A Low Power and Small Area Analog Voice Activity Detector Featuring a Time-Domain CNN as a Programmable Feature Extractor and a Sparsity-Aware Computational Scheme in 28nm CMOS

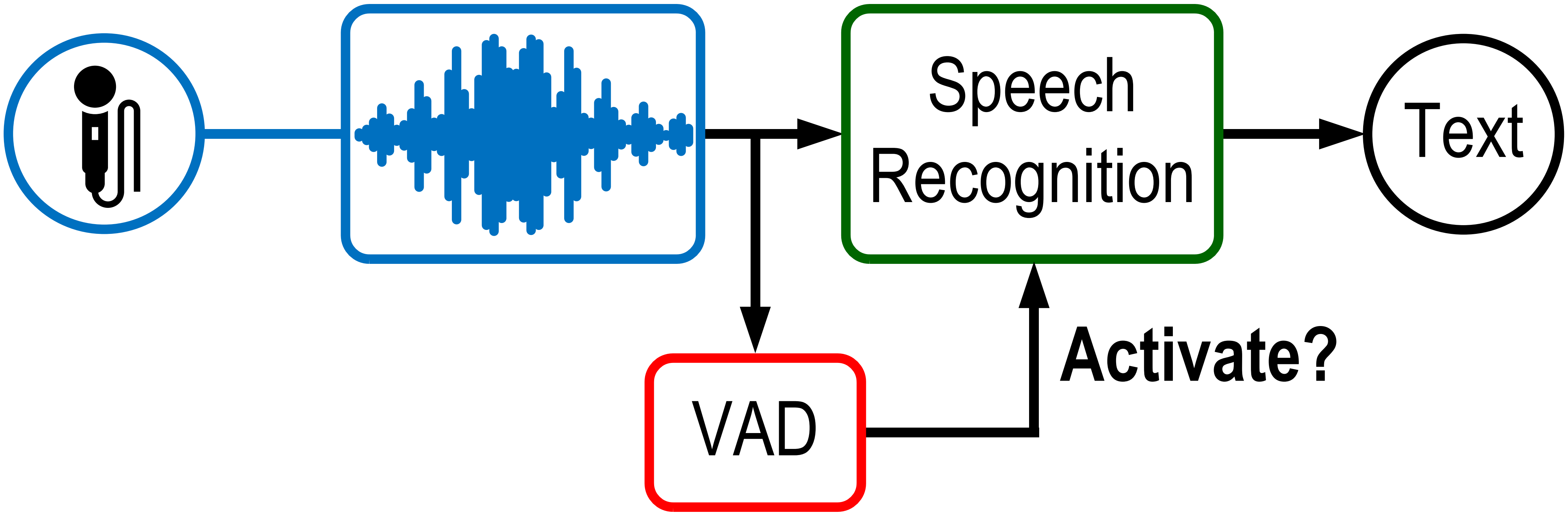
Ka-Fai Un, Assistant Professor
阮家煇 助理教授

State-Key Laboratory of Analog and Mixed-Signal VLSI
University of Macau, Macao, China

Outline

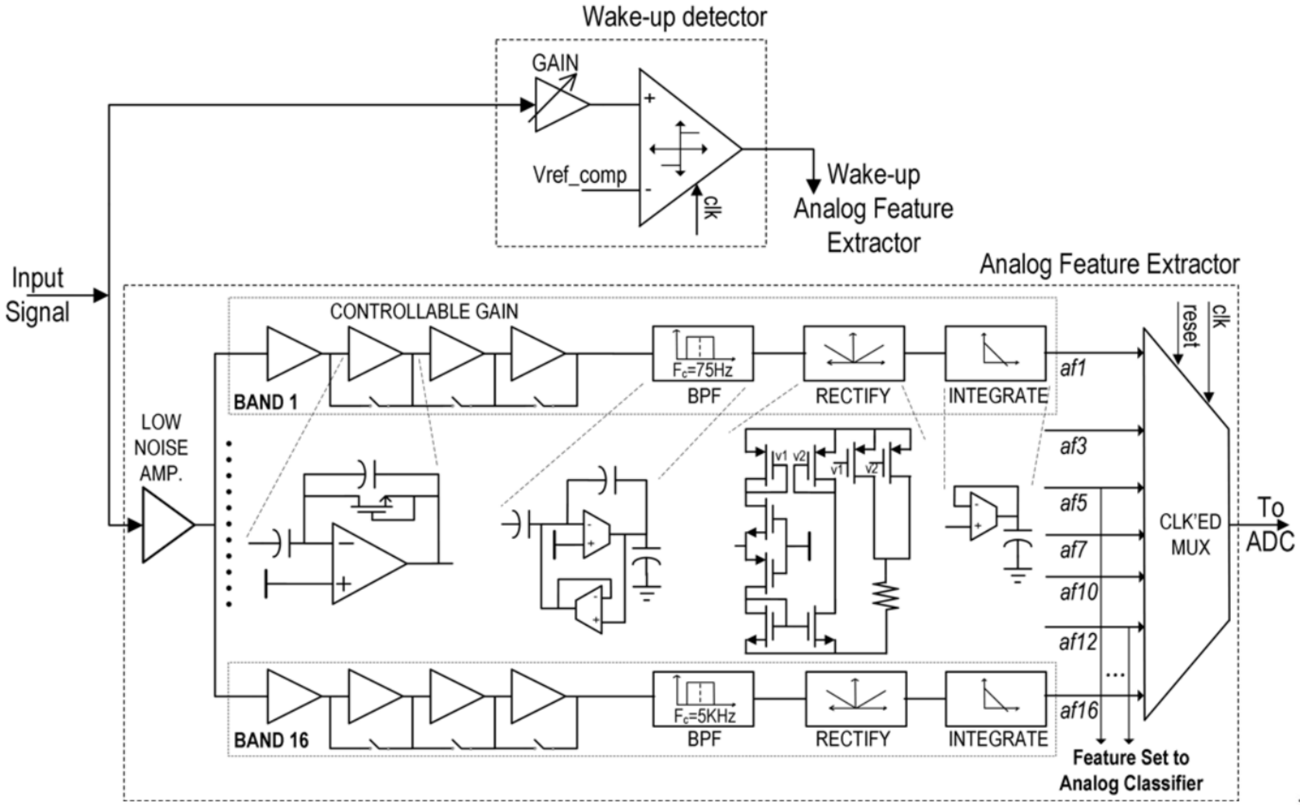
- **Motivation and Prior Arts**
- **Proposed Analog Voice Activity Detector (VAD)**
  - Time-Domain Convolutional Neural Network (TD-CNN)
  - Sparsity-Aware Computation (SAC)
  - Sparsified Quantization (SQ)
- **Measurement Results**
- **Conclusions**

Edge Computing VAD

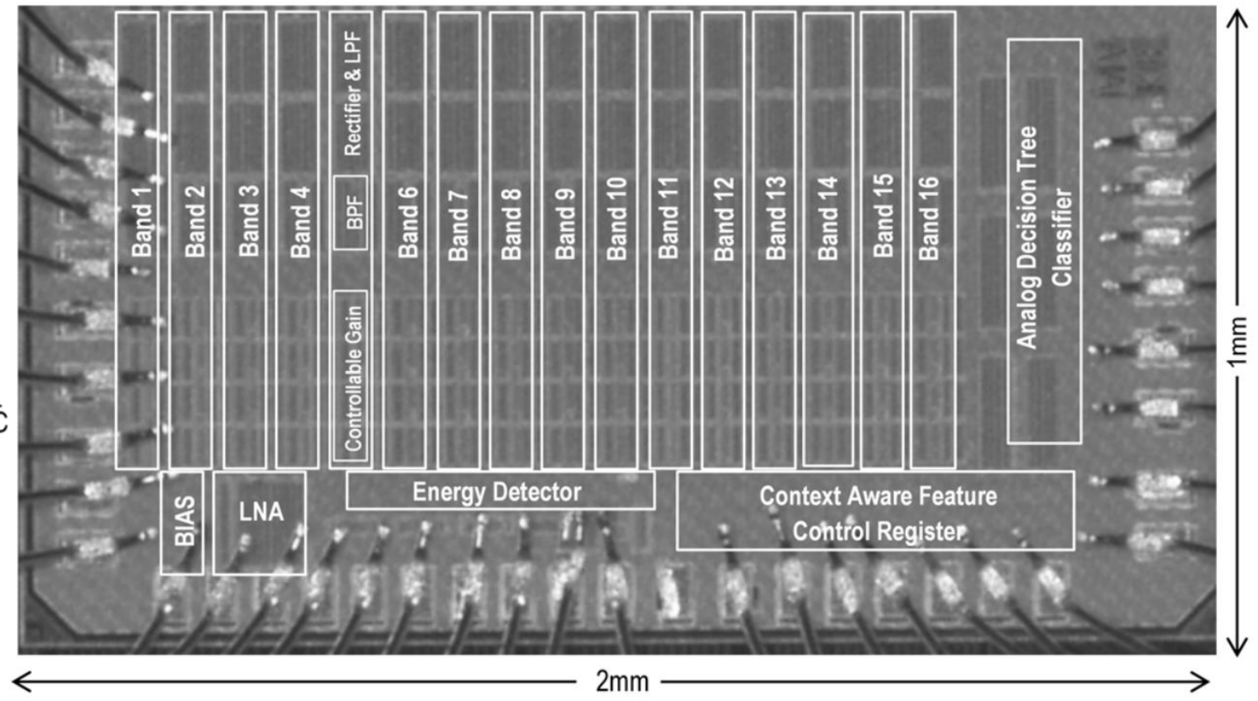


- Speech recognition(SR) for smart assistant, automatic subtitle, etc.
- Always on VAD activates SR when there is human voice
- Edge computation for privacy and reducing system power

Analog Feature Extractor (AFE) + Decision Tree (DT)-based Classifier

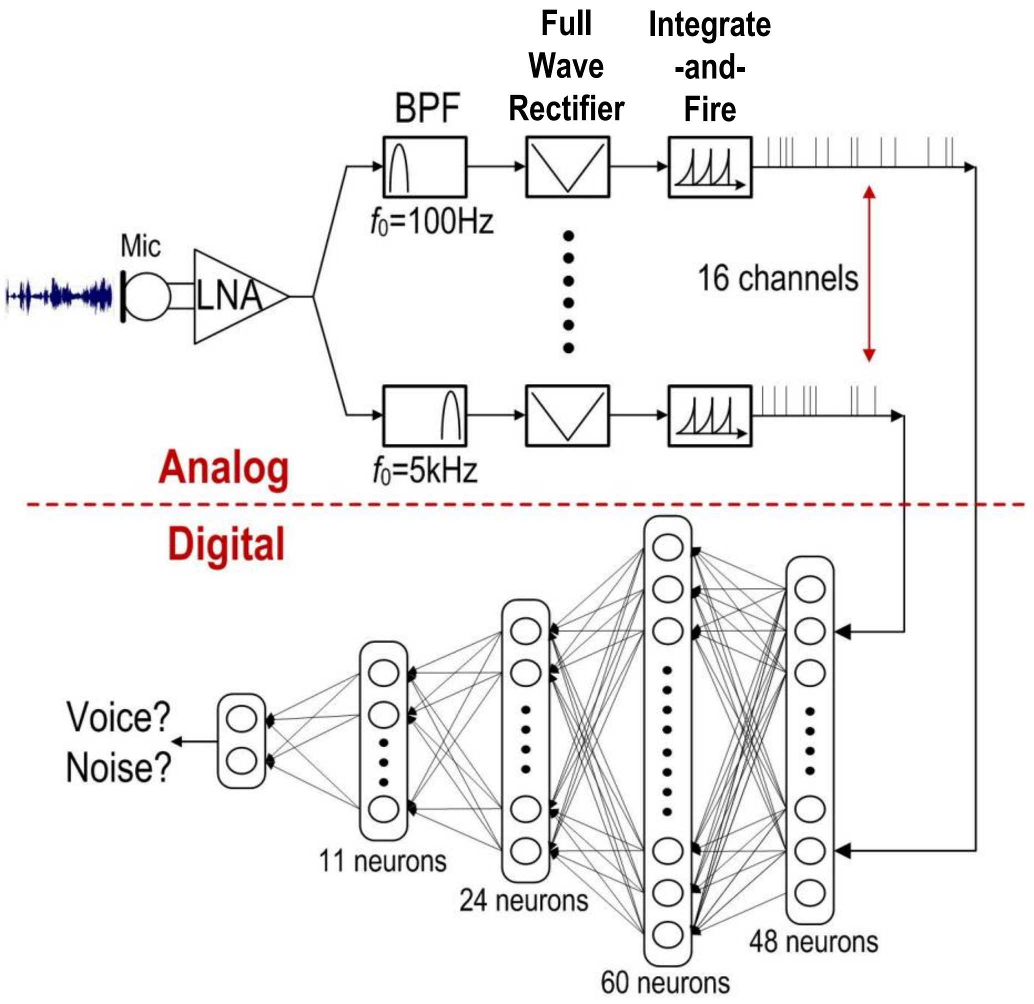


[K. Badami, et al. ISSCC'15]

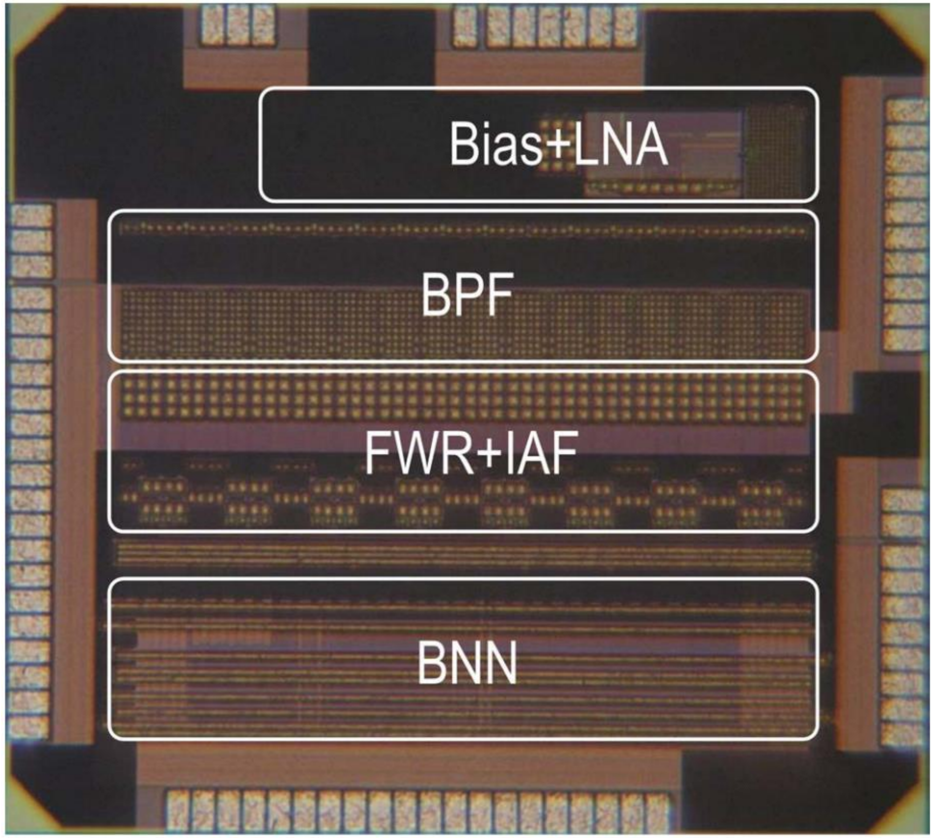


- ☒ **DT: inferior to Neural Network (NN)-based classifier**
- ☒ **AFE: power and area consuming**

AFE + BNN Classifier



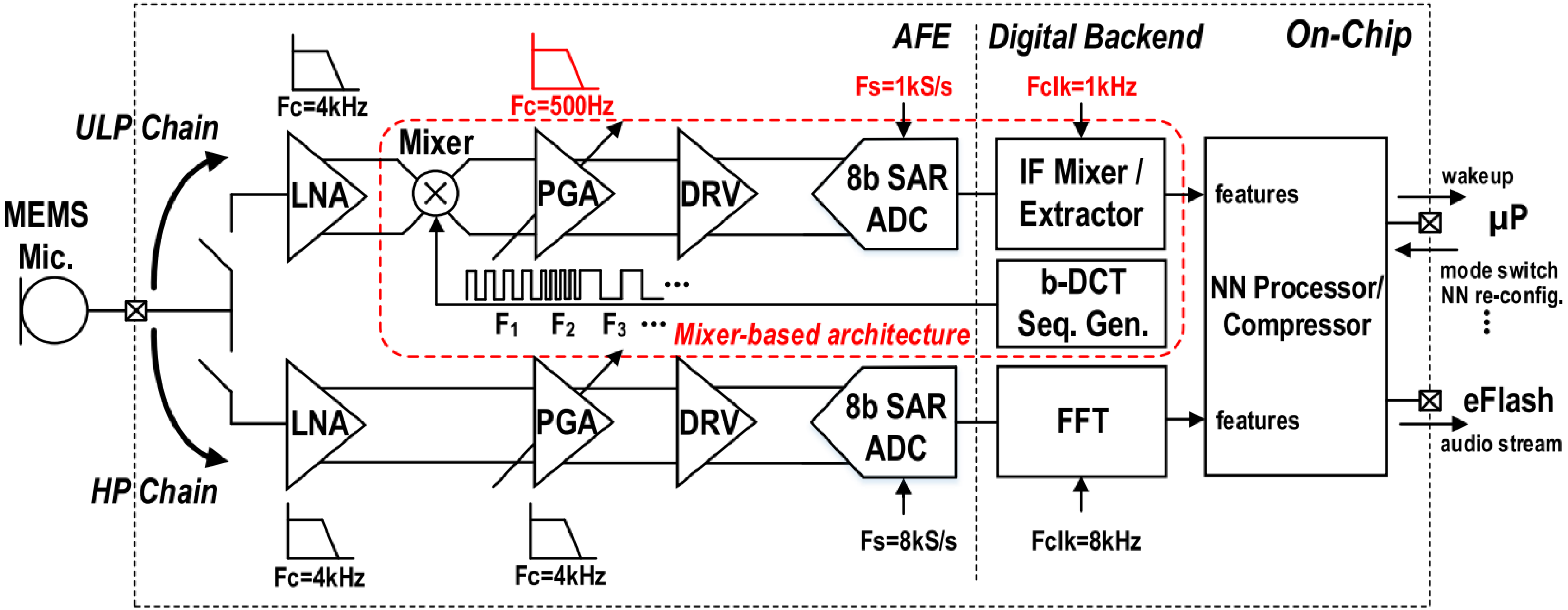
[M. Yang, et al. ISSCC'18]



☒ **AFE: power and area consuming** ☑ **BNN classifier**

☒ **Require high resolution ripple counter → High Power**

Mixer-based AFE + NN Classifier

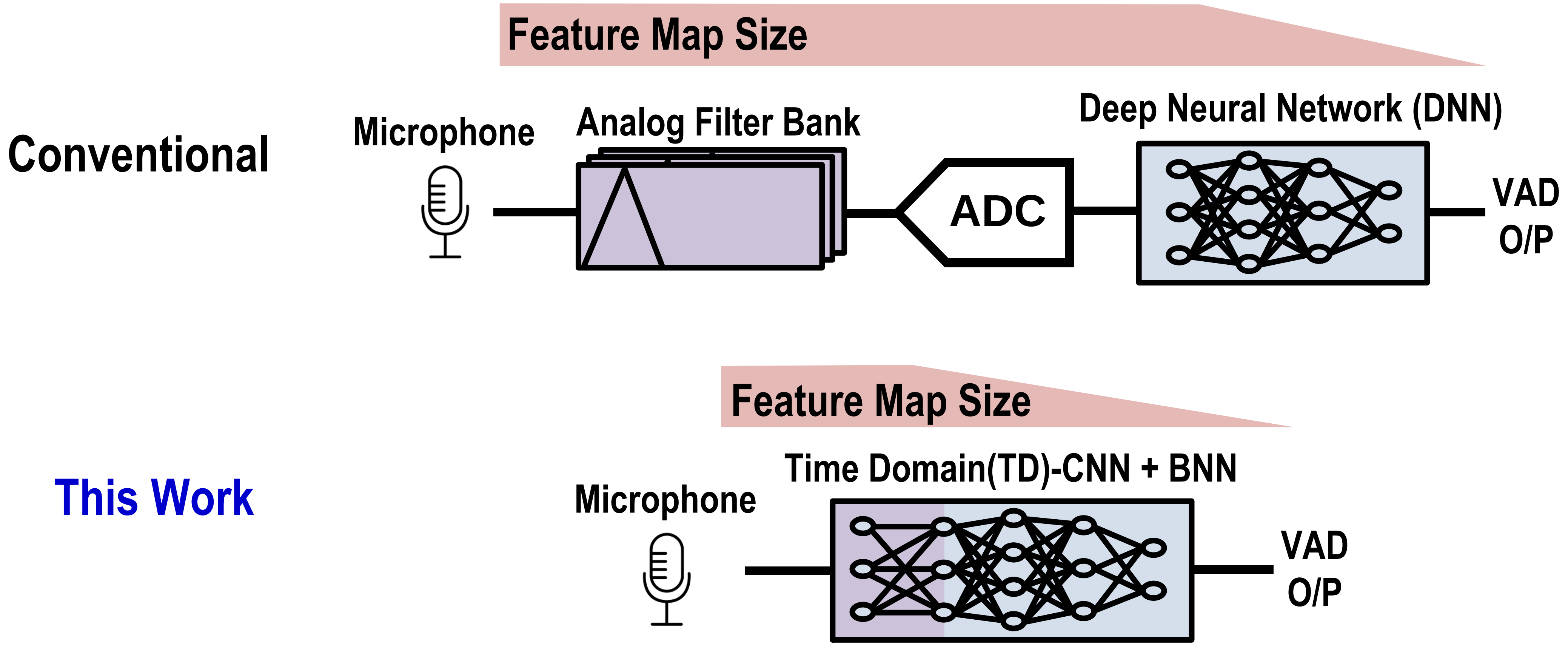


[M. Cho, et al. ISSCC'19]

Mixer-based TI-AFE: ☑ Power ☑ Area ☒ Latency (512ms)

☒ Incomplete extracted feature

Proposed VAD

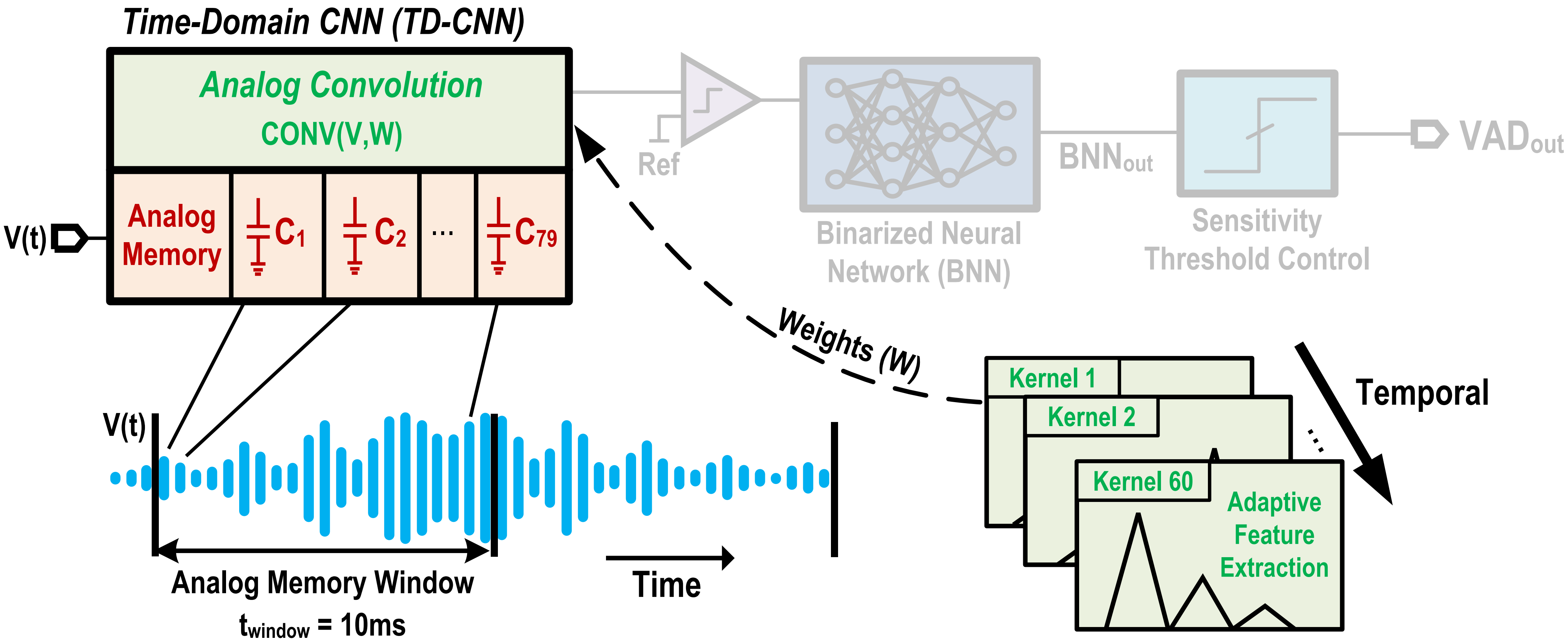


- ☑ **TD-CNN: low power SC-based feature extraction**
- ☑ **Binarized features: reduce A/D conversions**
- ☑ **BNN: small model size and simple operations**

Outline

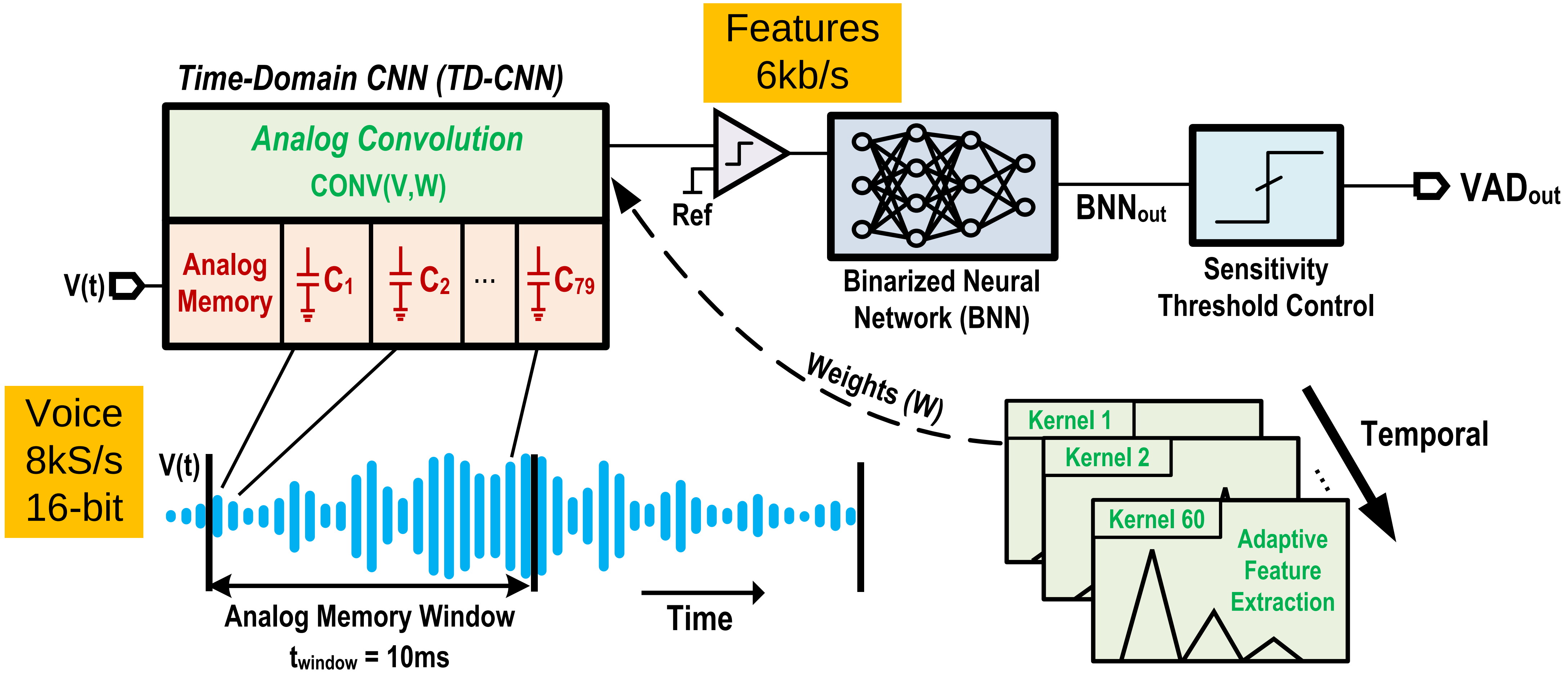
- Motivation and Prior Arts
- **Proposed Analog Voice Activity Detector (VAD)**
  - System Architecture
  - Time-Domain Convolutional Neural Network (TD-CNN)
  - Sparsity-Aware Computation
  - Sparsified Quantization
- Measurement Results
- Conclusions

System Architecture of VAD



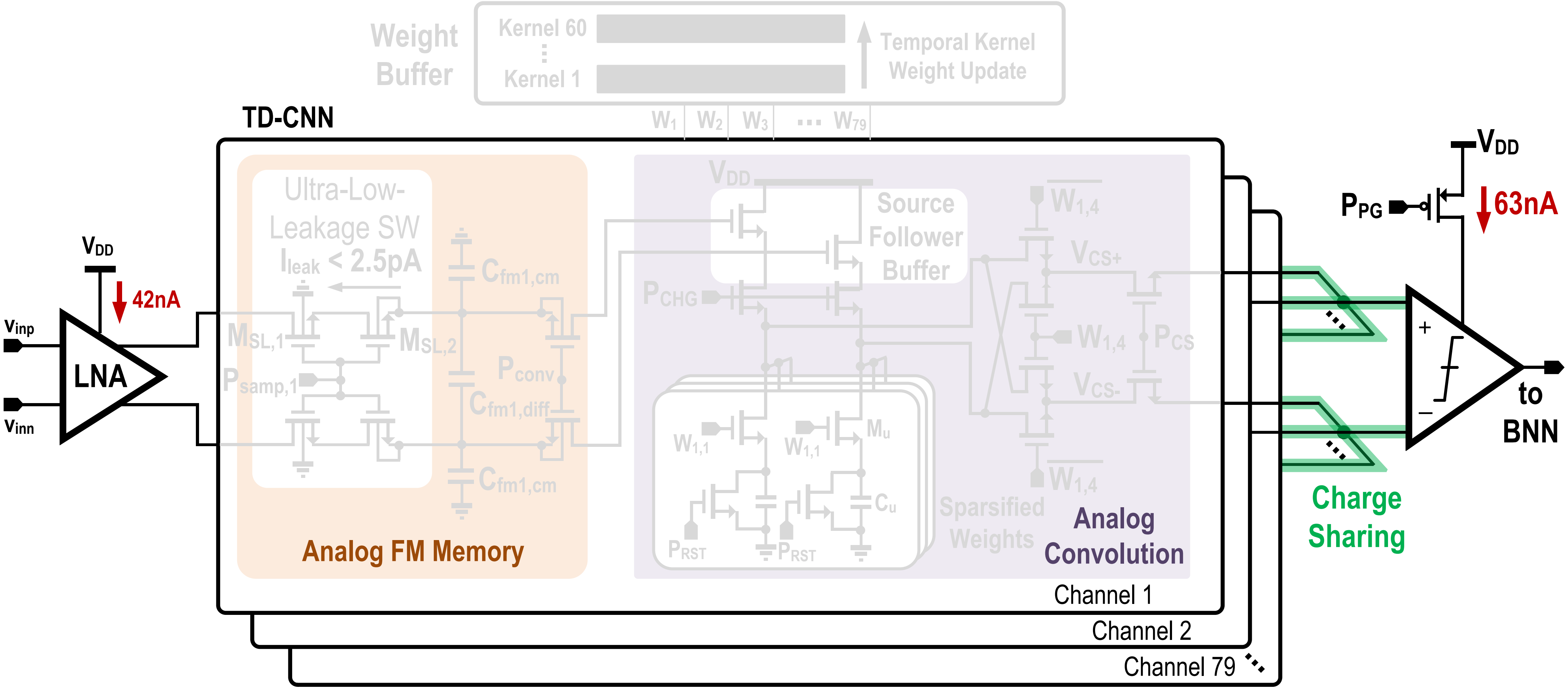
- Sampling capacitor array as analog memory
- Switched capacitor-based TD-CNN as feature extractor

System Architecture of VAD



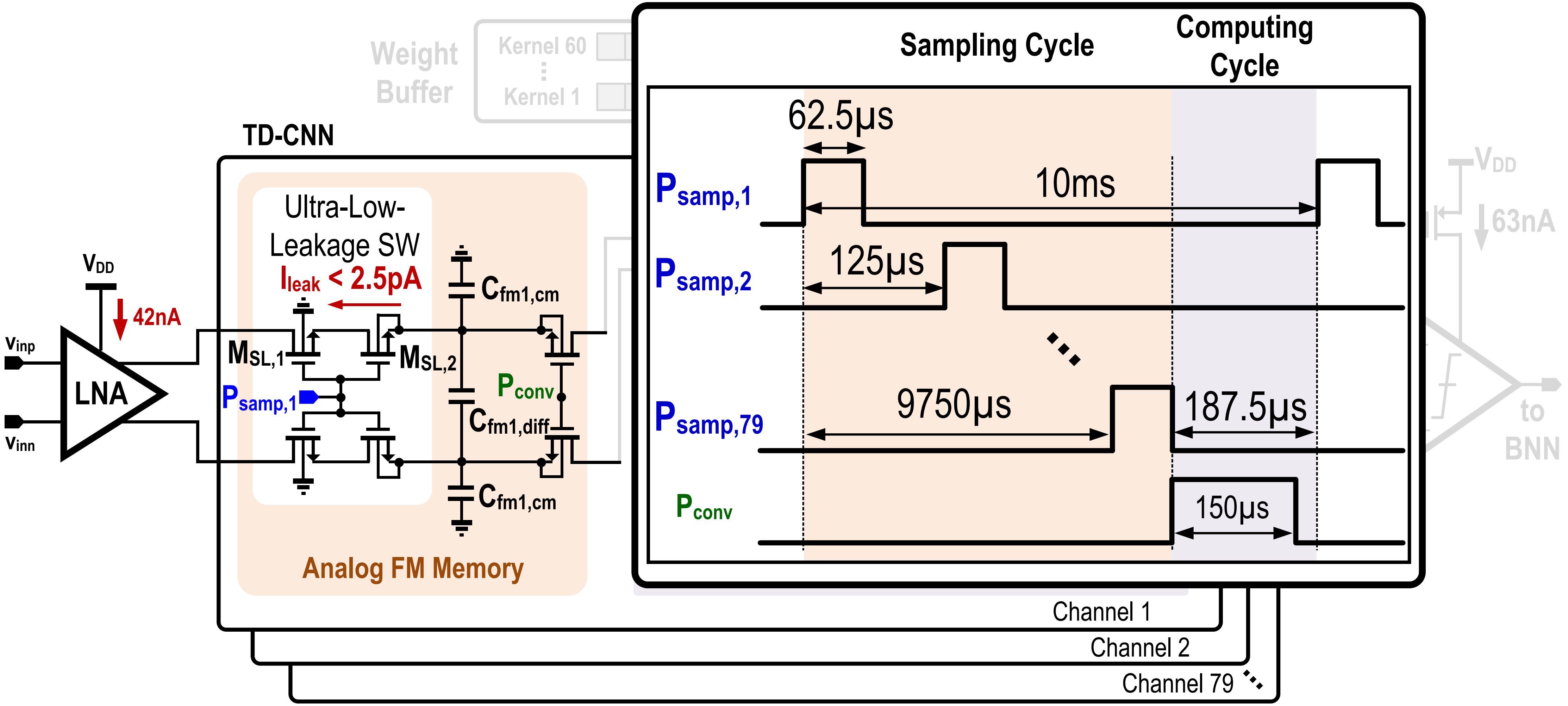
- Binarized features processed by BNN for VAD O/P
- Sensitivity threshold control to balance FP and FN

Time Domain-CNN - Overview



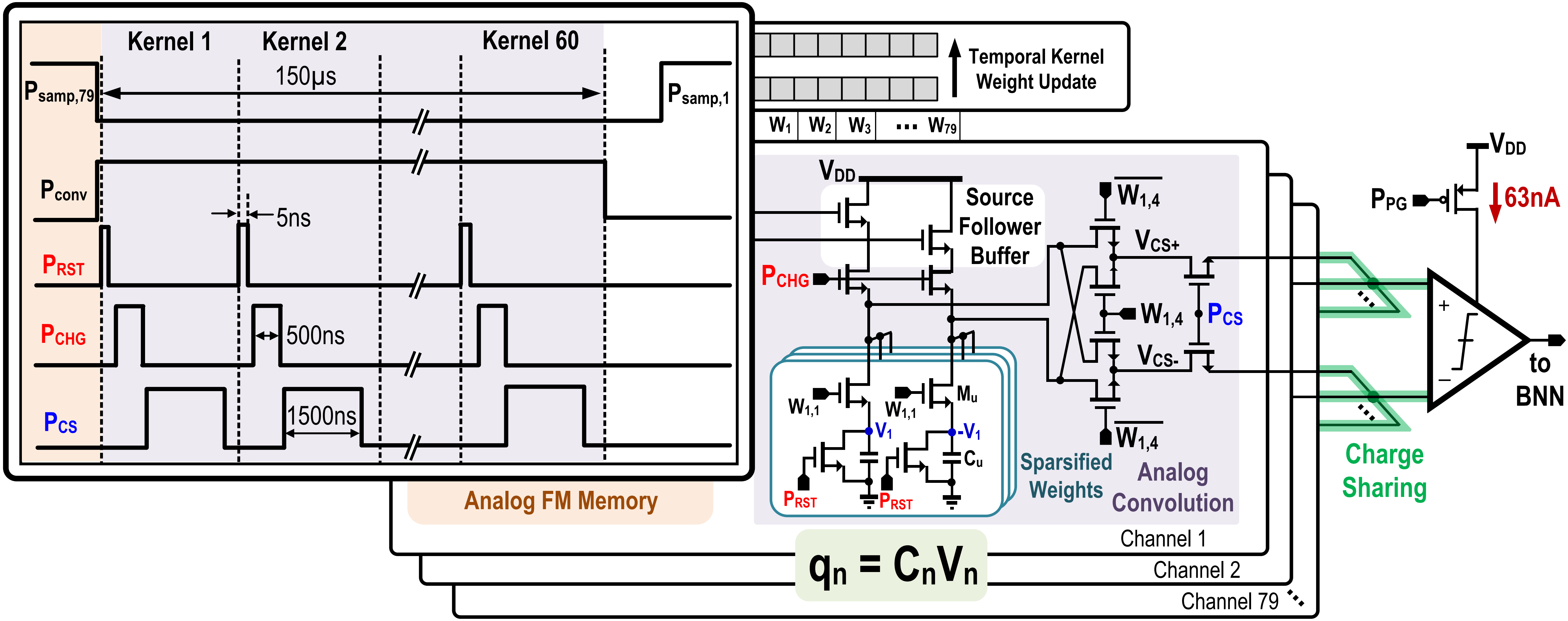
- Analog voltage is extracted as binary feature maps
- TD-CNN includes analog memory and convolution

TD-CNN – Sampling Cycle



- The 80th sample of each window is discarded
- $M_{SL,2}$ reduces the leakage of the analog memory

TD-CNN – Computing Cycle



- 60 kernels evaluated temporally in the 80th sample
- MAC output is 1-bit quantized for further classification

Sparsity-Aware Computation (SAC)

$$V_{out} = \sum_{n=1}^{N} W_n V_n + W_0$$

Sparsified quantization

$$|W_n| = C_n \Big/ \sum_{n=1}^{N} C_n \qquad \sum_{n=1}^{N} |W_n| = 1$$

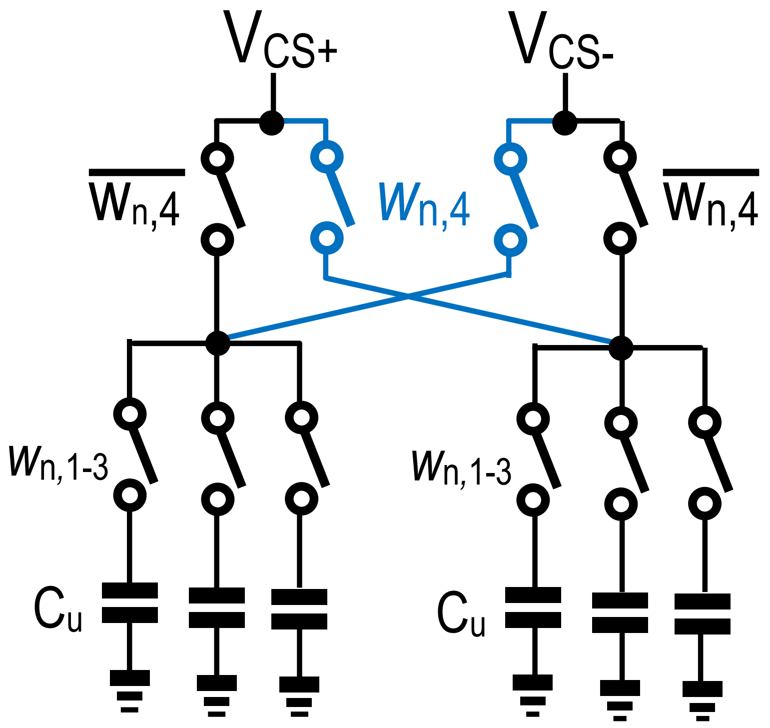
$w_{n,4}$=0 (+ Sign):
$\rightarrow$ Connected

$w_{n,4}$=1 (− Sign):
⤮ Connected

$$w_{n,1\text{-}3} = \begin{cases} 0 \rightarrow \text{Open} \\ 1 \rightarrow \text{Short} \end{cases}$$

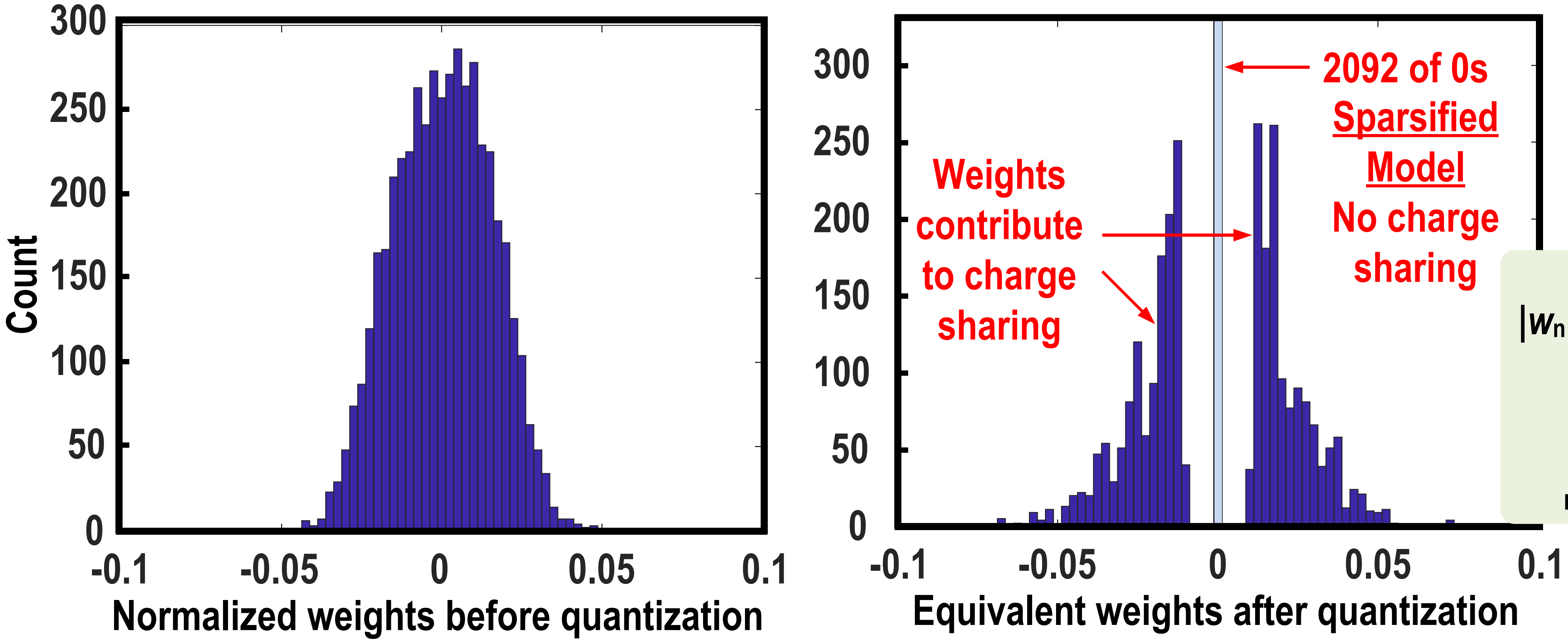
(Sparsity Aware)

Sparsified Weight Capacitors



- Top-plate sampling MAC unit (N=79)
- 0-weight is open to avoid charge sharing
- Increase signal swing of the MAC operation

Sparsified Quantization (SQ)

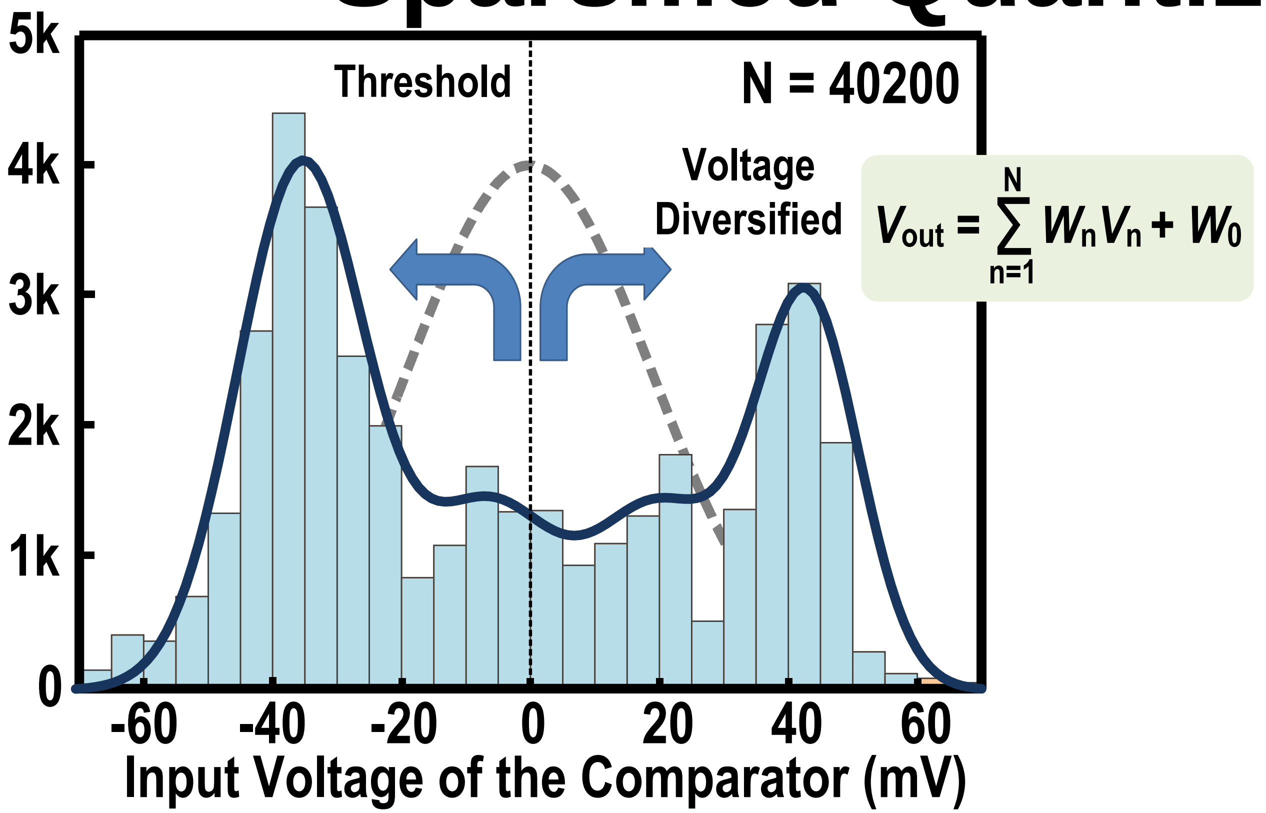


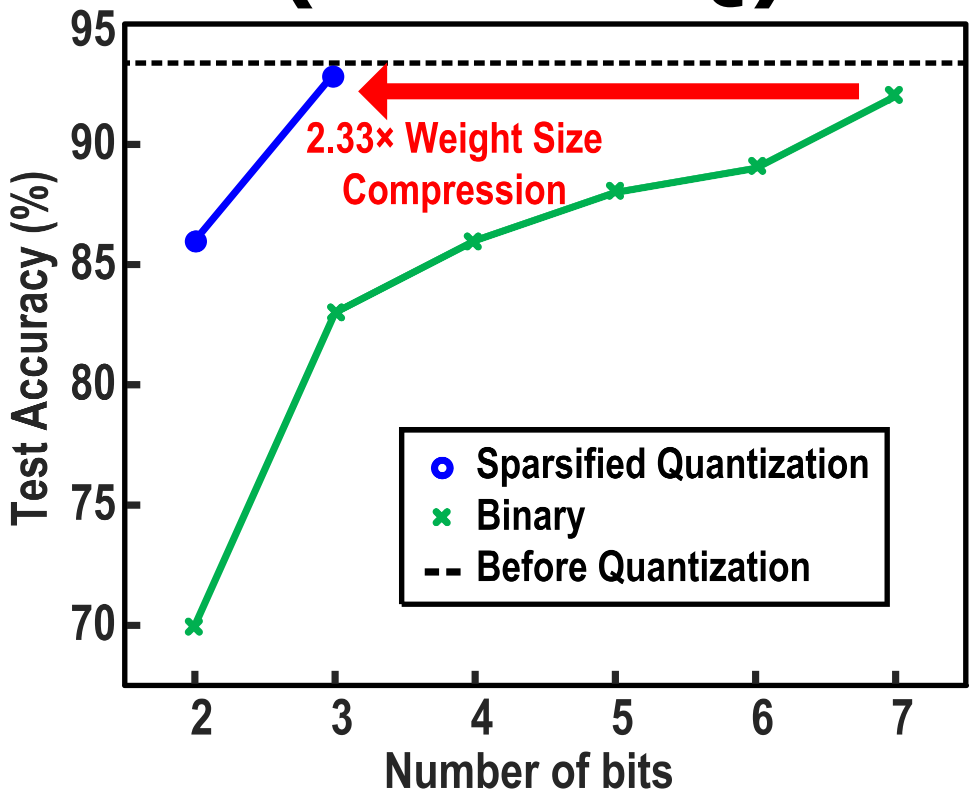
$$|w_n| = C_n \Big/ \sum_{n=1}^{N} C_n$$

$$\sum_{n=1}^{N} |w_n| = 1$$

- Quantization of weights within a kernel are **dependent**
- Quantized weight are "continuously" distributed

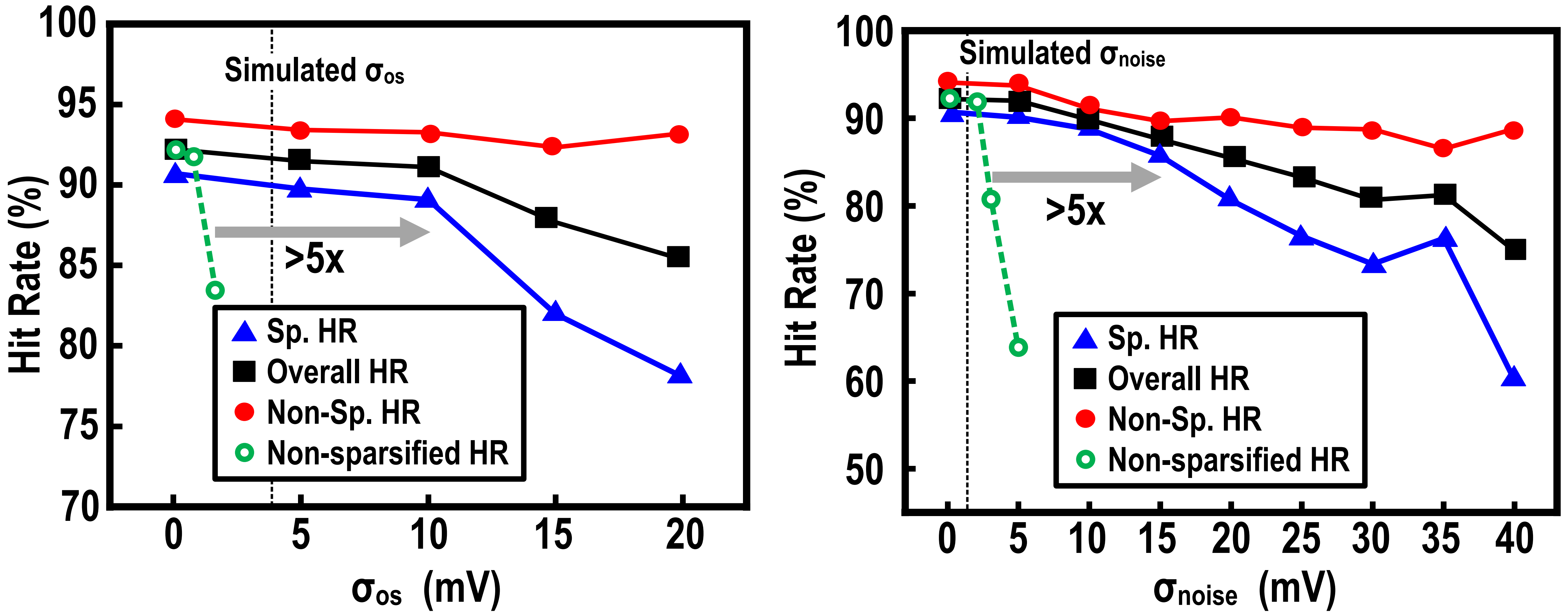
Sparsity-Aware Computation + Sparsified Quantization (SAC+SQ)





- Distribution of the output features are **diversified**
- Compatible to central limit theorem as kernel size is small (N=79)
- Test acc. of 3-bit SAC+SQ is higher than 7-bit binary quantization

Comparator Offset and Noise

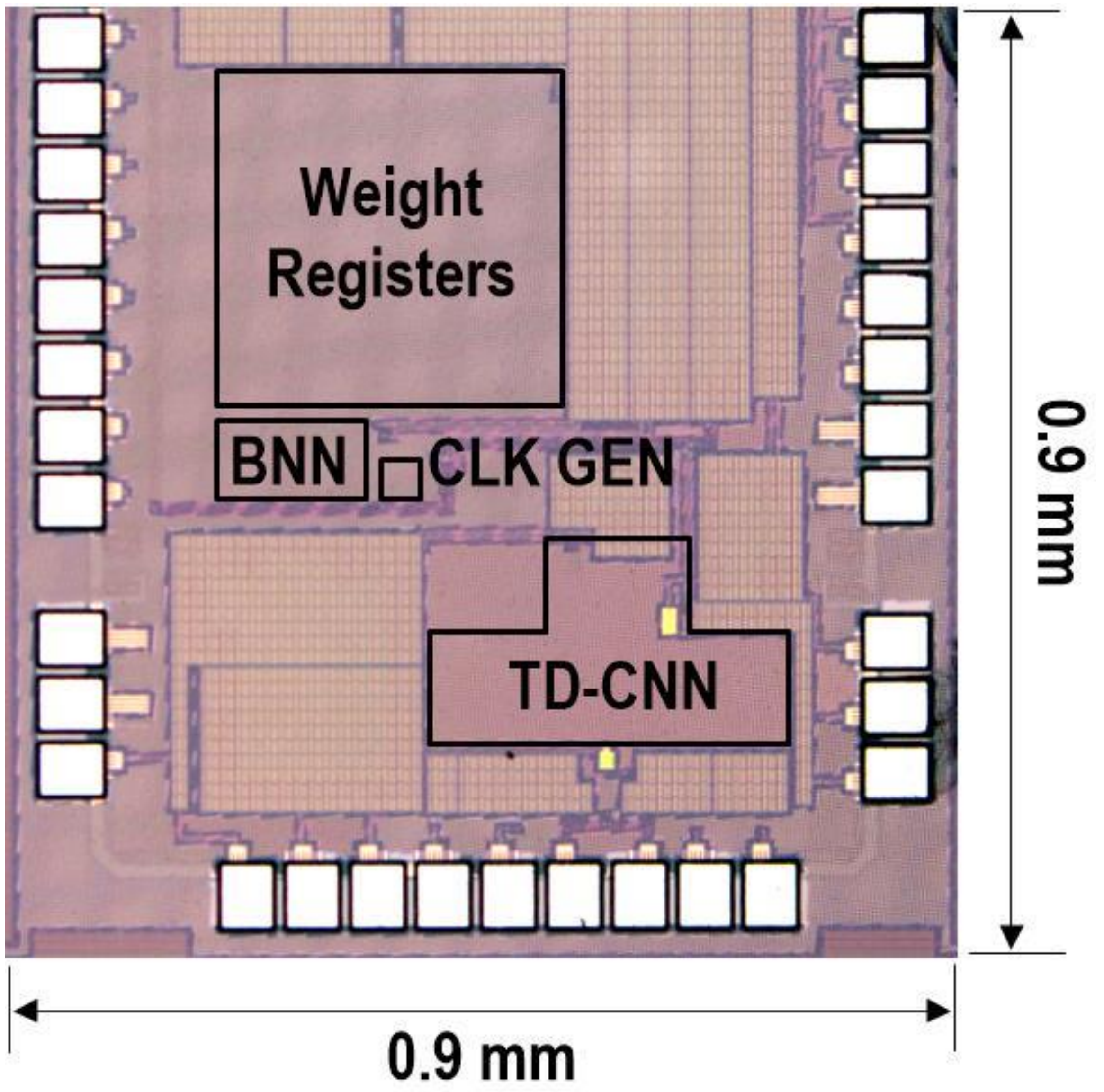


- VAD more robust to offset and noise with sparsified quantization

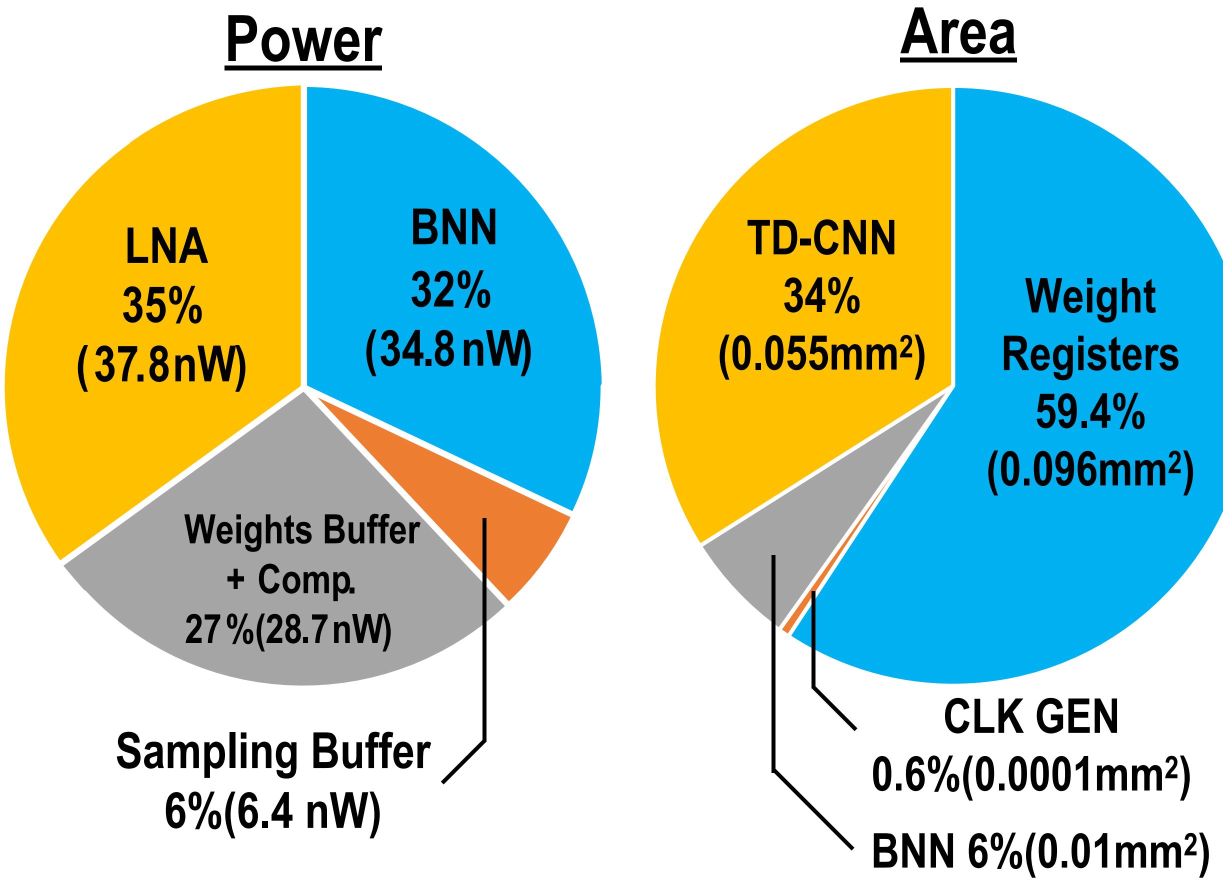
Outline

- Motivation and Prior Arts
- Proposed Analog Voice Activity Detector (VAD)
- **Measurement Results**
- Conclusions

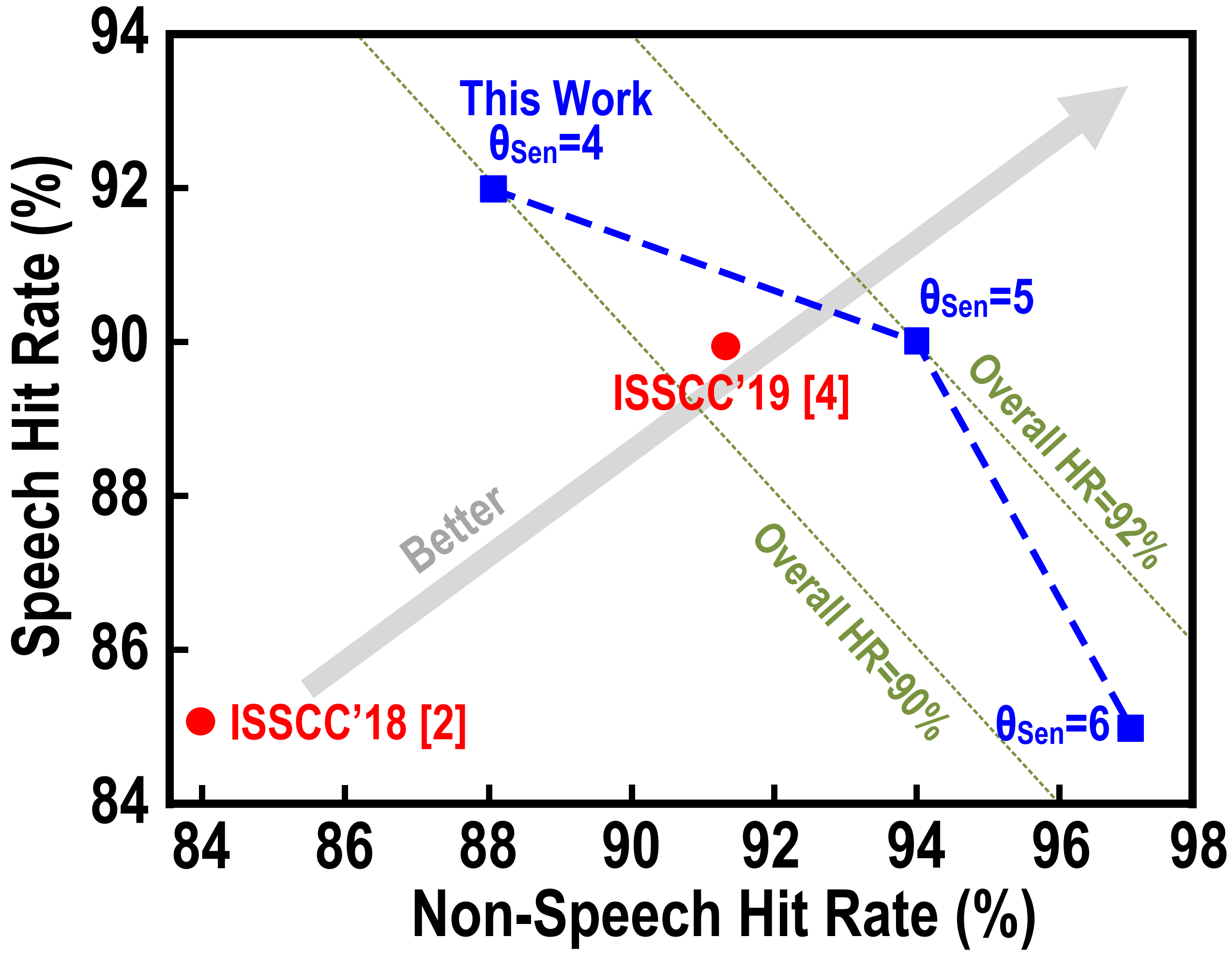
Chip Photo + Power/Area Breakdown

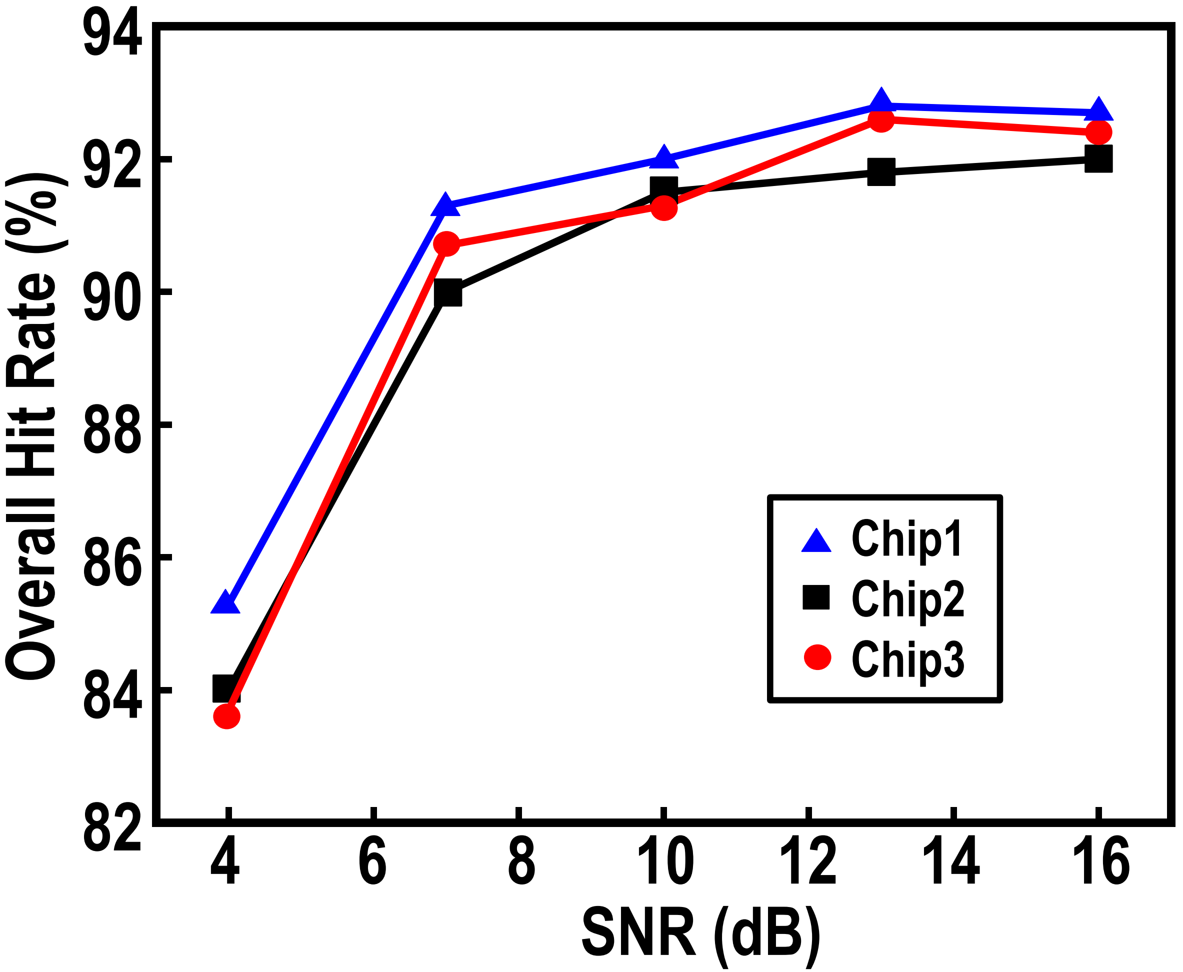


- 28nm CMOS
- Active area $0.8mm^2$



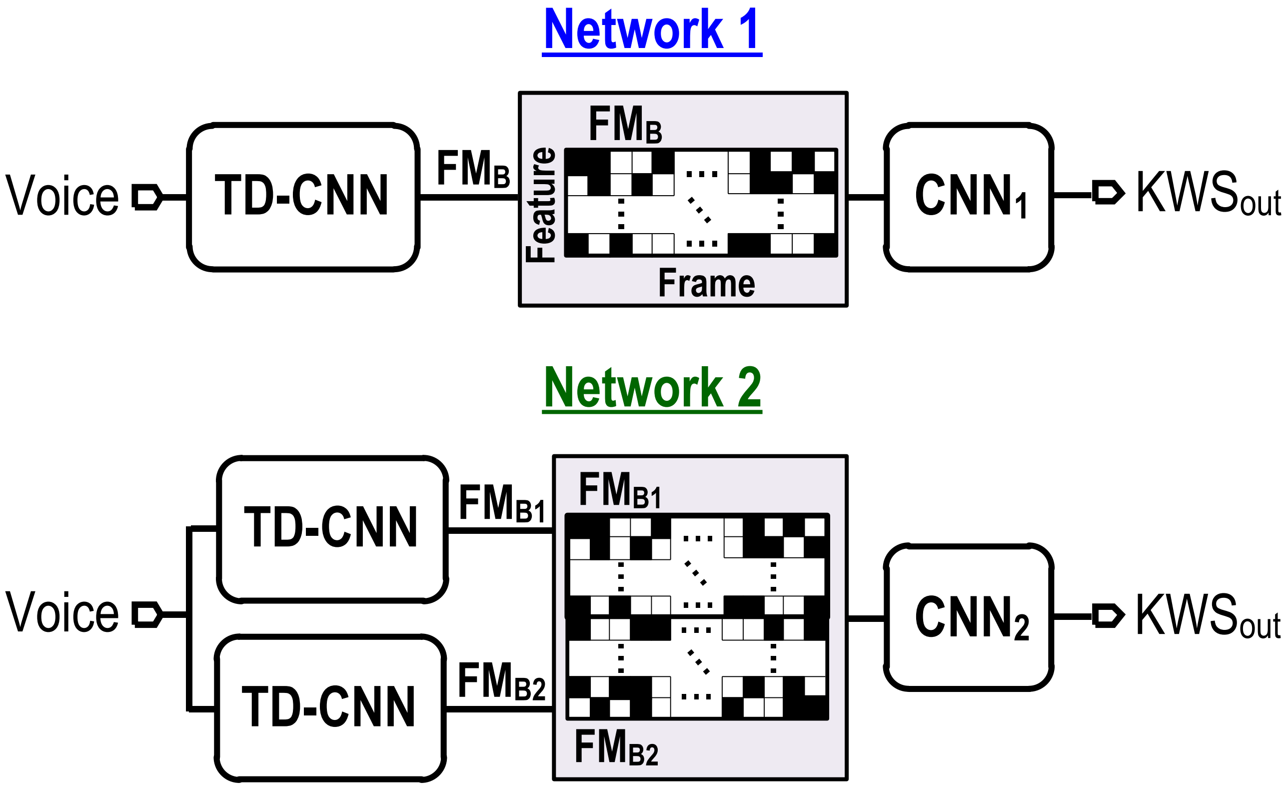
Measurement Results

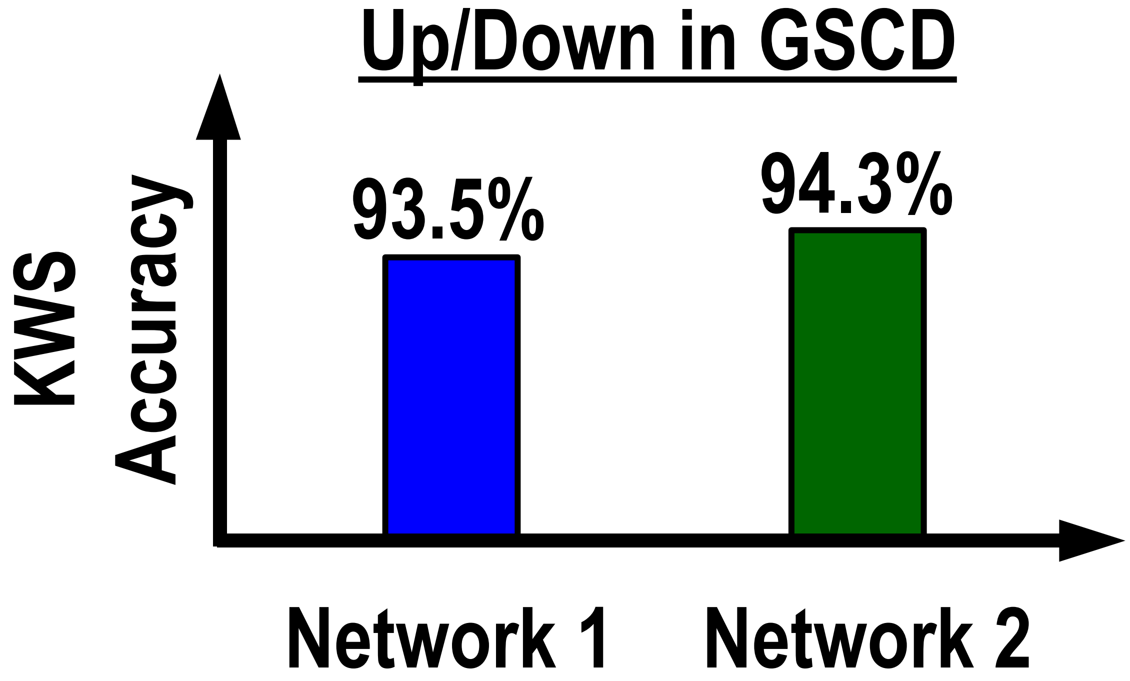




- Overall HR = 92%
- Overall HR > 91% over 3 chips

TD-CNN as Feature Extractor of KWS





- KWS (2 keywords) consists TD-CNN and a 2D-CNN
- Features extracted by TD-CNN are concatenated for the classifier
- 2 TD-CNNs in parallel to increase feature size for better accuracy

Comparison – Feature Extractor

Feature Extractor	This Work	ISSCC'19 [4]	ISSCC'18 [2]	ISSCC'15 [3]
Feature Extraction Topology	TD-CNN Feature Extraction	Time-Interleaved-Mixer-based Frequency Ext.	Analog-to-Event Filter Bank	Analog Filter Bank
Programmable Feature Extractor	**Yes (TD-CNN + Sparsity SC)**	Yes (b-DCT Sequence)	No	No
Channel Number	60 ✓	16-48	16	16
Frequency Range (Hz)	100 to 4k	75 to 4k	100 to 5k	75 to 5k
Feature Type	**Binary Adaptive**	Digital Frequency	Event-based Frequency	Analog Frequency
Power Consumption (nW)	73* ✓	60	380	6000
Area (mm^2)	**0.055** ✓	0.56	1.6	2.56

* *power consumed by the clock buffers for the sampling, charge-sharing buffers and LNA*

Comparison – Voice Activity Detector

Voice Activity Detector	This Work	ISSCC'19 [4]	ISSCC'18 [2]	ISSCC'17 [1]	JSSC'21[5]
Classifier	**TD-CNN + BNN**	Neural Network	BNN	Digital Fixed-Point Deep Neural Network	Analog SNR Decision Rule
Classification Rate (Hz)	100 ✓	2	100	100	31.25
Dataset	TIMIT + NOISEX-92	LibiSpeech + NOISEX-92	AURORA4 + DEMAND	AURORA2	Custom
Accuracy (SP HR/Non-SP HR)	90.1%/94% @ 10dB SNR ✓	91.5%/90% @ 10dB SNR	84%/85% @ 10 dB SNR	90% @ 7 dB SNR	Not Comparable
Power Consumption (nW)	**108** ✓	142	1000	22300	760
Energy/classification (nJ)	**1.08** ✓	73	10	223	24.32
Chip Area (mm^2)	0.8 ✓	17.6**	2.5	2.1	0.14 (Active)
Technology	28 nm CMOS	180 nm CMOS	180 nm CMOS	65 nm CMOS	180 nm CMOS

** *Area including an audio compressor and a processor*

Outline

- Motivation and Prior Arts
- Proposed Analog Voice Activity Detector (VAD)
- Measurement Results
- **Conclusions**

Conclusion

A 108nW 0.8mm^2 analog voice activity detector is implemented in 28nm CMOS

TD-CNN as Feature Extractor	Reduce area and power consumption Reduce A/D conversions More reconfigurable
Sparsity-Aware Computation	Increase signal swing of the MAC operation
Sparsified Quantization	Diversify the output to desensitize the output from mismatches and noise

Acknowledgments

- Multi-Year Research Grant of University of Macau
- Macao Science and Technology Development Fund (FDCT)