



多物理域异构融合芯片架构

陶耀宇
北京大学

一、个人简介



陶耀宇

北京大学

博雅青年学者

国家优青（海外）

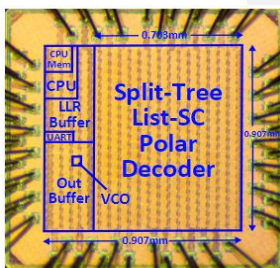
学校或单位	学习或任职
上海交通大学	电子与计算机工程（信号处理）学士
斯坦福大学	电子工程（模拟电路）硕士
美国甲骨文公司	VLSI实验室 模拟电路工程师
美国高通公司	无线研发部 高级工程师
美国密歇根大学	电子工程（芯片体系结构）博士
美国高通公司	无线研发部 主任研究科学家
北京大学	研究员、博雅青年学者

- 长期从事基于后摩尔先进器件的**多域异构融合芯片架构**与电路系统研究
- 研究成果发表**国际顶尖期刊与会议论文30余篇**（其中一作/通讯20余篇），包括**Nature Electronics、Nature Communications、芯片架构两大会议IEEE MICRO/HPCA、集成电路最高期刊IEEE JSSC、全球“芯片奥林匹克”会议IEEE ISSCC**，微电子器件与电路两大会议**IEEE IEDM/VLSI**、通信最高会议**IEEE Globecom**、FPGA最高会议**IEEE FPGA**等，**作为主要发明人申请/获批十余项中、美专利**
- 获2025年华为青年人才支持计划、2024中国电子学会青年年会专题报告奖、2024年中关村论坛“智能未来”分论坛特邀嘉宾、**2023年华为计算产品线优秀技术合作奖**、2023年Wiley Open Science Excellent Author、2022年北京智源人工智能青年科学家、**2021年世界通信大会（IEEE Globecom）最佳论文奖**、2019年至2021年**连续3年获高通技术明星奖**（Qualcomm QualStar Awards）、2019年大规模集成电路国际学术会议（IEEE VLSI）最佳论文提名奖、2018年Rackham科学研究奖、2013年电路与系统国际学术会议（IEEE ISCAS）最佳论文提名奖等荣誉
- 主持**国家自然科学基金委优青（海外）项目**、科技部国家重点研发计划项目“**物态调控**”专项（课题负责人）、科技部国家重点研发计划项目“**智能电网**”重大专项（子课题负责人）、**国家自然科学基金委重大研究计划（子课题负责人）**、北京市自然基金委非共识项目、华为计算产品线技术合作项目、纵慧芯光技术合作项目、武汉东湖-北大武汉人工智能研究院项目等，累计主持经费数千万元

二、科学研究

• 代表性科研成果

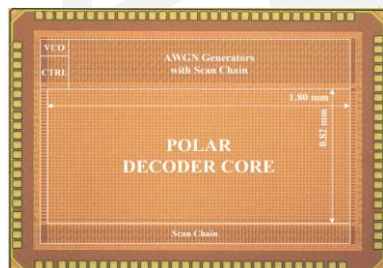
传统电压域计算芯片架构



SCL Polar体系结构

VLSI, JSSC

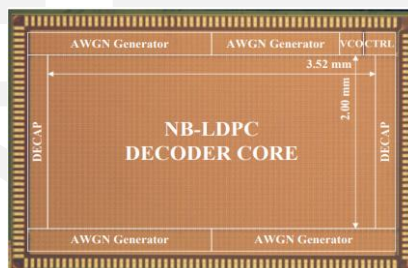
(均为一作, 最佳论文奖提名)



BP Polar体系结构

VLSI

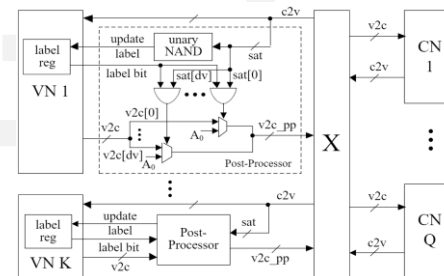
(共一)



NB-LDPC体系结构

ISCASx3, ISSCC, JSSC

(一作3篇, 最佳论文奖提名)

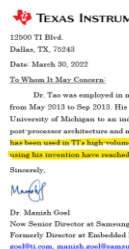


LDPC Post-Processor体系结构

TCAS-I, US Patent

(均为一作)

- 首个极化码并行分树架构芯片, 吞吐率达到当时最高的3.25Gbps
 - 发表JSSCx1、VLSIx2, 获VLSI 2019最佳论文奖提名
- 创新非二进制奇偶校验体系结构芯片, 解决了长期困扰LDPC实用化的“误码率墙”问题
 - 发表JSSCx1、ISSCCx1、ISCASx3、TCAS-Ix1、美国发明专利等, 获ISCAS 2013最佳论文提名、2019年高通技术明星奖 (落地产品实用)



相关专利应用在德州仪器芯片产品中, 累计量产超 1.5 亿颗



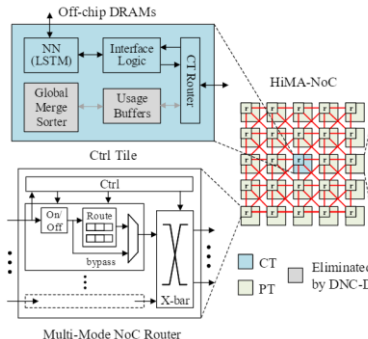
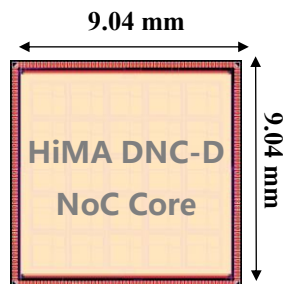
连续3年获高通技术明星奖

(Qualcomm QualStar Award)

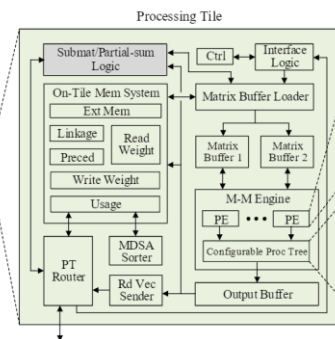
二、科学研究

• 代表性科研成果

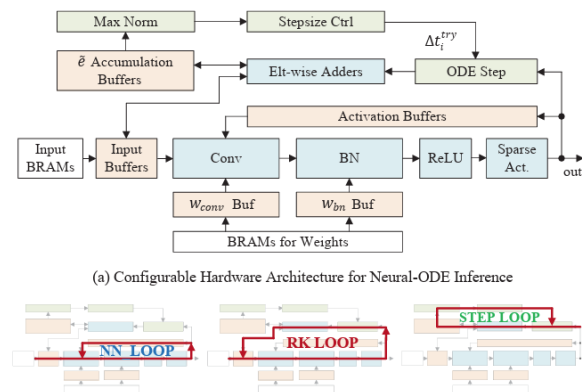
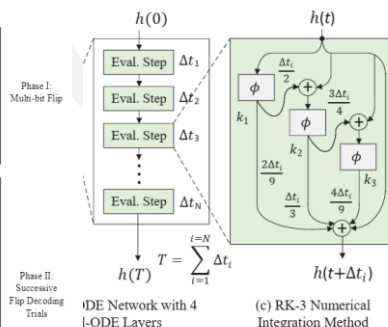
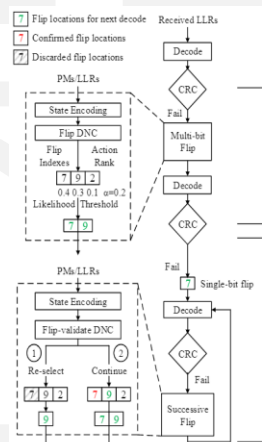
传统电压域计算芯片架构



HiMA DNC体系结构
MICRO (一作)

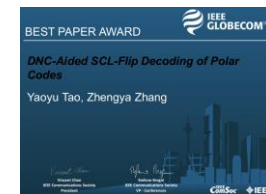


DNC辅助解码体系结构
Globecom (一作)(最佳论文奖)



Neural ODE体系结构
HPCA, FPGA (一作、通讯)

- 创新分布式DNC架构比当时最快AI芯片加速39.1X，首个DNC辅助解码架构实现54%延时降低
 - 发表MICROx1、Globecomx1，获2021 Globecom最佳论文奖、2020和2021年高通技术明星奖 (落地产品实用)
- 提前终止机制Neural ODE体系结构比当时最快AI芯片加速2.6X，解决Neural ODE高延迟问题
 - 发表HPCAx1、FPGAx1

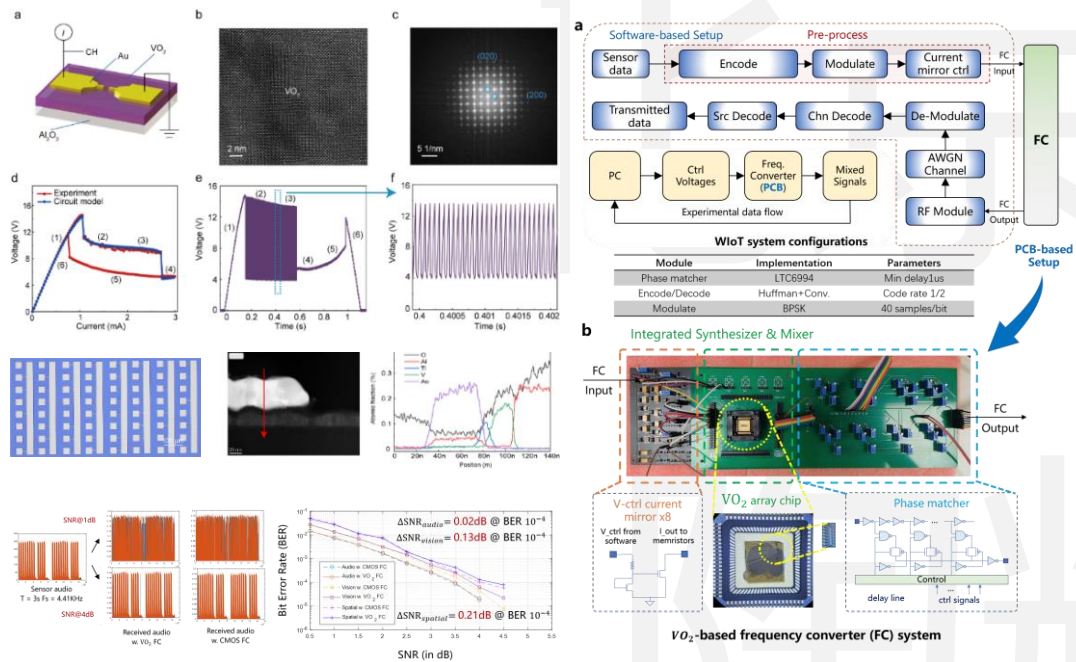


IEEE Globecom Best Paper Award
通信两大顶会Globecom最佳论文奖

二、科学研究

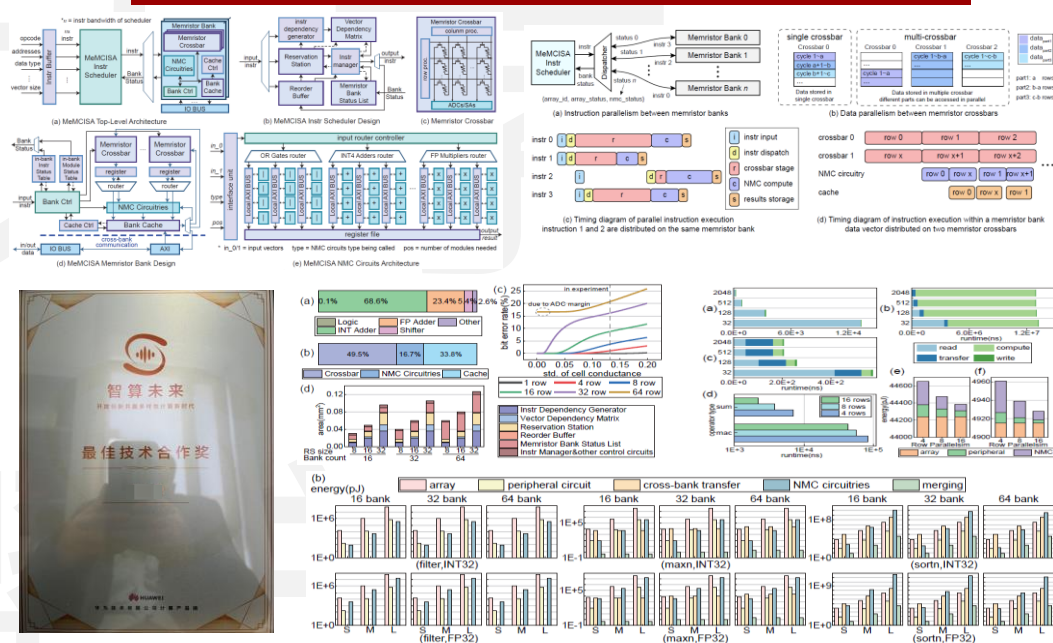
• 代表性科研成果

频率域处理架构/电路



- 首次提出利用电流控VO₂忆阻器阵列芯片原位生成不同频率的精准信号并可编程实现多路信号混频功能
- 端到端的WIoT性能测试能够降低1.45x~1.94x的能耗
- 发表于Nature Communications 2024 (通讯作者)

电流域存内/电压域近存异构融合架构

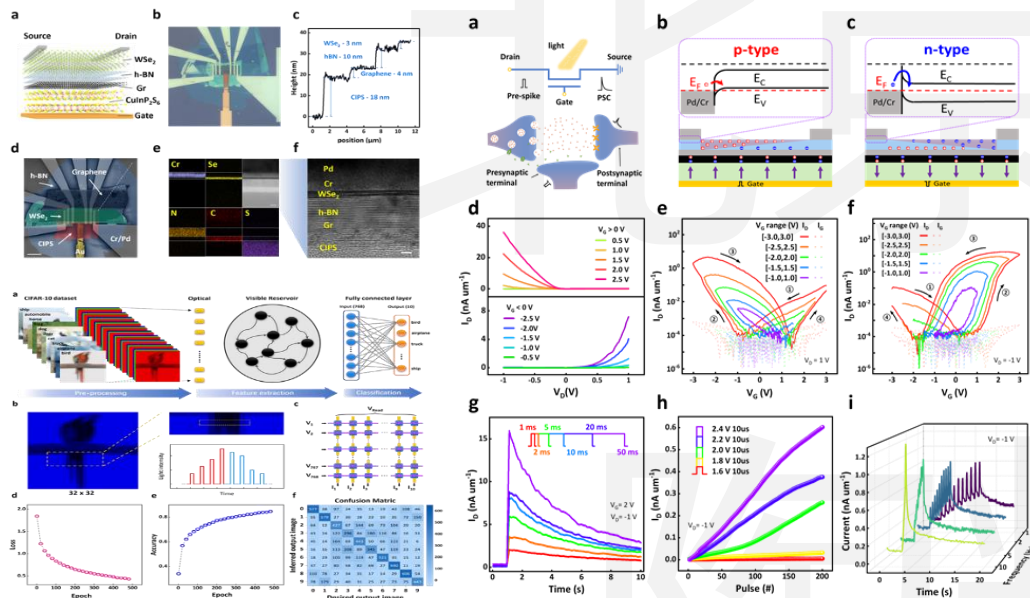


- 提出忆阻器存内、近存融合计算核架构，4类向量指令组成忆阻器存算一体指令集，端到端异构多核集成架构
- 华为计算产品线实际场景落地，SQL加速高达16.6X，获2023年度华为优秀技术合作奖
- 发表于IEEE/ACM MICRO 2024 (通讯作者)

二、科学研究

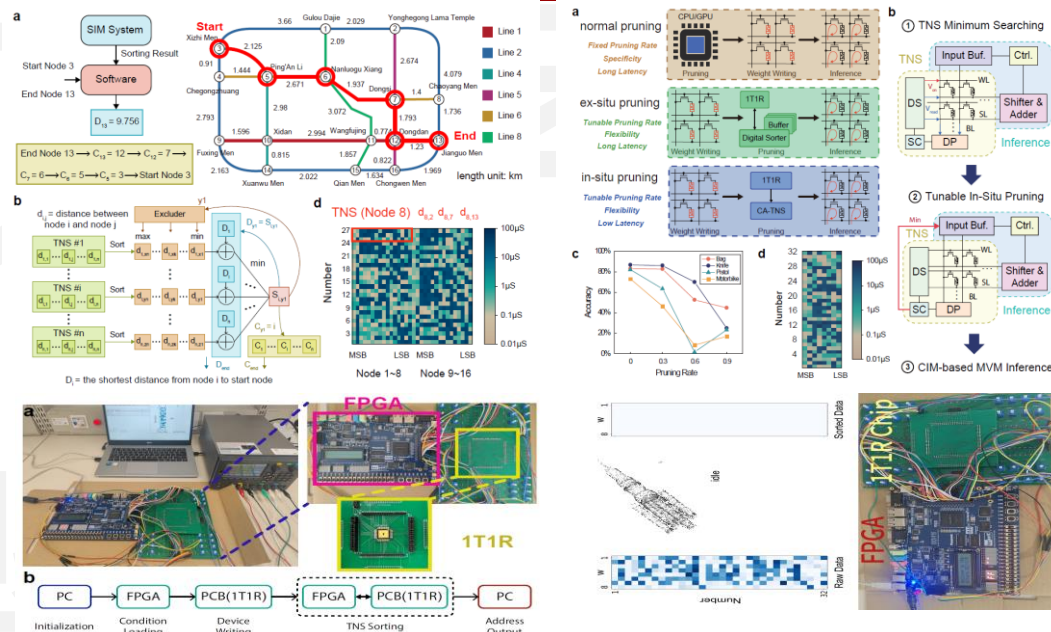
• 代表性科研成果

光学域信号处理架构



- 三端范德华 (vdW) 异质结的场效应晶体管 (FET)可执行彩色光学传感功能
- 光学存储器神经网络 (ORNN) 和由该光学突触装置组成的可见光通信系统 (VLC) 感存算一体架构
- 发表于Advanced Materials 2024 (通讯作者)

时间域大数据排序架构

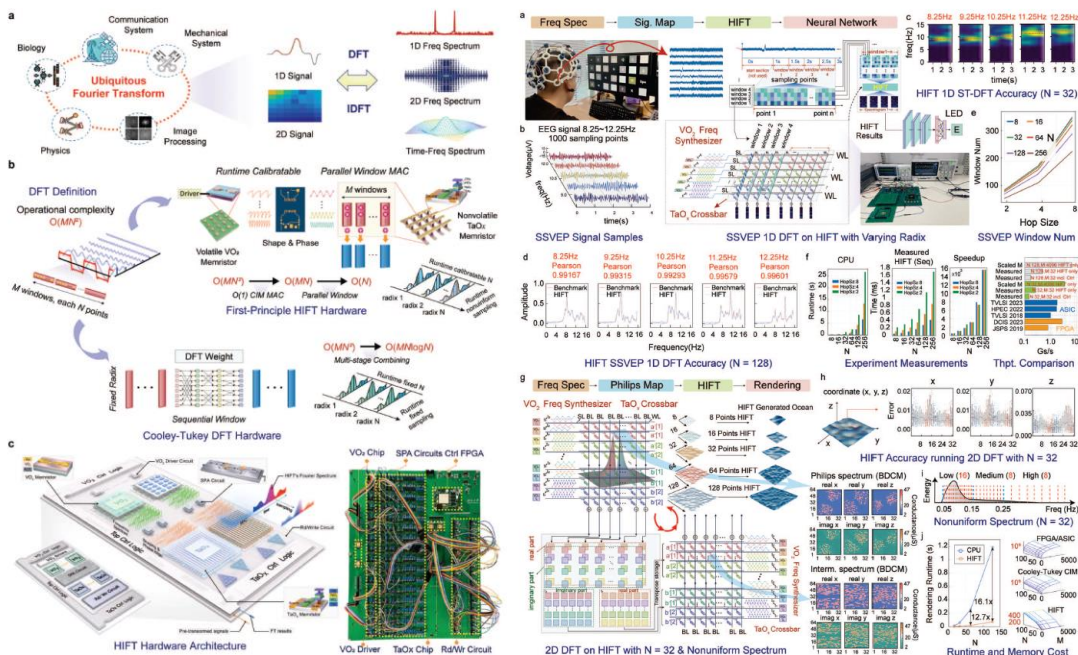


- 提出忆位读取新型阵列结构，节点跳跃式阵列操作方法、Multi-bank、Bit-Slice与Multi-level三种并行策略
- 具备与现有存内计算矩阵操作兼容的能力，实现3.3X~7.7X的加速、6.23X~183.5X的能效提升
- Nature Electronics 2025 (通讯作者)

二、科学研究

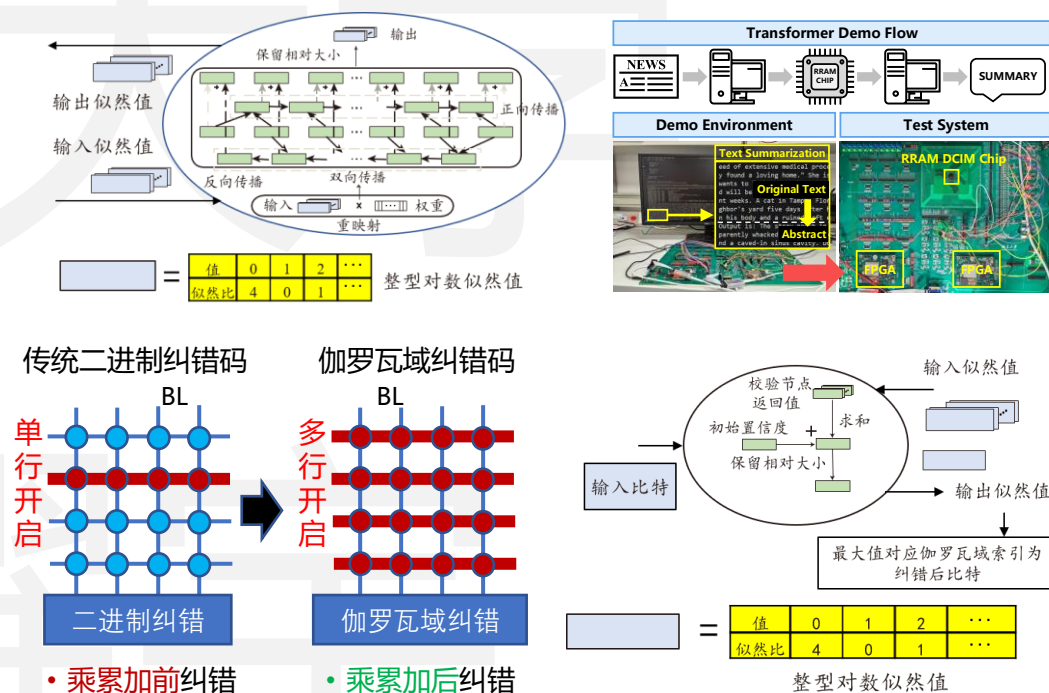
• 代表性科研成果

电流域/频率域融合FFT架构



- 首次提出利用**异质集成的易失与非易失器件**完成傅里叶变换的复杂计算
- 端到端的脑机信号处理性能测试展示速度、功耗提升效果
- **Nature Electronics 2025 (共一/通讯作者)**

电流域存算与纠错芯片架构



- **乘累加前纠错**
- **乘累加后纠错**
- 首次提出利用**非二进制LDPC编码**对存内计算乘累加后结果进行高精度纠错
- **多项发明专利 2024/2025 (第一发明人/主要发明人)**

二、工作影响力

芯片架构工作被多位领域内重要专家在高水平论文作为代表性工作引用

Lu, Y., Yang, Z., Tao, Y., Cai, L., Zhang, T., Yan, L., Huang, R. and Yang, Y., 2025. Energy-Efficient Online Training with In Situ Parallel Computing on Electrochemical Memory Arrays. *Advanced Intelligent Systems*, p.2401068.

北京大学黄如教授（中国科学院院士）在论文中引用候选人的工作为PCM存算芯片代表性工作

Medard M, et al., "Multi-code multi-rate universal maximum likelihood decoder using grand," IEEE 47th European Solid State Circuits Conference (ESSCIRC) 2021 Sep 13 (pp. 239-246).

美国麻省理工学院 Muriel Medard 教授（美国工程院院士）在论文中引用候选人工作作为极化码芯片代表性工作

A Universal Maximum Likelihood GRAND Decoder in 40nm CMOS

美国波士顿大学Rabia Yazicigil 教授（MIT科技评论 35 岁以下科技创新 35 人）将候选人工作作为高速极化码芯片的代表性工作

Arikan, E, et al., "A high throughput energy efficient implementation of successive cancellation decoder for polar codes using combinational logic," IEEE Transactions on Circuits and Systems I: Regular Papers, 2016 63(3), pp.436-447.

D. Markovic, et al, "A 2.267-Gb/s, 93.7-pJ/bit non-binary LDPC decoder with logarithmic quantization and dual-decoding algorithm scheme for storage applications," IEEE journal of solid-state circuits. 2018 May 22;53(8):2378-88.

primarily, and so on. Therefore, the implementation of online training in edge-terminal devices, which are capable of adjusting weights in the pretrained ANN accommodating to varying environments, can effectively improve the adaptivity and robustness of AI models.^{[6]-[8]} This is particularly important in the era of IoT, where IoT devices are expected to work properly in a diverse range of applications. However, high-performance online training based on traditional von Neumann architectures incurs high energy overhead and hardware costs,^{[9]-[11]} which are not suitable to be implemented in edge-terminal devices. Therefore, it poses a critical demand for novel computational architectures for high-performance ANN training at edge-terminal devices.

To solve this problem, researchers took inspiration from the human brain and proposed non-von Neumann computing

terminal devices have been proposed, while the core computing units are mostly based on emerging non-volatile memory (NVM), such as resistive random-access memory (RRAM),^{[12]-[14]} phase-change memory (PCM),^{[15]-[18]} magnetoresistive random-access memory (MRAM),^{[19]-[24]} ferroelectric field-effect transistor (FeFET),^{[25]-[28]} and electrochemical random-access memory (ECRAM).^{[29]-[31]} Based on the crossbar array built by the above NVMs, the matrix-vector multiplication (MVM), which are the dominant and most expensive computing kernel in ANN, can be physically accelerated by exploiting Ohm's law and Kirchhoff's law.^{[32]-[34]} In recent years, CMOS amplified computing-in-memory (CIM) architecture has achieved excellent results in the field of deep learning (DL), mainly due to their ability to achieve high-performance inference for a variety of ANN models with very low time and energy overheads.^{[35]-[37]}

to help guide error correction, which is called soft-decoding.

Since the 1950s, by co-designing codes with particular structure and corresponding, distinct code-specific decoders, a wide variety of codes have been developed and deployed for error correction. Examples of codes with dedicated hardware decoders include Reed-Muller (RM) codes [3], [4] with Majority Logic Decoding, Reed-Solomon codes [5] and Berlekamp-Massey decoding [6], [7]. A recent example of an important code structure that has a dedicated soft-decoding decoder is CRC-Assisted Polar (CA-Polar) codes, which will be used for all control channel communications in 5G NR. Radio, with CRC-Assisted Successive Cancellation List (CA-SCL) decoding [8]-[10]. Efficient hardware realization for all of these decoders have been reported, e.g. [11]-[14], but, as

in this work. In contrast to existing hardware that can only decode a handful of codes, the chip can accurately decode more than 2¹⁰⁰⁰ distinct codes. Moreover, a single chip can support efficient, hard-detection decoding of any product code of up to 16,384 bits with rate R=1/4, and a higher GECC than its 128-bit component codes [23].

II. GUESSING RANDOM ADDITIVE NOISE DECODING

Consider a binary codebook C_n , which is a set of 2^n strings in $\{0, 1\}^n$ where n is the code length, k is the number of information bits, and $R = k/n$ is the code rate. A transmitter sends a codeword x^n from the codebook C_n . Assume that a random noise effect N^n , which is independent of the channel input and also takes values in $\{0, 1\}^n$, additively alters the

a throughput of 122.6 Mbit/s. The chip allows seamless swapping between codebooks with no downtime, enabling use by multiple applications simultaneously.

Index Terms—GRAND, universal decoder, maximum likelihood decoding, hard-detection, error correcting codes

1. MOTIVATION

Transmitted data is subject to harmful corruption by noise present in a communication channel. By adding redundant bits to transmissions, error detection codes allow the data's integrity to be tested, while error-correcting codes enable the rectification of errors. The detection and correction of errors is done by involving decoding algorithms. Various codes with a particular structure have been developed and deployed for error correction as distinct code-specific decoders. Examples include Reed-Muller (RM) codes [3], [4] with Majority Logic Decoding, Reed-Solomon codes [5], [2], Berlekamp-Massey decoding [6], [7], and most recently, CRC-Assisted Polar (CA-Polar) codes [8]-[10].

1) pairing of restricted encoders with code-specific decoders; and 2) few constructions of codes at any given rate for any family of codes.

• High latency:

1) Emphasis on use of long codes such as LDPCs, even in newer developments such as Polar codes; and 2) reliance on interleavers to make the channel appear memoryless, thus do factor increasing the length of codes by two orders of magnitude.

GRAND is a universal decoding scheme that works efficiently with any code of moderate redundancy regardless of its length. In contrast to the state of the art, GRAND offers the following features to address the limitations of the conventional decoding approaches:

• Adaptability:

1) both to different technologies as well as to operating conditions; 2) universal ML decoding, allowing for a

Yazicigil, R, et al. "A Universal Maximum Likelihood GRAND Decoder in 40nm CMOS." 2022 14th International Conference on Communication Systems & Networks (COMSNETS).

linal decoders, and of the throughput gains obtained by hybrid-channel coding method that can provably achieve Shannon's channel capacity for any binary-input symmetric discrete memoryless channel. Apart from the intense theoretical interest in the subject, polar codes have attracted attention for their potential applications. There have been several proposals on hardware implementations of polar codes, which mainly focus on the efficient implementation of successive cancellation

I. INTRODUCTION

POLAR codes were proposed in [1] as a low-complexity channel coding method that can provably achieve Shannon's channel capacity for any binary-input symmetric discrete memoryless channel. Apart from the intense theoretical interest in the subject, polar codes have attracted attention for their potential applications. There have been several proposals on hardware implementations of polar codes, which mainly focus on the efficient implementation of successive cancellation

problem in polar decoding. One method that has been tried is Belief Propagation (BP) decoding, starting with [7]. In BP decoding, the decoder has the capability of making multiple bit decisions in parallel. Indeed, BP throughput of 2 Gb/s (with clock frequency 500 MHz) and 4.6 Gb/s (with clock frequency 200 MHz) are reported in [8] and [9], respectively. Generally speaking, the throughput advantage of BP decoding is observed at high SNR values, where correct decoding can be achieved after a small number of iterations. This advantage of BP decoders over SC decoders diminishes as the SNR decreases.

A second algorithmic approach to break the throughput bottleneck is to exploit the fact that polar codes are a class of generalized concatenated codes (GCC). More precisely, a

active savings [for example, $n_{\text{eff}} = 16$ for a GF(64) code in [4] or $n_{\text{eff}} = 32$ for a GF(256) code in [3]], a non-binary code defined over lower Galois field orders are still much more manageable while providing significant coding gain over binary LDPC codes. Unfortunately, the aforementioned technique does not apply to codes at lower GF(q), because there is no appropriate choice of n_{eff} for a small q that simultaneously maintains good coding gain and saves in hardware complexity, requiring us to pursue unique implementation techniques appropriate for lower GF(q).

Second, NB-LDPC decoder solutions provided thus far only apply to very short codes, which are not appropriate for storage systems. In fact, codes in storage systems are required to be very long (typically 1 KB and above) in order to ensure a sharp waterfall region, low error floor, and lack of undetectable errors, which gives rise to a new set of challenges in implementation, such as the large intermediate message storage requirement. A fully parallel architecture, as employed in [3] and [4], enables high throughput and avoids

are based on the Tanner graph of the code, constructed by representing the rows of H with M check nodes, the columns of H with N variable nodes, and the connecting check node m and variable node n with an edge, weighted by $h_{m,n}$, if $h_{m,n} \in \text{GF}(q) \setminus \{0\}$. Let I_n denote the set of variable nodes that are adjacent to check node n and J_m denote the set of check nodes adjacent to variable node n . Regular NB-LDPC codes have constant check and variable node degrees, denoted by $d_c = |J_m|$ and $d_v = |I_n|$, respectively. Let $L_n(a)$ be the a priori channel information of variable node n corresponding to $a \in \text{GF}(q)$, defined as

$$L_n(a) = \ln \left[\frac{\Pr(x_n = a | \text{channel})}{\Pr(x_n = 0 | \text{channel})} \right] \quad (1)$$

where \hat{a} is defined to be the most likely field element for variable node n , i.e., $\Pr(x_n = \hat{a})$ is maximum when $a = \hat{a}$. In iterative decoding algorithms of NB-LDPC codes, messages are passed back and forth between adjacent variable nodes and check nodes. Let $Q_{m,n}$ be the message from variable node n to check node m and $R_{m,n}$ be the message from check

IEEE Fellow、通信最高奖“香农奖”得主 Erdal Arikan 教授引用候选人工作作为极化码芯片代表性工作

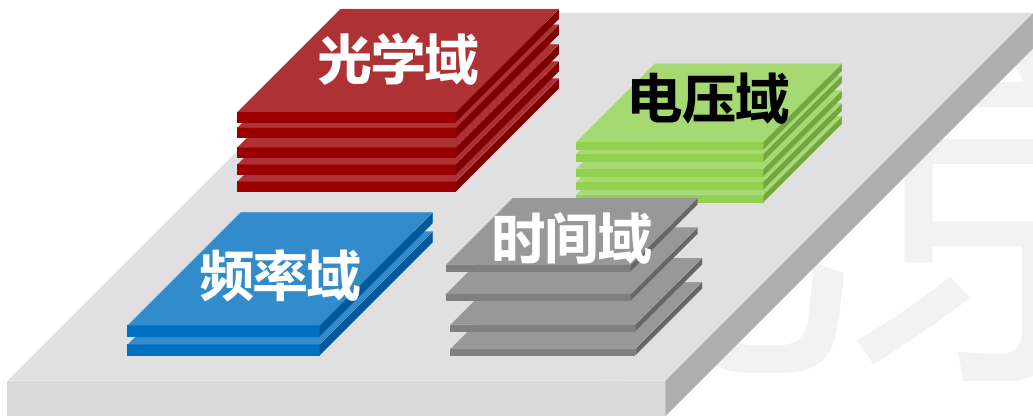
JSSC主编、UCLA Dejan Markovic 教授评价候选人提出的新架构使硬件效率提升且更容易实现



- **候选人获多个奖项，并在华为、高通、德州仪器等企业实现落地量产**
 - 2025年华为青年人才支持计划
 - 2024第26届中国电子学会青年年会专题报告奖
 - 2024年中关村论坛“智能未来”分论坛特邀嘉宾
 - **2023年华为计算产品线优秀技术合作奖**
 - 2023年Wiley Open Science Excellent Author
 - 2022年入选北京智源人工智能青年科学家俱乐部（青源学者）
 - **2021 IEEE Global Communication Conference最佳论文奖**
 - **2021年高通技术明星奖 Qualcomm QualStar Award**
 - **2020年高通技术明星奖 Qualcomm QualStar Award**
 - **2019 IEEE Symposium on VLSI最佳论文奖提名**
 - **2019年高通技术明星奖 Qualcomm QualStar Award**
 - **2018年密歇根大学Rackham科学研究STG奖**
 - 2015年斯坦福大学Analog Device模拟电路设计一等奖
 - **2013 IEEE ISCAS最佳论文奖提名**



三、多物理域融合芯片架构

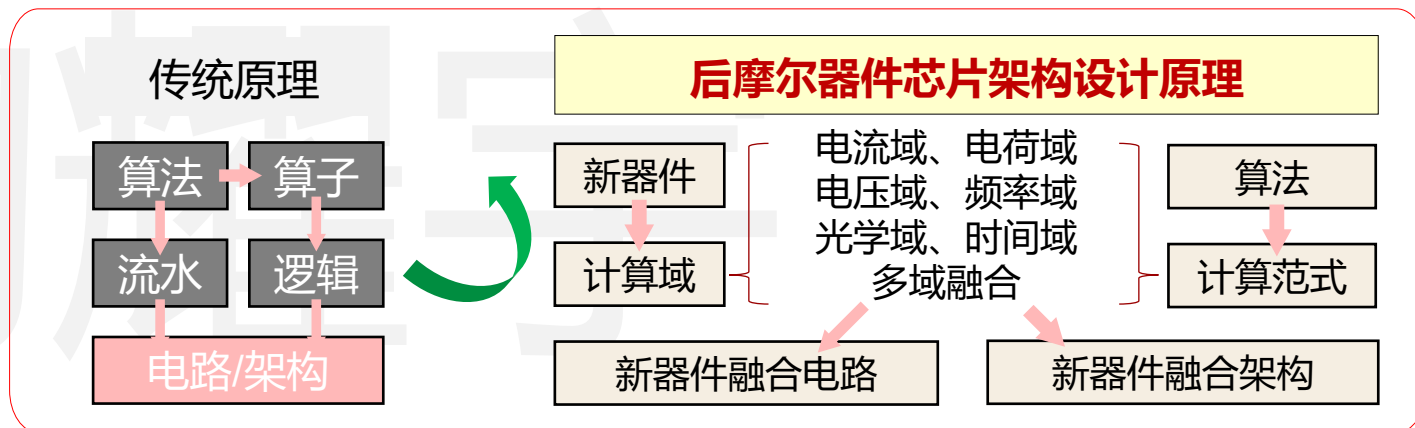


- **多物理融合芯片架构参考模型**，构建一个理论完备、层次清晰异构融合架构设计方案，涵盖各域功能划分、接口标准、数据流等
- **多物理域关键模块原型验证**，关键功能模块（如时间域编码处理单元、电光转换模块、跨域调度控制器）设计，形成多物理域融合芯片架构实际流片与板卡级演示
- **系统级多物理域异构融合仿真**，开发一套跨域异构计算仿真环境，用于评估计算任务性能、能耗、带宽瓶颈等关键指标

多物理域融合芯片架构

二维/三维异构

“算子-布尔逻辑-流水线”传统
芯片架构原理→“**算子-物理域-
多物理域融合**”芯片架构新原理





谢谢