A 104.8TOPS/W One-Shot Time-Based Neuromorphic Chip Employing Dynamic Threshold Error Correction in 65nm

Luke Everson, Muqing Liu, Nakul Pande, Chris Kim

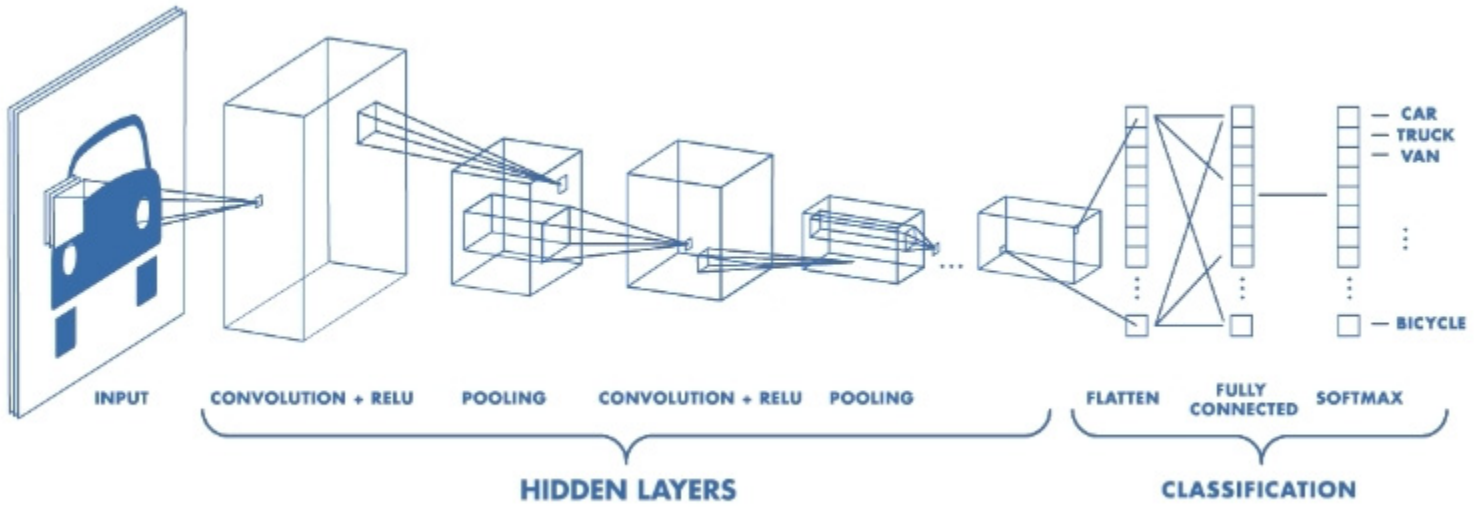
Department of Electrical & Computer Engineering

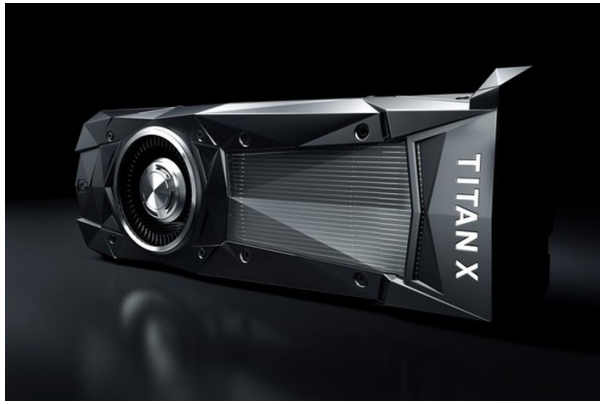
Univ. of Minnesota, Minnesota, USA

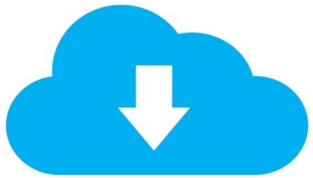
Outline

- Motivation
- Time Based Neural Network
- DTEC- Dynamic Threshold Error Correction
- Measurement results
- Conclusion

Motivation



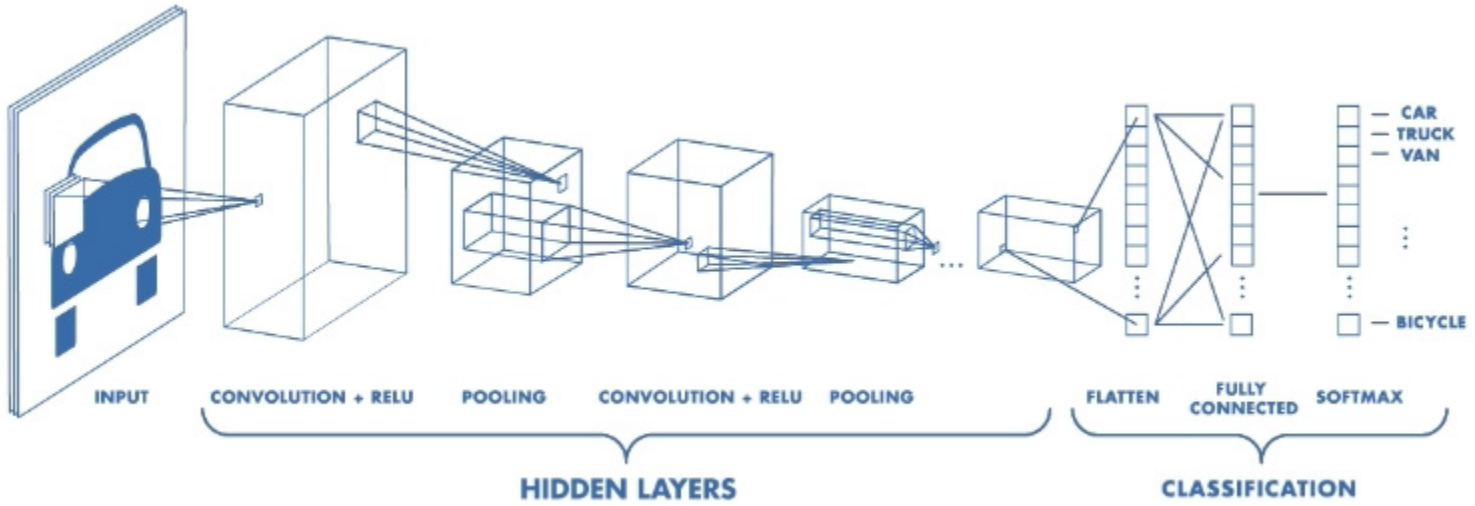


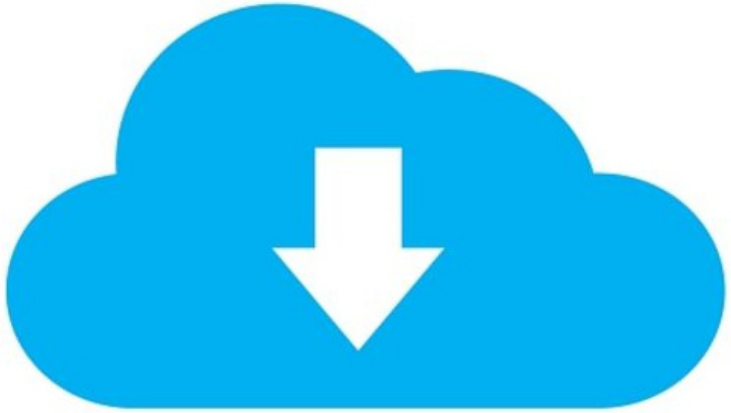


Send Data

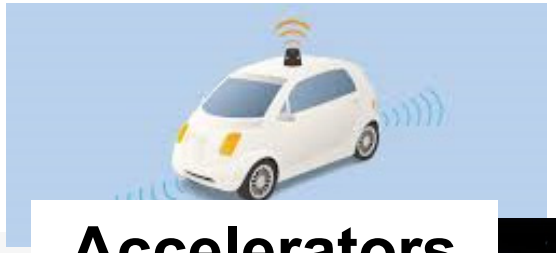
Receive Result

Motivation



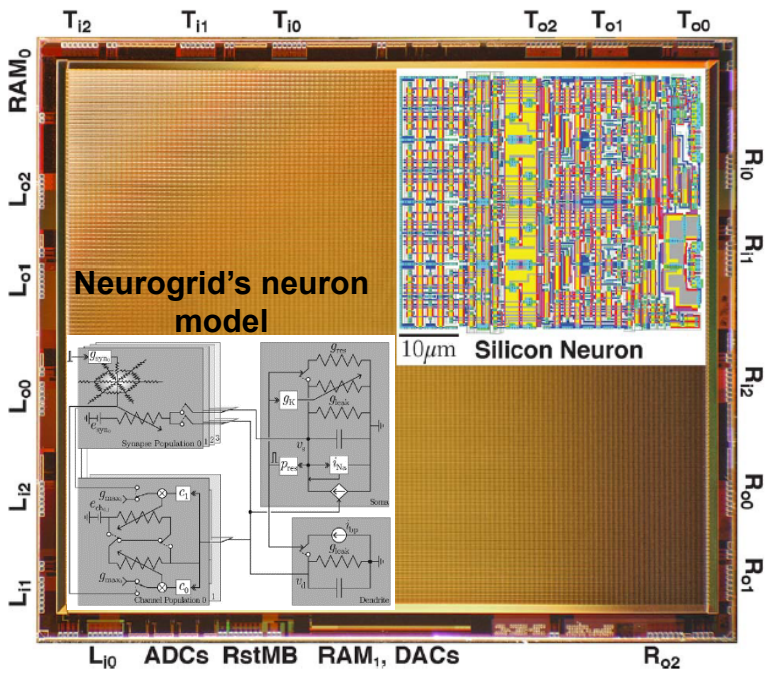


Send Result



Accelerators

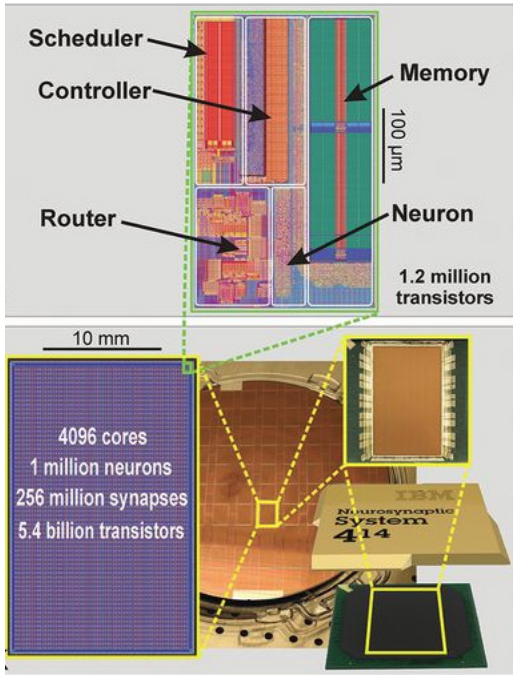
Motivation



Neurogrid (Analog neurons)[1]

+ Low area and power via subthreshold operation

- Sensitive to noise and PVT

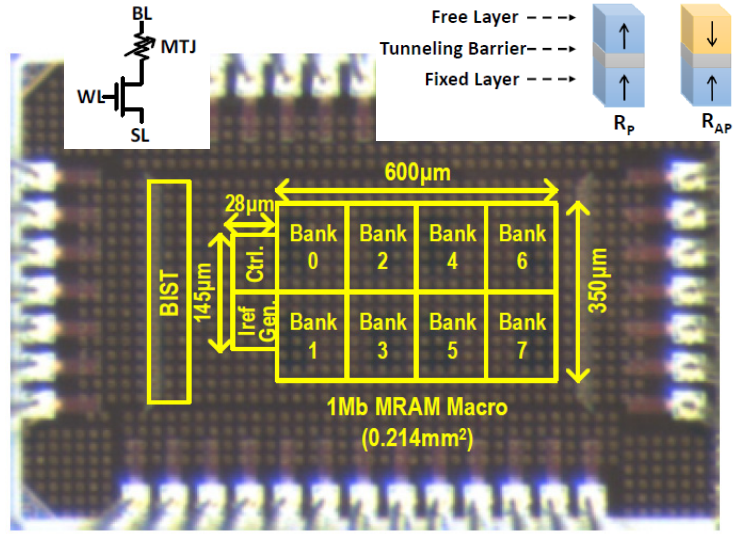


IBM TrueNorth (Digital neurons)[2]

+ Robust to PVT

+ Technology scaling

- Large area overhead



STT-MRAM(Emerging neurons)[3]

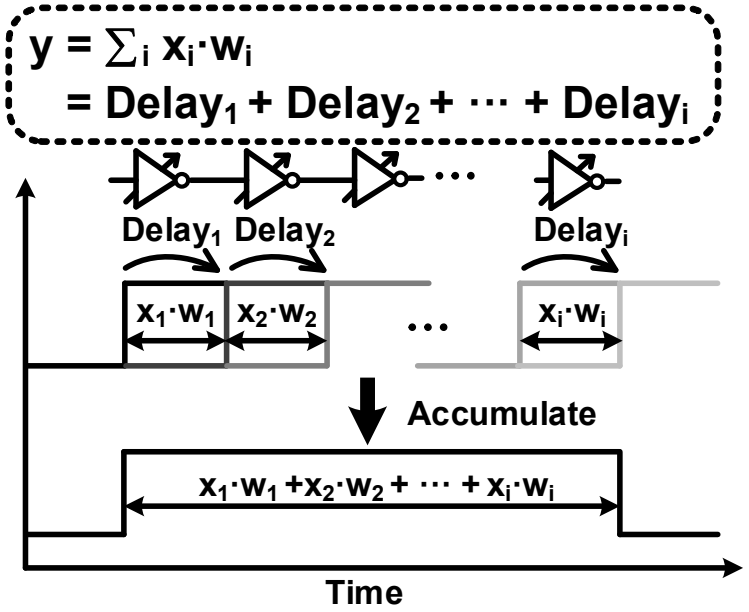
+ Compact, low write energy, scalable

- Beginning early production

[1]B.V. Benjamin, IEEE Proc., 2014. [2]P.A. Merolla, Science, 2014. [3]Q. Dong, ISSCC, 2018.

Time-Based Neuron

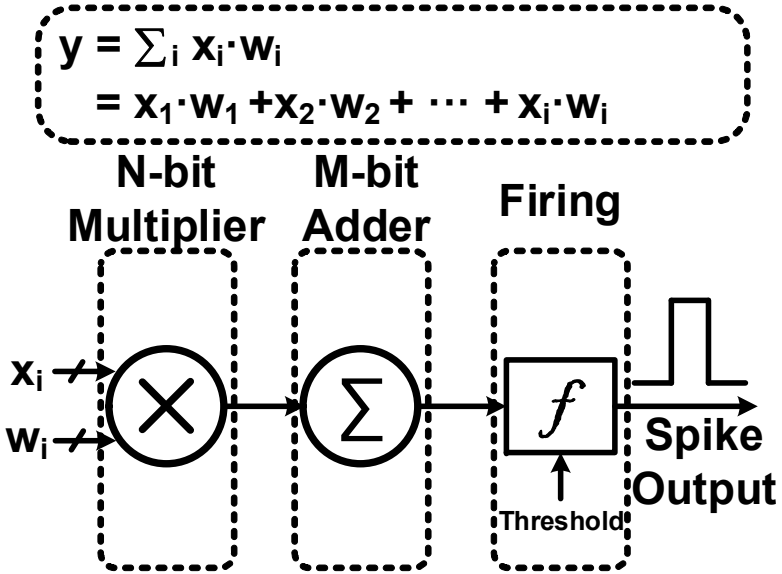
Time-based Neuron



Advantages of time-based circuits:

- Compact area
- Low power consumption
- MAC is intrinsic to structure
- High precision tunability

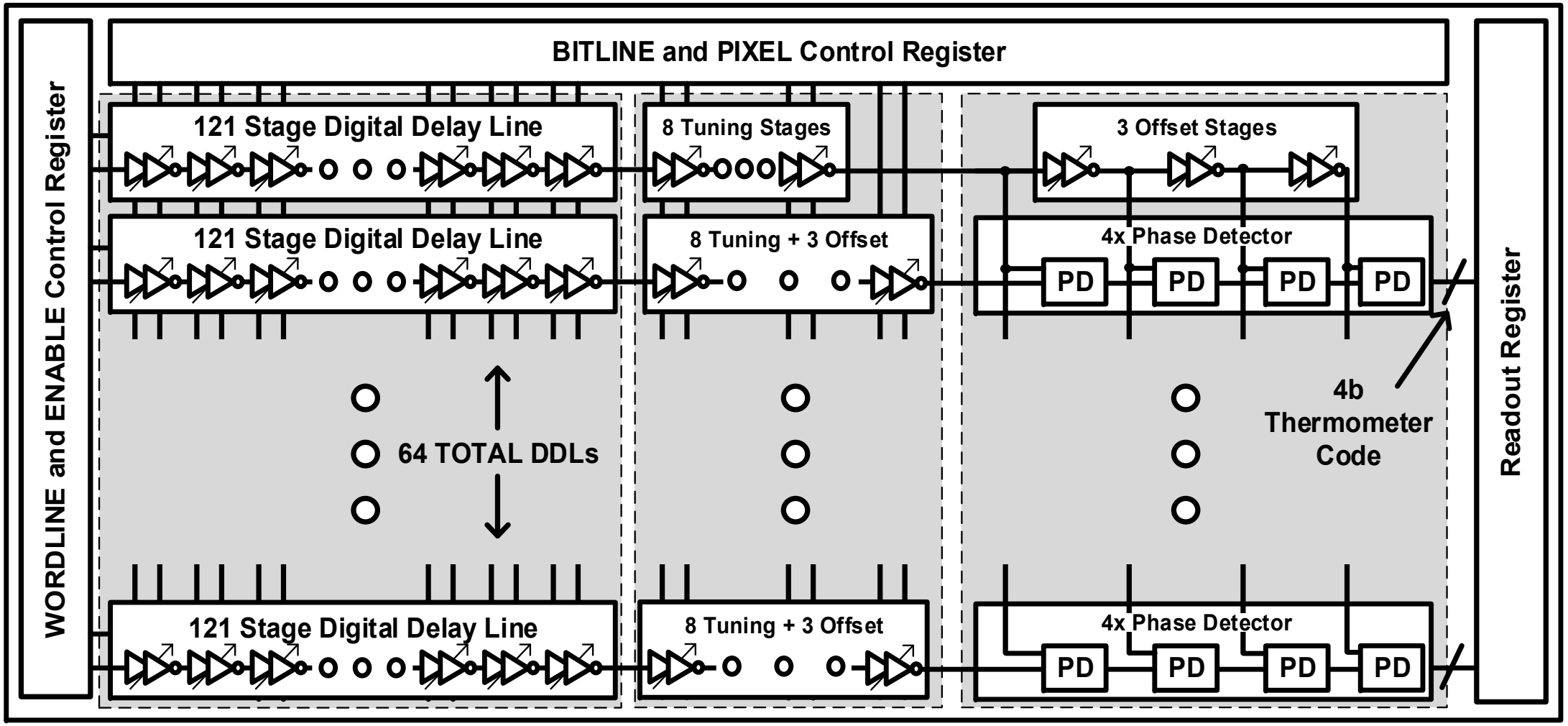
Digital Neuron



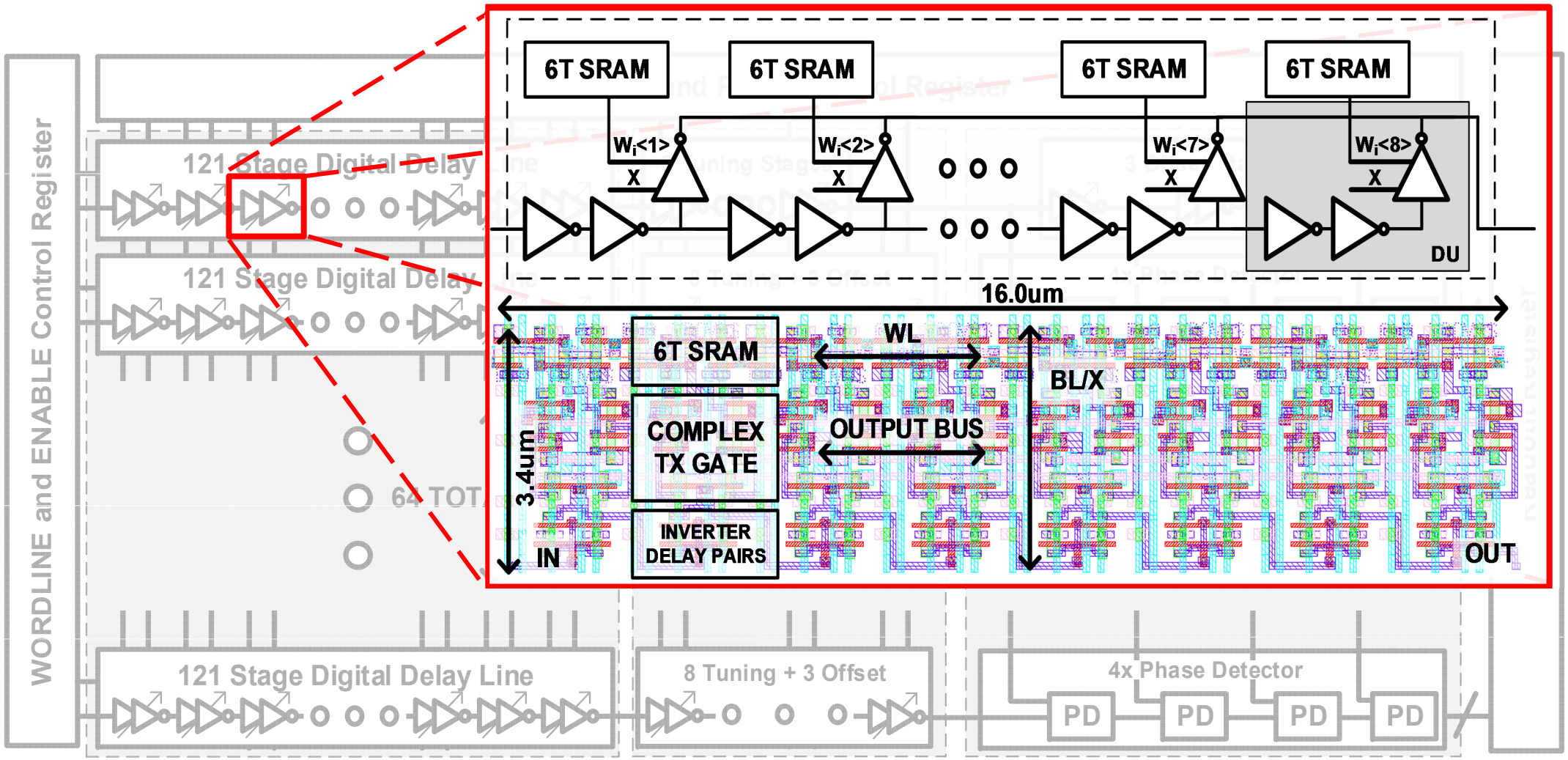
Advantages of digital arithmetic:

- Binary Representation
- Less "buy-in" required
- Existing IP for rapid SoC development
- No calibration

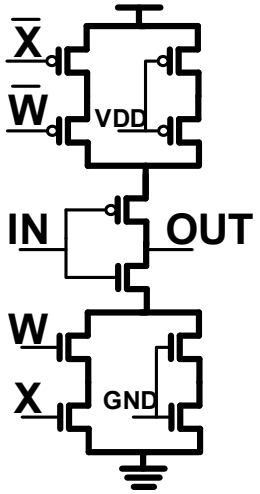
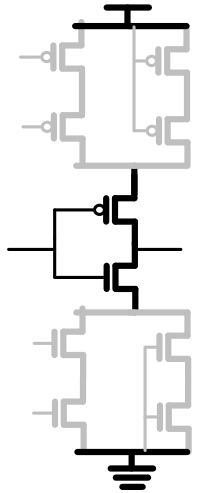
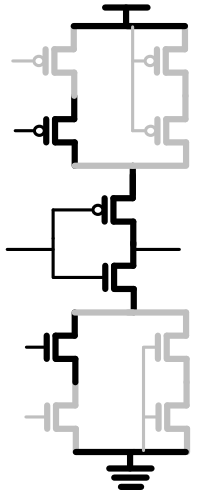
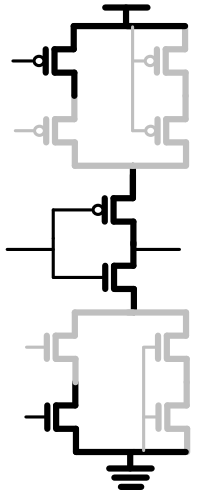
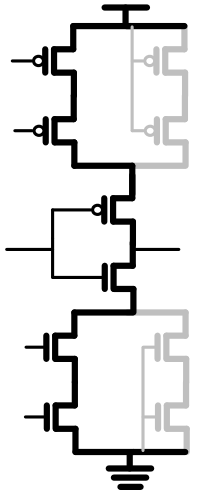
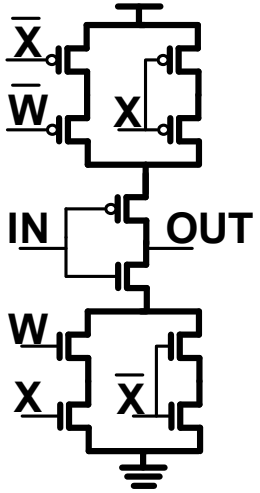
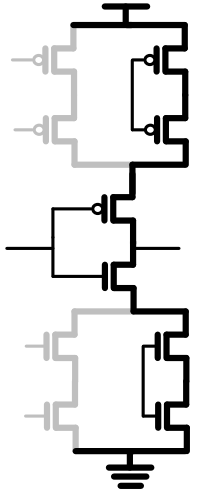
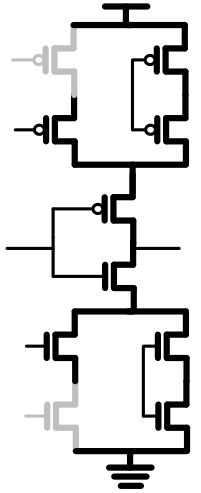
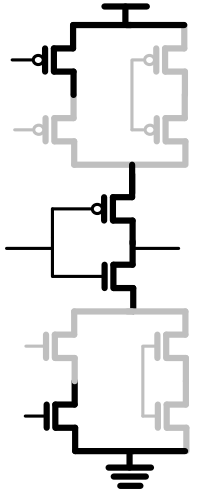
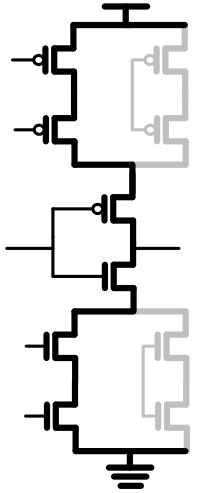
Top Level Schematic



Pixel Unit Detail

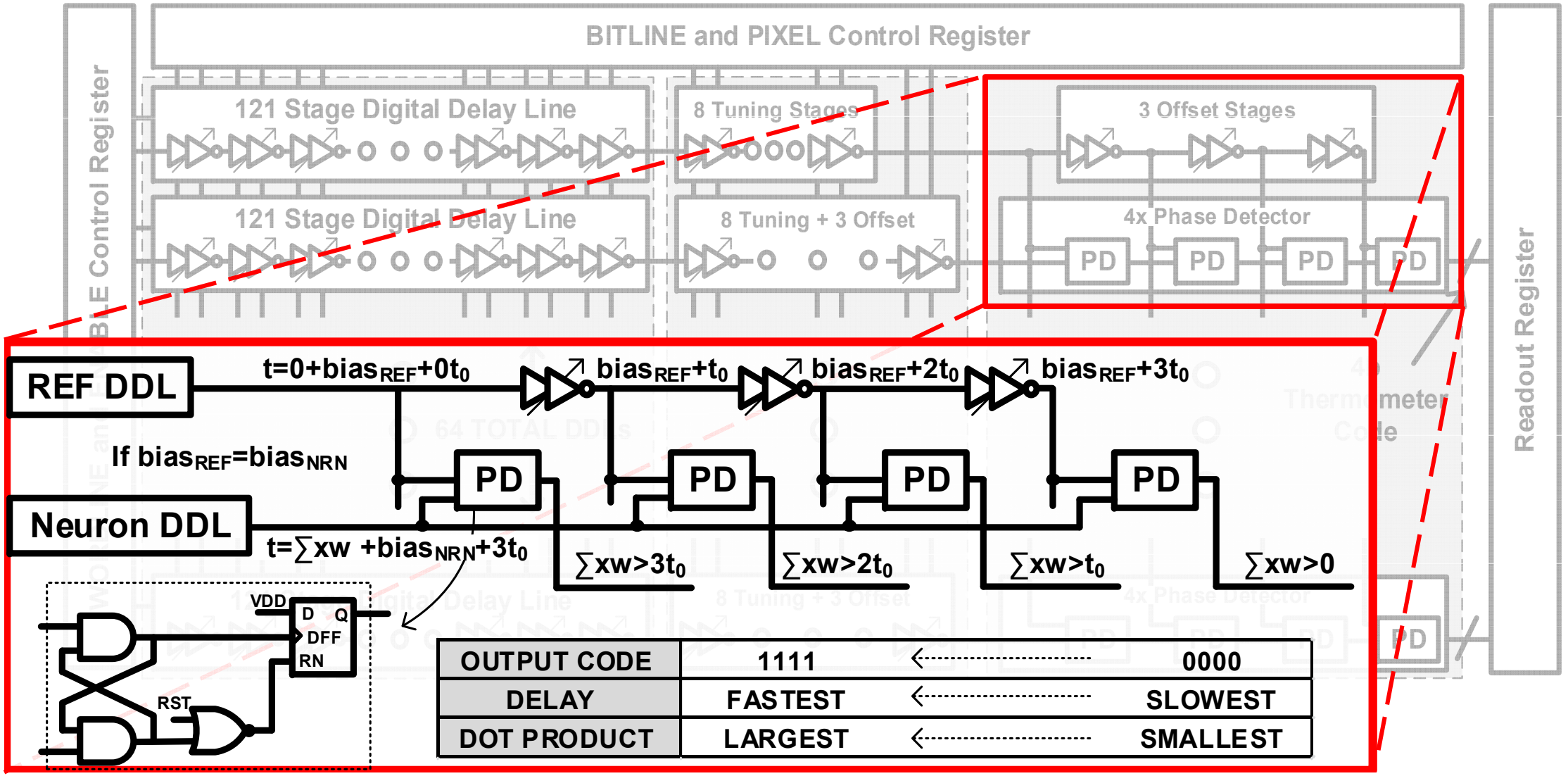


Complex Tristate

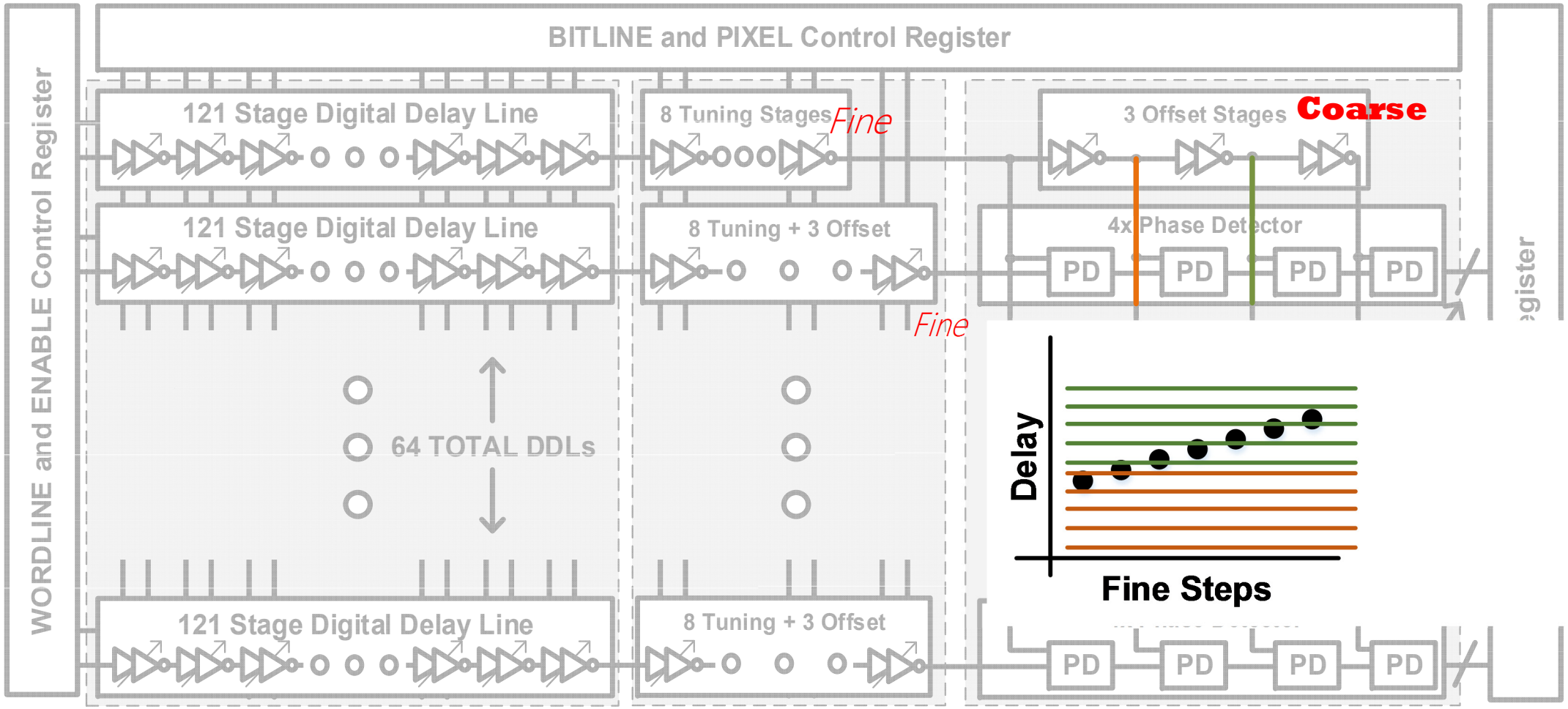
	Circuit	[X=0, W=0]	[X=0, W=1]	[X=1, W=0]	[X=1, W=1]
DUs [1-4,6-8]					
DU 5					

Weight	Delay
-3	8
-2	7
-1	6
0	5
1	4
2	3
3	2
4	1

Phase Detector Detail

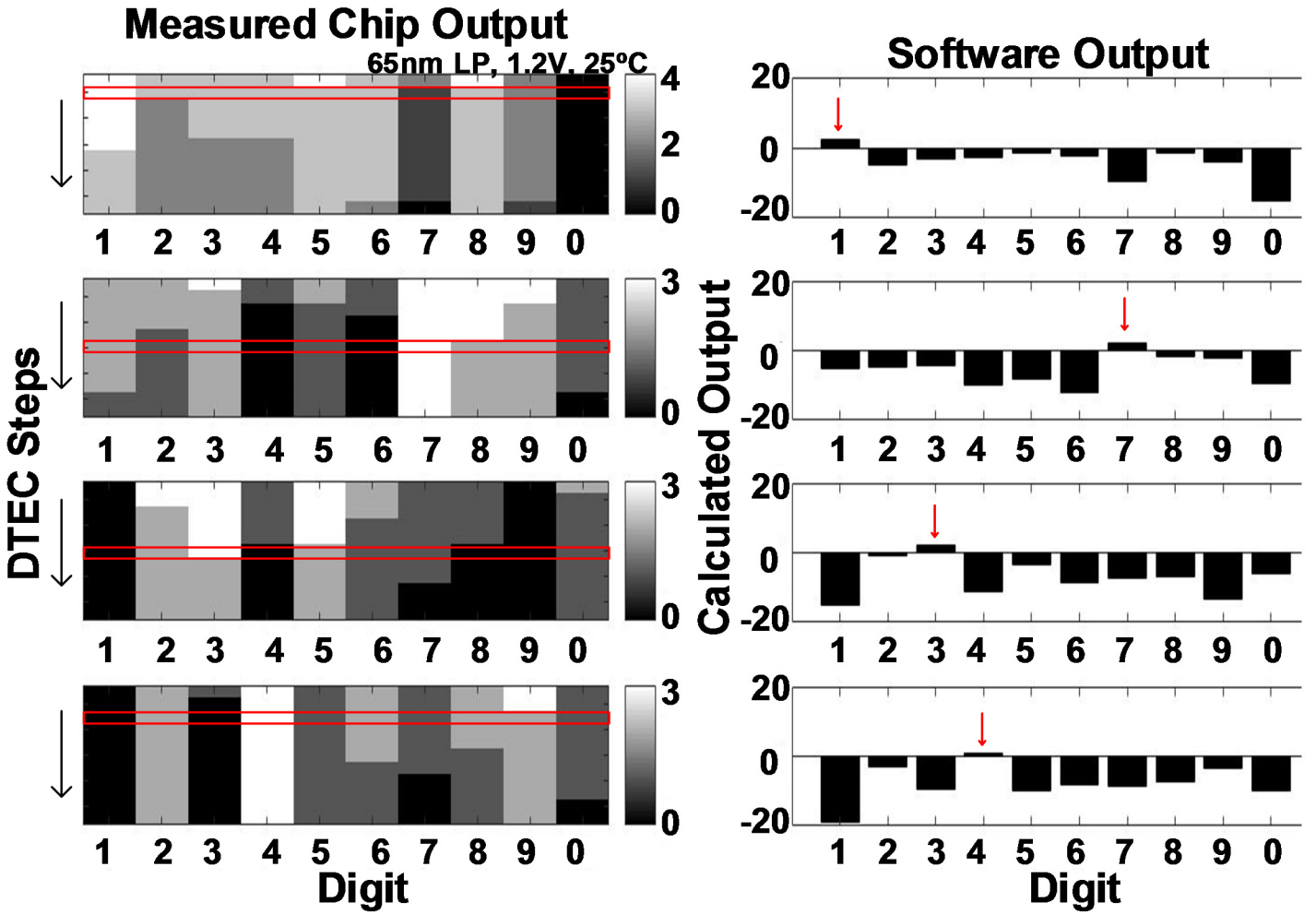


DTEC- Dynamic Threshold Error Correction

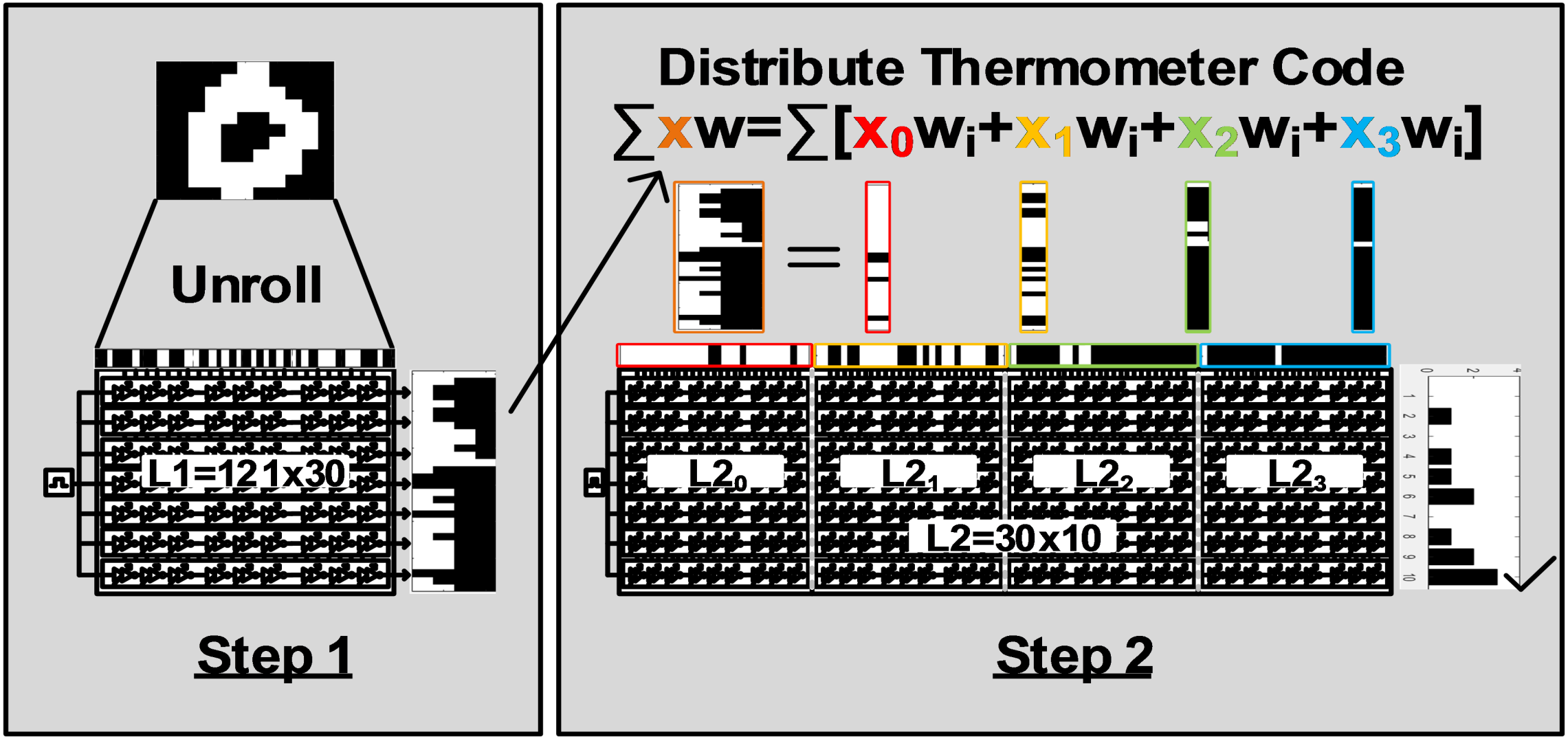


DTEC- Dynamic Threshold Error Correction

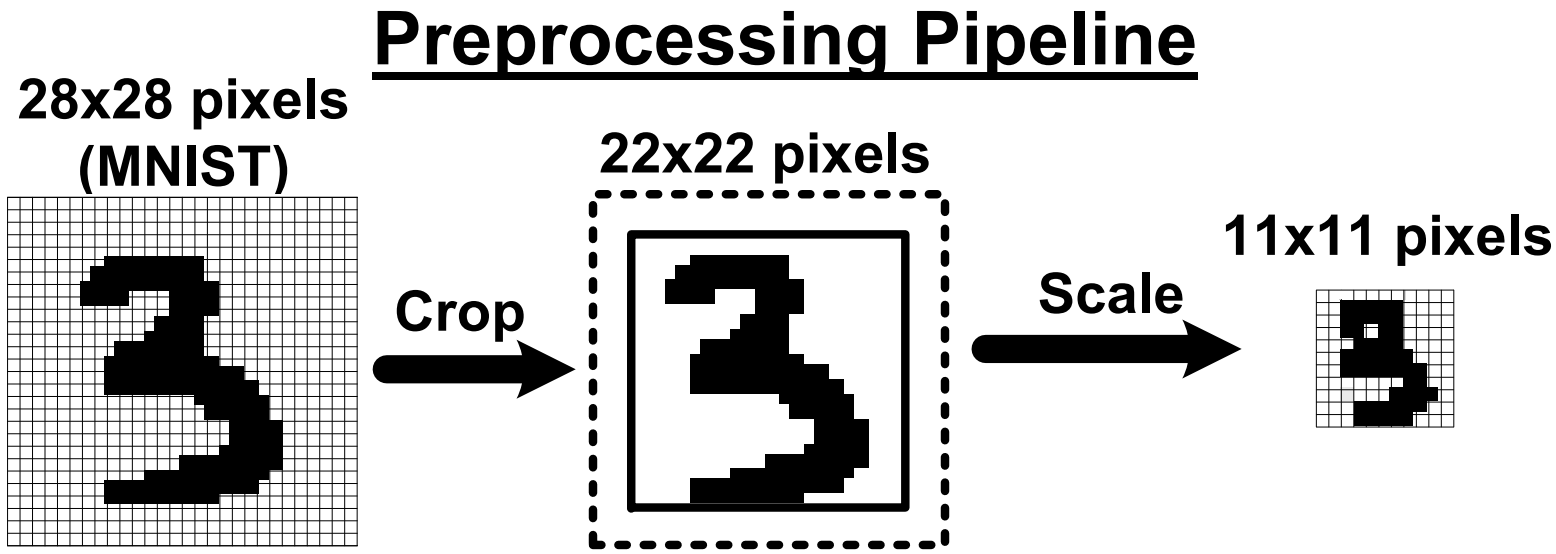
- DTEC applied to MNIST Application
  - Single Layer
  - 3b Weights
  - 11x11 image
- Trained with Tensorflow
- Coarse – 69.8% Accuracy
- 26% Ambiguous
- 1st Fine DTEC – 46% recovered
- 2nd Fine DTEC – 37% recovered
- Total Accuracy- 82%
- DTEC Overhead – 41%/image for 89% error recovery



Multi-Layer Dataflow

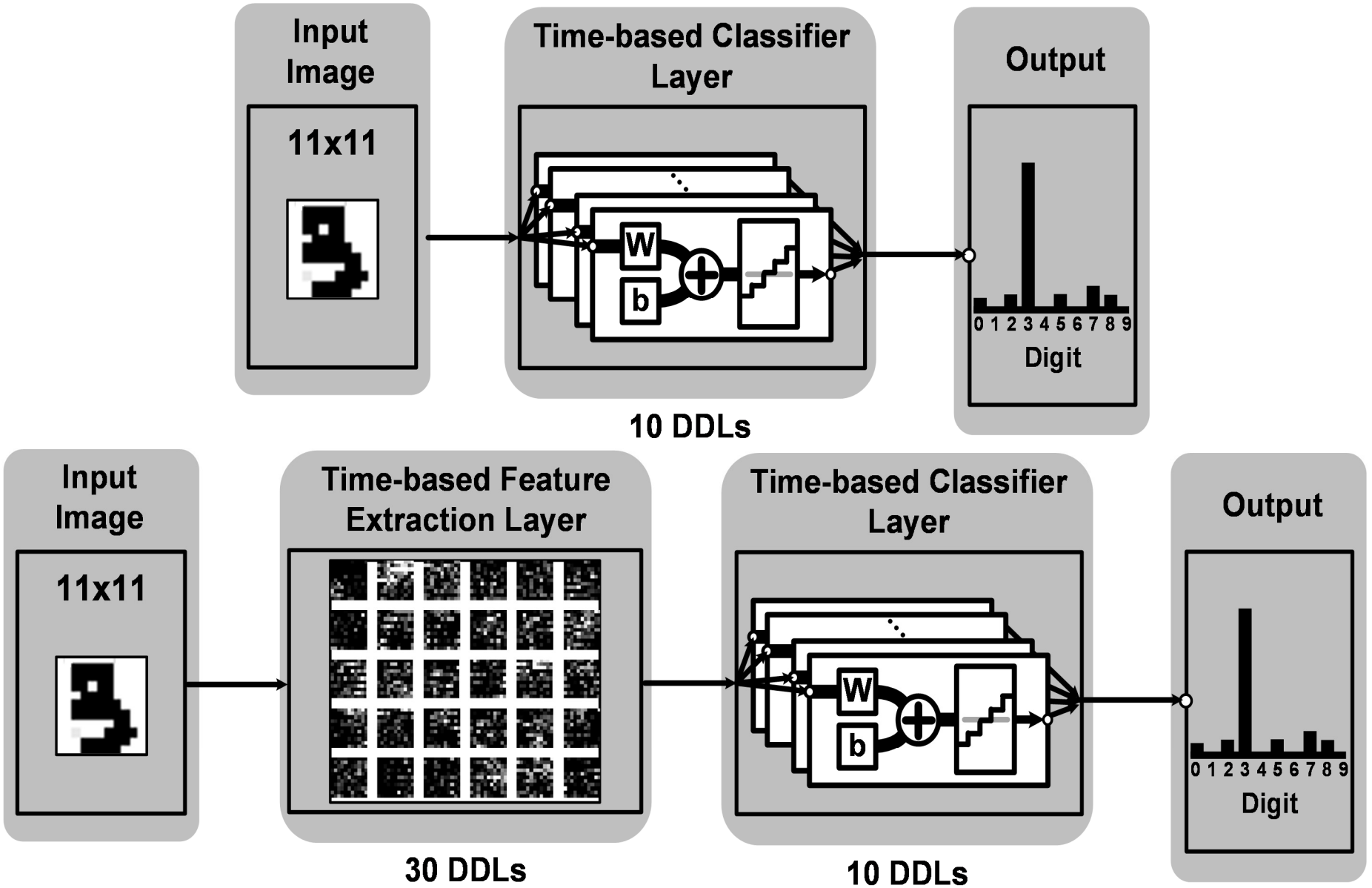


Handwritten Digit Recognition

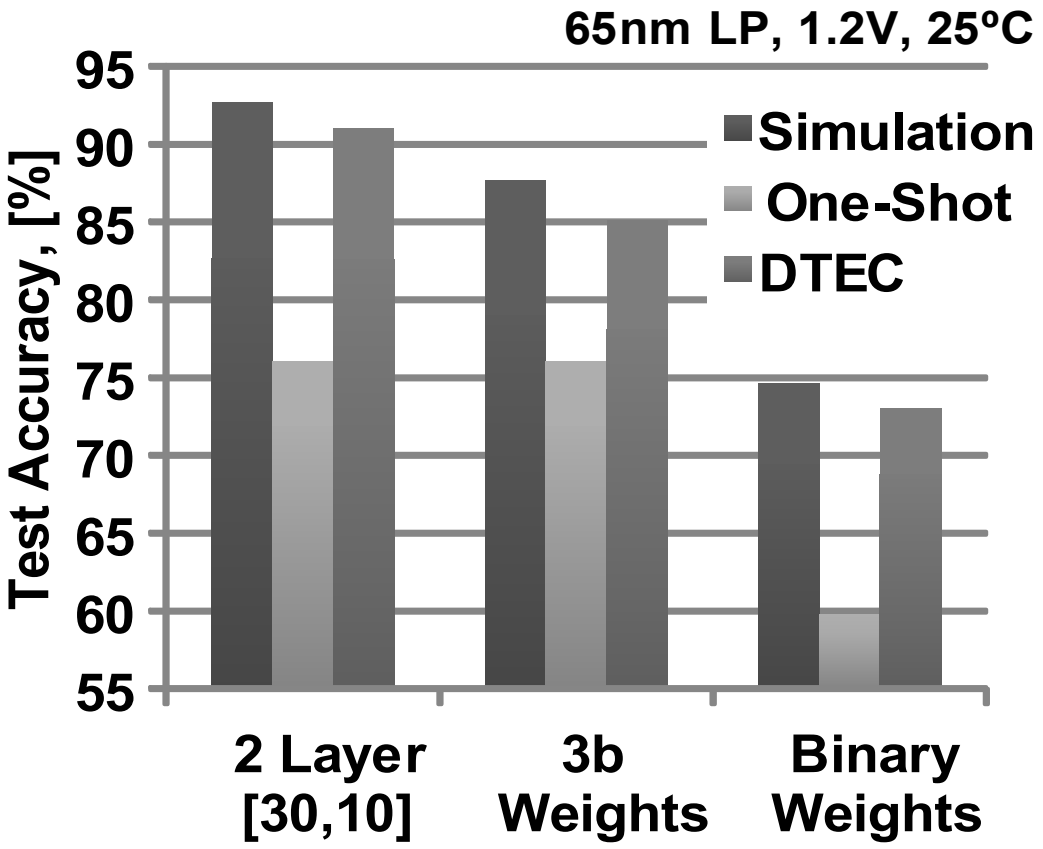


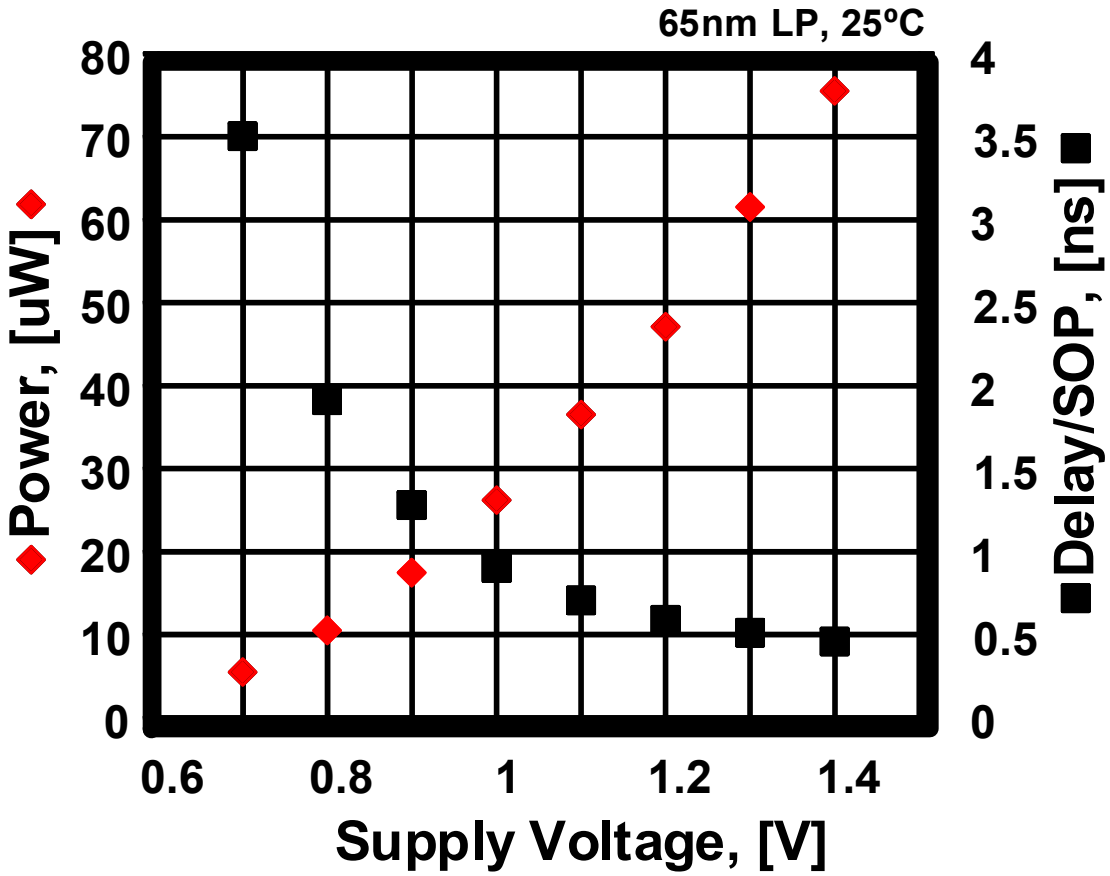
- Application – Handwritten digit recognition
- Training Network – Single Layer & MLP
- Learning Method – Supervised Learning
- Input database - MNIST

Classifier Networks

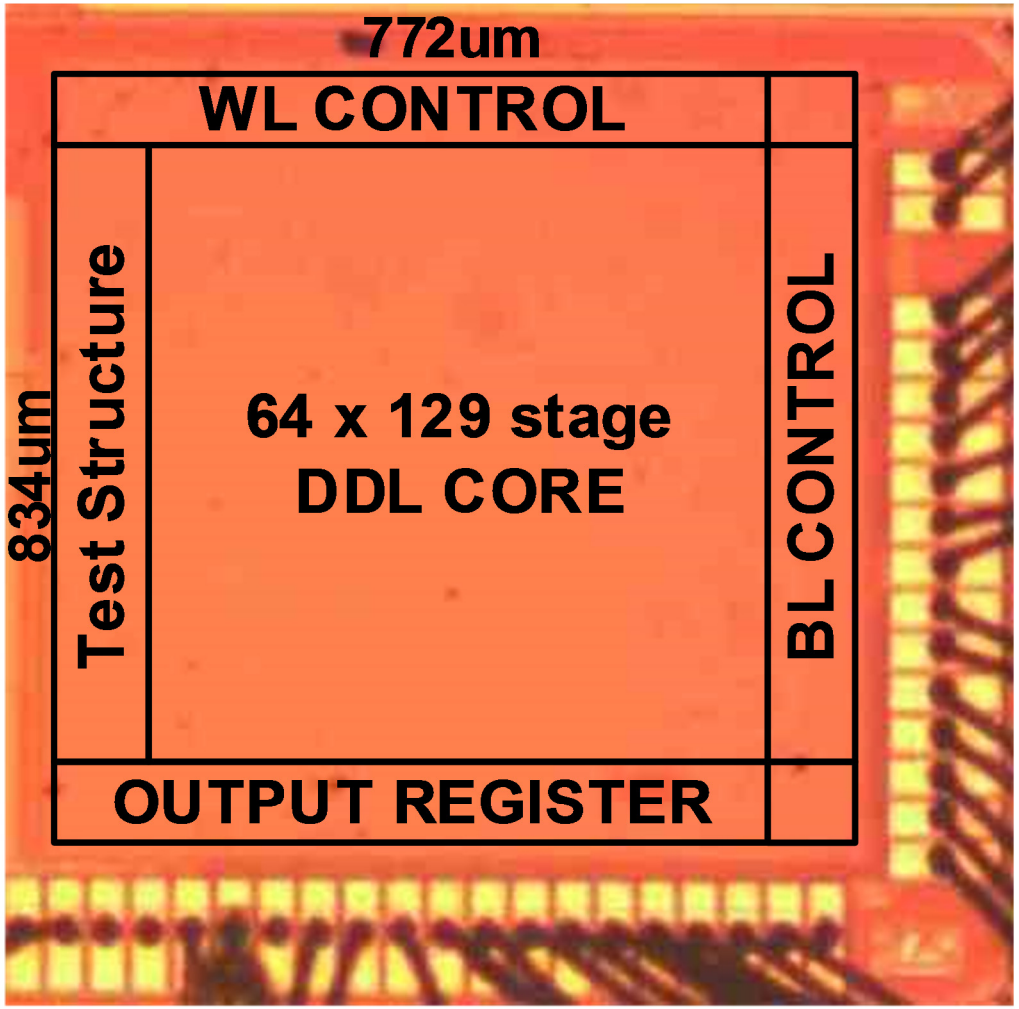


Measurement Results





Die Photo



Process	65nm LP CMOS
Core Area	0.644mm^2
VDD	0.7-1.4V
# of Neurons	64
# of Synapses	8256 (8K)
Throughput	1.7Gpixels/s/DDL
Power	47.1uW/DLL
Energy/SOP	27.6fJ

Comparison Table

	This Work		A-SSCC'16 [1]	CICC'17 [2]	ISSCC'17 [3]	ISSCC'17 [4]	ISSCC'16 [5]	ISSCC'16[6]	Science'14[7]
Chip Architecture	Time-Based		Time-Based	Time-Based	Digital	Digital	Digital	Sw. Cap	Digital
Algorithm Target	FCDNN & CNN		FCDNN & CNN	FCDNN & CNN	FCDNN & CNN	FCDNN & FFT	CNN	CNN & SGD	FCDNN & CNN
Technology [nm]	65		65	65	28 FDSOI	40	65	40	28
Chip Area [mm^2]	0.644		3.61	0.24	1.87	7.1	12.25	0.012	430
Precision* [b]	[B,T,2,3]		B	3	[4-16]	[6-32]	16	3	[B,T]
On-Chip SRAM [kB]	8.06		20	3	144	270	181.5	[-]	256MB
VDD [V]	1.2 (Nom.)	0.7 (E_{MAx})	1	1.2	0.6	0.65	0.82	1	0.85
Frequency [MHz]	1700	285	23041	792	200	19.3	250	1000	0.001
Energy Efficiency** [TSop/s/W]	36.2	52.4	48.2	2.47	5.0	0.19	.18	3.86	0.04
Hardware Efficiency [GE/PE][1]	38.4		76.5	33.2	7456	18269	50637	288	6.5

*B=Binary, T=Ternary **Synaptic Op=MAC

Conclusions

- Time-Based Neuromorphic Core in 65nm LP CMOS
  - 64 DDLs with 129 stages, 1 shared reference
- One-shot evaluation drives high energy efficiency
- Introduced DTEC to recover ambiguous predictions
- Evaluated on MNIST dataset and achieves ~1% difference in software accuracy
- 104.8TOp/S/W @ 0.7V with 3b = 19.1fJ/MAC