Flash Memory Summit

Achieving Robust Reliability Design in High Bit-Density 3D Flash

Jiezhi Chen

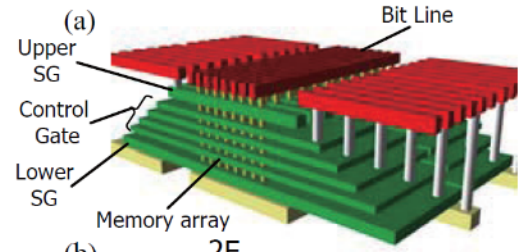
Shandong University, China

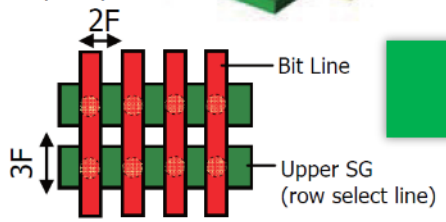
Email: chen.jiezhi@sdu.edu.cn

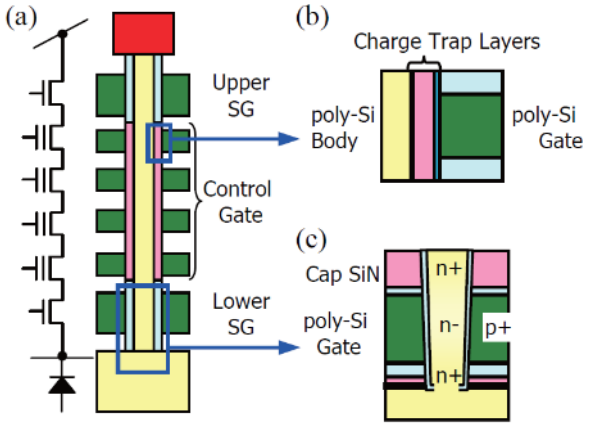
3D Technologies of Flash Memory

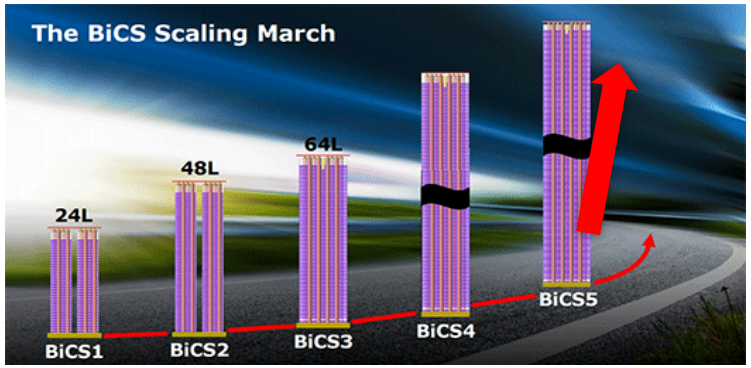
Flash Memory Summit

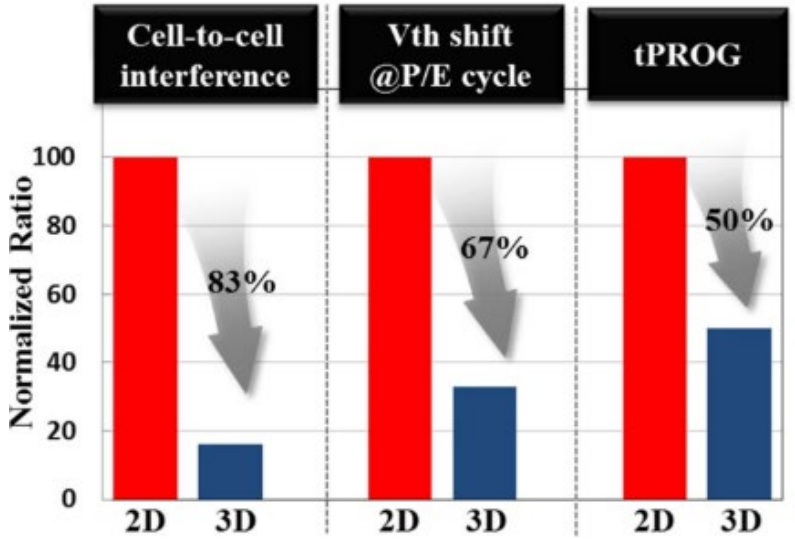
Toshiba, 2007



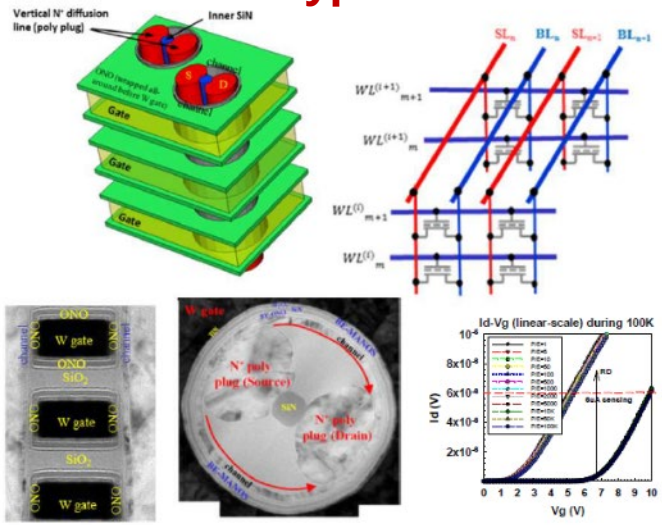


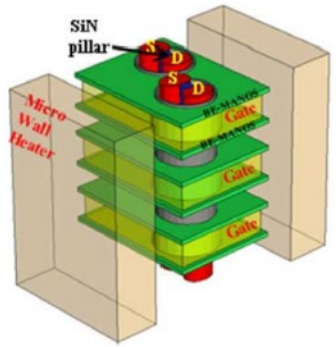


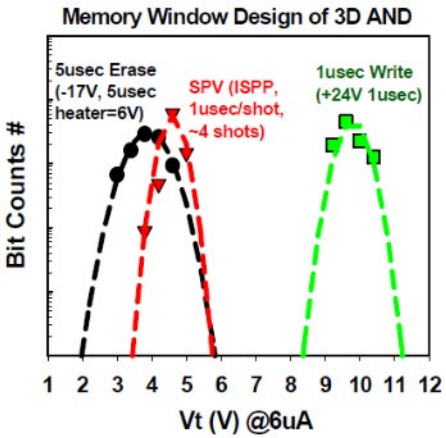




3D AND-type NOR Flash







@H. Tanaka et al., VLSI 2007; Kim H, Ahn S, et al. IMW 2017; H. T Lue, short course, VLSI 2021; open sources in websites

Outline

Flash Memory Summit

Sources of Error bits

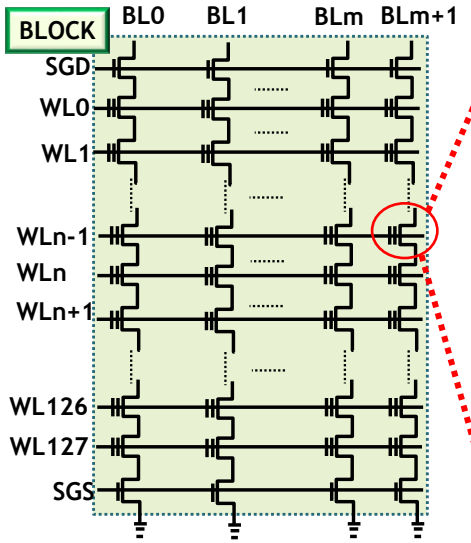
Optimization Strategies

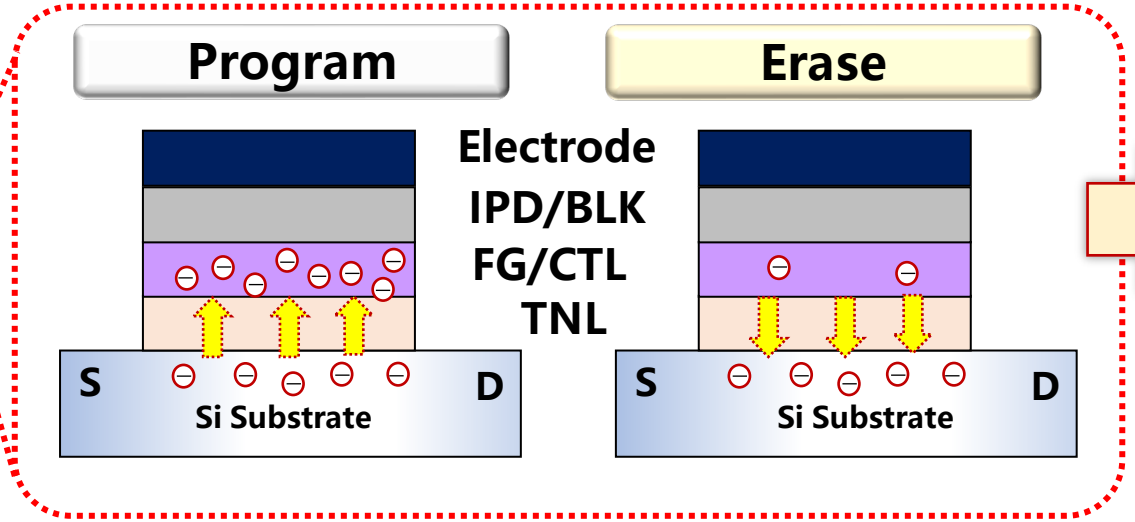
Future Technologies

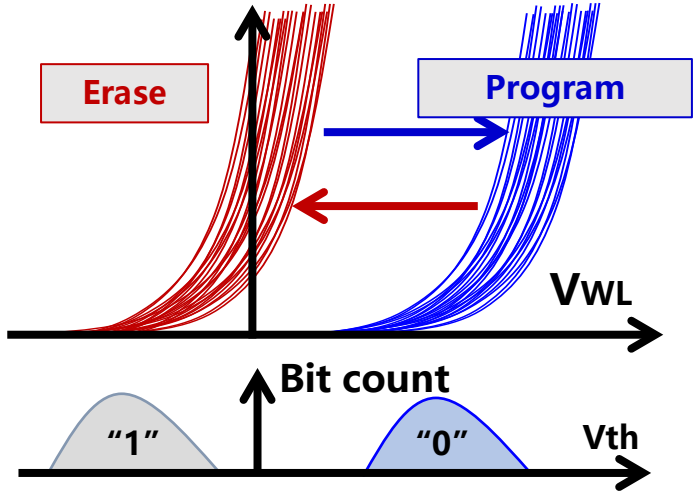
(I) Flash: Error Bits

Flash Memory Summit

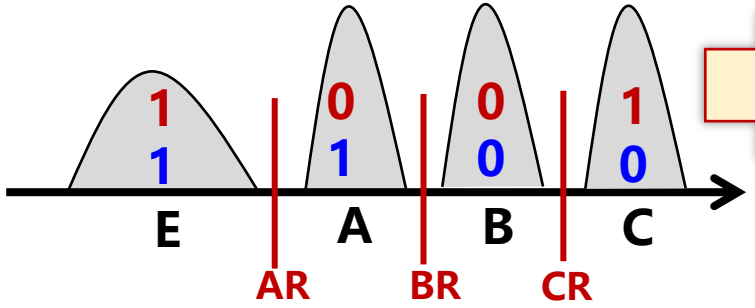
- Cells/arrays variations, noise, disturb, interference, instabilities…

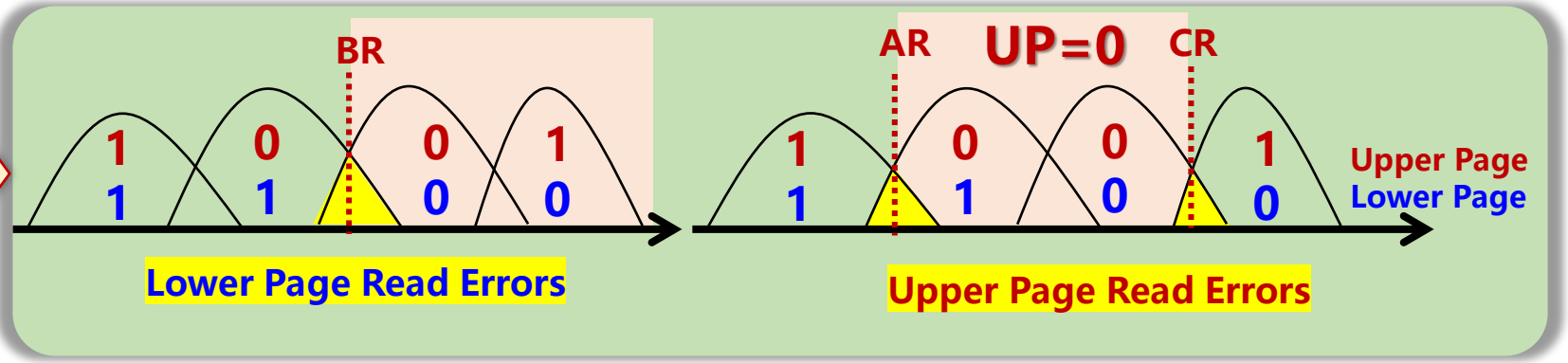






Ideal case w/ wide read margins

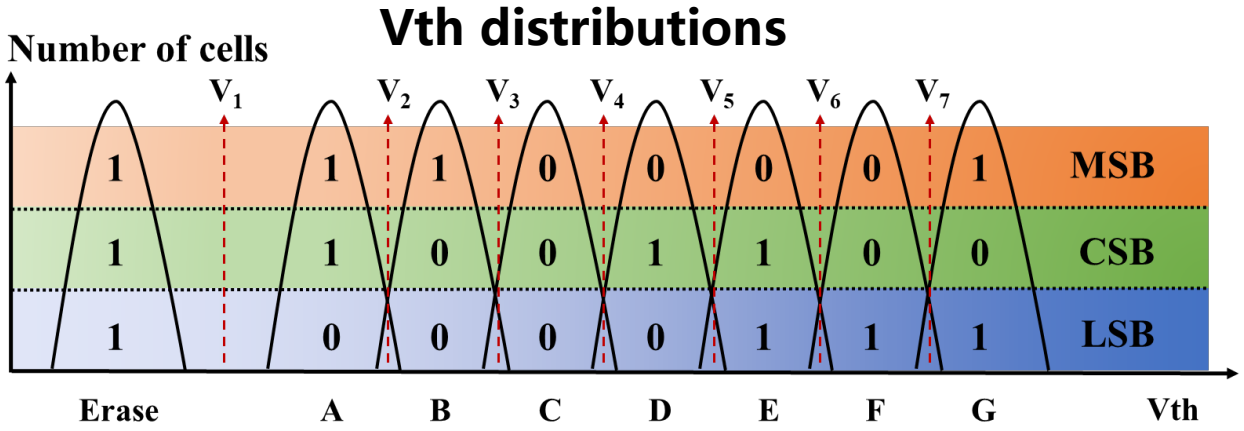




(II) Flash: Error bits classification

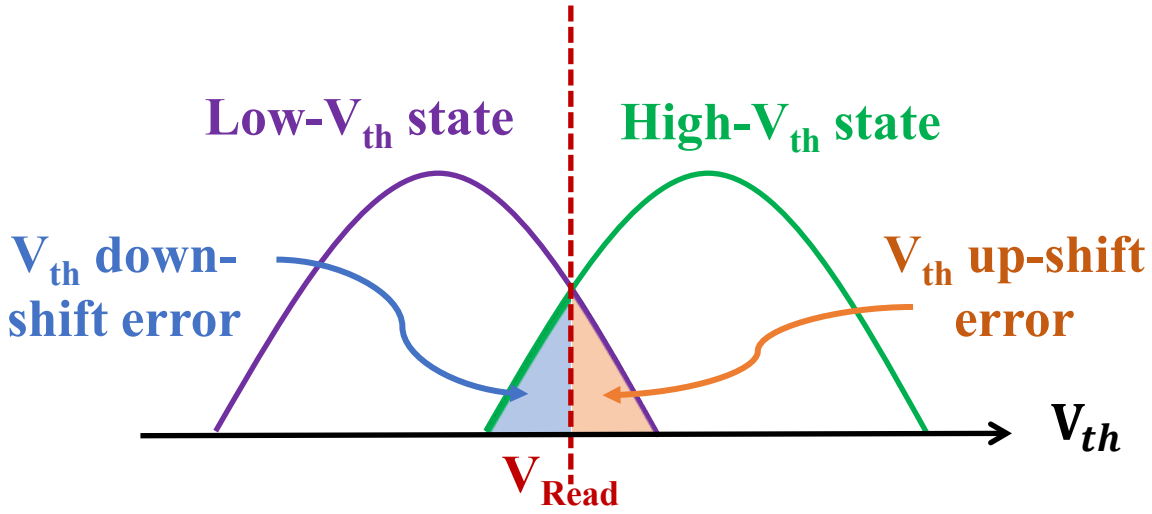
Flash Memory Summit

- **Overlap of Vth distributions exists between neighbor states**



Fail Bit Count (FBC)：Total fail bit count

$$Raw\ Bit\ Error\ Rate = \frac{Total\ error\ bits(FBC)}{Total\ data\ bits}$$



- ✓ **Down-shift Errors: Charge-loss**
- ✓ **Up-shift Errors: "unwanted" charges**

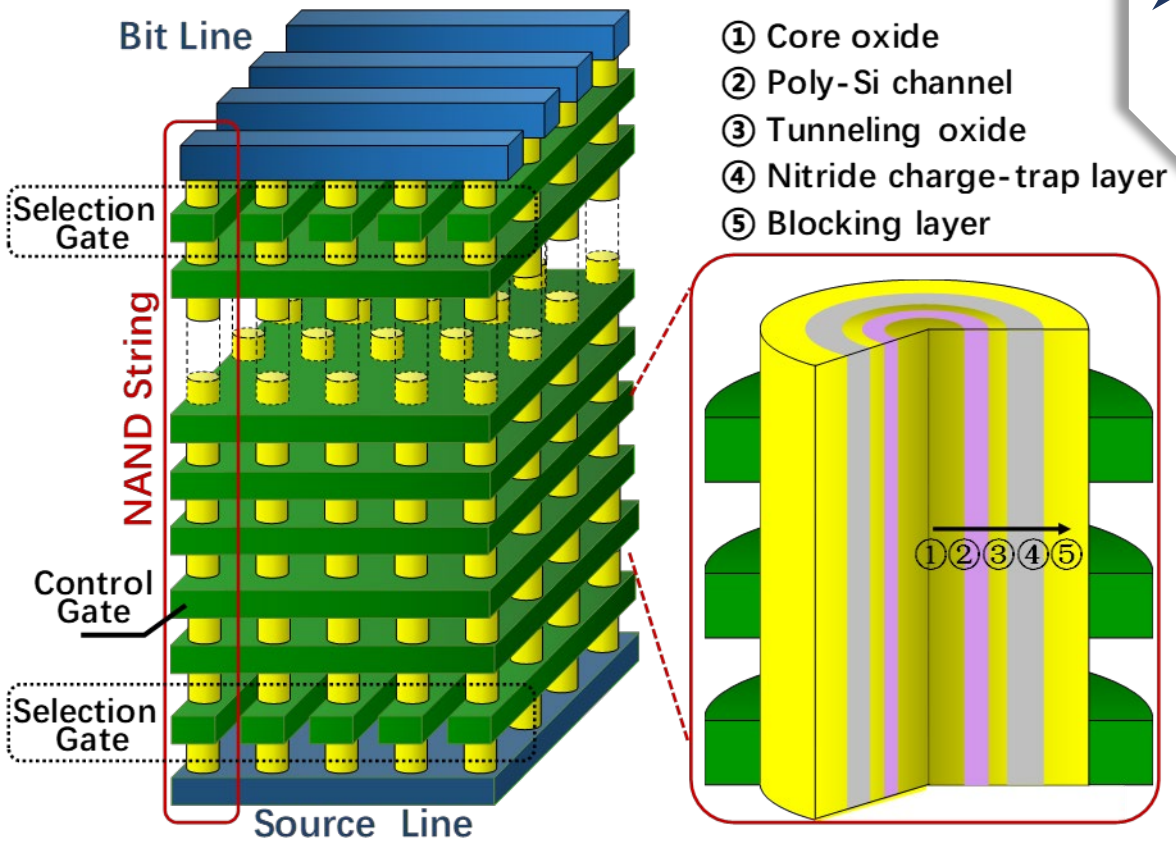
TLC → QLC → PLC → HLC →…

in a limited memory window

(III) 3D NAND: Structure & Material & Process

Flash Memory Summit

3D NAND is different from 2D NAND



- Poly-Si channel
  - Common CTL
- GAA Structure
  - 3D Integration

→ PE/DR/RD properties

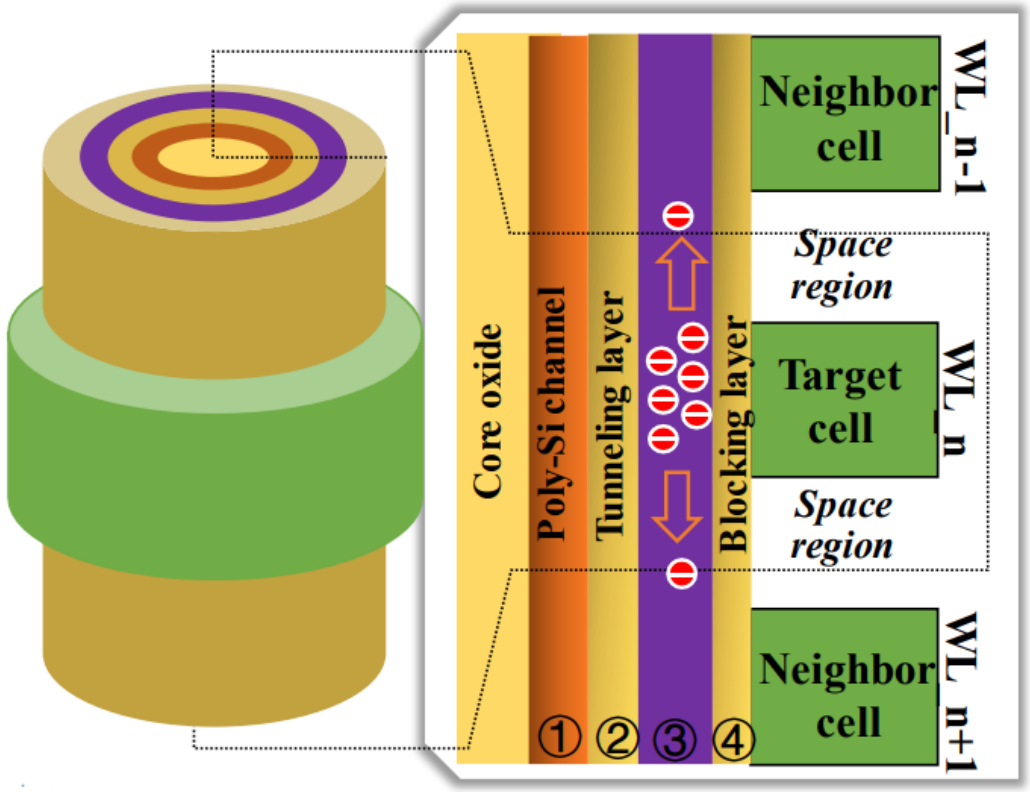
→ Temperature dependence

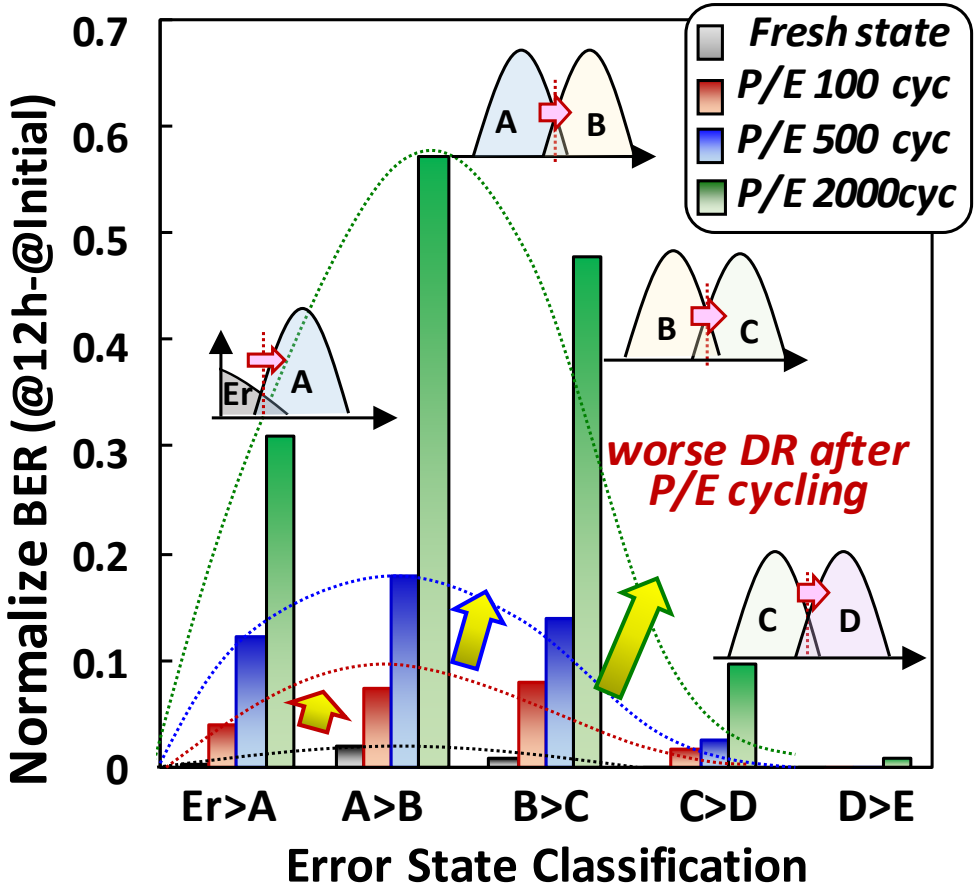
- **How to Improve Reliability?**
- **Lifetime prediction method?**

(IV) Lateral charge migration @CTL

Flash Memory Summit

- **Lateral charge migration increases the complexity of the charge movement, as well as the temperature dependence**

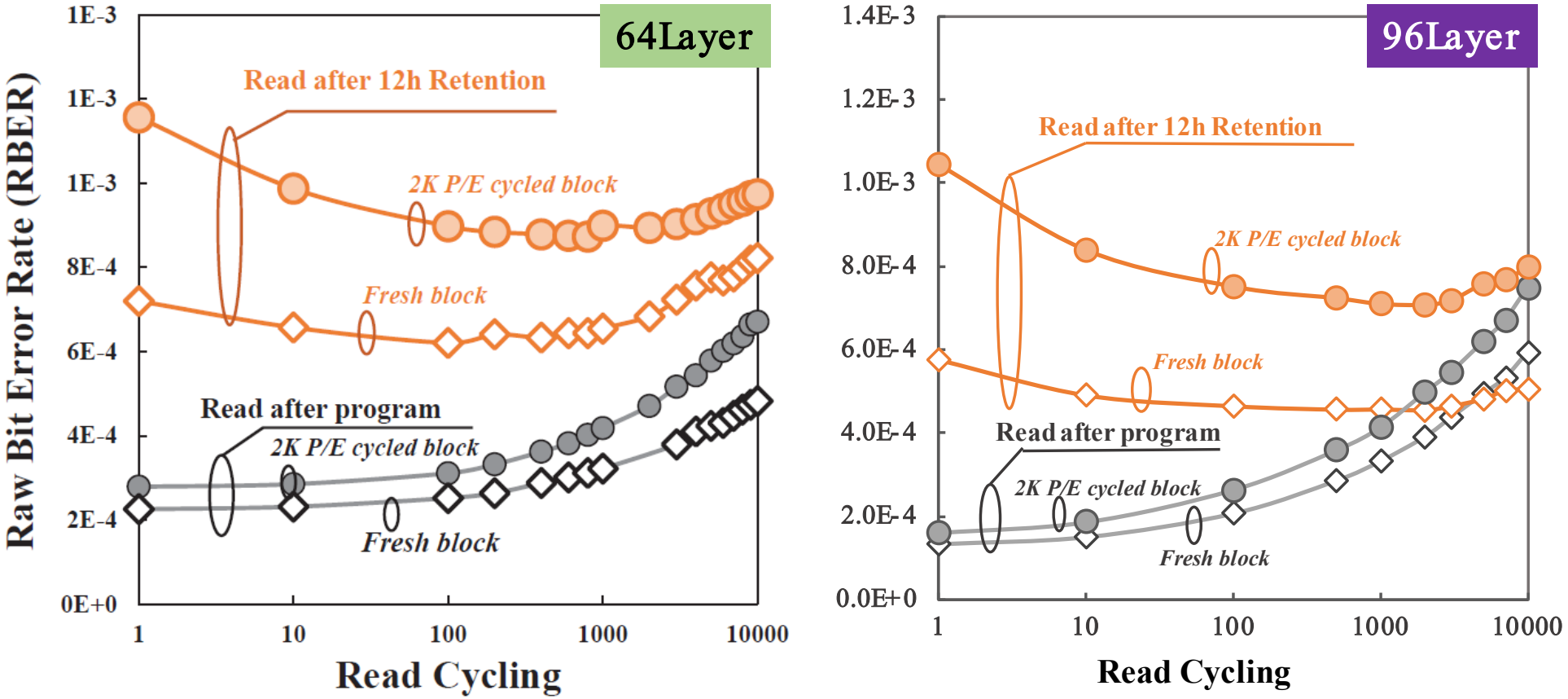




Ref: R. Cao, et al., IRPS 2019; K. Xie, et al., ICTA 2022

(V) DR-correlated RD

- **In 3D NAND, read disturb is strongly correlated to the retention w/ strong dependence on retention time.**

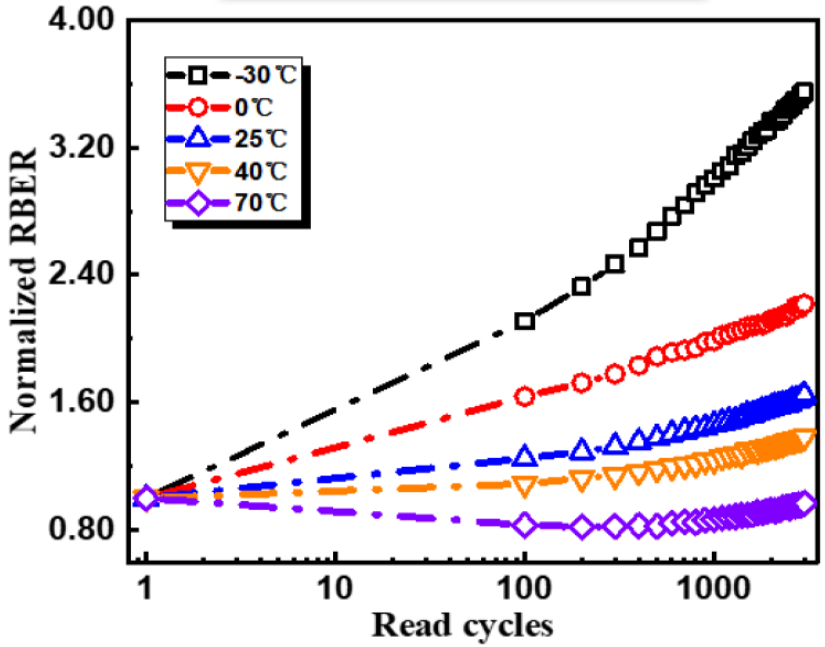


Ref: Y. Kong, et al., IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2020, 39(11): 4042-4051

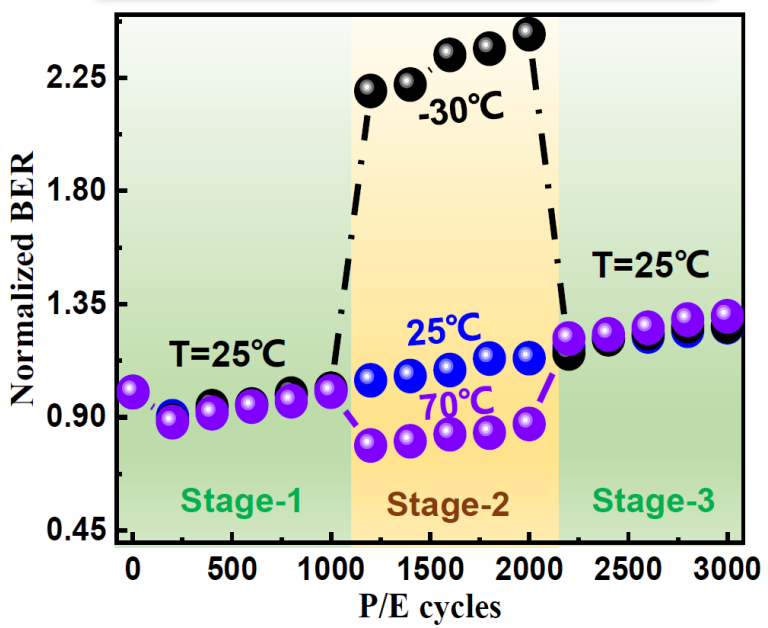
(VI) Temperature effects

- **PE/RD/DR are quite different from 2D NAND Flash**

RD vs Temperature



PE @Cross-Temperature



- **Traditional models in 2D NAND cannot be used in 3D NAND**

Ref: F. Chen, et al., micromachine 2021

Outline

Flash Memory Summit

Sources of Error bits

Optimization Strategies

Future Technologies

Achieve Robust Reliability in 3D NAND ?

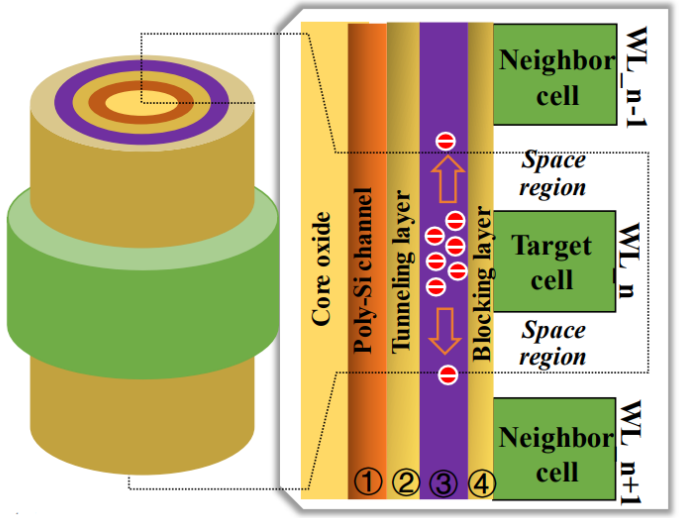
Flash Memory Summit

Process

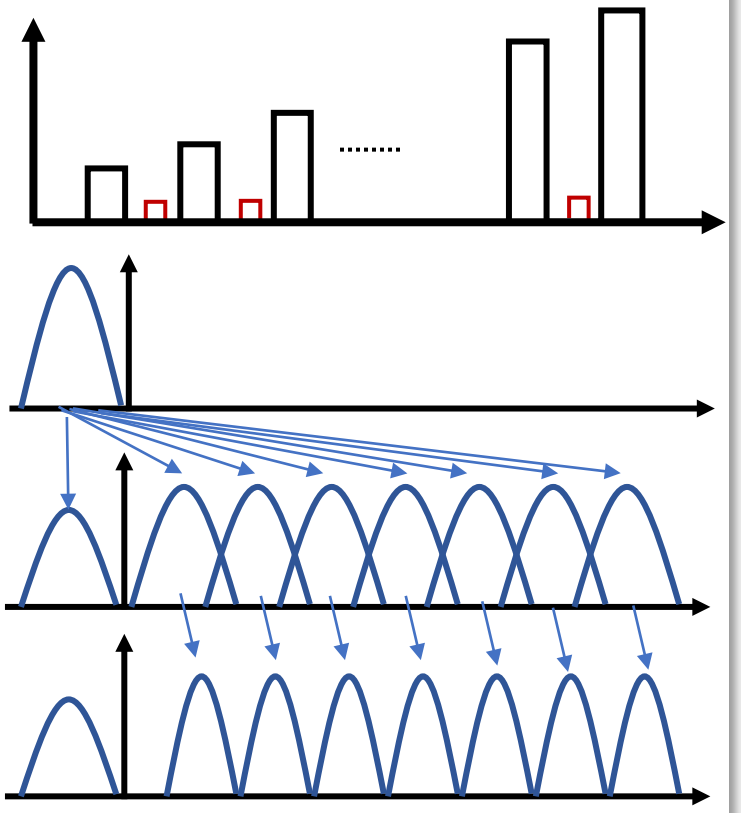
Materials

Flash cells

Memory array



Operation schemes (Para. trimming)



ECC Design

- Stand. LDPC
- Dynamic LDPC
- AI-assisted
- Error-sensitive
- .....

Storage System

Physics-based optimization strategies

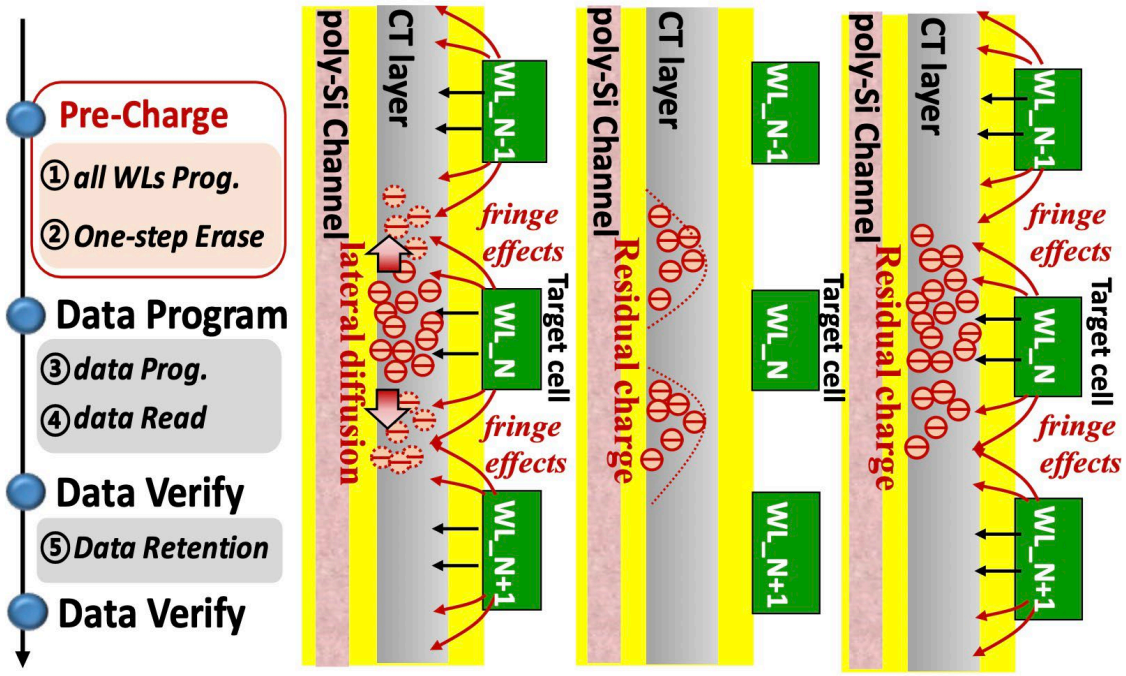
DTCO

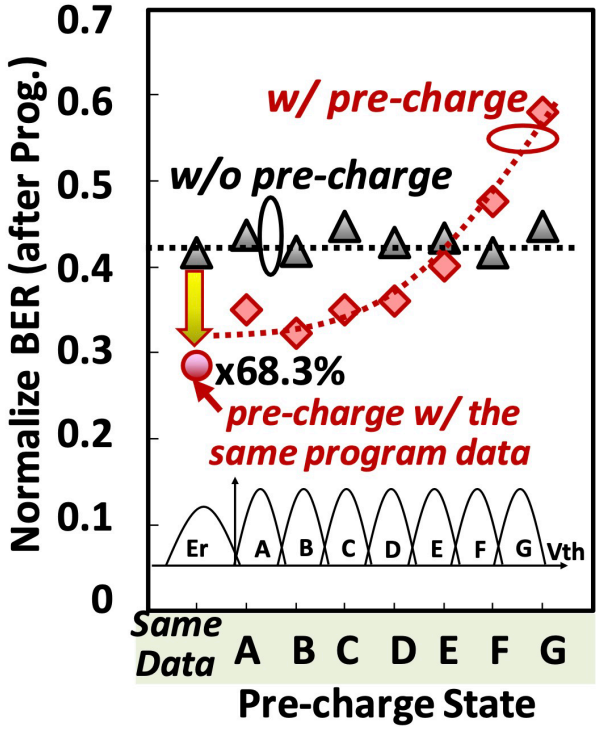
STCO

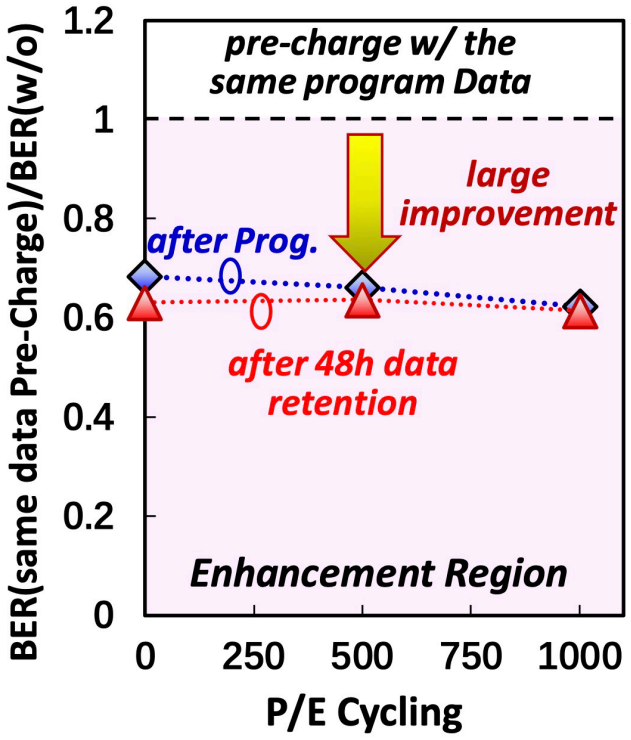
(I) Pre-charging Storage Layer

Flash Memory Summit

- **Pre-charging Storage Layer: >40% DR error bits could be suppressed**



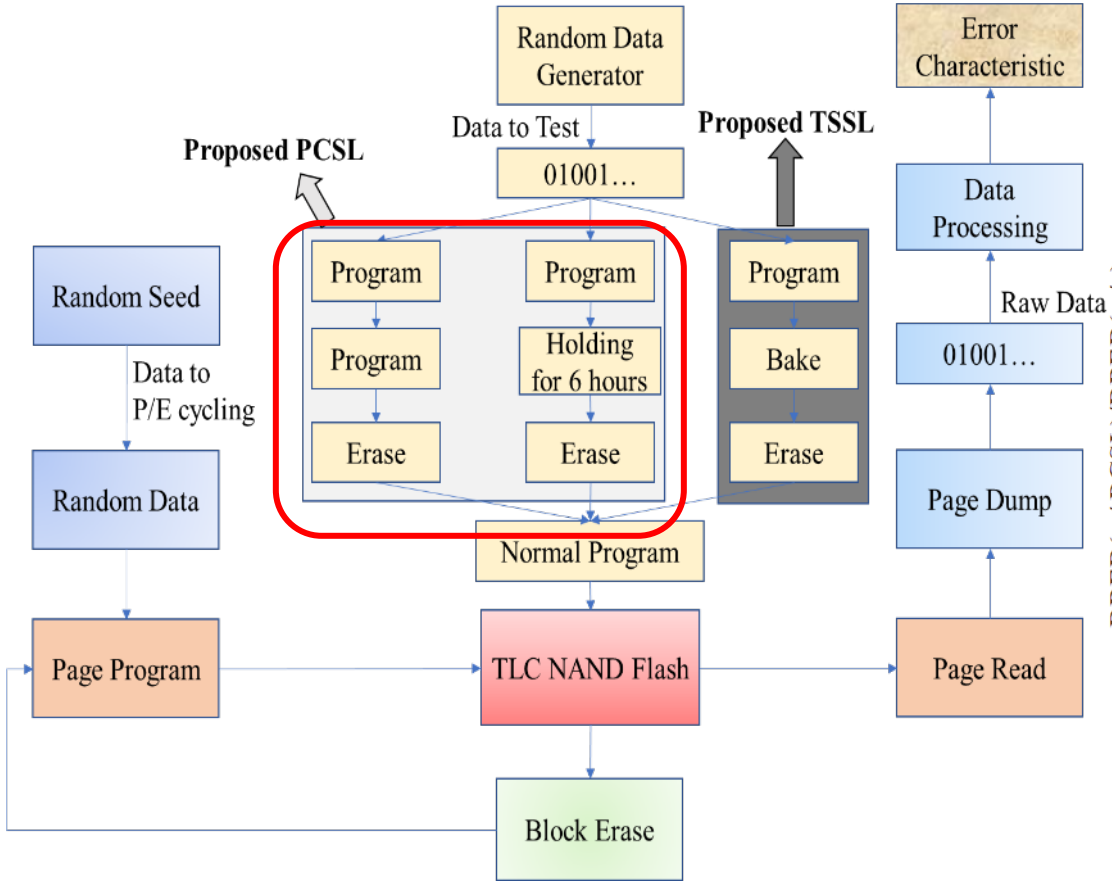


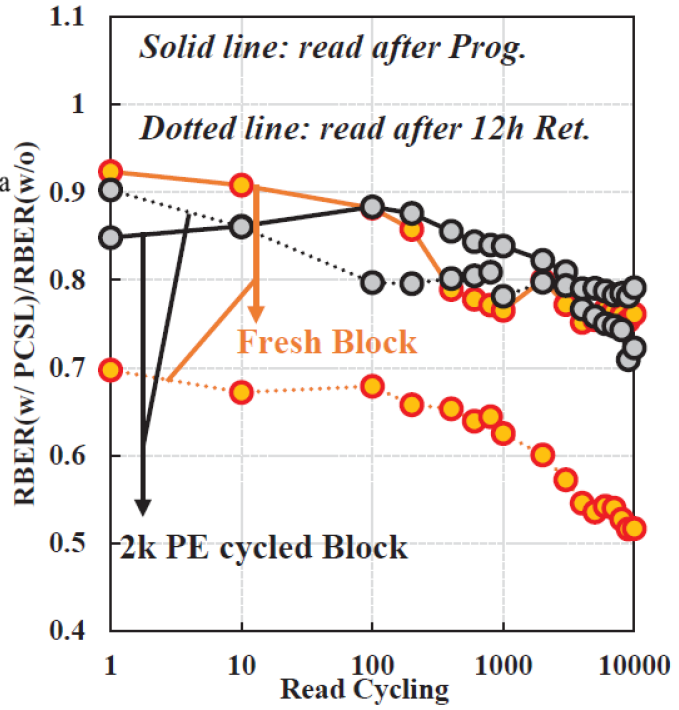


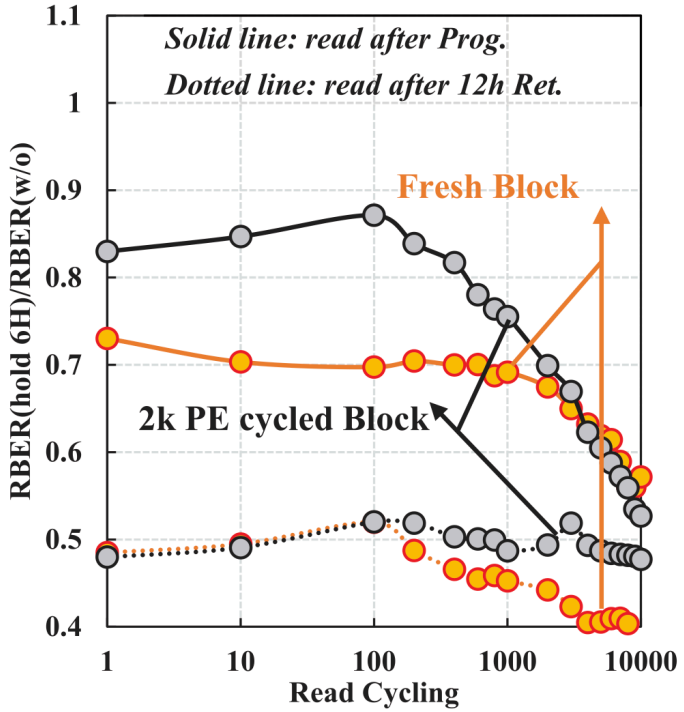
Ref: R. Cao, et al., IRPS 2019;

(I) Pre-charging Storage Layer

- **Double-Program for Pre-charging Storage Layer (PCSL),**
  >50% RD-related error bits could be suppressed



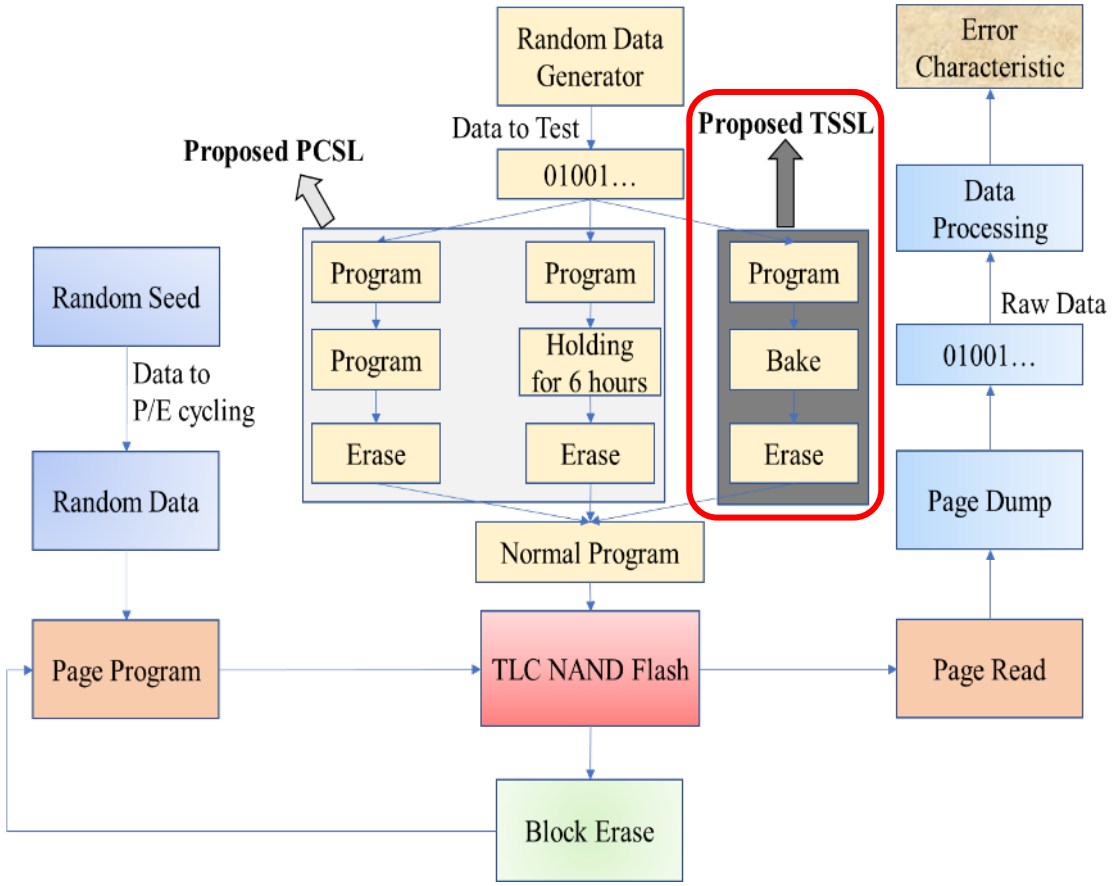


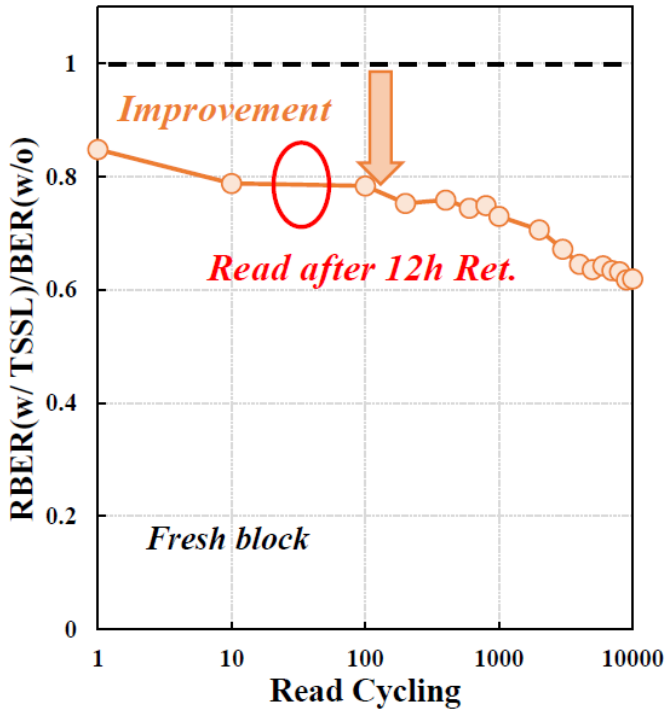


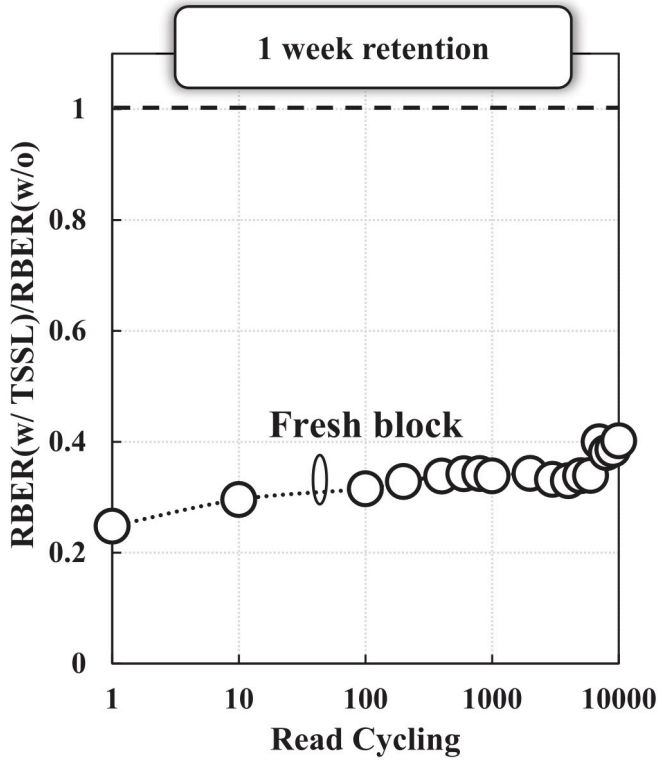
Ref: Y. Kong, et al., IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2020, 39(11): 4042-4051

(II) Thermal-assist Stabilization

- **Thermal-assist stabilization to the Storage Layer (TSSL) can suppress the RD-related error bits effectively after long-term retention.**







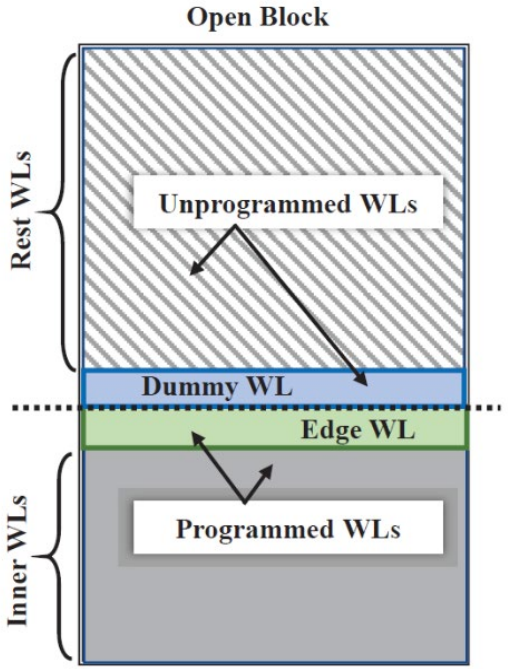
Ref: Y. Kong, et al., IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2020, 39(11): 4042-4051

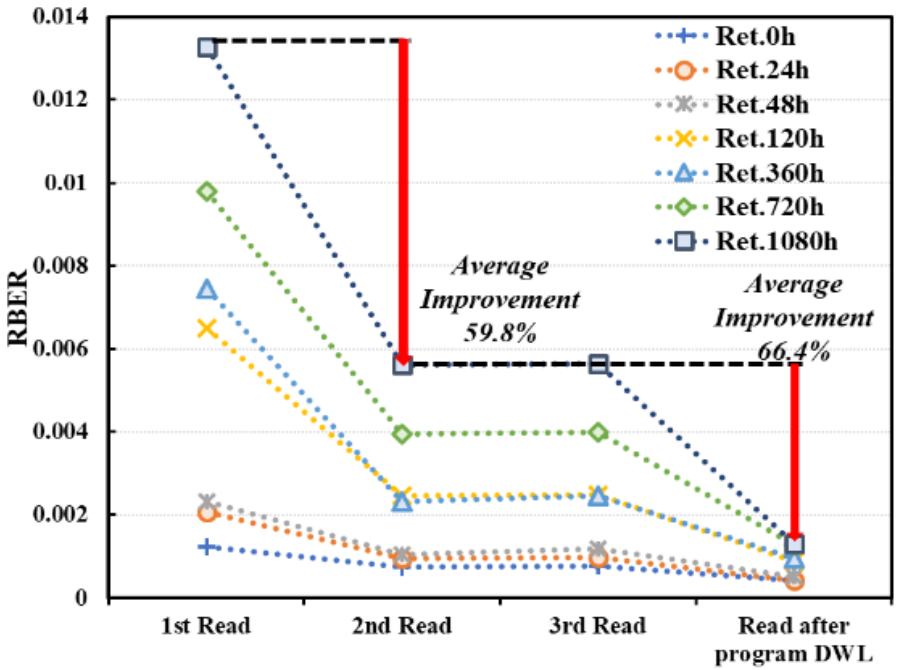
(III) Open-block Mode Operation

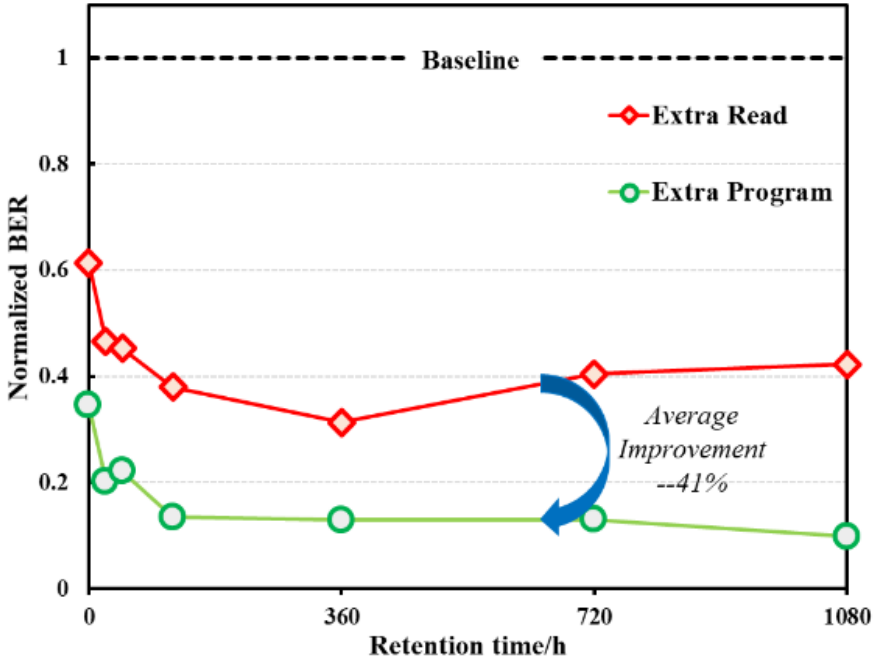
Flash Memory Summit

- 3D NAND Open-block operation: Under non-consecutive writing, both programmed and unprogrammed word lines exist in a block.
- Extra read (ER) and extra program (EP) scheme to compensate the charge loss by LCM.

With ER/EP, the bit error of edge word line is reduced by 59.8% and 86.5%, the error correction ability of LDPC decoding is improved by 1.92 and 4.76 times.



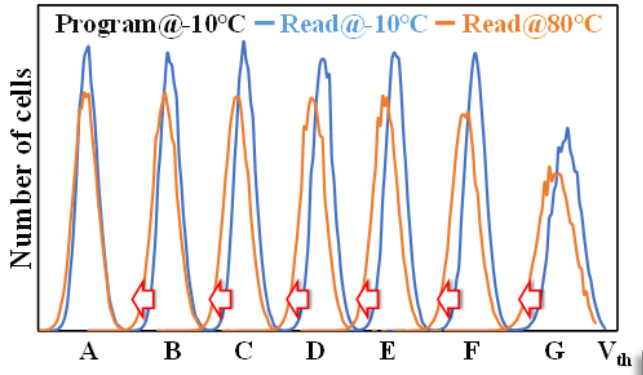


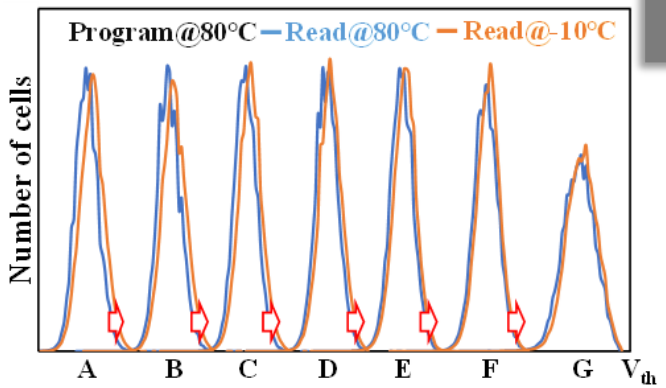


Ref: M. Jia, Y. Kong, et al., IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 2022, 41(11): 4797-4807

(IV) Modeling for Cross-temp Operation

- **Cross-temperature condition: optimal V_{opt} prediction by ΔT_{RP} can reduce the bit error and read-retry delay effectively .**





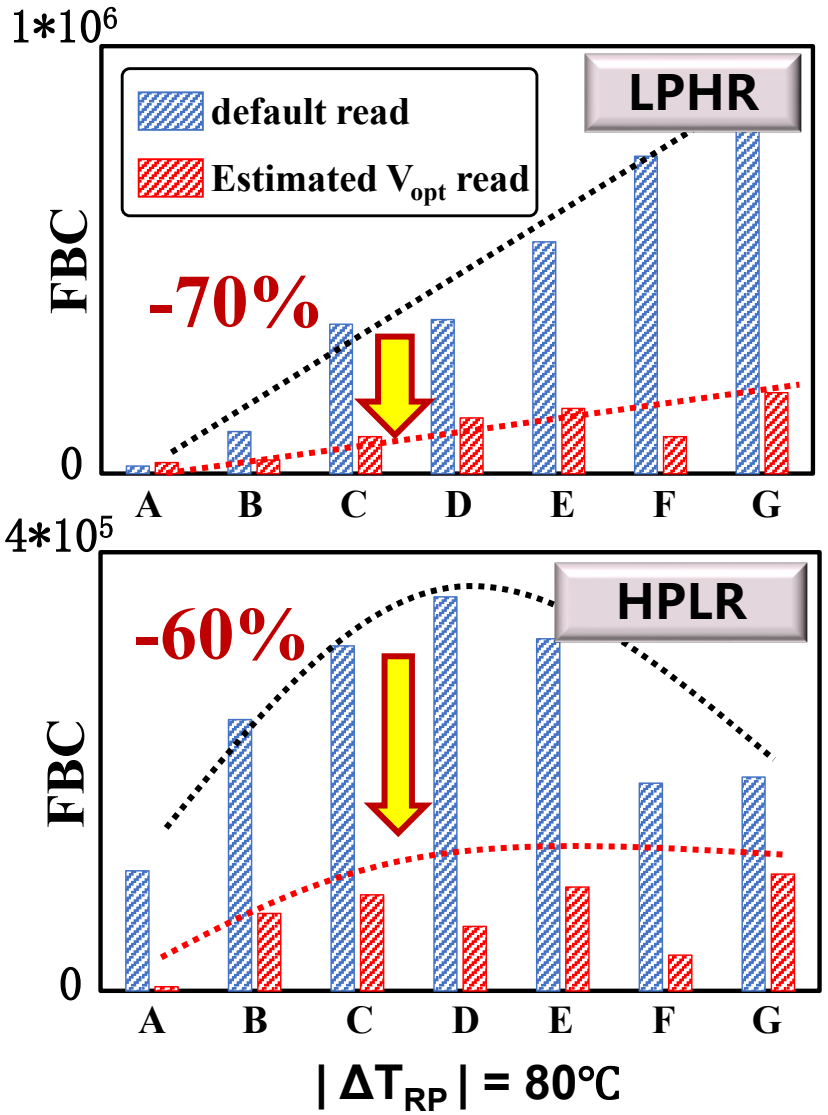
Record T_{prog} and T_{read}

Choose V_{temp} based on P/E cycles and ΔT_{RP}

Estimate V_{opt} by using the formula below

$$V_{opt} = V_{def} + V_{temp} * (T_{read} - T_{prog})$$

V_{temp} is related to temp. and cycling

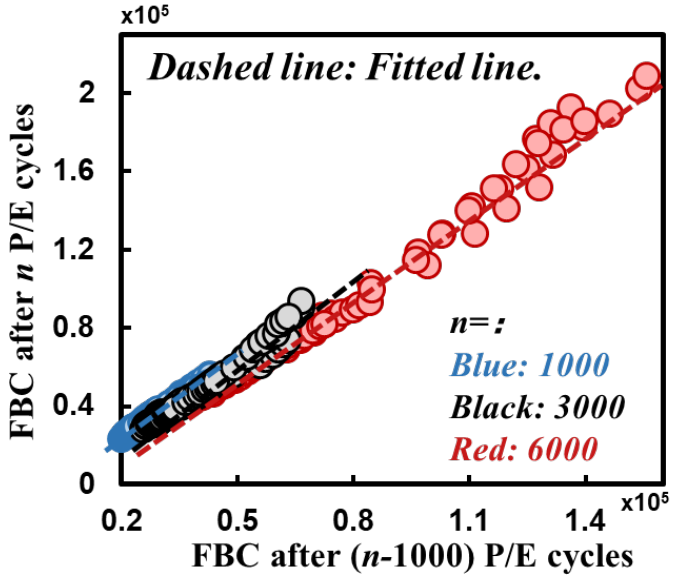


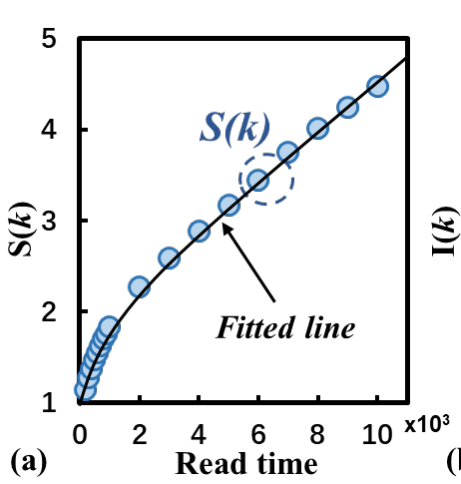
Ref: Y. Guo, et al., IEEE EDTM 2023

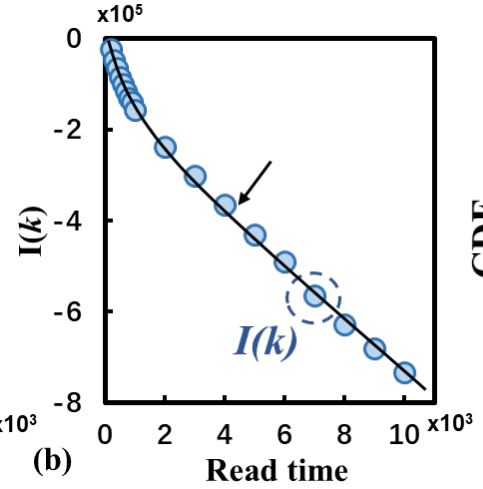
(V) 3D NAND Life Prediction @ Hot Data

Flash Memory Summit

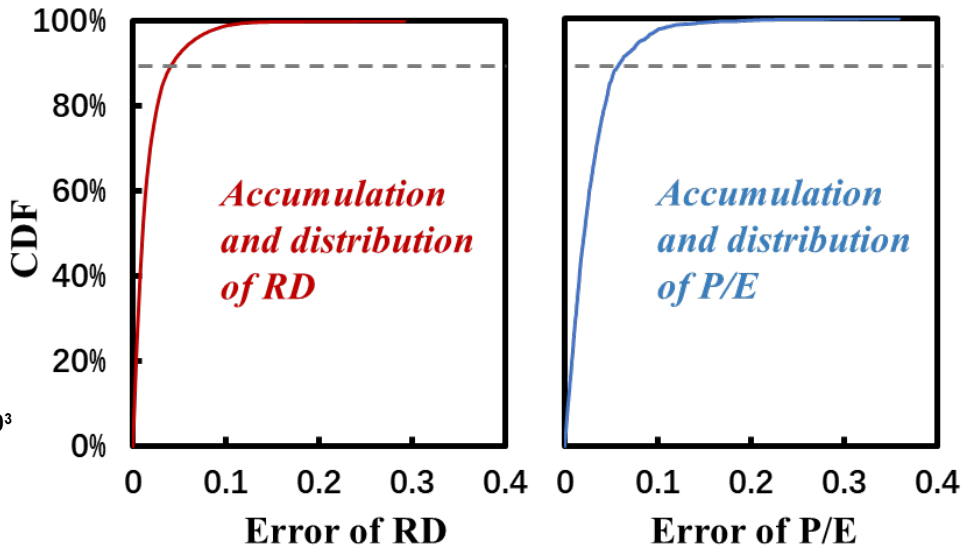
- Relationship between life stages: linear relationship for bit error rate
- A short-term life prediction model for hot data storage is proposed.







$$FBC_k^r = S(k) * FBC_{100}^r + I(k)$$

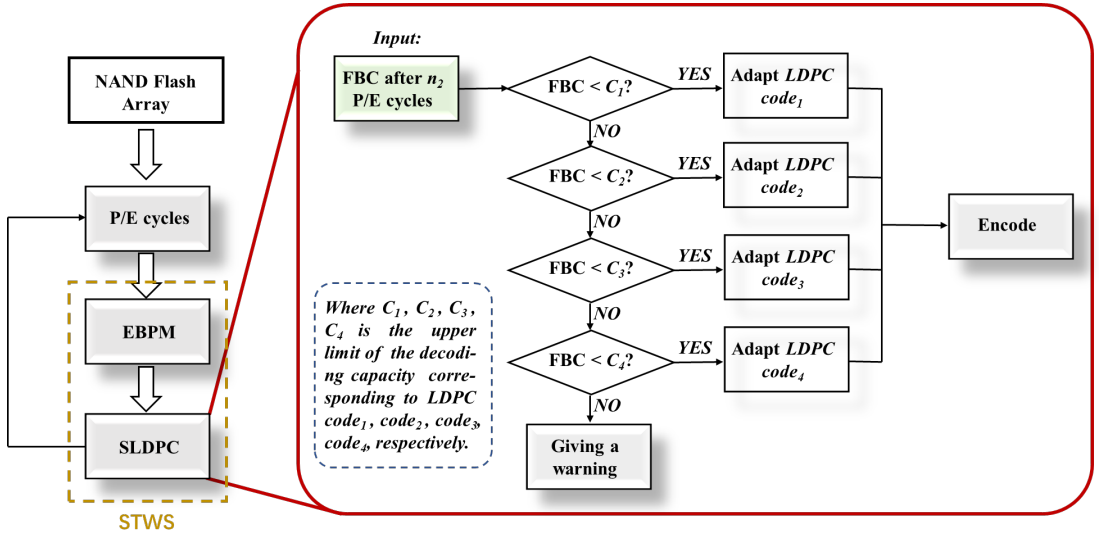


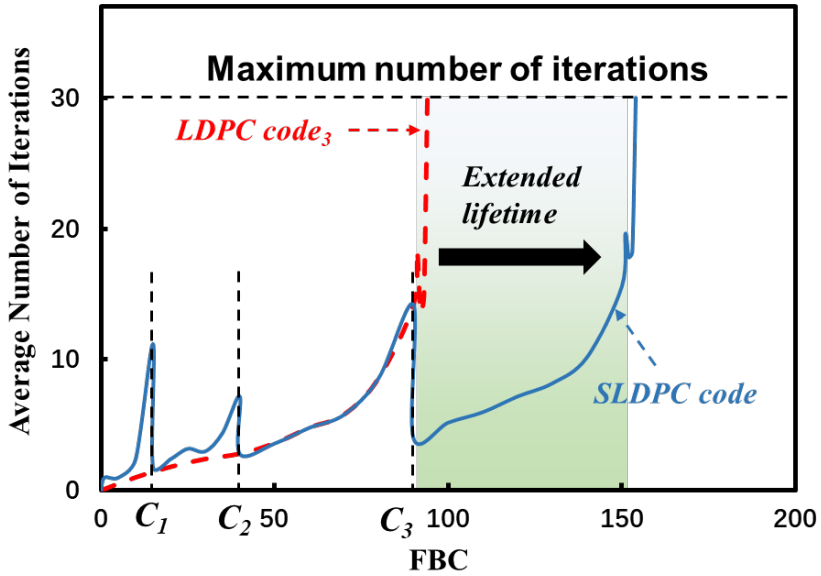
99.16% and 97.63% of the samples show less than 10% accuracy loss of RD and durability prediction models.

Ref: X. Fang, et al., IEEE TCAD, 2023, DOI: 10.1109/TCAD.2023.3240932

(V) 3D NAND Life Prediction @ Hot Data

- **Error bit prediction module (EBPM): Base on the short-term life prediction model, the FBC for the next forecast period is predicted.**
- **Adaptive LDPC code module (SLDPC): periodically adjust the LDPC bit rate of the next stage according to the EBPM forecast results.**





Short-term warning system (STWS) can save space for user data in early life and improve the lifetime of flash memory.

Ref: X. Fang, et al., IEEE TCAD, 2023, DOI: 10.1109/TCAD.2023.3240932

(VI) Data Searching in 3D NAND

Flash Memory Summit

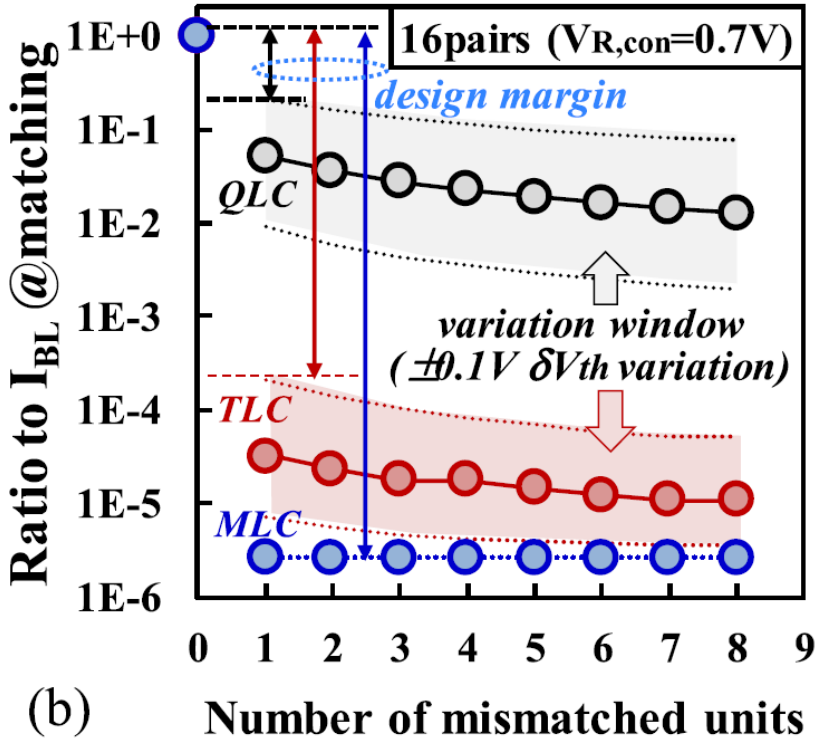
- How to lower the impacts from RD? *Minimize the read times*

~A complementary scheme to realize data search in ultra-densified NAND flash~

Set Conditions for Data Addressing in Pair Unit

$V_{th,d} + V_{th,c} = V_{con}$ $V_{comp} = V_{th,c} + V_{R,con}$ $V_{data} = V_{th,d} + V_{R,con}$

Operation Schemes	$V_{sear,d} = V_{data}$, $V_{sear,c} = V_{comp}$	$V_{sear,d} > V_{data}$, $V_{sear,c} < V_{comp}$	$V_{sear,d} < V_{data}$, $V_{sear,c} > V_{comp}$
Data Cell	Pass	Pass	Cut Off
Comp. Cell	Pass	Cut Off	Pass
Pair Unit	Pass data cell comp. cell	Cut Off data cell comp. cell X	Cut Off data cell comp. cell X



Ref: F. Wang, et al., IEEE EDL, 41(8), 1189-1192, 2020

Outline

Flash Memory Summit

Sources of Error bits

Optimization Strategies

Future Technologies

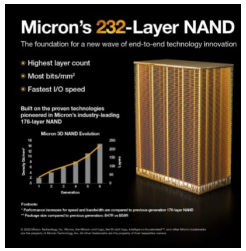
Scaling technologies as memory/storage

Flash Memory Summit

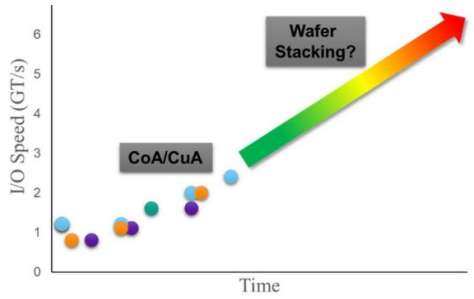
- **Scaling Vectors of 3D NAND**
  - ✓ Lateral Scaling
  - ✓ Architecture Scaling
  - ✓ Vertical Scaling
  - ✓ Logical Scaling
- **Flash Cell Scaling**
  - ✓ Poly-Si channel
  - ✓ Charge-trapping layer
  - ✓ Tunneling layer
  - ✓ Blocking layer

Industry

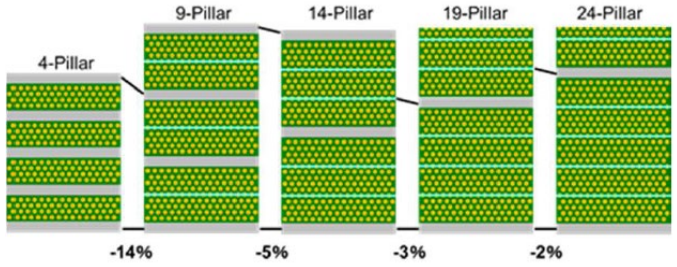
Split-Cell Structure



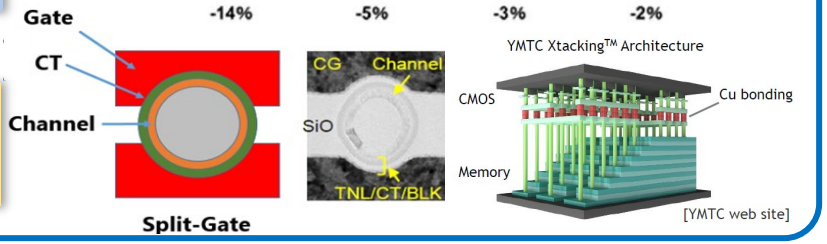
3D Vertical Stack >300 Layer



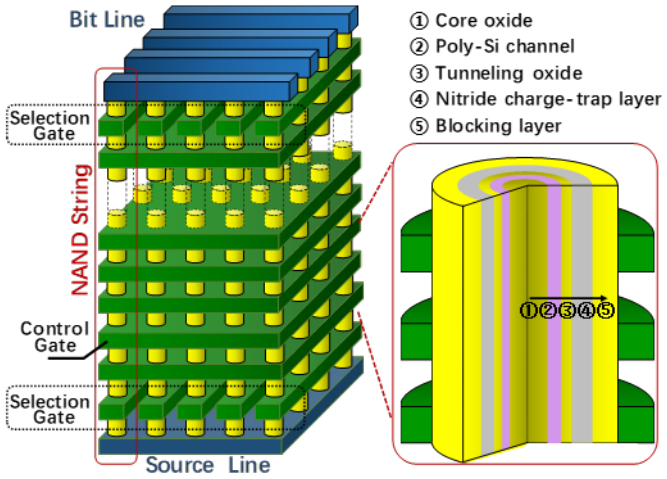
Super bits per cell 7bit/cell



3D Vertical Stack Xtacking



Academic

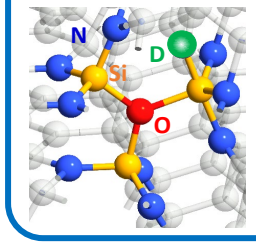


LTPS Process

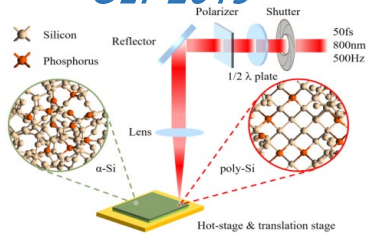
Ultra-thin Si-channel

Novel materials and structures
- Tunneling Layer
- Storage Layer
- Blocking Layer

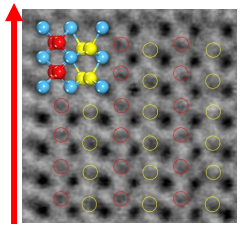
IEDM 2017



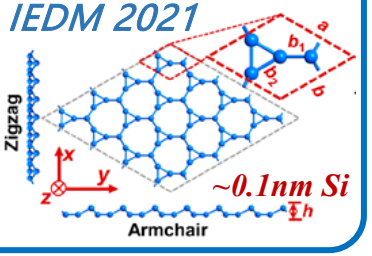
OLT 2019



APL 2019



ACS Mat. Lett 2021
IEDM 2021



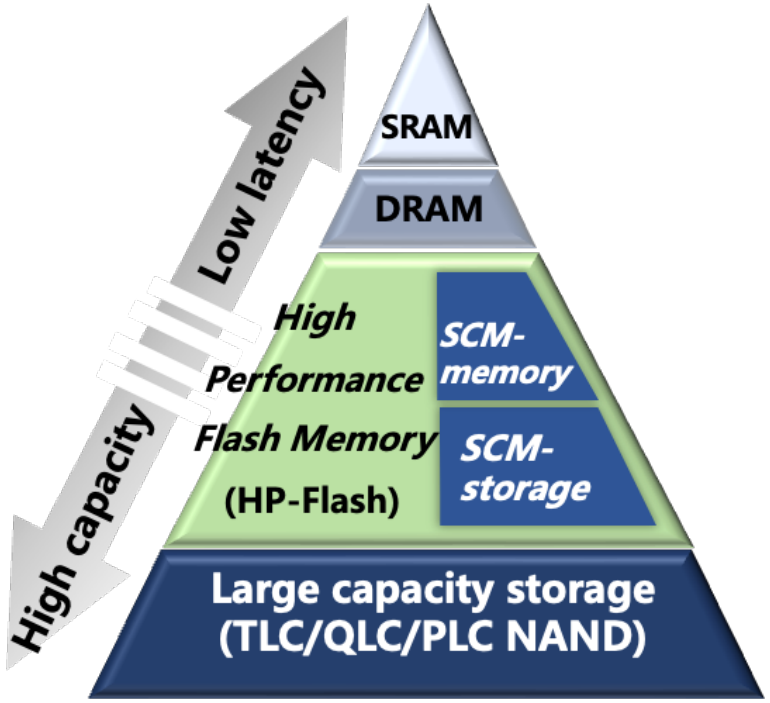
Refs: Reports from IEDM/VLSI/IMW/FMS conferences; open sources in websites

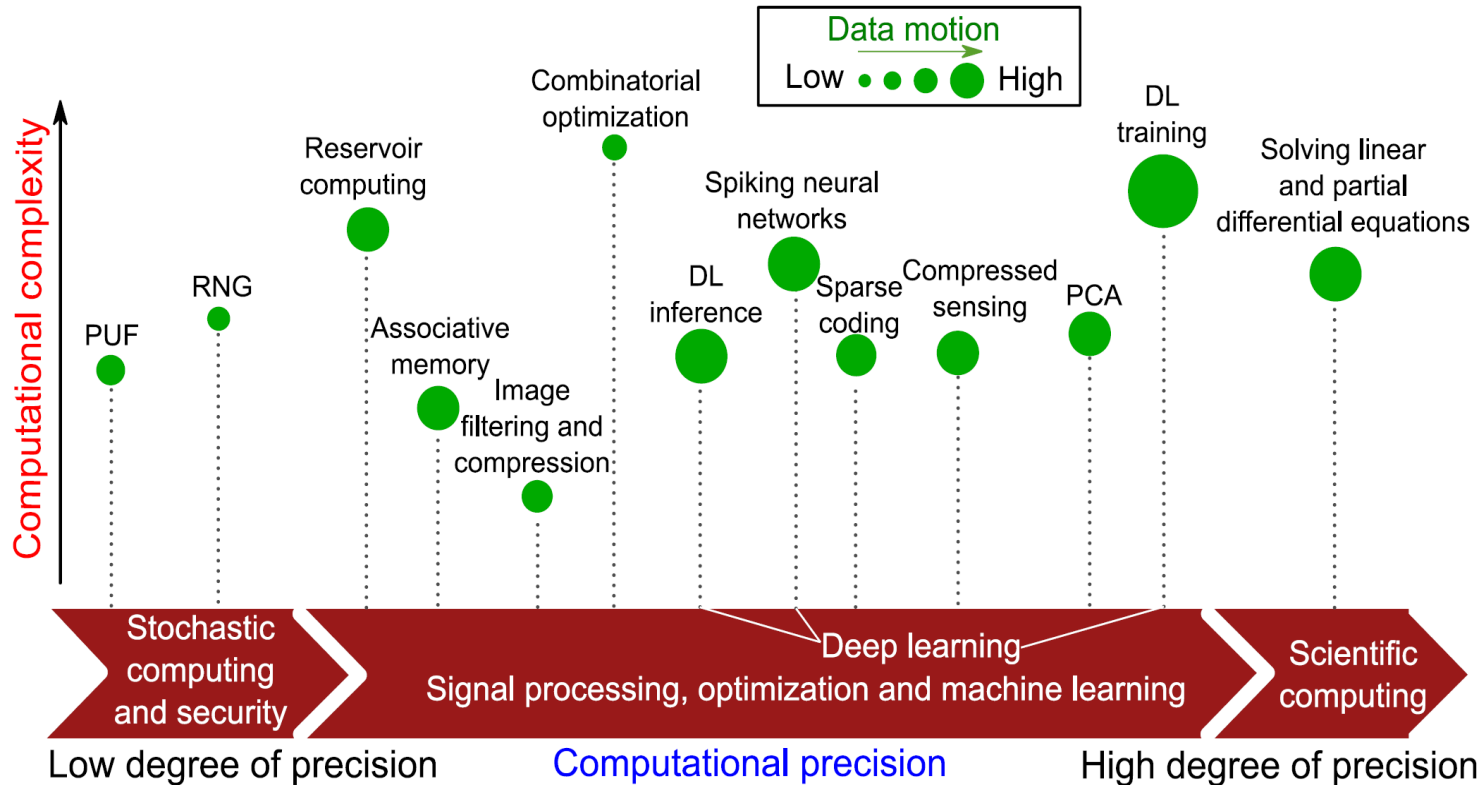
Applications @ Big Data

Flash Memory Summit

For robust reliability design: Simple is Better

Constructing Fully Flash computing systems ?





Refs: A. Sebastian, IEDM' 2019, Tutorial; A. Sebastian et al., Nature Nanotechnology 15, July, 2020:529-544

Applications @ Big Data

- ✓ **Ultra-big matrix with high bit density per cell**

  TLC(3bit/cell), QLC(4bit/cell), PLC(5bit/cell)

- ✓ **Mature integration technology**

  Floating Gate, Charge-trap, 2D, 3D

- ✓ **Highly compatible peri. circuits**

  NAND, NOR, Stand-alone, Embedded

- ✓ **Great Design flexibility**

Power efficiency?

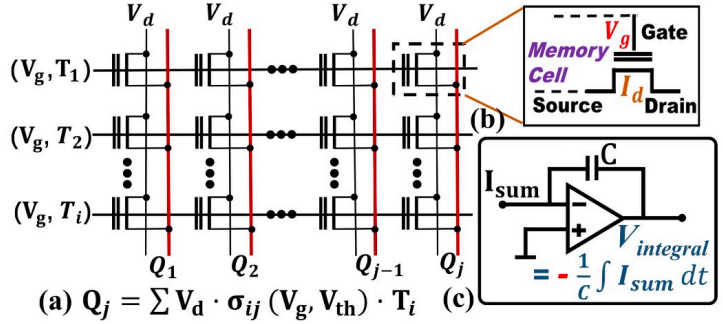
Limited lifetime?

Slow speed?

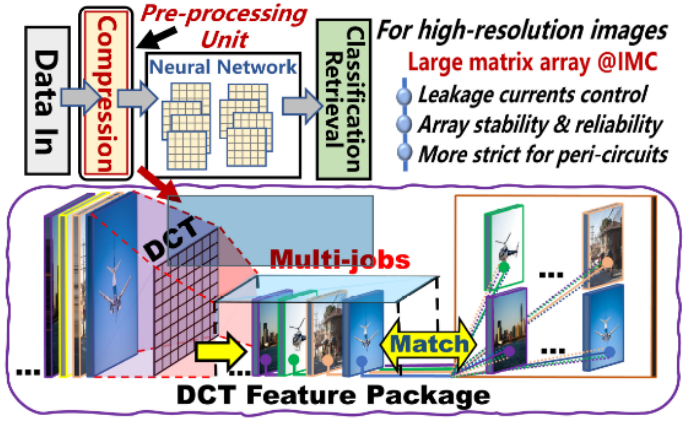
Flash-based CIM Designs (examples)

High Bit-density (~QLC), Large Flash array, High speed (<100ns) *

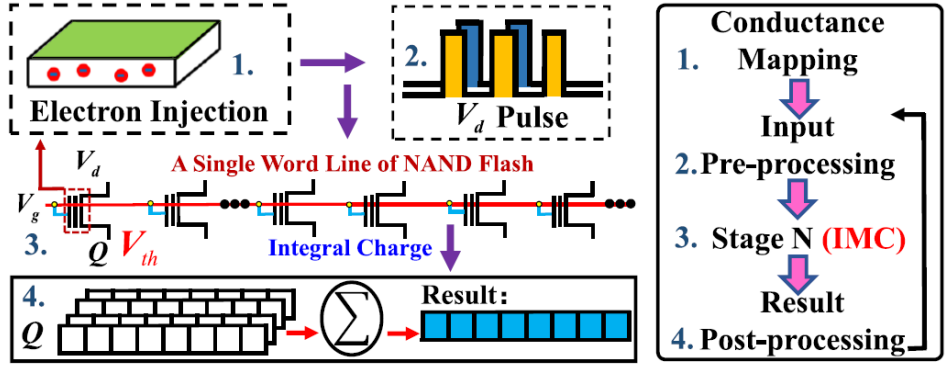
- ✓ NOR-Flash DCT CIM Design for Data Compression
- ✓ NAND-Flash FFT CIM Design for Signal Processing
- ✓ RTN-encrypted Flash-CIM Architecture Design



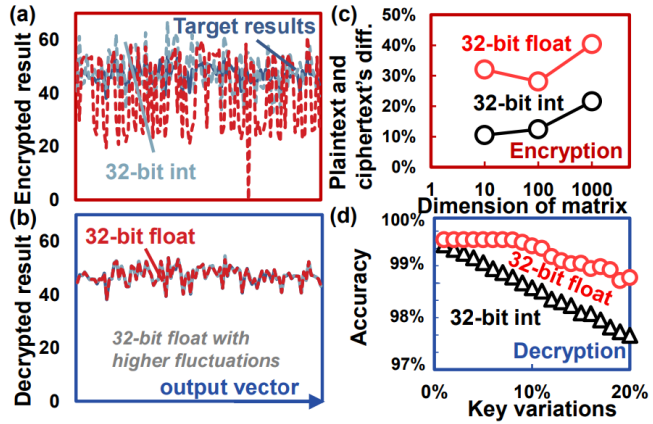
D. Zhang, et al., EDL, 42(11), 1603, 2021



D. Zhang, et al., EDL, 43(8), 1207, 2022



Y. Feng, et al., IEEE EDL, 43(9), 1455, 2022



* Ref: Y. Feng, et al., IEDM 2021, S12-1; Y. Feng, et al., IEEE TED, 70(2), pp. 461-467, 2023

Summary

- **3D NAND Reliability Issues?**
  - 3D NAND structure/process are different from 2D NAND, traditional models cannot be simply used in 3D NAND.
- **Reliability optimization strategies?**
  - Fundamental approach: Process & structure innovations
  - Efficient approach: Optimizations based on materials/cell/array properties
- **Diversified applications of flash memory**
  - Flash is still the most promising memory/storage, and can be used to process complex computation tasks with high-speed Flash and Flash-CIM architectures

Thank You !

Q & A

Email: chen.jiezhi@sdu.edu.cn