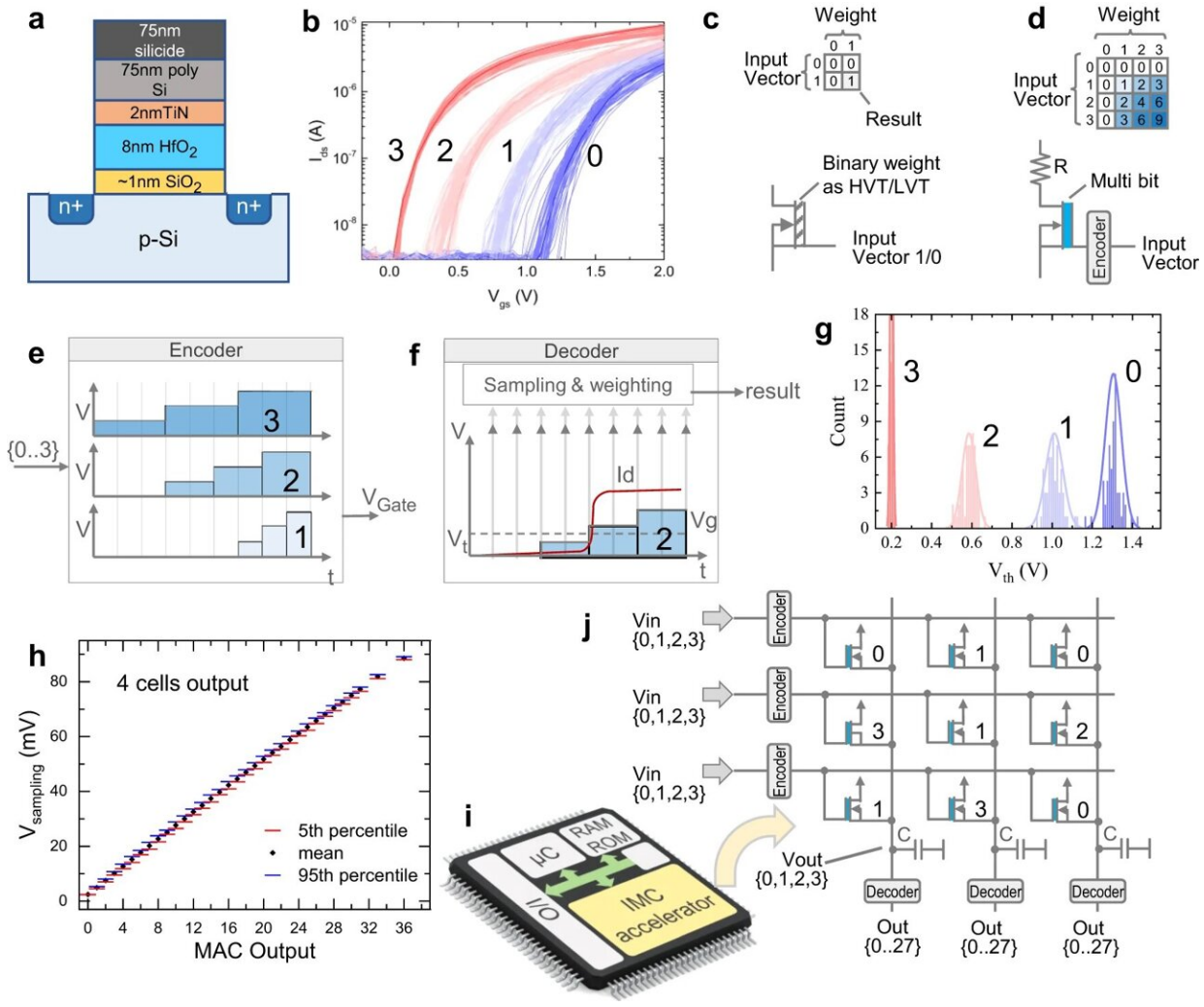
Exploring the details of an energy-saving AI chip

October 26 2023, by Andreas Schmitz



Overview of the proposed IMC macro for MAC operations. **a** The material stack of FeFETs. **b** The multi-bit FeFET can be programmed to different states to store the weight of the synapse. **c** Previous works[7, 8, 11] only considered binary AND or XNOR operations to compute a single-bit multiplication operation. **d**

Our proposed 2-bit multiplication operation with input encoding and 2-bit storage is shown. The corresponding output activates at different instances of time. **e** An encoder provides the gate voltage depending on the input value which changes between three levels at different instances of time. **f** The multiplication output of the input and stored state in the cell depends on the time at which one cell is activated which is accumulated and sampled using the decoder. **g** The V_{th} distribution of the four states for the $I_{ds} - V_{gs}$ curves is shown. **h** Depending on the activation time and the number of cells activated at a given time, the voltage across the capacitor connected to a column of cells is accumulated which corresponds linearly with the MAC output and has a minimal impact of the underlying device variation. **i** IMC accelerators facilitate MAC operations for AI workloads where our proposed design can be utilized. **j** The corresponding MAC operation is performed in the crossbar, accumulating the output in the capacitor voltage. Credit: *Nature Communications* (2023). DOI: 10.1038/s41467-023-42110-y

Hussam Amrouch has developed an AI-ready architecture that is twice as powerful as comparable in-memory computing approaches. As reported in the journal *Nature Communications*, the professor at the Technical University of Munich (TUM) applies a new computational paradigm using special circuits known as ferroelectric field effect transistors (FeFETs). Within a few years, this could prove useful for generative AI, deep learning algorithms and robotic applications.

The basic idea is simple: unlike previous chips, where only calculations were carried out on transistors, they are now the location of data storage as well. That saves time and energy. "As a result, the performance of the chips is also boosted," says Hussam Amrouch, a professor of AI processor design at the Technical University of Munich (TUM). The transistors on which he performs calculations and stores data measure just 28 nanometers, with millions of them placed on each of the new AI chips.

The chips of the future will have to be faster and more efficient than earlier ones. Consequently, they cannot heat up as quickly. This is essential if they are to support such applications as real-time calculations when a drone is in flight, for example. "Tasks like this are extremely complex and energy-hungry for a computer," explains the professor.

Modern chips: Many steps, low energy consumption

These key requirements for a chip are summed up mathematically by the parameter TOPS/W: "tera-operations per second per watt". This can be seen as the currency for the chips of the future. The question is how many trillion operations (TOP) a processor can perform per second (S) when provided with one watt (W) of power.

The new AI chip, developed in a collaboration between Bosch and Fraunhofer IMPS and supported in the production process by the US company GlobalFoundries, can deliver 885 TOPS/W. This makes it twice as powerful as comparable AI chips, including a MRAM chip by Samsung. CMOS chips, which are now commonly used, operate in the range of 10–20 TOPS/W.

In-memory computing works like the human brain

The researchers borrowed the principle of modern chip architecture from humans. "In the brain, neurons handle the processing of signals, while synapses are capable of remembering this information," says Amrouch, describing how people are able to learn and recall complex interrelationships. To do this, the chip uses "ferroelectric" (FeFET) transistors.

These are electronic switches that incorporate special additional characteristics (reversal of poles when a voltage is applied) and can store

information even when cut off from the power source. In addition, they guarantee the simultaneous storage and processing of data within the transistors.

"Now we can build highly efficient chipsets that can be used for such applications as deep learning, generative AI or robotics, for example where data have to be processed where they are generated," says Amrouch.

Market-ready chips will require interdisciplinary collaboration

The goal is to use the chip to run deep learning algorithms, recognize objects in space or process data from drones in flight with no time lag. However, the professor from the integrated Munich Institute of Robotics and Machine Intelligence (MIRMI) at TUM believes that it will be a few years before this is achieved.

He thinks that it will be three to five years, at the soonest, before the first in-memory chips suitable for real-world applications become available. One reason for this, among others, lies in the security requirements of industry. Before a technology of this kind can be used in the automotive industry, for example, it is not enough for it to function reliably. It also has to meet the specific criteria of the sector.

"This again highlights the importance of interdisciplinary collaboration with researchers from various disciplines such as computer science, informatics and electrical engineering," says the hardware expert Amrouch. He sees this as a special strength of MIRMI.

More information: Taha Soliman et al, First demonstration of in-memory computing crossbar using multi-level Cell FeFET, *Nature*

Communications (2023). DOI: 10.1038/s41467-023-42110-y