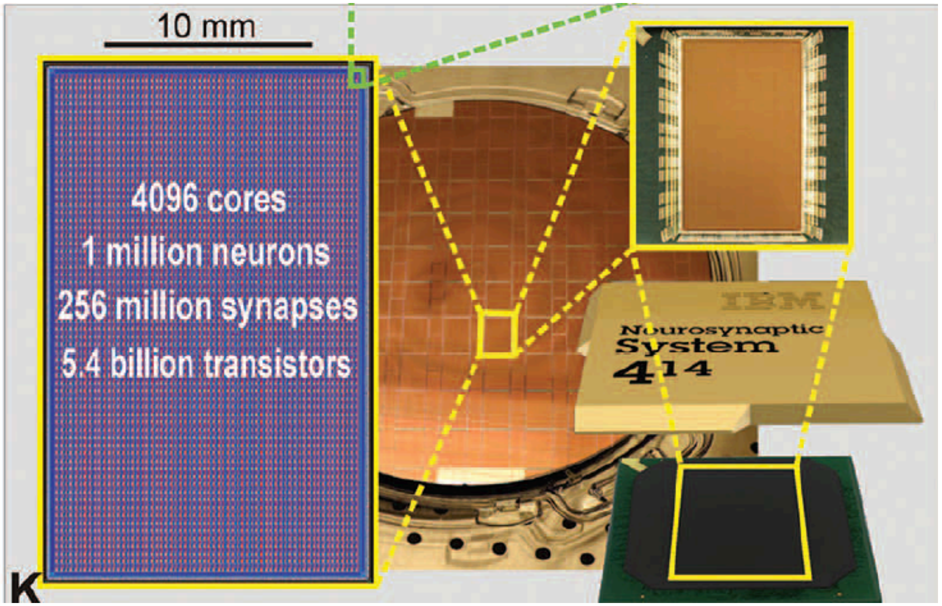
Digital Hardware Implementation of Neural Networks

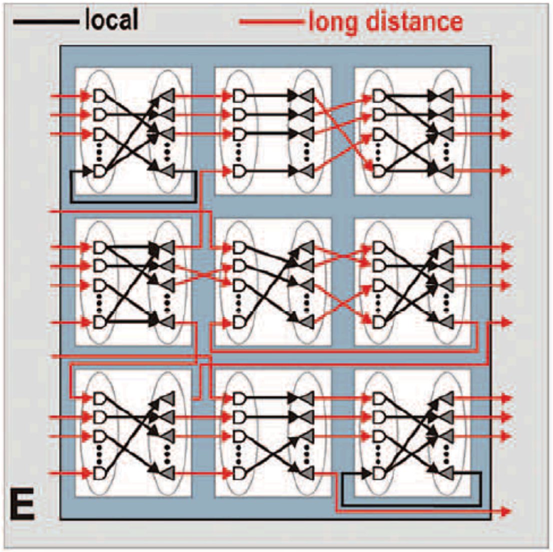
Yasmine Badr

12/9/2015

TrueNorth by IBM[1]



TrueNorth Chip



Each Core: 256x256 neurons

Neuflow [2]-intro

- For Real-time object detection and categorization
- Dataflow processor & more specific Convolutional Neural Network in FPGA and ASIC
- Compiler for flow-graph decription of algiruthms in **torch5 framework**

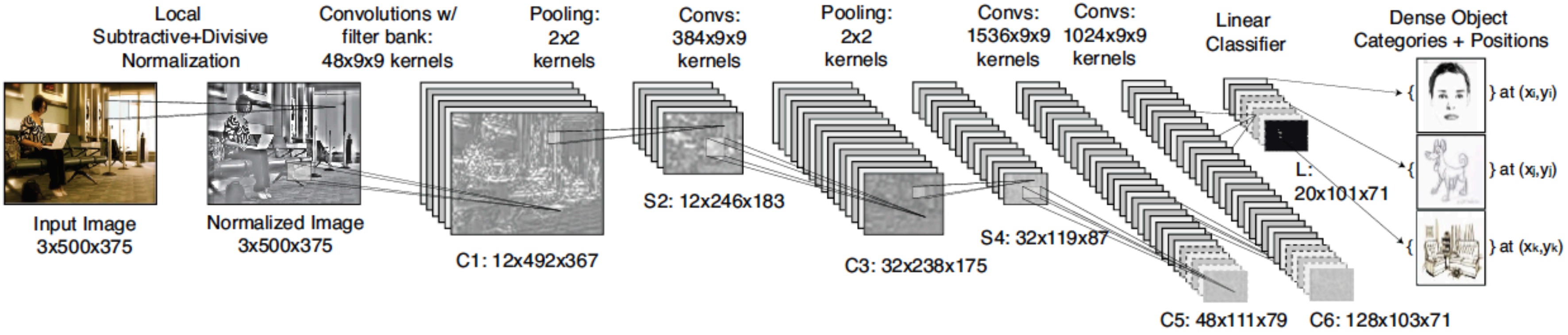
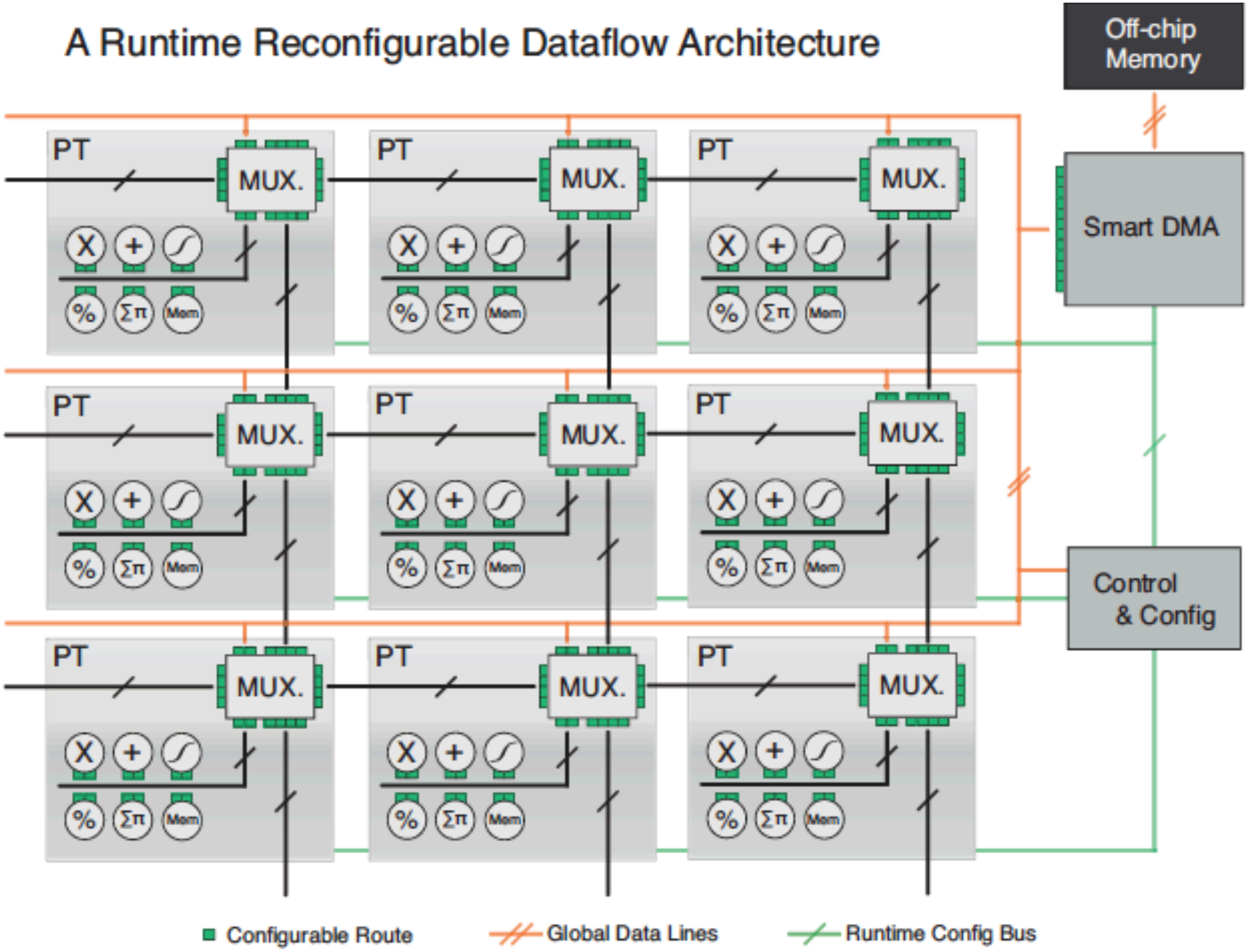


Figure 4. A convolutional network for street scene parsing.

Neuflow [2]- Runtime Configurable Data flow Processor



Neuflow[2]- configured architecture

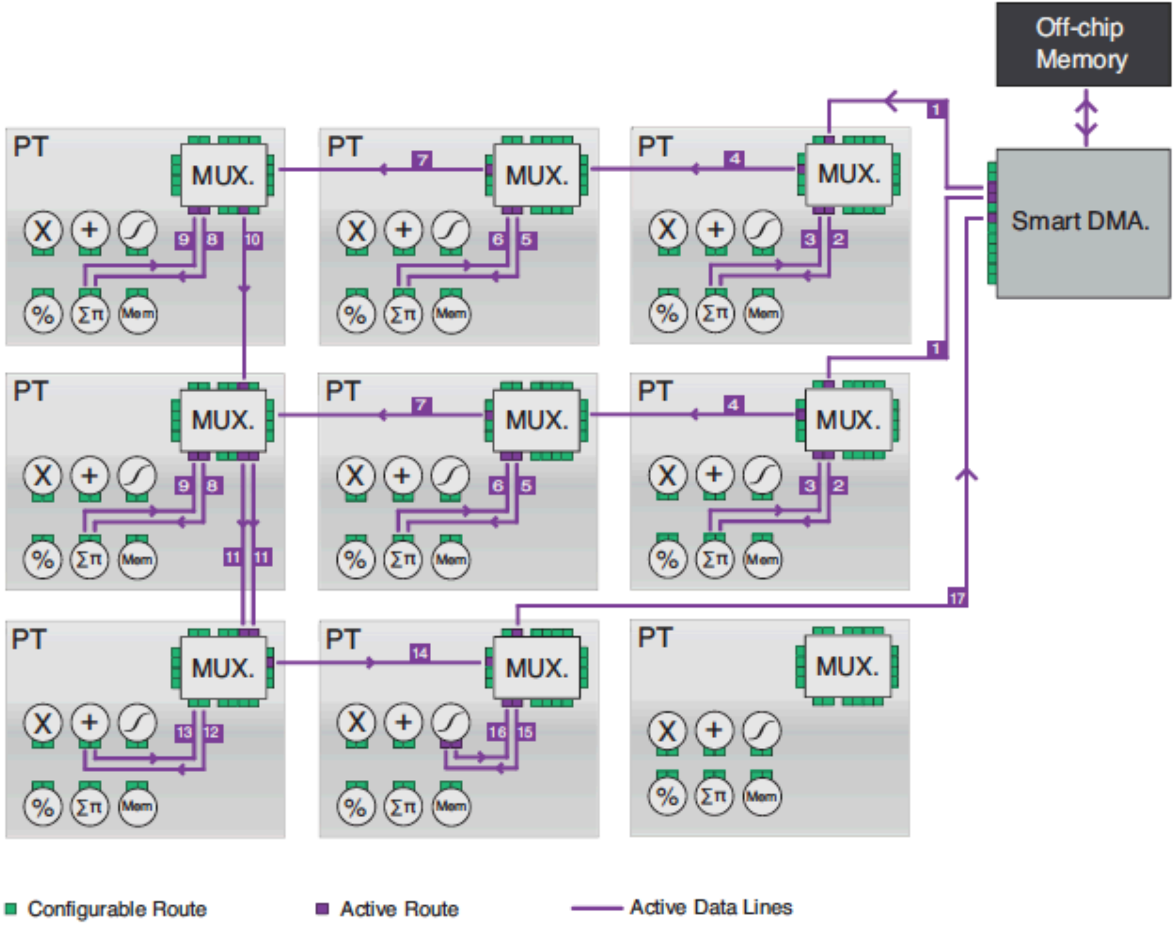


Figure 2. The grid is configured for a complex computation that involves several tiles: the 3 top tiles perform a 3 × 3 convolution, the 3 intermediate tiles another 3 × 3 convolution, the bottom left tile sums these two convolutions, and the bottom centre tile applies

$$y_{1,i,j} = Tanh(\sum_{m=0}^{K-1}\sum_{n=0}^{K-1} x_{1,i+m,j+n} w_{1,m,n} + \sum_{m=0}^{K-1}\sum_{n=0}^{K-1} x_{2,i+m,j+n} w_{2,m,n}).$$

Neuflow[2]- CNNs

- For DSP-oriented FPGA which have MAC units
- Specialized for CNNs
  - 80-90% computations are filtering

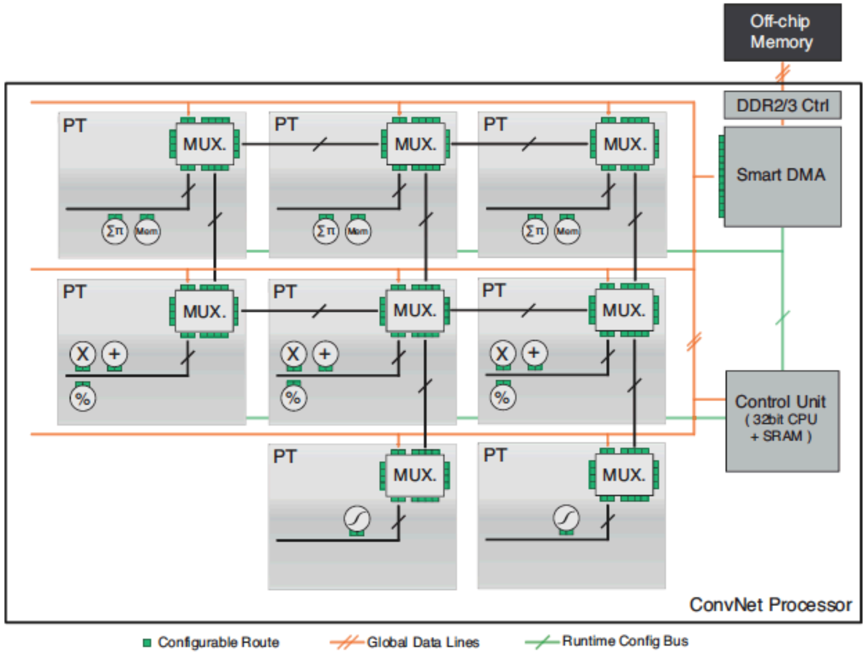


Figure 3. Optimizing the grid for filter-based systems. A grid of multiple full-custom Processing Tiles, and a fast streaming memory interface (Smart DMA).

Neuflow [2]- example

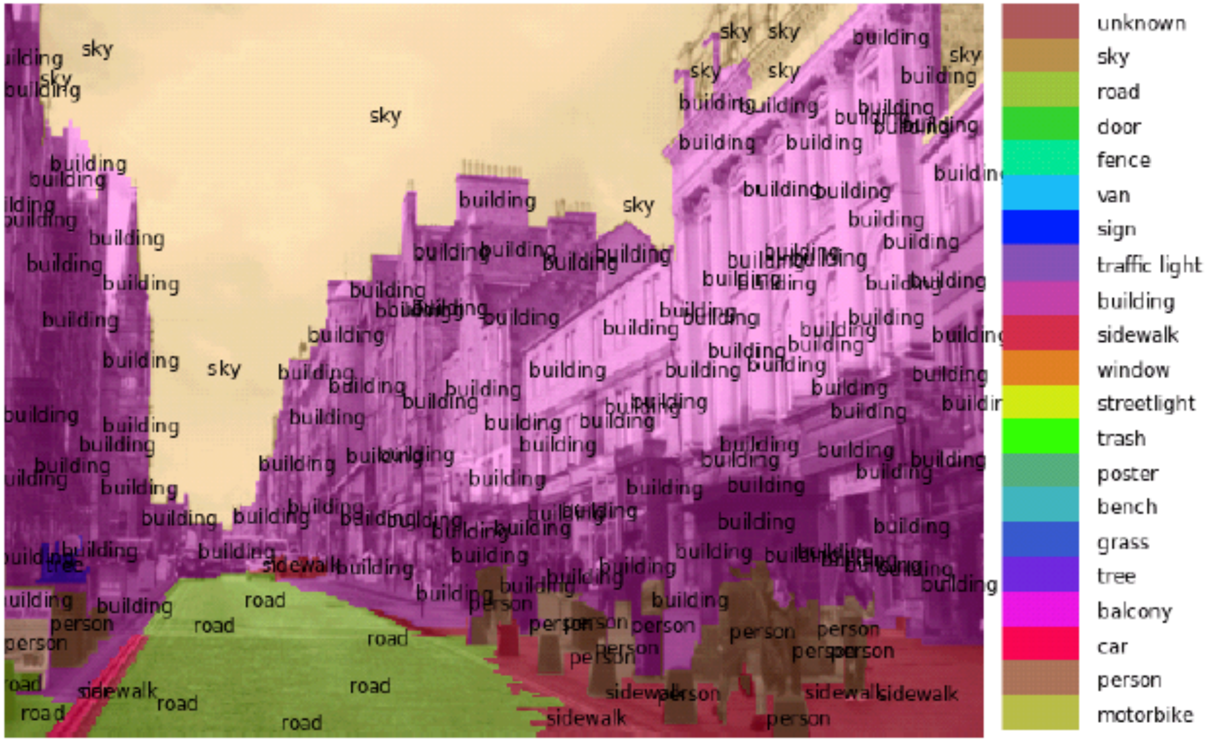


Figure 5. Street scene parsing: a convolutional network was trained on the LabelMe spanish dataset [26] with a method similar to [12]. The training set only contains photos from spanish cities; the image above is a picture taken in Edinburgh. The convolutional network is fully computed on neuFlow, achieving a speedup of about 100x (500x375 images are processed in 83ms, as opposed to 8s on a laptop).

Spiking Neural Network [Wikipedia]

- SNNs incorporate concept of **time**.
- Neurons do not fire at each propagation cycle (as it happens with typical multi-layer perceptron networks)
  - fire only when a membrane potential reaches a specific value.
- When a neuron fires, it generates a signal which travels to other neurons which, in turn, increase or decrease their potentials in accordance with this signal

CMOS Neuromorphic chip for Spiking Neural Networks [3]

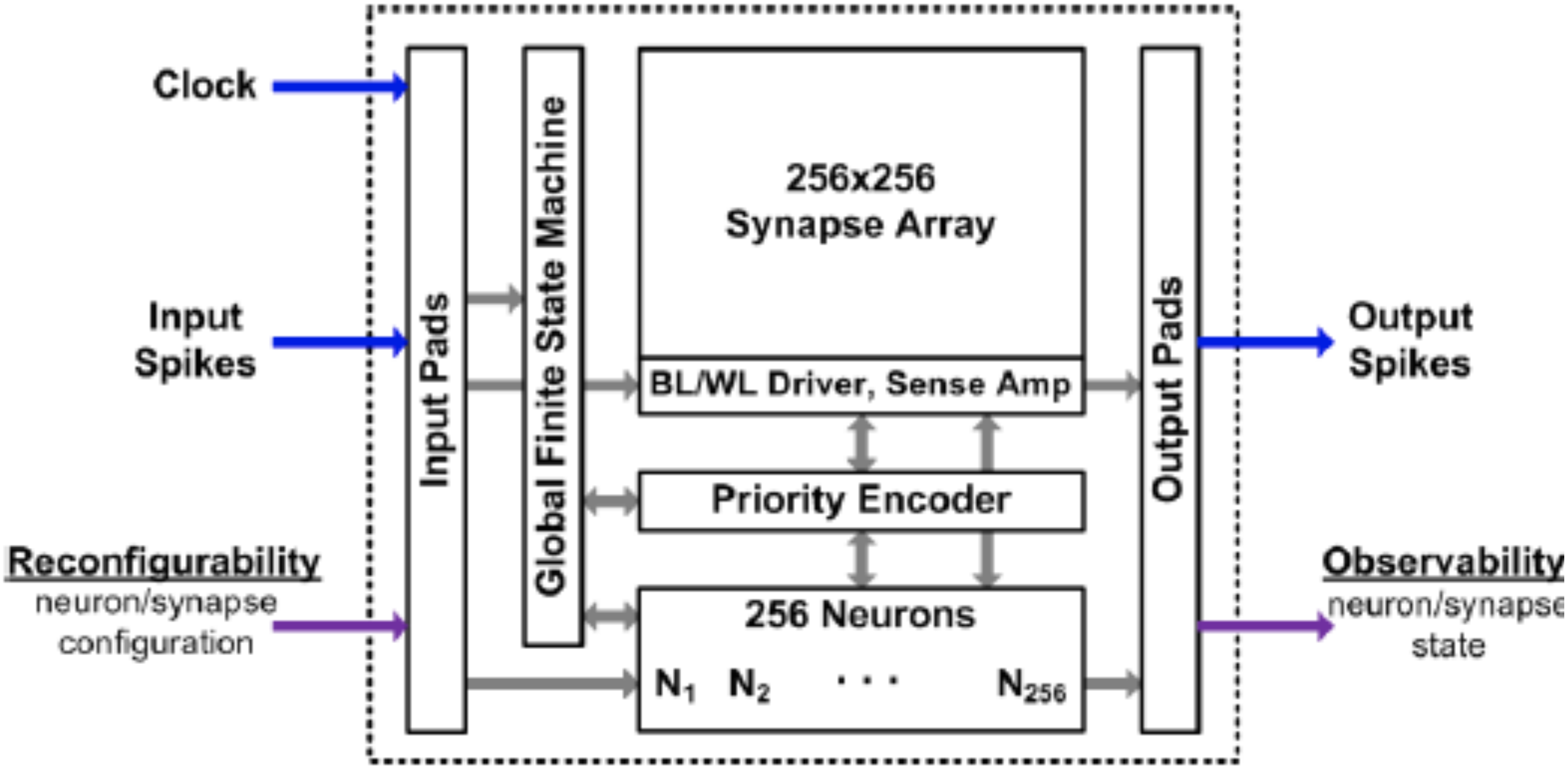


Fig. 1. Top-level block diagram of on-chip learning chip.

CMOS Neuromorphic chip for Spiking Neural Networks [3]

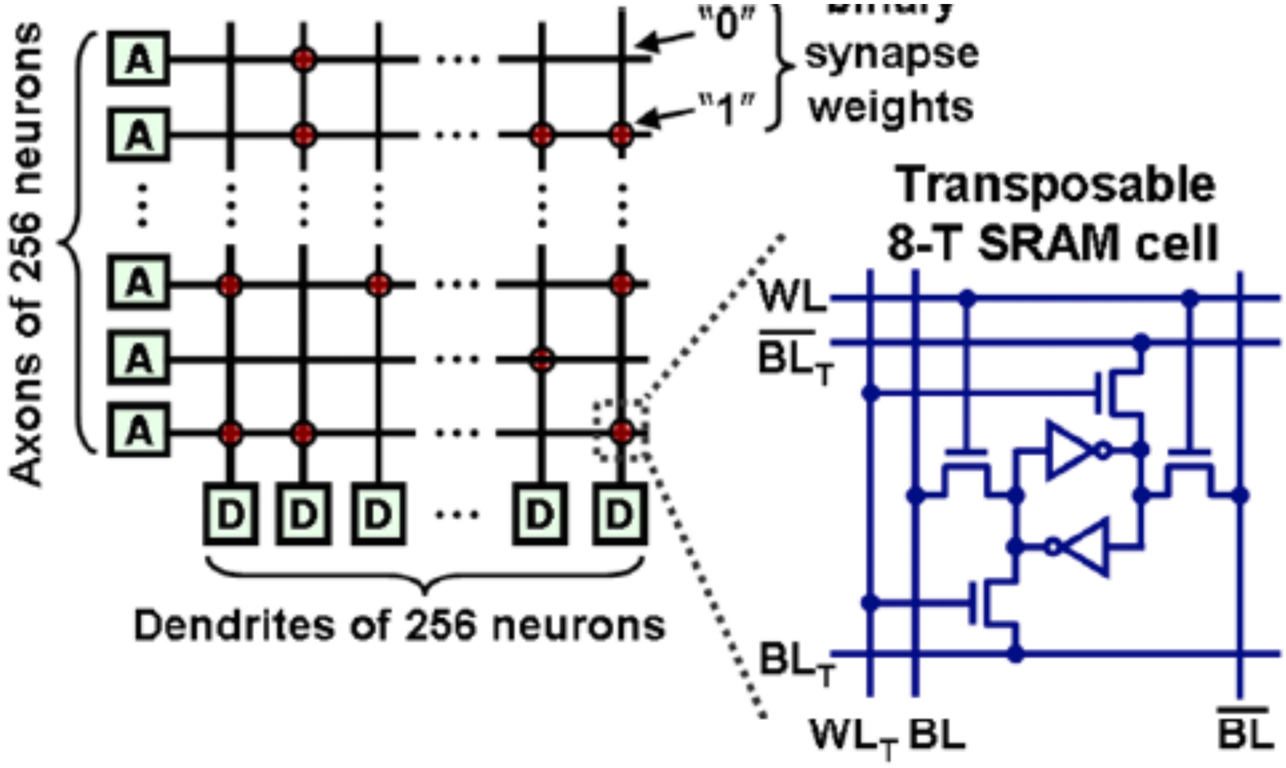


Fig. 2. Synapse array with transposable SRAM cells.

CMOS Neuromorphic chip for Spiking Neural Networks [3]

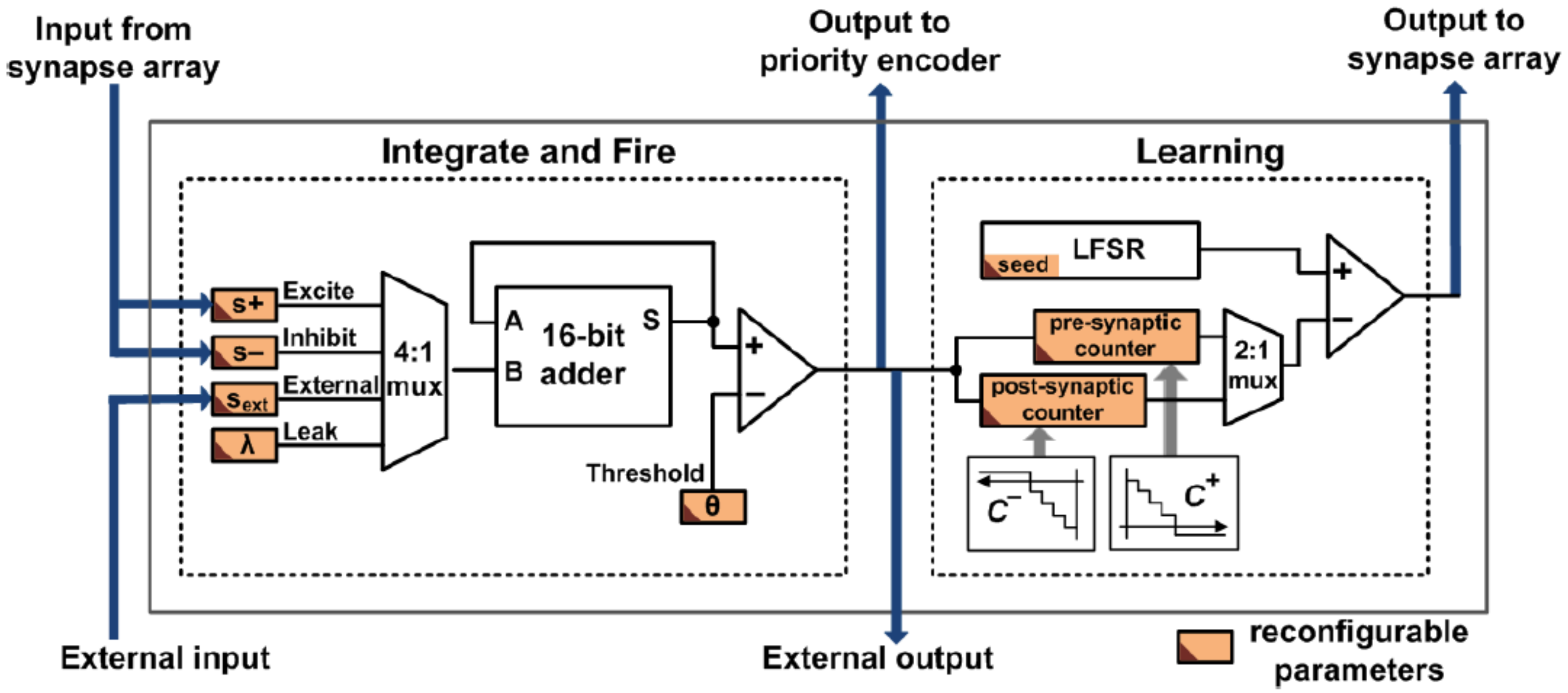
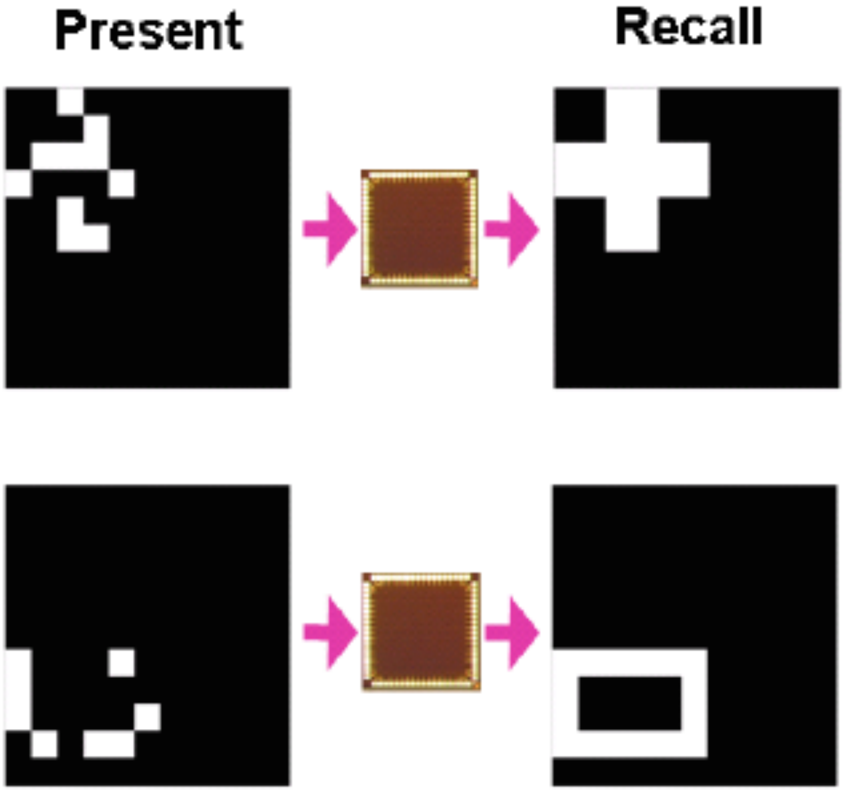


Fig. 3. Digital CMOS neuron with reconfigurability.

$$V(t) = V(t-1) + s_+ n_+(t) - s\_n\_(t) - \lambda$$

CMOS Neuromorphic chip for Spiking Neural Networks [3]- STDP learning



(a) present and recall of patterns

References

[1] Merolla, Paul A., et al. "A million spiking-neuron integrated circuit with a scalable communication network and interface." *Science* 345.6197 (2014): 668-673.

[2] Farabet, Clément, et al. "Neuflow: A runtime reconfigurable dataflow processor for vision." *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*. IEEE, 2011.

[3] Seo, Jae-sun, et al. "A 45nm CMOS neuromorphic chip with a scalable architecture for learning in networks of spiking neurons." *Custom Integrated Circuits Conference (CICC), 2011 IEEE*. IEEE, 2011.