2021

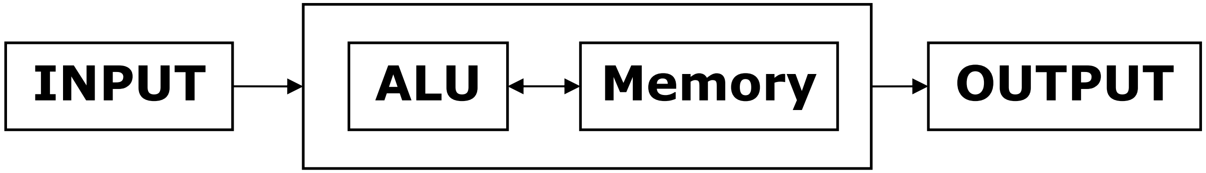
B11901027 王仁軒

PPT TemplateCredit : Jie-Hong Roland Jiang

Motivation & Background

- ML computing is data-centric
  - Energy in Traditional Von Neumann structure is mainly consumed by memory accesses
  - Computation in memory tackles this problem

- CIM structures focus on analog approach more
  - Analog designs lack accuracy (SNR $\propto \frac{1}{\sqrt{\mathrm{Bit}}}$)
  - Using digital structures can guarantee accuracy, also good for operations such as batch-norm, pooling

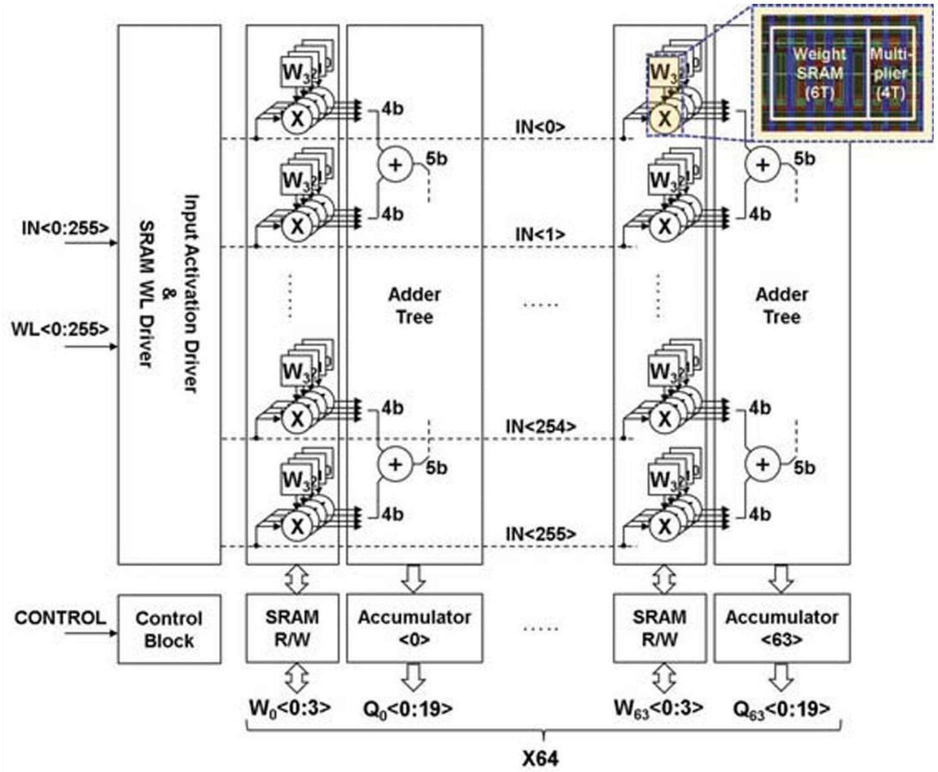


Features

- High bit-flexibility
  - Programmable bit-widths, signed/unsigned, and weights with 4, 8, 12, 16 bit widths..
- Parallel MAC operations, with high energy/area efficiency
- Allows simultaneous MAC and write operations
  - Time is wasted when updating weights

Structure

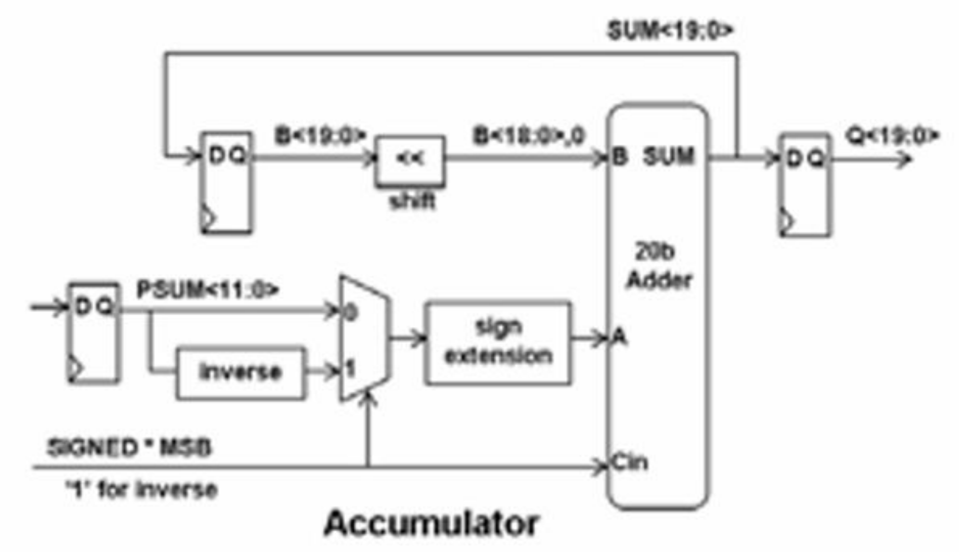
- 2 Mode: SRAM & CIM
  - SRAM: update weight
  - CIM : MAC operation, support 256×64 multiplication

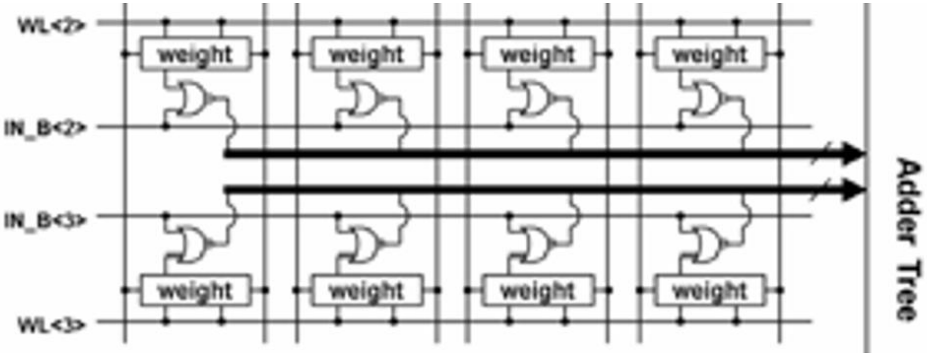


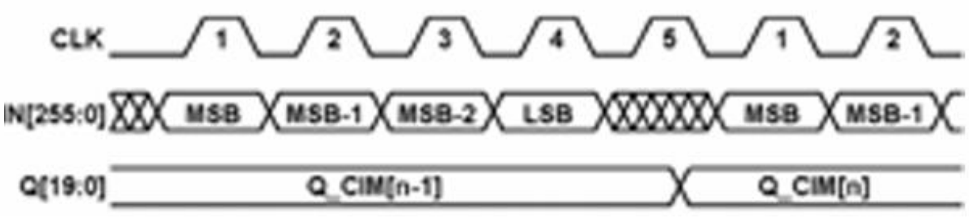
Structure

- 2 Mode: SRAM & CIM
  - SRAM: update weight
  - CIM : MAC operation , support 256×64 multiplication

- Require 5 cycles for 4-bit input activation

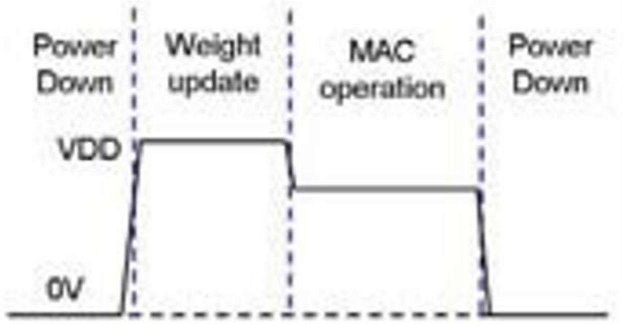




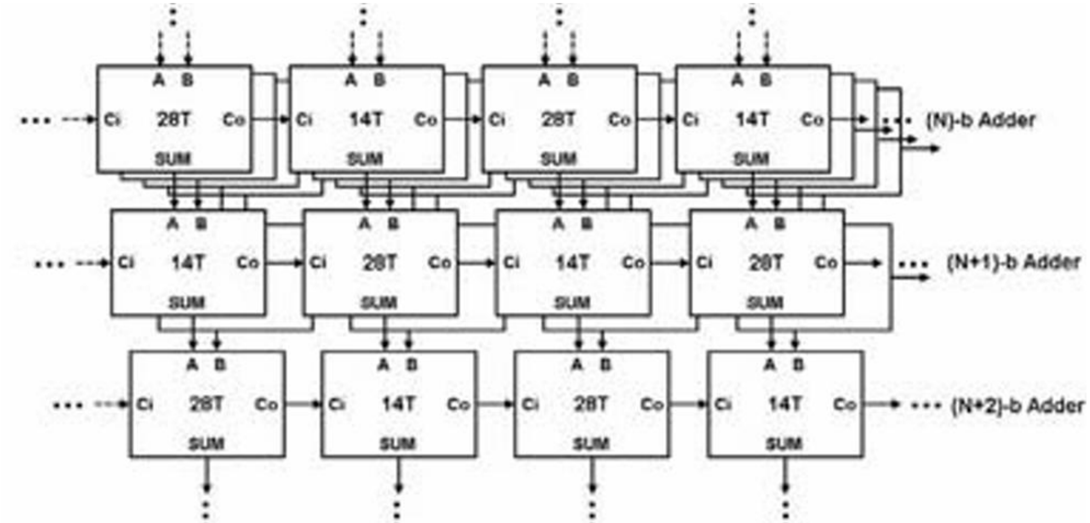


Structure

- Dynamic voltage scaling
  - 0.8V for weight updating
  - 0.68V for MAC operation
  - Better NM and less dynamic power

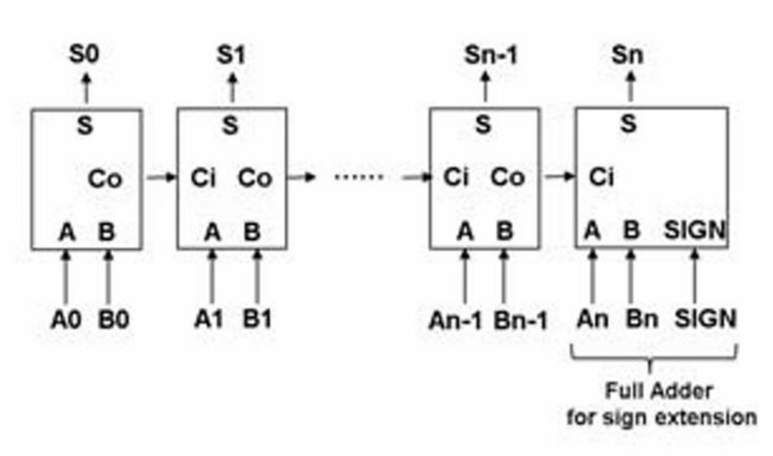


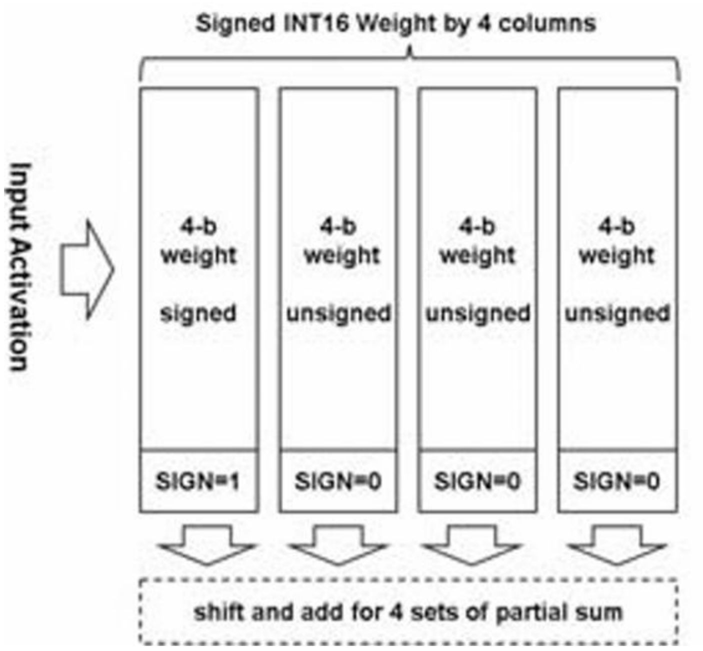
- Interleaved 28T & 14T Adder
  - Lower dynamic power



Structure

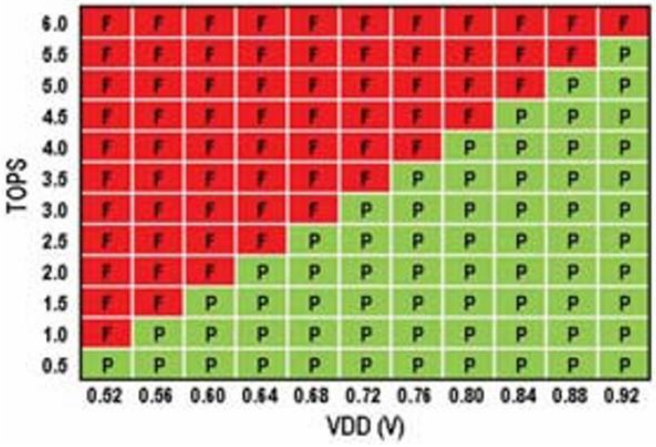
- Update weight and perform MAC simultaneously
  - When input = 0, Update weight
- Flexible bit width
  - Separate 16 bit into 4b signed And 3×4b unsigned



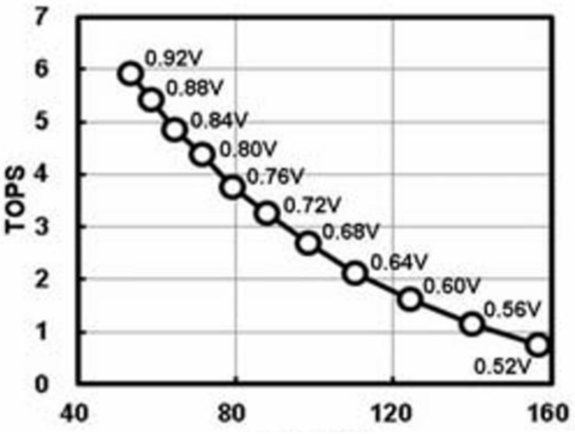


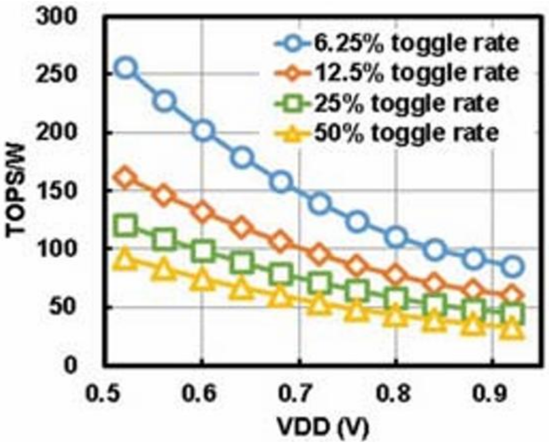
Result

- 89TOPS/W under 0.72V for a sparse pattern
  - □(18% input toggle rate, 50% 1s for weights)
- 52TOPS/W for a dense pattern
  - □(50% input toggle rate, 50% 1s for weights)
- 3.3TOPS from 0.72V at 25°C
  - □TOPS/W and TOPS/area improving by 2.8× and 19× under 5nm process compared to 22nm design



TOPS	0.52	0.56	0.60	0.64	0.68	0.72	0.76	0.80	0.84	0.88	0.92
6.0	F	F	F	F	F	F	F	F	F	F	F
5.5	F	F	F	F	F	F	F	F	F	F	P
5.0	F	F	F	F	F	F	F	F	F	P	P
4.5	F	F	F	F	F	F	F	F	P	P	P
4.0	F	F	F	F	F	F	F	P	P	P	P
3.5	F	F	F	F	F	F	P	P	P	P	P
3.0	F	F	F	F	F	P	P	P	P	P	P
2.5	F	F	F	F	P	P	P	P	P	P	P
2.0	F	F	F	P	P	P	P	P	P	P	P
1.5	F	F	P	P	P	P	P	P	P	P	P
1.0	F	P	P	P	P	P	P	P	P	P	P
0.5	P	P	P	P	P	P	P	P	P	P	P





Conclusion

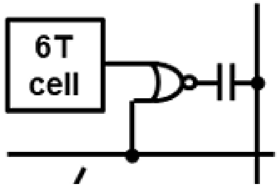
- Full precision
  - Digital adder tree structure with 256×64 bit sram array
- Low power
  - Dynamic voltage scaling (lower VDD for MAC)
  - Half of the 28T adders are replaced by 14T adders
- Highly programmable
  - Support different input bit-width and weight bit-width
  - Requires extra adder for sign extension and control blocks

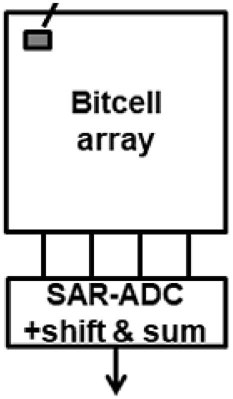
2022

PPT TemplateCredit : Jie-Hong Roland Jiang

Motivation & Background

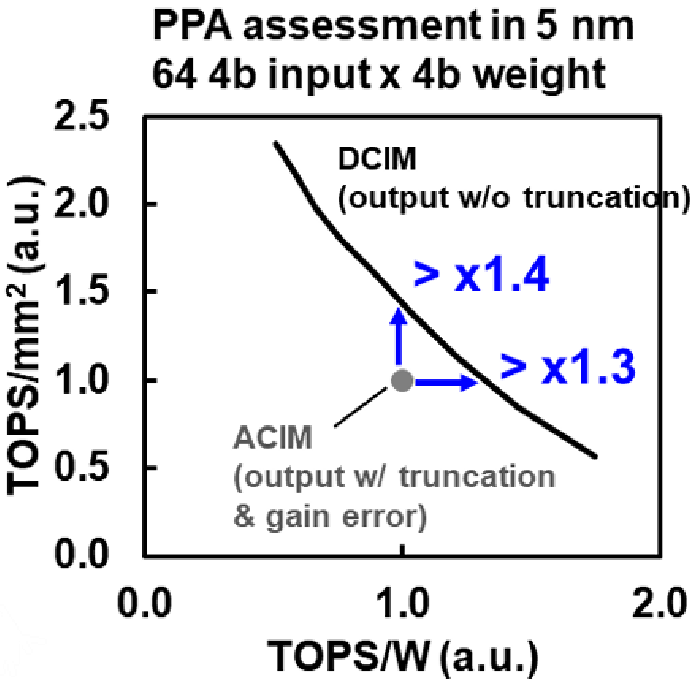
- Accuracy and scale of analog approach is limited
  - MOS PVT variation
  - ADC output range under dynamic voltage scaling
  - Truncation and gain error of ADC
  - Nonlinear Id due to drain-induced-barrier-lowering for current-based approach (lower Vt for short channel)
  - Smaller caps with technology-scaling for cap-approach
  - Analog components are not good for testing





Motivation & Background

- Digital approach achieves better performance
  - >1.3x power efficiency
  - >1.4x performance/area efficiency

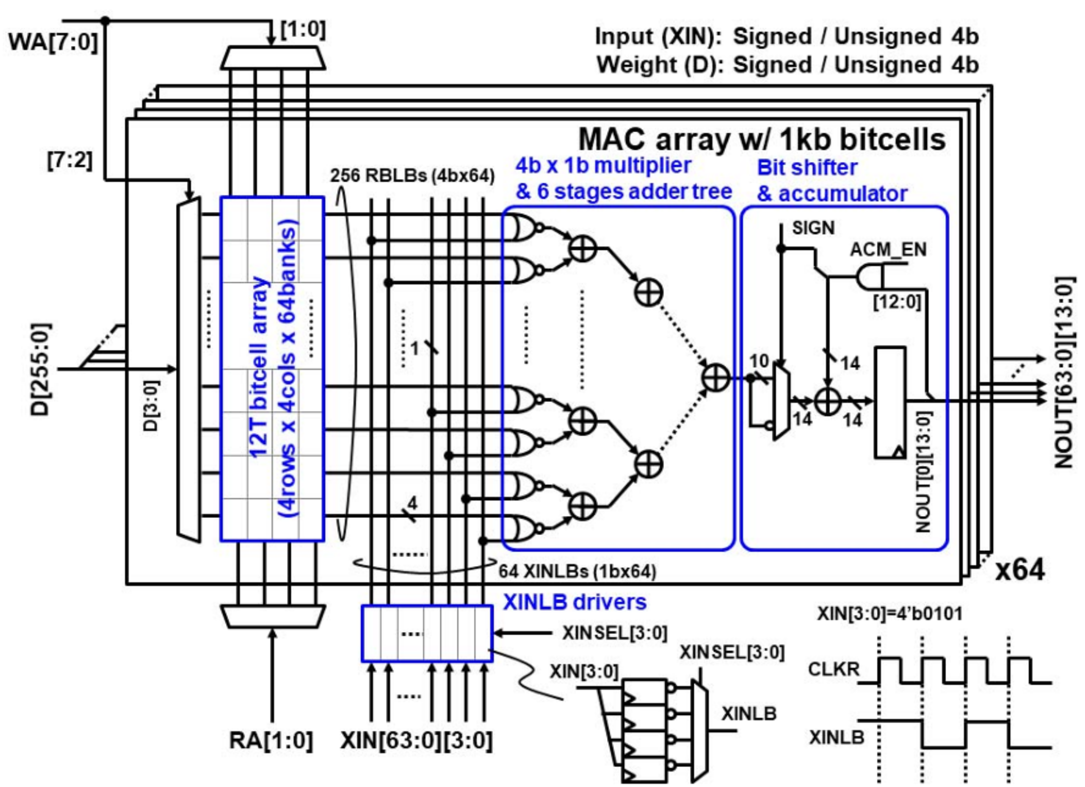


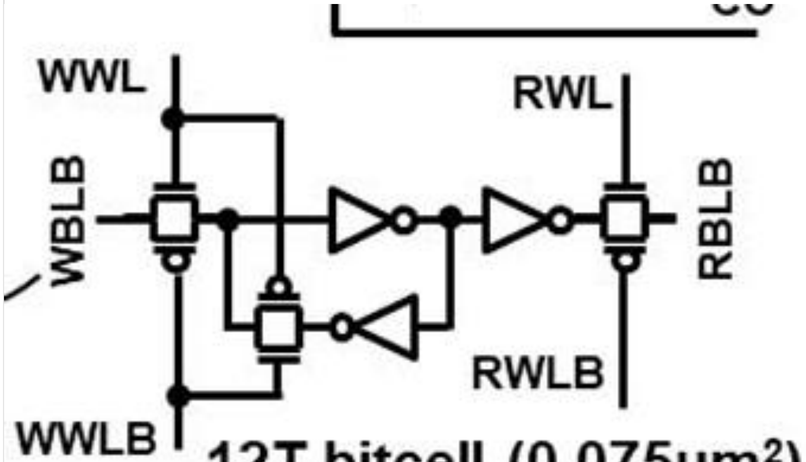
Features

- Allows simultaneous MAC and write operations
  - Time is wasted when updating weights
- Small area
- Low power
- Flexible bit width
  - 4b, 8b, … weight and input

Structure

- 64 MAC arrays with 64kb of 12T sram cell
- 64 banks, each bank contains 4b×4b
- MAC operation requires 5 clock cycles

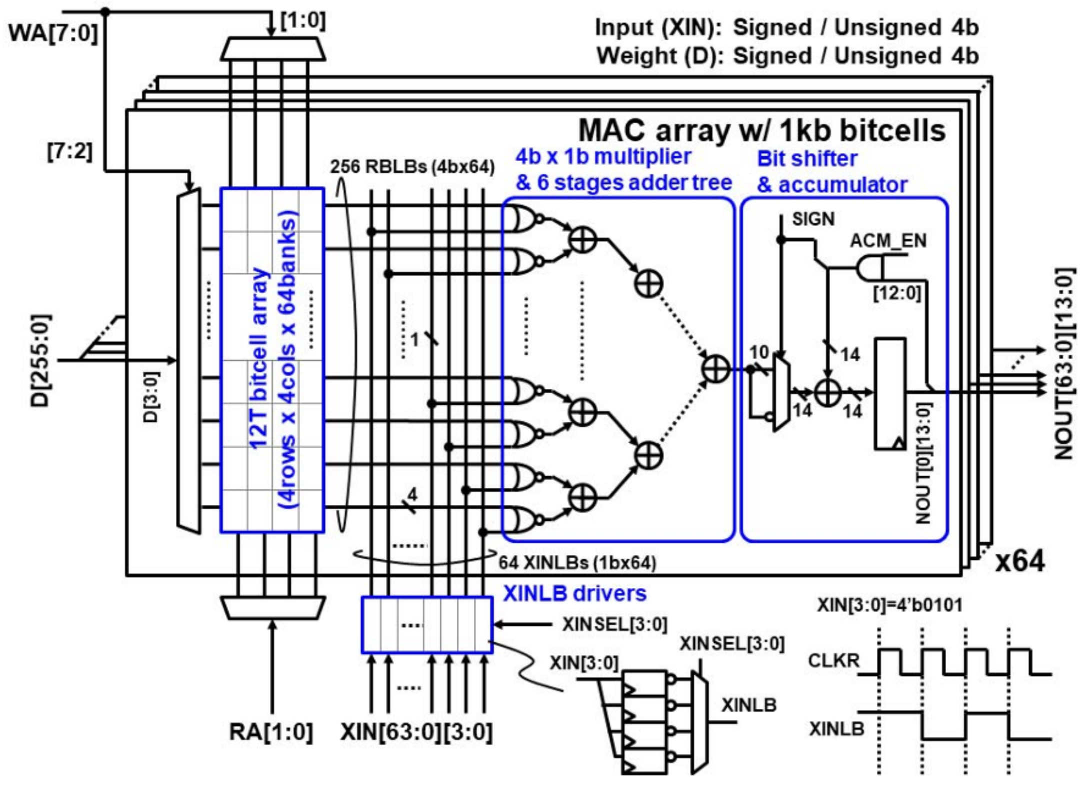


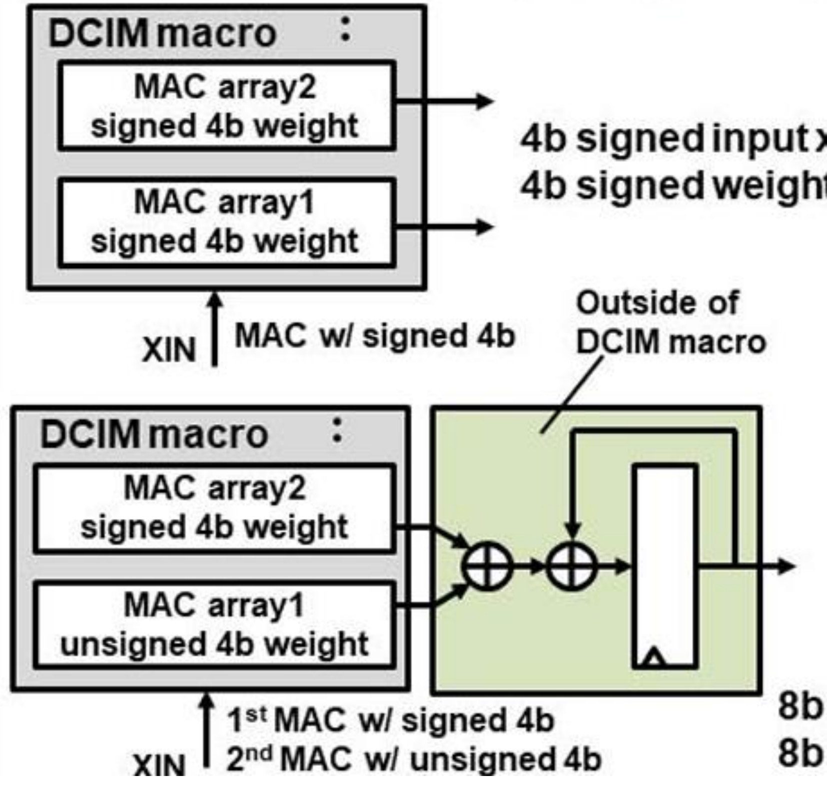


12T bitcell (0.075um^2)

Structure

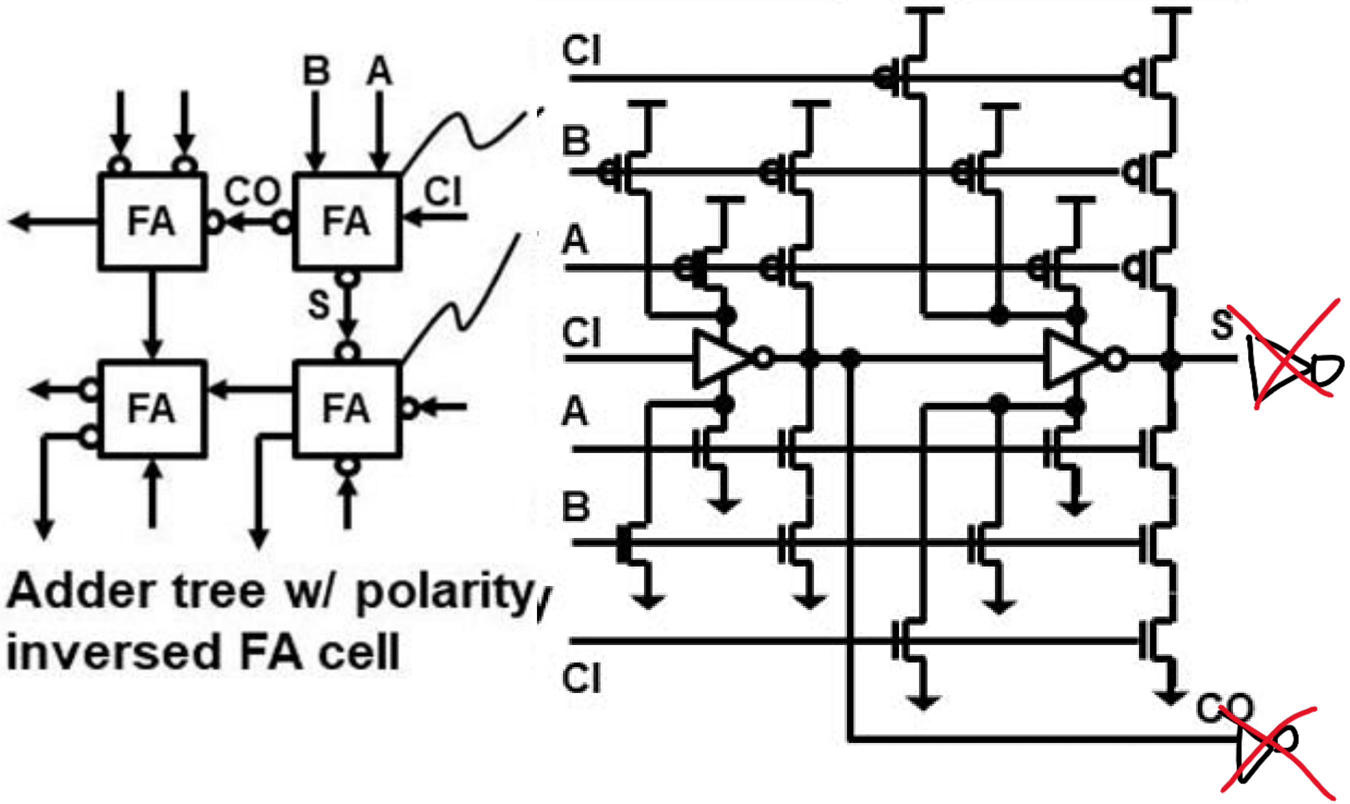
- Read operations reads data from the cell to NOUT
- Accumulator/Adder does sign extension if needed
  - Can support 8b data or more





Structure

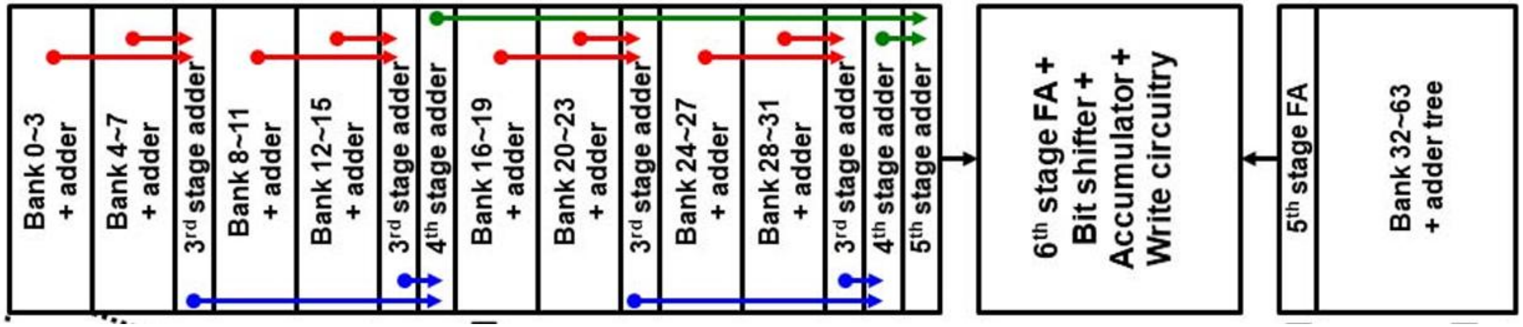
- Inverse FA
  - 12.5% smaller area, 15 % less total MAC power
  - FA($\overline{A}$, $\overline{B}$, $\overline{Cin}$) = ($\overline{S}$, $\overline{Cout}$)



Adder tree w/ polarity inversed FA cell

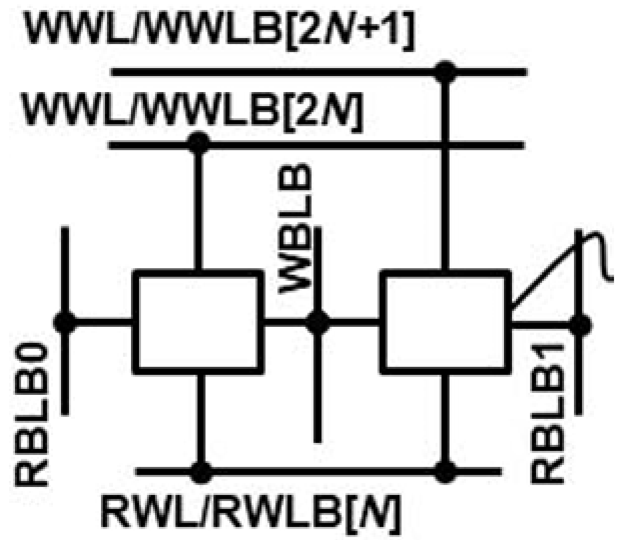
Structure

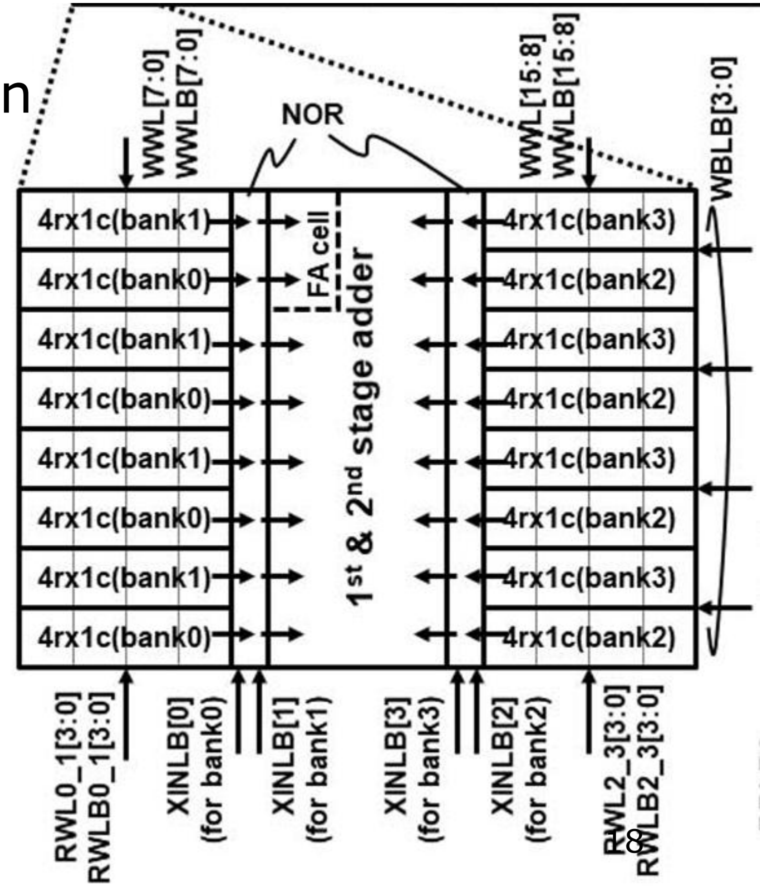
- Place adder closer to sram cell
  - Easier for routing
  - Shorter wire
- Merge 12T sram with logic part
  - No dummy space for transition region between sram/add
  - Space can be less than 6T sram design



Structure

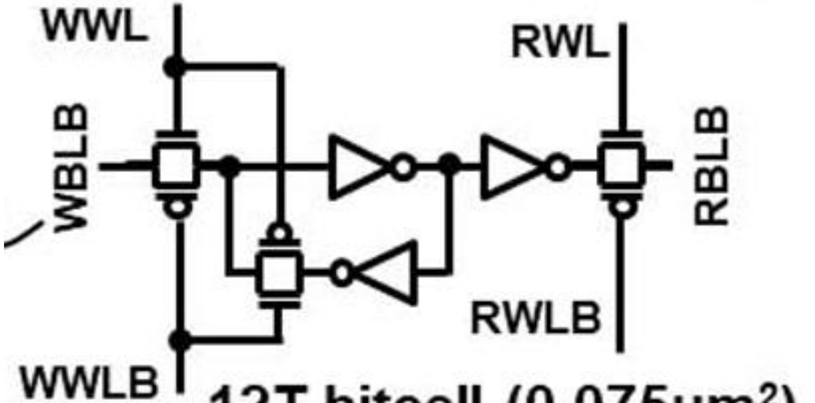
- Interleaved write WL
  - Easier for routing in x-direction
- Shared Read WL
  - Easier for routing in y-direction





Structure

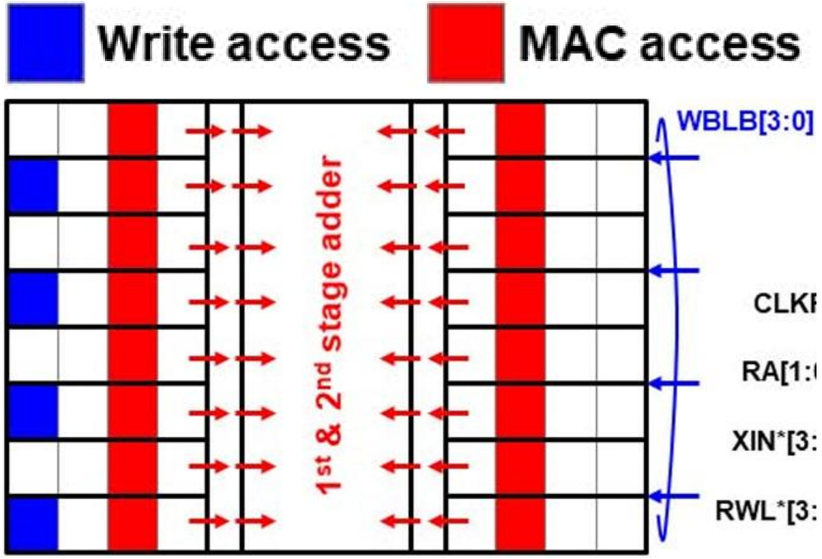
- No RBL precharge
  - CMOS output of 12 T sram does not need precharge
  - Saves power

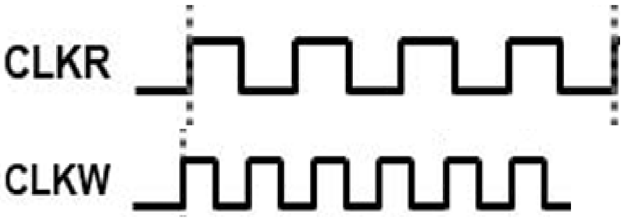


12T bitcell (0.075um^2)
w/ interleaved WWL

Structure

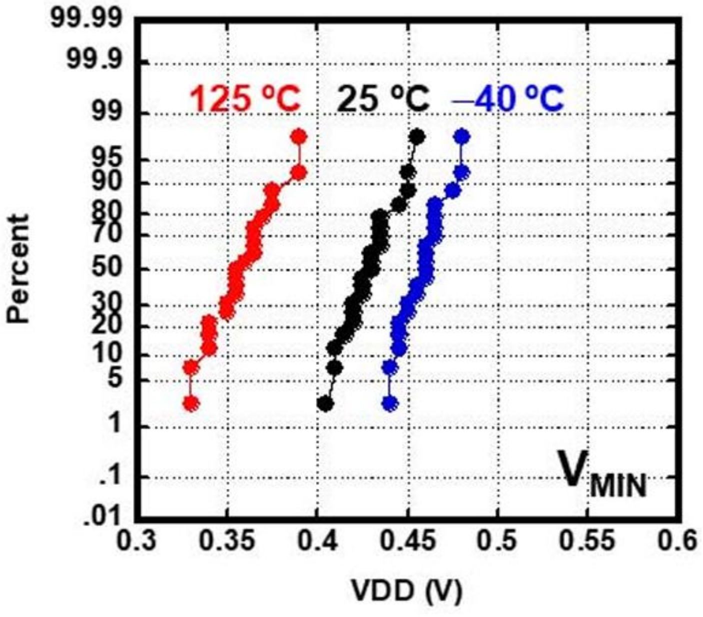
- Update weight when MAC is not using the cell
  - Time of updating can be faster than MAC if faster clock for write operation is used

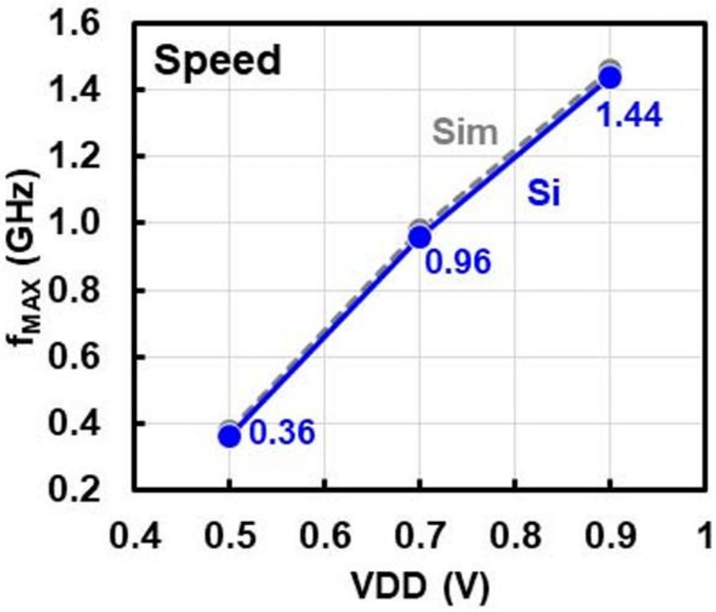




Result

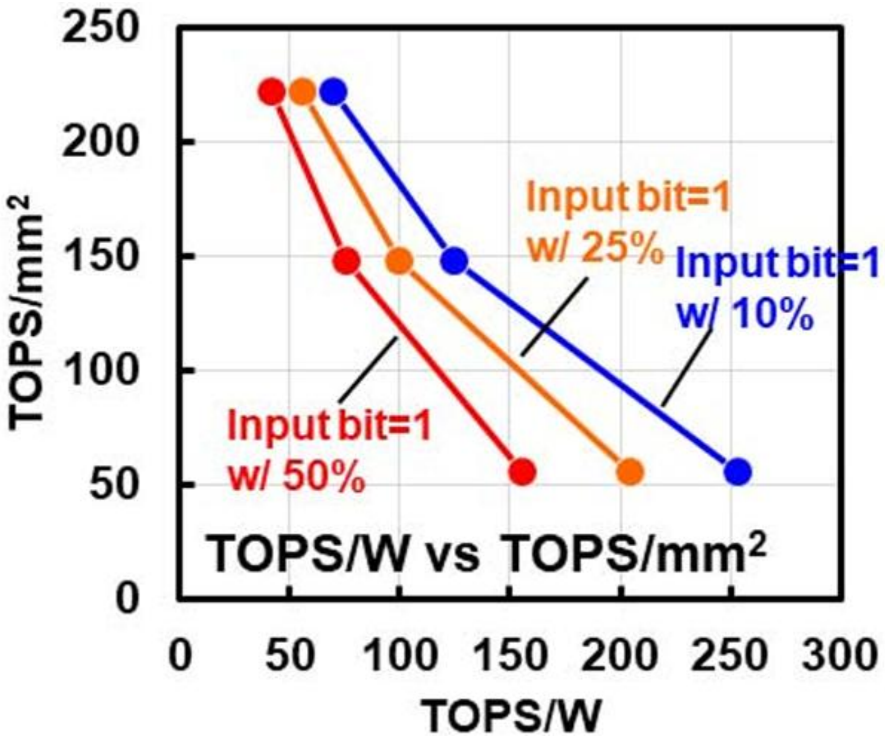
- Area: 12208 µm2
  - 3× smaller than using a similar digital architecture
- Vmin: 0.5V at −40°C with a 95% yield.
- Fmax: 0.36, 0.96, 1.44GHz at 0.5, 0.7, 0.9V





Result

- TOPS/W vs TOPS/mm2 chart:
  - VDD = 0.5 V ~ 0.9V
  - 253.5, 205.0 and 155.2TOPS/W with 10, 25 and 50% input bit sparsity under 0.5 V



Conclusion

- Allows simultaneous MAC and write operations
  - Update weight while not used for MAC
  - Use a faster clock for updating
- Small area
  - 12T sram is better than 6T sram
  - Inversed FA
  - Shared WL/ interleaved WL
- Low power
  - 12T sram does not need BL precharge
  - Inversed FA
- Flexible bit width
  - Additional shift and add control