ELEC-L352002 - Postgraduate Course in Electronic Circuit Design II

Time-Domain Neural Networks

Ahmed M. Mohey

Aalto University
School of Electrical
Engineering

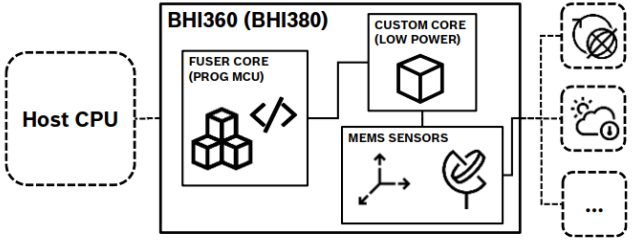
Outline

- Motivation
- Introduction to Time-Domain MAC
- GRO-Based Time-Domain MAC
- GDL-Based Time-Domain MAC
- Conclusion

Motivation

Target Application

- Minimizing both the area and energy consumption of neural network accelerators to enable advanced capabilities to edge devices operating with strict power budget and form factor.
- An example of an edge device that can benefit from these capabilities is the smart sensor from Bosh sensortec.
- The sensor composed of integrated MEMS motion sensor (3-axis accelerometer and gyroscope), a programmable controller unit (FUSER), and a custom IP (accelerator).



Bosh Sensortec Self-learning AI Smart Sensor with Integrated IMU [1]

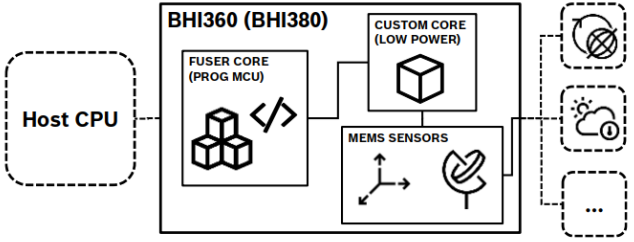
Motivation

Neural Network Accelerator

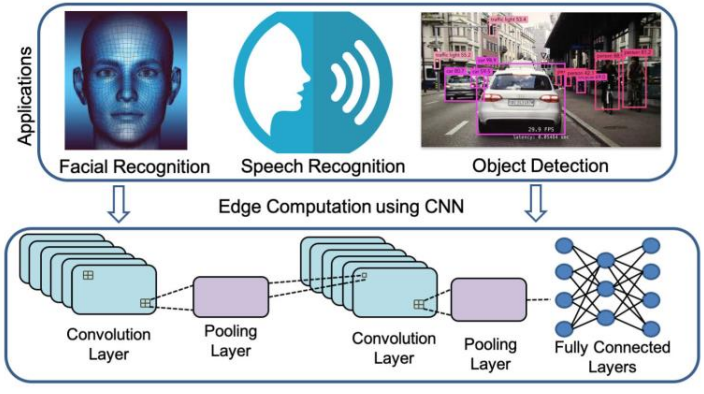
- The major operation in neural network accelerator is the multiply-and-accumulate MAC operations, which is computed by performing a dot product of weight matrix and an input matrix.

$$MAC = \sum_{i=1}^{N} X_i \times W_i$$

- Machine learning accelerators can require billions of MAC operations. Such enormous number of operation consume a significant fraction of the total power budget. Thus, the energy efficiency of MAC operations is extremely important.



Bosh Sensortec Self-learning AI Smart Sensor with Integrated IMU [1]

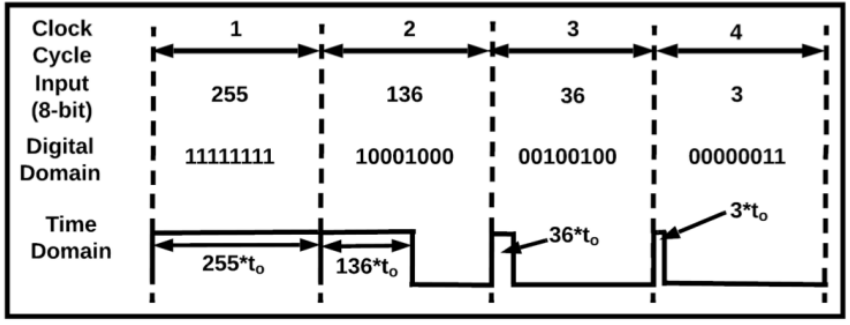


Edge Computing with Neural Network [2]

Introduction to Time-Domain MAC

- In the time-domain computing approach, data can be represented by pulse width modulation PWM, pulse position modulation PPM, or a combination of PWM and PPM, or alternatively data can be expressed by time delay, frequency, or phase.

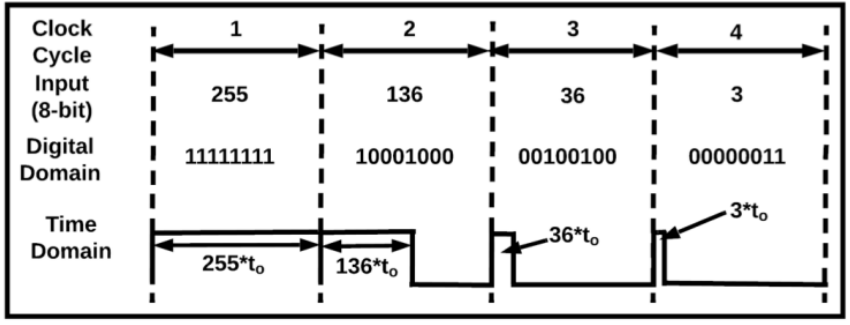
- In [1]-[2], a multi-bit input is represented as a single PWM signal with its pulse width representing the magnitude of the input data. This results in multi-bit data to single data-signal compaction reducing the overall dynamic power consumption.



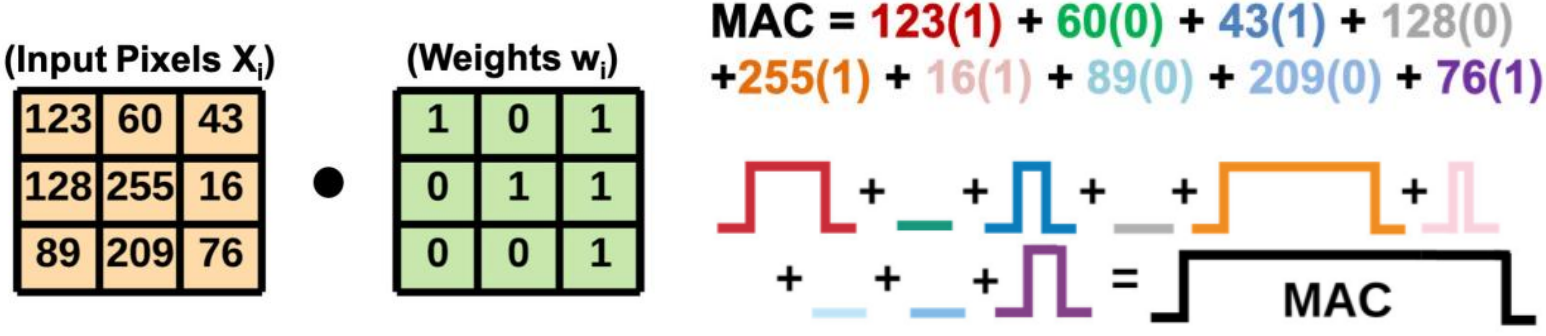
PWM time-domain representation of 8-bit digital input [2]

Introduction to Time-Domain MAC

- In the time-domain computing approach, data can be represented by pulse width modulation PWM, pulse position modulation PPM, or a combination of PWM and PPM, or alternatively data can be expressed by time delay, frequency, or phase.
- In [2]-[3], a multi-bit input is represented as a single PWM signal with its pulse width representing the magnitude of the input data. This results in multi-bit data to single data-signal compaction reducing the overall dynamic power consumption.



PWM time-domain representation of 8-bit digital input [2]

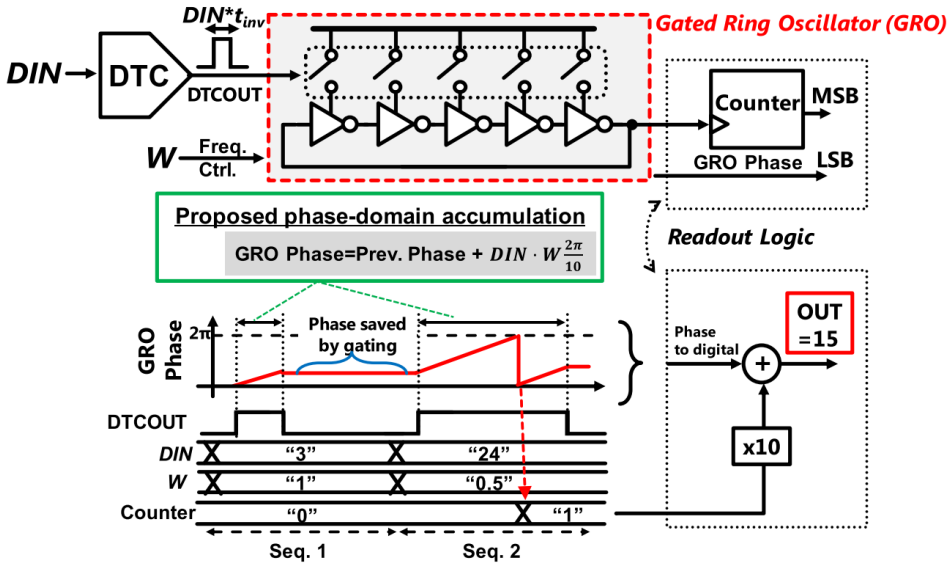


Time-Domain MAC Operation [2]

GRO-Based Time-Domain MAC

Phase-domain MAC Components

- A gated ring oscillator GRO which oscillates when it is enabled by the switches otherwise its phase information. Additionally, its oscillation frequency is controlled by the weight signal W.
- A digital-to-time converter which generates a pulse $DTCOUT$ proportional to the input signal DIN (PWM).
- A counter that increment when the GRO phase returns to 0.
- A readout logic which samples the counter output, and the GRO phases to generates the MAC operation Result

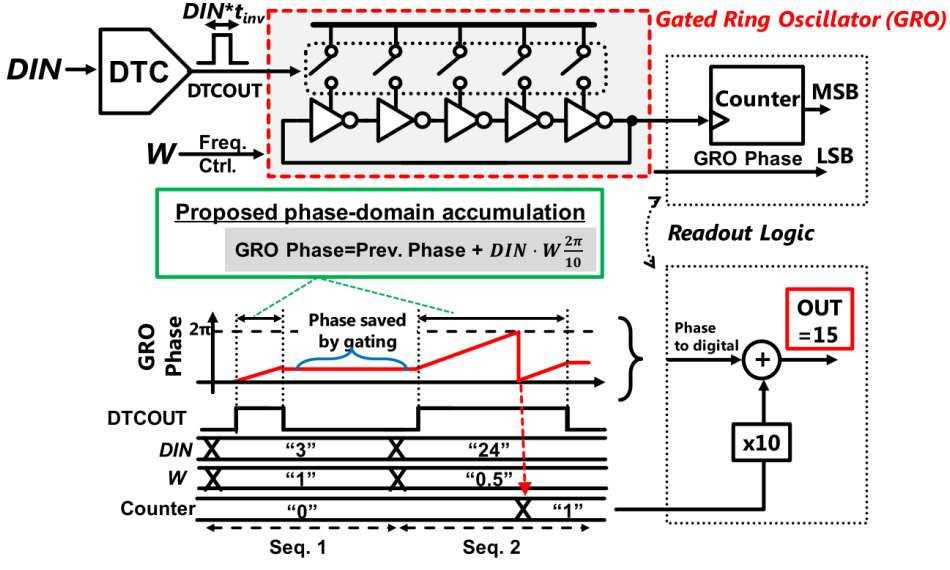


Concept of Phase-Domain MAC [3]

GRO-Based Time-Domain MAC

Phase-domain MAC Operation

- Phases advances by $DIN \cdot \mathrm{W} \cdot \frac{2\pi}{10}$ (in case of 5-stages oscillator). For example, first operation in the figure has $DIN = 3,\ W = 1$. In this case the GRO phase advances by 0.6π.
- The oscillator phase is held until the next operation is triggered.
- When the phase state returns to the initial condition (0), the event is caught by the counter.
- In this case, phase-domain accumulation is realized by the GRO, and it can continuously operate until the counter saturates.

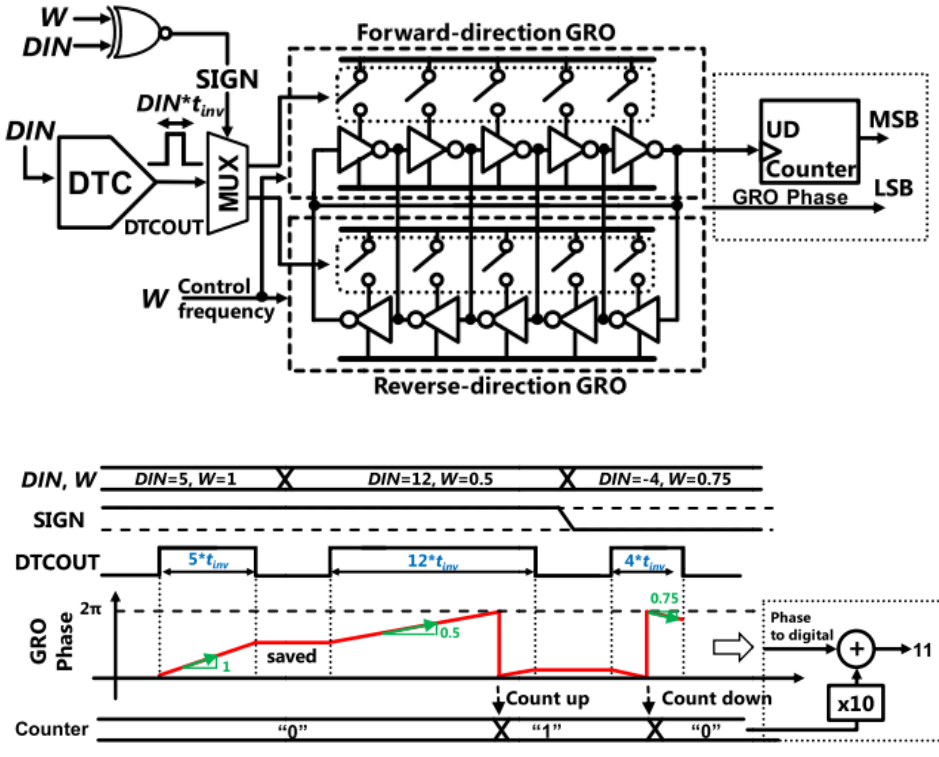


Concept of Phase-Domain MAC [3]

GRO-Based Time-Domain MAC

Circuit Implementation: *Signed Accumulation*

- Signed accumulation can be realized by bi-direction GRO which shares the phase information by connection the oscillator taps of 2 GROs: forward-direction GRO, and reverse-direction GRO.
- The output of the DTC is provided to the forward-direction GRO if the sign is positive, and to the reverse-direction if the sign is negative, allowing the phase to advance in both negative and positive directions.
- The number of rotations of the phase is counted by U/D counter. By reading the counter and the phase information, the result of the sum of products is obtained.

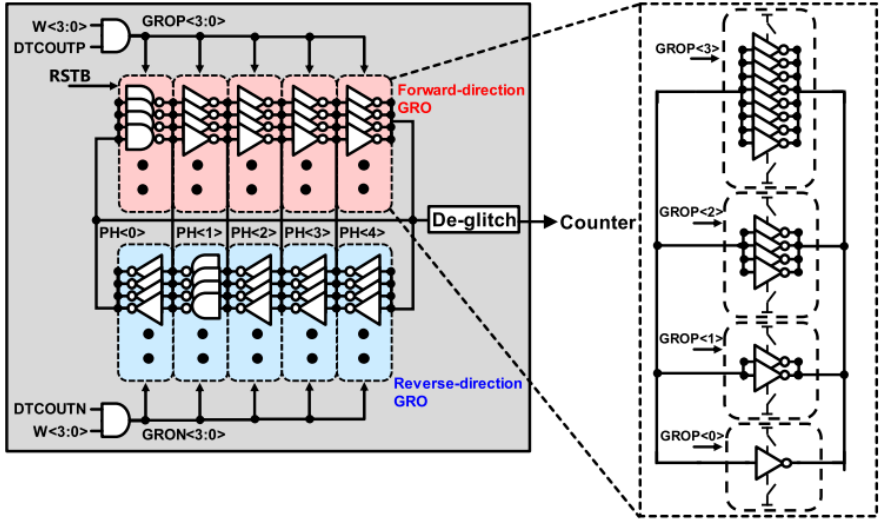


Bi-direction GRO [3]

GRO-Based Time-Domain MAC

Circuit Implementation: *Multi-Bit Weights (Frequency Control)*

- The weigh and frequency resolution is realized by assigned the LSB $W[3:0]$ and the MSB $W[6:4]$ to two individual GROs.
- If $W[i] = 1$, $GROP < i >$ inverters will turn ON, otherwise, the inverters will turn OFF allowing weighted frequency control as required.

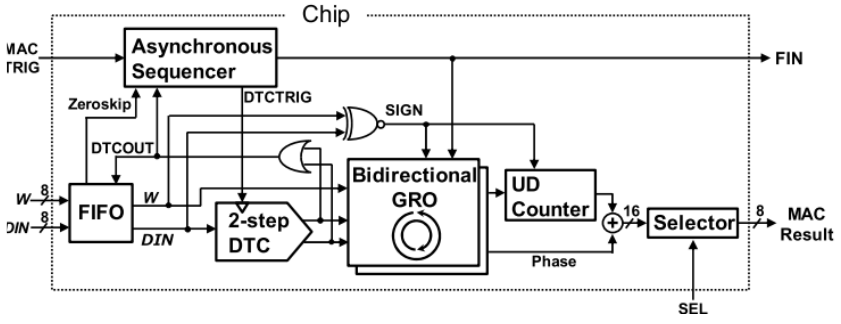


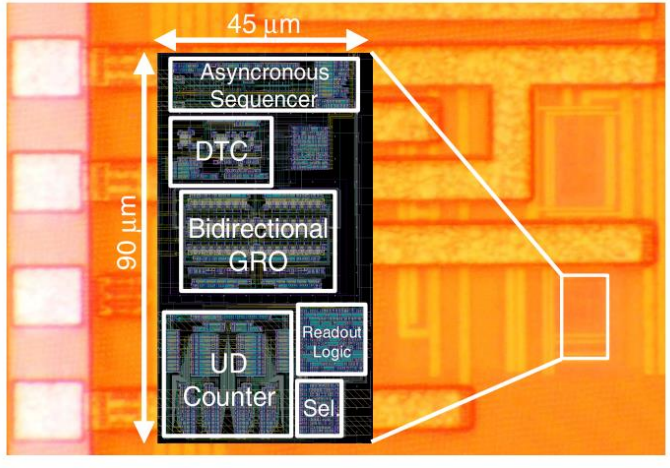
GRO Implementation [3]

GRO-Based Time-Domain MAC

Bi-Directional GRO Uni-Directional GRO

	This Work ASSCC [15]		This Work VLSI [14]	
Domain	Phase		Phase	
Process	28 nm		28 nm	
Resolution	8 bit		8 bit	
MAC Area [μm^2]	960[a]		1200	
Normalized MAC Area[b] [μm^2]	960[a]		1200	
Supply Voltage [V]	0.7		0.7	
Application	MNIST	Anomaly Detection	MNIST	Anomaly Detection
MAC rate [MHz]	753	675	780	700
Efficiency [TOPS/W]	12.4	10.3	14	11.6
Efficiency [TOPS/W*Bit]	99.2	82.4	112	92.8
Area Efficiency [TOPS/mm^2]	1.57	1.41	1.30	1.17





Chip Architecture and Photograph [3]

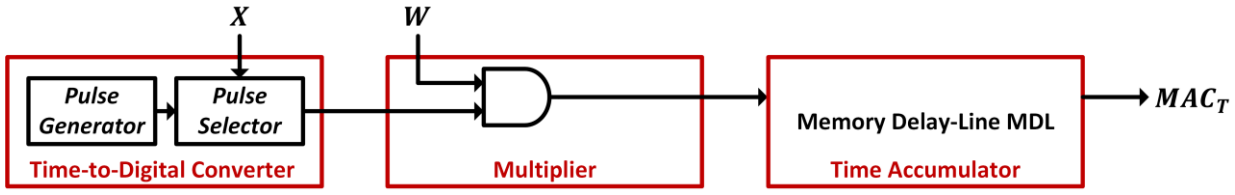
Aalto University
School of Electrical
Engineering

GRO-Based Time-Domain MAC

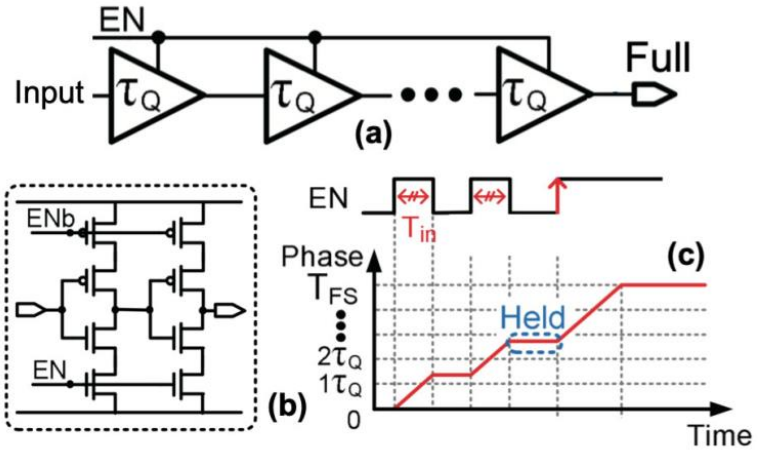
Bi-Directional GRO Uni-Directional GRO

	This Work ASSCC [15]		This Work VLSI [14]		Synthesized Digital		JSSC [26]	JSSC [10]	JSSC [6]	JSSC [5]
Domain	Phase		Phase		Digital		Digital	Charge	Charge	Time
Process	28 nm		28 nm		28 nm		28nm SOI	40 nm	28 nm	65 nm
Resolution	8 bit		8 bit		8 bit		8 bit	3 bit	1 bit	1 bit
MAC Area [μm^2]	960[a]		1200		900		2083	12000	4600	13000
Normalized MAC Area[b] [μm^2]	960[a]		1200		900		2083	5880	4600	6370
Supply Voltage [V]	0.7		0.7		0.7	0.9	0.8	1	0.8	N.A.
Application	MNIST	Anomaly Detection	MNIST	Anomaly Detection	-	-	-	CIFAR 10	CIFAR 10	MNIST
MAC rate [MHz]	753	675	780	700	437	800	1000	1000	10	N.A.
Efficiency [TOPS/W]	12.4	10.3	14	11.6	4.4	2.8	3.2	8.77	532	77
Efficiency [TOPS/W*Bit]	99.2	82.4	112	92.8	34.9	22.4	25.6	26.3	532	77
Area Efficiency [TOPS/mm^2]	1.57	1.41	1.30	1.17	0.97	1.76	0.96	0.17	0.42	N.A.

GDL-Based Time-Domain MAC



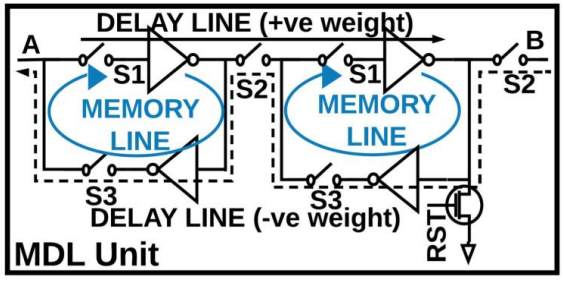
- The property of time addition of two different input pulses is utilized to perform accumulation by utilizing a gated-delay line GDL.
- The phase increases when the EN is high and held constant when the EN signal is low. Hence, when an input pulse is received, the phase of the GDL is advanced by the amount of the input pulse width (high) and held when its low.
- When the input pulse reaches the end of the GDL (phase reaches its full-scale value), a full-scale signal is asserted.



GDL Phase Accumulation [2]

GDL-Based Time-Domain MAC

Forward Delay Line

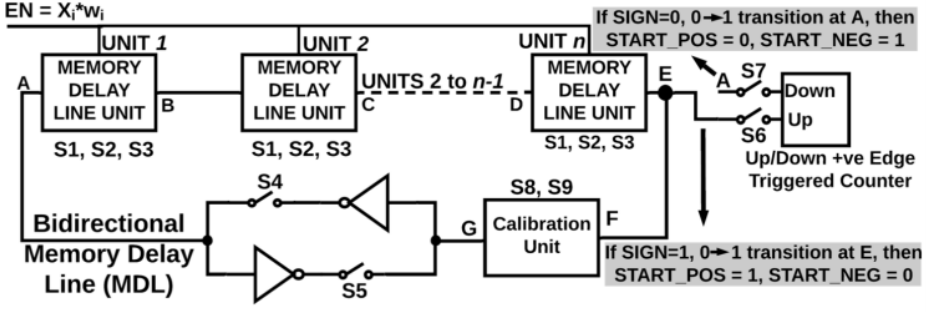


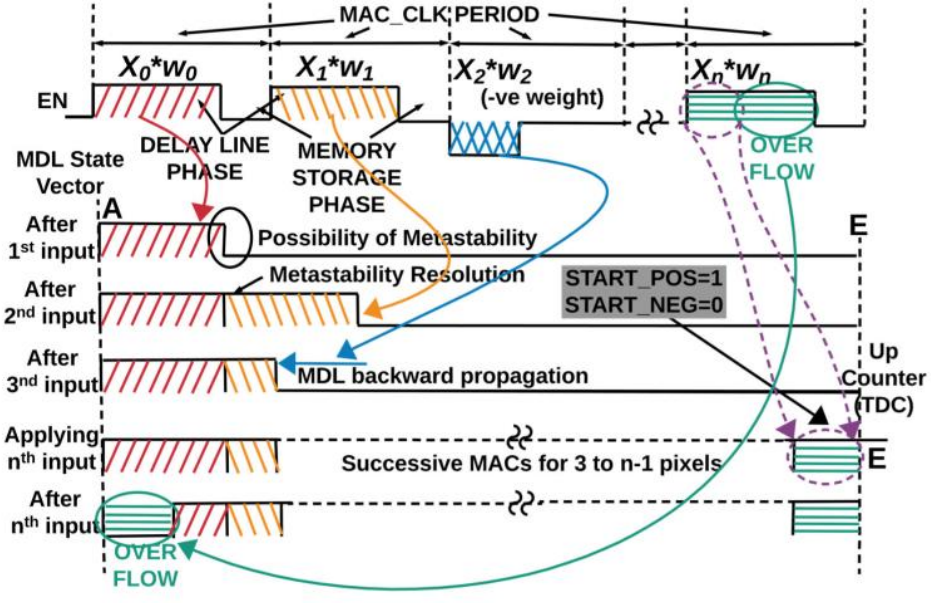
Backward Delay Line

Switch	Logic Equation
S1	$\overline{EN} + (EN.SIGN)$
S2	EN
S3	$\overline{EN} + (EN.\overline{SIGN})$
S4, S8	$SIGN$
S5, S9	$\overline{SIGN}$
S6	$SIGN.\overline{START\_NEG}$
S7	$\overline{SIGN}.\overline{START\_POS}$

If EN=1 (MDL → DELAY LINE)
else (MDL → MEMORY)

If WEIGHT == +1 (SIGN=1)
else (SIGN=0)

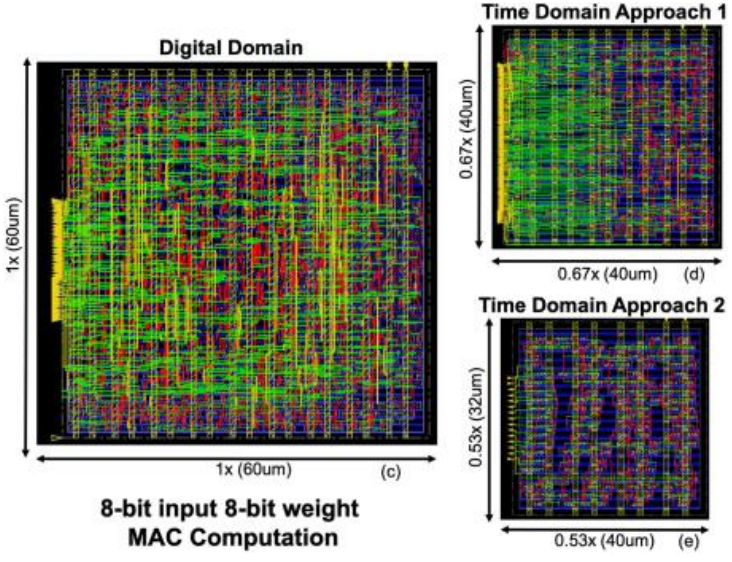




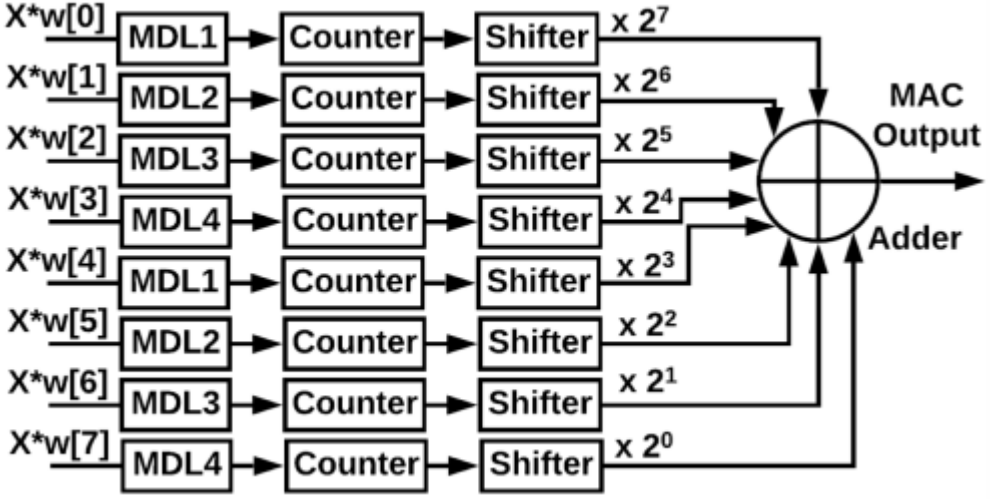
Aalto University
School of Electrical Engineering

*figures adapted from [2]

GDL-Based Time-Domain MAC



8-bit input 8-bit weight MAC Computation



Supporting Multi-bit Weights (Time-Domain Approach 1)

Supporting Multi-bit Weights (Time-Domain Approach 2)

Aalto University
School of Electrical
Engineering

*figures adapted from [2]

GDL-Based Time-Domain MAC

Reference	Tech. (nm)	Circuit Type	Input/ Weight Size	Core Size (mm^2)	Pool-ing	Low Vcc Op.	Cap. or ADCs	Accu-racy	ML/CNN Architecture	Dataset	Through put (GOPS)	Power (μW)	Energy Efficiency (TOPS/W)
ISSCC'16 [14]	65	Digital	16bits	16.00	Yes	No	No	98.3%	LeNet-5	MNIST	64	4.51E+4	1.42^{α}
ISSCC'17 [15]	28	Digital	1-16bits	1.87	No	Yes	No	-	AlexNet/VGG	ImageNet	76	7.60E+3	$10^{\alpha\#}$
ISSCC'18 [16]	65	Analog	8bits	1.44	No	No	Yes	96.0%	SVM	MIT CBCL	-	-	3.125
JSSC'17 [17]	130	Analog	5/1bits	0.267	No	No	Yes	~90%	Classifier	MNIST	-	-	-
ISSCC'18 [18]	65	Analog	6/1bits	0.067	No	No	Yes	99.0%	LeNet-5	MNIST	10.70	380.7	28.10
CICC'17 [19]	65	Frequency	8/3bits	0.24	No	Yes	Yes	91.0%	Multi-layer Perceptron	MNIST##	0.396	2.05E+4	0.019
ISSCC'18 [20]	55	Frequency	6/6bits	3.125	No	Yes	No	-	Reinforcement Lear.	-	2.152	690	3.12
A-SSCC'16 [21]	65	Time	1/1bit	3.61	No	No	No	98.5%	LeNet-5	MNIST	-	-	48.20**
This work (MDL CNN)	40	Time	4/1*bits	0.124	Yes	Yes	No	98.42%$	LeNet-5	MNIST	0.365	30.17	12.08

Reference

[1] Bosch Sensortec, Smart sensor: BHI380 [online] https://www.bosch-sensortec.com/products/smart-sensor-systems/bhi380/#documents

[2] A. Sayal, S. S. T. Nibhanupudi, S. Fathima, and J. P. Kulkarni, “A 12.08-TOPS/W All-Digital Time-Domain CNN Engine Using Bi-Directional Memory Delay Lines for Energy Efficient Edge Computing”, IEEE Journal of Solid-State Circuits, vol. 55, no. 1, pp. 60–75, 2020.

[3] Y. Toyama, K. Yoshioka, K. Ban, S. Maya, A. Saiand K. Onizuka, “An 8 Bit 12.4 TOPS/W Phase-Domain MAC Circuit for Energy-Constrained Deep Learning Accelerators”, IEEE Journal of Solid-State Circuits, vol. 54, no. 10, pp. 2730–2742, 2019.

Assignment Question

- What are the advantages/disadvantages of utilizing time-domain MAC?

Thank You