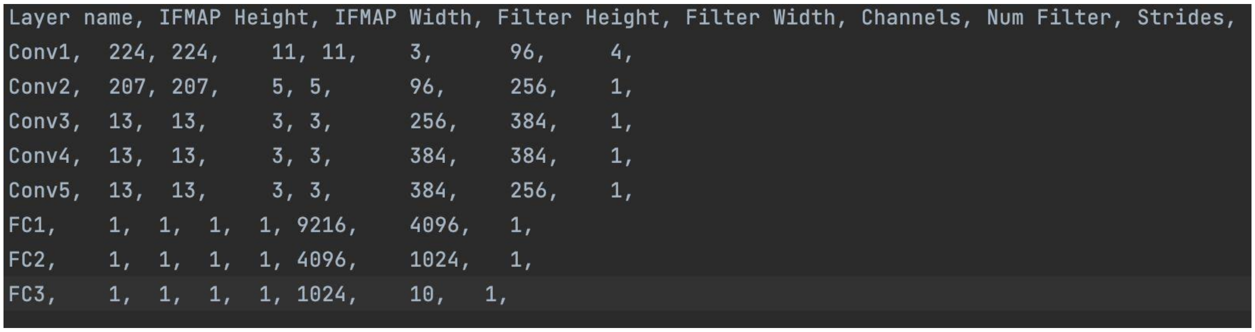
Lab 4 Report

Aaryan Agarwal aaryagar

Part 1

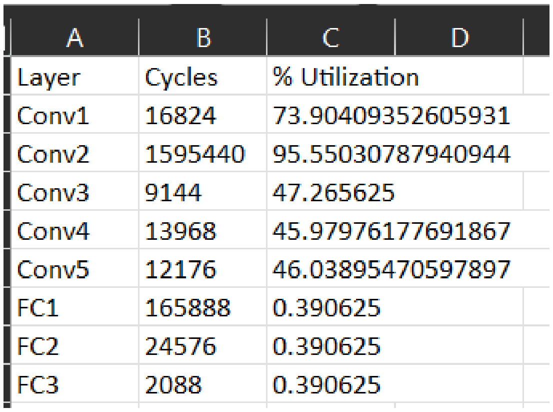
Graph for Alexnet TPU with os is and ws dataflow.

Alexnet structure



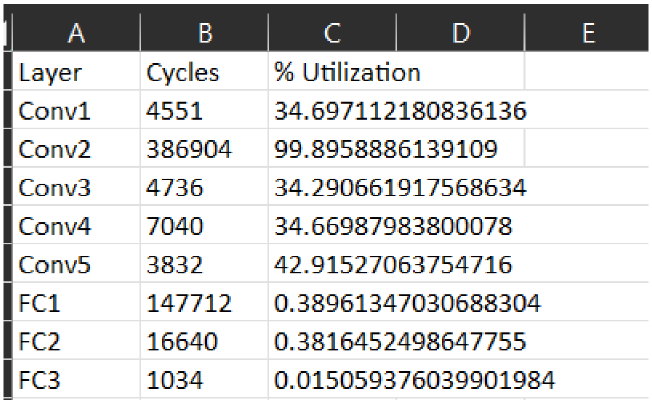
```
Layer name, IFMAP Height, IFMAP Width, Filter Height, Filter Width, Channels, Num Filter, Strides,
Conv1,  224, 224,    11, 11,    3,      96,     4,
Conv2,  207, 207,    5, 5,      96,     256,    1,
Conv3,  13,  13,     3, 3,      256,    384,    1,
Conv4,  13,  13,     3, 3,      384,    384,    1,
Conv5,  13,  13,     3, 3,      384,    256,    1,
FC1,    1,  1,  1,  1, 9216,    4096,   1,
FC2,    1,  1,  1,  1, 4096,    1024,   1,
FC3,    1,  1,  1,  1, 1024,    10,   1,
```

TPU IS



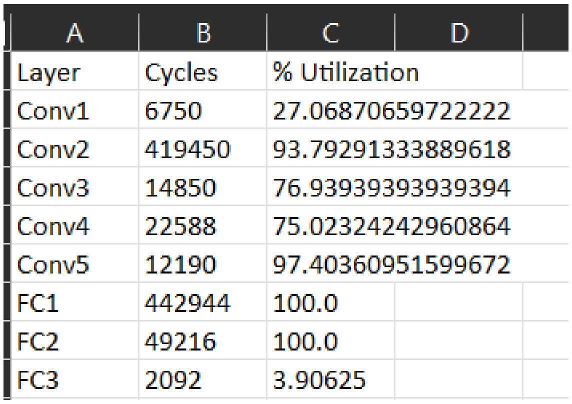
A	B	C	D
Layer	Cycles	% Utilization	
Conv1	16824	73.90409352605931	
Conv2	1595440	95.55030787940944	
Conv3	9144	47.265625	
Conv4	13968	45.97976177691867	
Conv5	12176	46.03895470597897	
FC1	165888	0.390625	
FC2	24576	0.390625	
FC3	2088	0.390625	

TPU OS

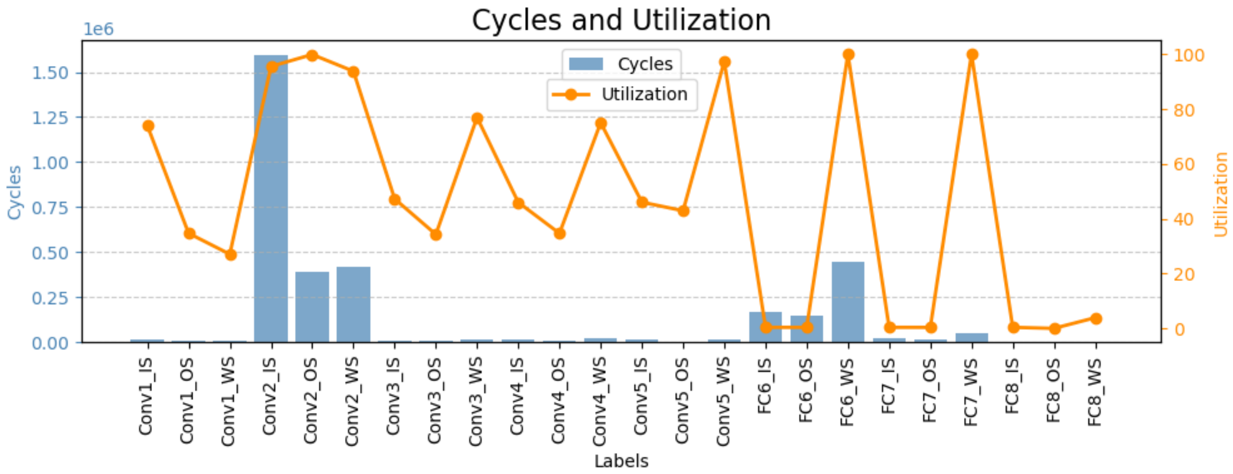


A	B	C	D	E
Layer	Cycles	% Utilization		
Conv1	4551	34.697112180836136		
Conv2	386904	99.8958886139109		
Conv3	4736	34.290661917568634		
Conv4	7040	34.66987983800078		
Conv5	3832	42.91527063754716		
FC1	147712	0.38961347030688304		
FC2	16640	0.3816452498647755		
FC3	1034	0.015059376039901984		

TPU WS



A	B	C	D
Layer	Cycles	% Utilization	
Conv1	6750	27.06870659722222	
Conv2	419450	93.79291333889618	
Conv3	14850	76.93939393939394	
Conv4	22588	75.02324242960864	
Conv5	12190	97.40360951599672	
FC1	442944	100.0	
FC2	49216	100.0	
FC3	2092	3.90625	

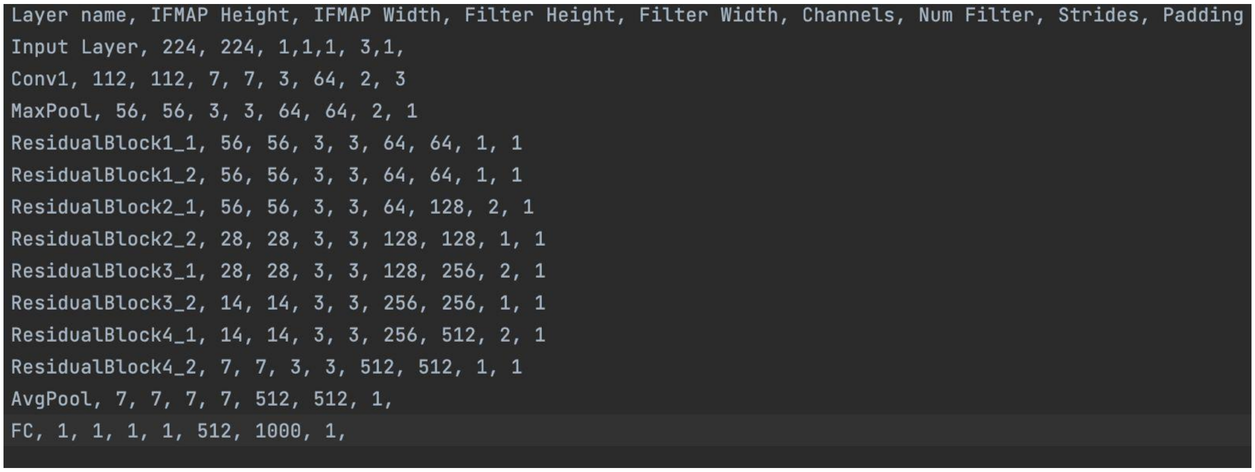


The second layer has the most cycles and takes the maximum amount of time to execute. TPU has parallelization because of which the utilization in convolution layer is more than the fully connected layers. We can see that on average utilization in FC is lesser than the convolution layers because of lack of parallelization.

Second convolution layer is the most complex which is why it takes more cycles.

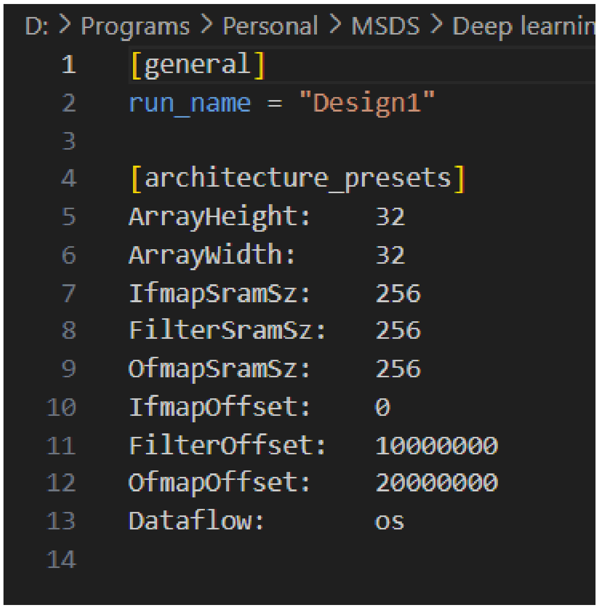
Part 2

Resnet structure:



```
Layer name, IFMAP Height, IFMAP Width, Filter Height, Filter Width, Channels, Num Filter, Strides, Padding
Input Layer, 224, 224, 1,1,1, 3,1,
Conv1, 112, 112, 7, 7, 3, 64, 2, 3
MaxPool, 56, 56, 3, 3, 64, 64, 2, 1
ResidualBlock1_1, 56, 56, 3, 3, 64, 64, 1, 1
ResidualBlock1_2, 56, 56, 3, 3, 64, 64, 1, 1
ResidualBlock2_1, 56, 56, 3, 3, 64, 128, 2, 1
ResidualBlock2_2, 28, 28, 3, 3, 128, 128, 1, 1
ResidualBlock3_1, 28, 28, 3, 3, 128, 256, 2, 1
ResidualBlock3_2, 14, 14, 3, 3, 256, 256, 1, 1
ResidualBlock4_1, 14, 14, 3, 3, 256, 512, 2, 1
ResidualBlock4_2, 7, 7, 3, 3, 512, 512, 1, 1
AvgPool, 7, 7, 7, 7, 512, 512, 1,
FC, 1, 1, 1, 1, 512, 1000, 1,
```

Design 1



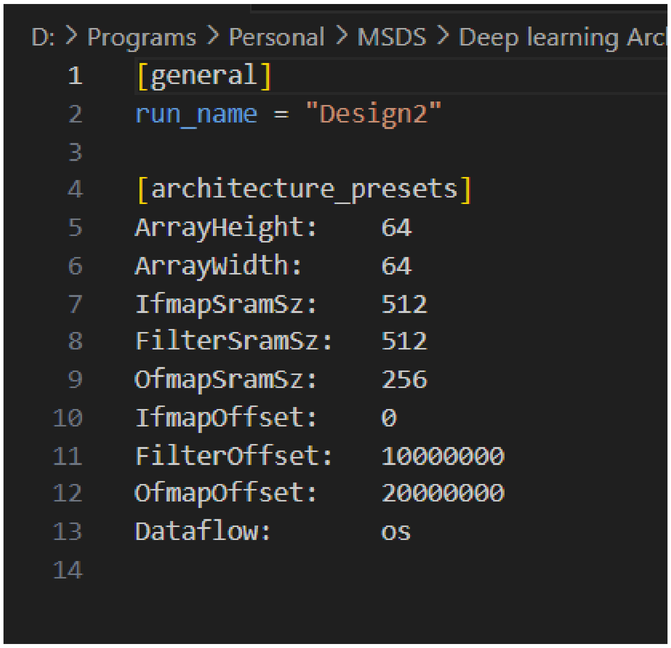
D: > Programs > Personal > MSDS > Deep learnin

```
[general]
run_name = "Design1"

[architecture_presets]
ArrayHeight:    32
ArrayWidth:     32
IfmapSramSz:    256
FilterSramSz:   256
OfmapSramSz:    256
IfmapOffset:    0
FilterOffset:   10000000
OfmapOffset:    20000000
Dataflow:       os

```

Design 2

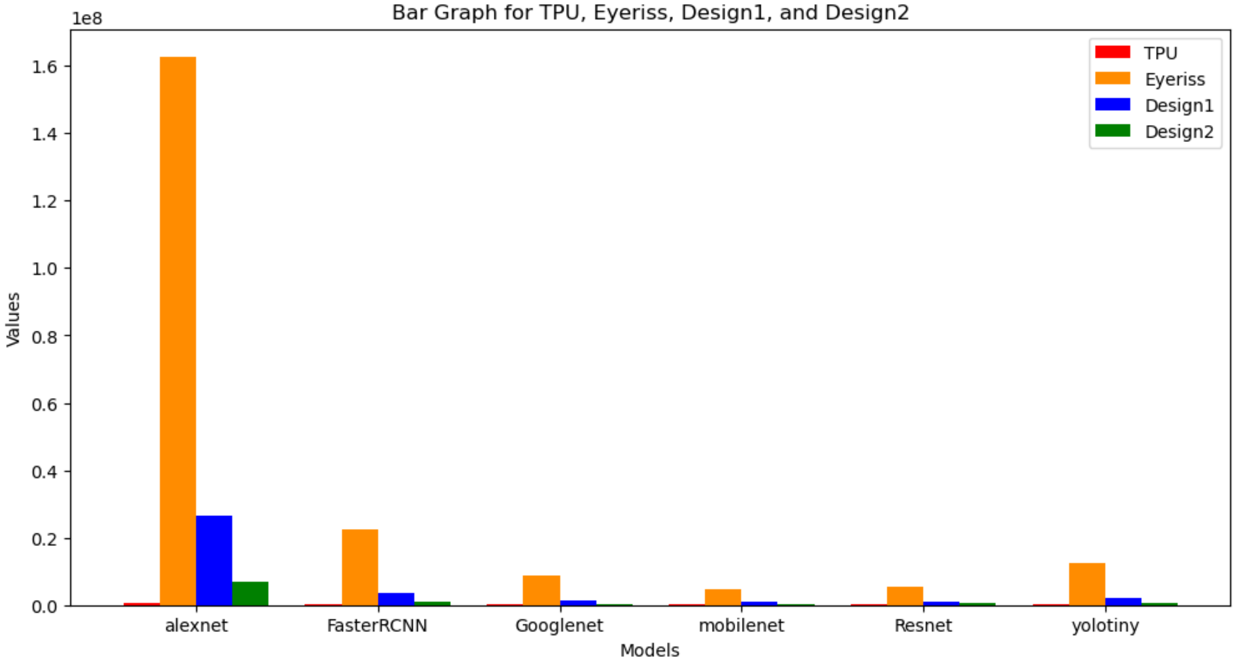


```
[general]
run_name = "Design2"

[architecture_presets]
ArrayHeight:    64
ArrayWidth:     64
IfmapSramSz:    512
FilterSramSz:   512
OfmapSramSz:    256
IfmapOffset:    0
FilterOffset:   10000000
OfmapOffset:    20000000
Dataflow:       os

```

Graph for cycles of each algorithm with different configurations



TPU takes the least amount of time on all the algorithms and eyeriss takes the most amount of time to execute. Design 1 and design 2 are somewhere in between the two and design 1 has more execution time than design 2. This is due to the factor of difference in the size of the systolic arrays. Design 1 uses 32x32 systolic arrays and design 2 uses 64x64 systolic arrays hence design 2 performs faster.