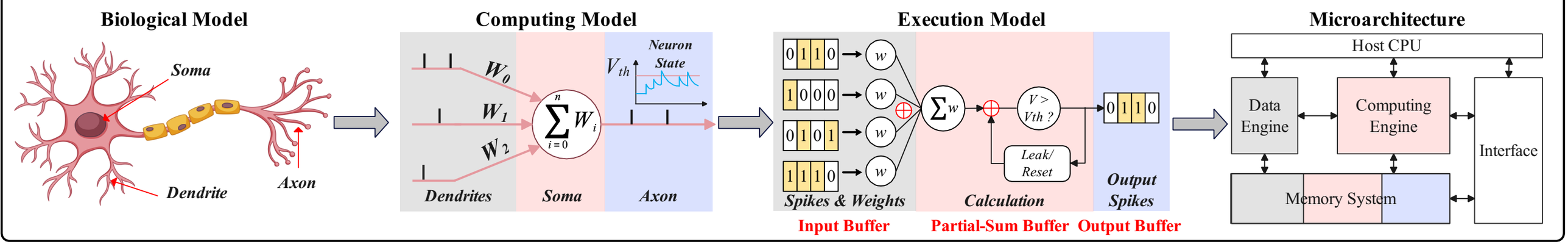
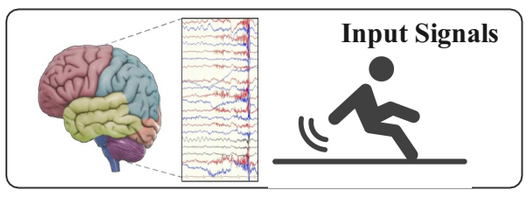
ExSpike-S: An Area-Efficient Neuromorphic Processor with Dense–Sparse Scheduling and a Shared Dual-Mode PE Array

Yuehai Chen and Farhad Merchant
Bernoulli Institute and CogniGron, University of Groningen, the Netherlands

From Biological Neurons to Neuromorphic Microarchitecture



Lightweight SNN (L-SNN)



Spiking Encoder

- 1D Depthwise Conv
- Batch Normalization
- Spiking Neurons
- 1D Pointwise Conv
- Batch Normalization
- Spiking Neurons
- Max Pooling

Depthwise Separable (DS) Block 1

- 1D Depthwise Conv
- Batch Normalization
- Spiking Neurons
- 1D Pointwise Conv
- Batch Normalization
- Spiking Neurons
- Max Pooling

DS Block 2

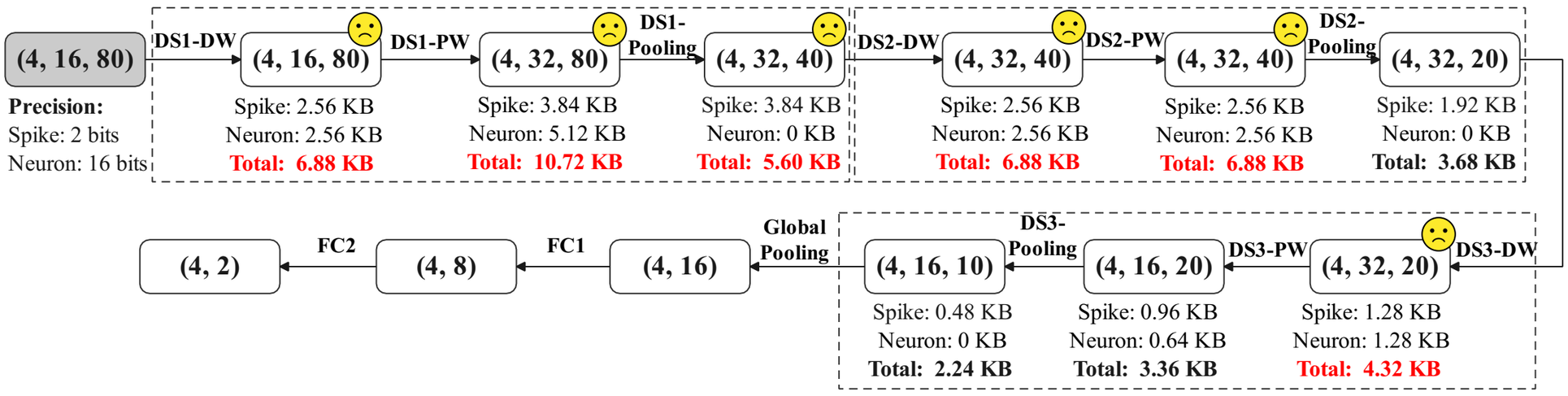
- 1D Depthwise Conv
- Batch Normalization
- Spiking Neurons
- 1D Pointwise Conv
- Batch Normalization
- Spiking Neurons
- Max Pooling

DS Block 3

- FC Layer 1
- FC Layer 2

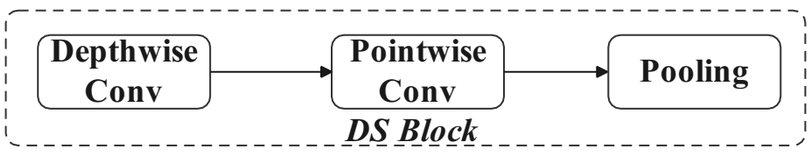
No. of Layers	11
Input Shape	(4, 16, 80)
Parameters	2,787
Precision	INT4
Memory Size	1.76 KB

Detailed Layer-Wise Memory Analysis

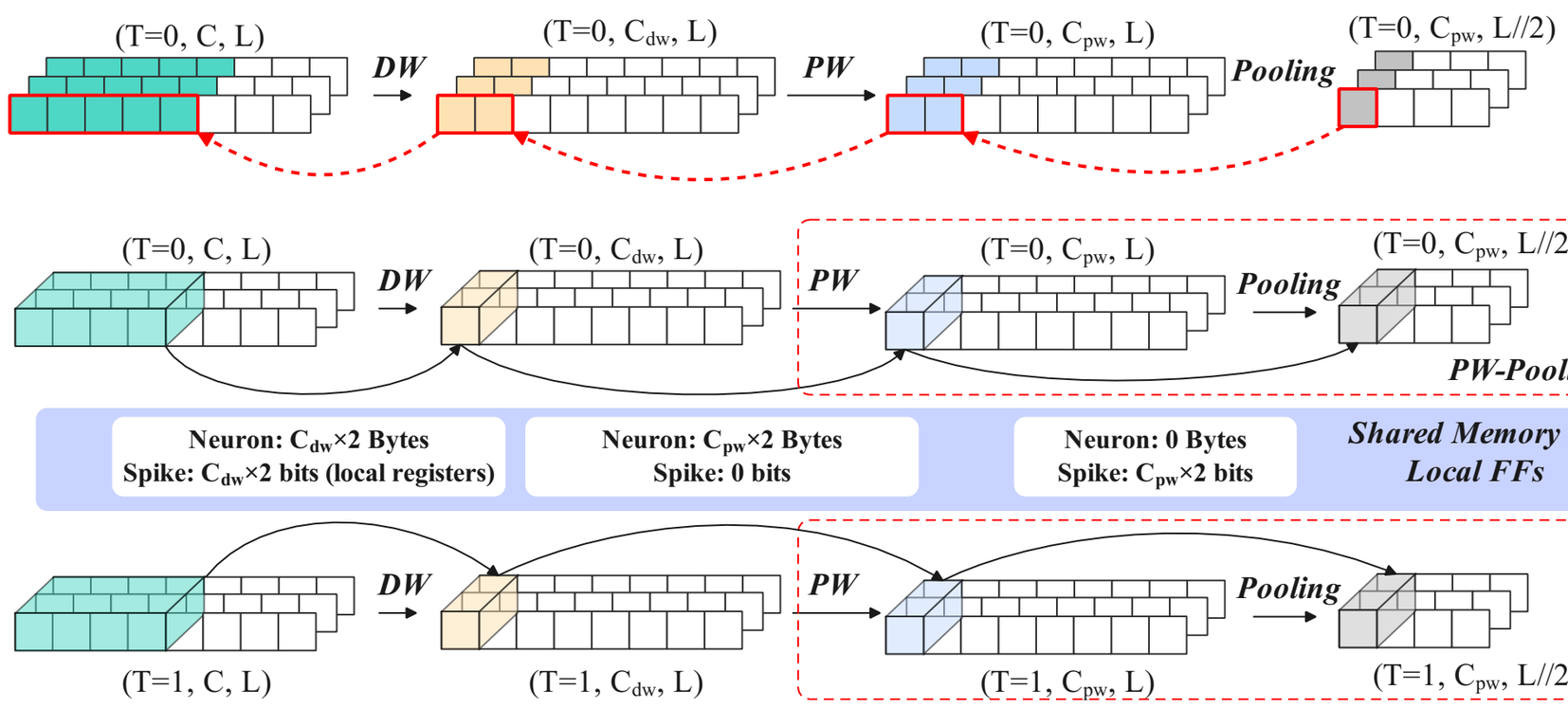


Challenge: Under a specific memory budget (e.g., 4 KB), only a few layers of the L-SNN can be mapped at a time, primarily the final layers and pooling layers.

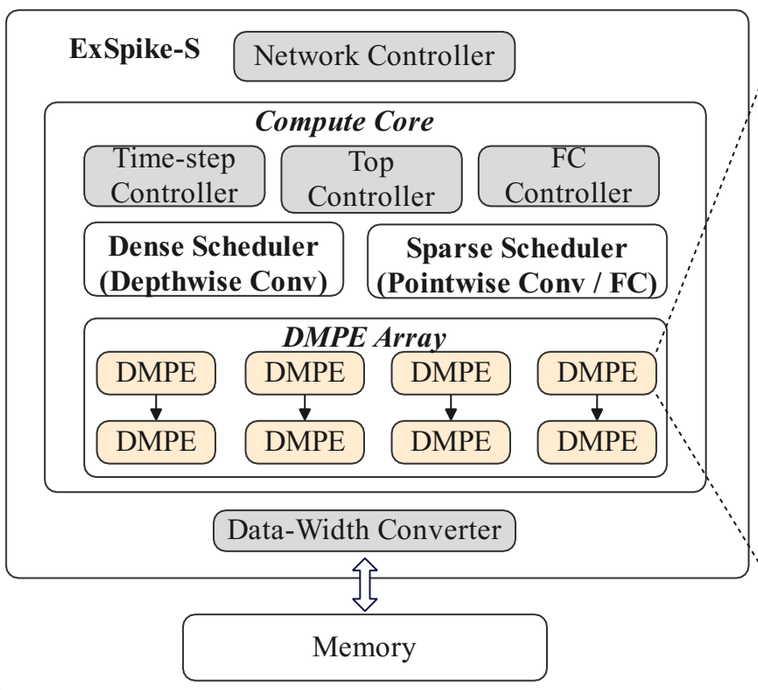
Inter-DS-Block Pipelining with Intra-Block Operation Fusion

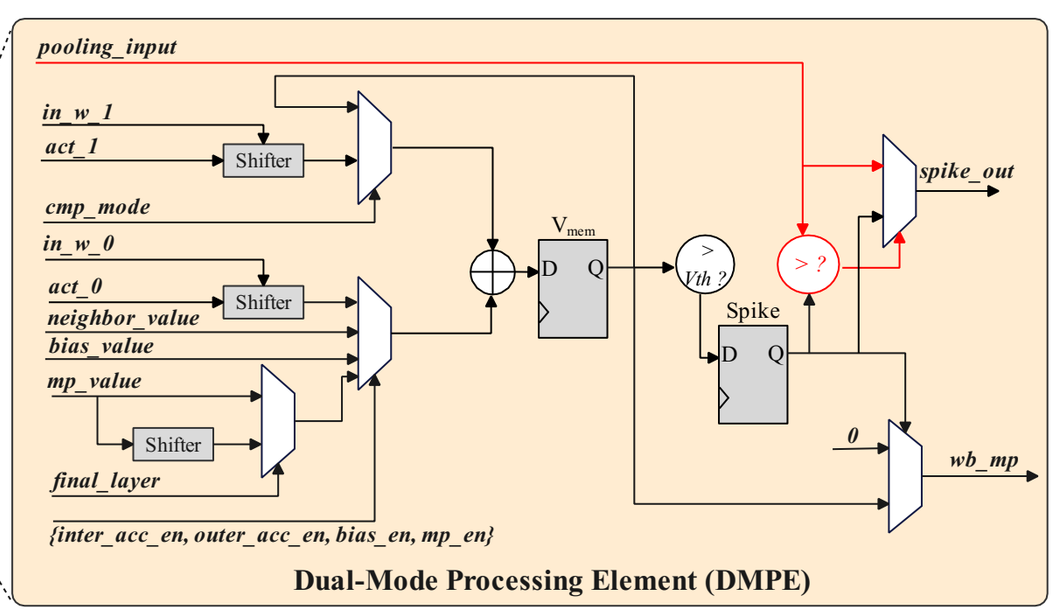


Solution: Based on the computation flow within a DS block, once a layer generates an output token, the token can be sent directly to the next layer, reducing the neuron-state memory overhead.



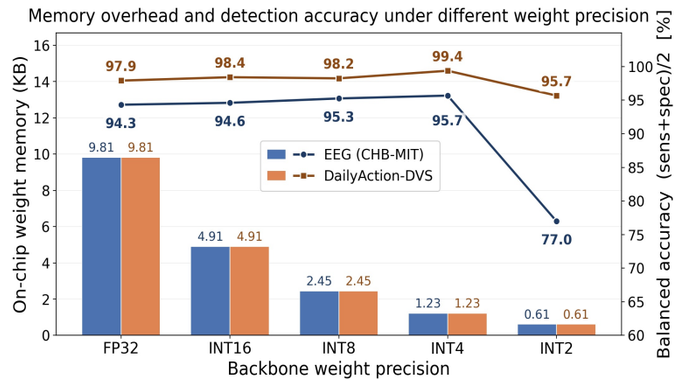
Architecture: Separate Dense and Sparse Schedulers with a Shared Dual-Mode PE Array





Evaluation

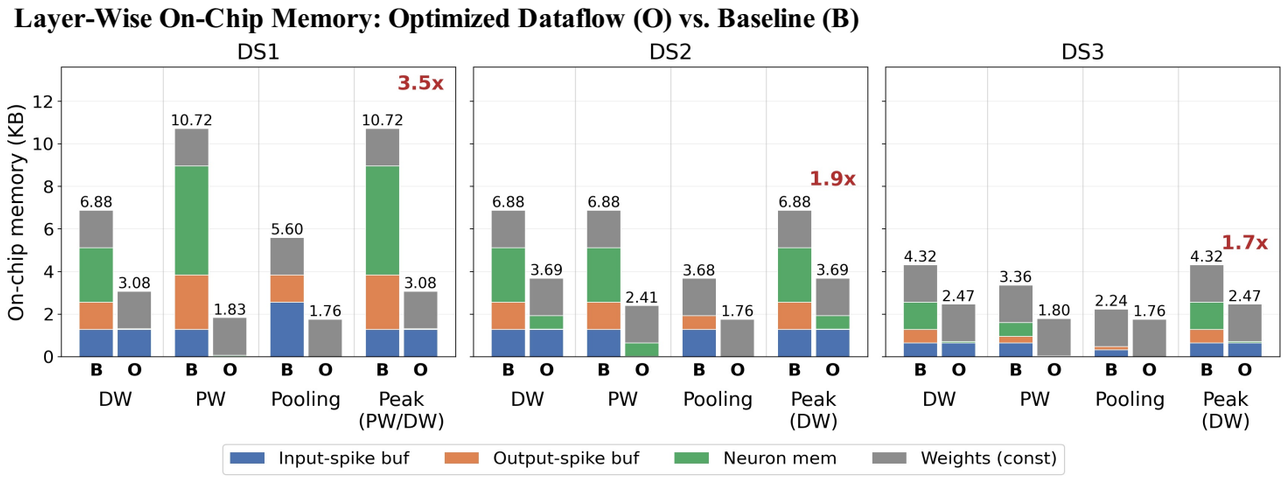
Quantization at Different Precisions



Resource Breakdown of the FPGA Prototype

	Dense Scheduler	Sparse Scheduler	PE Array	Total
LUTs	191	382	712	1966
FFs	168	110	144	625

Layer-Wise On-Chip Memory: Optimized Dataflow (O) vs. Baseline (B)



	ISSCC'23 [1]	JSSC'24 [2]	TVLSI'26 [3]	ExSpike-S
Tech	40 nm	180 nm	40 nm	130 nm
Data Width	10/16 bit	10 bit	8 bit	4 bit
Area	6 mm^2	0.3 mm^2	0.21 mm^2	0.12 mm^2
Freq.	4 MHz	N/A	5 MHz	100 MHz
Power	N/A	17.6µW	73.71µW	20.7 mW
Paradigm	DL	ML	SNN	SNN

References:

[1] C.-W. Tsai et al., "SciCNN: A 0-shot-retraining patient-independent epilepsy-tracking SoC," in ISSCC.
[2] A. Dabbaghian and H. Kassiri, "Modular DR- and CMR-boosted artifact-resilient EEG headset with distributed pulse-based feature extraction and neuro-inspired boosted-SVM classifier," in JSSC.
[3] H. Zhang et al., "QSNNA: An Energy-Efficient Quaternary Spiking Neural Network Accelerator for Seizure Detection," in TVLSI.

Paper & Code
(ExSpike FPL'2026)