A chat at Graduate School of Advanced Technology at National Taiwan University
May/29/2023

Scaling in Electron Devices and Electronic Systems

A decadal projection in memory technologies

DerChang Kau (高德昌)
Intel Fellow, Director of External Technology Pathfinding
Corporate Planning Group

Abstract

Since 1897's discovery of electron by Sir Joseph John Thomson (Nobel Laureate 1906), electron devices started the journey of the 2nd industrial evolution onward. Under the guidance of Moore's Law and Dennard Scaling, electron devices' density increase at cadence and electronic systems strive maintaining power density while improving performance. This talk invites participants first travel down the time tunnel of electron devices. We will jointly revisit researches in material synthesis and device exploration in atomistic switching physics to disrupt electrostatic switching mechanism. Memory technology will be used to project decadal electronic system advances in homogenous and heterogenous integration through VLSI fabrication and chiplet interconnect processes.

About DerChang Kau

An Intel Fellow and a Director of External Technology Pathfinding, exploring chip manufacturing and system integration technologies aligning with product roadmap. For more than 3 decades, DerChang has been involved in various technology pathfinding and scaling deployment, including logic, memory, flash, embedded memory, mixed-signal, radio and advanced packaging. He also led Intel's efforts to define strategic direction for atomistic memory switching and threshold switching devices and drove the product initiatives of Optane Memory Technology. He holds MS degrees in Electrical Engineering from OSU.

Outline

- Travel down the time tunnel of electron devices –
  - A Centennial Journey of Electrical Charges
- State of Art of Memory Technologies –
  - System hierarchy and incumbent deployment
  - A holistic view of Memories: Cell, Array and Subsystem
- Memories in next decade –
  - Cache, Memory and Storage
  - Near Memory and Storage Class Memory

A chat at Graduate School of Advanced Technology at National Taiwan University
May/29/2023

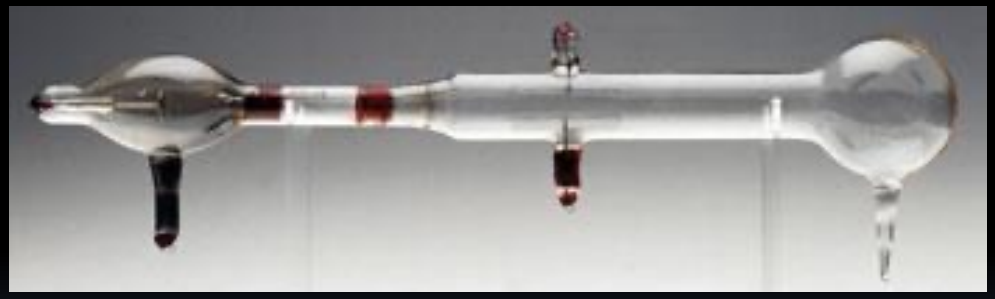
TRAVEL DOWN THE TIME TUNNEL OF ELECTRON DEVICES

A Centennial Journey of Electrical Charges

Discoveries of Electron

1895 Wilhelm Conrad Röntgen (1845 - 1923) discovered X rays, for which he received the first Nobel Prize for physics in 1901. The "hand" was taken in 1896 with cathode-ray discharge. He later demonstrated the metallurgical and medical use of X rays.

https://www.nobelprize.org/prizes/physics/1901/summary/

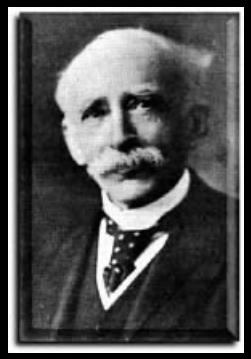


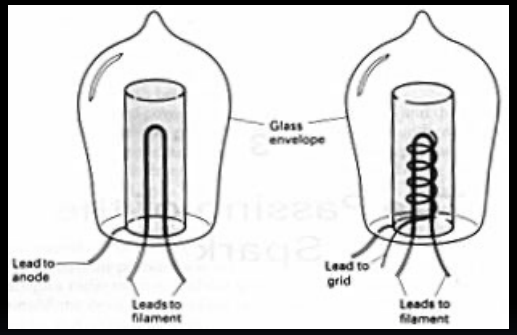
1897 Sir Joseph John Thomson (1856 - 1940) discovered and identified the electron, for which he received the Nobel Prize for physics in 1906. Thomson demonstrated that cathode rays were actually units of electrical current made up of negatively charged particles of subatomic size as an integral part of all matter and theorized the "plum pudding" model of atomic structure in which a quantity of negatively charged electrons was embedded in a sphere of positive electricity, the two charges neutralizing each other.

https://www.nobelprize.org/prizes/physics/1906/summary/

Innovations of Vacuum Electron Devices

1905 Sir John Ambrose Fleming (1849 - 1945) made the first diode tube, the Fleming valve.

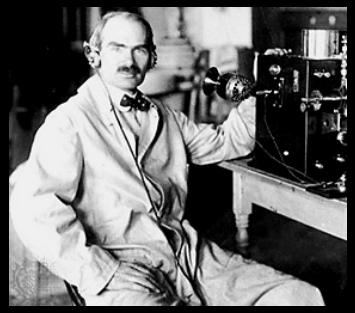




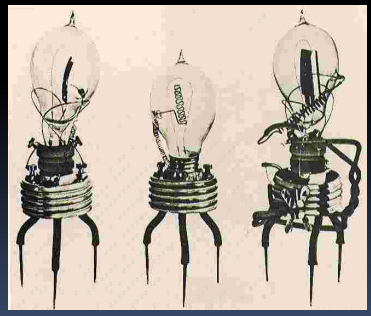
The device had three leads, two for the heater and the cathode and the other for the plate.

https://madeupinbritain.uk/Vacuum_Tube

1907 Lee De Forest (1873 - 1961) added a grid electrode to Flemings' valve and created the triode tube,



later improved and called the Audion.



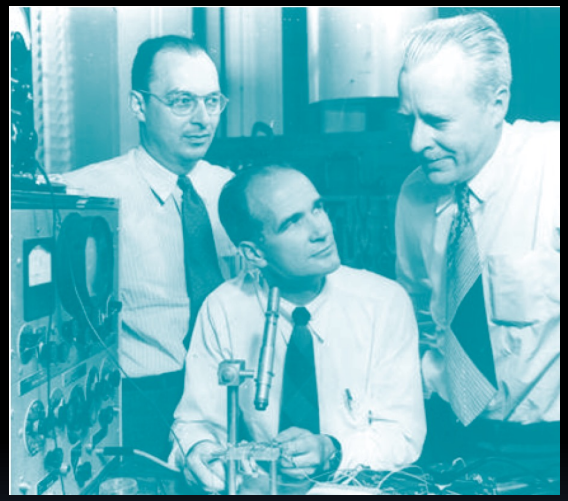
He was a prolific inventor, and was granted more than 300 patents in Electronics

https://www.britannica.com/biography/Lee-de-Forest

Revolution of Electronics –

From Vacuum Science to Solid State Physics

In 1948, John Bardeen (1908 - 1991), **William Bradford Shockley** (1910 - 1989), **Walter Houser Brattain** (1902 - 1987) invented the transistor for Bell labs, sharing the 1956 Nobel Physics prize for the invention.



Shockley's insight of "minority carrier injection" became the bible of the new era

[1] W. Shockley, "The Theory of *P-N* Junctions in Semiconductors and *P-N* Junction Transistors." Nokia Bell Lab, July 1949

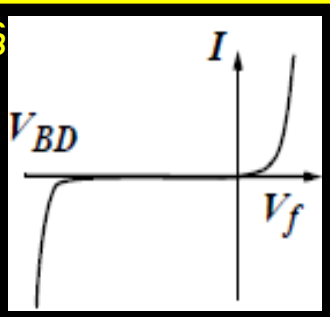
[2] W. Shockley, "*Electrons and Holes in Semiconductors: With Applications to Transistor Electronics",* Van Nostrand, Jan, 1050

Thermionic emission (single-/dual-carrier) and beyond

Diode: two terminals and exhibiting a nonlinear I-V, *IEEE Standard Dictionary of Electrical and Electronics Terms, 1980*

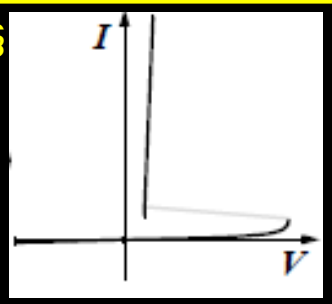
Unidirectional
Asymmetric
Unipolar

§



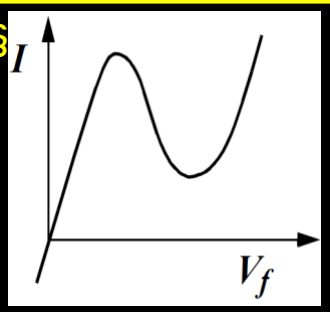
L-Shape

§



S-Shape

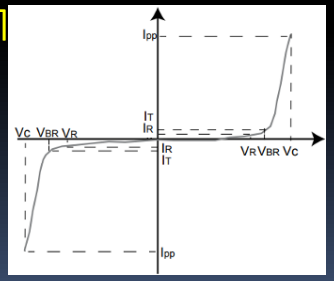
§



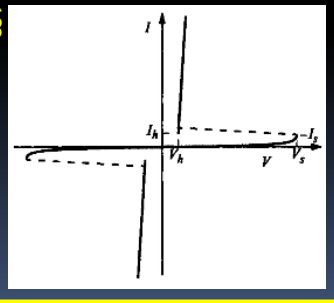
N-Shape

Bidirectional
Symmetric
Ambipolar

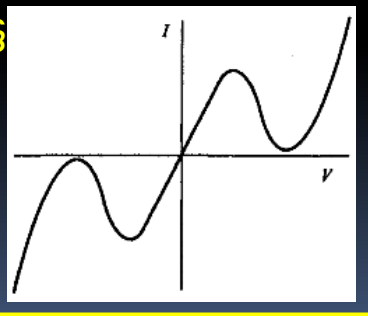
¶



§

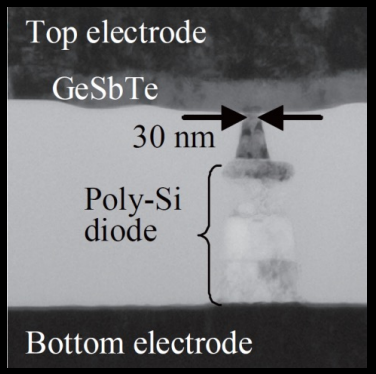


§

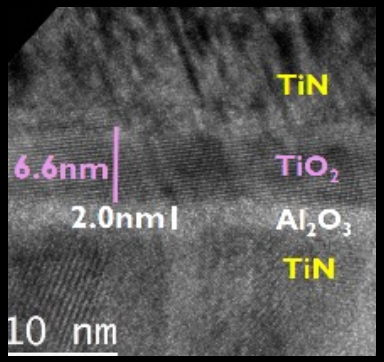


§ : Kowk K. Ng, Complete guide to semiconductor devices, 2nd ed., 2002
¶ :https://www.digikey.com/en/articles/design-in-tvs-diode-protection-enhance-can-bus-reliability

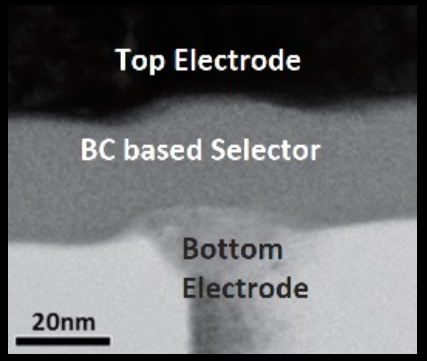
Thin Film Diode Candidates for Cross Point Memory



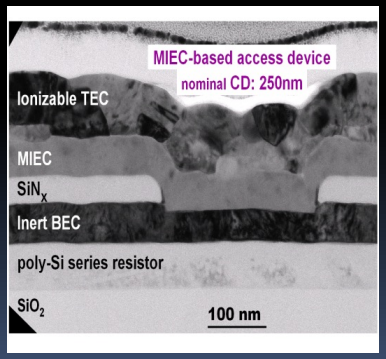
Y. Sasago, *et. al.*,
VLSI '09. T2B-1



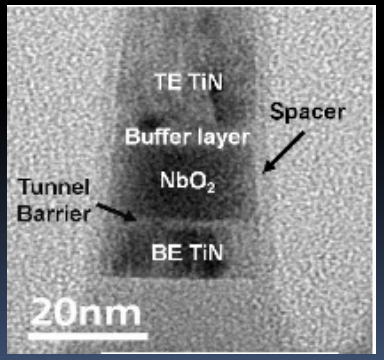
B. Govoreanu, *et.al.*,
IEDM'13. S10.2



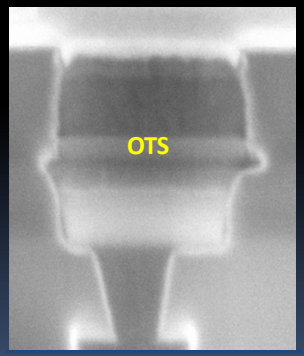
S. Yasuda, *et.al.*,
VLSI, 2017. T2-4



K. Gopalakrishnan, *et. al.*,
VLSI '10. TS19.41

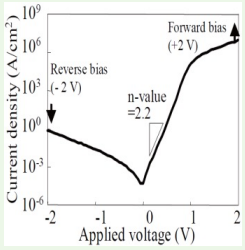
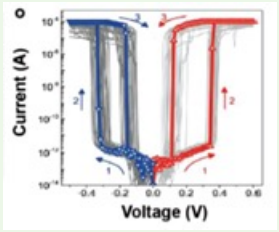
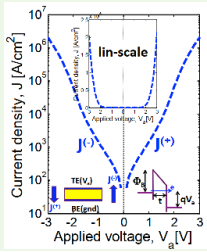
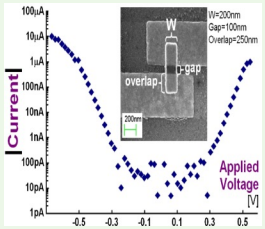
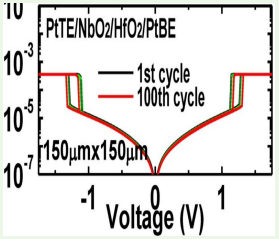
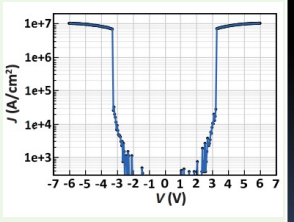


S.G. Kim, *et. al.*,
IEDM '15. S10.3



D. Kau, *et. al.*,
IEDM '09. S27.1

Threshold Switching Phenomenology & Mechanisms

Mechanism	Thermionic	Filamentation	Tunneling	MIEC	MIT	OTS
Construct.	P-N or M-S Jx	Ion (Ag^+) in Ox	MIM	Cu^+ in SE	NbO_2	Chalcogenide
Switching	Electronic	Atomistic	Electronic	Atomistic	Atomistic	Electronic
Polarity	Unidirectional	Bidirectional	Bidirectional	Bidirectional	Bidirectional	Bidirectional
τ_{switch}	sub nsec	ns ~ 100s ns	ps or faster	ns ~ 100s ns	ns to 10s ns	sub ns
J_{MAX}	< 10MA/cm²	1-10MA/cm²	< 1MA/cm²	~10MA/cm²	> 10MA/cm²	> 10MA/cm²
$J_{Inhibit}$	<1A/cm²	<1KA/cm²	<1KA/cm²	<1KA/cm²	<1KA/cm²	<1KA/cm²
$V_{Inhibit}$	< 3V	< 1V	<3V	< 1V	< 3V	< 3V
I-V						
Reference	Y. Sasago, *et. al., VLSI '09*	J. Yang, *et.al., Adv Func Mtls (2018)*	B. Govoreanu, *et.al., IEDM'13*. P10.2	K. Gopalakrishnan, *et.al., VLSI Symposium* '10.	X. Liu *et.al., EDL* Oct.'14	S. Yasuda, *et.al., VLSI symposium*, '17,

A chat at Graduate School of Advanced Technology at National Taiwan University
May/29/2023

STATE OF ART OF MEMORY TECHNOLOGIES

System hierarchy and incumbent deployment
A holistic view of Memories: Cell, Array and Subsystem

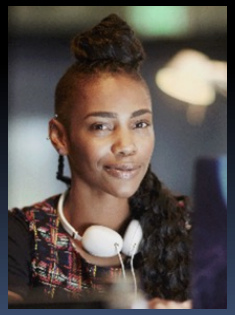
We are in a data-centric world

Data are

stored, processed, and analyzed

Every Day Data is EXPLODING

1.5 GB
AVERAGE INTERNET USER



50 GB
AUTONOMOUS DRIVING

3 TB
SMART HOSPITAL

40 TB
AIRPLANE DATA

1 PB
SMART FACTORY

50 PB
PUBLIC SAFETY

https://www.cisco.com/c/dam/m/en_us/service-provider/ciscoknowledgenetwork/files/547_11_10-15-DocumentsCisco_GCI_Deck_2014-2019_for_CKN__10NOV2015_.pdf

intel®

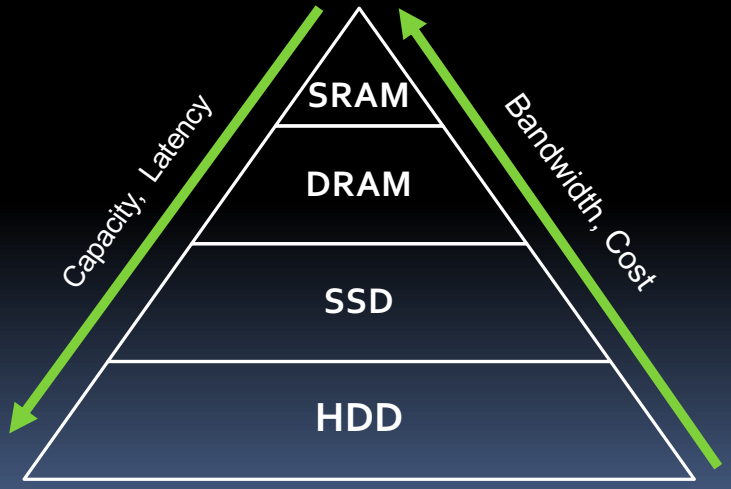
Incumbent Technologies in Memory/Storage Hierarchy

each of which has greater capacity than the preceding but which is less quickly accessible

*"**The Memory Organ –** **Ideally** one would desire an **indefinitely large memory capacity** such that any particular … **word** would be **immediately** … available. … It does not seem possible physically to achieve such a capacity. …We are therefore forced to recognize the possibility of constructing a **hierarchy of memories**, each of which has greater capacity than the preceding but which is less quickly accessible."*

Preliminary Discussion of the Logical Design of an Electronic Computing Instrument

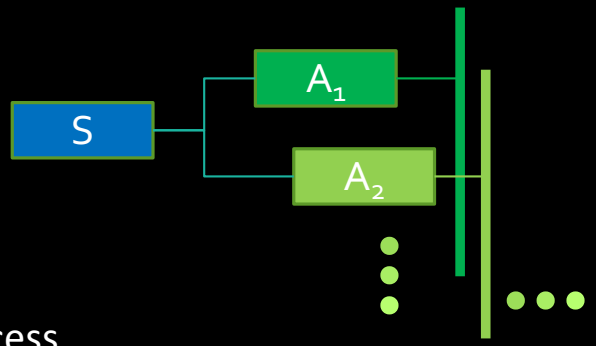
Arthur Burks, Herman Goldstine and John von Neumann, June/28/1946



intel®

A memory cell consists one storage element and one or more access switch(es)

- Storage Elements:
  - Latch: Circuit based or Thyristor device
  - Electrostatic memory switch: Two terminal capacitor, floating gate/plate, charge traps devices
  - Atomistic memory switch: Phase change, Magneto Tunnel Junction, Ferroelectric, Oxy-Vacancy (Filamentation)
- Access Devices:
  - 3 terminal Transistor:
    - Pass logic (1T) is typical used for read/write access.
    - Gated common-source amp (2T) is used for read only access
    - Biased Source-follower (1T) is used for read only access
  - Two-terminal diodes: Thermionic (PN junction, Schottky barrier, DIAC & etc) , MIT, OTS MIEC, volatile filamentation and etc.

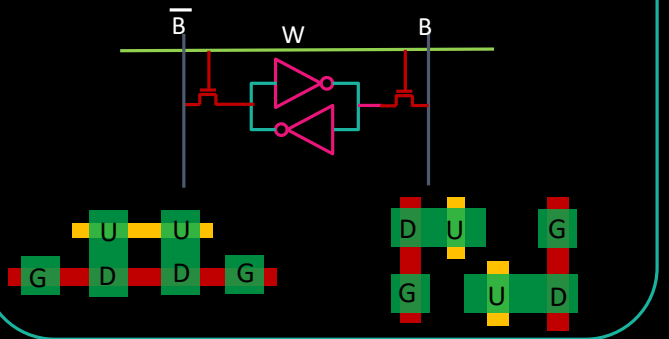


Brain teaser: multiple memory storage elements share one access device

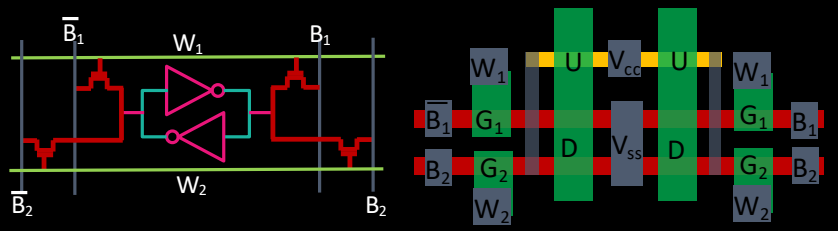
intel

Multi-ported SRAM Symmetry

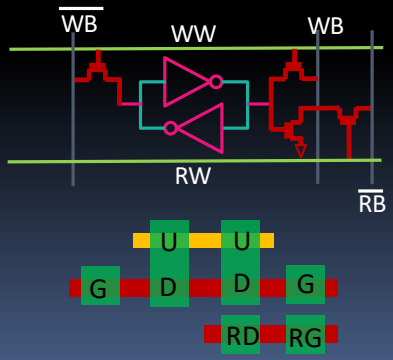
1RW



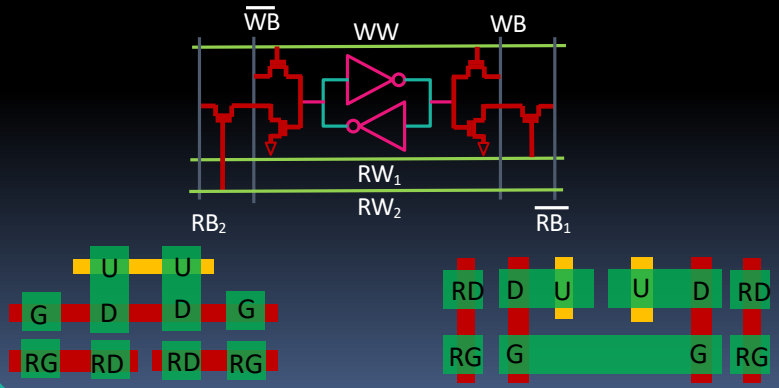
2RW



1R1W



2R1W



intel

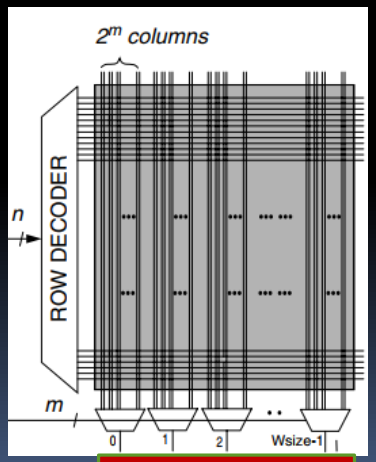
Why "Multi-Port Memories"

- Multi-Port Memories have all the features of single-port memories
- Used for multi-tasking
- Parallel Processing – increase bandwidth without cycle time compressing
- "Simultaneously" read and write data or read-while-write
- Examples
  - Dedicated Memory – In a branch-target-buffer memory to keep track branch prediction and decision
  - Shared Memory – multiple requests for data from multi-cores or networked processors
- Analysis and Synthesis
  - Multi-port (MP) SRAM is a commonly used for cache design as it allows simultaneous R/W accesses.
  - For CISC CPUs, such as the X86 architecture, MP cache such as mRnW configuration is used
  - For RISC CPUs, such as the ARM architecture, SP cache such as 1RW with multi-bank (MB) configuration is used.
  - Overall, the choice of cache memory organization and multiport SRAM configuration depends on the specific design goals, performance requirements, and trade-offs between cost, power consumption, and performance.

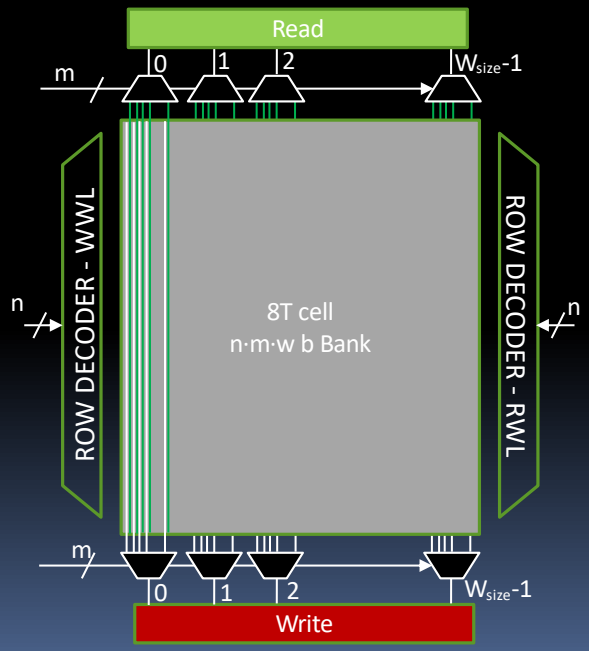
Array Capacity: $n \cdot m \cdot w$ bits with 1RW, 1R1W, 2R2W cells

6T cell
1RW

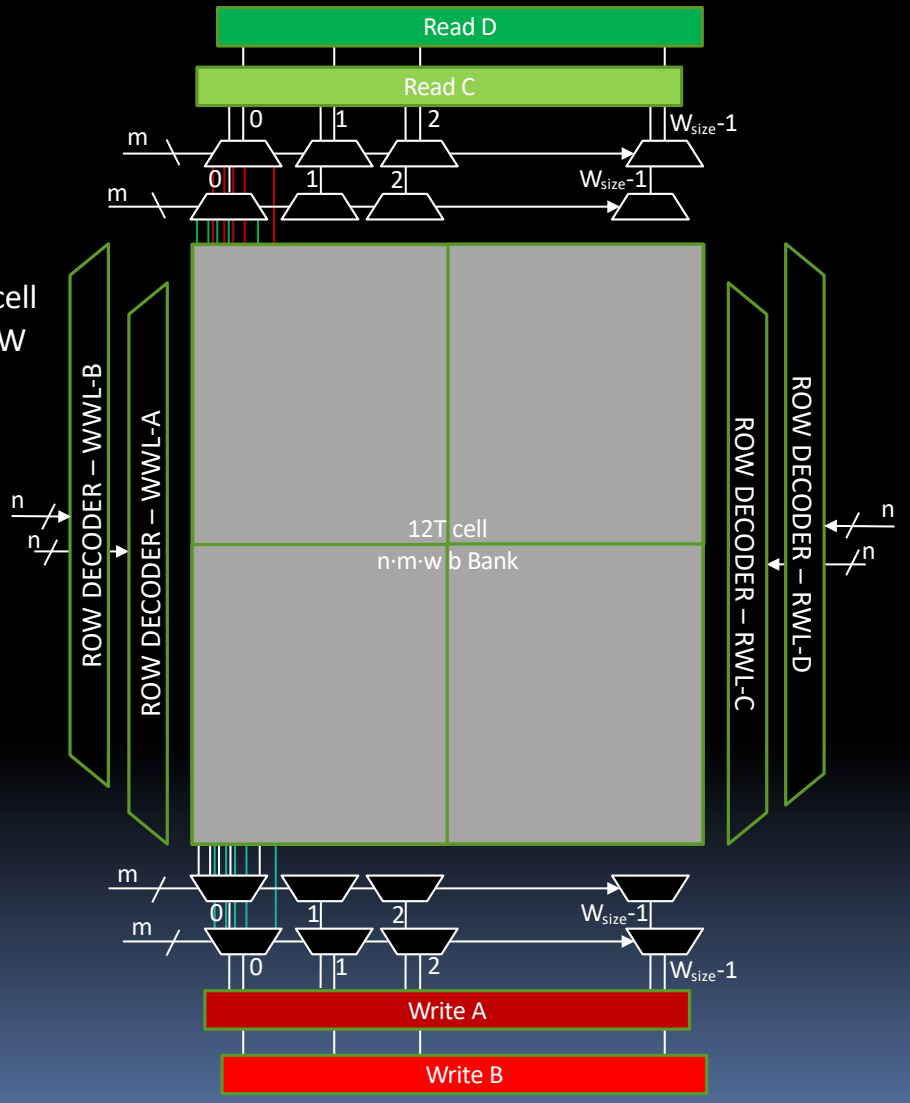
Bruce Jacob et.al, Memory System, p.265



8T cell
1R1W



16T cell
2R2W

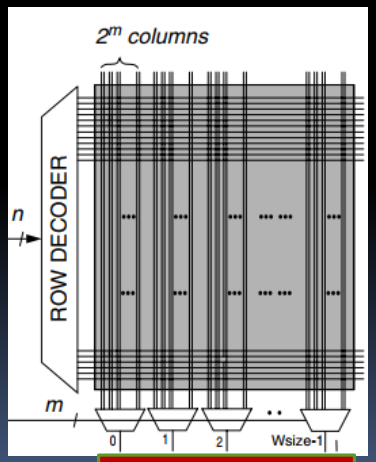


intel

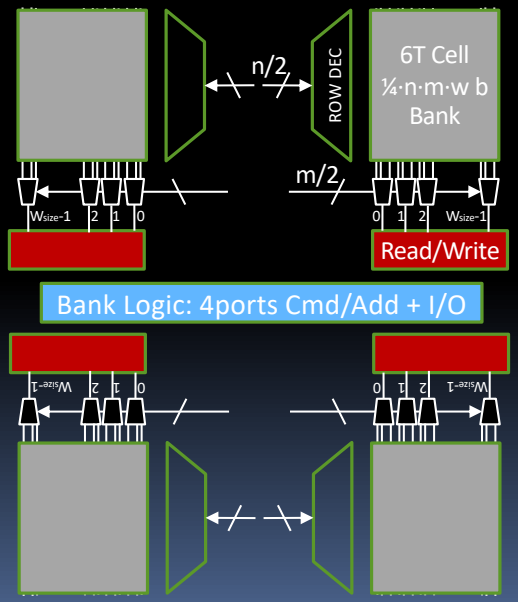
$n \cdot m \cdot w$-bit array
4 RW ports vs. 2R2W ports

6T cell
1 RW port
1Bank

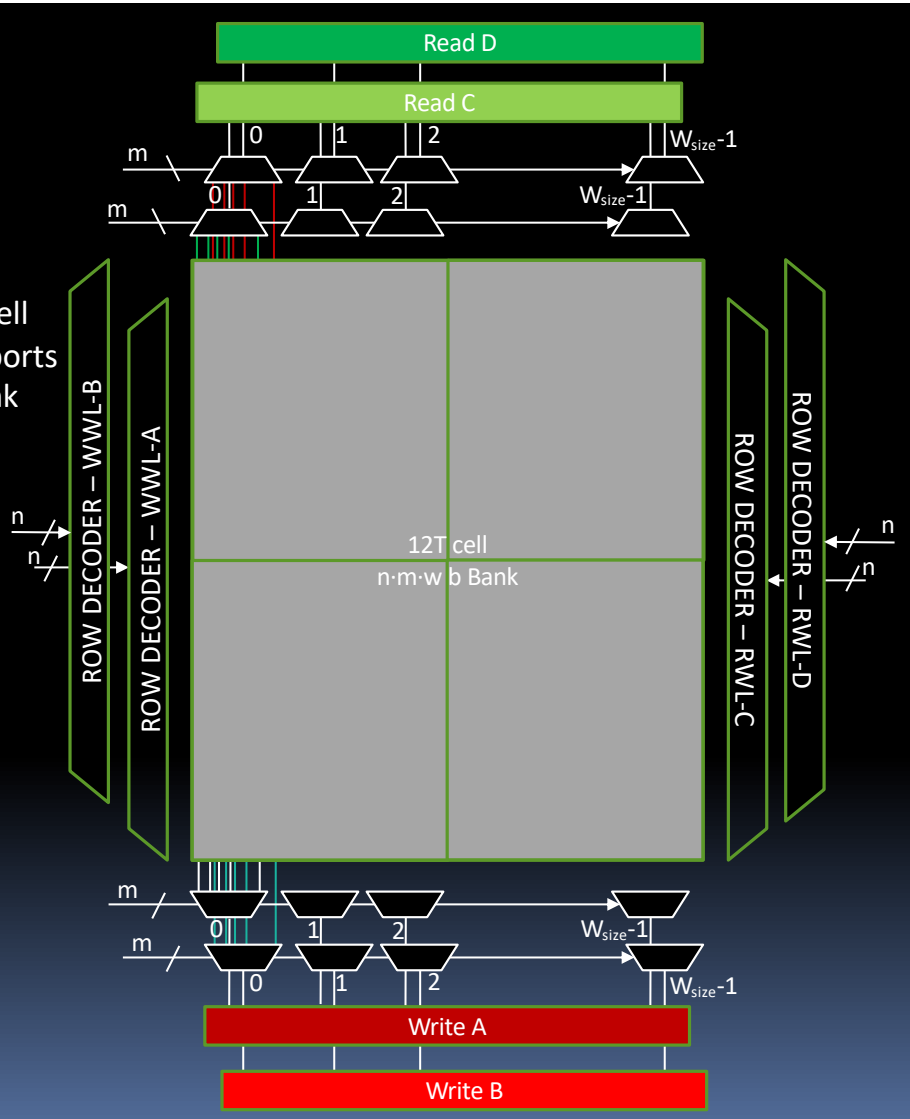
Bruce Jacob et.al, Memory System, p.265



6T cell
4 RW ports
4 banks



16T cell
2R2W ports
1Bank



intel

A chat at Graduate School of Advanced Technology at National Taiwan University
May/29/2023

MEMORIES IN NEXT DECADE

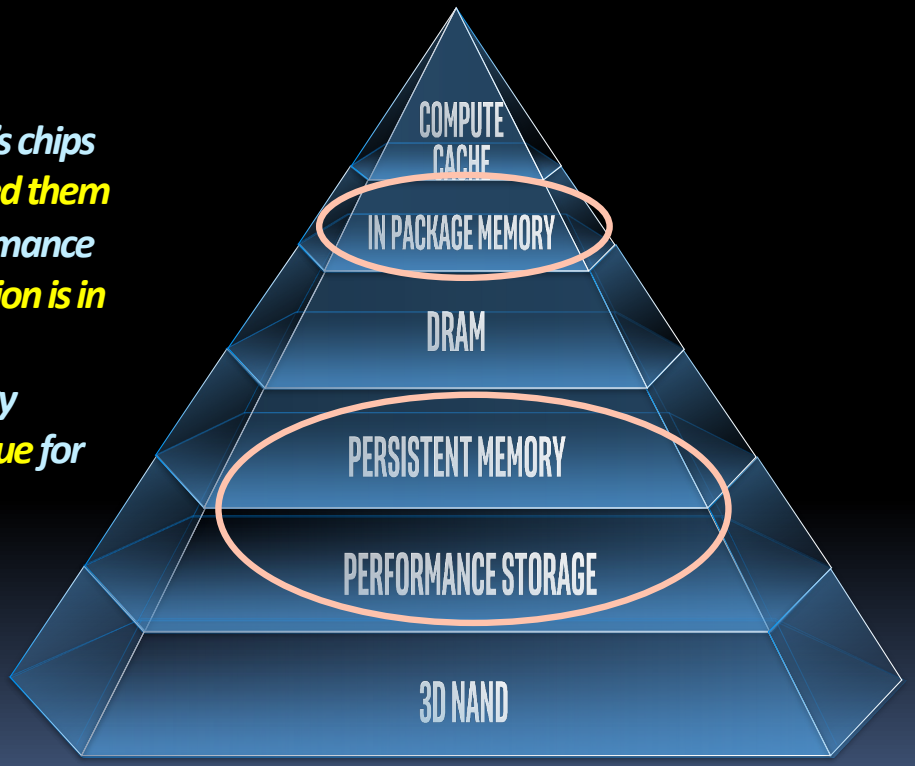
Cache, Memories and Storage
Near Memory and Storage Class Memory

Gaps to fill between Caches, Main Memory & Storage beyond scaling

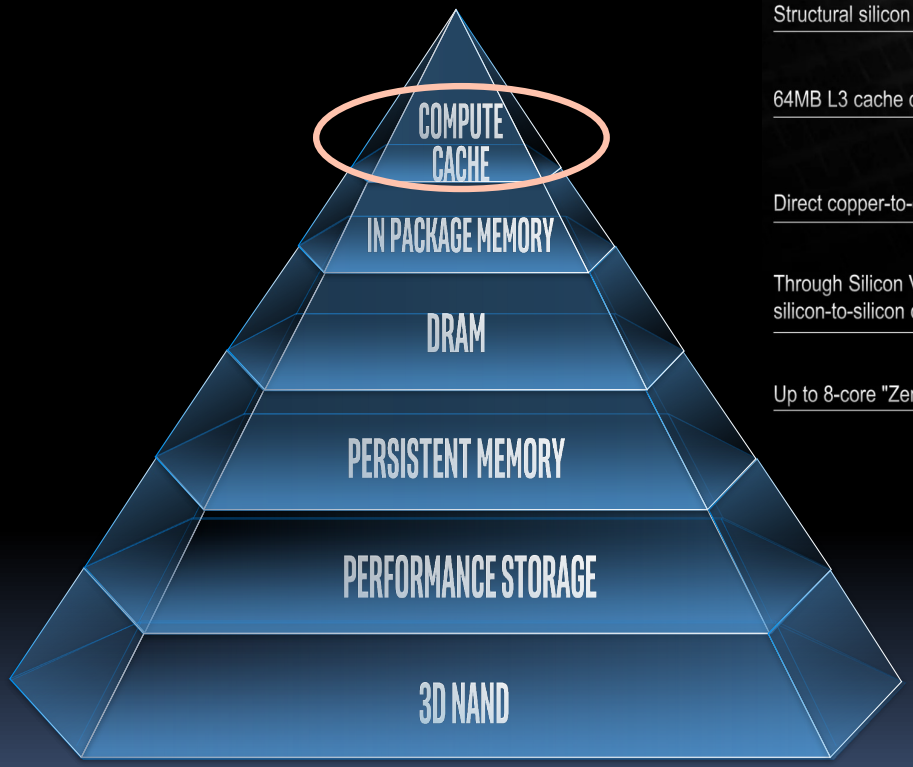
"It's the Memory, Stupid!

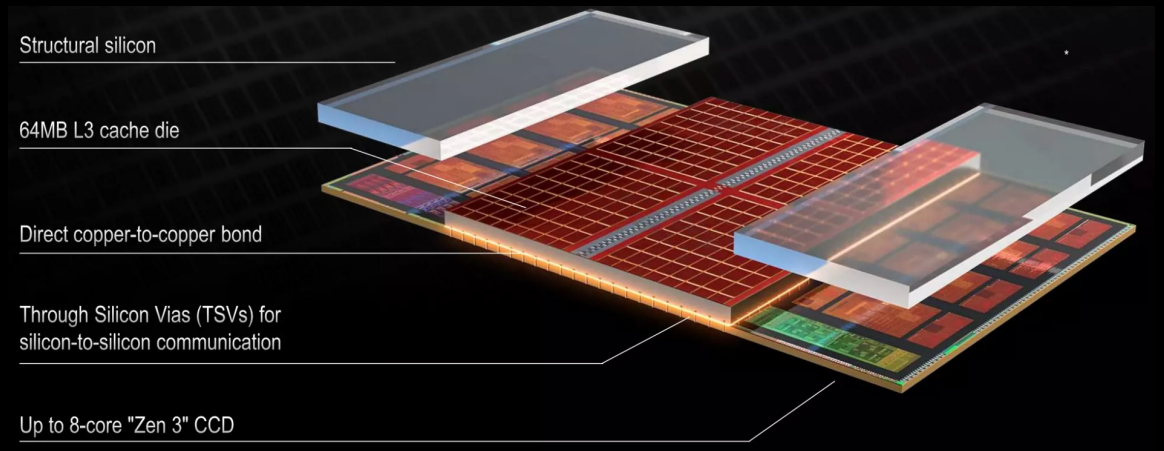
... Processor Speed Has Outstripped Memory ... today's chips are largely able to execute code faster than we can feed them with instructions and data. There are no longer performance bottlenecks in the floating- point ... The real design action is in memory subsystems—caches, buses, bandwidth, and latency...I expect that over the coming decade memory subsystem design will be the only important design issue for microprocessors."

Architects Look to Processors of Future,
Microprocessor Report, V10, No. 10, Aug/5/1996
Dick Sites, Digital Equipment Corp



Cache Memory in next decade



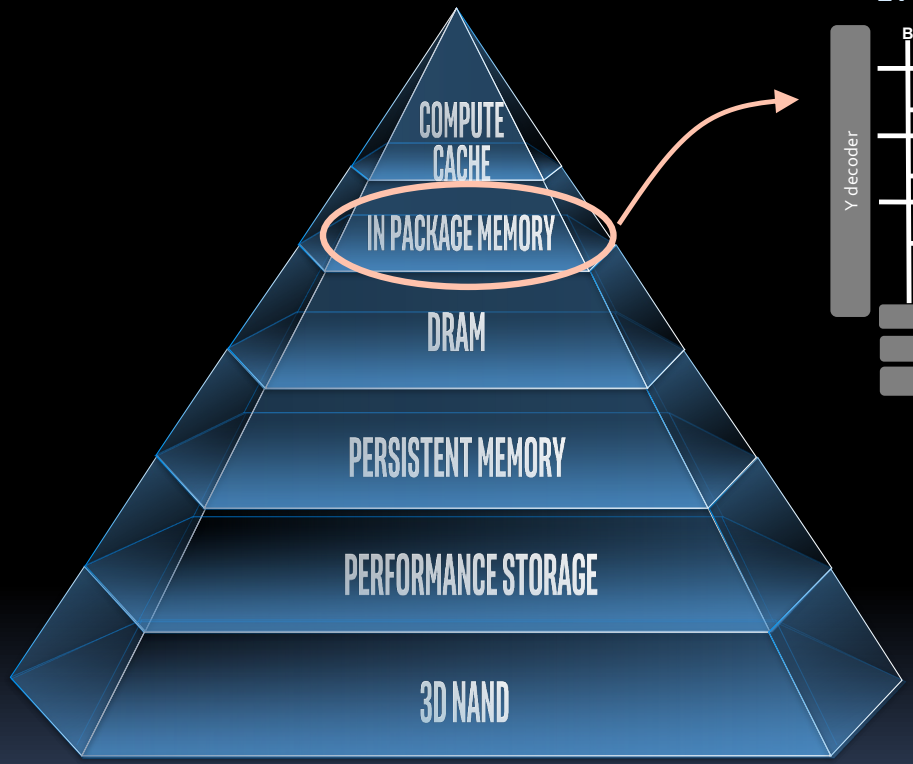


*L. Su, "High-Performance Computing: Services and Products Essential to our Daily Lives" Computex, 2021.

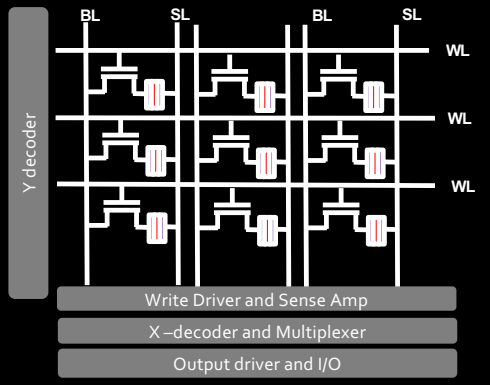
- Scaling trajectory of mainstream logic technology yet optimized for SRAM
- IP folding to increase Cache capacity : Core
  - Enabling technology: TSV, back-side metallization, reconstructed wafer, wafer-wafer (hybrid) bonding,
- Key Challenges: Power and Thermal

intel®

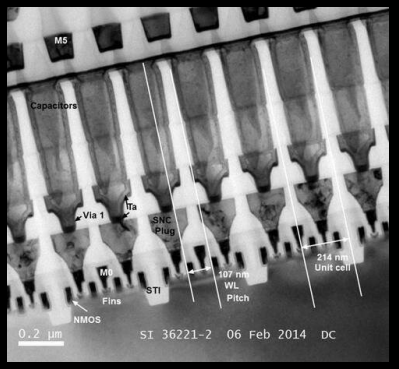
Emerging Memories for In Package Memory candidates



1T- 1X Array Architecture



Intel Gen 1 eDRAM [1T1C]



Access Transistor

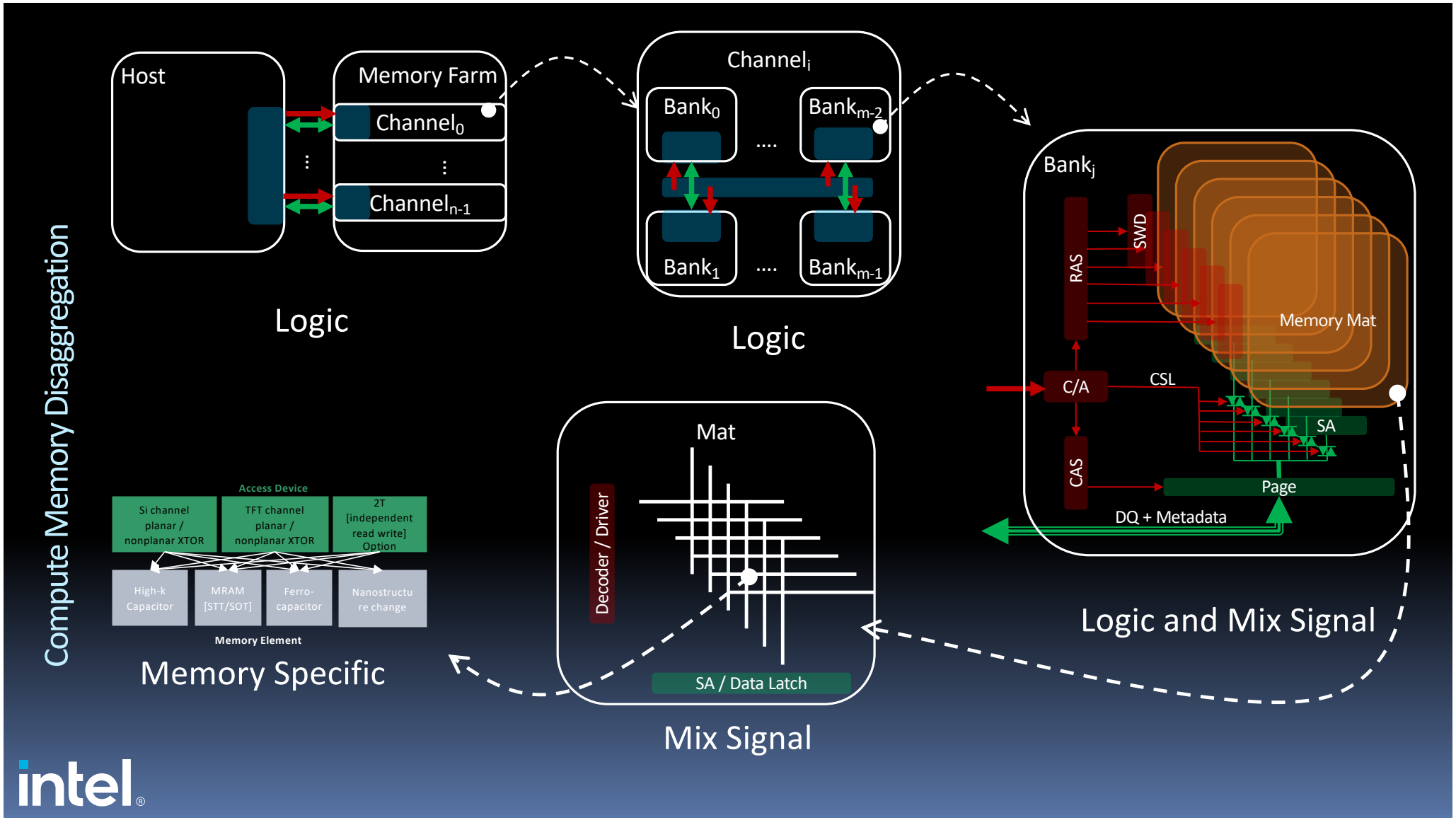
Si channel planar / nonplanar XTOR	TFT channel planar / nonplanar XTOR	2T [independent read write] Option

Memory Element

High-k Capacitor	MRAM [STT/SOT]	Ferro-capacitor	Nano-structure change

Key Challenge: BW, Granularity

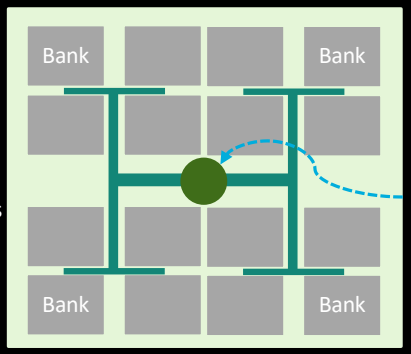
intel®


Compute Memory Disaggregation
Host
Memory Farm
Channel0
Channeln-1
Logic
Channeli
Bank0
Bankm-2
Bank1
Bankm-1
Logic
Bankj
RAS
SWD
Memory Mat
C/A
CSL
SA
CAS
Page
DQ + Metadata
Logic and Mix Signal
Mat
Decoder / Driver
SA / Data Latch
Mix Signal
Access Device
Si channel planar / nonplanar XTOR
TFT channel planar / nonplanar XTOR
2T [independent read write] Option
High-k Capacitor
MRAM [STT/SOT]
Ferro-capacitor
Nanostructure change
Memory Element
Memory Specific

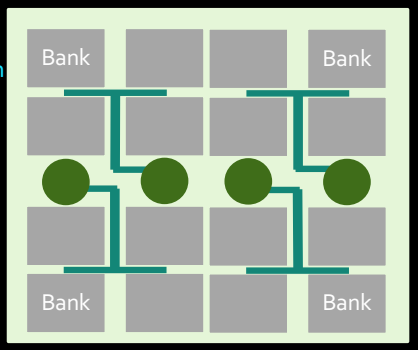
Chiplet stacking to improve packing density, Power and Bandwidth

M. O'connor, et.al., ACM/MICRO 2017

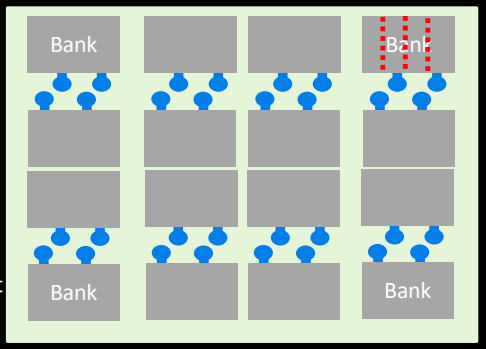
HBM2 Channels:
- Shared 16 GB/s Channel
- 16 banks per channel
- BW of idle banks is "wasted"



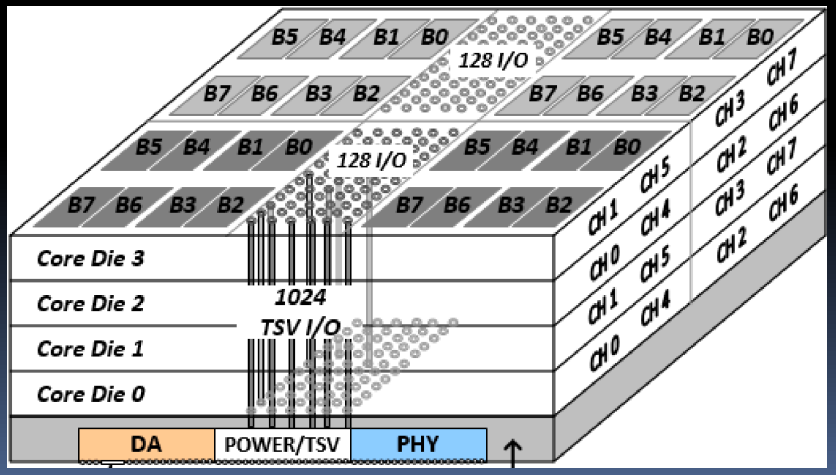
Quad-Bandwidth HBM Channels:
- Four 16 GB/s Channels
- Shared inter-bank bus
- High datapath & array act. energy



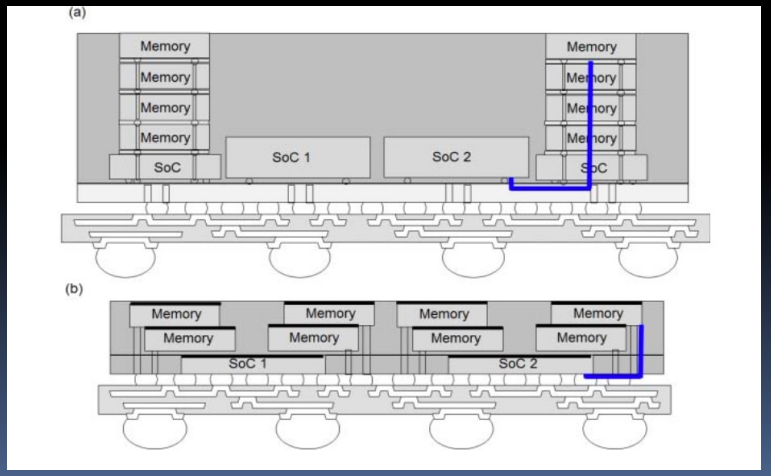
Fine-Grained DRAM Arch:
- Partition Bank into narrower pseudo-banks
- 32 Dedicatedm Local 2 GB/s Channel
- Local, parallel I/O per bank at lower energy



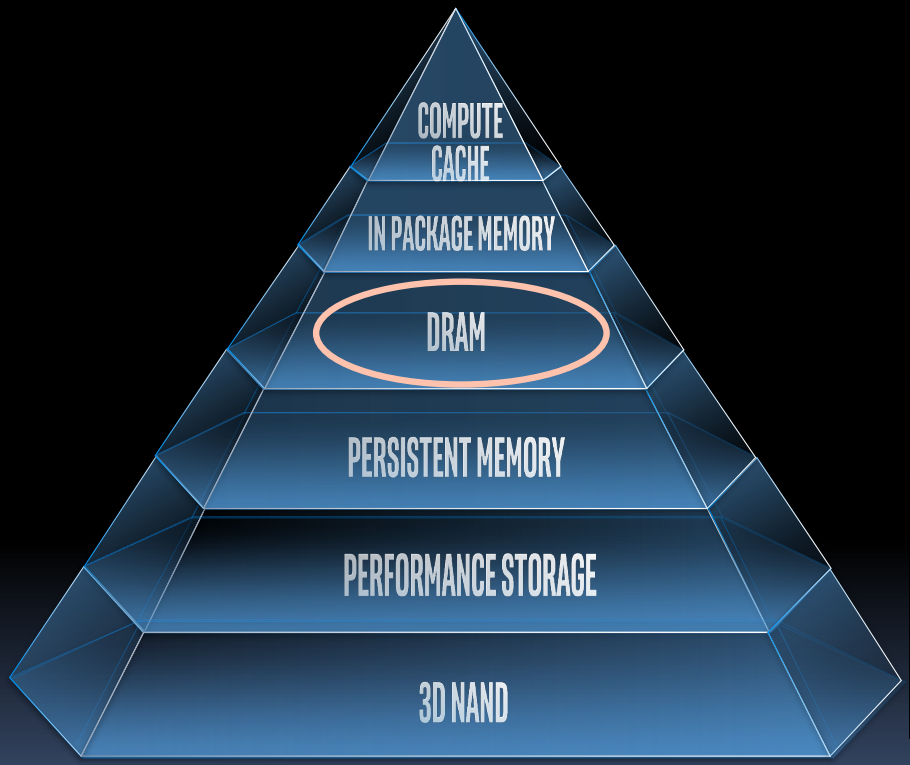
HBM, H Jun, et.al., IMW'2017

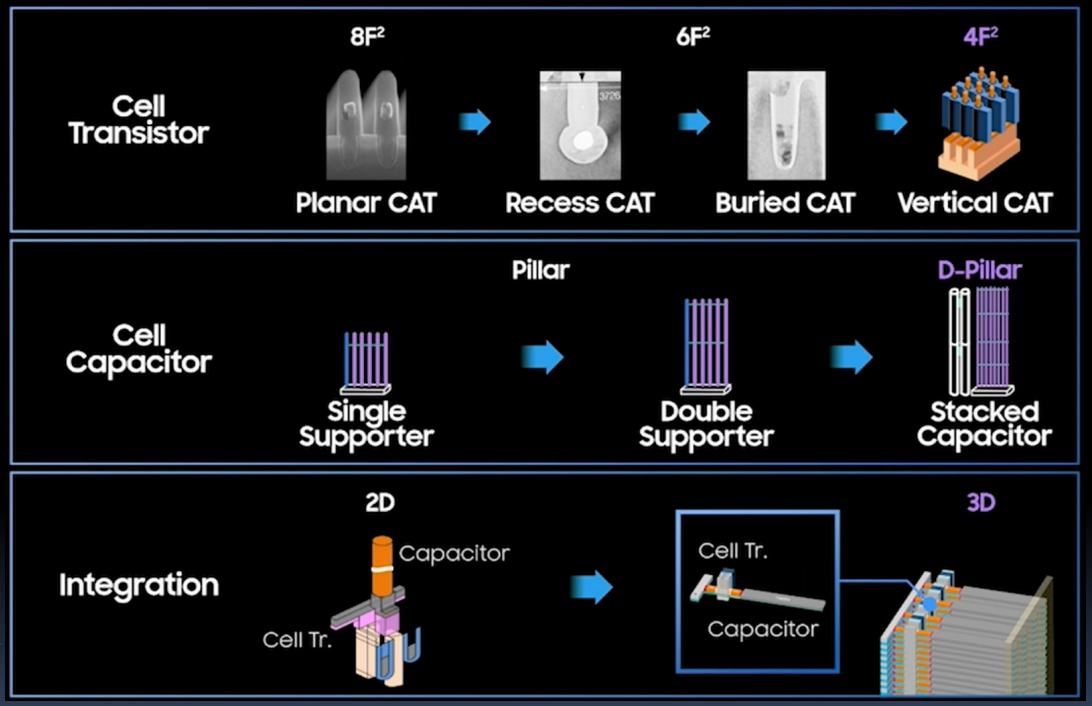


A. Su, et.al., ECTC2019



DRAM's Journey beyond Trench Capacitor

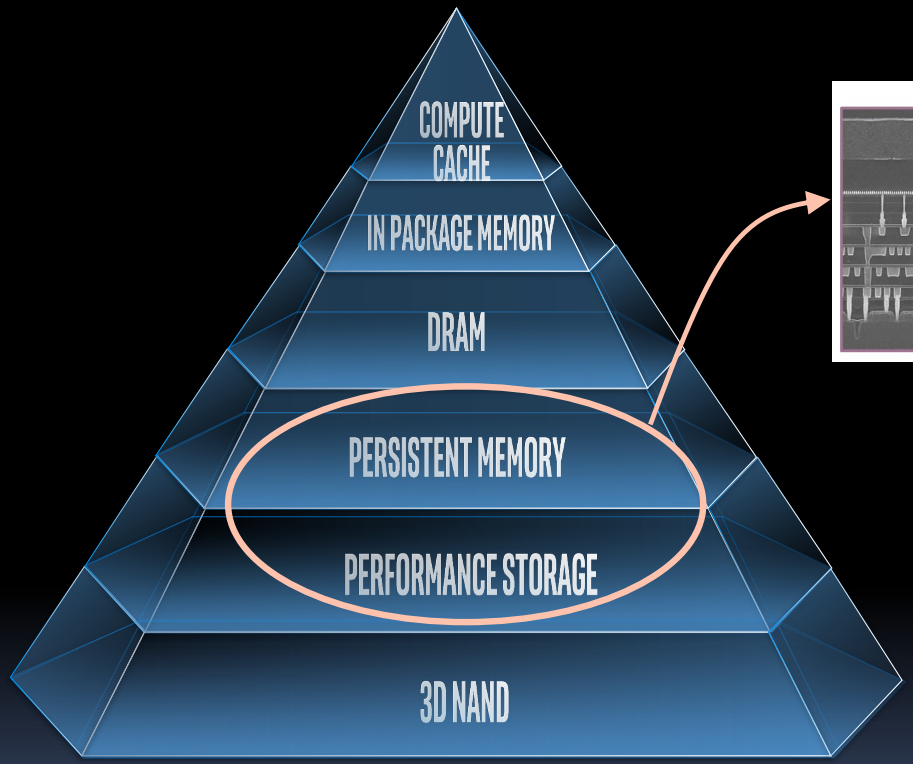




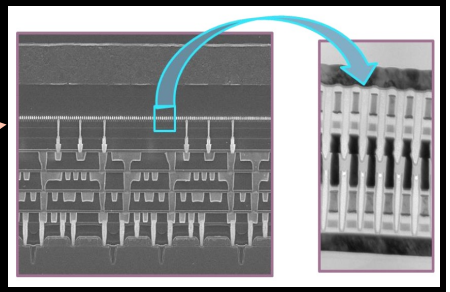
Kinam Kim, IEDM'21, Plenary Session 1.1

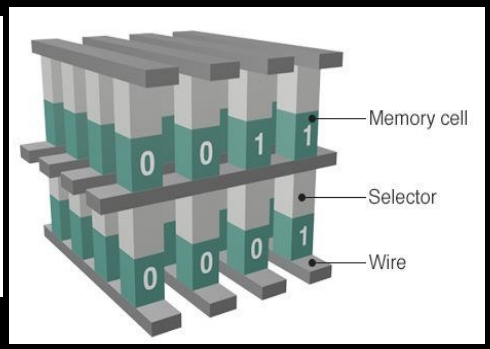
intel

Emerging Memories for Persistent Memory in 3D Cross Point Architectures



Intel Gen 1 3D XPoint™





Access Diode

Thermionic [diode, MIEC, ..]

Volatile filament

Metal-Insulator Transition

Ovonic Threshold

Magnetic [MRAM, STT(SOT)MRAM]

Resistance Change ReRAM[1D, 2D], PCM [3D]

Ferroelectrics

Memory Element

Key Challenge: Matching Memory & Selector

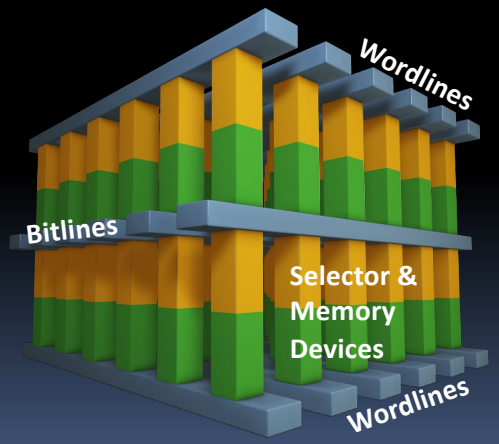
intel®

Consideration of architecting a cross point memory technology

Desirable Attributes: Nonvolatile, Low Cost, High Performance

- **Simple scalable, 3D stackable construction; low D·T, BEOL compatible with mainstream IC manufacturing**
- **Memory switching in atomistic state (bulk); superior scalability than electrostatic memory switching physics**
- **Individual cell access, small granularity**
- **Fast switching materials + local low resistance metal interconnect**

Challenge: to "mate" selector and memory in achieving a non-linear I-V for array access



Potential Memory Options:

- Spin polarization
- Phase change/segregation
- Ferroelectric hysteresis
- Interfacial barrier modulation
- Oxygen vacancy relocation
- Ionic transport
- NRAM (?)

Potential Selector Options:

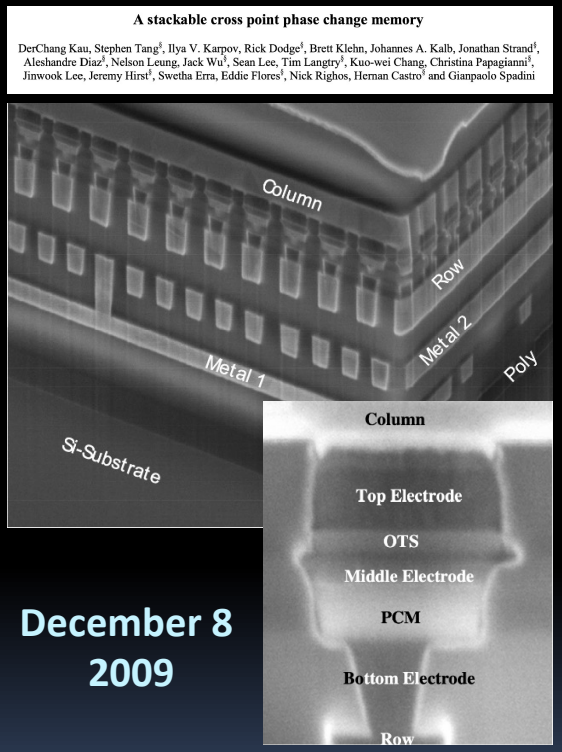
Unidirectional
- Homo/Heterogenous P/N Jx
- Schottky Barrier

Bidirectional
- Mixed ionic-electronic conduction
- Tunneling barrier
- Metal–insulator transitions
- Ovonic threshold switching

In Pursuit of a Dream …

September 28 1970

December 8 2009



August 13 2020

A Half Century Endeavor

D. Kau, VLSI-TSA 2021

Long gestation of Disruptive Memory Technologies

Ideation

1970

2003

PoC: Electron Devices

4Kb
200nm feature,
4-mask
Sublitho Heater
Damascene PM
Subtractive SD

2007

2010

PoC: Cell and Array

64Mb
180nm ½ Pitch
6-Mask Sublitho Heater
Damascene PM,
Subtractive SD

64Mb
100nm feature
2-Mask
Self-Aligned,
Subtractive Cell

PoC: Process Architecture

2012

2013

Research Graduation

64Mb
50nm ½ Pitch,
2-Mask
Dual Deck,
Self Aligned
Subtractive Cell

128Gb
41nm Pitch
Dual Deck
Self Aligned
Subtractive Cell

1st Production Generation

2015

2019

2nd Production Generation

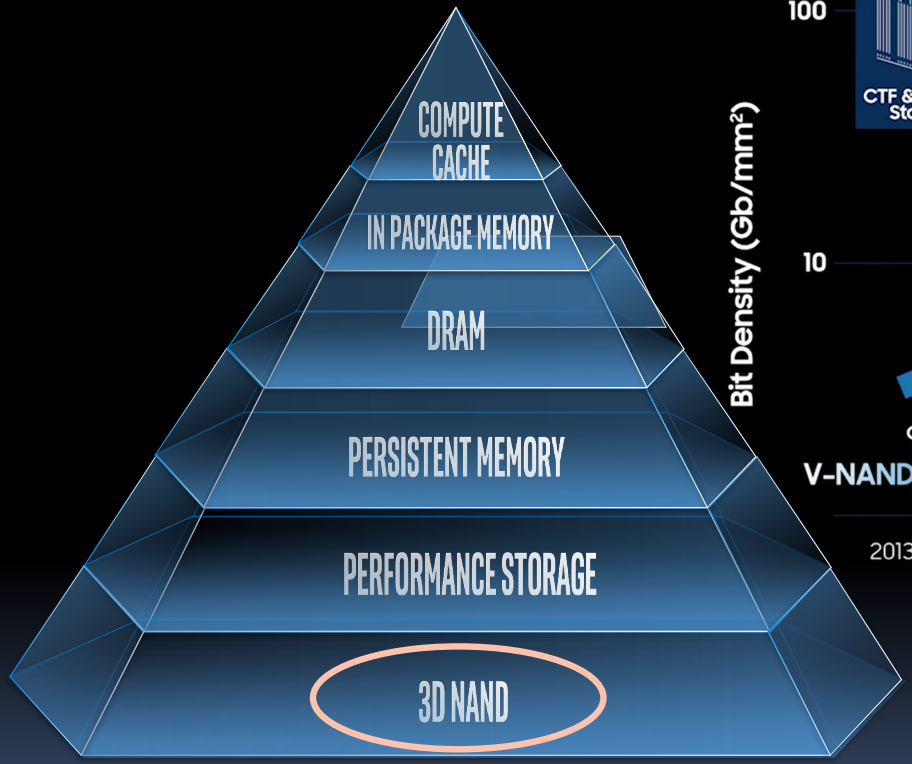
256Gb
41nm Pitch
Quad Deck
Self Aligned
Subtractive Cell

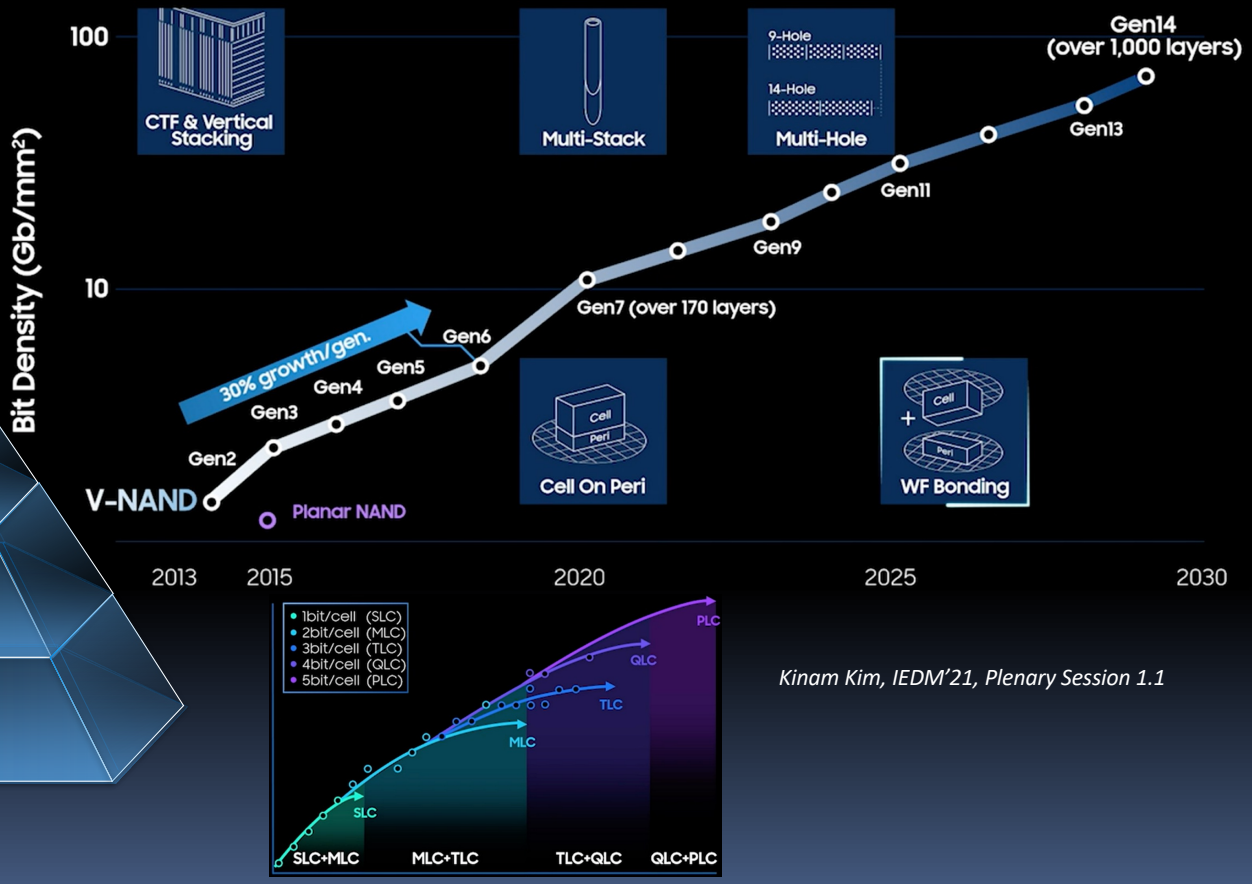
Scaled
Line/Space

3rd Production Generation

intel®

3D NAND in next decade will continue to drive $/bit on par with HDD





Kinam Kim, IEDM'21, Plenary Session 1.1

intel

Tiering/Decking engineering + multi-level placement algorithm

Takeaway of Advances of Memory Technologies

- Universal memory
  - Universal memory often sought, but hierarchy always existed and likely always will be
  - FoM must be benchmarked with respect to the corresponding *hierarchy of memories*
- Disrupt memory technology
  - Typically long gestation of Disruptive Memory Technologies
  - Mainstream memories hard to establish; hard to displace or to replace;
    DRAM replaced Core memory in 1960's and SSD displacing spin drive in 2000's
  - Memory technology research to productization can take decades and once productized, achieving mainstream is still not guaranteed
- Memory Subsystem
  - Modular, Modular, Modular
  - Power/Thermal conscious: Energy from I/O and from moving bits back/forth becoming limiting
  - 3D (monolithic, wafer-wafer) and packaging of multi-chip
    → TTM and Cost driven
  - Pair memory tech research with compute architecture/hierarchy research.
  - Software (often ignored) enabling to alleviate adoption barriers or to unlock value