NOEMA:

A Massive-Scale Brain Activity Decoding Chip

Ameer Abdelhadi **Eugene Sha** **Andreas Moshovos**

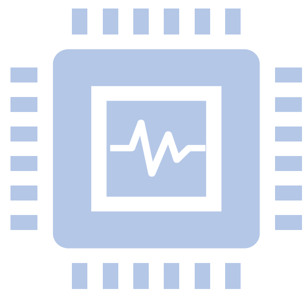
University of Toronto

August, 2022

Brain Machine Interfaces (BMIs)

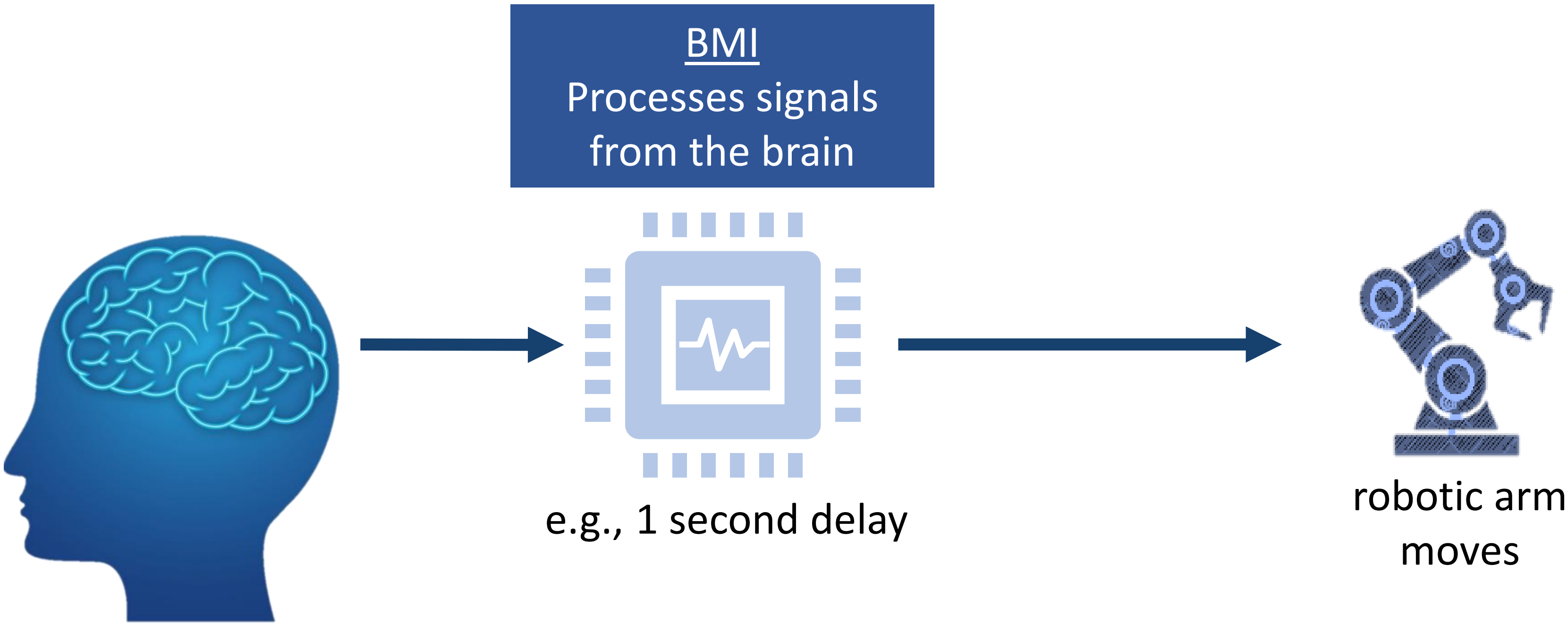
BMI

Processes signals from the brain

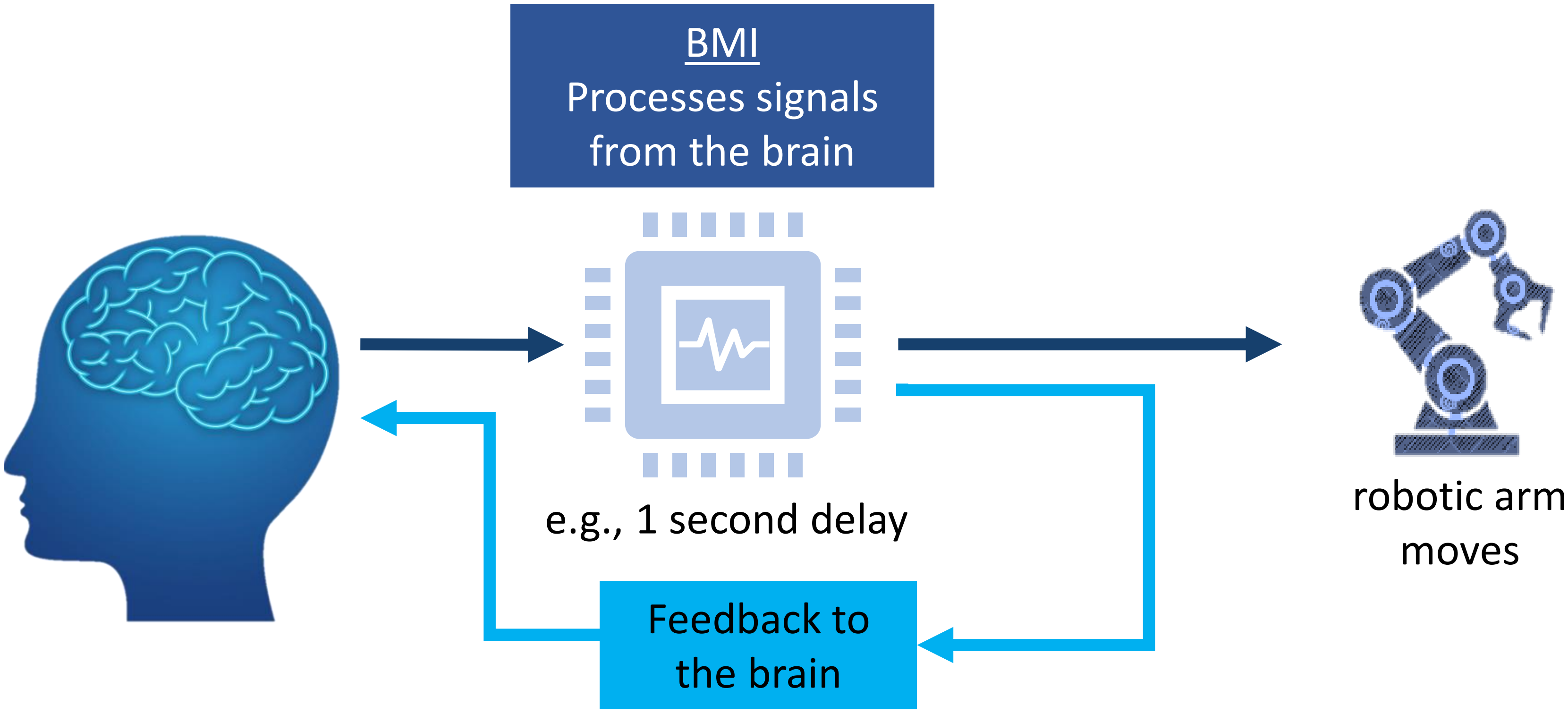


e.g., 1 second delay

Brain Machine Interfaces (BMIs)



Brain Machine Interfaces (BMIs)



BMIs at the edge

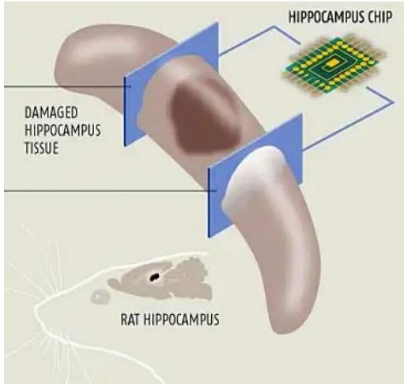
What if we can detect patterns of neuron activity in real-time?

Detect, in real-time, memories, decisions, emotions, and experiences

Applications

Repair brain function

Interface brain regions which no longer connect, e.g. Alzheimer's



Replacement of damaged hippocampus with a chip [1]

[1] https://www.newscientist.com/article/dn3488-worlds-first-brain-prosthesis-revealed/ (Hippocampus repair)

BMIs at the edge

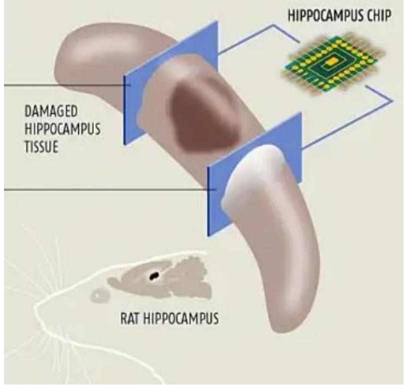
What if we can detect patterns of neuron activity in real-time?

Detect, in real-time, memories, decisions, emotions, and experiences

Applications

Repair brain function

Interface brain regions which no longer connect, e.g. Alzheimer's



Replacement of damaged hippocampus with a chip [1]

Drive effectors

Greater accuracy and dexterity, e.g. robotic limbs

Woman controls robotic arm with 100-channel Utah array [2]

[1] https://www.newscientist.com/article/dn3488-worlds-first-brain-prosthesis-revealed/ (Hippocampus repair)
[2] https://continuum.utah.edu/web-exclusives/the-bionics-man/ (Utah Array)

BMIs at the edge

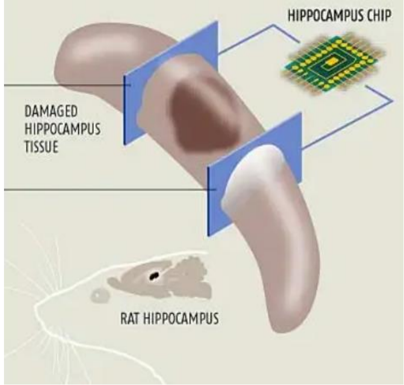
What if we can detect patterns of neuron activity in real-time?

Detect, in real-time, memories, decisions, emotions, and experiences

Applications

Repair brain function

Interface brain regions which no longer connect, e.g. Alzheimer's



Replacement of damaged hippocampus with a chip [1]

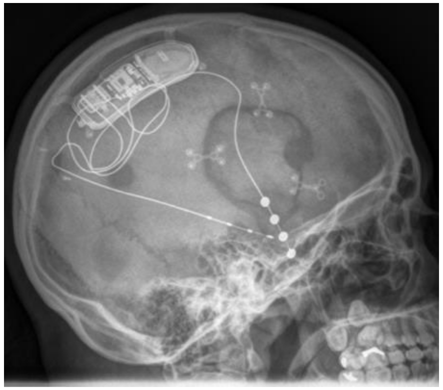
Drive effectors

Greater accuracy and dexterity, e.g. robotic limbs

Woman controls robotic arm with 100-channel Utah array [2]

Anticipate and prevent harmful neural activity

e.g. epilepsy

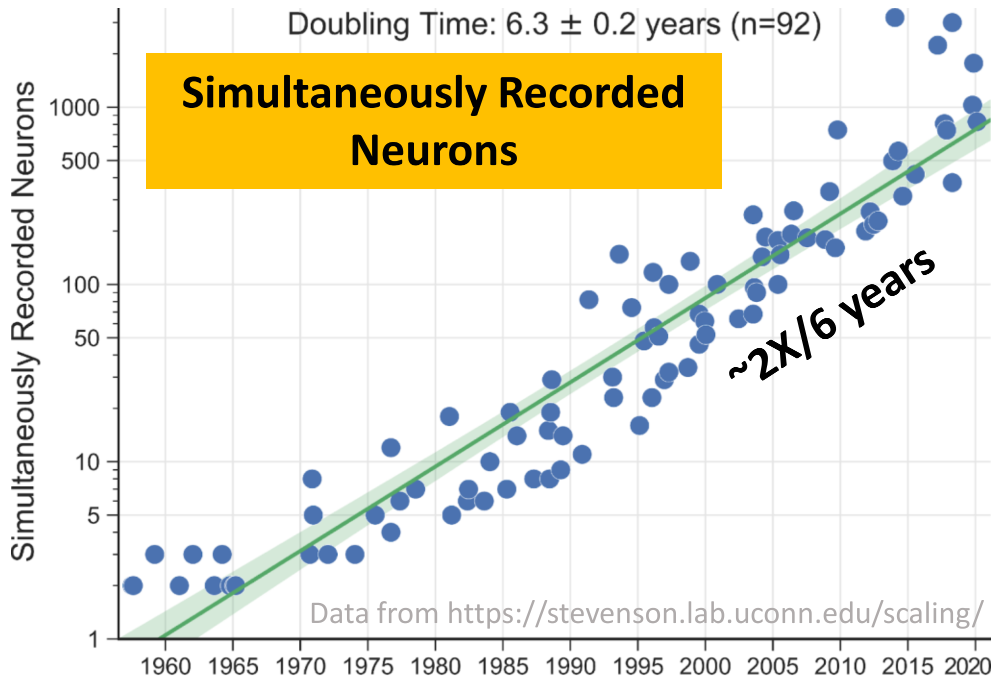


Responsive neurostimulator system for epilepsy [3]

[1] https://www.newscientist.com/article/dn3488-worlds-first-brain-prosthesis-revealed/ (Hippocampus repair)
[2] https://continuum.utah.edu/web-exclusives/the-bionics-man/ (Utah Array)
[3] Critical review of the responsive neurostimulator system for epilepsy (Thomas and Jobst, 2015)

The Challenge and Opportunity

Capture Capability Growing Exponentially



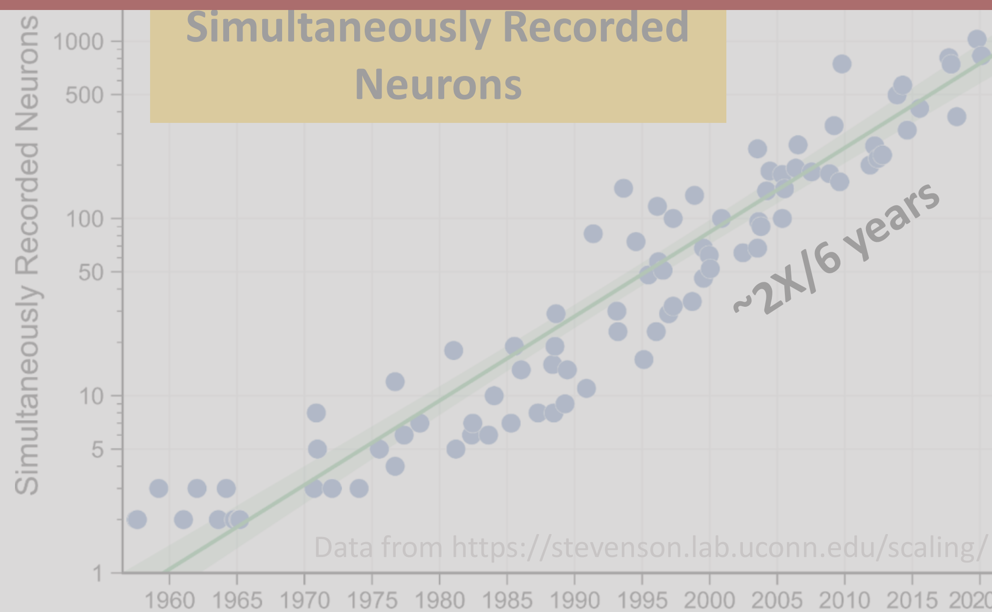
Constraints for a *portable implanted device*

1. Fast (real-time, <5ms detection latency)
2. Low-power & low-area
3. Scalable

Data quickly outpacing analysis techniques

Capture Capability Growing Exponentially

Existing solutions can’t cope



Constraints for a *portable implanted device*

1. Fast (real-time, <5ms detection latency)
2. Low-power & low-area
3. Scalable

Data quickly outpacing analysis techniques

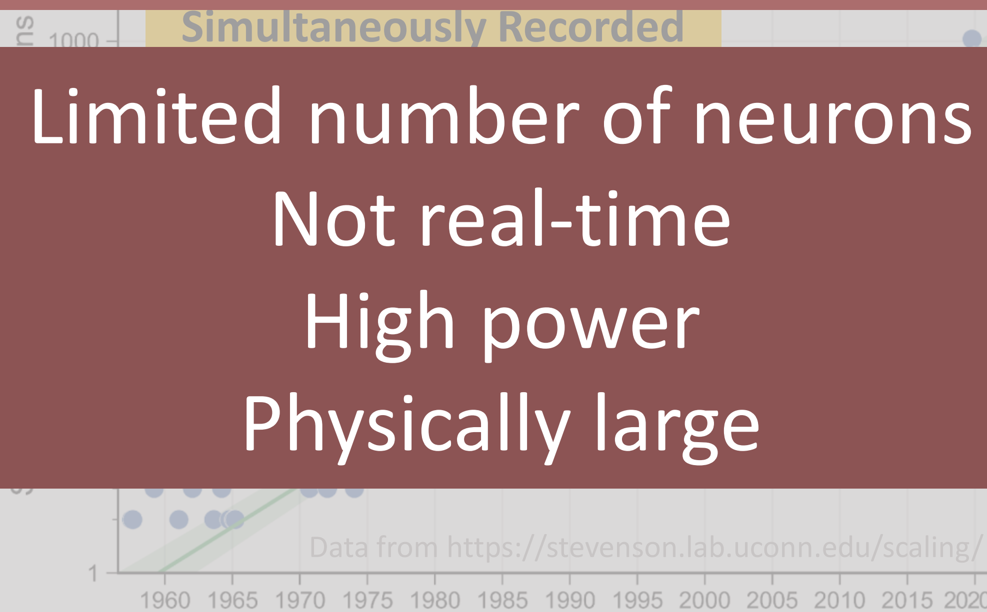
Existing solutions can't cope

Limited number of neurons

Not real-time

High power

Physically large



Constraints for a *portable implanted device*

1. Fast (real-time, <5ms detection latency)
2. Low-power & low-area
3. Scalable

Data quickly outpacing analysis techniques

Existing solutions can't cope

Limited number of neurons

Not real-time

High power

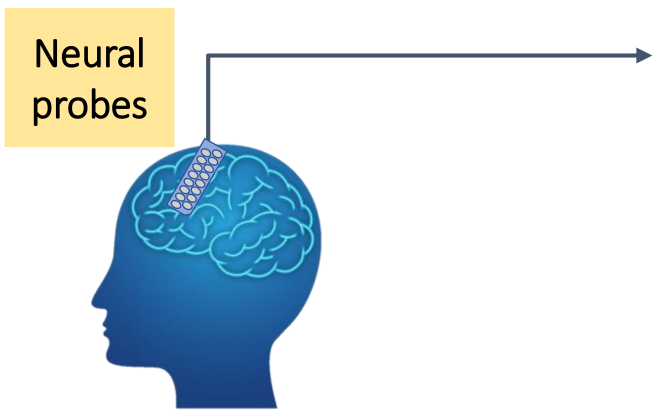
Physically large

Brain activity decoding is
memory intensive &
computationally expensive

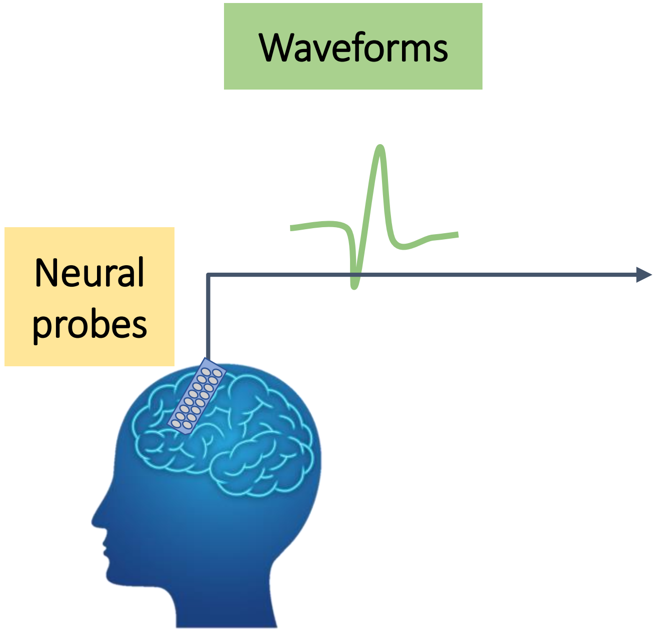
Roadmap to NOEMA

- Input to the system
- Template matching
- Baseline design & Noema
- Results

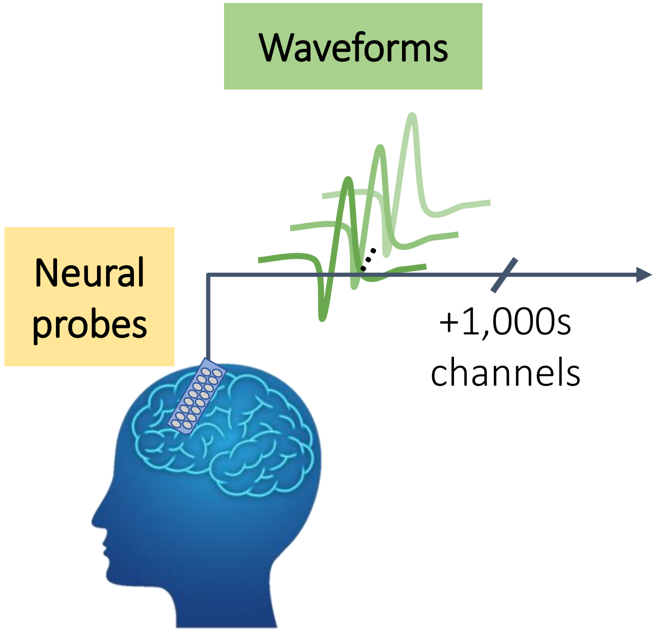
The Raw Input Data



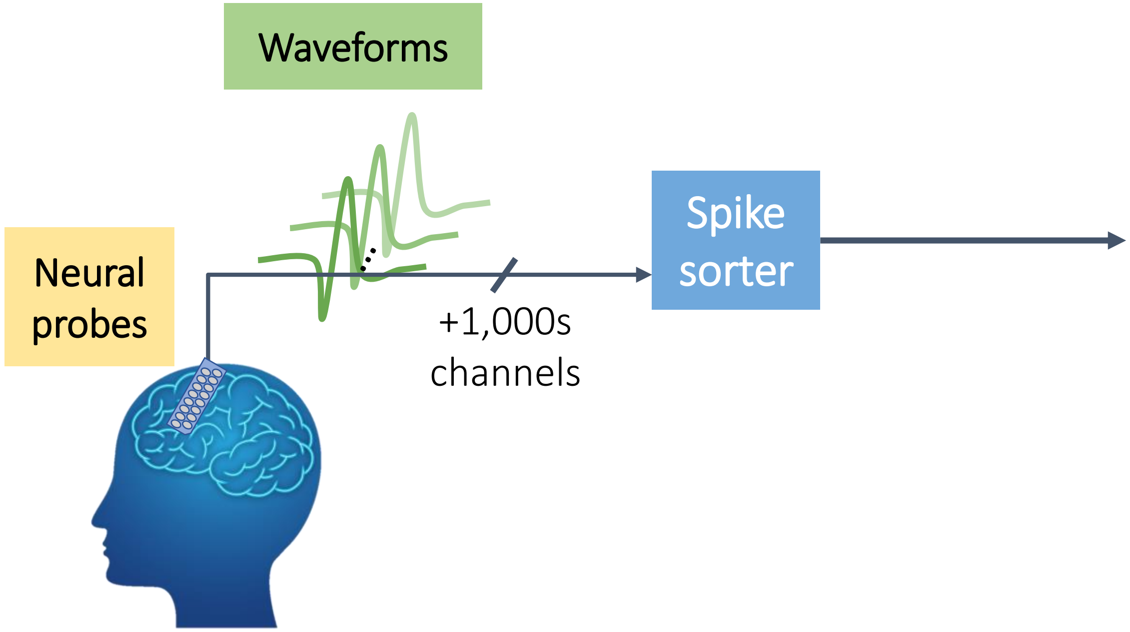
The Raw Input Data



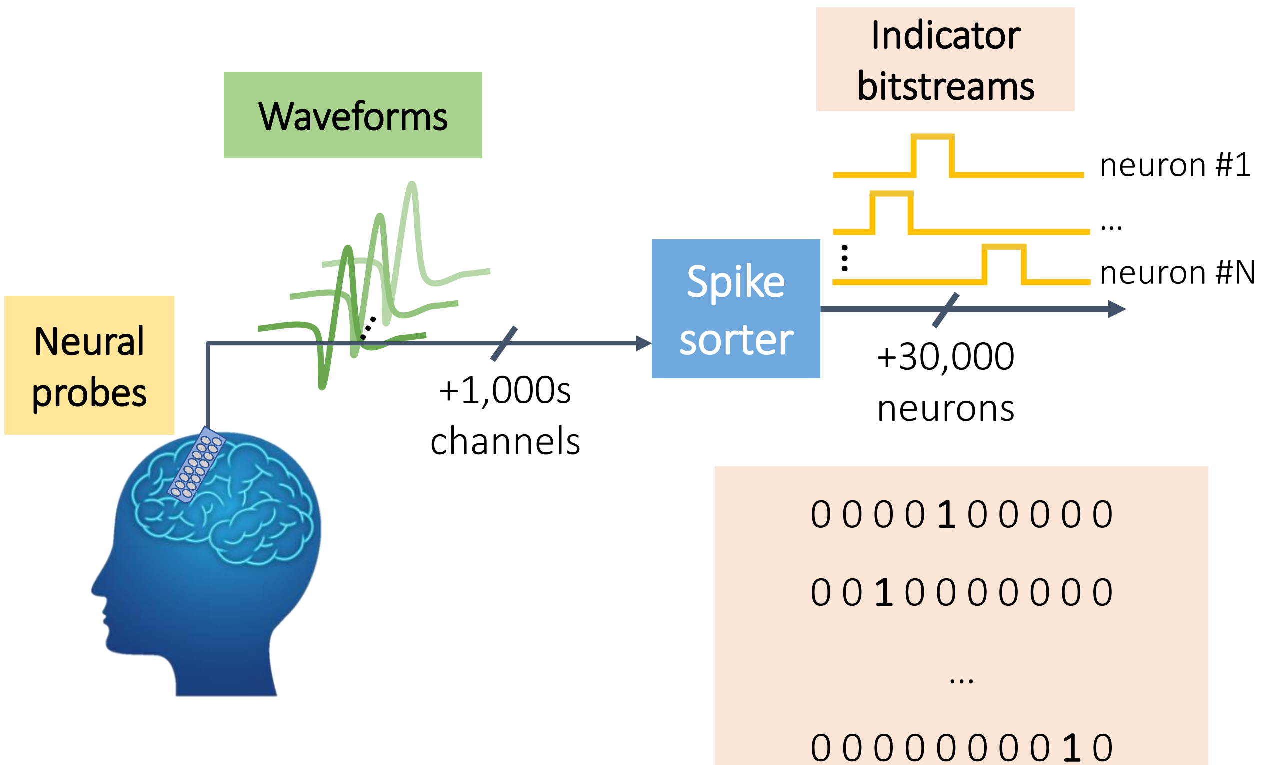
The Raw Input Data



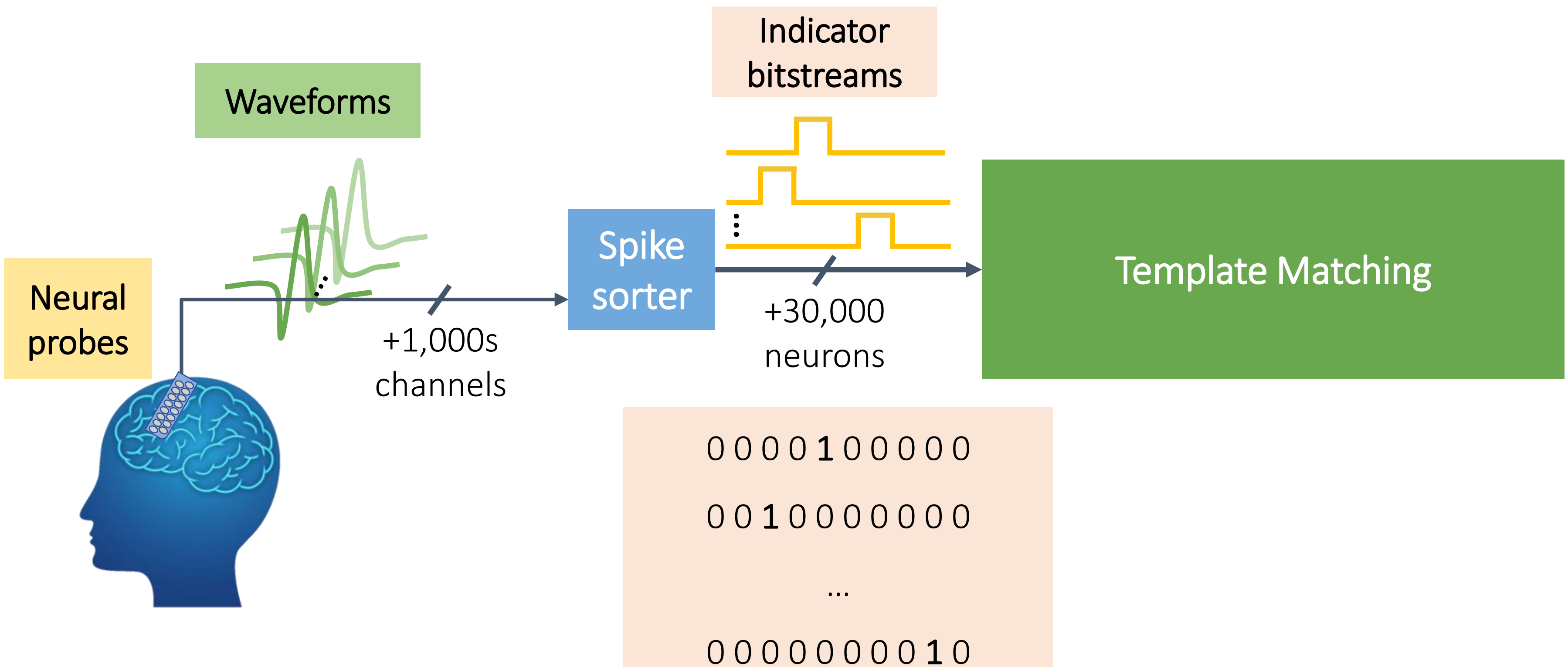
The Raw Input Data



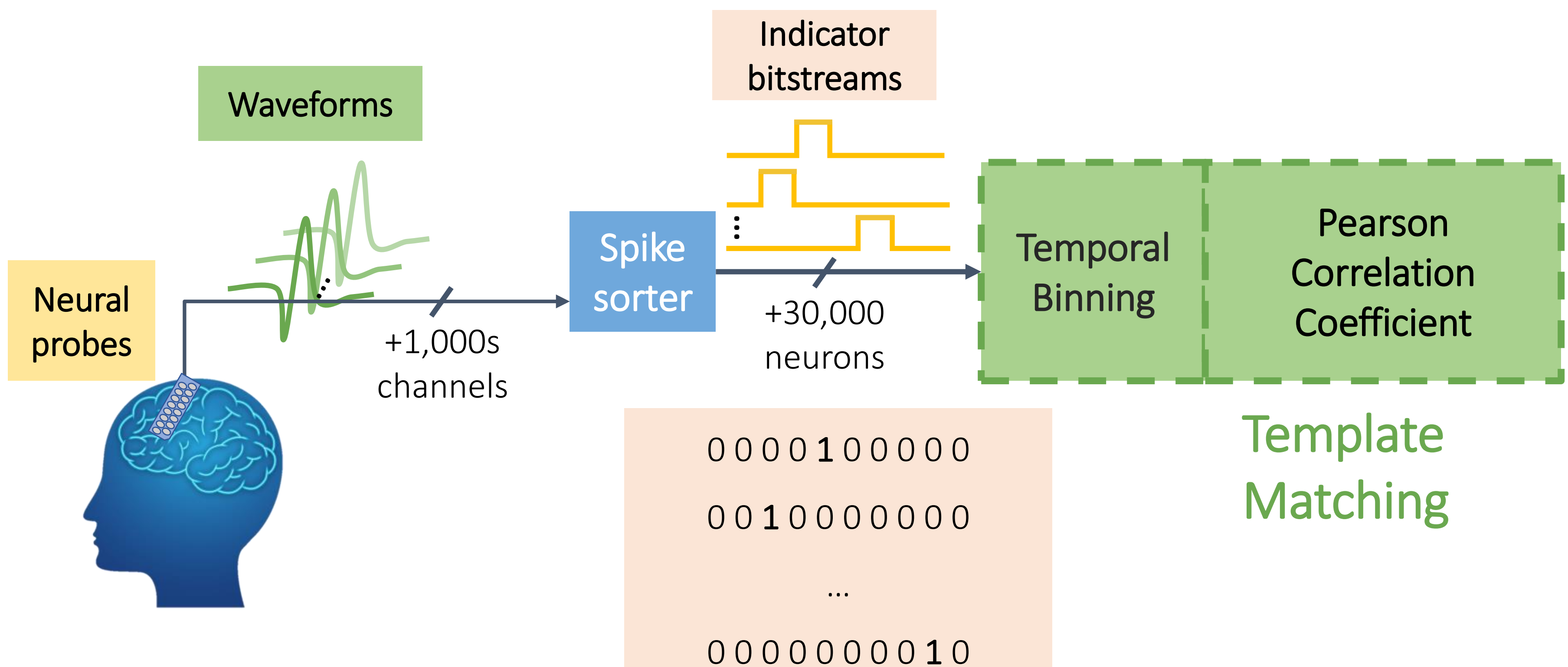
Processing Pipeline



Processing Pipeline

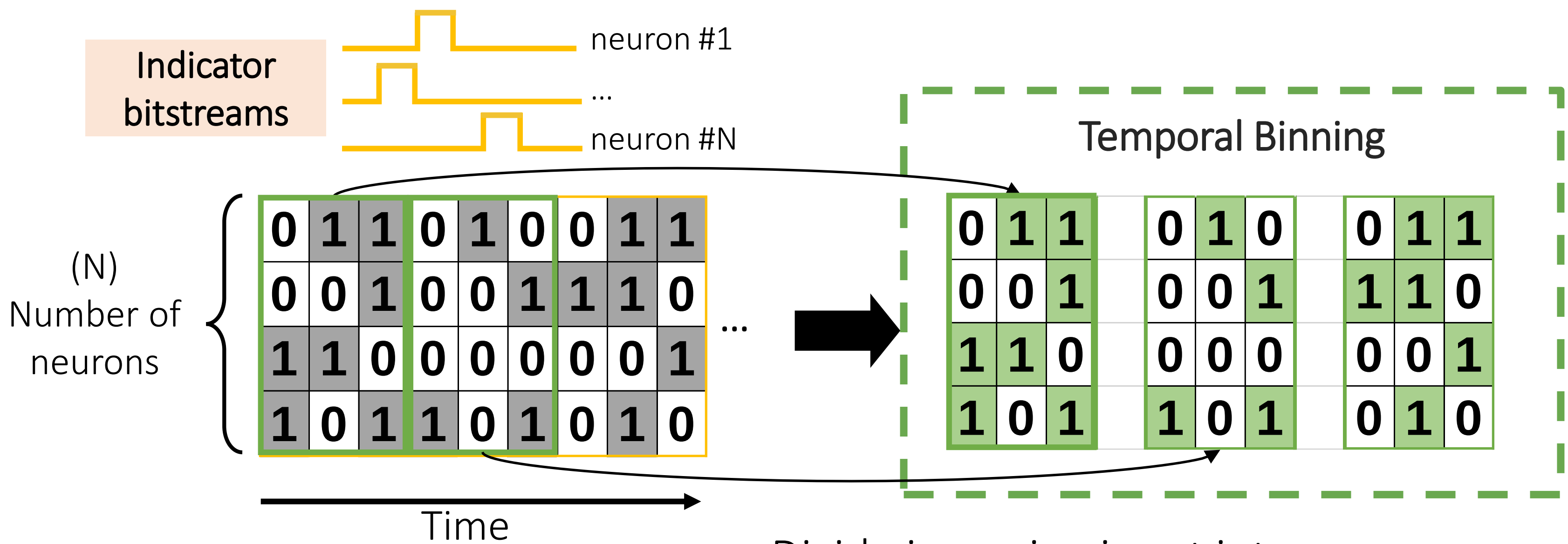


Processing Pipeline



Temporal Binning

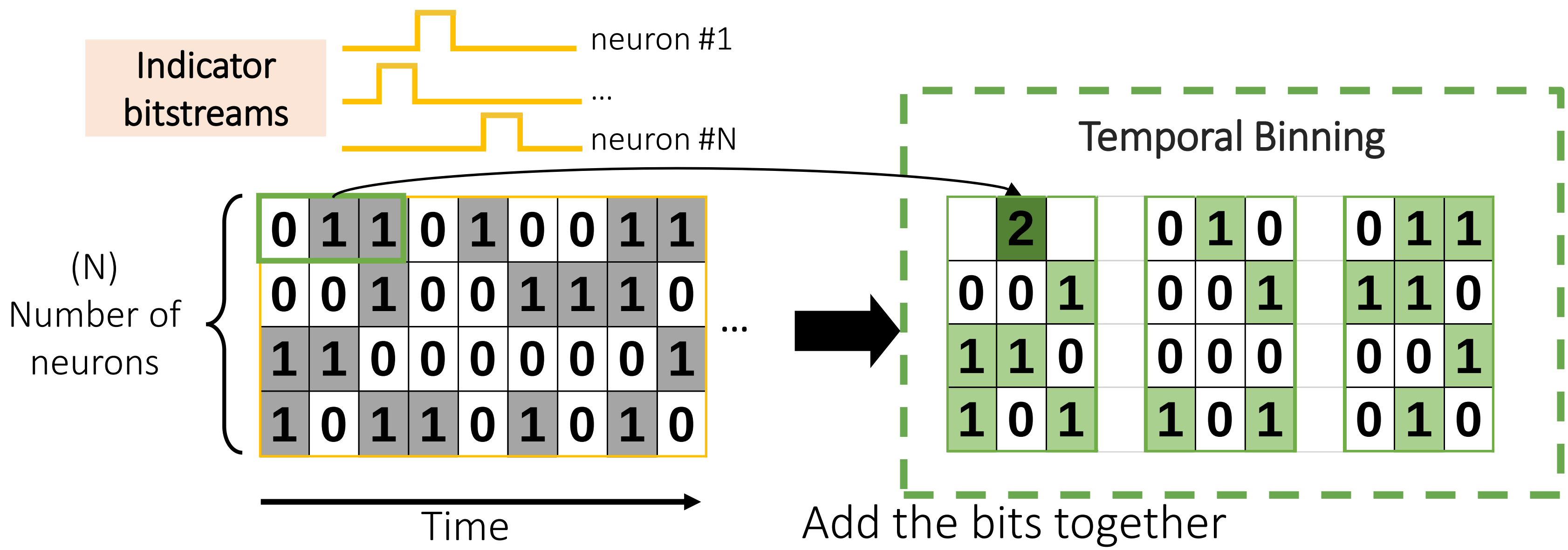
Data "smoothing"



Divide incoming input into groups of 3 (example bin size)

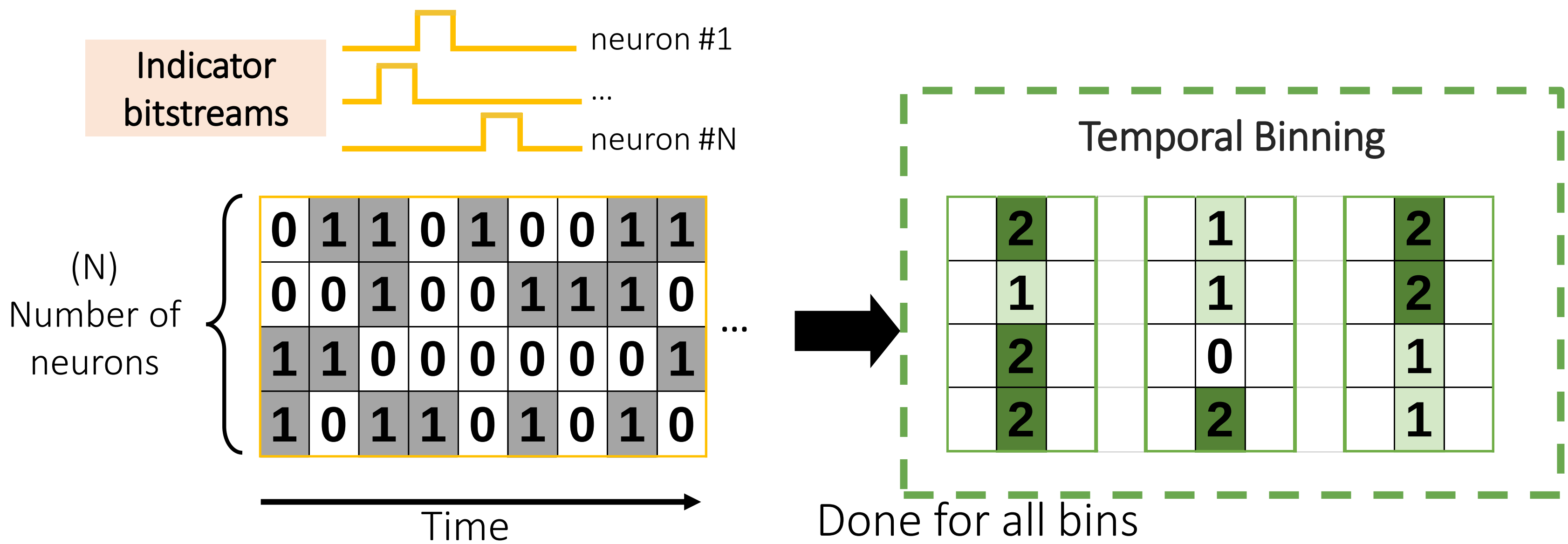
Temporal Binning

Data "smoothing"



Temporal Binning

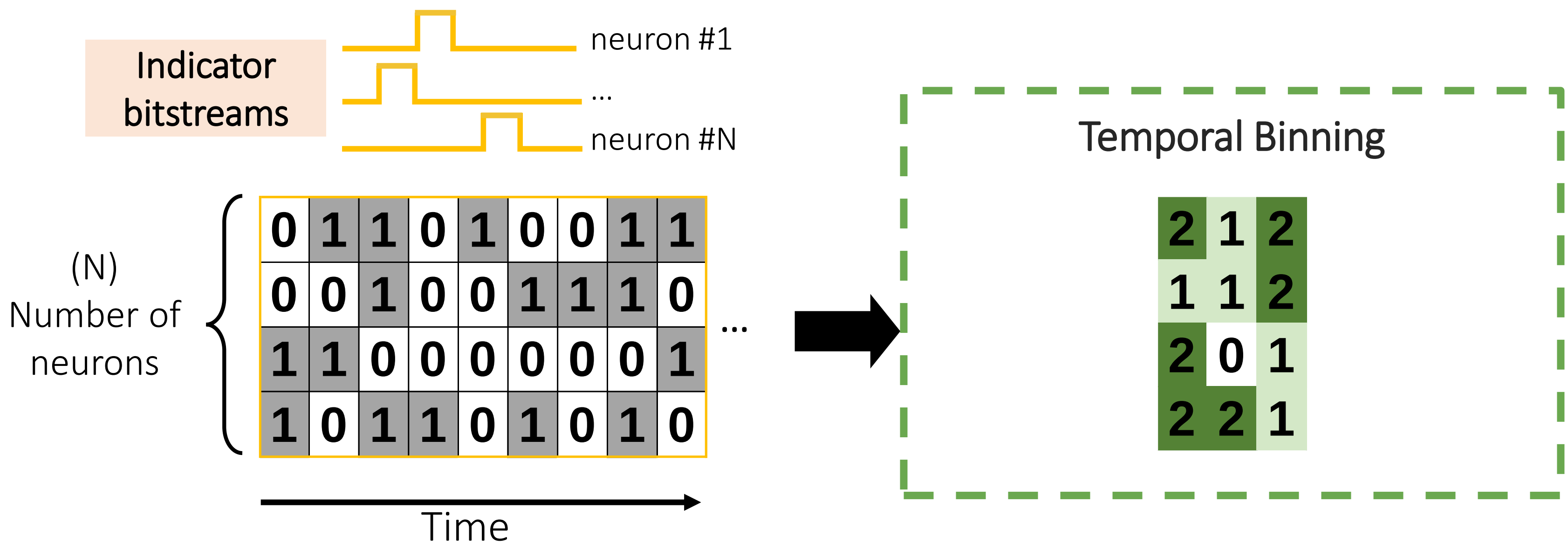
Data "smoothing"



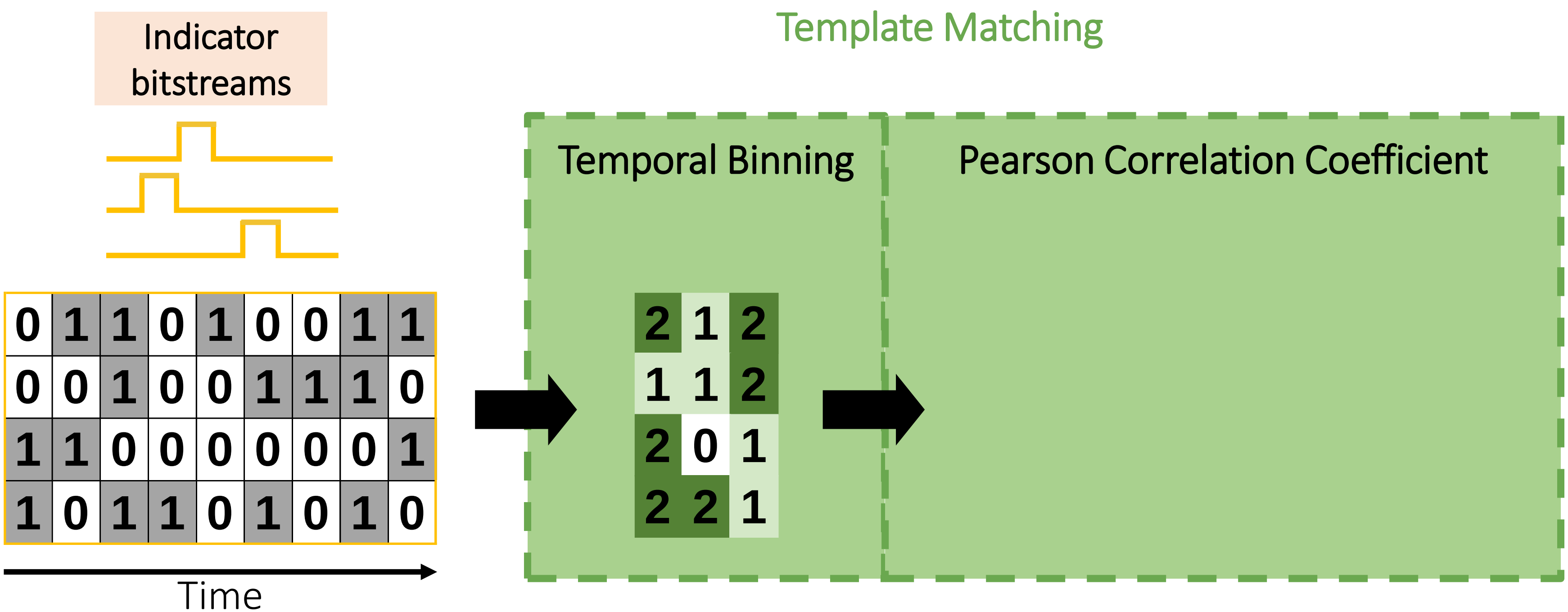
Done for all bins in a template

Temporal Binning

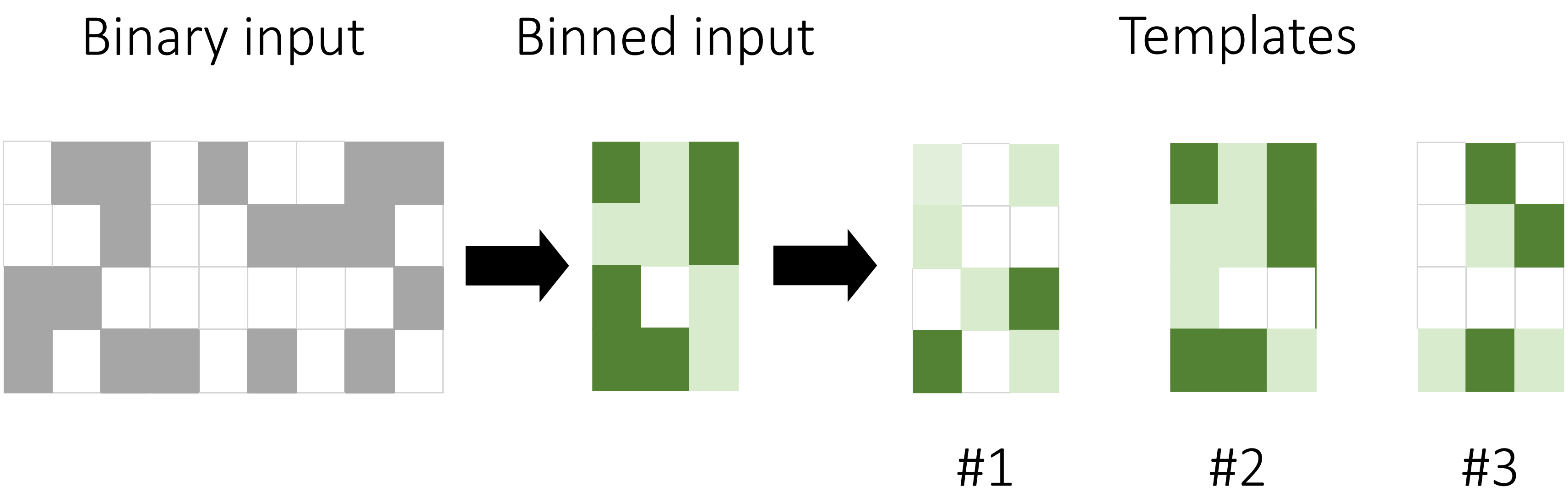
Data "smoothing"



Template Matching

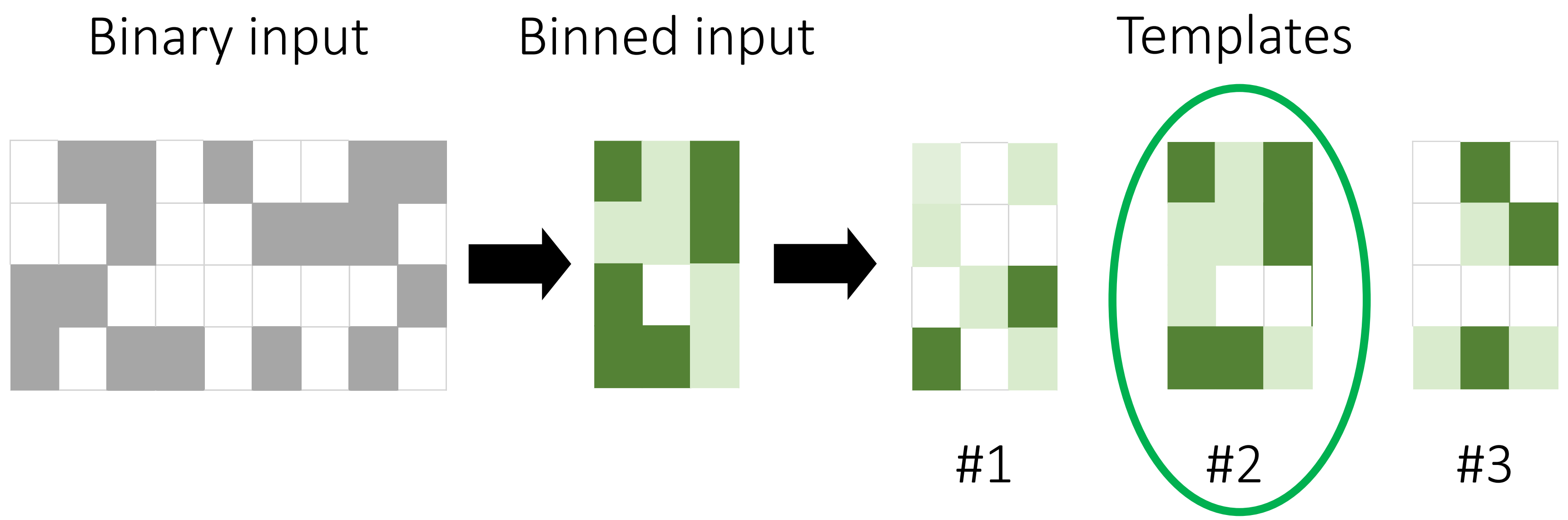


Template Matching



Which template does the input most closely resemble?

Template Matching



How do neuroscientists determine this?

Pearson Correlation Coefficient (PCC)

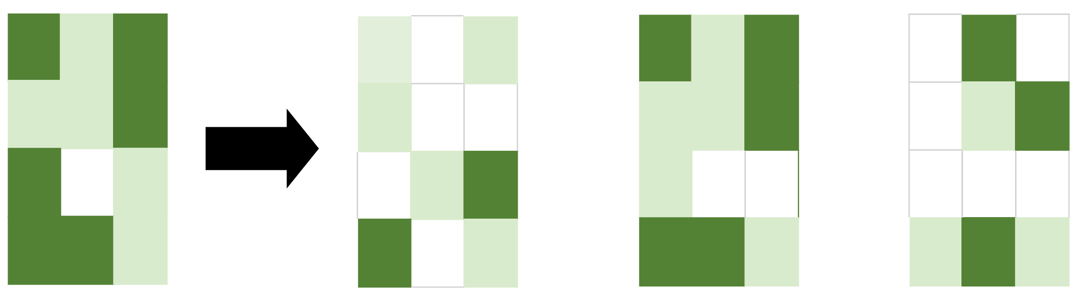
Widely used metric to measure
the "closeness" of two matrices

$$r(X,Y) = \frac{\sum_{i=1}^{L}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{L}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{L}(y_i - \overline{y})^2}}$$

PCC Example

Binned input

Templates



#1
Move
right arm

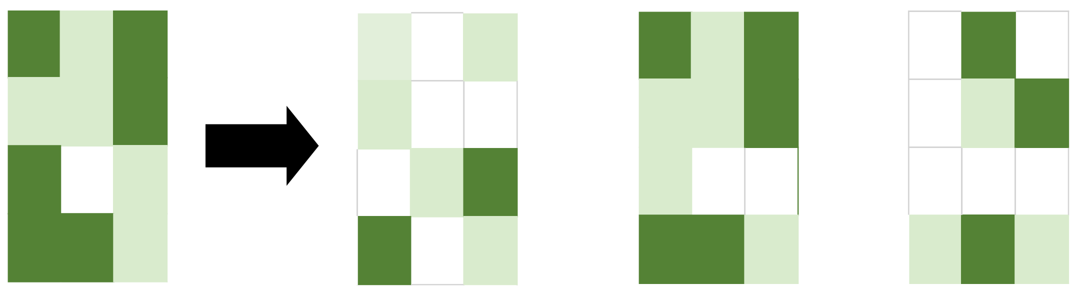
#2
Move
left arm

#3
Move
left leg

PCC Example

Binned input

Templates



#1
Move
right arm

0.135

#2
Move
left arm

0.857

#3
Move
left leg

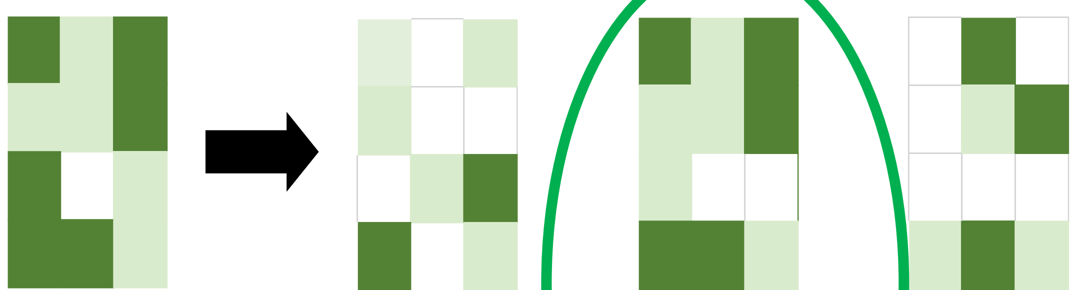
0.196

PCC scores (r)

PCC Example

Binned input

Templates



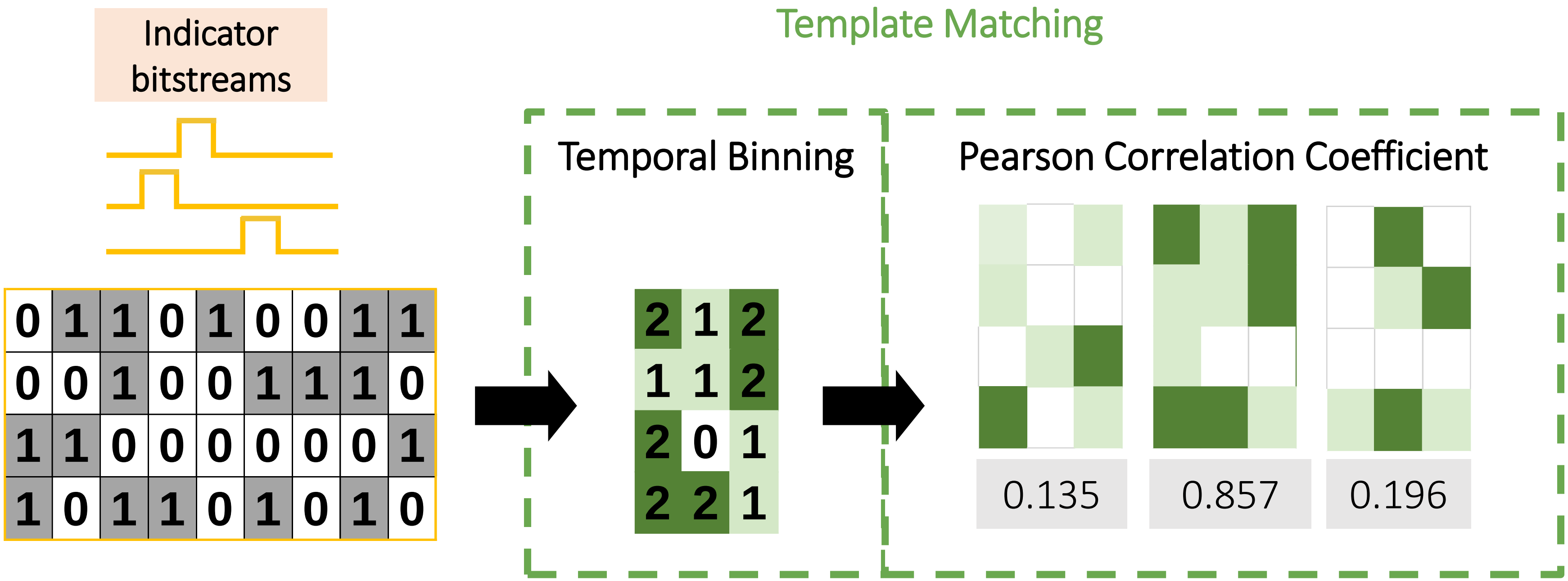
#1
Move
right arm
0.135

#2
Move
left arm
0.857

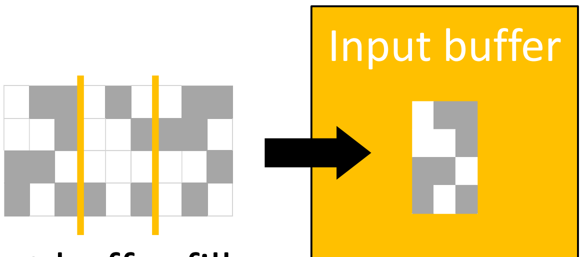
#3
Move
left leg
0.196

PCC scores (*r*)

Template Matching Overview



Costs of baseline template matching design

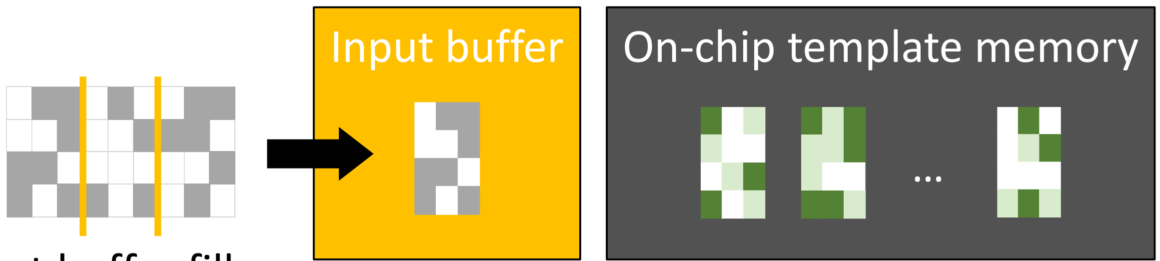


Entire input buffer fills before compute begins
→ High latency

<u>Most difficult requirement</u>

5ms for real-time

Costs of baseline template matching design



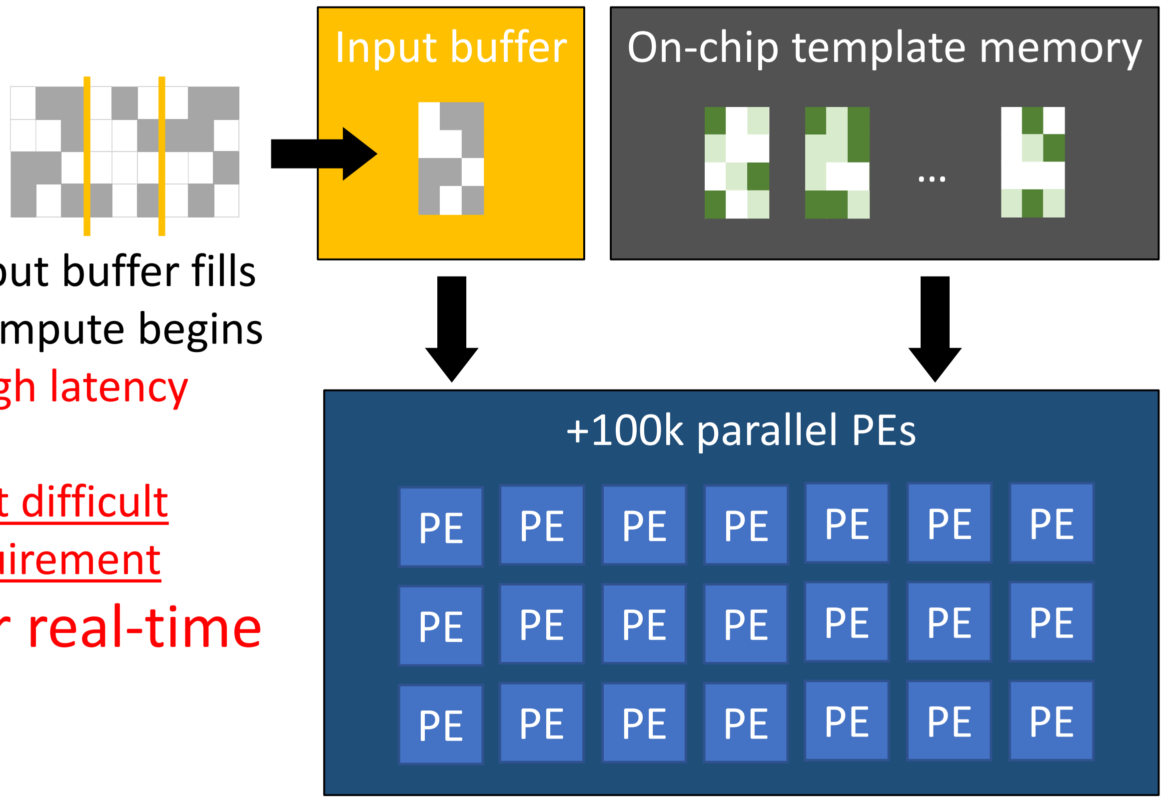
Entire input buffer fills before compute begins
→ High latency

<u>Most difficult requirement</u>

5ms for real-time

Storage of input + templates
→ Large memory cost
e.g. +1.24 Gb each

Costs of baseline template matching design



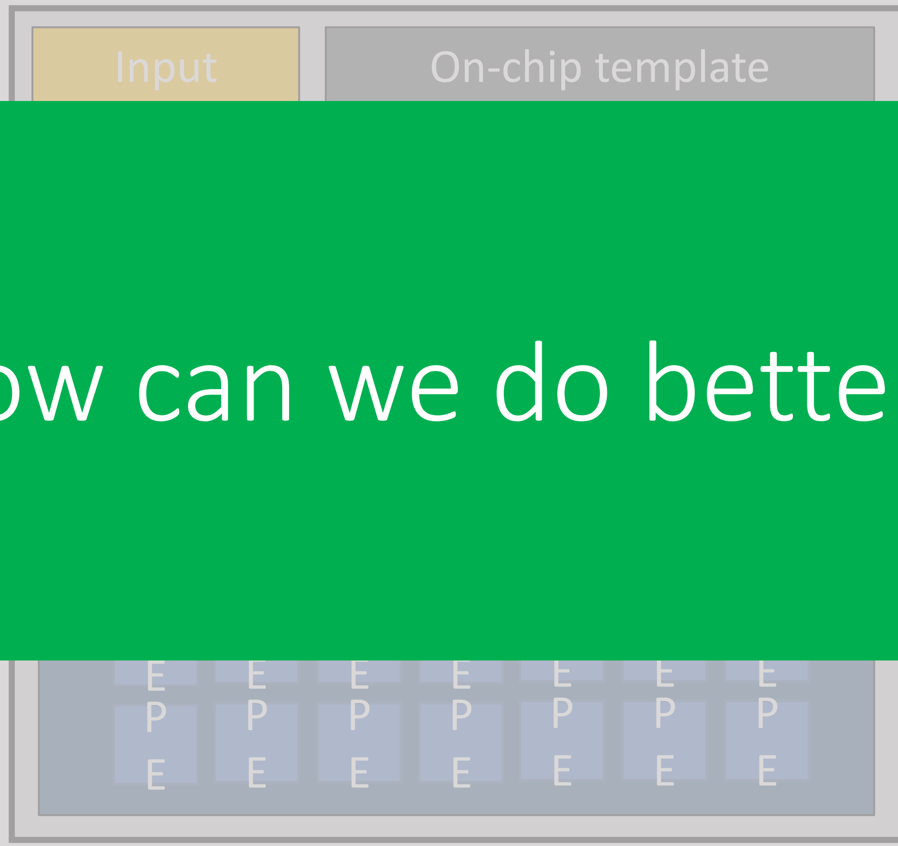
Entire input buffer fills before compute begins
→ High latency

<u>Most difficult requirement</u>
5ms for real-time

Storage of input + templates
→ Large memory cost
e.g. +1.24 Gb each

Many processing elements
→ Large area cost

Costs of baseline template matching design



Storage of input + templates

→ Large area cost

How can we do better?

NOEMA [MICRO'21, Patented]:
Brain Interfaces at the Edge

A multidisciplinary collaboration effort in analyzing and developing a custom hardware platform to decipher the brain neural activity

NOEMA [MICRO'21, Patented]:
Brain Interfaces at the Edge

A multidisciplinary collaboration effort in analyzing and developing a custom hardware platform to decipher the brain neural activity

Enabling truly portable systems for processing high-resolution brain activity signals for treatment, augmentation, and repair of brain functions

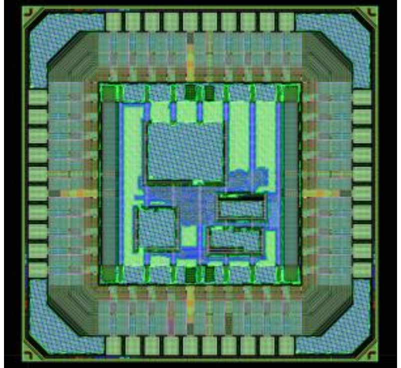
NOEMA [MICRO'21, Patented]:
Brain Interfaces at the Edge

A multidisciplinary collaboration effort in analyzing and developing a custom hardware platform to decipher the brain neural activity

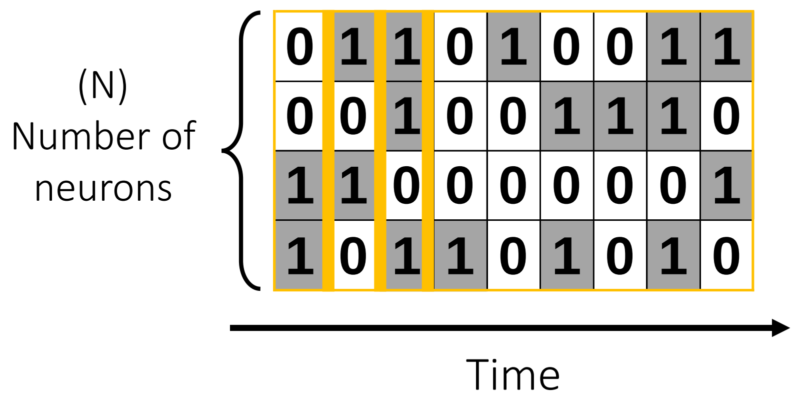
Enabling truly portable systems for processing high-resolution brain activity signals for treatment, augmentation, and repair of brain functions

NOEMA's Prototype Chip

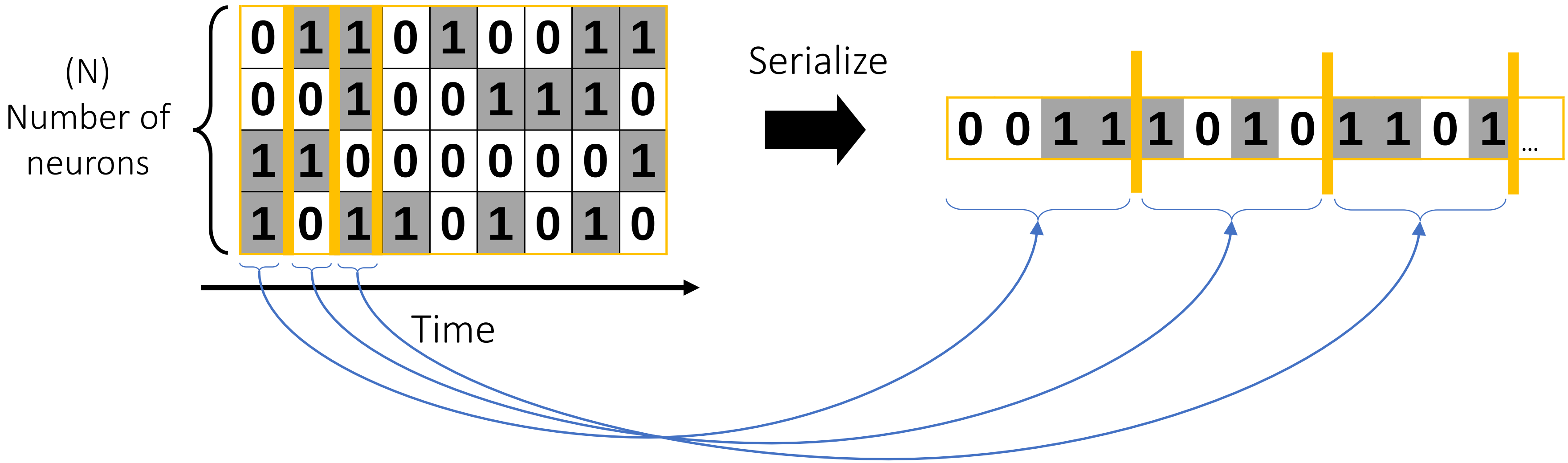
- Fabricated with **TSMC 65nm GP** technology
- Only **24µsec** latency!
- **5 sec** experience, **1K** neurons @ **0.73 mW**
- Scales to **30K** neurons, **10**× more than have ever been recorded
- Scales to meet ***future*** demand!



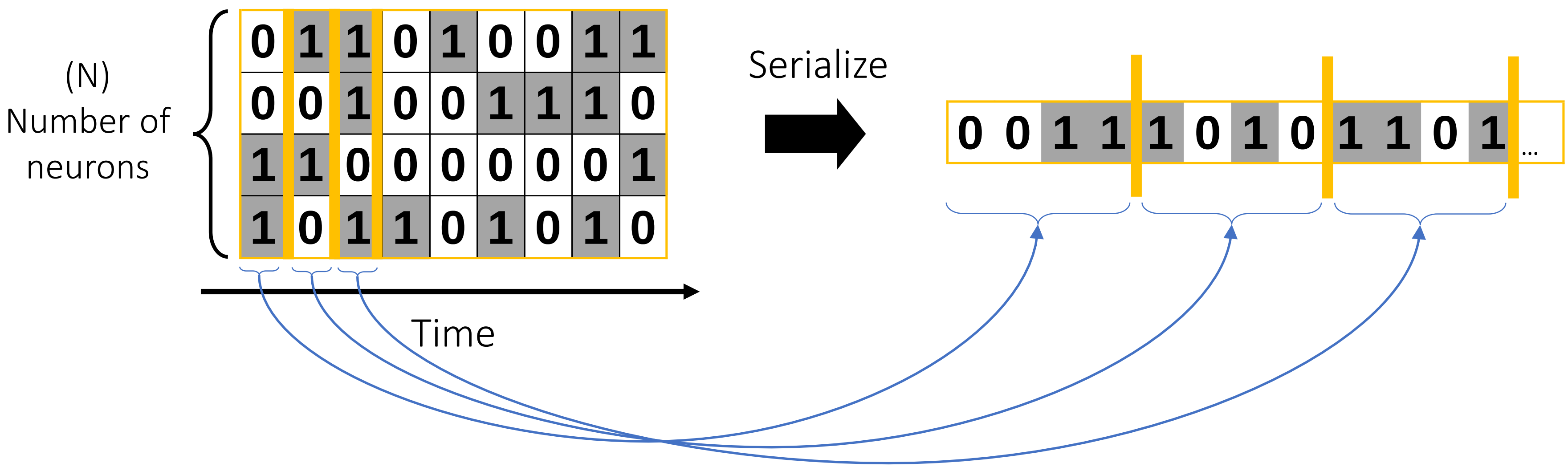
Input Serialization & PCC Reformulation



Input Serialization & PCC Reformulation



Input Serialization & PCC Reformulation



$$r(X,Y) = \frac{\sum_{i=1}^{L} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{L} (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{L} (y_i - \overline{y})^2}}$$

Reformulation

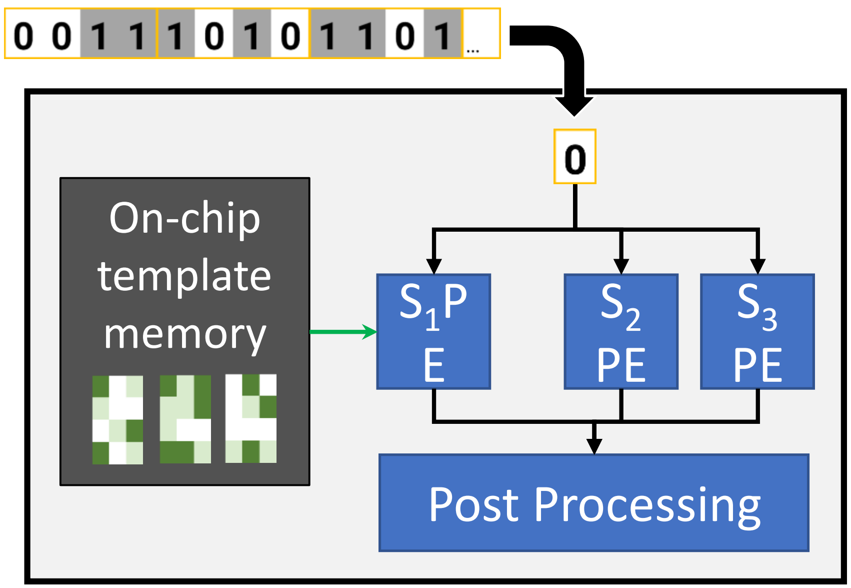
$$r[t]^2 = \frac{(C_1 S_1[t] - C_2 S_2[t])^2}{C_3\left(C_1 S_3[t] - S_2[t]^2\right)}$$

NOEMA's innovations

$$r[t]^2 = \frac{(C_1 S_1[t] - C_2 S_2[t])^2}{C_3\left(C_1 S_3[t] - S_2[t]^2\right)}$$

<u>Bit-serial input</u>

- No buffering overhead
- Compute immediately when received

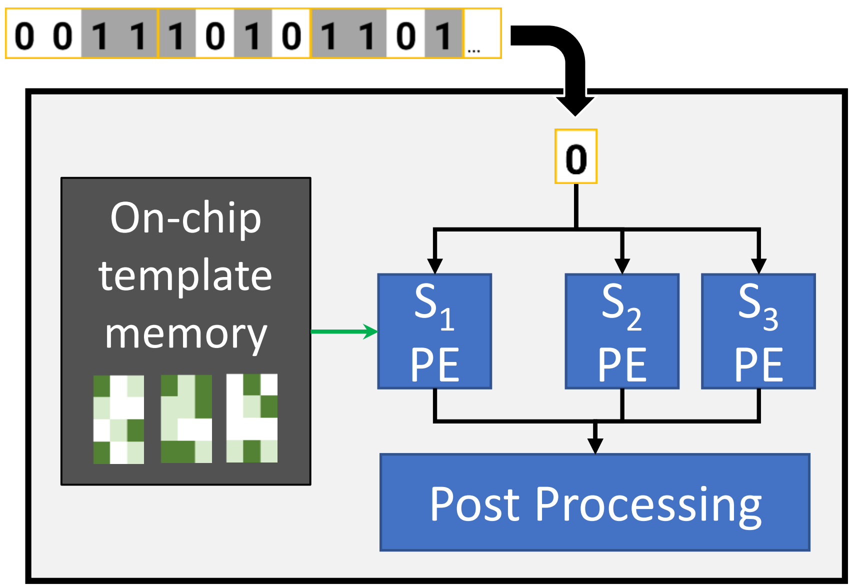


NOEMA's innovations

$$r[t]^2 = \frac{(C_1 S_1[t] - C_2 S_2[t])^2}{C_3\,(C_1 S_3[t] - S_2[t]^2)}$$

Bit-serial input

- No buffering overhead
- Compute immediately when received



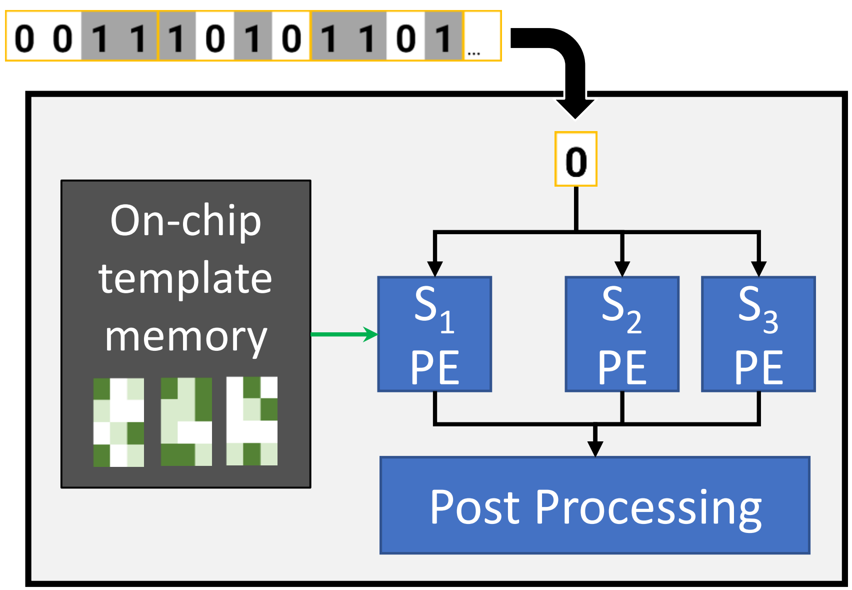
Near-memory bit-serial PEs

- Based on reformulated PCC
- Tiny, easy to scale

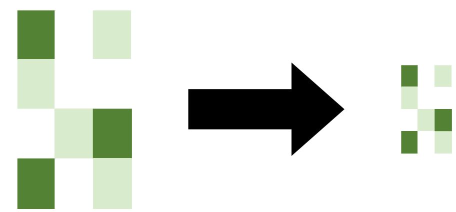
NOEMA's innovations

$$r[t]^2 = \frac{(C_1 S_1[t] - C_2 S_2[t])^2}{C_3\,(C_1 S_3[t] - S_2[t]^2)}$$

Bit-serial input
- No buffering overhead
- Compute immediately when received



Simple memory compression (~2.8x)

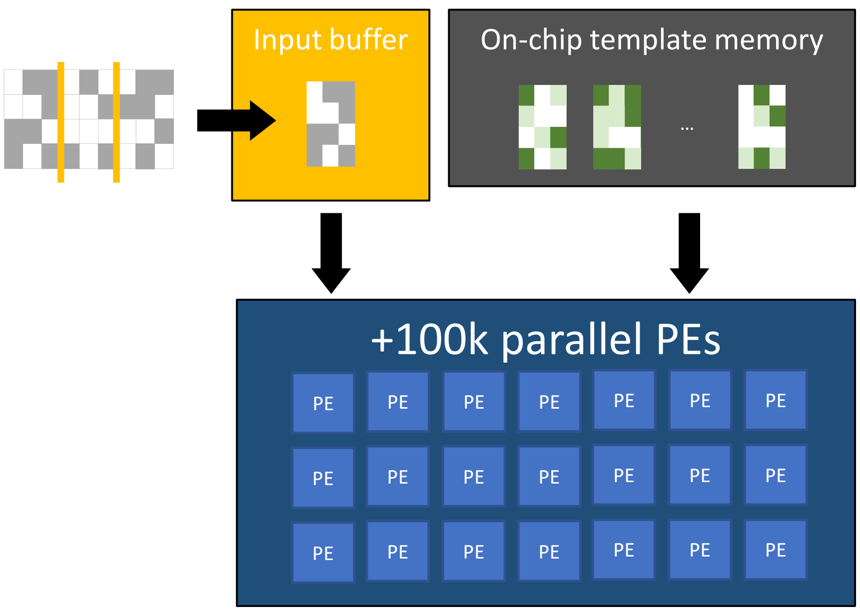


Near-memory bit-serial PEs
- Based on reformulated PCC
- Tiny, easy to scale

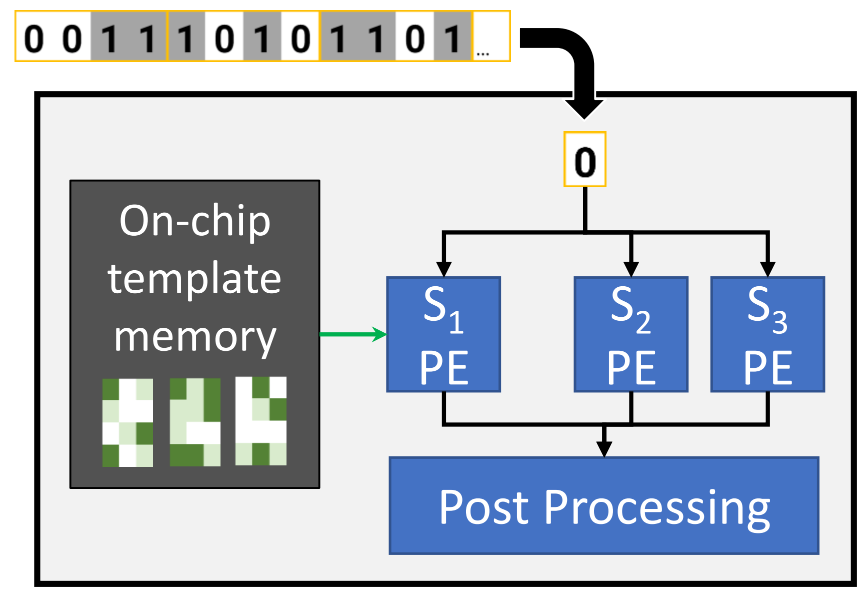
Fits well with existing probe interfaces (time-multiplexed ADC out)

Baseline to NOEMA Overview

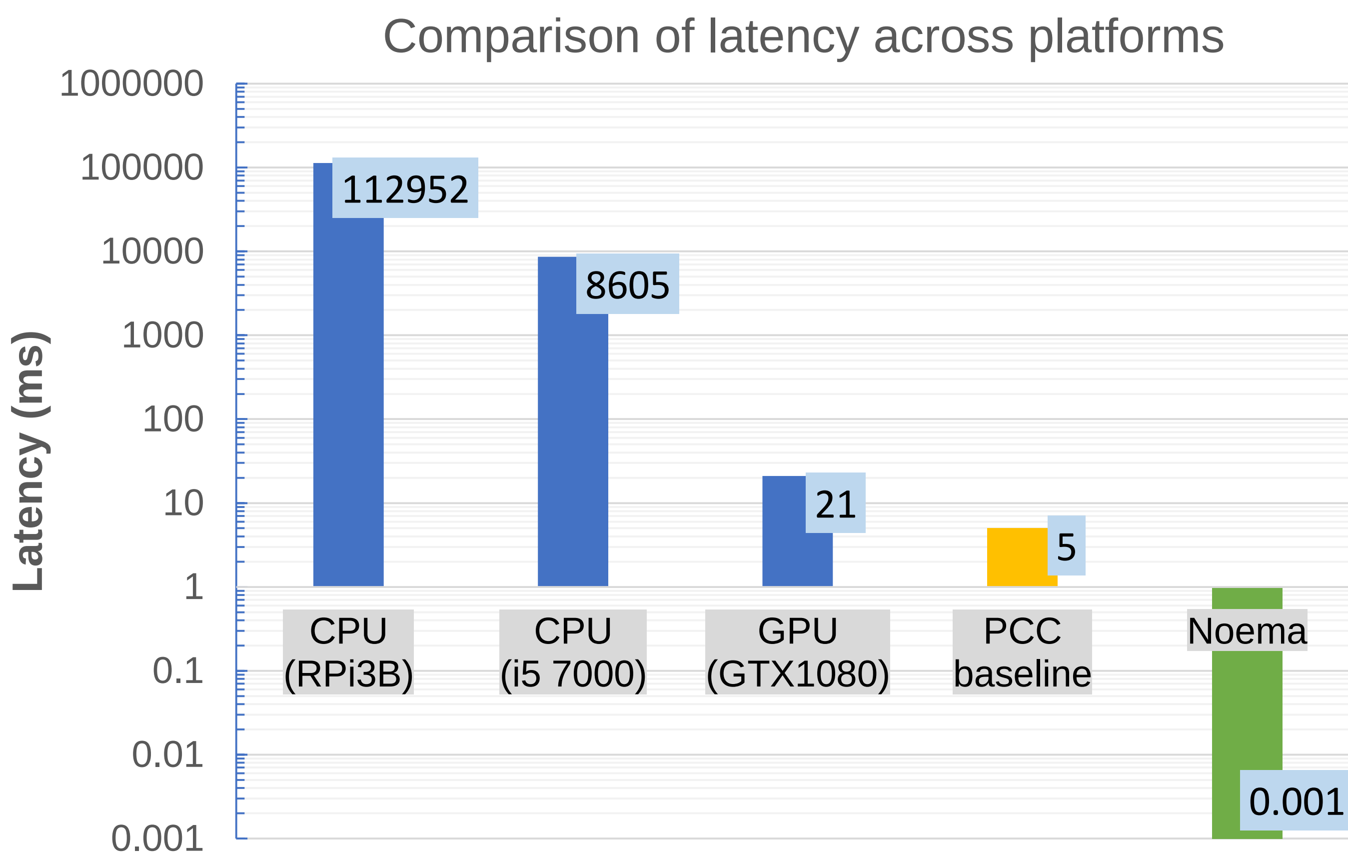
Baseline



NOEMA

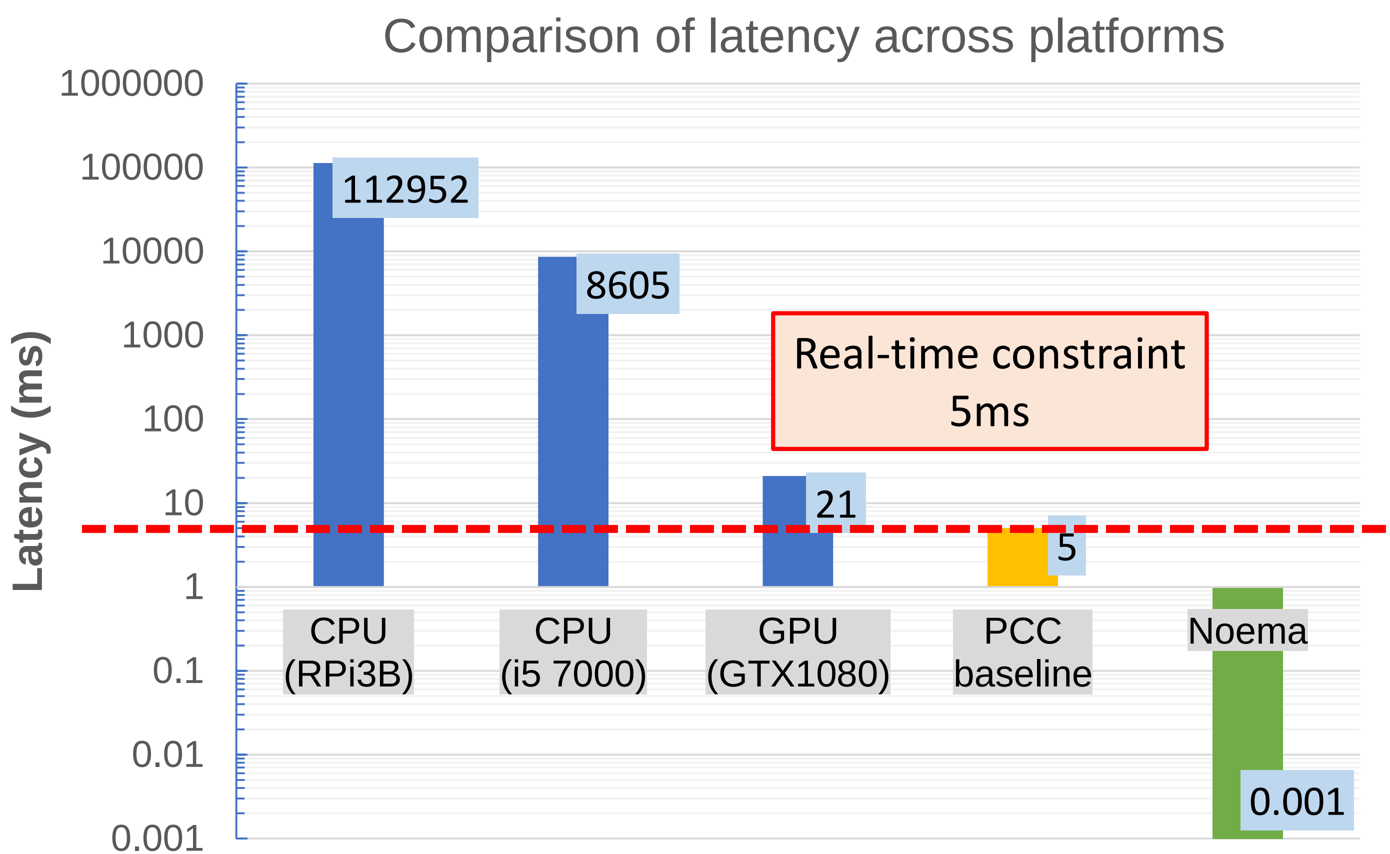


Performance Results



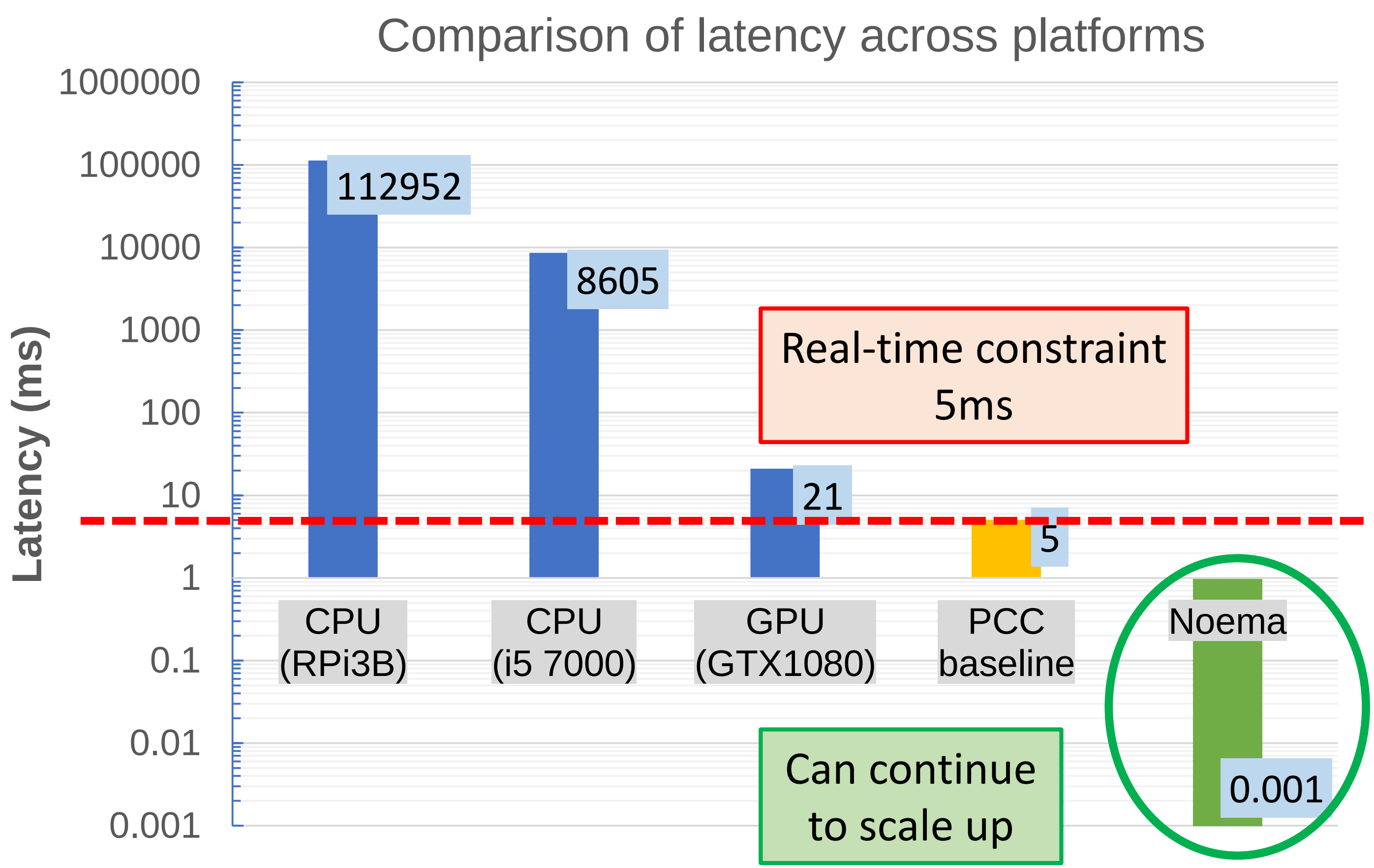
* For the most demanding configuration tested (9 sec experience, 30K neurons)

Performance Results



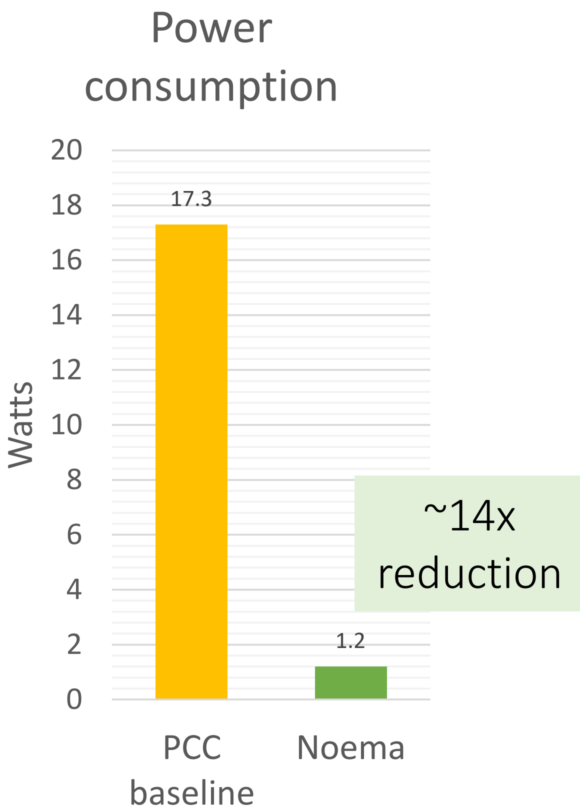
* For the most demanding configuration tested (9 sec experience, 30K neurons)

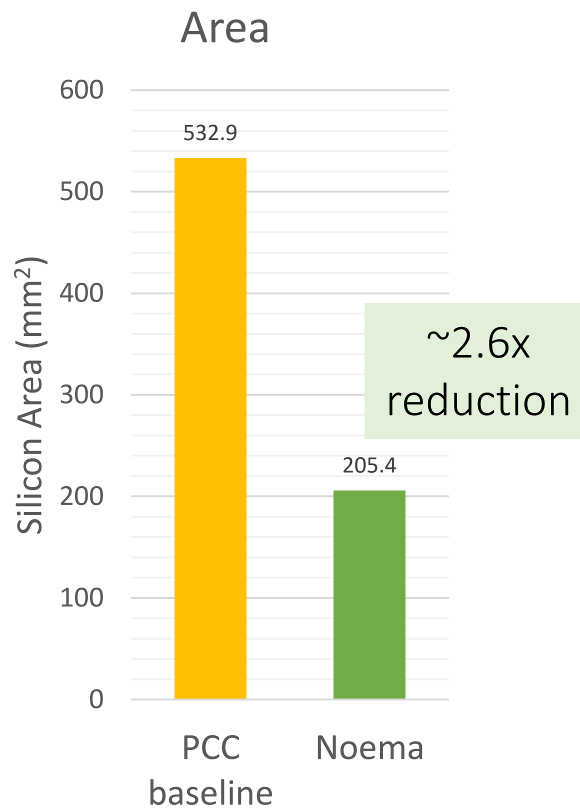
Performance Results



* For the most demanding configuration tested (9 sec experience, 30K neurons)

Power & Area Results





* For the most demanding configuration tested (9 sec experience, 30K neurons)

The NOEMA Family

Device	F_{max} (MHz)	Neurons (thousands)	Templates	Duration[1] (seconds)	Resolution[2] (milliseconds)	Requirements[3]		Implementation	
						Compute (GOPs)	Memory (Mb)	FPGA[4]	ASIC[5]
NOEMA01K1T05S250	30	1	1	5	250	0.6	0.3	✓	✓

1. Duration of the decoded experience
2. Resolution window of the incoming activities.
   Activities within this windows are binned (averaged).
3. If executed on commodity hardware.
4. Intel's Stratix 10 FPGA
5. TSMC 65nm GP

The NOEMA Family

Device	F_{max} (MHz)	Neurons (thousands)	Templates	Duration[1] (seconds)	Resolution[2] (milliseconds)	Requirements[3]		Implementation	
						Compute (GOPs)	Memory (Mb)	FPGA[4]	ASIC[5]
NOEMA01K1T05S250	30	1	1	5	250	0.6	0.3	✓	✓
NOEMA10K2T05S005	300	10	2	5	5	628.0	114.4	✓	Planned

1. Duration of the decoded experience
2. Resolution window of the incoming activities.
   Activities within this windows are binned (averaged).
3. If executed on commodity hardware.
4. Intel's Stratix 10 FPGA
5. TSMC 65nm GP

The NOEMA Family

Device	F_{max} (MHz)	Neurons (thousands)	Templates	Duration[1] (seconds)	Resolution[2] (milliseconds)	Requirements[3]		Implementation	
						Compute (GOPs)	Memory (Mb)	FPGA[4]	ASIC[5]
NOEMA01K1T05S250	30	1	1	5	250	0.6	0.3	✓	✓
NOEMA10K2T05S005	300	10	2	5	5	628.0	114.4	✓	Planned
NOEMA20K3T09S250	600	20	3	9	250	64.8	33.0	✗[6]	Planned

1. Duration of the decoded experience
2. Resolution window of the incoming activities.
   Activities within this windows are binned (averaged).
3. If executed on commodity hardware.
4. Intel's Stratix 10 FPGA
5. TSMC 65nm GP
6. Not applicable; device can't meet target frequency.

The NOEMA Family

Device	F_{max} (MHz)	Neurons (thousands)	Templates	Duration[1] (seconds)	Resolution[2] (milliseconds)	Requirements[3]		Implementation	
						Compute (GOPs)	Memory (Mb)	FPGA[4]	ASIC[5]
NOEMA01K1T05S250	30	1	1	5	250	0.6	0.3	✓	✓
NOEMA10K2T05S005	300	10	2	5	5	628.0	114.4	✓	Planned
NOEMA20K3T09S250	600	20	3	9	250	64.8	33.0	✗[6]	Planned
NOEMA30K4T09S005	900	30	4	9	5	6786.4	1236.0	✗[6]	Planned

1. Duration of the decoded experience
2. Resolution window of the incoming activities.
   Activities within this windows are binned (averaged).
3. If executed on commodity hardware.
4. Intel's Stratix 10 FPGA
5. TSMC 65nm GP
6. Not applicable; device can't meet target frequency.

NOEMA's ASIC Devices

Device	Silicon Area (mm^2)			Power (mW)			Latency (µs)	Chip Status
	Memory	Logic	Total	Memory	Logic	Total		
NOEMA01K1T05S250	0.36	0.07	0.43*	0.30	0.43	0.73	23.9	In lab+#

* Core only; 2.1mm² total silicon area.
+ Fabricated with TSMC 65nm GP
# Also tested on Intel's Stratix 10 FPGA

NOEMA's ASIC Devices

Device	Silicon Area (mm^2)			Power (mW)			Latency (µs)	Chip Status
	Memory	Logic	Total	Memory	Logic	Total		
NOEMA01K1T05S250	0.36	0.07	0.43*	0.30	0.43	0.73	23.9	In lab+#
NOEMA10K2T05S005	28.46	1.35	29.81	89.78	84.28	174.06	2.8	Simulated#

* Core only; 2.1mm^2 total silicon area.
+ Fabricated with TSMC 65nm GP
# Also tested on Intel's Stratix 10 FPGA

NOEMA's ASIC Devices

Device	Silicon Area (mm²)			Power (mW)			Latency (µs)	Chip Status
	Memory	Logic	Total	Memory	Logic	Total		
NOEMA01K1T05S250	0.36	0.07	0.43*	0.30	0.43	0.73	23.9	In lab+#
NOEMA10K2T05S005	28.46	1.35	29.81	89.78	84.28	174.06	2.8	Simulated#
NOEMA20K3T09S250	6.26	0.09	6.25	18.55	9.68	28.23	1.5	Simulated

* Core only; 2.1mm² total silicon area.
+ Fabricated with TSMC 65nm GP
# Also tested on Intel's Stratix 10 FPGA

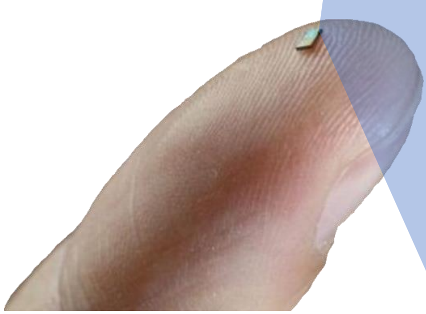
NOEMA's ASIC Devices

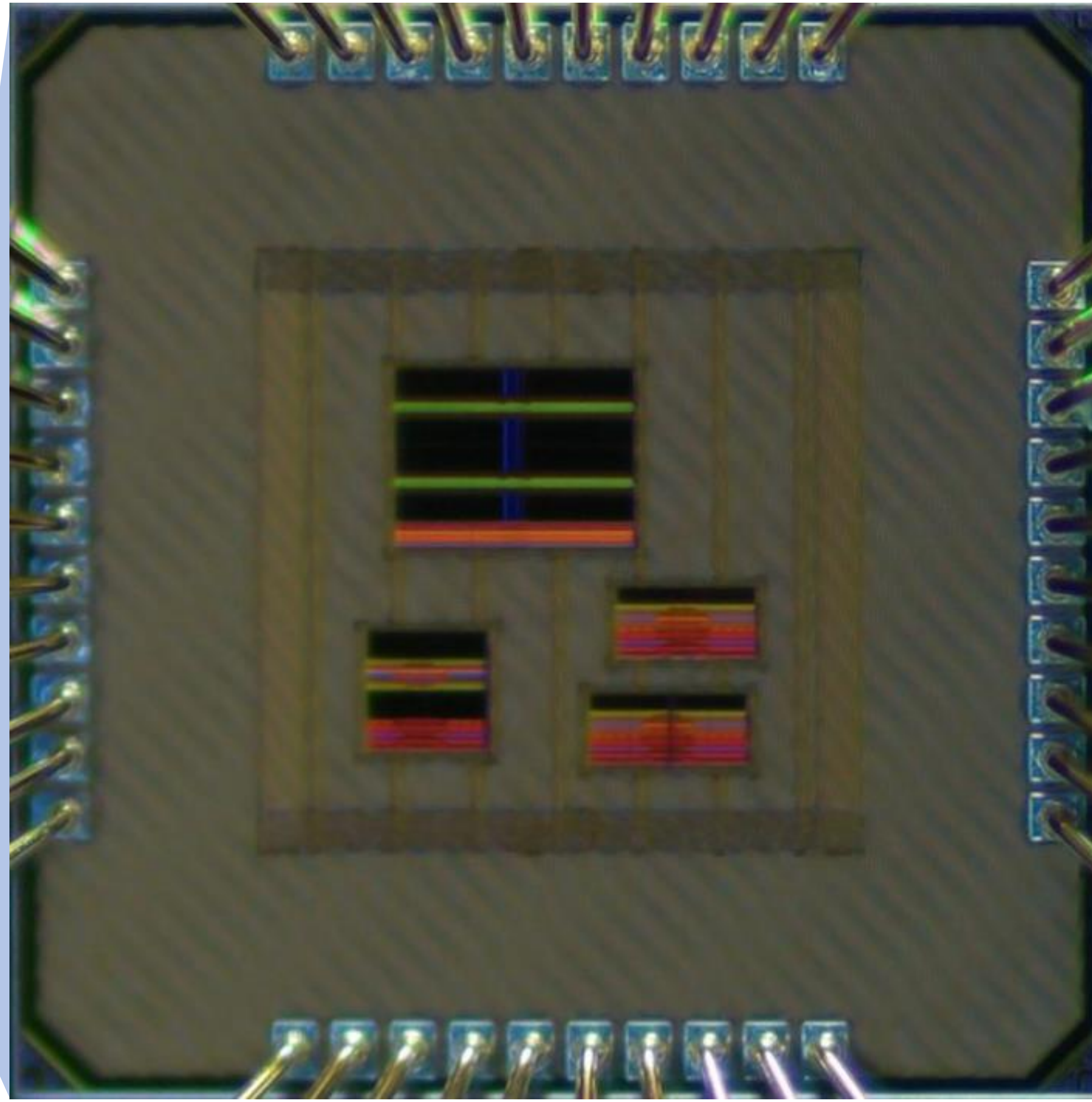
Device	Silicon Area (mm^2)			Power (mW)			Latency (µs)	Chip Status
	Memory	Logic	Total	Memory	Logic	Total		
NOEMA01K1T05S250	0.36	0.07	0.43*	0.30	0.43	0.73	23.9	In lab+#
NOEMA10K2T05S005	28.46	1.35	29.81	89.78	84.28	174.06	2.8	Simulated#
NOEMA20K3T09S250	6.26	0.09	6.25	18.55	9.68	28.23	1.5	Simulated
NOEMA30K4T09S005	202.00	3.42	205.42	682.70	522.76	1205.46	1.0	Simulated

* Core only; 2.1mm² total silicon area.
+ Fabricated with TSMC 65nm GP
# Also tested on Intel's Stratix 10 FPGA

NOEMA01K05S250MS

- TSMC **65nm** GP
- **24µsec** latency
- **1K** neurons (scales to **30K**)
- 5sec experience
- Consumes **0.73mW**
- Equivalent of 600MOPs 32bit-FP



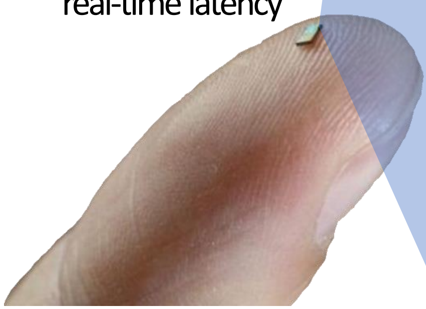


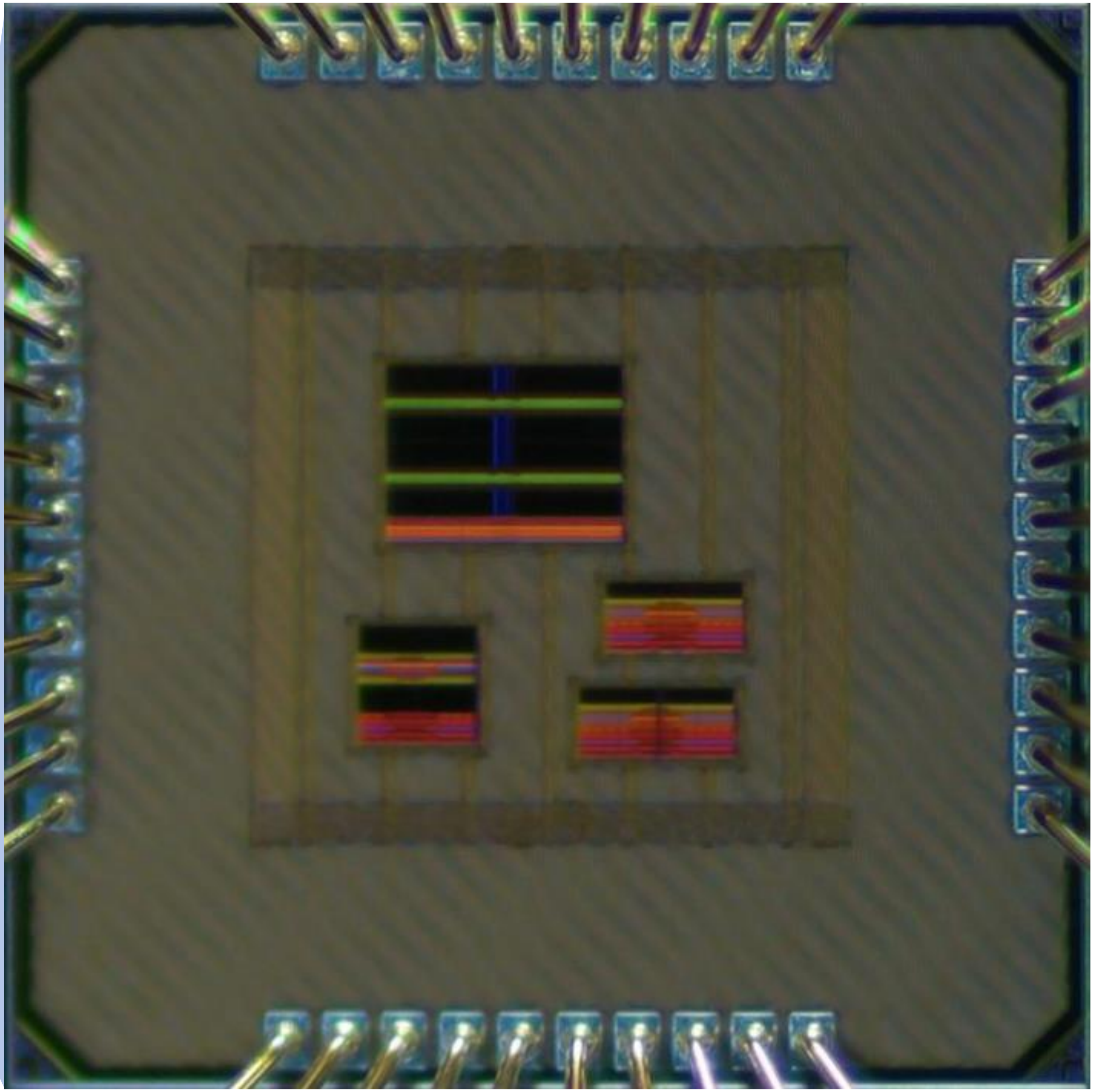
NOEMA01K05S250MS

- TSMC **65nm** GP
- **24µsec** latency
- **1K** neurons (scales to **30K**)
- 5sec experience
- Consumes **0.73mW**
- Equivalent of 600MOPs 32bit-FP

By Comparison:

- **Nvidia Jetson Nano**
  - Consumes **10W**
  - Barely meets **5ms** real-time latency
- **Intel i5-7000**
  - **63ms** latency
  - Fails to meet real-time latency



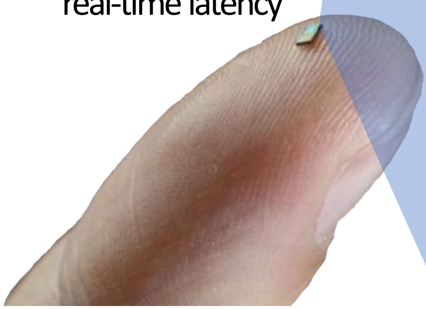


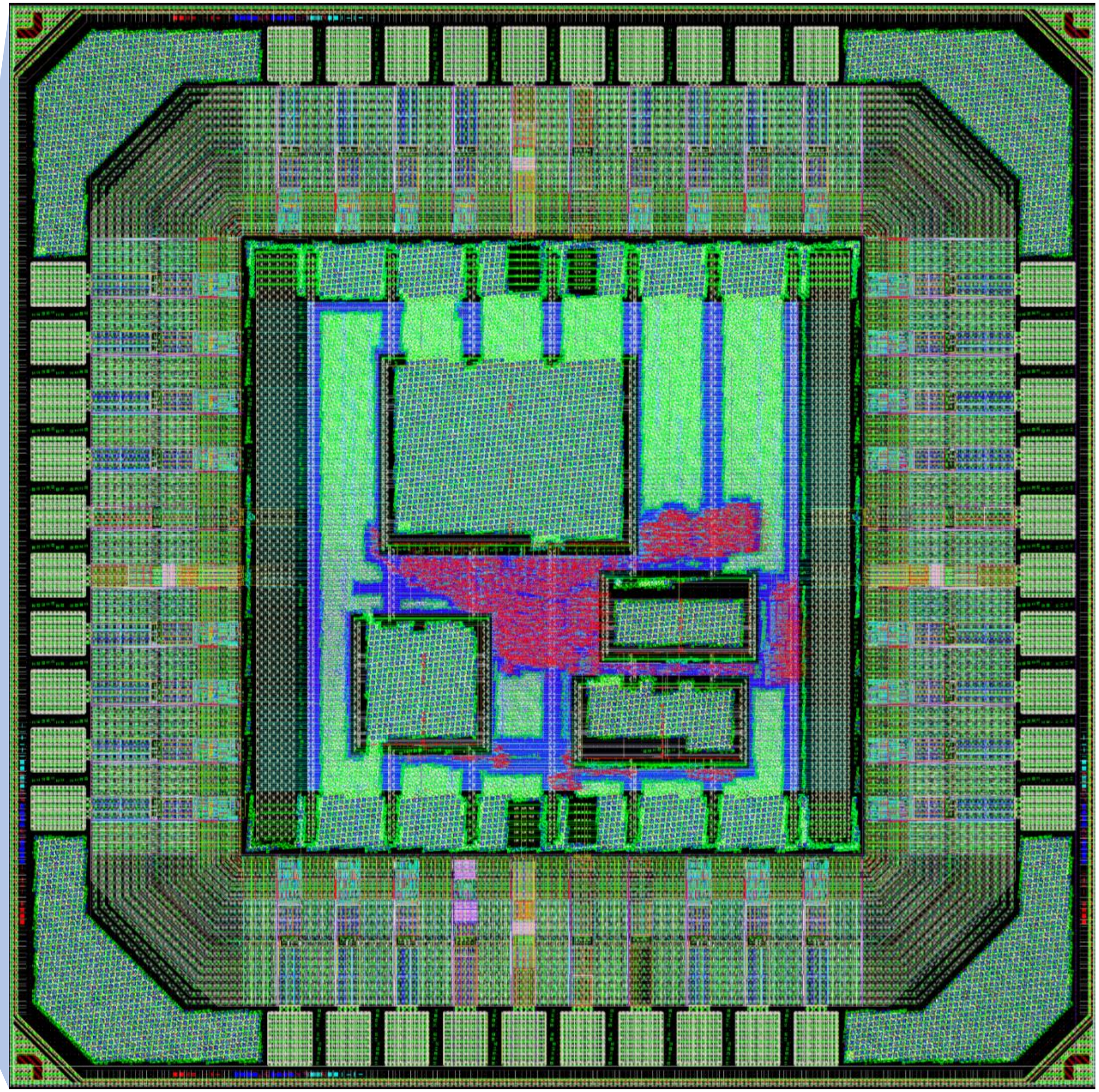
NOEMA01K05S250MS

- TSMC **65nm** GP
- **24µsec** latency
- **1K** neurons (scales to **30K**)
- 5sec experience
- Consumes **0.73mW**
- Equivalent of 600MOPs 32bit-FP

By Comparison:

- **Nvidia Jetson Nano**
  - Consumes **10W**
  - Barely meets **5ms** real-time latency
- **Intel i5-7000**
  - **63ms** latency
  - Fails to meet real-time latency



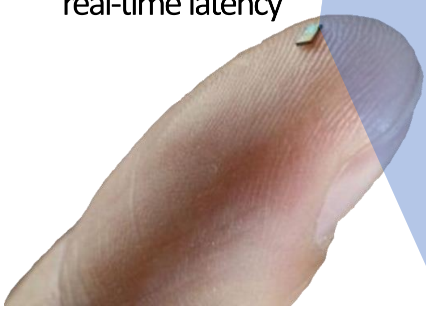


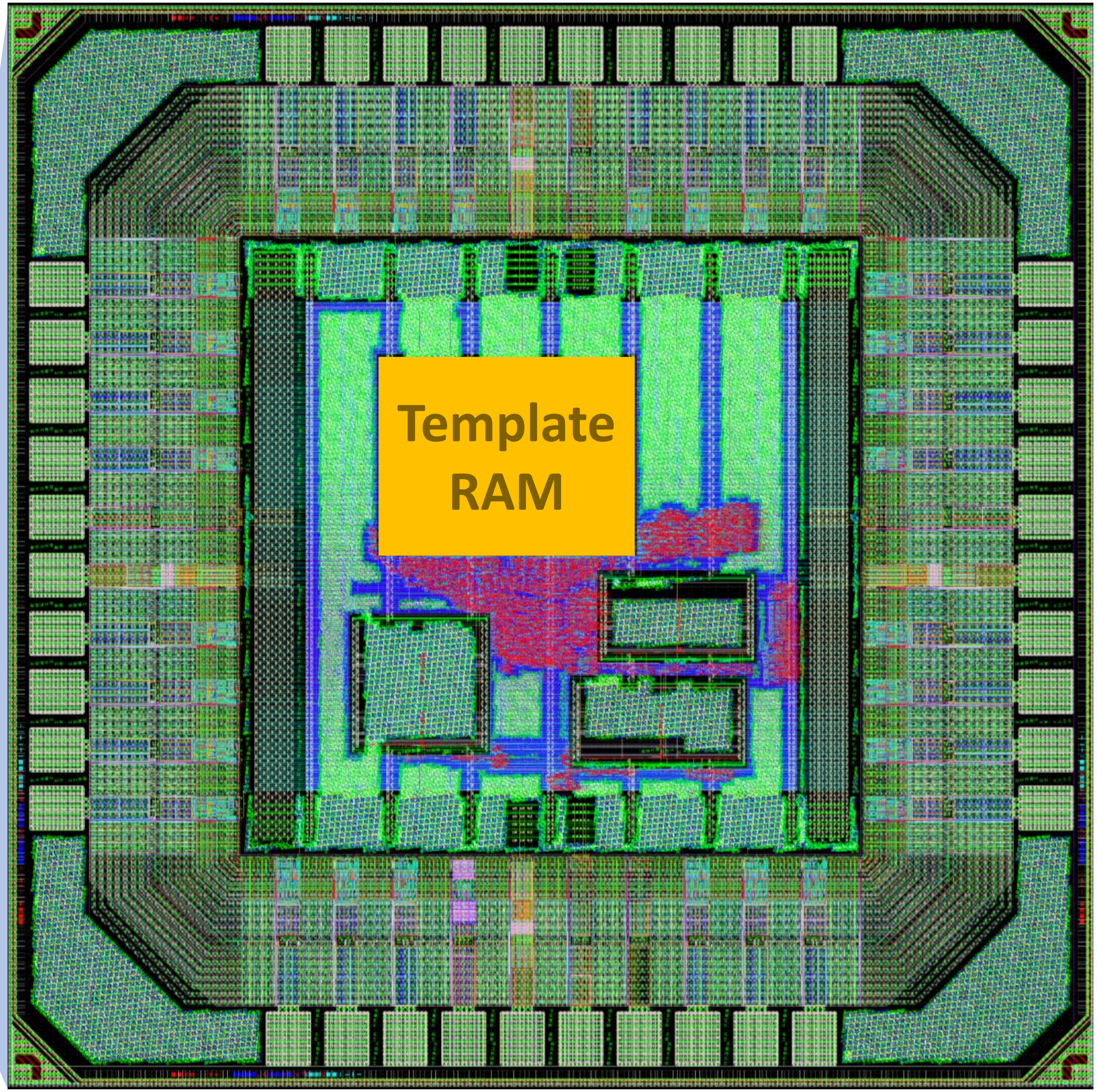
NOEMA01K05S250MS

- TSMC **65nm** GP
- **24µsec** latency
- **1K** neurons (scales to **30K**)
- 5sec experience
- Consumes **0.73mW**
- Equivalent of 600MOPs 32bit-FP

By Comparison:

- **Nvidia Jetson Nano**
  - Consumes **10W**
  - Barely meets **5ms** real-time latency
- **Intel i5-7000**
  - **63ms** latency
  - Fails to meet real-time latency



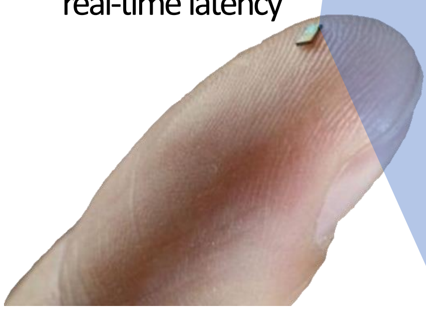


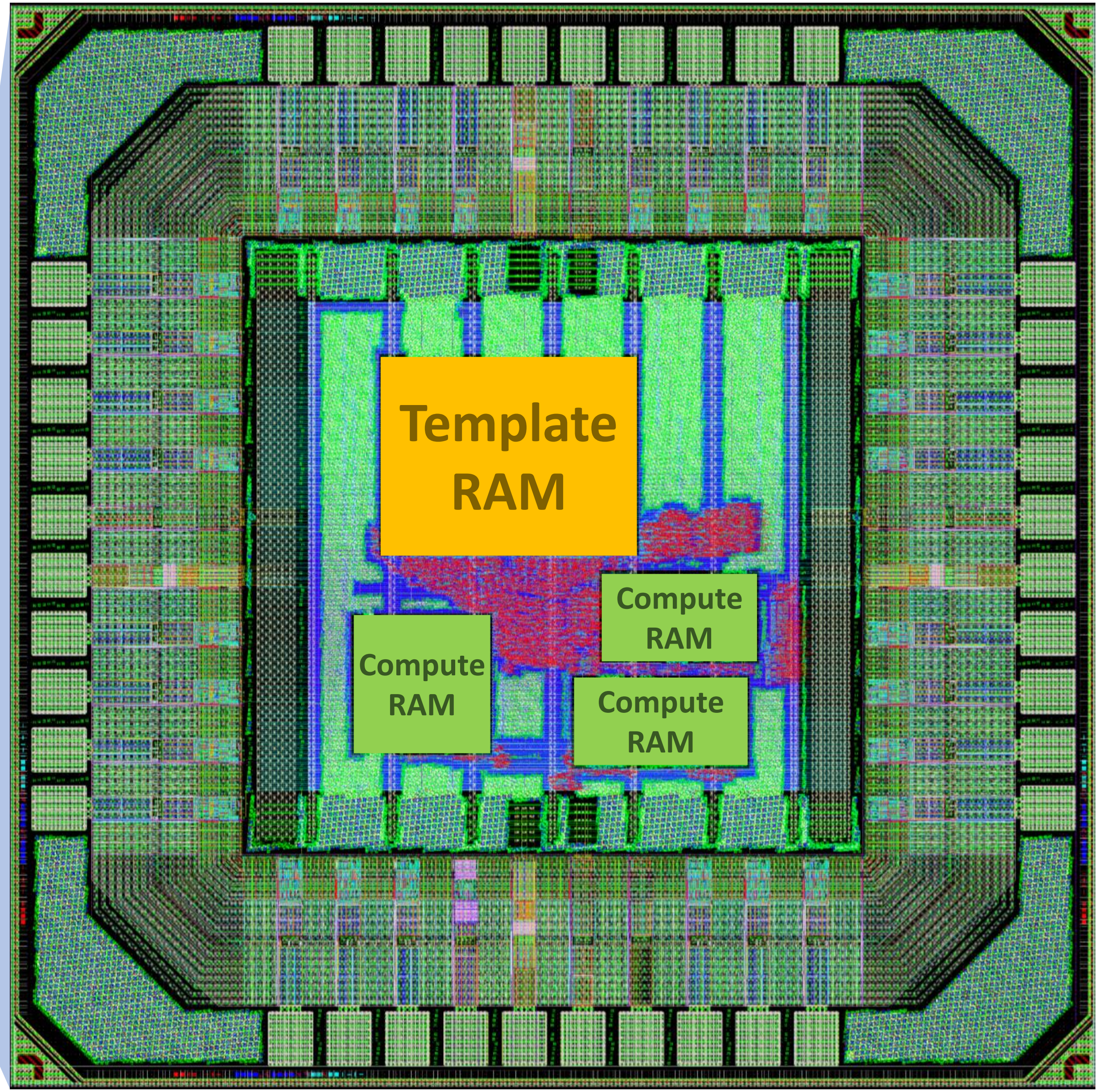
NOEMA01K05S250MS

- TSMC **65nm** GP
- **24µsec** latency
- **1K** neurons (scales to **30K**)
- 5sec experience
- Consumes **0.73mW**
- Equivalent of 600MOPs 32bit-FP

By Comparison:

- **Nvidia Jetson Nano**
  - Consumes **10W**
  - Barely meets **5ms** real-time latency
- **Intel i5-7000**
  - **63ms** latency
  - Fails to meet real-time latency



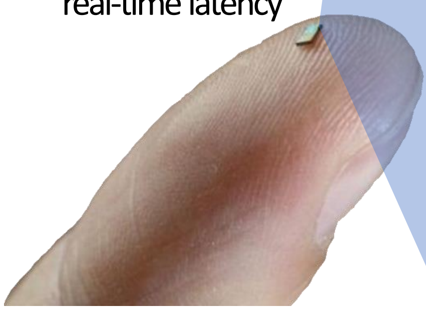


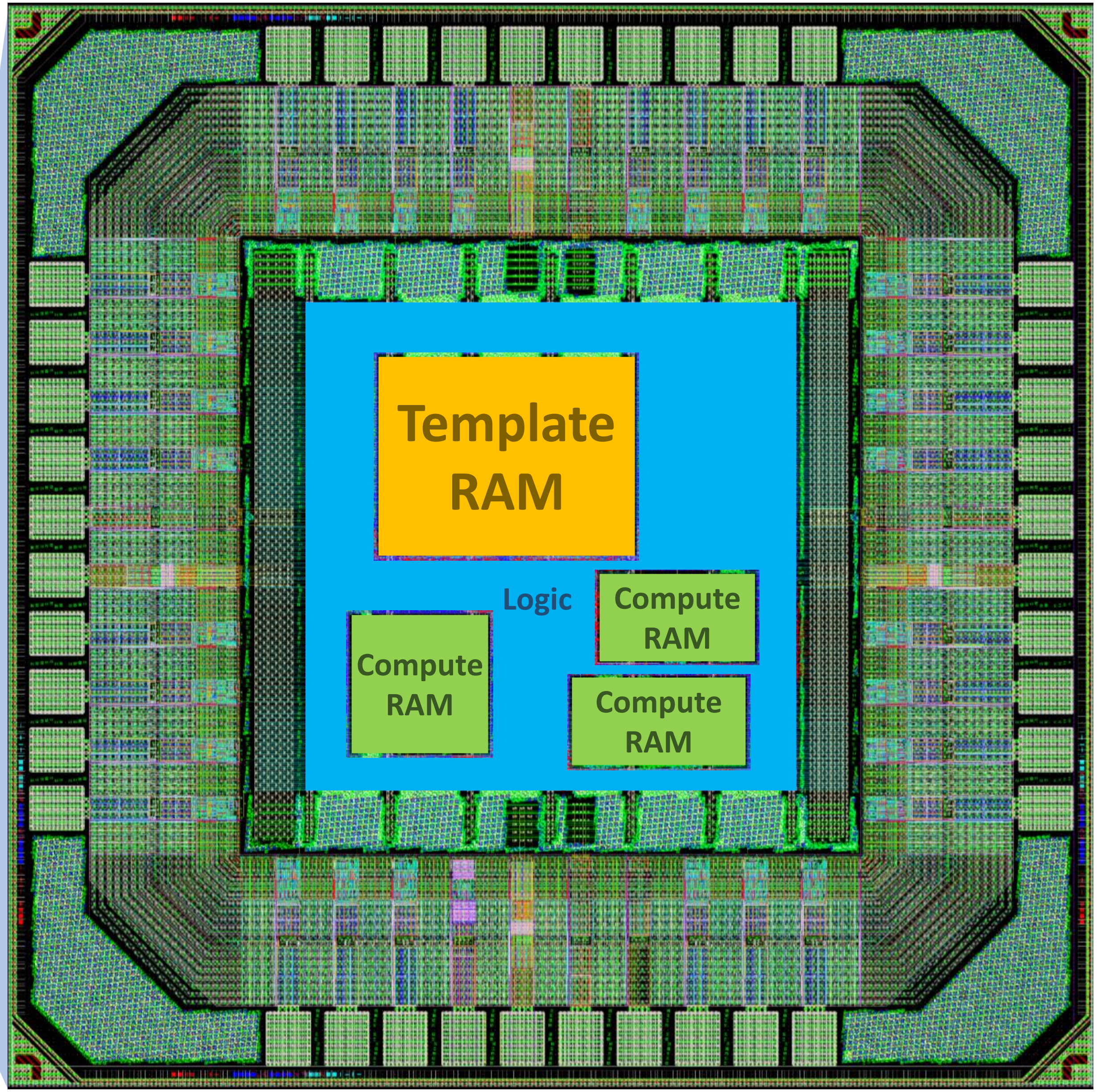
NOEMA01K05S250MS

- TSMC **65nm** GP
- **24μsec** latency
- **1K** neurons (scales to **30K**)
- 5sec experience
- Consumes **0.73mW**
- Equivalent of 600MOPs 32bit-FP

By Comparison:

- **Nvidia Jetson Nano**
  - Consumes **10W**
  - Barely meets **5ms** real-time latency
- **Intel i5-7000**
  - **63ms** latency
  - Fails to meet real-time latency





NOEMA

Key Takeaways

Brain machine interfaces:

- ✗ Exponential growth in data
- ✗ Current solutions are not sufficient

NOEMA's key innovation:

- ✓ Uses simple, low-cost, area- and energy efficient bit-serial and integer arithmetic units
- ✓ Enables computations to proceed progressively as data is received
- ✓ Scales to meet ***future*** demand
  - 14x less power, 2.6x smaller, order of µsec latency

Thank you!