# Flash Endurance and Retention Monitoring

## An in-situ and Field Testing Method

Steven R. Hetzler
IBM Fellow, Mgr. Storage Architecture Research
Blog: http://drhetzler.com/smorgastor

# Predicting Wearout and Retention

- Wear leveling prevents direct measurement of endurance and retention
  - Block virtualization obscures PE cycle, age, location
- Beneficial to measure the behavior at the system level
  - Can perform cost-effective preventative maintenance
  - Can adjust usage to favor strong devices over weak
  - Can perform incoming parts test and verify device behavior
    - Not all devices behave the same

FlashMemory SUMMIT

# Monitor Data Defined

- A selected set of reserved blocks
  - To be written with known monitor data
    - Good idea to include a real-time stamp!
  - Not wear leveled
  - Read/erase/write all controlled directly
    - Raw read (ECC off) – allows testing beyond ECC
    - Because some sectors are over the legal limit!
  - Always at PE cycles > user data PE cycles
- Measures end of life limits early
  - Time until reliability limit reached with current workload
- Enables 100% incoming parts test

# User Data

Access Distributions

Blocks (Arbitrary Units)

0.6
0.5
0.4
0.3
0.2
0.1
0.0

User Data

User Data Access

0
1,000
2,000
3,000
4,000
5,000

PE Cycles

- Distribution of erase blocks as a function of PE cycles at some time

# Monitor Data Histogram

Access Distributions

User Data
Monitor Data

Monitor Data Set

Blocks (Arbitrary Units)

0.6
0.5
0.4
0.3
0.2
0.1
0.0

0
1,000
2,000
3,000
4,000
5,000

PE Cycles

- Monitor data access distribution is at higher PE cycle count
- Provides information on device behavior in advance of user data
- Always kept ahead of the user distribution
  - Cycled and aged to sample error rate behavior

# Qualification Data Histogram

Access Distributions

Blocks (Arbitrary Units)

0.6
0.5
0.4
0.3
0.2
0.1
0.0

User Data
Monitor Data
Qual Data

Qual Data Set

0
1,000
2,000
3,000
4,000
5,000

PE Cycles

- Can also have a small set of factory qualification data
  - Pre-cycled to expected limit
  - Can measure longer retention times
  - Can put cycle past expected limit to test headroom

# Monitor Data Behavior

Access Distributions

Blocks (Arbitrary Units)

0.6
0.5
0.4
0.3
0.2
0.1
0.0

0
1,000
2,000
3,000
4,000
5,000

PE Cycles

User Data
Monitor Data
Qual Data

- After further PE cycles, user distribution move to right
  - Monitor data cycled to remain ahead of user distribution

**These are canary blocks *ahead* of the user data**

# Practical Aspects

- Impact can be minimal
  - 0.1% monitor data should be adequate
    - 256MB on a 256GB device
  - 0.1% impact on write and erase to keep ahead
  - Monitor data cycles once for every 1,000 data PE
- Qualification blocks
  - Pre-cycled at incoming test
  - @ 1s PE cycle time per set = 80 minutes

# Sector Error Limits

- System designed to a sector loss rate
  - RAID designed to correct for this

| | SAS HDD | eSSD |
|---|---|---|
| NRRE bit interval spec | 1e16 | 1e18* |
| IOPS (4kB) | 250 | 30,000 |
| Mean Y/sector loss | 320 | 260 |

Your target is higher!

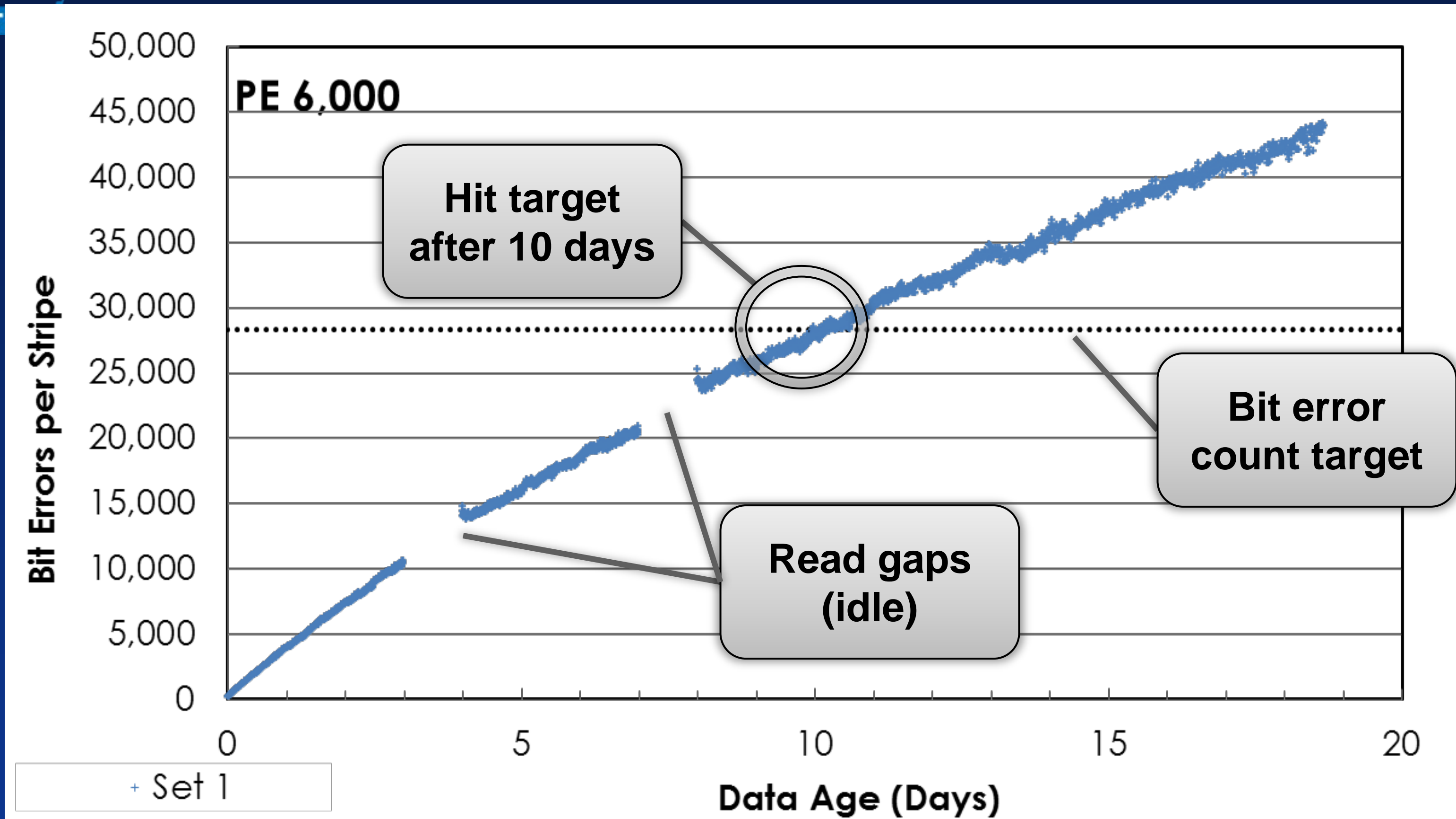
If your device is faster…

- *JESD218 spec is same as SAS HDD = 2.6Y MTTDL
  - Important spec is time to fail, not bits to fail
- See http://drhetzler.com/smorgastor for more details

# Computing BER at Retention Limit

- Compute limiting BER based on device ECC
- For 3xnm device data here, 15 bit BCH
  - pFail = 4291/1e18 = 4.1e-15
  - BERtgt = 2.11e-4 (just invert the binomial)
  - You can do this for any ECC
  - Convert to a bit error count/MonitorData sample
    - Here, 16x1MB erase blocks (128Mbits)
    - BCH 15: 28,300 error bits
    - BCH 29: 91,268 error bits
    - BCH 60: 146,297 error bits

# Snapshot View of Device Data

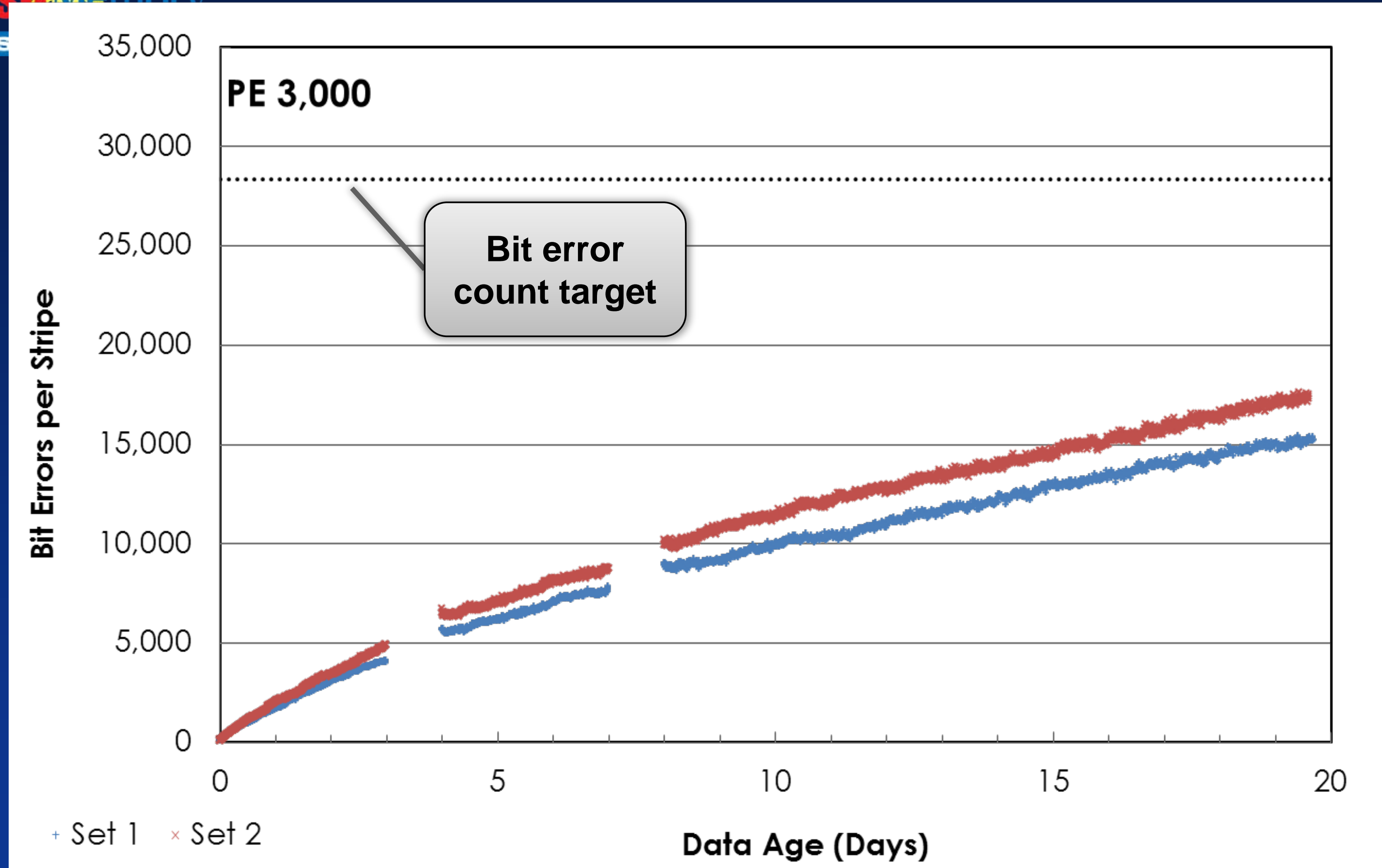



- Monitor data set cycled to PE 6,000
  - Data collected for 18 days – read 10 times per hour
  - 2x 24 hour gaps with no reads (to measure read disturb)
  - Directly measured the retention limit – here 10 days

# The Tests

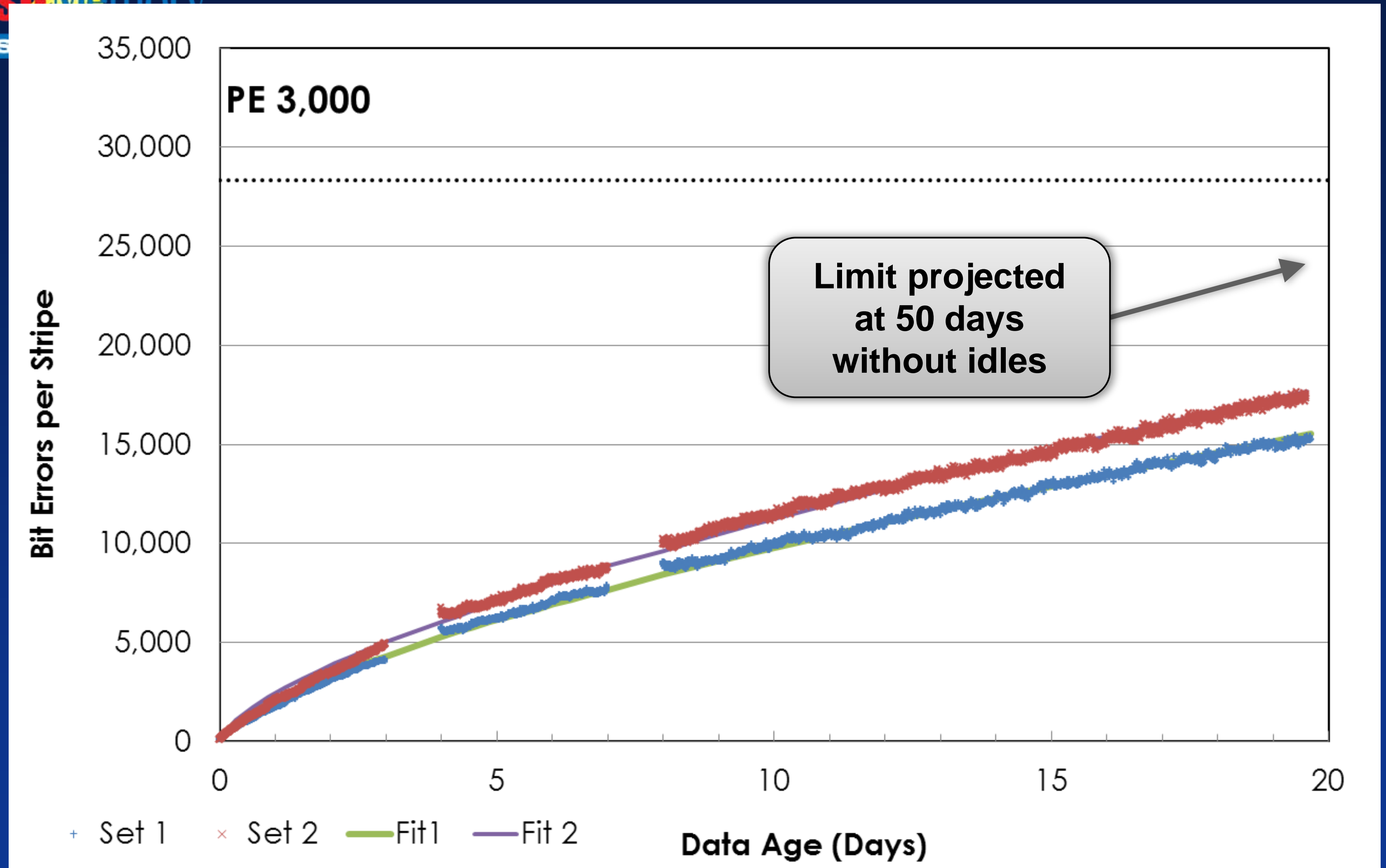
- 3xnm SSDs
  - Spec 1 year retention @ 5,000 PE @ 60C
  - Raw data shown here as 70C to limit aging
    - Effect is the same, values will change somewhat
  - Device supports monitor data interface
- 2 erase block stripes per MD sample
  - 256Mbits per sample
  - 3 samples in the set (total 768Mbits)
  - PE Cycles
    - 3,000, 4,000, 6,000
  - Real time aging

# 3xnm Data at PE 3,000



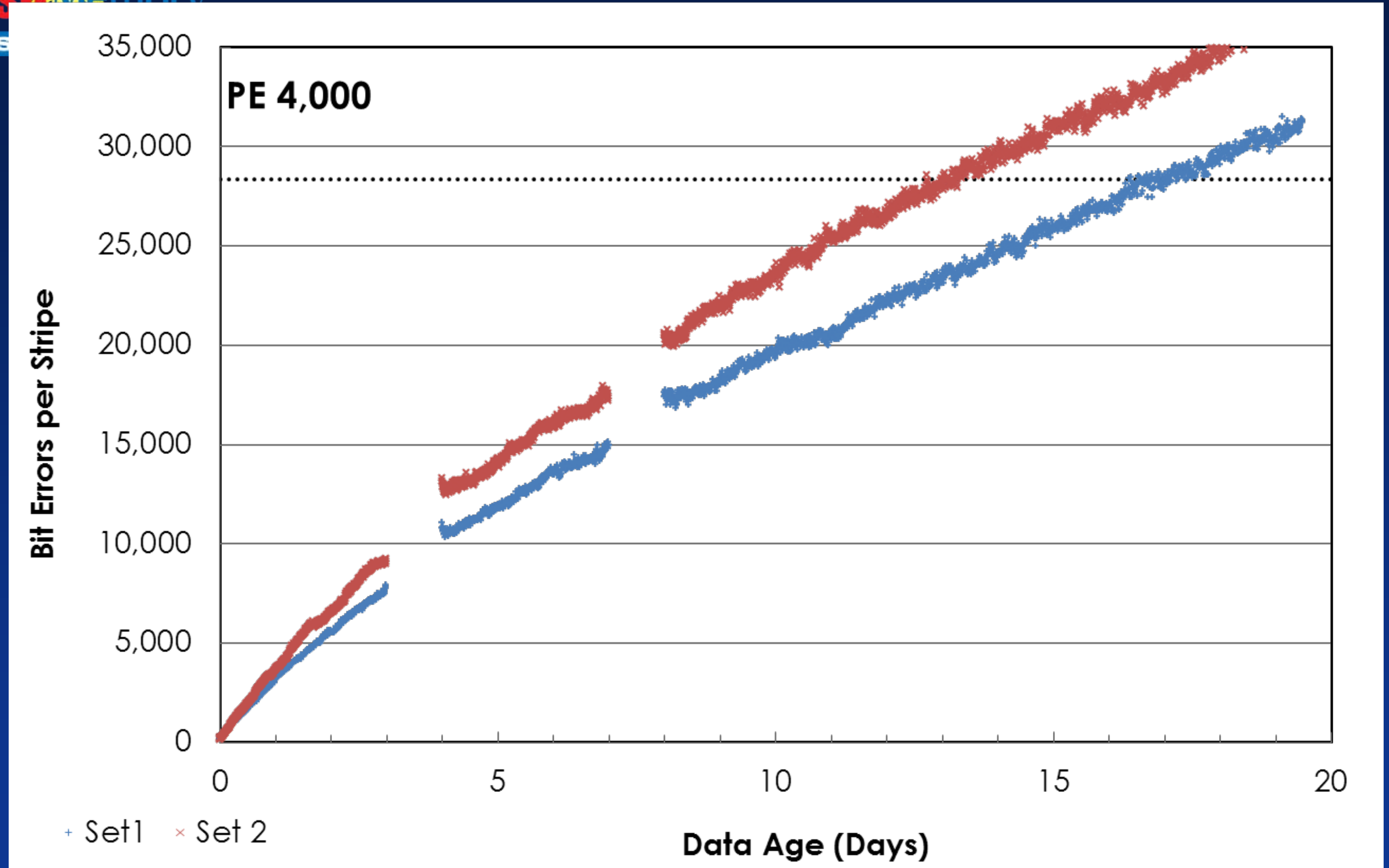

- Can estimate variance from data
  - Can also combine with ECC correction counts from user data

# 3xnm Data at PE 3,000

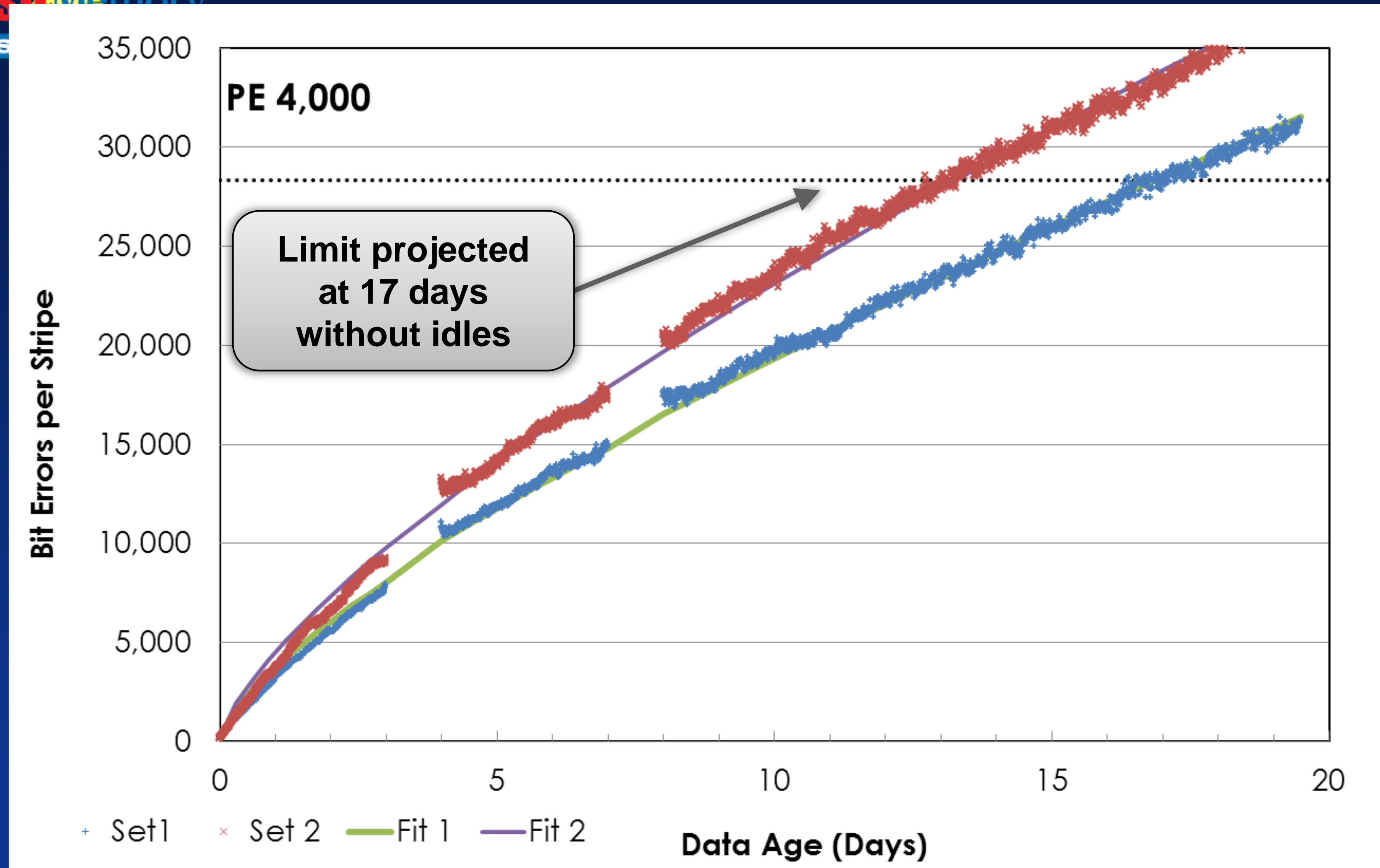



- Can do a functional fit: E=h(Age^k)(Reads^g) + b
  - Allows projection to different read rates

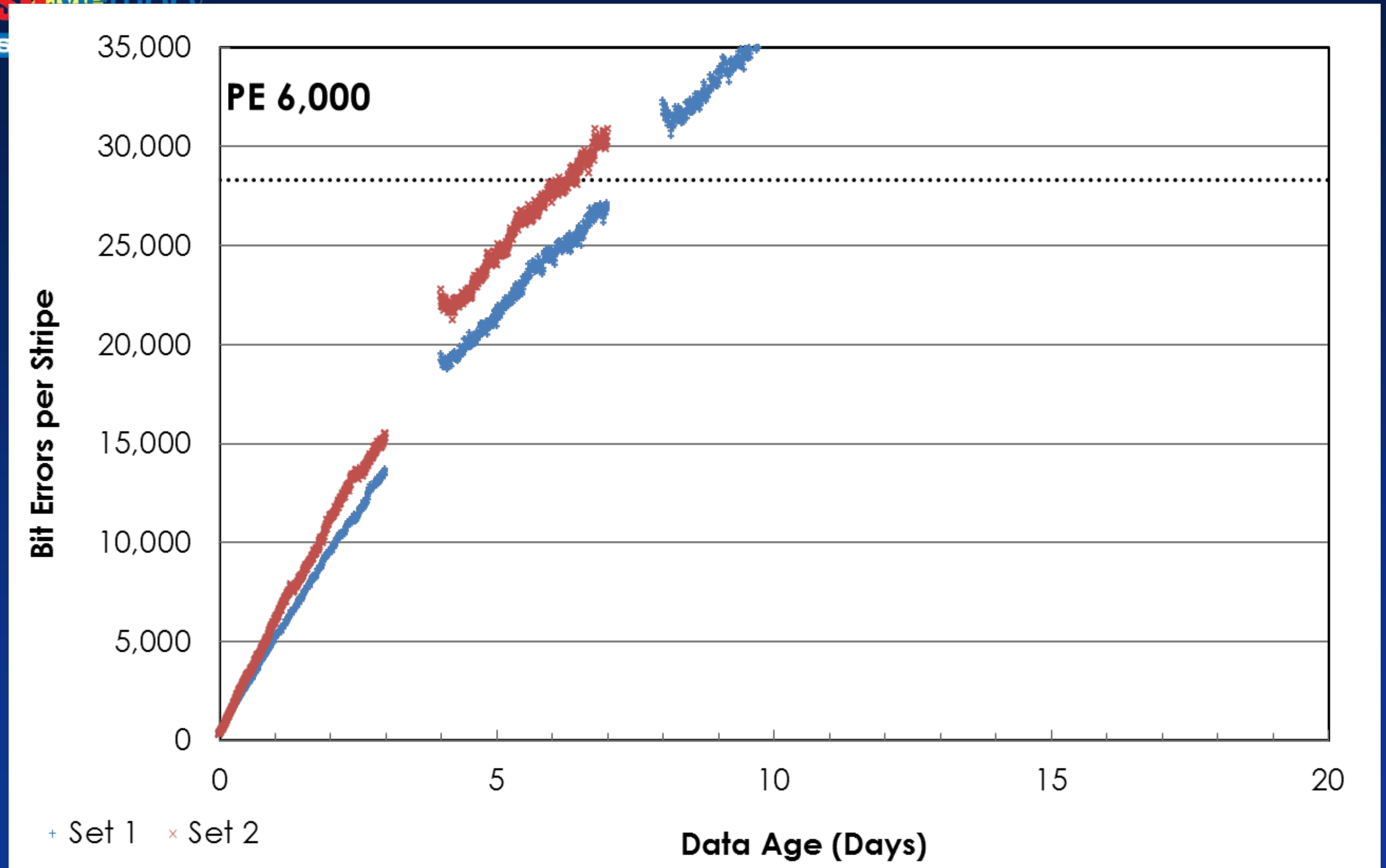
# 3xnm Data at PE 4,000



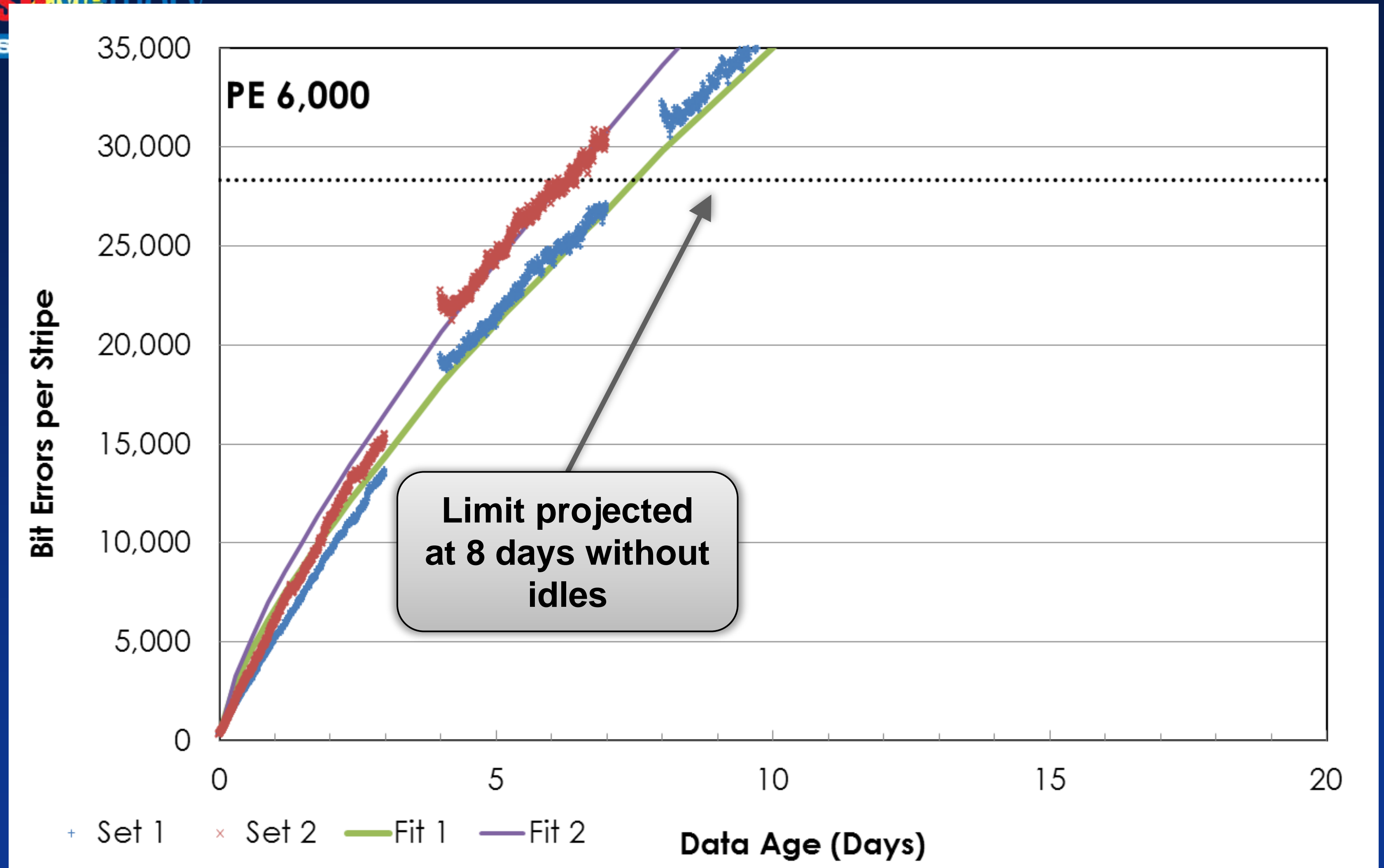
# 3xnm Data at PE 4,000



# 3xnm Data at PE 6,000

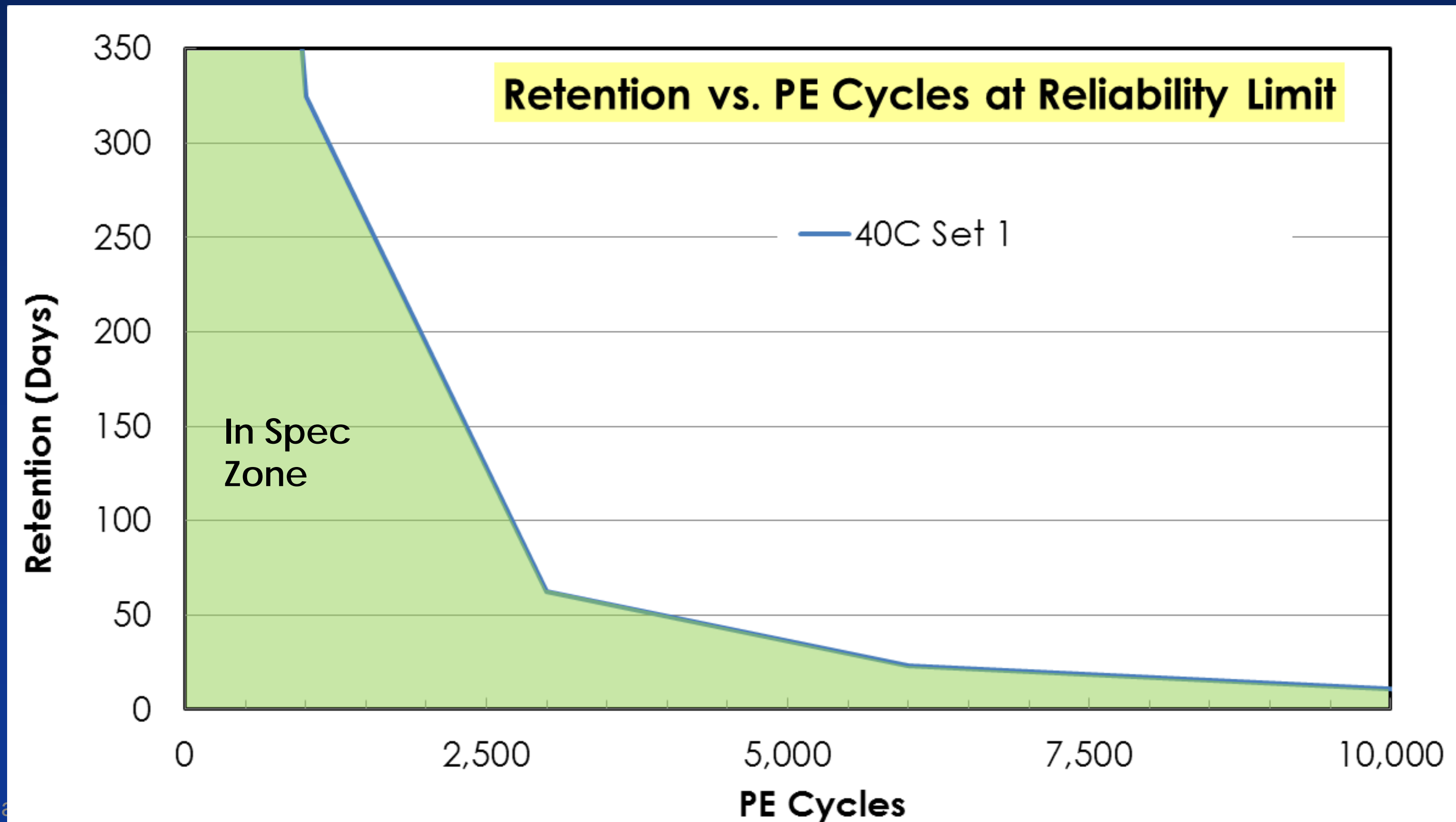


# 3xnm Data at PE 6,000



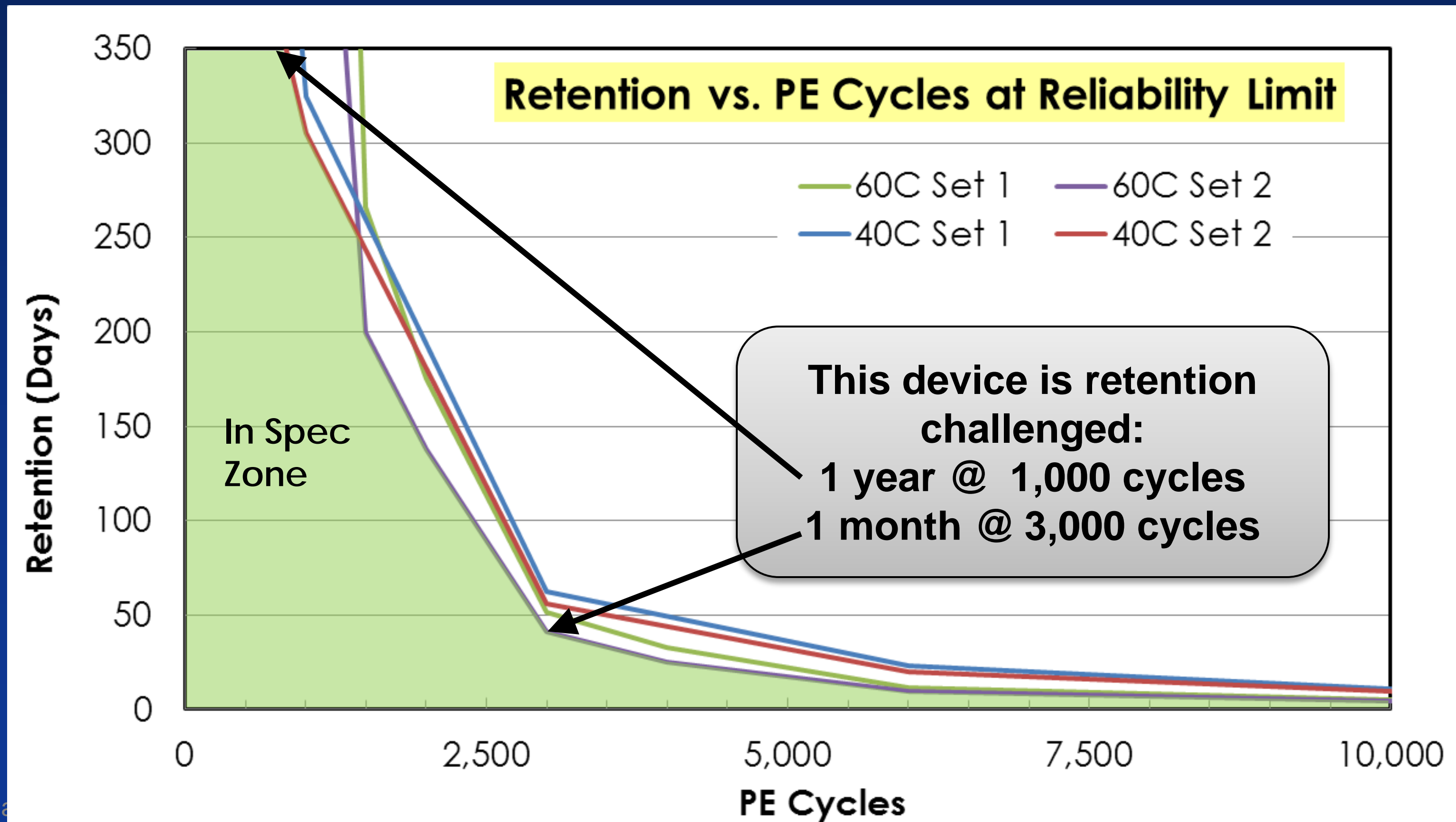
# The RPE Chart

- Retention vs. PE cycles at reliability limit
  - Combines data from Monitor Data sets
  - Blocks in the “In Spec Zone” should be fine



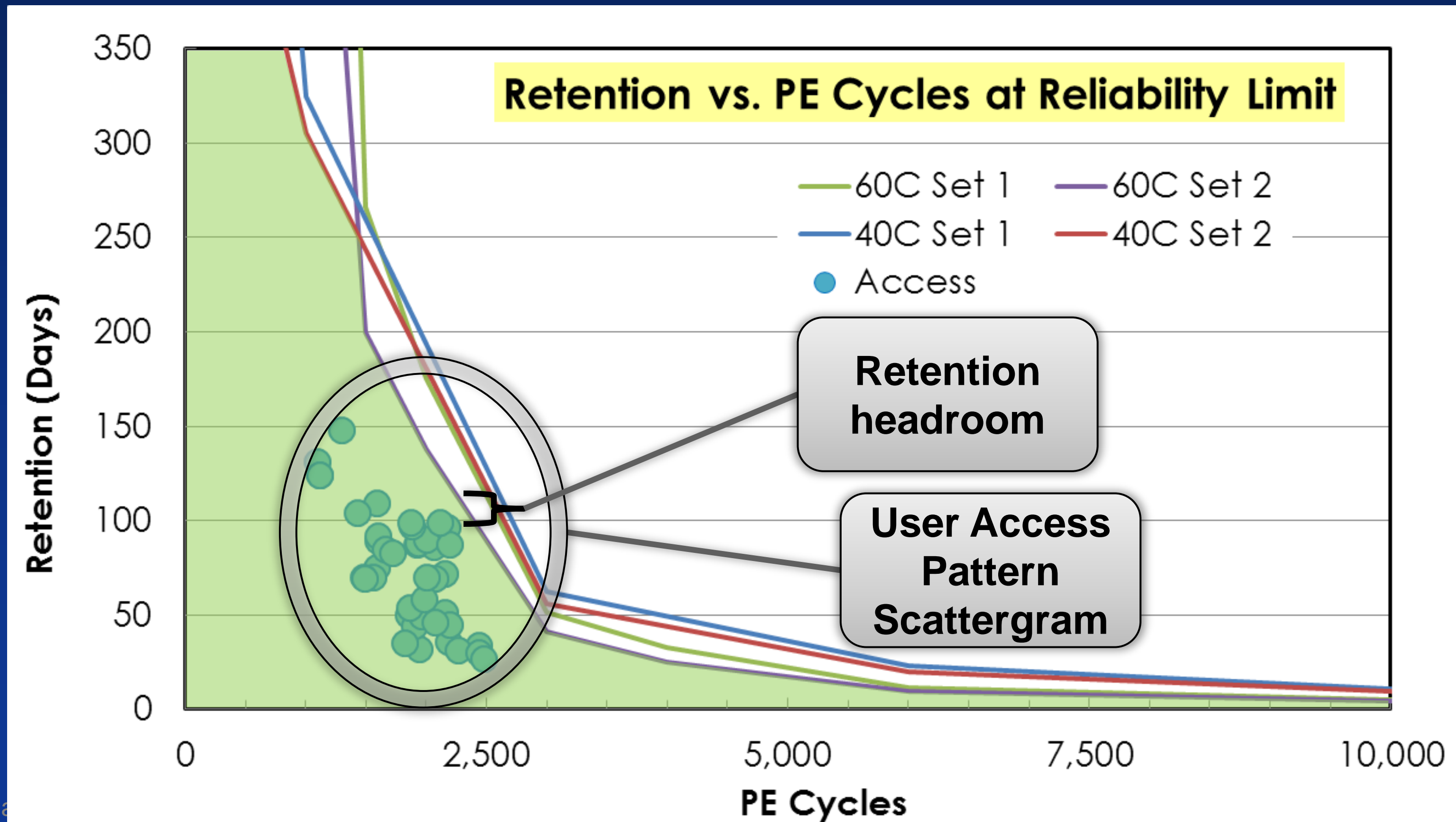
# RPE Chart From Combined Data

- Combined temperature curves here from lab
  - Likely to see aggregate temperatures in the field



# RPE Chart From Combined Data

- We can plot the scatter-gram of the user access patterns
  - Retention headroom can be measured for each block



# Summary

- Monitor data makes it possible to measure endurance and retention
  - Provide an interpolated end of life measurement
  - Within a device, or at the system level
- System level interface changes:
  - Select set of blocks, direct read (raw)/write/erase
- Integrators can verify devices against specs
  - Parts can be tested, sorted and used to maximum capability
- Need to estimate the variance
  - Weakest blocks responsible for most errors
- Systems and devices can take remedial action
  - Feedback to data management (e.g. refresh policy, load balance, maintenance, block retirement, etc.)