Analysis of the trade-offs of dealing with sparsity in a custom DNN accelerator

Adrián Alcolea Moreno, Jesús Javier Resano Ezcaray,
Javier Olivito, Hortensia Mecha
{alcolea, jresano, jolivito}@unizar.es, horten@ucm.es

Index

- Convolutional Neural Networks (CNNs)
  - Sparsity
- Architecture Analysis
  - Compression
  - Identification of useful operations
  - Sparse model on an FPGA
- Experimental Results
- Conclusions: Is it worth exploiting sparsity?

Index

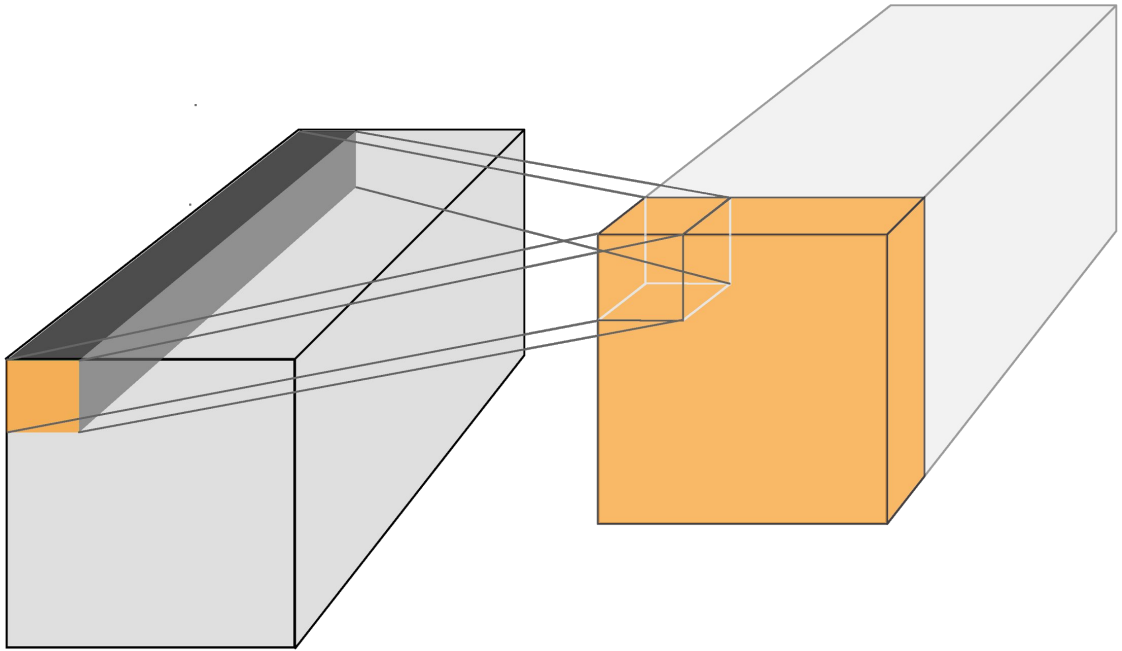
- Convolutional Neural Networks (CNNs)
  - Sparsity
- Architecture Analysis
  - Compression
  - Identification of useful operations
  - Sparse model on an FPGA
- Experimental Results
- Conclusions: Is it worth exploiting sparsity?

Convolutional Neural Networks (CNNs)

- CNNs are one of the most used machine learning techniques for images automatic analysis and classification.
- They achieve great accuracy, but the models tend to be very big and they need a lot of calculations during inference.
- The convolution algorithm is computationally intensive and in these networks it is performed repeatedly.

Convolutional Neural Networks (CNNs)

- Two concepts for later: ‘activation’ and ‘filter’



Index

- Convolutional Neural Networks (CNNs)
  - Sparsity
- Architecture Analysis
  - Compression
  - Identification of useful operations
  - Sparse model on an FPGA
- Experimental Results
- Conclusions: Is it worth exploiting sparsity?

Sparsity

- CNNs activations had a lot of zeros generated by the activation function.
- There are also a lot of works dedicated to generate zeros in the filters by pruning techniques.
- The presence of zeros in the network architecture allows to compress it saving space…
- … but also permits to avoid a great amount of operations.

Index

- Convolutional Neural Networks (CNNs)
  - Sparsity
- Architecture Analysis
  - Compression
  - Identification of useful operations
  - Sparse model on an FPGA
- Experimental Results
- Conclusions: Is it worth exploiting sparsity?

Compression

Uncompressed

0	0	7	0	2
4	3	0	0	0
0	2	0	8	9
5	0	0	0	0
0	6	0	1	0

a) Uncompressed

Values array

7	2	4	3	2	8	9	5	6	1

Column indices

2	4	0	1	1	3	4	0	1	3

Row pointers

0	2	4	7	8	10

a) CSR format representation

Values array

7	2	4	3	2	8	9	5	6	1

Indices matrix

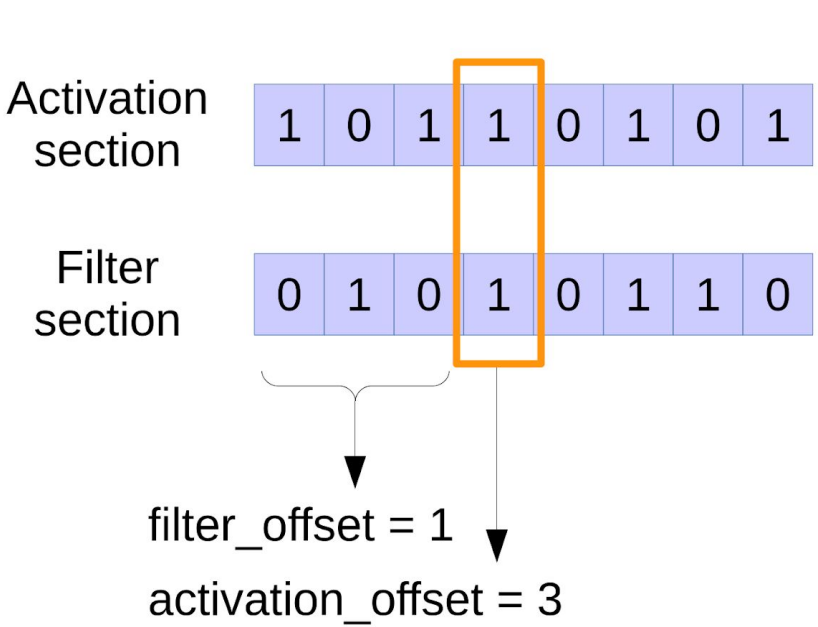
0	0	1	0	1
1	1	0	0	0
0	1	0	1	1
1	0	0	0	0
0	1	0	1	0

b) Indices matrix representation

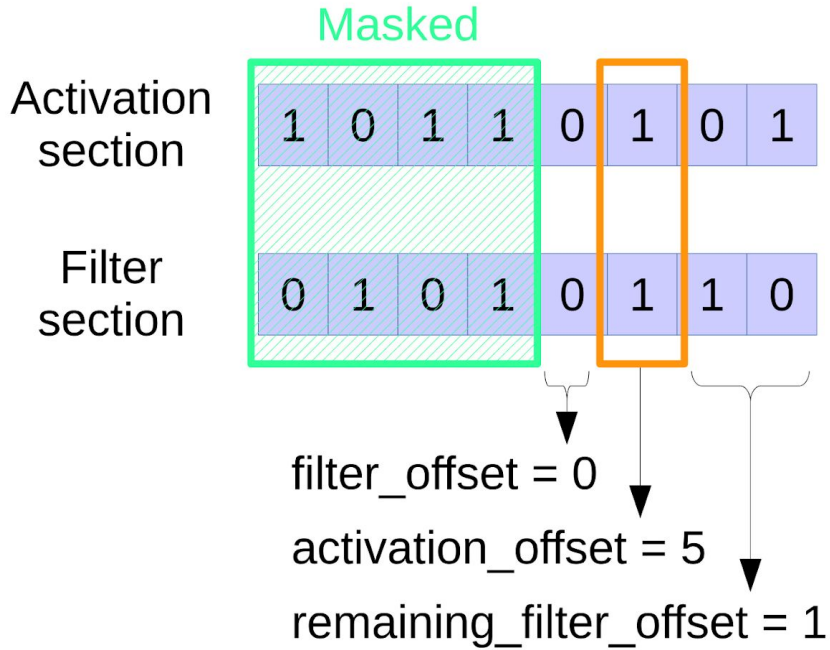
Index

- Convolutional Neural Networks (CNNs)
  - Sparsity
- Architecture Analysis
  - Compression
  - Identification of useful operations
  - Sparse model on an FPGA
- Experimental Results
- Conclusions: Is it worth exploiting sparsity?

Identification of useful operations



a) Iteration one



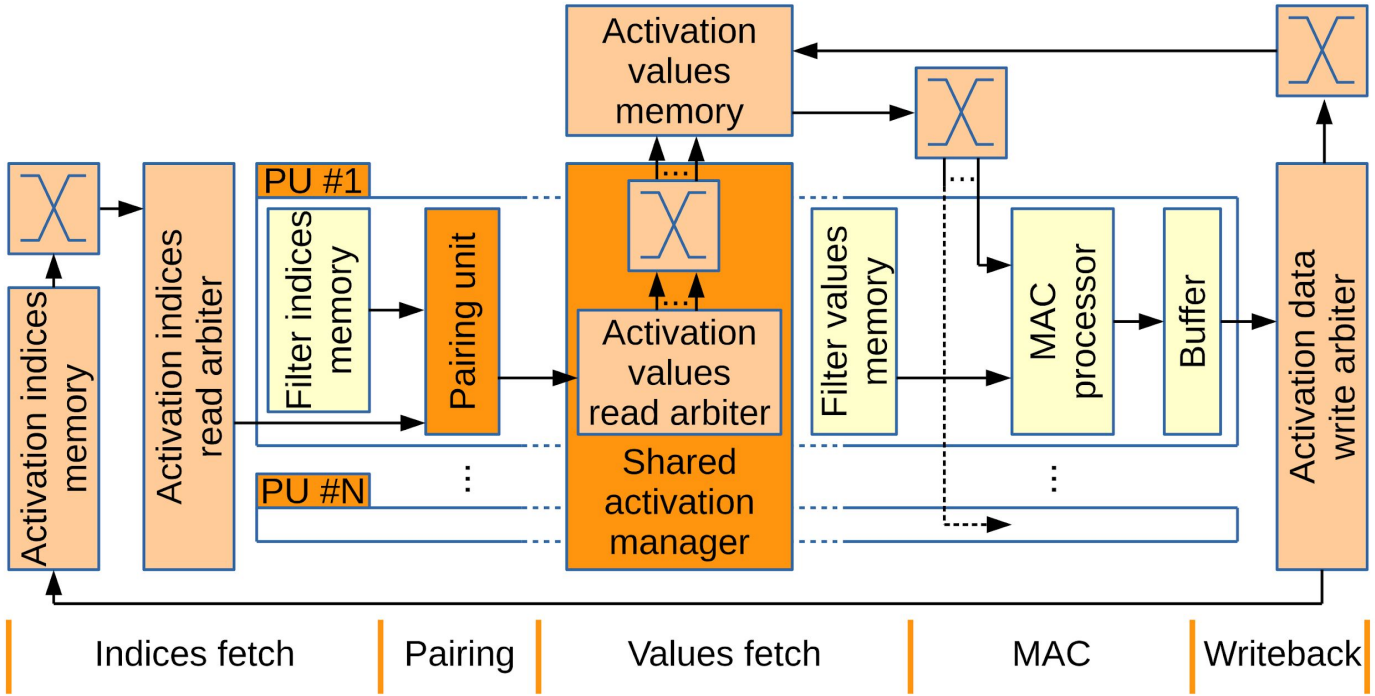
b) Iteration two

Index

- Convolutional Neural Networks (CNNs)
  - Sparsity
- Architecture Analysis
  - Compression
  - Identification of useful operations
  - Sparse model on an FPGA
- Experimental Results
- Conclusions: Is it worth exploiting sparsity?

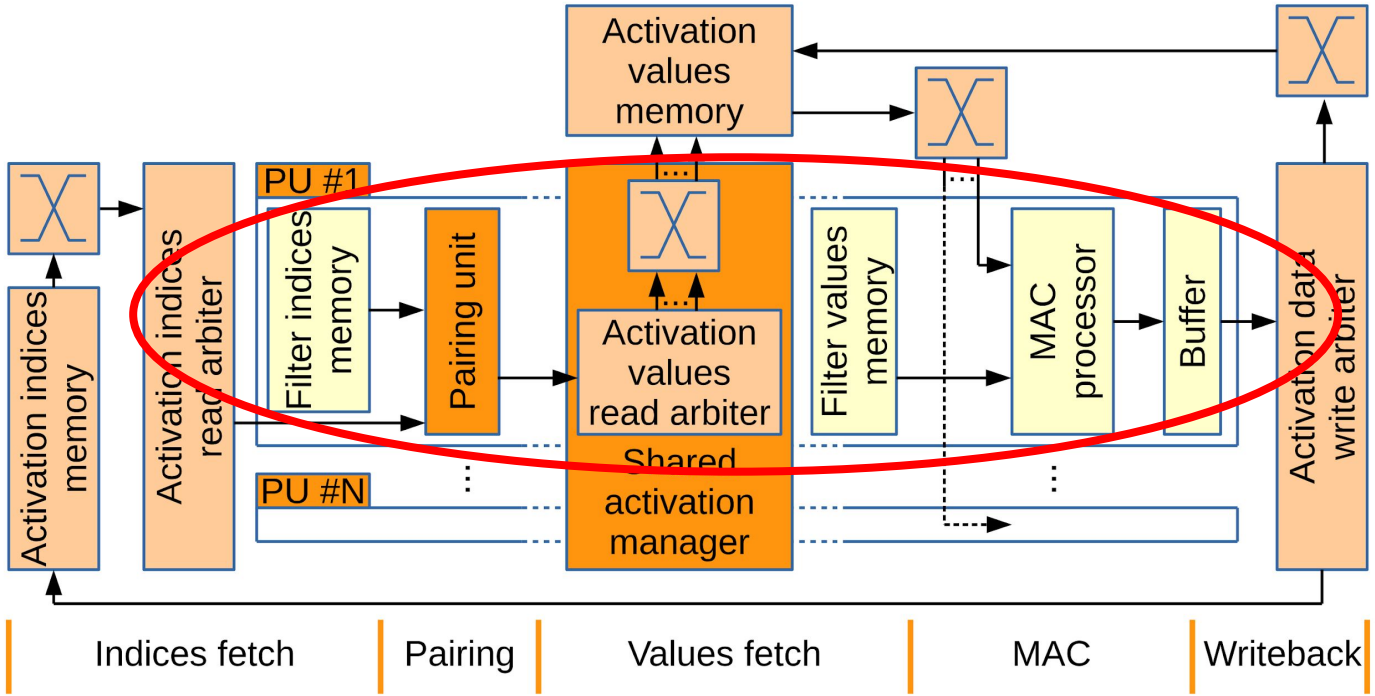
Sparse model on an FPGA

- We designed a five-stage pipeline architecture.



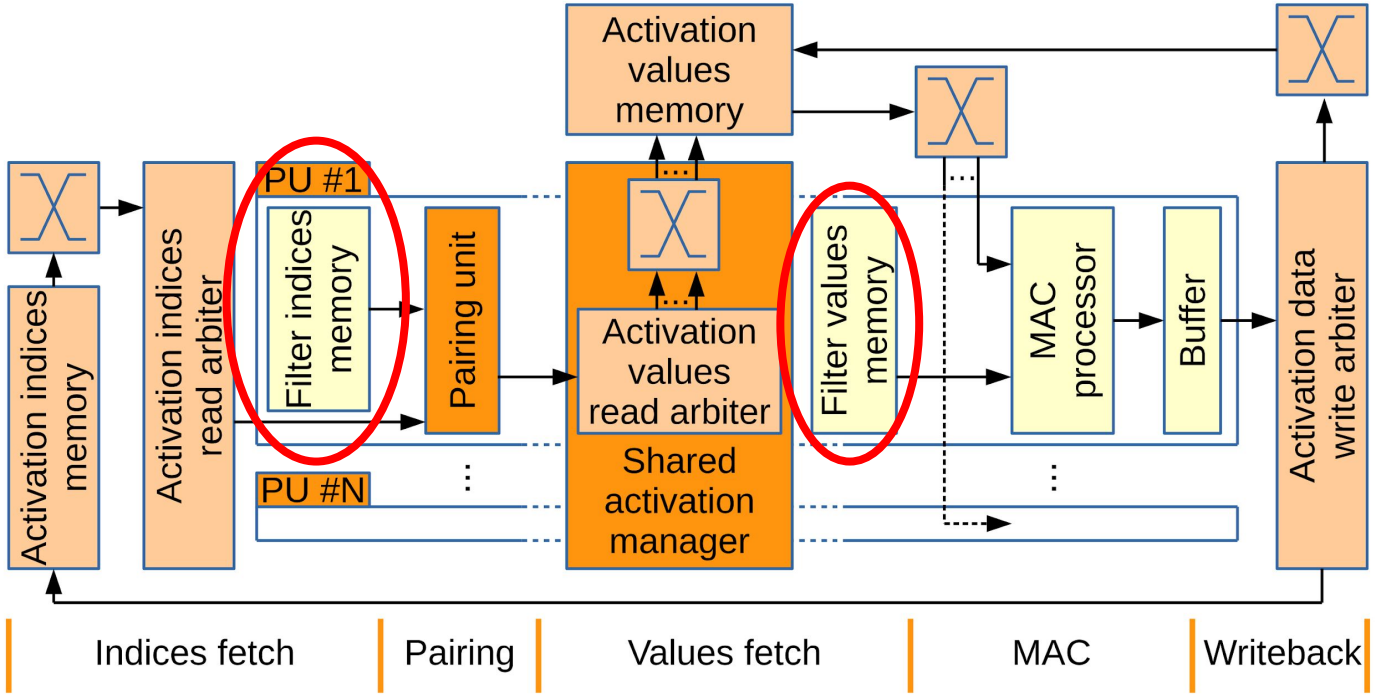
Sparse model on an FPGA

- It consists of a series of PUs working in parallel in different filters.



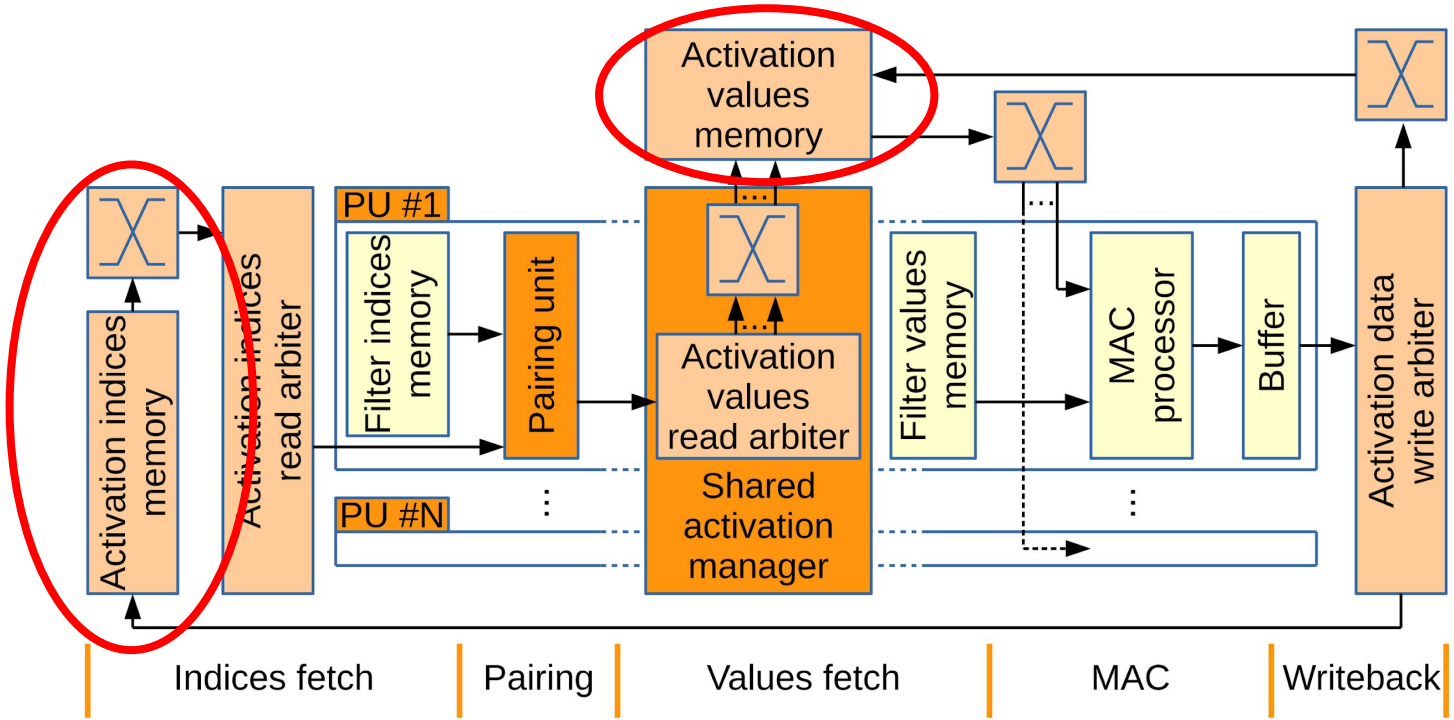
Sparse model on an FPGA

- Each PU works with its own private filter memory.



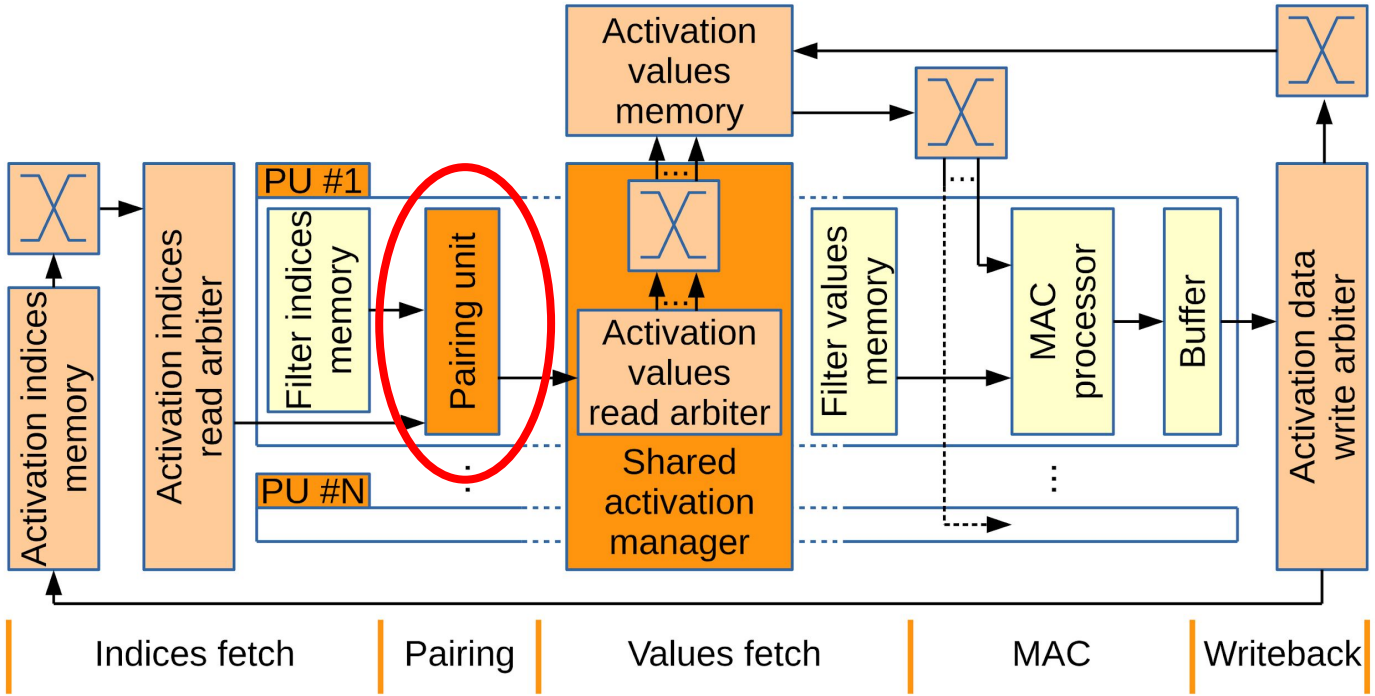
Sparse model on an FPGA

- Activation memory is shared among all the PUs.



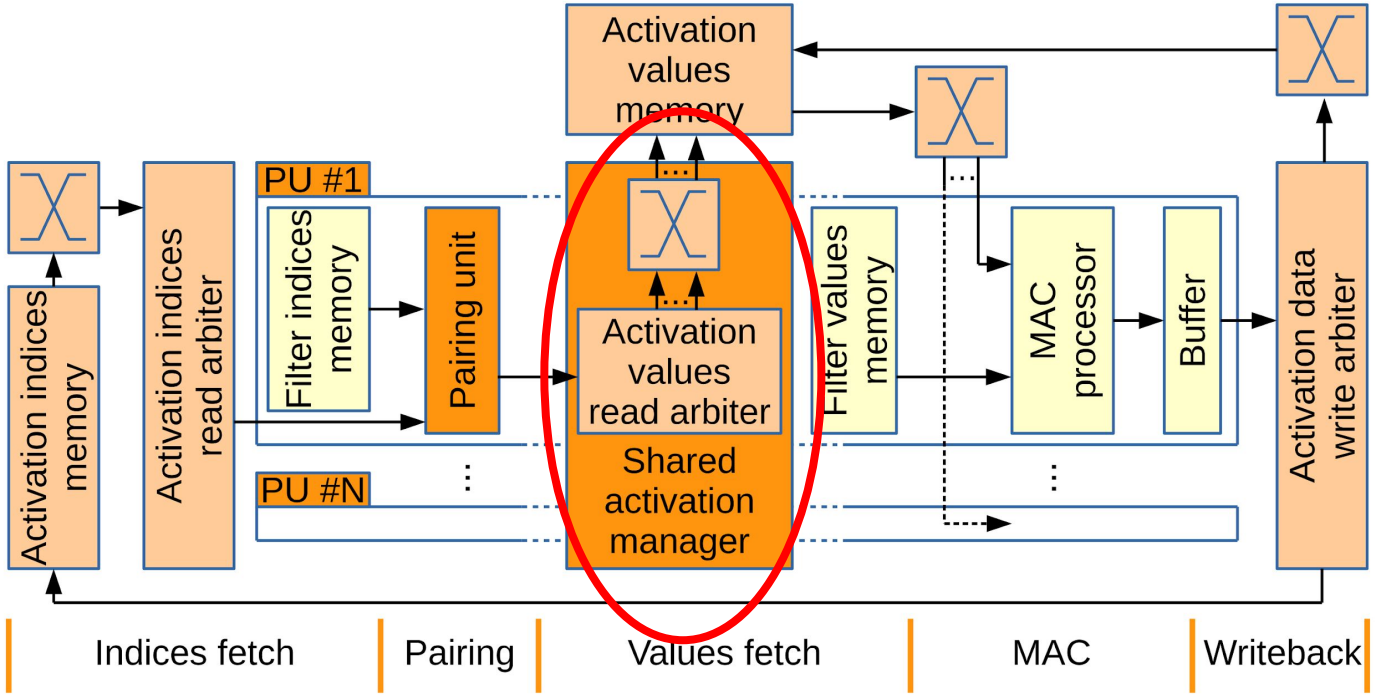
Sparse model on an FPGA

- Address generation only for useful operations.

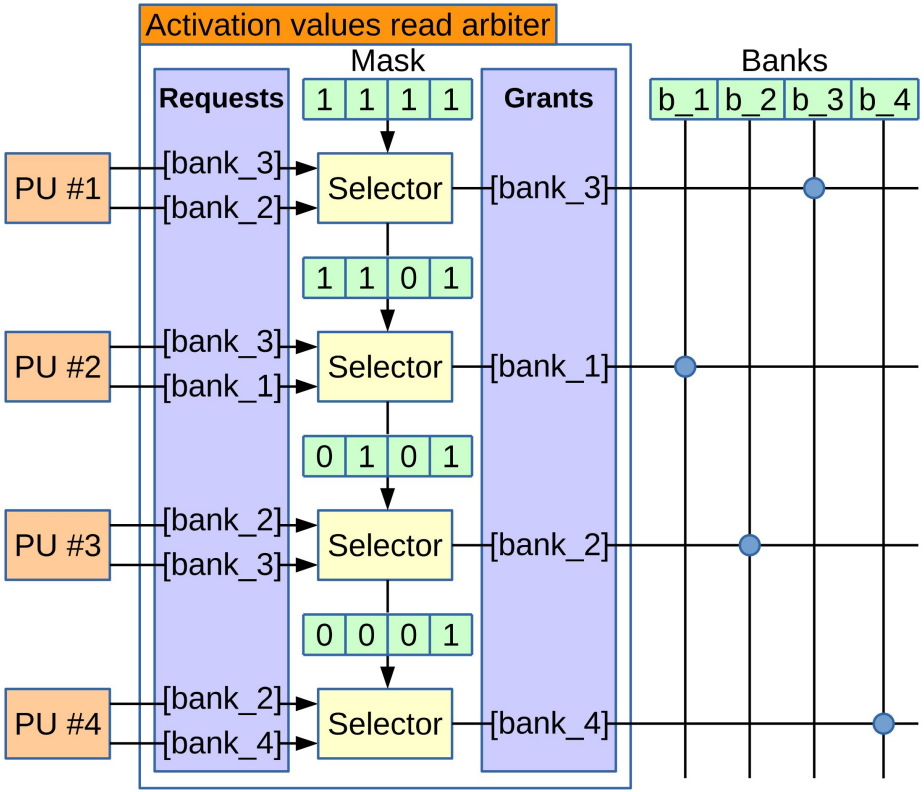


Sparse model on an FPGA

- Shared (and unordered) accesses to the activation memory.



Sparse model on an FPGA

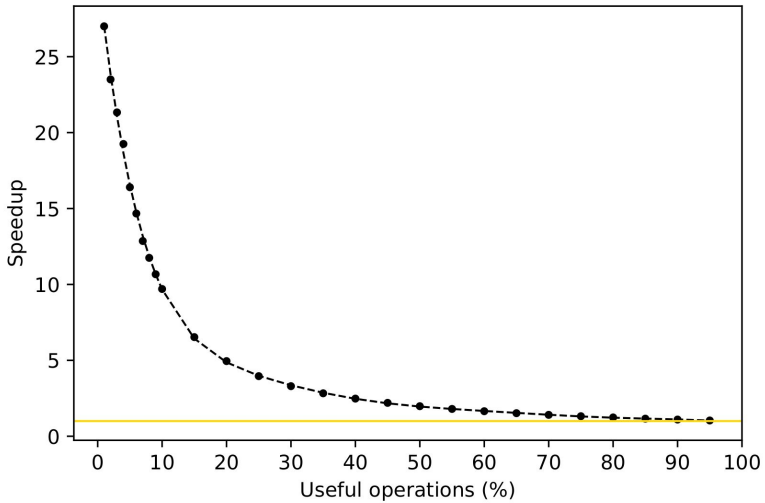


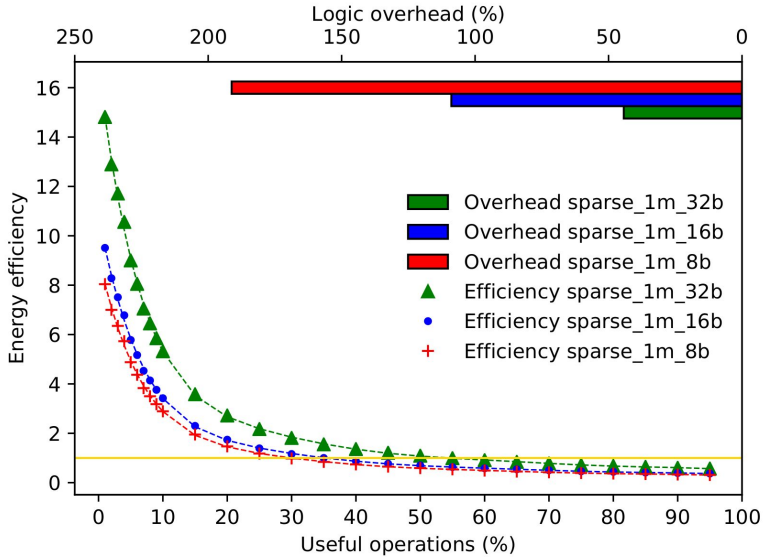
Index

- Convolutional Neural Networks (CNNs)
  - Sparsity
- Architecture Analysis
  - Compression
  - Identification of useful operations
  - Sparse model on an FPGA
- Experimental Results
- Conclusions: Is it worth exploiting sparsity?

Experimental Results

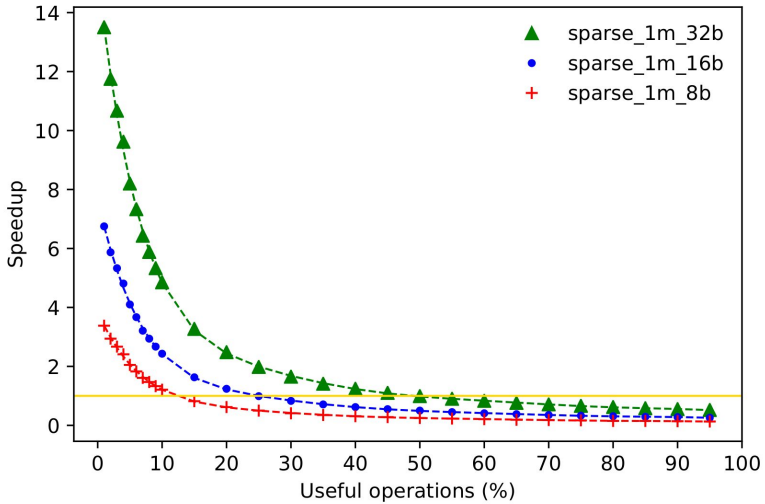
- Avoiding useless operations saves energy and time…
- … but it requires extra hardware to manage sparsity.

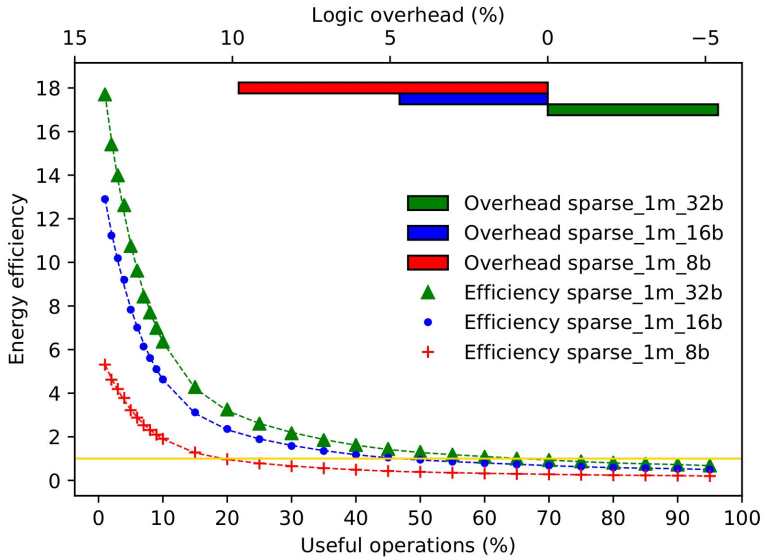




Experimental Results

- We decided to provide our dense design with more arithmetic resources until both designs where similar in area.





Index

- Convolutional Neural Networks (CNNs)
  - Sparsity
- Architecture Analysis
  - Compression
  - Identification of useful operations
  - Sparse model on an FPGA
- Experimental Results
- Conclusions: Is it worth exploiting sparsity?

Conclusions: Is it worth exploiting sparsity?

- Sparsity is essential when evaluating an accelerator. But the arithmetics bitwidth is also important. According to our figures:
- For 32-bit arithmetics, the benefits are clear.
- For 8-bit arithmetics, it is difficult to take greater advantage than that obtained by increasing the number of multipliers.
- For 16-bit arithmetic, it depends on the number of useful operations (<50% for energy efficiency advantages and <25% for performance advantages).

The End

Thank you