Low Power FPGA and ASIC Focussed Convolutional Neural Networks

James Garland

SFI Project 12/IA/1381
Trinity College Dublin

2016

Agenda

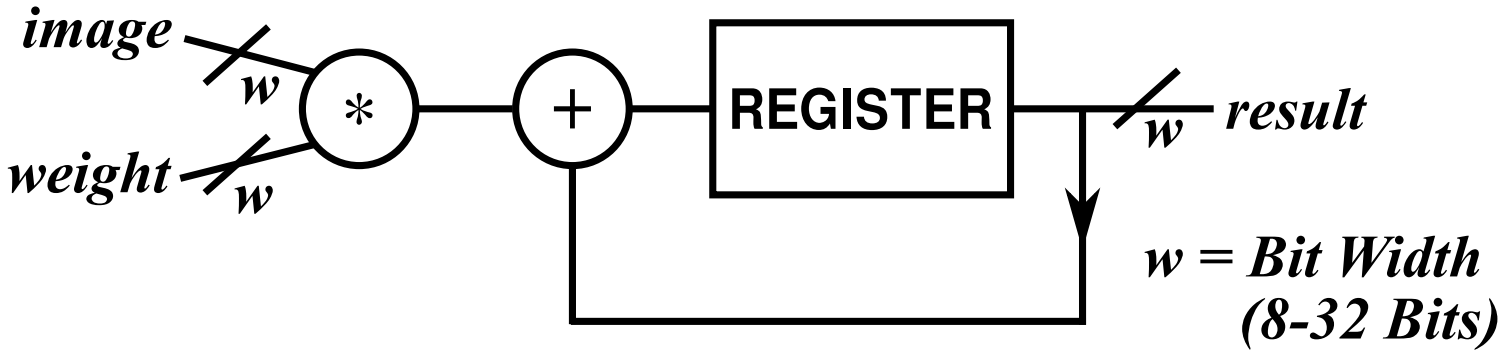
- About me
- My Present Research
- Next Steps in My Research

About Me

- I come from Birmingham, UK.
  - I'm a Brummie, not Australian!
- M.Sc. in System Design (Microelectronic) (Birmingham City University, 1995)
- 21 years industrial experience in embedded hardware design and test in various industries.
  - 10 in ASIC design.
  - 11 in FPGA design and test.
- 8 years part time lecturing experience.
- 2 years research experience in Trinity College Dublin.
- Started PhD research in March 2016.
  - Low power machine learning / training in hardware.

Background Research

- CNNs have lots (100,000's or more) multiply and accumlates.

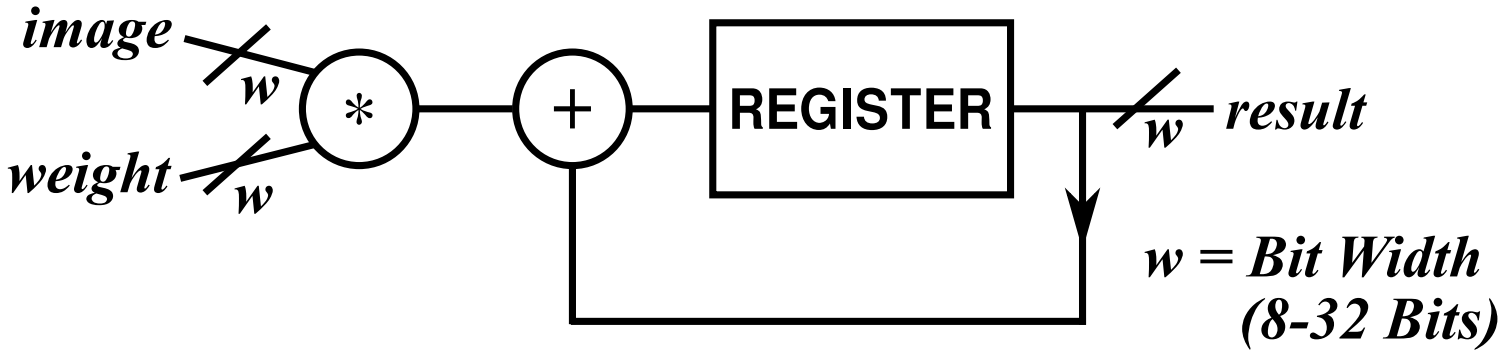


- Hardware Multipliers are large and power hungry.

- Existing approach show relatively few **weight** multiplicand values are needed.

Background Research

- CNNs have lots (100,000's or more) multiply and accumlates.

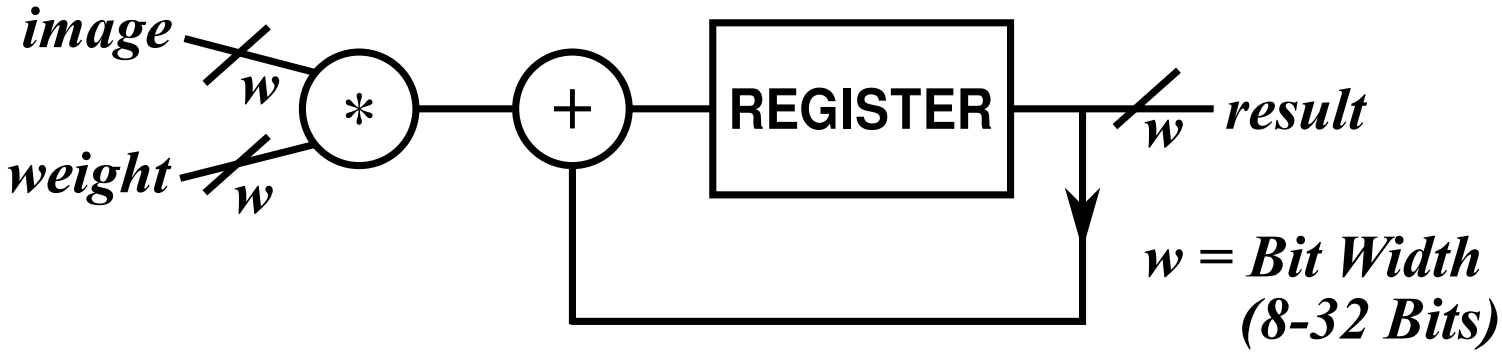


- Hardware Multipliers are large and power hungry.

- Existing approach show relatively few **weight** multiplicand values are needed.
- We Rewrite the multiply accumulate into a series of accumulators followed by a multiplier.

Background Research

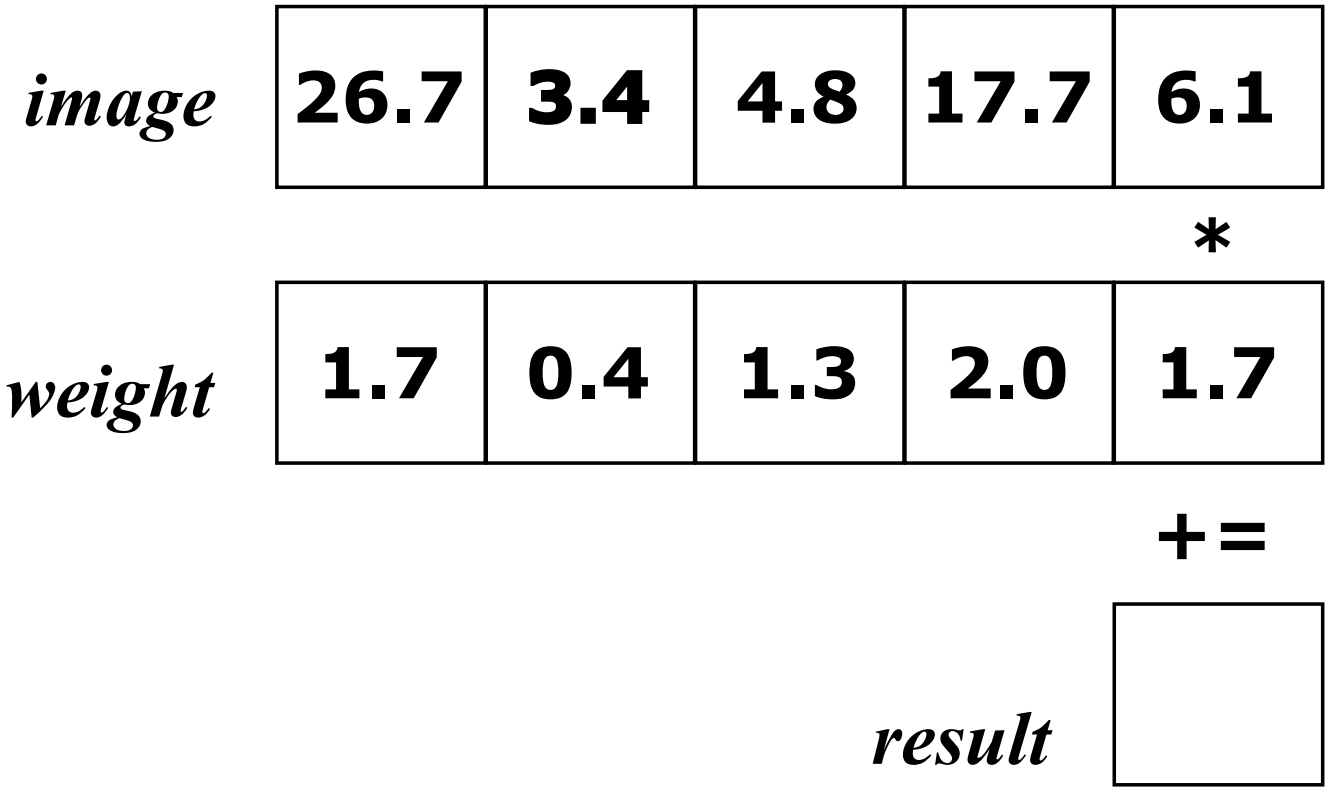
- CNNs have lots (100,000's or more) multiply and accumlates.



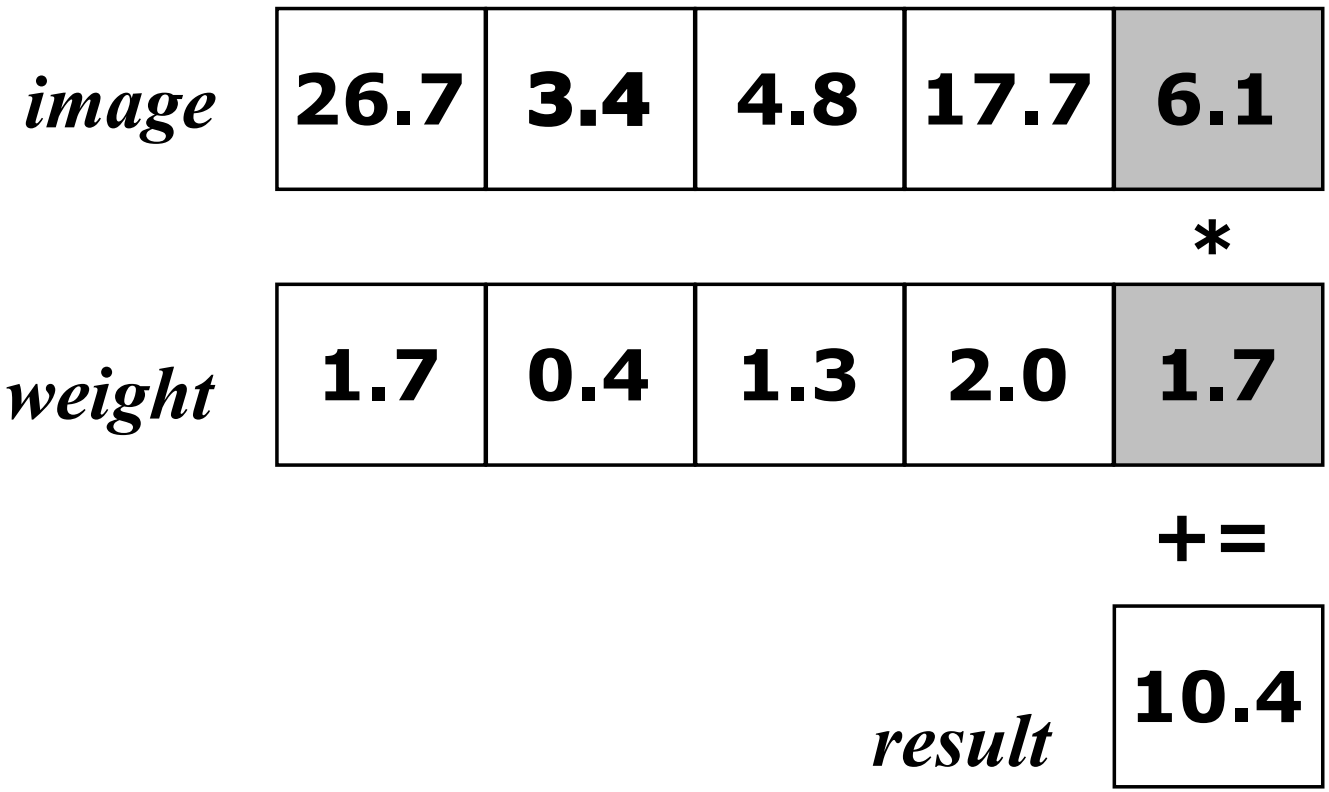
- Hardware Multipliers are large and power hungry.

- Existing approach show relatively few **weight** multiplicand values are needed.
- We Rewrite the multiply accumulate into a series of accumulators followed by a multiplier.
- Lower power and smaller area in hardware.

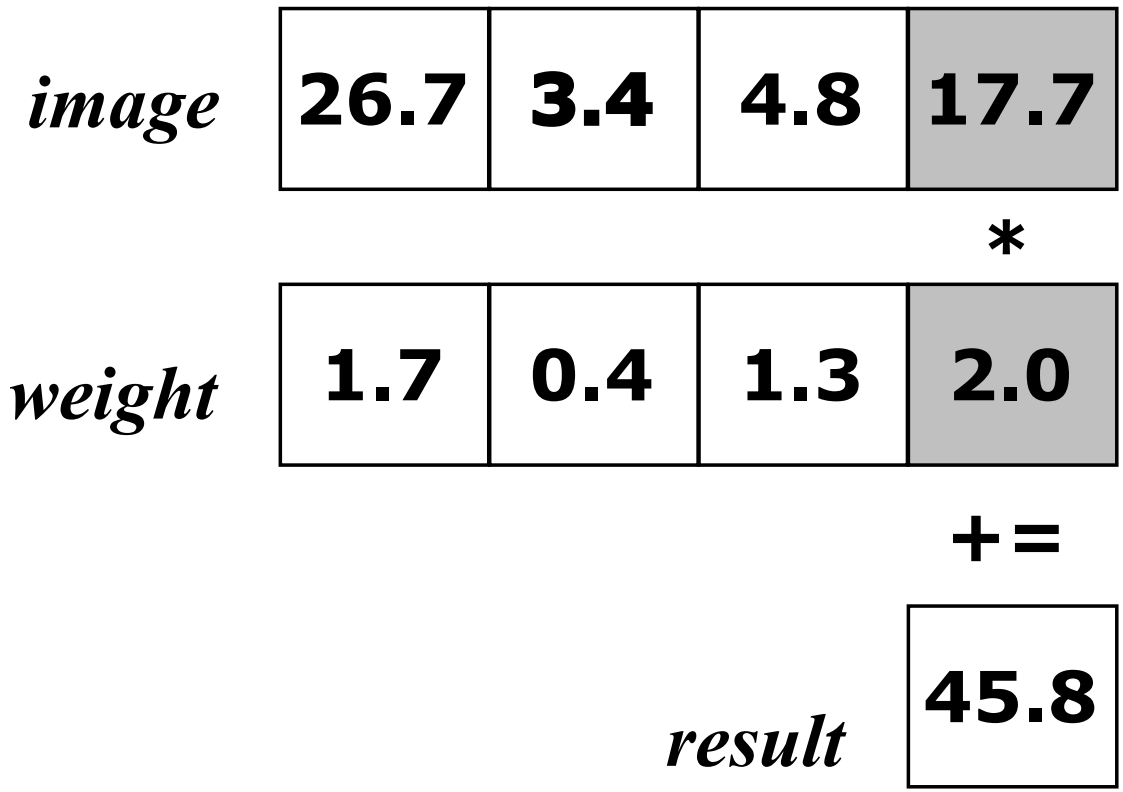
How a Multiply Accumulate (MAC) Works



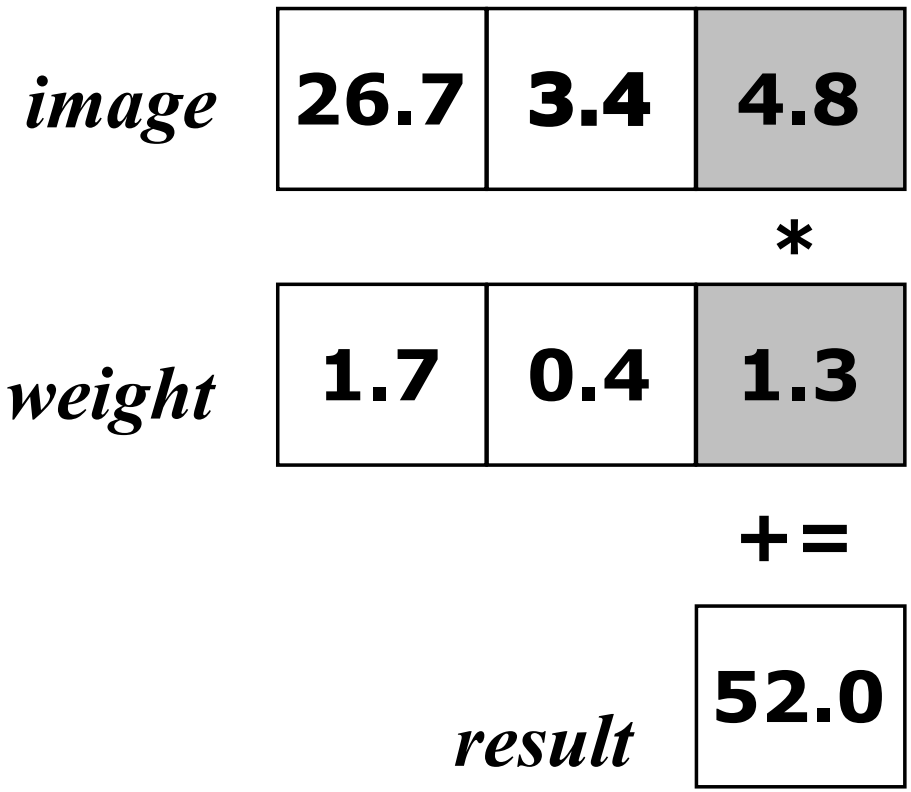
How a MAC Works



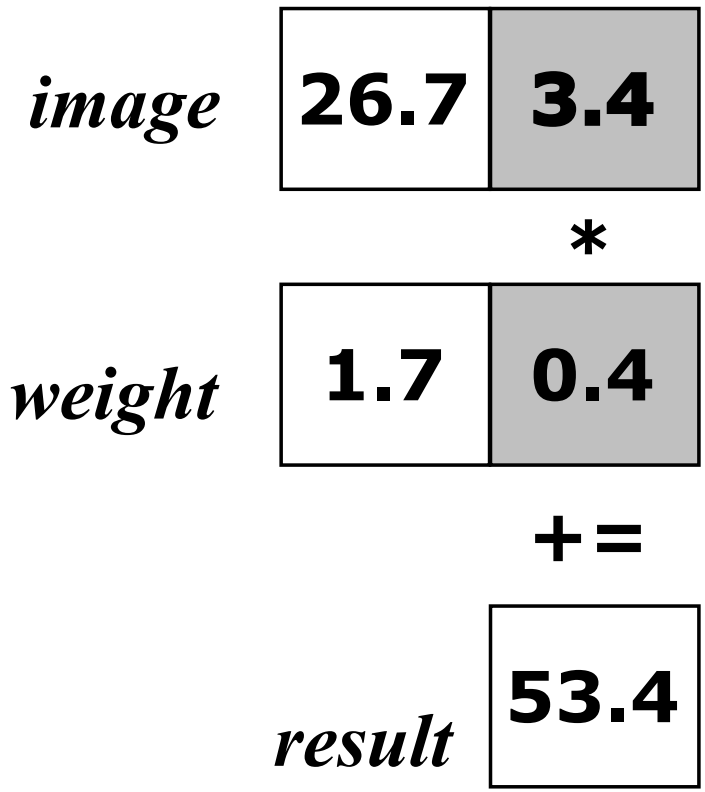
How a MAC Works



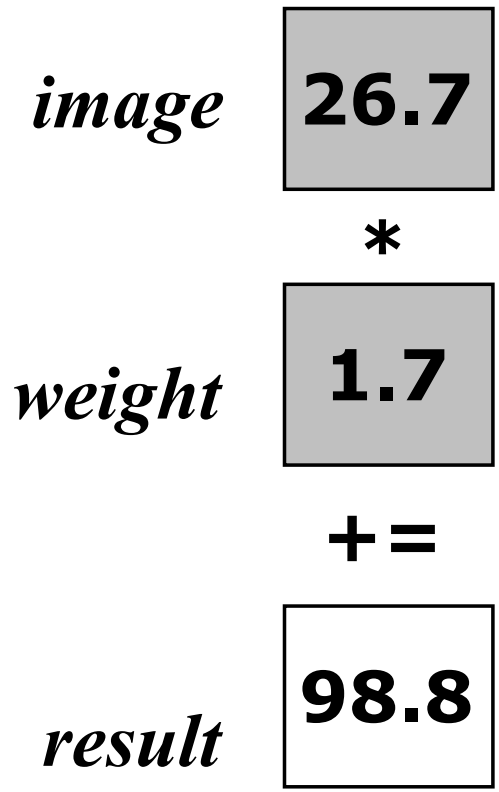
How a MAC Works



How a MAC Works

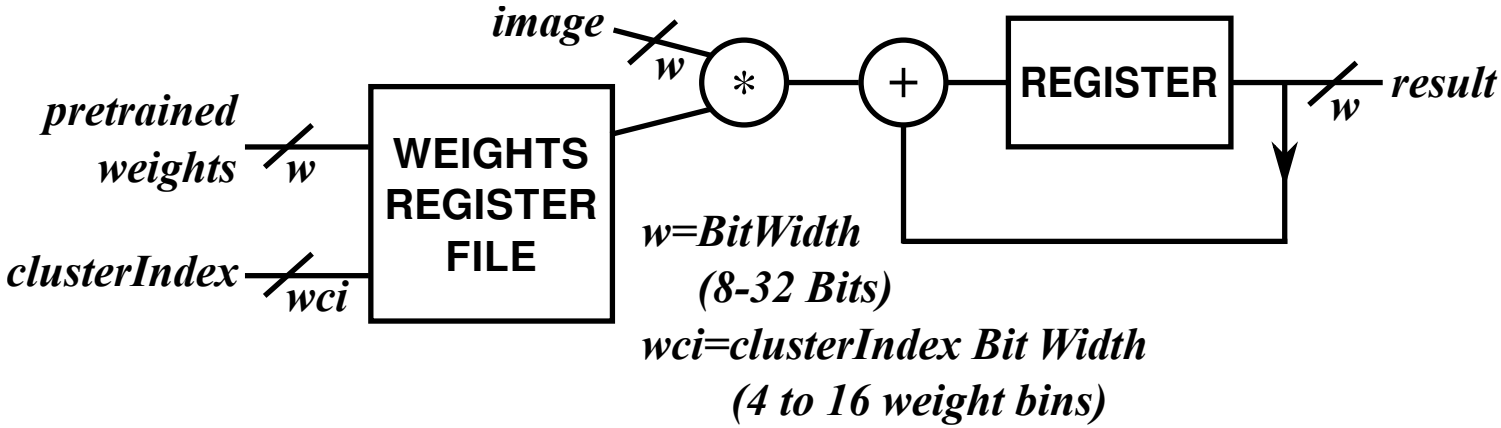


How a MAC Works



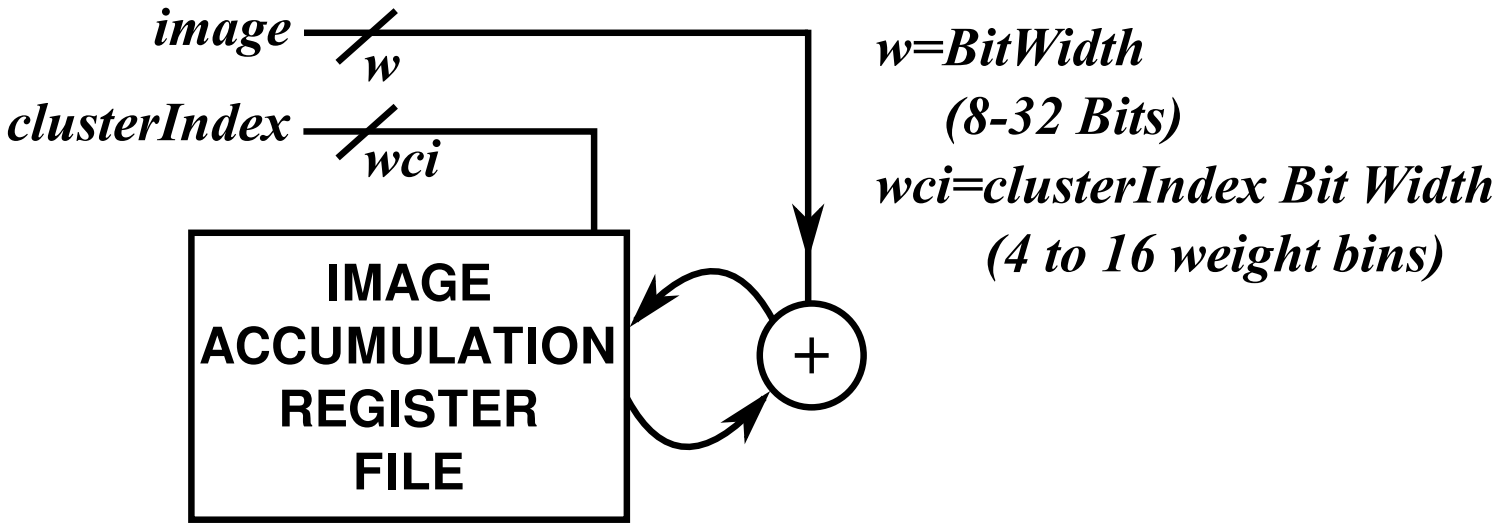
Weight Shared MAC

- Han *et al.* propose an accelerator architecture for implementing Convolutional Neural Networks (CNNs) with their weight-sharing scheme.
- Simple MAC - pre-trained **weight** values are stored in a weights register file.
- Values are indexed and retrieved by **clusterIndex** and multiplied by the corresponding **image** value.

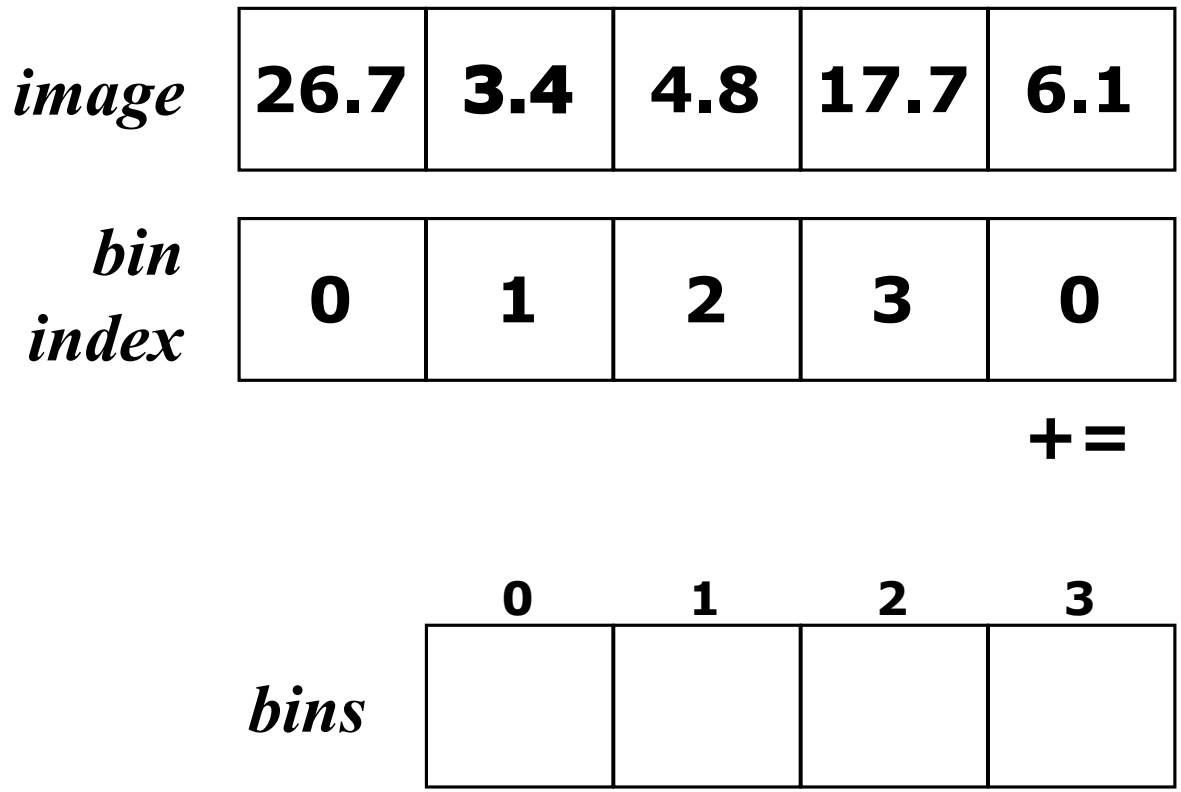


Proposed PAS

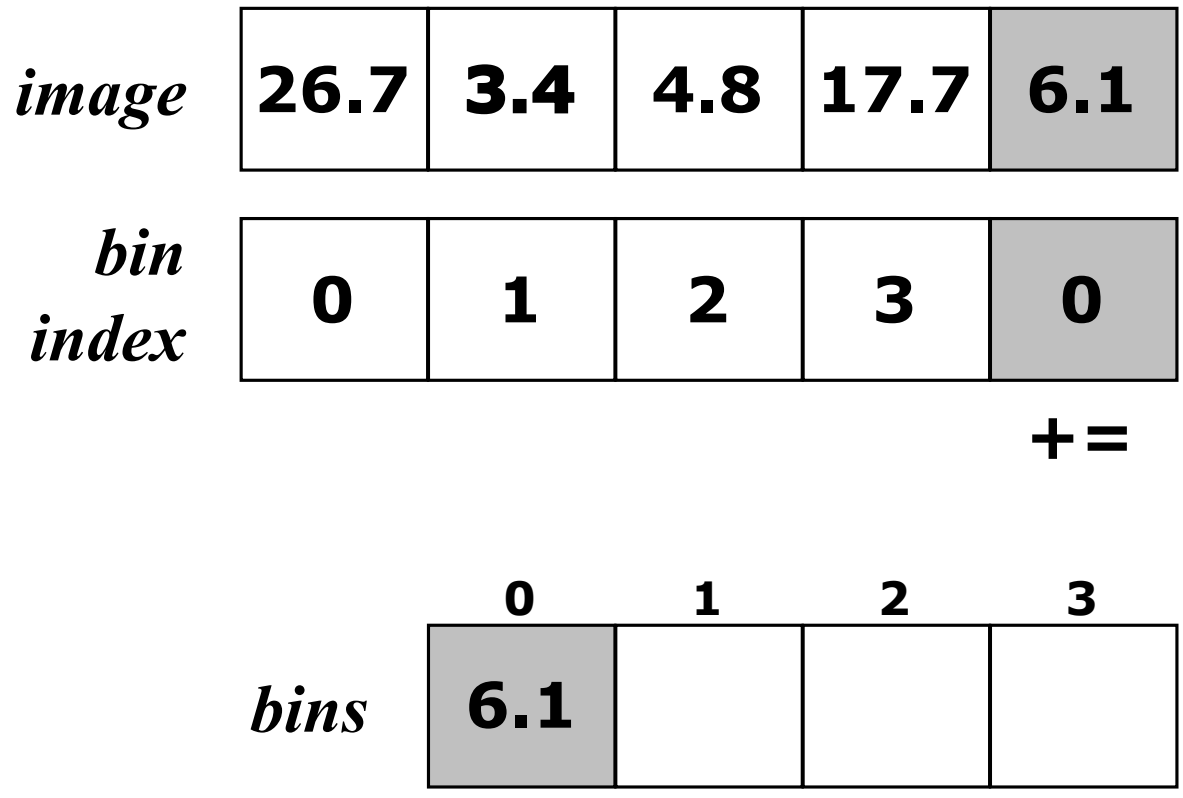
- We propose multiple Parallel Accumulate and Store (PAS) units followed by **one** shared MAC.
- A Parallel Accumulate and Store (PAS) accumulating an **image** of w bits wide indexed into the image accumulation register file with **clusterIndex** $b = 2^{wci}$ bins deep.



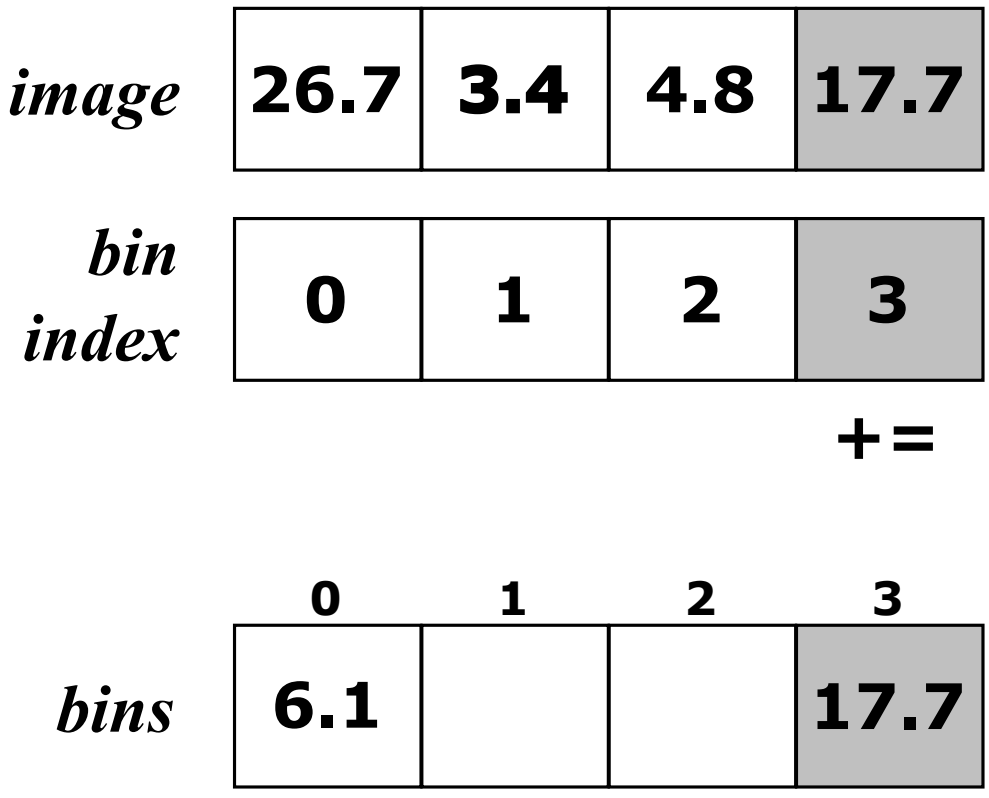
How the PAS Works



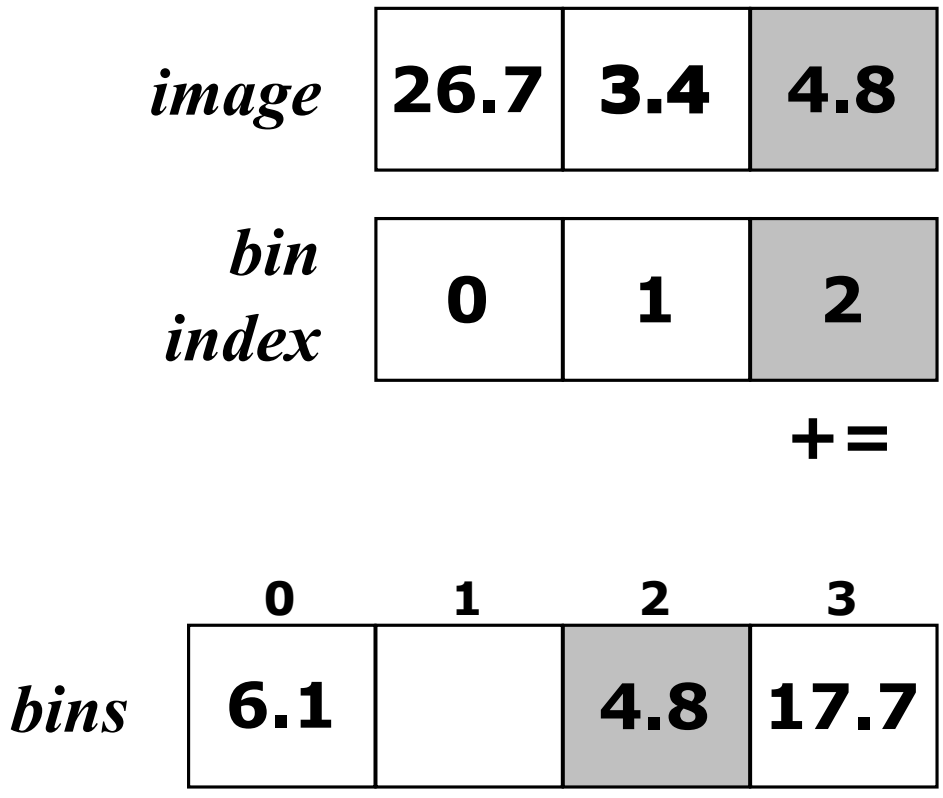
How the PAS Works



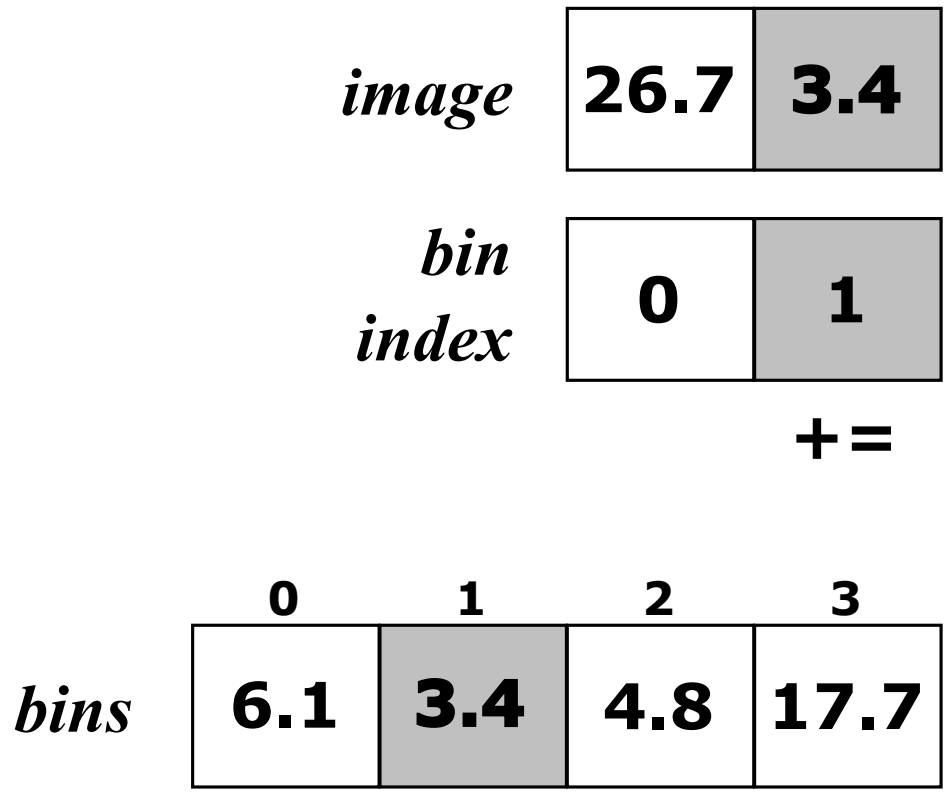
How the PAS Works



How the PAS Works



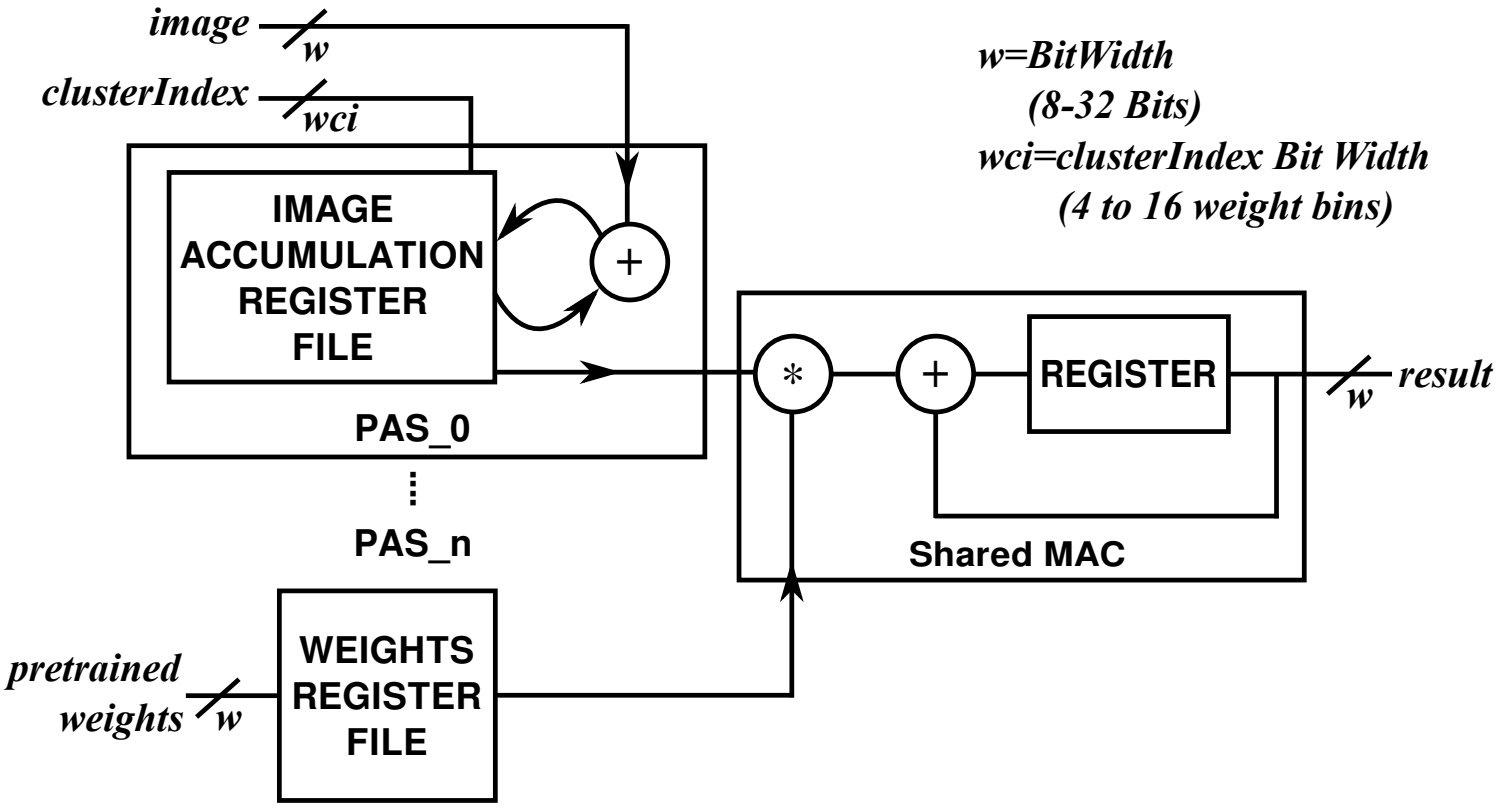
How the PAS Works



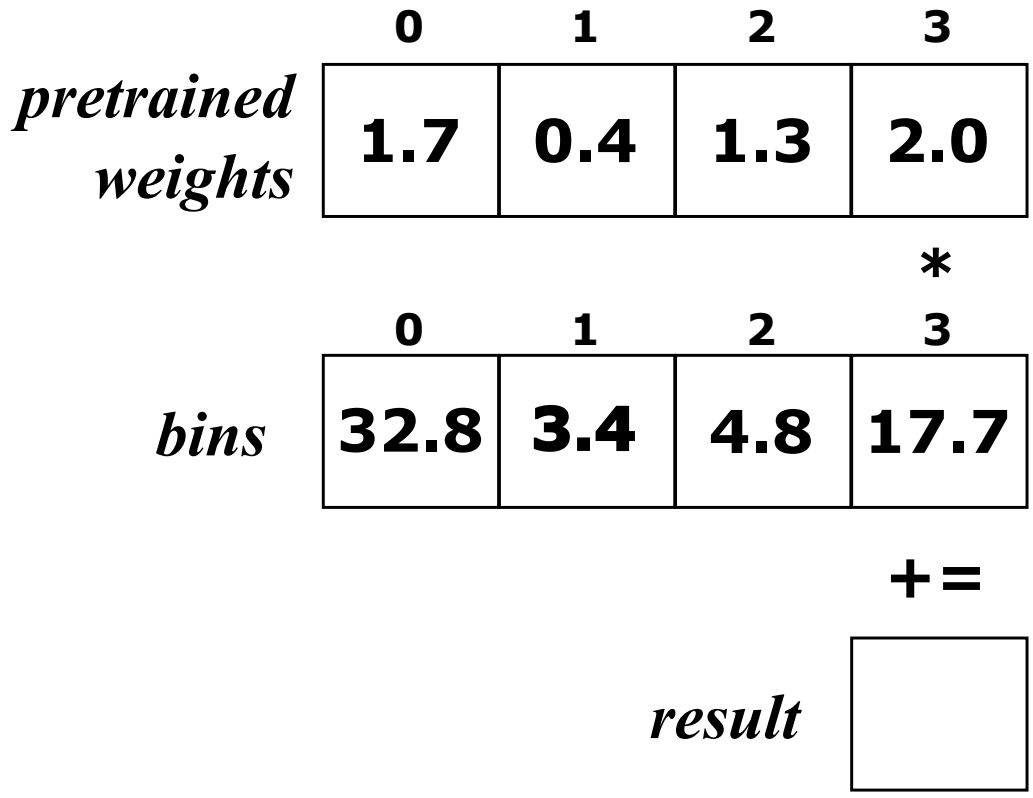
How the PAS Works

Proposed PASM

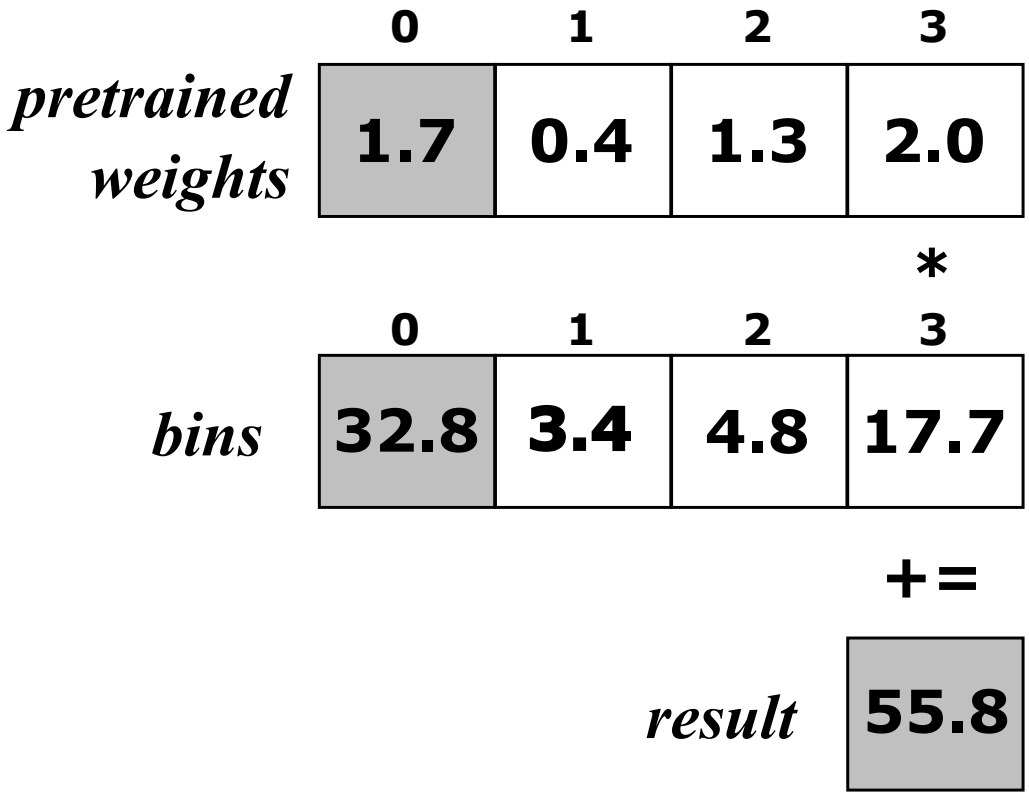
- Multiple PAS-Shared-MAC (Parallel Accumulate Shared MAC (PASM))
- The post-pass multiply-accumulate phase would multiply the weights with clustered image values indexed by **clusterIndex**.



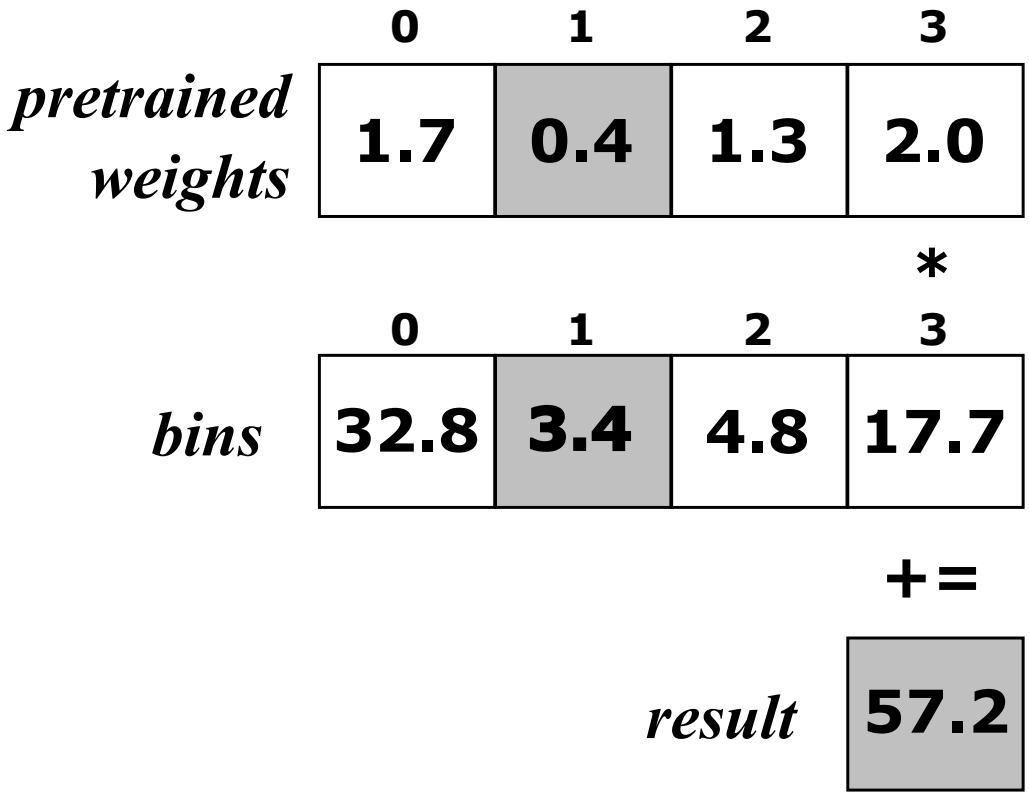
How the PASM Works



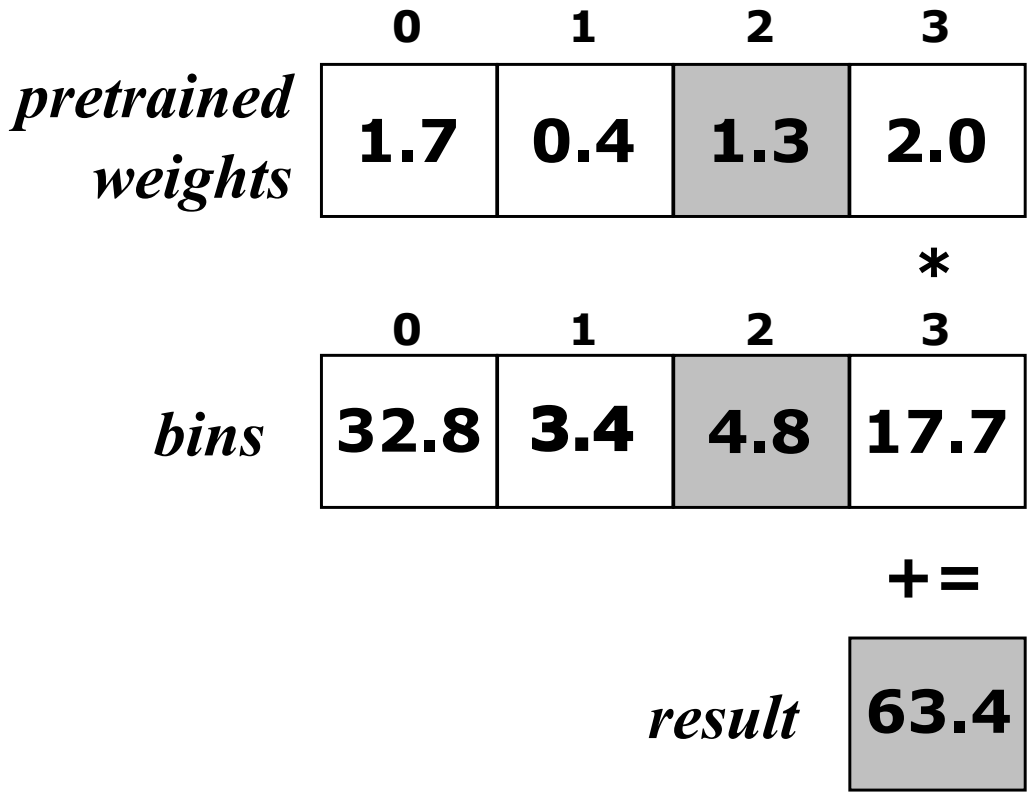
How the PASM Works



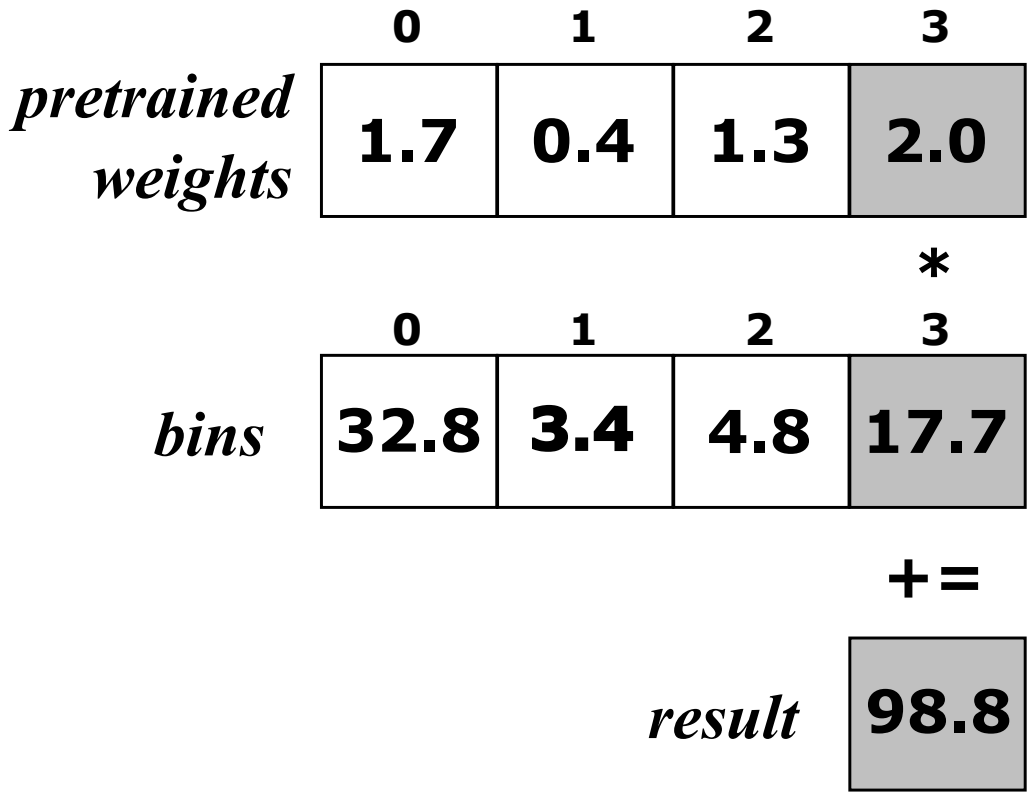
How the PASM Works



How the PASM Works

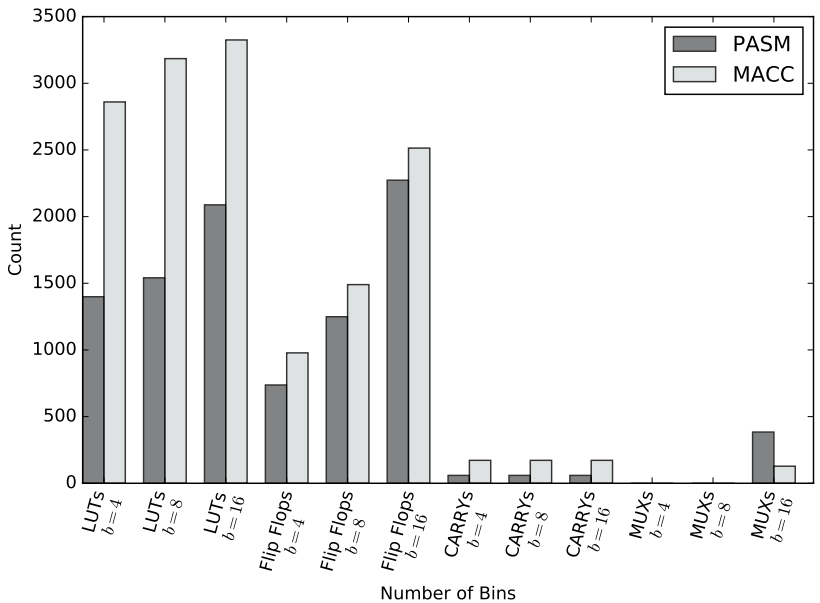


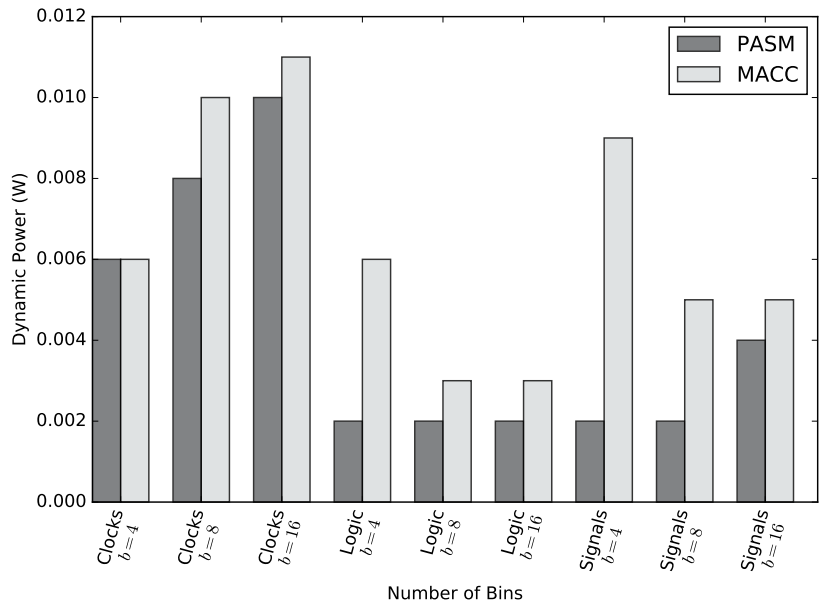
How the PASM Works



PASM - Results

- Utilization results show more efficient logic gate count for the Multiple PAS-Shared-MAC (PASM)
- Power consumption results show a lower power consumption for the Multiple PAS-Shared-MAC (PASM).





Paper Submitted to arXiv and IEEE CAL

- Short (4page) paper submitted for publication on 30 August 2016.
- Accepted on arXiv.
- Under review for publication on IEEE Computer Architecture Letters.
  - First round reviews encouraging.
  - Preparing major revision for resubmission in 6 weeks.

Immediate Future Research

- How can differing data types of weights affect efficiency?
- How can differing data types / widths of values for each layer affect efficiency and training?
- How can differing data types / widths of values for each layer affect efficiency and prediction accuracy?

Future Research

- How can training of CNNs be accelerated on Field Programmable Gate Array (FPGA)?
- How can training of CNNs be more efficient on FPGA or Application Specific Integrated Circuit (ASIC)?
- Focus on Energy efficiency and performance.