2nd EU-Japan Digital Week: Semiconductor Workshop: “Japan-EU Cooperation on Advanced Computing, Advanced Functionalities and Semiconductor Value Chain”, Haseko Kuma Hall, U. Tokyo, March 24 (2026)

Physical Reservoir Computing Utilizing HfO_2-based Ferroelectric Devices for Edge-AI Applications

Shinichi Takagi
Teikyo University, Tokyo, Japan

Kasidit Toprasertpong, Eishin Nako, Shin-Yi Min, Rikuo Suzuki
Mitsuru Takenaka, Ryosho Nakane
The University of Tokyo, Tokyo, Japan

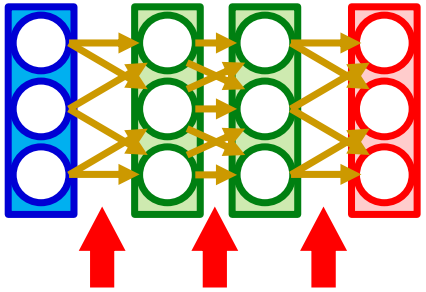
TEIKYO 帝京大学 Teikyo University

UTokyo

What is reservoir computing?

Deep Neural Network

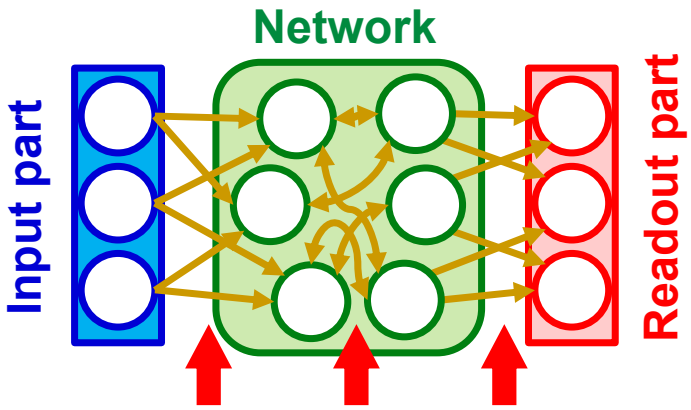
Suitable for
static data processing



Train all weights
(High learning cost)

Recurrent Neural Network

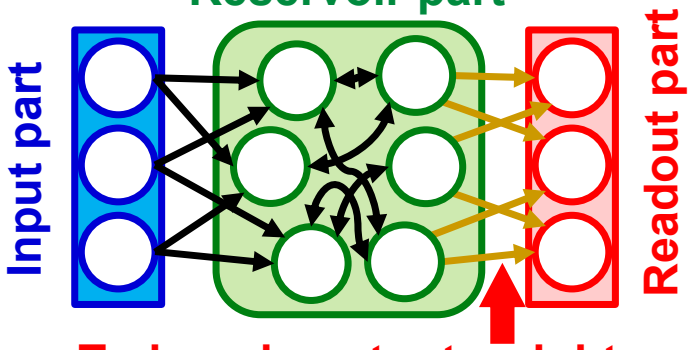
Suitable for **time series data processing**



Train all weights

(High learning cost)

Reservoir computing



Train only output weights

(Low learning cost)

- RNN for time series data processing
- Reservoir computing is RNN that trains only output weights

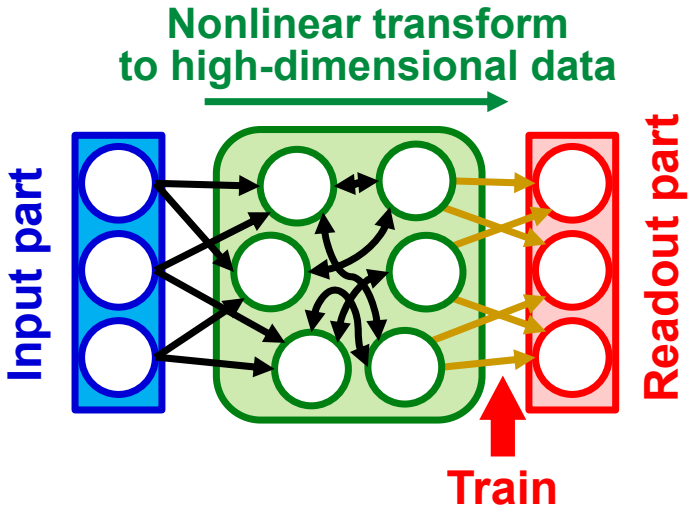
⇒ Training with high speed and low energy consumption is obtained

Physical reservoir computing

Reservoir computing

Eco state network

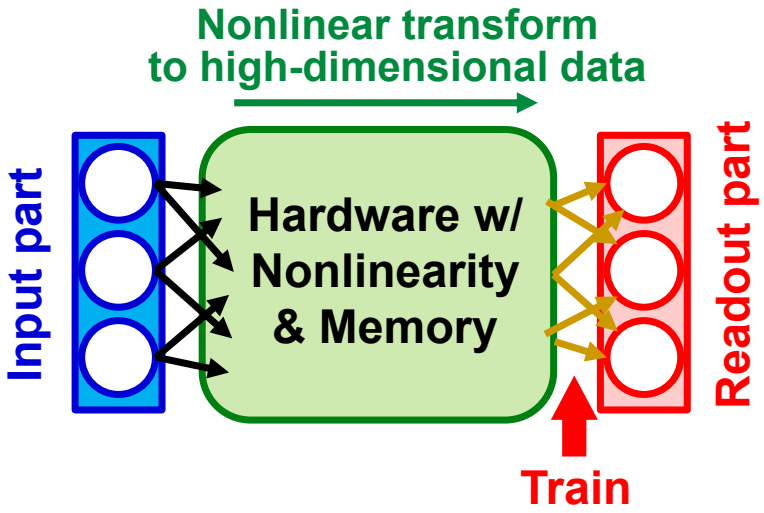
Reservoir part = **software NN**



W. Maass *et al.*, *Neural Comp.* **14**, 2531 (2002)
H. Jaeger, *GMD Report* **148** (2001)

Physical Reservoir computing

Reservoir part = **nonlinear hardware**

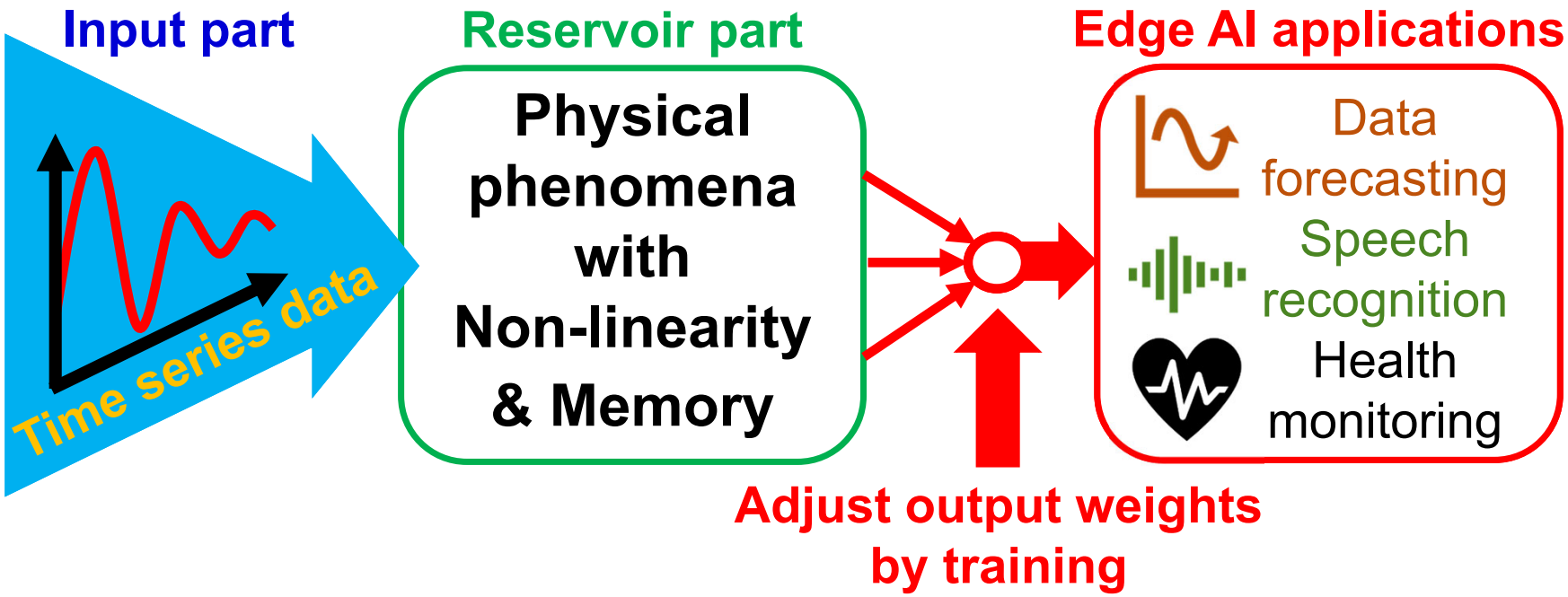


G. Tanaka et al., Neural Networks 115, 100 (2019)

- Reservoir part can be realized by hardware with nonlinear dynamics
- Physical reservoir computing can reduce computational and hardware cost

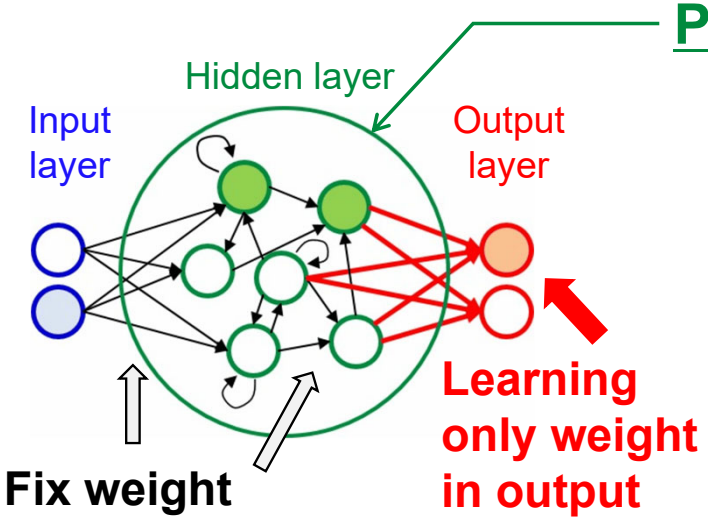
Application of physical reservoir computing

G. Tanaka et al., Neural Networks 115, 100 (2019)



- Physical reservoir computing is expected to offer efficient online learning for edge AI applications to time-series data processing with high energy efficiency
- Potentially important edge AI applications include data forecasting, speech recognition, and health monitoring

Requirements for physical reservoir

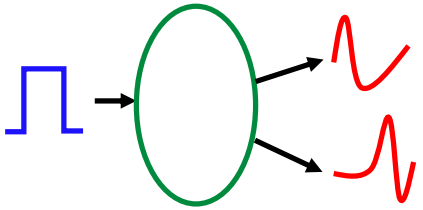


Properties needed for reservoirs

Short term memory ⇒ Next state is dependent on present state and input

Non-linearity ⇒ Non-linear mapping from input to output data

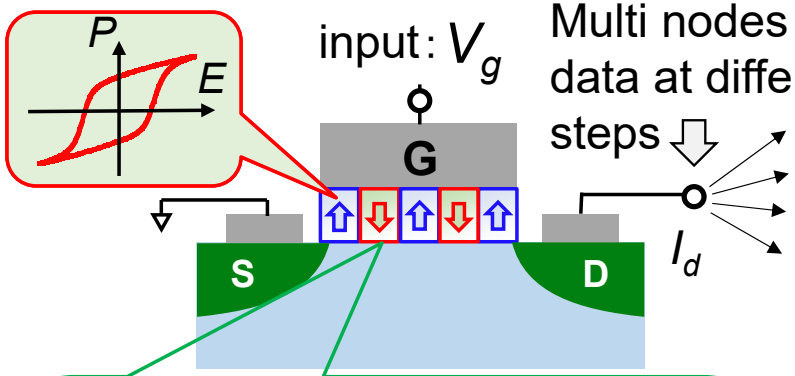
High dimensionality of mapped data



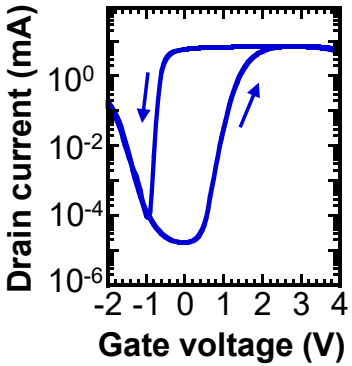
Mapping into higher dimensional space

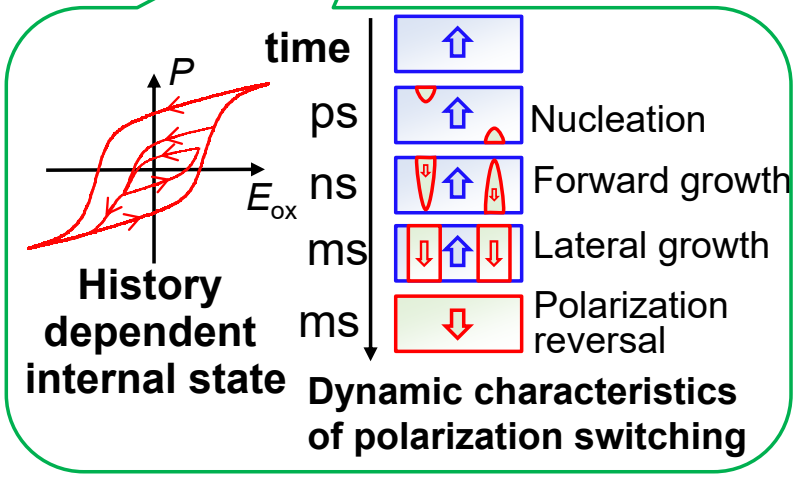
- Only weights of output layer are trained
  ⇒ Training with high speed and low energy consumption
- Reservoir can be implemented by a physical system with short term memory and non-linear and high-dimensional-mapping functions
  ⇒ further reduces computational and hardware energy cost
  ⇒ suitable for processing time-series data in edge computing
- CMOS-friendly reservoir is preferable for total system integration

Expectation for FeFET reservoir computing



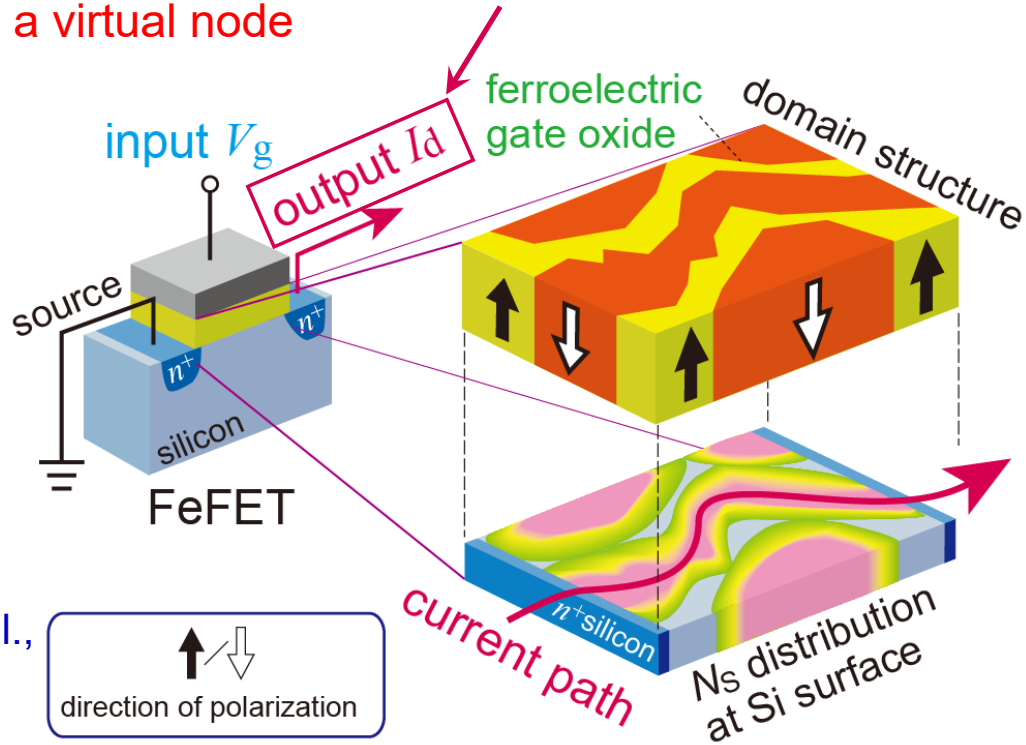
Multi nodes by using data at different time steps





Nako et al., VLSI Symp. (2020) TN1.6;
Toprasertpong et al., Comm. Eng. 1 (2022) 21

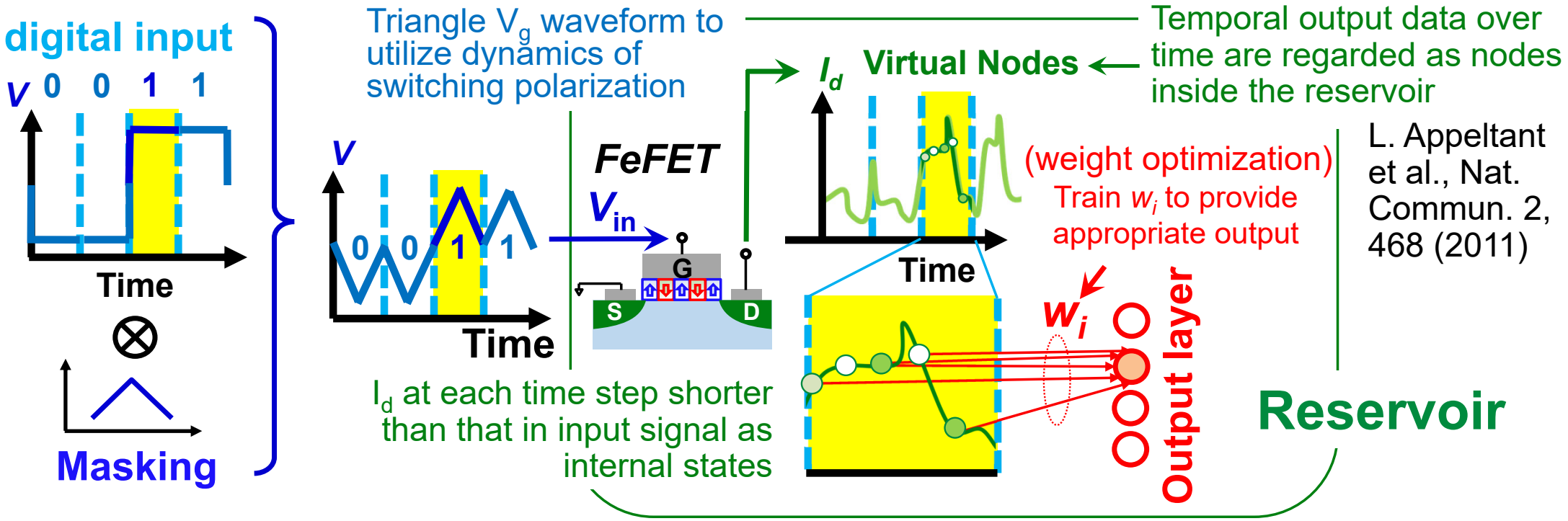
Time response of FeFET currents is utilized as a virtual node



- Memory function due to polarization and rich non-linearity due to complex time responses of polarization domains, the complex spatial distributions of domains, and the interaction between domains

Operation of reservoir computing using FeFET

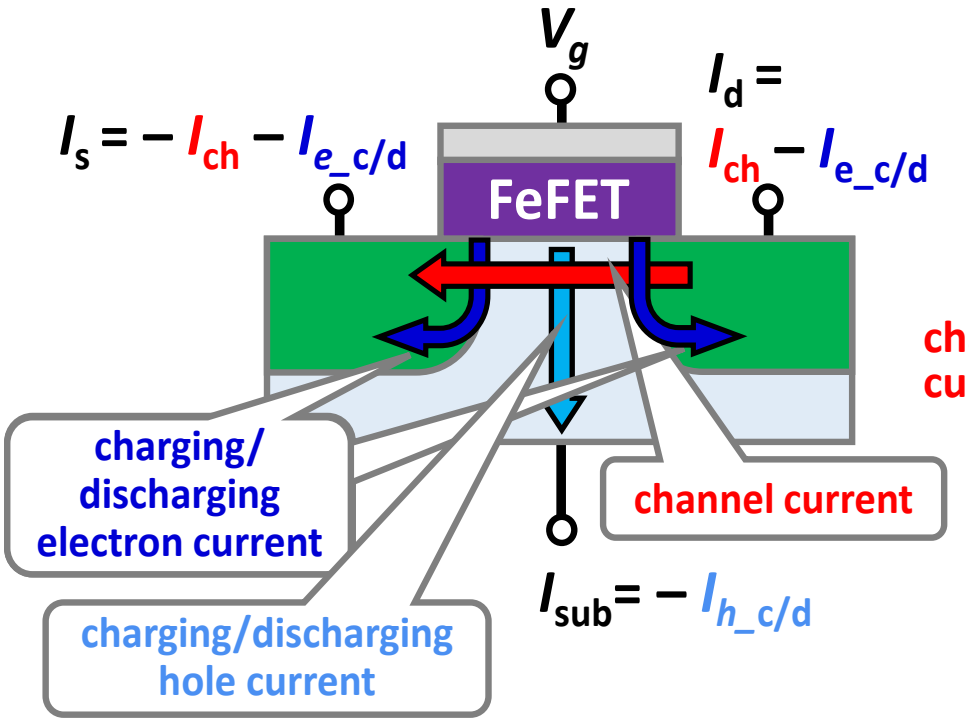
Nako et al., VLSI Symp. (2020) TN1.6; Toprasertpong et al., Comm. Eng. 1 (2022) 21

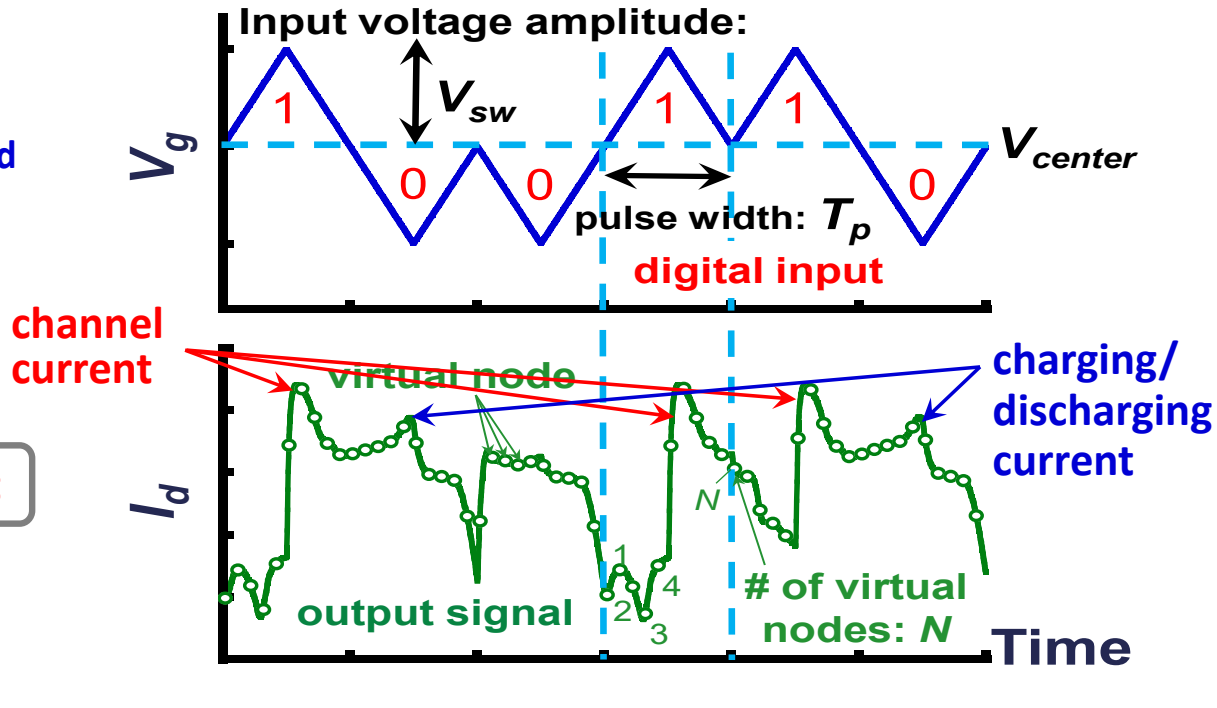


- Inference is performed by training the weights between these virtual nodes and the output layer
- How to extract rich information from output signal is a key to successful reservoir computing

Pattern recognition with current waveform

Nako et al., VLSI Symp. (2020) TN1.6; Toprasertpong et al., Comm. Eng. 1 (2022) 21



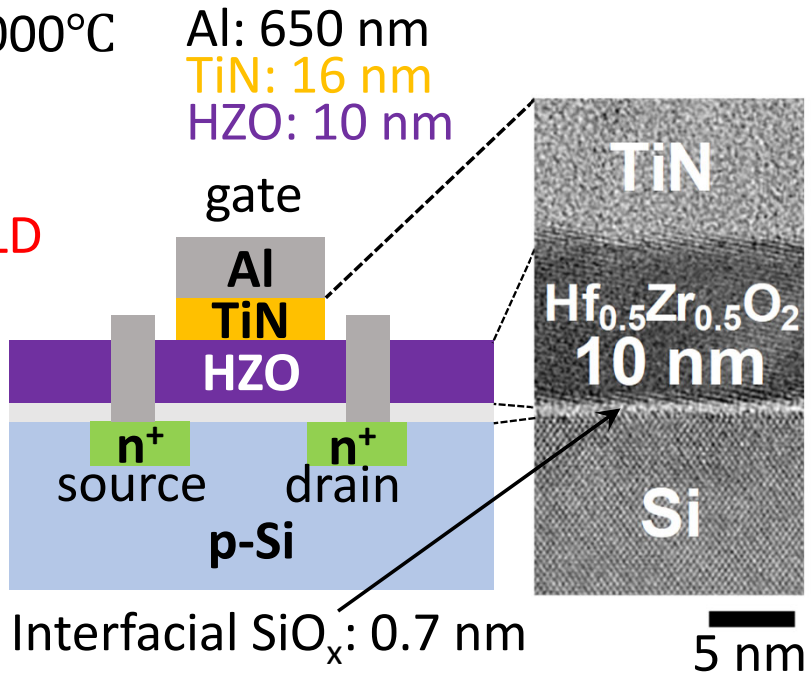


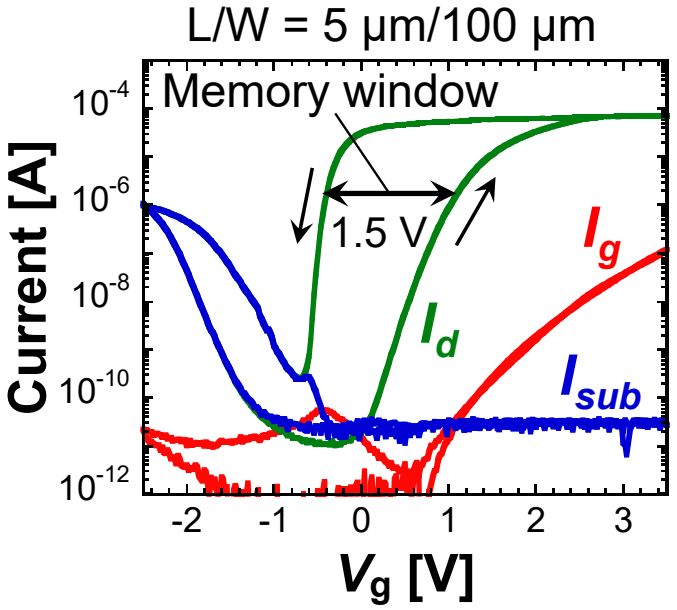
- Polarization switching can affect both channel current and charging/ discharging current
- The time-series current waveform patterns of FeFETs, which depend on the input history, are utilized for inference

FeFET used for reservoir computing

K. Toprasertpong et al., IEDM (2019) 570, VLSI Symp. (2020)TF1.5, APL 116 (2020) 242903, EDL 41 (2020) 1588

- Ion implantation for S/D
- Thermal activation @1000°C
- Remove SiO_2 hardmask
- SiO_x 0.7 nm with SC2
- $Hf_{0.5}Zr_{0.5}O_2$ 10 nm by ALD
- TiN 16 nm
- Al 650 nm
- Gate patterning
- Al for S/D Contact
- Al Back contact
- PMA 400°C 30 s

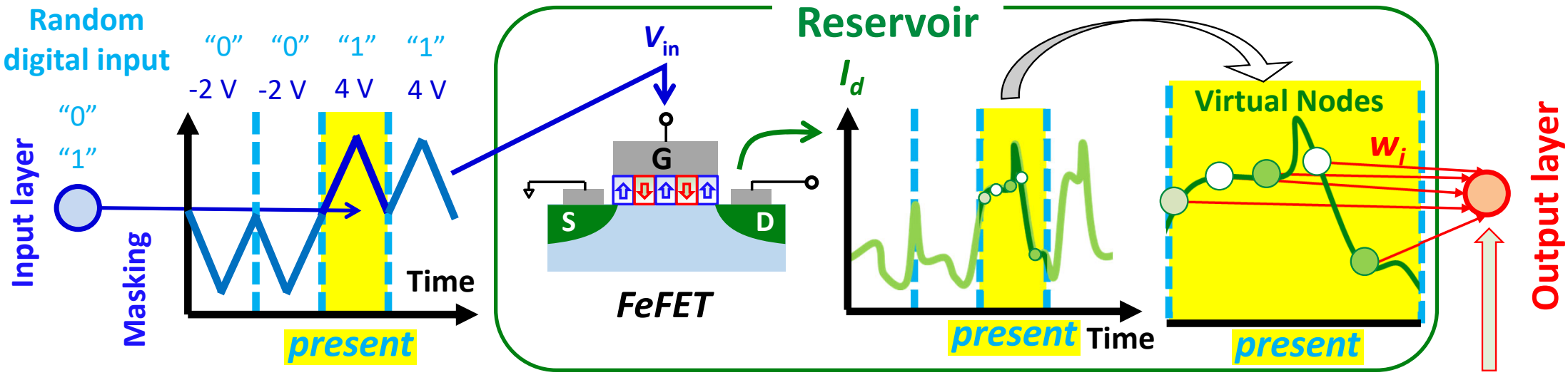




- $TiN/Hf_{0.5}Zr_{0.5}O_2$ (10 nm)/SiO_2/Si FeFETs were used for reservoir computing
- Typical device size: gate length/gate width = 5 μm/100 μm
- Ferroelectric hysteresis is observed

Evaluation of computing performance by basic tasks

Nako et al., VLSI Symp. (2020) TN1.6; Toprasertpong et al., Comm. Eng. 1 (2022) 21



T_{delay} ··· 2 1 0 (time step)

"0" "0" "1"

XOR ("0" "0" "1")

Task 1: Short-term memory

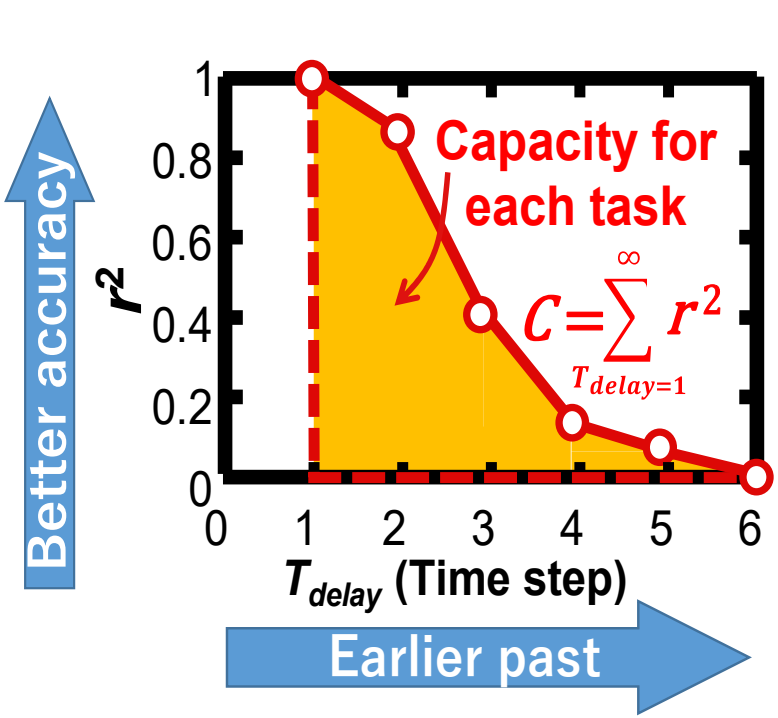
Task 2: Parity check
(needs non-linearity and memory)

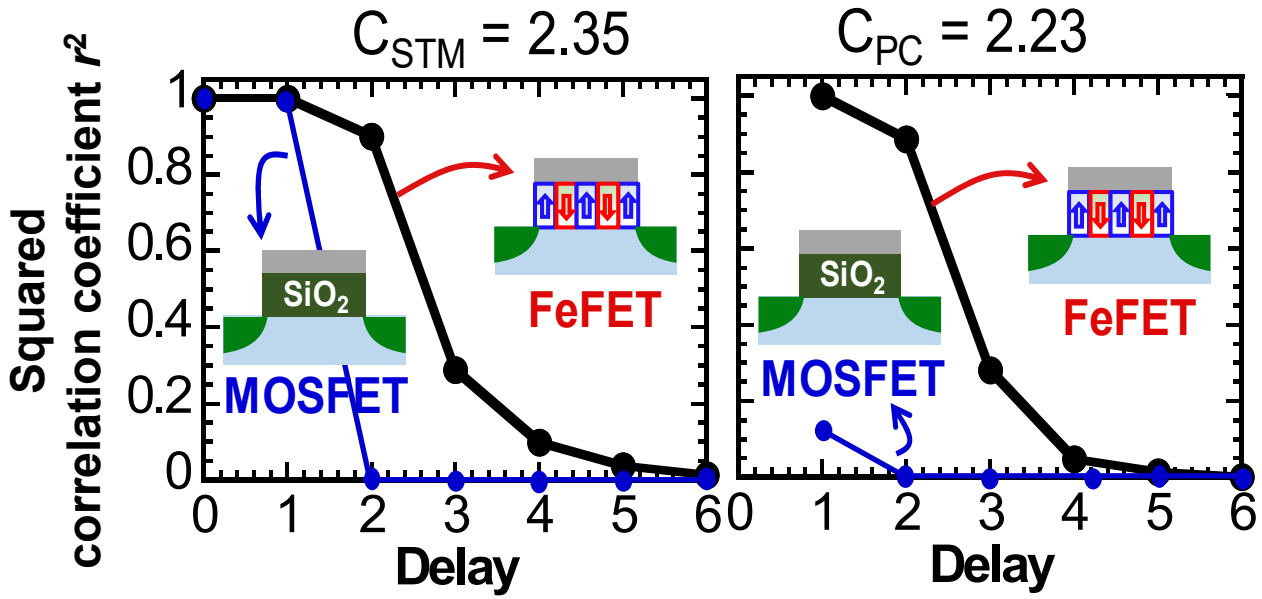
Inference performance for short-term memory and parity check tasks is evaluated as a function of T_{delay}

- Two basic tasks are performed to evaluate the "short-term memory" and "non-linearity" of the FeFET reservoir as a function of the time delay step

Basic task performance of a single FeFET reservoir

Nako et al., VLSI Symp. (2020) TN1.6; Toprasertpong et al., Comm. Eng. 1 (2022) 21

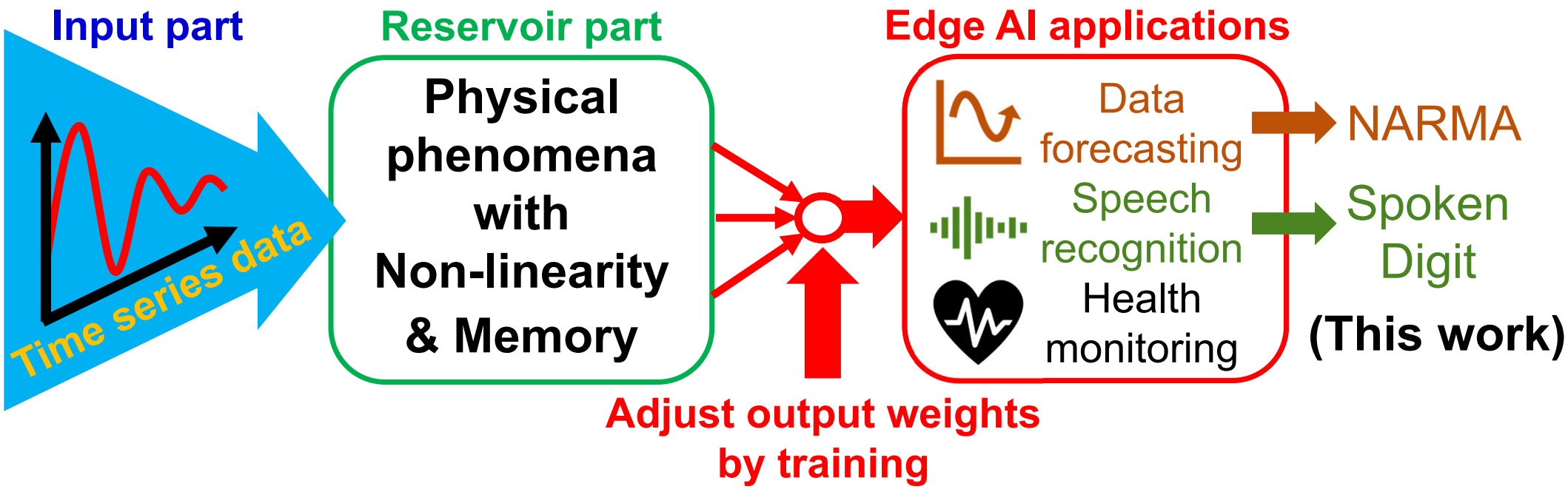




- Reservoir computing performance is estimated by correlation coefficient as a function of the time delay step and the integrated value (capacity)
- MOSFETs have no reservoir computing performance
- FeFETs exhibit much higher performance, attributed to polarization in HZO

Application of physical reservoir computing in this study

G. Tanaka et al., Neural Networks 115, 100 (2019)



- Potentially important edge AI applications include data forecasting, speech recognition, and health monitoring
- In this study, we apply FeFET reservoir computing to two applications, NARMA as a data forecasting application and spoken digit as a speech recognition application

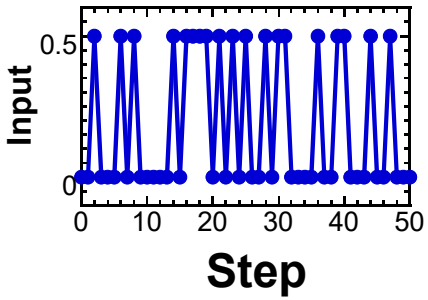
Prediction of time-series output from NARMA

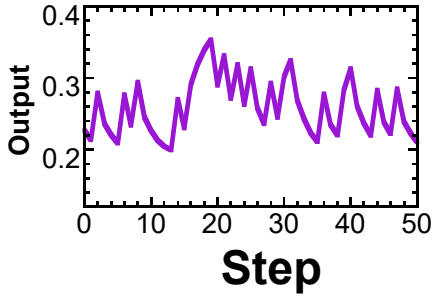
NARMA-2 $d(n) = 0.4d(n-1) + 0.4d(n-2) + 0.6u^3(n) + 0.1$

NARMA-N (N>=3)

$d(n) = 0.3d(n-1) + 0.05d(n-1)[d(n-1) + ... + d(n-N)] + 1.5u(n)u(n-N+1) + 0.1$

A. F. Atiya et al., IEEE Trans. Neural Netw. 11, 697 (2000)





Nth-order dynamical system

Input $u(n)$

Model output $d(n)$

prediction

Computing system output $y(n)$

W

FeFET reservoir

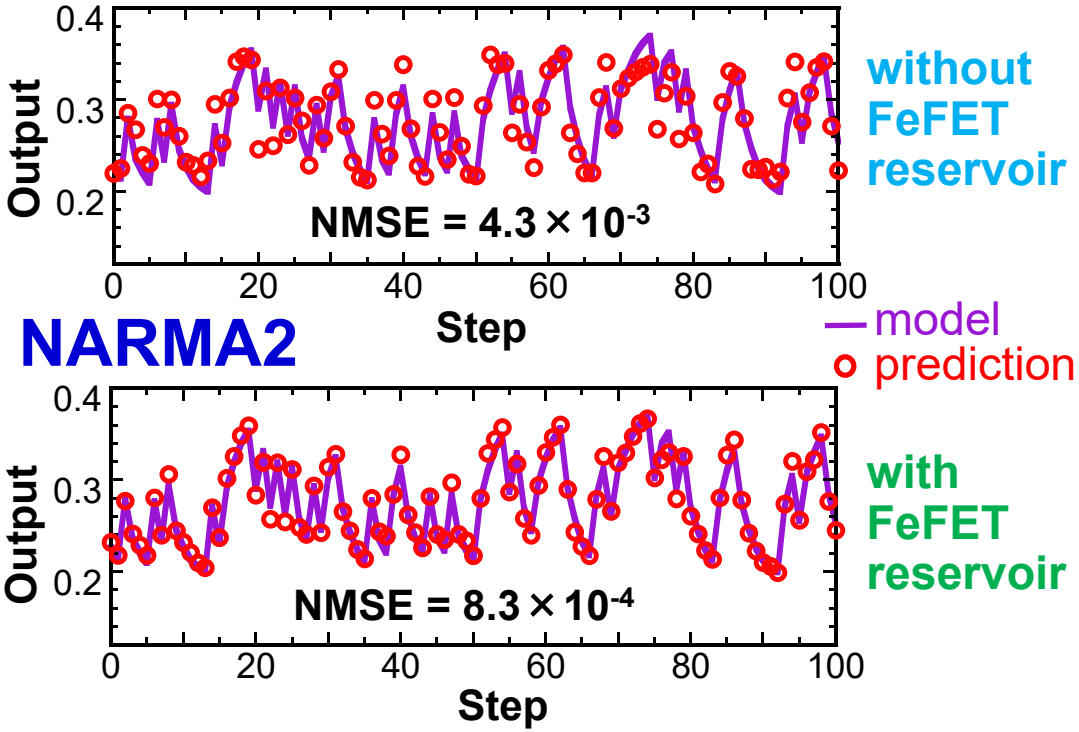
Learning to predict (reproduce) $d(n)$

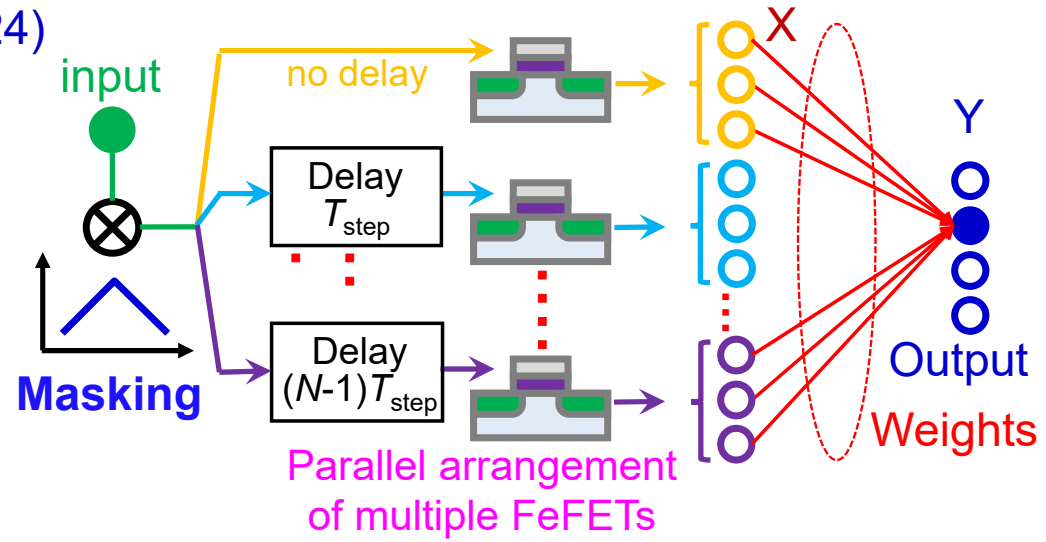
NARMA: nonlinear autoregressive moving-average time series

- Performance of the prediction of time-series output from an Nth-order nonlinear dynamic system (NARMA-N) is examined

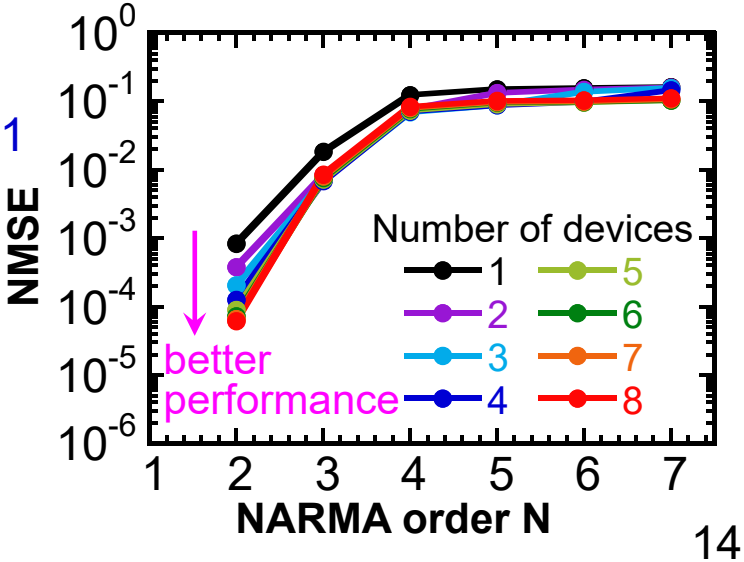
NARMA prediction results by FeFET reservoir

S. Takagi et al., IEDM, 22.2 (2024)





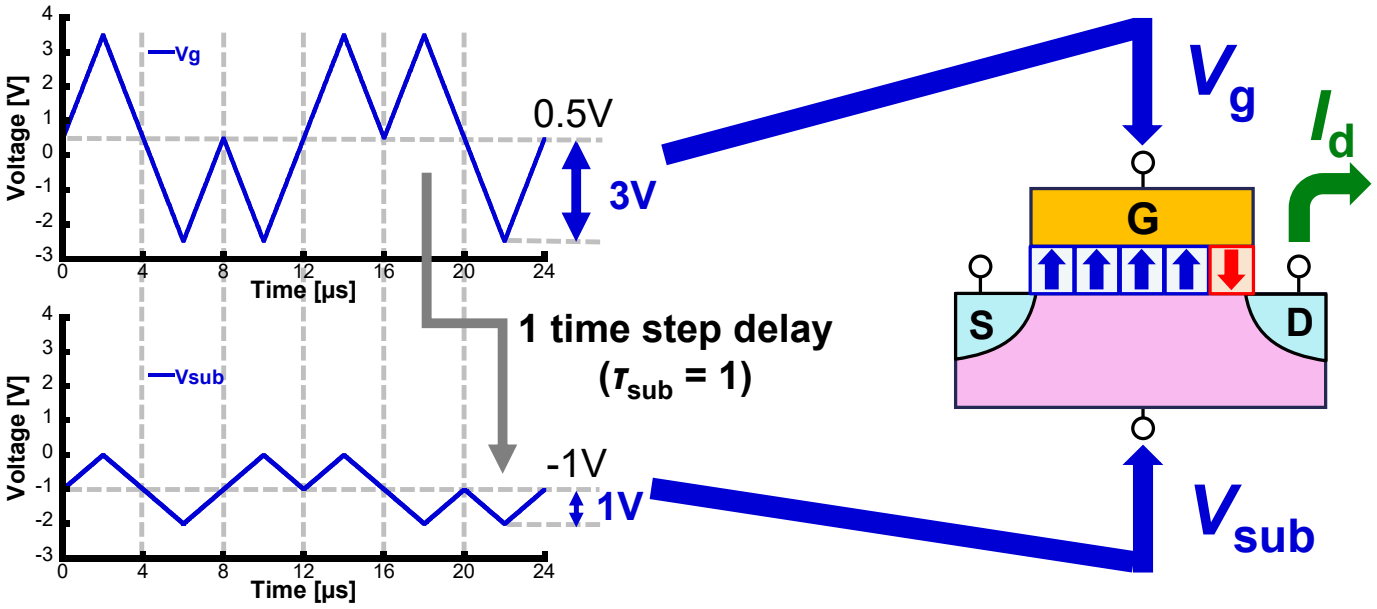
Toprasertpong et al., Comm. Eng. 1 (2022) 21

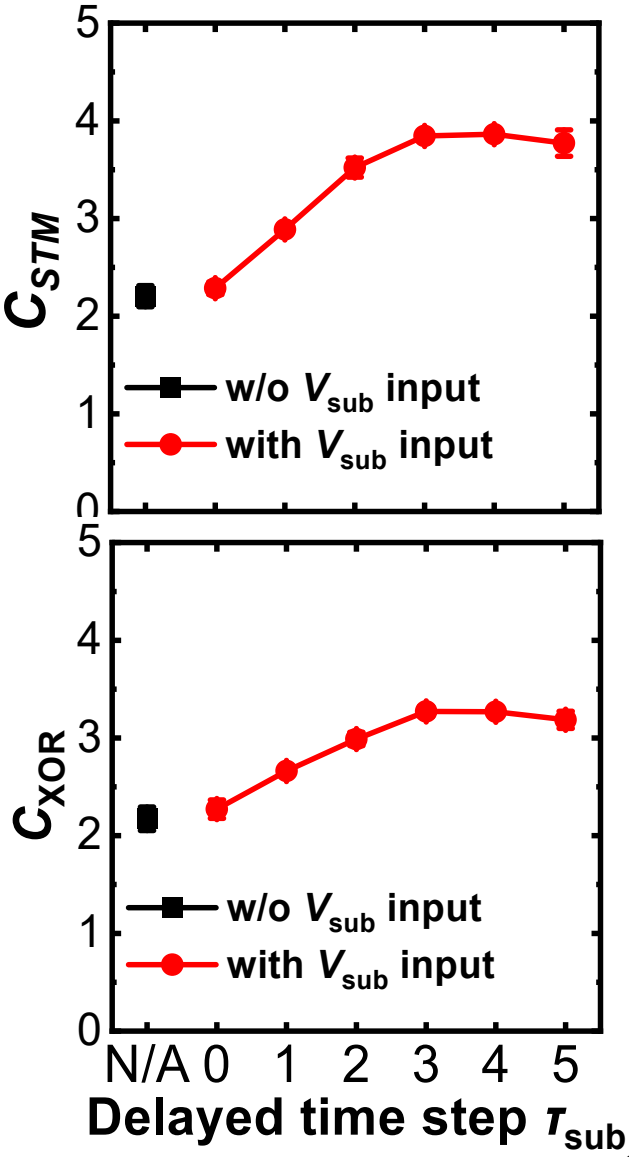


- High prediction accuracy (low normalized mean square error (NMSE)) is obtained by FeFET reservoir
- Prediction accuracy is enhanced by multiple FeFETs in parallel with past data input to the gate to improve short-term memory by a circuit with a time delay

Improvement of reservoir performance by combing gate and substrate input to FeFET

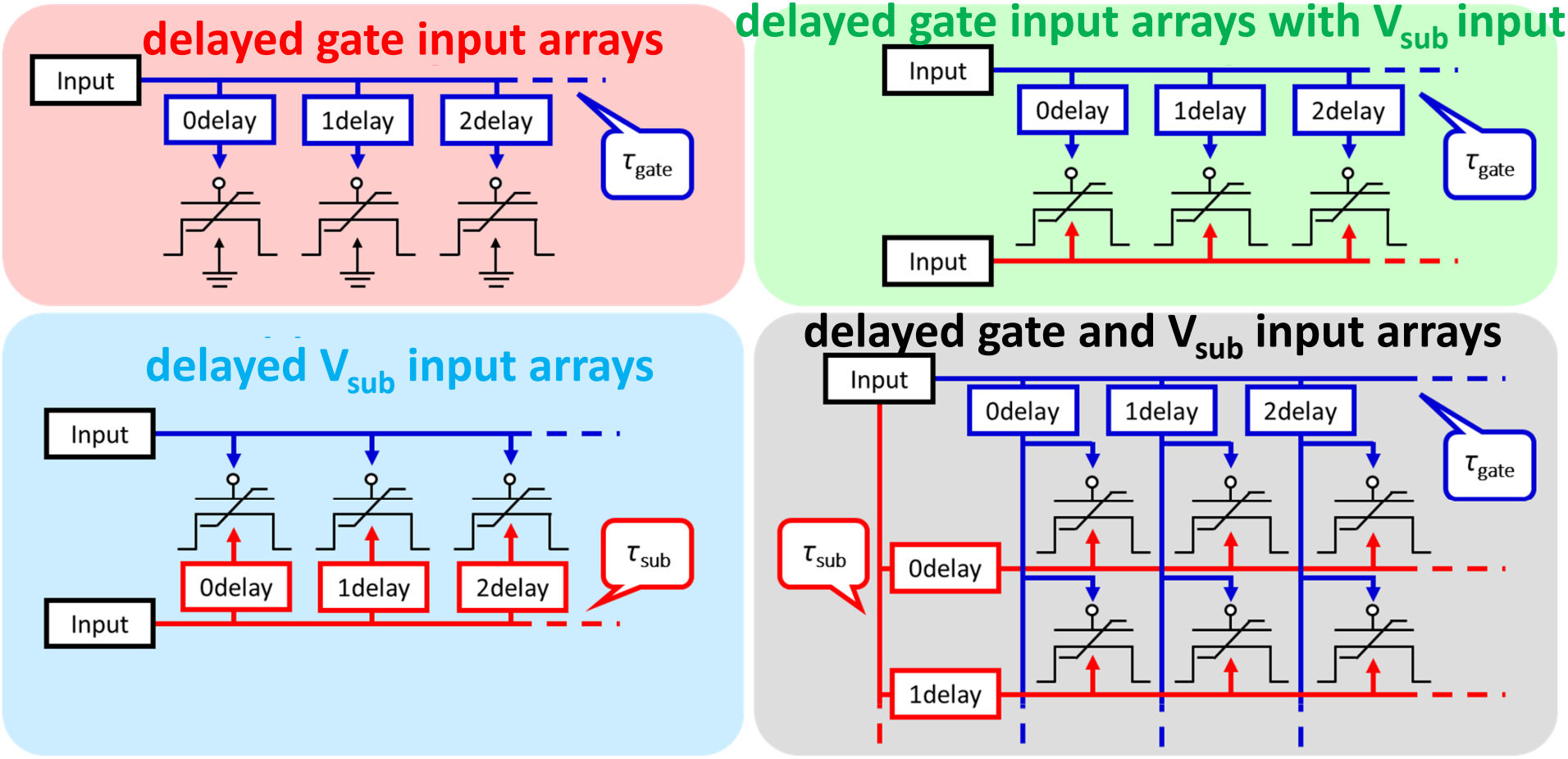
E. Nako et al., published in APL Machine Learning (2026)





- When the input delayed relative to the gate input is applied in reverse phase to substrate of FeFETs, STM and XOR capacities are enhanced

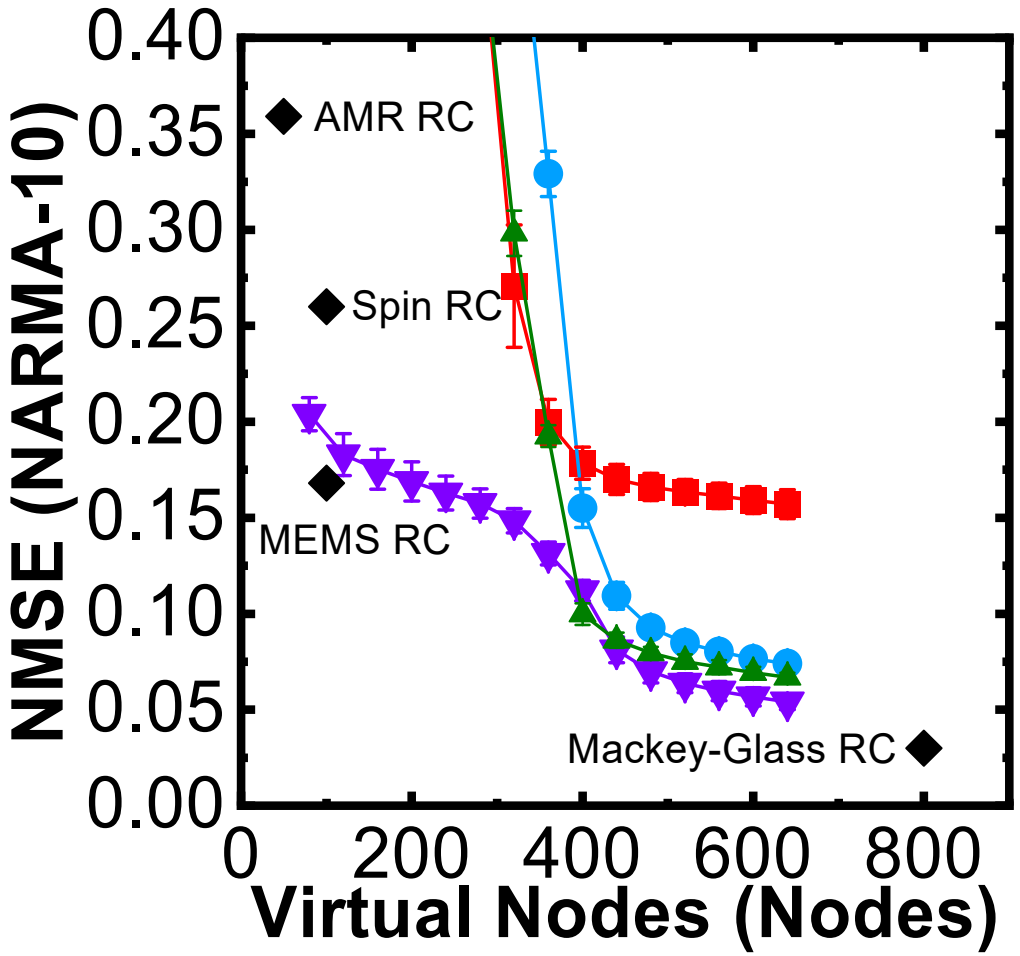
FeFET reservoir arrays with different combinations of delayed substrate and gate inputs



- The gate and substrate input and the delay times can be adjusted independently, allowing us to propose various combinations

E. Nako et al., published in APL Machine Learning (2026)

Benchmark of NMSE in NARMA-10



E. Nako et al., published in APL Machine Learning (2026)

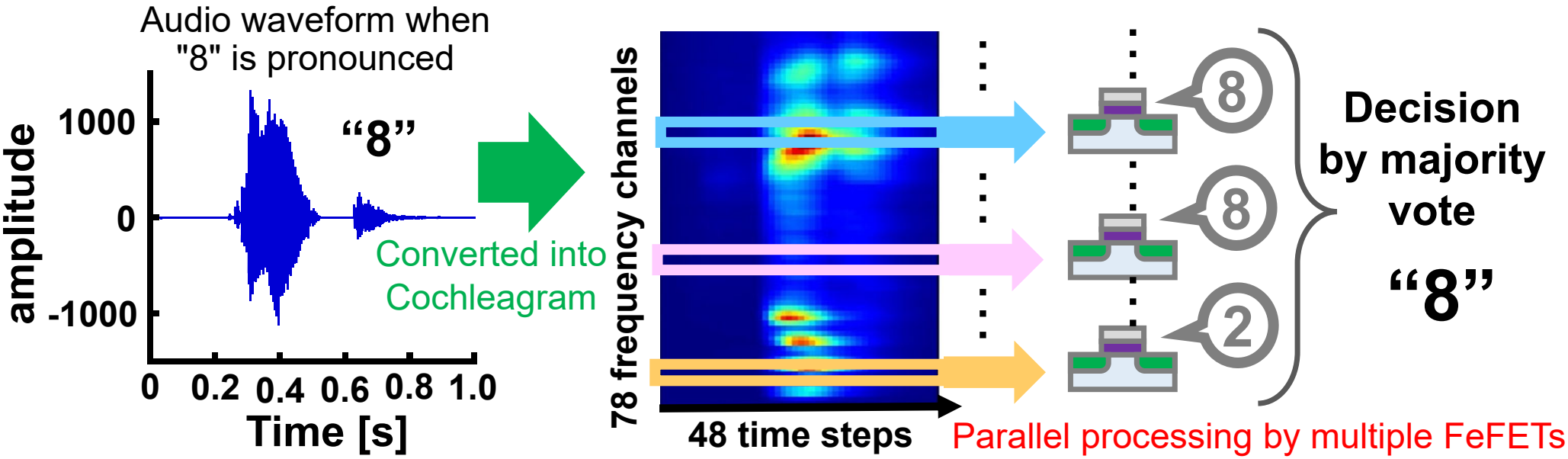
- **delayed gate input arrays**
- **delayed V_{sub} input arrays**
- **delayed gate input arrays with V_{sub}**
- **delayed gate input arrays with V_{sub} + (τ_{gate}, τ_{sub}) = (0, 9)**
- ◆ Other RC system

[AMR] I. T. Vidamour et al., Commun. Phys. (2023)
[MEMS] B. Bazarani et al., J. Microelectromechanical Syst. (2020)
[Spin RC] W. Namiki et al., Adv. Intell. Syst. (2023)
[Mackey-Glass RC] L. Appeltant et al., Sci. Rep. (2014)

• FeFET reservoirs with delayed substrate and gate input arrays can deliver NARMA-10 performance comparable to other physical reservoirs

Spoken digit recognition by FeFET reservoir computing

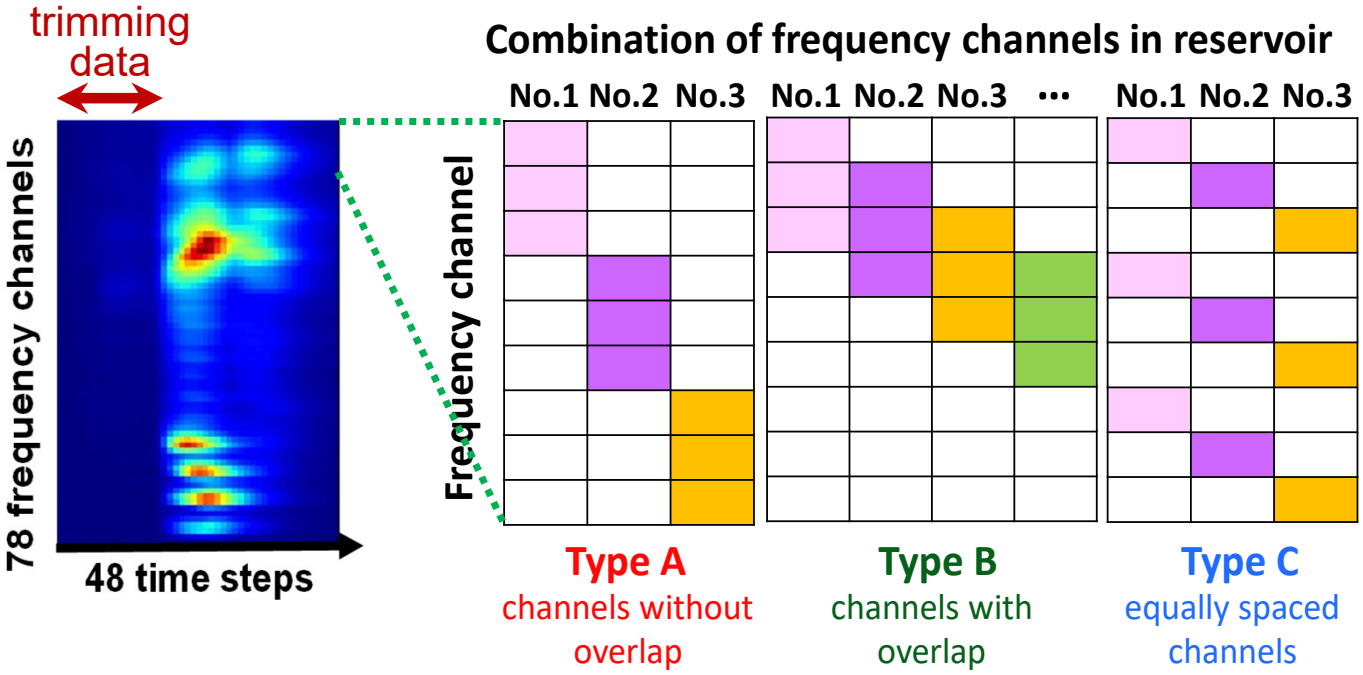
Nako et al., VLSI Symp., 220 (2022); IEEE TED, 70, 5657 (2023); S. Takagi et al., IEDM, 22.2 (2024)

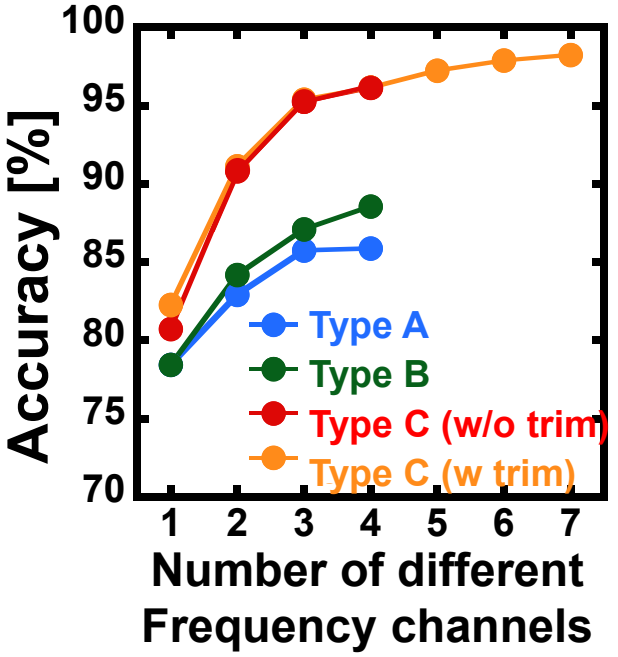


- Spoken 0-9 digit speech data are converted into cochleagram, which is composed of time series data with multi-frequency channels
- We have proposed a reservoir computing scheme using parallel processing by multiple FeFETs for spoken digit recognition
- A final decision is made by a majority vote of inference by multiple FeFETs

Improvement by combination of different frequency channels

Nako et al., VLSI Symp., 220 (2022); IEEE TED, 70, 5657 (2023); S. Takagi et al., IEDM, 22.2 (2024)

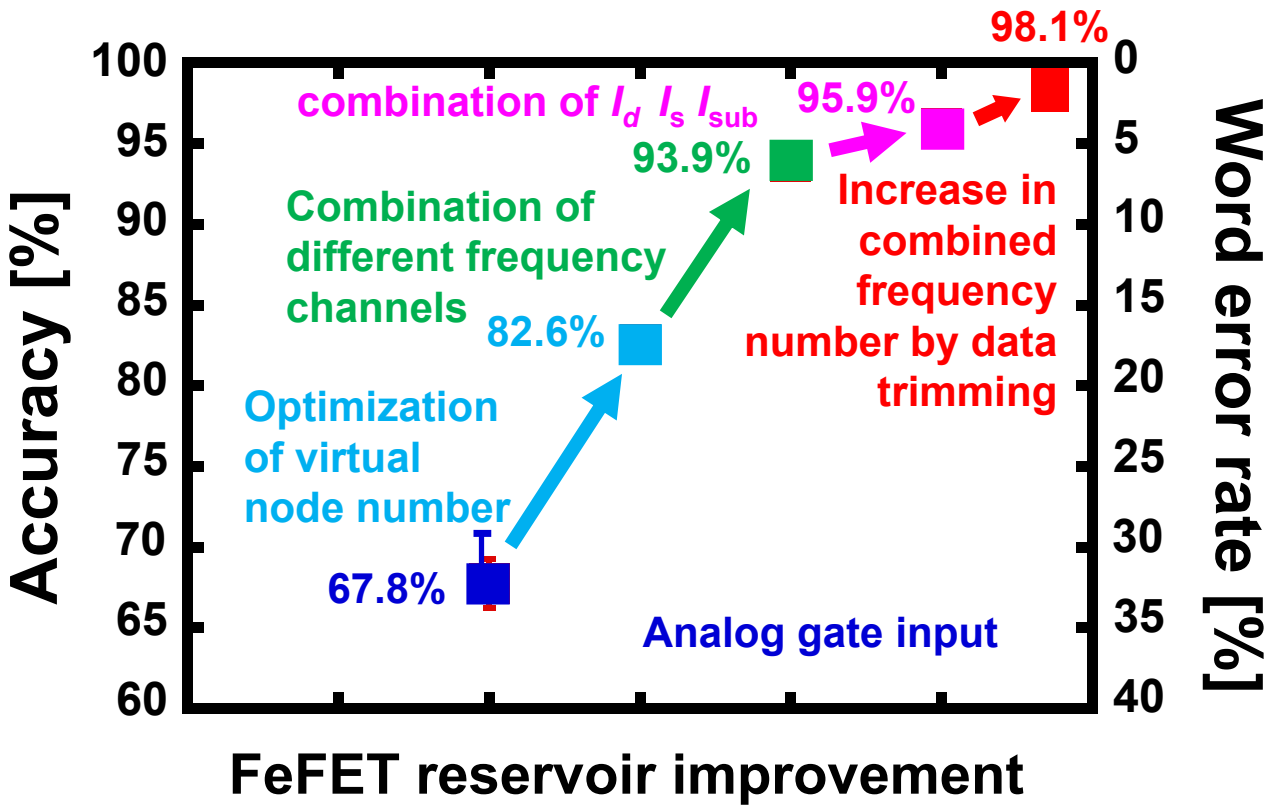




- A combination of different frequency channels and optimization of how to arrange their combinations are effective for improving the accuracy
- The recognition accuracy is enhanced by maximally combining distant frequency channels after trimming uninformative data in cochleagram

Classification accuracy of speech recognition

Nako et al., VLSI Symp., 220 (2022); IEEE TED, 70, 5657 (2023); S. Takagi et al., IEDM, 22.2 (2024)



- Each improvement additively contributes to an increase in recognition accuracy
- FeFET reservoir computing achieves 98.1% classification accuracy

Summary

- We have proposed and demonstrated a Si-CMOS-friendly physical reservoir computing scheme by using HZO/Si FeFETs, where the time response of currents of Si FeFETs is utilized as the virtual nodes
- We experimentally presented the fundamental properties of FeFET reservoir computing performance and highlighted our several unique approaches that can be used to build high-performance FeFET reservoir computing systems
- This scheme has been successfully applied to two typical AI tasks; prediction of nonlinear time-series data and speech recognition
- High performance of a prediction task of NARMA data has been realized by reservoir computing using FeFET arrays with delayed substrate and gate input
- We experimentally show a classification accuracy of 98.1 % in a speech recognition task to classify the audio waveforms of ‘0’ to ‘9’ spoken digits
- The proposed FeFET reservoir has a high potential to provide effective reservoir computing platforms for versatile applications by incorporating various ideas, including the use of a large number of FeFETs and the combination of existing CMOS circuits and new circuit technology

Acknowledgement

This work was supported by JST CREST (JPMJCR20C3) and JSPS KAKENHI (21H01359)

Thank you for your attention!