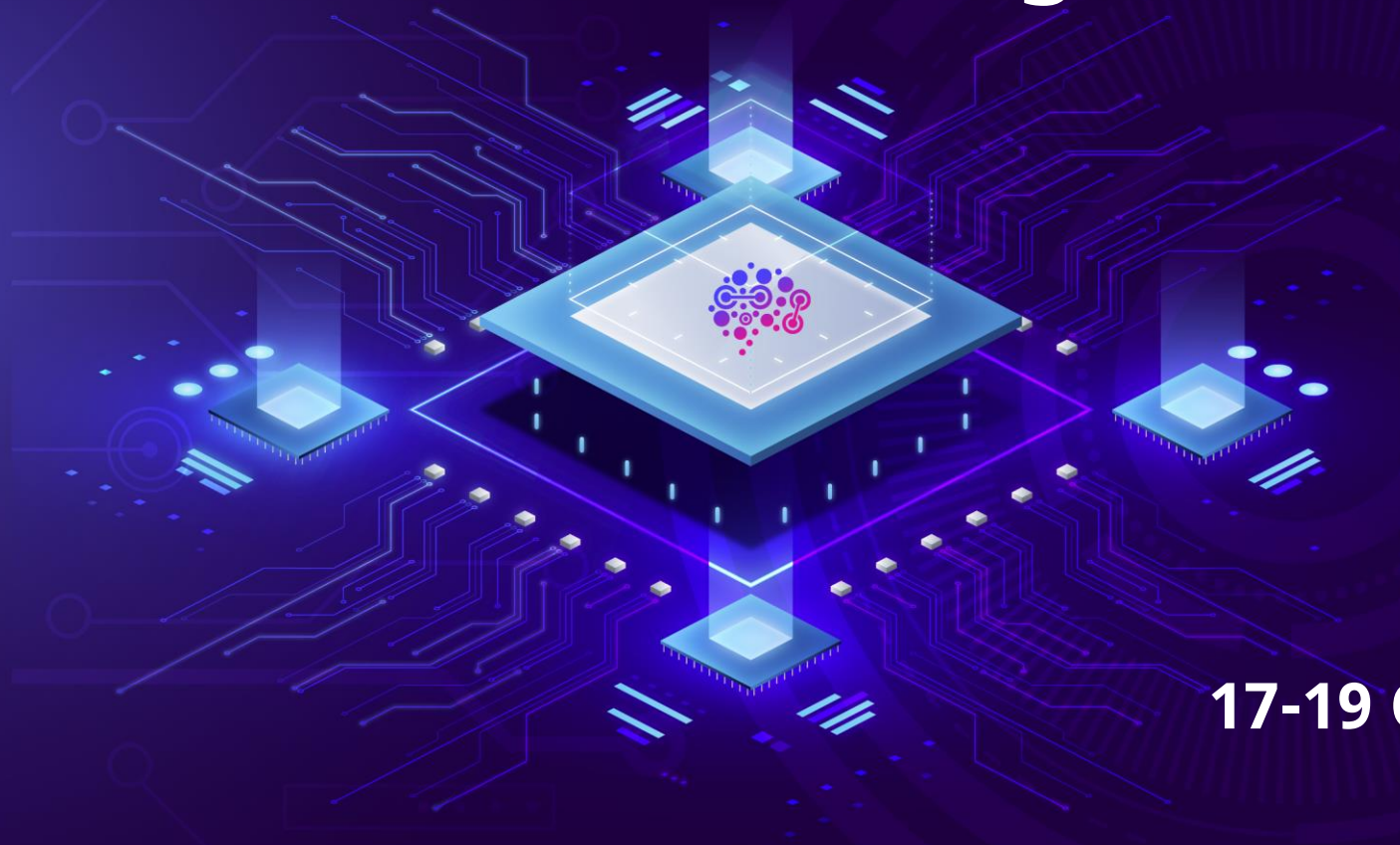
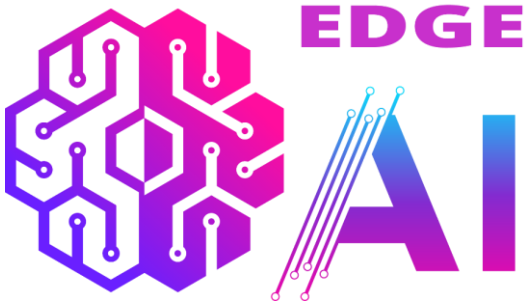
European Conference on EDGE AI Technologies and Applications - EEAI



Advancing emerging edge AI technologies and driving next generation intelligent applications.

17-19 October 2023, Athens, Greece

European Conference on EDGE AI Technologies and Applications - EEAI



SENECA: Building a fully digital neuromorphic processor, design trade-offs and challenges

Guangzhi Tang, imec, The Netherlands

17-19 October 2023 Athens, Greece

Copyright © 2023

Neuromorphic and Hardware Efficient AI Research

@ imec the Netherlands

Holst Centre
Powered by imec & TNO

µBrain [1]
- Asynchronous
- Lightweight

Argus [2]
- Multicore
- Sensor-fusion

SENECA [3]
- Flexible
- Scalable

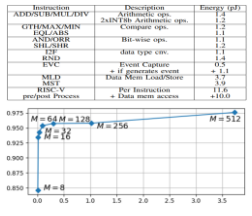
Increase Sensor and Compute Complexity

Lightweight Sensors

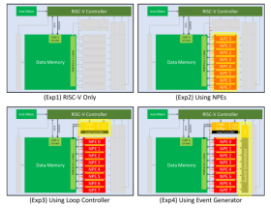
High-resolution Sensors

[1] Stujit, et al., Frontiers in Neuroscience, 2021.
[2] Detterer, et al., in submission.
[3] Tang, et al., Frontiers in Neuroscience, 2023.

HW-Algorithm Co-optimization

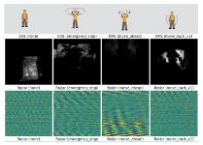


Instruction-level benchmarking [4]

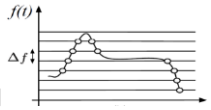


Design space exploration [5]

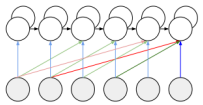
Neuromorphic Algorithms & Datasets



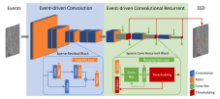
Sensor Fusion Dataset [6]



Delta DNN [8]

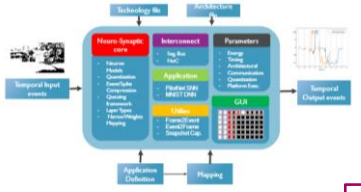


Synaptic Delay [7]



Event RNN [9]

Software Design



Event-driven Simulator [10]

SDK & Mapper

[4] Tang, et al., ISCAS, 2023.
[5] Shidqi, et al., under review.
[6] Muller, et al., RadarConf, 2023.
[7] Patiño-Saucedo, et al., ISCAS 2023.
[8] Yousefzadeh, et al., IJCNN, 2022.
[9] Wang, et al., in submission.
[10] Nembhani, et al., under review.

Copyright © 2023

Neuromorphic and Hardware Efficient AI Research

@ imec the Netherlands

Holst Centre
Powered by imec & TNO

µBrain [1]
- Asynchronous
- Lightweight

Argus [2]
- Multicore
- Sensor-fusion

SENECA [3]
- Flexible
- Scalable

Increase Sensor and Compute Complexity

Lightweight Sensors

High-resolution Sensors

[1] Stujit, et al., Frontiers in Neuroscience, 2021.
[2] Detterer, et al., in submission.
[3] Tang, et al., Frontiers in Neuroscience, 2023.

HW-Algorithm Co-optimization

Instruction-level benchmarking [4]

Design space exploration [5]

Neuromorphic Algorithms & Datasets

Sensor Fusion Dataset [6]

Delta DNN [8]

Synaptic Delay [7]

Event RNN [9]

Software Design

Event-driven Simulator [10]

SDK & Mapper

Outlines

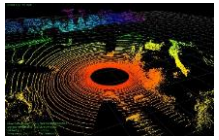
1. Event-based neuromorphic computing and design tradeoffs
2. SENECA neuromorphic architecture
3. Optimizing event-based neural network on SENECA

[4] Tang, et al., ISCAS, 2023.
[5] Shidqi, et al., under review.
[6] Muller, et al., RadarConf, 2023.
[7] Patiño-Saucedo, et al., ISCAS 2023.
[8] Yousefzadeh, et al., IJCNN, 2022.
[9] Wang, et al., in submission.
[10] Nembhani, et al., under review.

Copyright © 2023

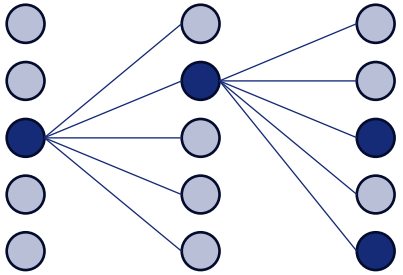
Event-based Neuromorphic Computing

Event Vision



Point Cloud

Event-based Neural Network



- **Event-driven**: Sensory input and activation sparsity.
- **Data-flow**: Data directly move to consuming location.
- **Near-memory**: Decrease data-movement distance.

Loihi2

TrueNorth

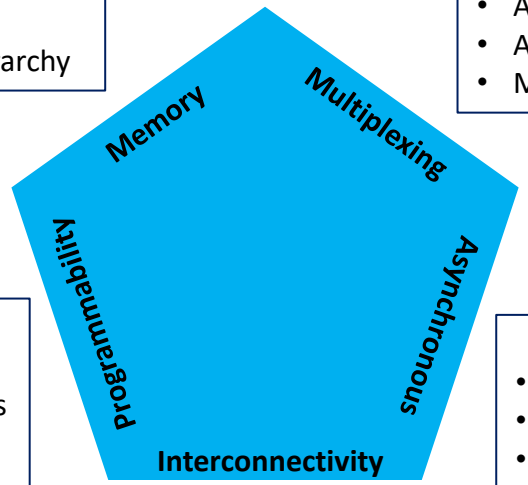
SpiNNaker

Tianjic

Tradeoffs of digital neuromorphic designs

Memory
- Size and flexibility
- Technology and hierarchy

Multiplexing
- Axon and neuron
- Area efficiency
- Memory and control cost



Programmability
- Events and precision
- Synaptic and neural models
- Architecture
- Learning

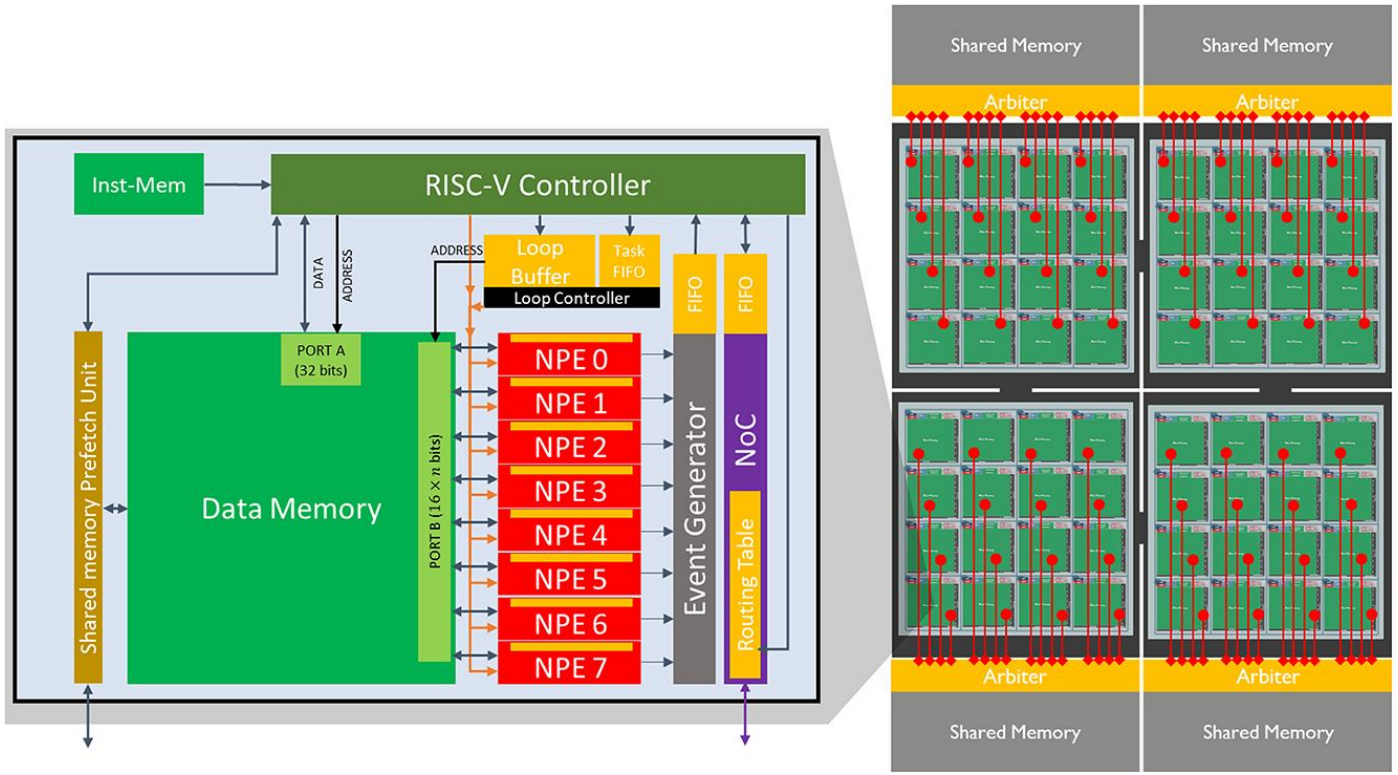
Asynchronous
- Full asynchronous
- Self-clocked cores, global synchronous
- Synchronous cores, global asynchronous

Interconnectivity
- Multicasting
- Compression

Tang, et al., Frontiers in Neuroscience, 2023.

Copyright © 2023

SENECA - Scalable Energy efficient NEuromorphic Computer Architecture



Goals

1. Hardware optimization on event-based computing
2. Flexible architecture for optimizations

- **Memory** – Unified, 3-level hierarchy with register, SRAM, shared memory
- **Multiplexing** – Axon and neuron
- **Programmability** – Fully programable for synapse, neuron, architecture, learning
- **Asynchronous** – Core-to-core asynchrony
- **Interconnectivity** – Multicasting NoC, software compression

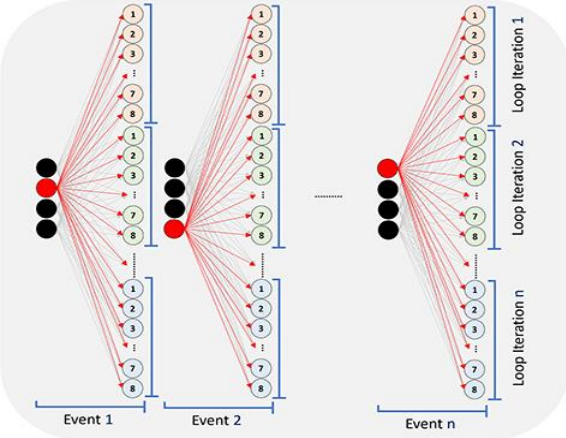
Tang, et al., Frontiers in Neuroscience, 2023.

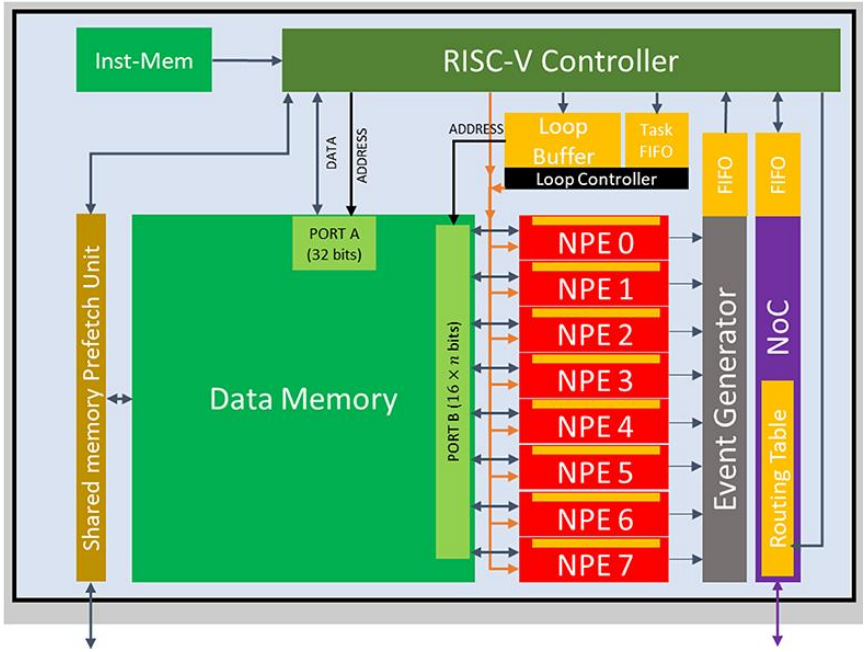
Copyright © 2023

Hardware optimization on event-based computing

Tang, et al., Frontiers in Neuroscience, 2023.

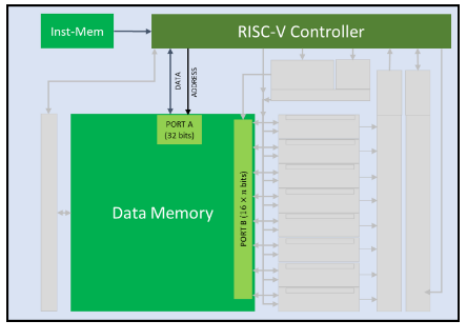
Event-driven Network
(Input event integrate to all post-synaptic states)



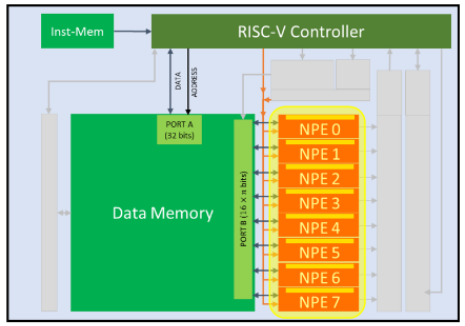


- Event Receiving
- Task Decoding
- Neural Processing
- Event Generation

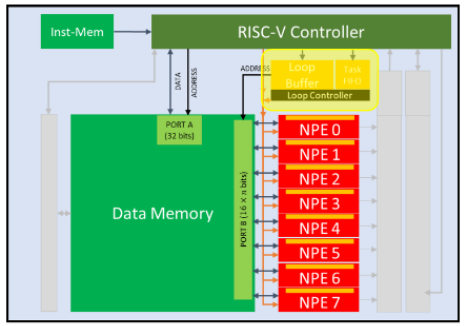
Shidqi, et al., under review.



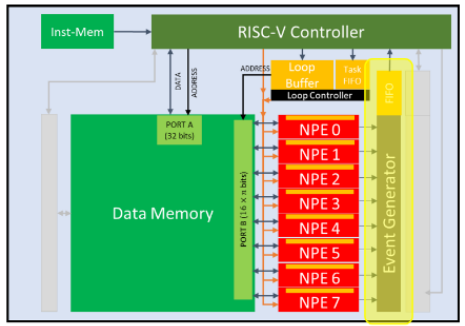
(Exp1) RISC-V Only



(Exp2) Using NPEs



(Exp3) Using Loop Controller



(Exp4) Using Event Generator

Experiment	Latency (µS)	Energy (µJ)
RISC-V Only	6625	34.0
+NPEs	1098	6.71
+Loop Controller	541	2.81
+Event Generator	**400**	**2.10**

>15x HW Improvement

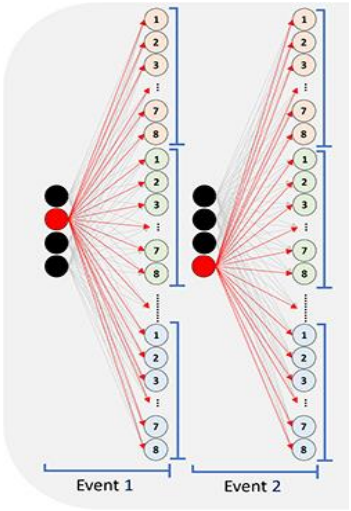
** Results generated by Cadence Xcelium and JOULES ** KWS Network: 390-256-256-29*

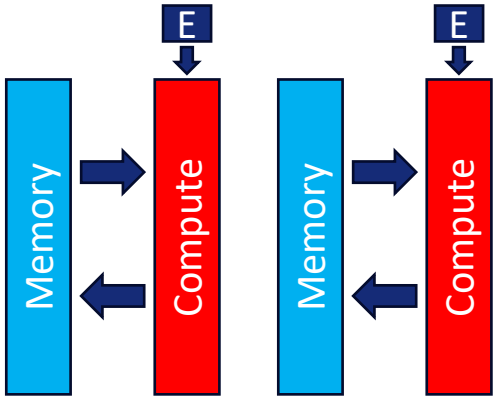
Copyright © 2023

Software optimization on event-based computing

Shidqi, et al., under review.

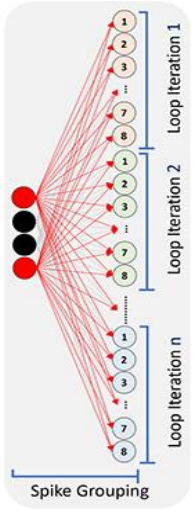
Event-by-event

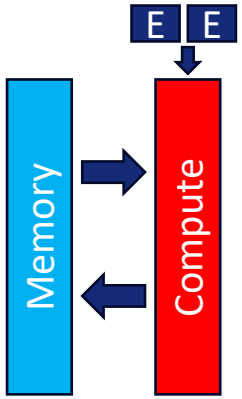




Two state memory access for each event

Spike-Grouping





Two state memory access for event batch

Optimizing single synaptic operation on SENECA

Experiment	SOp Time (µS)	SOp Energy (pJ)
Graded Event	228	14.2
Binary Spike	180	13.3
+Quantization	109	6.7
+Spike-Grouping	**58**	**2.7**

Keyword Spotting (KWS) Power Benchmark

Hardware	Accuracy	Inference/sec	Energy (µJ)
Jetson TX1	92.7%	419	11885
Loihi	93.8%	296	372
SpiNNaker2	93.8%	1000	7.1
SENECA	**93%**	**4587**	**1.2**

** Results generated by Cadence Xcelium and JOULES*

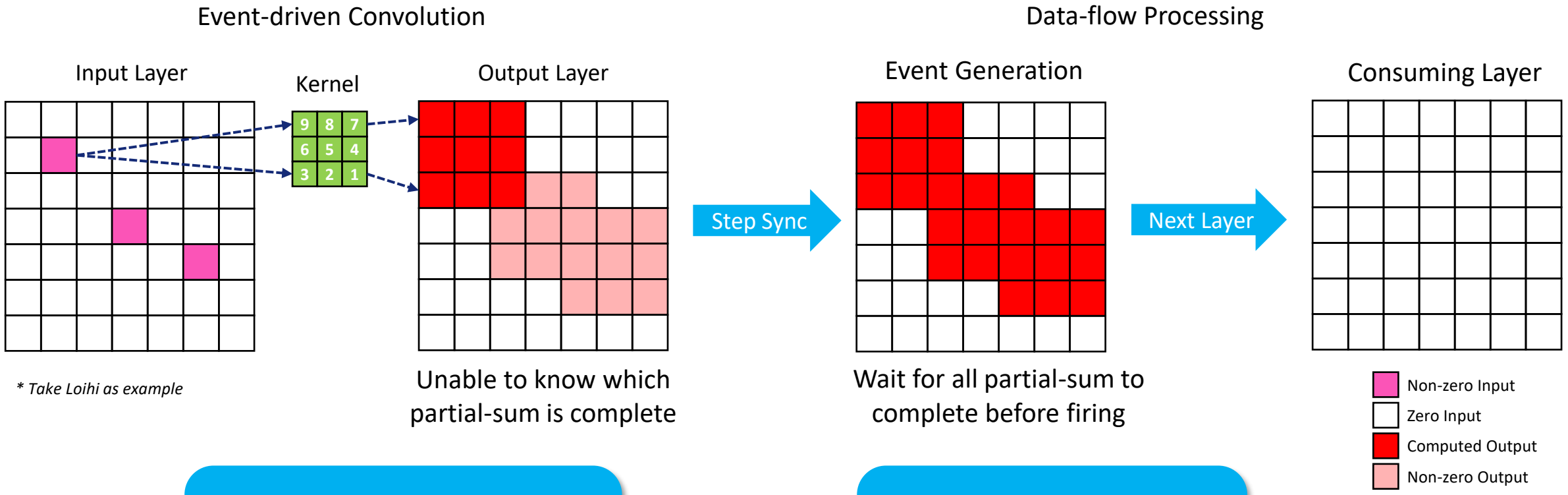
*** KWS Network: 390-256-256-29*

300x vs Loihi

6x vs SpiNNaker2

Copyright © 2023

Problems on event-driven CNN on neuromorphic HW



Problem 1: State (Partial-Sum) Memory

- Store all states on-chip
- Unbearable for high-res

Problem 2: Layer Latency

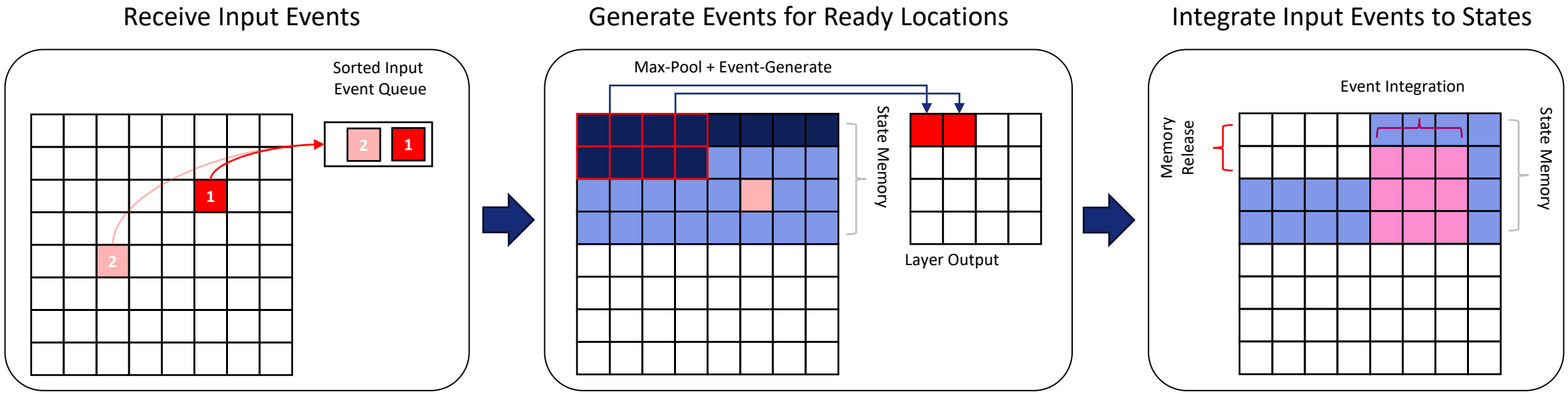
- Event generate after step sync
- Additional latency per layer

Shidqi, et al., under review.

Copyright © 2023

Optimizing event-driven CNN – Event-driven depth-first convolution

3x3 Convolution + 2x2 Max-Pooling



- Know completed neuron
- Immediately fire neuron

- Release memory after firing
- Reuse memory space

Neural state not in memory
Neural state in memory
Neural state complete input integration
Neural state ready to max-pool and fire
Output event location
Neural state integrate input events to partial sum

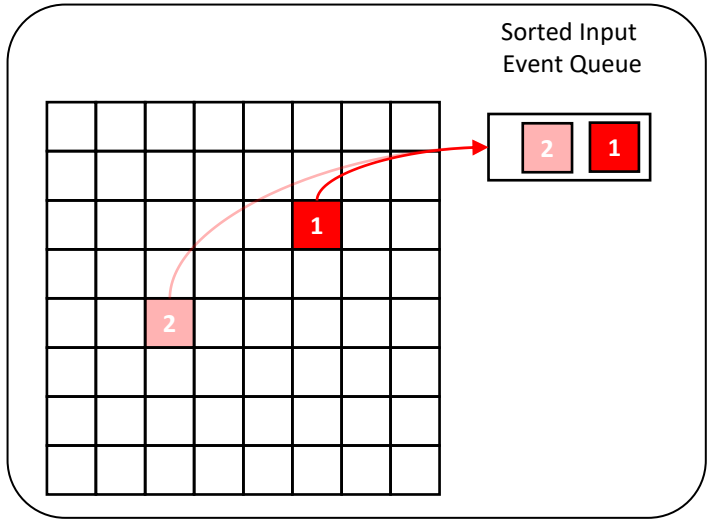
Shidqi, et al., under review.

Copyright © 2023

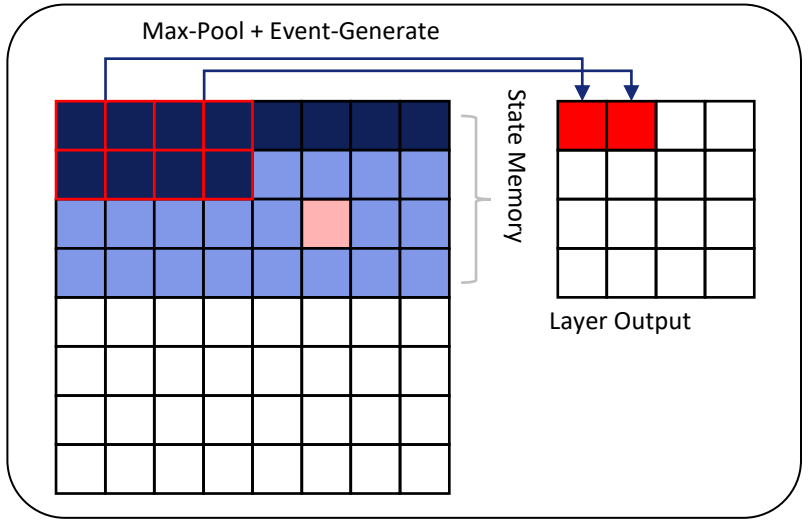
Optimizing event-driven CNN – Event-driven depth-first convolution

3x3 Convolution + 2x2 Max-Pooling

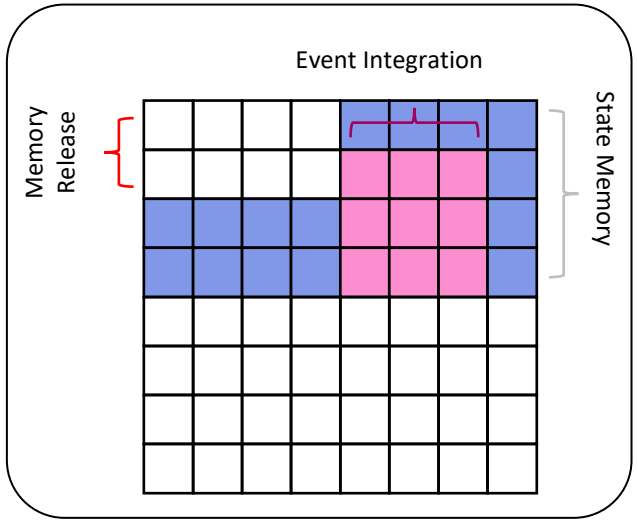
Receive Input Events



Generate Events for Ready Locations



Integrate Input Events to States



Improve vs	Energy	Latency	Area
Loihi	>50x	>3x	>3x
TrueNorth	>9x	>3x	>100x

** Results generated by Cadence Xcelium and JOULES*

*** 5-layer CNN: (8c3k2p)-(16c3k2p)-(32c3k2p)-128-class*

- Know completed neuron
- Immediately fire neuron

Reduce layer latency

Full layer to **First event**

- Release memory after firing
- Reuse memory space

Reduce state memory cost

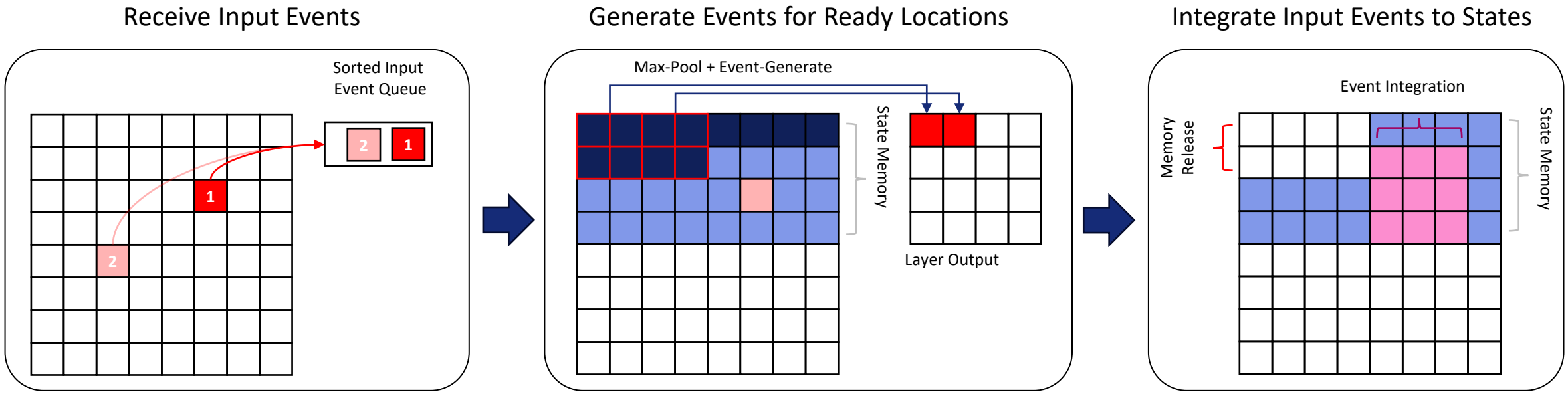
X*Y*C to **min(X, Y)*C*4**

Shidqi, et al., under review.

Copyright © 2023

Optimizing event-driven CNN – Event-driven depth-first convolution

3x3 Convolution + 2x2 Max-Pooling



Limitations:

1. **Input Preprocess Overhead**: Overheads on buffering and sorting input events.
2. **Multi-core Mapping**: Difficulties for synchronizing a multi-core single-layer mapping in asynchronous system.
3. **Stateful Neuron**: Only non-stateful convolutional layer can benefit on state memory reduction.

Copyright © 2023

Future Hardware-Algorithm Co-optimizations

Benchmarking event-driven neuromorphic algorithms using SENECA

Open the box of digital neuromorphic HW [1]

Instruction Energy

Instruction	Description	Energy (pJ)
ADD/SUB/MUL/DIV	Arithmetic ops. 2xINT8b Arithmetic ops.	1.4 1.2
GTH/MAX/MIN EQL/ABS	Compare ops.	1.2 1.1
AND/ORR SHL/SHR	Bit-wise ops.	1.1 1.2
I2F RND	data type cnv.	1.1 1.4
EVC	Event Capture + if generates event	0.5 + 1.1
MLD MST	Data Mem Load/Store	3.7 3.9
RISC-V pre/post Process	Per Instruction + Data mem access	11.6 +10.0

Hebbian Learning

Micro-kernel 1: ***Weight Update.*** *See Eq. (4).*

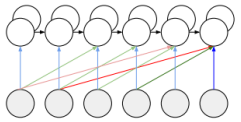
```
MLD(R0, ADD1, 0) //load weight w_ij
MLD(R1, ADD2, 1) //load trace{s_in,j}
MUL(R1, R1, R2) //trace{s_out,i} × trace{s_in,j}
MUL(R1, R1, R3) //η×R1, R3← η
ADD(R0, R0, R1) //update weight
MST(ADD1, R0, 1) //store R0 in w_ij
```

Micro-kernel 2: ***Spike Trace Update.*** *See Eq. (5).*

```
MLD(R0, ADD1, 0) //load trace
MLD(R1, ADD2, 1) //load input s
MUL(R0, R0, R2) //β × trace, R2← β
MUL(R1, R1, R3) //(1 − β) × s_i, R3← (1 − β)
ADD(R0, R0, R1) //update trace
MST(ADD1, R0, 1) //store R0 in trace
```

Help HW-Algorithm Co-design

Delay Network [2]

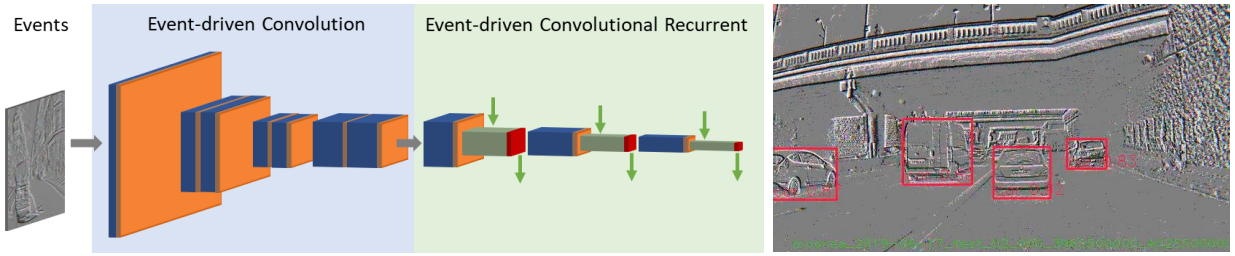


Recurrent vs Delay

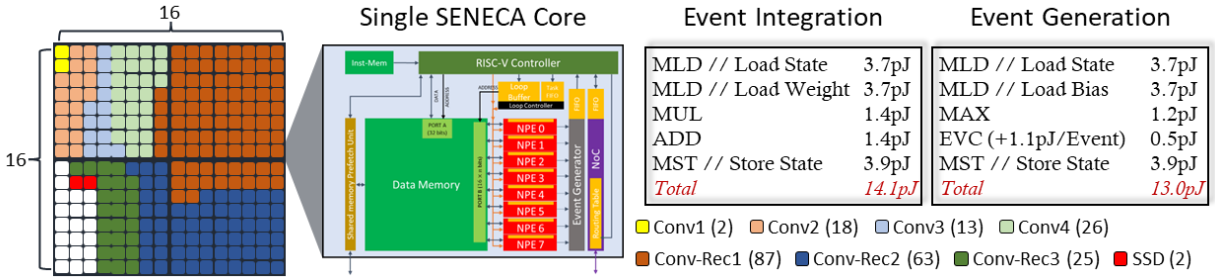
Measurement	R1	R2	D1	D2
neurons per hidden layer	128	48	8	8
number of delays	1	1	10	5
avg spk/timestep, layer1	8.678	6.725	1.894	1.686
avg spk/timestep, layer 2	4.582	3.456	1.772	2.539
max spk/timestep, layer 1	-	-	7	7
max spk/timestep, layer 2	-	-	7	8
test set accuracy	81.020	80.200	82.170	80.510
Neurosynaptic cost estimation				
energy (uJ)	20.213	7.390	2.304	1.745
memory (param. count)	141588	41684	7876	6756

- HW-algorithm co-optimization on novel algorithms
- HW IPs to further optimize event-based processing

Event-based Object Detection Network [3]



Our object detection prediction on Prophesee's 1Mpix Dataset



Scale the neuromorphic solution to real-world problems

[1] Tang, et al., ISCAS, 2023.
[2] Patiño-Saucedo, et al., ISCAS 2023.
[3] Wang, et al., in submission.

Copyright © 2023

Hardware Efficient AI Team @ imec the Netherlands

ORIGINAL RESEARCH article

Front. Neurosci., 23 June 2023
Sec. Neuromorphic Engineering
Volume 17 - 2023 |
https://doi.org/10.3389/fnins.2023.1187252

This article is part of the Research Topic
Spike-based learning application for neuromorphic engineering
View all 16 Articles ›

SENECA: building a fully digital neuromorphic processor, design trade-offs and challenges

Guangzhi Tang[1] Kanishkan Vadivel[1] Yingfu Xu[1] Refik Bilgic[2] Kevin Shidqi[1]
Paul Detterer[1] Stefano Traferro[1] Mario Konijnenburg[1] Manolis Sifalakis[1]
Gert-Jan van Schaik[1] Amirreza Yousefzadeh[1*]

[1] Imec, Eindhoven, Netherlands
[2] Imec, Leuven, Belgium

This technology is partially funded and initiated by the Netherlands and European Union's Horizon 2020 research and innovation projects **TEMPO** (ECSEL Joint Undertaking under grant agreement No 826655), **ANDANTE** (ECSEL Joint Undertaking under grant agreement No 876925), **DAIS** (Key Digital Technologies Joint Undertaking under grant agreement No 101007273) and **REBECCA** (Key Digital Technologies Joint Undertaking under grant agreement No 101097224).

Team Members

Amirreza

Guangzhi

Kanishkan

Paul

Gert-Jan

Manolis

Yingfu

Kevin

Alexandra

Students

Refik

Prithvish

Shenqi

Cina

Roy

...

Collaborations

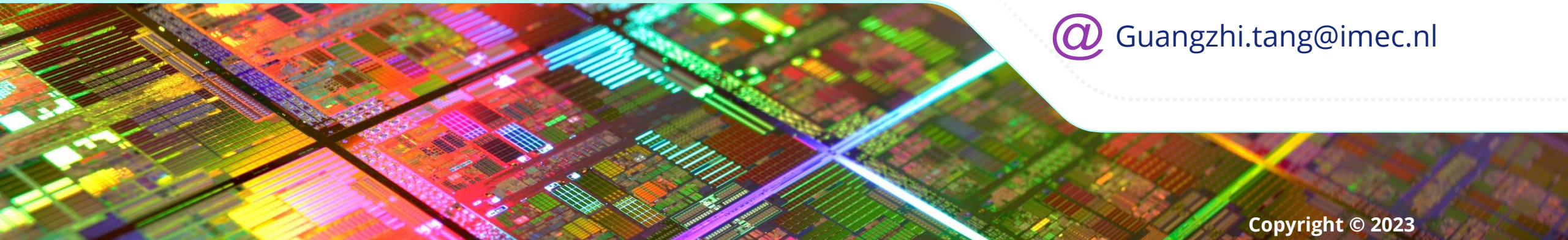
TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY
TU Delft
Radboud Universiteit

Copyright © 2023

EDGE AI

Thank You

For your attention



@ Guangzhi.tang@imec.nl

Copyright © 2023