How the Common Retention Acceleration Method of 3D NAND Flash Memory Goes Wrong?

Qiao Li, Min Ye, Tei-Wei Kuo, Chun Jason Xue
City University of Hong Kong
National Taiwan University

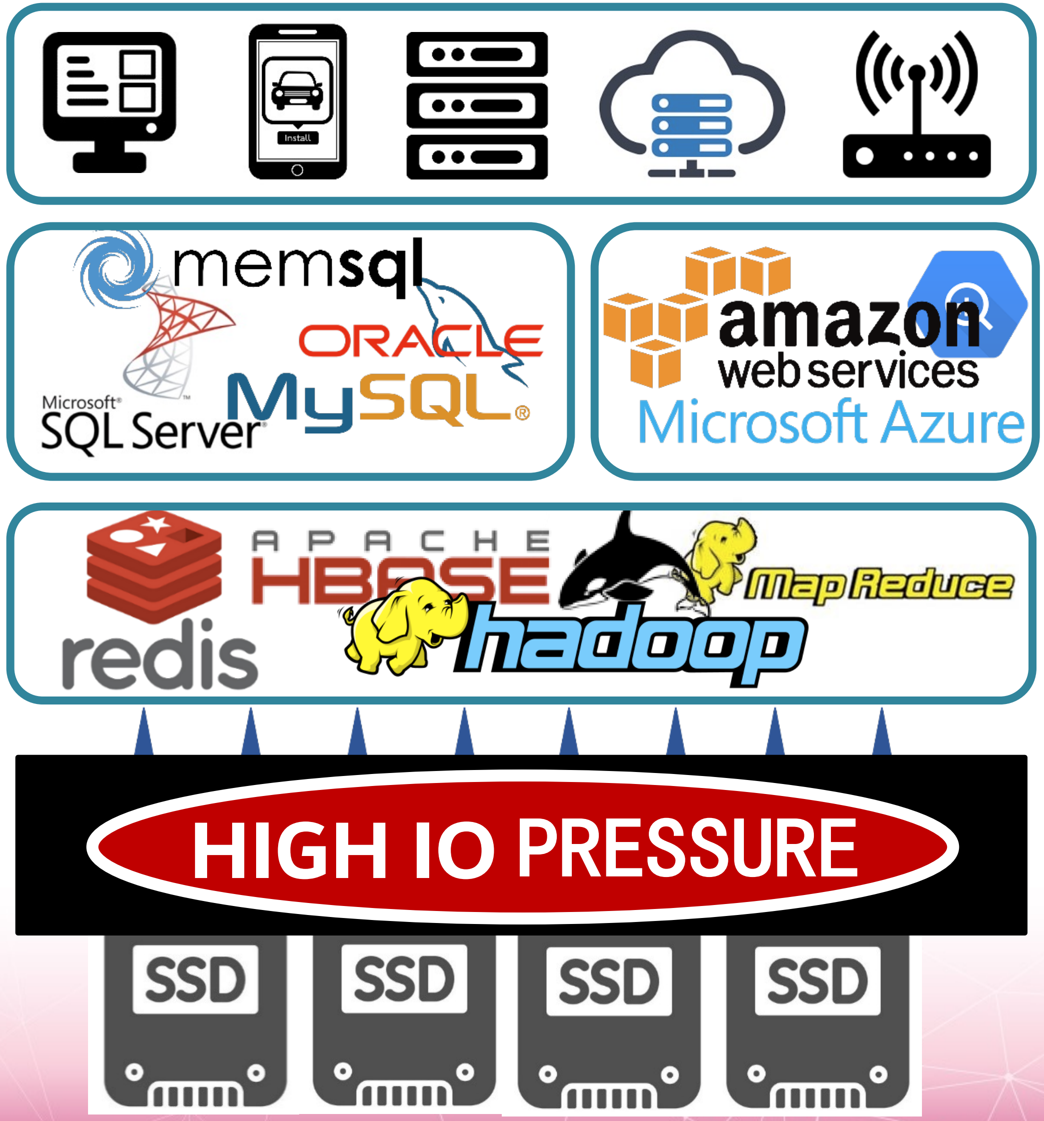
HotStorage 2021
The 13th ACM Workshop on Hot Topics In Storage and File Systems
July 27-28
Virtual

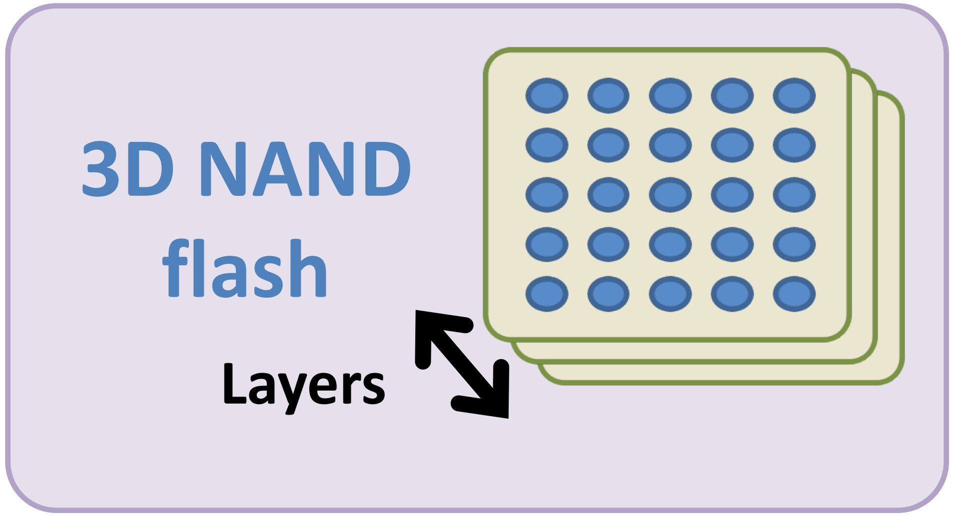
CityU

Outline

- Flash Retention Acceleration
- How 3D Flash Retention Acceleration Goes Wrong?
  - Evaluation Setup
  - New Findings
- Real Long-Retention Errors and Discussions
- Conclusions

3D NAND Flash

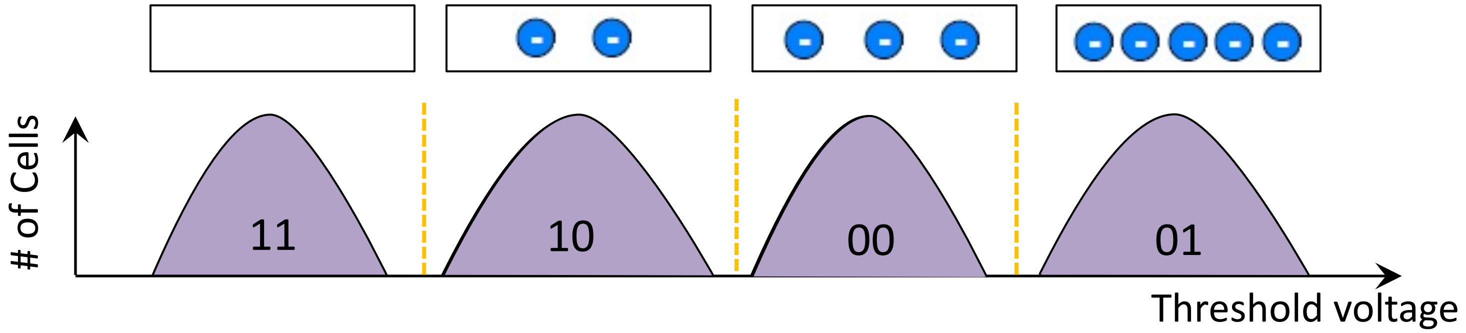




- 3D NAND flash is prevailing
- More vertical layers for higher density

Flash Retention Acceleration

- Retention error is the main error source in NAND flash
  - Caused by charge leakage over retention time
- High temperatures accelerate retention effects
  - High temperatures increase the mobility of electron



Flash Retention Acceleration

- Arrhenius' law
  - Baking flash chips at a high temperature can achieve the equivalent retention effect to that of a low temperature

Retention time at T_1

$$AF = \frac{t_1}{t_2} = exp(\frac{E_a}{k} * (\frac{1}{T_1} - \frac{1}{T_2}))$$

Retention time at T_2

A high temperature

A low temperature

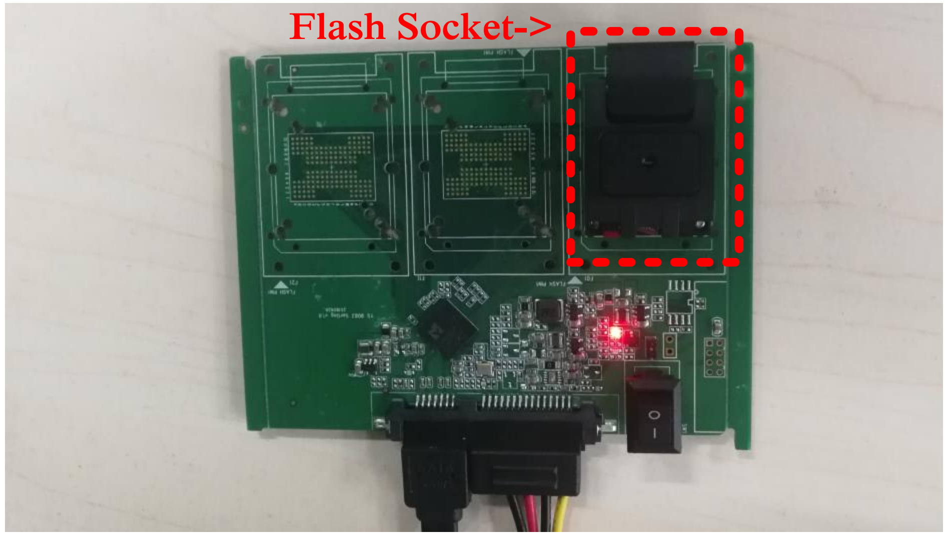
Baking hours to obtain equivalent retention effects of 25℃ for 1 year

High temperature (℃)	Baking hours
60	97.65
80	11.16
100	1.61
120	0.28

CityU

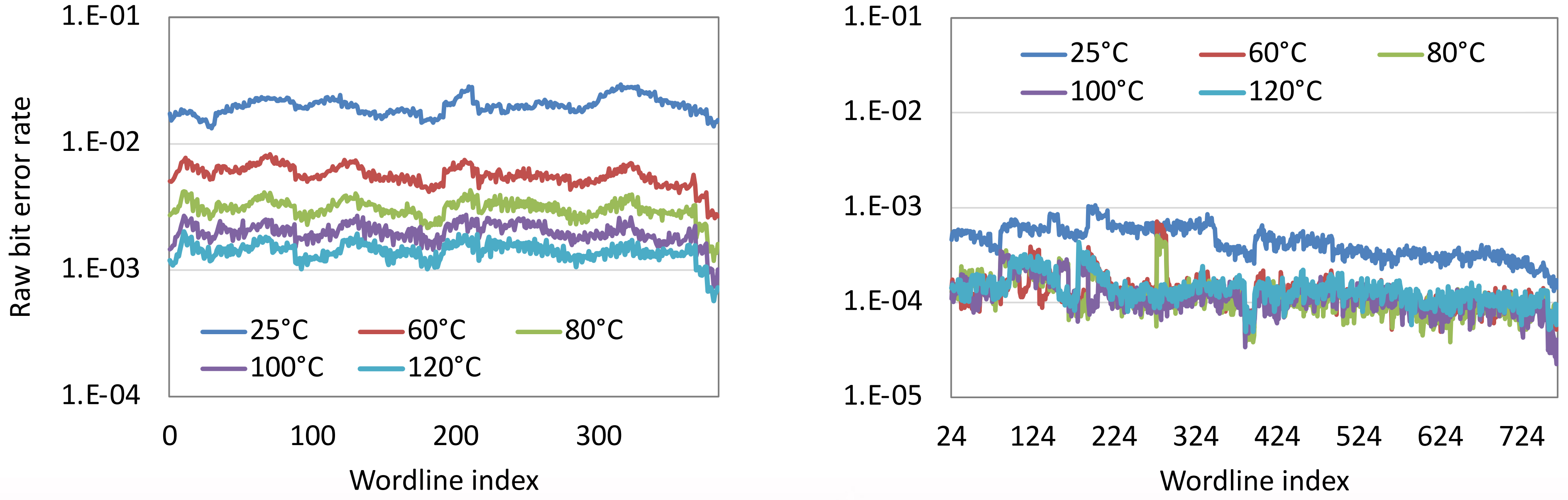
Evaluation Setup

- YEESTOR 9083 flash testing platform
- Four 3D flash chips
  - 3 TLC and 1 QLC
  - 2 FG and 2 CT
- Room temperature for one year
  - Air condition at 25°C
- High temperature baking



New Finding 1

- The errors through baking with high temperatures significantly **underestimate** the real long-retention errors.
  - 3D NAND flash memories have worse retention characteristics.

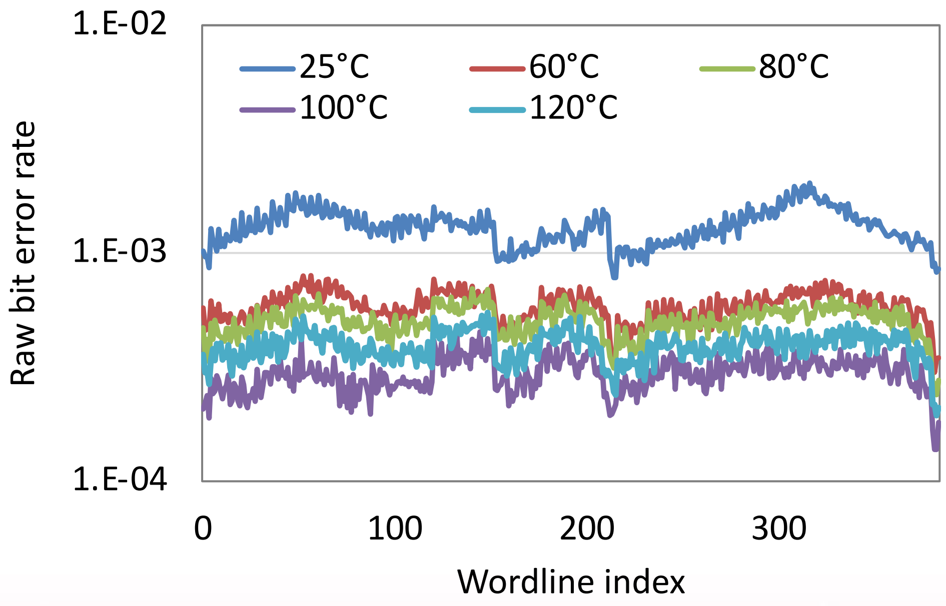


Charge trap flash

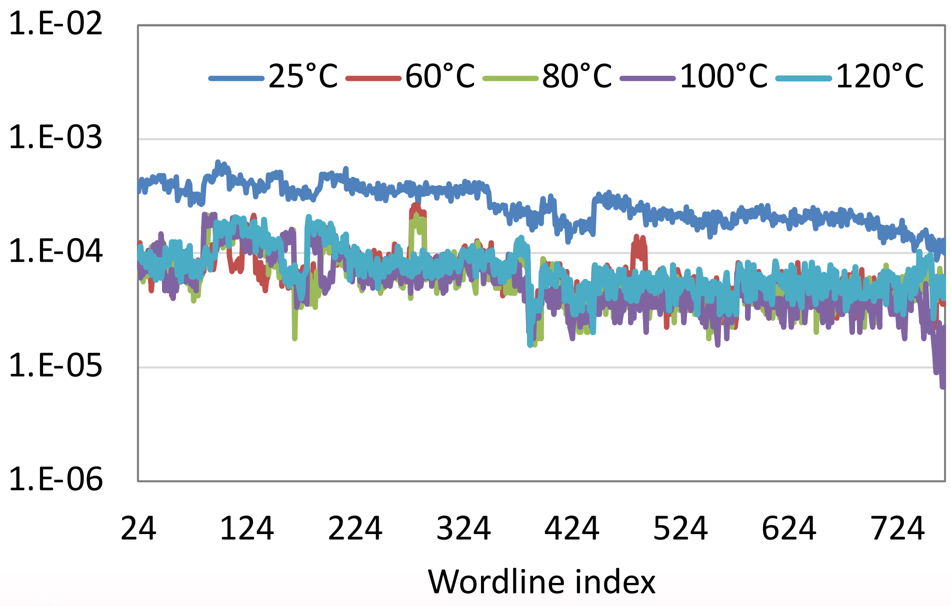
Floating gate flash

New Finding 2

- The difference between the real and estimated retention errors on 3D FG flash is smaller than that of 3D CT flash.
  - CT flash memories have different characteristics from FG flash memories.



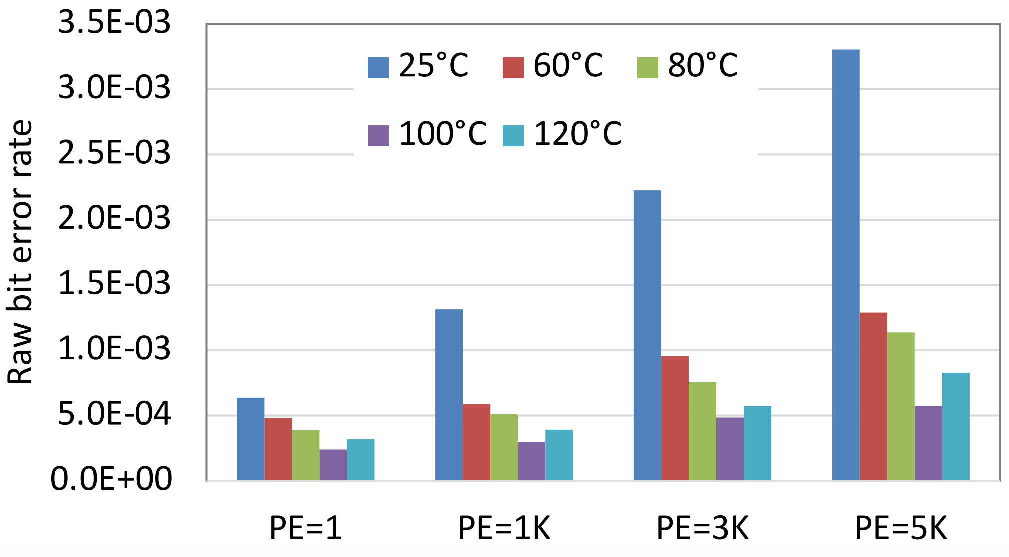
Charge trap flash



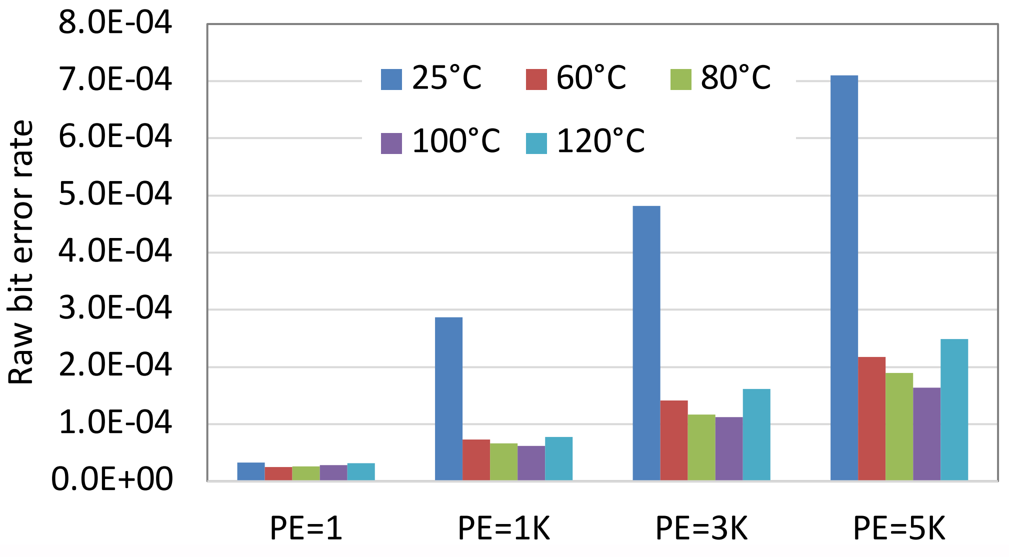
Floating gate flash

New Finding 3

- The difference between the real and estimated retention errors changes with the number of P/E cycles.
  - P/E operations introduce oxide degradation and cell damage.



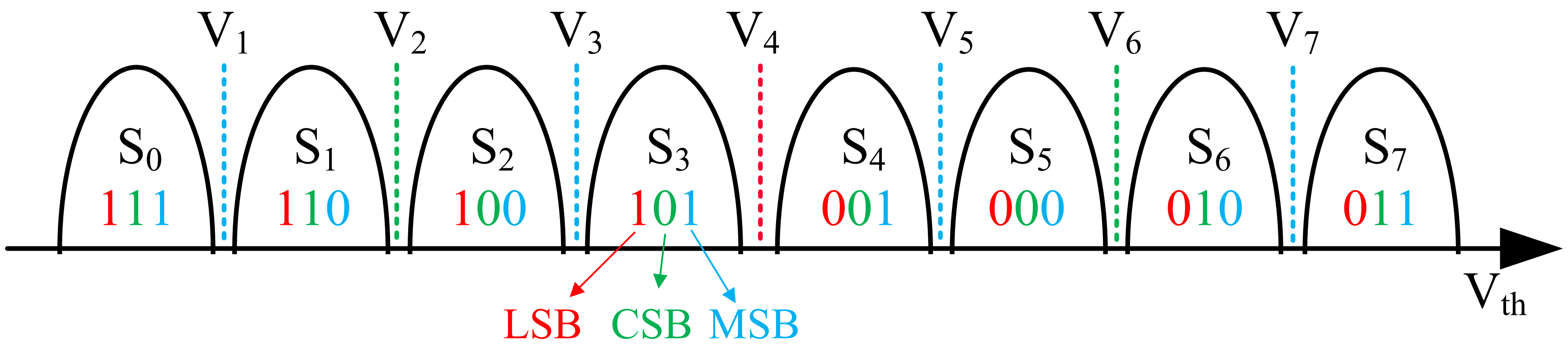
Charge trap flash



Floating gate flash

Read Voltages and Pages of TLC flash

- Each TLC cell stores three bits
- 7 read reference voltages to differentiate 8 states

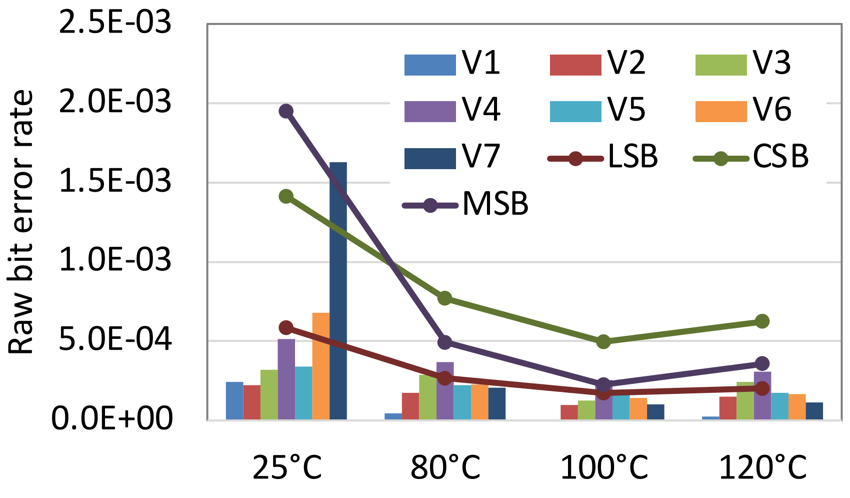


LSB	1	1	1	1	0	0	0	0
CSB	1	1	0	0	0	0	1	1
MSB	1	0	0	1	1	0	0	1

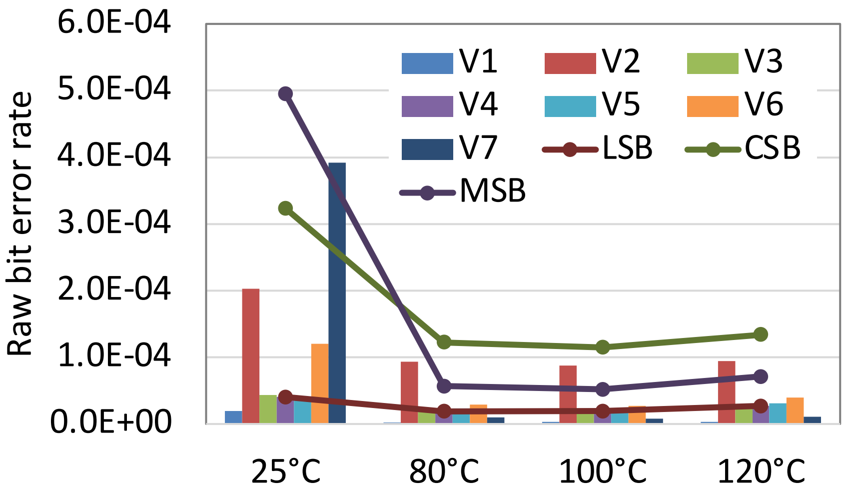
V1 V2 V3 V4 V5 V6 V7

New Finding 4 - TLC

- The temperature affects the RBER variations among read voltages, as well as pages.
  - Different voltage states suffer from retention errors in different degrees.



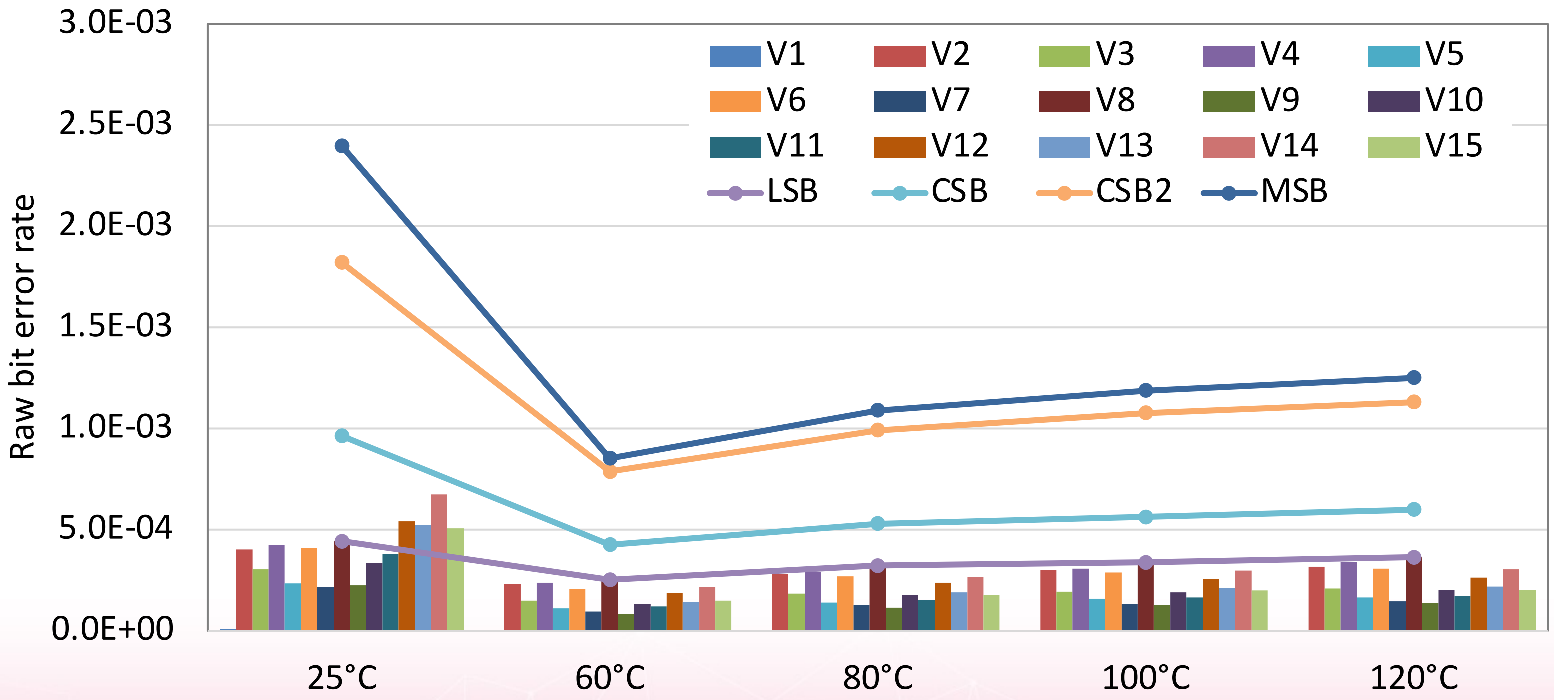
Charge trap flash



Floating gate flash

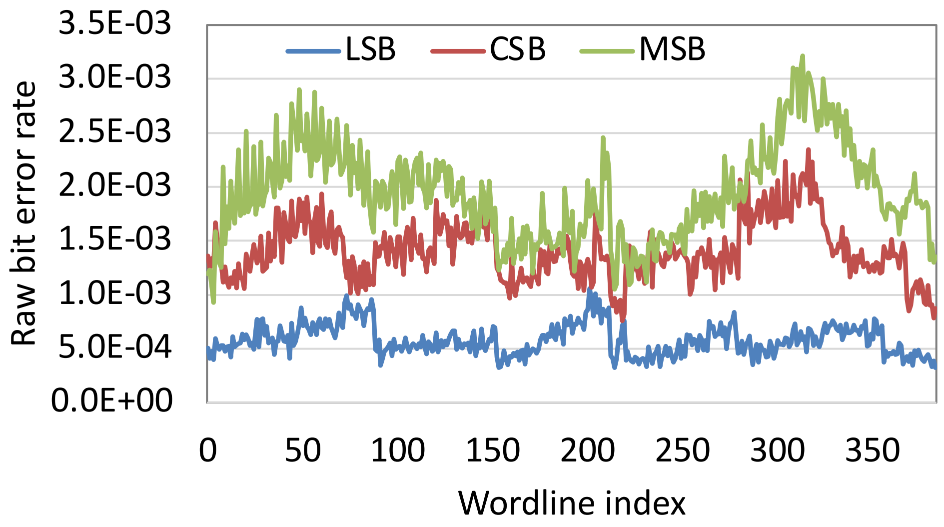
New Finding 4 - QLC

- The temperature affects the RBER variations among read voltages, as well as pages.
  - Different voltage states suffer from retention errors in different degrees.

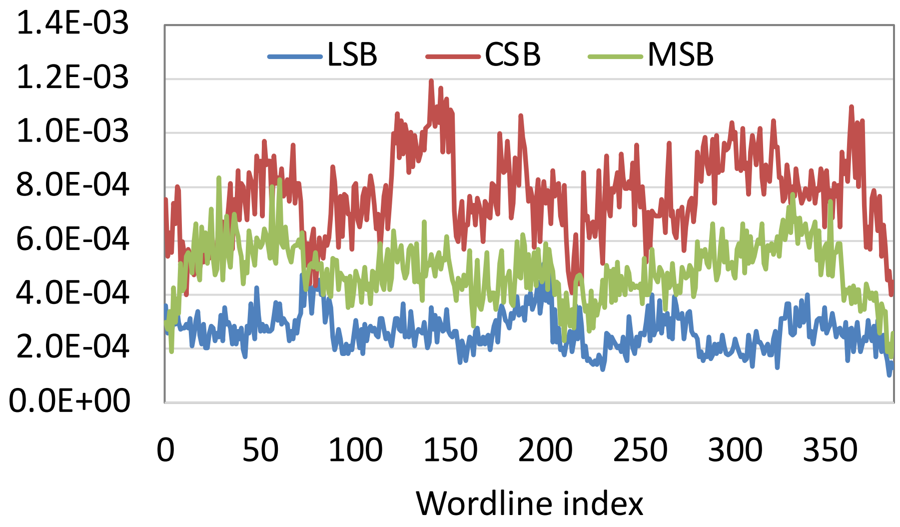


New Finding 5

- The temperature affects the RBER variations among wordlines.
  - Variations among layers/wordlines come form the physical structure of 3D flash blocks.



1 year at 25°C



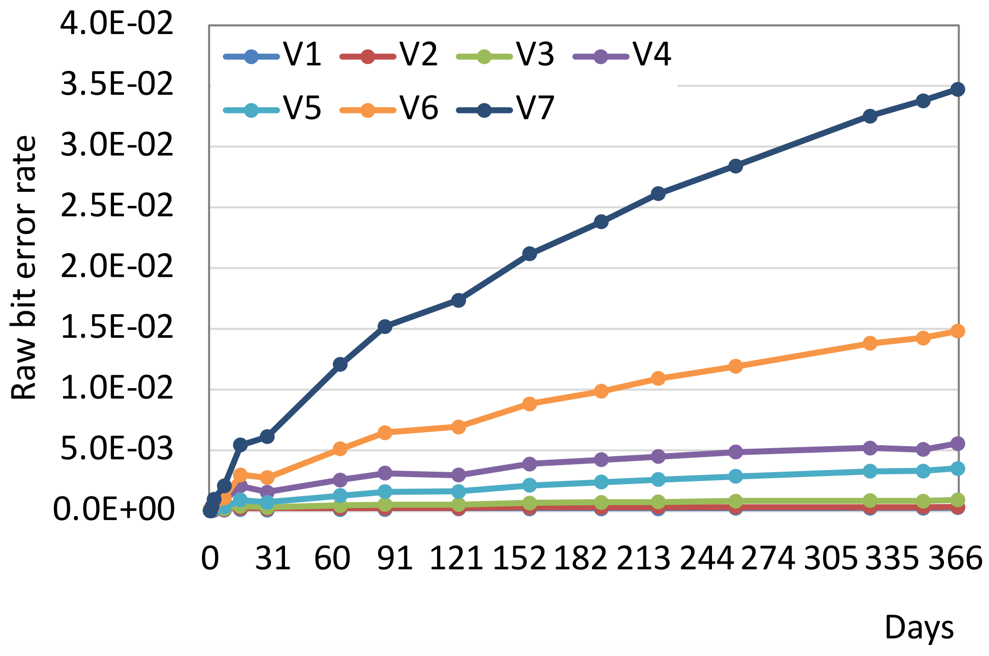
11.2 hours at 80°C

Conclusions of New Findings

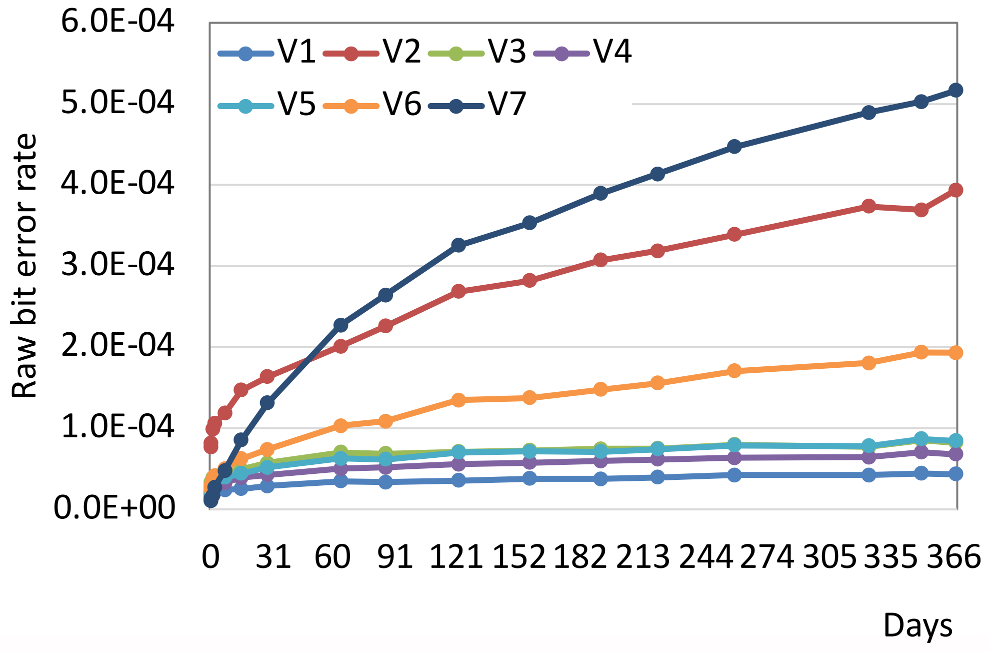
- The error behaviors at high temperatures are different from that of the room temperature
  - Underestimation
  - Flash type (floating-gate and charge-trap)
  - Number of P/E
  - Variations among pages
  - Variations among wordlines/layers

Real Long-Retention Errors

- The error rates introduced by higher voltage states are higher.
- CT flash has much higher error rates.



Charge trap flash



Floating gate flash

Discussions

- Characterization of retention errors
  - New retention characterization method is required for 3D flash.
- Influence on the reliability
  - Inaccurate retention characterization may introduce data loss.
- Optimizations of the reliability
  - Revisions are required for current reliability optimization methods.

Conclusions

- Retention acceleration through baking is widely applied.
- A one-year retention evaluations on real 3D NAND flash chips.
- The real error behaviors at low temperatures are different from that through baking with high temperatures.
- We provide several insights to explore the long retention errors, along with real long-retention errors on 3D flash memory.

How the Common Retention Acceleration Method of 3D NAND Flash Memory Goes Wrong?

Qiao Li, Min Ye, Tei-Wei Kuo, Chun Jason Xue

HotStorage 2021

The 13th ACM Workshop on Hot Topics In Storage and File Systems

July 27-28 Virtual

Thank you!