memBrain™

A Microchip Technology Company

memBrain™ Technology for Edge AI/ML Acceleration

Presenter: Hieu Tran, SST/Microchip

August 7, 2024

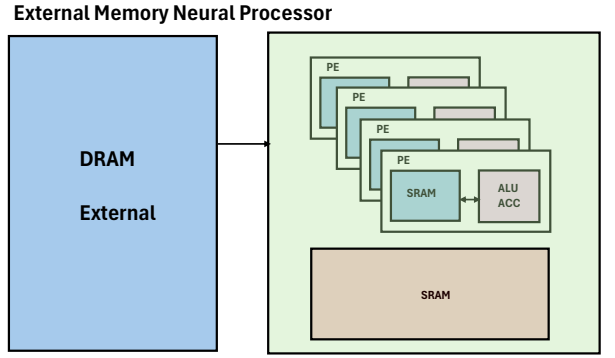
FMS the Future of Memory and Storage

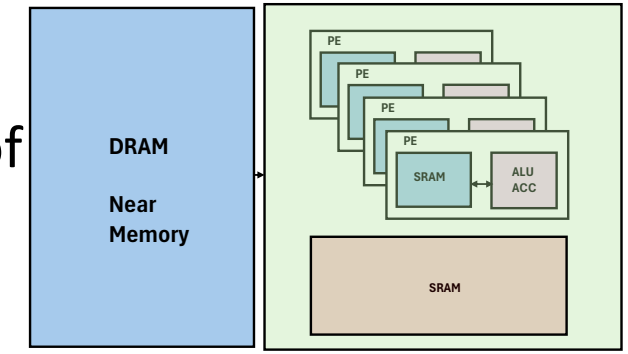
Microchip

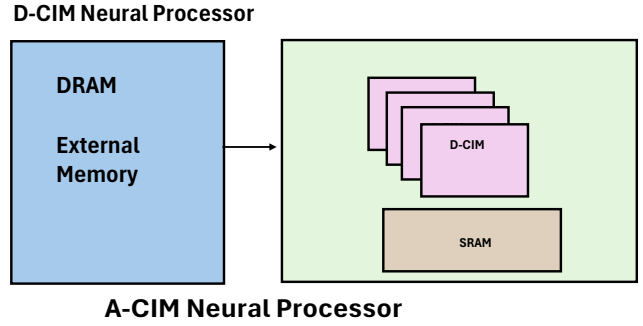
Introduction

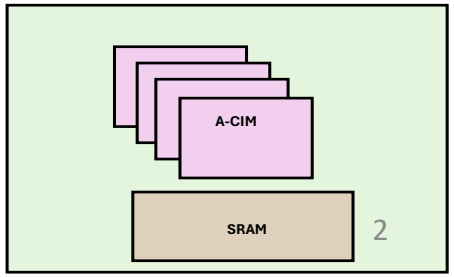
Current hardware solutions for edge AI/ML

- AI/ML Accelerator (AI chip) is a device to speed up AL/ML applications by accelerating computation often at reduced energy. Main computation task is dot product operation for vectors and matrices.
- Current HW accelerators for edge AI/ML encompass multiple application space having a wide range of weight capacities from tens of KW, tens of MWs, hundreds of MWs, and few BWs.
- Digital Neural Processor: Typically, tens to thousands of PE (processing element which execute dot product op), local SRAM; external memory (DRAM and NVM) or near memory (SRAM, DRAM)
- Digital CIM: SRAM cell store binary weights, ADC based summation or logic summation (built-in adders). External DRAM and NVM.
- Analog CIM: VM or NVM array-based MAC operation.









Comparison summary for Digital Accelerator, D-CIM, & A-CIM

	Power Efficiency	Area	Performance
Digital Acc	✓	✓	✓
Digital CIM	✓✓	✓-	✓✓
Analog CIM	✓✓✓	✓✓	✓✓✓

Comparison summary for A-CIM for different NVM technologies

- **ReRAM:**
  - Only 4 levels per cell at most
  - Reliability, Repeatability
  - Technology still immature
  - Availability: Limited
  - High Power (cell current ~20-40uA)
- **MRAM:**
  - Only Binary
  - Availability: Limited
  - High Power (cell current ~20 uA)
- **SRAM:**
  - Only Binary
  - High Power (cell current tens of uA)
    - Volatile
    - Need external NVM memory

- **ESF FG A-CIM Solution:**
  - **In Production**
  - **Extreme Low Power (cell current 0-100nA)**
  - **Master multi-level sub threshold region**
  - **> 32 Levels per cell**
  - **Current Solution: 100K Weights to 100M Weights**
  - **Current Development: improved Tile, chiplet, 3D/2.5D packaging**
  - **Available at all major Foundries.**

memBrain™

What IS **memBrain**?

- memBrain IS IP!
- SST ESF3 SuperFlash bitcell
- Analog Compute in Memory (aCIM)
- “Single cycle” analog MAC operation performed inside the SuperFlash array
- Full set of drop-in H/W IP:
  - Array
  - Analog eFlash Control
  - Algorithms
  - DAC/ADC
  - Simulation models
  - Test programs
  - Documentation

- Software IP tool flow to build & load models
- 40nm and 28nm foundry IP
- Can use alongside standard ESF3 code/data macros

Application:

AI/ML edge inference

memBrain™

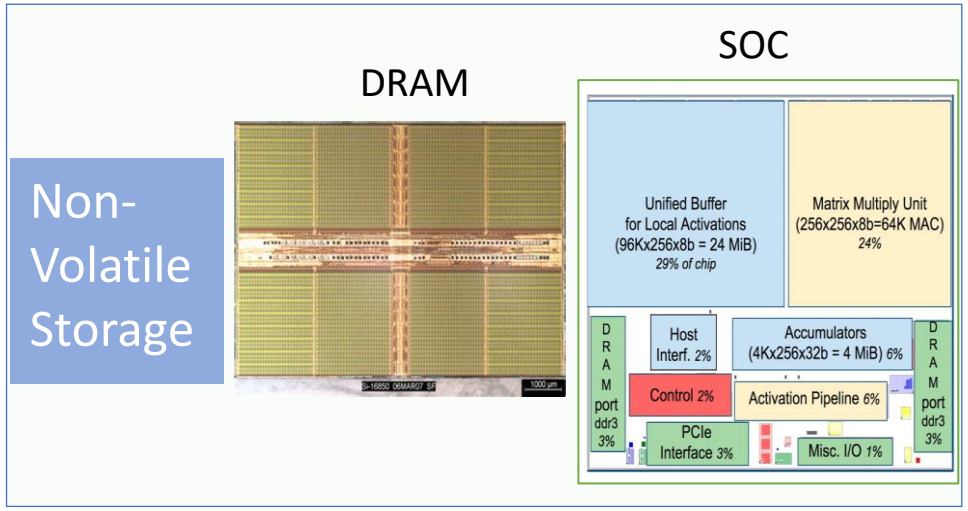
Why use **memBrain** IP? Where to apply **memBrain** IP?

- Why:
  1. **Low cost** (40nm, 28nm)
  2. **Low power** (<5W, mW)
  3. **Low heat** dissipation
  4. Handles 100K to 100M+ weights
  5. Programmable to standard models
  6. Similar performance to GPU + DRAM
  7. Same ESF3 bitcell (no process change)
  8. Just IP (from SST)
  9. Available NOW
  10. No “magic memory” needed
  11. Integrates well with MCUs/IoT/Sensors
  12. Extensible chiplet architecture

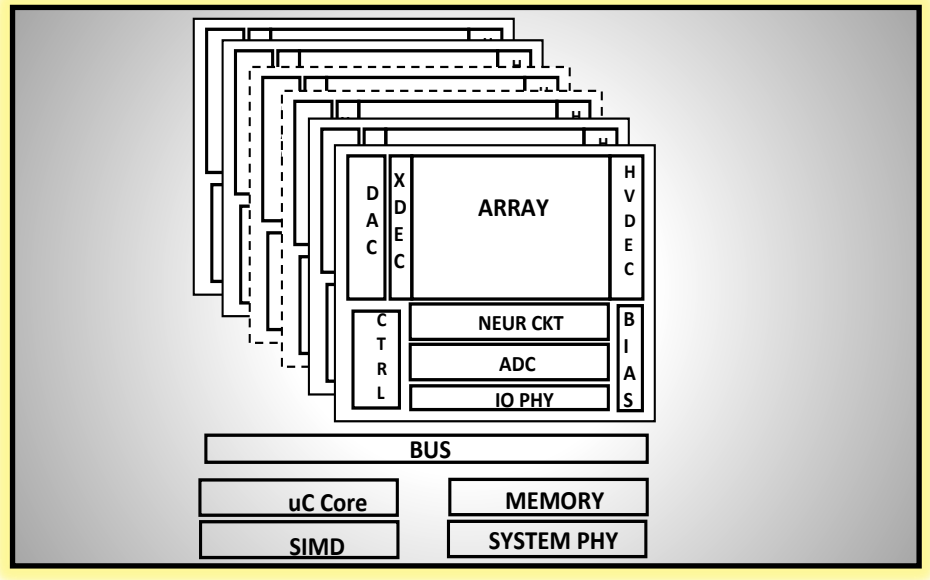
- Where:
  - A. Smart Speakers
  - B. Headphones/Earbuds
  - C. Doorbell cameras
  - D. Home automation
  - E. Smartphones
  - F. Drones
  - G. Security cameras
  - H. Threat detection
  - I. AR/VR headsets
  - J. Video upscalers
  - K. Medical diagnostics
  - L. IoT devices

memBrain™ Technology

- A Power and Area Effective A-CIM solution
- Extreme Low Power utilizing sub-threshold region
- Single chip solution without external DRAM or NVM
- Cover tens of KW to billions of weight applications
- Area Efficiency
  - Analog multilevel technology
  - Sub-threshold region
- Cost efficiency
  - DRAM-less
  - MLC array
- Power Efficiency ~ >10X
- Architecture: Homogenous or Heterogenous, hybrid Architecture with A-CIM and/or D-CIM and/or Dig Acc



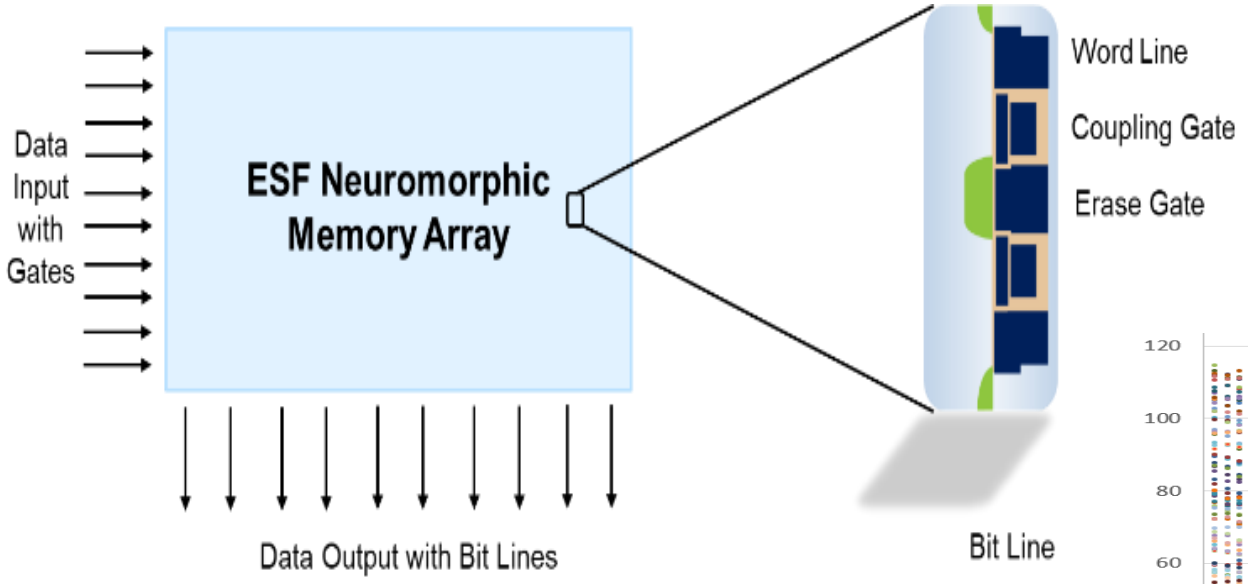
A Digital DL Accelerator

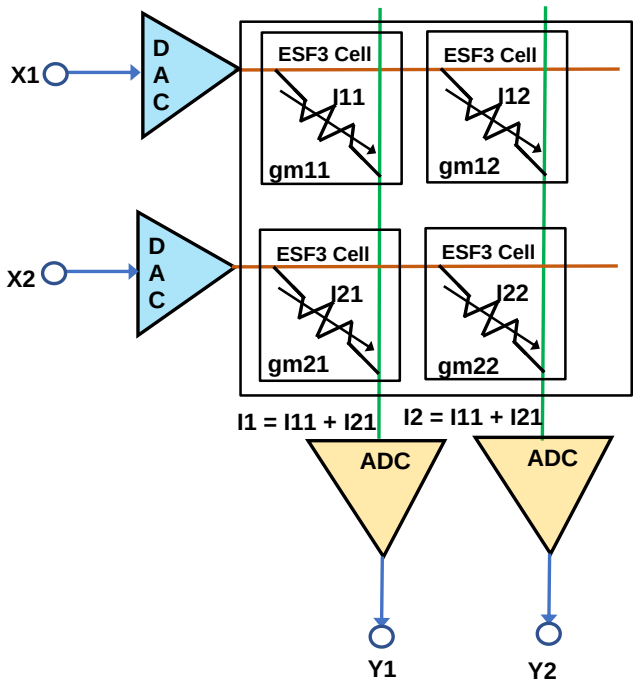


An Analog DL Accelerator using ESF

memBrain™ Technology

- CIM memBrain™ implements the dot product (MAC) through memory array operation
- Output cell current equals to its conductance multiplied by the input voltage
- Yj = ∑ (Xi * Wij)

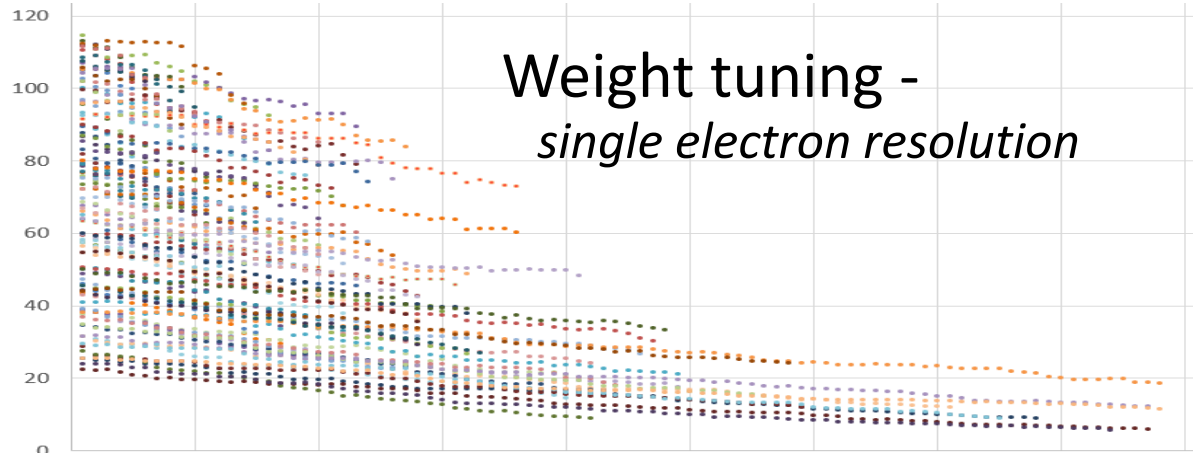




dot product:

$$I_j = \sum_{i=0} W_{ji} V_i$$

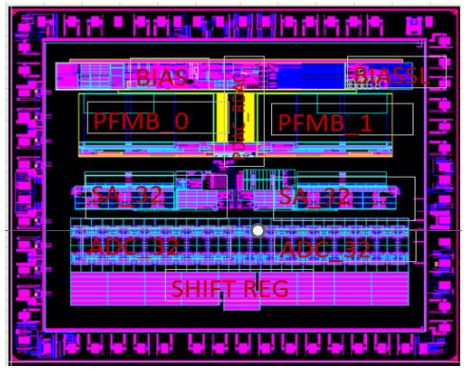
Where:
W is a weight matrix and Wji is an element of that matrix
Vi is an input from the input vector
Ij = output of the dot product

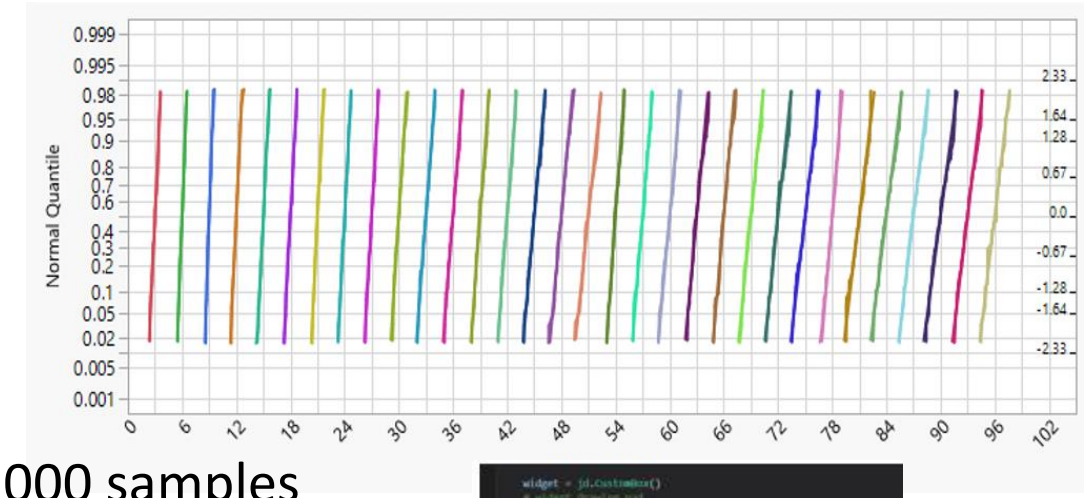


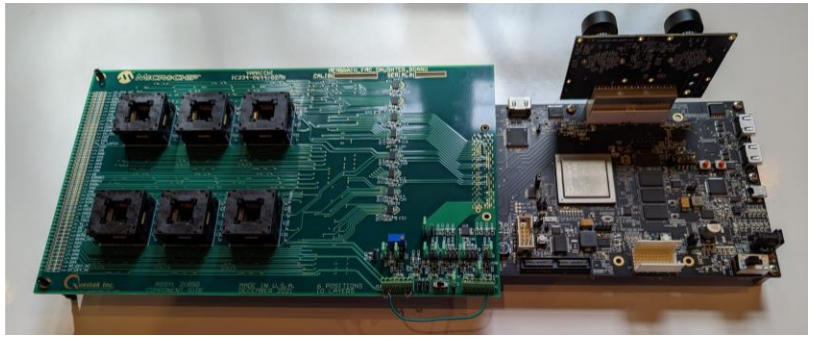
Weight tuning -
single electron resolution

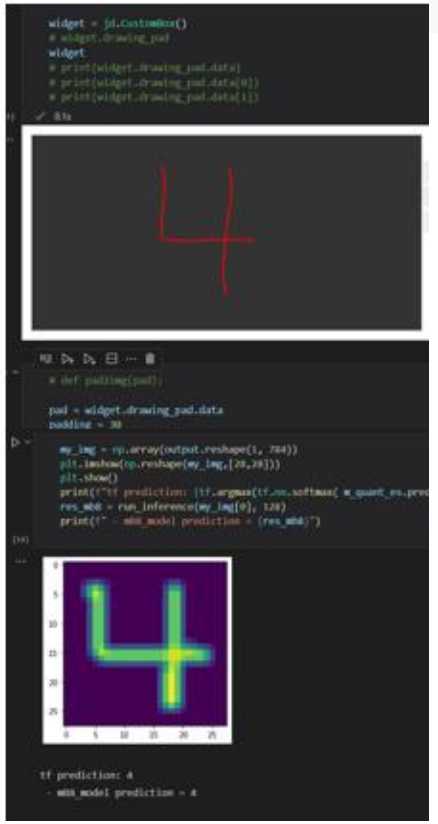
ESF3-28 **memBrain** Silicon 1

- ESF3-28 TC1:
  - 1.3MW, -40C to 85C
  - Yield ~ 95% 1st Silicon
- Inference Result:
  - ESF3-40 TC1: 3.5% loss
    - MNIST MLP: 93.0% vs. 96.5% TF baseline, 1,000 samples
  - ESF3-28 TC1: 0.6% to 2.1% loss
    - MNIST MLP: 94.4% vs. 96.5% TF baseline, 1,000 samples
    - MNIST ConvNet: 97.7% vs. 98.3%TF baseline, 1,000 samples





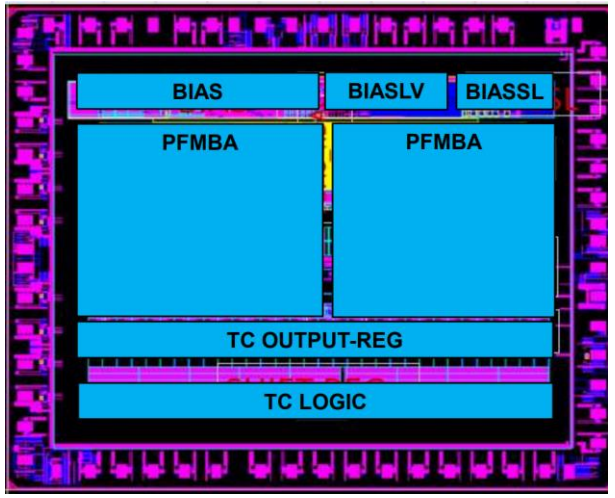


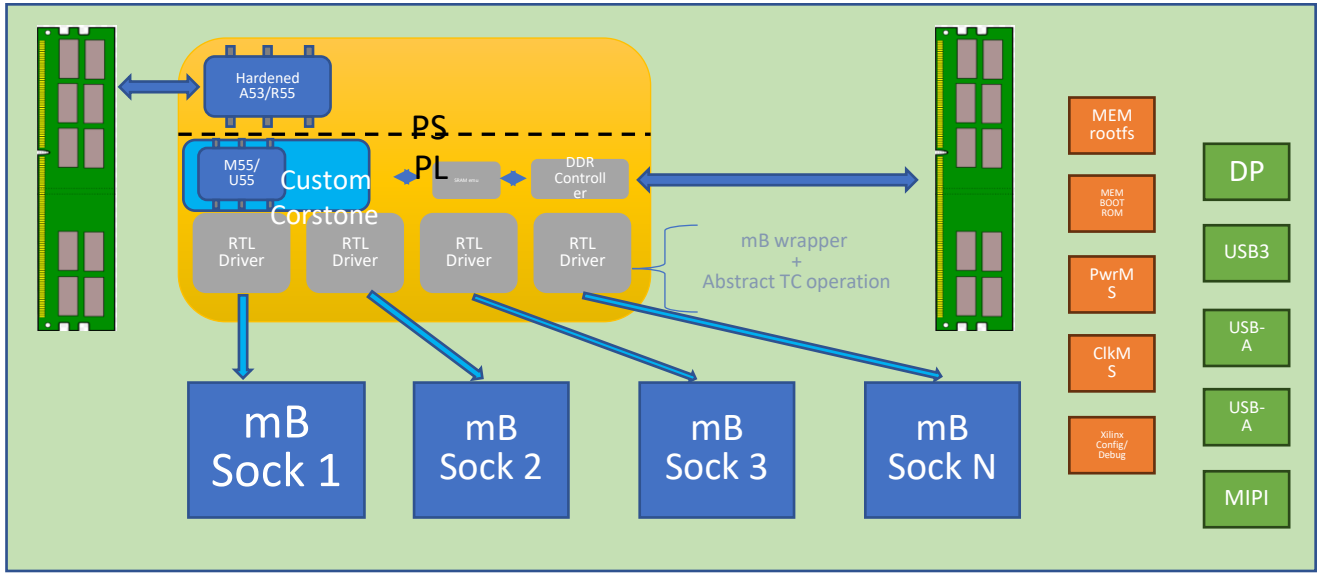


ESF3-28 **memBrain** Silicon 2

- Macro Level Features
  - 1.3MW, - 40C to 105C
  - 8b DAC/ADC – Fast speed ~us
  - 3-stage pipelined data path
  - Redundancy
- Silicon in Characterization

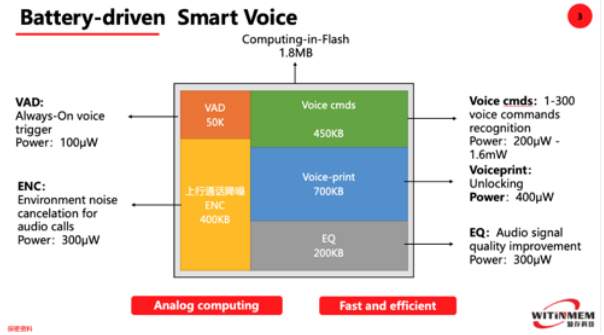
- TC2 Demo Board:
  - 10 TCs on one board
  - Target neural nets:
    - E.g., Object Det.
      Image classifier

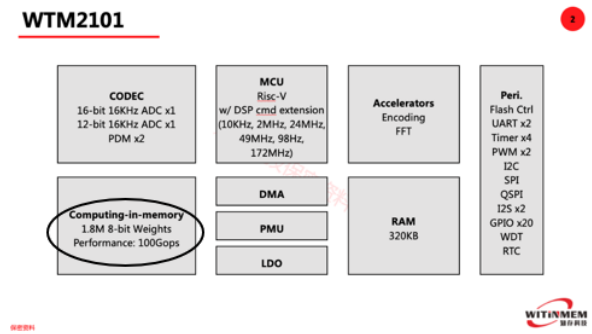




Commercial examples for A-CIM **memBrain**

- **Customer A** Audio Processor



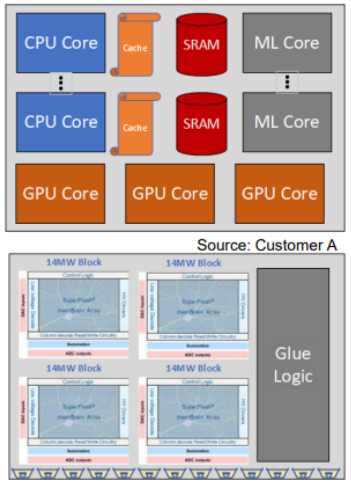


Production on UMC 40nm of Audio "Smart Voice" device

7 sq mm die very similar in function to Syntiant's latest but much lower power

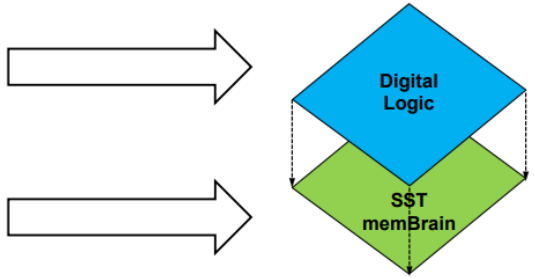
Multiple China based customers designed in the product

Video Processor



Source: Customer A

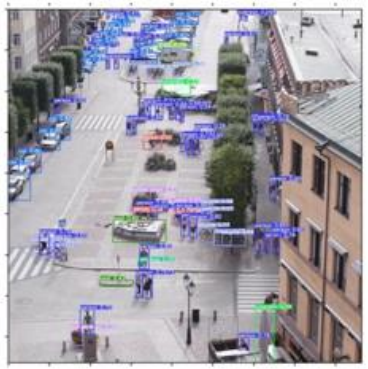
- Wafer-to-wafer bonding 12nm digital chiplet to 28nm memBrain chiplet (same size)



- 64M analog weight capacity
- Upscales 4K video to 8K video and other video processing capabilities
- 2x better power/perf than competitor part
- Targets TVs and Mobile phones

- **Customer B**
  - Object detection
  - Pose detection

Source: Mythic



YOLOv5s6 @1408x1408

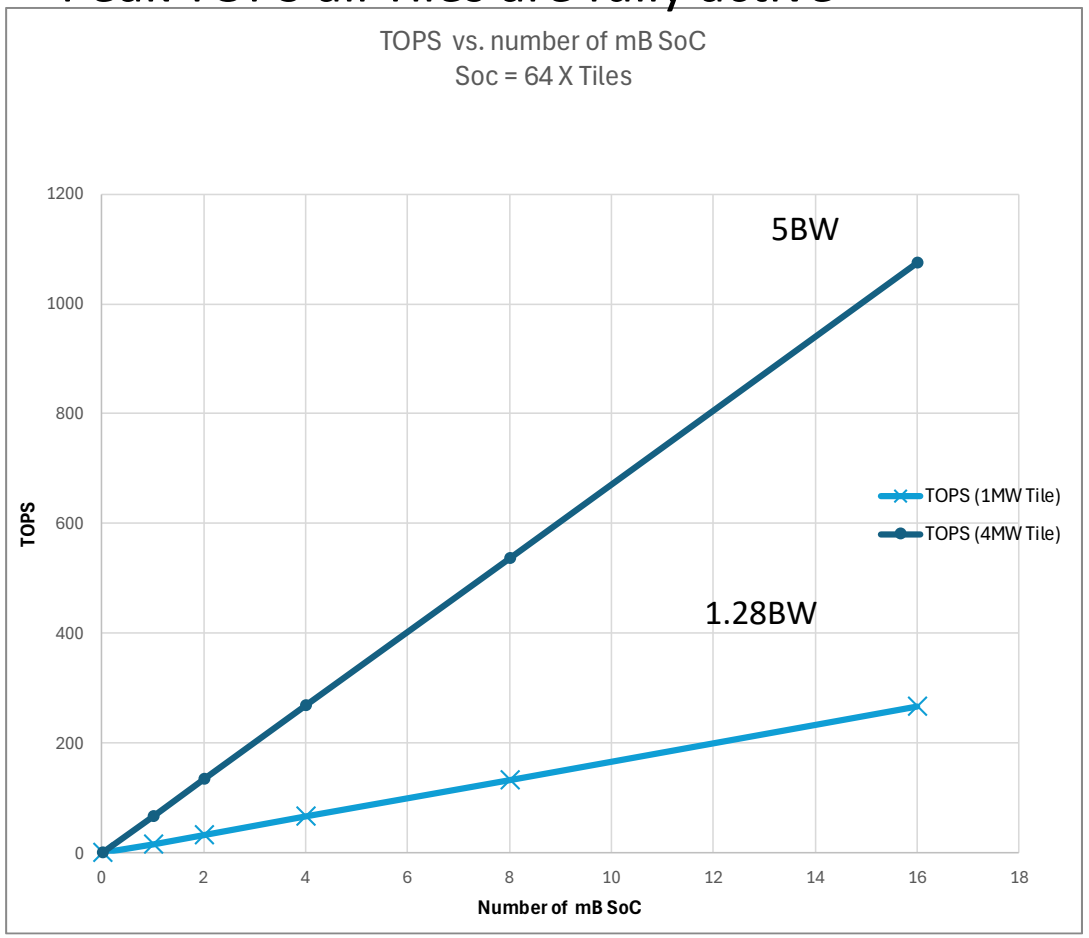
OpenPose Body25 @90fps

memBrain Chiplet Architecture

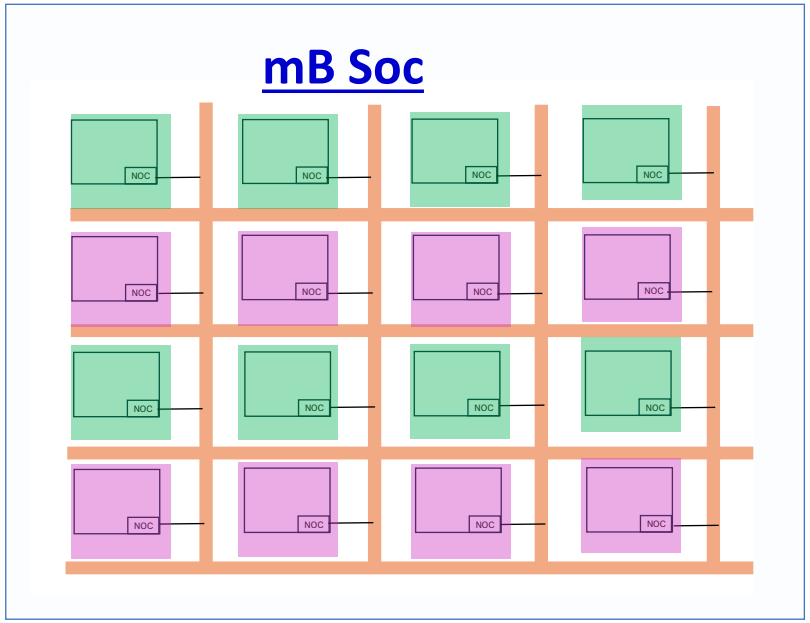
- Tile modulization:
  - Allow system performance scaling-up or scaling-down and functional aggregate
  - Easiness of extending weight capacity for different applications
- Homogenous technology for functional aggregate
- Heterogeneous technology for different chiplet, e.g., finfet for uC and non-finfet for A-CIM
- Less thermal concern due to low power nature of the memBrain technology
- Mixed tile architecture (A-CIM, D-CIM, Digital Accelerator)
- D-CIM and/or Dig Acc for dynamic weight matrix and A-CIM for fixed weight matrix such as for Transformer network
- D-CIM and/or Dig Acc for low dimension tensor and A-CIM for high dimension tensor operation

memBrain Chiplet Architecture

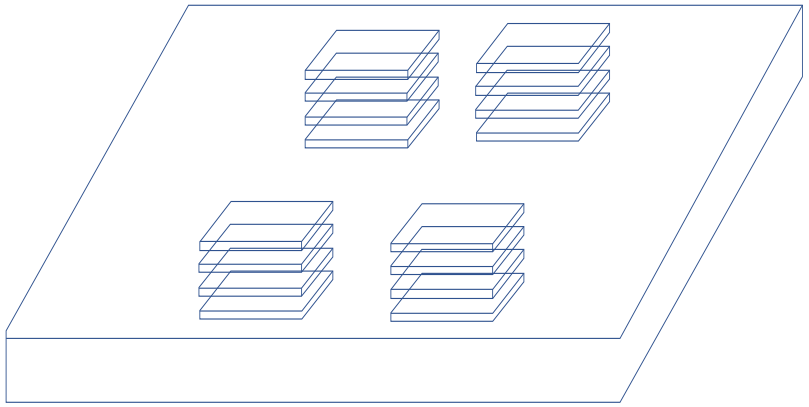
Peak TOPS all Tiles are fully active



mB Soc



mB Sys (mB SoP)



ESF3 **memBrain** Roadmap

Available | **In** development | In plan/forecast

Market	Technology	2020	2021	2022	2023	2024	2025	2026	2027	2028
		Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4	Q1 Q2 Q3 Q4
Artificial Intelligence	memBrain™ IP	**memBrain** ESF1-180 / **memBrain** ESF3-40		**memBrain** ESF3-28		**memBrain ESF3-1x FinFET**				
Artificial Intelligence	memBrain™ Reference Board				**memBrain Ref Board ESF3-28**					
Artificial Intelligence	memBrain™ System					**memBrain sys ESFx**				

SST SuperFlash® applied to AI/ML edge inference