14.4 All-Digital Time-Domain CNN Engine Using Bidirectional Memory Delay Lines for Energy-Efficient Edge Computing

Aseem Sayal, Shirin Fathima, S. S. Teja Nibhanupudi, Jaydeep P. Kulkarni

University of Texas, Austin, TX

Convolutional Neural Networks (CNN) provide superior classification accuracy in a variety of machine learning applications, such as image/speech/sensor data processing. However, CNNs require intensive compute and memory resources making it challenging to employ in energy-constrained edge-computing devices. Specifically, Multiply-and-Accumulate (MAC) operations consume a significant portion of the total CNN energy [1].

Various analog compute techniques using charge manipulation schemes and A/D converters, as well as frequency-modulation-based approaches have been proposed to realize efficient MAC computations in a CNN accelerator (Fig. 14.4.6) [1-2, 5-6]. However, finite voltage headroom is required in analog MAC designs, whereas accurate frequency control is necessary for prior frequency-domain MAC approaches. This limits the voltage scalability of analog approaches and performance scalability of prior frequency-domain techniques degrading the MAC energy efficiency – critical for CNNs used in edge-compute devices. In this paper, we demonstrate an energy-efficient CNN engine implemented in a 40nm CMOS (Figs. 14.4.1, 14.4.7) featuring: 1) Bi-directional Memory Delay Lines (MDL) performing time-domain MAC operations; 2) multi-precision filter weight support (signed/unsigned 1-8b); 3) 16 filters each supporting 2×2 sub-sampling (max. pooling) and averaging; 4) all-digital, technology scalable design without requiring any capacitors, A/D converters, and/or frequency generators/modulators; and 5) near-threshold voltage operation supporting 16× speed-up with 4 input encoding modes.

The CNN data flow (Fig. 14.4.1) consists of: 1) training input data using TensorFlow/Keras software framework; 2) feeding trained filter weights and test input data to the test-chip using LabVIEW-PXI data acquisition instruments; 3) performing on-chip MAC, averaging, and pooling operations for the test input data; and 4) fully connected layers and soft-max computation in software (TensorFlow). The time-domain MAC computations are realized using the proposed MDL which accumulates the dot product of a weight bit and the time encoded input pulse-width (Fig. 14.4.2). It is derived from the concept of a 'time-register' used in high precision time-to-digital converters, which can perform time addition and time storage using a string of delay cells controlled by an enable (EN) signal [7]. Each MDL unit comprises two cross-coupled inverter pairs, S1-S4 switches, and a reset control (RST). The time-encoded dot product of input (X_i) and 1b filter weight (w_i) acts as an EN pulse and controls the MDL operating mode. During the accumulation phase, EN=1 and the MDL acts either as a forward delay line (for +ve dot product, S1 and S2 are ON) or a backward delay line (for -ve dot product, S2, S3 and S4 are ON), enabling bi-directional data flow emulating signed dot products. When EN goes low, the MDL acts as a memory storage line and retains the MDL state vector using cross-coupled inverters (S1, S3 and S4 are ON). The metastability risk during an EN falling transition is resolved by the next incoming EN pulse, as the MDL is transformed into a chain of cascaded delay cells. When the MDL state vector string progresses towards the either end of the MDL (node A or node E), an up-down counter is triggered, which translates time-domain dot product accumulation information into digital bits acting as a time-to-digital converter. If the accumulated dot product pulse-width exceeds the full-scale MDL delay, an overflow condition is detected, and the propagating edge is inverted (using S5-S6) and applied at the beginning node A of MDL (or trailing end node E). Thus, a finite length MDL can be used to perform long-duration time-domain accumulation using an up-down counter. The calibration unit consists of additional delay cells which can be added to the MDL to mitigate delay mismatch in the presence of process variations.

The bi-directional MDL forms the core of the CNN engine implementing 16 filters. Each filter consists of 4 bi-directional MDLs, a weight shift register, a shared pulse generator/selector, an up/down counter, a bi-directional barrel shifter, and pooling comparators (Fig. 14.4.3). 8b input data (X) is represented in the time domain as a Pulse-Width Modulated (PWM) signal as multiples of input clock period ($2t_0$). A pulse generator/selector circuit is designed to generate 0-255t_0 PWM signals using a two-stage approach [1]. As 4 MSBs (*X[7-4]*) correspond to a maximum of value of 240 (out of the 255 full-scale value of an 8b input), an MSB_EN signal is asserted for 240t_0 duration in the first stage generating T0-T15 output pulses in increments of 16t_0 duration. In the second step, the MSB_EN signal is de-asserted and 4 LSBs (*X[3-0]*) PWM signals are generated as outputs T0-T15. Four precision modes are implemented to support 1-to-16× speed-up in the input PWMs by quantizing 4 LSBs. As shown in the pulse generator timing diagram (Fig. 14.4.3), varying duration T0-T15 pulses are generated based on the precision mode and concatenated with *X[7-4]* MSB pulses generated in the first step. The LSB pulse-width quantization steps are chosen to limit the quantization error to ±0.5*speed-up ratio. As the pulse generator operates continuously in every MAC_CLK period, pulse gating AND logic is implemented to ensure that only a single PWM input pulse is applied to the MDL in each dot product computation. Next, the single PWM input is multiplied with a 1b filter weight (w_i) stored in a 25b shift register (for a 5×5 filter size) producing the $X_1 \bullet w_i$ dot product. The bi-directional MDL then performs signed accumulation, time-to-digital conversion (20b-up/down counter) and averaging/scaling (20b-bi-directional barrel shifter performing up to 7b shifts). For multi-bit filter weights, multiple instances of MDLs with each weight shift register initialized with one bit of the weight vector can be used. A sub-sampling operation using max pooling across a 2×2 window is implemented to reduce the convolution layer output data size by 75%. This is achieved by 4 concurrent MDL operations and feeding the MDL counter outputs to 8b MAX comparators. The pooled output from each filter is stored off-chip and reused as the input to the next convolution layer.

Figure 14.4.4 shows oscilloscope-captured waveforms from the 40nm CMOS test-chip (Fig. 14.4.7) confirming MDL functionality with delay phase and storage phase for different MDL lengths. The pulse generator and the pulse-gating logic functionality is verified with the correct toggling of MSB_EN, T15, T8, T4 outputs and the pulse-gating control signals. 1-to-16× speed-up in the PWM input representation is validated with multiple precision modes for a test-case input of 214. The measured classification accuracy (on 100 images) for LeNet-5 using the MINIST dataset is ~2% lower relative to software counterpart (Fig. 14.4.5). 16× speed-up mode resulted in lower accuracy because of input quantization and increased sensitivity of MDL residue. Simulation results using the proposed MDL-based CNN for AlexNet with 2-class ImageNet dataset (cats *vs.* dogs) with signed 8b weights, shows 13% lower classification accuracy compared to software (16b floating-point weights). The CNN engine is operable down to 375mV with more than 90% MNIST classification accuracy. For the LeNet-5 case, both C1- and C3-layer throughput increases with the higher speed-up mode and with the increasing supply voltage achieving a peak throughput of 0.38 (0.128) GOPS for the C3 (C1) layer at 585mV. The energy efficiency peaks with increasing supply voltage scaling and reaches a maximum of 13.46 (4.61) TOPS/W for the C3 (C1) layer at 496mV. Figure 14.4.6 tabulates LeNet-5 parameters, test-chip characterization results (at optimal voltage of 537mV), and compares with earlier approaches [1-6].

Acknowledgements:
Authors thank Vignesh Radhakrishnan and Jacob Rohan for helping with the test-chip measurements. Authors also like to thank TSMC University shuttle program for the test-chip fabrication support.

References:
[1] A. Biswas, et al., "Conv-RAM: An Energy-Efficient SRAM With Embedded Convolution Computation For Low-Power CNN-Based Machine Learning Applications," *ISSCC*, pp. 488-490, 2018.
[2] S. K. Gonugondla, et al., "A 42pJ/decision 3.12TOPS/W Robust In-Memory Machine Learning Classifier With On-Chip Training," *ISSCC*, pp. 490-492, 2018
[3] J. Sim, et al., "A 1.42TOPS/W Deep Convolutional Neural Network Recognition Processor for Intelligent IoE Systems," *ISSCC*, pp. 264-265, 2016.
[4] B. Moons, et al., "A 0.3–2.6 TOPS/W Precision-Scalable Processor for Real-Time Large-Scale ConvNets," *IEEE Symp. VLSI Circuits*, pp. 178-179, 2016.
[5] M. Liu, et al., "A scalable time-based integrate-and-fire neuromorphic core with brain-inspired leak and local lateral inhibition capabilities," *IEEE CICC*, pp. 1-4, 2017.
[6] A. Amravati, et al., "A 55nm Time-Domain Mixed-Signal Neuromorphic Accelerator With Stochastic Synapses and Embedded Reinforcement Learning For Autonomous Micro-Robots," *ISSCC*, pp. 124-126, 2018.
[7] K. Kim, et al., "A 9 bit, 1.12ps Resolution 2.5b/Stage Pipelined Time-to-Digital Converter in 65 nm CMOS Using Time-Register," *IEEE JSSC*, vol. 49, no. 4, pp. 1007-1016, 2014.

 978-1-5386-8531-0/19/$31.00 ©2019 IEEE

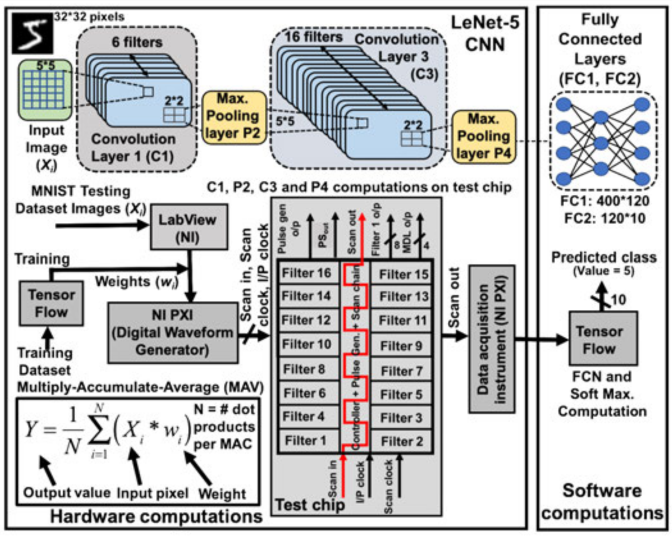


Figure 14.4.1: LeNet-5 CNN engine data flow and its interface with the test-chip which implements MAC, averaging, and pooling operations.

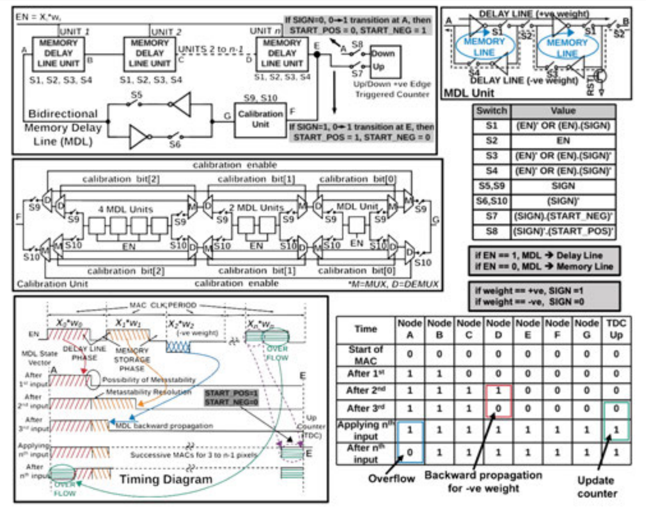
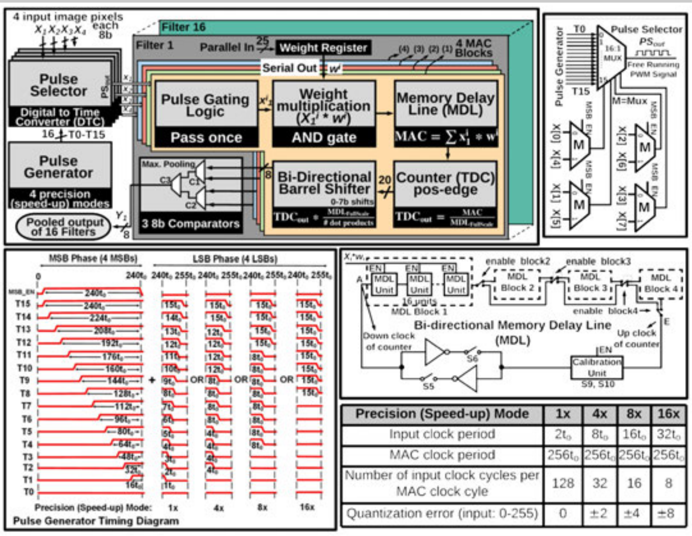


Figure 14.4.2: Proposed bi-directional Memory Delay Line concept (MDL) for time-domain, signed MAC computation: circuit topology, switch-selection logic, calibration unit, and timing diagram.



Precision (Speed-up) Mode	1x	4x	8x	16x
Input clock period	$2t_0$	$8t_0$	$16t_0$	$32t_0$
MAC clock period	$256t_0$	$256t_0$	$256t_0$	$256t_0$
Number of input clock cycles per MAC clock cyle	128	32	16	8
Quantization error (input: 0-255)	0	±2	±4	±8

Figure 14.4.3: Overall architecture of the proposed MD- based CNN engine with input PWM pulse generator supporting 1-to-16× speed-up modes.

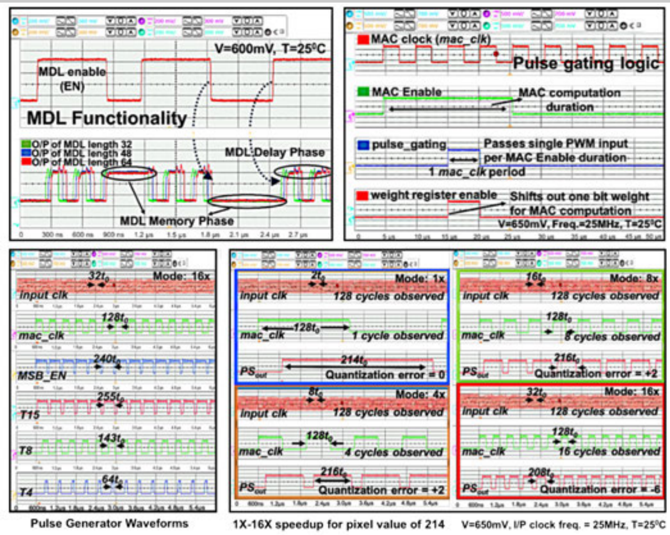


Figure 14.4.4: Experimental demonstration of the MDL functionality, pulse generator/pulse gating logic, and 1-to-16× speed-up with input quantization.

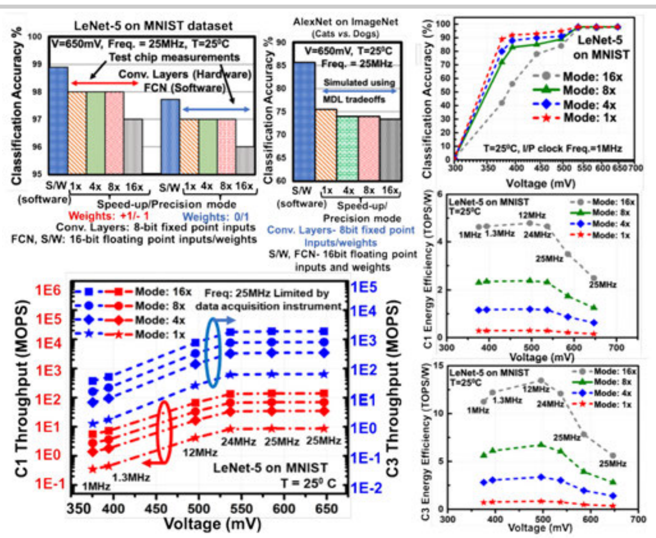


Figure 14.4.5: Measured throughput, energy efficiency, and classification accuracy for 1-to-16× speed-up modes on LeNet-5 using MNIST dataset.

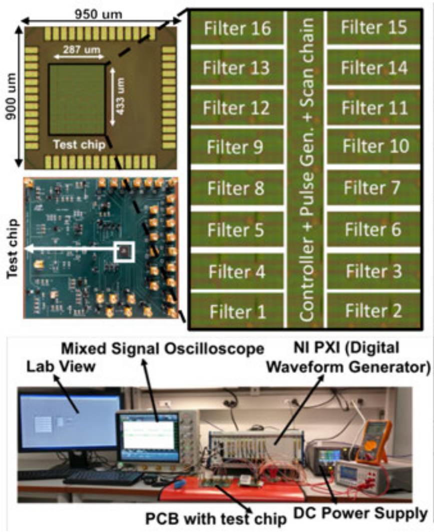
Parameters for LeNet-5	C1+P2	C3+P4
Filter Size	5*5*1*6	5*5*6*16
Input/Filter Size	8bits/ 1bit	8bits/ 1bit
Input Size	32*32*1	14*14*6
Output Size	14*14*6	5*5*16
#Filters	6	16
#Operations/ convolution*	(25*4*6) *2	(150*4* 16)*2
#MAC clock cycles/conv	28	158

LeNet-5 Results/Metrics	Convolution Layer C1 and Pooling Layer P2				Convolution Layer C3 and Pooling Layer P4			
Precision (Speed-up) Mode	1x	4x	8x	16x	1x	4x	8x	16x
Input clock frequency (MHz)	24.0	24.0	24.0	24.0	24.0	24.0	24.0	24.0
MAC clock frequency (MHz)	0.19	0.75	1.50	3.00	0.19	0.75	1.50	3.00
Convolution Cycle Time (us)	149.3	37.33	18.67	9.33	842.67	210.67	105.33	52.67
Operating Voltage (V)	0.537	0.537	0.537	0.537	0.537	0.537	0.537	0.537
Power (uW)	28.67	28.67	28.67	28.67	30.17	30.17	30.17	30.17
Throughput (GOPS)	0.008	0.032	0.064	0.128	0.023	0.091	0.183	0.365
Energy Efficiency (TOPS/W)	0.29	1.16	2.33	4.65	0.76	3.02	6.04	12.08

*Assuming 1 Multiply-Accumulate-Average, and 1 Pooling = 2 operations
**Scalable to multi-bit weights

Reference	Tech. (nm)	Circuit Type	Input/ Weight Size	Chip Size (mm^2)	Pooling	Low Vcc Support	Capaci-tors or ADCs	Classifi-cation Accuracy	Throughput (GOPS)	Power (uW)	Energy Efficiency (TOPS/W)
ISSCC'18 [1]	65	Analog	6bits/1bit	0.067	No	No	Yes	96%	10.70	380.7	28.10
ISSCC'18 [2]	65	Analog	8bits	1.440	No	No	Yes	96%	-	-	3.125
ISSCC'16 [3]	65	Digital	16bits	16.000	Yes	No	No	98.3%	64	4.51E+4	1.42
VLSI'16 [4]	40	Digital	6bits/4bits	2.400	No	Yes	No	98%	102	3.9E+4	2.60
CICC'17 [5]	65	Time	8bits/3bits	0.24	No	Yes	Yes	91%	0.396	2.05E+4	0.019
ISSCC'18 [6]	55	Time	6bits/6bits	3.125	No	Yes	No	-	2.152	690	3.12
This work (MDL CNN)	**40**	**Time**	**8bits/ 1bit****	**0.124**	**Yes**	**Yes**	**No**	**97%**	**0.365**	**30.17**	**12.08**

Figure 14.4.6: Performance summary for C1+P2 and C3+P4 convolutional + pooling layers; proposed MDL comparison with prior approaches.



Technology	40nm CMOS
Area	$0.124mm^2$
# standard cells	~40,000
# of filters supported	16
Data type supported	Signed/ unsigned
Voltage Range	0.375-1.1V
Input clock frequency	25MHz
MAC clock frequency	0.2-3.1MHz
Energy Efficiency @537mV	12.08 TOPS/W
Throughput @537mV	0.365 GOPS
Power @537mV	$30.17\mu W$
LeNet-5 MNIST classification accuracy	97%

Figure 14.4.7: Test-chip die micrograph, characterization setup, and measurements summary table.