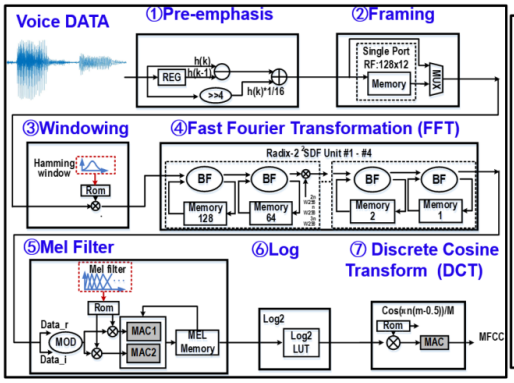
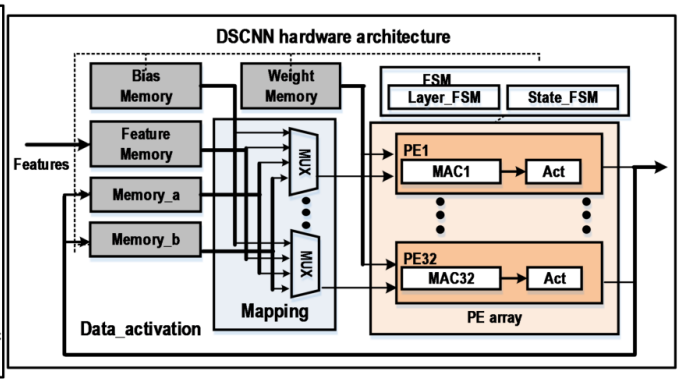
A 510nW, 0.41V Low-Memory, Low-Computation Keyword Spotting Chip using Serial FFT based MFCC and Binarized Depthwise Separable Convolutional Neural Network in 28nm CMOS

Keyword spotting system: front end+ back end
– Front end: LNA/filter + ADC+ feature extraction
– Back end: Classifier (neural network)



MFCC feature ext.

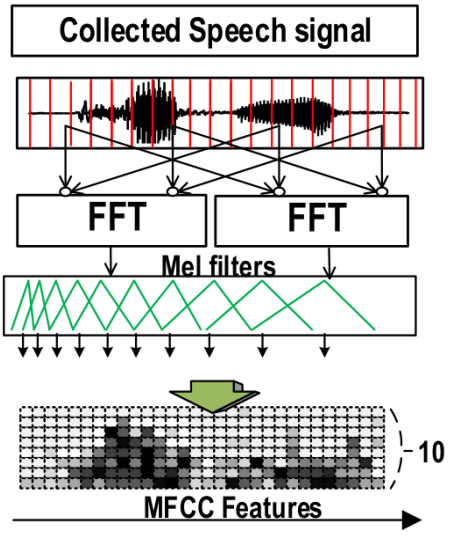


DSCNN classifier

MFCC Feature Extraction Architecture

Mel frequency cepstrum coefficient (MFCC): to extract features from speech signal.

Input: 16bits, output: 10 dimensions of 8-bit data.

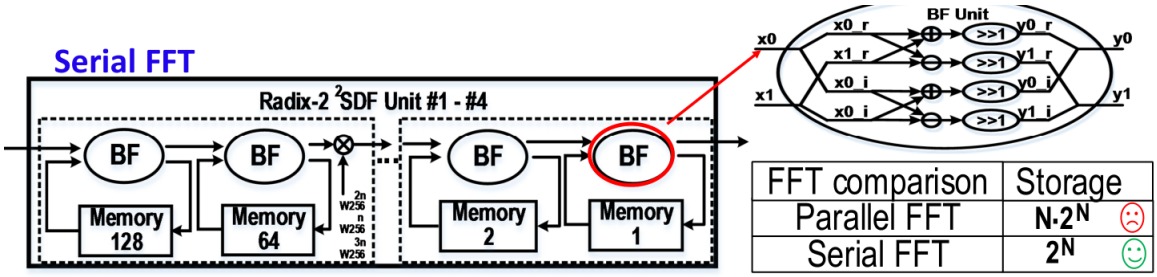


- Google speech command dataset
- 8kHz sampling rate
- 32ms/frame with a 16ms step
- 256-point/frame for FFT

- FFT: to transfer voice signals from the time domain to the frequency domain
- FFT consumes most of the power

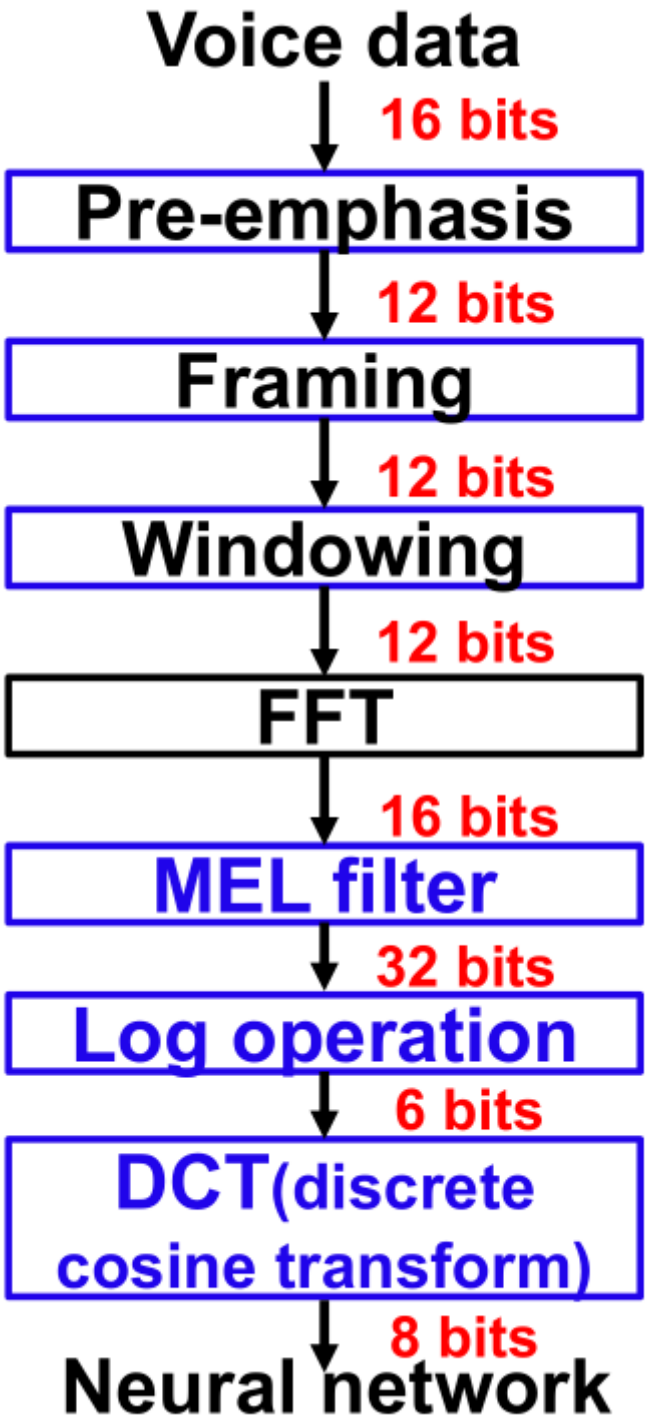
FFT Architecture in MFCC

Main idea: using **serial FFT**

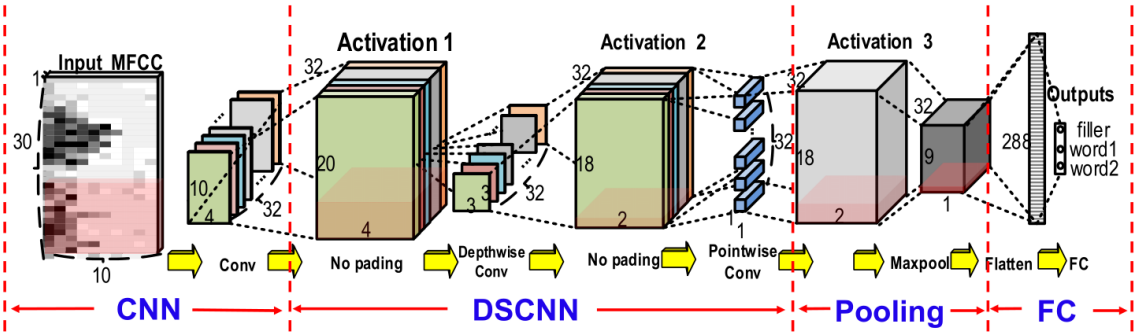


FFT comparison	Storage
Parallel FFT	$N \cdot 2^N$ ☹
Serial FFT	2^N ☺

Precision optimization per block



DSCNN Architecture



- One or two keyword detection
- 4-layer network
- Binarized NN:
  - Input features: 8b
  - All weights: 1b
  - All activations: 1b

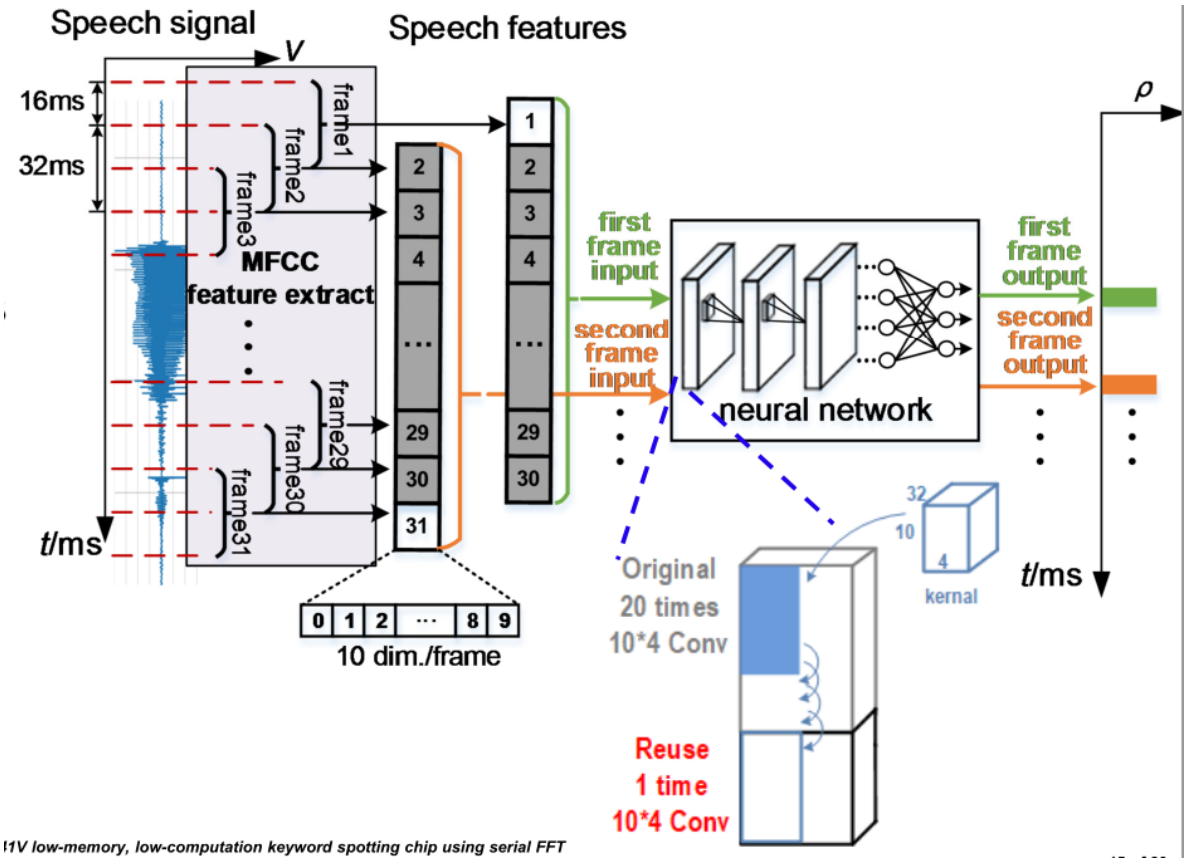
Weights	Computations
3,456 bits	202,400 8-bit MACs (Input layer matters) 48,096 1-bit MACs

Memory/Computation Reuse

Multiple concatenated frames enter a classifier as input

Input layer: 30 frames×10 dimensions × 8 bits;

Our idea: reuse of frame computations and activation data storage over inferences



11V low-memory, low-computation keyword spotting chip using serial FFT

Method-1: frame computation reuse

Method-2: Hardware-aware NN algorithm modification

Near-threshold voltage circuit design

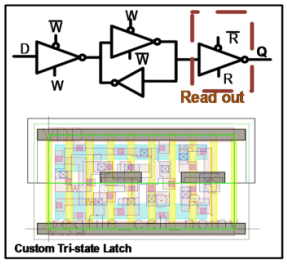
40kHz --> Leakage dominates! Especially in near-threshold voltage

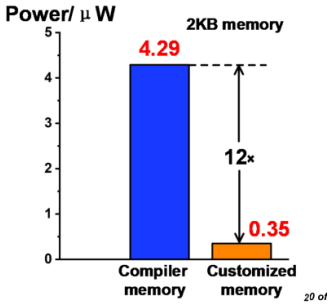
☐ **Memory design is a challenge**
- Four dual-port & five single-port memory
- SRAM leakage power is dominant

☐ **SRAM provided by a foundry is not for NTV**
- Supply voltage of 0.9V. Needs a separate power domain
- The leakage of 2kB SRAM: 4.29μW >> 1μW

☐ **Custom NTV memory**
- Ultra-low leakage EHVT cell
- Latch-like memory for NTV
- Leakage power reduces by 12×

Memory blocks		
Size/bit		Type
MFCC	128*32	Single port
	128*16	Dual port
	128*16	Single port
	64*32	Dual port
	32*32	Dual port
NN	110*8	Dual port
	99*32	Single port
	20*32	Single port
	16*32	Single port





14.1: A 510nW, 0.41V low-memory, low-computation keyword spotting chip using serial FFT based MFCC and binarized depthwise separable convolutional neural network in 28nm CMOS

	ISSCC 2017 [1]	VLSI 2018 [2]	VLSI 2019 [3]	ESSCIRC 2018 [4]	This work
Tech.	40 nm	28 nm	65 nm	65 nm	28 nm
Algorithm	DNN	CNN	LSTM	LSTM	DSCNN
Voltage	0.63-0.9V	0.57-0.9V	0.6V	0.575V	0.41V
Memory	270kB	52kB	65kB	32kB	2kB
Core Size	7.1mm²	1.29mm²	2.56mm²	1.04mm²	0.23mm²
Frequency	1.9MHz	2.5MHz	250kHz	250kHz	40kHz
Latency	6.5ms	0.5-25ms	16ms	16ms	64ms
Keyword Num	10 words	1 word	10 words	4 words	1~2 words
Power	288μW	141μW	16.1μW*	5μW**	0.51μW
Dataset	NA	TIDIGIT	GSCD	NA	GSCD
Accuracy	NA	96%	90.87%	91.2%	98@1 word; 94.6%@2 words

Smallest voltage
Smallest memory
Smallest area
Lowest freq.
Lowest power

* Only for digital classifiers
** MFCC feature extraction was excluded