A Deep dive into 3D-NAND Silicon Linkage to Storage System Performance & Reliability

Jung Yoon, Ranjana Godse, Andy Walls
IBM Systems

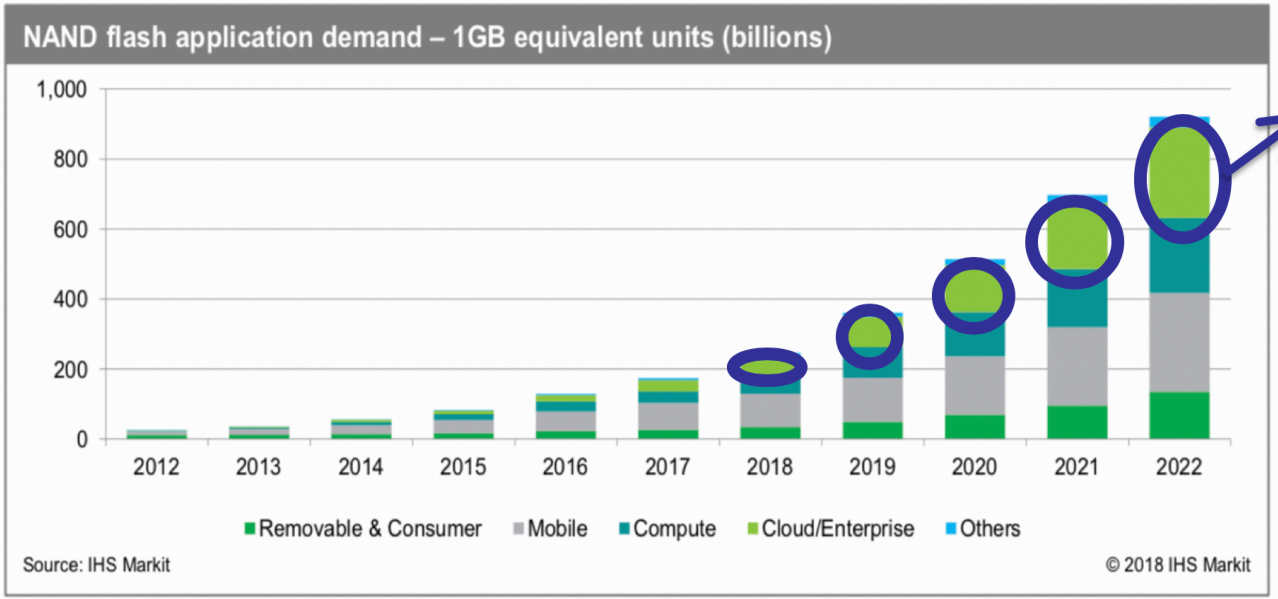
Flash Memory Summit

Agenda

1. 3D-NAND scaling – driving force for exponential flash market growth
2. Deep dive into 3D NAND Silicon Technology
    - Layer Count Scaling outlook vs challenges
    - Charge Trap vs Floating Gate key considerations
    - 3D-NAND Manufacturing & Reliability
3. 3D NAND scaling – Key factors from AI & Cloud System perspective

Flash Memory Summit

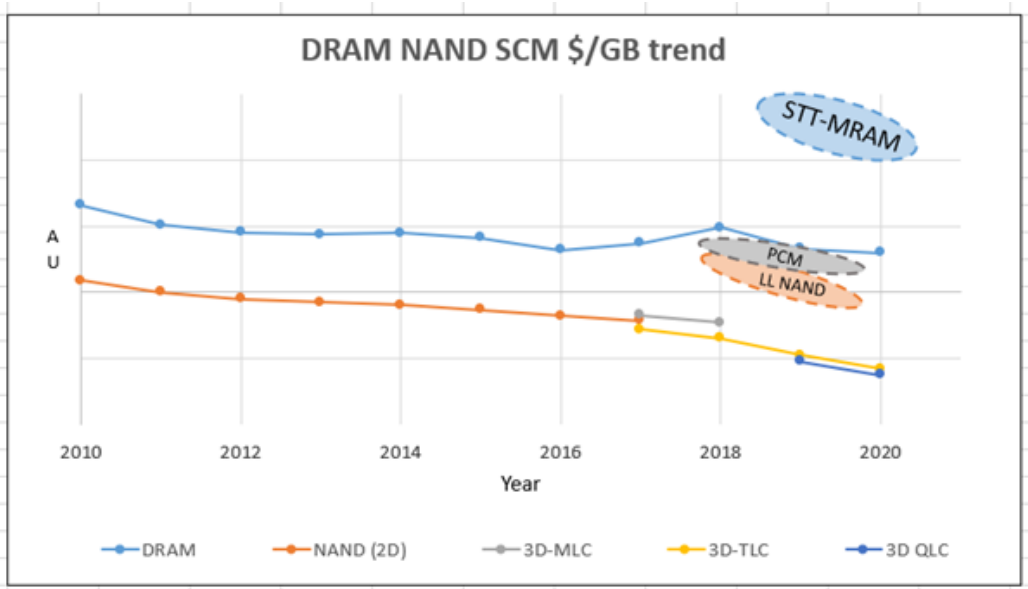
NAND Continues its Explosive Growth



Large growth in Enterprise and Cloud.

- 3D NAND driven scaling enabling aggressive $/GB reductions thru 2022+
- 3D TLC endurance gains enabling significant Enterprise & Hyperscale SSD market growth
- 3D QLC will continue to drive capacity growth for expanding capacity and workloads

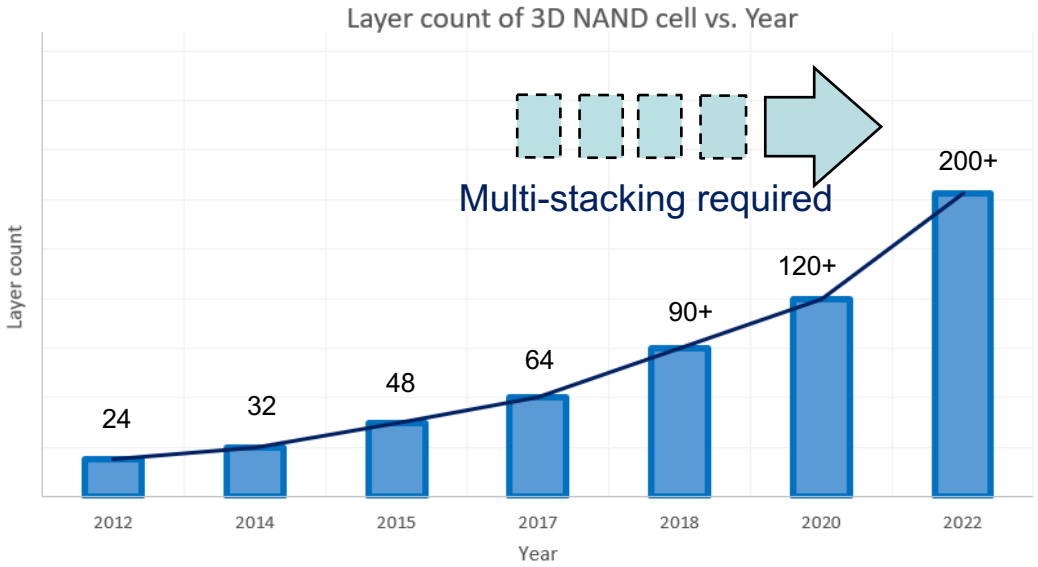
Memory Bit Cost Curve vs Scaling

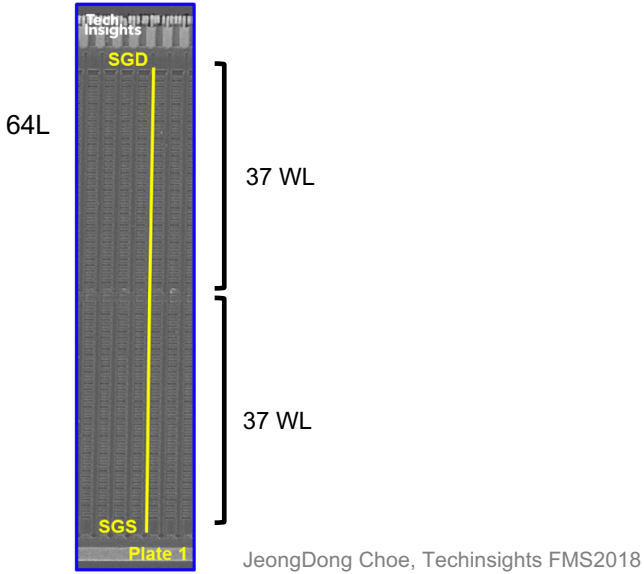


Source – IBM, Gartner, IDC, DeDios,

- DRAM, NAND $/GB reduction in driven by technology scaling and supply/demand dynamics
- Flash bit cost reduction driven by i) 3D-NAND layer count (N) scaling (48L > 64L >96L > 120L), ii) Array efficiency, iii) lithographic (x-y) process scaling
- Flash Market Demand growth – fueled by high density & low cost 3D NAND, Enterprise SSD, Mobile applications growth

3D-NAND Scalability considerations



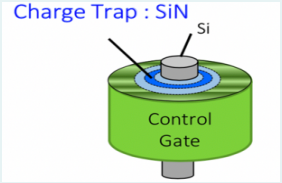
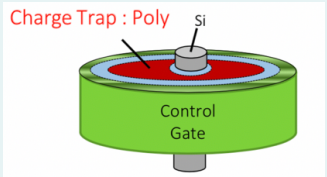


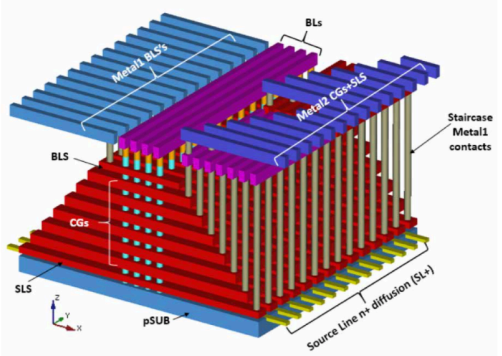
JeongDong Choe, Techinsights FMS2018

- Channel etch AR challenges driving need for Multi Stacking – all suppliers expected to transition to 2 high Multi-stacking process at >120 Layer 3D-NAND generation,
- Additional process steps (lithography, thin film deposition, etch etc.) adds cost
- Overlay accuracy of Upper and Lower channel critical

Flash Memory Summit

3D-NAND Scalability considerations

	Charge Trap	Floating Gate
Cell		
Cell Size	1	~1.3
Scalability	+	-
Retention	+	++
Endurance	+	-



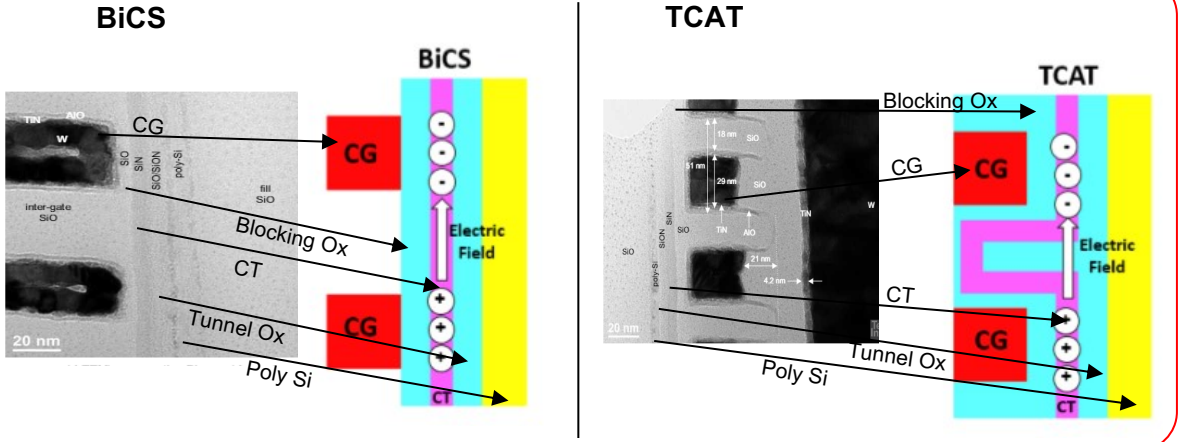
J. Jang, Proc IEEE Symp VLSI Tech, 2009, pp192-193

- FG has larger 3D Pillar due due to presence of Floating Gate poly-Si – lower scalability, but more stable charge in storage layer (better retention)
- Lithographic (x-y) scaling key factor in bit cost reduction for >120 Layer continued 3D-NAND scaling, Bit-line pitch, staircase contact process

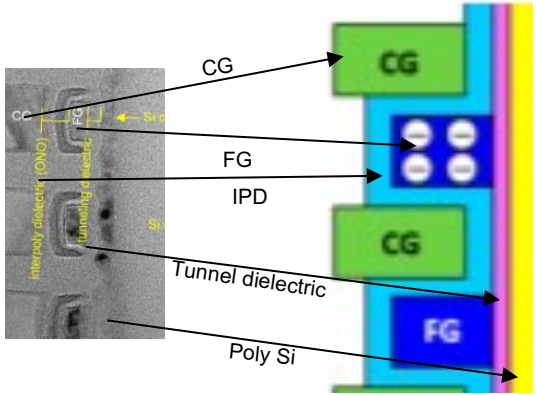
Charge Trap vs. Floating Gate

Flash Memory Summit

Charge Trap



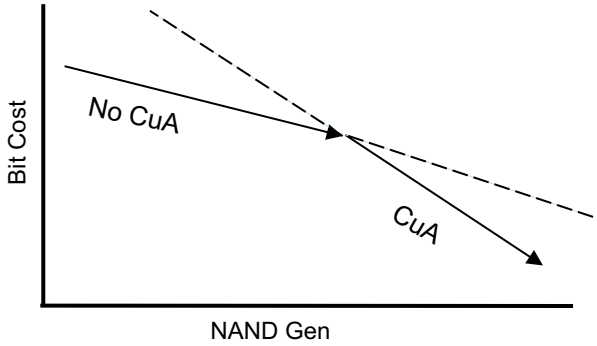
Floating Gate

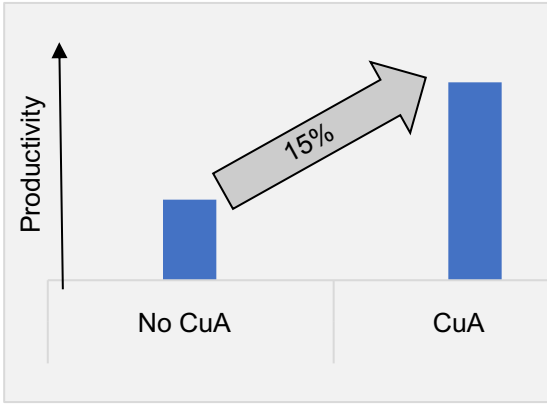


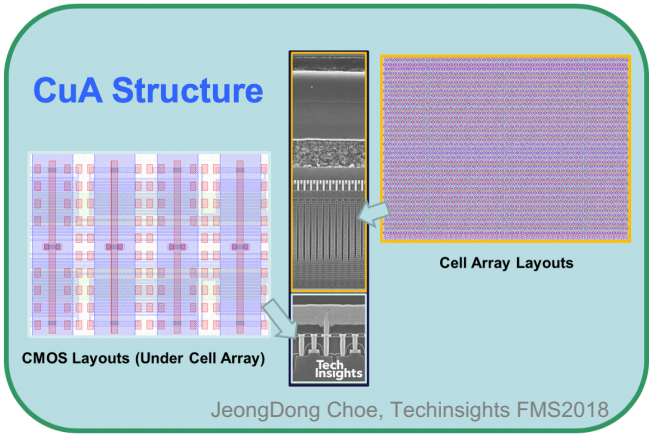
1) Micheloni, R. 3D Flash Memories; Springer: Dordrecht, The Netherlands, 2016.
2) Techinsight 2019

- CT cell experiences charge spreading resulting in early data retention loss effects, temperature dependent charge loss mechanism
- 3D-NAND Data retention mechanisms and transient effects are complex and dependent on cell structure and material – strong dependence on 3D-NAND cell (Charge Trap vs Floating gate)

3D-NAND Scalability considerations





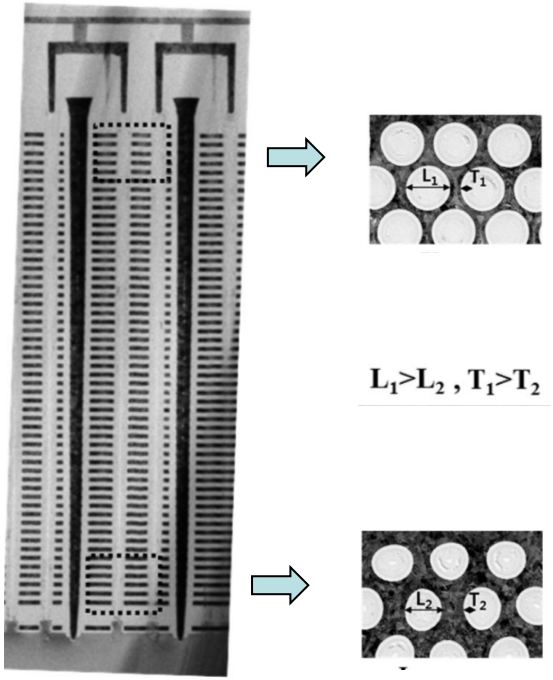


- CMOS Under Array (CuA) technology is key enabler of die cost reduction via die size reduction – key for 3D-NVM technologies (3D-NAND, 3DXP)
- Efficient layout design and peripheral transistor scaling
- Overall die cost determined by net die per wafer (die area) vs additional process steps needed for CuA
- Thermal budget management critical in during 3D-Cell integration – lower Contact Resistance (Rc) required for high speed IO

3D-NAND Manufacturing & Reliability Mechanisms

Z-directional cell characteristics

- Channel hole size and Gate stack thickness causing difference in Program/Erase/Read speed, and retention characteristics
- Program/Erase speed of lower WL is faster due to higher coupling ratio originated by smaller channel hole CD
- Retention characteristics of lower WL poor due to insufficient thin film step coverage resulting in thinner gate stack
- Difference in characteristics of WL leads to decrease in read Vth window



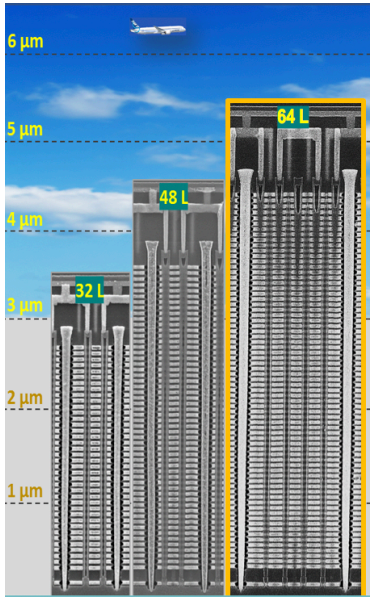
H. Kim, Samsung, 978-1-5090-3274-7/17 IEEE 2017

3D-NAND Manufacturing & Reliability Mechanisms

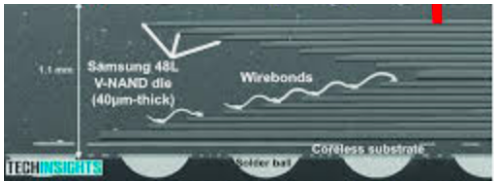
Flash Memory Summit

Z scaling & Cell-to-Cell interference

- Thickness of unit cell with 3D-NAND generations decreases – leading to Cell to Cell Interference challenges similar to 2D-NAND
- Total 3D-NAND stack height to reach 10um at > 200 Layer generations
- High Aspect Ratio Channel Etch & multi string stacking - key in future 3D-NAND scaling
- Stacked package (16DP, 32DP) requires Si thinning <25 um thickness. 3D-NAND overall stack height drives need for reduction of WL and dielectric mold thickness
- 3D-NAND reliability, performance and power scaling will face difficulties due to higher cell-to-cell interference, and heavier WL RC loading, and precise staircase contact placement accuracy



JeongDong Choe, Techinsights FMS 2018



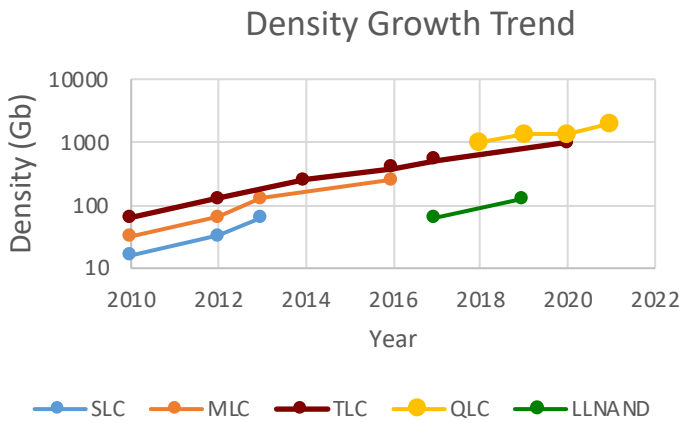
3D-NAND Manufacturing & Reliability Mechanisms

- 3D-NAND Transient Effects/Mechanisms – requiring background read refresh. Effect to system performance consistency if refresh frequency gets too high
  - Grain Boundary Trap in Poly-Si Channel – Floating body effect
  - Electron transport mobility
- RTN (Random Telegraph Noise) – random placement of discrete dopant atoms and trap at tunnel oxide/channel interface
- Stress management in integration of complex 3D-NAND structure
- Fab process controls – thin film deposition, etch, lithography alignment (staircase contact, multi-stacking)
- Fab particle control – in line defect metrology, early life quality & reliability

3D-NAND Scaling – Key Factors from Systems Perspective (AI & Cloud)

Flash Memory Summit

1. **Density**
   - TLC, QLC multi Terabit density – can allow for lots of data to be accessed with low latency
   - 3D-Flash is ideal for Cloud – data center power, cooling, footprint advantage
   - For a given capacity Si die, need more performance per die – multi plane read/write features become important (i.e., more writes & reads in parallel) to reduce Write Amplification
   - Stacked package >16DP, TSV driving industry's highest Density/mm^3 package, however may face technology limits in Si wafer thinning with 3D-NAND scaling, and Bus bottlenecks limiting throughput



Flash Memory Summit

3D-NAND Scaling – Key Factors from Systems Perspective (AI & Cloud)

2. Flash Architecture

- Increasing number of Pages/Block, increasing Block size
- Cell current for sense margin decrease with 3D-NAND layer count driving increased Block size
- Increasing Block size driving large amount of data in Flash Translation Layer with Block retirement scheme
- Page retirement data structure is complex. Driving complexity in garbage collection and increase in Over Provisioning

3. Cost

- 3D-NAND scaling via layer count increase
- Memory Array efficiency (CuA) – Die size vs added process steps
- TLC vs QLC cost factors

Flash Memory Summit

3D-NAND Scaling – Key Factors from Systems Perspective (AI & Cloud)

4. Performance

- Real time AI & workloads require ingestion of large amount of data at high rate and very high throughput (Read throughput #1, Write throughout #2)
- As datasets used for training ML/DL are growing, Flash with its low latency and high throughput is optimal for AI storage
- Flash enables high IOPS with low read latency, lower power and smaller footprint compared to HDD

5. Reliability

- PE Cycling endurance, Vth window, 3D-NAND process controls, consistency & uniformity
- Data Retention, Complex 3D-NAND charge loss mechanisms (Charge Trap & Floating Gate)
- AI driving read intensive workloads – understanding of read disturb mechanisms & process improvements

6. Power

- Program & Read energy/Byte scaling vs 3D-NAND generation critical
- AI workload & utilization, Cloud system power linkage with 3D-NAND Vcc, Icc, tProg/tRead key parameters

Summary

Flash Memory Summit

- 3D-NAND scaling will continue to fuel exponential flash market growth thru 2022+. This will be enabled by continued $/GB reduction via 3D-NAND cell layer scaling, CuA technologies, and lithographic shrinks.

- Process technology & Materials innovations will be needed to continue 3D-NAND scaling at >200 Layers, managing cell-to-cell interference, channel current, read/write latency, and power challenges

- AI & Cloud growth will be accelerated by 3D-NAND technology scaling via high throughput density, lower power consumption, and smaller system footprints - Flash process, materials, design & architecture innovations will be needed

- 3D-NAND scaling driven cost reductions, throughput density increases, and lower power consumption will enable significant flash growth focused on AI and Cloud workloads and applications