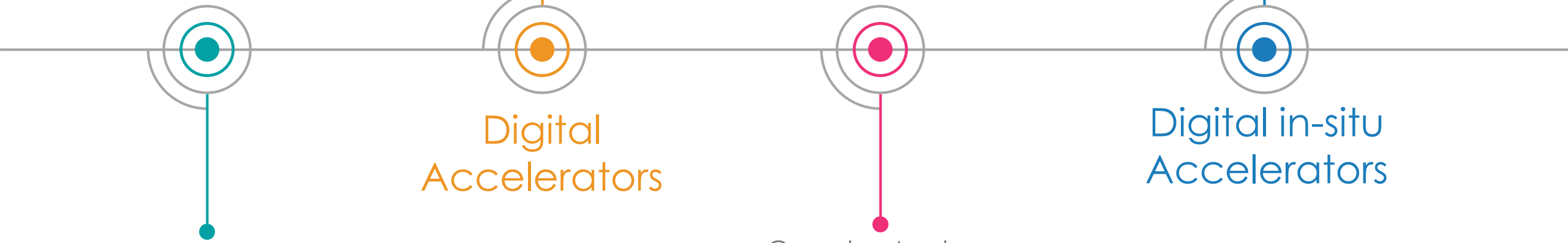
Moving CNN Accelerator Computations Closer to Data

Sumanth Gudaparthi
Surya Narayanan
Rajeev Balasubramonian

Evolution of CNN Accelerators



SRAM based In-Situ Computation Accelerator

DA vs SISCA

Perform Computations In-Situ

AIA vs SISCA

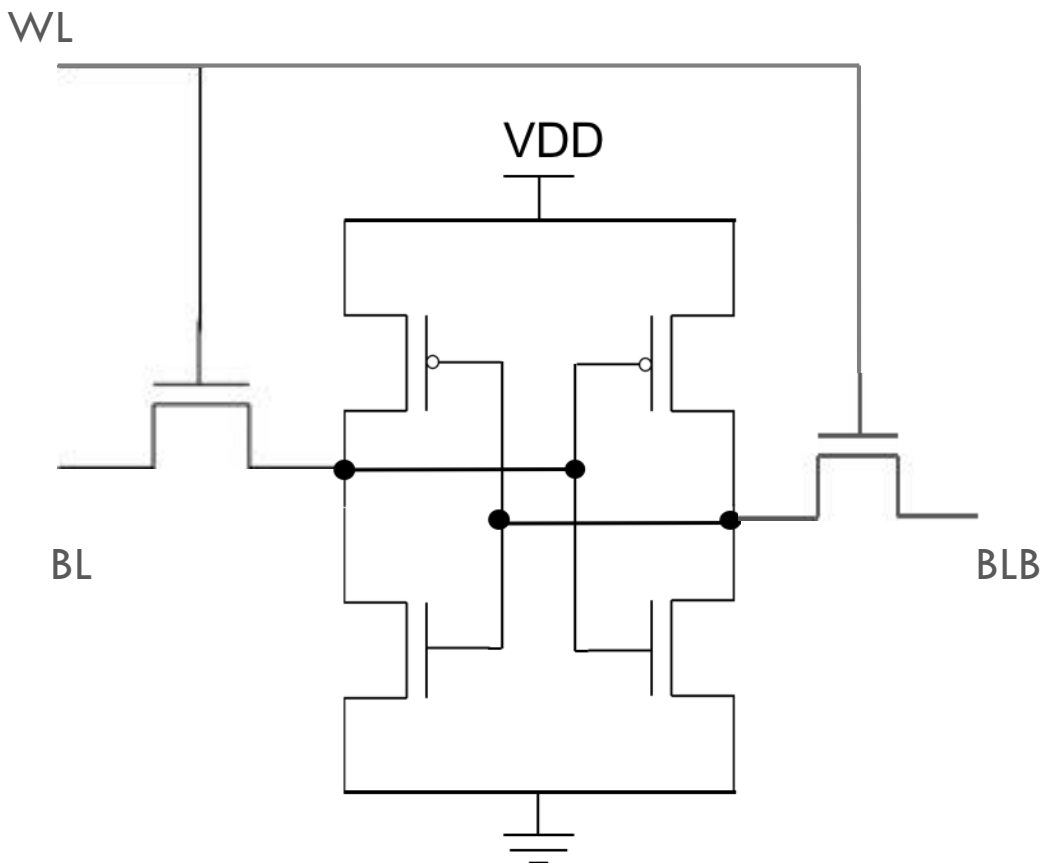
Use SRAM cells to perform In-Situ Computations

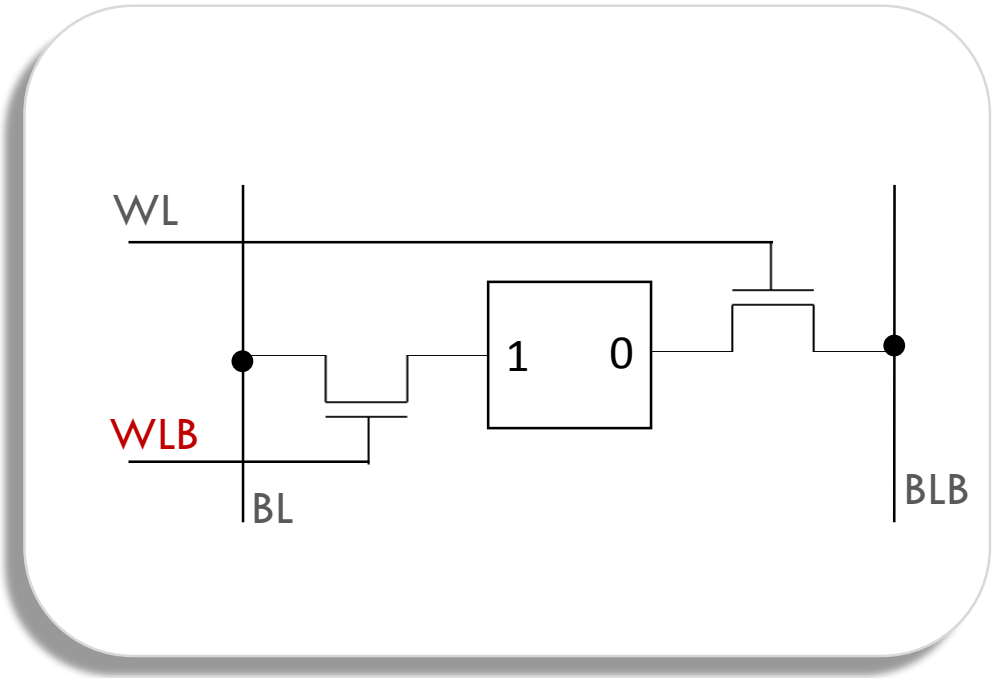
DIA vs SISCA

Modify the LLC. Trivial overhead on baseline Cache operations

SISCA: Proposed Accelerator DA: Digital Accelerators AIA: Analog In-situ Accelerators DIA: Digital In-situ Accelerators

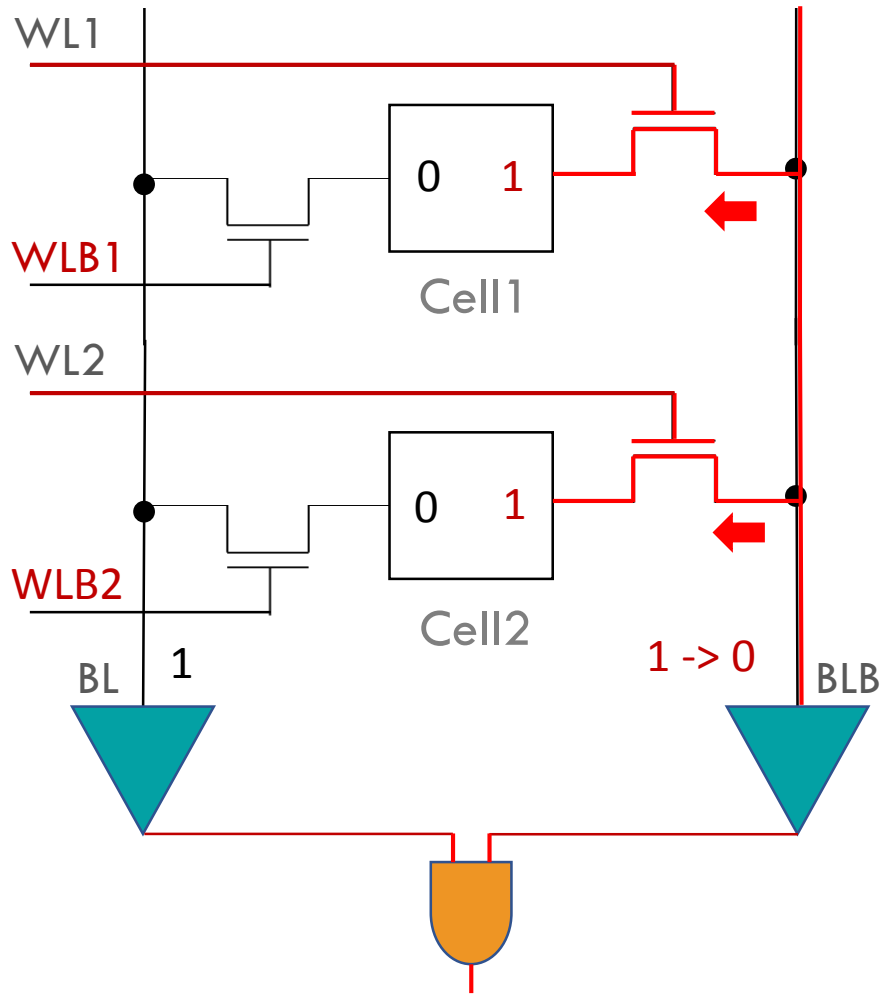
Logic-In-Memory





Jeloka et al., 2016

Logic-In-Memory



Pre-charge the bit-lines

Activate the word-lines

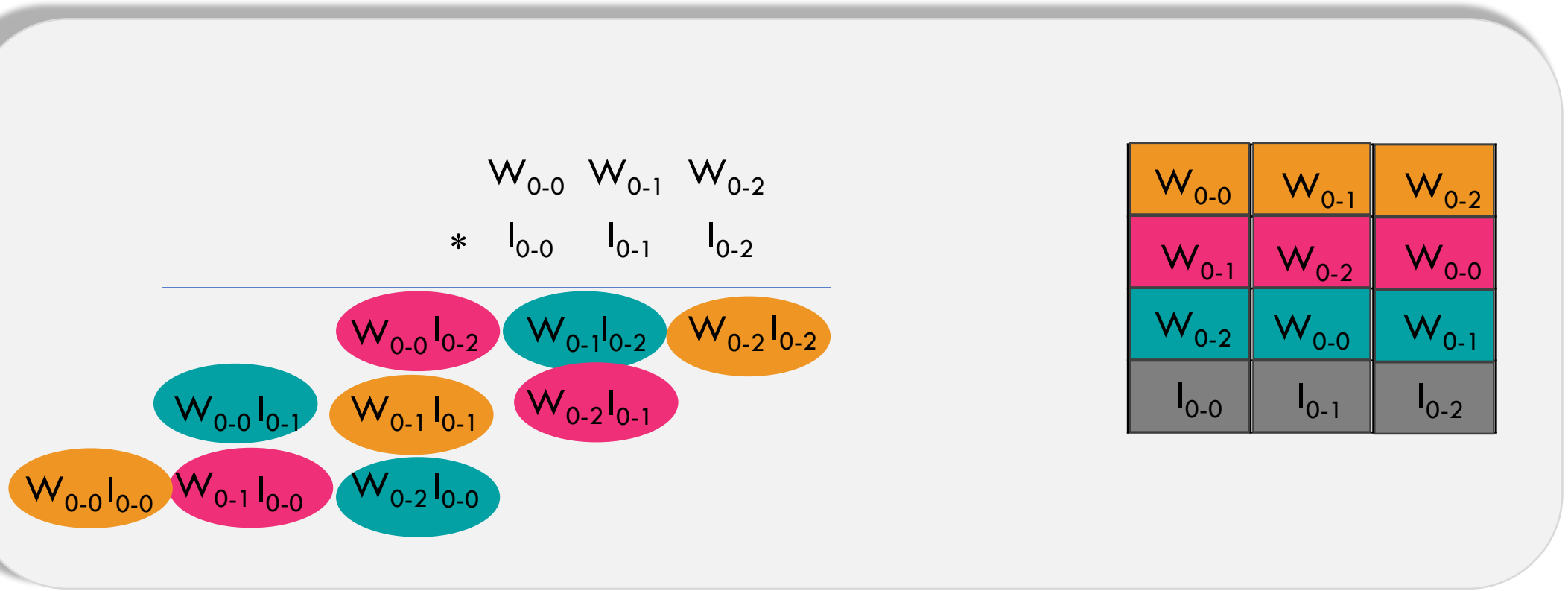
Discharge of bit-line voltage through Cell1

Discharge of bit-line voltage through both Cells

Bit-line stays Pre-charged

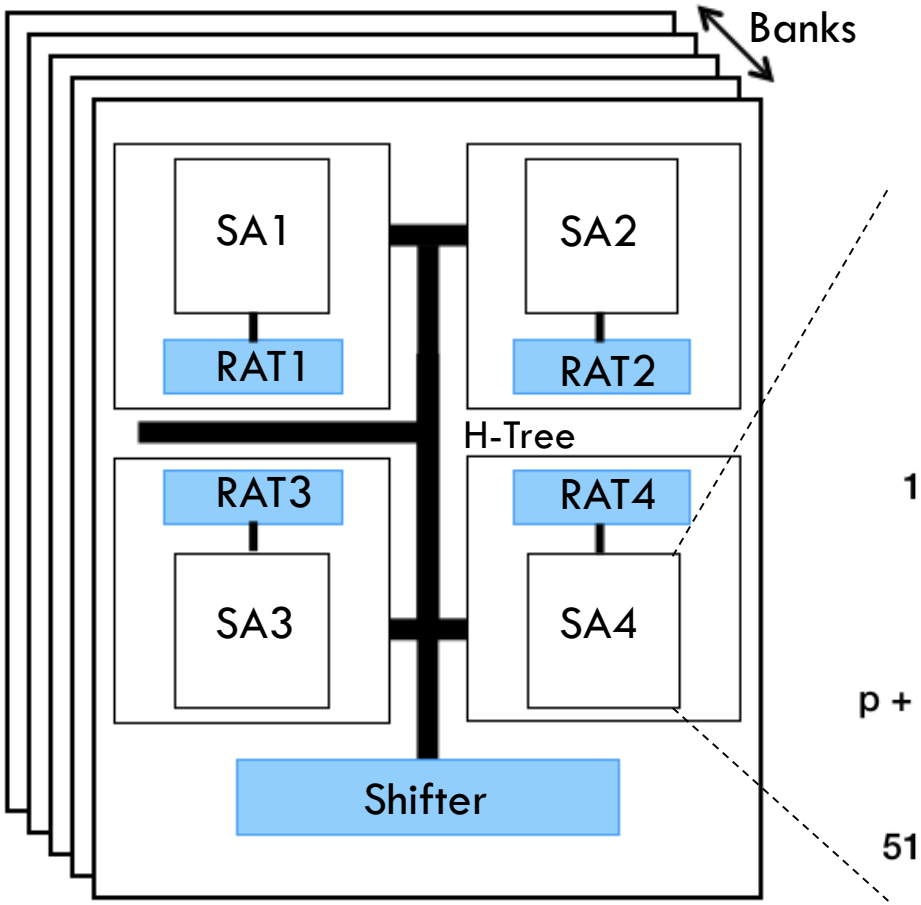
Jeloka et al., 2016

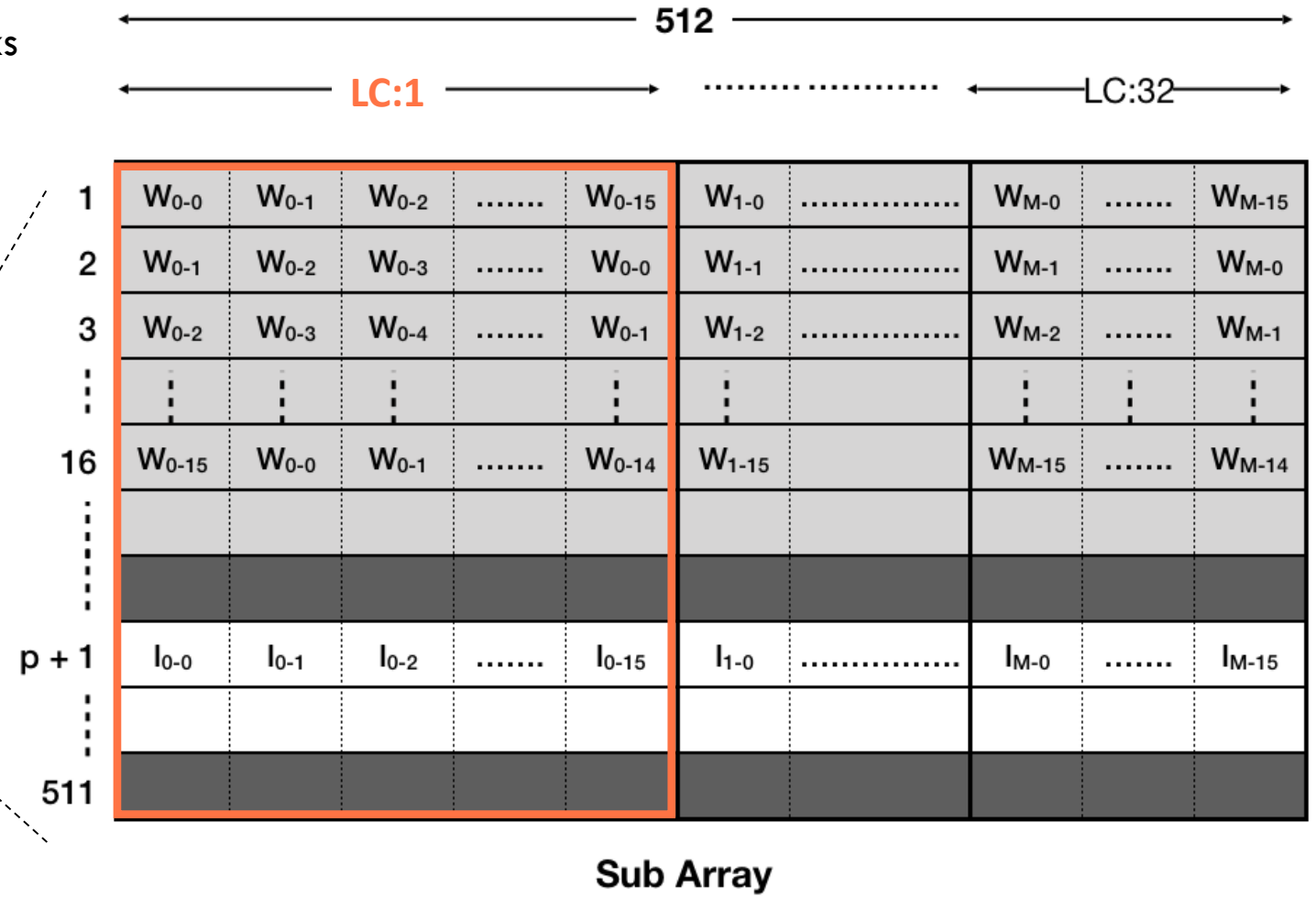
Enabling In-Situ Multiplication in Caches



$C_{a\text{-}b}$: bit number-b in a^{th} variable of C

SISCA Organization





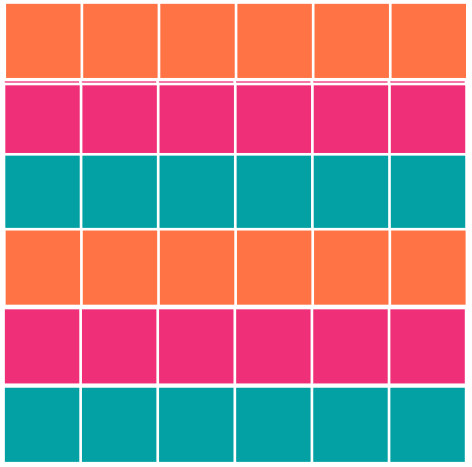
$C_{a\text{-}b}$: bit number-b in a^{th} variable of C

Unused Sub-array Entries

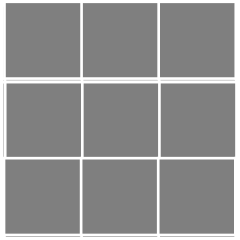
Kernel Entries

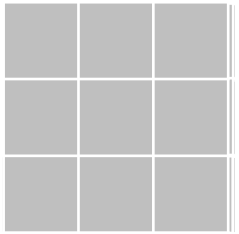
Feature Map Entries

SISCA Dataflow

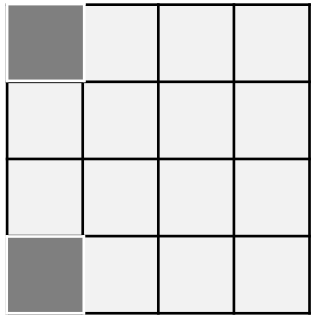


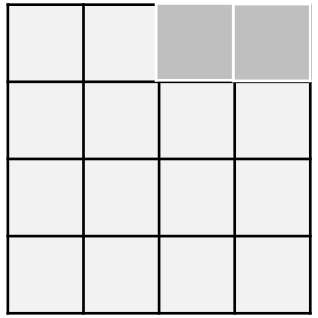
Input Feature Map
(6x6)



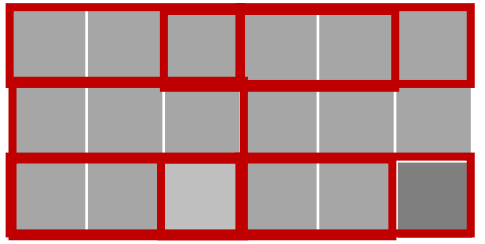


Kernel Maps
2x(3x3)

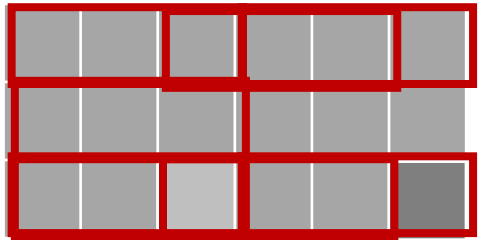




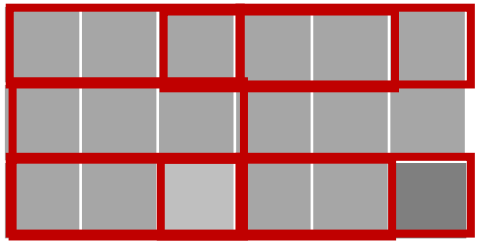
Output Feature Maps
2x(4x4)



Sub Array 1

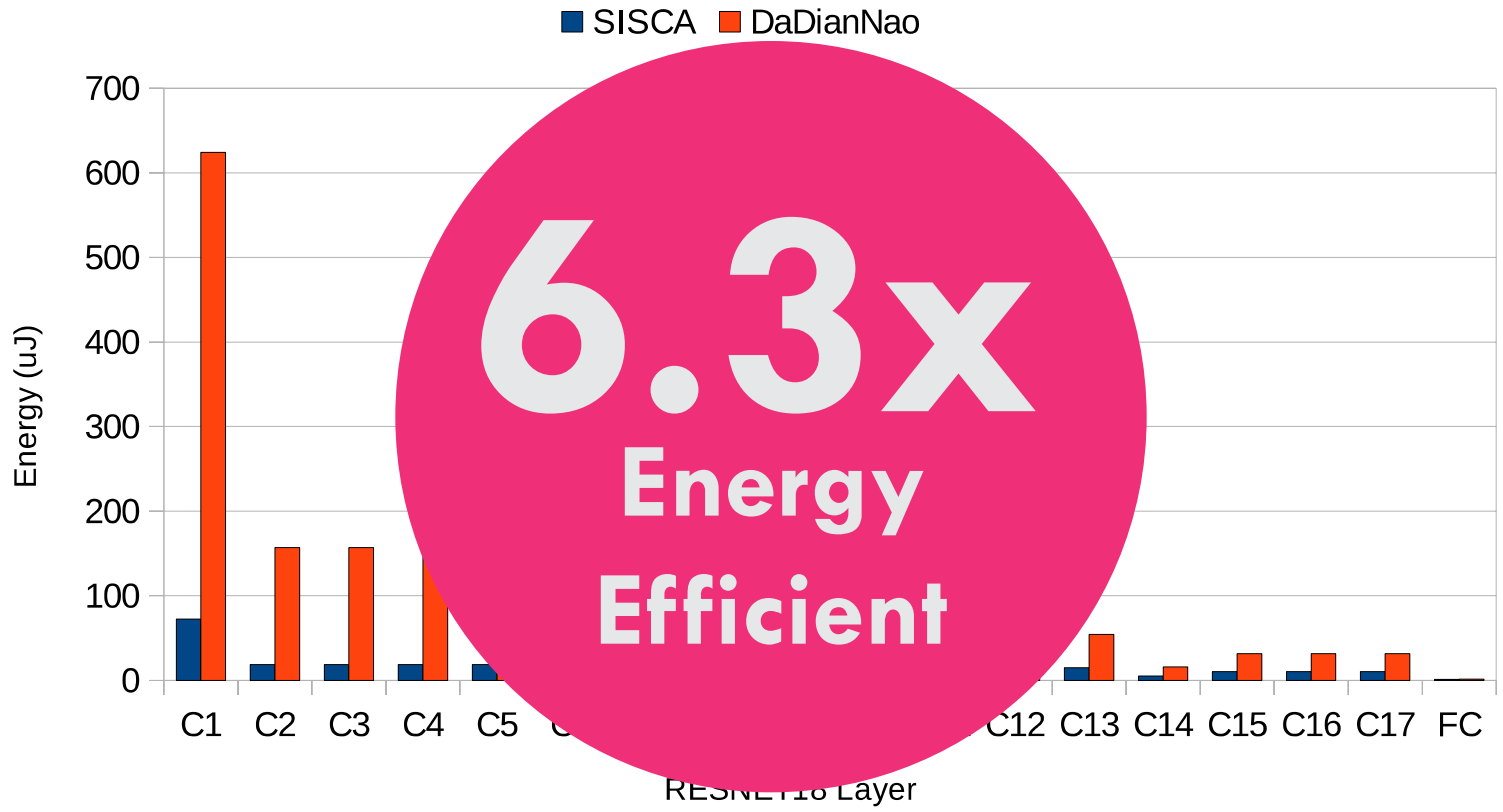


Sub Array 2

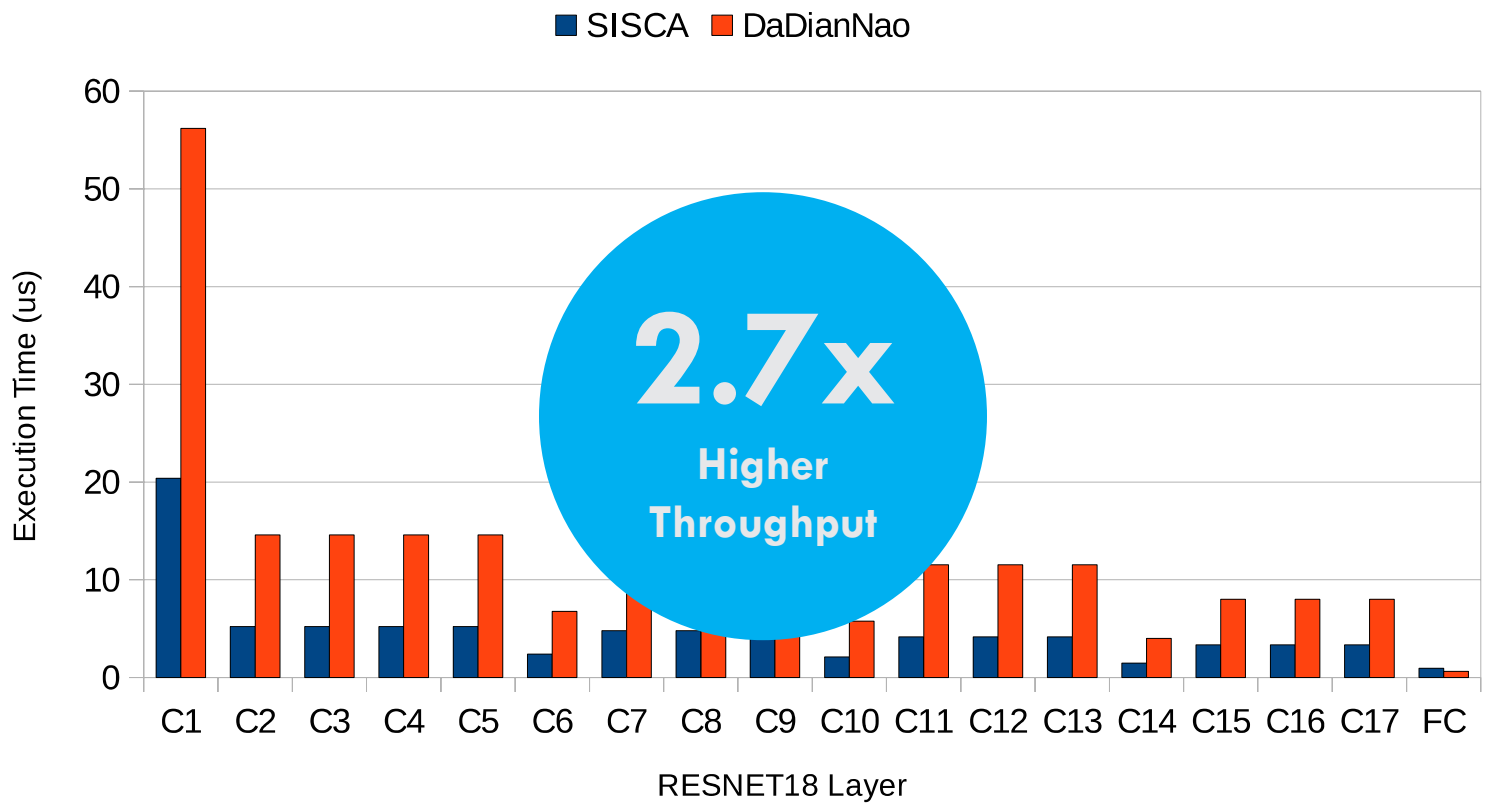


Sub Array 3

Energy Improvements



Performance Improvements



Conclusions and Future Work

- SISCA is an SRAM in-situ computation Engine for Convolution Neural Networks
- Uses on-chip Last Level Cache (LLC) to perform computations
- SISCA is 6.3x Energy efficient, and has 2.7x higher throughput than DaDianNao
- Better dataflow and mapping mechanisms can further improve the Energy and Throughput.
- Need to work on better scheduling mechanisms to distribute the general purpose workload, and CNN data across the Cache.

Questions?