Docket #: S23-084

Pipelined chip architecture for low-cost, energy efficient machine learning on edge devices

Researchers in the Murmann Mixed Signal Group have developed a pipelined chip architecture with inverted residual and linear bottlenecks-based networks for energy efficient Machine Learning inference on edge devices. ML (machine learning) models "at the edge" on resource-constrained devices is difficult, due to limited memory and a strict power budget. The Stanford researchers used quantized and heavily pruned bottleneck-based networks with dense custom latch arrays (CLAs), to create a tight, integrated compute and memory device. Their low-read-energy, dense CLAs and high degree of parallelism of their dense compute fabric facilitate dataflow with inputs and outputs only read/write once. This eliminates the need for a data buffer in the memory hierarchy, saving silicon area and memory accesses. The resulting, energy efficient, near ideal device is as dense as possible and cheaper to read than conventional standard-cell-based latch array.

Stage of Development – Prototype
The Murmann group tested their custom chip prototype and verified the end-to-end performance and power/energy per-inference estimates. Their CLA implementation in 28nm achieves 60x lower read energy (1.6x higher density) than iso-port width SRAM macros of the same capacity made by a memory compiler, and more than 5x lower energy (2x higher density) than a latch array synthesized from standard cells in the same technology. Further chip testing and architecture analysis is underway.

Applications

- **IoT devices** and high-end consumer electronics such as **VR glasses, phones, and security cameras**.

- **Edge computing devices** that perform **machine learning** inference in power/energy-constrained environments.
- **IC design** to run larger ML models on power-constrained devices and to achieve higher compute performance given low power/energy budget devices.

Advantages

- **More energy efficient and lower cost**:
  - **60x lower read energy** (1.6 x higher density) than iso-port width SRAM macros of the same capacity made by a memory compiler.
  - **More than 5x lower energy** (2x higher density) than a latch array synthesized from standard cells.

Innovators

- Massimo Giordano
- Rohan Doshi
- Boris Murmann

Licensing Contact

Luis Mejia

Senior Licensing Manager, Physical Sciences

Email