Evaluating pseudo-random SRAM for AI applications in GPU cache

Kwaku Ofosu Asare

June 2024

1 Popular Science Summary

AI has become unavoidable in daily life; it is used to recommend services to us, in virtual assistants, healthcare, finance, and more. The meteoric rise of AI and its applications can be attributed to the technological advancements that have allowed processing power to catch up with the computational requirements of AI training.
These AI/ML models require large compute resources and are still limited by hardware, as current models are designed to take advantage of all the available hardware resources. Graphics Processing Units (GPUs) are the prevalent choice of processors for specific repetitive workloads with parallelism due to the GPU structure. This has led to a great increase in the computing power of graphics cards (GPUs) which has eclipsed the growth of computing power in Central Processing Units (CPUs). Performance increases in GPUs can be attributed to advancements in manufacturing processes and an increase in onboard memory.