AI Inference and Storage

Tom Coughlin
Coughlin Associates, Inc.

Most machine learning for artificial intelligence (AI) applications is done in large data centers using curated training data. This is because the computing and energy resources required are large. As a recent paper points out, "it was found that AI models that use neural architecture search emit the carbon dioxide equivalent of nearly five times the lifetime emissions of an average American car."[1]

However, the application of trained AI at the edge of networks or at endpoints involves inference engines. These inference engines are embedded semiconductor devices, which store the weighting functions from the AI training for AI applications in the field that often work in harsh constrained environments with limited energy supplies. As a consequence, large-scale use of AI applications requires new levels of integration and technologies to reduce energy consumption.

One way to reduce device energy consumption is to replace volatile memory with nonvolatile memory. In particular, replacing static random access memory (SRAM) memory with fast nonvolatile memory could save energy and, since SRAM takes up a lot of real-estate on the die, could also increase the amount of memory that can be put into a given area on a semiconductor die.[2]

Not Or (NOR) flash memory is often used as a nonvolatile memory in embedded applications, but NOR flash faces limitations on how small it can get. NOR flash probably cannot be economically manufactured with lithographic features less than about 28 nm. As the memory requirements for AI-enabled applications increases, NOR flash built into embedded devices faces severe limitations.

Digital Object Identifier 10.1109/MCE.2019.2940743
Date of current version 6 December 2019.

A good candidate to replace slower SRAM and NOR flash for embedded applications is magnetic random access memory (MRAM).[3] MRAM is nonvolatile and maintains its data even when the power is turned OFF. Current spin tunnel torque MRAM has very high write endurance (up to 10^{15}) and read and write times that can match slower SRAM used in higher level caching applications (e.g., L3 and L4).

Being able to turn a device OFF when it is not in active use, and turn it ON and being able to pick up where the machine left OFF, can save considerable energy and enable battery powered applications. This could help with many mobile and remote AI-powered consumer and industrial products.

Furthermore, while SRAM memory cells have six to eight transistors per cell, MRAM has one-transistor per memory cell. Thus, for a given die area, MRAM can provide 10× or greater the memory than SRAM. An embedded chip with MRAM replacing some of the slower MRAM and embedded NOR flash is depicted in Figure 1. The power savings and memory density increase possibly by using MRAM in embedded products.

Published by the IEEE Consumer Electronics Society

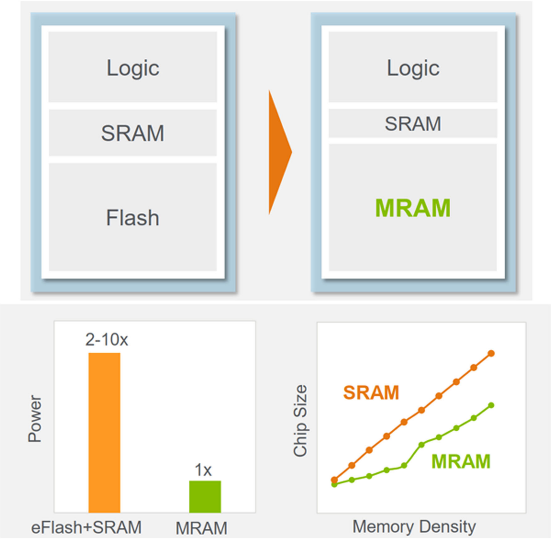

Figure 1. MRAM used to replace NOR flash and slower SRAM, lowering power consumption and increasing memory density.[4]

Because of these power and space advantages for replacing fast volatile memory with nonvolatile memory, such as MRAM, all the major semiconductor foundries are offering MRAM as an optional embedded product memory. Some early products using MRAM are now available including the 22 nm ASIC Gyrfalcom MRAM Engine (Lightspeeur 2802M) with 40 MB of MRAM and meant for image classification, voice identification, text to speech, facial recognition, voice commands, and other AI applications in the field.

Emerging memories, such as MRAM, are enabling the next generation of AI applications used in consumer as well as industry devices, particularly in power constrained and harsh environments. The major semiconductor foundries are geared up to include MRAM and other nonvolatile memories in their embedded products. You can expect to buy long lasting AI-enabled devices using these nonvolatile memories in the near future.

REFERENCES

1. E. Strubell *et al.*, "Energy and policy considerations for deep learning in NLP," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, Jul. 2019. [Online]. Available: https://arxiv.org/abs/1906.02243
2. T. Coughlin, "When memory starts to think," *IEEE Consum. Electron. Mag.*, vol. 8, no. 3, pp. 90–91, May 2019.
3. T. Coughlin, "Crossing the chasm to new solid-state, storage architectures," *IEEE Consum. Electron. Mag.*, vol. 5, no. 1, pp. 133–142, Jan. 2016.
4. K. Moraes, "Accelerating new memory technologies for the IoT and cloud computing," in *Proc. Appl. Mater. Presentation Semicond.*, slides 7 and 8, Jul. 2019.

Tom Coughlin is the President of Coughlin Associates, Inc. He has been a Digital Storage Analyst and Business and Technology Consultant for more than 37 years. He is an IEEE Fellow and the President of IEEE-USA. Contact him at tom@tomcoughlin.com.