MAX78002: Ultra-Low-Power 1,2,3,4,8-Bit Convolutional Neural Network (CNN) Accelerator

Zurich University
of Applied Sciences

School of Engineering
InES Institute of
Embedded Systems

Scope:

Convolutional Neural Networks are brought to the edge via ultra-low-power accelerators integrated in microcontrollers.

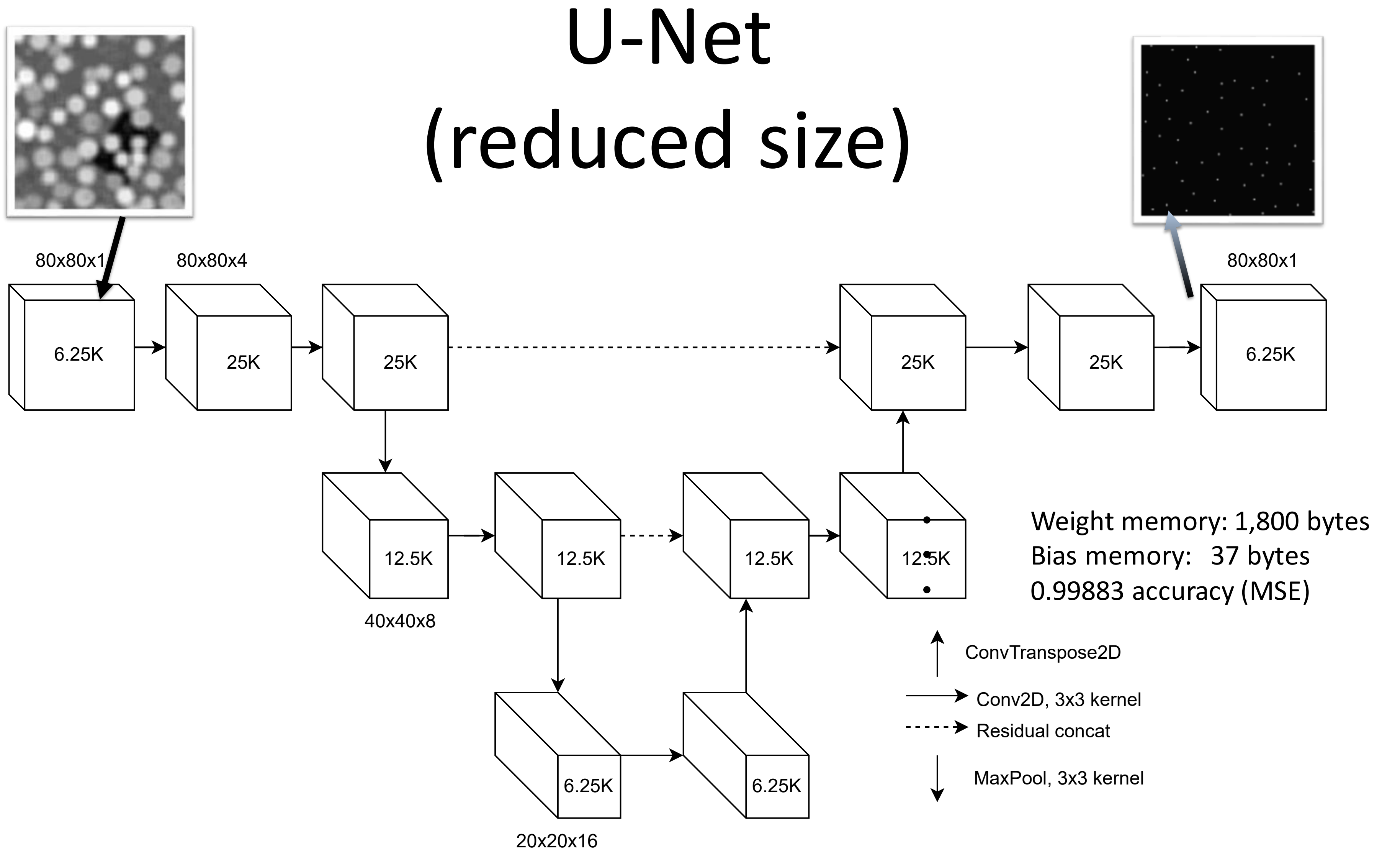
Application:

Deployment of a highly reliable low-power object counting artificial intelligence. The model must resist noise and light changes, and recognize the proper round object.

The inference result must contain a single dot for each valid object found. The dots should be located approximately at the center of the found objects and the integration of them all must result in the count of relevant objects observable in the input picture.

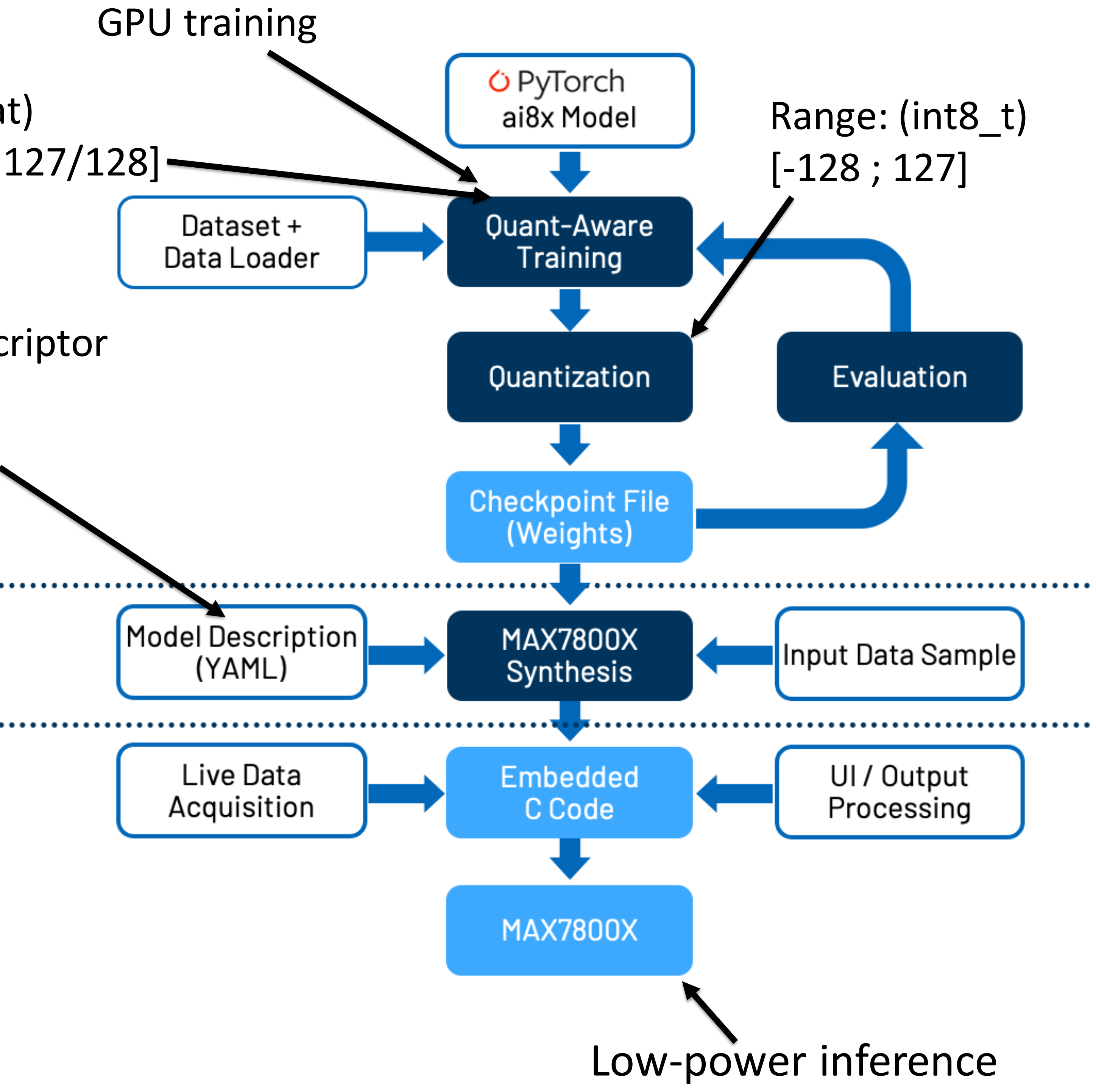
Each centered dot is the result of a regression, where probabilities are produced via ReLU in the numerical range of [0, 127]. Similarly, as with logistic regression, a threshold for a valid point must be set.

U-Net (reduced size)

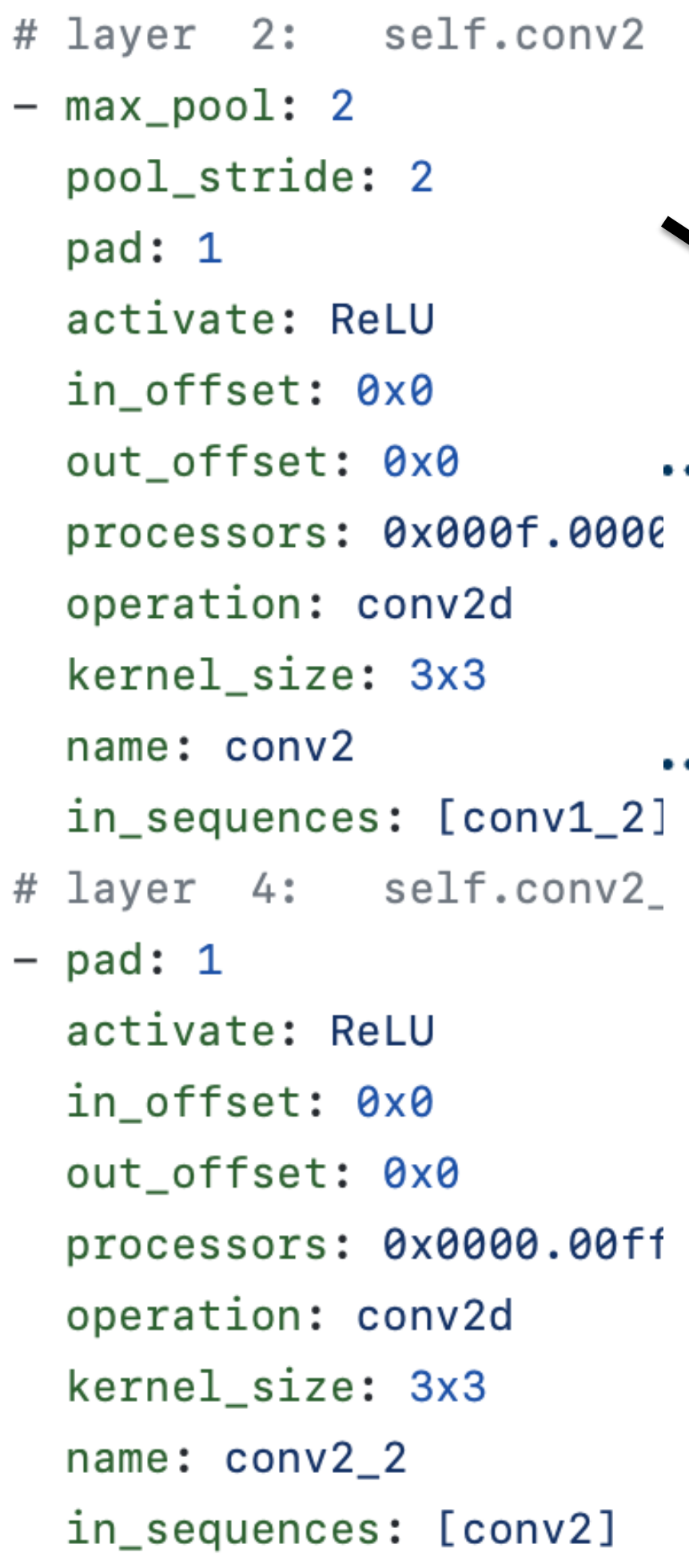


Weight memory: 1,800 bytes
Bias memory: 37 bytes
0.99883 accuracy (MSE)

Workflow



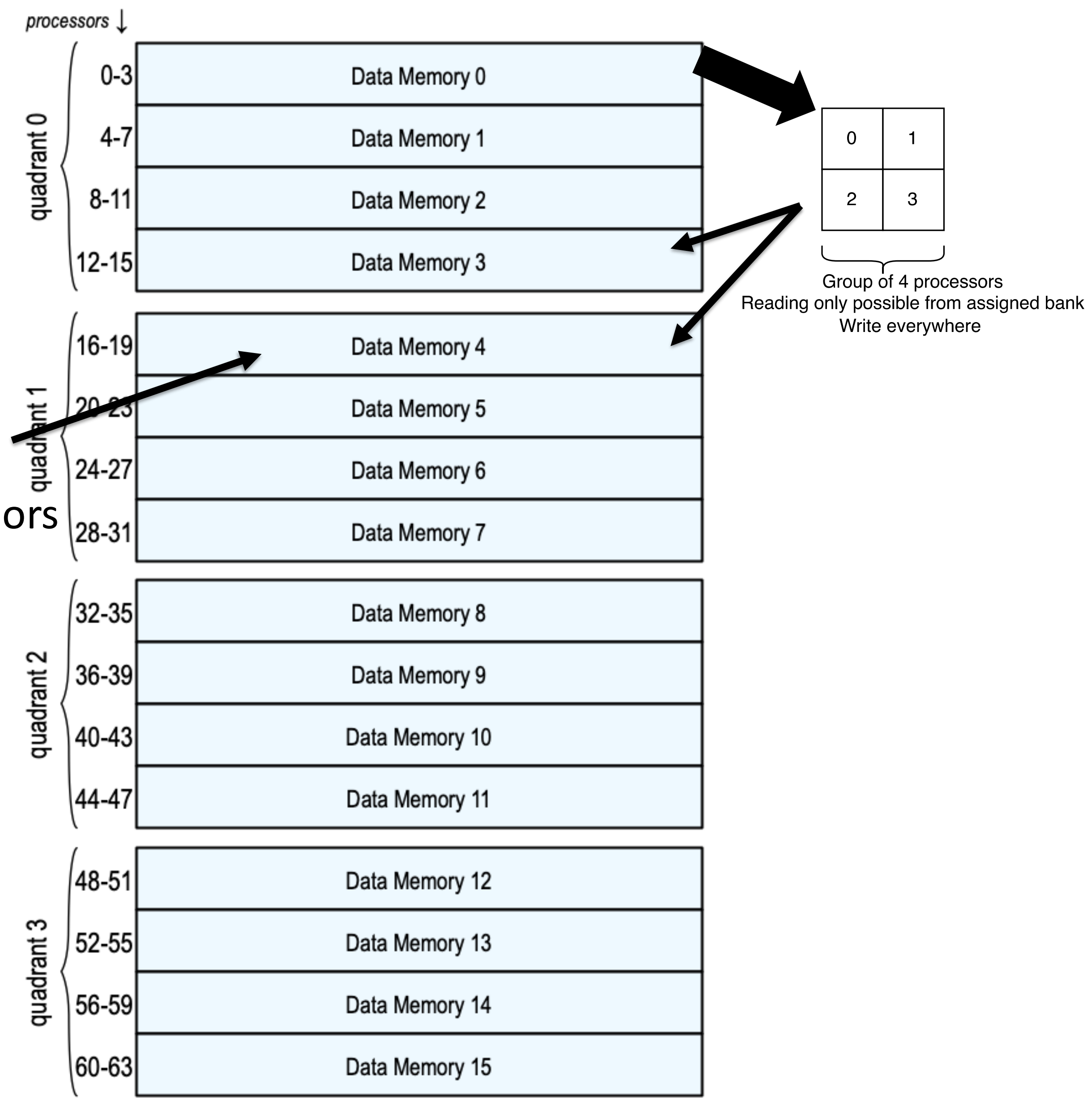
Model to HW descriptor



```
# layer  2:   self.conv2
- max_pool: 2
  pool_stride: 2
  pad: 1
  activate: ReLU
  in_offset: 0x0
  out_offset: 0x0
  processors: 0x000f.0000
  operation: conv2d
  kernel_size: 3x3
  name: conv2
  in_sequences: [conv1_2]
# layer  4:   self.conv2_
- pad: 1
  activate: ReLU
  in_offset: 0x0
  out_offset: 0x0
  processors: 0x0000.00ff
  operation: conv2d
  kernel_size: 3x3
  name: conv2_2
  in_sequences: [conv2]
```

Memory layout

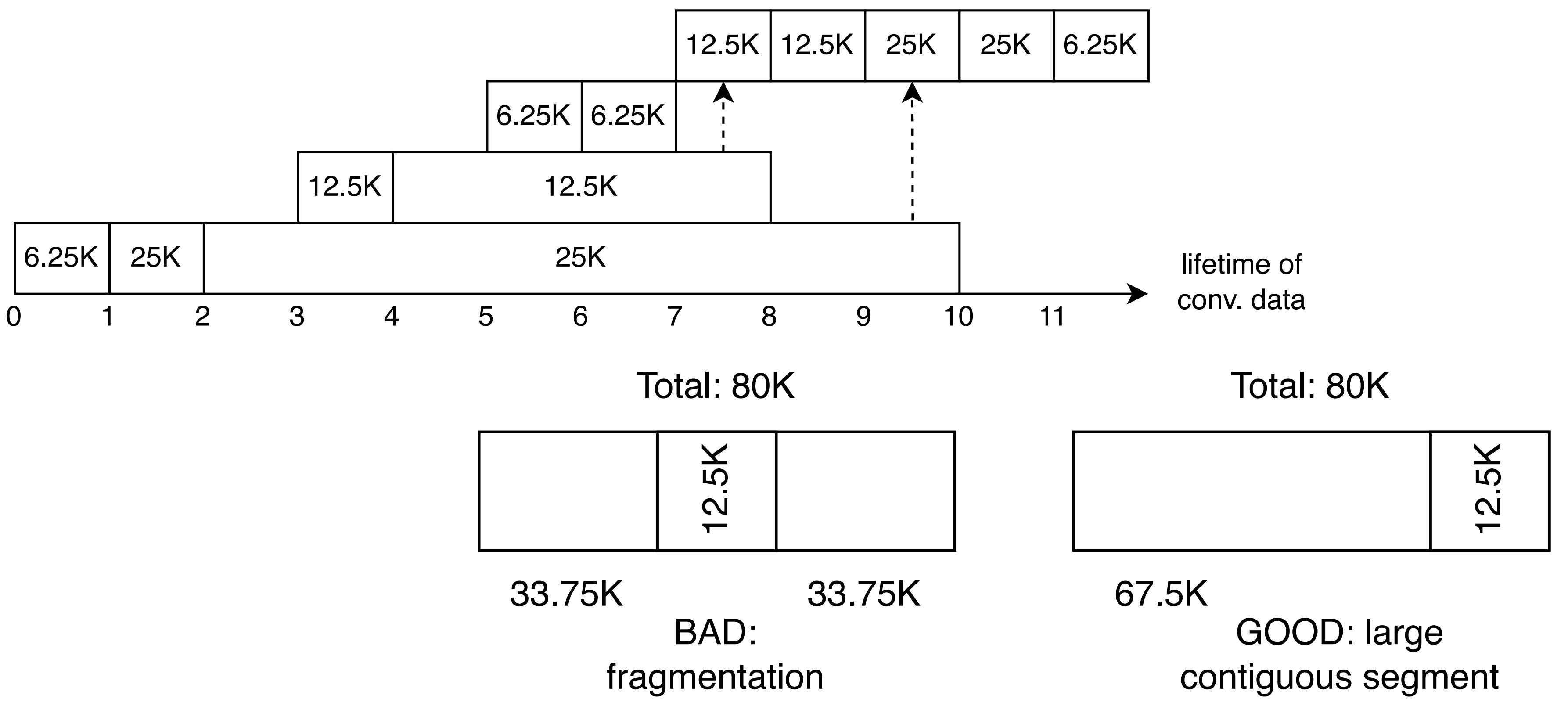
80Kbytes, 16 instances
Shared across 4 processors
A total of 64 processors



Inference memory allocation

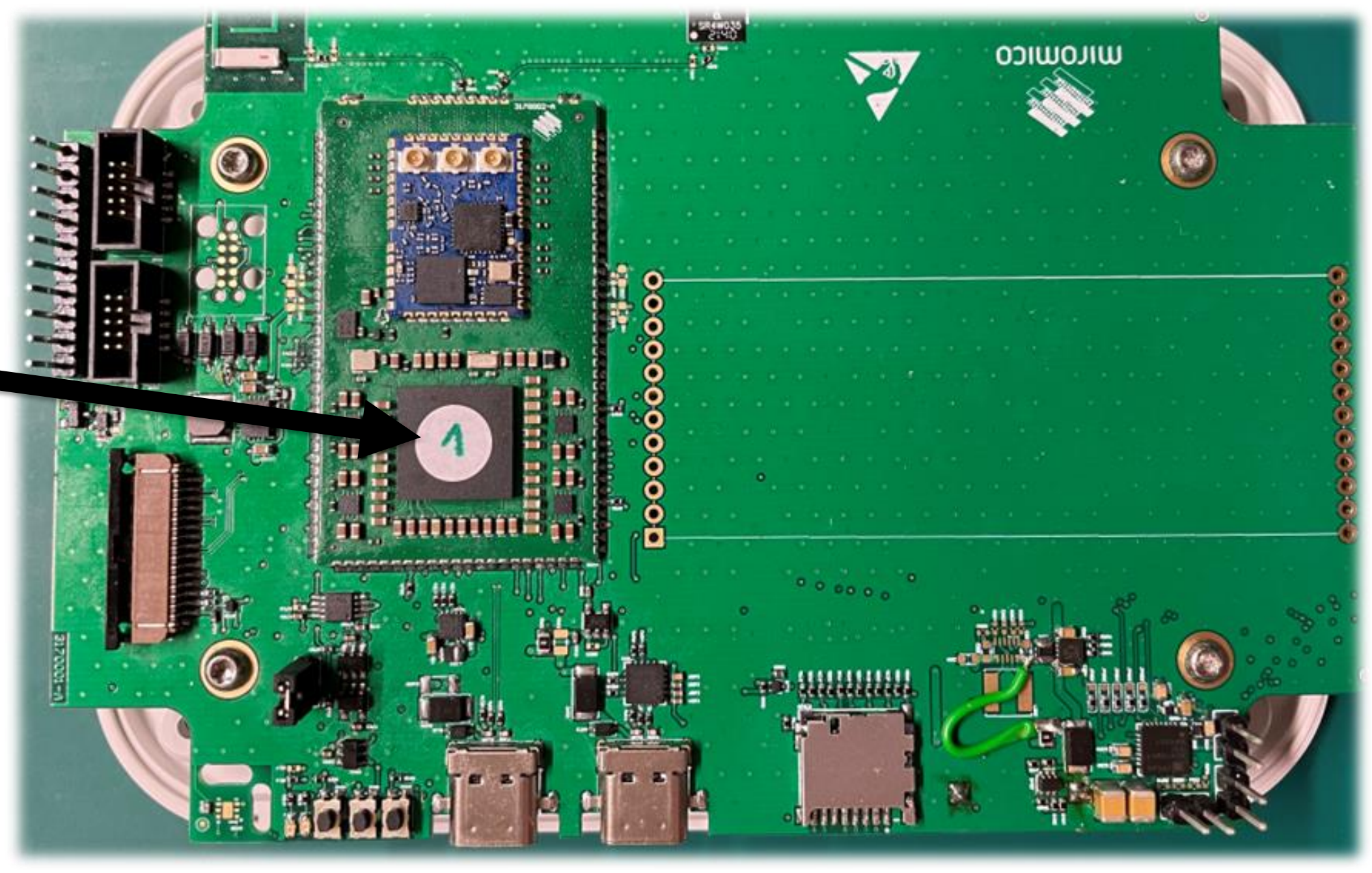
Memory management is a great challenge in the MAX78002 CNN accelerator. Memory frames will vary in size and number during the inference process. The lifetime of each frame varies based on the deployed model. Each channel requires an available CNN processor. Four processors share one 80K data memory block, and they can only read from this dedicated memory block. Writing is possible in all memory blocks. A processor is only available if its input data is in the correct memory location and alignment. Also, each of the output convolution processors must have enough memory available in their readable memory bank in order to store the input for the next inference step.

Memory fragmentation quickly leads to the inability to use a required processor.
In case an intermediate output of 10 Kbytes for 64 channels must be stored, each memory bank must have a free contiguous range of 40 Kbytes. Thus, not having badly-placed residuals is extremely important.



Overview of the MAX78002:

- Arm Cortex-M4 Processor with FPU up to 120 MHz
- 32-Bit RISC-V Coprocessor up to 60MHz
- 2.5 MB Flash, 64 KB ROM and 384 KB SRAM
- Highly Optimized for Deep CNNs
- 2 M 8-Bit Weight Capacity with 1,2,4,8-Bit Weights
- 1.3 MB CNN Data Memory
- 64 CNN Coprocessors



ZHAW Institute of Embedded Systems
Dávid Isztl
CH-8401 Winterthur
iszt@zhaw.ch
+41 58 934 74 80