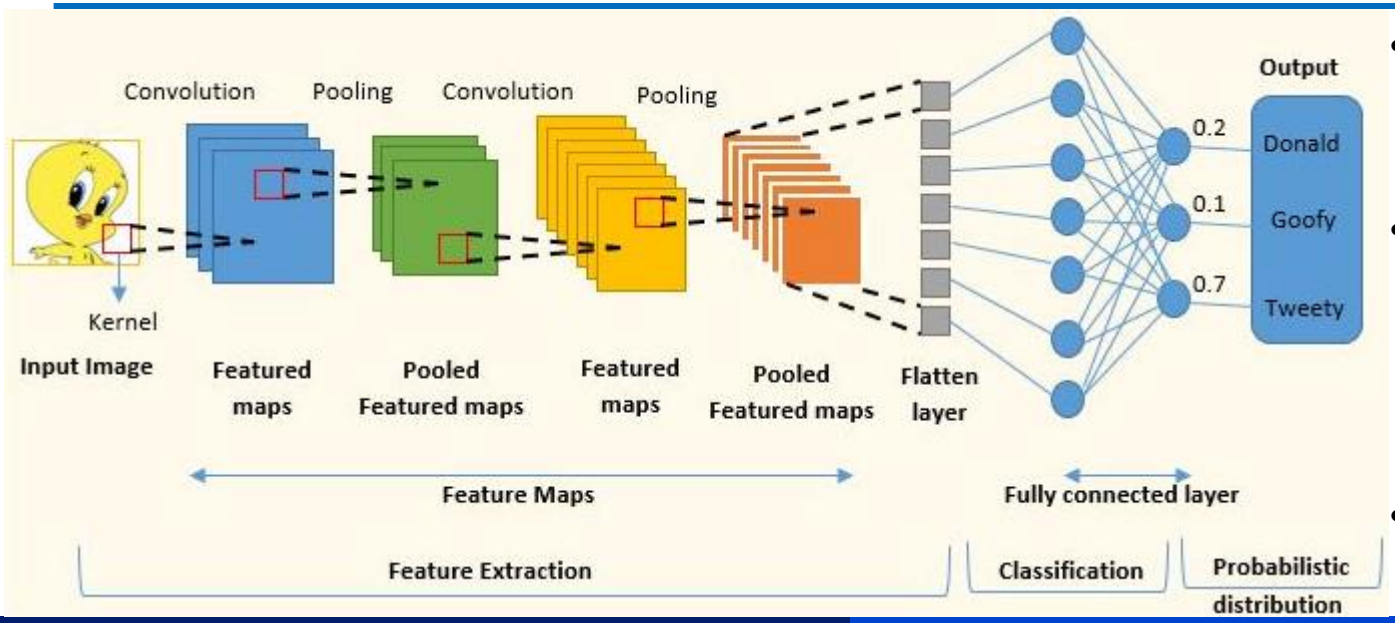
Flash Memory Summit

Memory solutions for secure and sustainable server infrastructure

Presenters: Ju Jin An, Jung Yoon, IBM Infrastructure

Memory consumption of deep learning neural network architectures – CNN(Convolutional Neural Network)



- **Significant areas of memory consumption**
  - Store **Weights, Biases (model parameters)**
  - Store **Activations** in the feature maps
- **Top 3 factors to impact the number of parameters**
  - Number of filters in convolutional layer
  - Depth of the network (number of layers)
  - The size of the input data and the number of neurons in the fully connected layers
- **Memory consumption (Training vs. Inference)**
  - Training consumes 2-3 more memory (backpropagation)

Convolutional Layer

- Feature extraction & detection
  - Deconstructing image into details
  - Filters act as feature detectors
- Example of features
  - Edge of a license plate number
- Dot product between two matrices to generate a feature map (activation)

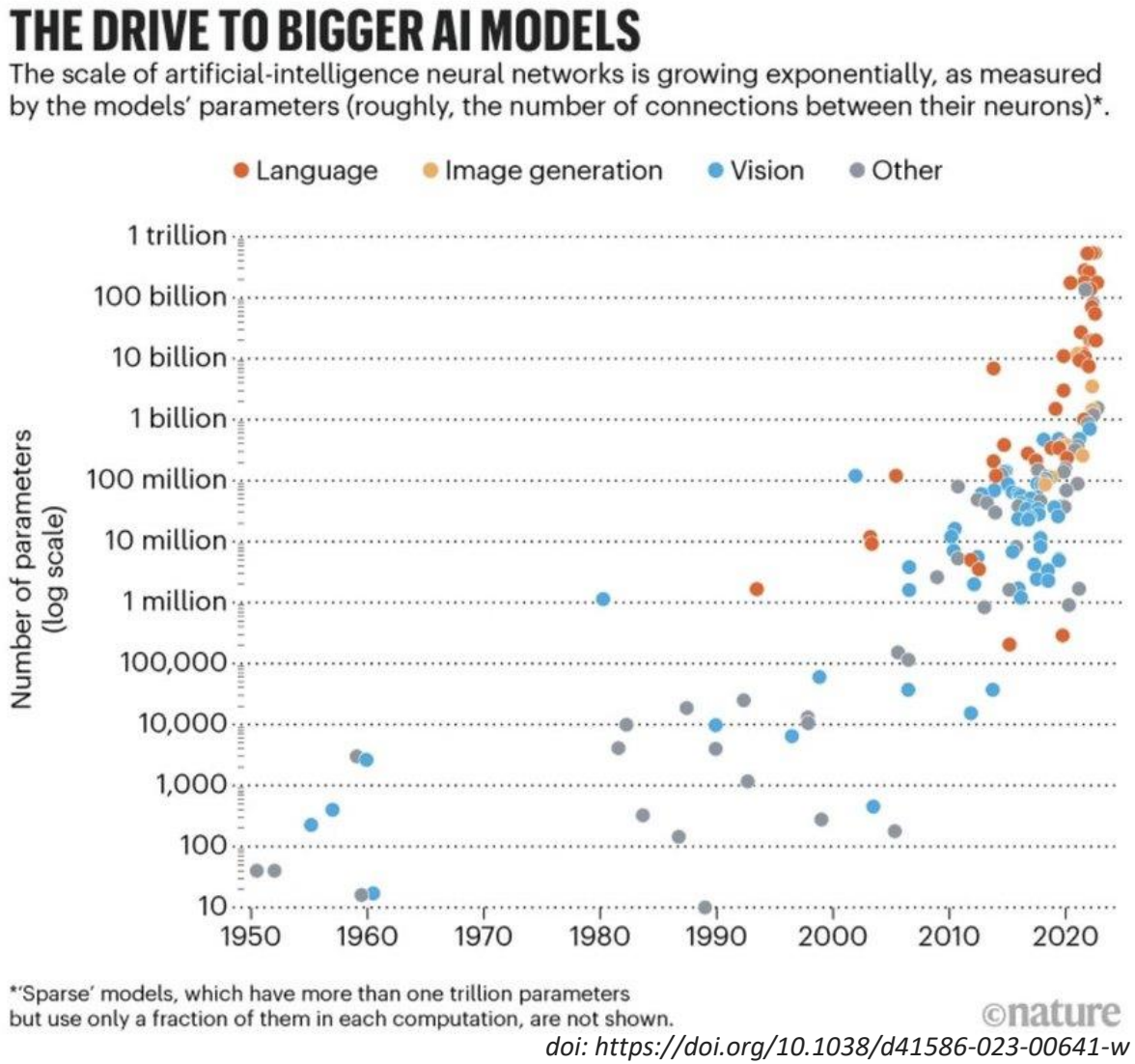
Pooling layer

- Figure out whether a particular region in the image has the feature we are interested in or not
- Pooling reduces only the height & width of the feature map
- Non training layer, not requiring memory consumption

Fully connected layer

- Performs image classification by tying all features together to see the total feature
- Input from previous layer feeds into the neurons in its layer and applies weights to predict the correct label

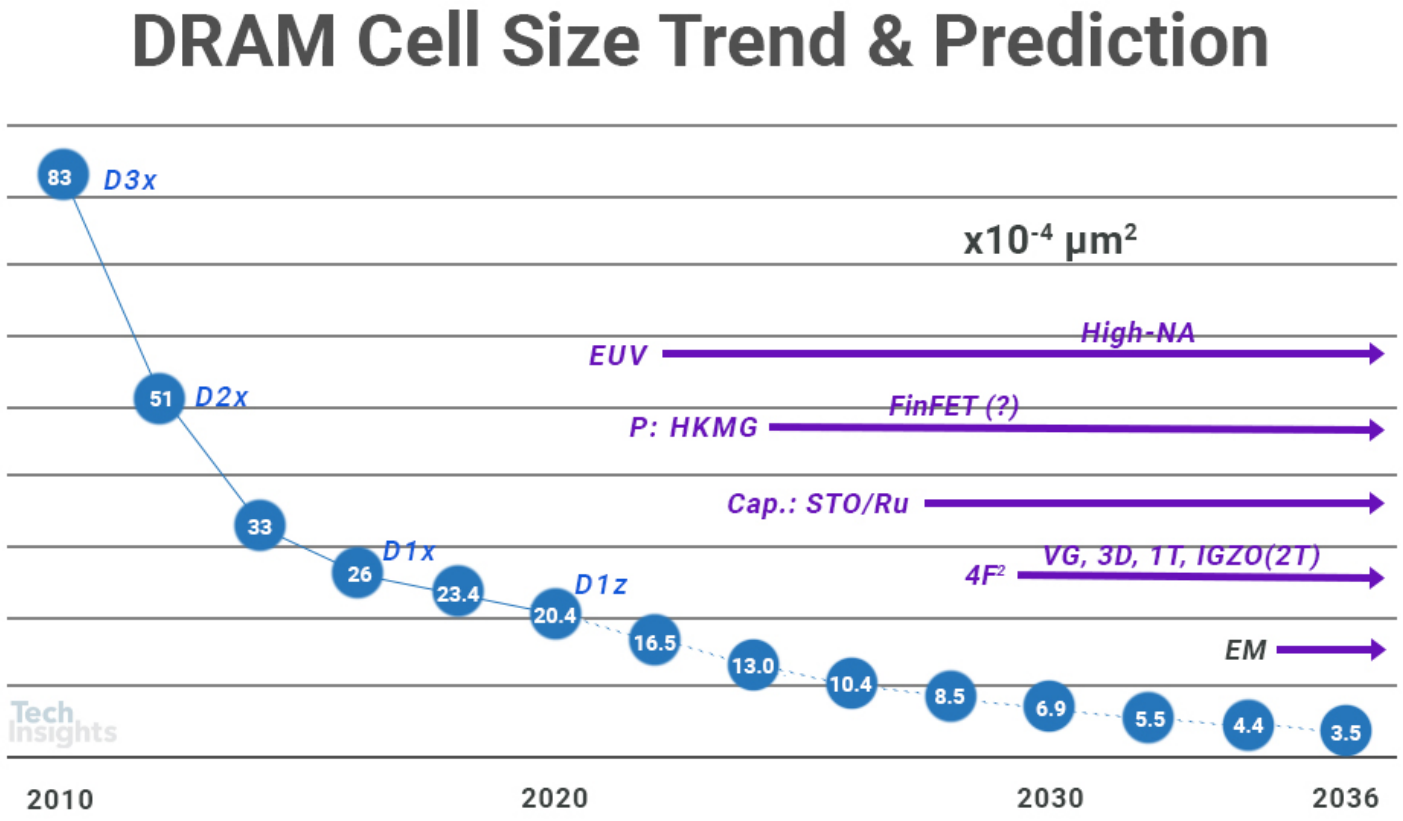
Growing parameters in deep neural networks and Impact on memory



doi: https://doi.org/10.1038/d41586-023-00641-w

- **Exponential increase in AI model size**
  - Larger parameters allow the model to capture finer details and nuances
  - More efficient pre-training
- **Impact on memory footprint / consumption**
  - High memory capacity
  - High memory bandwidth
  - Low latency
  - Low cost of ownership, sustainability
  - High level of security

DRAM scaling challenge and memory innovation

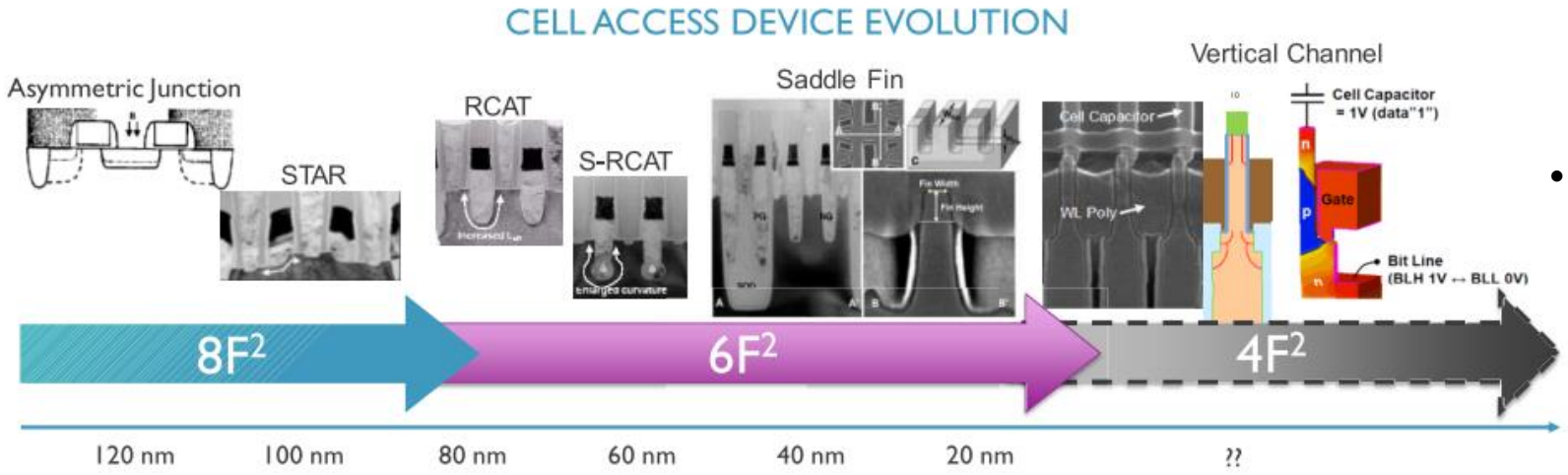


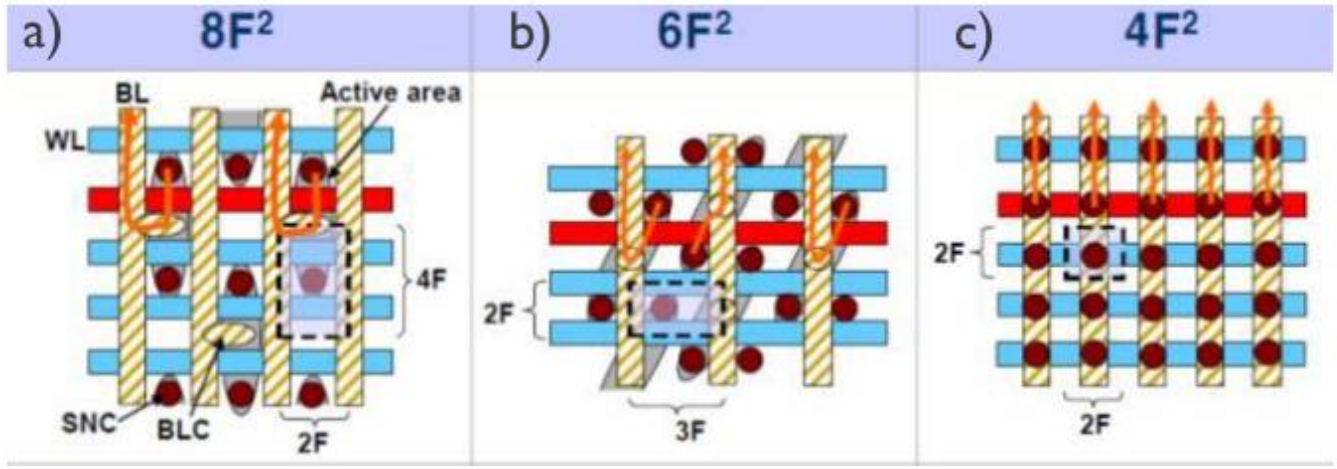
Source : https://www.techinsights.com/blog/dram-scaling-trend-and-beyond (Jeongdong Choe)

- **DRAM scaling has provided tremendous benefits**
  - Reduction in physical CD, Idd (power consumption), cost per bit
  - Reason for DRAM to become main memory
- **2D lateral scaling is losing steam**
  - Physical space between memory cells becomes too close to cause unnecessary interference
  - EUV increases overall manufacturing cost
  - Gate dielectric becoming too thin causing leakage
  - Vth variation (Wafer to Wafer, Within Wafer)
- **Innovation in memory technology needs to catch up with CPU/xPU improvement to prevent system performance bottleneck**

3D-DRAM

6F2 → 4F2 transition in cell architecture

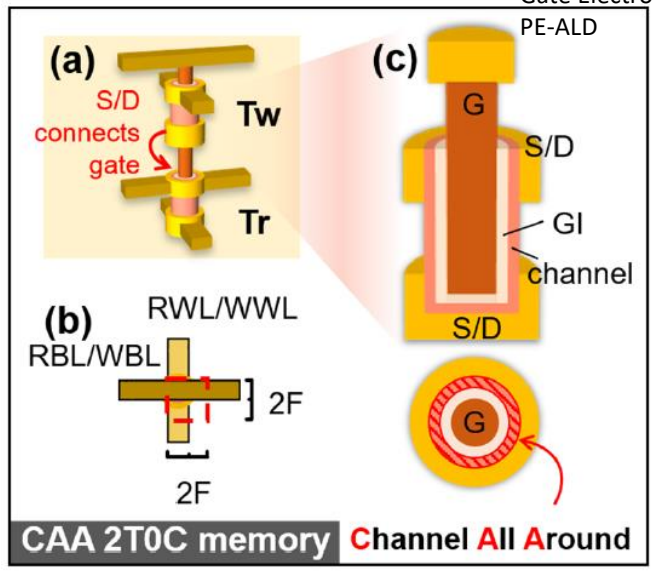




A. Spessot and H. Oh, "1T-1C Dynamic Random Access Memory Status, Challenges, and Prospects," in *IEEE Transactions on Electron Devices*, vol. 67, no. 4, pp. 1382-1393, April 2020, doi: 10.1109/TED.2020.2963911.

- **Current commercial DRAM products - 1T1C 6F2**
  - Diagonal isolation pattern to save additional space
- **4F2 cell architecture – 30% reduction in die area compared to 6F2**
  - Only a few papers have been published
  - 1T0C, 2T0C

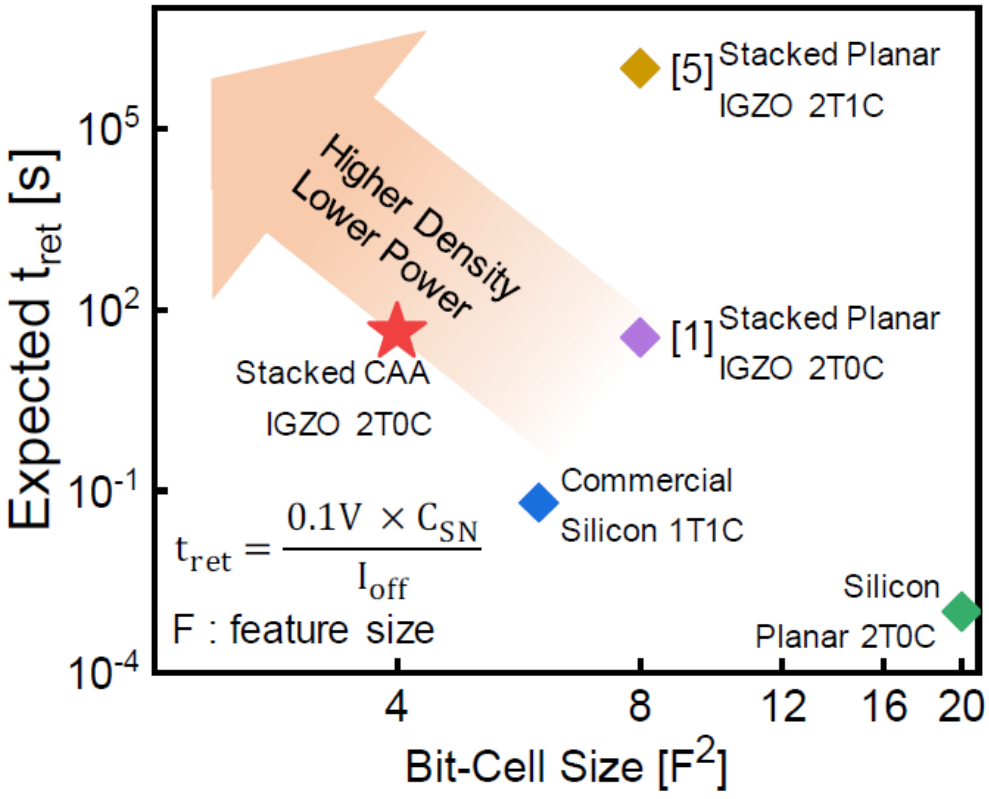
Channel – IGZO
Gate Insulator – Al2O3
Gate Electrode – IZO
PE-ALD



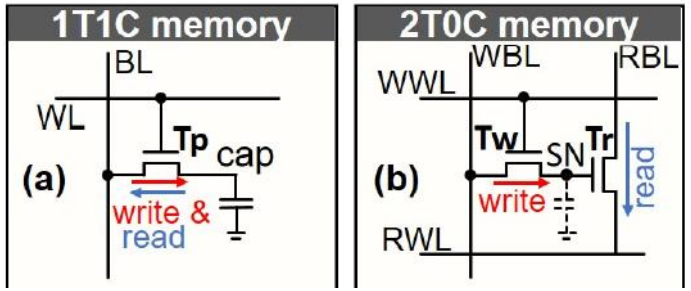
Source : X. Duan *et al.*, "Novel Vertical Channel-All-Around (CAA) In-Ga-Zn-O FET for 2T0C-DRAM With High Density Beyond 4F2 by Monolithic Stacking," in *IEEE Transactions on Electron Devices*, vol. 69, no. 4, pp. 2196-2202, April 2022, doi: 10.1109/TED.2022.3154693.

3D-DRAM

4F2 2T0C Vertical CAA IGZO (Wide bandgap amorphous oxide)



	Commercial 1T1C	Planar 2T0C	[1] Planar 2T0C	[5] Planar 2T1C	This Work
Channel Material	Silicon	Silicon	IGZO	CAAC-IGZO	**IGZO**
Area per transistor	$6F^2$	$8F^2$	$8F^2$	$8F^2$	**$4F^2$**
Read Transistor I_{on} (μA/μm)	100~300	500 ~1000	~10 (V_{DS}=0.8V)	~100 (V_{DS}=2.5V)	**30 (V_{DS}=1V)**
Retention	~64 ms	<1ms	400 s	> 1 hour	**300 s**
BEOL compatible	no	no	yes	yes	**yes**
Read while write	Unable	Enable	Enable	Enable	**Enable**

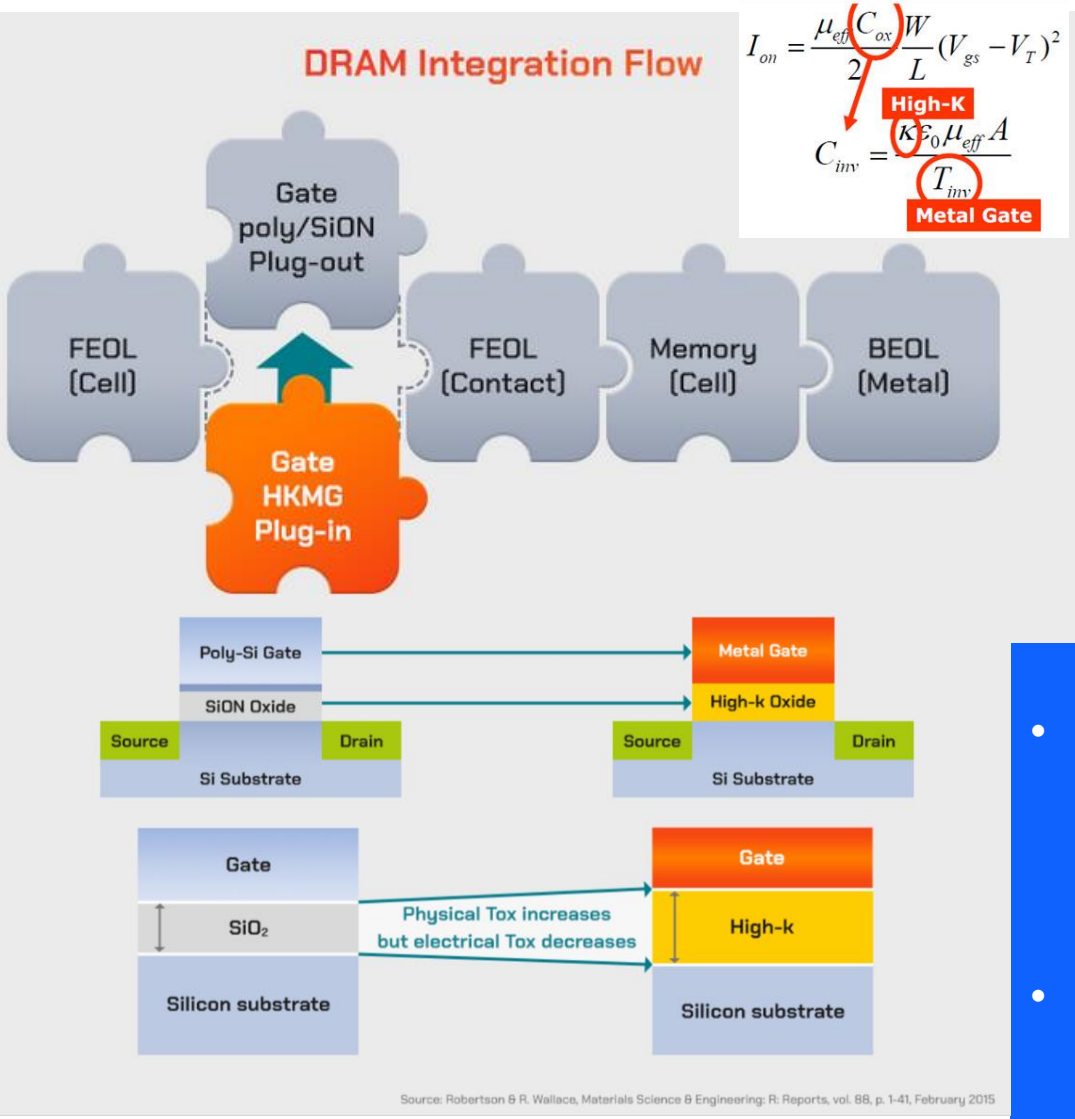


- **4F2 2T0C in latest papers**
  - Excellent retention time
- **Areas of improvement**
  - Low Ion
  - Operation speed of DRAM cells
  - Potential failure mode (unknown)

X. Duan et al., "Novel Vertical Channel-All-Around(CAA) IGZO FETs for 2T0C DRAM with High Density beyond 4F2 by Monolithic Stacking," 2021 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2021, pp. 10.5.1-10.5.4, doi: 10.1109/IEDM19574.2021.9720682.

HKMG (High K Metal Gate) in Peripheral

Enable power reduction / faster speed, Gox thickness scaling



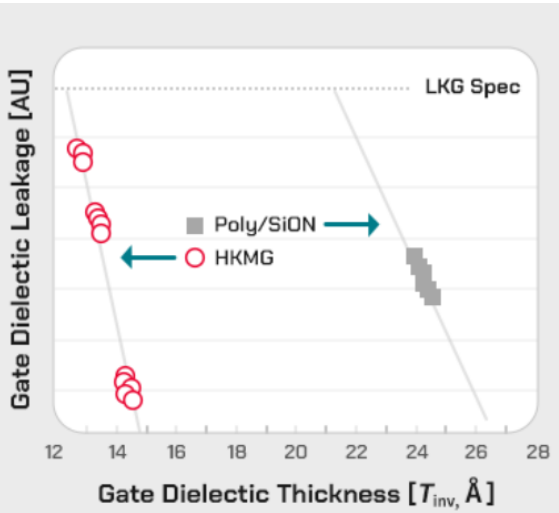
$$I_{on} = \frac{\mu_{eff} C_{ox}}{2} \frac{W}{L} (V_{gs} - V_T)^2$$

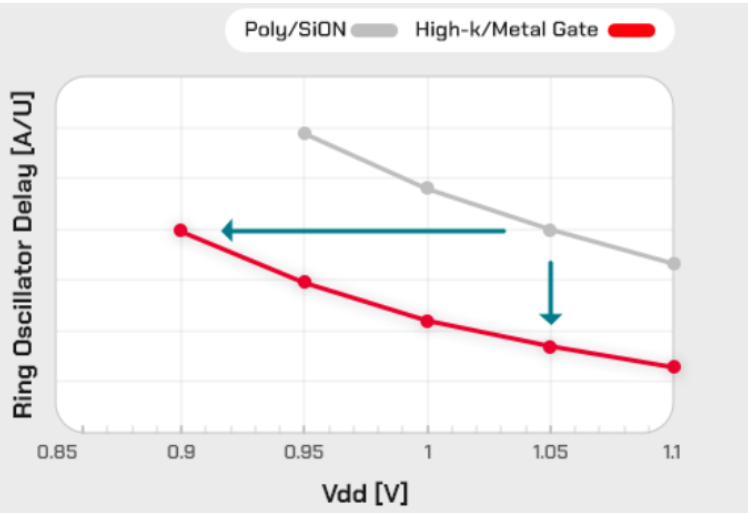
High-K

$$C_{inv} = \frac{\kappa \varepsilon_0 \mu_{eff} A}{T_{inv}}$$

Metal Gate

Source: Robertson & R. Wallace, Materials Science & Engineering: R: Reports, vol. 88, p. 1-41, February 2015



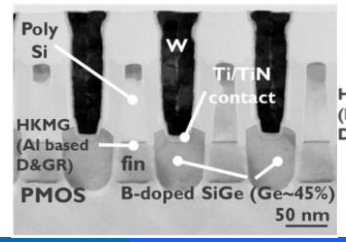


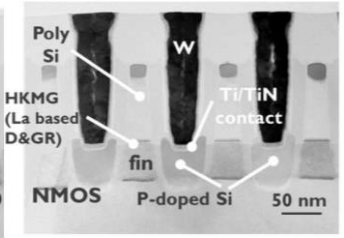
Source (Left): M. Sung et al., "Gate-first high-k/metal gate DRAM technology for low power and high performance products." 2015 IEEE International Electron Devices Meeting (IEDM), 2015, pp. 26.6.1-26.6.4

- **High-K Metal Gate process technology has been developed for logic devices decades ago**
  - DRAM recently adopted HKMG process in DDR5
  - Gate first (more compatible with existing DRAM process)
- **High-K Metal Gate is one of the innovations that enables DRAM scaling**
  - Lower power consumption (related to sustainability), faster speed

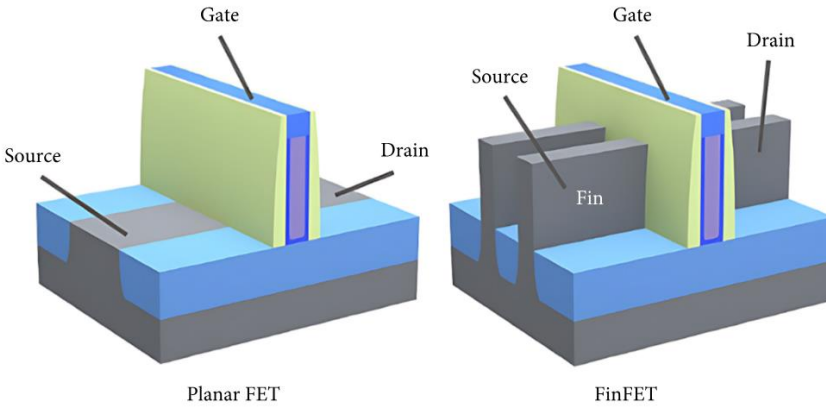
FinFET in Peripheral

Enable high performance and energy efficiency

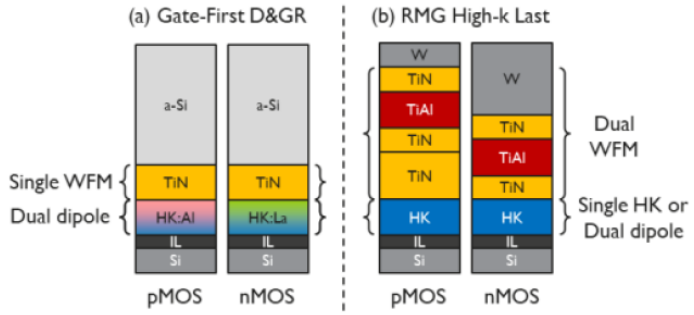




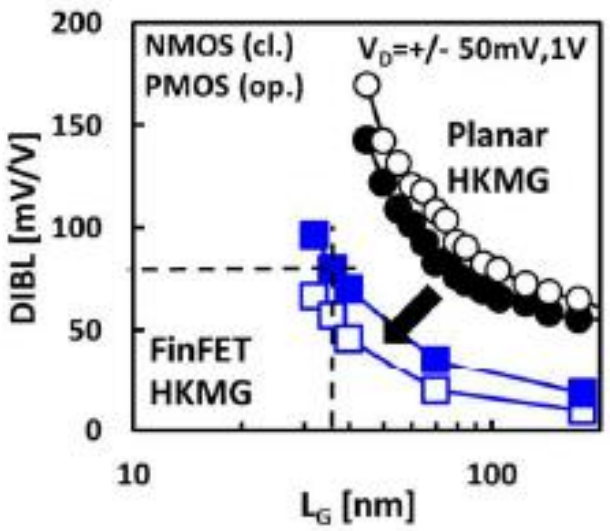
Flash Memory Summit

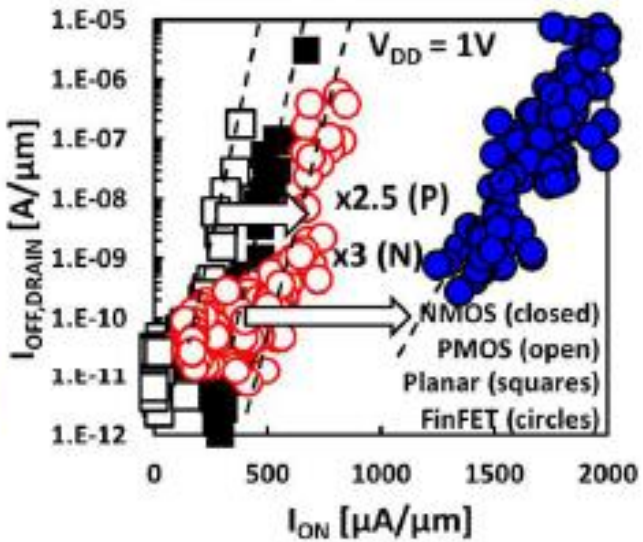


M. V. N. Rao *et al.*, "Design and Development of Efficient SRAM Cell Based on FinFET for Low Power Memory Applications," *Journal of Electrical and Computer Engineering*, vol. 2023, p. 7069746, Jun. 2023



E. Capogreco *et al.*, "FinFETs with Thermally Stable RMG Gate Stack for Future DRAM Peripheral Circuits," *2022 International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2022, pp. 26.2.1-26.2.4, doi: 10.1109/IEDM45625.2022.10019422.





R. Ritzenthaler *et al.*, "High Performance Thermally Resistant FinFETs DRAM Peripheral CMOS FinFETs with VTH Tunability for Future Memories," *2022 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, Honolulu, HI, USA, 2022, pp. 306-307

Improved Control over Leakage Current

Better Drive Current (higher performance)

Reduced Short-Channel Effects

- DIBL and sub-threshold leakage are reduced in FinFETs, enhancing the transistor's reliability and stability
- Maintain good electrostatic control even at smaller dimensions

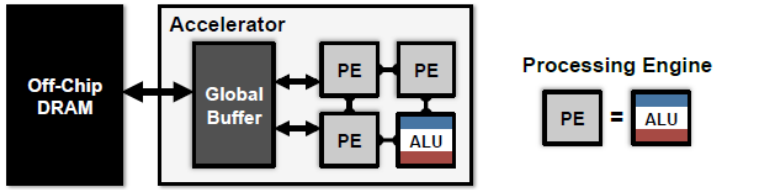
Faster Switching Speed

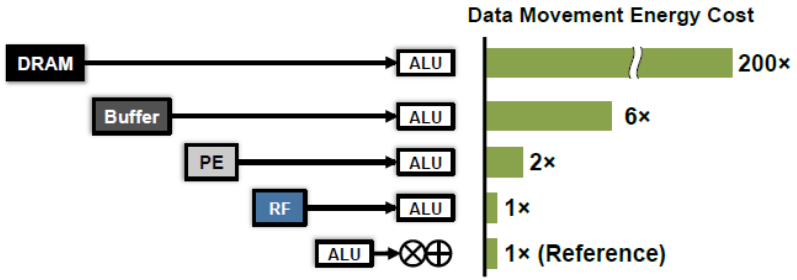
Latest papers (2022)

- **D&GR (Gate first)** – use single WFM and relies on dual interface dipoles to differentiate eWF of P/N MOS
- **RMG (Gate last)** – increased freedom of WFM engineering between P/N MOS

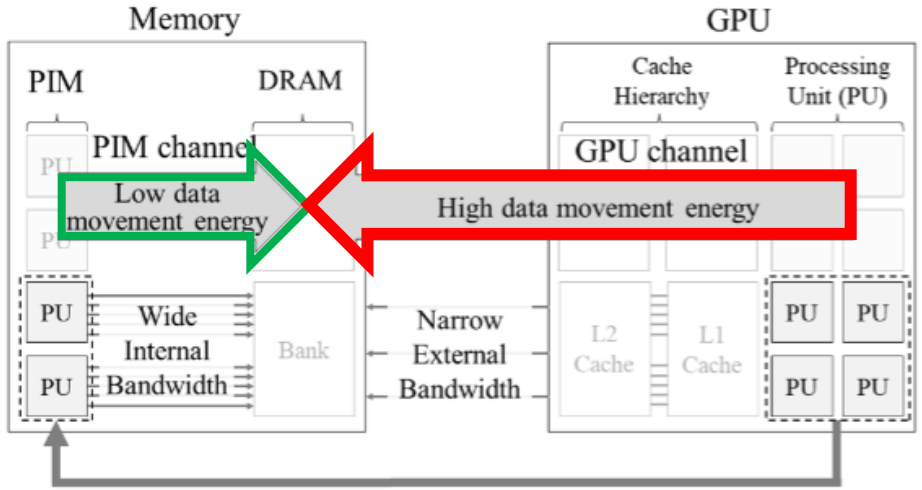
PIM (Processing In Memory)

Efficient data movement for sustainability





V. Sze, *et al.*, "Hardware for machine learning: Challenges and opportunities," *2017 IEEE Custom Integrated Circuits Conference (CICC)*, Austin, TX, USA, 2017, pp. 1-8, doi: 10.1109/CICC.2017.7993626.

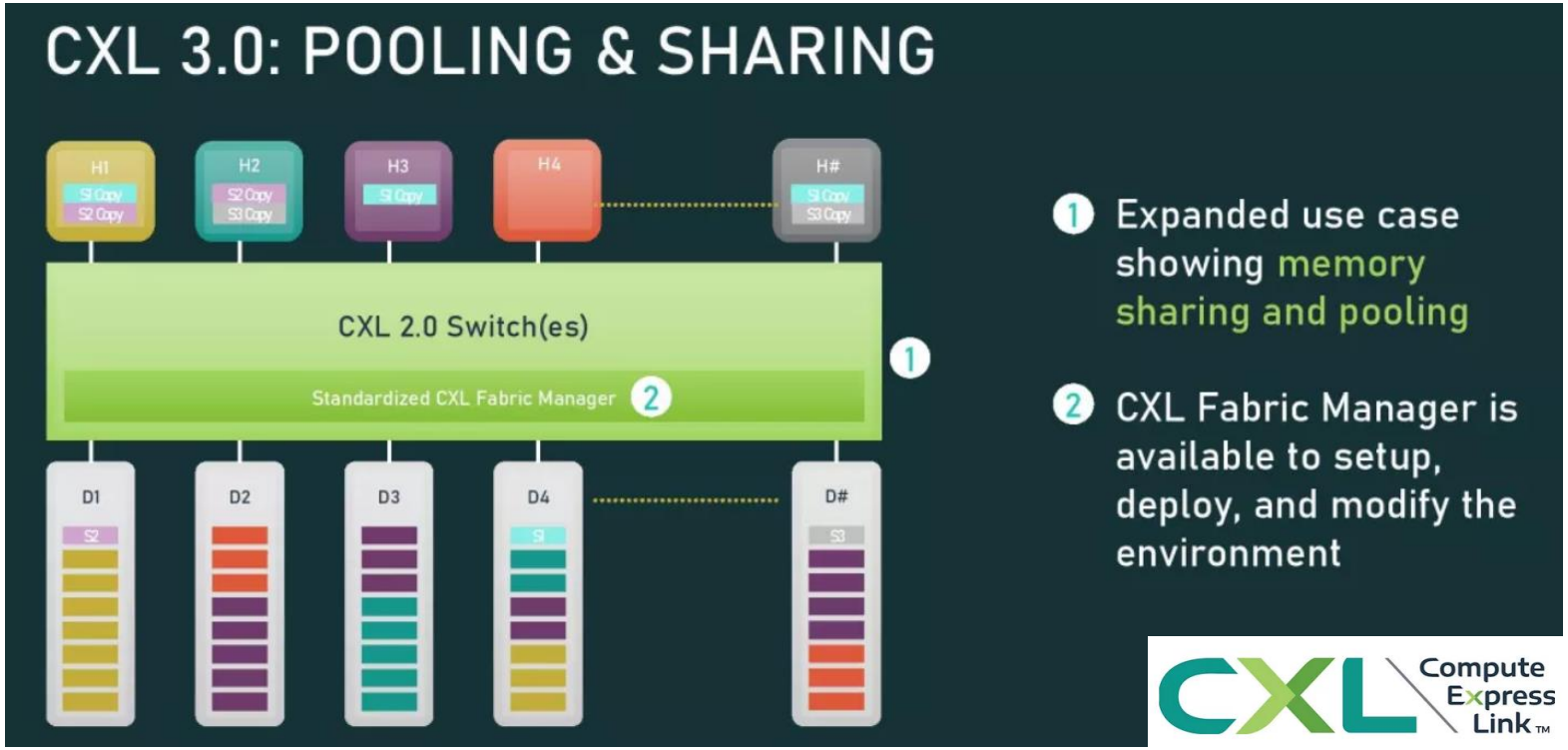


S. Kim et al., "Signal Integrity and Computing Performance Analysis of a Processing-In-Memory of High Bandwidth Memory (PIM-HBM) Scheme," in IEEE Transactions on Components, Packaging and Manufacturing Technology, vol. 11, no. 11, pp. 1955-1970, Nov. 2021, doi: 10.1109/TCPMT.2021.3117071

- **Off-chip DRAM access to CPU/GPU (data movement) costs 200x higher energy than accessing data from RF(Register File)**
  - Short-distance transfers (communicating with a neighbor PE) are charged a lower energy cost than longer-distance transfers (global buffer access) due to smaller wiring capacitance and simpler NOC (on chip network)
- **Processing In Memory in HBM**
  - Offload memory-intensive computation to PIM

CXL (Compute Express Link)

Efficient resource utilization & data movement for sustainability



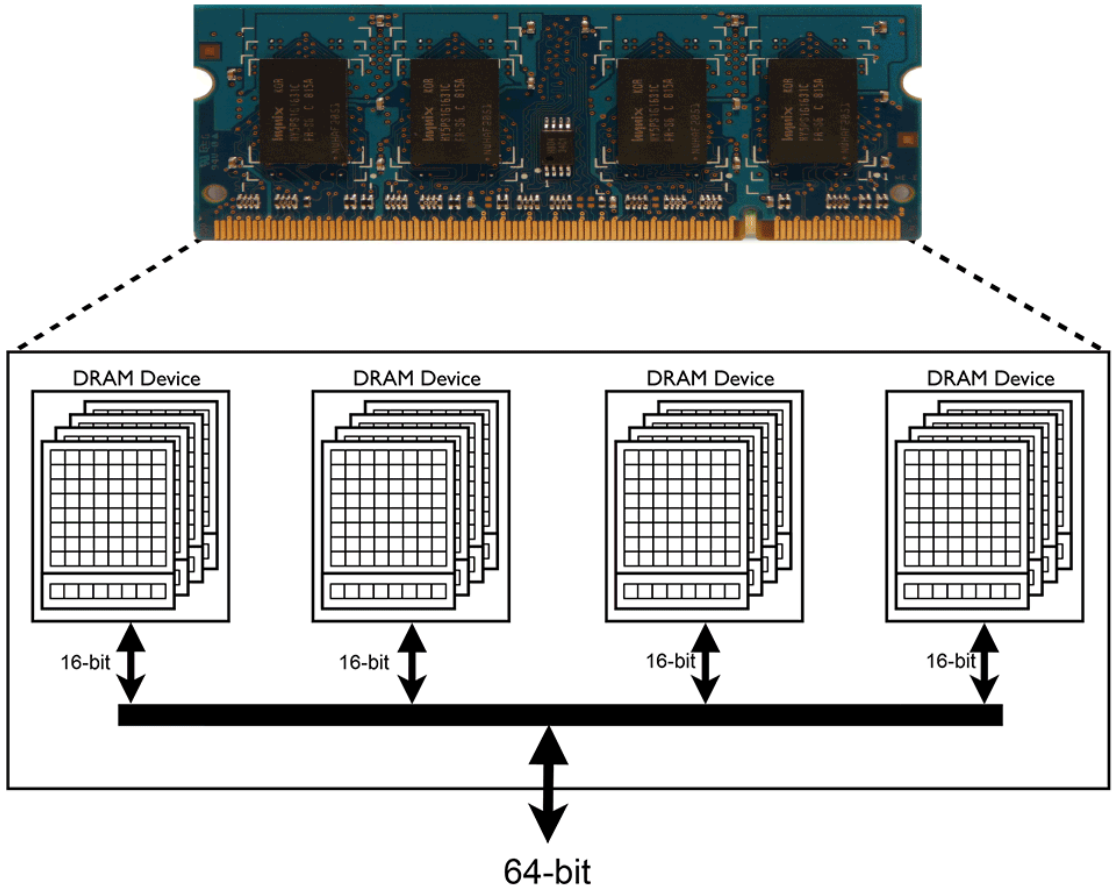
Memory Pooling

- A memory device can be partitioned off to multiple hosts/ accelerators, enabling better utilization of memory resources

Memory Sharing

- A given region of memory can be simultaneously accessible by more than one host in the coherency domain
- Need for redundant data storage and data transfers between different components is reduced by minimizing unnecessary data movement and data duplication

RowHammer as a significant threat to system security



Source : http://apt.cs.manchester.ac.uk/projects/ARMOR/RowHammer/

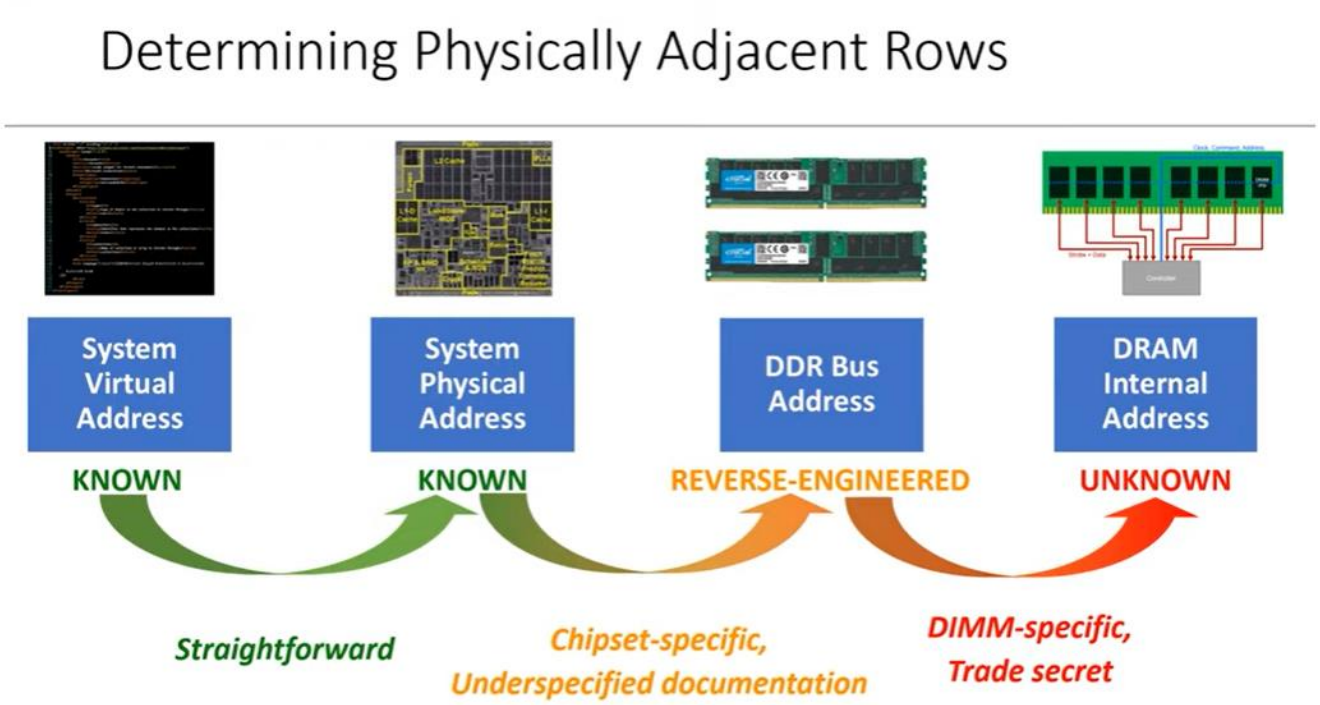
Single or multi-bit flip can be induced on adjacent cells through repeated activations in a target row, breaking memory isolation → threat to system security

- By carefully selecting rows to hammer, an attacker can induce bitflips in sensitive data stored in DRAM
- Privilege Escalation attacks by corrupting PTEs (page-table entries) – through the bit flip in the right location of a PTE, gain write access to page table

RowHammer Mechanisms

- Electron injection & capture
- Capacitive crosstalk

Mitigation of RowHammer

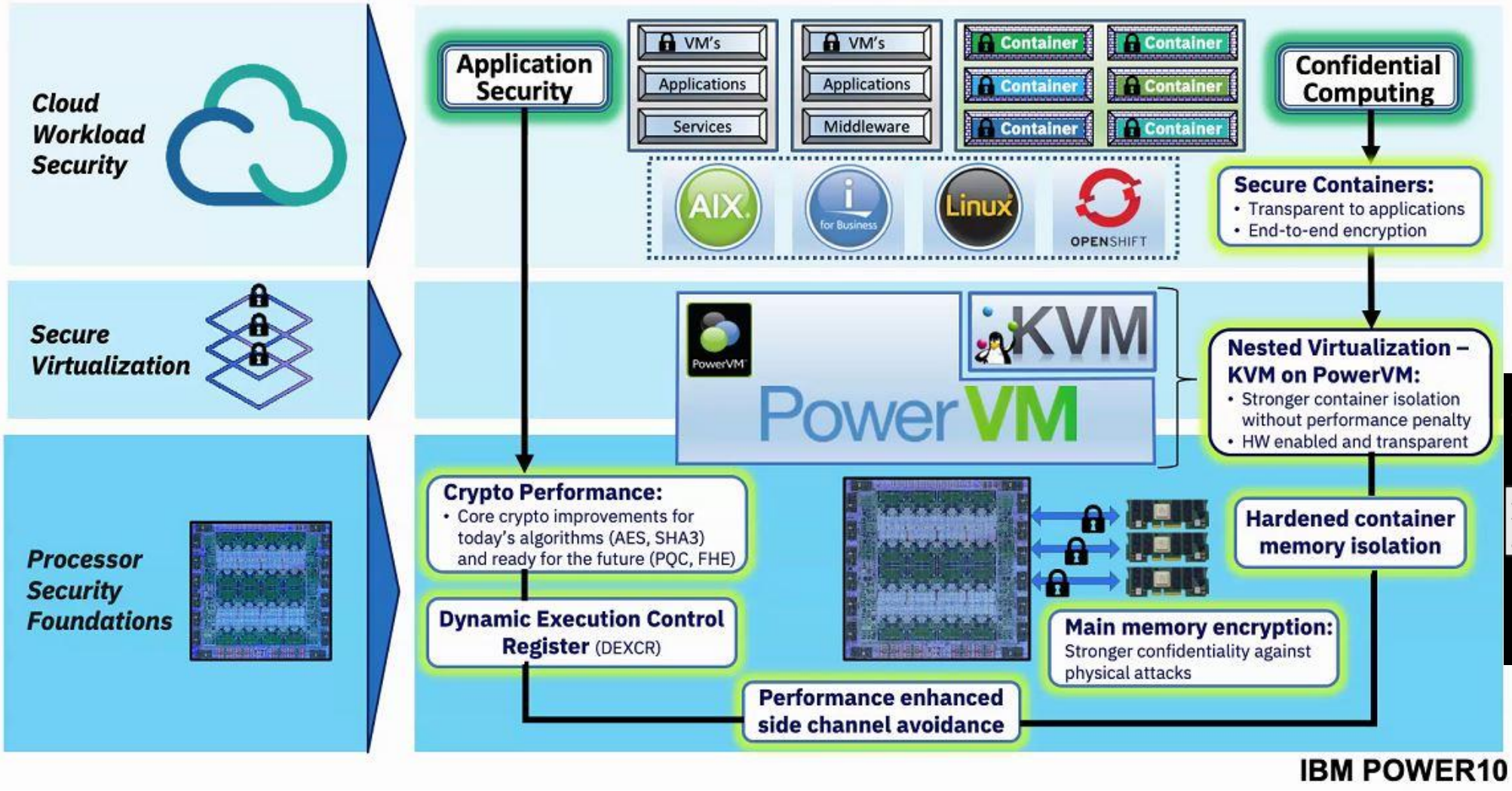


Source : https://www.youtube.com/watch?v=XP1SvxmJoHE&ab_channel=IEEESymposiumonSecurityandPrivacy

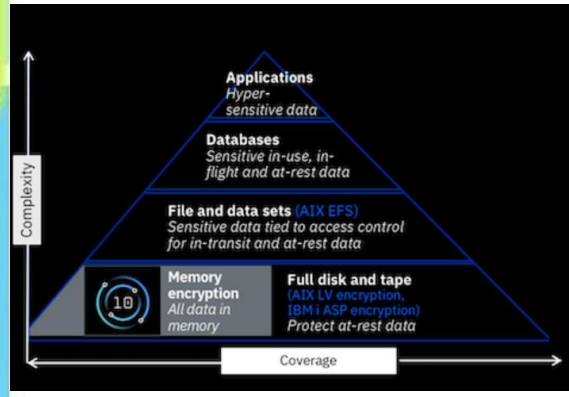
- **In-DRAM mitigation – TRR (Target Row Refresh)**
  - DRAM supplier proprietary, Undocumented
  - Reverse engineering of TRR was published (U-TRR)
  - Tight control of chip to chip retention time variation
- **JEDEC Data Integrity TG (TG42.8)**
  - **RFM** – Host to issue RFM commands every N activations
  - **ARFM** - Supports 3 separate RFM rates based on RowHammer risk level (Host/HW/FW)
  - **DRFM** - Host to provide additional TRR to suspicious addresses
  - **PRHT** - Deterministic tracking of per row address activates
- **Additional System Level Mitigation**
  - Disabling deferred refresh
  - 2x refresh with performance impact
  - Other proprietary system level mitigation to make address mapping challenging

Protect data from Core to Cloud

Security : End-to-End for the Enterprise Cloud



- Enable end to end security with full stack encryption
- The entire memory is encrypted, with no performance penalty or management set-up



Summary

- **Significant amount of memory is consumed during deep learning neural network**
  - Exponential increase in the number of AI model parameters for accuracy and better throughput
  - Memory capacity, bandwidth, latency improvement is required along with maintaining **sustainability** and elevated level of **security**
- **2D DRAM scaling is losing steam**
  - Innovations such as High K Metal Gate, FinFET are required to reduce power consumption, affecting sustainability
  - 3D DRAM (4F2) development is in pipeline
- **Efficient Data movement helps maintain sustainability**
  - Smart data movement via PIM (Processing in Memory) and CXL (Computer Express Link) consumes less power
- **RowHammer is a significant threat to system security**
  - DRAM suppliers developed TRR (undocumented) to mitigate RowHammer
  - JEDEC TG42 is making standard to mitigate RowHammer (RFM/ARFM/DRFM/PRHT)
  - Additional system level mitigation is needed to prevent RowHammer

When you interact with IBM, this serves as your authorization to Flash Memory Summit or its vendor to provide your contact information to IBM in order for IBM to follow up on your interaction.

IBM's use of your contact information is governed by the IBM Privacy Policy.