Enhancing Generative AI with 3D DRAM and Advanced Memory Architectures

Presenter : Ju Jin An

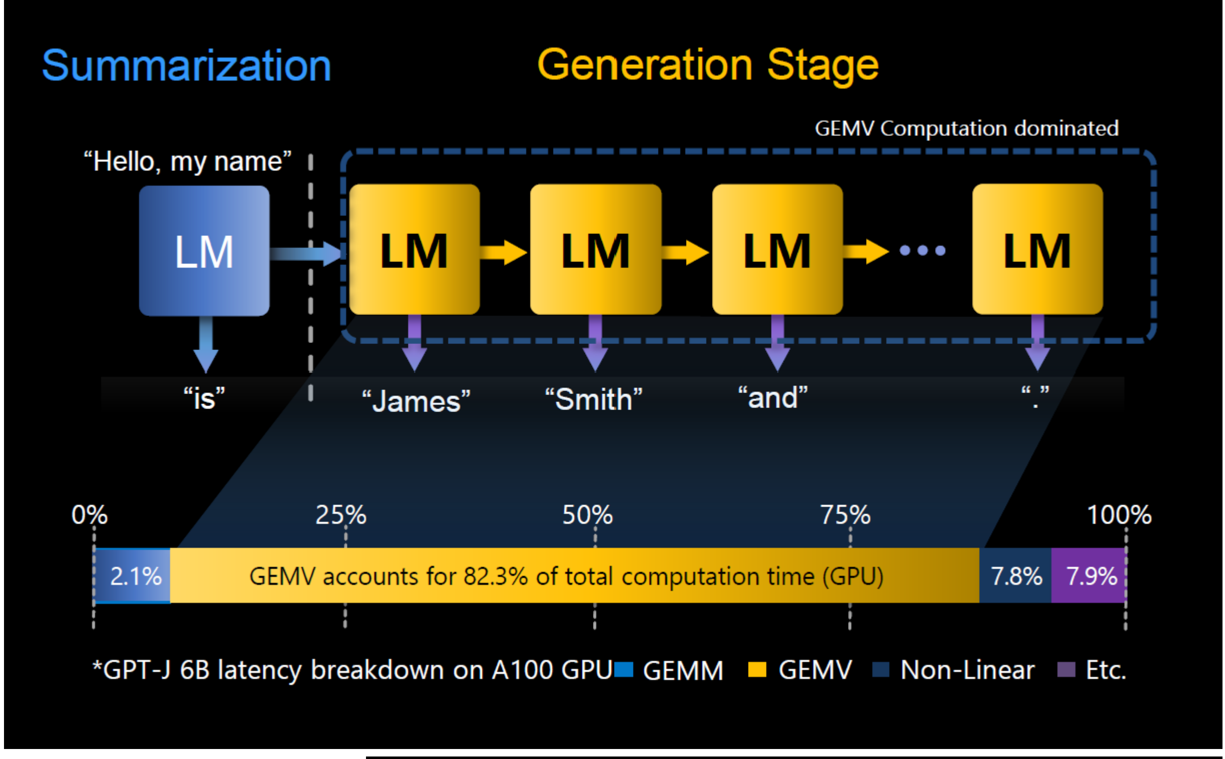
Supply Chain Engineering, Infrastructure, IBM

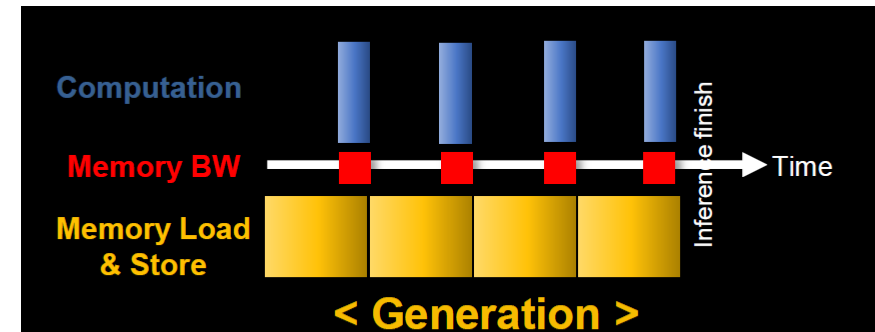
Outline

1. Generative AI and higher memory capacity requirement
2. Approximate computing without increasing memory capacity
3. How can we increase DRAM capacity?
   - TSV, Hybrid bonding
   - COP (Cell On Peri)
   - 1T1C 3D DRAM (4F2 IGZO VCT, VS CAT - Vertical BL, Vertical WL)
   - Capacitor-less 3D DRAM (3 STAR, GCT, 2T0C, X-DRAM)

Generative AI and higher memory capacity requirement

Illustration of transformer-based text generation





Inference stage (generation of responses) requires the highest memory capacity in generative AI

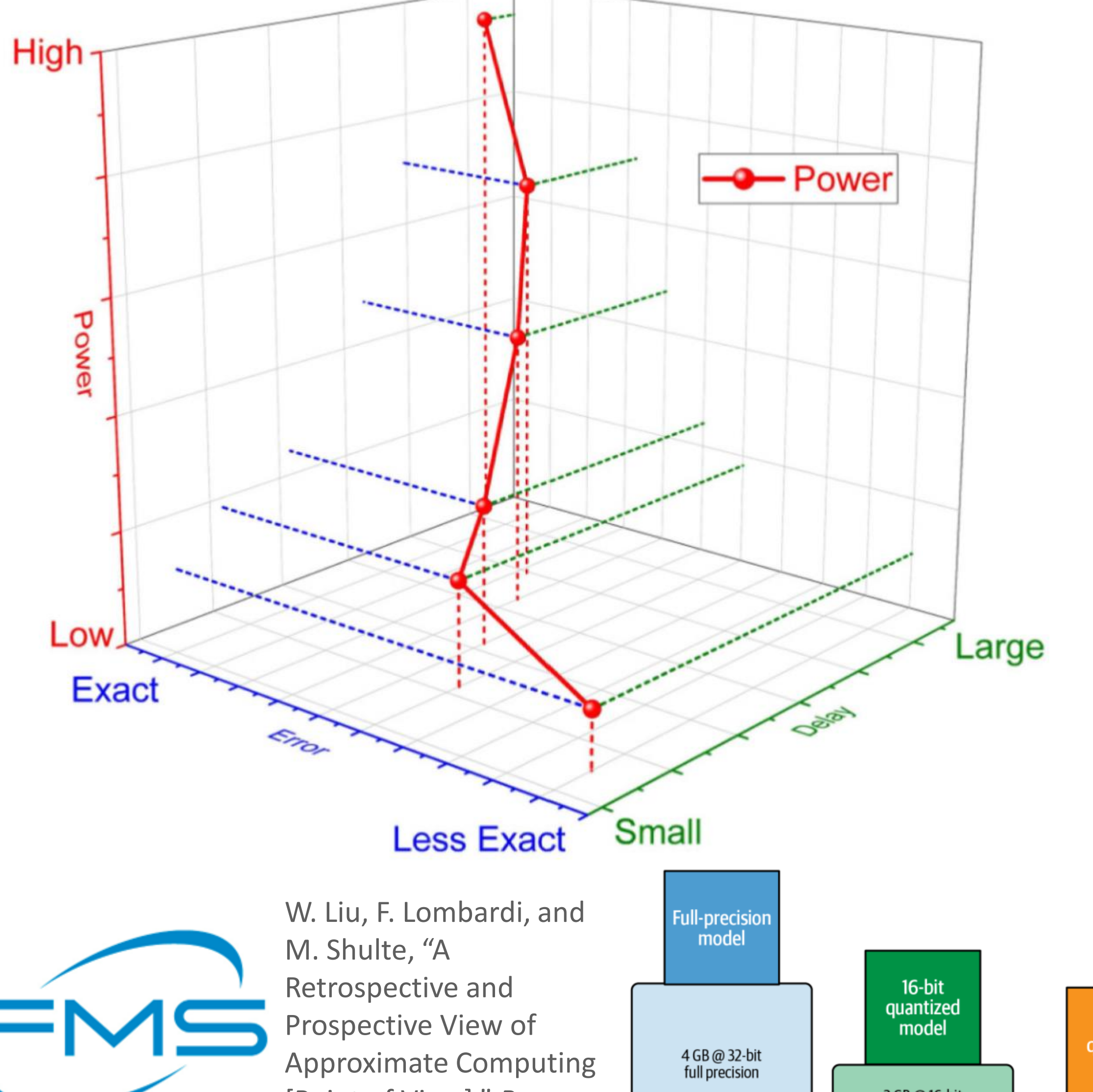
- Need to store large models
- Manage intermediate activations
- Handle long context windows
- Optimize the response generation process

Training stage

- Training stage is also memory-intensive due to backpropagation and gradient storage
- Usually performed from distributed computing and specialized hardware unlike inferencing

Approximate computing - providing significant benefits without memory capacity increase

Three-dimensional design space of approximate computing



W. Liu, F. Lombardi, and M. Shulte, “A Retrospective and Prospective View of Approximate Computing [Point of View],” *Proc. IEEE*.

Full-precision model

4 GB @ 32-bit full precision

16-bit quantized model

2 GB @ 16-bit half precision

8-bit quantized model

1 GB @ 8-bit precision

Not all AI applications require exact computation

- Many can tolerate some degree of approximation without significantly impacting the application’s functionality
- Many machine learning models can tolerate approximations in both training and inference phases

Trade-off between accuracy and memory / power consumption

- Challenge (Error Management) : Ensuring the introduced approximations do not lead to unacceptable levels of error is a critical challenge

Approximate computation in the design of IBM AIU

- Leaner bit formats - from 32-bit floating point arithmetic to bit-formats holding a quarter as much information
- Simplified format cuts down the amount of number crunching needed to train and run an AI model without sacrificing accuracy

How can we increase DRAM bit density to maximize memory capacity?

1. 2D shrink, High NA EUV

On-going

But, running out of steam

V. T. Pham *et al.*, “Patterning optimization for single mask bit-line-periphery and storage-node-landing-pad DRAM layers using 0.33NA EUV lithography at the resolution limit,” Apr. 2024, p. 29.

2. 3D stacking (TSV, COP, IGZO VCT, VS-CAT, 2T0C, GCT, 3D X-DRAM, etc.)

Technology is not mature yet

(limited number of publications)

TSV and hybrid bonding

Category	Item	HBM1	HBM2	HBM2E	HBM3	HBM3E	HBM4
	Time	2014	2018	2020	2022	2024	2026
General	Die Density	2Gb	8Gb	16Gb	16Gb	24Gb	24Gb
	Max. Bandwidth	128GB/s	0.3TB/s	0.5TB/s	0.7TB/s	1.18TB/s	1.65TB/s
Configuration	Max. Stack Height	4HI	8HI	8HI	12HI	12HI	16HI
	Capacity	1GB	8GB	16GB	24GB	36GB	48GB
	Total IO/Cube	1024	1024	1024	1024	1024	2048
Power	vDDC	1.2V	1.2V	1.2V	1.1V	1.1V	1.05V
	vPPE	2.5V	2.5V	2.5V	1.8V	1.8V	1.8V
	vDDQ	1.2V	1.2V	1.2V	1.1V	1.1VV	0.8V

	TC-NCF	MR-MUF	Hybrid bonding	Remark
Image				
Bonding Material	Micro Solder bump	Micro Solder bump	Cu-Cu	
Max Stack Height	Up to 12 Hi	Up to 12 Hi	4 ~ 16 Hi	HBM cube height : 720um
Min. Bump Pitch	~20um	~20um	<20um	
Relative thermal resistance	1	0.5~0.64	0.4~0.5	Depending upon metal portion, gap-fil material conductivity and thickness, and so on

- **TSV technology enables DRAM package level density up to 48GB per cube (HBM4) with 16 die stacking**
- **Challenges in HBM TSV package technology – trapped heat, leading to thermal degradation of DRAM**
  - HBM base die transition from DRAM-based to logic-based die to cut down power consumption
  - Advanced packaging solution (MR-MUF) provides excellent heat dissipation characteristic
- **Cu-Cu Hybrid bonding**
  - Eliminates micro solder bumps, allowing up to 16 die stacking

K. Kim and M. Park, “Present and Future, Challenges of High Bandwith Memory (HBM),” in *2024 IEEE IMW*, pp. 1–4.

COP (Cell On Core/Peri), Wafer Bonding

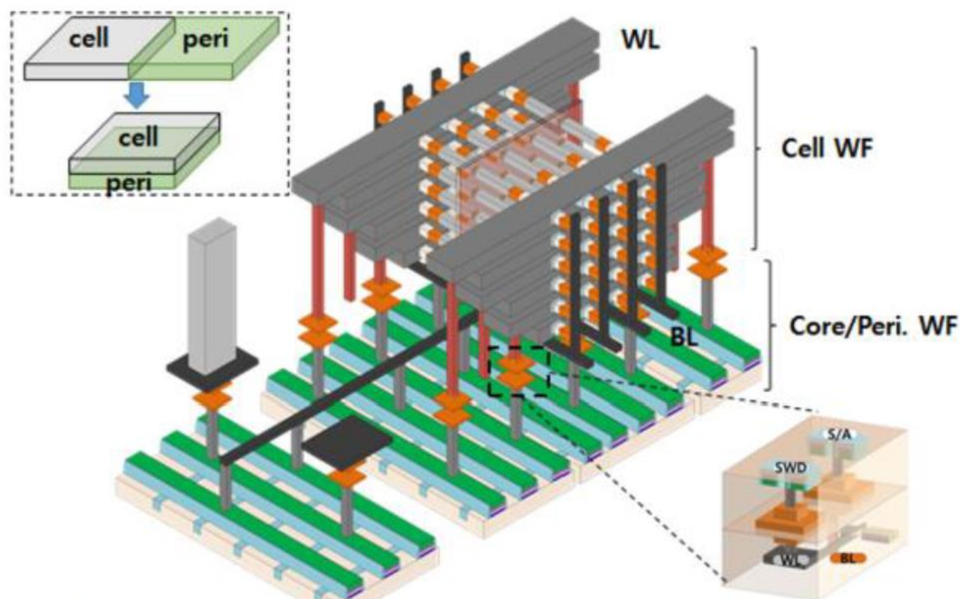
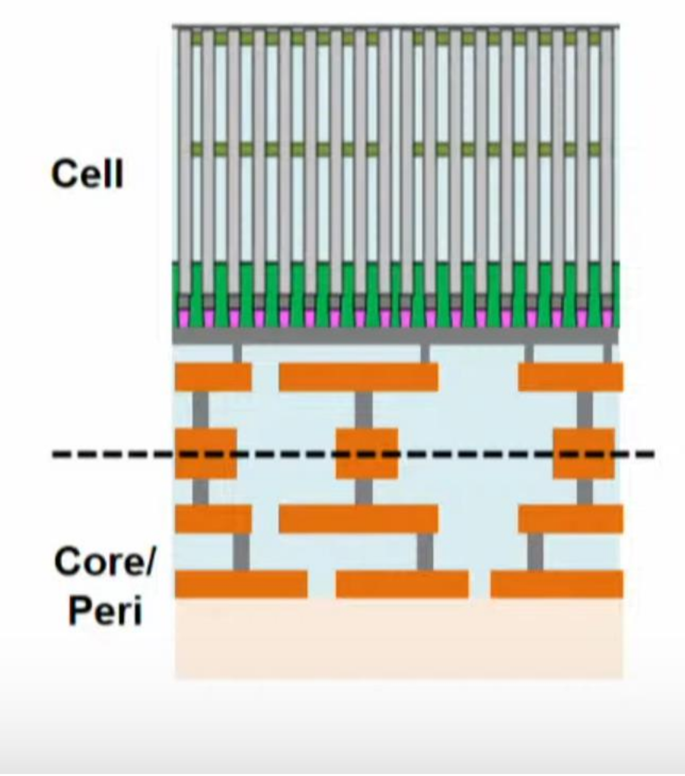


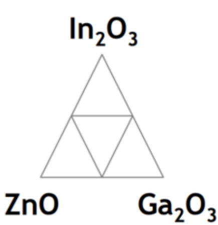
Fig. 8 A conceptual illustration of a cell-on-peri architecture with the peripheral layer placed beneath the cell layer. Hybrid copper bonding scheme connects the cell and core/peripheral interfaces.

J. W. Han *et al.*, "Ongoing Evolution of DRAM Scaling via Third Dimension - Vertically Stacked DRAM -," in *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits),* pp. 1–2.



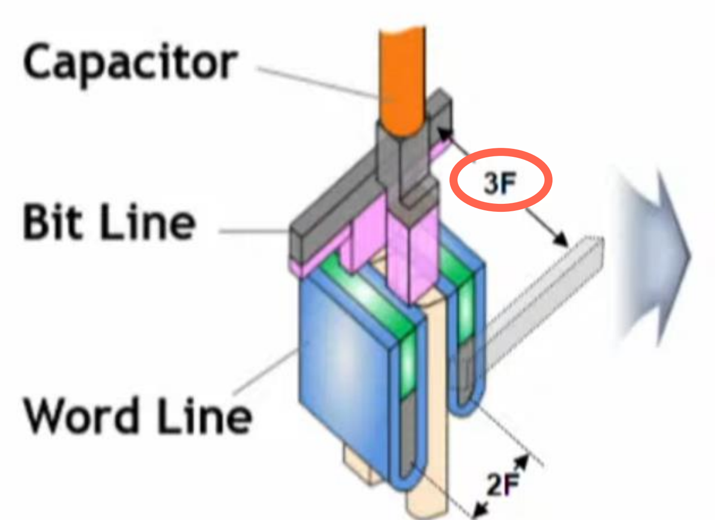
- **Maximize Cell array area by removing Core/Peri from the same plane**
- **Wafer bonding helps avoid incompatibility in fab processing & thermal budget between Cell and Peri**

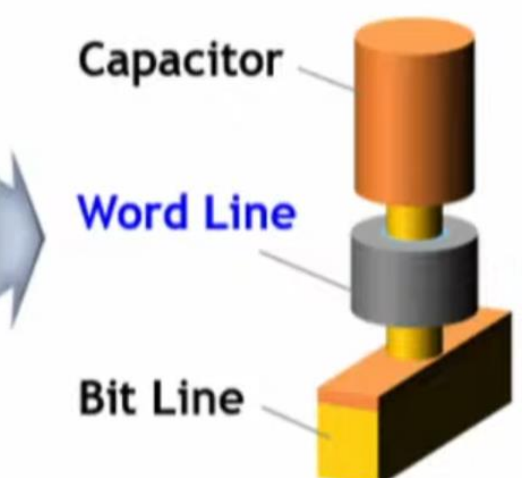
Transition from 6F2 RCAT to 4F2 VCT

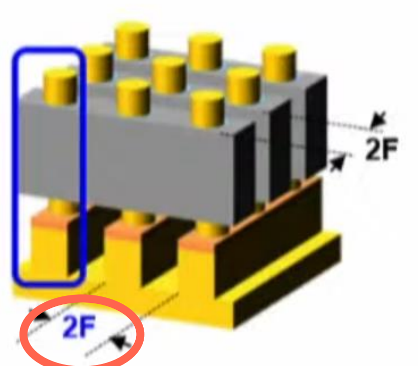


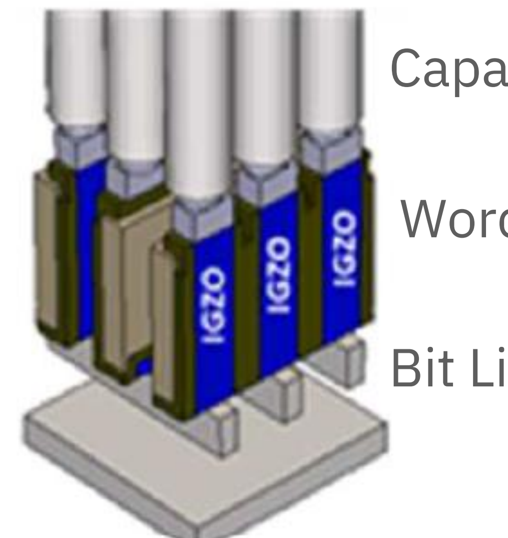
Conv. BCAT ($6F^2$)

Vertical Channel Transistor ($4F^2$) (e.g., IGZO VCT)



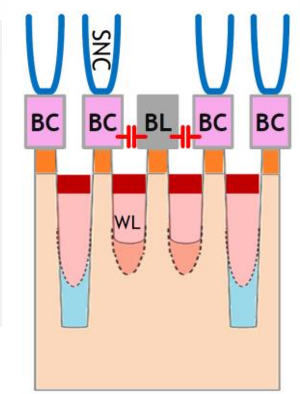






BL and SN contacts are in the same plane

- 3F Bit Line pitch is needed to avoid signal interference between BL and SN



BL and SN contacts are placed out of plane

- Elongates the space between BL and SN contact, enabling 4F2

Immune to Rowhammer

- In 4F2, adjacent cells do not share active region

Floating Body Effect → Charge in the capacitor is lost due to transient leakage

- Floating Body Effect is suppressed by IGZO channel thanks to negligible hole generation and a large bandgap

1) RCAT (Recessed Channel Array Transistor)
2) VCT (Vertical Channel Transistor)

1T1C 3D DRAM– VS CAT (Vertical BL, Vertical WL)

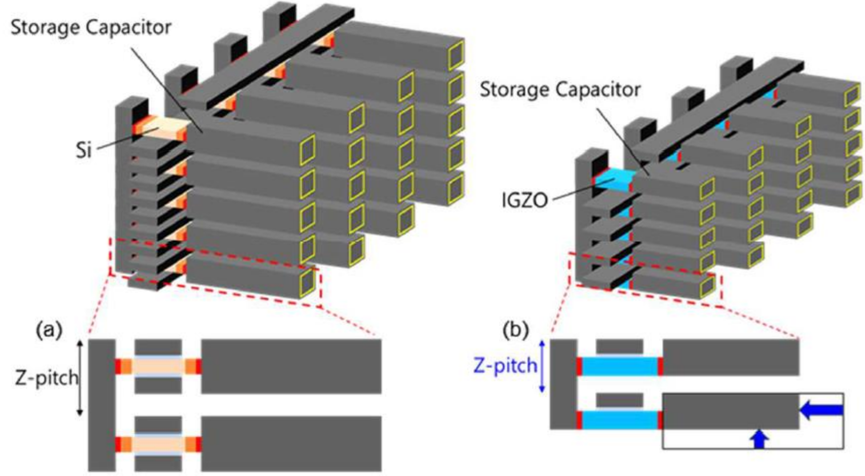
	3D 1T1C: vertical BL	3D 1T1C: vertical WL
Schematic	Z X Y; Metal BL; To WL driver; cap; Z-dir WL separation; BL plugs	Metal BL; To WL driver; cap; Tight-pitch BL staircase
Physics	1T1C charge-sharing sensing	1T1C charge-sharing sensing
Capacitor	with	with
Staircase	WL	BL
BEOL complexity	modest	difficult tight-pitch (X) staircase BL process for layer sensing
X-pitch scalability	good	poor (WL cut, staircase BL)
Y-pitch scalability	poor (with large foot print of cap)	poor (with large foot print of cap)
Z-pitch scalability	poor (Z-dir WL separation)	good
Ideal cell size/pitch	X-pitch~40nm; Y-pitch>1um; Z-pitch >100nm	X-pitch>80nm; Y-pitch>1um; Z -pitch ~40nm

W.-C. Chen *et al*, "A Highly Pitch-Scalable Capacitor-less 3D DRAM Using Cross-bar Selection with Gate-Controlled Thyristor (GCT) Featuring High Endurance and Free Read-Disturb," in *2023 IEDM*, pp. 1–4.

FMS

1) VS CAT (Vertically Stacked Cell Array Transistor)

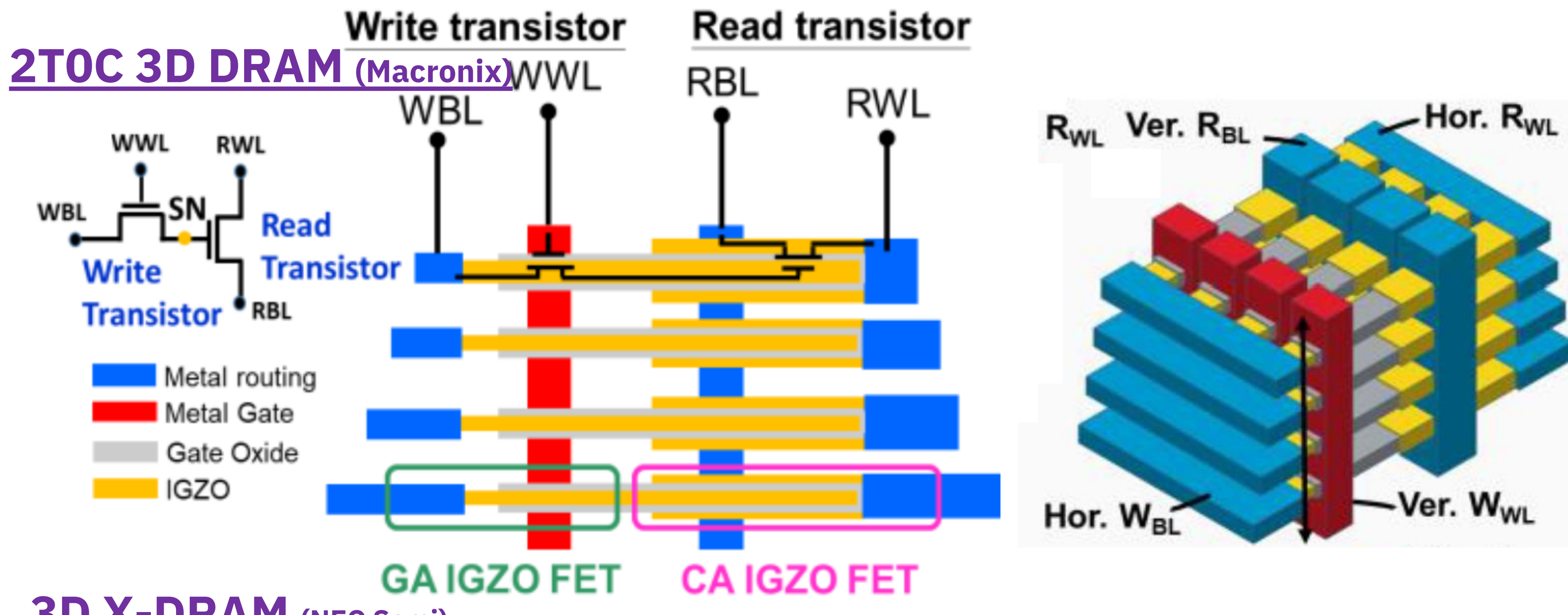
- **Large footprint of capacitor**
  - IGZO (new channel material, extremely low Ioff) can reduce capacitor size
- **Vertical BL scheme occupies less footprint than vertical WL scheme, but requires taller cell block height due to more components in vertical direction**
- **Seed layer for epitaxial silicon channel to guarantee high quality stacked channel**
  - IGZO can eliminate the need of epi Si



D. Ha *et al.*, "Exploring Innovative IGZO-channel based DRAM Cell Architectures and Key Technologies for Sub-10nm Node," in *2024 IEEE IMW*, pp. 1–4

Capacitorless DRAM - Bit cost scalable 3D 2T0C IGZO DRAM, 3D X-DRAM

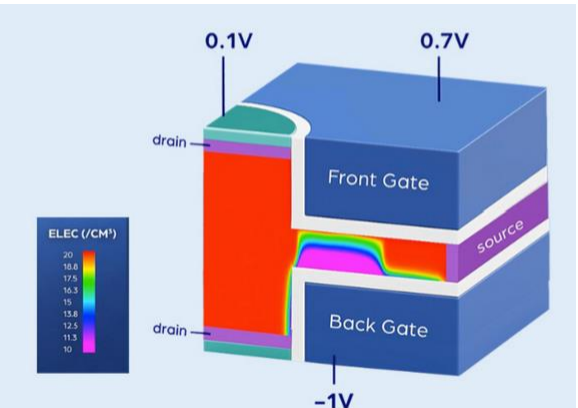
2T0C 3D DRAM (Macronix)



- **IGZO - high mobility, ultra low leakage, and low process temperature**
- **Conventional 2T0C IGZO memory devices are non-cost-scalable**
- **IGZO gate and IGZO channel to form 3D 2T0C DRAM**
- **Prototype device showed long retention (180s)**

F.-M. Li *et al*, "Bit-cost-scalable 3D DRAM Architecture and Unit Cell First Demonstrated with Integrated Gate-around and Channel-around IGZO FETs," in 2024 VLSI.

3D X-DRAM (NEO Semi)



* Based on TCAD Simulations

	3D X-DRAM	2D FBC	2D DRAM
Sensing window	20 uA	2 uA	10 mV
Retention Time (85C)	200 ms	5 us	> 64 ms
Operation Voltage	1V	3V	1V
Write time (cell level)	200 ps	200 ps	5 ns
Endurance Cycles	> 10^{16}	n/a	> 10^{16}
Maximum Density	>> 128 Gb	n/a	< 48 Gb

- **Unique dual gate (Front/Back) to resolve high coupling issue between front gate to floating cell**
  - Back gate attracts holes in the floating cell, increasing data retention time and sensing window
  - Large sensing window allows the cell to achieve high read speed and high noise immunity
- **Thin body cell structure**
- **Only TCAD simulation results are published (2024 IMW)**

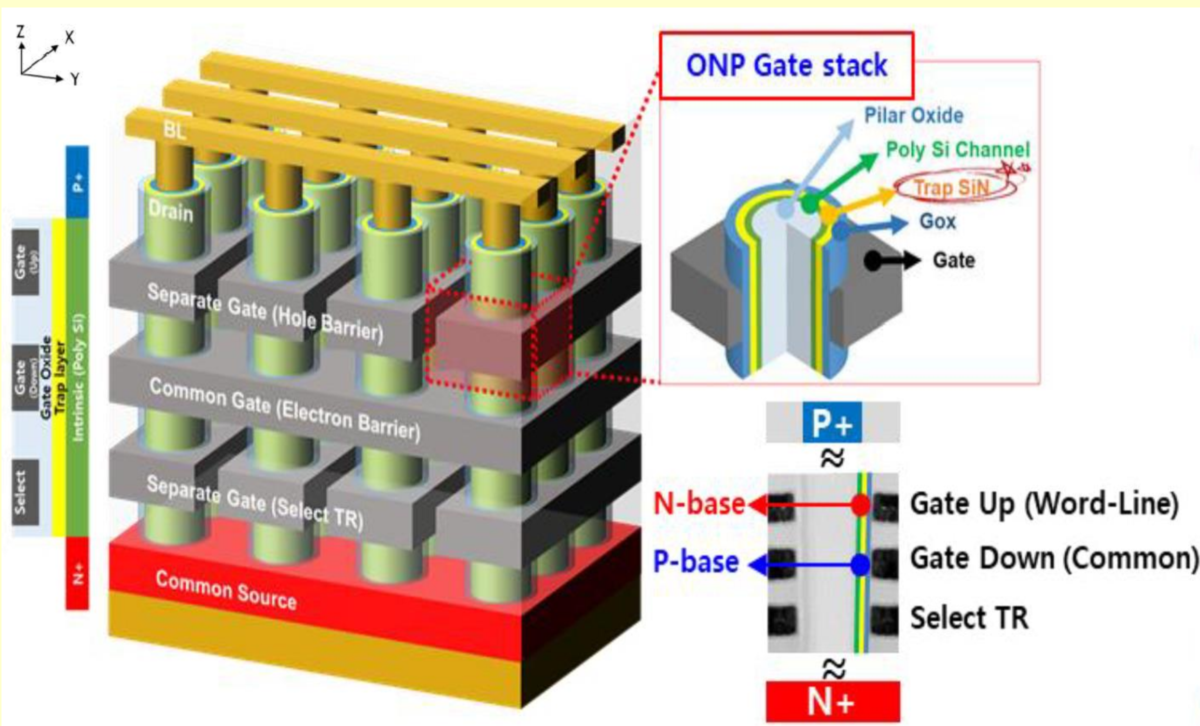
FMS

1) GA FET – Gate Around FET
2) CA FET – Channel Around FET

F.-C. Hsu *et al.*, "3D X-DRAM: A Novel 3D NAND-like DRAM Cell and TCAD Simulations," in *2024 IEEE IMW*, pp. 1–4.

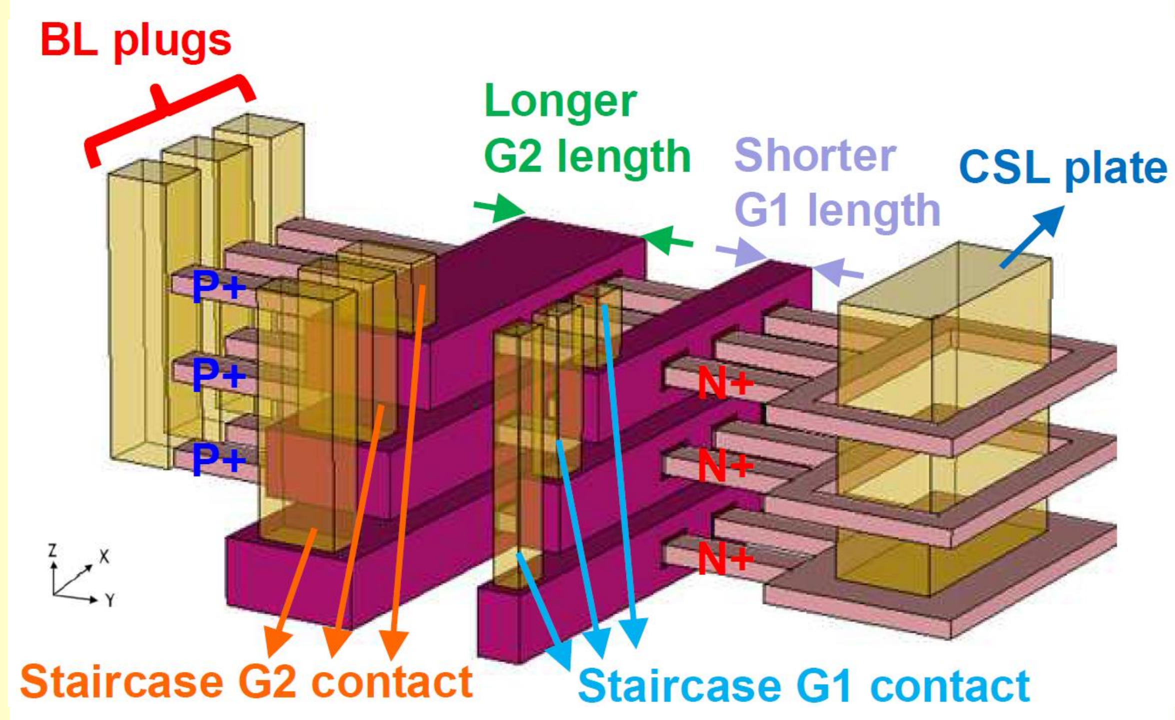
Capacitorless DRAM using Thyristor Positive Feedback FET - 3 STAR, 3D X-Bar GCT

3-STAR (Samsung)



K. Lee *et al.*, "3-STAR: A Super-steep switching, Stackable, and Strongly Reliable Transistor Array RAM for Sub-10nm DRAM and beyond," in *2023 IEDM*, pp. 1–4.

3D X-Bar GCT (Macronix)



W.-C. Chen *et al*, "Improved 3D DRAM Design Based on Gate-Controlled Thyristor Featuring Two Asymmetrical Horizontal WL's and Vertical BL for Better Cell Size Scaling and Array Selection," in *2024 IEEE IMW*, pp. 1–4.

Excellent Retention (100s) – longer carrier lifetime

- Low recombination probability of carriers stored in the SiN trap

No need to grow epitaxial Si for vertical stacking

- 3 STAR uses deposition poly silicon

However, separate gate/common gate increases process complexity

Nanosecond-level read/write speeds

Retention exceeding 10sec, endurance exceeding 1E10 cycles

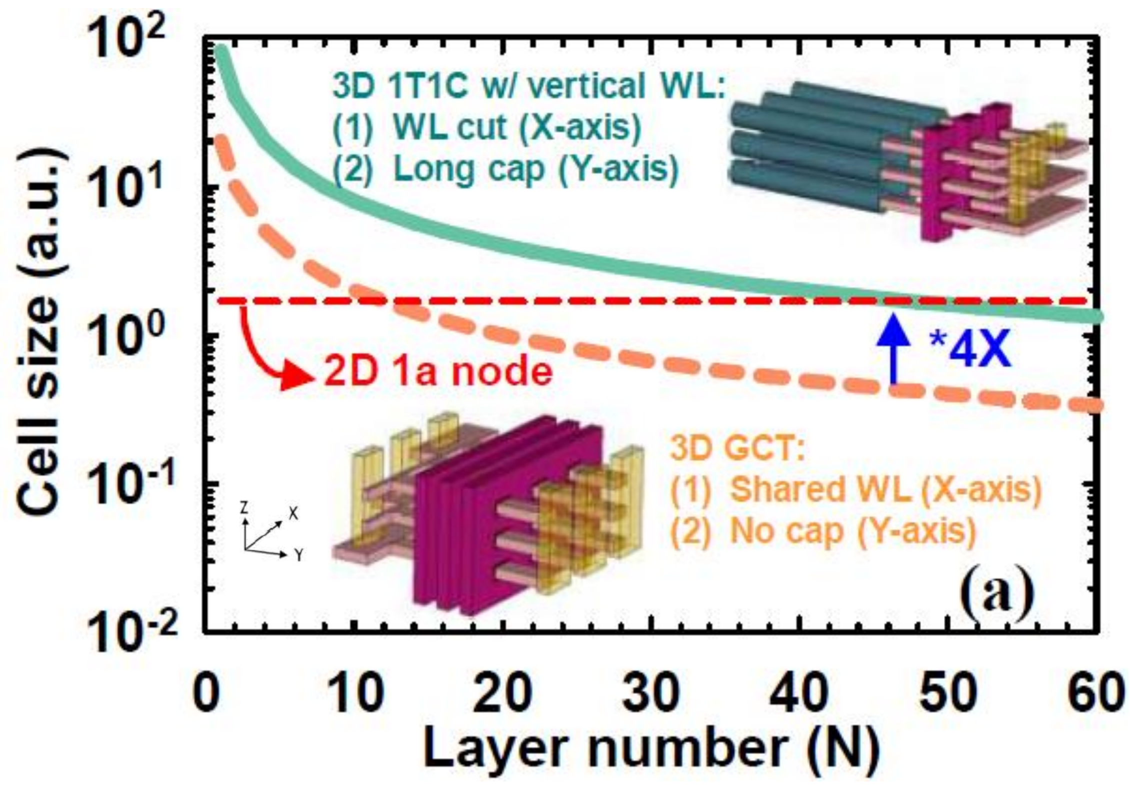
Virtual junction is controlled by gate biases

- Instead of complex physical junction dopant engineering

FMS

1) 3-STAR - Super-steep switching, Stackable, and Strongly Reliable Transistor Array RAM
2) 3D Cross Bar GCT (Gate Controlled Thyristor)

How many layers/stacks are needed to overcome 2D DRAM scaling challenge?



W.-C. Chen, *et al*, "A Highly Pitch-Scalable Capacitor-less 3D DRAM Using Cross-bar Selection with Gate-Controlled Thyristor (GCT) Featuring High Endurance and Free Read-Disturb," in *2023 IEDM*, pp. 1–4.

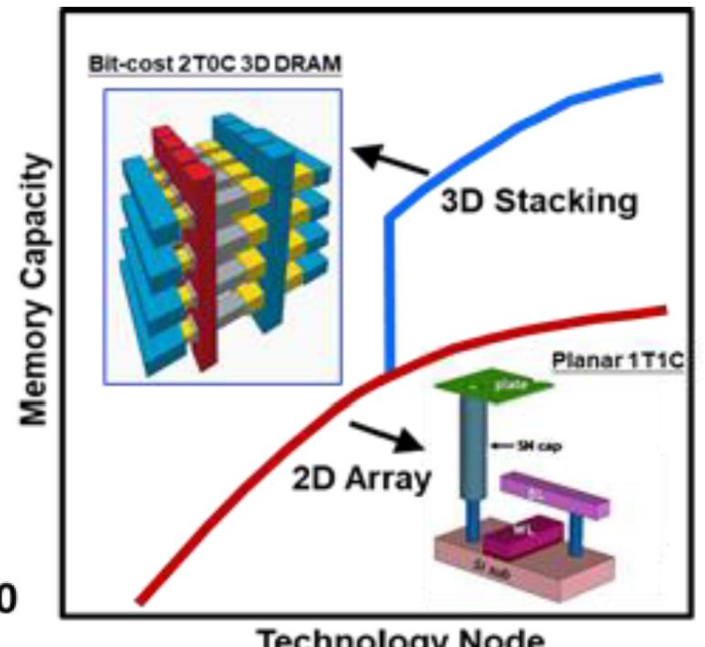
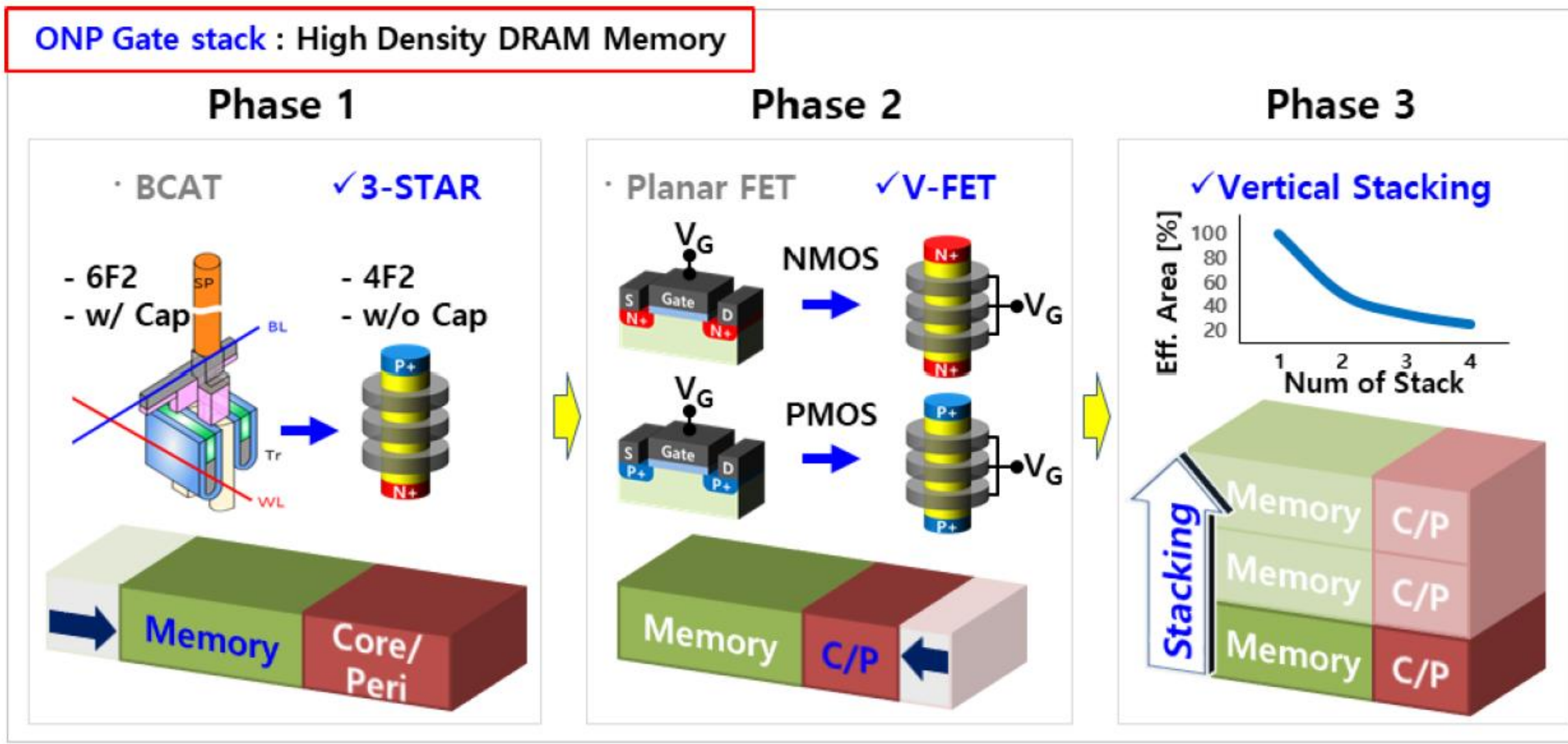


Fig.17 The scaling trend forecast of 2T0C IGZO 3D DRAM cell.

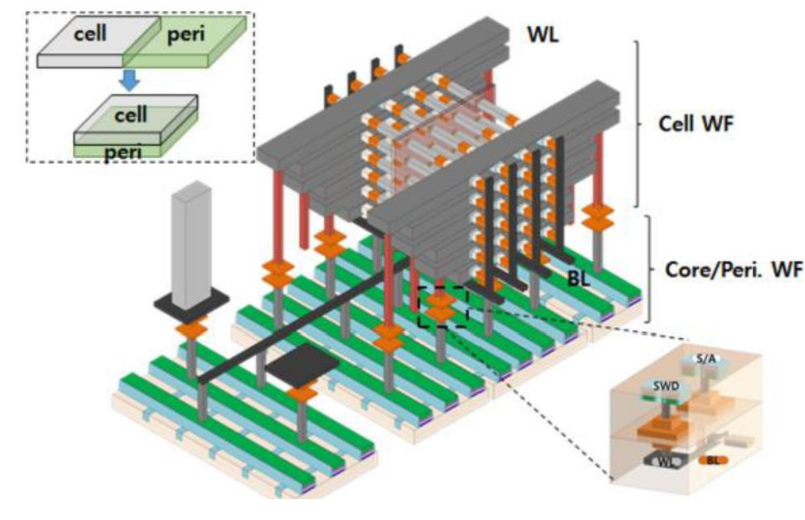
F.-M. Li *et al*, "Bit-cost-scalable 3D DRAM Architecture and Unit Cell First Demonstrated with Integrated Gate-around and Channel-around IGZO FETs," in 2024 VLSI.

- **3D 1T1C VS-CAT requires 50+ layers to catch up with conventional 1a nm technology**
  - Capacitor takes up too much space
  - IGZO channel can provide longer retention, potentially reducing the size of capacitor
- **Capacitorless GCT or bit-cost scalable 2T0C 3D DRAM may require less number of layers (+20~30 layers)**

Future DRAM scaling



COP/Wafer bonding



- **4F2**
- **New channel material (IGZO)**
- **3D DRAM (VCT, VS-CAT, GCT, 2T0C, X-DRAM)**
- **Wafer bonding (COP) and/or Peri V-FET**

K. Lee *et al.*, "3-STAR: A Super-steep switching, Stackable, and Strongly Reliable Transistor Array RAM for Sub-10nm DRAM and beyond," in *2023 IEDM*, pp. 1–4.

Reference

[1] W. Liu, F. Lombardi, and M. Shulte, “A Retrospective and Prospective View of Approximate Computing [Point of View,” Proc. IEEE.

[2] V. T. Pham et al., “Patterning optimization for single mask bit-line-periphery and storage-node-landing-pad DRAM layers using 0.33NA EUV lithography at the resolution limit,” Apr. 2024, p. 29. doi: 10.1117/12.3010934.

[3] J. W. Han et al., “Ongoing Evolution of DRAM Scaling via Third Dimension -Vertically Stacked DRAM -,” in 2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits), 2023, pp. 1–2. doi: 10.23919/VLSITechnologyandCir57934.2023.10185290.

[4] W.-C. Chen, H.-T. Lue, M.-H. Wu, Y.-T. Lin, K.-C. Wang, and C.-Y. Lu, “A Highly Pitch-Scalable Capacitor-less 3D DRAM Using Cross-bar Selection with Gate-Controlled Thyristor (GCT) Featuring High Endurance and Free Read-Disturb,” in 2023 International Electron Devices Meeting (IEDM), 2023, pp. 1–4. doi: 10.1109/IEDM45741.2023.10413828.

[5] K. Kim and M. Park, “Present and Future, Challenges of High Bandwith Memory (HBM),” in 2024 IEEE International Memory Workshop (IMW), 2024, pp. 1–4. doi: 10.1109/IMW59701.2024.10536972.

[6] K. Lee et al., “3-STAR: A Super-steep switching, Stackable, and Strongly Reliable Transistor Array RAM for Sub-10nm DRAM and beyond,” in 2023 International Electron Devices Meeting (IEDM), 2023, pp. 1–4. doi: 10.1109/IEDM45741.2023.10413741.

[7] D. Ha et al., “Exploring Innovative IGZO-channel based DRAM Cell Architectures and Key Technologies for Sub-10nm Node,” in 2024 IEEE International Memory Workshop (IMW), 2024, pp. 1–4. doi: 10.1109/IMW59701.2024.10536968.

[8] F.-C. Hsu et al., “3D X-DRAM: A Novel 3D NAND-like DRAM Cell and TCAD Simulations,” in 2024 IEEE International Memory Workshop (IMW), 2024, pp. 1–4. doi: 10.1109/IMW59701.2024.10536979.

[9] C. Chen et al., “First Demonstration of Stacked 2T0C-DRAM Bit-Cell Constructed by Two-Layers of Vertical Channel-All-Around IGZO FETs Realizing 4F2 Area Cost,” in 2023 International Electron Devices Meeting (IEDM), 2023, pp. 1–4. doi: 10.1109/IEDM45741.2023.10413790.

[10] W.-C. Chen, H.-T. Lue, M.-H. Wu, Y.-T. Lin, K.-C. Wang, and C.-Y. Lu, “Improved 3D DRAM Design Based on Gate-Controlled Thyristor Featuring Two Asymmetrical Horizontal WL’s and Vertical BL for Better Cell Size Scaling and Array Selection,” in 2024 IEEE International Memory Workshop (IMW), 2024, pp. 1–4. doi: 10.1109/IMW59701.2024.10536917.

FMS