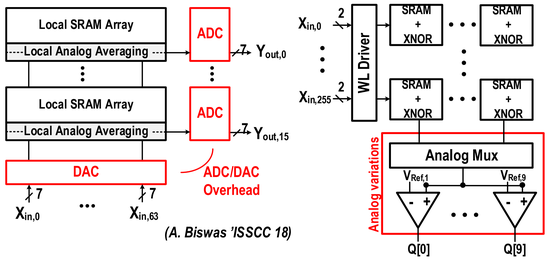
Precision Reconfigurable Digital Compute-In-Memory for Embedded Neural Network Processing

Prof. Bongjin Kim (bjkim@ntu.edu.sg), **Hyunjoon Kim** (kimh0003@e.ntu.edu.sg)
Nanyang Technological University, Singapore, Singapore

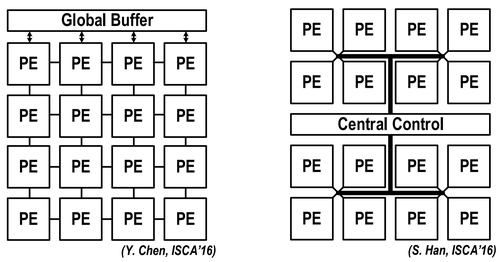
Design Features

- ✓ Fully digital (scalable, robust to variations)
- ✓ SRAM based Compute-In-Memory (CIM) architecture
- ✓ Precision reconfigurability (weight, activation and input)

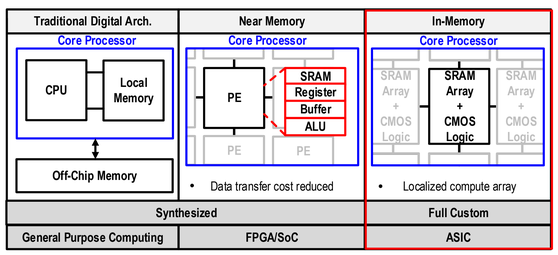
Background



- Analog SRAM based design concerns: ADC/DAC overhead, PVT variation and SRAM RD/WR disturb issues

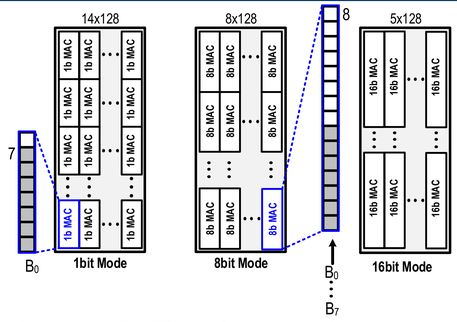


- Previous digital designs are not CIM, larger area than analog and have limited reconfigurability

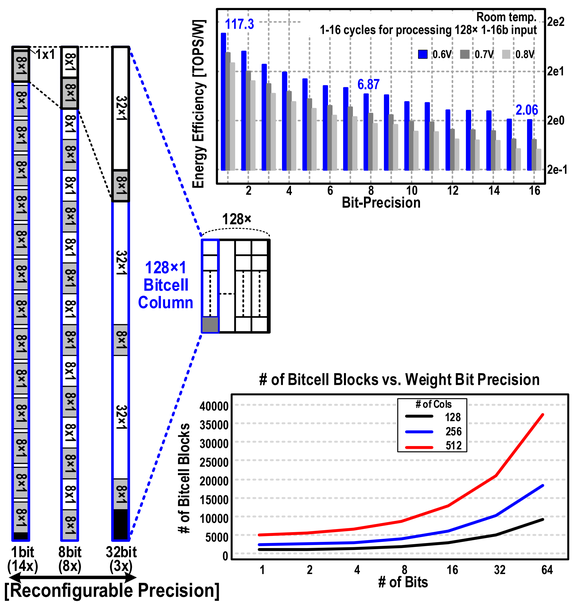


- Compared to other digital architectures, CIM architecture needs to improve on **Reconfigurability**, **Scalability** and **Design Flow**

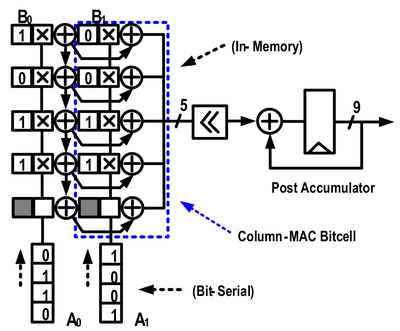
Reconfigurability Implementation



- # of white cells (Type-A) = weight bit precision
- # of grey cells (Type-B) = fixed based on the # of columns



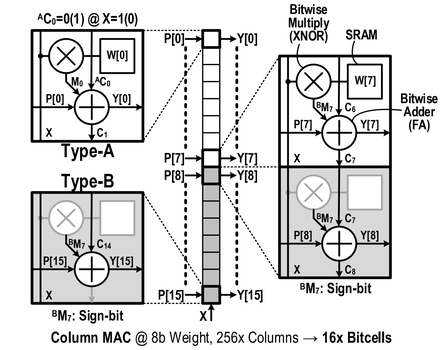
Digital CIM Macro Architecture



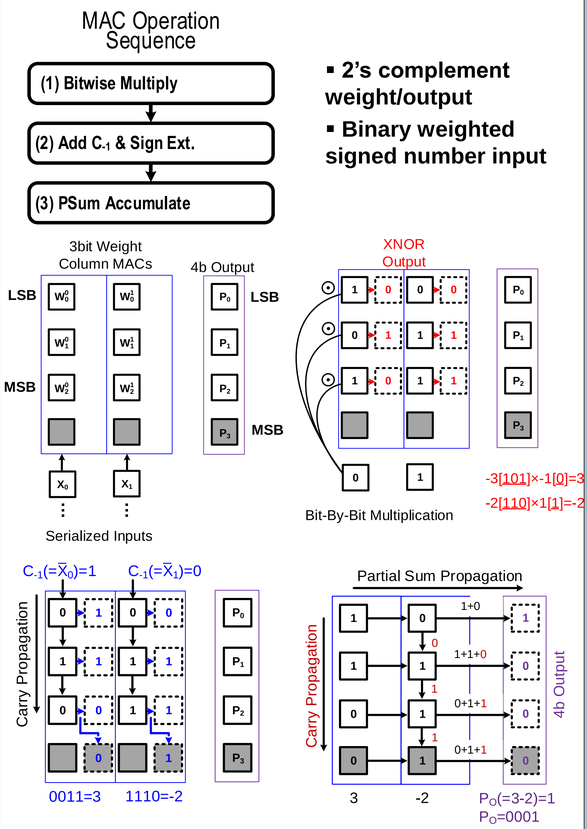
- Input/output clock cycles determine input/output bit precision/activation

Macro Bitcell Configuration

- Regular column structure multiply-and-accumulate (MAC) array composed of two (Type-A and Type-B) different configuration unit bitcell blocks
- Different control signals applied to alter the configuration of the unit cell based on its position in the column-MAC

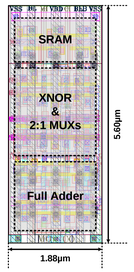


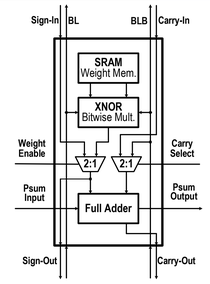
3-bit Operation Example



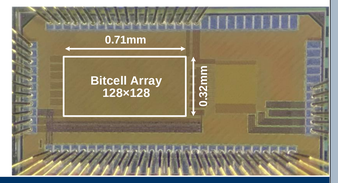
Bitcell Macro Test Chip

	[3]Envision ISSCC'17	[5]UNPU ISSCC'18	[7] VLSI'18	[Our Work] ESSCIRC'19
Multiply Precision	1-16/N bit (N=1,2,4)	1-to-16bit	FP16b INT8/16b	**1-to-16bit**
Accumulate Precision	48/N bit	32bit	FP32b INT24/48b	**8-to-23bit**
Reconfigurability	Reconfig. Multiplier	Bit-Serial	Fixed Bits (8,16,24,48)	**Column MAC Bit-Serial**
Energy per MAC [pJ/MAC]	N/A	0.055(1b) 1.26(16b)	N/A	**0.017(1b) 0.78(16b)**
Min. Energy Eff. [TOPS/W]	0.26	3.08(16b)	0.55(16b)	**2.06(16b)**
Max. Energy Eff. [TOPS/W]	10	50.6(1b)	11.3(8b)	**117.3(1b)**

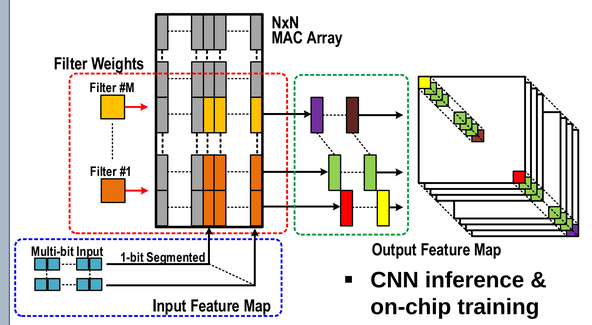




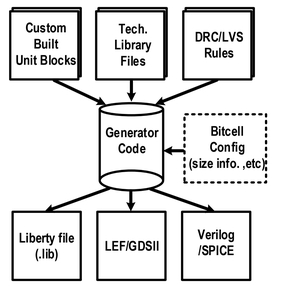
- Custom built circuit using 65nm LP



Future Development



- CNN inference & on-chip training



- Compiler design for design automation & SoC system design