3D NAND based Compute-in-memory Technology for Energy-efficient Processing of Huge AI Models

Department of Electrical and Information Engineering,

Seoul National University of Science and Technology (SeoulTech)

Wonbo Shim

2023. 04. 04.
The 17th ROK-USA Forum on Nanotechnology
Plaza Hotel, Seoul

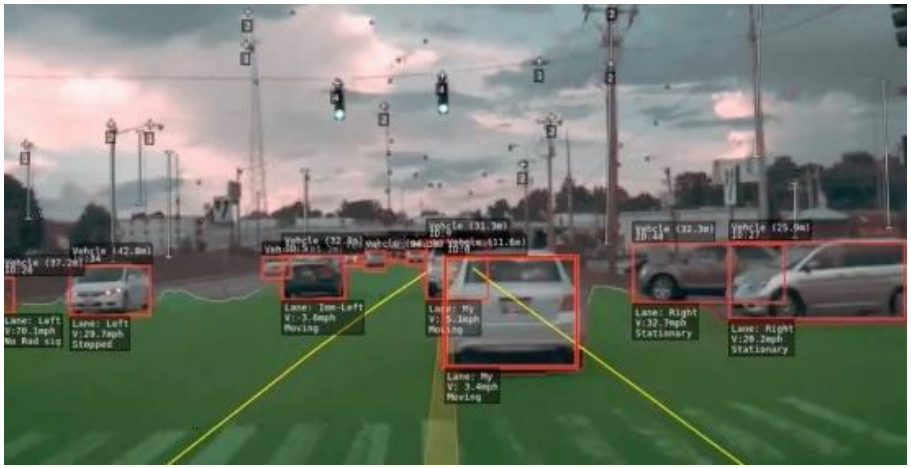
SEOULTECH
SEOUL NATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY

Bio

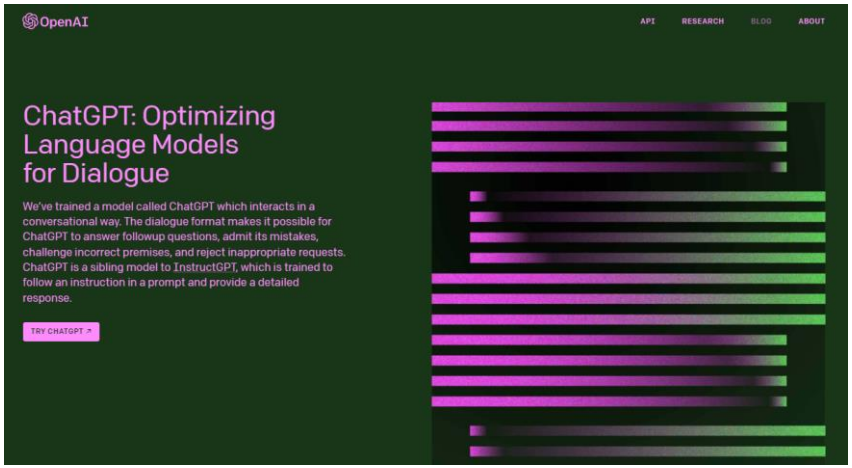
- **B.S. Seoul National University (2007)**
  - **Electrical Engineering**
- **Ph. D. Seoul National University (2013)**
  - **Electrical and Computer Engineering**
- **Samsung Electronics (2013~2019)**
  - **Flash Design Team**
- **Postdoctoral research fellow, Georgia Tech (2019~2021)**
- **Assistant Professor, SeoulTech (2021~present)**
  - **Department of Electrical and Information Engineering**

Motivation of Compute-in-memory (CIM)

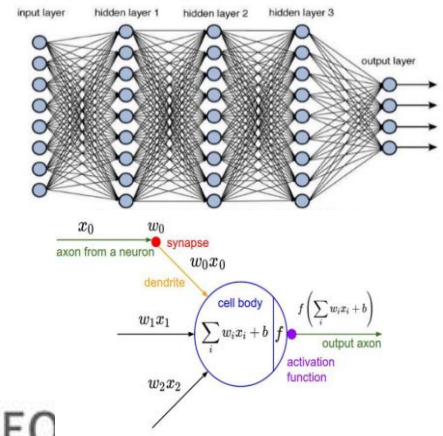
- **Recently, AI is widely used in computer vision (e.g. image classification), natural language processing (e.g. language generation), etc.**



Tesla, Autopilot



OpenAI, ChatGPT



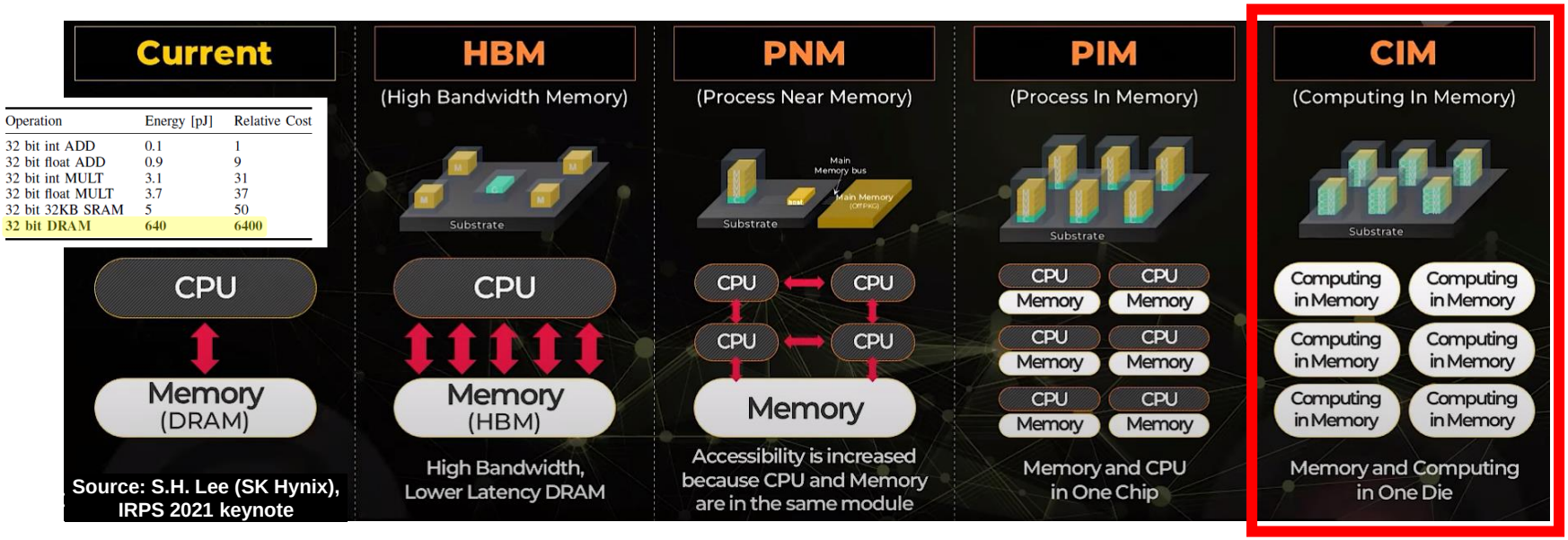
But,

- **Requires lots of computations**
- **Computing hardware consumes huge power**
  - Lee Se-dol (20W) vs Alphago (1MW)

Using AI is not free!

Motivation of Compute-in-memory (CIM)

- **In CPU or GPU, most of the energy is consumed for “data movement” between processor and memory.**
- **To replace von-Neumann architecture, various memory-centric systems have been proposed.**



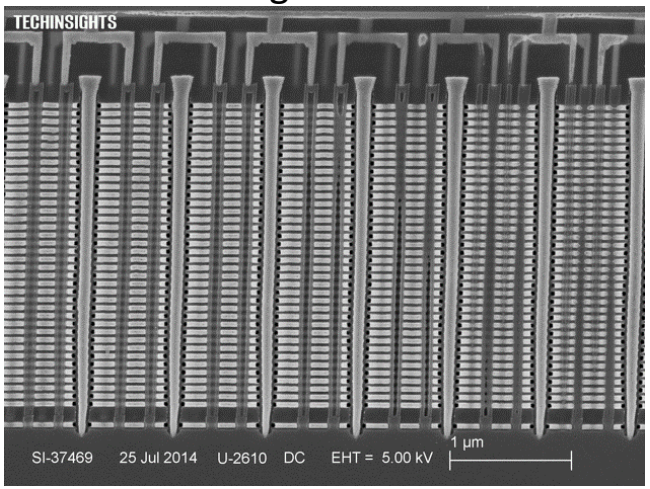
Compute-in-memory (CIM) technology : > x100 energy efficiency

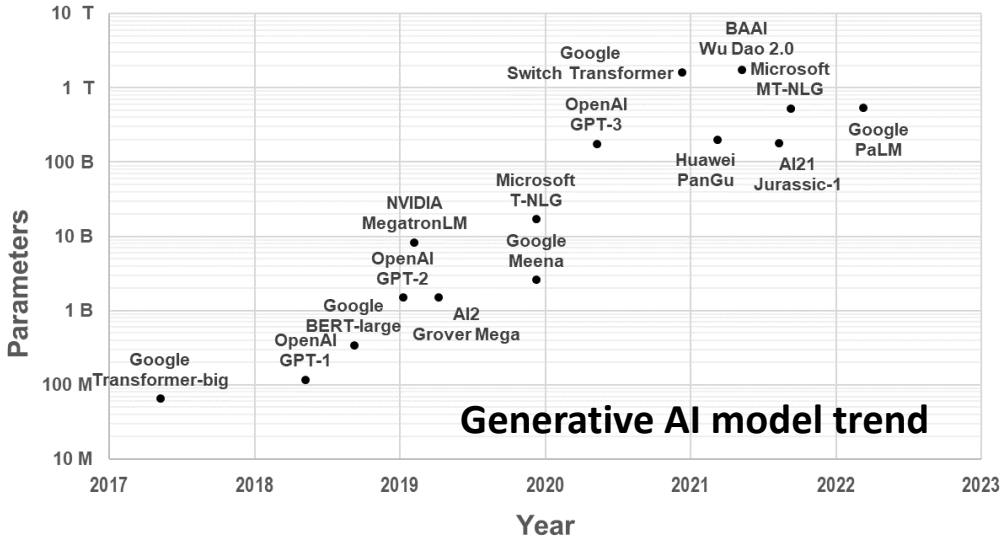
All kinds of memory devices can be a candidate, but still don’t know what is best.

3D NAND based CIM

- **3D NAND – Ultra-high density, low cost**
  - **Conventional 3D NAND : Mass storage application**
  - **3D NAND CIM : energy-efficient processing of huge AI model**
    - **GPT-3 (175 Billion parameters)**

Samsung's 3D NAND

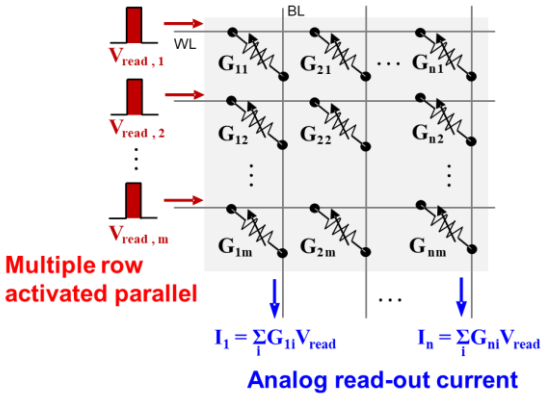


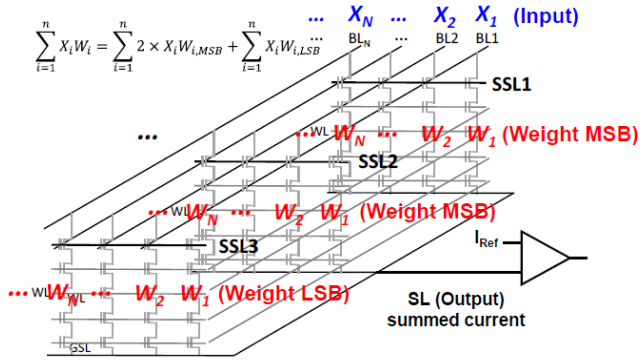


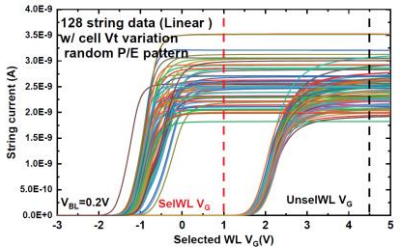
Conversion from GPU to "3D NAND CIM" can save power & cost

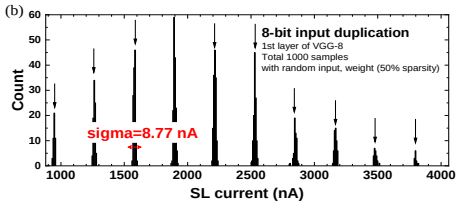
3D NAND based CIM

Q. How does it work?
A. Analog computation. (Then, what about accuracy?)

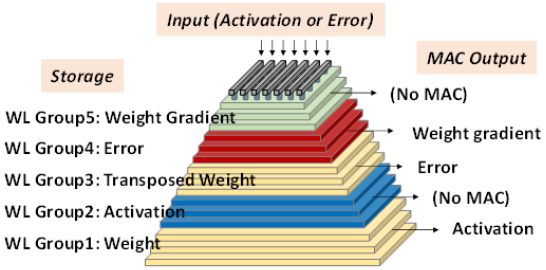


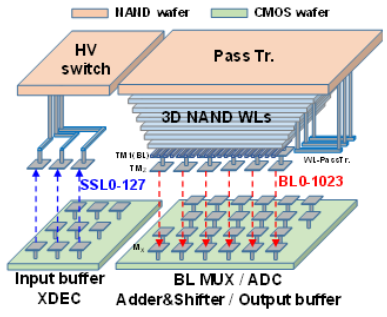


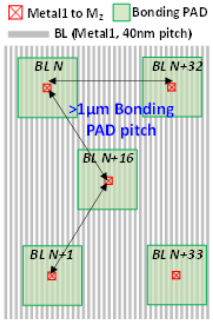


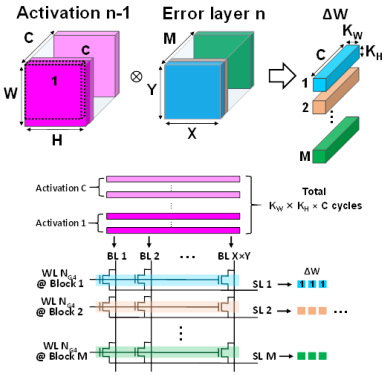


Q. How does it achieve good energy efficiency and latency?
A. Optimized architecture, operation method design.









KOR-U.S. Collaboration

- **Previous collaboration**
  - **Georgia Tech, Arizona State Univ.**

962 IEEE TRANSACTIONS ON ELECTRON DEVICES, VOL. 67, NO. 3, MARCH 2020

Drain-Erase Scheme in Ferroelectric Field Effect Transistor—Part II: 3-D-NAND Architecture for In-Memory Computing

Panni Wang, *Student Member, IEEE*, Wonbo Shim, Zheng Wang, *Student Member, IEEE*, Jae Hur, Suman Datta, *Fellow, IEEE*, Asif Islam Khan, *Member, IEEE*, and Shimeng Yu, *Senior Member, IEEE*

160 IEEE ELECTRON DEVICE LETTERS, VOL. 42, NO. 2, FEBRUARY 2021

Technological Design of 3D NAND-Based Compute-in-Memory Architecture for GB-Scale Deep Neural Network

Wonbo Shim and Shimeng Yu, *Senior Member, IEEE*

500 IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS, VOL. 12, NO. 2, JUNE 2022

GP3D: 3D NAND Based In-Memory Graph Processing Accelerator

Wonbo Shim and Shimeng Yu, *Senior Member, IEEE*

- **I hope & expect..**
  - **Korea (memory fabrication leader) + U.S. (processor design leader) could advance the development of *"Compute-in-memory"* technology.**

Thank you
Q&A