HeatWatch

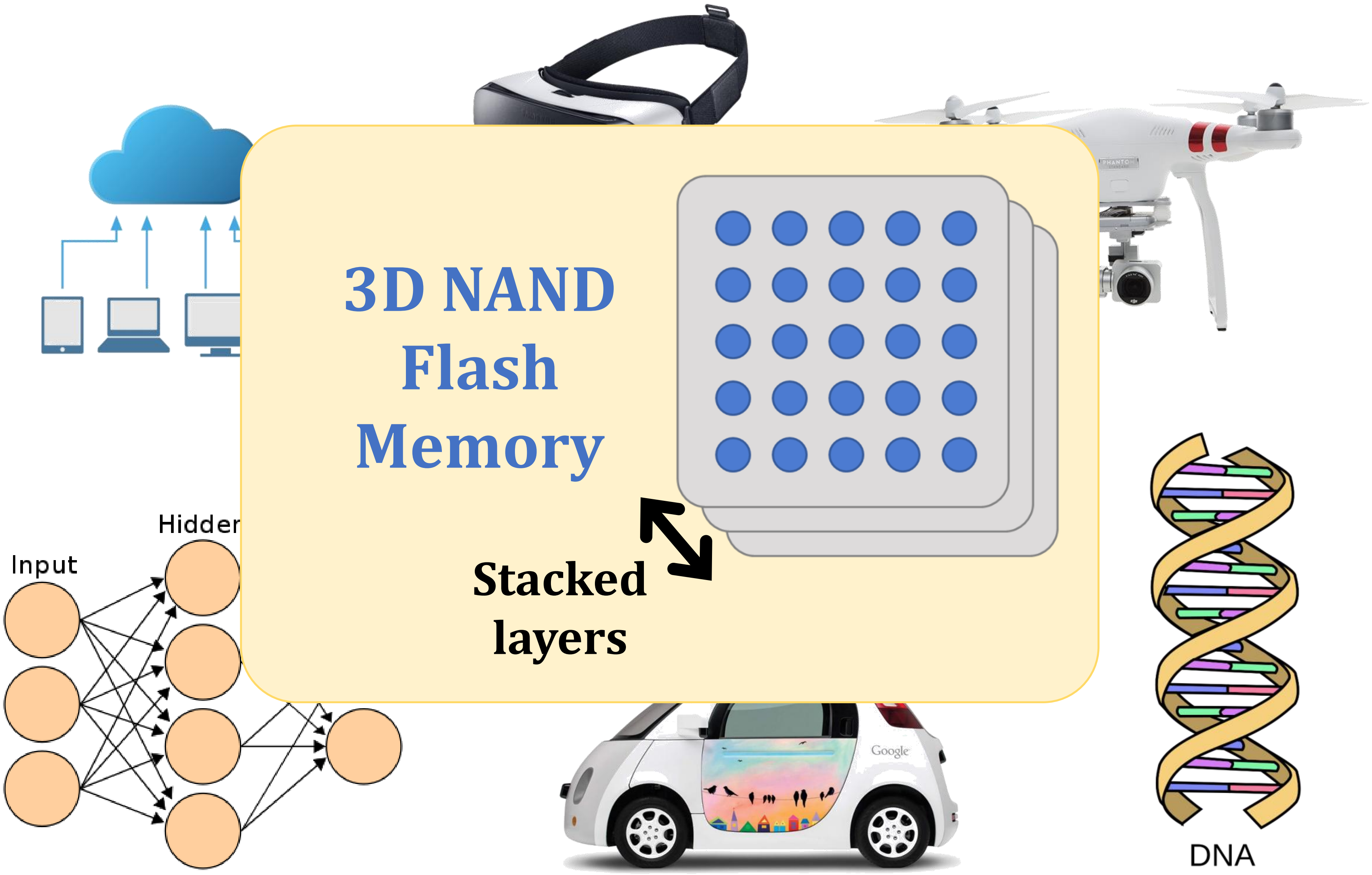
Improving 3D NAND Flash Memory Device Reliability by Exploiting Self-Recovery and Temperature Awareness

Yixin Luo **Saugata Ghose** **Yu Cai** **Erich F. Haratsch** **Onur Mutlu**

Carnegie Mellon SK hynix ETH Zürich

SAFARI SEAGATE

Storage Technology Drivers - 2018



3D NAND Flash Memory

Stacked layers

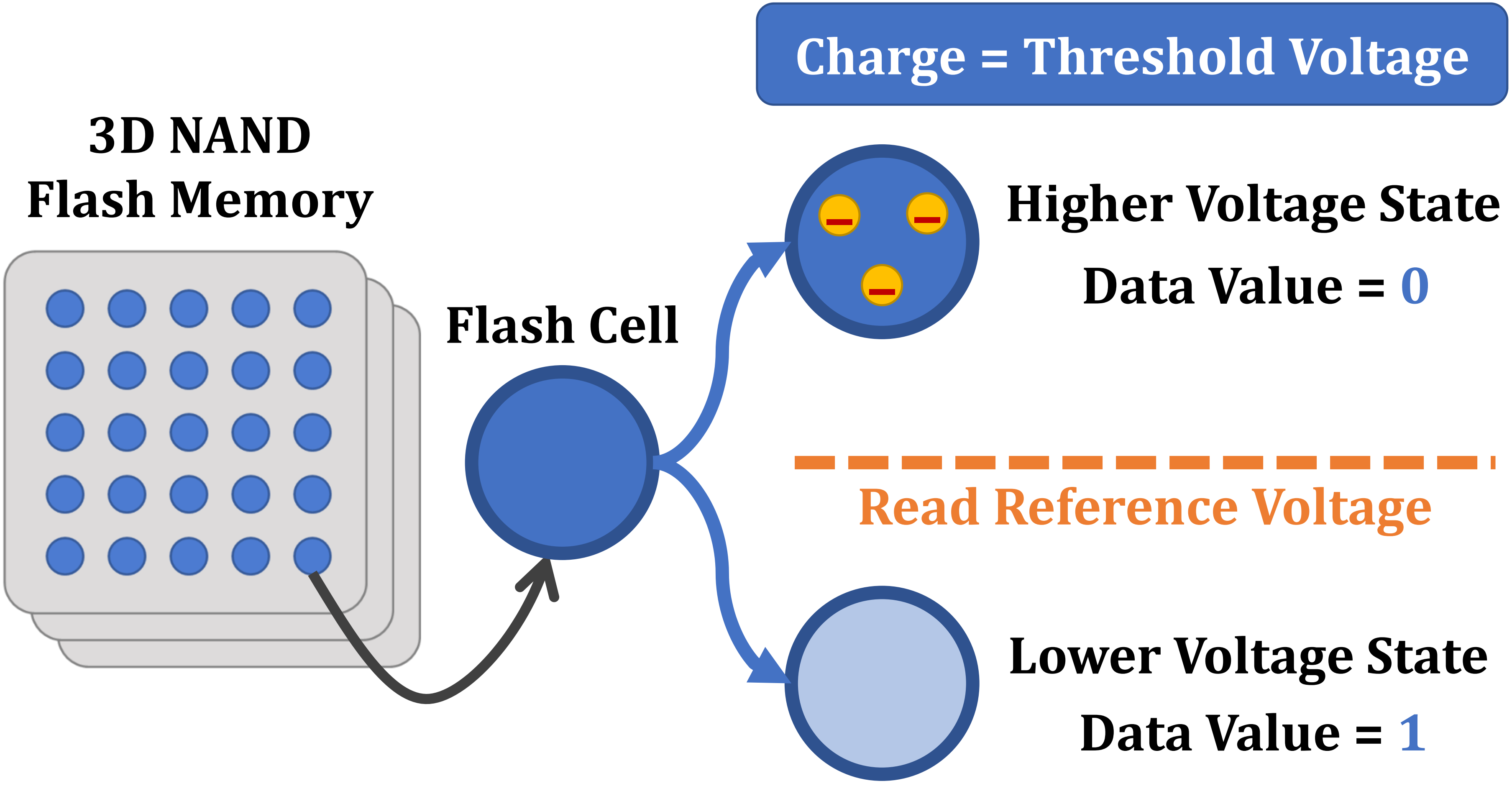
Executive Summary

- 3D NAND flash memory susceptible to **retention errors**
  - Charge leaks out of flash cell
  - Two unreported factors: ***self-recovery*** **and** ***temperature***
- We study *self-recovery* and *temperature* effects
  - **Experimental characterization** of *real* 3D NAND chips
  - **Unified Self-Recovery and Temperature (URT) Model**
    - Predicts impact of retention loss, wearout, self-recovery, temperature on **flash cell voltage**
    - **Low prediction error rate: 4.9%**
- We develop a new technique to improve flash reliability
  - **HeatWatch**
    - Uses URT model to find optimal read voltages for 3D NAND flash
    - **Improves flash lifetime by 3.85x**

Outline

- Executive Summary
- **Background on NAND Flash Reliability**
- Characterization of Self-Recovery and Temperature Effect on Real 3D NAND Flash Memory Chips
- URT: Unified Self-Recovery and Temperature Model
- HeatWatch Mechanism
- Conclusion

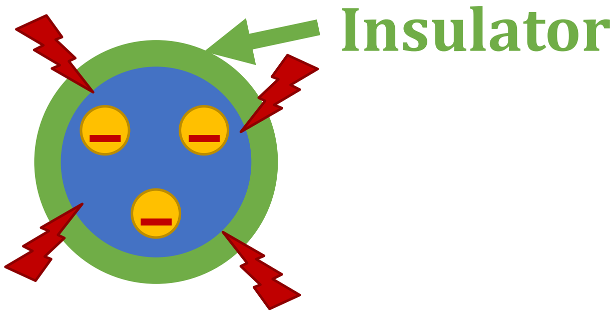
3D NAND Flash Memory Background



Charge = Threshold Voltage

Flash Wearout

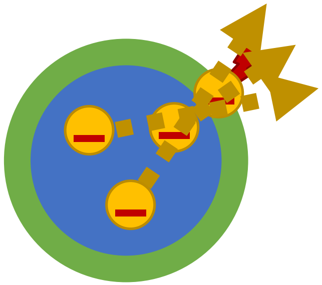
Program/Erase (P/E) → Wearout



Wearout Effects:

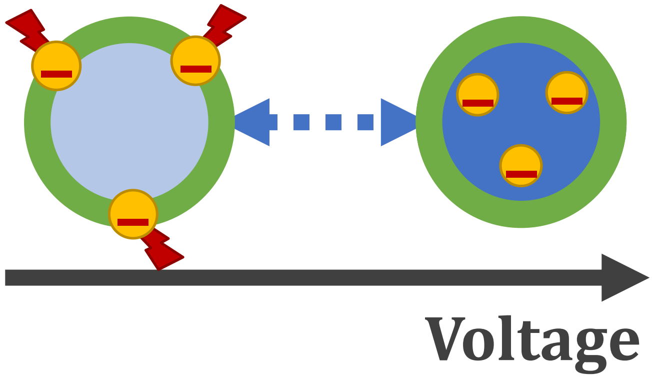
1. Retention Loss

(voltage shift over time)



2. Program Variation

(init. voltage difference b/w states)



Wearout Introduces Errors

Improving Flash Lifetime

Errors introduced by wearout
limit flash lifetime
(measured in P/E cycles)

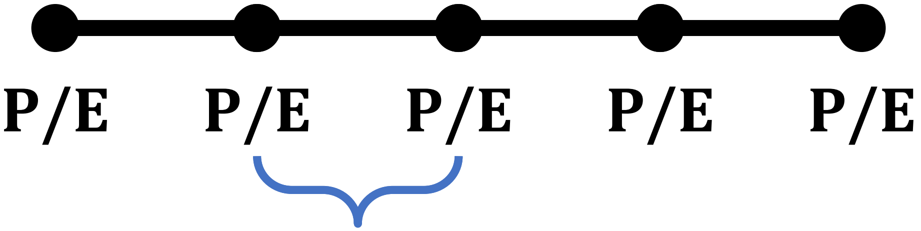
Two Ways to Improve Flash Lifetime →

Exploiting the Self-Recovery Effect

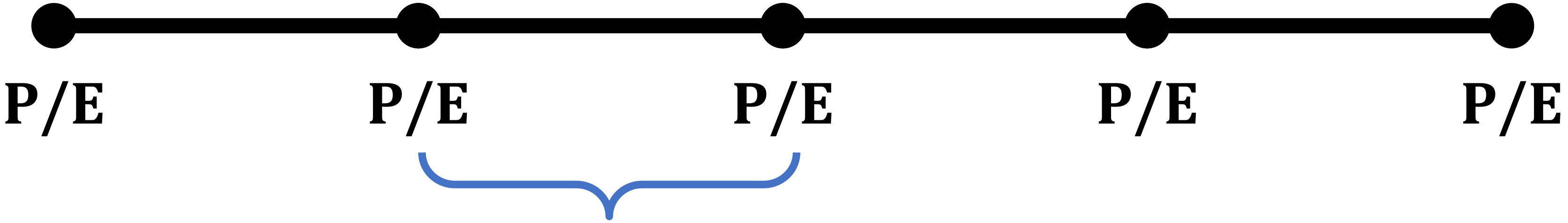
Exploiting the Temperature Effect

Exploiting the Self-Recovery Effect

Partially repairs damage due to wearout



Dwell Time: Idle Time Between P/E Cycles

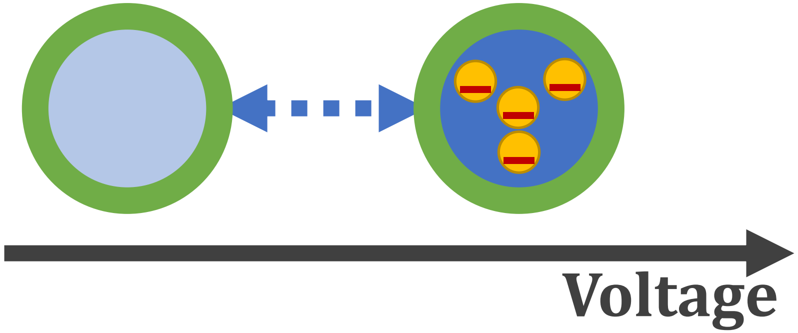


Longer Dwell Time: More Self-Recovery

Reduces Retention Loss

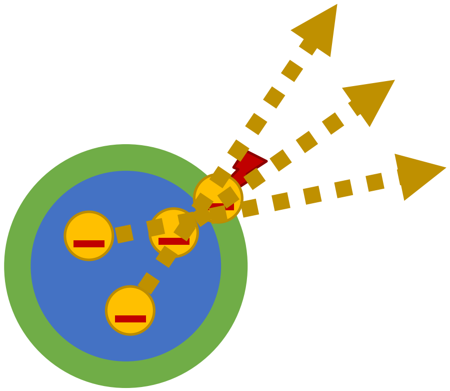
Exploiting the Temperature Effect

High Program Temperature



Increases Program Variation

High Storage Temperature



Accelerates Retention Loss

Prior Studies of Self-Recovery/Temperature

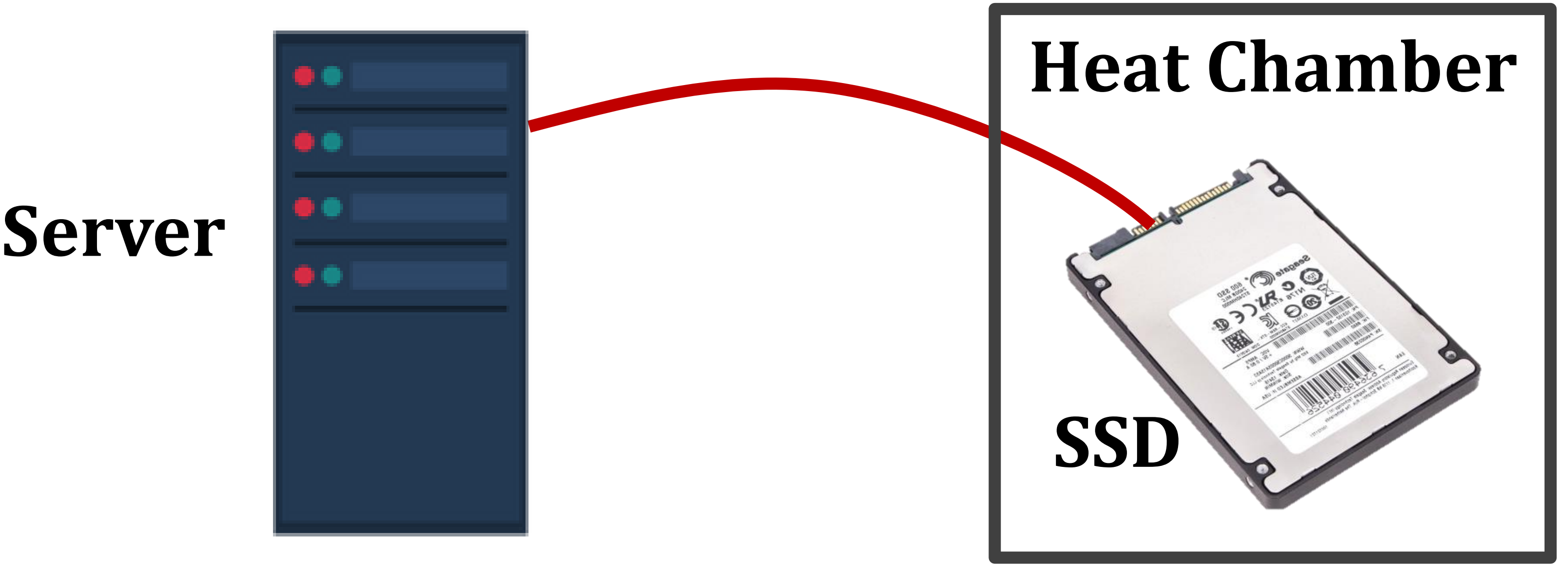
	Planar (2D) NAND	3D NAND
Self-Recovery Effect	✓ Mielke 2006	✗
Temperature Effect	✓ JEDEC 2010 (no characterization)	✗

Outline

- Executive Summary
- Background on NAND Flash Reliability
- **Characterization of Self-Recovery and Temperature Effect on Real 3D NAND Flash Memory Chips**
- URT: Unified Self-Recovery and Temperature Model
- HeatWatch Mechanism
- Conclusion

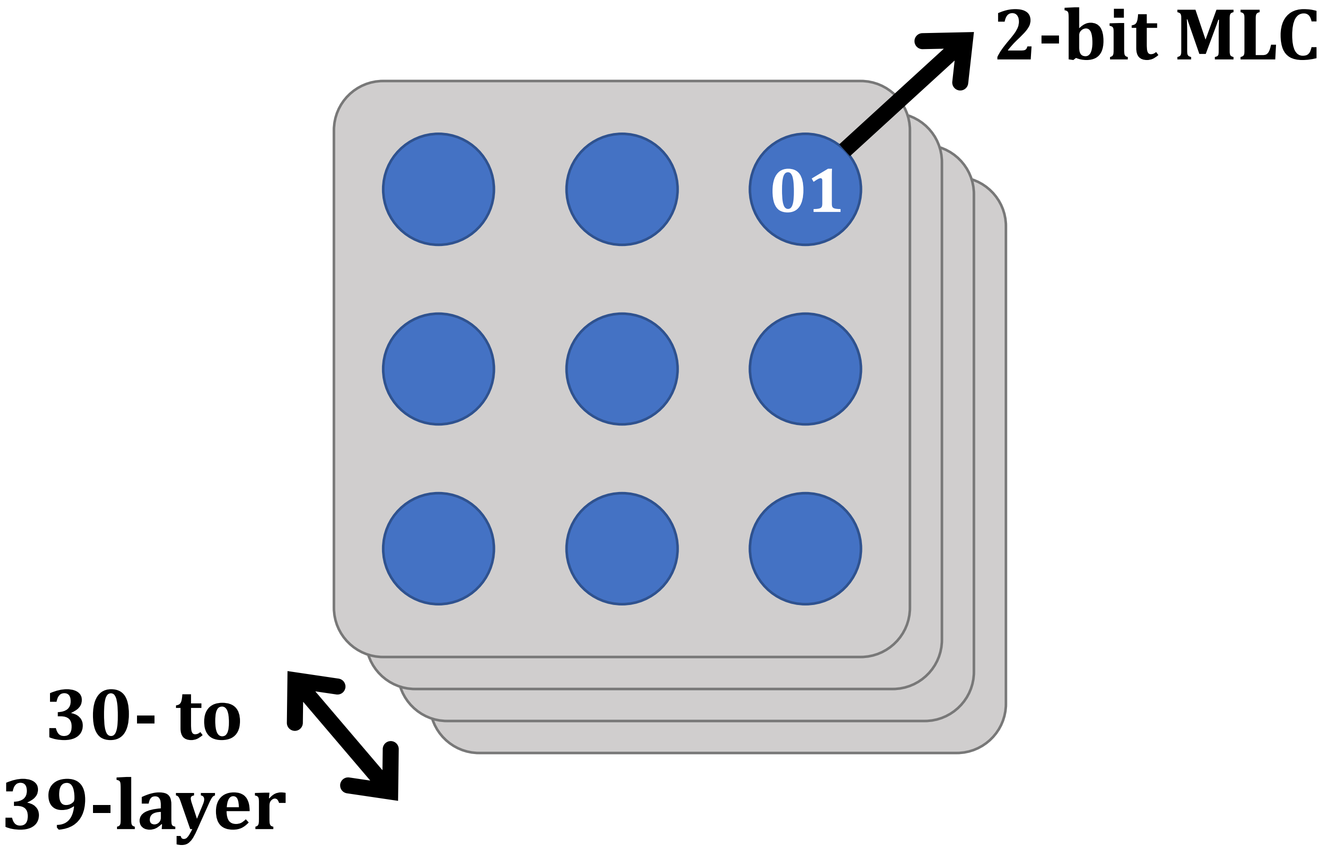
Characterization Methodology

- Modified firmware version in the flash controller
  - Control the read reference voltage of the flash chip
  - Bypass ECC to get raw NAND data (with raw bit errors)
- Control temperature with a heat chamber

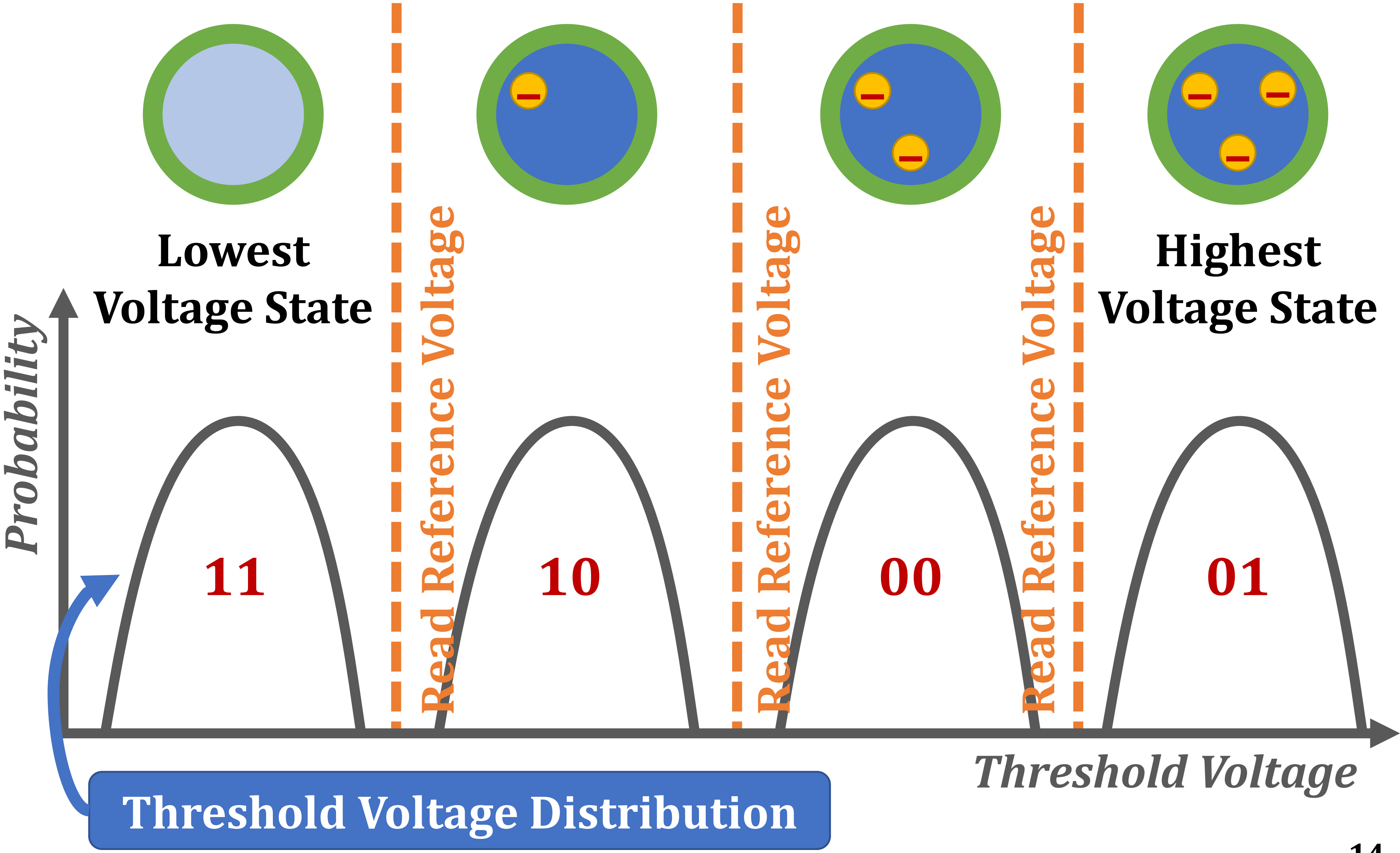


Characterized Devices

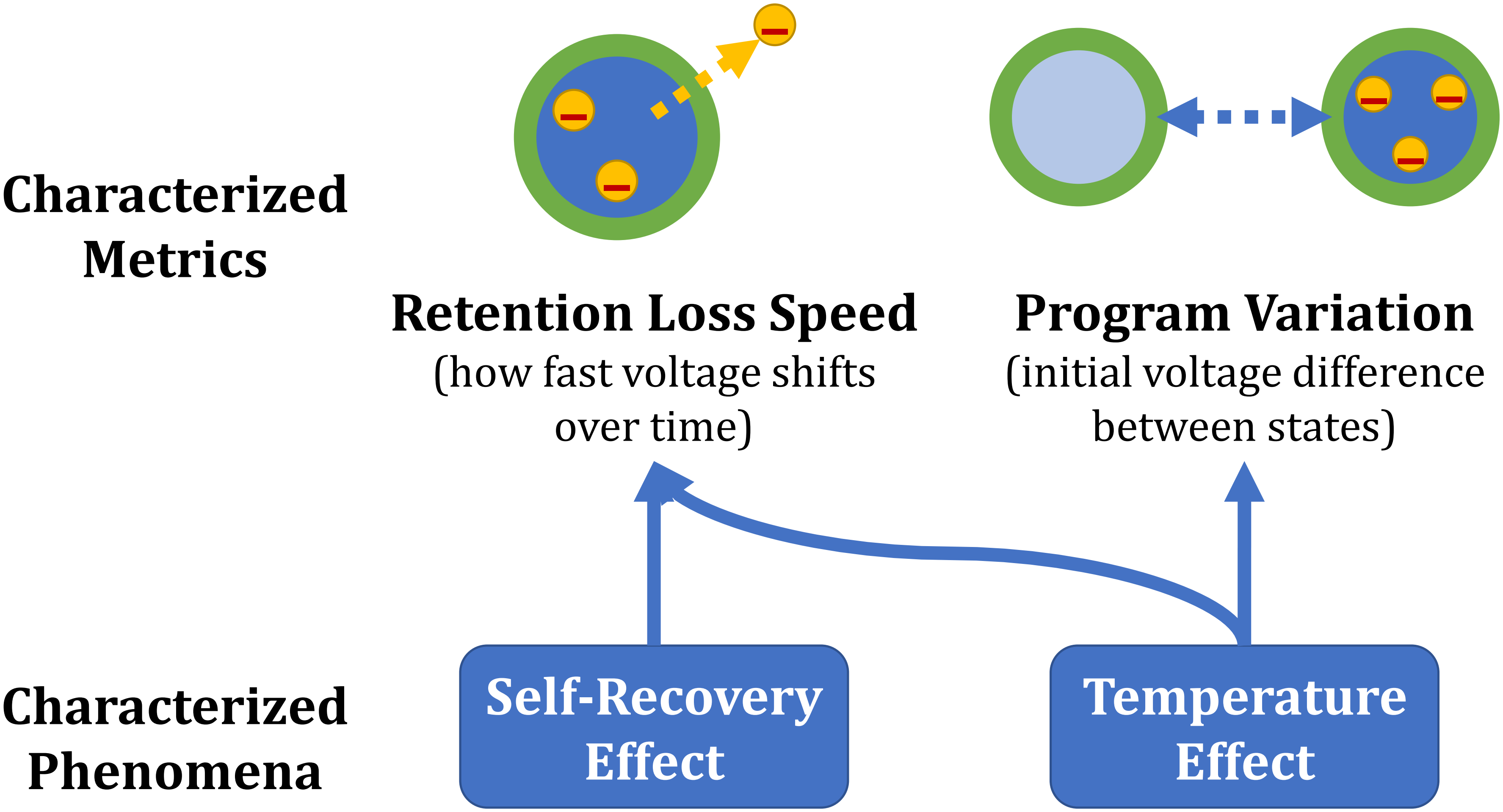
Real 30-39 Layer 3D MLC NAND Flash Chips



MLC Threshold Voltage Distribution Background



Characterization Goal



Characterized Metrics

Retention Loss Speed
(how fast voltage shifts over time)

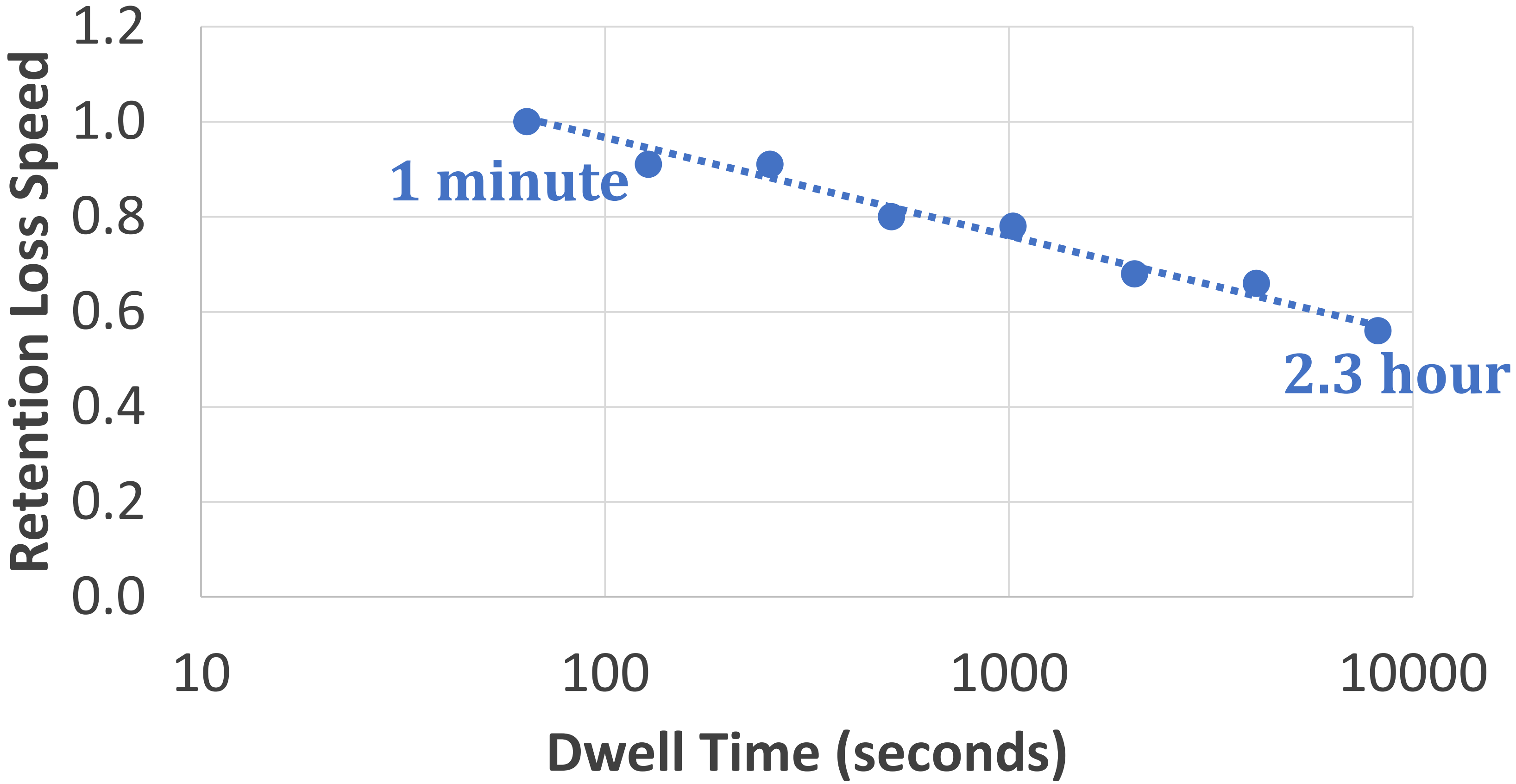
Program Variation
(initial voltage difference between states)

Characterized Phenomena

Self-Recovery Effect

Temperature Effect

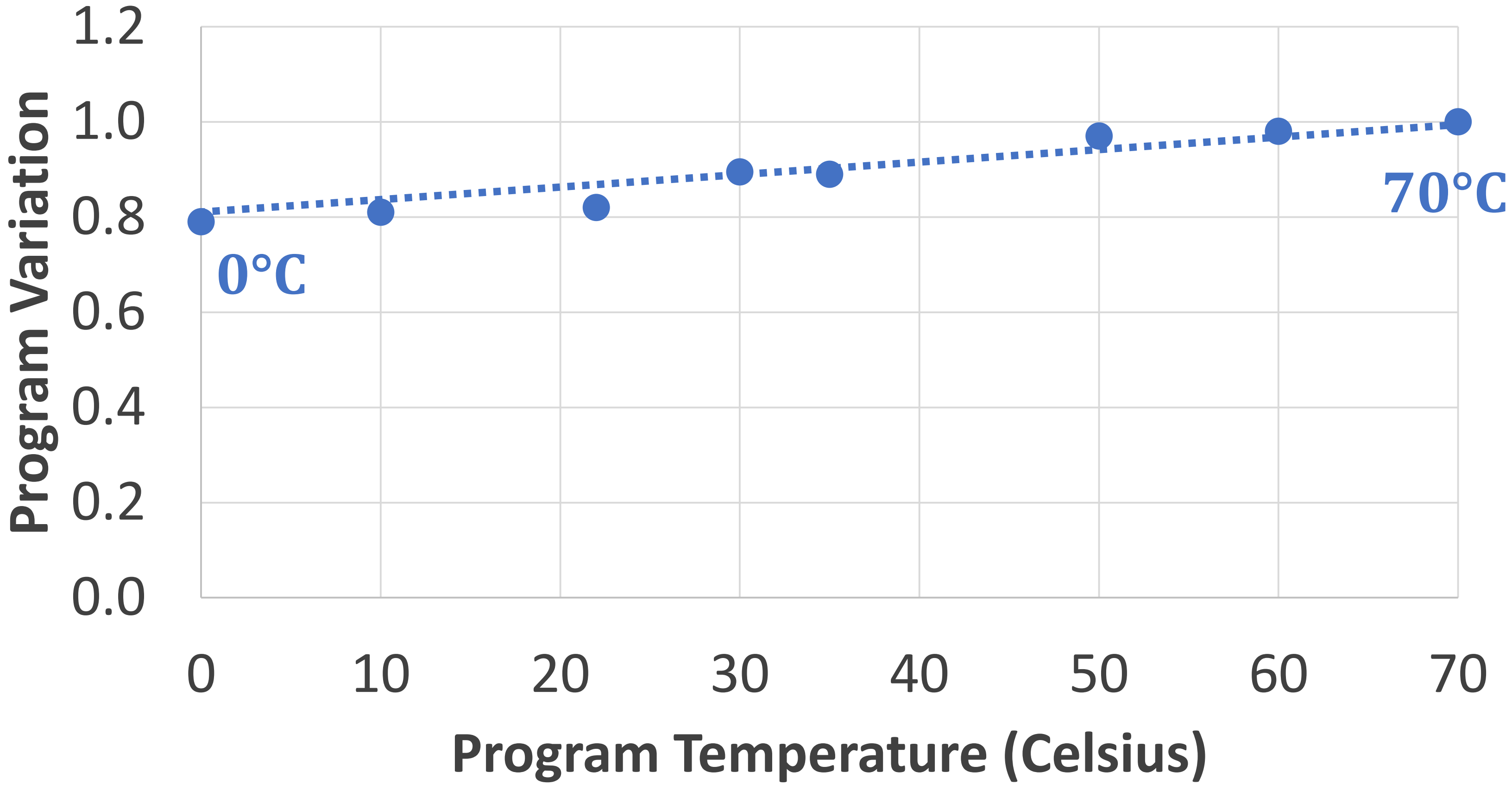
Self-Recovery Effect Characterization Results



Dwell time: Idle time between P/E cycles

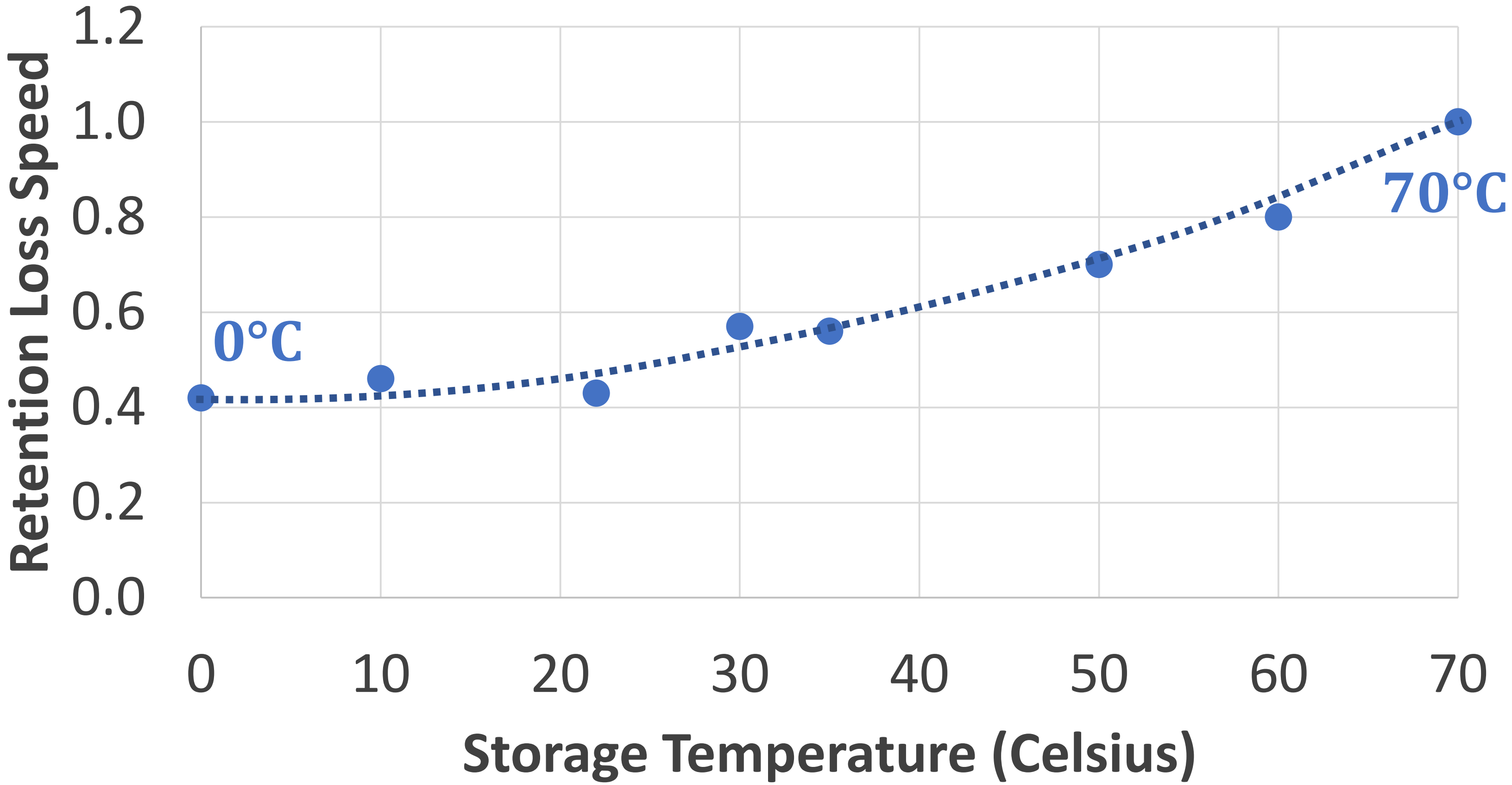
Increasing dwell time from 1 minute to 2.3 hours slows down retention loss speed by 40%

Program Temperature Effect Characterization Results



Increasing program temperature from 0°C to 70°C improves program variation by 21%

Storage Temperature Effect
Characterization Results



Lowering storage temperature from 70°C to 0°C slows down retention loss speed by 58%

Characterization Summary

Major Results:

- *Self-recovery* affects retention loss speed
- Program *temperature* affects program variation
- *Storage temperature* affects retention loss speed

Unified Model

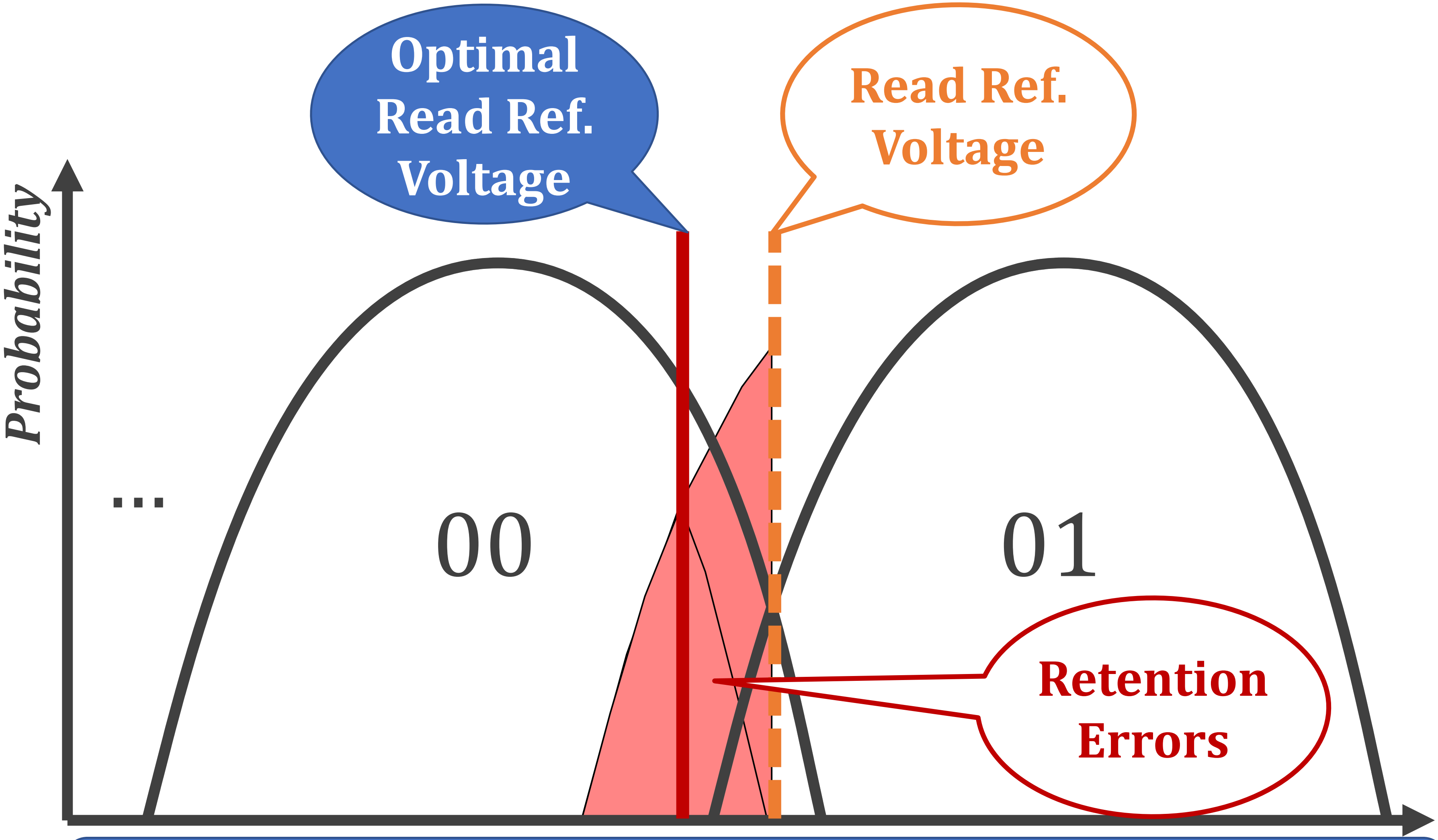
Other Characterizations Methods in the Paper:

- More detailed results on self-recovery and temperature
  - Effects on error rate
  - Effects on threshold voltage distribution
- Effects of recovery cycle (P/E cycles with long dwell time) on retention loss speed

Outline

- Executive Summary
- Background on NAND Flash Reliability
- Characterization of Self-Recovery and Temperature Effect on Real 3D NAND Flash Memory Chips
- **URT: Unified Self-Recovery and Temperature Model**
- HeatWatch Mechanism
- Conclusion

Minimizing 3D NAND Errors

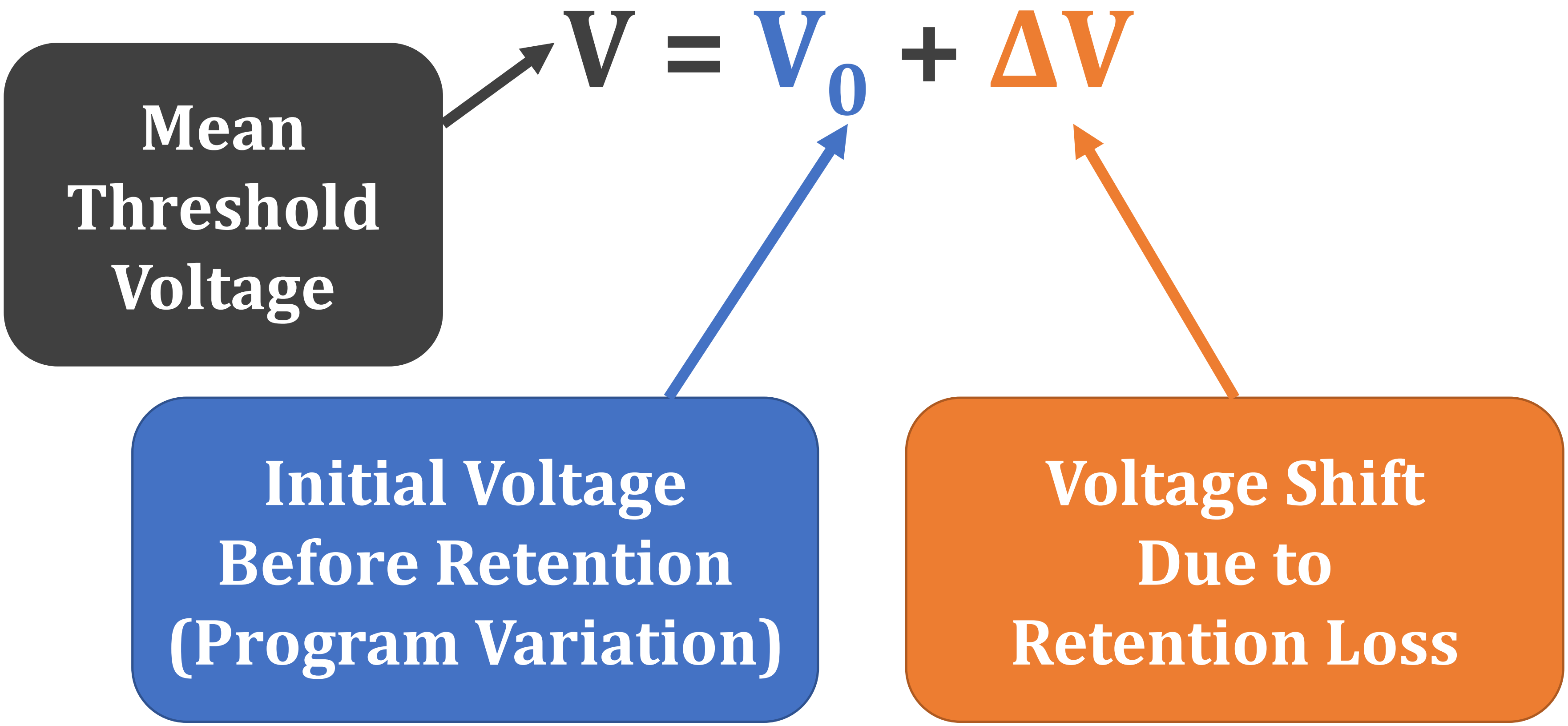


Optimal read reference voltage minimizes 3D NAND errors

Predicting the Mean Threshold Voltage

Our URT Model:

$$V = V_0 + \Delta V$$

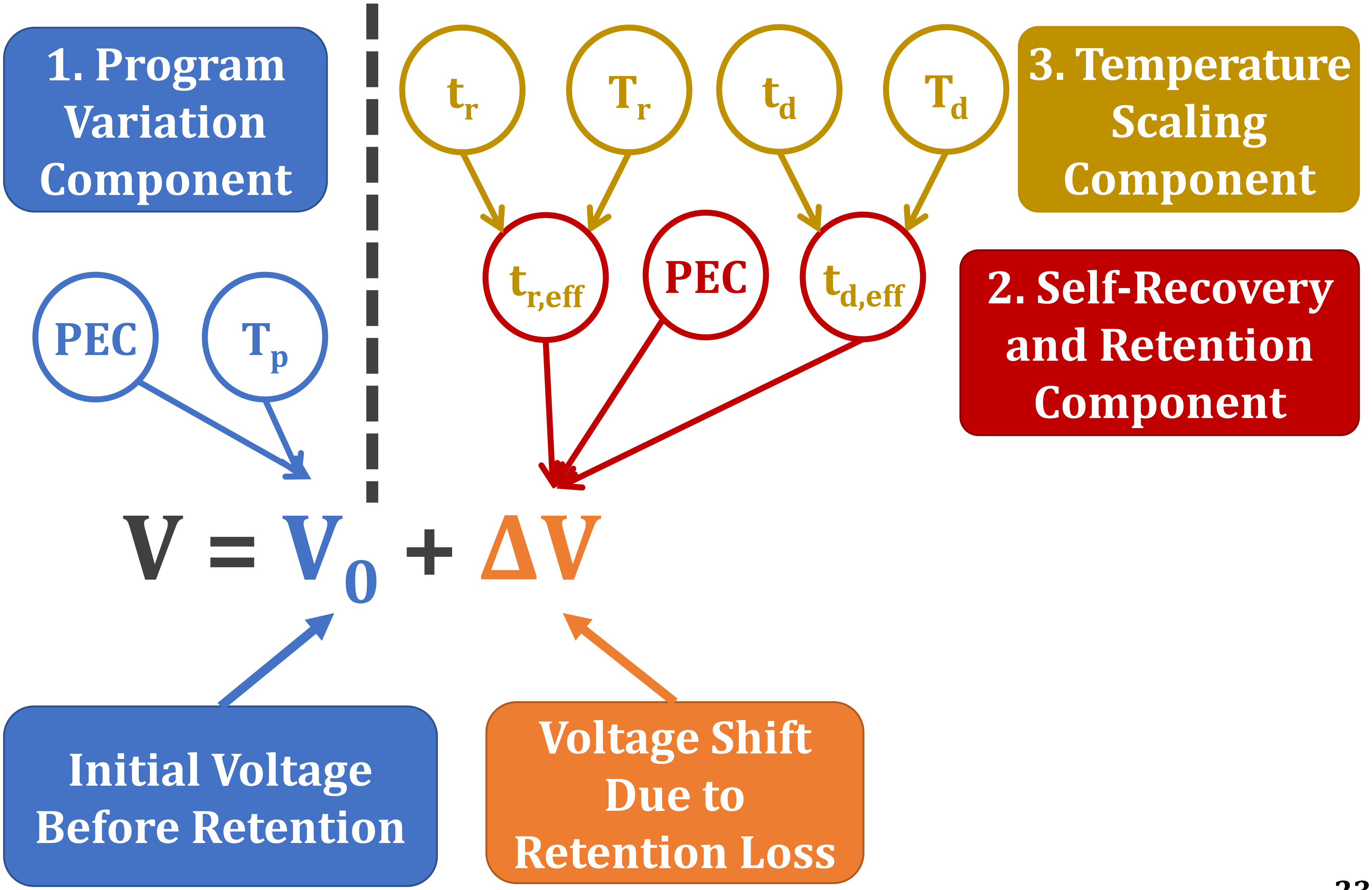


Mean Threshold Voltage

Initial Voltage Before Retention (Program Variation)

Voltage Shift Due to Retention Loss

URT Model Overview



1. Program Variation Component

PEC

T_p

3. Temperature Scaling Component

t_r T_r t_d T_d

$t_{r,eff}$ PEC $t_{d,eff}$

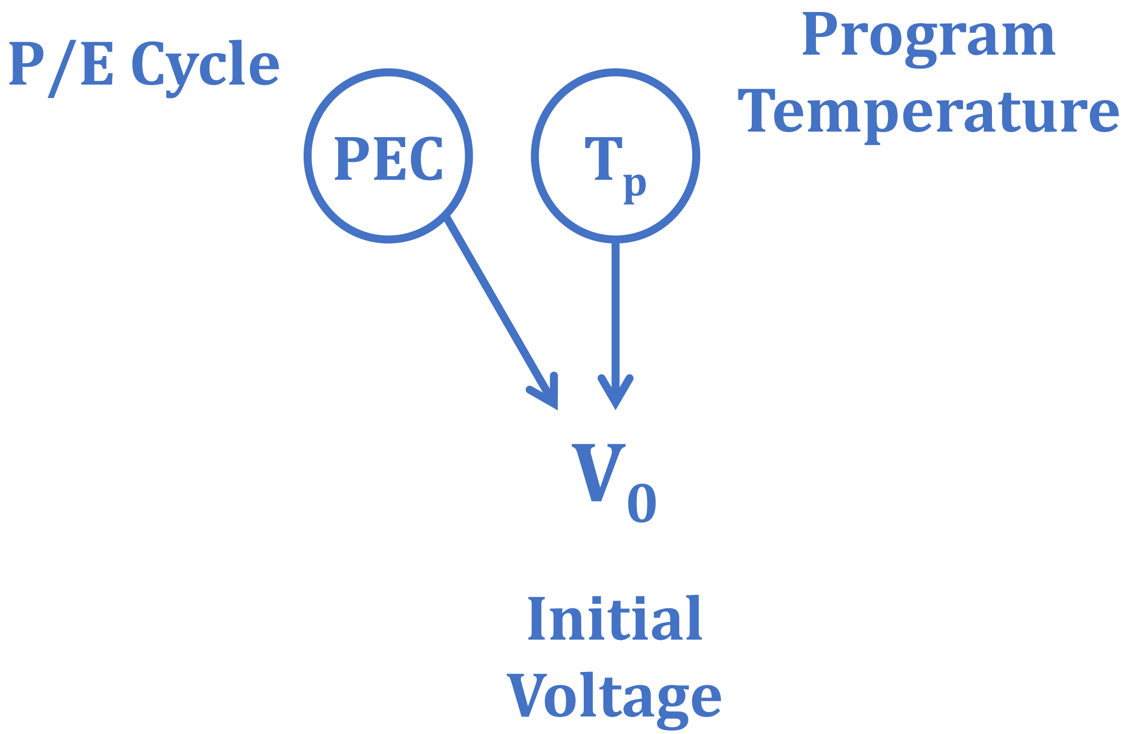
2. Self-Recovery and Retention Component

$$V = V_0 + \Delta V$$

Initial Voltage Before Retention

Voltage Shift Due to Retention Loss

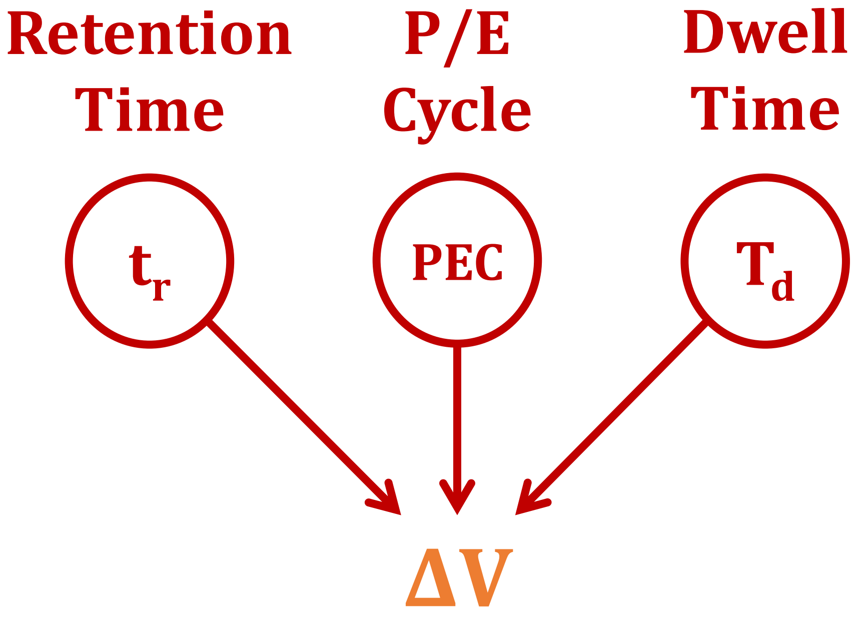
1. Program Variation Component



$$V_0 = A \cdot T_p \cdot PEC + B \cdot T_p + C \cdot PEC + D$$

✓ **Validation: R^2 = 91.7%**

2. Self-Recovery and Retention Component



Retention Time

P/E Cycle

Dwell Time

t_r

PEC

T_d

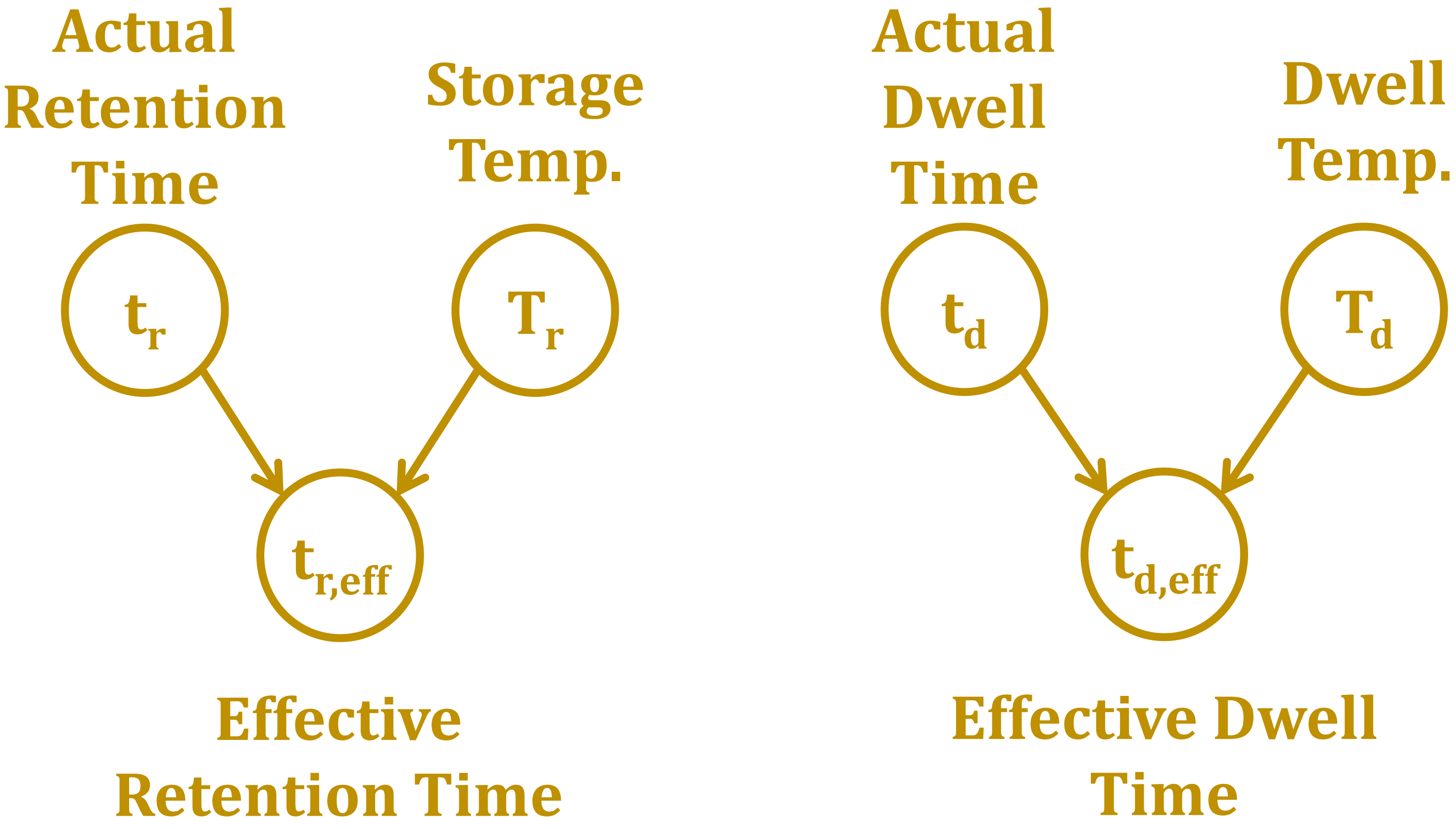
ΔV

Retention Shift

$$\Delta V\,(t_{er}, t_{ed}, PEC) = b \cdot (PEC + c) \cdot \ln\left(1 + \frac{t_{er}}{t_0 + a \cdot t_{ed}}\right)$$

Validation: 3x more accurate than state-of-the-art model

3. Temperature Scaling Component



Actual Retention Time

Storage Temp.

t_r

T_r

$t_{r,eff}$

Effective Retention Time

Actual Dwell Time

Dwell Temp.

t_d

T_d

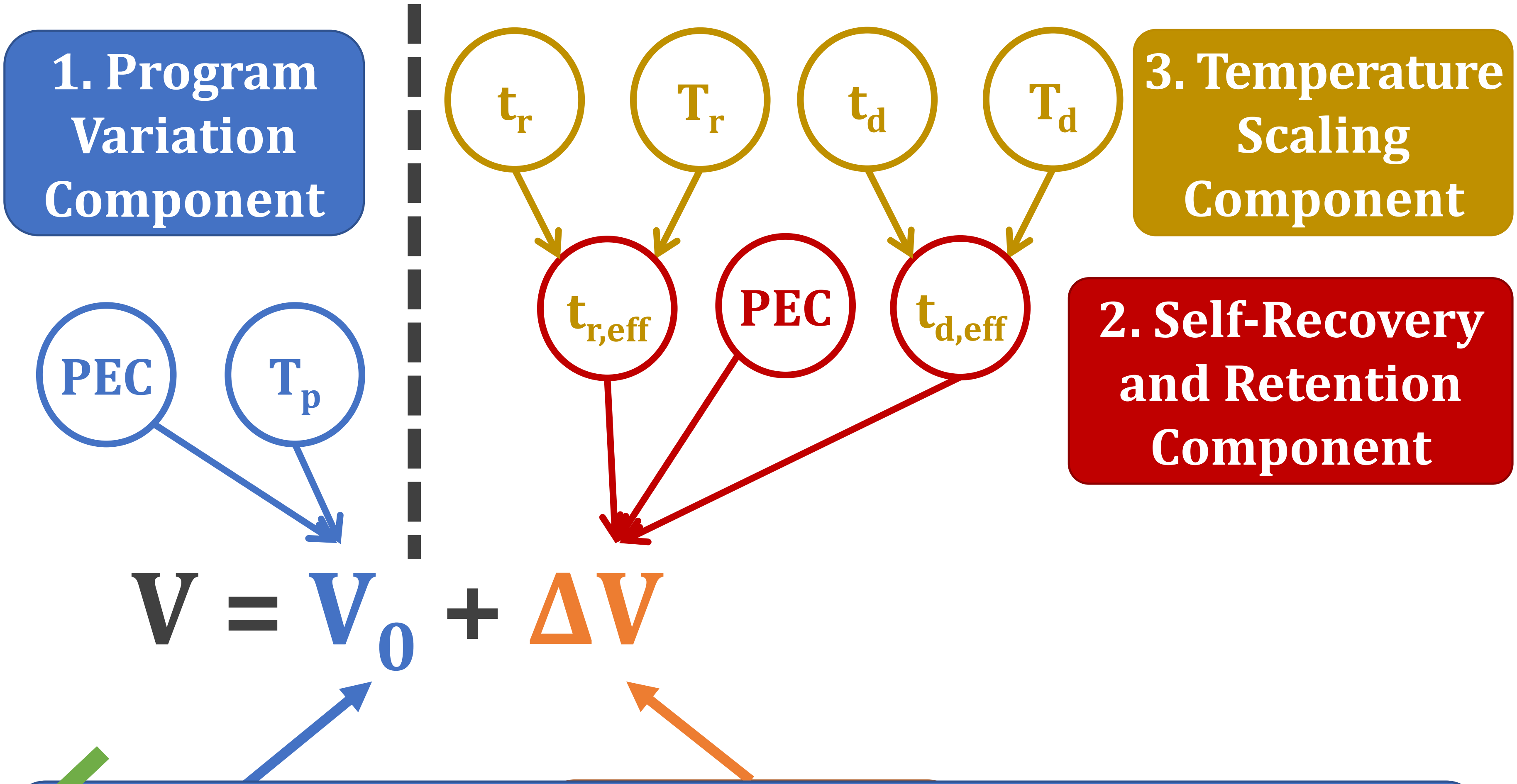
$t_{d,eff}$

Effective Dwell Time

Arrhenius Equation:
$$AF = \frac{t_{real}}{t_{room}} = \exp\left(\frac{E_a}{k_B}\cdot\left(\frac{1}{T_{real}} - \frac{1}{T_{room}}\right)\right)$$

Validation: Adjust an important parameter, E_a, from 1.1 eV to 1.04 eV

URT Model Summary


1. Program Variation Component
PEC
T_p
t_r
T_r
t_d
T_d
3. Temperature Scaling Component
$t_{r,eff}$
PEC
$t_{d,eff}$
2. Self-Recovery and Retention Component

$$V = V_0 + \Delta V$$

Validation:
Prediction Error Rate = 4.9%

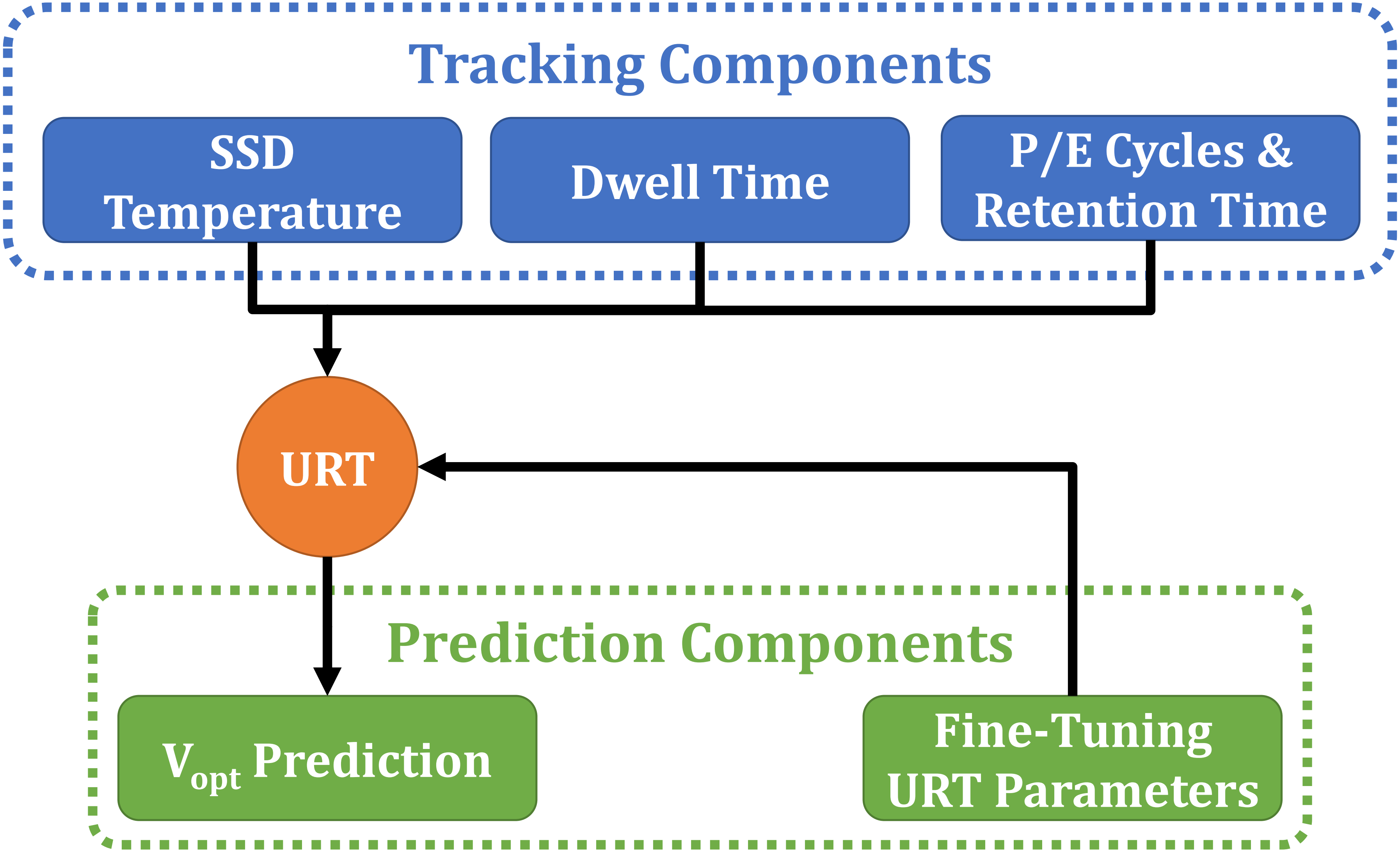
Outline

- Executive Summary
- Background on NAND Flash Reliability
- Characterization of Self-Recovery and Temperature Effect on Real 3D NAND Flash Memory Chips
- URT: Unified Self-Recovery and Temperature Model
- **HeatWatch Mechanism**
- Conclusion

HeatWatch Mechanism

- **Key Idea**
  - **Predict change in threshold voltage distribution** by using the URT model
  - **Adapt read reference voltage to near-optimal** (V_{opt}) based on predicted change in voltage distribution

HeatWatch Mechanism Overview



Tracking Components

SSD Temperature

Dwell Time

P/E Cycles & Retention Time

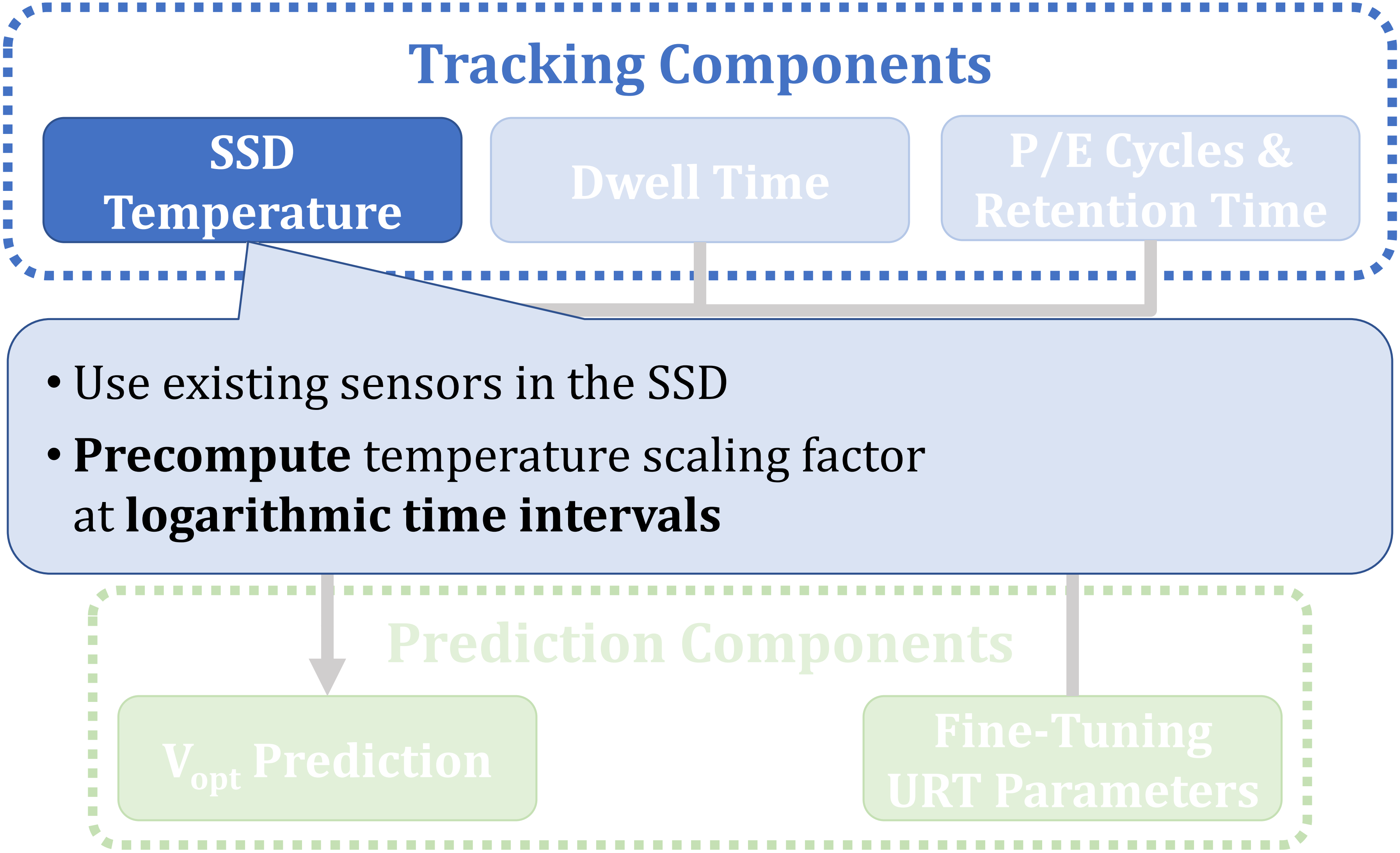
URT

Prediction Components

V_{opt} Prediction

Fine-Tuning URT Parameters

Tracking SSD Temperature



Tracking Components

SSD Temperature

Dwell Time

P/E Cycles & Retention Time

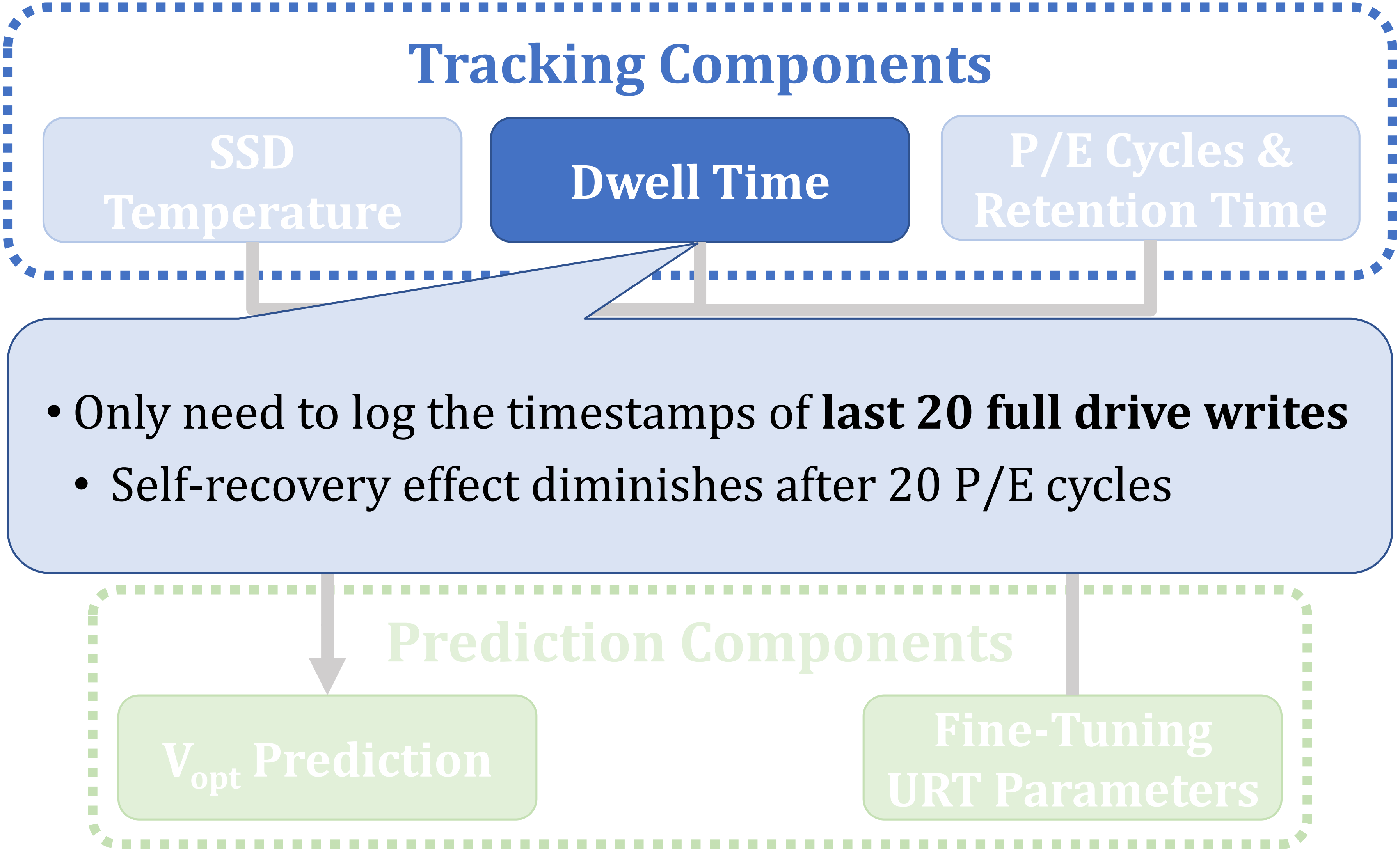
- Use existing sensors in the SSD
- **Precompute** temperature scaling factor at **logarithmic time intervals**

Prediction Components

V_{opt} Prediction

Fine-Tuning URT Parameters

Tracking Dwell Time



Tracking Components

SSD Temperature

Dwell Time

P/E Cycles & Retention Time

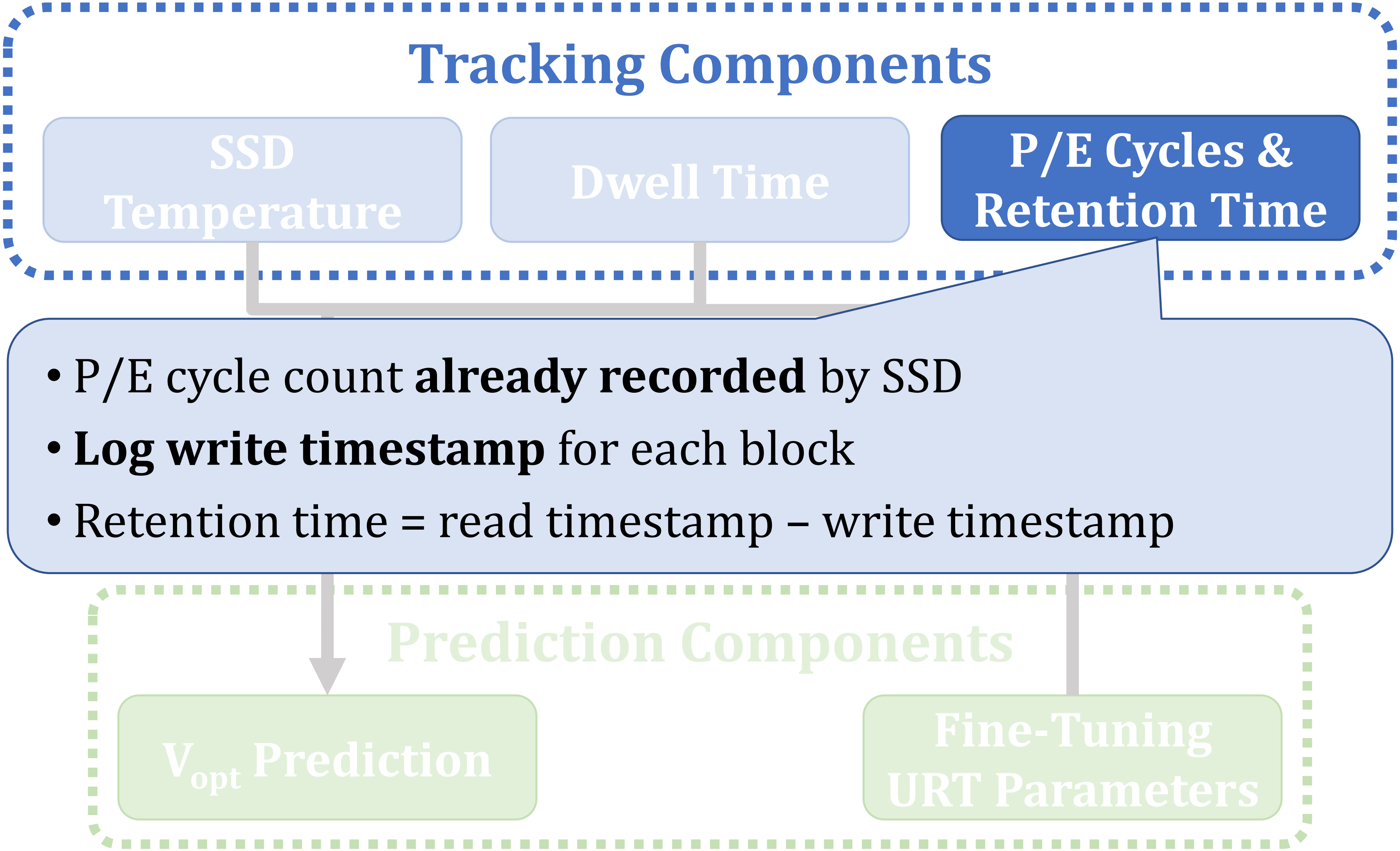
- Only need to log the timestamps of **last 20 full drive writes**
  - Self-recovery effect diminishes after 20 P/E cycles

Prediction Components

V_{opt} Prediction

Fine-Tuning URT Parameters

Tracking P/E Cycles and Retention Time



Tracking Components

SSD Temperature

Dwell Time

P/E Cycles & Retention Time

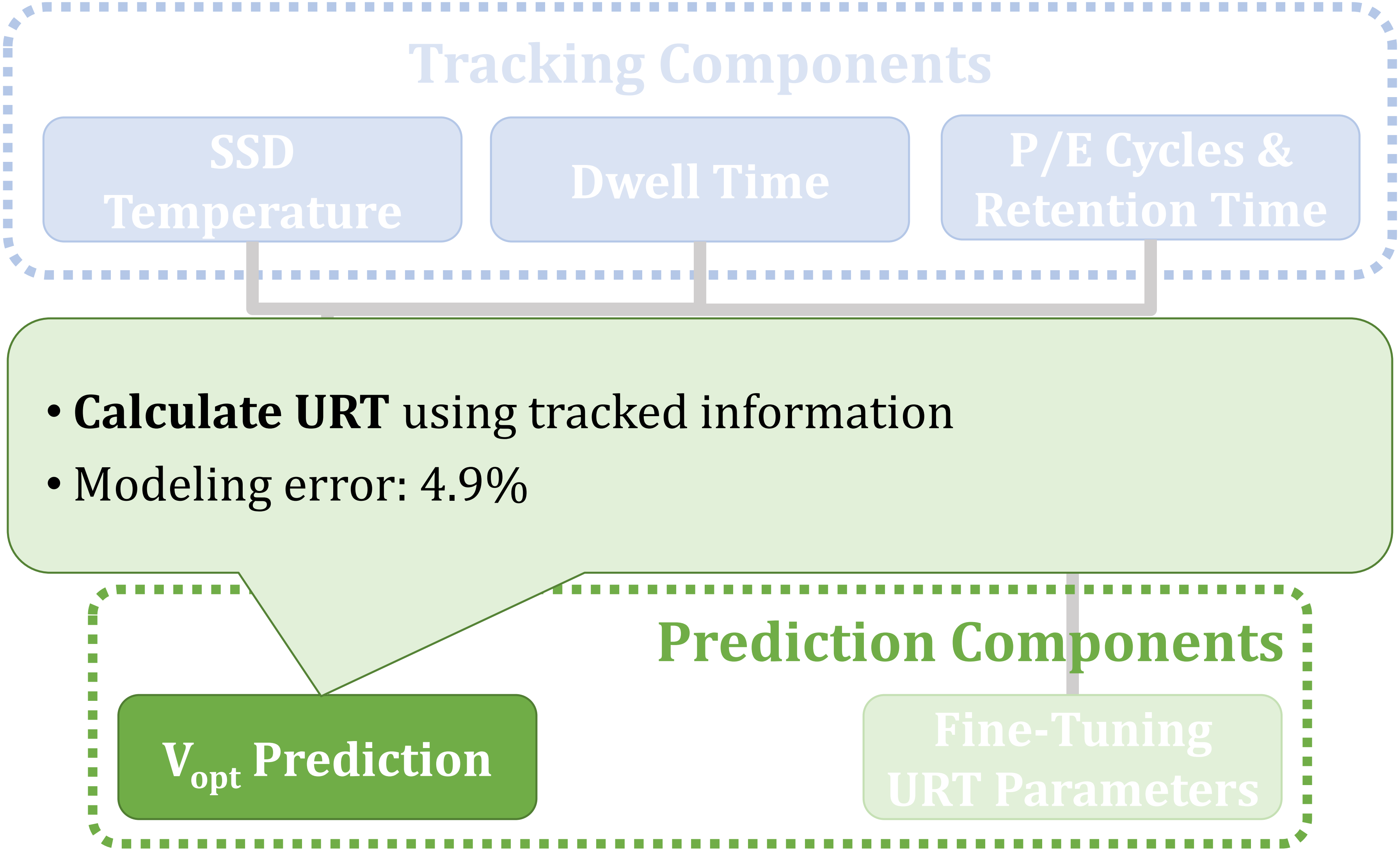
- P/E cycle count **already recorded** by SSD
- **Log write timestamp** for each block
- Retention time = read timestamp – write timestamp

Prediction Components

V_{opt} Prediction

Fine-Tuning URT Parameters

Predicting Optimal Read Reference Voltage



Tracking Components

SSD Temperature

Dwell Time

P/E Cycles & Retention Time

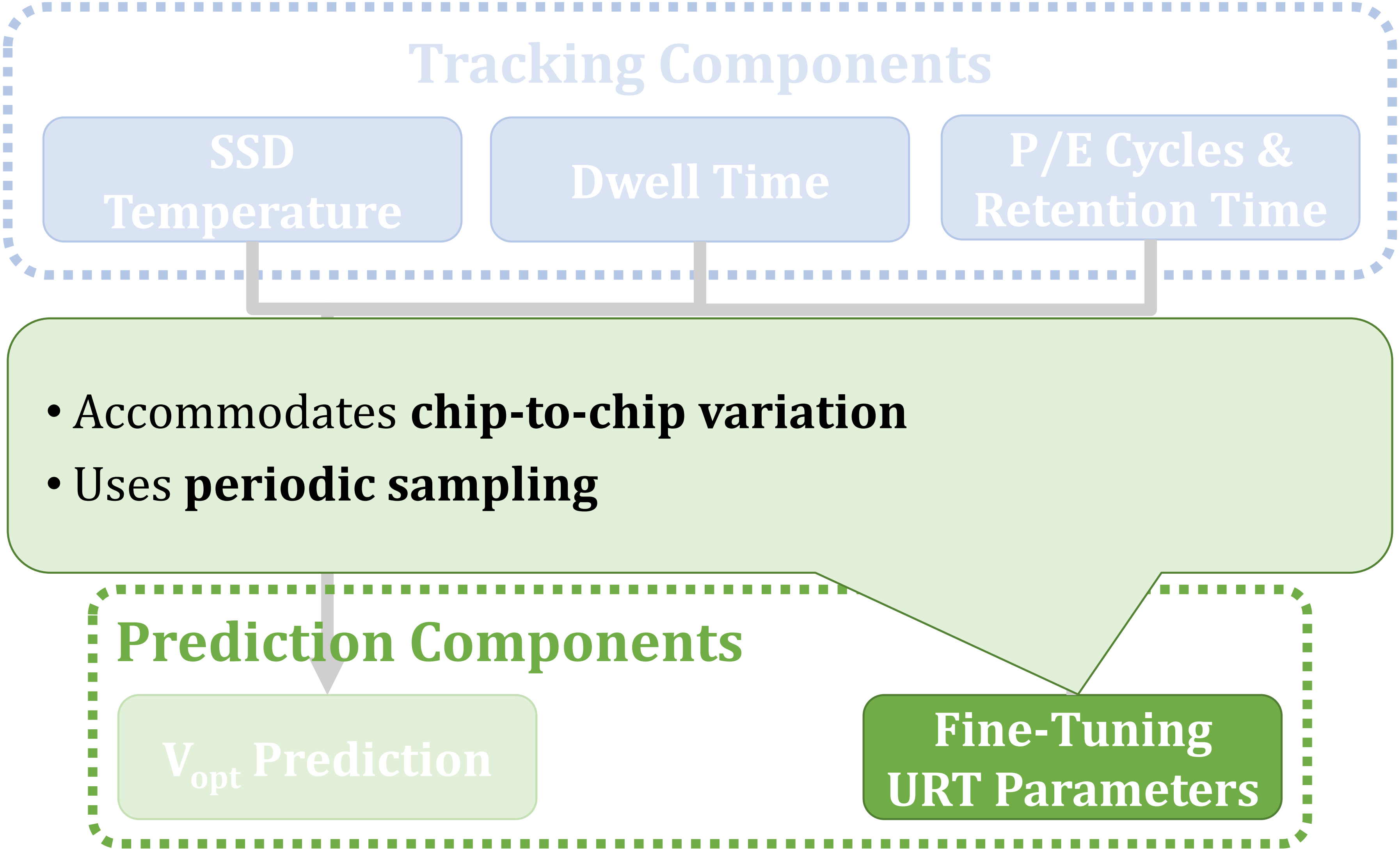
- **Calculate URT** using tracked information
- Modeling error: 4.9%

Prediction Components

V_{opt} Prediction

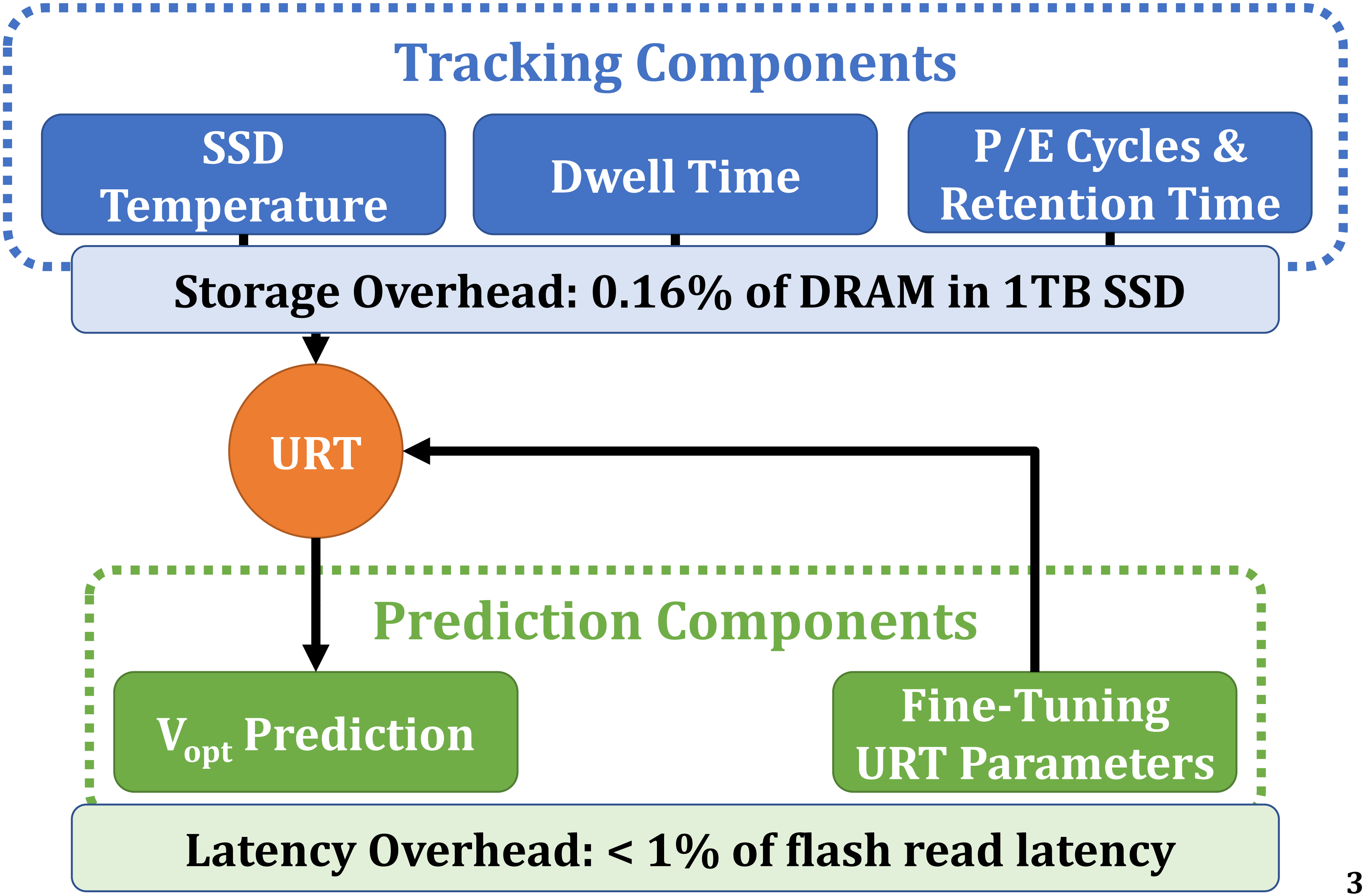
Fine-Tuning URT Parameters

Fine-Tuning URT Parameters Online



- Accommodates **chip-to-chip variation**
- Uses **periodic sampling**

HeatWatch Mechanism Summary



Tracking Components

SSD Temperature

Dwell Time

P/E Cycles & Retention Time

Storage Overhead: 0.16% of DRAM in 1TB SSD

URT

Prediction Components

V_{opt} Prediction

Fine-Tuning URT Parameters

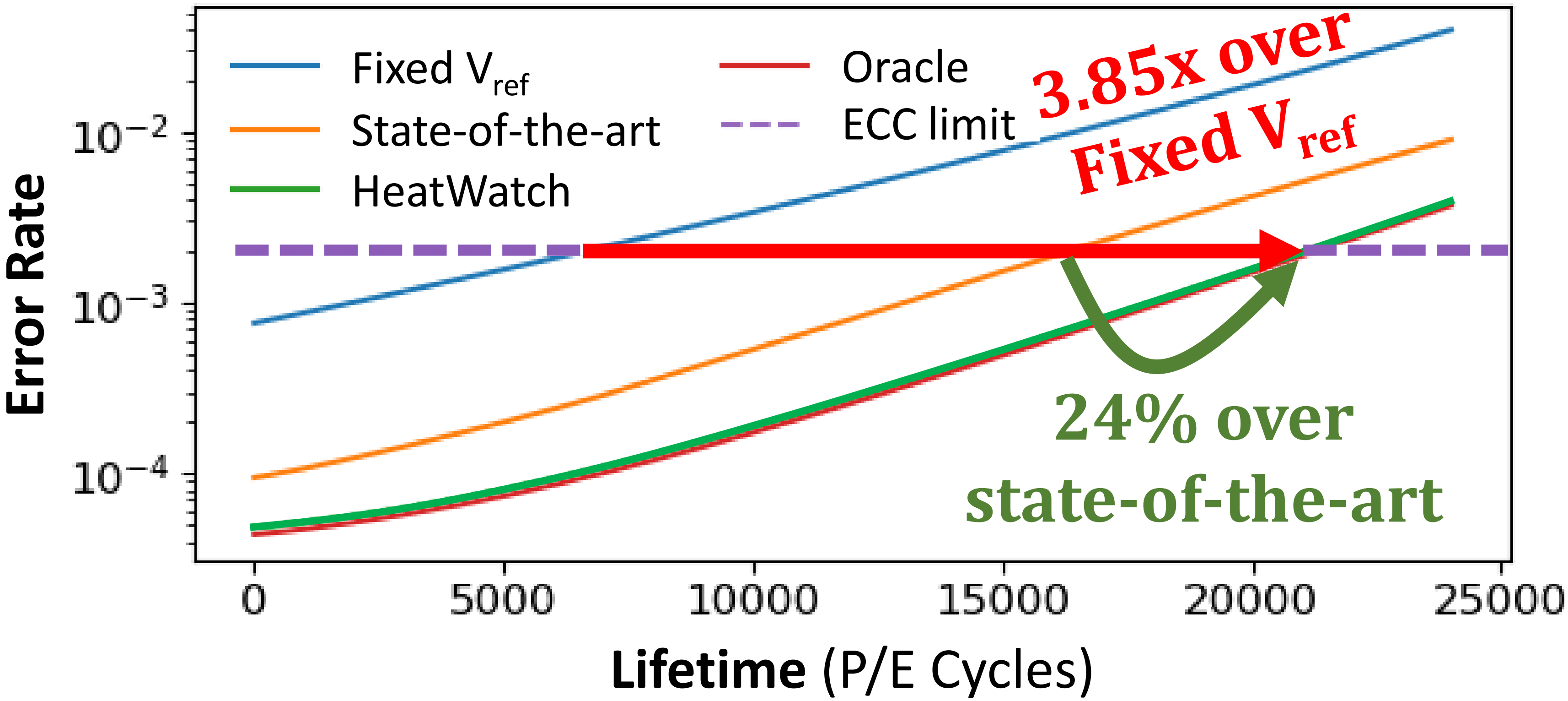
Latency Overhead: < 1% of flash read latency

HeatWatch Evaluation Methodology

- **28 real workload storage traces**
  - MSR-Cambridge
  - We use **real dwell time, retention time values** obtained from traces

- **Temperature Model:**
  Trigonometric function + Gaussian noise
  - Represents periodic temperature variation in each day
  - Includes small transient temperature variation

HeatWatch Greatly Improves Flash Lifetime



3.85x over Fixed V_{ref}

24% over state-of-the-art

HeatWatch improves lifetime by capturing the effect of retention, wearout, self-recovery, temperature

Outline

- Executive Summary
- Background on NAND Flash Reliability
- Characterization of Self-Recovery and Temperature Effect on Real 3D NAND Flash Memory Chips
- URT: Unified Self-Recovery and Temperature Model
- HeatWatch Mechanism
- **Conclusion**

Conclusion

- 3D NAND flash memory susceptible to **retention errors**
  - Charge leaks out of flash cell
  - Two unreported factors: ***self-recovery* and *temperature***
- We study *self-recovery* and *temperature* effects
  - **Experimental characterization** of *real* 3D NAND chips
  - **Unified Self-Recovery and Temperature (URT) Model**
    - Predicts impact of retention loss, wearout, self-recovery, temperature on **flash cell voltage**
    - **Low prediction error rate: 4.9%**
- We develop a new technique to improve flash reliability
  - **HeatWatch**
    - Uses URT model to find optimal read voltages for 3D NAND flash
    - **Improves flash lifetime by 3.85x**

HeatWatch

Improving 3D NAND Flash Memory Device Reliability by Exploiting Self-Recovery and Temperature Awareness

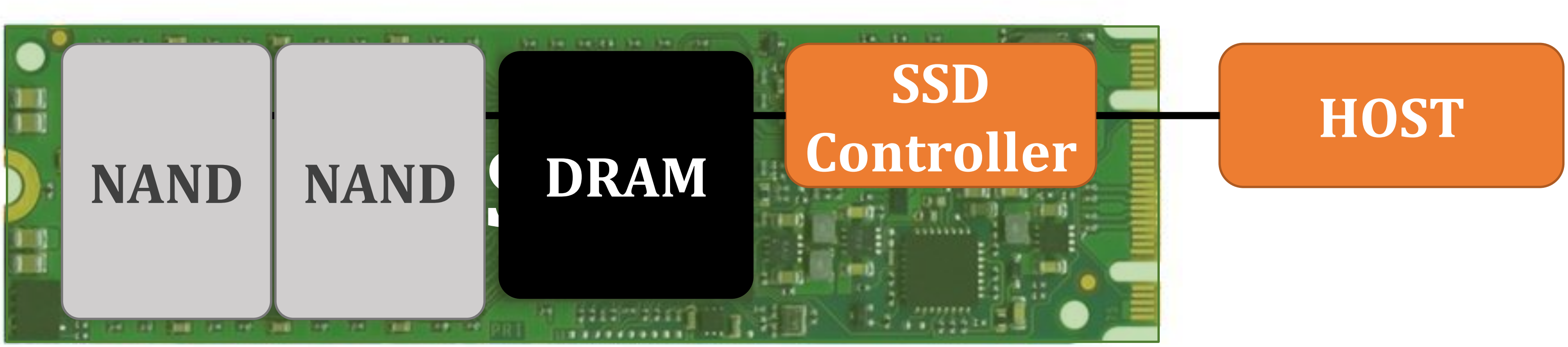
Yixin Luo **Saugata Ghose** **Yu Cai** **Erich F. Haratsch** **Onur Mutlu**

Carnegie Mellon SK hynix ETH Zürich

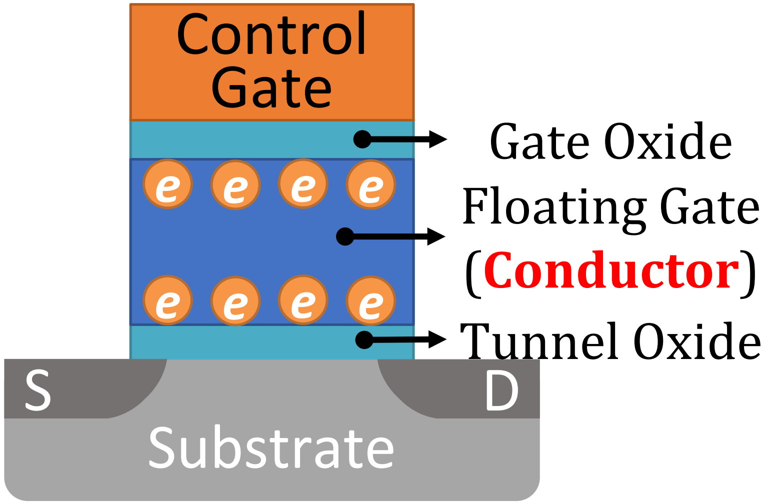
SAFARI SEAGATE

Backup Slides

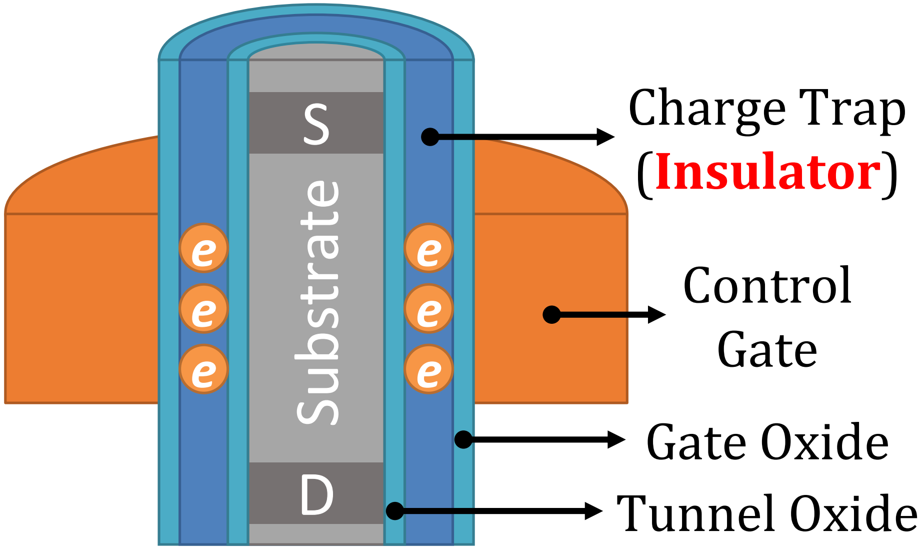
SSD Architecture



3D vs. 2D Flash Cell Design



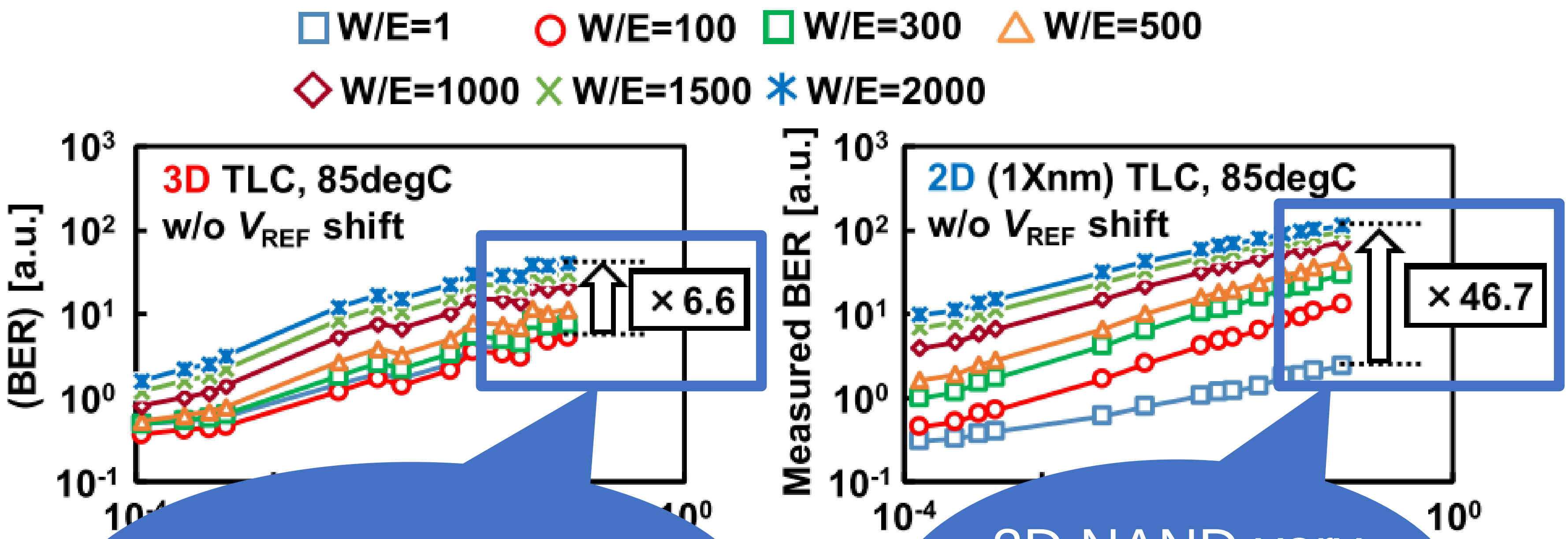
Floating-Gate Cell



3D Charge-Trap Cell

Charges stored in insulator, thinner tunnel oxide
→ Faster data retention

3D vs. 2D Retention Characteristics



3D NAND uniformly affected by wearout

2D NAND very sensitive to wearout

Source: K. Mizoguchi, et al., “Data-Retention Characteristics Comparison of 2D and 3D TLC NAND Flash Memories,” IMW, 2017.

Limitations

- Vendor-to-vendor variation
  - Self-recovery and temperature effect should be similar for 3D charge trap NAND (Samsung, Hynix, Toshiba, Sandisk)

- Chip-to-chip variation
  - Each of our experiments takes several months
  - Expect future large-scale study on 3D NAND errors

Not our limitation:

- Any process variation within a chip
  - Our results include tens of randomly selected flash blocks
  - ~1 million cells

Generalizability of Results

- Should apply to other 3D NAND flash memory that uses charge trap cells (Samsung, Hynix, Toshiba, Sandisk)

Self-Recovery and Temperature in Planar NAND

- UDM [Mielke 2006]
- Only models retention shift, no initial voltage
- Exponential P/E cycle effect
- Activation energy for planar NAND

- 3 other works propose mechanism and speculate different lifetime improvements
  - 211x [Mohan+ HotStorage10]
  - 5.8x [Wu+ HotStorage11]
  - 2.8x [Lee+ FAST12]

Novelty vs. UDM

- 3D charge trap cells are more resilient to P/E cycling than floating-gate cells in planar NAND
- Different activation energy
- Program temperature effect not discussed in planar NAND

Ideal SSD Temperature

- It depends!
  - High program temperature increases program variation (good)
  - High dwell temperature accelerates self-recovery (good)
  - High retention temperature accelerates retention loss (bad)

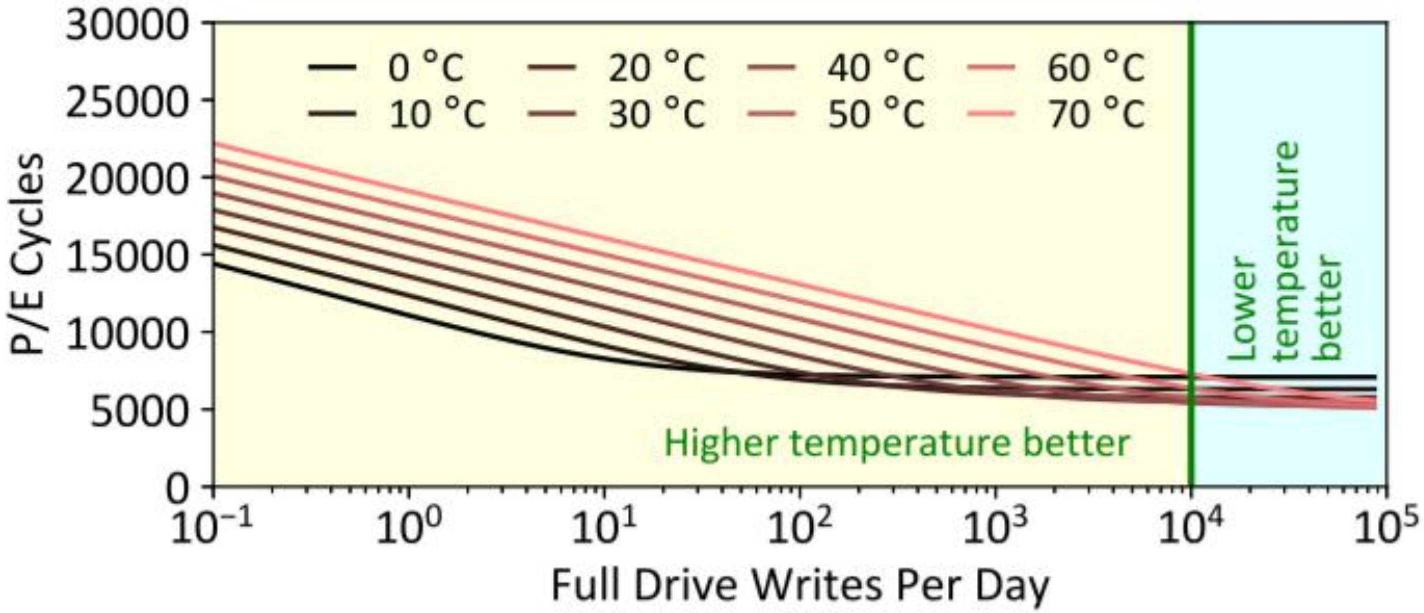


Figure 12: Change in flash lifetime due to write intensity and environmental temperature (t_r = 3 months).

URT Fine Tuning

- Randomly sample 10 wordlines in each chip
- Learn V_{opt} by sweeping V_{ref}
- Fit URT model with newly learned V_{opt}

HeatWatch Overhead

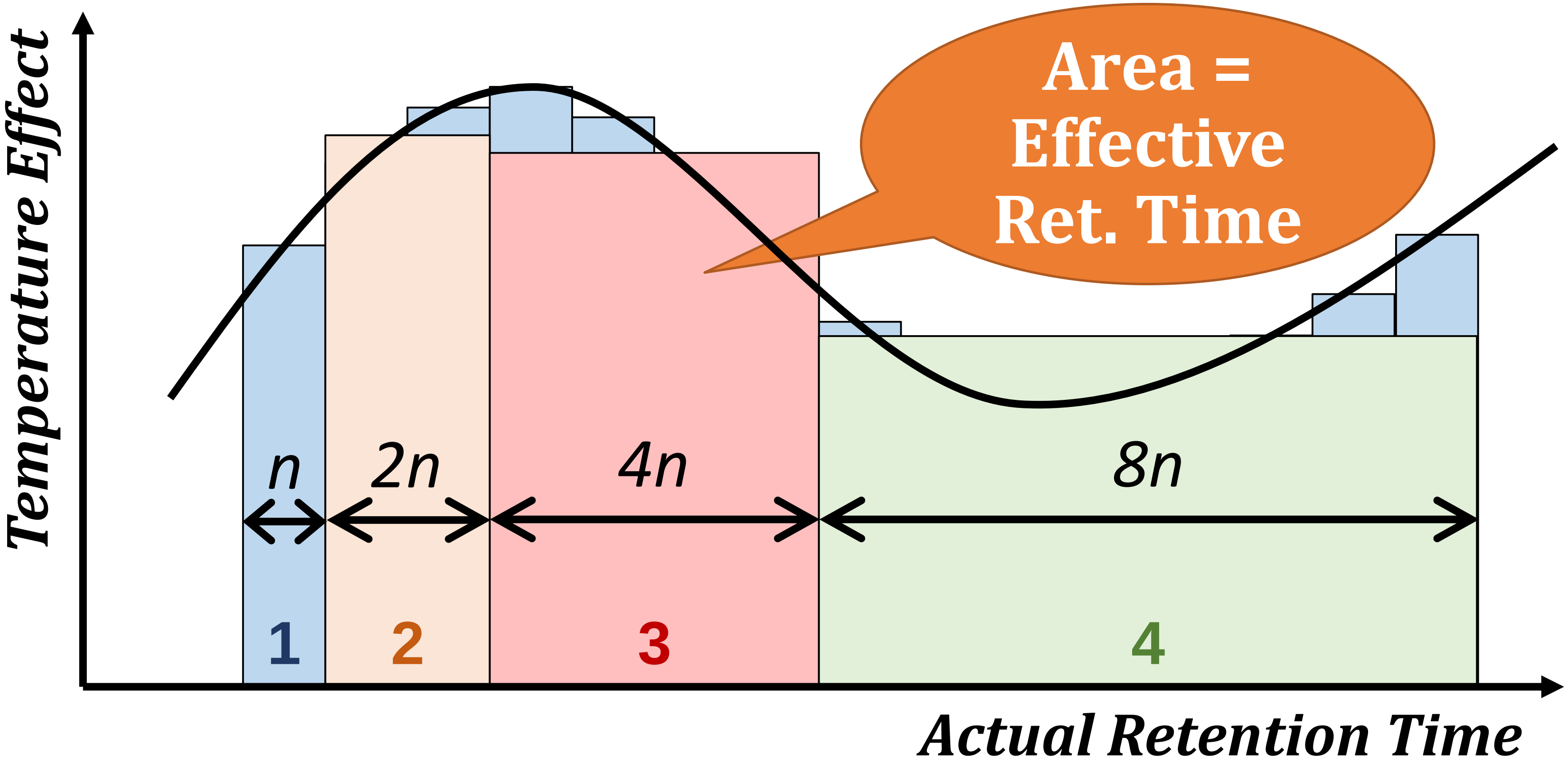
Storage Overhead:

- Tracking SSD Temperature
  - 26 logarithmic intervals
  - 208 B
- Program temperature, dwell time, program time per block
  - 1.5 MB
- Dwell time
  - Timestamp for last 20 full drive writes
  - 85 B

- Latency Overhead:
  - <1% of flash read latency (25 us)

HeatWatch: Tracking Components

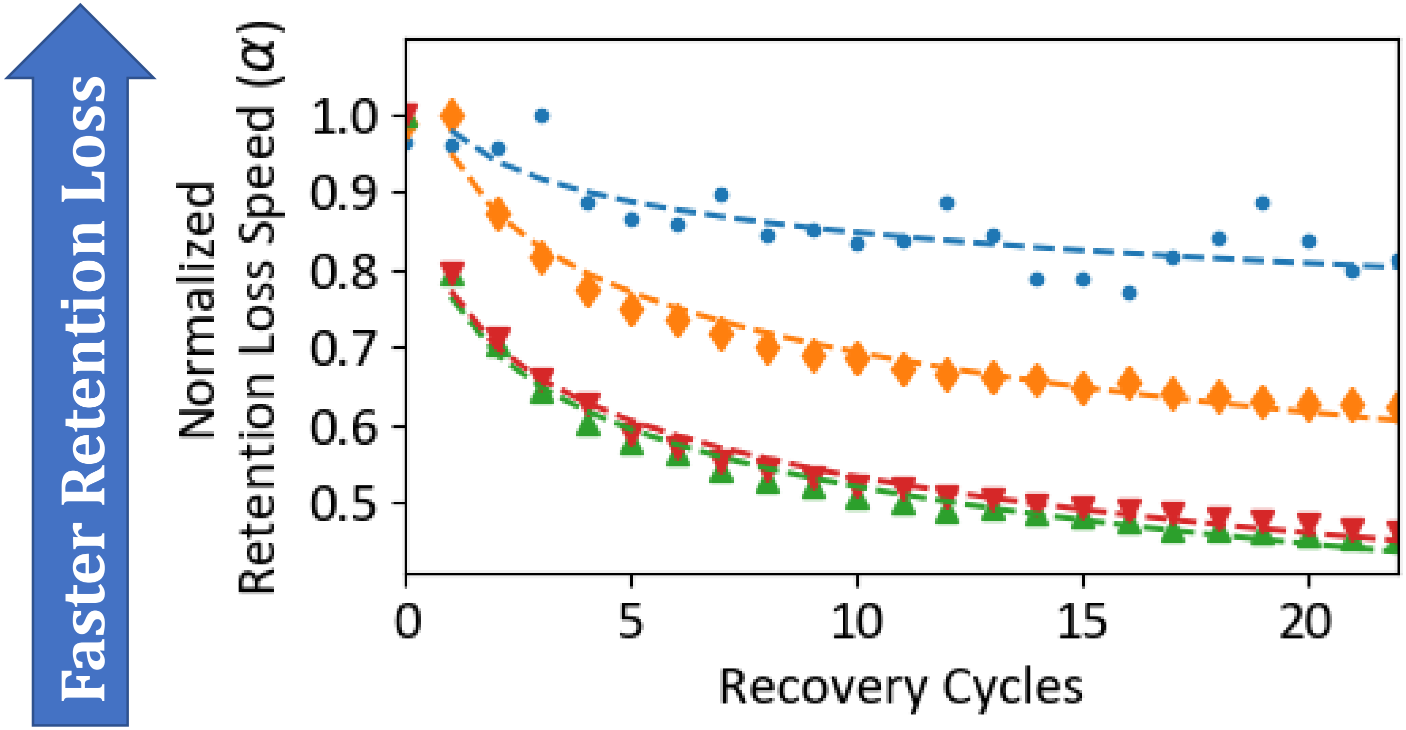
1. Tracking SSD temperature
   - *Use existing sensors in the SSD*
   - ***Precompute*** *temperature scaling factor at* ***logarithmic time intervals***



HeatWatch: Tracking Components

2. Tracking dwell time

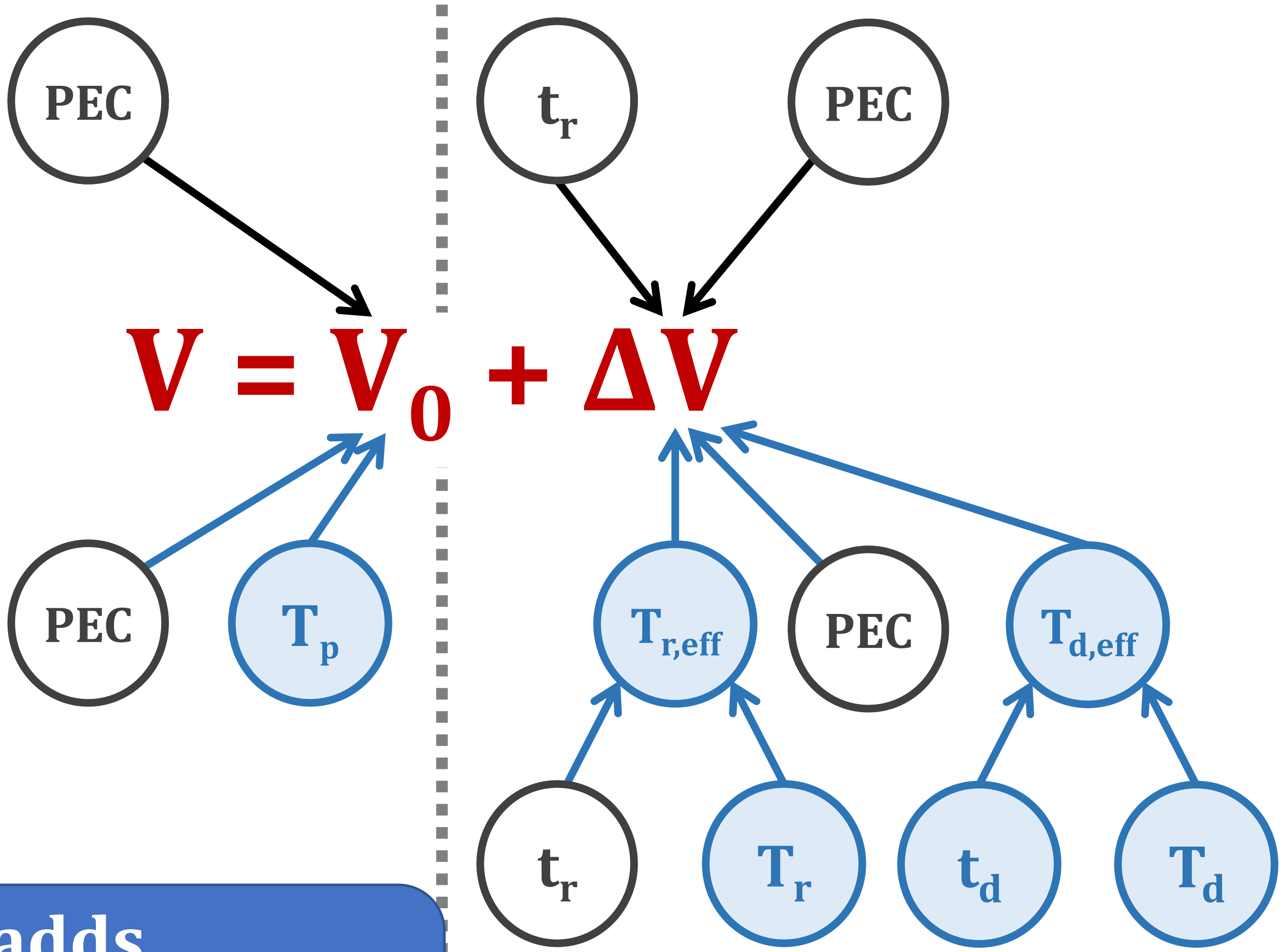
- *Only need to track write frequency* ***for last 20 P/E cycles***



Self-recovery effect plateaus after 20 P/E cycles

URT vs. Conventional Model

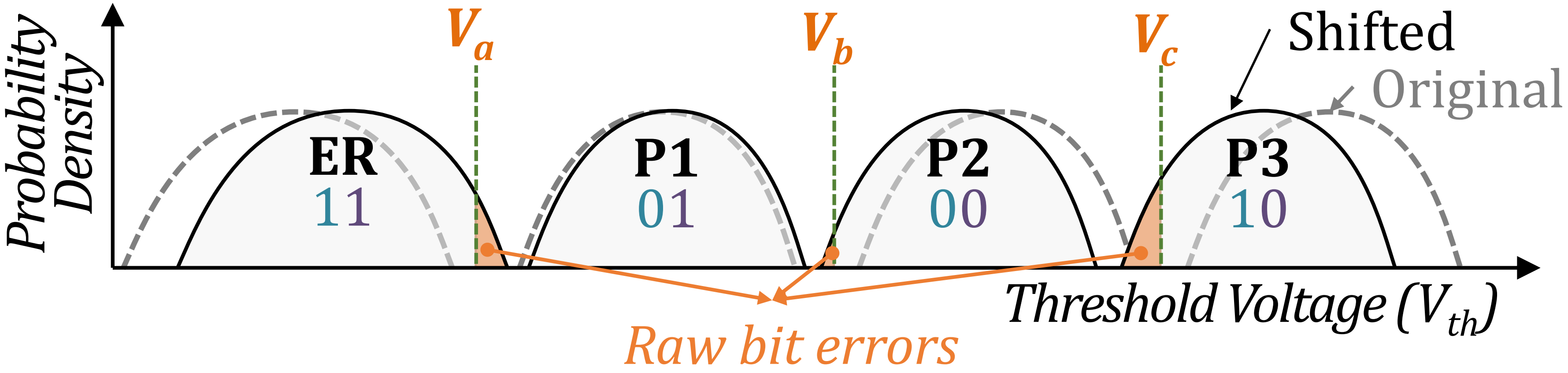
Conventional



URT

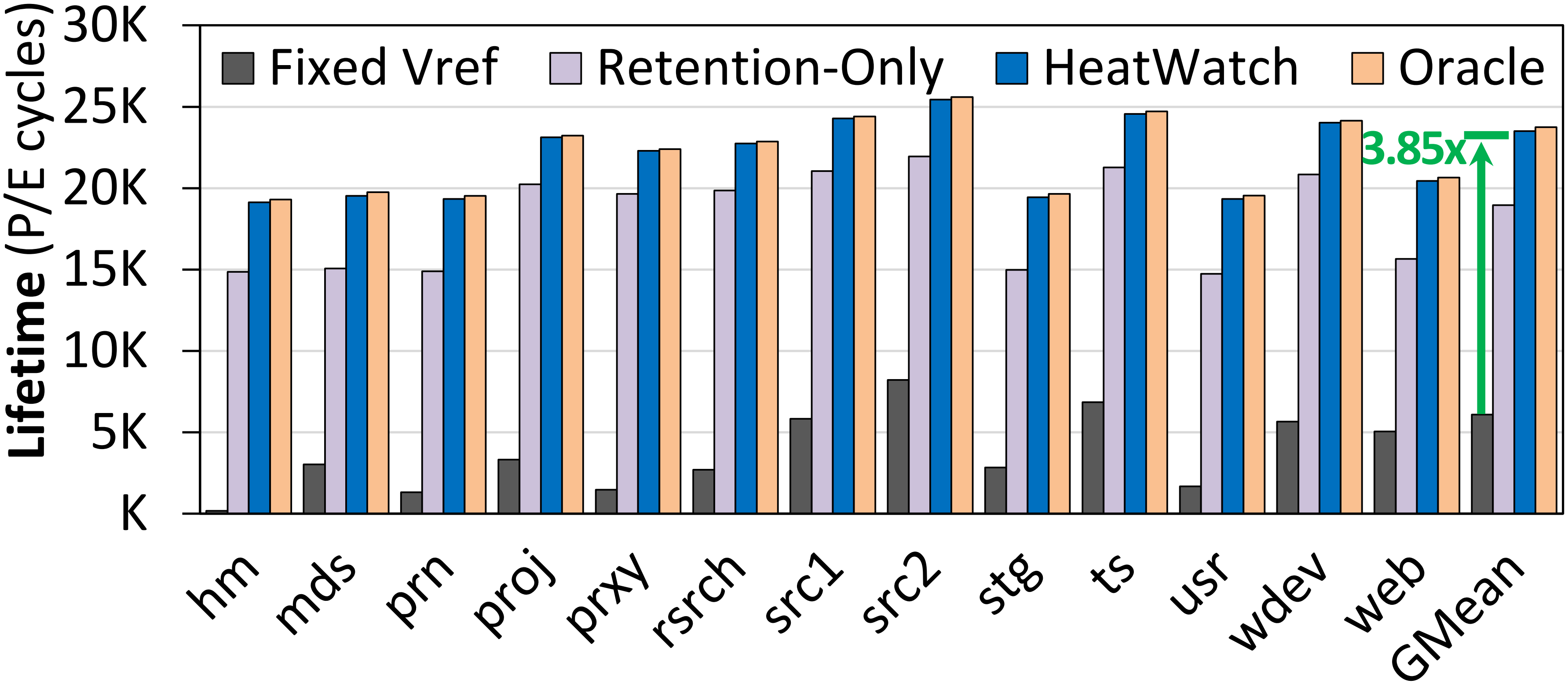
URT adds self-recovery, temperature to conventional model

Threshold Voltage Distribution Shifts

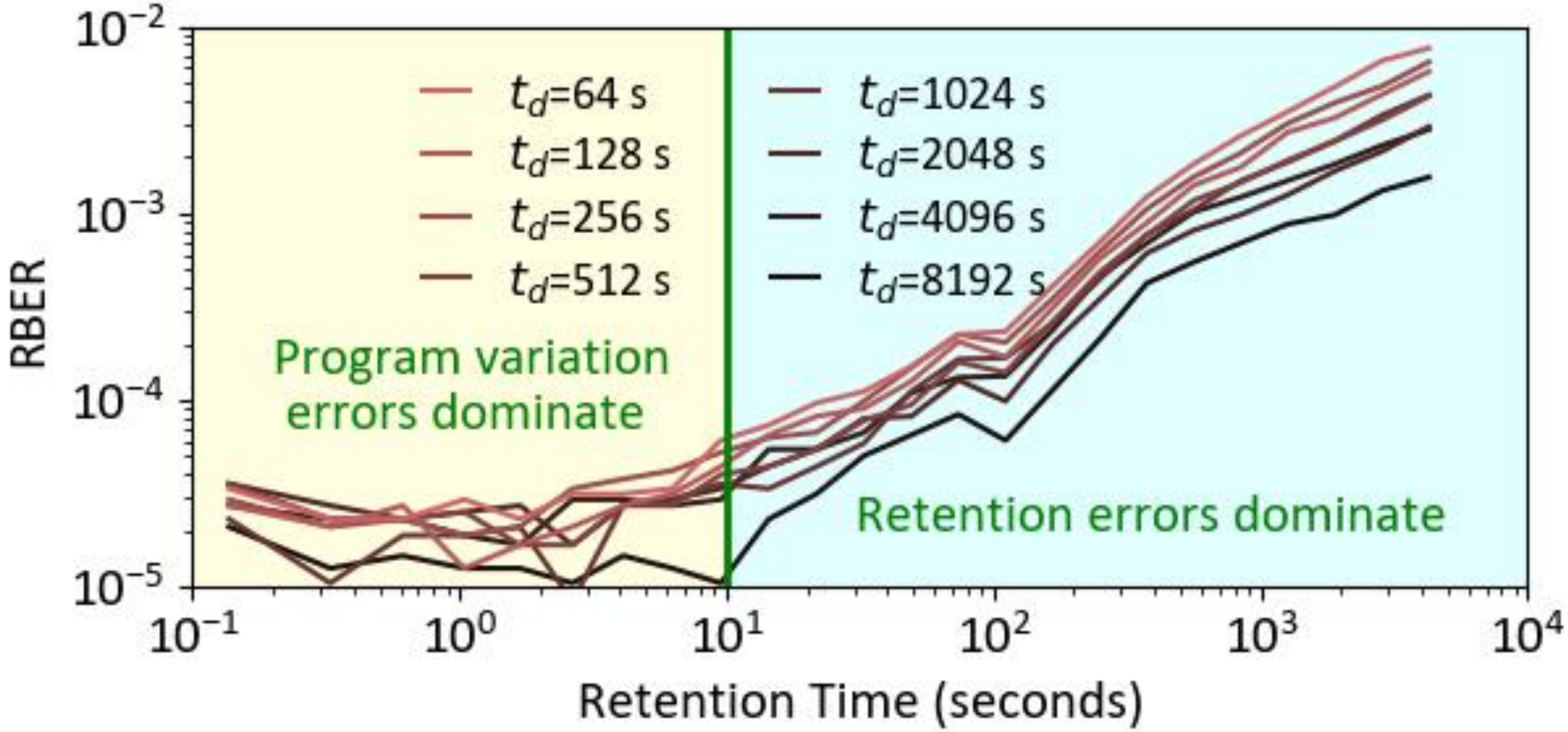


- Shifts occur over time due to multiple factors (e.g., retention)
- Can cause distribution of one state to cross over the read reference voltage boundary
  - Some cells get misread
  - Introduces raw bit errors

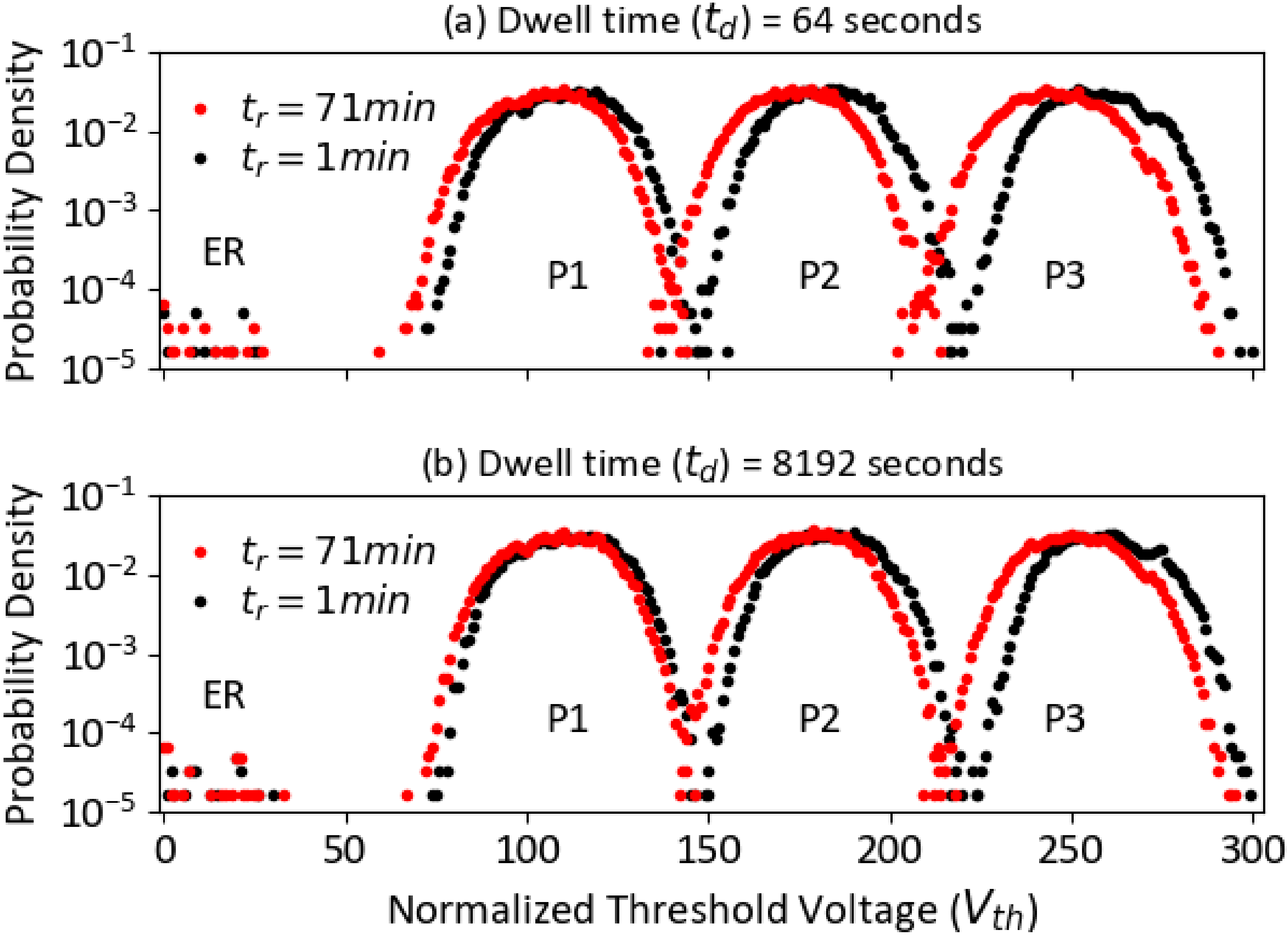
Per-Workload Flash Lifetime Improvements



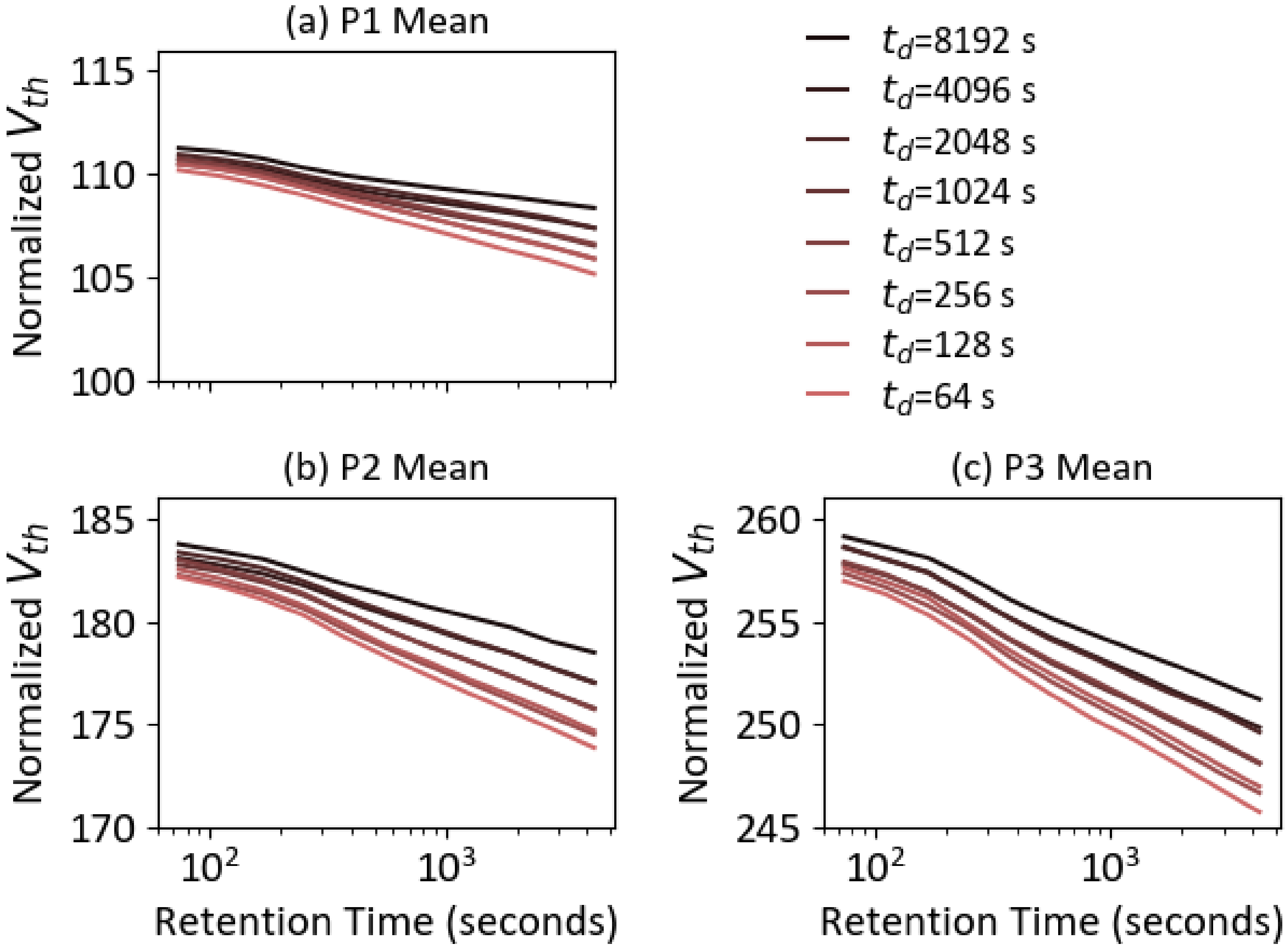
Dwell Time Impact on Error Rate After Retention



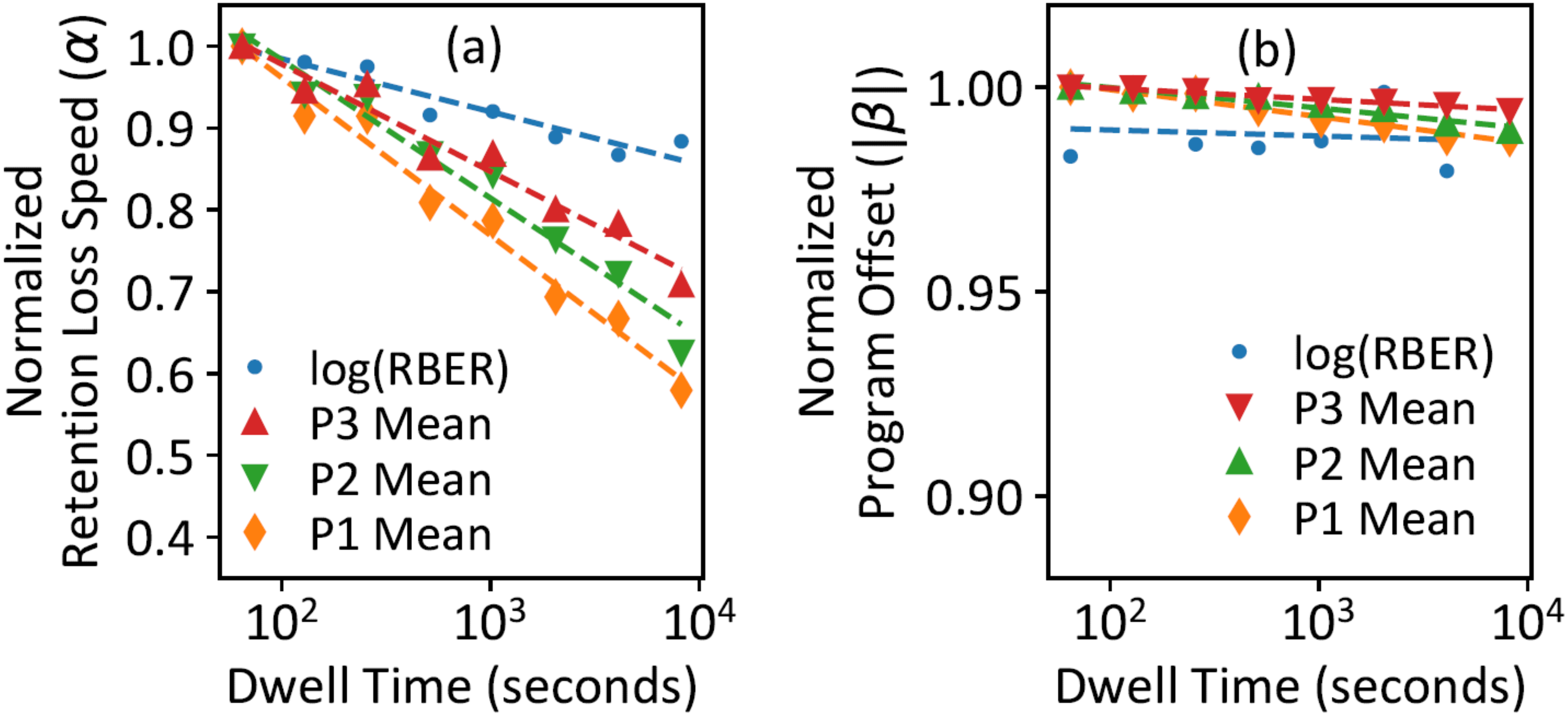
Dwell Time Impact on Threshold Voltage Distributions



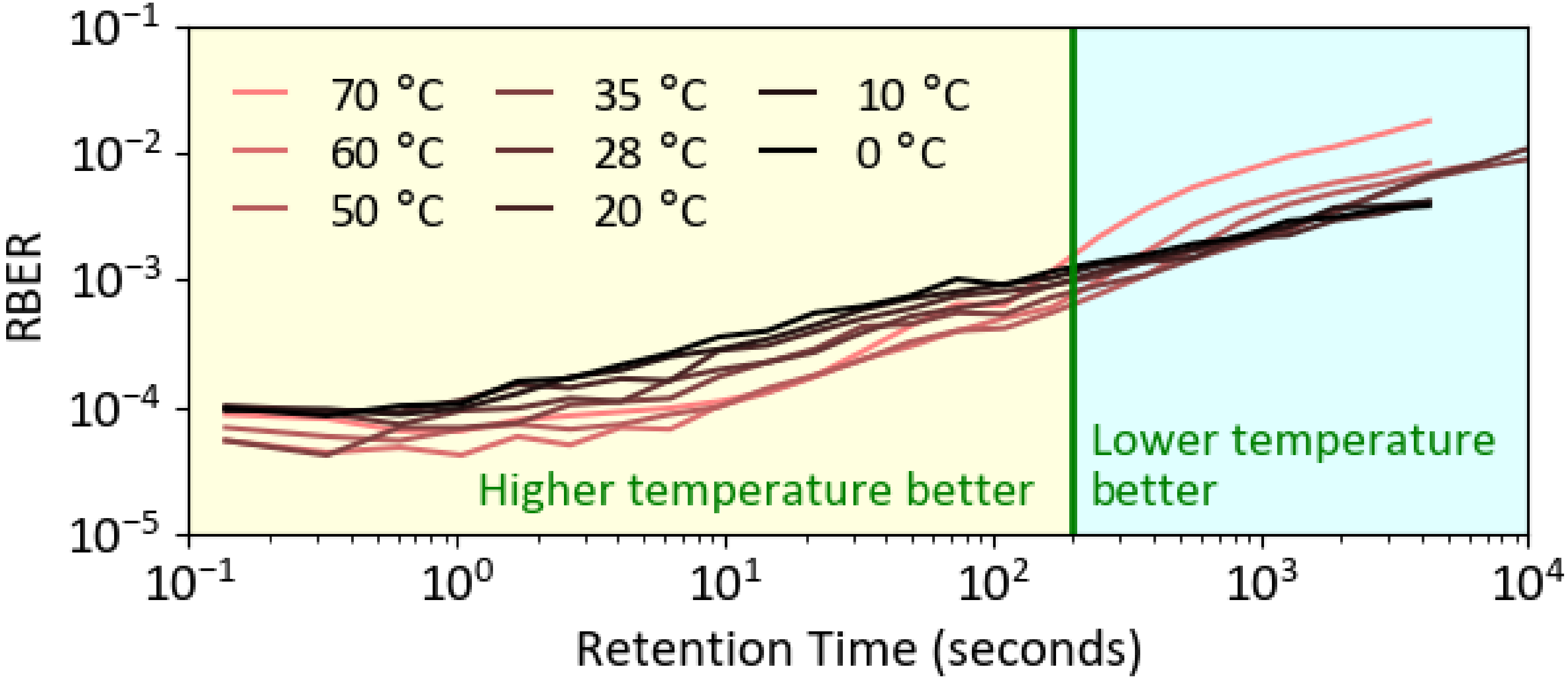
Mean Distribution Voltage vs. Retention for Different Dwell Times



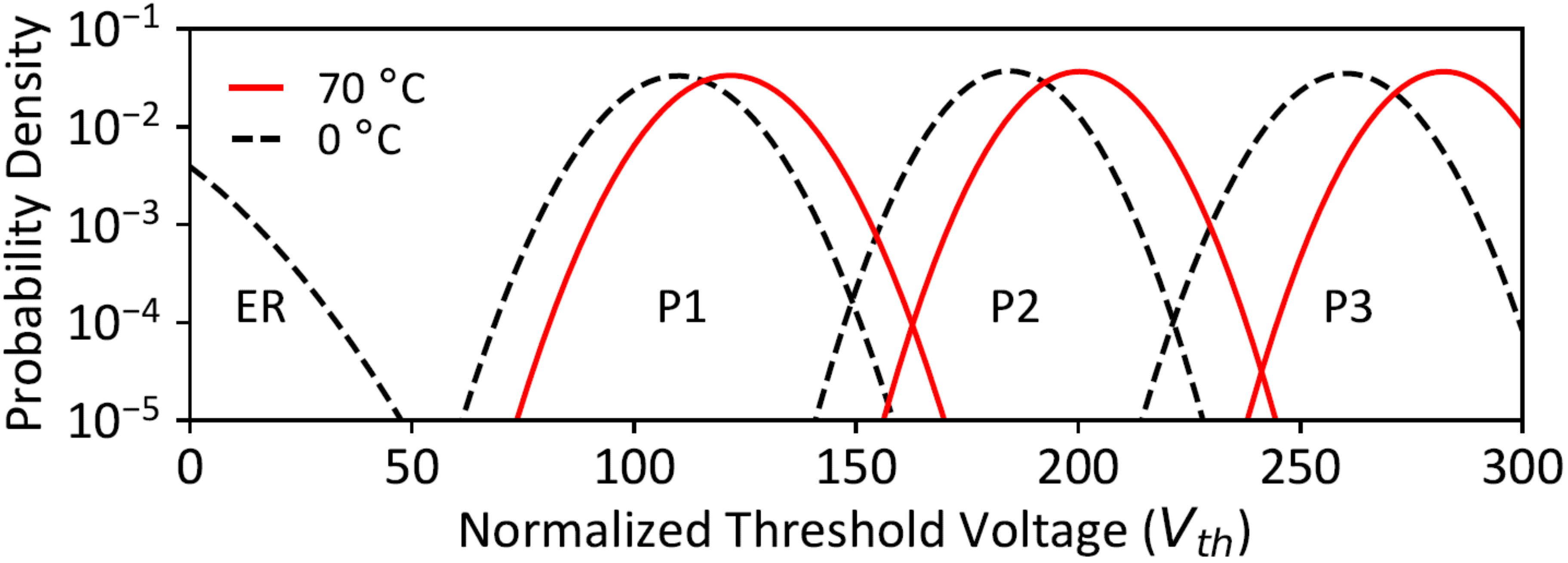
Impact of Dwell Time on Error Rate and Threshold Voltage Distribution Means



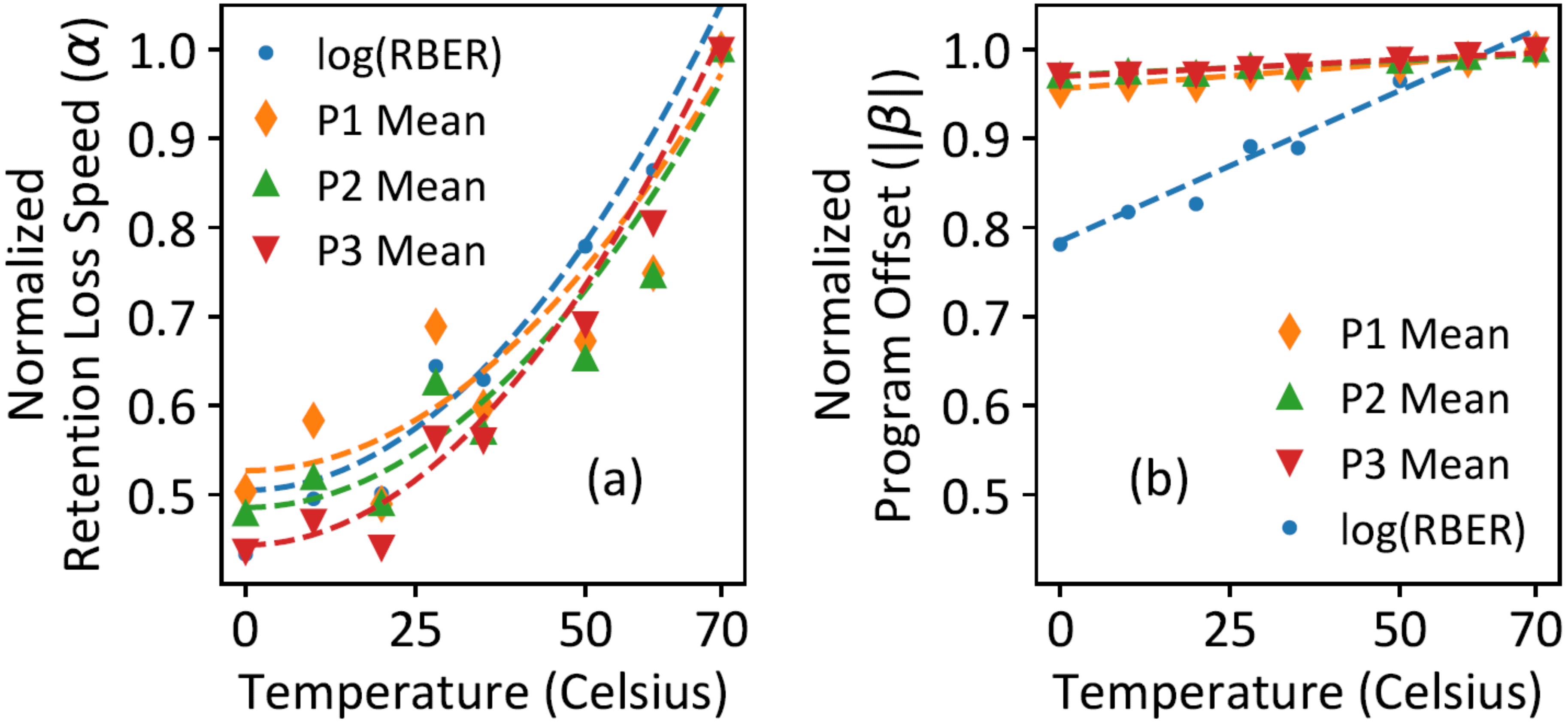
Temperature Impact on Error Rate After Retention



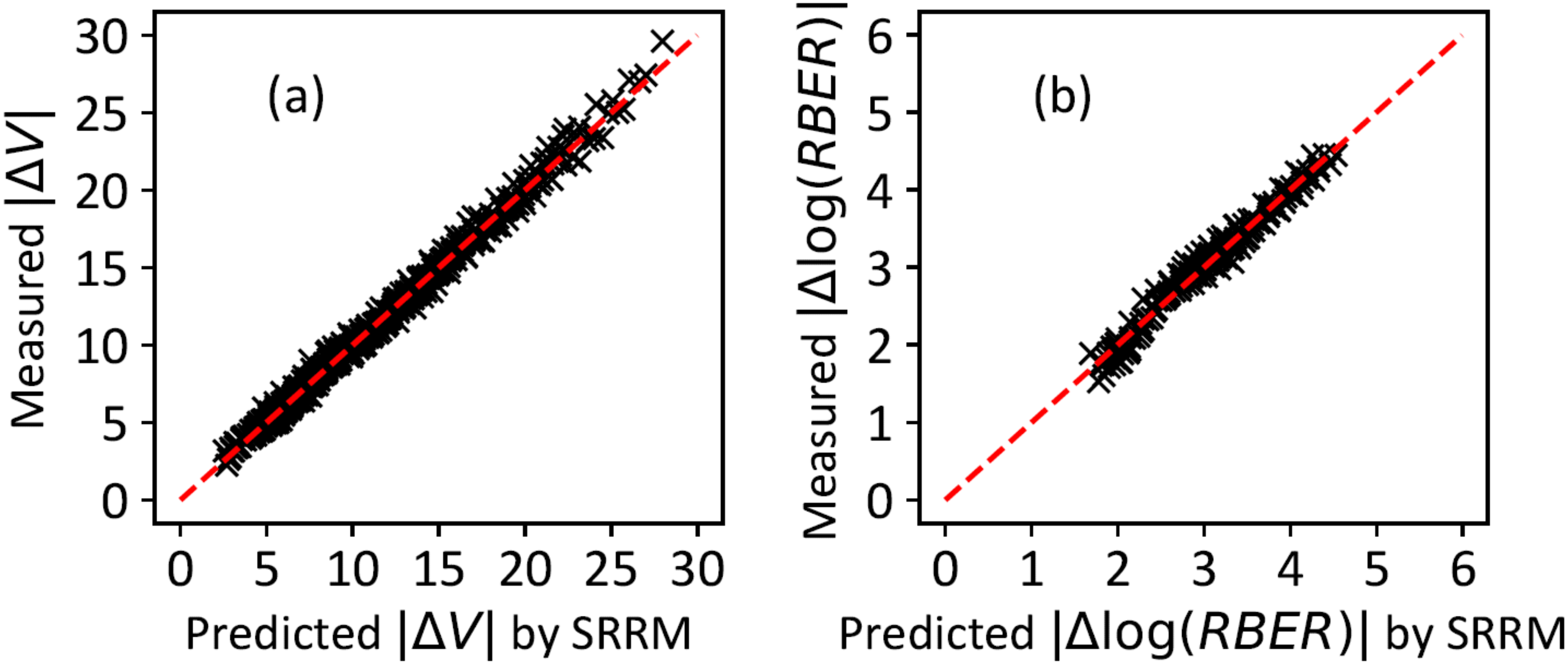
Impact of Programming Temperature on Threshold Voltage Distributions



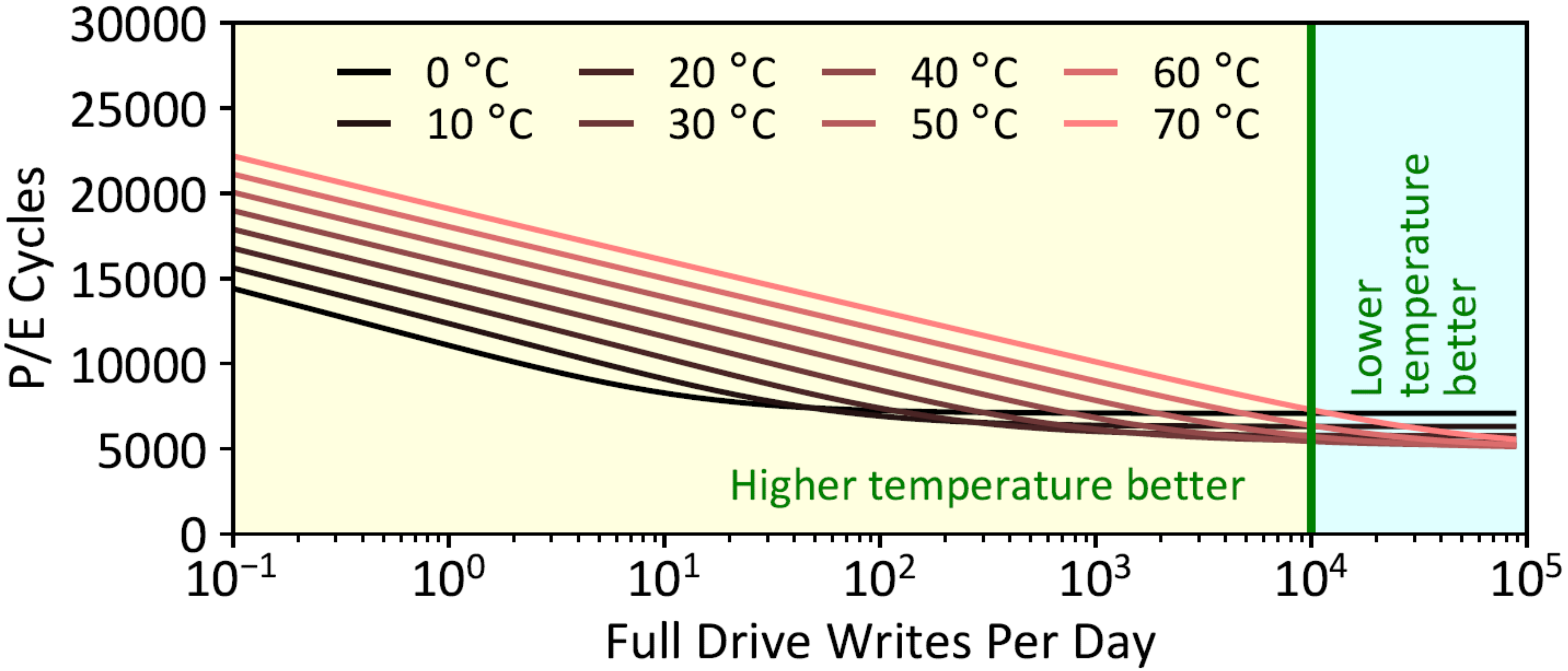
Impact of Programming Temperature on Error Rate and Threshold Voltage Distribution Means



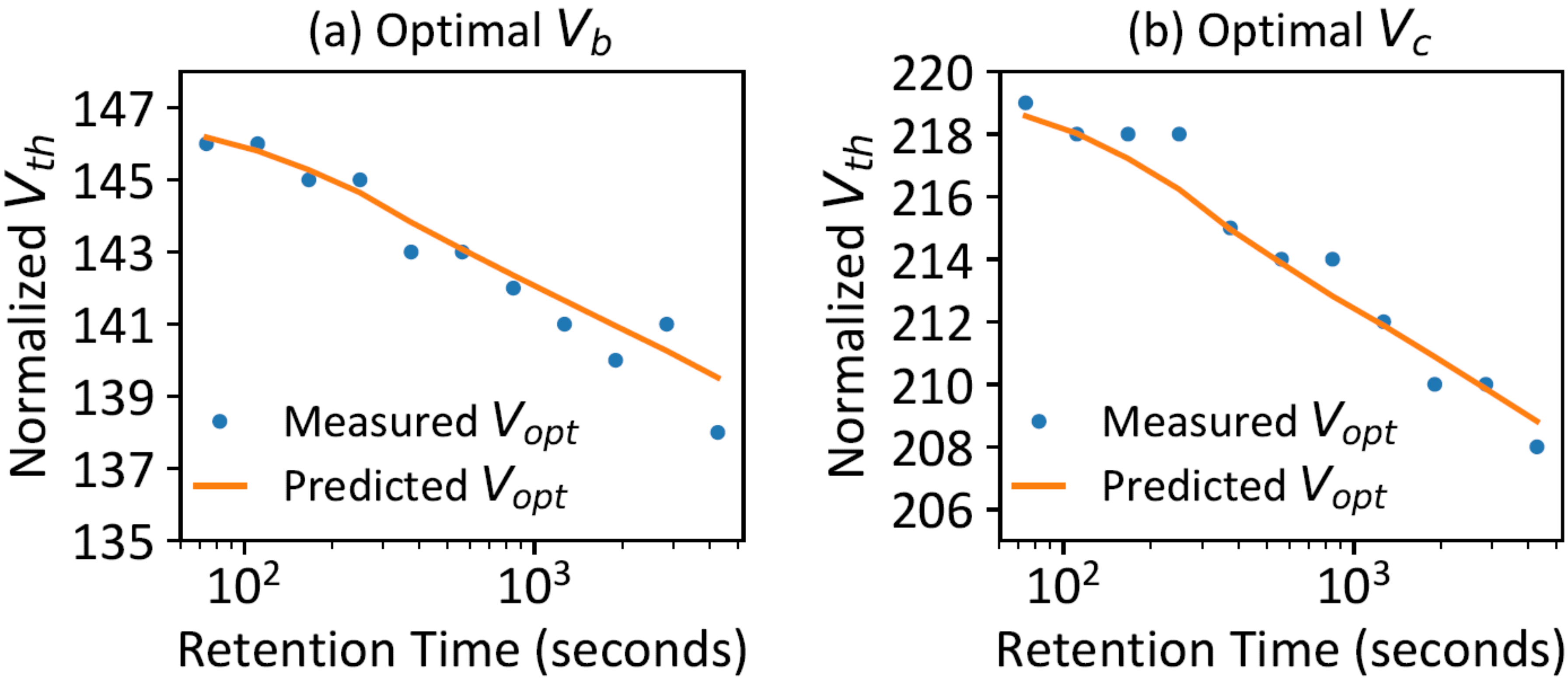
SRRM Prediction Accuracy



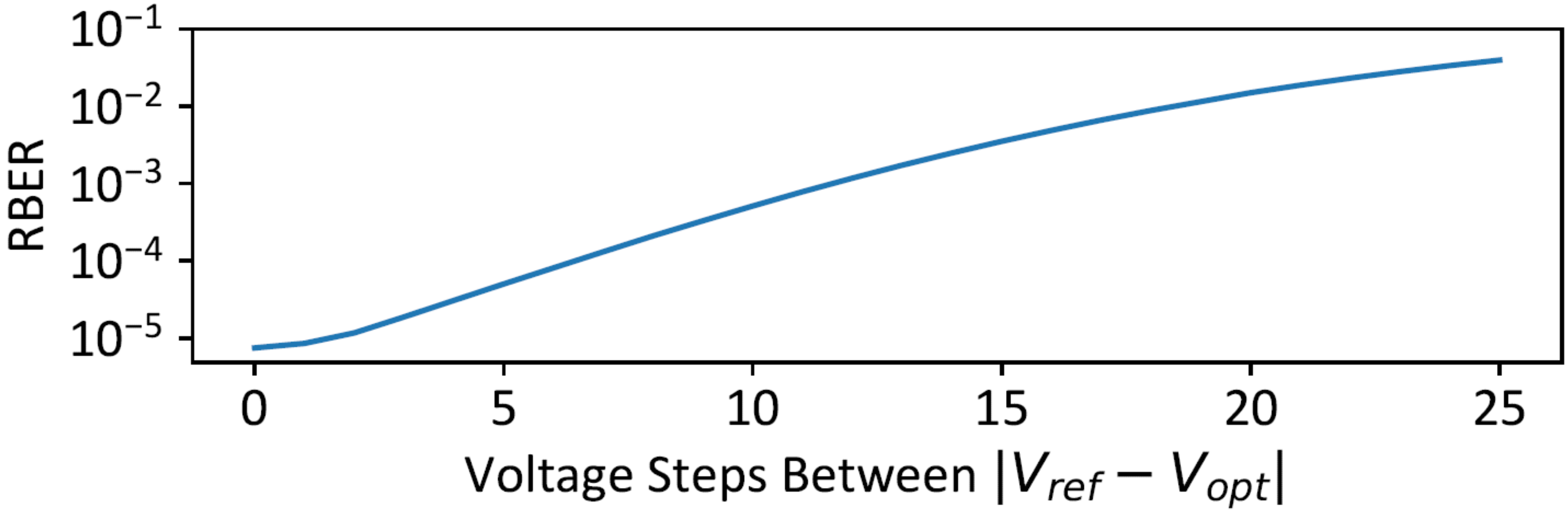
Change in Flash Lifetime Due to Programming Temperature and Write Intensity



Optimal Read Reference Voltage: Measured vs. Predicted by URT



Inaccurate Read Reference Voltages Increase Error Rate



10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
RBER
0
5
10
15
20
25
Voltage Steps Between $|V_{ref} - V_{opt}|$