IDEC Chip Design Contest

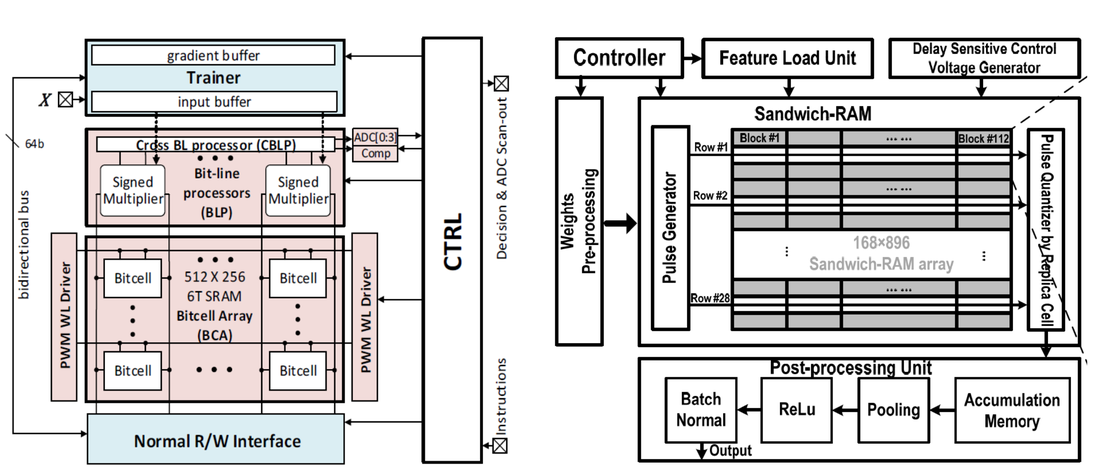
Modified SRAM based Computation-In-Memory for CNNs

Jongho Kim, and Jintae Kim
Konkuk University, Korea
Mixed Signal Electronics Lab (MSEL)
jh.kim@msel.konkuk.ac.kr
The chip fabrication was supported by the IC Design Education Center(IDEC), Korea.

Abstract

We present an 8-bit precision 6T-SRAM based mixed-signal multiply-and-accumulate (MAC) processor for convolutional neural network (CNN) algorithms. The chip has been fabricated in 65nm CMOS process using a 16kb SRAM and measured power dissipation is approximately 3.2mW when running at 100MHz, achieving energy efficiency of about 4TOPS/W. The backend A/D converter consumes power of less than 5% of the total power, which is well amortized along the parallel bitline column processors.

Previous Studies on In-Memory-Computing and Our Proposed Scheme



(a) (b)

Fig. 1 (a) Gonugondla et al., ISSCC18, (b) Jun Yang et al., ISSCC19

	Our chip	Reference [1]	Reference [2]
Architecture	In memory (6T SRAM)	Switched Capacitor	In-memory (6T SRAM)
Application	MobileNet	MNIST	SVM
Resolution [bit]	8	8	8
Output Rate[MHz]	100	2.4	32
Efficiency[TOPS/W]	4	9.61	3.12
Area [um x um]	675 x 605	113 x 102	1200x1200

Fig. 2 Comparison Table

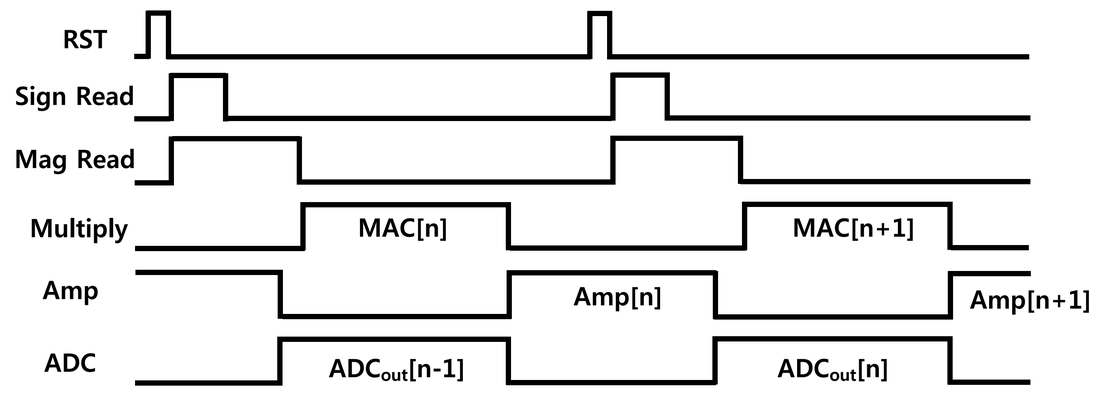


Fig. 3 Clock timing diagram

Recently, in-memory computing design paradigm which aims to minimize power consumption by reducing memory access and through massive parallel computing is widely explored. Previous studies on SRAM based MAC architecture suffers from nonlinearity in MAC output due to the parasitic load on bitlines (BL) and the non-ideal behavior of bitcells as weight D/A converters (DAC).

In our proposed scheme, we improved the MAC linearity by customizing bitcells. Also, by exploiting the sparsity in the weights of CNNs, we reduced the BL switching power which dominates the overall power consumption without lowering the MAC precision. The process and temperature variation have been compensated using foreground calibration via reference generator.

Chip Layout View

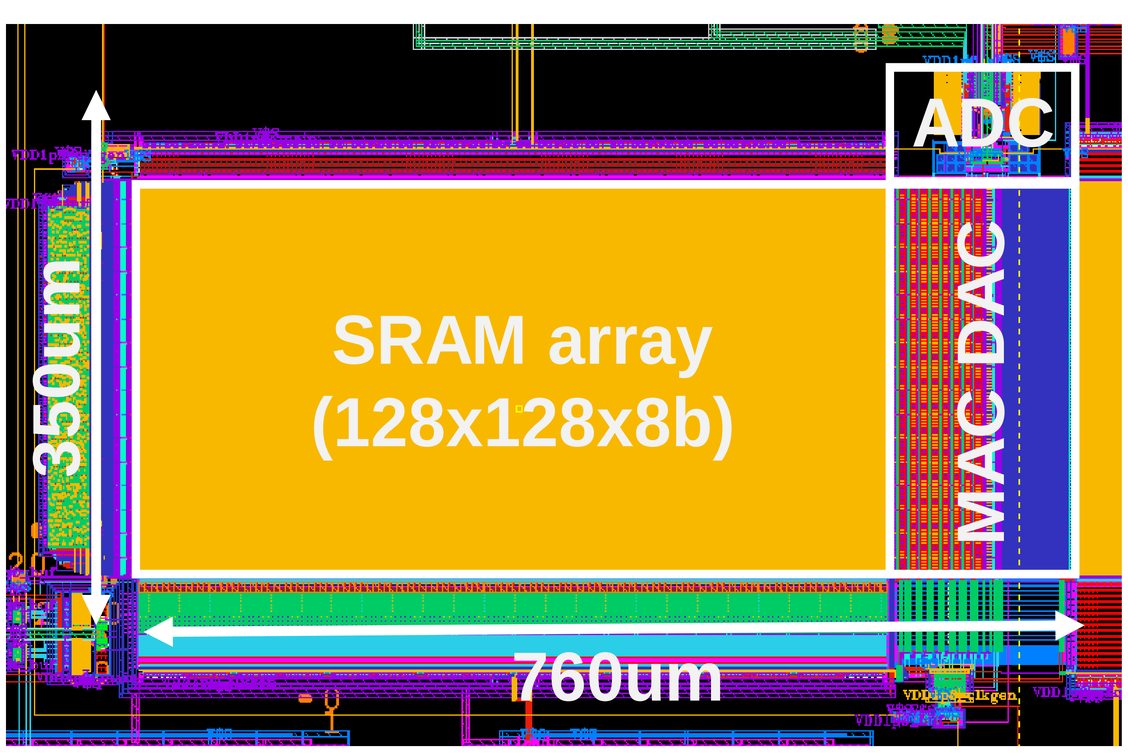


Fig. 4 Top level layout view of CIM macro

Simulation Results

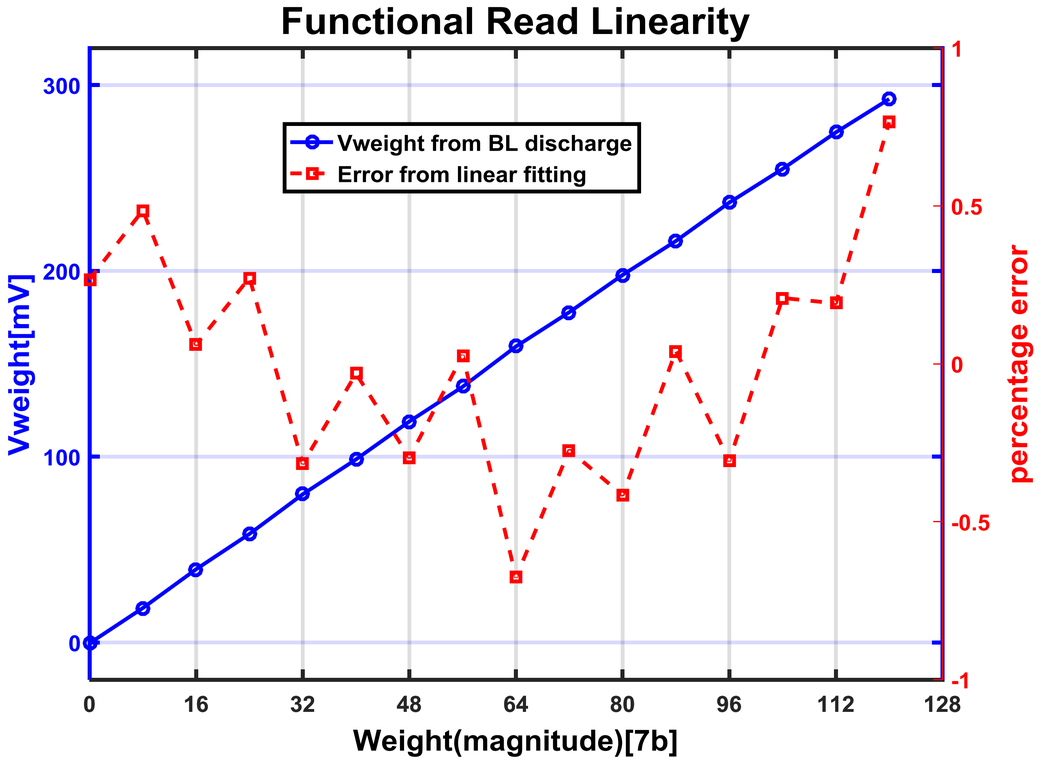


Fig. 5 Simulated plot of the weight DAC linearity