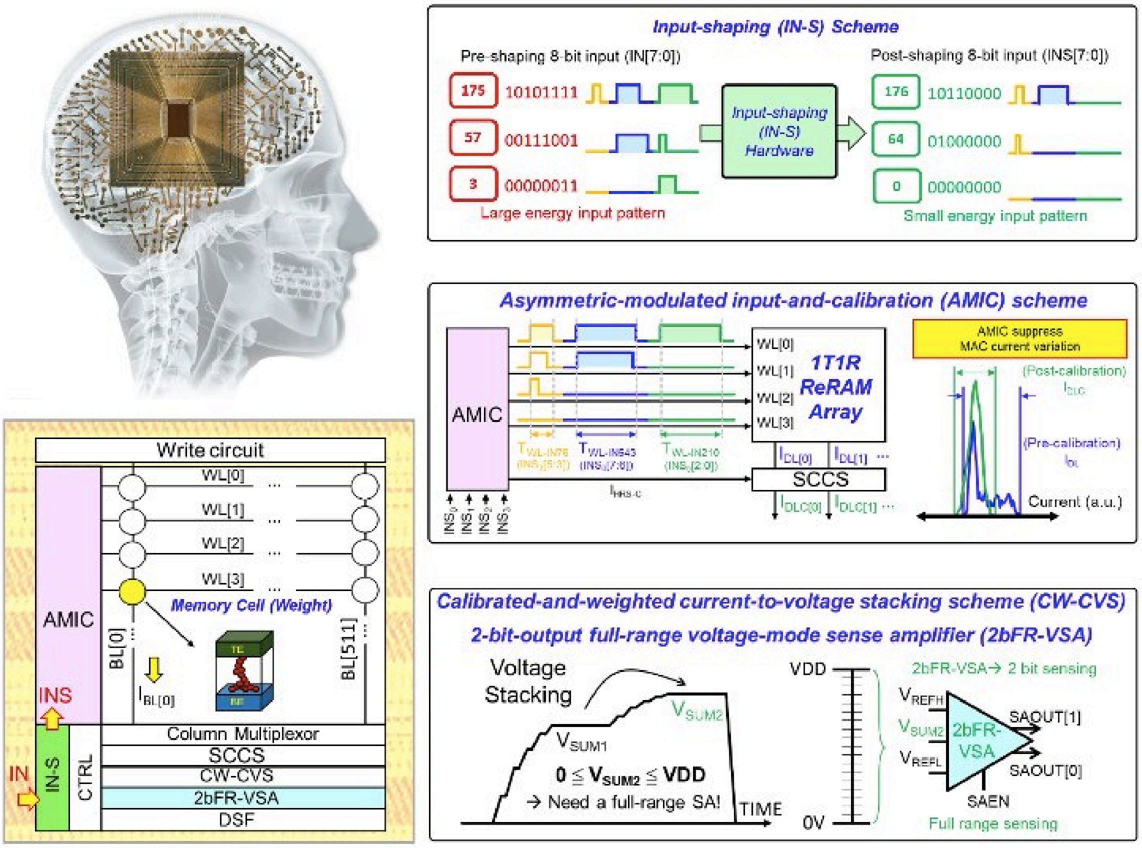
TechXplore

A four-megabit nvCIM macro for edge AI devices

January 27 2022, by Ingrid Fadelli


Input-shaping (IN-S) Scheme
Pre-shaping 8-bit input (IN[7:0])
175 10101111
57 00111001
3 00000011
Large energy input pattern
Input-shaping (IN-S) Hardware
Post-shaping 8-bit input (INS[7:0])
176 10110000
64 01000000
0 00000000
Small energy input pattern
Asymmetric-modulated input-and-calibration (AMIC) scheme
AMIC
WL[0]
WL[1]
WL[2]
WL[3]
1T1R ReRAM Array
SCCS
AMIC suppress MAC current variation
(Post-calibration)
(Pre-calibration)
Current (a.u.)
Write circuit
AMIC
INS
IN
IN-S
CTRL
BL[0]
BL[511]
Memory Cell (Weight)
Column Multiplexor
SCCS
CW-CVS
2bFR-VSA
DSF
Calibrated-and-weighted current-to-voltage stacking scheme (CW-CVS)
2-bit-output full-range voltage-mode sense amplifier (2bFR-VSA)
Voltage Stacking
VDD
0V
TIME
0 ≤ VSUM2 ≤ VDD
→ Need a full-range SA!
2bFR-VSA→ 2 bit sensing
SAOUT[1]
SAOUT[0]
SAEN
Full range sensing

Die photo and macro structure along with the proposed scheme (input-shaping scheme, asymmetric-modulated input-and-calibration scheme, calibrated-and-weighted current-to-voltage stacking scheme, and 2-bit-output full-range voltage-mode sense amplifier). Credit: Hung et al.

Edge AI devices, systems that combine artificial intelligence (AI) and edge computing techniques, are becoming an essential part of the rapidly growing Internet of Things (IoT) ecosystem. These devices include smart speakers, smart phones, robots, self-driven cars, drones and data-processing surveillance cameras.

While these technologies have become increasingly advanced over the past few years, most of them exhibit limited energy efficiencies, inference accuracies, and battery lifetimes. Non-volatile computing-in-memory (nvCIM) architectures, an emerging class of approaches that minimize the movement of data between processors and memory components, could help to significantly reduce the latency and energy consumption associated with complex AI computations.

Researchers at the Taiwan Semiconductor Manufacturing Company (TSMC) recently developed a new four-megabit (4Mb) nvCIM approach that could help to improve the overall performance of edge AI devices. Their proposed architecture, presented in a paper published in *Nature Electronics*, combines memory cells with peripheral circuitry based on complementary metal-oxide semiconductor (CMOS) technology.

"The computing latency and energy consumption of neural networks operating for AI applications using conventional von Neumann computing architectures are dominated by the movement of data between the processing element and memory, creating a performance bottleneck known as the memory wall," Meng-Fan Chang, one of the researchers who carried out the study, told TechXplore. "NvCIM may help to overcome the memory-wall bottleneck for battery-powered AI edge devices by allowing analog operations for vector-matrix multiplication, which is the major computing operation in the neural network during the inference stage."

NvCIM architectures can significantly reduce the amount of data that is transferred between processors and memories in AI edge devices, particularly while the devices are performing inference and power-on operations on-chip. This can in turn lead to better energy efficiencies and extended battery lifetimes.

Chang and his colleagues have been developing computing-in-memory (CIM) devices for almost 10 years. In their past studies, they used a variety of different memory components, including SRAM, STT-MRAM, PCM, ReRAM, and NAND-Flash, to assess the resulting performance.

"Over the past five years, we presented 40 papers related to CIM at top Microelectronics conferences (ISSCC, IEDM and DAC)," Chang explained. "Our recent work builds on our long-term research on CIM, which outlined technical background of memory circuit design, the system-level chip design of neural networks, and AI algorithms."

The new 4Mb nvCIM architecture created by the researchers is based on 22-nm-foundry resistive random-access memory (ReRAM) devices, also known as memristors. Remarkably, Chang and his colleagues found that it can perform high-precision dot-product operations involving an 8-bit input, 8-bit weight and 14-bit output with little latency and high energy efficiencies.

"We developed a hardware-based input-shaping circuit, using software-hardware co-design methods to improve energy efficiency without degrading the system-level inference accuracy," Chang said. "To reduce computing latency and improve readout accuracy, we develop an asymmetrically modulated input-and-calibration (AMIC) scheme."

To reduce their device's computing latency, the researchers built a calibrated and weighted current-to-voltage stacking circuit with a 2-bit output and full-range voltage-mode sense amplifier. This circuit also ensures a good readout yield for the most significant bits (MSBs), reducing the architecture's overall readout energy.

The architecture created by Chang and his colleagues can tackle complex computing tasks across a variety of application scenarios. In addition, compared to other nvCIM architectures proposed in the past, it is more precise, has a higher computing throughput and a larger memory capacity, consumes less energy, and has a lower computing latency.

"We also focused on software-hardware co-design to further improve the chip-level performance," Chang said. "Existing advanced edge devices for AI and AI-enabled Internet of Things (AIoT) applications commonly adopt nvCIM for power-off data storage to suppress power consumption in standby mode and light computing tasks during wake-up."

In the future, the architecture developed by this team of researchers could be used to enhance the performance and energy efficiency of different edge AI devices, ranging from smart phones to more sophisticated robotic systems. Among other things, it can support basic vector-matrix multiplications (VMMs) performed by various neural network models, including convolution neural networks (CNNs) for image classification or deep neural network (DNNs).

"Circuit level optimization, nvCIM architecture novelty, improvement of specification, and performance of nvCIM macro are definitely next on our roadmap," Chang added. "Software-hardware co-design is also one of our future research topics, we aim at developing nvCIM-friendly neural network algorithms to further maximize the performance of nvCIM macro. Beyond that, our goal is to integrate the nvCIM macro and other necessary digital circuits into a chip-level system design for the next generation AI chips."

More information: Je-Min Hung et al, A four-megabit compute-in-memory macro with eight-bit precision based on CMOS and resistive random-access memory for AI edge devices, *Nature Electronics* (2021). DOI: 10.1038/s41928-021-00676-9

Citation: A four-megabit nvCIM macro for edge AI devices (2022, January 27) retrieved 8 March 2024 from https://techxplore.com/news/2022-01-four-megabit-nvcim-macro-edge-ai.html