A 65nm Image Processing SoC Supporting Multiple DNN Models and Real-Time Computation-Communication Trade-off via Actor-Critical Neuro-Controller

Ningyuan Cao[1], Baibhab Chatterjee[2], Minxiang Gong[1], Muya Chang[1], Shreyas Sen[2], Arijit Raychowdhury[1]
Georgia Institute of Technology[1], Purdue University[2]

Abstract

This paper presents a 65nm wireless image processing SoC for real-time computation-communication trade-off on resource-constrained edge devices. The test-chip includes (1) an all-digital, near-memory, reconfigurable and programmable neural-network (NN) based systolic image processor at 1.05TOPS/W (peak), (2) a digitally-adaptive RF-DAC based transceiver with Tx energy-efficiency of 768pJ/b and (3) a mixed-signal, time-based, actor-critic neuro-controller with compute-in-memory (CIM) and in-place weight updates that provides online learning and adaptation at 0.59pJ/MAC for efficiently controlling the computation, communication blocks separately as well as jointly.

Introduction and Motivation

The wide spread proliferation of smart sensors has led to hardware that enable edge intelligence (EI) with extreme energy-efficiencies. This decreases the volume of data that is transmitted to the cloud, thus reducing: (1) processing latency, (2) communication energy and (3) network congestion. However, this comes with an added cost of computation at the edge node [1-3] (Fig. 1(a)). The cost (energy/latency) of edge computation and the cost of communication to the cloud vary widely depending on operating conditions, that include (1) information content in the data, (2) algorithm selection, (3) channel conditions (noise, path-loss etc.), (4) network size, available bandwidth and (5) resources at the cloud, as shown in Fig. 1(b). We call the number of NN layers processed at the edge, processing-depth (PD). Increasing edge-computation increases PD, but reduces the volume of data to be transmitted. This not only provides an opportunity to efficiently configure computation and communication blocks but also trade-off between computation and communication in real-time to meet system targets.

SoC Architecture and Circuit

The system architecture, shown (Fig. 1 (c)), illustrates (1) an analytics engine with 9 processing elements (PEs) each with 8 ALUs and 8KB of SRAM, (2) digitally controlled transceiver, (3) actor-critic neuro-controller and (4) peripheral circuits including frame-buffer, scan, interfaces, instruction/data cache.

The chip features a 3 x 3 array of PEs with (1) retention-enabled, local 8KB SRAM (2) 8 programmable ALUs and (3) input buffers/controllers, as is shown in Fig. 2. It supports (1) fully connected (FC) layers, (2) weight-stationary CONV layers and (3) sparse networks. For sparse networks, the weights (w) and the indices (d) are stored in separate sub-banks. Control bits allow choice of NN model, optimal PD, and enables retention mode for un-accessed sub-banks. The energy-efficient and adaptive RF subsystem (Fig. 3), contains a reconfigurable RF-DAC based Inverse class-D PA with adaptive closed loop control on data rate (8 bits), and error correction coding (ECC – 1 bit control for [8,4] Hamming Code) and Output power (3 bits). Information about the channel and network conditions are received through a low-power OOK receiver with 2 stages of RF gain (~20 dB), an Envelope detector (ED) and 2 stages of BB VGA (~20-40 dB). The controls on data rate and ECC are employed in the digital baseband in the Tx, while the output power control (along with digital amplitude pulse-shaping) is performed using an on-chip tapped capacitor matching network (MN) with tunable capacitor banks for low MN loss and high back-off efficiency. The 4 LO phases are generated using an injection locked inverter based quadrature LO generator, which are fed to I and Q-paths (Fig. 3).

The large control space across computation and communication is learnt using a low overhead (~5% power) actor-critic NN (AC-NN) controller (Fig. 4). The AC-NN takes both design targets and sensed variables as inputs and learns to optimally control the control knobs. These are listed in Fig. 1(b). The motivation for actor-critic NN on the edge is its real-time controllability as well as ability to model the SoC dynamics and environmental variables while providing an optimal policy through a single inference. The controller features (1) 4 10 x 10 memory sub-banks, and (2) a NN controller. The actor-critic NNs store 8b thermometer encoded data and enables time-based compute-in-memory (CIM). During inference, digital to time converters (DTCs) allow pulse width modulated word-lines (WLs) (input signals) to be turned on sequentially such that the falling edge of one row triggers the rising edge of the next. The partial products are accumulated on the BL as long as V_{BL} is greater than a threshold (V_L) to avoid read disturb. However, if the operands are large and V_{BL} reaches V_L then the process is stopped, the ADC converts to a 6b word, the BL pre-charged and the sequence restarts. The differential bitcell and ADC allows both positive and negative weights by discharging either BL (positive) or BL_bar (negative). The thermometer encoding of data enables a weight update to be a left or a right shift (sign of the update), and that the duration of shift process (magnitude of update) is controlled by the DTC (Fig. 4). The array can be read both row as well as column-wise providing a seamless design for transposing the weight matrix during back-propagation. This also enables in-place online learning without requiring reads and write-backs (baseline designs).

Measurement Results

The measured power-performance of the processing engine (Fig. 5) shows V_{MIN} of 0.5 V and F_{MAX} of 760 MHz. Peak arithmetic energy-efficiency of 1.05 TOPS/W (0.43 TOPS/W, 0.18 TOPS/W) is measured for CONV (FC, sparse) networks at 210 MHz (0.575 V). The RF subsystem, shows a maximum Tx efficiency of 30.3% at -0.3 dBm, with back-off efficiencies of 19.2% (7.8%) at -6.5 (-13.7dBm) with QPSK. At 1 Mbps, the Tx energy efficiency is 768 pJ/bit with 1 V supply (-0.3 dBm output power. The measured energy-efficiency for the OOK Rx is 207 (124) pJ/bit at 1 (0.8) V supply, with a sensitivity of -72 dBm for a BER of 10^{-3} at 1 Mbps. An [8,4] Hamming Code on the Tx improves the sensitivity to -78 dBm but halves the number of information bits. The measured performance of the neuro-controller is shown in Fig. 5. The CIM consumes a measured 305.2 pJ (training) and 156.8 pJ (inference) at 0.7V with less than 0.6lsb of non-linearity error. The peak measured energy efficiency is 0.59 pJ/MAC and 0.4 pJ for each weight update which are 2.2x and 4.75x lower than a digital counterparts (simulated). The full system is deployed and neuro-controller is allowed to learn online from emulated signals from the cloud and energy meters. Then it is tested for varying noise power and network sizes and the system autonomously determines the optimal PD to minimize energy, latency or EDP. The online adaptation allows the system to learn and choose the CTRL parameters optimally. We test across various conditions of path-loss, number of edge nodes (i.e., available bandwidth) and obtain a 2.44x (1.47x) improvement in average energy (latency) for a BER of 10^{-3} compared to the baseline cases while running a modified AlexNet that maps to the SoC. Benchmarking across other designs [1-6] show competitive figures-of-merit. The design presents a vertically integrated SoC featuring the first real-time NN based adaptation for computation, communication and their trade-offs in energy constrained systems.

Acknowledgement

This project was supported by the Semiconductor Research Corporation under grant no. 2720.001 and JUMP CBRIC task ID 2777.006.

References

[1] N. Cao et. al., ISSCC, pp. 222–224, 2019.
[2] D. Shin et. al., ISSCC, pp. 240–241, 2017.
[3] S. Bang et. al., ISSCC, pp. 250–251, 2017.
[4] A. Paidimarri et. al., JSSC, vol. 48, no. 4, pp. 1042–1054, 2013.
[5] M. Lee et. al JSSC, vol. 54, no. 6, pp. 1541-1552, 2019.
[6] T. Karnik et. al., ISSCC, pp. 46–48, 2018.

978-1-7281-9942-9/20/$31.00 ©2020 IEEE

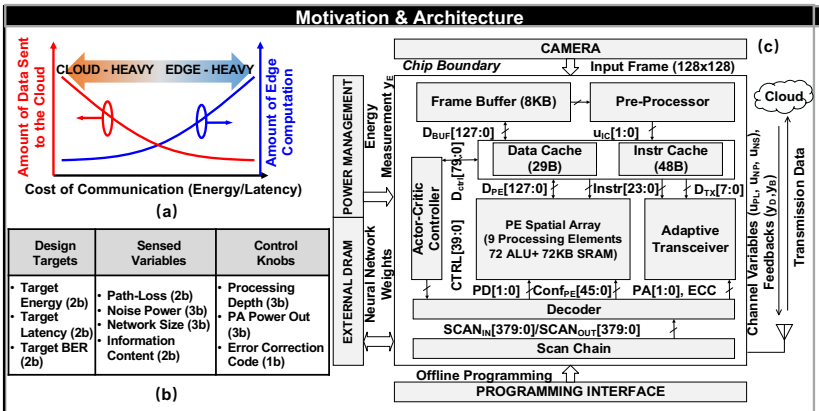


Fig.1: (a) Design Motivation (b), Sources of dynamic variations, and (c) the SoC architecture

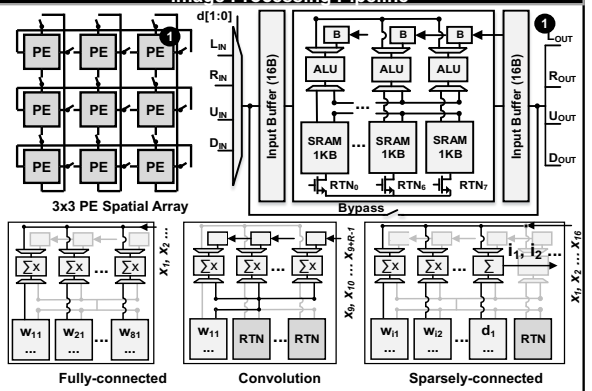


Fig.2: Computation: PE array and algorithm-driven re-configuration

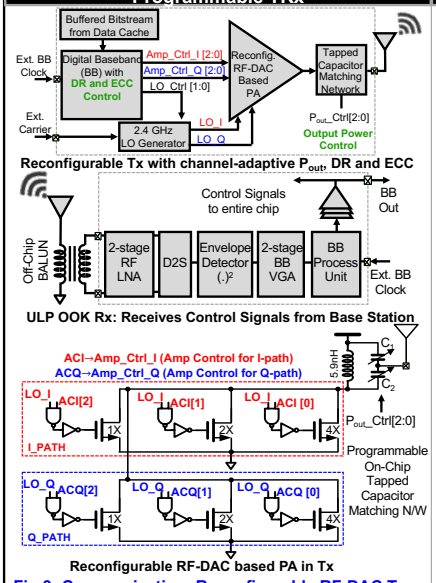


Fig.3: Communication: Reconfigurable RF-DAC Tx and ULP OOK Rx

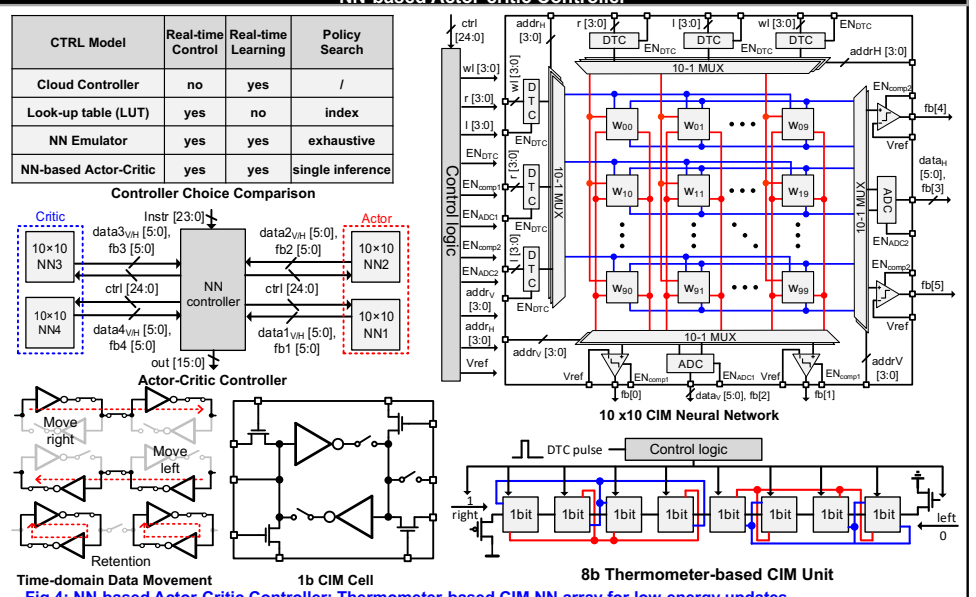
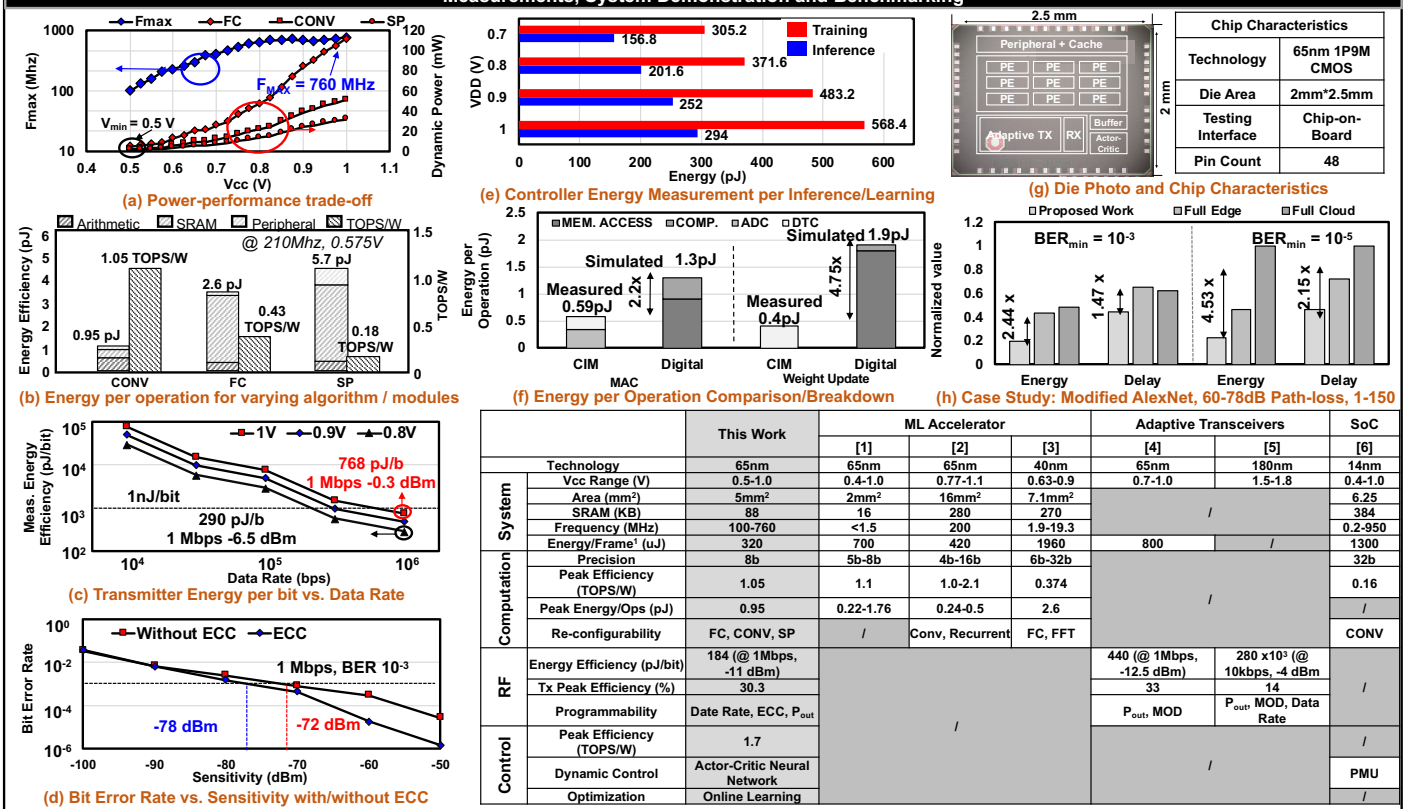


Fig.4: NN-based Actor-Critic Controller: Thermometer-based CIM NN array for low-energy updates



		This Work	ML Accelerator [1]	ML Accelerator [2]	ML Accelerator [3]	Adaptive Transceivers [4]	Adaptive Transceivers [5]	SoC [6]
	Technology	65nm	65nm	65nm	40nm	65nm	180nm	14nm
System	Vcc Range (V)	0.5-1.0	0.4-1.0	0.77-1.1	0.63-0.9	0.7-1.0	1.5-1.8	0.4-1.0
	Area (mm²)	5mm²	2mm²	16mm²	7.1mm²			6.25
	SRAM (KB)	88	16	280	270	/		384
	Frequency (MHz)	100-760	<1.5	200	1.9-19.3			0.2-950
	Energy/Frame[1] (uJ)	320	700	420	1960	800	/	1300
Computation	Precision	8b	5b-8b	4b-16b	6b-32b			32b
	Peak Efficiency (TOPS/W)	1.05	1.1	1.0-2.1	0.374			0.16
	Peak Energy/Ops (pJ)	0.95	0.22-1.76	0.24-0.5	2.6	/		/
	Re-configurability	FC, CONV, SP	/	Conv, Recurrent	FC, FFT			CONV
RF	Energy Efficiency (pJ/bit)	184 (@ 1Mbps, -11 dBm)				440 (@ 1Mbps, -12.5 dBm)	280 x10³ (@ 10kbps, -4 dBm	
	Tx Peak Efficiency (%)	30.3				33	14	/
	Programmability	Date Rate, ECC, P_out	/			P_out, MOD	P_out, MOD, Data Rate	
Control	Peak Efficiency (TOPS/W)	1.7						/
	Dynamic Control	Actor-Critic Neural Network				/		PMU
	Optimization	Online Learning						/

Fig.5: Measured (a) power-performance, (b) energy/op for computation, (c-d) TX energy and BER w/ and w/o ECC, (e-f) Controller energy / op, (g) die-shot and chip characteristics, (f) system study for AlexNET architecture with dynamic environmental conditions and Benchmarking Table.