东南大学国家ASIC中心在ISSCC 2020发表AI芯片亮点论文

发布者：李震 发布时间：2020-03-03 浏览次数： 5048

【东大新闻网3月3日电】（通讯员 单伟伟）2月16日-20日，第67届IEEE国际固态电路峰会（ISSCC 2020）于美国旧金山召开。这是世界上规模最大、最权威、水平最高的固态电路国际会议，被称为集成电路行业的芯片奥林匹克大会。此次ISSCC的主题是“Integrated Circuits Powering the AI ERA”，彰显了人工智能时代集成电路的重要地位。

本次会议新设了两个机器学习分会，所有的四个主题演讲都与人工智能有关。两个机器学习分会分别以高性能和低功耗为主题。其中“Low-Power Machine Learning”分会(Session 14)面向的是可移动终端等对功耗有极致要求的领域，共收录3篇论文，都来自于中国大陆，这展示了中国在AI芯片设计的学术领域已处于世界领先水平。亮点论文来自东南大学电子科学与工程学院国家ASIC中心的单伟伟、杨军、时龙兴团队。另两篇论文来自清华大学刘勇攀教授团队和湃方科技。

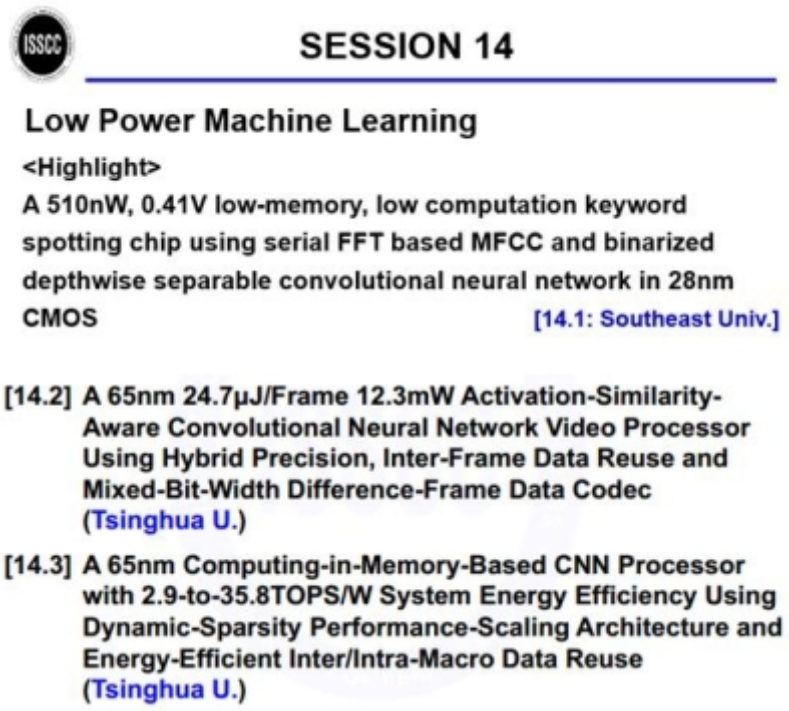

图1低功耗机器学习分会的3篇论文

东南大学的亮点论文是关于语音关键词唤醒的低功耗AI加速器的论文：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。该语音唤醒智能芯片从算法、芯片架构和电路三个层次统筹优化，如下图所示，算法级采用基于串行FFT的MFCC特征提取和深度可分离卷积神经网络，极大降低了计算量和存储量；架构级提出了语音数据的逐帧数据复用方法。这两个级别联合使得芯片可工作在极低频率40kHz下，进而促成了全芯片的近阈值设计和超低漏电的定制存储器电路，最终现了史上功耗最低的关键词唤醒电路，功耗仅为510纳瓦。

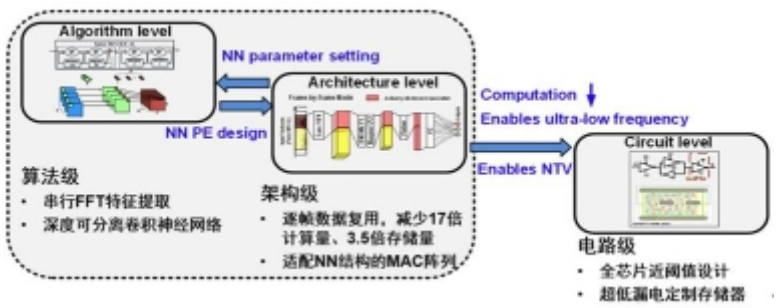

图2语音唤醒智能芯片的算法、芯片架构和电路三个层次统筹优化

独特的创新技术如下：

（1）提出并实现了基于串行FFT的梅尔频率倒谱系数（MFCC）特征提取电路，同时用混合量化逐层降低硬件实现代价。FFT是特征提取中计算最复杂、功耗最大的模块，与传统并行FFT相比，提出的串行FFT电路的存储量降低8×，功耗降低11×；

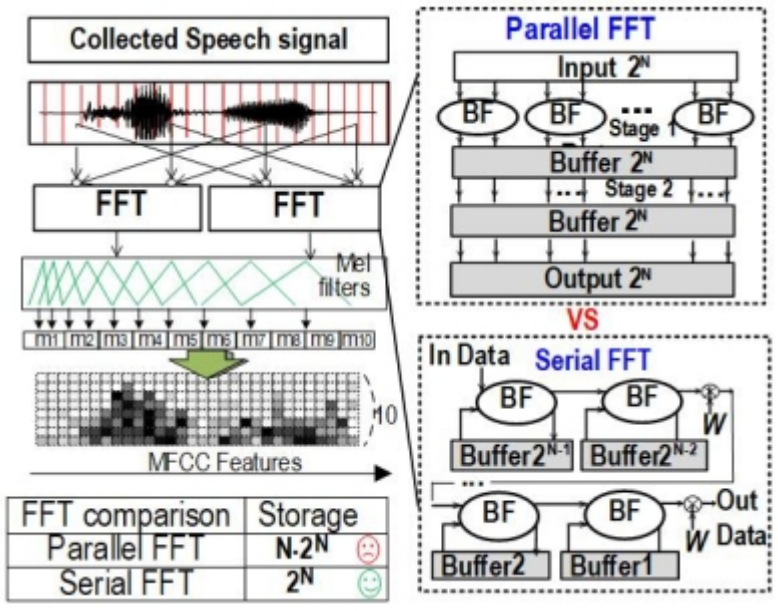

图3 MFCC特征提取电路结构及亮点

（2）深度可分离卷积神经网络的二值化轻量级神经网络，与CNN相比存储量和计算量均降低7×；基于此设计了契合算法的神经网络硬件架构，由计算单元（PE）阵列（含32个乘累加MAC单元）、存储模块、数据映射模块及控制状态机组成；

（3）提出了逐帧数据复用技术，利用语音应用中相邻两帧输入数据的计算存在大量重复导致的卷积计算中存在大量重复计算的特点，对神经网络中的数据的存储和计算量进行压缩，使得计算量降低17.4×，中间存储量减少3.5×；

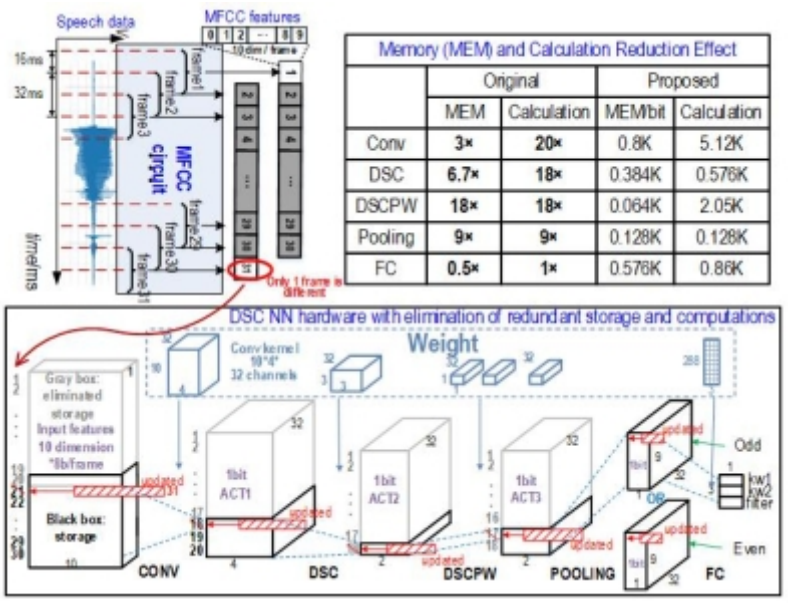

图4深度可分离卷积神经网络架构及逐帧数据复用技术

（4）全芯片近阈值设计与定制的极低漏电存储器，再次降低功耗。前面算法级和架构级的双重优化，使得整体神经网络加速器仅需640个周期就能完成一轮推理，在16ms的帧间隔内完成即可，因此工作频率仅需40kHz下，这就促成了全芯片可采用近阈值设计。近阈值设计的难点有二：存储器和漏电控制。我们定制了能工作在低电压下、且具有低漏电的latch型存储器，实现神经网络与MFCC所需的片上多块、多类型的小容量存储。定制Memory比工艺厂提供的SRAM编译器生成的同等大小的SRAM模块的漏电低了12倍，且可与其他数字电路一起工作在0.41V。

这些技术应用在T28nm的神经网络加速器，实现了史上最低的关键词唤醒电路芯片,0.41V电压下整体功耗仅0.51uW，比国际同类研究降低了10到564倍，见对比表格。

	ISSCC2017[2]	VLSI2018[3]	VLSI2019[4]	ESSCIRC18[5]	This work
Tech.	40 nm	28 nm	65 nm	65 nm	28 nm
Algorithm	DNN	CNN	LSTM	LSTM	DSCNN
Voltage	0.63-0.9 V	0.57-0.9 V	0.6 V	0.575 V	0.41 V
Memory	270 KB	52 KB	65 KB	32 KB	2 KB
Core Size	7.1 mm^2	1.29 mm^2	2.56 mm^2	1.04 mm^2	0.23 mm^2
Frequency	1.9 MHz	2.5 MHz	250 KHz	250 KHz	40 KHz
Latency	6.5 ms	0.5-25 ms	16 ms	16 ms	16 ms
Keyword Num	10 words	1 word	10 words	4 words	2 words
Power	288 μW	141 μW	16.1 μW*	5 μW**	0.51 μW
Dataset	NA	TIDIGIT	GSCD	NA	GSCD
Accuracy	NA	96%	90.87%	91.2%	93.6%

* 16.1 μW refers to the power of digital KWS system in [4], not including Analog Front-End.
** Ref. [5] does not have MFCC circuit.

供稿：电子科学与工程学院

（责任编辑：翟梦杰 审核：李小男）