	LATENCY (ms)
AVAILABLE			432	360	141120	70560	
Conv2 (96X27x27)	UNOPTIMIZED	TOTAL	828	12	3835	6805	8969
		UTILIZATION	191%	3%	2%	9%	
	OPTIMIZED	TOTAL	27	11	10637	10855	47.66 (188.2x)
		UTILIZATION	6%	3%	7%	15%	
Conv3 (256x13x13)	UNOPTIMIZED	TOTAL	973	12	3499	7174	3001
		UTILIZATION	225%	3%	2%	10%	
	OPTIMIZED	TOTAL	47	20	11217	17091	11.57 (259.4x)
		UTILIZATION	10%	5%	7%	24%	
Conv4 (384x13x13)	UNOPTIMIZED	TOTAL	260	5	2727	4863	252
		UTILIZATION	60%	1%	1%	6%	
	OPTIMIZED	TOTAL	223	200	9146	15378	5.252 (47.98X)
		UTILIZATION	51%	55	6	21	
Conv5 (384x13x13)	UNOPTIMIZED	TOTAL	197	5	2722	4834	168
		UTILIZATION	45%	1	1	6	
	OPTIMIZED	TOTAL	172	69	5634	12260	4.362 (38.51x)
		UTILIZATION	39%	19	3	17	

Fully Connected - 1

		BRAM_18K	DSP	FF	LUT	LATENCY (ms)
UNOPTIMIZED	AVAILABLE	432	360	141120	70560	
	TOTAL	34581	2	2761	4151	1133
	UTILIZATION	8004%	~0%	1%	5%	
OPTIMIZED	TOTAL	44	32	3925	5441	379 (3x)
	UTILIZATION	10%	8%	2%	7%	

- FC1 dimensions = 9216 x 4096
- Tiling buffer dimension = 32 x 256 (input depth x output depth)
- Pipelining on input depth.
- Further tried to optimize latency by incorporating array partitioning in the weights and output array. But saw no improvement in latency, with an increase in resource utilization.

			BRAM_18K	DSP	FF	LUT	LATENCY (ms)
		AVAILABLE	432	360	141120	70560	
FC2 4096 x 4096	UNOPTIMIZED	TOTAL	29720	6	2124	3374	336
		UTILIZATION	6879%	1%	1%	4%	
	OPTIMIZED	TOTAL	27	16	3892	5255	167 (2x)
		UTILIZATION	6%	4%	2%	7%	
FC3 4096 x 10	UNOPTIMIZED	TOTAL	51	18	4319	5257	1.23
		UTILIZATION	11%	5%	3%	5%	
	OPTIMIZED	TOTAL	8	16	4376	5104	0.459 (2.7x)
		UTILIZATION	1%	4%	3%	7%	
FLAT 9216	UNOPTIMIZED	TOTAL	24	0	1017	2167	0.277
		UTILIZATION	5%	~0%	~0%	3%	
	OPTIMIZED	TOTAL	4	0	1126	3230	0.092 (3x)
		UTILIZATION	~0%	~0%	~0%	4%	

*Flat is the connecting layer b/w CNN and ANN layers to flatten data from 3 dimension to 1 dimension

Results – Maxpool/ LRN

		BRAM_18K	DSP	FF	LUT	LATENCY (ms)
AVAILABLE		432	360	141120	70560	
Maxpool1	UNOPTIMIZED	510%	1%	1%	5%	4.304
	OPTIMIZED	44%	1%	9%	27%	2.968 (1.45x)
Maxpool2	UNOPTIMIZED	330%	0%	1%	4%	2.732
	OPTIMIZED	~0%	0%	27%	81%	1.885 (1.45x)
Maxpool5	UNOPTIMIZED	80%	0%	0%	4%	0.617
	OPTIMIZED	~0%	0%	6%	21%	0.441 (1.4x)
LRN1	UNOPTIMIZED	124%	0%	2%	5%	8.713
	OPTIMIZED	~0%	0%	1%	3%	2.9 (3x)
LRN2	UNOPTIMIZED	40%	1%	0%	3%	11.213
	OPTIMIZED	~0%	0%	0%	2%	1.866 (6x)

Unoptimized Binding Results

```
=====
== Utilization Estimates
=====
* Summary:
```

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	2	-
FIFO	-	-	-	-	-
Instance	3532	67	14855	28759	0
Memory	55118	-	0	0	0
Multiplexer	-	-	-	1533	-
Register	-	-	116	-	-
Total	58650	67	14971	30294	0
Available SLR	1440	2280	788160	394080	320
Utilization SLR (%)	4072	2	1	7	0
Available	2880	4560	1576320	788160	640
Utilization (%)	2036	1	~0	3	0

Product family: virtexuplus

Target device: xqvu7p-flra2104-1-I

Unoptimized Latency = 15.6 sec

<16, 4> fixed type data used

Optimized Binding Results

=====						
== Utilization Estimates						
=====						
* Summary:						
+-----+-----+-----+-----+-----+						
Name	BRAM_18K	DSP	FF	LUT	URAM	
+-----+-----+-----+-----+-----+						
DSP	-	-	-	-	-	
Expression	-	-	0	284	-	
FIFO	-	-	-	-	-	
Instance	666	212	27459	50412	0	
Memory	1194	-	0	0	0	
Multiplexer	-	-	-	4391	-	
Register	-	-	203	-	-	
+-----+-----+-----+-----+-----+						
Total	1860	212	27662	55087	0	
+-----+-----+-----+-----+-----+						
Available SLR	1440	2280	788160	394080	320	
+-----+-----+-----+-----+-----+						
Utilization SLR (%)	129	9	3	13	0	
+-----+-----+-----+-----+-----+						
Available	2880	4560	1576320	788160	640	
+-----+-----+-----+-----+-----+						
Utilization (%)	64	4	1	6	0	
+-----+-----+-----+-----+-----+						

Product family: virtexuplus

Target device: xqvu7p-flra2104-1-I

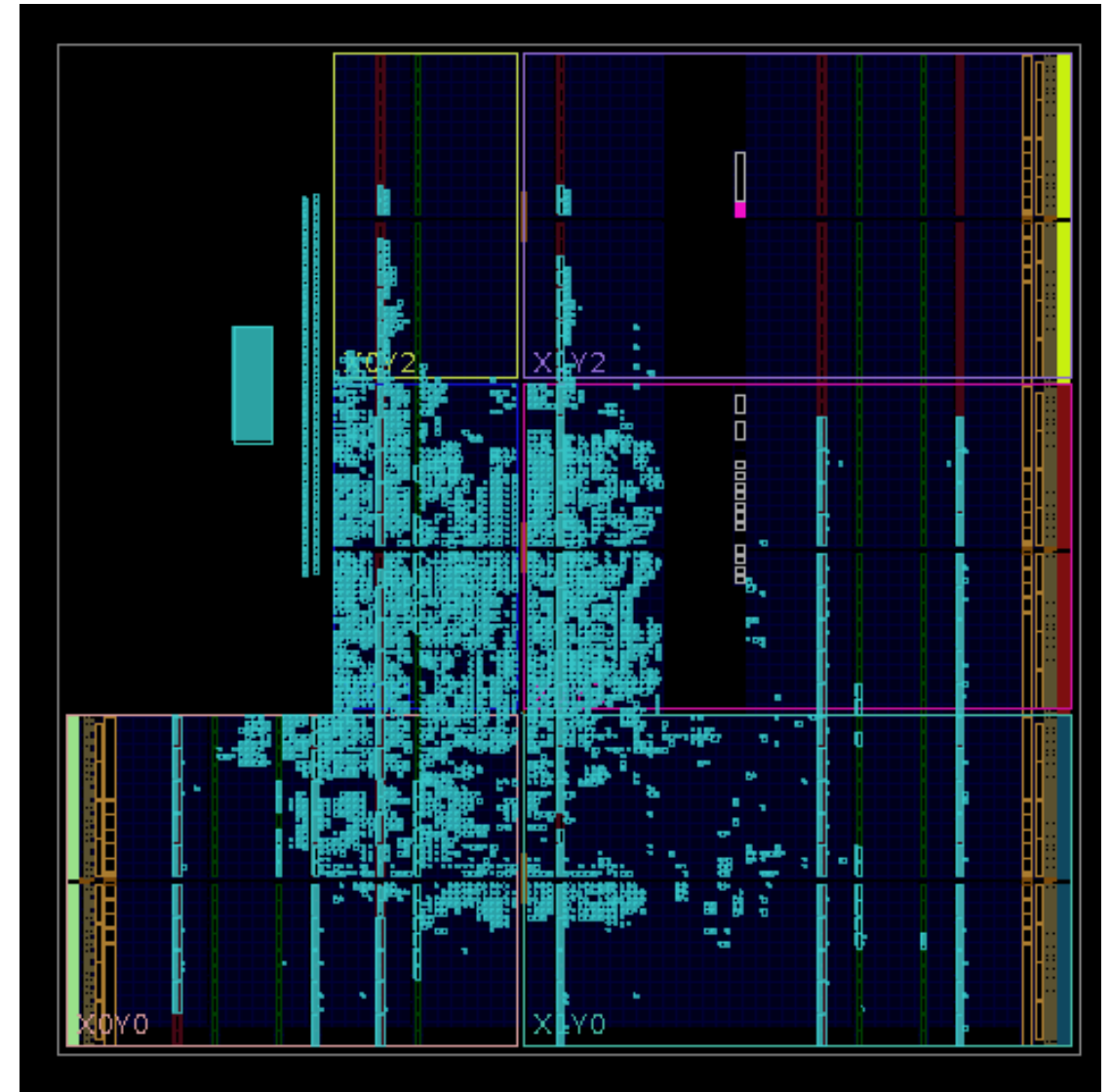
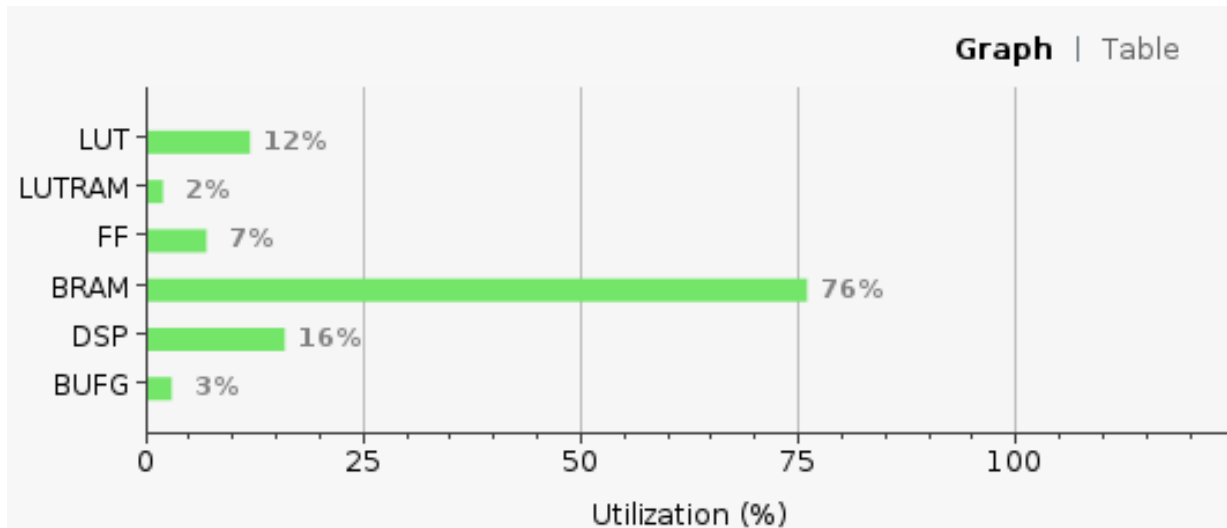
Optimized Latency =1.877 sec

Speedup = 15.6/ 1.877
= 8.3x

<16, 4> fixed type data used

Deployment

Successfully deployed CONV1 on
Pynq-Z2 (xc7z020clg400-1)



DSE and Learnings

- Data Quantization
 - Tested different fixed type dimensions $\langle 16, x \rangle$, $\langle 32, x \rangle$
 - Checked the latency, utilization, accuracy
 - Used $\langle 16, 4 \rangle$ for binding layers.
- On board storage
 - FPGA resources limited
 - During binding, could not store intermediate outputs on board (sacrificed latency)
- Used “[hls_math.h](#)” library for synthesis of LRN and FC3 layer on the Vitis HLS tool.

Thank You

Appendix – conv1 <16,4>

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	154	-
FIFO	-	-	-	-	-
Instance	148	21	4797	9509	0
Memory	11	-	0	0	0
Multiplexer	-	-	-	645	-
Register	-	-	339	-	-
Total	159	21	5136	10308	0
Available	432	360	141120	70560	0
Utilization (%)	36	5	3	14	0

Appendix – conv1 <32,8>

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	157	-
FIFO	-	-	-	-	-
Instance	258	43	5358	10642	0
Memory	22	-	0	0	0
Multiplexer	-	-	-	645	-
Register	-	-	341	-	-
Total	280	43	5699	11444	0
Available	432	360	141120	70560	0
Utilization (%)	64	11	4	16	0

Appendix – conv2

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	145	-
FIFO	-	-	-	-	-
Instance	18	11	10302	10157	0
Memory	9	-	0	0	0
Multiplexer	-	-	-	553	-
Register	-	-	335	-	-
Total	27	11	10637	10855	0
Available	432	360	141120	70560	0
Utilization (%)	6	3	7	15	0

Appendix – conv3

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	25	-
FIFO	-	-	-	-	-
Instance	34	20	10940	16365	0
Memory	13	-	0	0	0
Multiplexer	-	-	-	701	-
Register	-	-	277	-	-
Total	47	20	11217	17091	0
Available	432	360	141120	70560	0
Utilization (%)	10	5	7	24	0

Appendix – conv4

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	22	-
FIFO	-	-	-	-	-
Instance	197	200	8872	14637	0
Memory	26	-	0	0	0
Multiplexer	-	-	-	719	-
Register	-	-	274	-	-
Total	223	200	9146	15378	0
Available	432	360	141120	70560	0
Utilization (%)	51	55	6	21	0

Appendix – conv5

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	23	-
FIFO	-	-	-	-	-
Instance	159	69	5362	11536	0
Memory	13	-	0	0	0
Multiplexer	-	-	-	701	-
Register	-	-	272	-	-
Total	172	69	5634	12260	0
Available	432	360	141120	70560	0
Utilization (%)	39	19	3	17	0

Appendix – maxpool1

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	66	18	-
FIFO	-	-	-	-	-
Instance	194	5	13332	19353	0
Memory	-	-	-	-	-
Multiplexer	-	-	-	18	-
Register	-	-	8	-	-
Total	194	5	13406	19389	0
Available	432	360	141120	70560	0
Utilization (%)	44	1	9	27	0

Appendix – maxpool2

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	97	25	-
FIFO	-	-	-	-	-
Instance	2	1	39236	57798	0
Memory	-	-	-	-	-
Multiplexer	-	-	-	18	-
Register	-	-	12	-	-
Total	2	1	39345	57841	0
Available	432	360	141120	70560	0
Utilization (%)	~0	~0	27	81	0

Appendix – maxpool5

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	83	22	-
FIFO	-	-	-	-	-
Instance	2	0	9647	15225	0
Memory	-	-	-	-	-
Multiplexer	-	-	-	18	-
Register	-	-	10	-	-
Total	2	0	9740	15265	0
Available	432	360	141120	70560	0
Utilization (%)	~0	0	6	21	0

Appendix – fc1

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	97	-
FIFO	-	-	-	-	-
Instance	43	32	3588	5177	0
Memory	1	-	0	0	0
Multiplexer	-	-	-	167	-
Register	-	-	337	-	-
Total	44	32	3925	5441	0
Available	432	360	141120	70560	0
Utilization (%)	10	8	2	7	0

Appendix – fc2

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	97	-
FIFO	-	-	-	-	-
Instance	26	16	3555	4991	0
Memory	1	-	0	0	0
Multiplexer	-	-	-	167	-
Register	-	-	337	-	-
Total	27	16	3892	5255	0
Available	432	360	141120	70560	0
Utilization (%)	6	4	2	7	0

Appendix – fc3

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	0	46	-
FIFO	-	-	-	-	-
Instance	8	16	4055	4852	0
Memory	0	-	32	3	0
Multiplexer	-	-	-	203	-
Register	-	-	289	-	-
Total	8	16	4376	5104	0
Available	432	360	141120	70560	0
Utilization (%)	1	4	3	7	0

Appendix – lrn1

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	-	-	-
FIFO	-	-	-	-	-
Instance	4	-	1337	1875	0
Memory	-	-	-	-	-
Multiplexer	-	-	-	370	-
Register	-	-	142	-	-
Total	4	0	1479	2245	0
Available	432	360	141120	70560	0
Utilization (%)	~0	0	1	3	0

Appendix – lrn2

Name	BRAM_18K	DSP	FF	LUT	URAM
DSP	-	-	-	-	-
Expression	-	-	-	-	-
FIFO	-	-	-	-	-
Instance	2	-	788	1144	0
Memory	-	-	-	-	-
Multiplexer	-	-	-	370	-
Register	-	-	142	-	-
Total	2	0	930	1514	0
Available	432	360	141120	70560	0
Utilization (%)	~0	0	~0	2	0