Flexon: A Flexible Digital Neuron for Efficient Spiking Neural Network Simulations

Dayeol Lee[†], Gwangmu Lee[*], Dongup Kwon[*],
Sunghwa Lee[*], Youngsok Kim[*], and Jangwoo Kim[*]

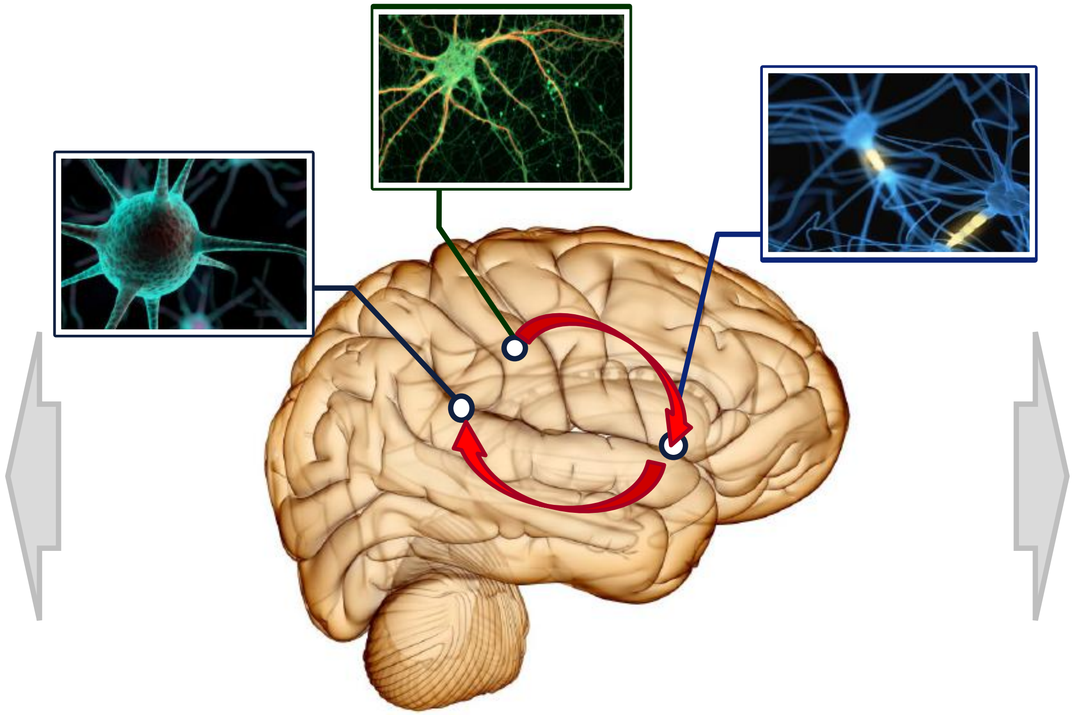
[*]Dept. of Electrical and Computer Engineering, Seoul National University
[†]Dept. of Electrical Engineering and Computer Sciences, University of California, Berkeley

Degeneration

Parkinson's Disease,
CJD, Dementia

Consciousness

Morality,
Social Value



Recognition

Object detection,
Classification

Emotion

Sympathy,
Happiness, Apathy

Neuroscience

The Study of **Neurons** and **Brains**

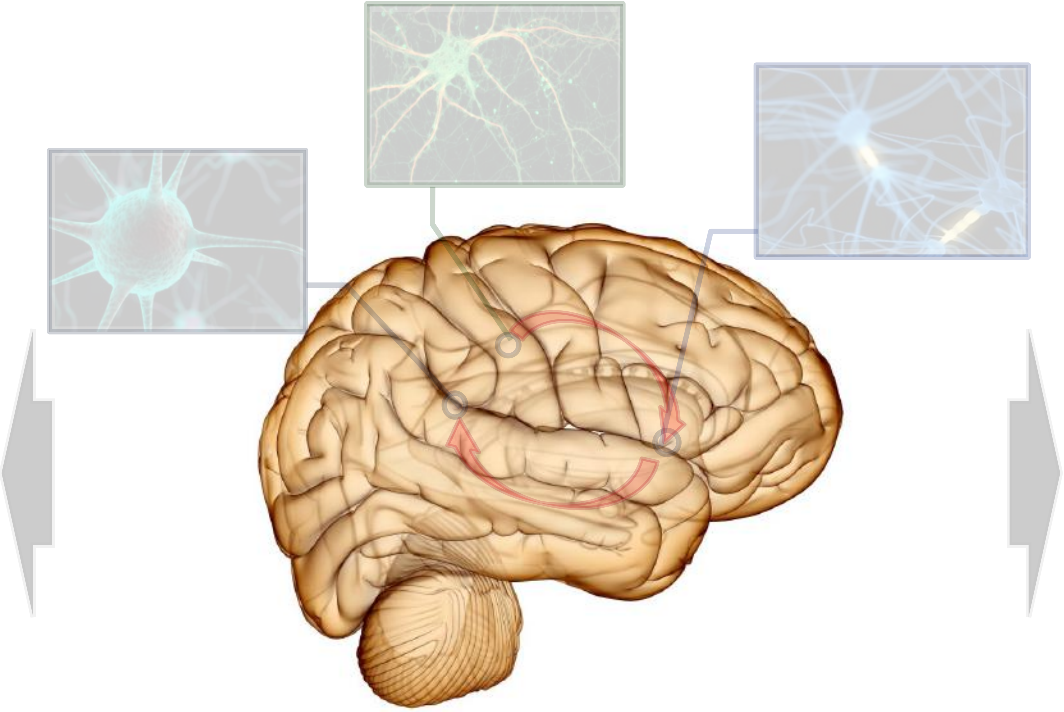
✓ **DNN** did this

Recognition

Object detection, Classification

Degeneration

Parkinson's Disease, CJD, Dementia



Consciousness

Morality, Social Value

Emotion

Sympathy, Happiness, Apathy

Neuroscience

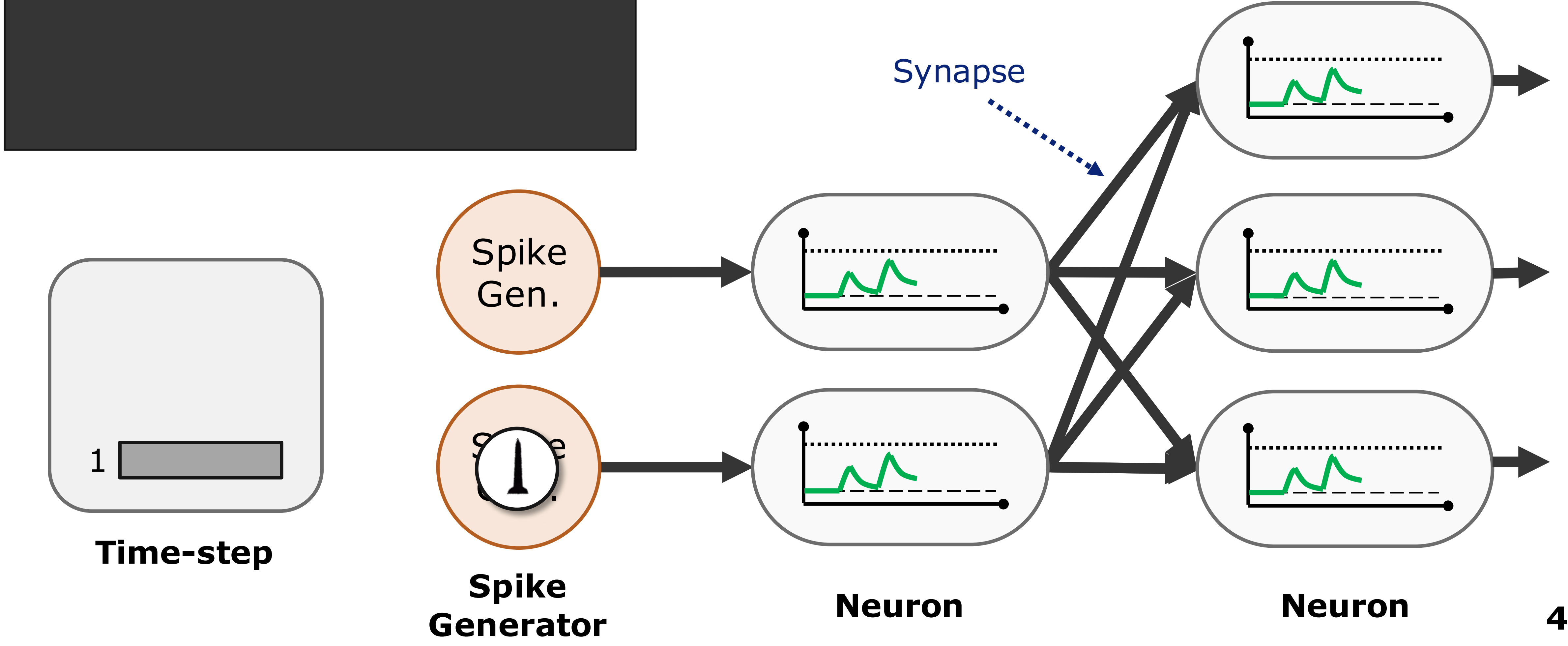
The Study of **Neurons** and **Brains**

? **Unexplored yet**

Modeling a Brain: Spiking Neural Network

How to compute spiking neural network?

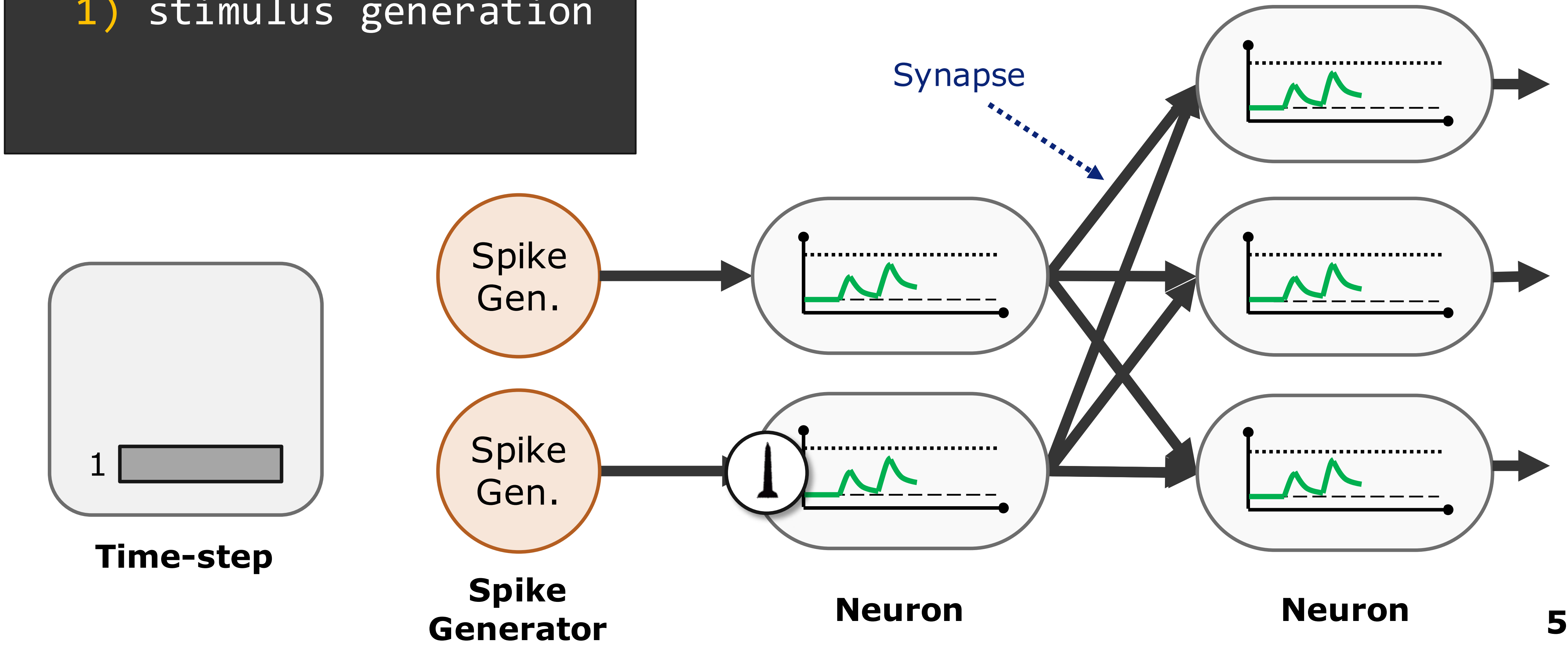
```
for each time-step:
```



Modeling a Brain: Spiking Neural Network

How to compute spiking neural network?

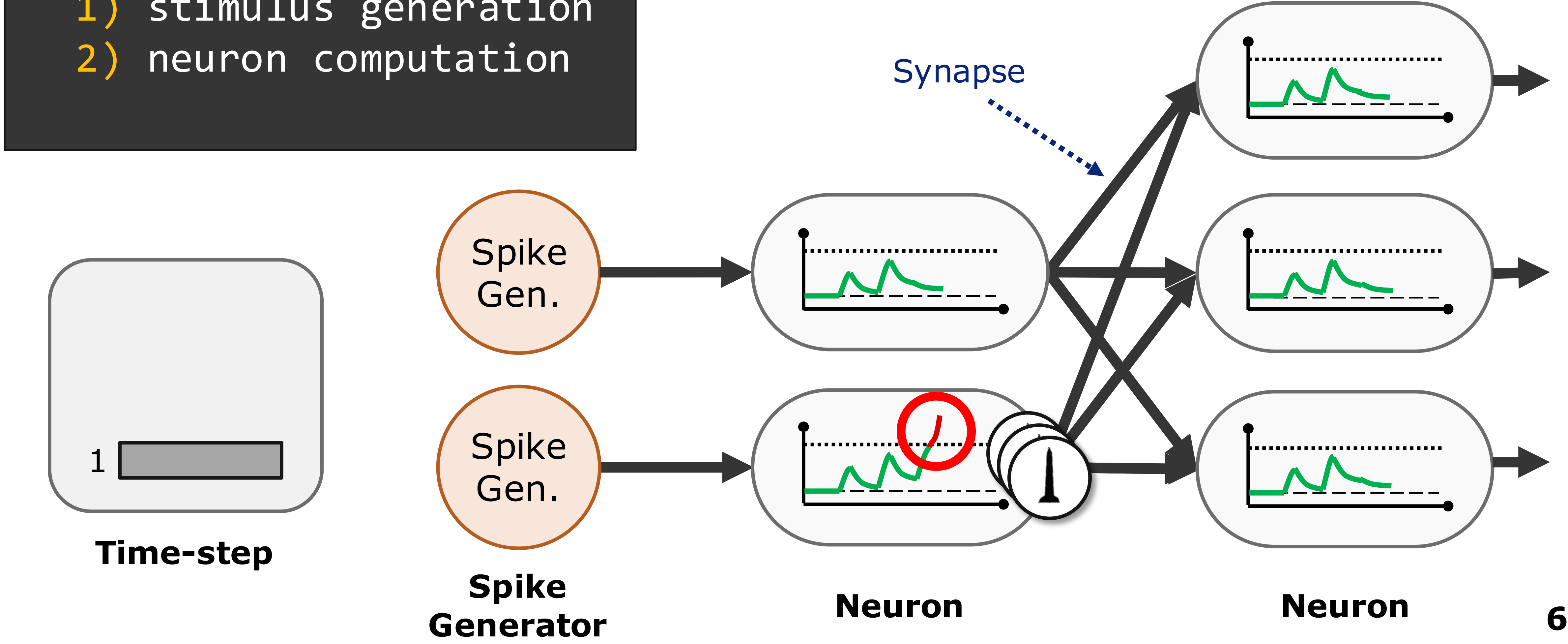
```
for each time-step:
  1) stimulus generation
```



Modeling a Brain: Spiking Neural Network

How to compute spiking neural network?

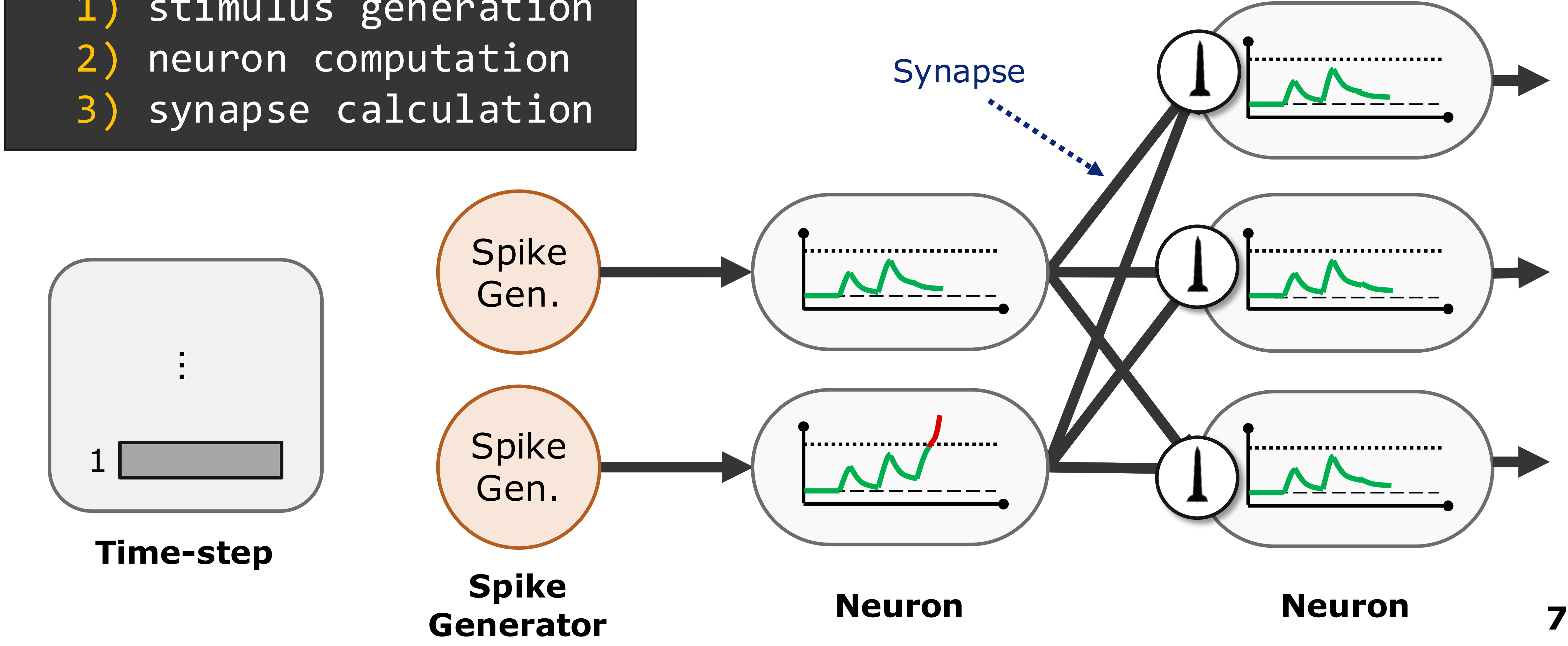
```
for each time-step:
  1) stimulus generation
  2) neuron computation
```



Modeling a Brain: Spiking Neural Network

How to compute spiking neural network?

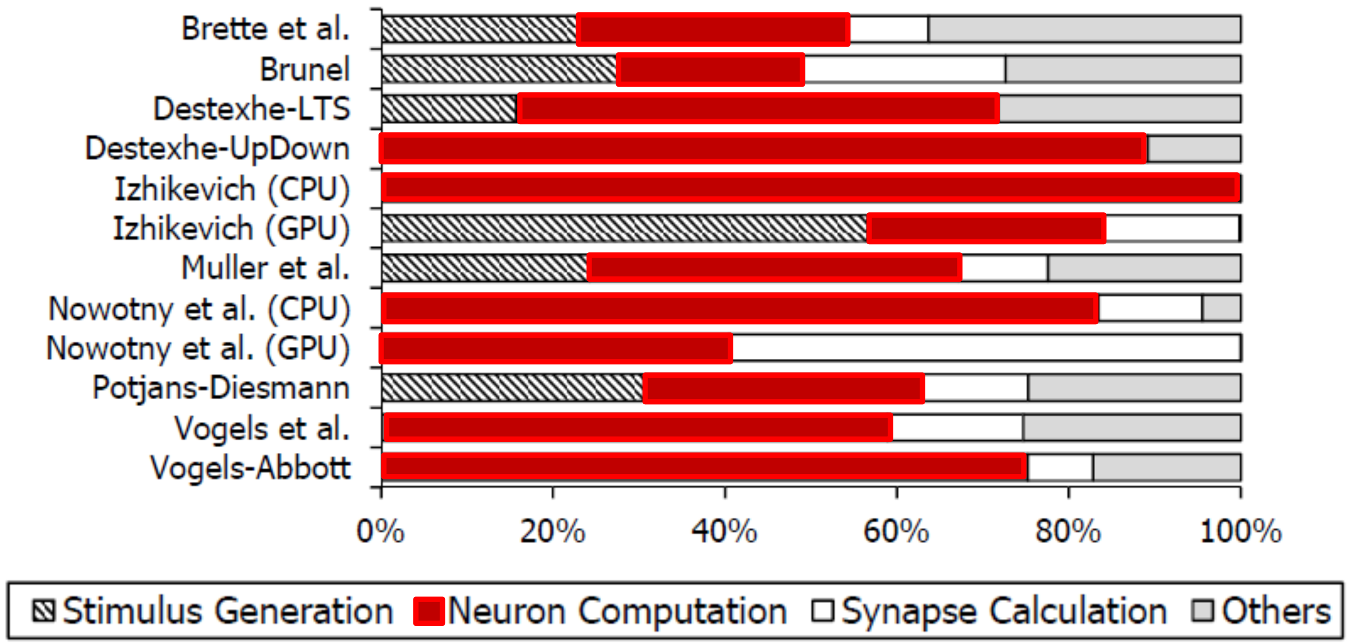
```
for each time-step:
  1) stimulus generation
  2) neuron computation
  3) synapse calculation
```



Where Does Time Go?

10 Representative Benchmarks on CPU/GPU

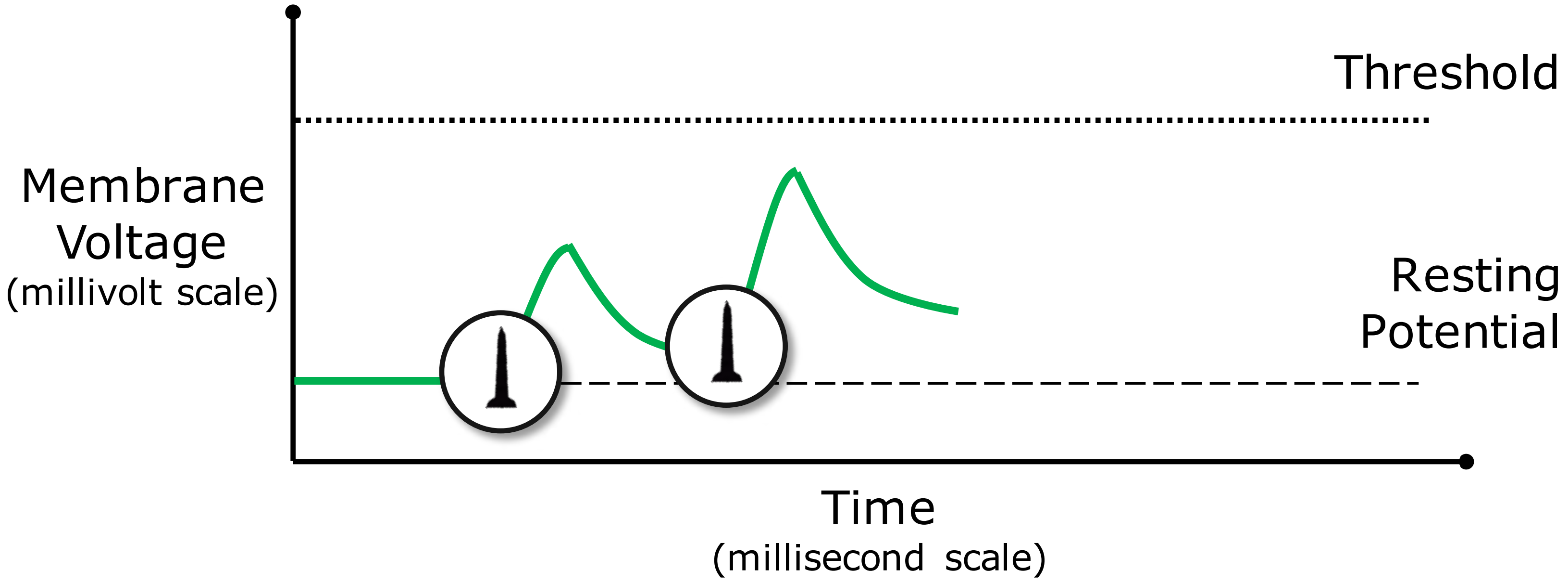
CPU: Intel Xeon E5-2630 v4 CPU (12-core, 2.2 GHz) / GPU: NVIDIA Titan X (Pascal) GPU

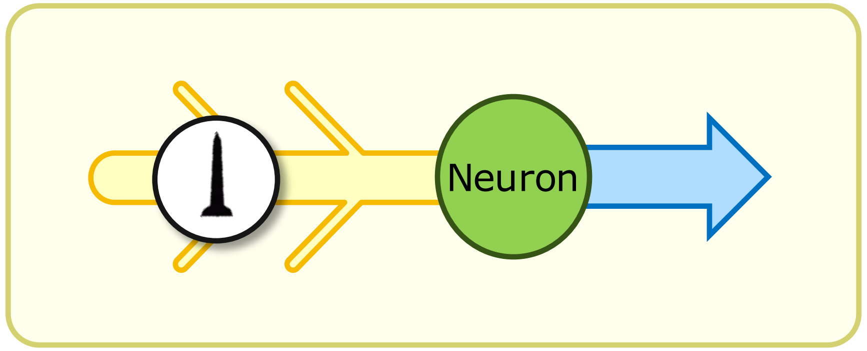


~50% of overheads coming from

Neuron Computation

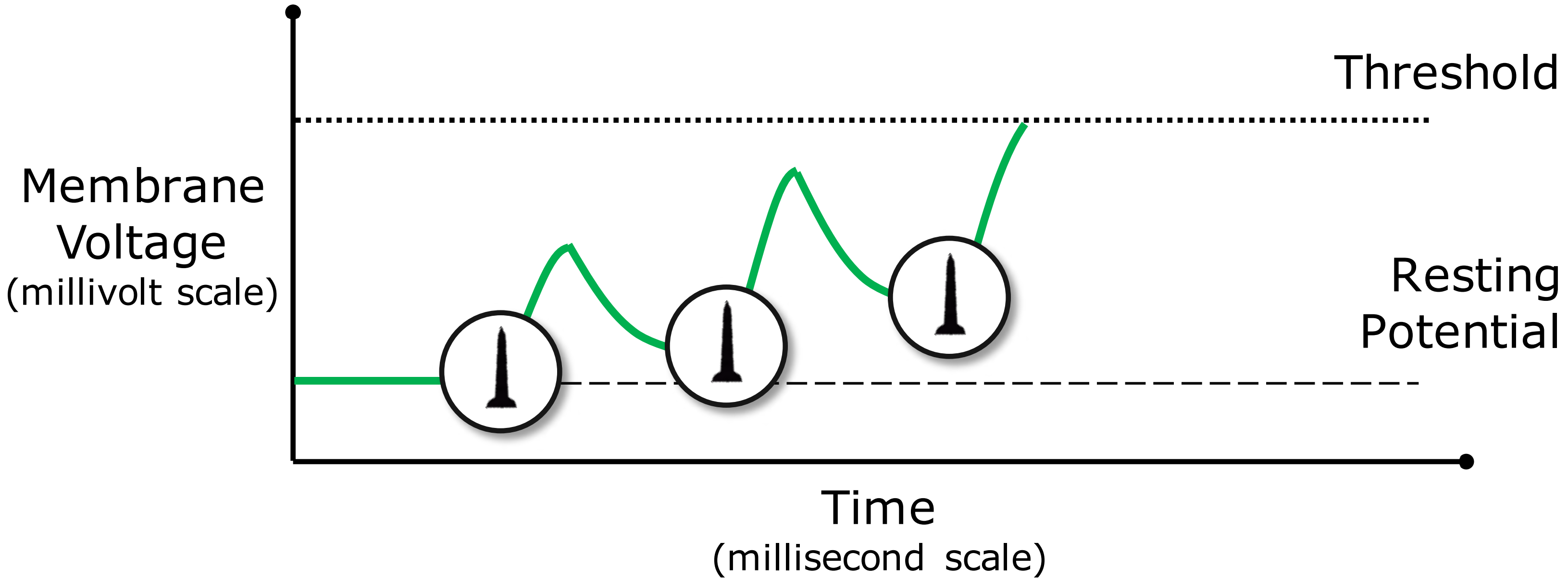
How Does a Neuron Behave?

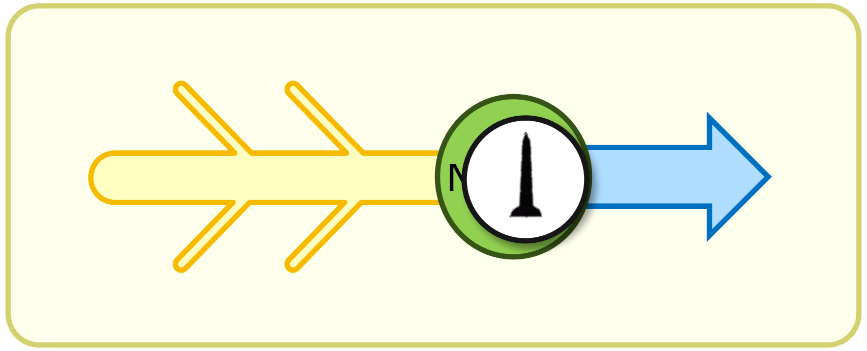




Biological Neuron

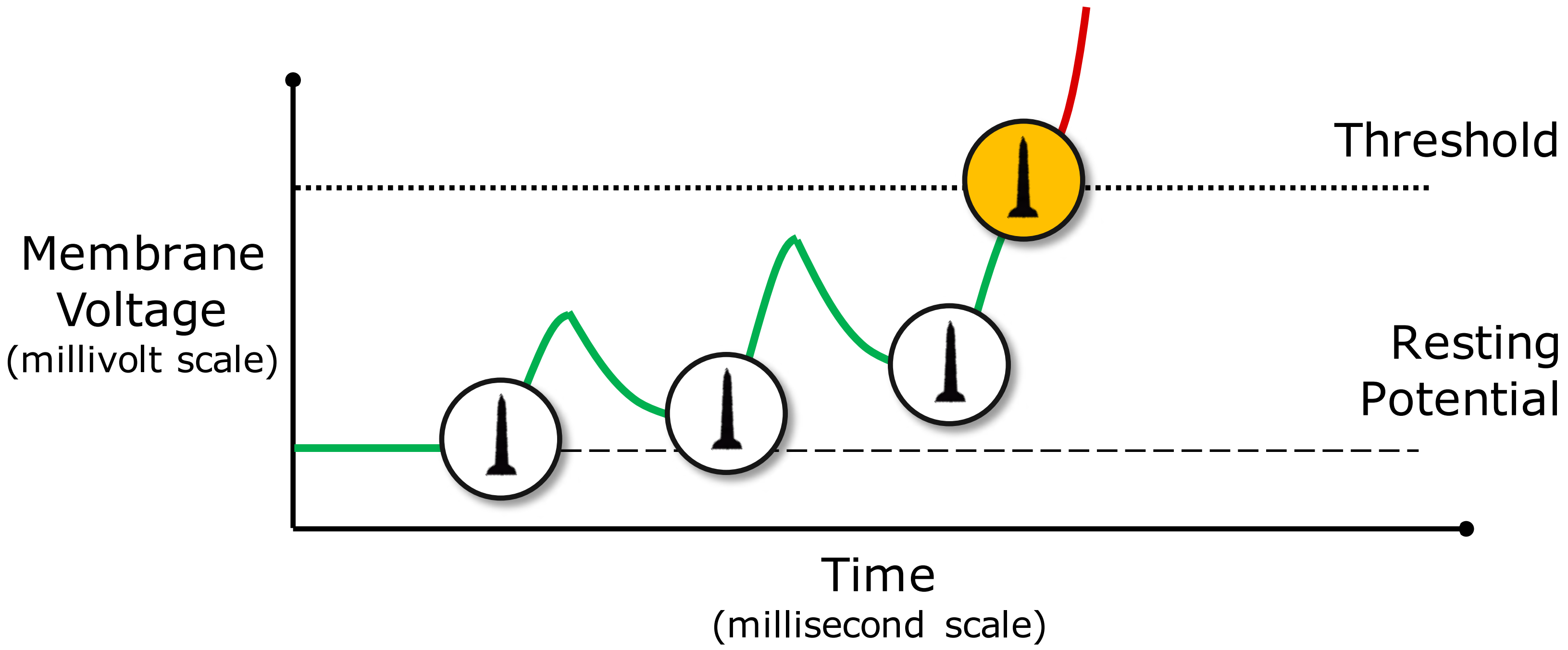
How Does a Neuron Behave?

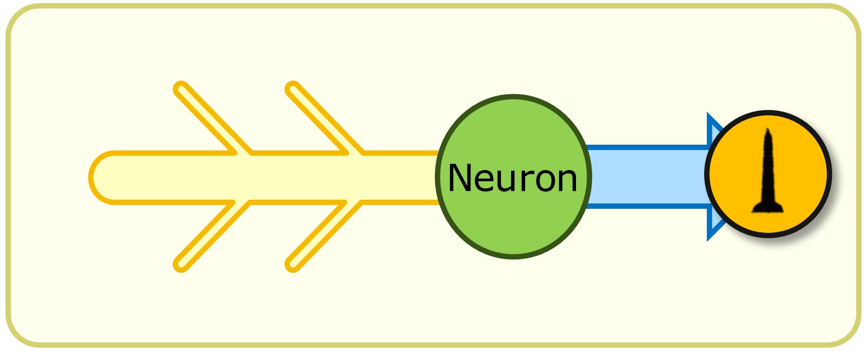




Biological Neuron

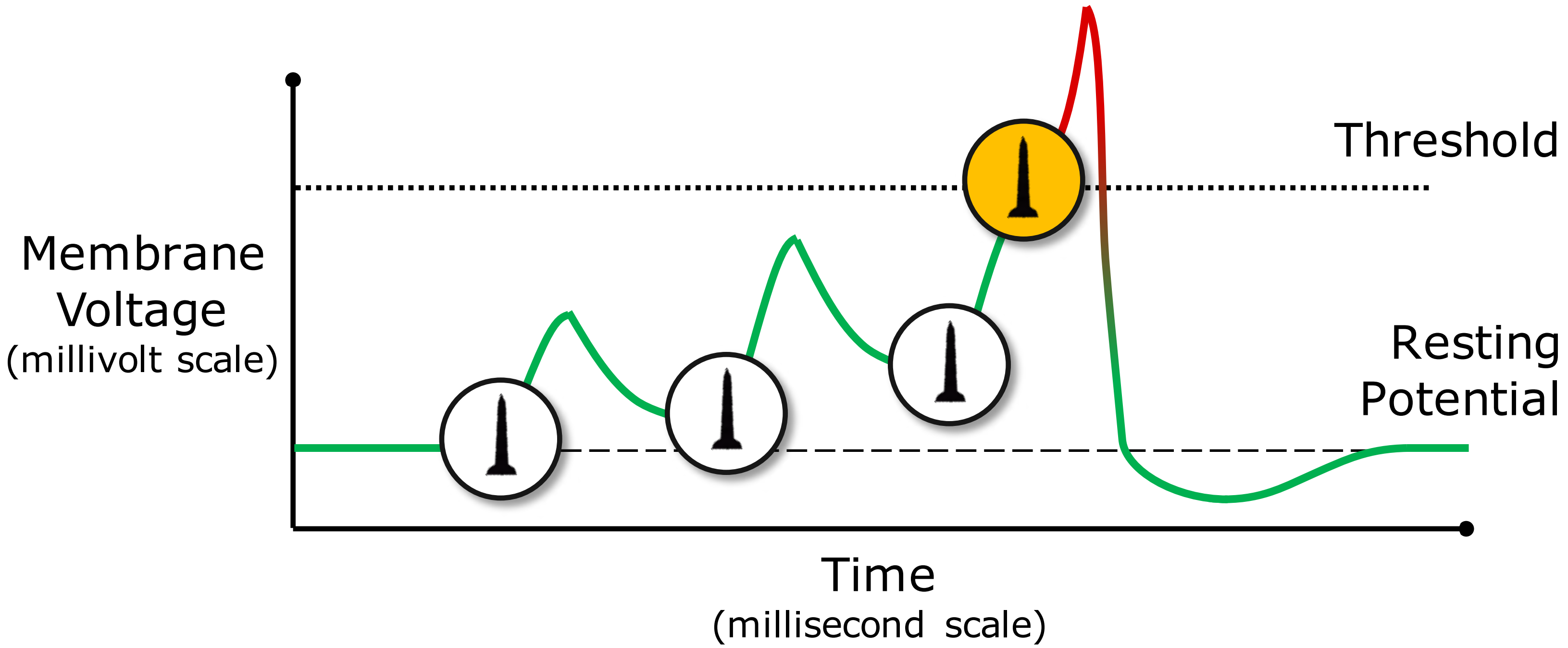
How Does a Neuron Behave?

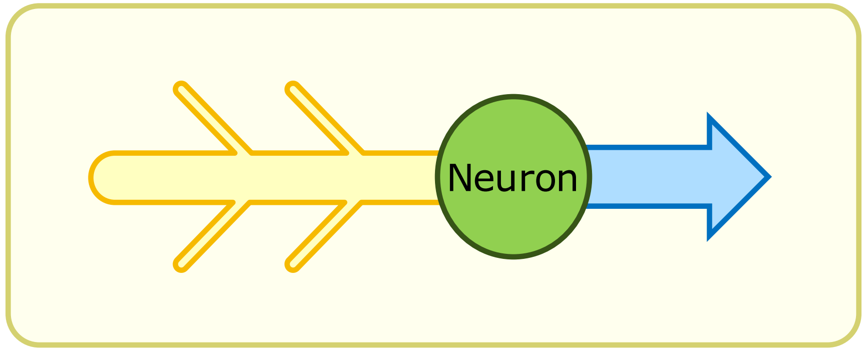




Biological Neuron

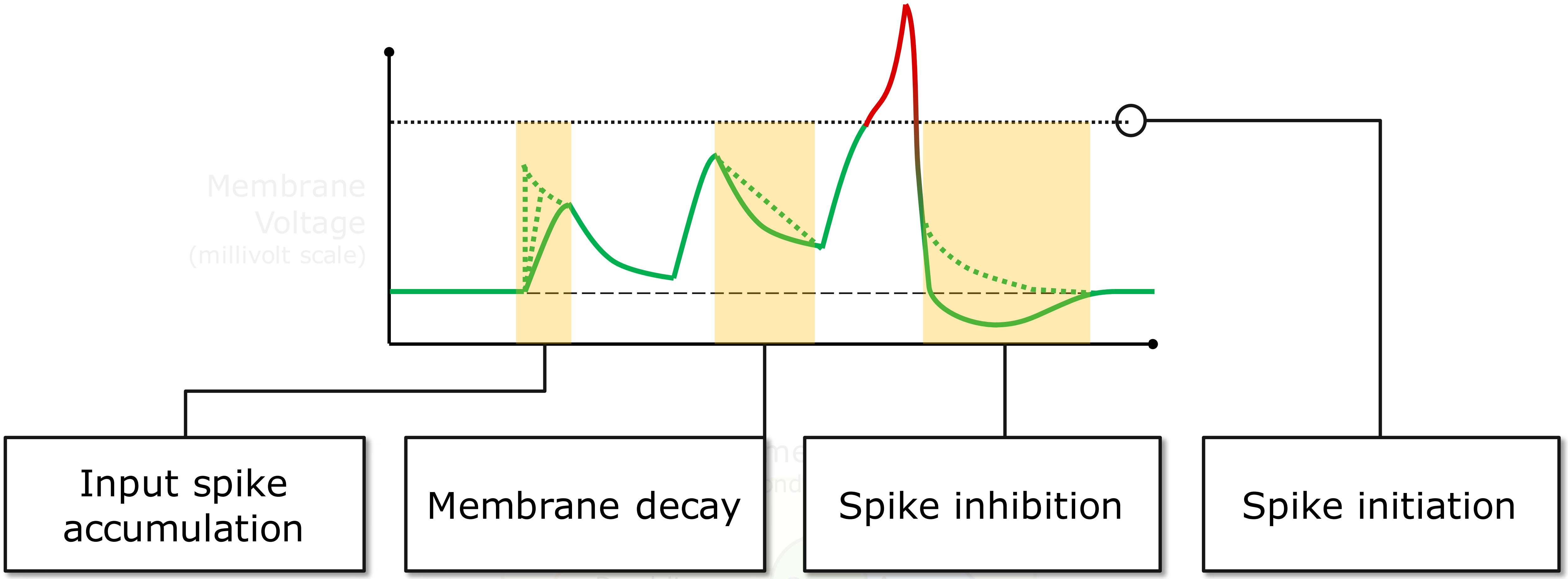
How Does a Neuron Behave?





Biological Neuron

Various Neuron Behaviors



Input spike accumulation

Membrane decay

Spike inhibition

Spike initiation

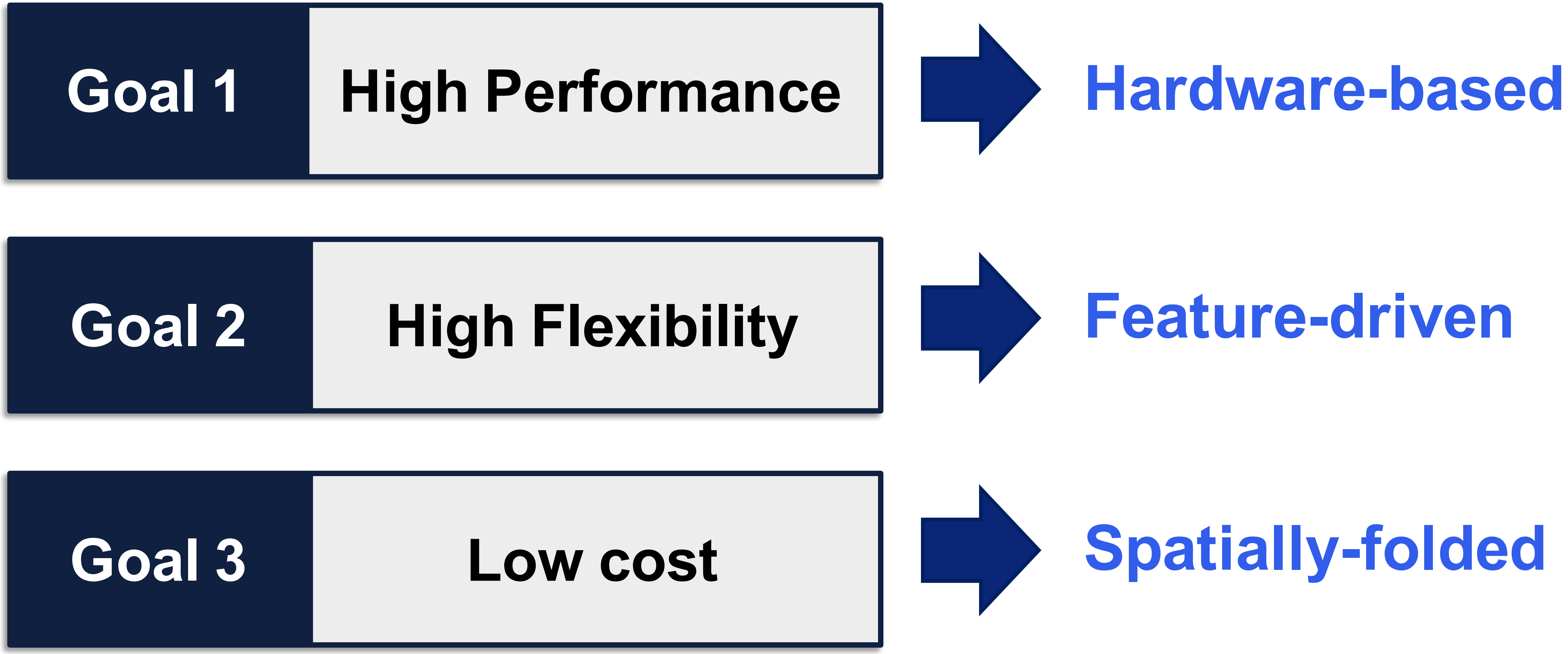
Tons of variants exist, depending on their **feature set.**

We need to support **various features** for **accurate** brain simulations.

Solutions and Limitations

Software Simulation		Custom Hardware Framework
☑	Flexibility	☒
☑	Accuracy	☒
☒	High Performance	☑
☒	Low Energy	☑

Design Goals & Key Ideas

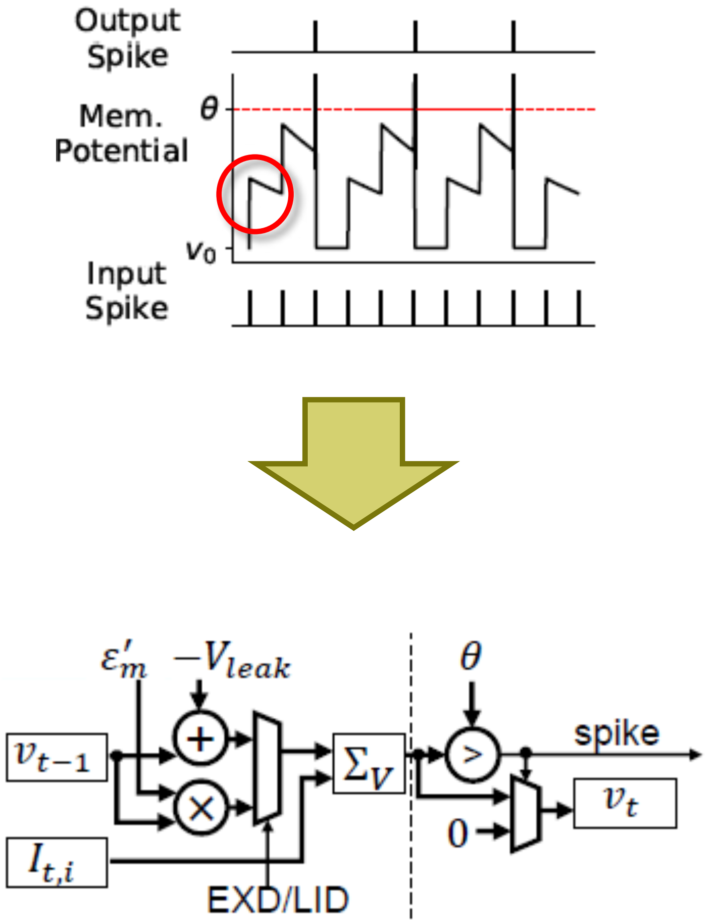


Goal 1 | High Performance → Hardware-based

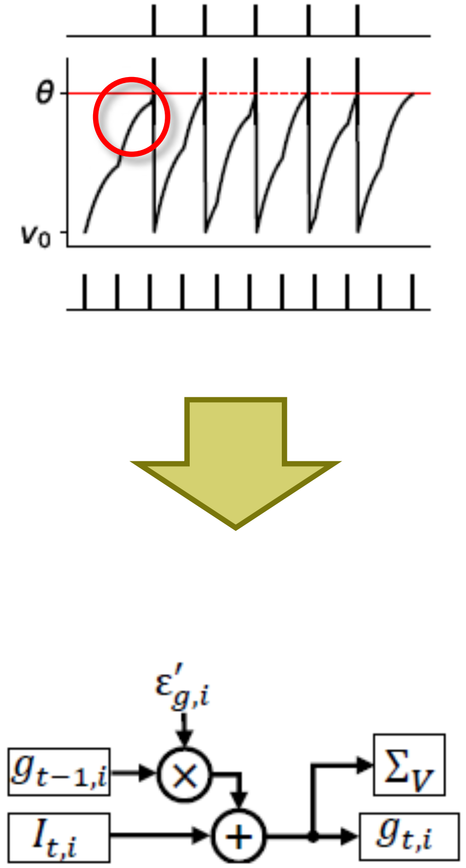
Goal 2 | High Flexibility → Feature-driven

Goal 3 | Low cost → Spatially-folded

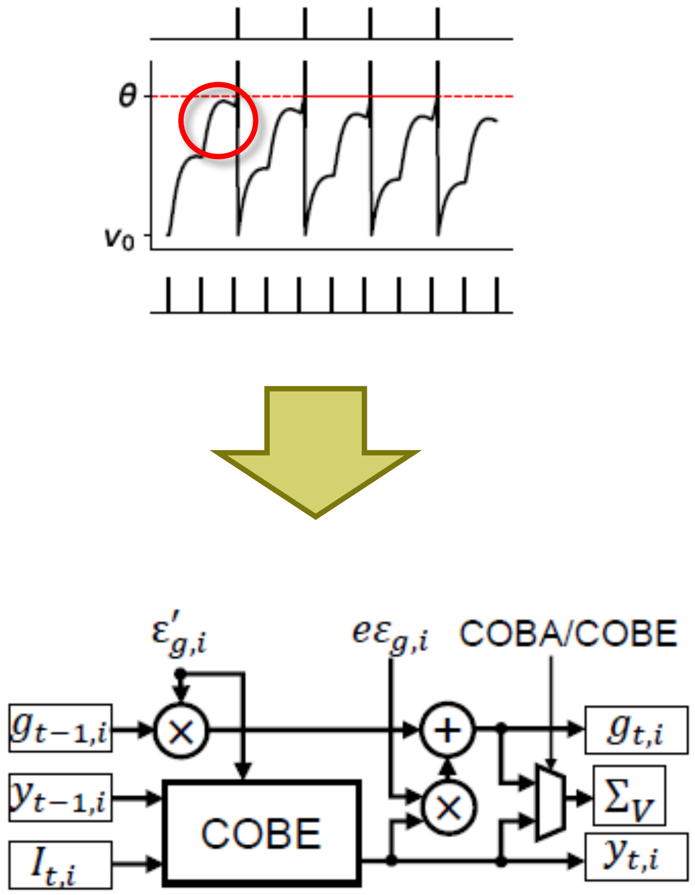
Neuron Feature #1: **Input Spike Accumulation**



Current-based

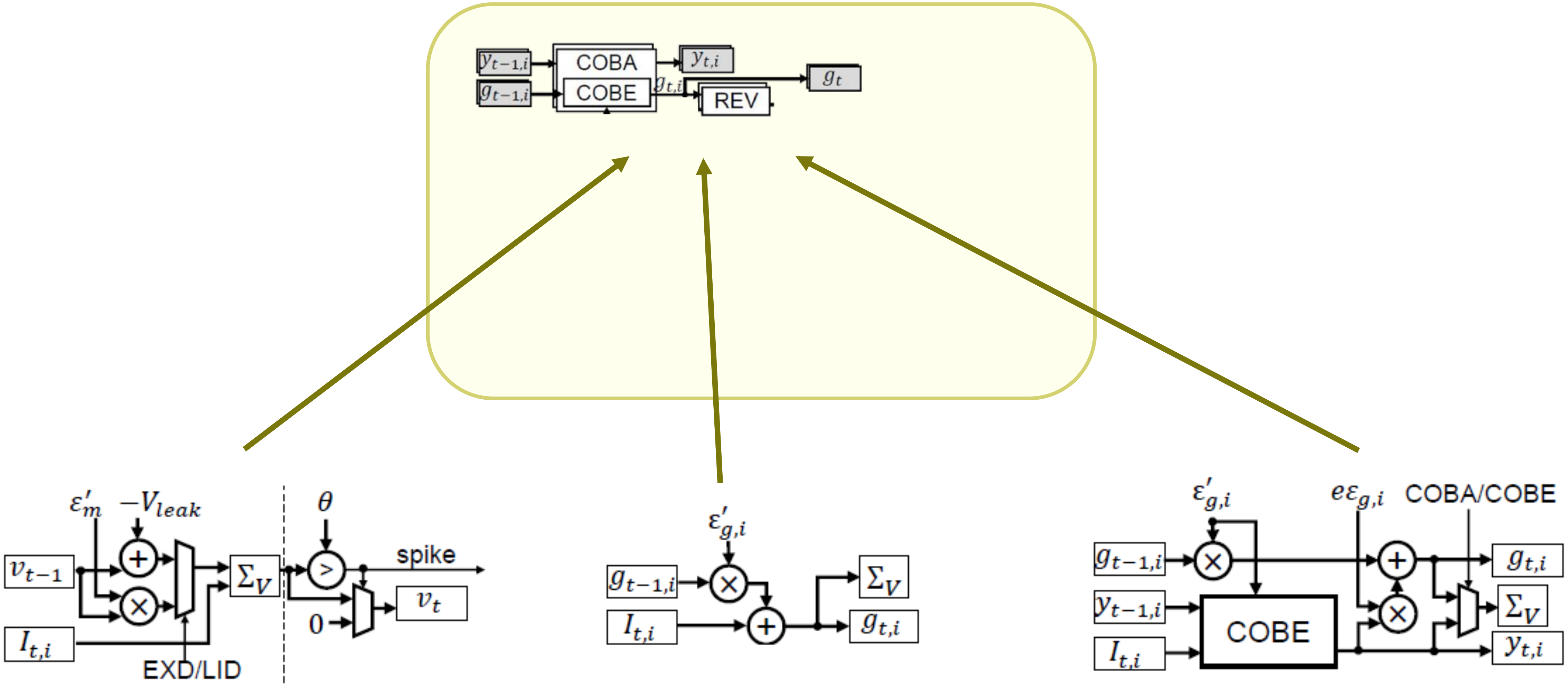


Conductance-based
(Exponential-shaped)



Conductance-based
(Alpha function-shaped)

Neuron Feature #1: **Input Spike Accumulation**

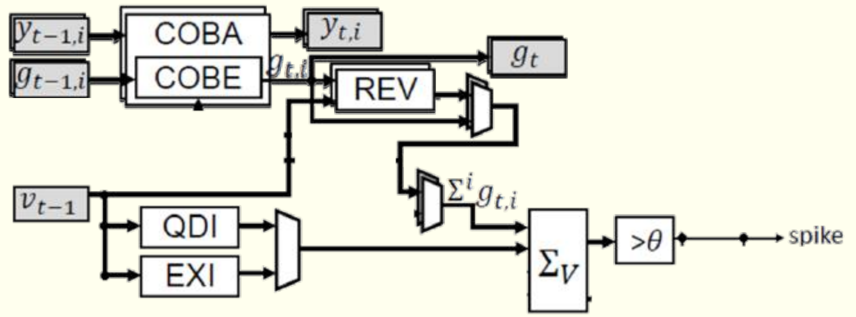


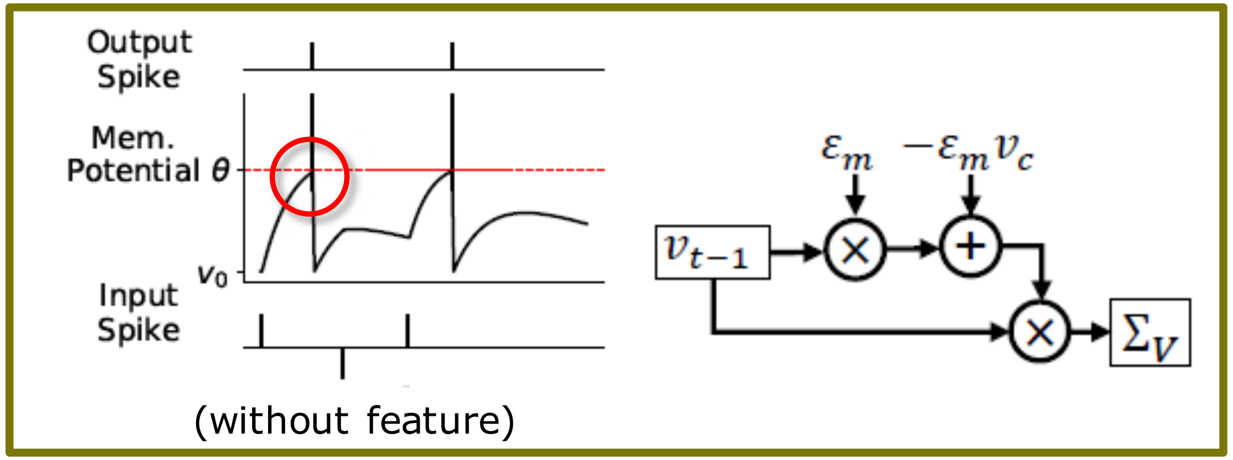
Current-based

Conductance-based
(Exponential-shaped)

Conductance-based
(Alpha function-shaped)

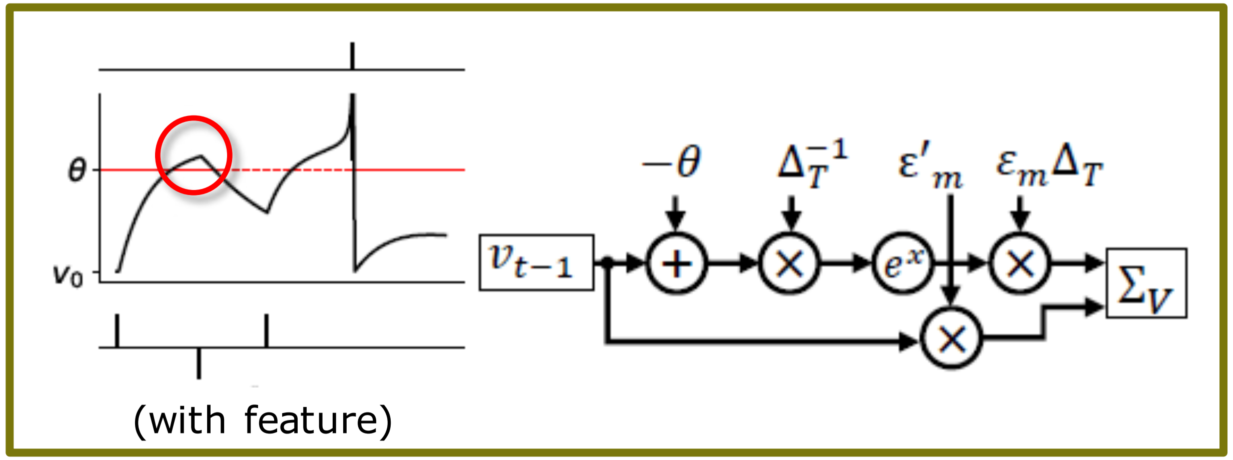
Neuron Feature #2: **Spike Initiation**





(without feature)

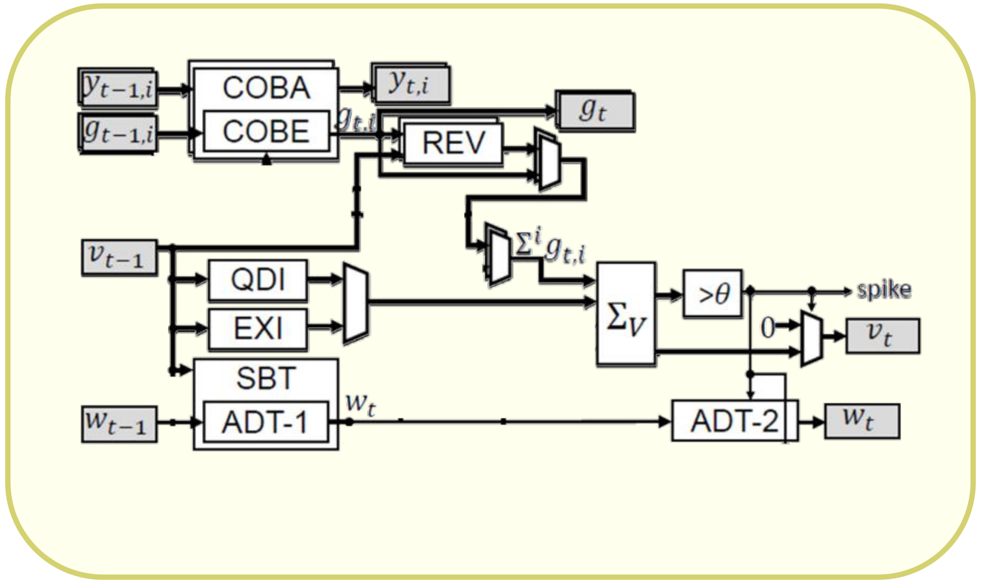
Quadratic

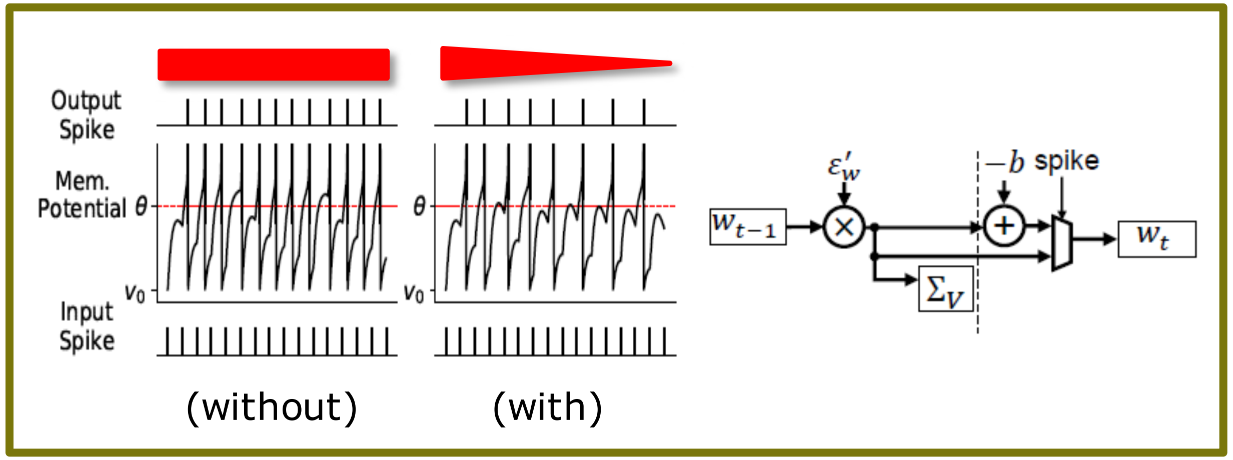


(with feature)

Exponential

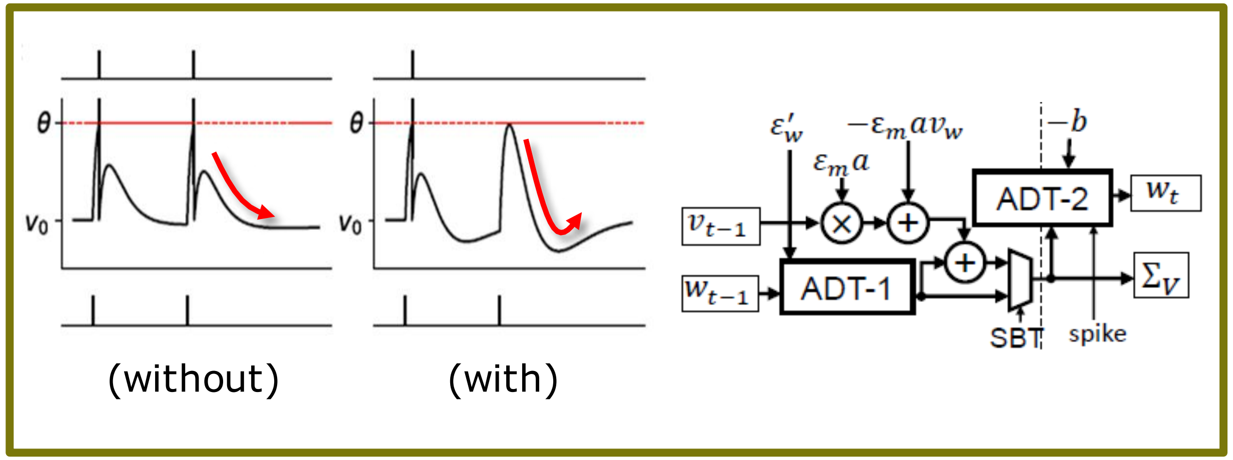
Neuron Feature #3: **Spike-triggered Current**





(without) (with)

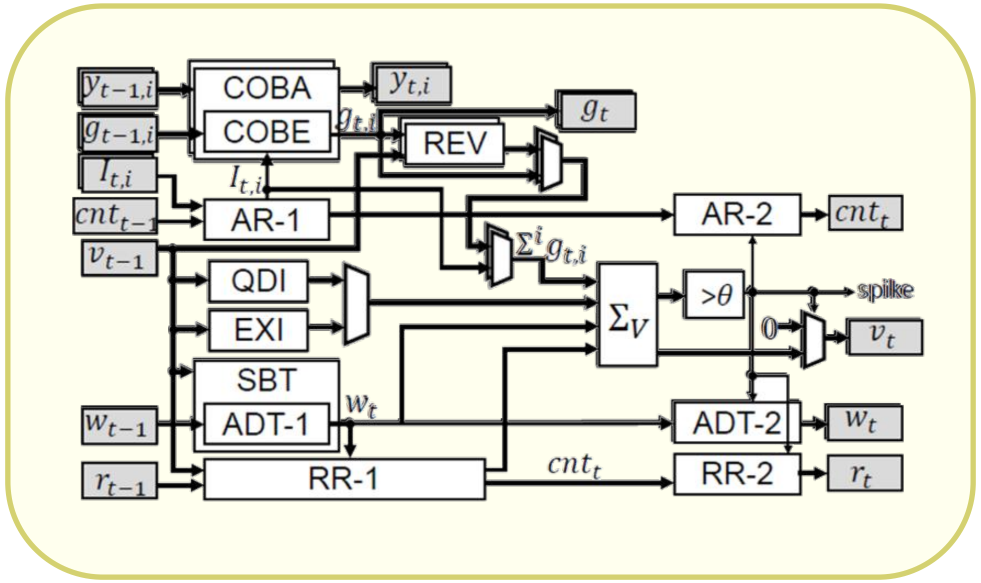
Adaptation

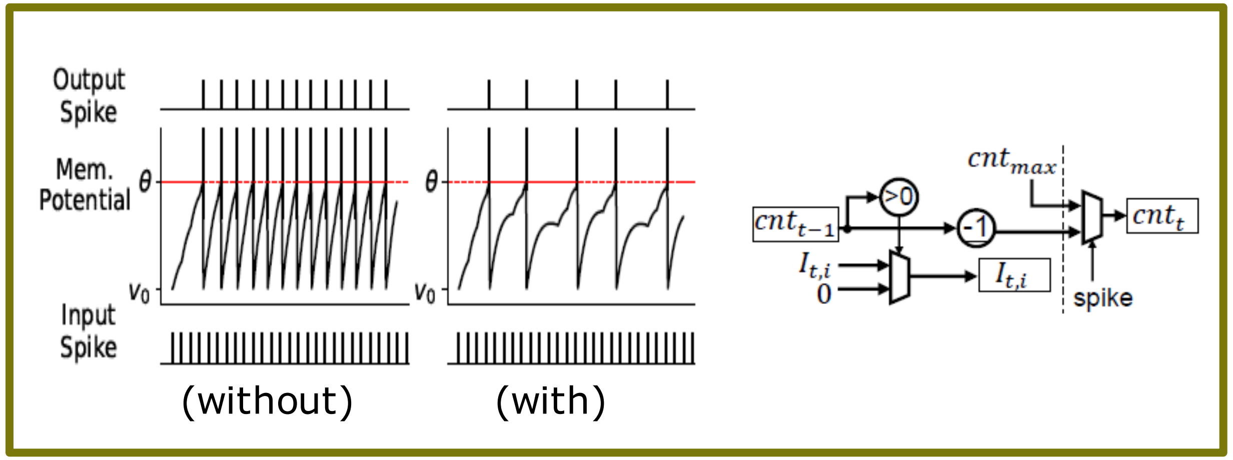


(without) (with)

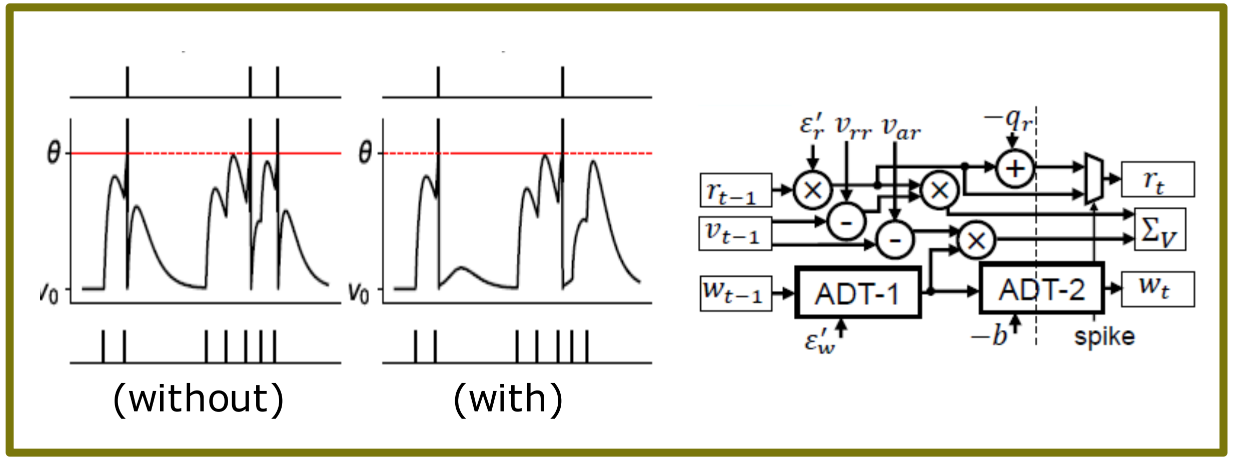
Sub-threshold Oscillation

Neuron Feature #4: **Refractory Period**



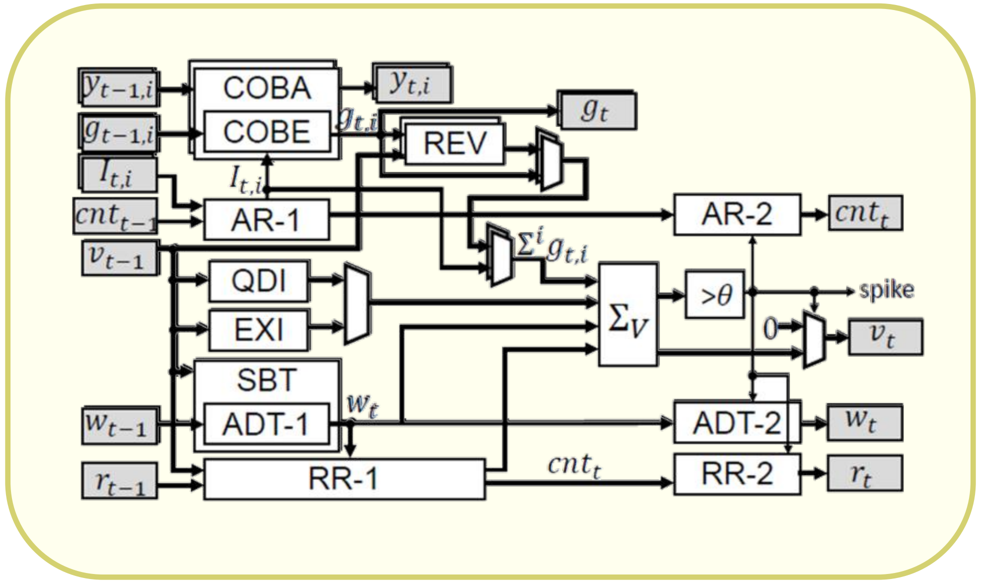


Absolute

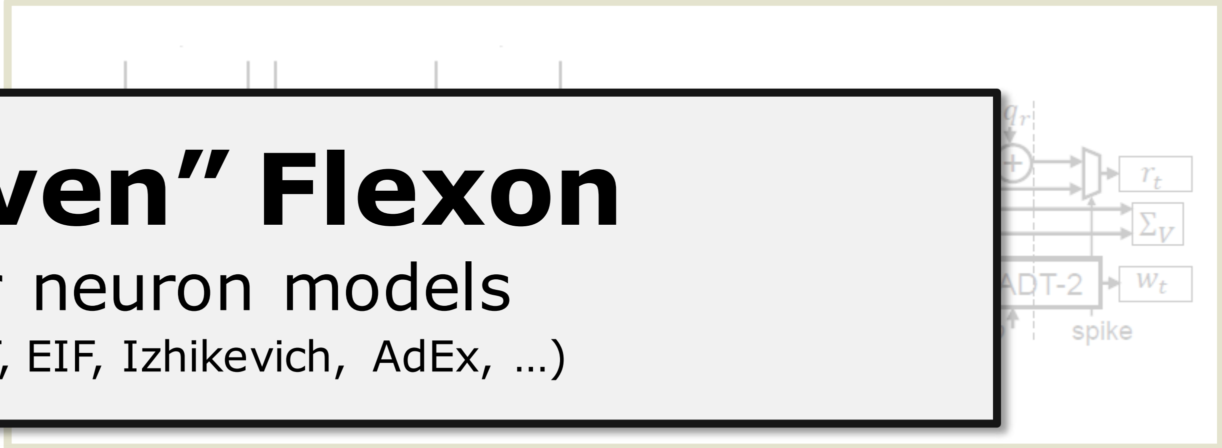


Relative

Neuron Feature: **Flexible Feature Support**



Absolute



Relative

“Feature-driven” Flexon

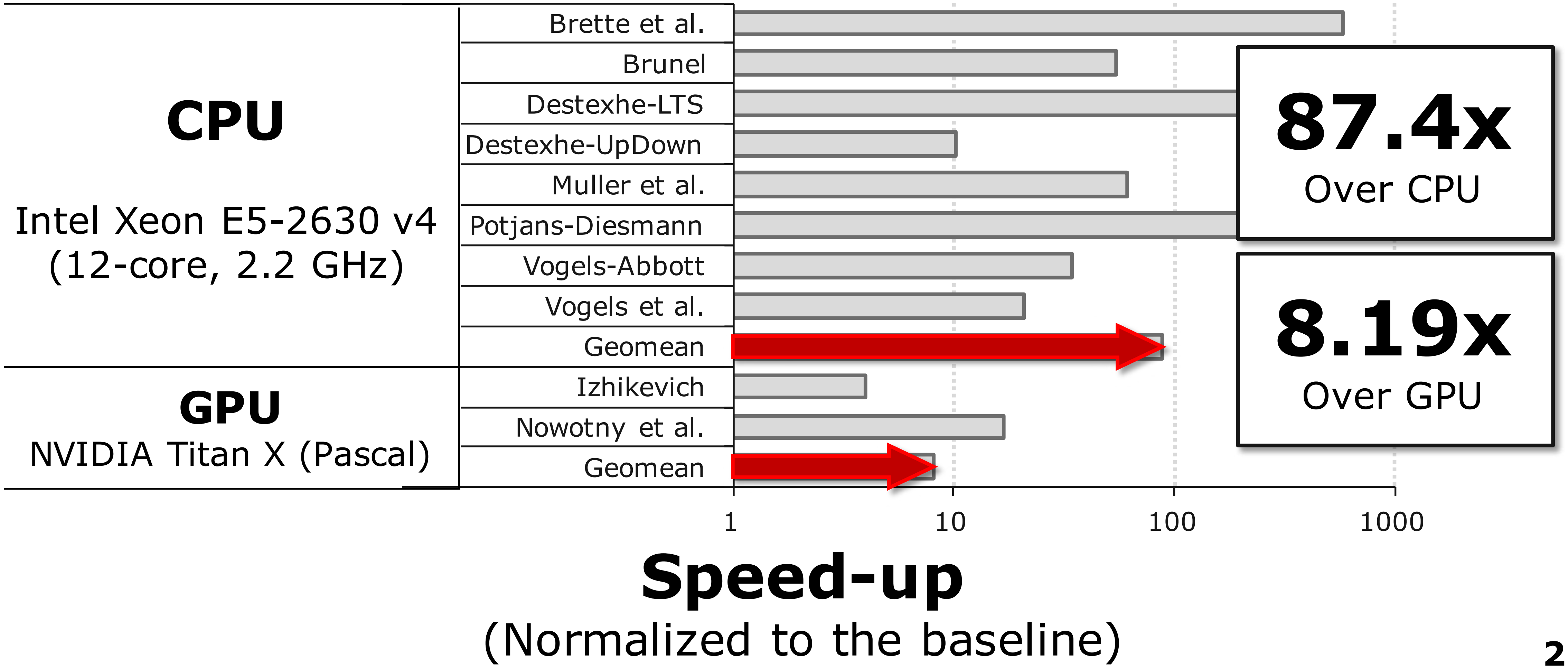
supports 11 major neuron models

(LLIF, SLIF, DSRM0, DLIF, QIF, EIF, Izhikevich, AdEx, …)

Evaluation *(12x Feature-driven Design)*

8 CPU + 2 GPU Representative Benchmarks

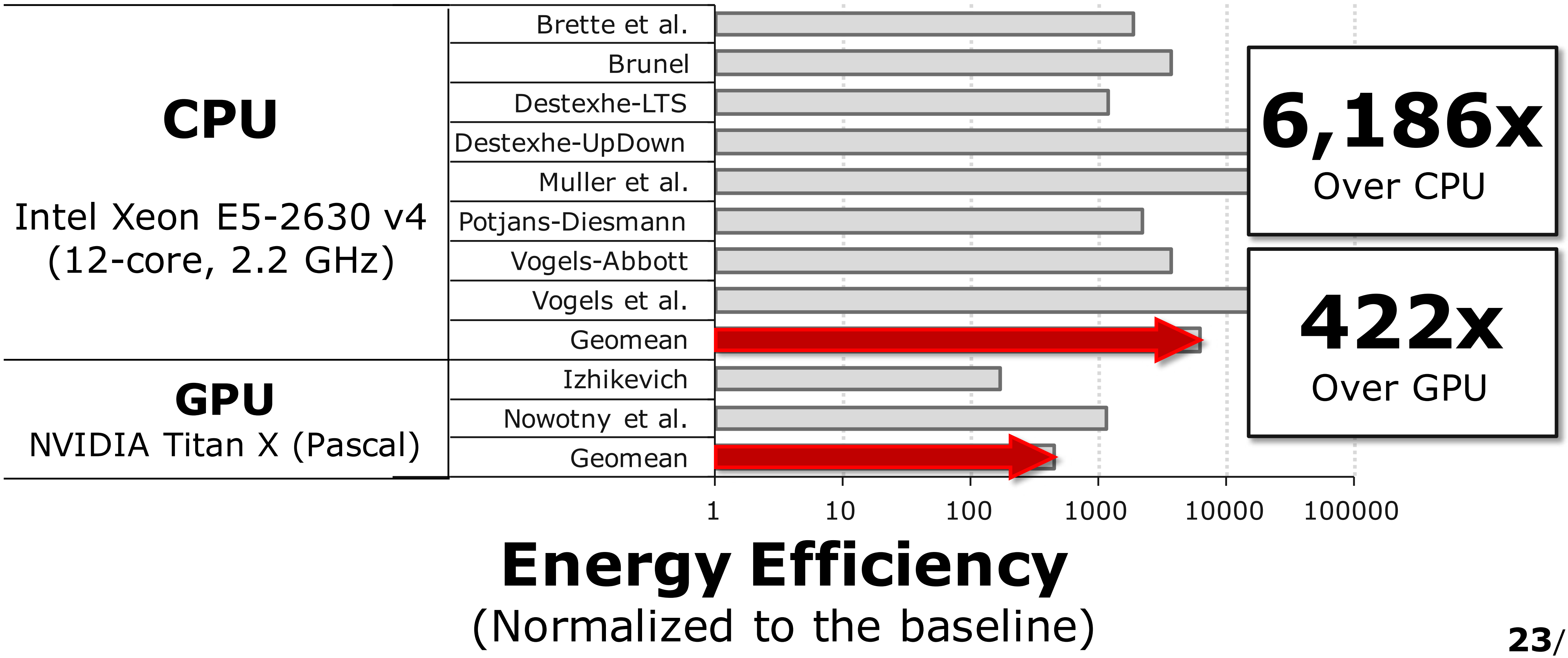
Flexon: TSMC 45nm, Synopsys Design Compiler (neuron), CACTI 6.5 (SRAM)



Evaluation *(12x Feature-driven Design)*

8 CPU + 2 GPU Representative Benchmarks

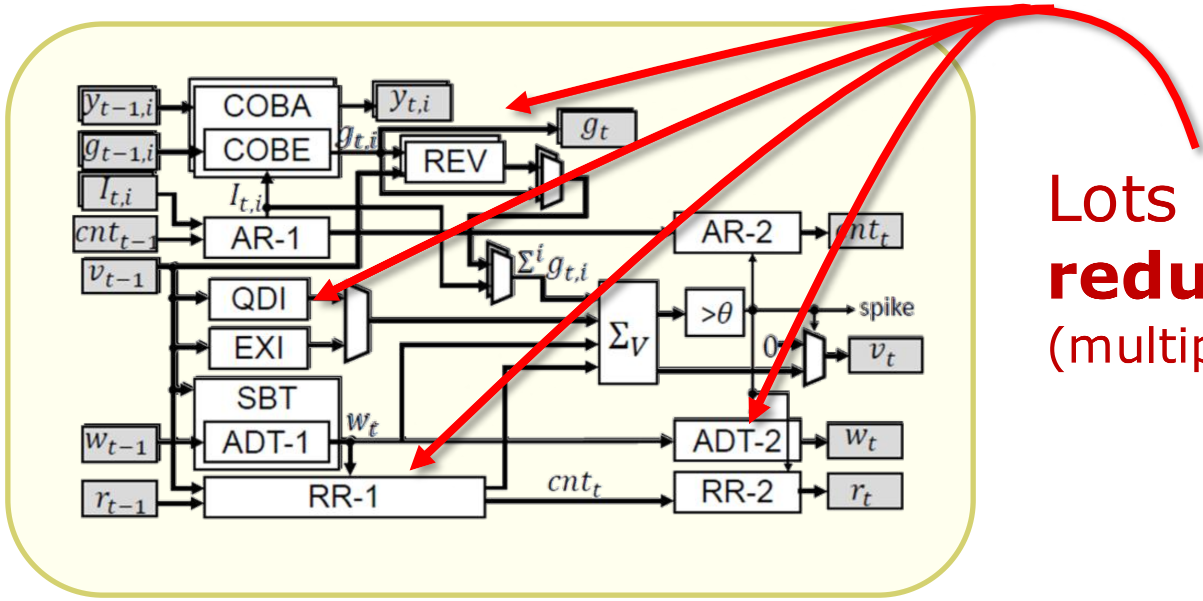
Flexon: TSMC 45nm, Synopsys Design Compiler (neuron), CACTI 6.5 (SRAM)



6,186x Over CPU

422x Over GPU

Intrinsic Space-inefficiency



Lots of **redundant** units (multiplier, adder, …)

“Feature-driven” Flexon

supports 11 major neuron models

(LLIF, SLIF, DSRM0, DLIF, QIF, EIF, Izhikevich, AdEx, …)

Outp Spik

Mem Potent

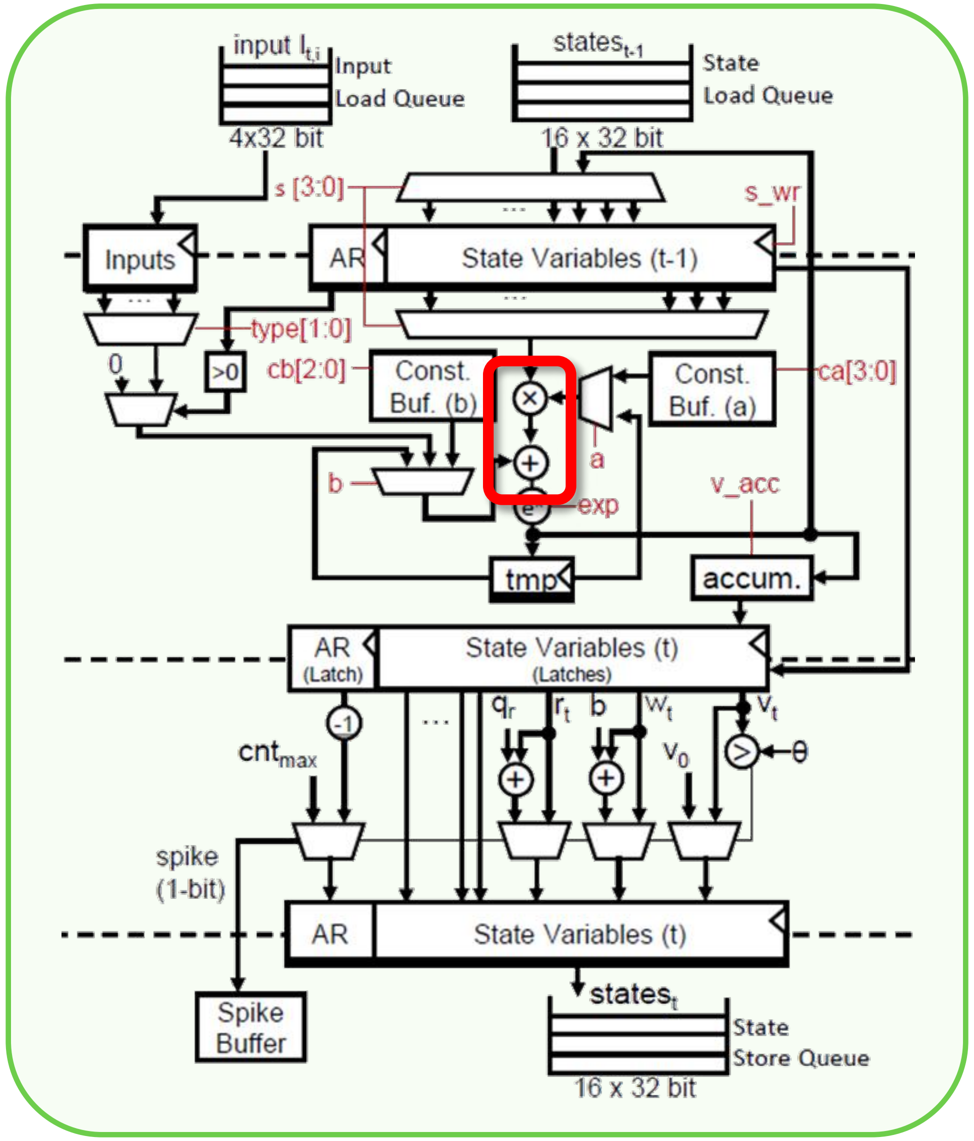
Inpu Spik

(without) (with)

Absolute

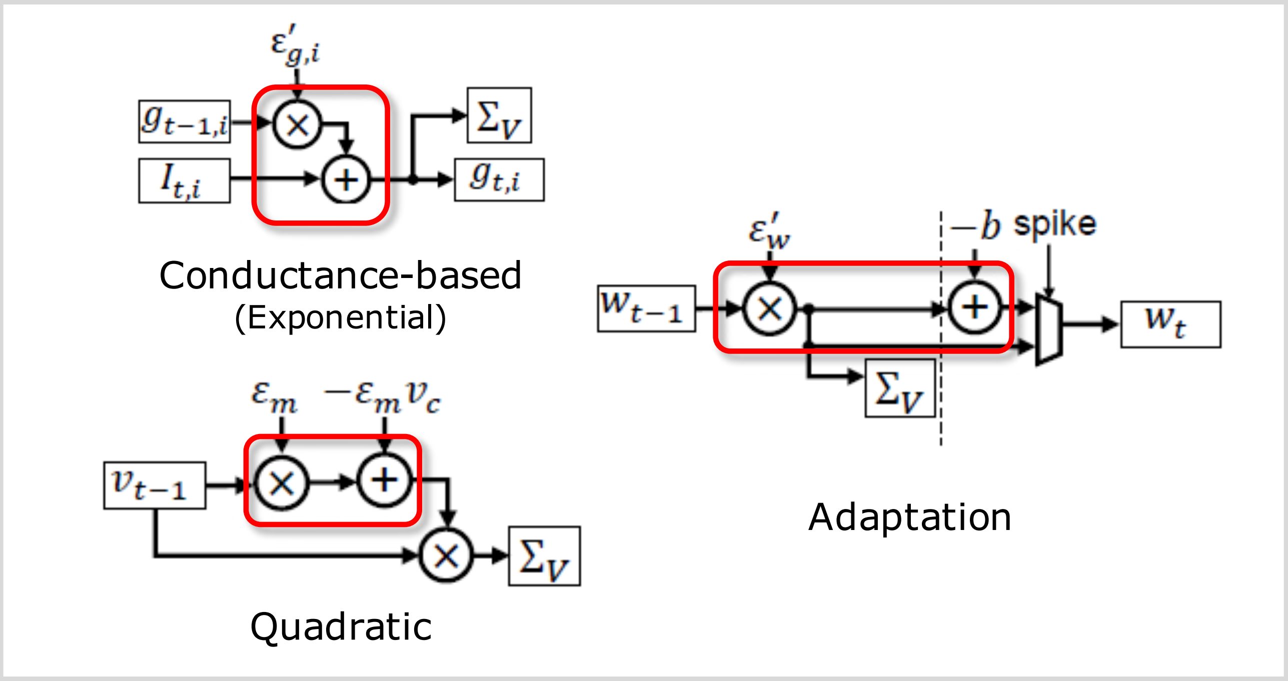
Relative

Constructing *Spatially-folded* Flexon

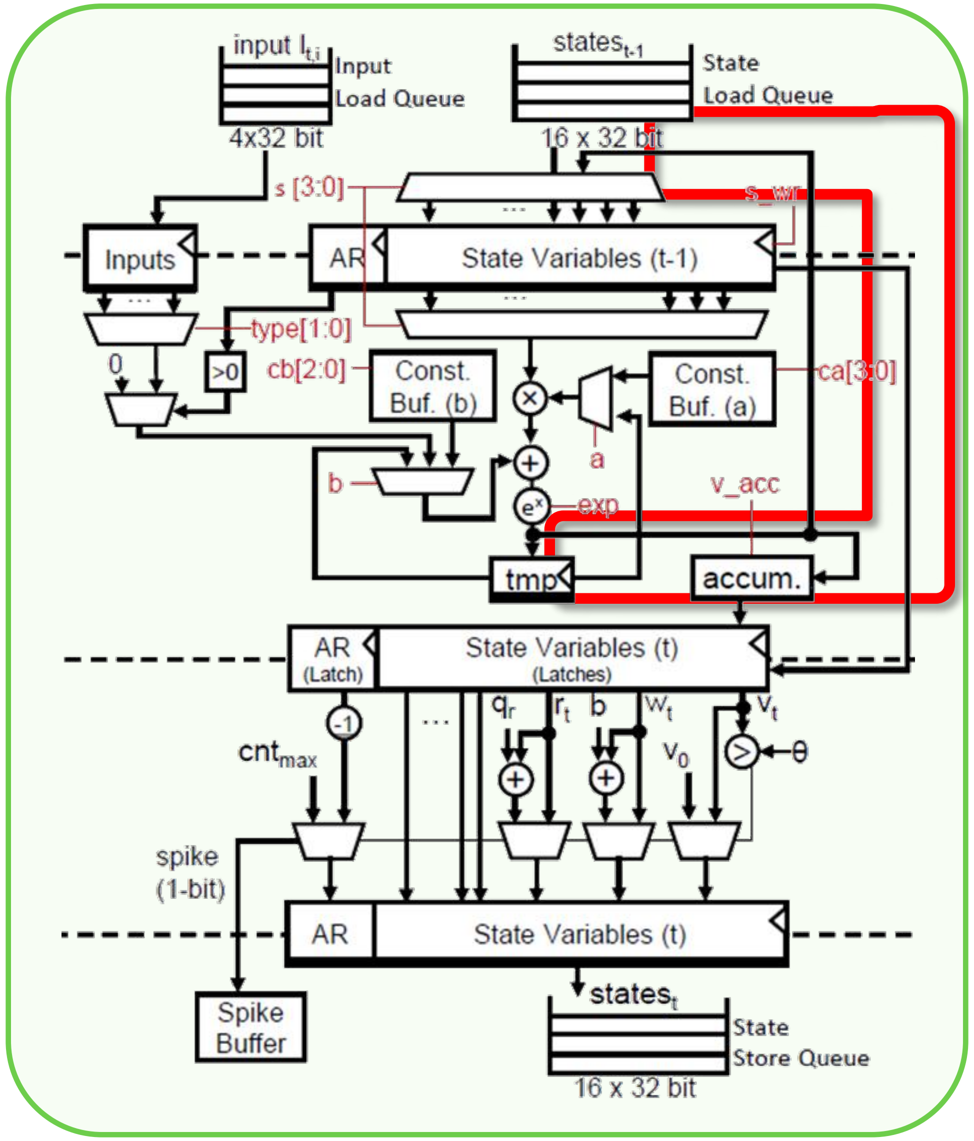


Spatially-folded design ➔ reduce area

- Remove redundant MAC operators



Constructing *Spatially-folded* Flexon

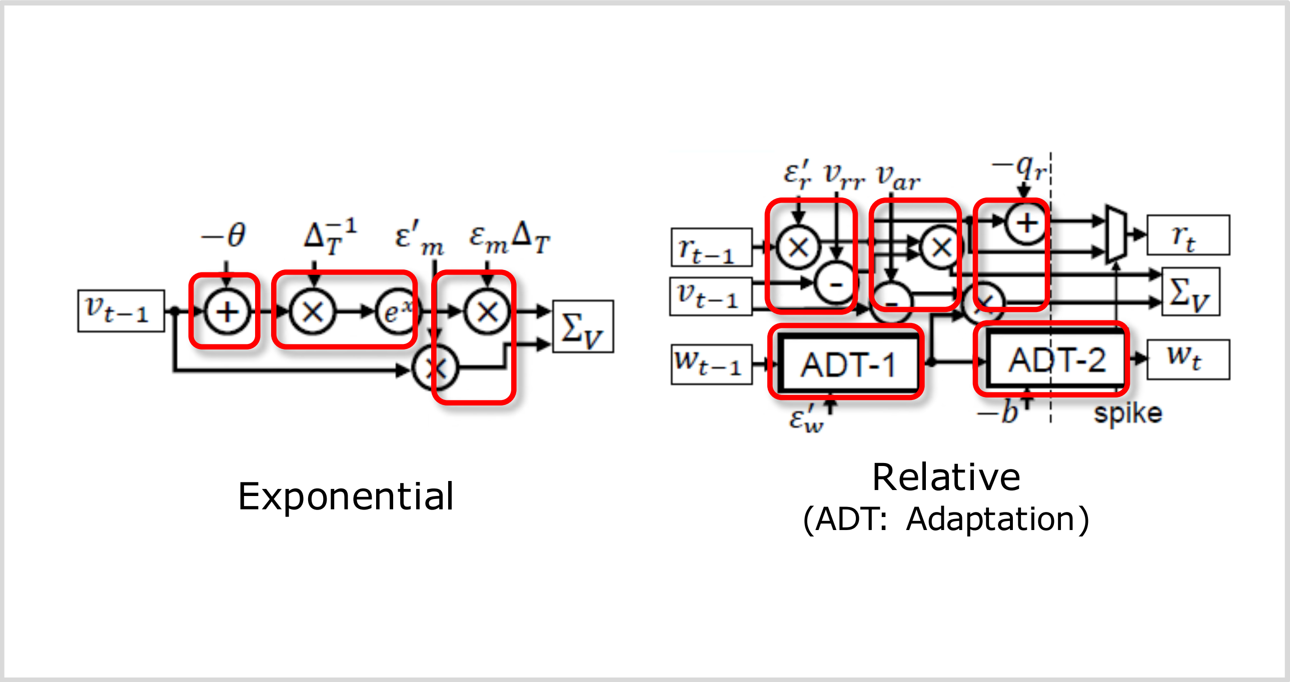


Spatially-folded design ➔ reduce area

- Remove redundant MAC operators

Modifications from the baseline

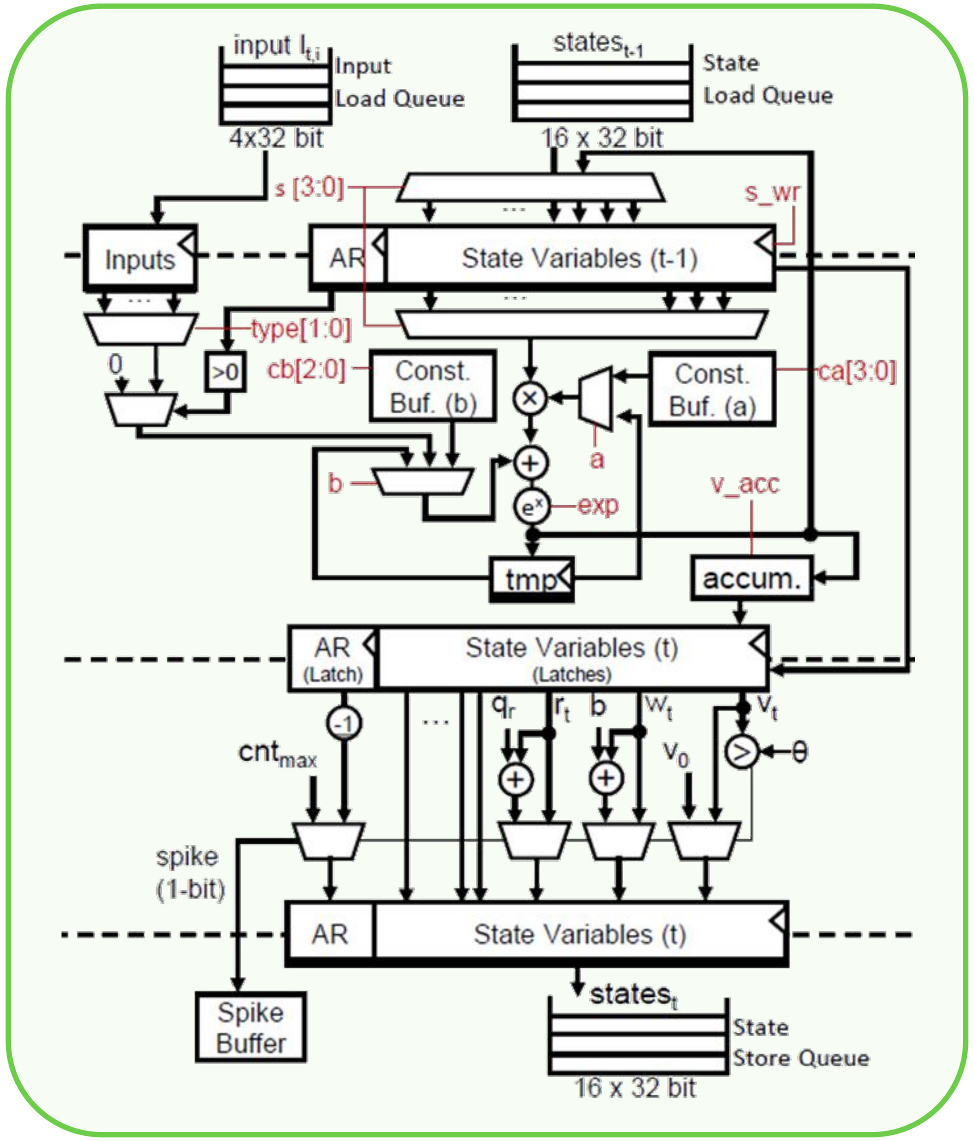
- 2-stage pipeline, multi-cycle implementation



Exponential

Relative
(ADT: Adaptation)

Constructing *Spatially-folded* Flexon



Spatially-folded design ➔ reduce area

- Remove redundant MAC operators

What we should change

- 2-stage pipeline, multi-cycle implementation

"Spatially-folded" Flexon

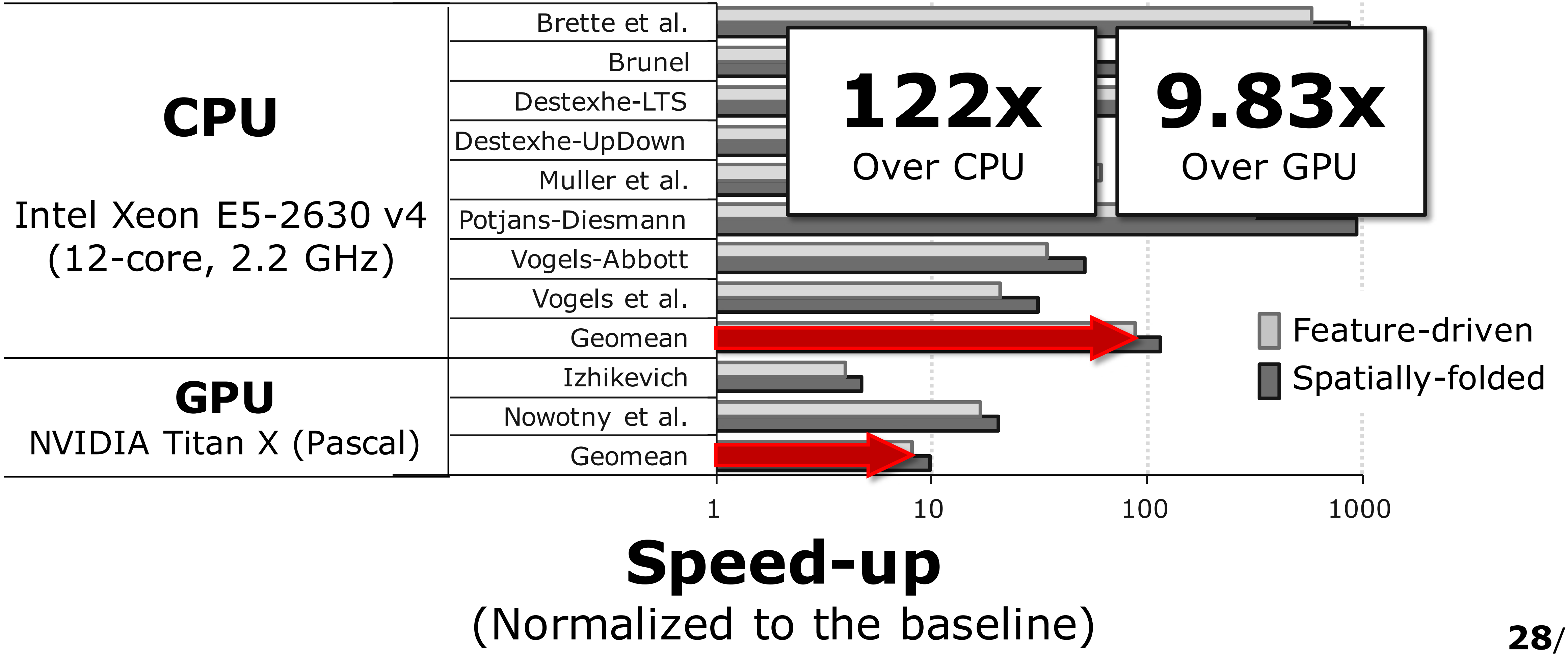
supports various major neuron models

6x area saving

Evaluation (*72x Spatially-folded* Design)

8 CPU + 2 GPU Representative Benchmarks

Flexon: TSMC 45nm, Synopsys Design Compiler (neuron), CACTI 6.5 (SRAM)



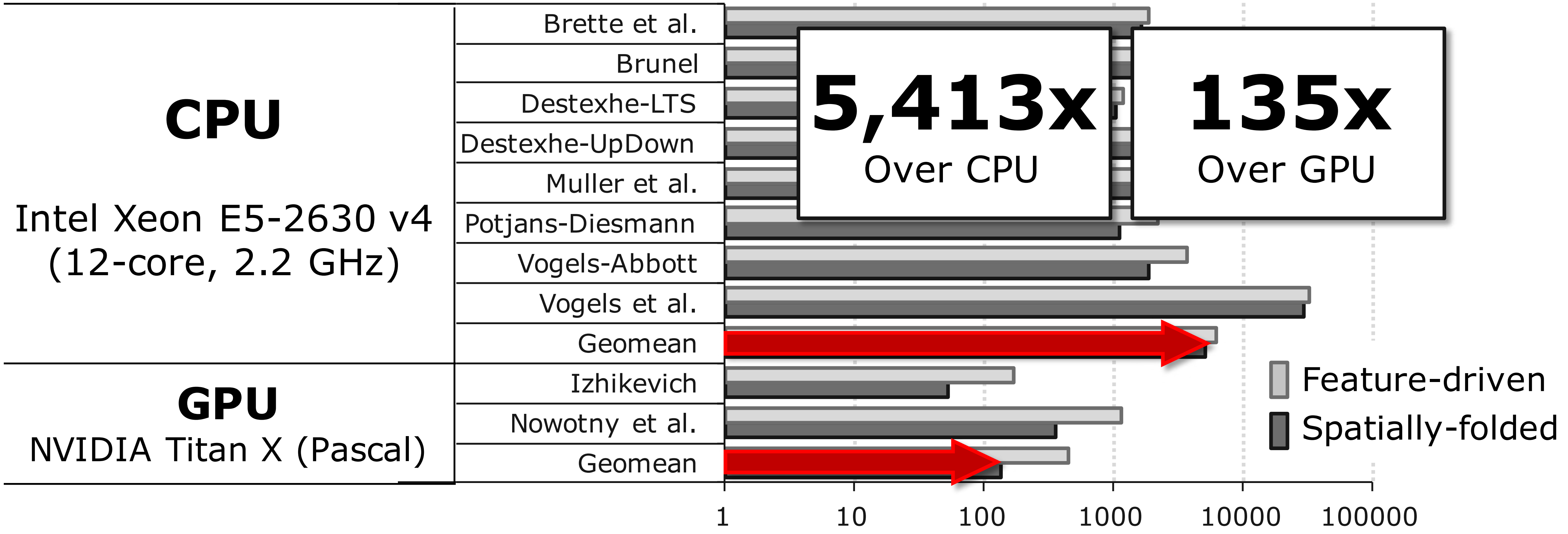
122x Over CPU

9.83x Over GPU

Evaluation (*72x Spatially-folded* Design)

8 CPU + 2 GPU Representative Benchmarks

Flexon: TSMC 45nm, Synopsys Design Compiler (neuron), CACTI 6.5 (SRAM)



Energy Efficiency
(Normalized to the baseline)

Baseline Flexon vs. Spatially-folded Flexon

- Baseline "Feature-driven" Flexon
  - Fast: 87.4x over CPUs, 8.19x over GPUs
  - Energy-efficient: **6,186x** over CPUs, **422x** over GPUs

- "Spatially-folded" Flexon
  - Fast: **122x** over CPUs, **9.83x** over GPUs
  - Energy-efficient: 5,413x over CPUs, 135x over GPUs

Conclusion

- **Flexon** is a flexible feature-driven digital neuron design, capable of realizing various major neuron models.
  - Flexible & **power-efficient** (6,186x over CPU)

- **Spatially-folded Flexon** makes features share units, reducing 6x circuit area.
  - Flexible & **fast** when integrated (122x over CPU)

Flexon

A Flexible Digital Neuron for Efficient Spiking Neural Network Simulations

Thank you for listening