DATE 22
DESIGN, AUTOMATION & TEST IN EUROPE
14 – 15 March 2022 · on-site event
16 – 23 March 2022 · online event
The European Event for Electronic System Design & Test

SNE: an Energy-Proportional Digital Accelerator for Sparse Event-Based Convolutions

Alfio Di Mauro†, Arpan Suravi Prasad†, Zhikai Huang†, Matteo Spallanzani†, Francesco Conti †‡, Luca Benini†‡

†Dept. of Information Technology and Electrical Engineering, ETH Zurich, Switzerland

‡Dept. of Electrical, Electronic and Information Engineering, University of Bologna, Italy

ETH zürich

Outline

- **Introduction**
- **SNE overview**
- **Execution model**
- **Accelerator Architecture**
- **Results**
- **Conclusion**

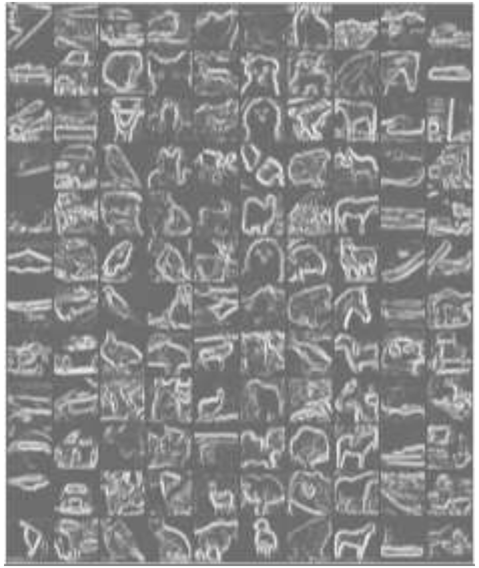
Event-based sensors

- **Event-based visual sensors transmit only the brightness change**
- **Information is encoded with low redundancy**
- **Such sensors are Low-power, fast responsive, low bandwidth**

CIFAR-10 Data set samples

https://www.cs.toronto.edu/~kriz/cifar.html

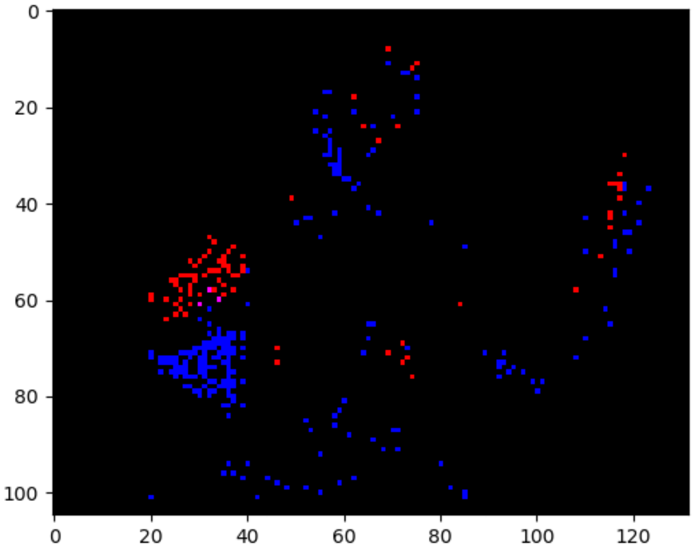
CIFAR-10-DVS Data set samples

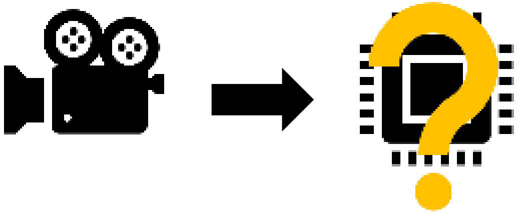


CIFAR10-DVS: An Event-Stream Dataset for Object Classification. Li Hongmin, Liu Hanchao, Ji Xiangyang, Li Guoqi, Shi Luping

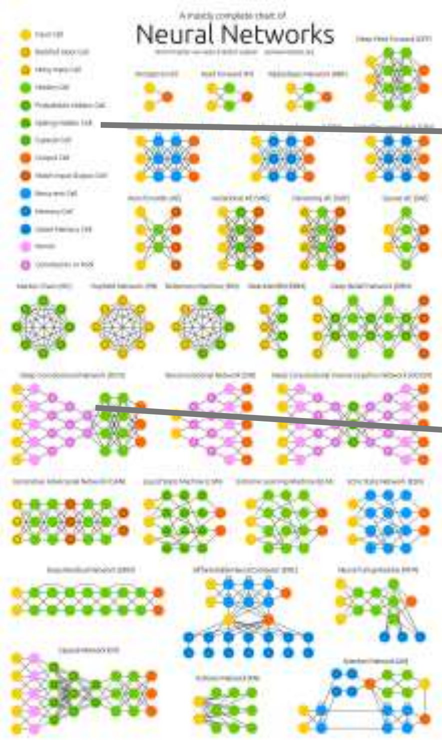
Energy-proportional computing

- In this framework, the information quantum is represented by a single event
- Additional information is carried by the inter-event time
- The amount of event is proportional to the activity
- Event-based sensor data are characterized by <u>unstructured</u> sparsity
- Conventional DNNs are not a good candidate to process the events: CPU or GPU can profit almost exclusively from <u>structured</u> sparsity





Convolutional Spiking Neural Network (C-SNN)



https://www.asimovinstitute.org/neural-network-zoo/

△ Spiking hidden cell

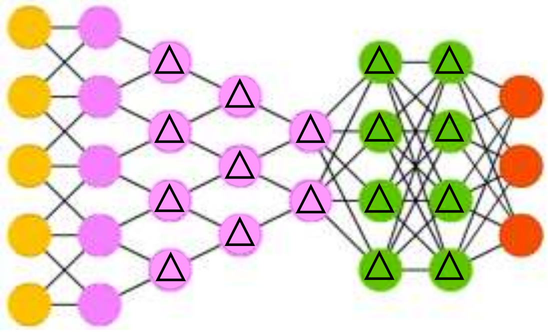
△ Spiking convolutional hidden cell

- Stateful behavior

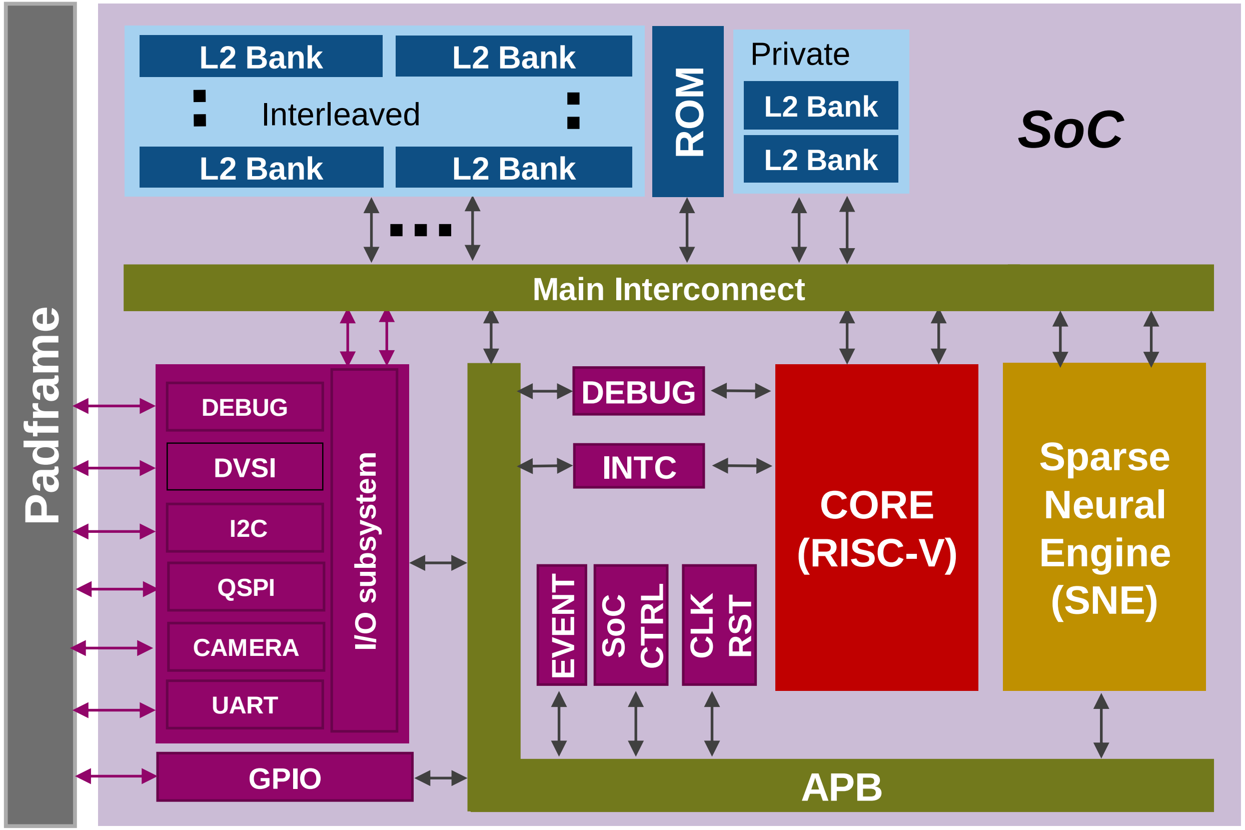
uted features

ous”

Deep Spiking convolutional network



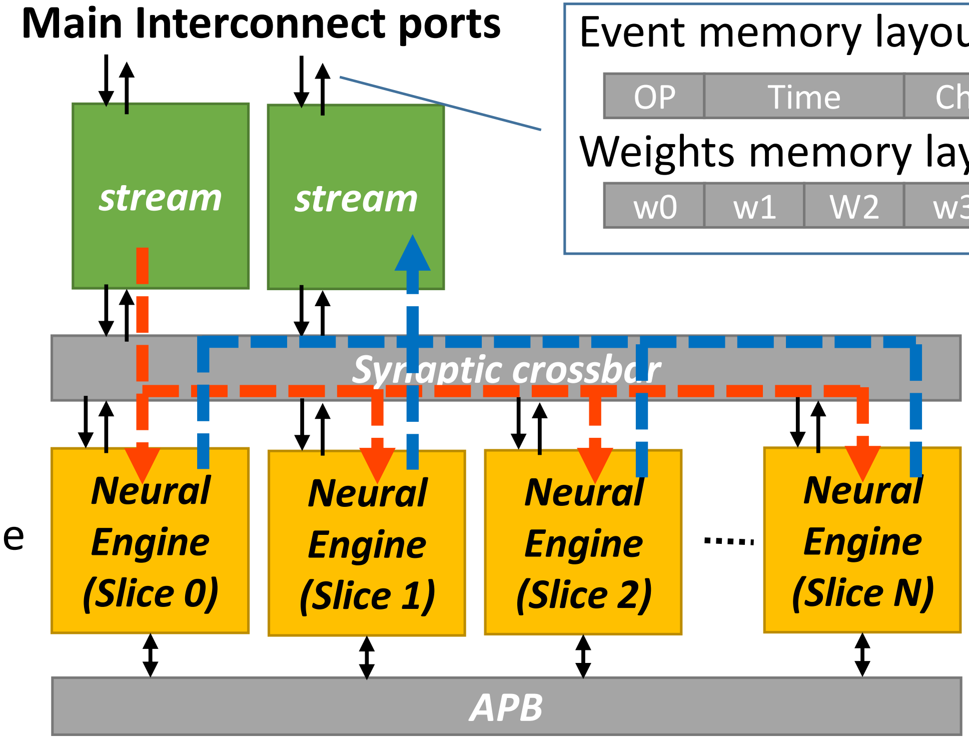
Heterogeneous SoCs with support for C-SNN



SNE architecture

Input event stream

- Fetched from the main memory
- Broadcasted to the LIF units



Event memory layout

OP	Time	Ch	Y	X

Weights memory layout

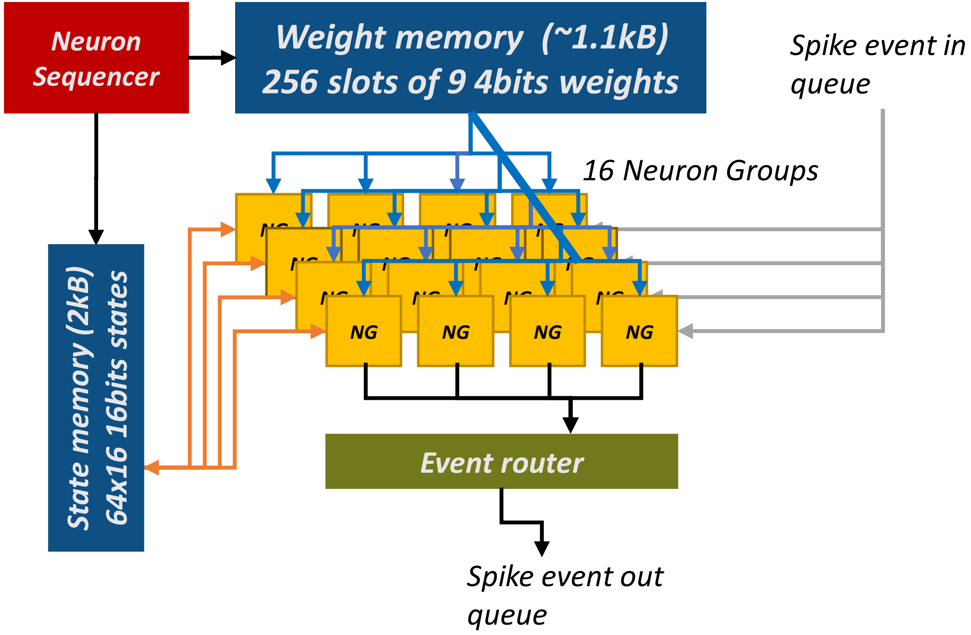
w0	w1	W2	w3	w4	w5	w6	w7

Output event stream

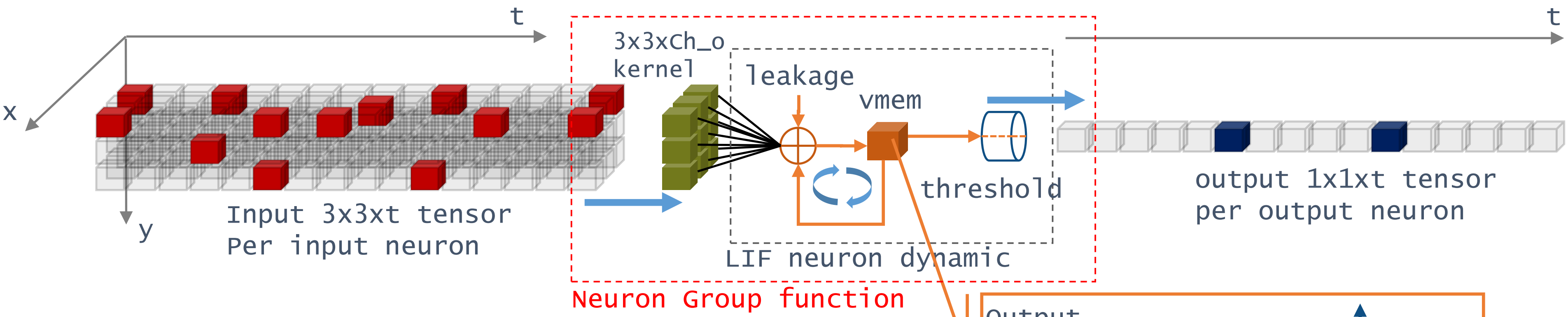
- Pushed back to the main memory

SNE computing slice

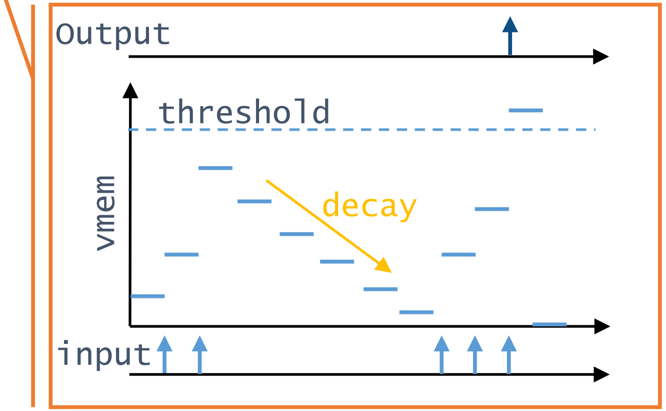
- Each engine has 16 **LIF neuron** data paths, and can execute (sustained performance) 16 Synaptic operation per cycle (SOP)
- Up to 256 convolutional kernels are stored locally
- update 64x16 neuron in 64 cycles (sustained performance)



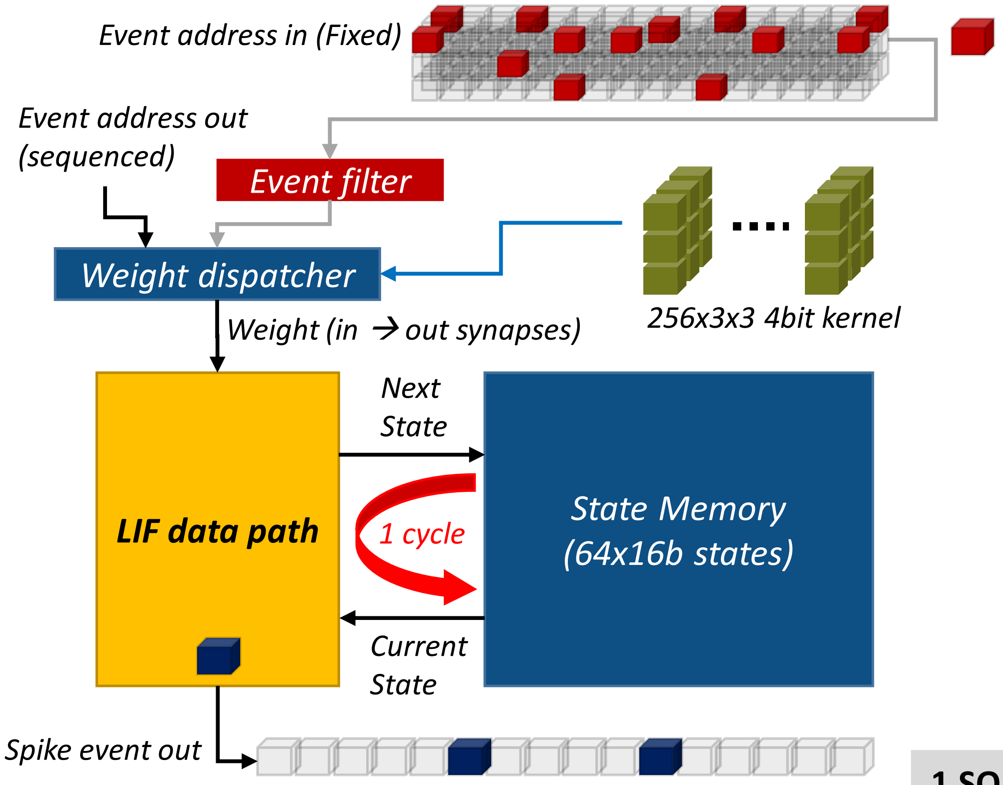
Single Neuron Operation (Conv)



- Each neuron has a 3x3 input receptive field
- “Empty” event operations are inherently skipped
- The membrane potential (vmem) is stored locally
- Linear vmem decay is obtained by iteratively subtracting a constant factor



Neuron Group Architecture



- 1 physical LIF neuron data path
- 64 neurons implemented in Time-Domain Multiplexing (TDM)
- Combinational data path, single cycle neuron state update
- 1 SOP implements:
  - weight accumulation on the neuron membrane potential
  - the leaky decay since the last membrane update
  - spike generation
  - membrane potential reset

1 SOP = 1 4b-ADD + 1 8b-MUL + 1 8b-SUB + 1 8b-COMPARE

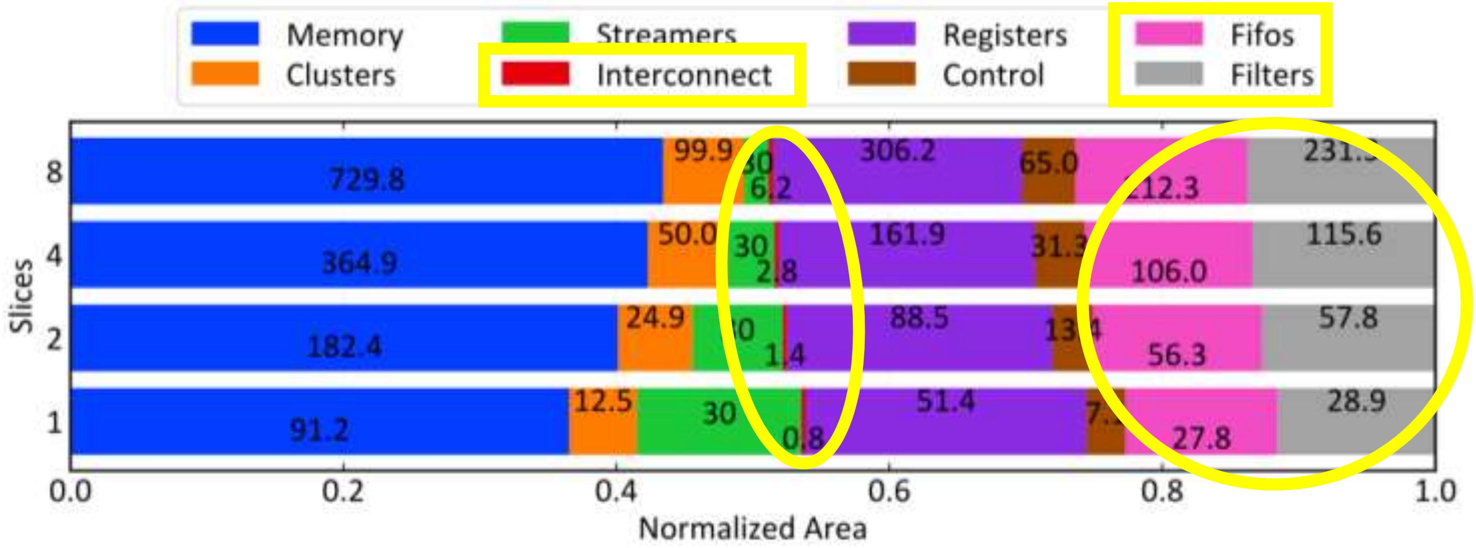
Experimental conditions

- We performed the exploration for 1, 2, 4, 8 slices, implementing, 1024, 2048, 4096, 8192 LIF neurons

- Logic Synthesis Tool: Synopsys Design Compiler 2020.09
- Technology node GlobalFoundries 22nm FDX process, 8T, 20, 24, 28, L, and SL voltage threshold cells
- Sign-off corner: SSG, 0.72V nominal supply voltage, -40C
- Target frequency: 400MHz

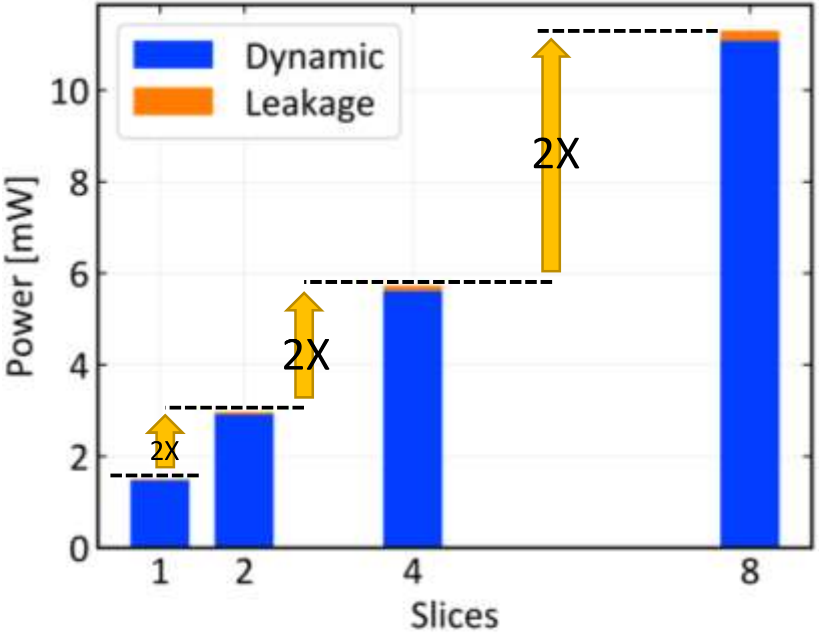
Power consumption estimates have been performed at the target 400MHz clock frequency, TT corner, 0.8V supply voltage, 25C, by using Synopsys Prime-Power 2019.12

Area results

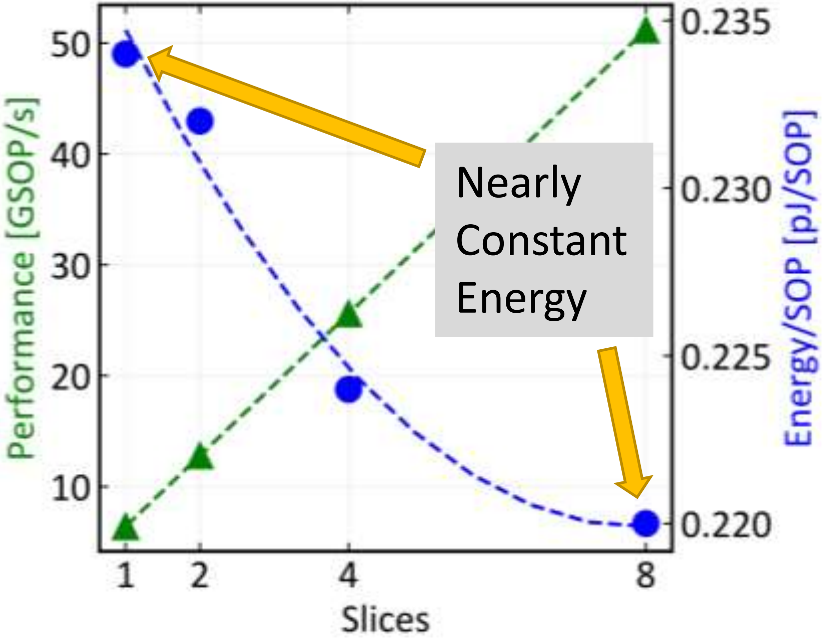


Linear area scaling, Including interconnect and event routing circuitry

Power and energy consumption results



Linear power scaling



Nearly Constant Energy

Linear Performance scaling

Comparison with the SoA

Name	Tech	Neuron Model	Type	Neuron Number	Energy pJ/SOP	Eff. TOP/s/W	Bits	Freq
SNE (this work)	**Dig. 22nm**	**LIF**	**Conv SNN**	**8192**	**0.221**	**4.54**	**4**	**400MHz**
Tianjic	Dig. 28nm	-	Hybrid	40000	6.18	1.28	8	300MHz
Truenorth	Dig. 28nm	EXP LIF	SNN	1M	27	0.046	1	Asynch
Loihi	Dig. 14nm	LIF+	SNN	131072	23	-	1-64	Asynch
Spinnaker 2	Dig. 22nm	Prog.	DNN/SNN	-	1700	3.6	Var.	200MHz

Conclusion

In this work we presented SNE, a Sparse Convolutional Neural Network accelerator for event-based computation.

We demonstrated an architectural solution that is inherently energy-proportional and that allows to increase the weight and input event reuse.

We demonstrated that SNE achieves a SoA Energy/SOP (221 fJ – 235fJ), while scaling proportionally it’s size.

Thank you