



Flash Memory Summit

# Advanced NAND Management Using Machine Learning

Ramyakanth Edupuganti, Applications Engineering  
Microchip Technology Inc.

# Agenda



Flash Memory Summit

- Advanced NAND Management Challenges
- Valley Search Usage Model
- Example Valley Search Usage Model for NAND
- Using Machine Learning Engine for Vt Tuning
- MLE – Data Collection and Training Phase
- NAND Bit Error Rate (BER) Analysis

# Agenda



Flash Memory Summit

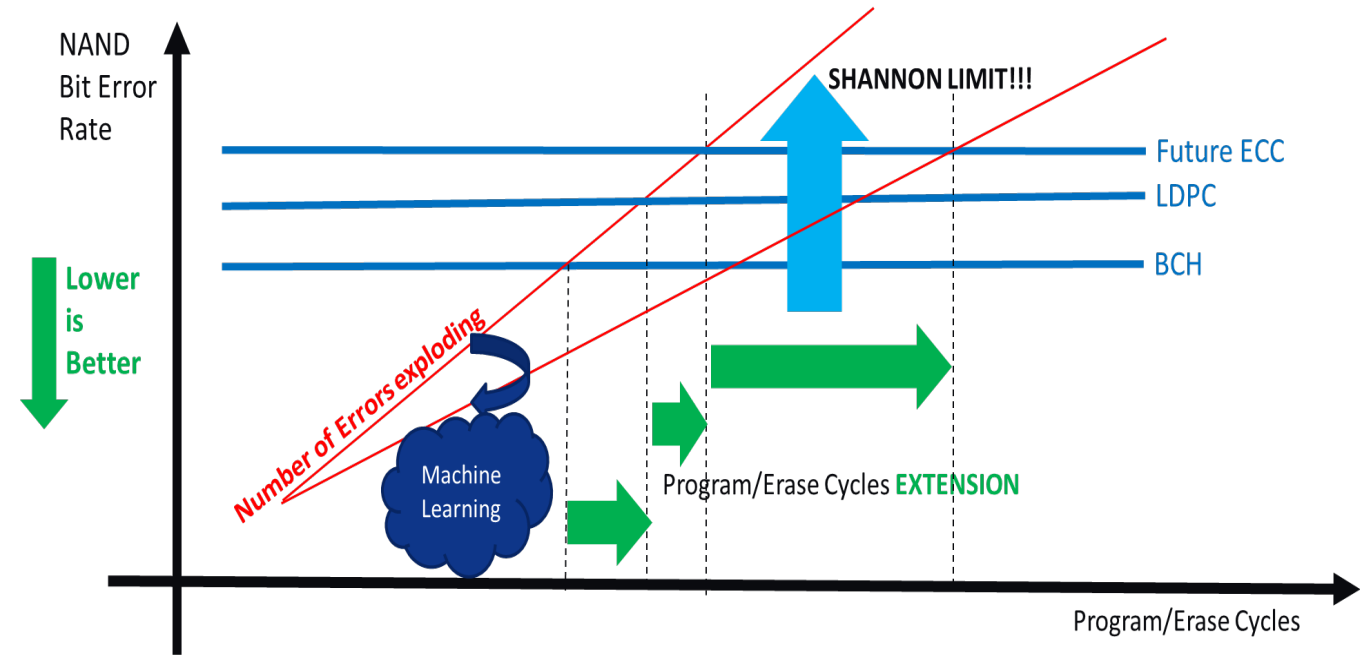
- Advanced NAND Management Challenges
- Valley Search Usage Model
- Example Valley Search Usage Model for NAND
- Using MLE for Vt Tuning
- MLE – Data Collection and Training Phase
- NAND BER Analysis

# Advanced NAND Management Challenges

As NAND technology advances, effective NAND management is becoming challenging:

- Layer count increasing
- Number of bits/cell increasing

Need to improve error correction and develop techniques to reduce errors to take advantage of newer, denser, and cost-efficient NAND

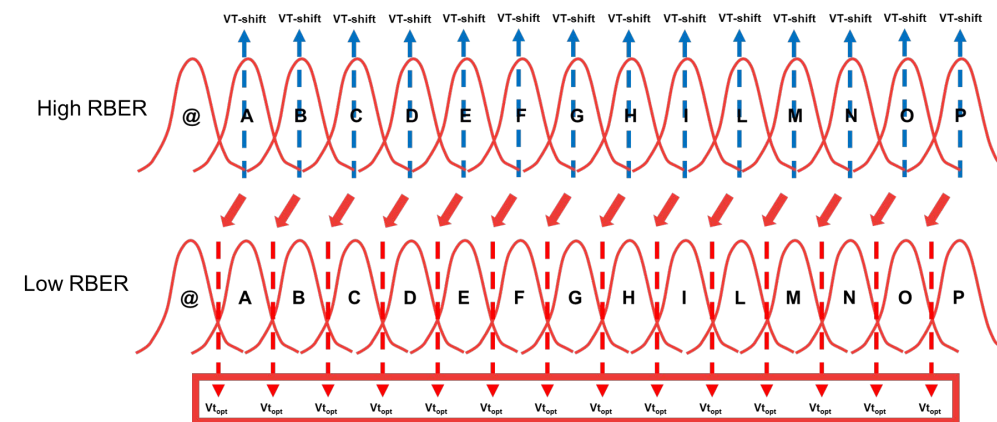
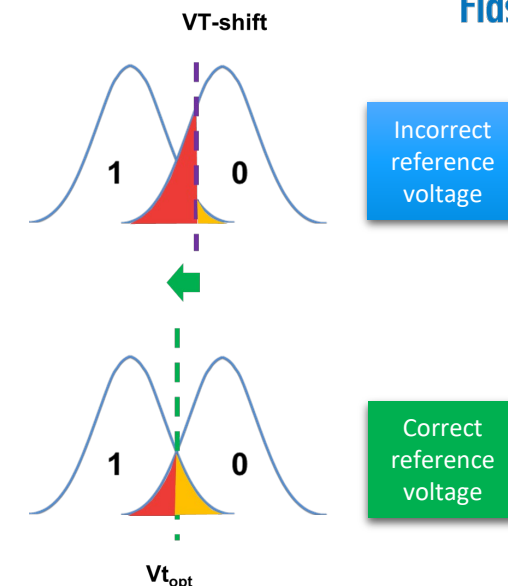


# Increased NAND Layer Count



Flash Memory Summit

- Layer count increase means Bit Error Rate (BER) will increase
  - Most bit read errors are unrelated to physical damage
  - Read errors mostly due to tighter Vt distributions in denser cells
  - Need to re-try the read with the correct reference voltage
  - A QLC-NAND (4 bits per cell) may require up to 15 Vt reference voltages for 16 states to retry!
- Vt Shift and Read-Retry based algorithms exist for error recovery Require extensive characterization to optimize
  - Trial-and-error based, less precise
  - Impact read bandwidth with background read overhead
  - Very challenging to optimize – the Vt distribution can change between every layer
  - NAND is at 128L today and growing!
- Need a solution to optimize the Vt reference voltage and reduce correctable errors to improve read QoS



# Increased Number of Bits per Cell



Flash Memory Summit

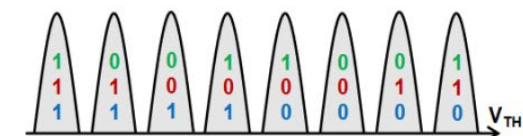
- Bits per cell increase means cell reliability will degrade faster
  - Errors occur when a 'worn out' block is accessed, and FW ECC fails to correct the block
  - Need to initiate soft correction and/or RAID operations to recover the user data
- NAND Block Program/Erase counters and retention times are some of the metrics used by FW to track NAND reliability state
  - FW structures don't recognize if a block is "more sensitive" and wears out faster than expected – Reactive only
  - All reliability states must be rebuilt if FW structures are corrupted before NAND access is enabled
  - Error correction overhead can impact bandwidth, latency and QoS
- Need a solution to measure and track the 'effective' reliability of each NAND block to optimize accesses with minimal error corrections requirements



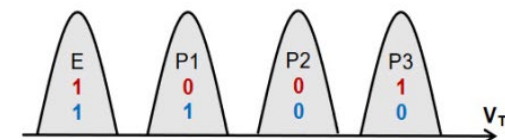
## Data representation in NAND flash



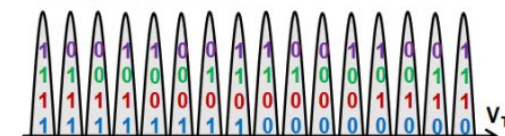
**Single Level Cell (SLC):** 2 States (1 Erase + 1 Pgm)  
= 1 bit of information per cell



**Triple Level Cell (TLC):** 8 States (1 Erase + 7 Pgm)  
= 3 bits of information per cell  
= 1.5x capacity of MLC



**Multi Level Cell (MLC):** 4 States (1 Erase + 3 Pgm)  
= 2 bits of information per cell  
= 2x capacity of SLC



**Quad Level Cell (QLC):** 16 States (1 Erase + 15 Pgm)  
= 4 bits of information per cell  
= 1.3x capacity of TLC

A Penta Level Cell (PLC)  
packs up to 5 bits per cell!

# Vt Reference Tuning Comparison with Existing Solutions

SOLUTION	DESCRIPTION	NAND CHAR EFFORT	QOS IMPACT	PRECISION	SCALABILITY (QLC, PLC, High Layer Count)
Read Retry	Defined by Flash vendors	High	High	Low	Low
Vt-shift	Custom Reference Voltages	High	High	Low	Low
BRP-like	Background Reference Positioning	Med	0	Good	Poor because of the background reads
Neural Network	VT Reference Tuning with Machine Learning	Med	0	Excellent	Excellent

# Agenda



Flash Memory Summit

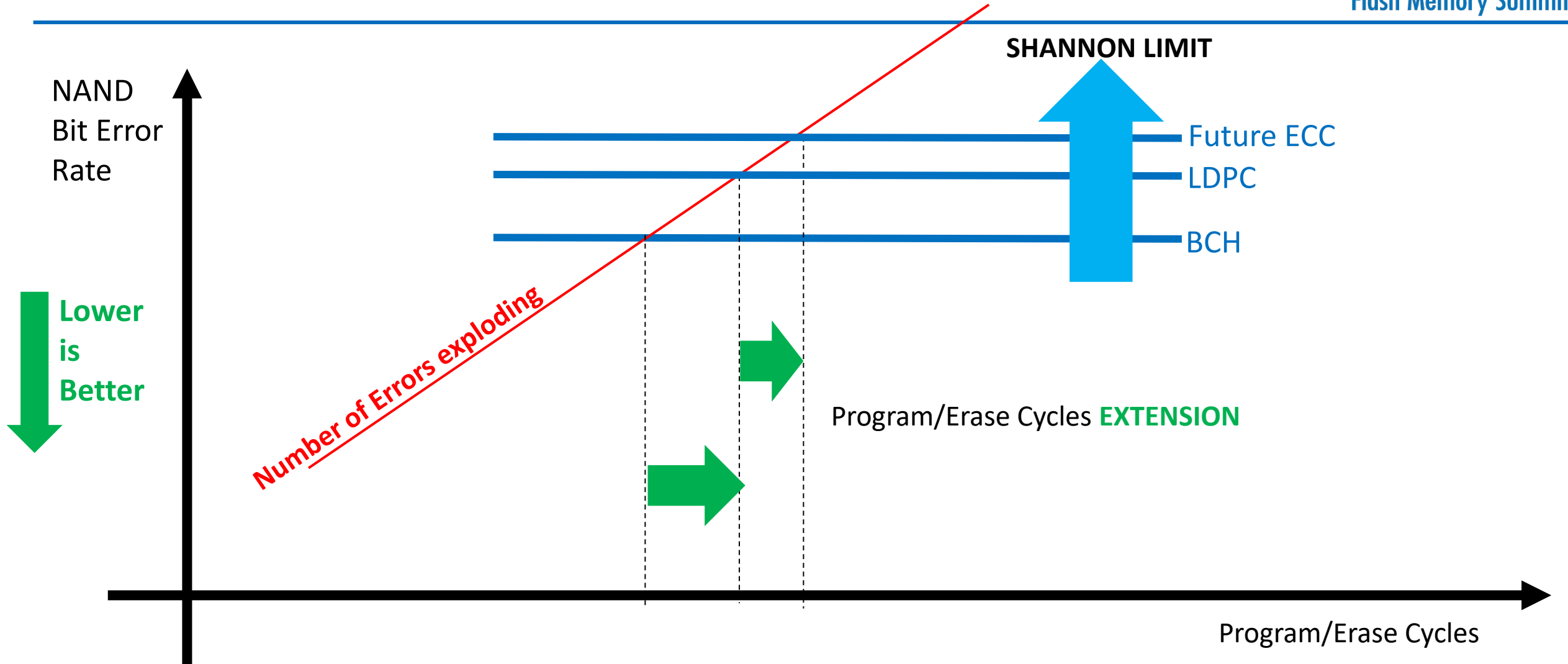
- Advanced NAND Management Challenges
- Valley Search Usage Model
- Example Valley Search Usage Model for NAND
- Using MLE for Vt Tuning
- MLE – Data Collection and Training Phase
- NAND BER Analysis



# How to Address the NAND BER Increase? ECC

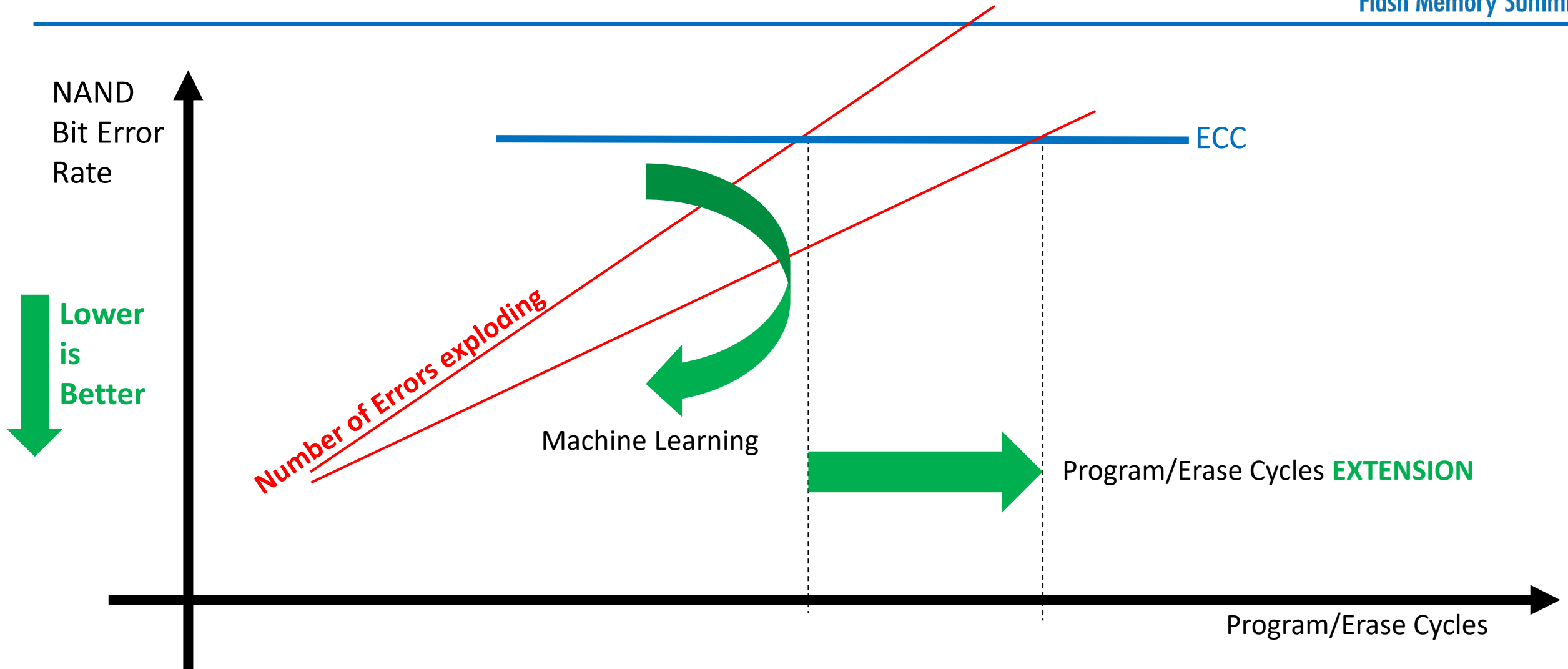


Flash Memory Summit





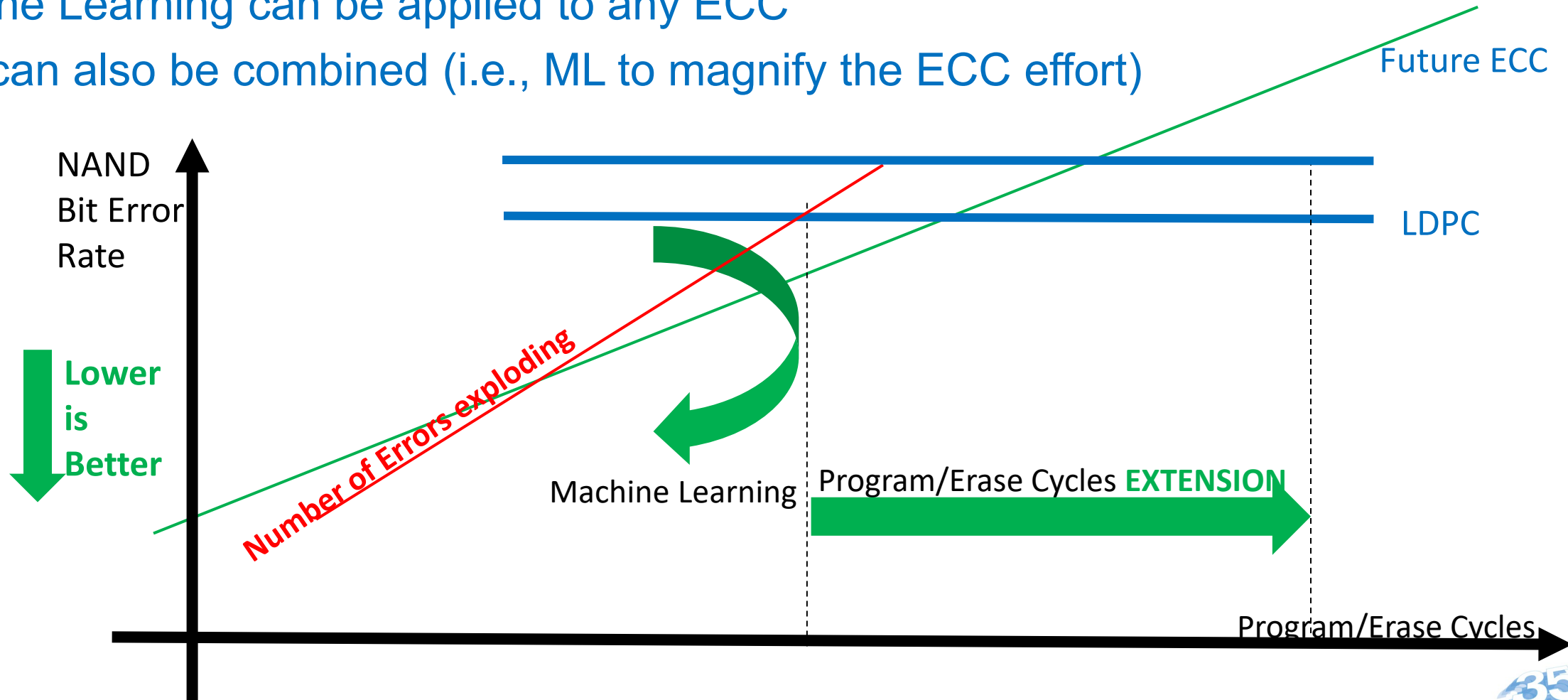
# How to Address the NAND BER Increase? Machine Learning





# Considerations

- Machine Learning and ECC are orthogonal to each other
- Machine Learning can be applied to any ECC
- They can also be combined (i.e., ML to magnify the ECC effort)



# Agenda



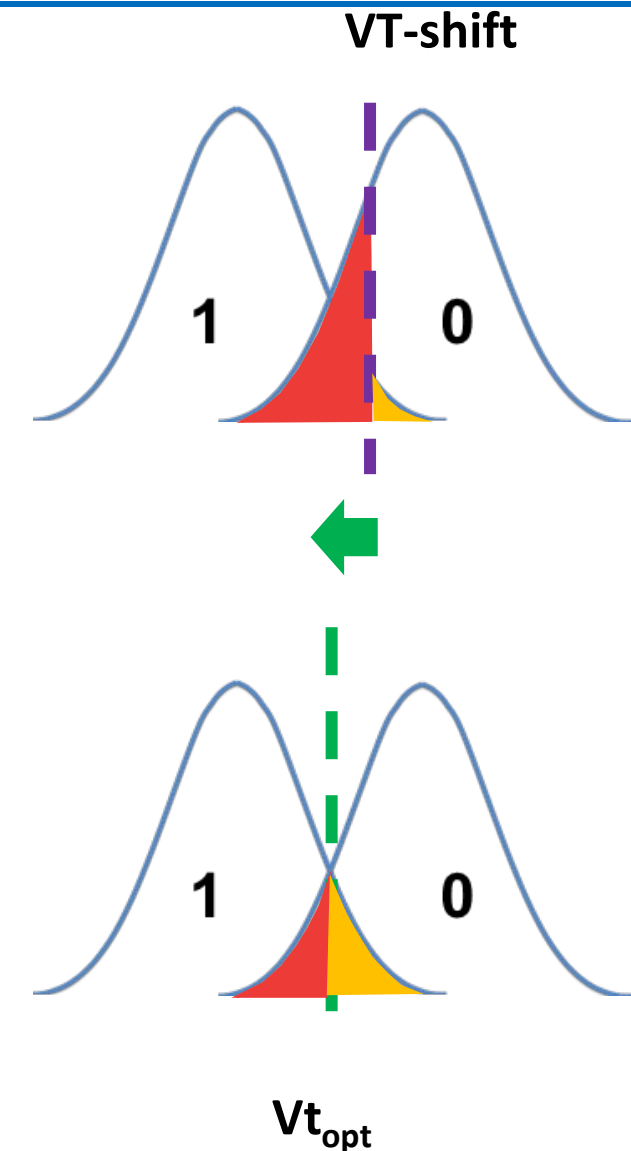
Flash Memory Summit

- Advanced NAND Management Challenges
- Valley Search Usage Model
- **Example Valley Search Usage Model for NAND**
- Using MLE for Vt Tuning
- MLE – Data Collection and Training Phase
- NAND BER Analysis



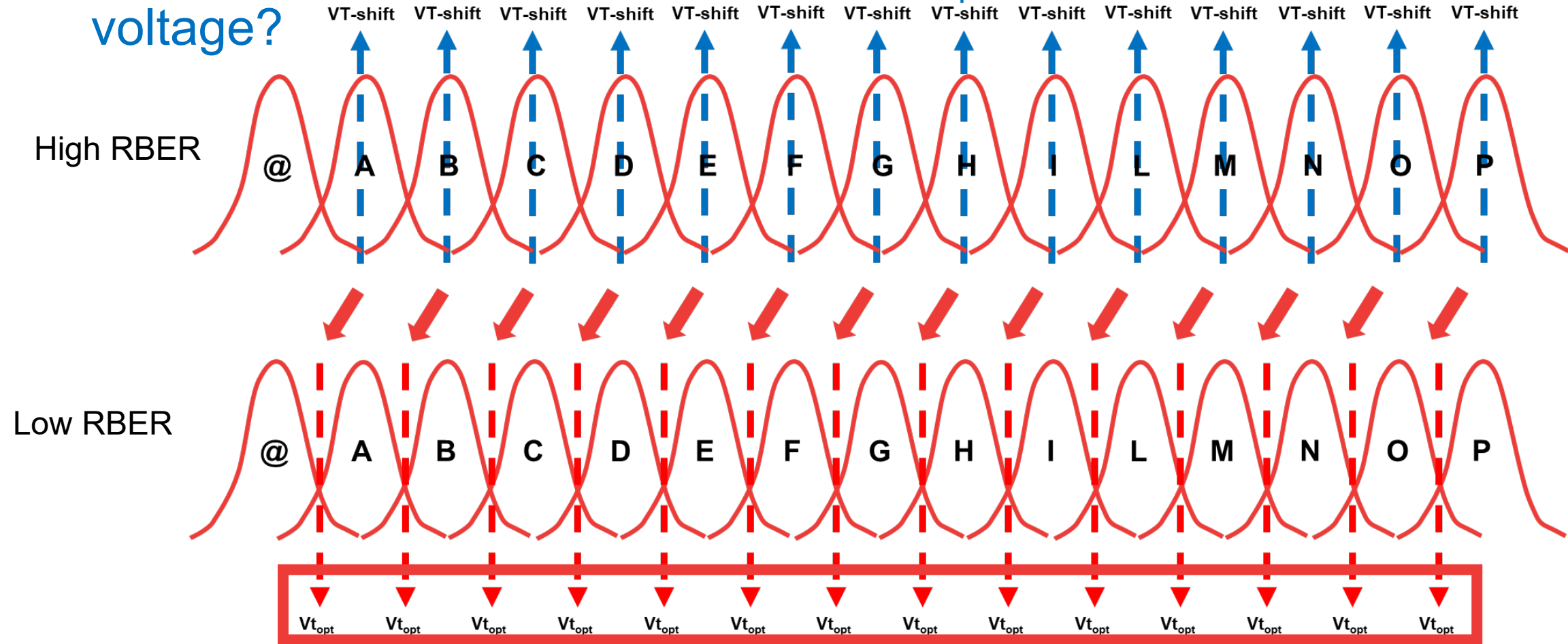
# NAND RBER vs. Read Reference Voltage

- NAND raw BER is a function of PE, time, read disturb
- In 3D NAND most of the errors are recoverable; in other words, they are not related to a physical damage
- To minimize the RBER it is necessary to “center” the reference voltage
- Min RBER is when all the Reference Voltages (7 for TLC, 15 for QLC, 31 for PLC) are in the corresponding  $V_{t_{opt}}$



# Problem Statement

- How to predict the best placement ( $V_{t_{opt}}$ ) for each NAND reference voltage?



# Agenda



Flash Memory Summit

- Advanced NAND Management Challenges
- Valley Search Usage Model
- Example Valley Search Usage Model for NAND
- Using MLE for Vt Tuning
- MLE – Data Collection and Training Phase
- NAND BER Analysis

# What is Vt Shift Algorithm ?



Flash Memory Summit

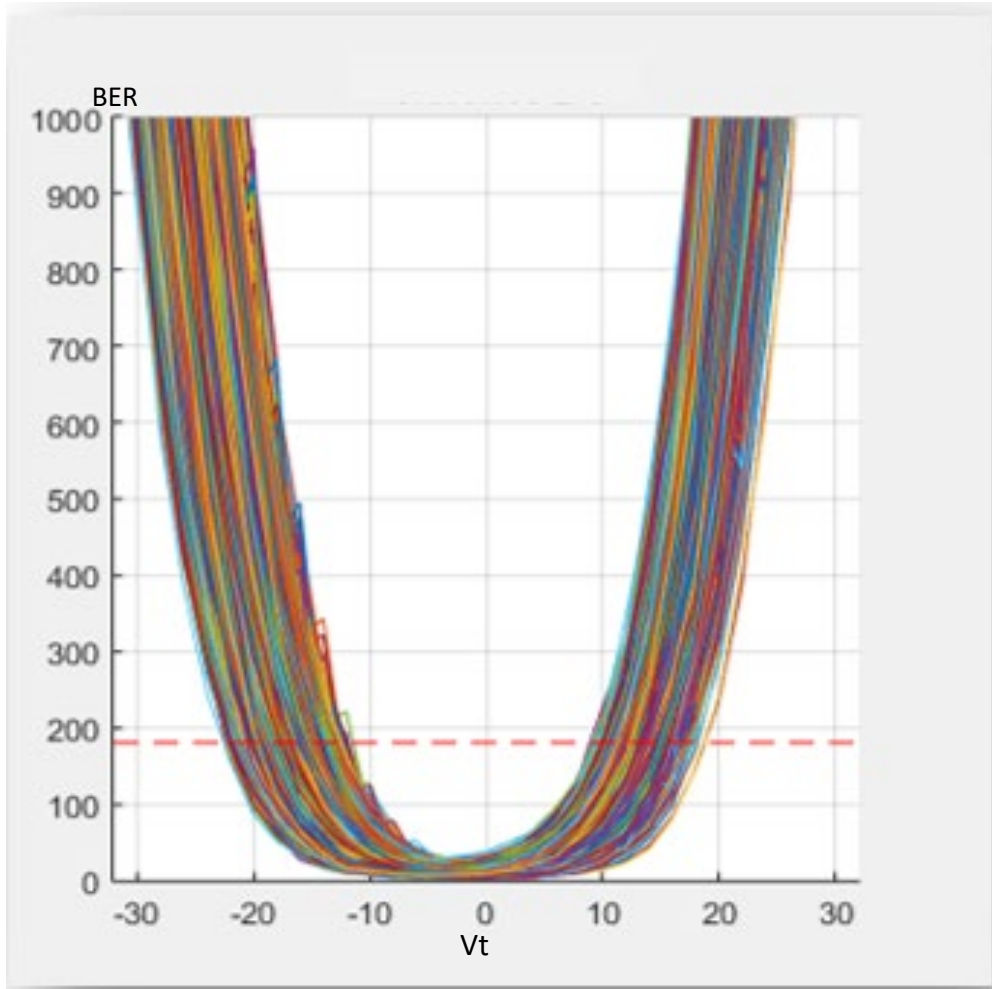
- Vt shift algorithm uses small number of samples – error data(from 3 VTs) to rebuild continuous valley (9 VTs) and find lowest point of the valley that corresponds to Vt with smallest number of errors
- Vt shift utilizes interaction between NAND management SW and MLE HW
- MLE HW is based on Neural Network and requires pre-trained Neural Network (NN) models for operation
- Reliability states are defined based on Endurance, Retention and Read disturb parameters



# What is Vt Shift Algorithm ?



Flash Memory Summit



- **Training phase**

- Pick 3 samples below the error threshold
- The training is done on the target of (for example) 9 VTs
- After the training you get the model

- **Inference phase**

- Read the 3 VT samples, use the model, obtain the bottom of the parabola
- From these numbers calculate the minimum values
- The output are the minimum related VT and the second minimum VT

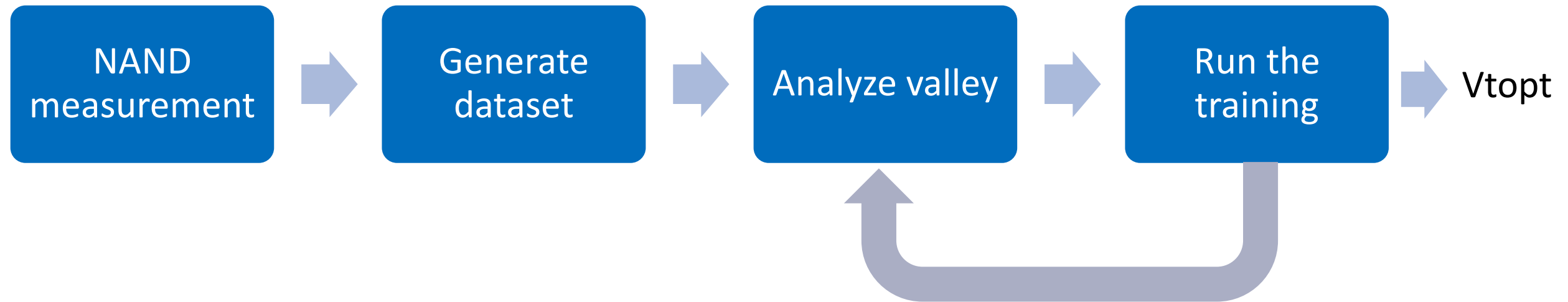
# Agenda



Flash Memory Summit

- Advanced NAND Management Challenges
- Valley Search Usage Model
- Example Valley Search Usage Model for NAND
- Using MLE for Vt Tuning
- **MLE – Data Collection and Training Phase**
- NAND BER Analysis

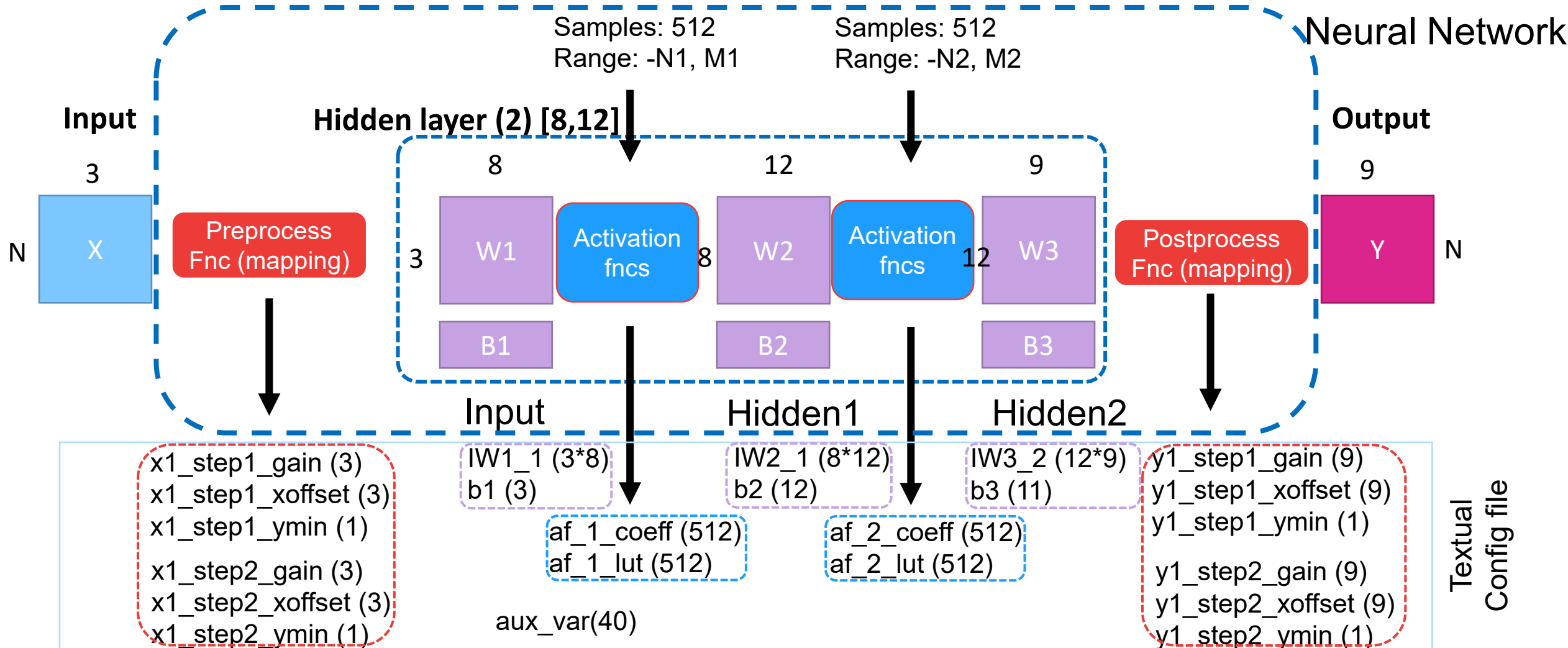
# Data Collection and Training Phase Steps



# Training Results



Flash Memory Summit



# Agenda



Flash Memory Summit

- Advanced NAND Management Challenges
- Valley Search Usage Model
- Example Valley Search Usage Model for NAND
- Using MLE for Vt Tuning
- MLE – Data Collection and Training Phase
- **NAND BER Analysis**

# “Classic” Flash Management (BRP-Like)

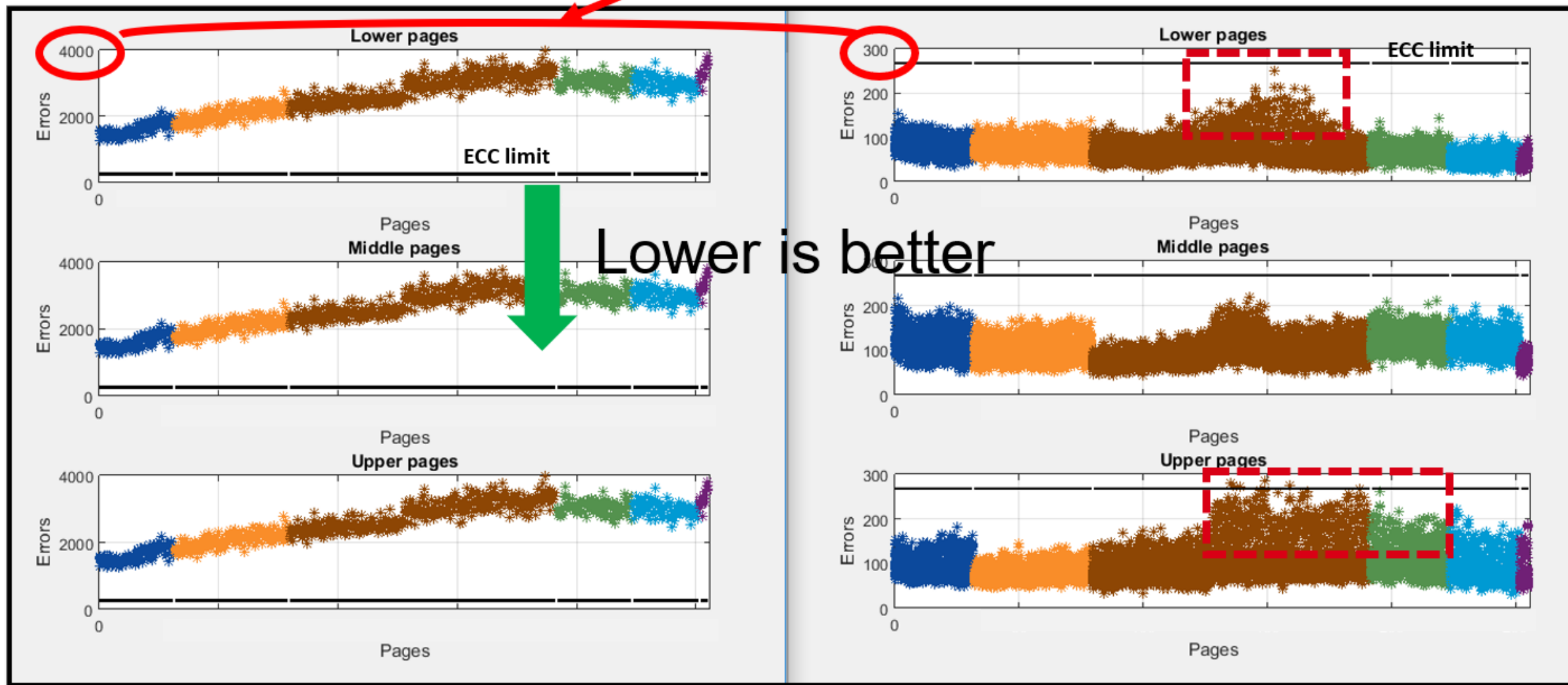


Flash Memory Summit

Native NAND BER

Look at the scale!

NAND BER w/ Classic Flash Management

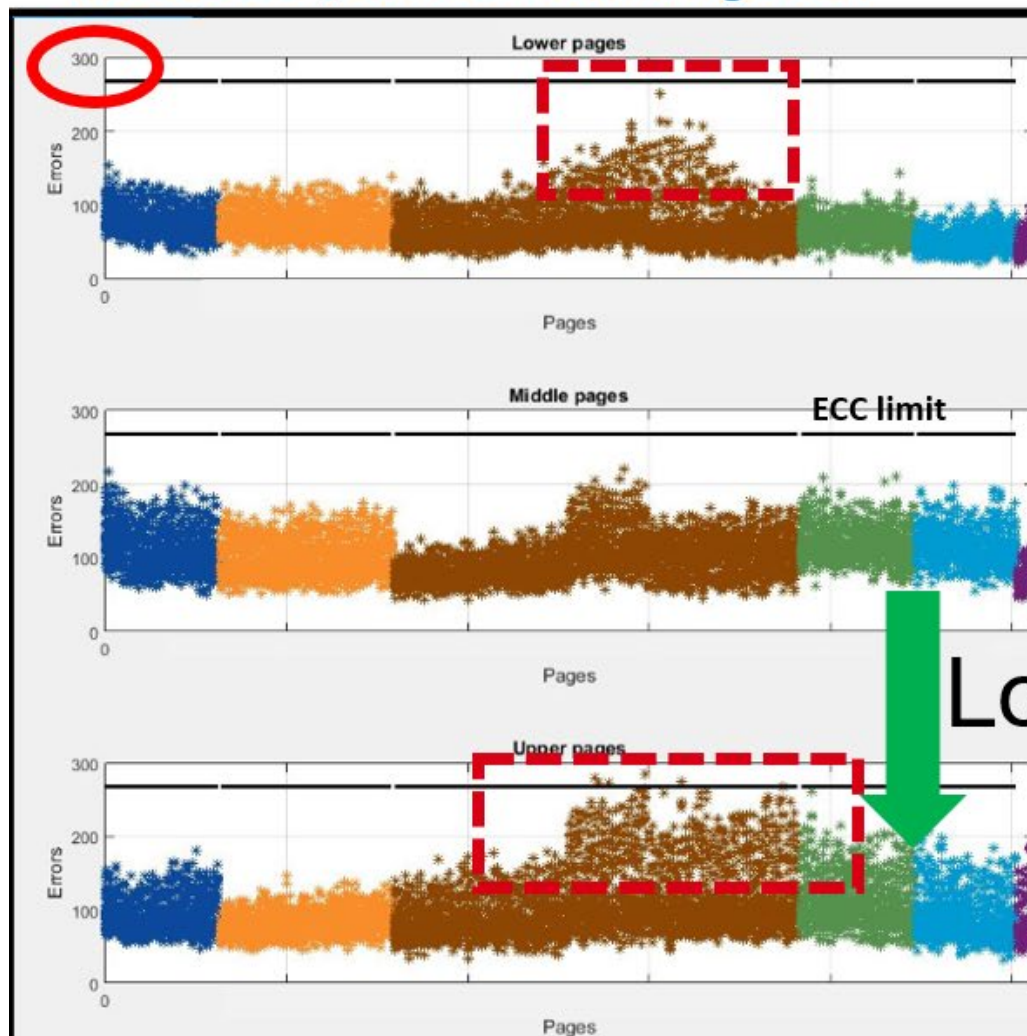


# NAND BER After Applying Machine Learning

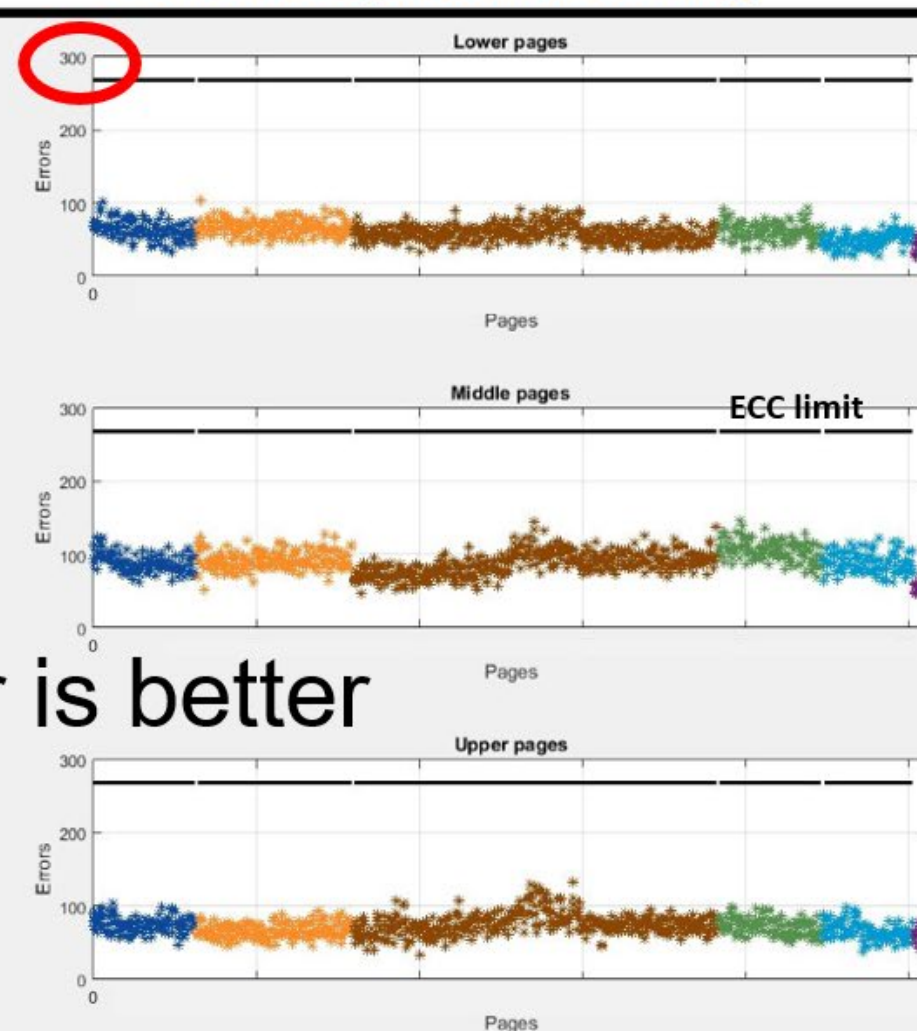


Flash Memory Summit

## NAND BER w/ Classic Flash Management



## NAND BER w/ Machine Learning



Lower is better



Thank you !  
Visit Microchip Booth # 613