Department of Electrical
and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

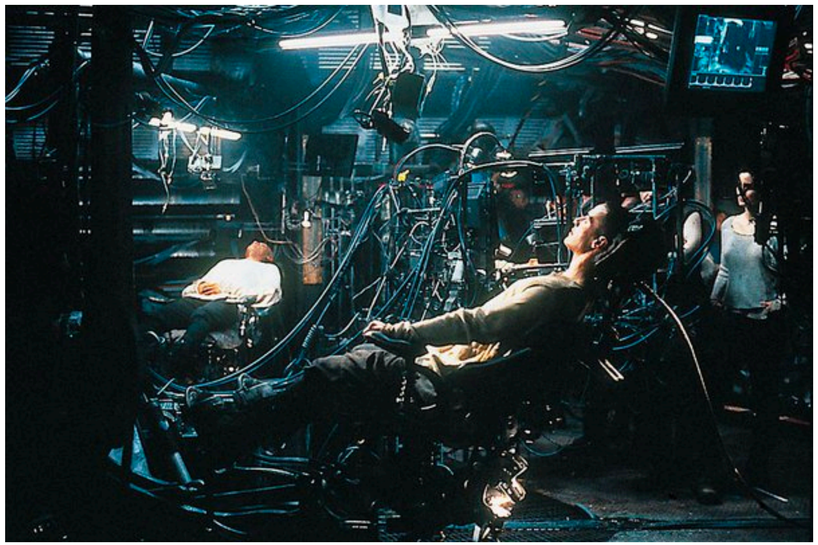
uBrain: A Unary Brain Computer Interface

Di Wu, Jingjie Li, Zhewen Pan, Younghyun Kim, Joshua San Miguel

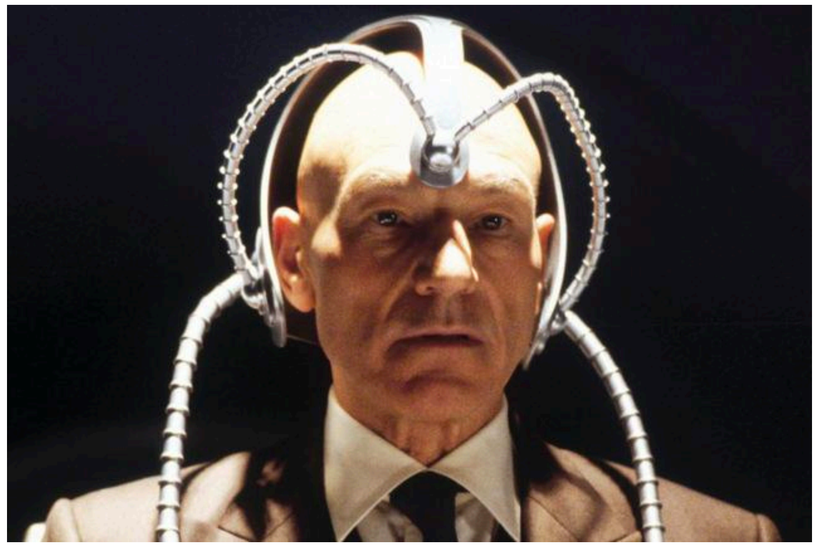
The 49th International Symposium on Computer Architecture
ISCA 2022, New York, USA

Brain computer interface

“The Matrix”, 1999



“X-Men”, 2000



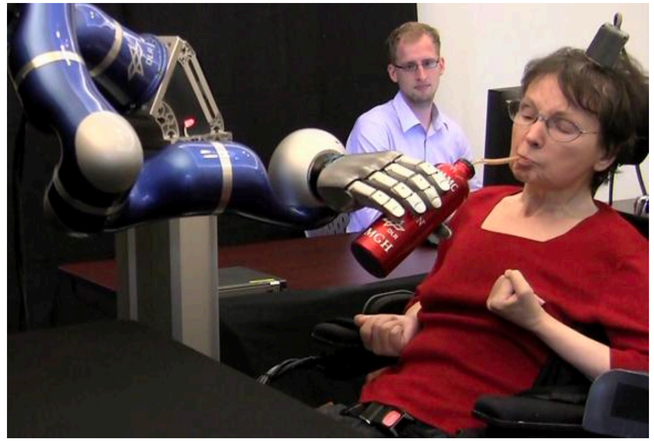
“Iron-Man”, 2008

Outline

- **Background**
- Algorithm
- Architecture
- Evaluation
- Conclusion

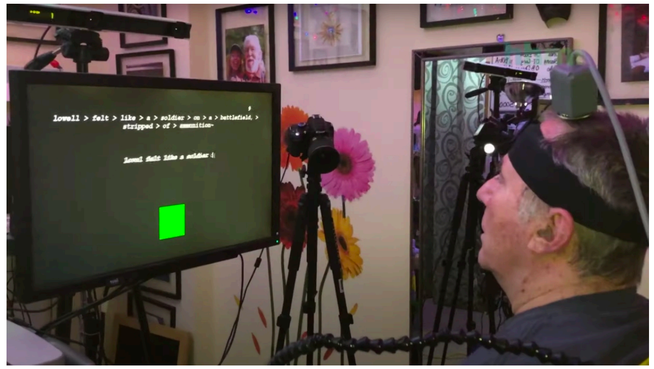
Brain computer interface

Prosthetic control



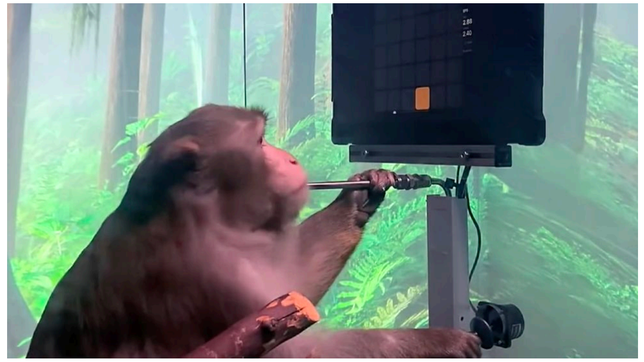
Brown

Text imagery



Stanford

Video gaming



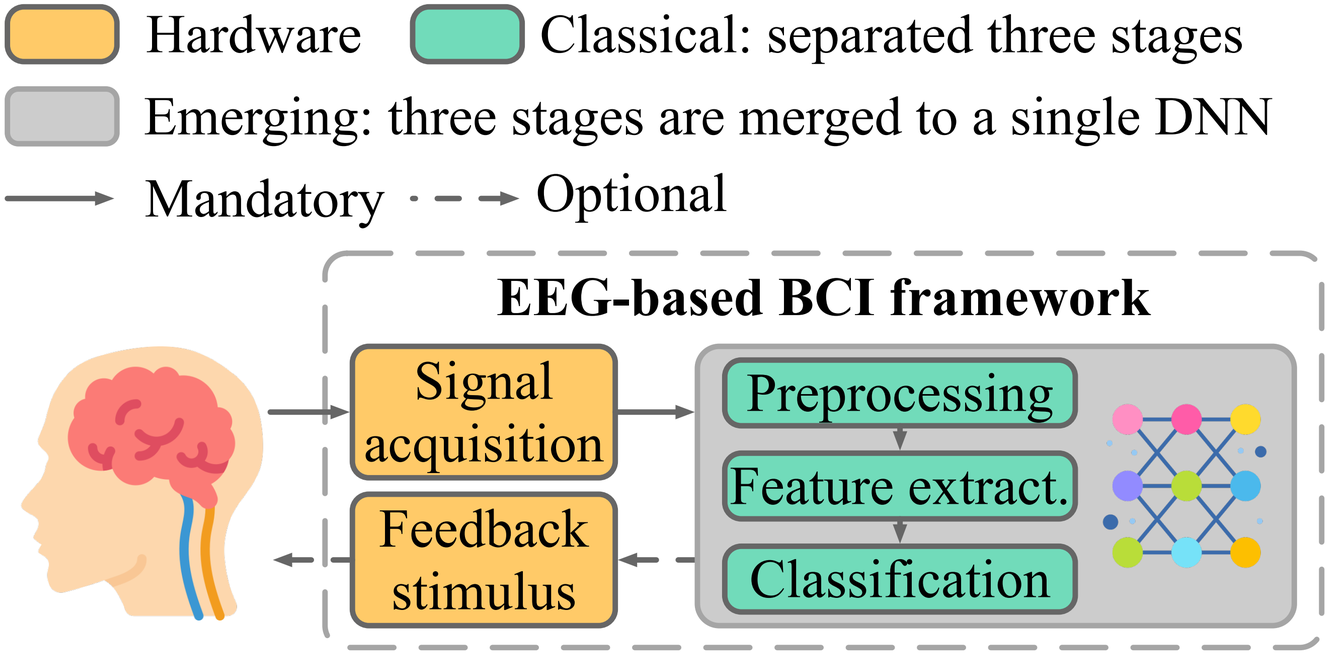
Neuralink

Brown, https://news.brown.edu/articles/2012/05/braingate2
Stanford, https://www.hhmi.org/news/brain-computer-interface-turns-mental-handwriting-into-text-on-screen
Neuralink, https://www.bbc.com/news/technology-56688812

Brain computer interface

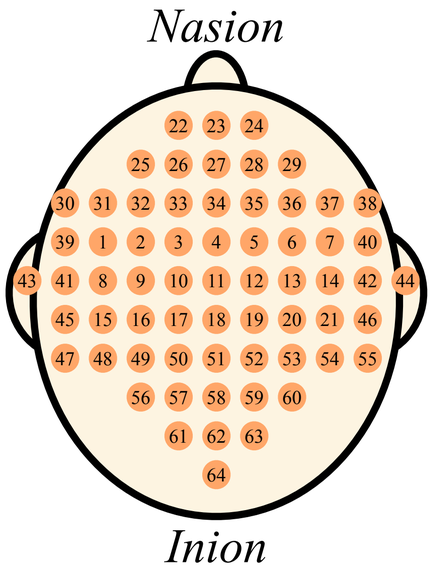
Electroencephalogram (EEG): collected at the scalp w/o surgery

https://www.medgadget.com/2017/02/non-invasive-brain-computer-interface-completely-locked-patients-interview-dr-ujwal-chaudhary.html



Brain computer interface

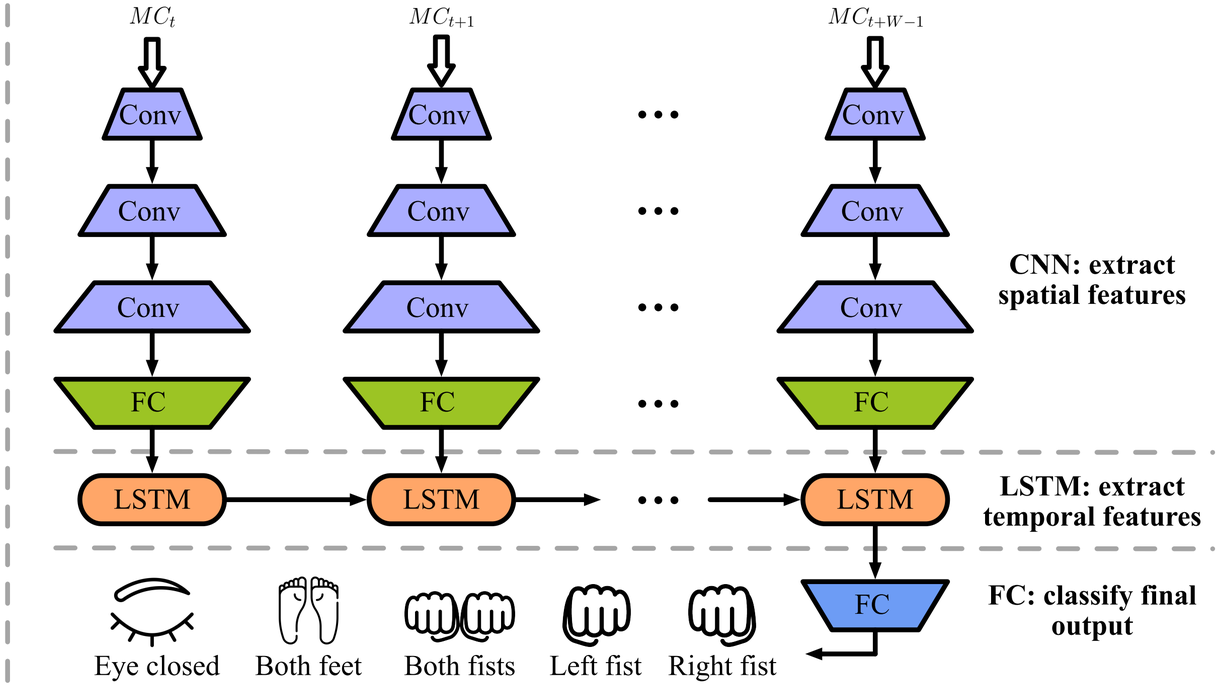
- Cascaded DNN model



Electrode placement

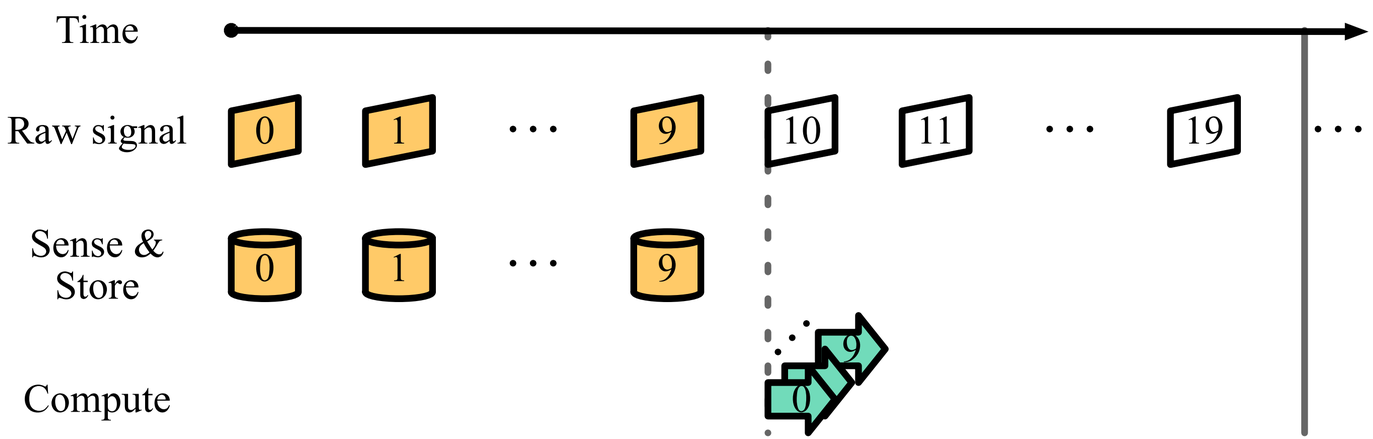
$$MC_t = \begin{bmatrix} 0 & 0 & 0 & 0 & s_t^{22} & s_t^{23} & s_t^{24} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & s_t^{25} & s_t^{26} & s_t^{27} & s_t^{28} & s_t^{29} & 0 & 0 & 0 \\ 0 & s_t^{30} & s_t^{31} & s_t^{32} & s_t^{33} & s_t^{34} & s_t^{35} & s_t^{36} & s_t^{37} & s_t^{28} & 0 \\ 0 & s_t^{39} & s_t^{1} & s_t^{2} & s_t^{3} & s_t^{4} & s_t^{5} & s_t^{6} & s_t^{7} & s_t^{40} & 0 \\ s_t^{43} & s_t^{41} & s_t^{8} & s_t^{9} & s_t^{10} & s_t^{11} & s_t^{12} & s_t^{13} & s_t^{14} & s_t^{42} & s_t^{44} \\ 0 & s_t^{45} & s_t^{15} & s_t^{16} & s_t^{17} & s_t^{18} & s_t^{19} & s_t^{20} & s_t^{21} & s_t^{46} & 0 \\ 0 & s_t^{47} & s_t^{48} & s_t^{49} & s_t^{50} & s_t^{51} & s_t^{52} & s_t^{53} & s_t^{54} & s_t^{55} & 0 \\ 0 & 0 & 0 & s_t^{56} & s_t^{57} & s_t^{58} & s_t^{59} & s_t^{60} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & s_t^{61} & s_t^{62} & s_t^{63} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & s_t^{64} & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Positional mapping

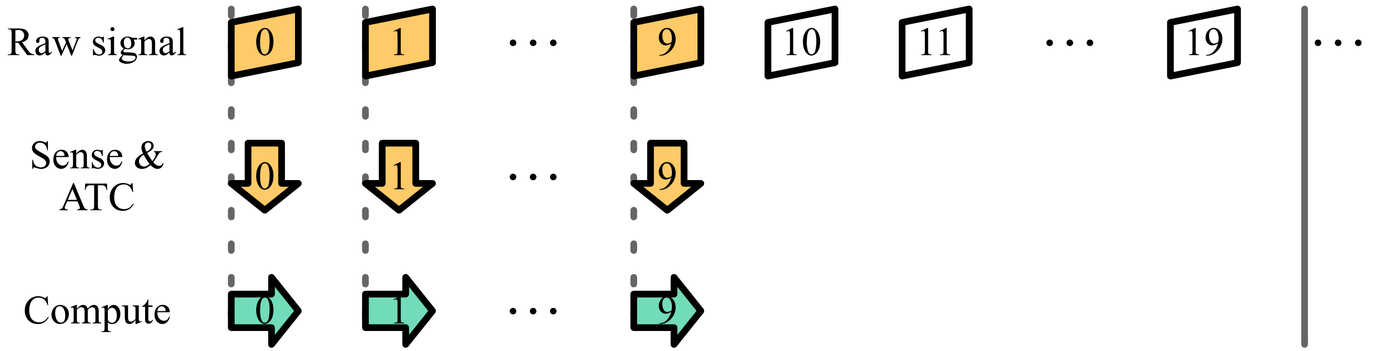


Brain computer interface

- Dataflow
  - Classical



  - Immediate processing (immediate signal processing after sensing)
    - Longer runtime
    - Lower frequency
    - Lower power



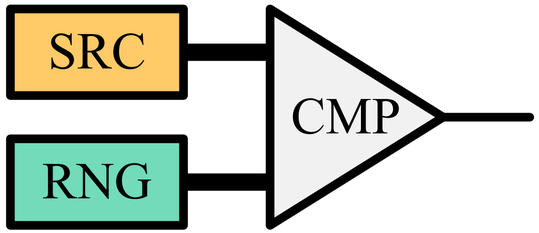
Unary computing

- Data: serial bitstreams

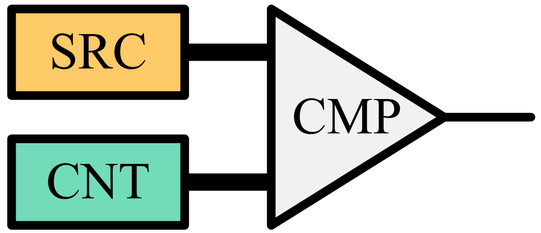
V=0.5 (8/16)

Rate coding	A	1100101010100110
Temporal coding	B	0000000011111111

- Bitstream generation



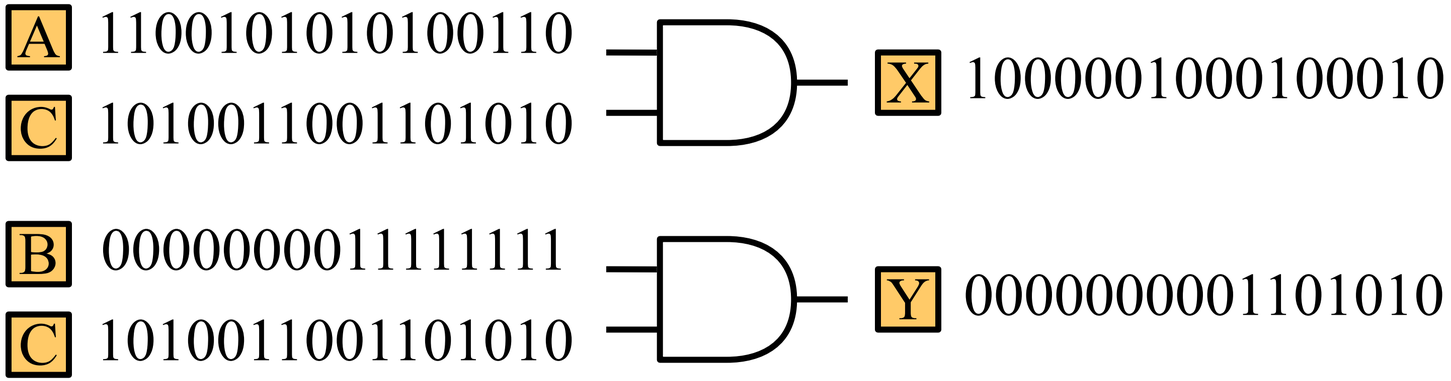
Rate coding



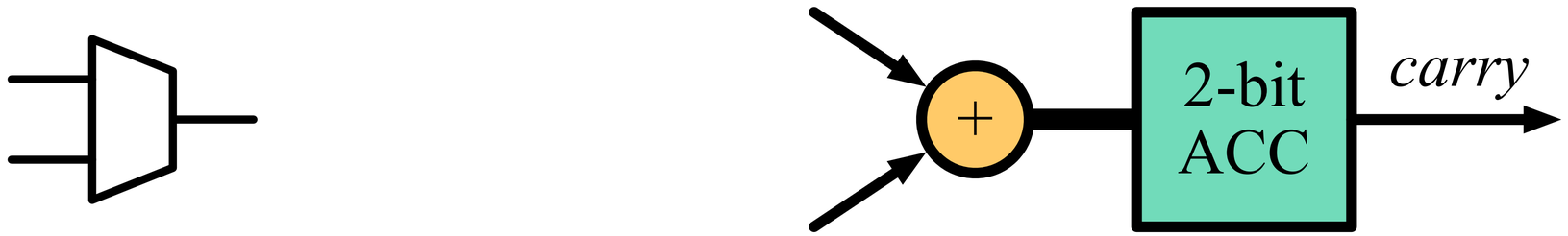
Temporal coding

Unary computing

- Multiplier



- Adder



uBrain overview

- A BCI compute engine with algorithm-hardware co-design
  - **minimum accuracy loss** via customized DNN with piece-wise linear activations
  - **high power efficiency** via immediate processing (immediate signal processing after sensing)
    - Inter-layer hardware time-division multiplexing
    - Analog-to-Temporal Conversion (ATC)
    - Low-cost temporal multiplier

uBrain overview

- Comparison from the algorithm perspective

Platform	Accuracy	Compatible operations	Power efficiency
CPU	67~98%	All	Low
HALO [1]	67~85%	SVM, FFT, etc.	Medium
SC-SVM [2]	67~75%	SVM	High
uBrain (ours)	91~95%	CNN, RNN	High

[1] Ioannis Karageorgos et al., Hardware-Software Co-Design for Brain-Computer Interfaces. ISCA 2020.
[2] Kaining Han et al., A Low Complexity SVM Classifier for EEG Based Gesture Recognition Using Stochastic Computing. ISCAS 2020.

Outline

- Background
- **Algorithm**
- Architecture
- Evaluation
- Conclusion

Target task

Task	Output category
Motor imagery	Eye closed Both feet Both fists Left fist Right fist
Seizure prediction	On-set Not on-set

Customized DNN

- Minimized model size

Layer		Shape		Activation
Type	Name	Input	Output	
CNN	Conv1	(10, 1, 10, 11)	(10, 16, 10, 11)	*ReLU*
	Conv2	(10, 16, 10, 11)	(10, 32, 10, 11)	*ReLU*
	FC3	(10, 32*10*11)	(10, 256)	*ReLU*
RNN	MGU4	(10, 256)	(10, 64)	*Sigmoid, Tanh*
Head	FC5	64	5	*Tanh*
	FC6	64	2	*Tanh*

Customized DNN

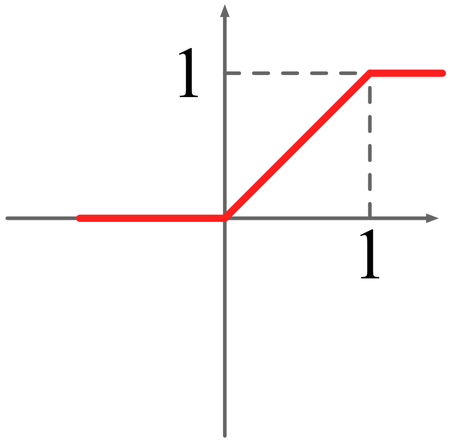
- Simplified activation for unary computing

Layer		Shape		Activation
Type	Name	Input	Output	
CNN	Conv1	(10, 1, 10, 11)	(10, 16, 10, 11)	*HardReLU*
	Conv2	(10, 16, 10, 11)	(10, 32, 10, 11)	*HardReLU*
	FC3	(10, 32*10*11)	(10, 256)	*HardReLU*
RNN	MGU4	(10, 256)	(10, 64)	*HardSigmoid, HardTanh*
Head	FC5	64	5	*HardTanh*
	FC6	64	2	*HardTanh*

Piece-wise linear activation

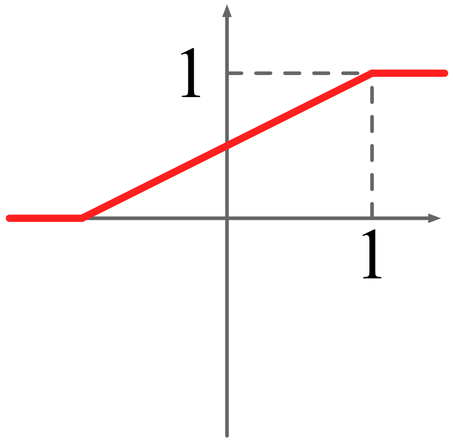
- Minimum hardware requirement

HardReLU



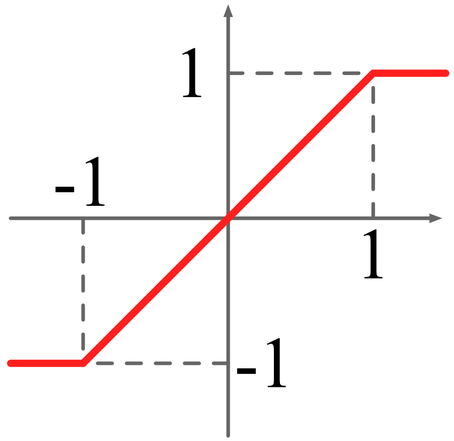
$\max(0, \min(1, x))$

HardSigmoid



$\max(0, \min(1, (x+1)/2))$

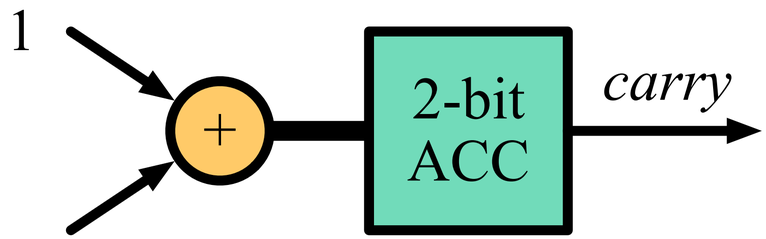
HardTanh



$\max(-1, \min(1, x))$

Piece-wise linear *HardSigmoid*

- Minimum hardware requirement



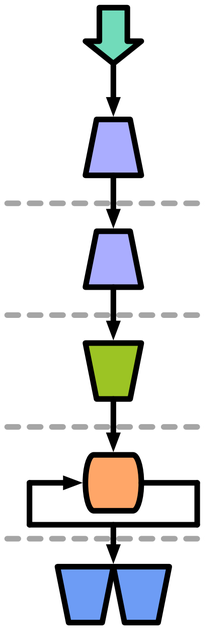
Outline

- Background
- Algorithm
- **Architecture**
- Evaluation
- Conclusion

Architecture

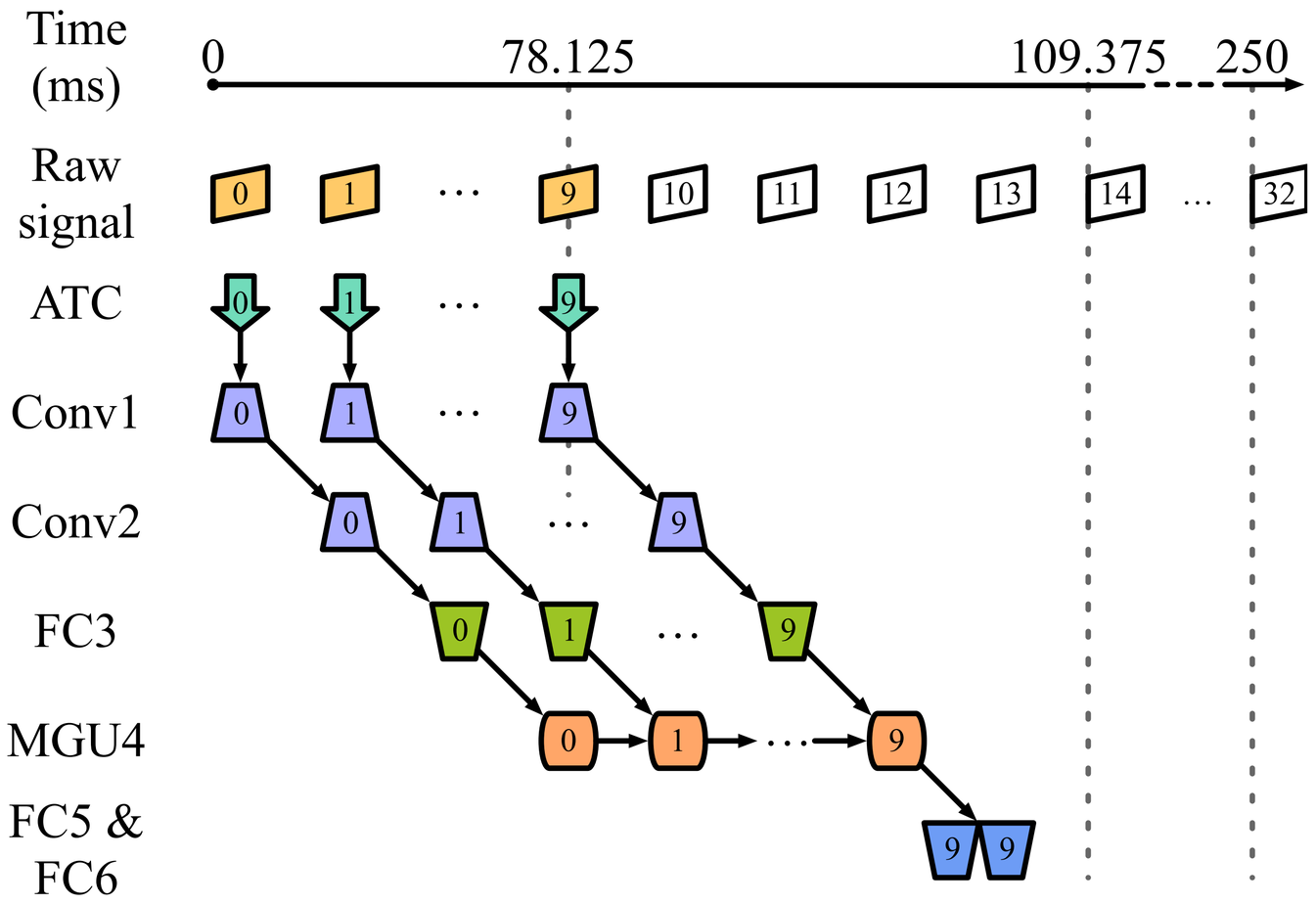
- Immediate processing (immediate signal processing after sensing)

Block	Input coding	Weight coding
ATC	Temporal	
Conv1	Temporal	Rate
Conv2	Rate	Temporal
FC3	Rate	Temporal
MGU4	Rate	Rate
FC5 & FC6	Rate	Temporal



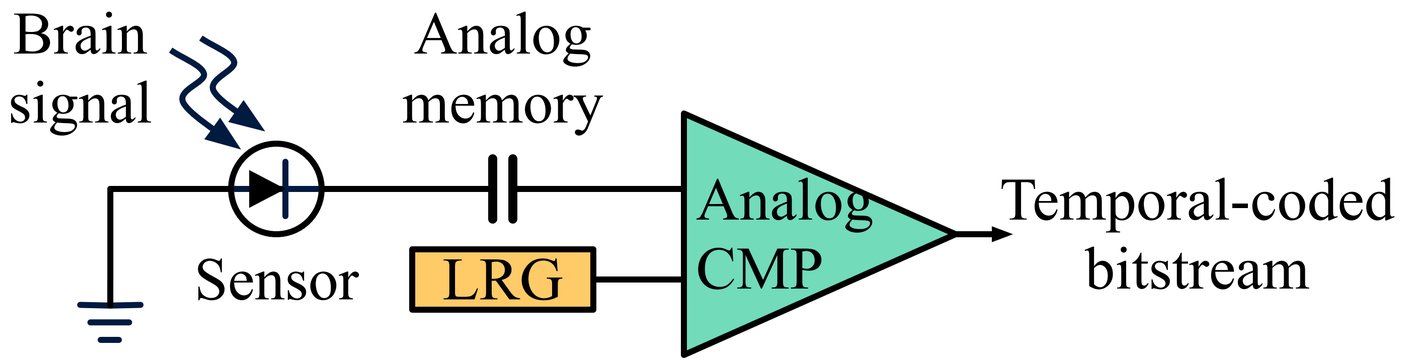
Sequential dataflow

- Inter-layer hardware time-division multiplexing



Micro-architecture optimization

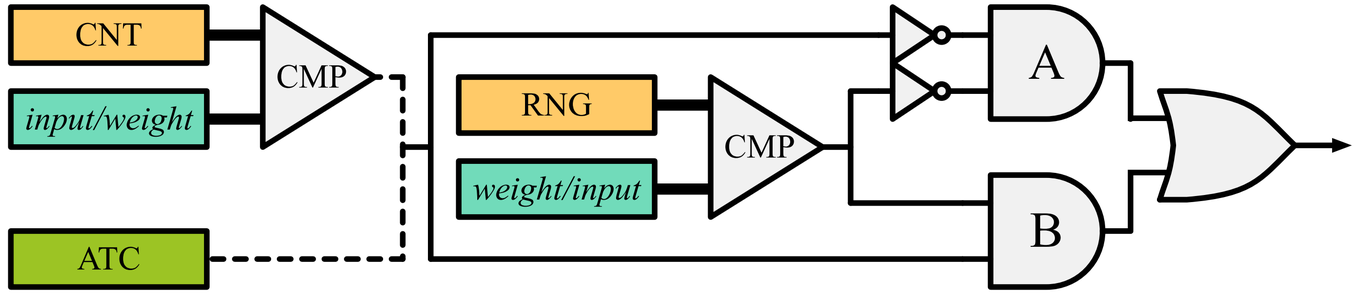
- Analog-to-Temporal Conversion (ATC)
  - Buffer a raw analog signal in an analog memory
  - Compare the buffered signal with a linear ramp signal
  - Output a temporal bitstream



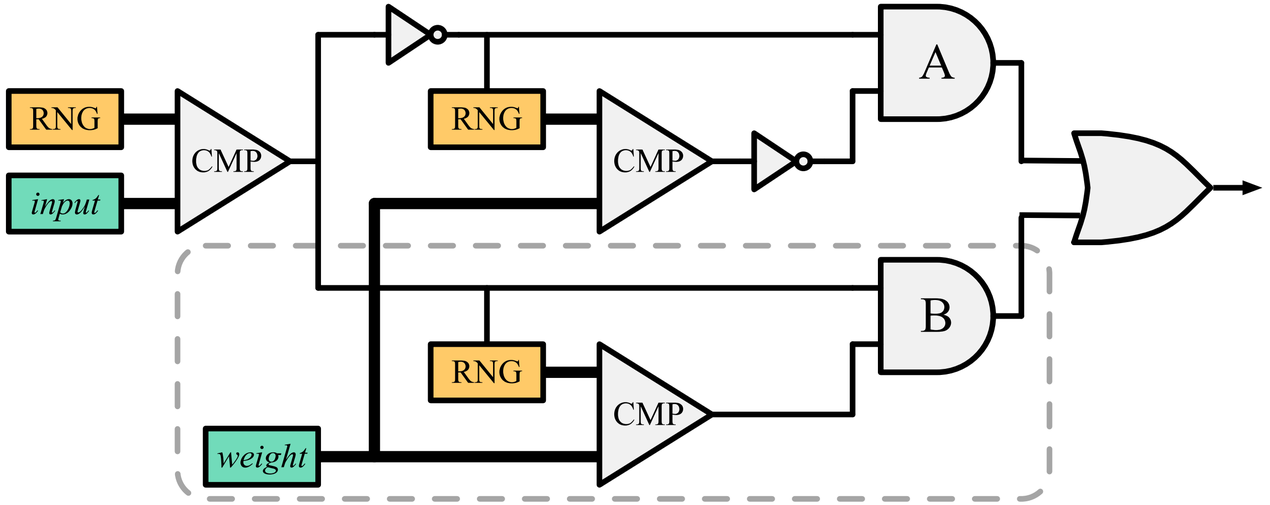
Micro-architecture optimization

- Temporal multiplier

Proposed temporal-coded multiplier for area savings



Existing SC rate-coded multiplier [1]



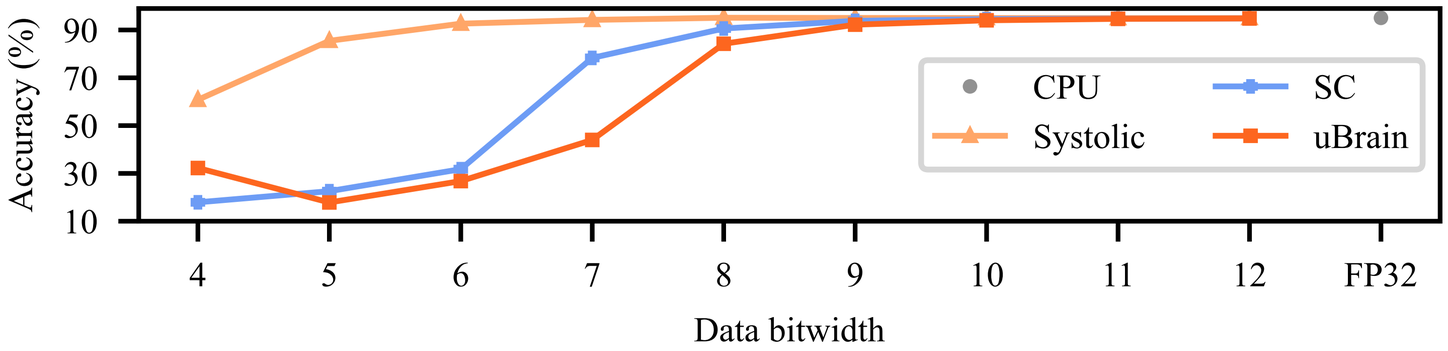
[1] Di Wu et al., uGEMM: Unary Computing Architecture for GEMM Applications. ISCA 2020.

Outline

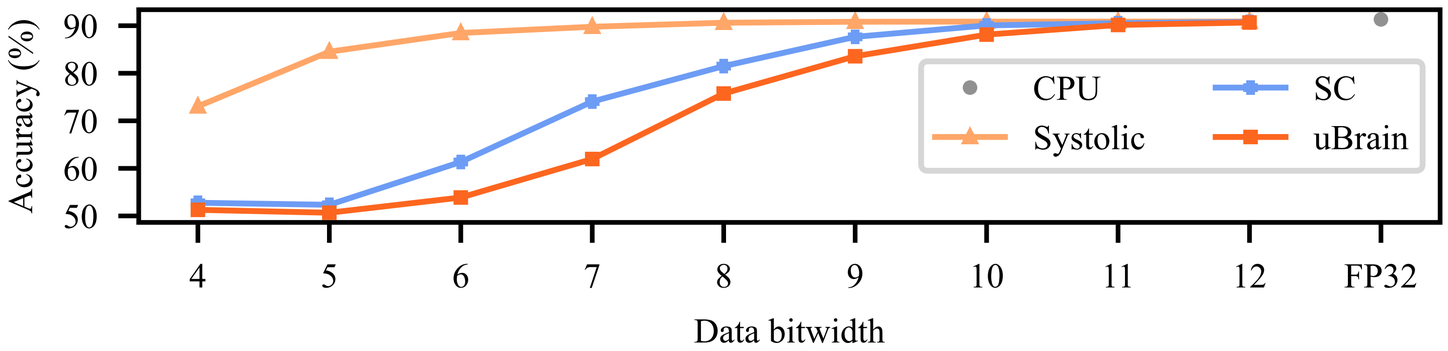
- Background
- Algorithm
- Architecture
- **Evaluation**
- Conclusion

Accuracy

- Customized DNN
  - Motor imagery



  - Seizure prediction

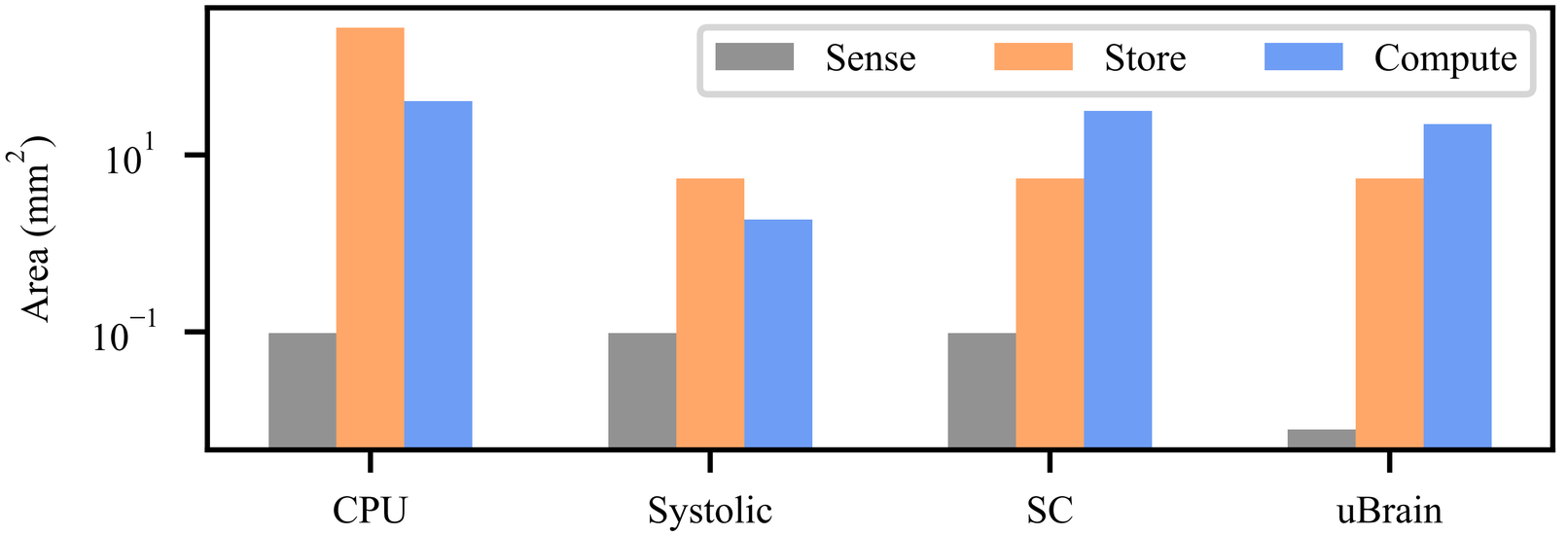


Evaluated hardware

Design	Compute	Memory	Frequency
CPU	ARM A57 (FP32)	4GB LPDDR4	Maximum 1400MHz
Systolic array	12-by-14 PEs (8-bit)	16MB DDR3	Maximum 400MHz
Stochastic computing	SC multiplier (10-bit)		~4MHz
uBrain	Temporal multiplier (10-bit)		

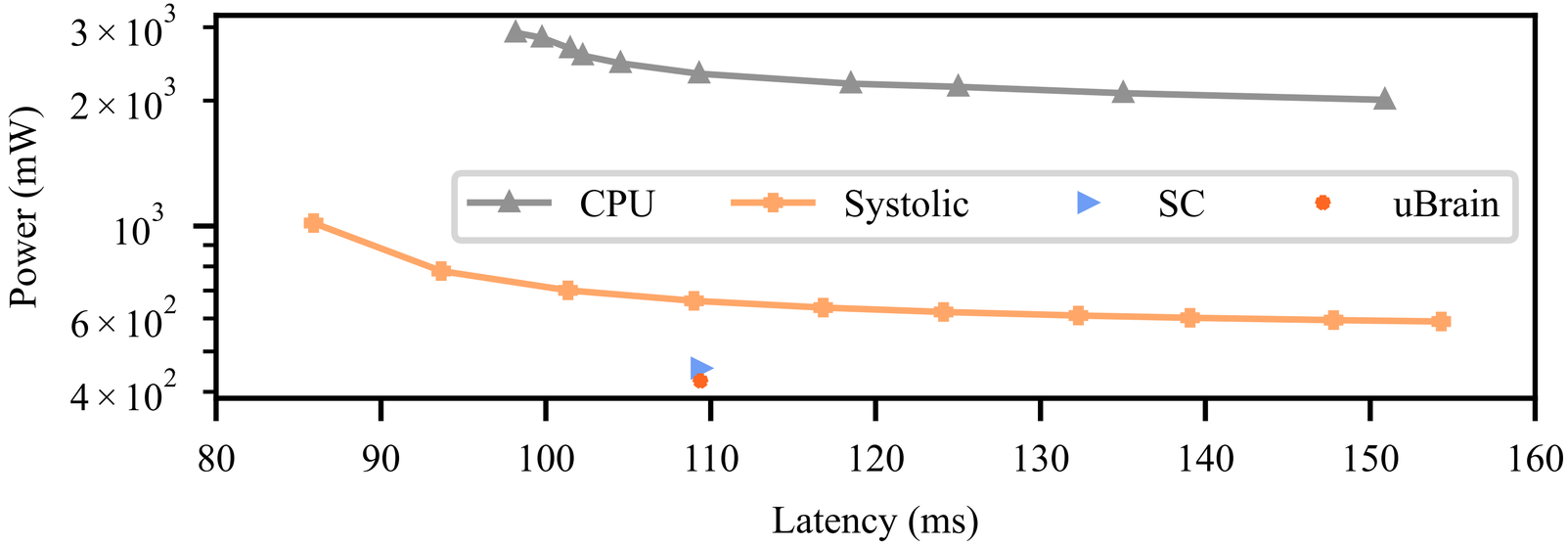
Area

- Total area



Power

- Total power
  - uBrain is 5.5X, 1.6X and 1.1X better than CPU, systolic array and SC BCIs.



Outline

- Background
- Algorithm
- Architecture
- Evaluation
- **Conclusion**

Conclusion

- uBrain introduces
  - **minimum accuracy loss** via customized DNN
  - **high power efficiency** via immediate processing
    - Inter-layer hardware time-division multiplexing
    - Analog-to-Temporal Conversion (ATC)
    - Low-cost temporal multiplier

Department of Electrical
and Computer Engineering
UNIVERSITY OF WISCONSIN–MADISON

THANK YOU
Q & A

Di Wu, Jingjie Li, Zhewen Pan, Younghyun Kim, Joshua San Miguel

The 49th International Symposium on Computer Architecture
ISCA 2022, New York, USA