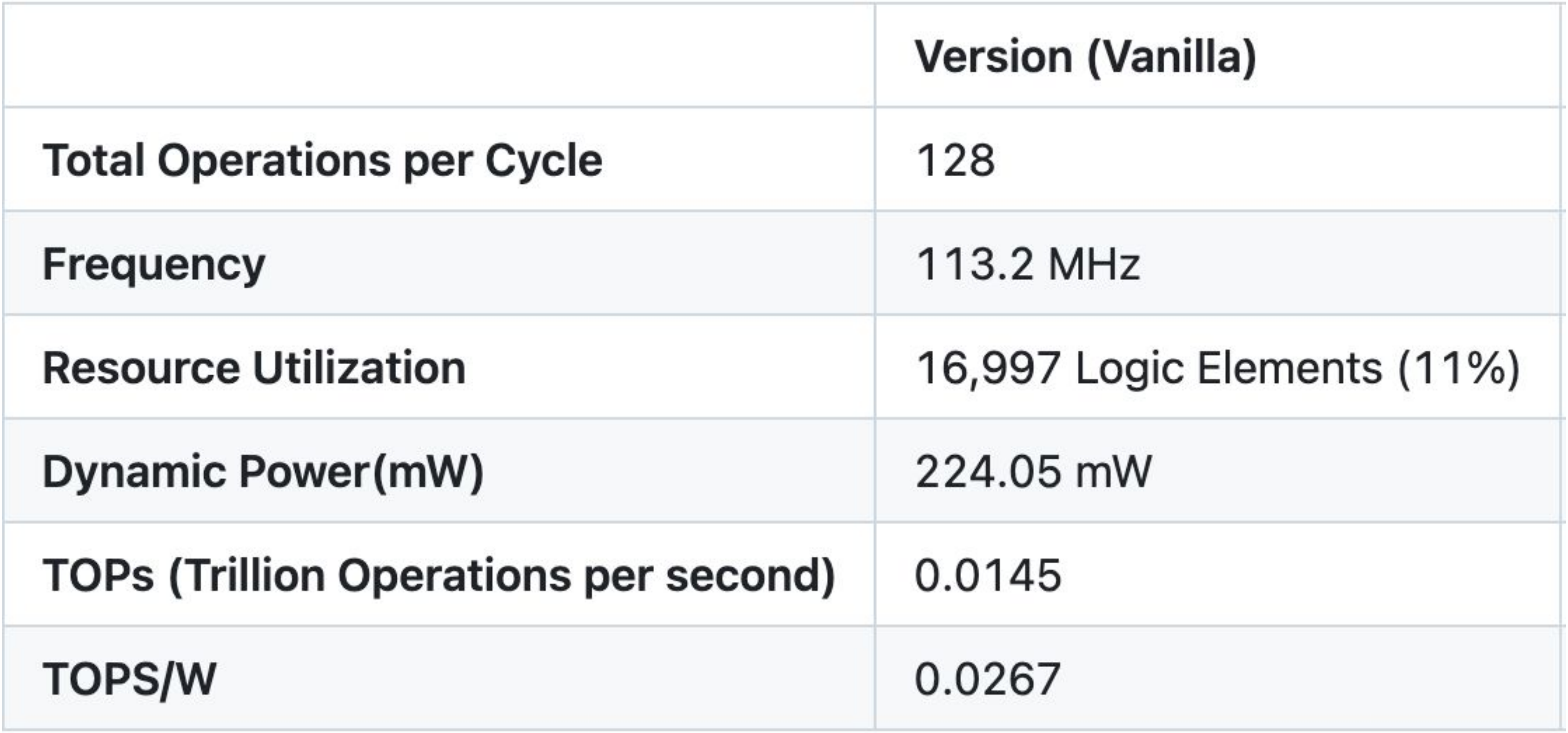
2D Reconfigurable Systolic Array with Physical Board Mapping

Velociraptor Team

Wuqiong Zhao, Yijie He, Matthew Jarecky, Haochen Jiang, Achuth Krishna

	VGGNet (4b weight, 4b activation)	VGGNet (4b weight, 2b activation)
Accuracy	90%	89.03%
Quantization Error	0.00000032691	0.00000022487

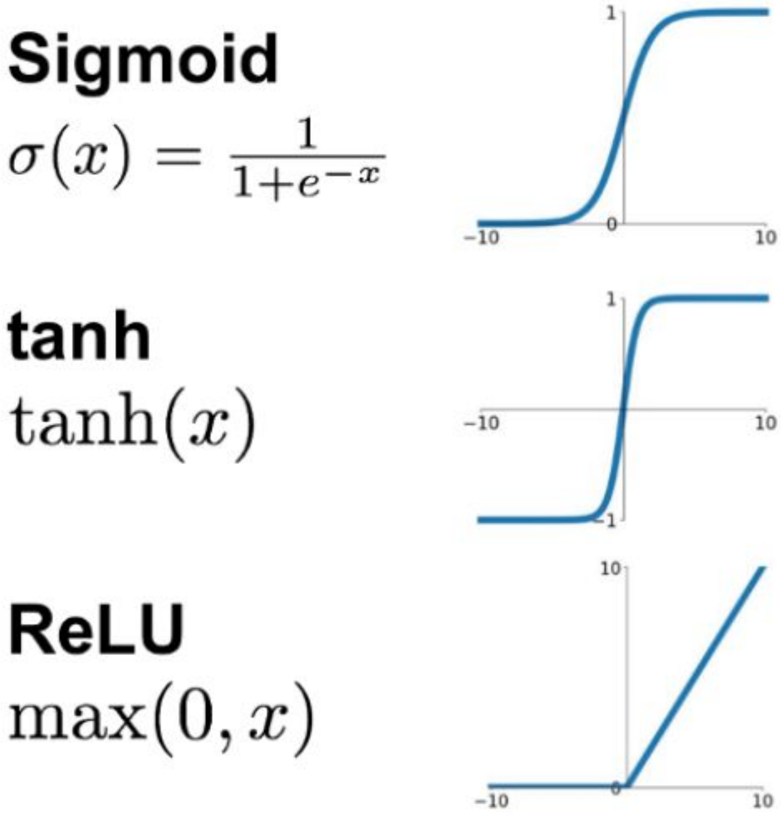
Mapping on FPGA (Cyclone)



	Version (Vanilla)
Total Operations per Cycle	128
Frequency	113.2 MHz
Resource Utilization	16,997 Logic Elements (11%)
Dynamic Power(mW)	224.05 mW
TOPs (Trillion Operations per second)	0.0145
TOPS/W	0.0267

Alpha 1: Reconfigurable SFP for Activation Functions using LUT

- Use lookup tables to efficiently compute different activation functions in the SFP
- Because of quantization, LUT makes it efficient to look up values rather than execute expensive computations
- In addition to ReLU, parameter allows to select Sigmoid or Tanh activation functions
- A value of ‘0’ indicates no activation function used in the SFP



Sigmoid

$\sigma(x) = \frac{1}{1+e^{-x}}$

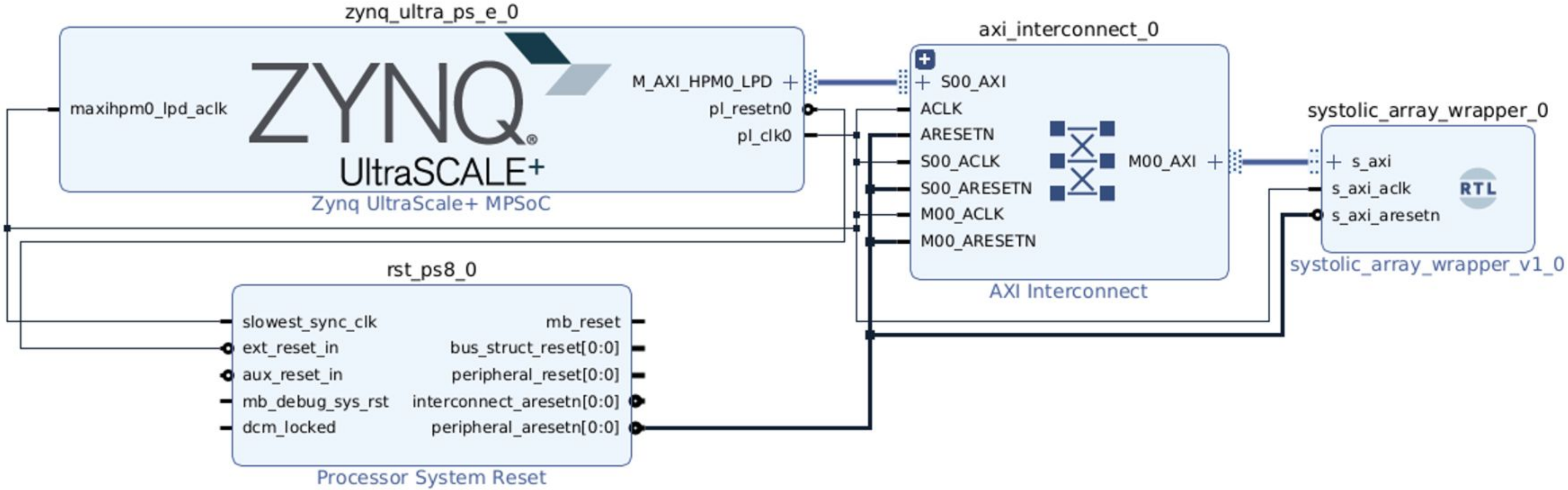
tanh

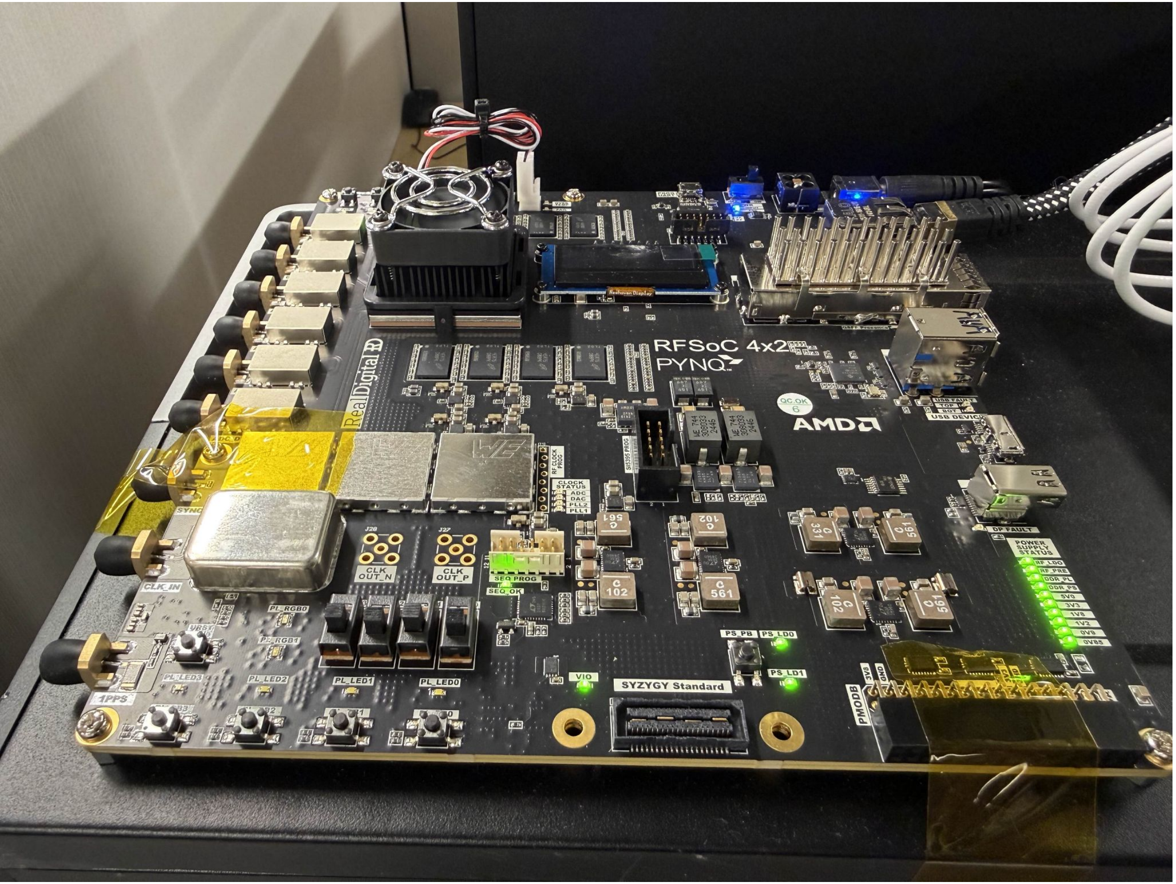
$\tanh(x)$

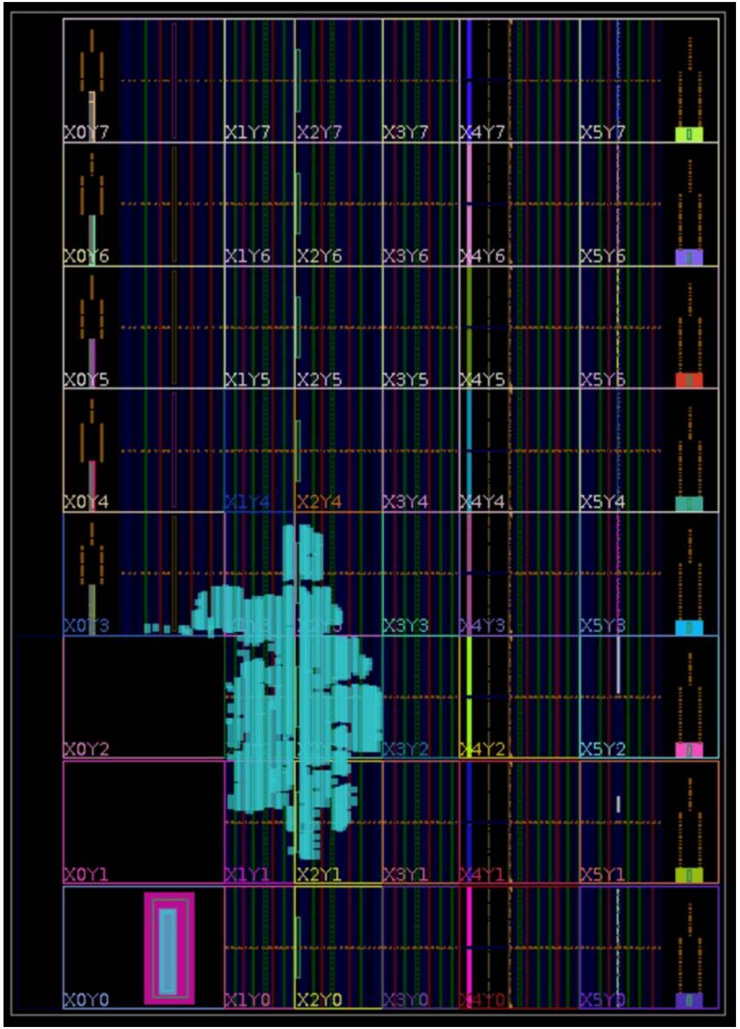
ReLU

$\max(0, x)$

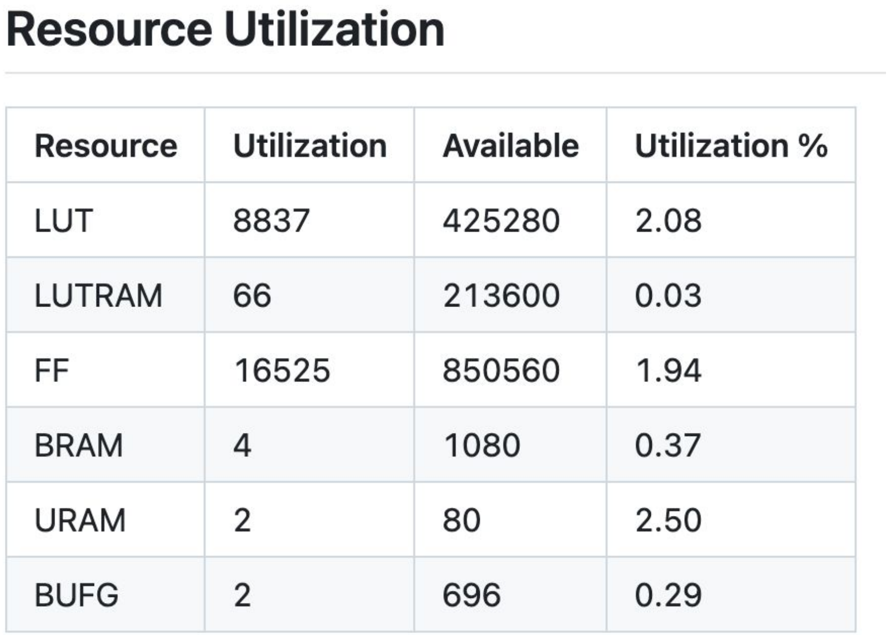
Alpha 2: Physical FPGA (Pynq Board) Mapping




RFSoC 4x2
PYNQ
AMD
RealDigital
QC OK
SYZYGY Standard
CLK OUT_N
CLK OUT_P
J28
J27
CLK_IN
SYNC
1PPS
PL_RGB0
PL_RGB1
PL_LED3
PL_LED2
PL_LED1
PL_LED0
PS_PB
PS_LD0
PS_LD1
VIO
SEQ PROG
SEQ OK
PMODB
3V3
GND
CLOCK STATUS
ADC
DAC
PLL2
PLL1
POWER SUPPLY STATUS
DP FAULT
USB FAULT
USB DEVICE
331
561
102

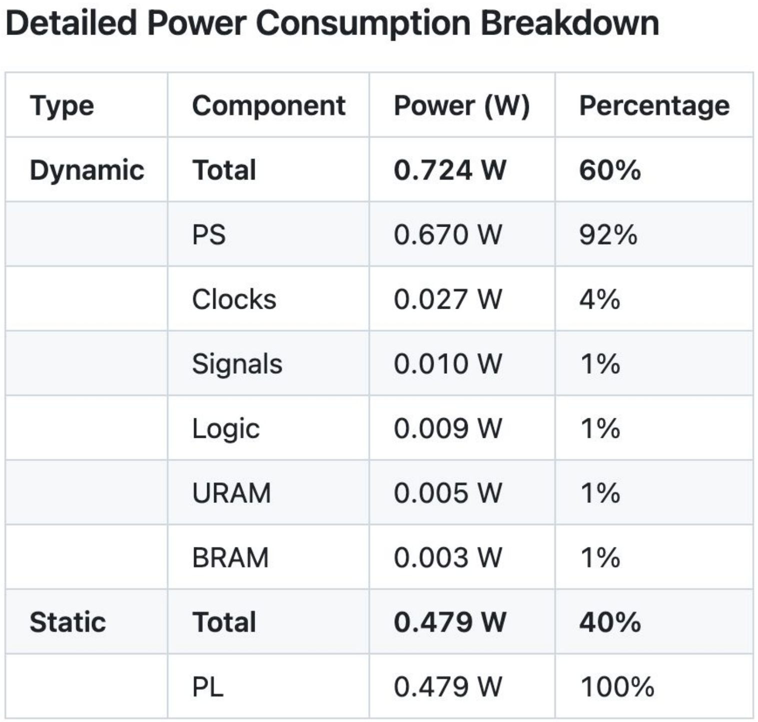

X0Y7
X1Y7
X2Y7
X3Y7
X4Y7
X5Y7
X0Y6
X1Y6
X2Y6
X3Y6
X4Y6
X5Y6
X0Y5
X1Y5
X2Y5
X3Y5
X4Y5
X5Y5
X0Y4
X1Y4
X2Y4
X3Y4
X4Y4
X5Y4
X0Y3
X3Y3
X4Y3
X5Y3
X0Y2
X3Y2
X4Y2
X5Y2
X0Y1
X1Y1
X2Y1
X3Y1
X4Y1
X5Y1
X0Y0
X1Y0
X2Y0
X3Y0
X4Y0
X5Y0

- Mapped systolic array to physical FPGA board
- Stored weights in BRAM
- Functionally verified with part 1 (4 bit weights, 4 bit activations, weight-stationary mapping)



Resource Utilization

Resource	Utilization	Available	Utilization %
LUT	8837	425280	2.08
LUTRAM	66	213600	0.03
FF	16525	850560	1.94
BRAM	4	1080	0.37
URAM	2	80	2.50
BUFG	2	696	0.29

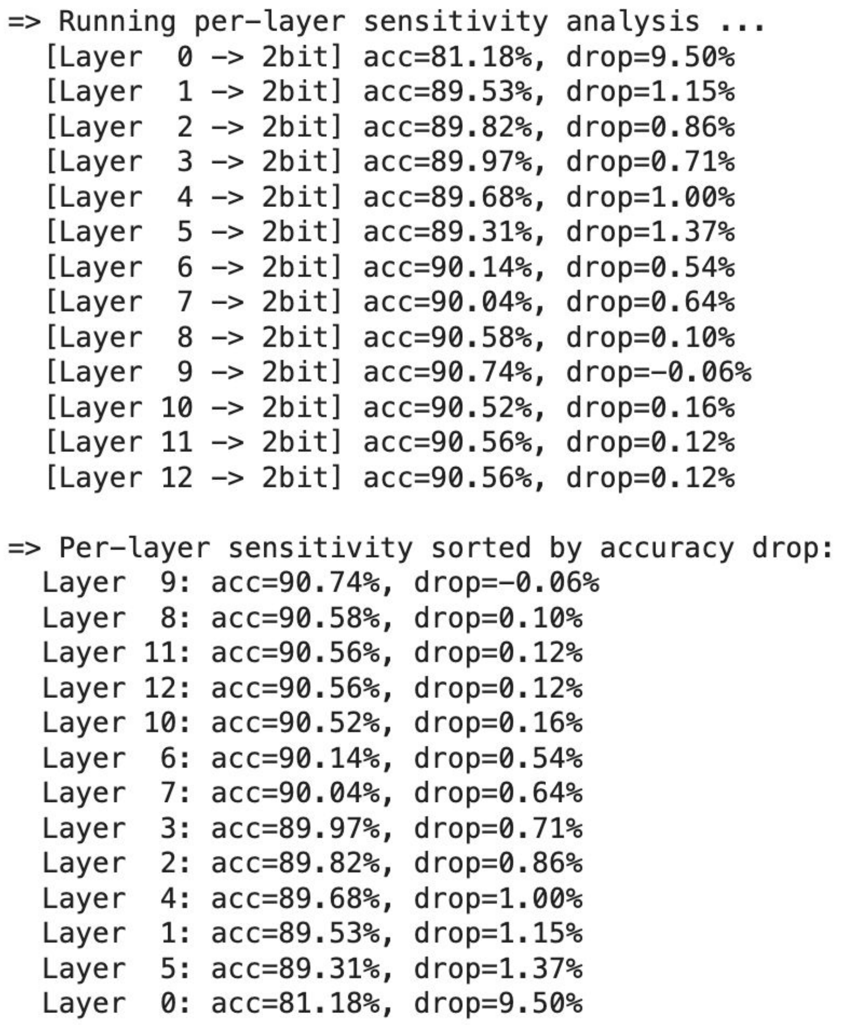


Detailed Power Consumption Breakdown

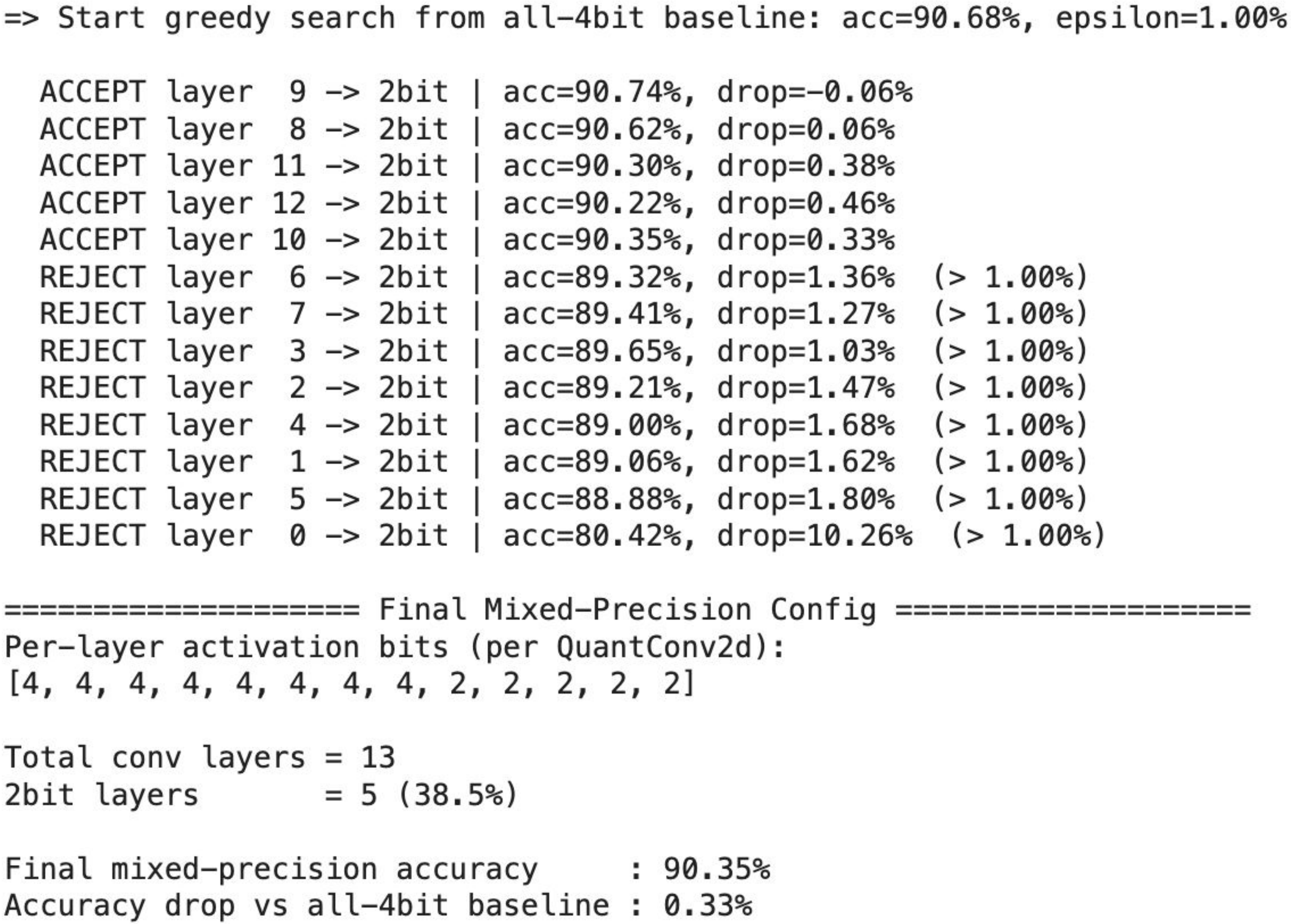
Type	Component	Power (W)	Percentage
Dynamic	**Total**	**0.724 W**	**60%**
	PS	0.670 W	92%
	Clocks	0.027 W	4%
	Signals	0.010 W	1%
	Logic	0.009 W	1%
	URAM	0.005 W	1%
	BRAM	0.003 W	1%
Static	**Total**	**0.479 W**	**40%**
	PL	0.479 W	100%

Alpha 3: VGG16 Layer Quantization Study

- Goal: determine which QuantConv2d layers are most sensitive to activations being quantized to 2 bits
- Weights stay at 4 bits
- Measure accuracy drop for each layer
- Compare to baseline accuracy (no further activation quantization) of 90.86%
- Greedy algorithm with threshold epsilon

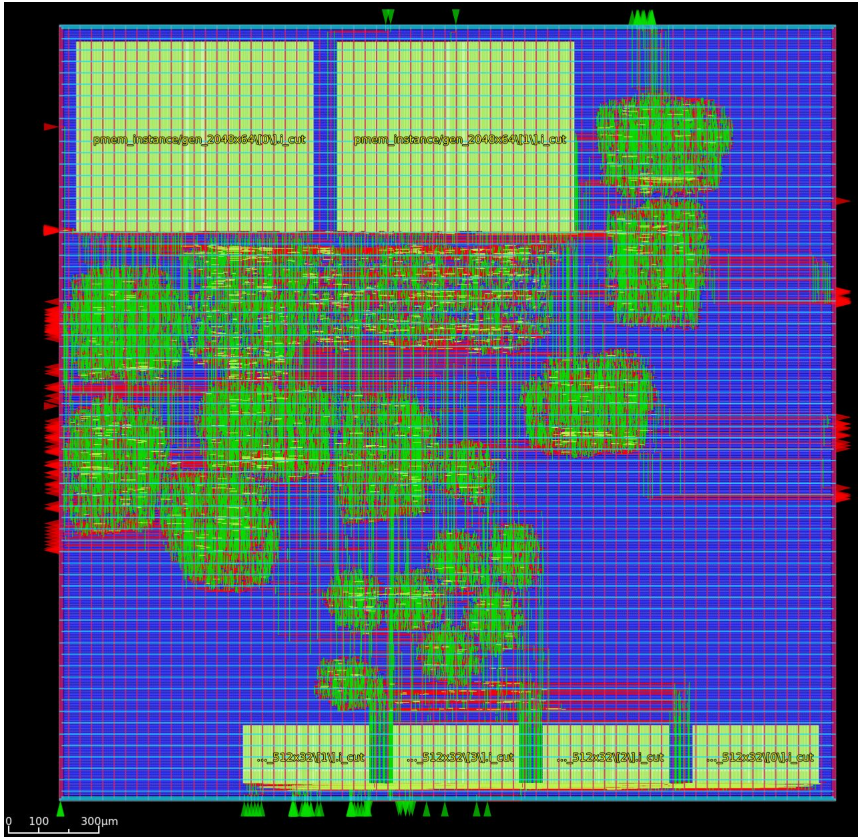


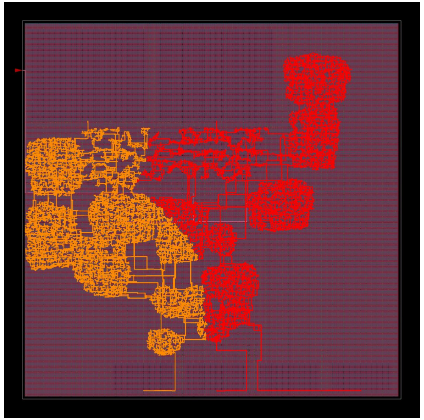
```
=> Running per-layer sensitivity analysis ...
  [Layer  0 -> 2bit] acc=81.18%, drop=9.50%
  [Layer  1 -> 2bit] acc=89.53%, drop=1.15%
  [Layer  2 -> 2bit] acc=89.82%, drop=0.86%
  [Layer  3 -> 2bit] acc=89.97%, drop=0.71%
  [Layer  4 -> 2bit] acc=89.68%, drop=1.00%
  [Layer  5 -> 2bit] acc=89.31%, drop=1.37%
  [Layer  6 -> 2bit] acc=90.14%, drop=0.54%
  [Layer  7 -> 2bit] acc=90.04%, drop=0.64%
  [Layer  8 -> 2bit] acc=90.58%, drop=0.10%
  [Layer  9 -> 2bit] acc=90.74%, drop=-0.06%
  [Layer 10 -> 2bit] acc=90.52%, drop=0.16%
  [Layer 11 -> 2bit] acc=90.56%, drop=0.12%
  [Layer 12 -> 2bit] acc=90.56%, drop=0.12%

=> Per-layer sensitivity sorted by accuracy drop:
  Layer  9: acc=90.74%, drop=-0.06%
  Layer  8: acc=90.58%, drop=0.10%
  Layer 11: acc=90.56%, drop=0.12%
  Layer 12: acc=90.56%, drop=0.12%
  Layer 10: acc=90.52%, drop=0.16%
  Layer  6: acc=90.14%, drop=0.54%
  Layer  7: acc=90.04%, drop=0.64%
  Layer  3: acc=89.97%, drop=0.71%
  Layer  2: acc=89.82%, drop=0.86%
  Layer  4: acc=89.68%, drop=1.00%
  Layer  1: acc=89.53%, drop=1.15%
  Layer  5: acc=89.31%, drop=1.37%
  Layer  0: acc=81.18%, drop=9.50%
```

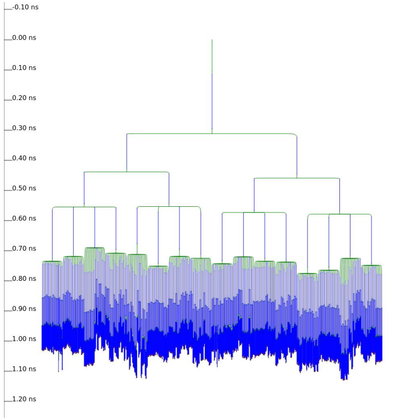


```
=> Start greedy search from all-4bit baseline: acc=90.68%, epsilon=1.00%

  ACCEPT layer  9 -> 2bit | acc=90.74%, drop=-0.06%
  ACCEPT layer  8 -> 2bit | acc=90.62%, drop=0.06%
  ACCEPT layer 11 -> 2bit | acc=90.30%, drop=0.38%
  ACCEPT layer 12 -> 2bit | acc=90.22%, drop=0.46%
  ACCEPT layer 10 -> 2bit | acc=90.35%, drop=0.33%
  REJECT layer  6 -> 2bit | acc=89.32%, drop=1.36%  (> 1.00%)
  REJECT layer  7 -> 2bit | acc=89.41%, drop=1.27%  (> 1.00%)
  REJECT layer  3 -> 2bit | acc=89.65%, drop=1.03%  (> 1.00%)
  REJECT layer  2 -> 2bit | acc=89.21%, drop=1.47%  (> 1.00%)
  REJECT layer  4 -> 2bit | acc=89.00%, drop=1.68%  (> 1.00%)
  REJECT layer  1 -> 2bit | acc=89.06%, drop=1.62%  (> 1.00%)
  REJECT layer  5 -> 2bit | acc=88.88%, drop=1.80%  (> 1.00%)
  REJECT layer  0 -> 2bit | acc=80.42%, drop=10.26%  (> 1.00%)

==================== Final Mixed-Precision Config ====================
Per-layer activation bits (per QuantConv2d):
[4, 4, 4, 4, 4, 4, 4, 4, 2, 2, 2, 2, 2]

Total conv layers = 13
2bit layers       = 5 (38.5%)

Final mixed-precision accuracy     : 90.35%
Accuracy drop vs all-4bit baseline : 0.33%
```

Alpha 4: Synthesis and PnR flow using OpenROAD







- Weight stationary systolic array core implemented with 130nm open source technology node (IHP-SG13).
- SRAM Verilog rewritten to realize SRAM macros in tech node.
- Core utilization: 40%
- Placement density: 0.5
- Total chip area: **6.41M u^2**
- Total power: **0.14 W**
- Max clock frequency: **144.3MHz** (max path delay reg to reg is 6.93ns)
- Not fully optimized. Still many issues caused by SRAM blocks and chip size can be reduced by increasing cell density.