Health-Binning

Maximizing the Performance and the Endurance of Consumer-Level NAND Flash

Roman Pletka, Saša Tomić
IBM Research – Zurich

Systor 2016, Haifa, Israel

June 6, 2016

Outline and Motivation

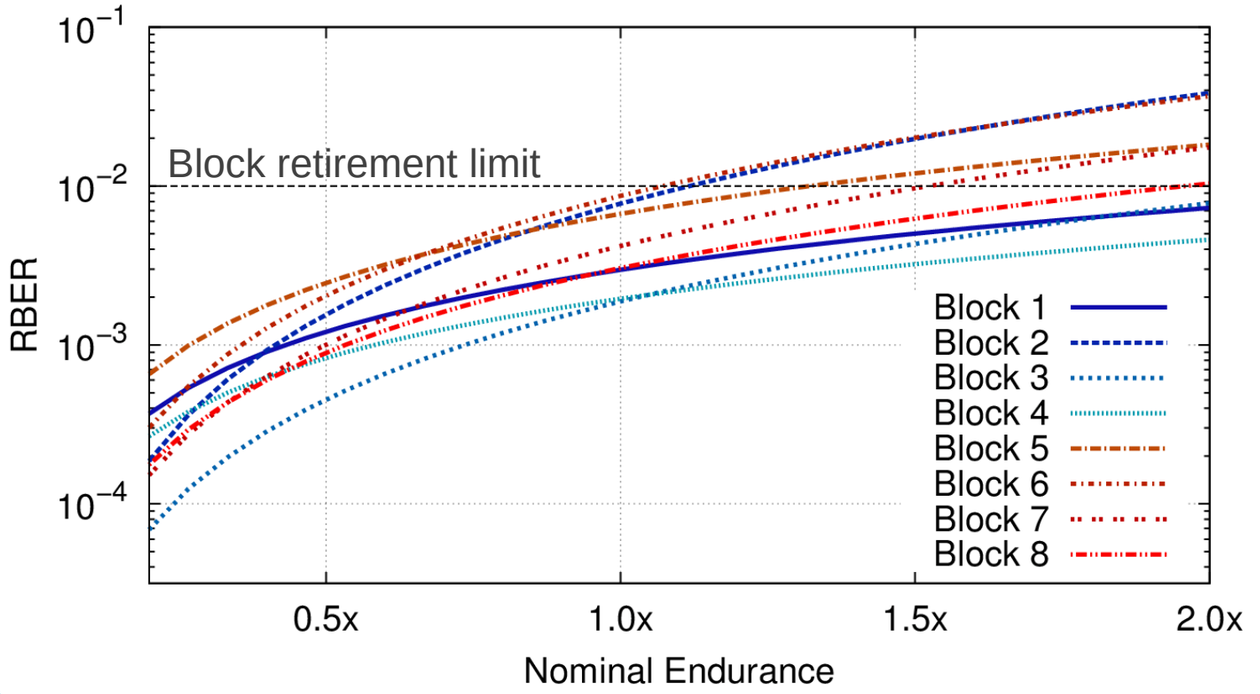
- Introduction
  - Consumer-level NAND Flash Characteristics
  - Wear Leveling
- Block Management with Health Binning
- Evaluation Environments
  - Simulator and hardware platforms
  - Flash model
- Endurance Evaluation
  - Health Distributions during operation and towards end-of-life
  - Endurance gains and segregation granularity
- Conclusion

Disclaimer: Results and parameters in this presentation are not specific to a particular product or a Flash memory vendor

Typical consumer-level Flash block characteristics

- Significant differences in the RBER over time can be observed in consumer-level NAND Flash:
  - Some blocks have almost twice the endurance of others
  - Low RBER at early life does not indicate a good block, and an early high RBER not a weak one

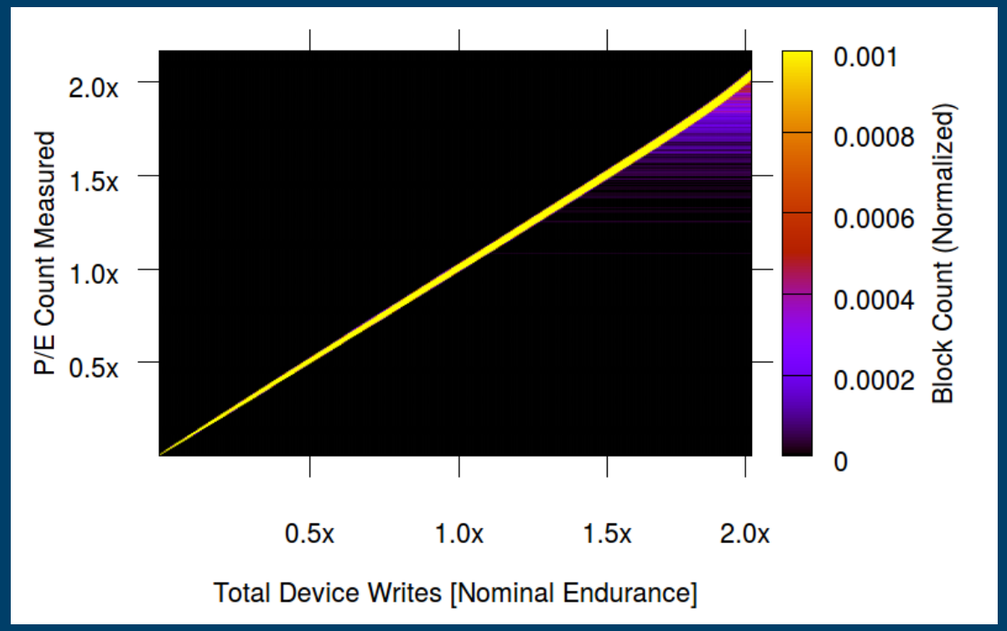
RBER of different consumer-level flash blocks in the same device as a function of P/E cycles

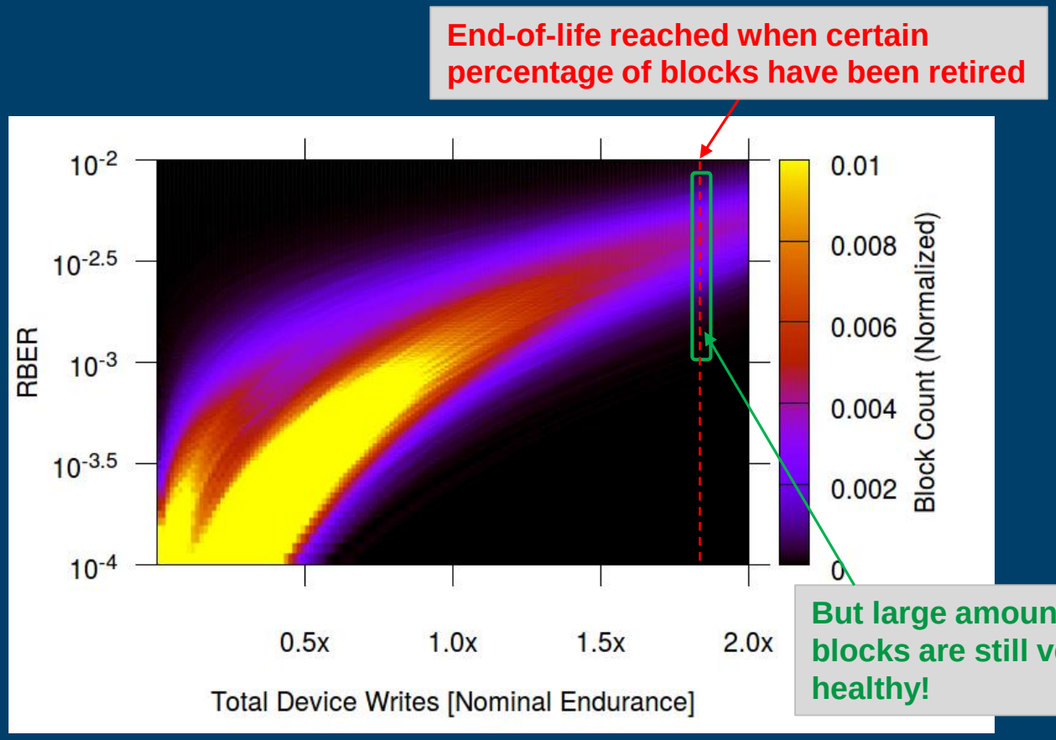


How does Workload Skew affect P/E Cycle and Health Distributions ?

Characteristics of conventional wear leveling using P/E cycle balancing under a Zipfian 95/20 workload :

- Measured P/E count distribution very narrow despite the workload skew
- First blocks get retired after ~1.1x of the nominal endurance (horizontal lines)
- Significant variations in block wear throughout the device lifetime
- Less than 60% of available endurance achieved when end-of-life reached (i.e., given percentage of blocks retired)





From Traditional Wear Leveling to Health Binning

Dynamic WL

Balance P/E cycles across blocks upon overwrites and relocations. Typically uses the least worn available block to place new data.

- Introduce data placement with stream segregation
- Use better blocks for hotter data

Static WL

Identifies the least worn blocks holding static data in the background. Still valid data is relocated to another block causing an increase in write amplification.

- Reduce Static WL to address retention and read disturb limitations of Flash
- Perform relocations instead of block swapping

P/E Cycle-based WL

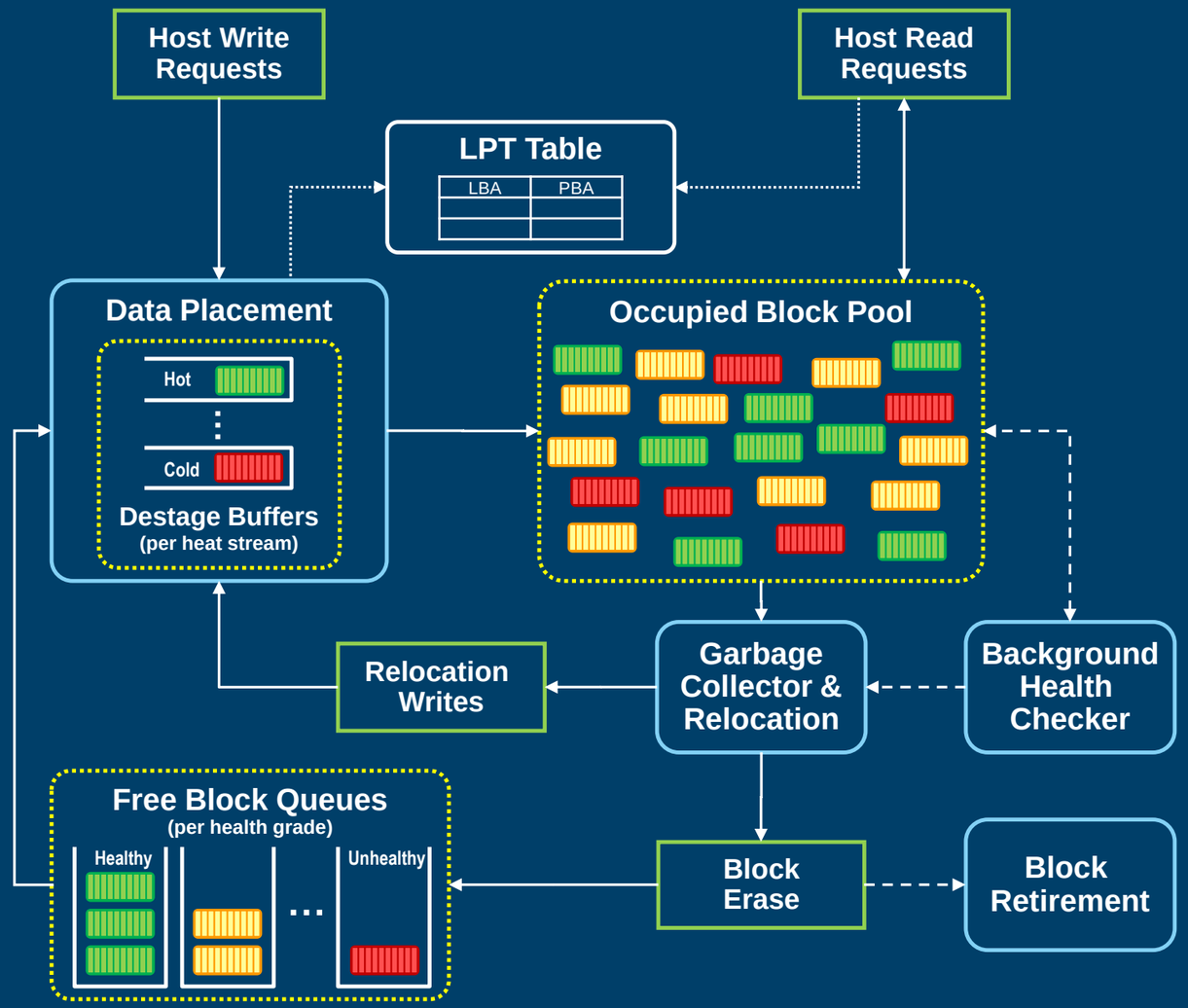
Balances wear of blocks based on their program-erase cycle count only.

- Background grading of blocks based on RBER
- RBER estimation based on ECC feedback

Block Management Overview

Block Management functions enabling Health Binning:

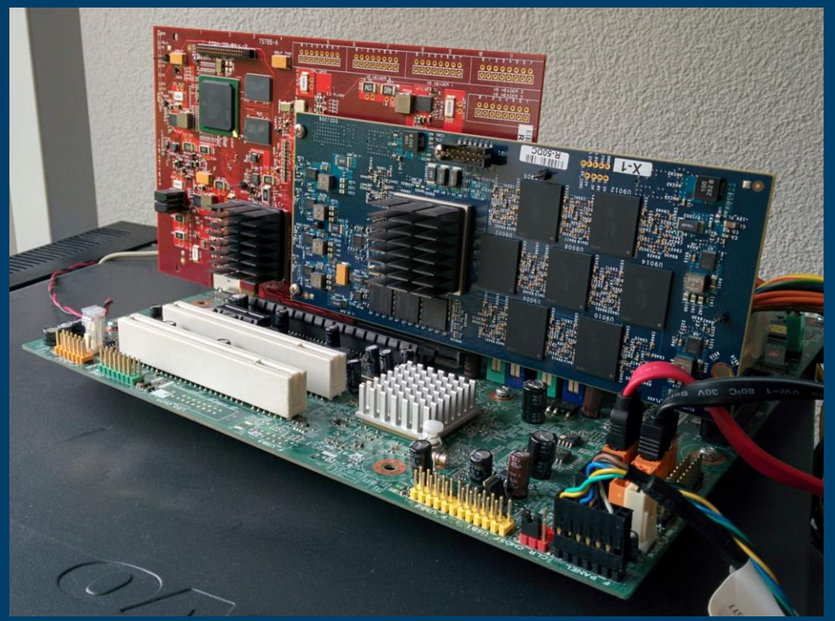
- Background Health Checker
  - Continuous health monitoring of each block holding data in the background
  - Block health is determined by the worst page in the block
- Free Block Queues
  - Maintain separate block queues for each health grade
  - Healthy block may overtake less healthy ones
- Data Placement
  - Segregate writes into “streams” according to their update frequency at LBA granularity
  - Maintain separate destage buffers for each stream, hotter streams get better blocks



Simulation and Hardware Environment

Simulation Environment :

- Simulation environment performs all block management functions but reads and writes do not transfer data
- Flash model that emulates ECC decoder output (RBER)
  - Large-scale characterization data of 19nm and 16nm cMLC flash devices from different manufacturers used to build device-specific flash models
  - The flash model is based on a Gaussian mixture model [1]
  - More than 10s of thousands of flash block parameter generated



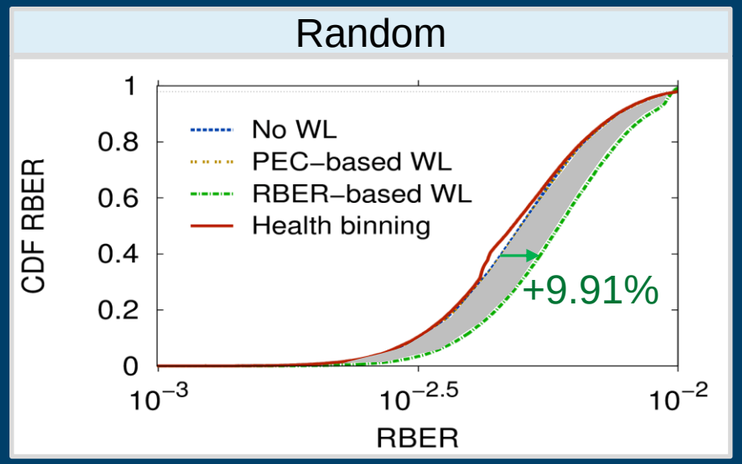
Hardware Setup:

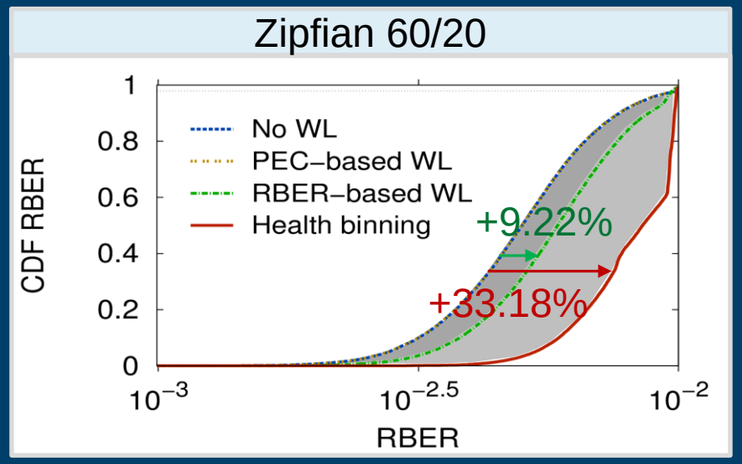
- Full-fledged FPGA Flash controller and GPP for flash management with integrated ECC encoder/decoder
- Utilized for verification of simulation results and product development
- Possibility to bypass access to flash and replace it with our flash model

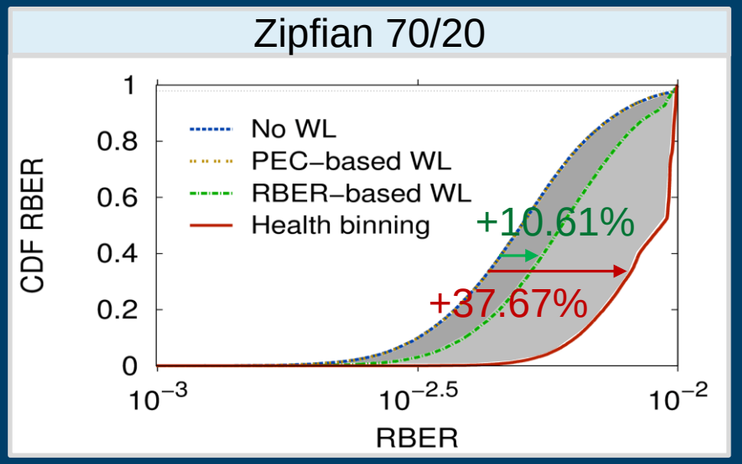
[1] T. Parnell, N. Papandreou, T. Mittelholzer, and H. Pozidis. Modelling of the threshold voltage distributions of sub-20nm NAND flash memory. GLOBECOM ’14

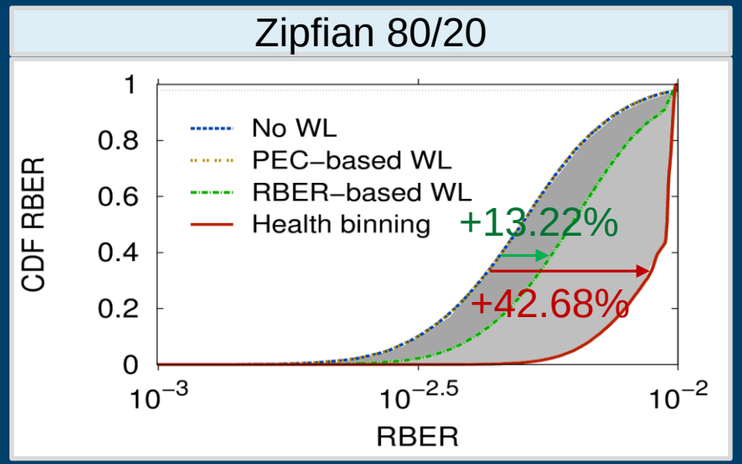
Comparison of Block Wear at End-of-life

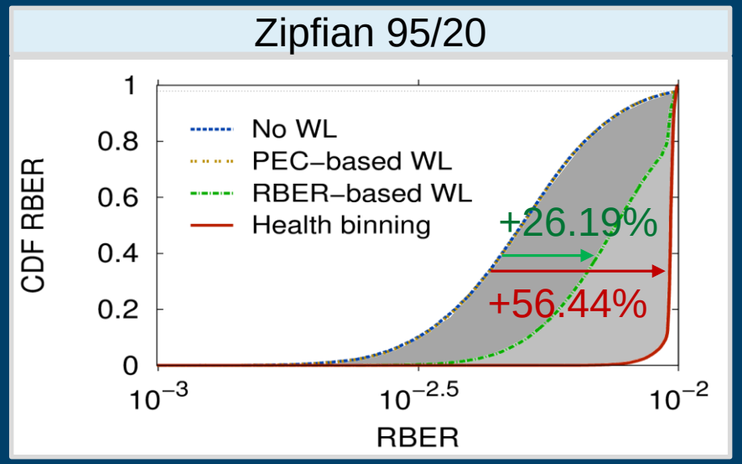
CDF of block wear at end of life for different workloads for a 19nm CMLC device for different workloads :

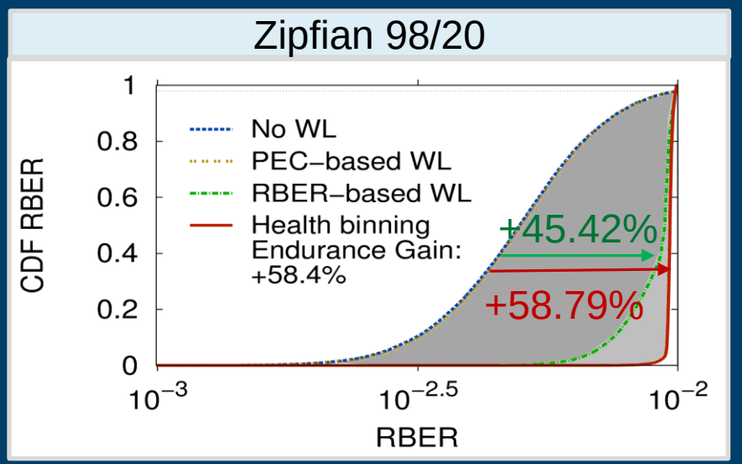








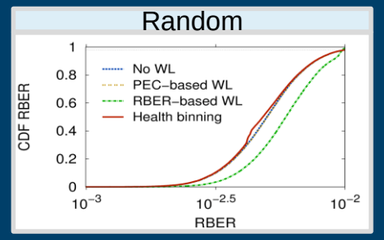




Comparison of Block Wear at End-of-life (cont.)

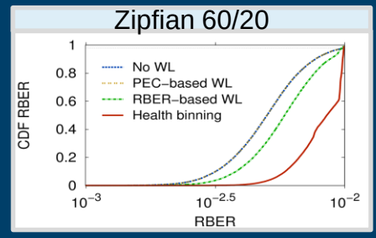
Uniform Random:

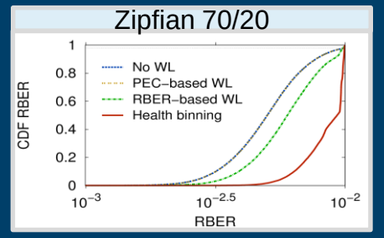
- RBER-based WL performs best with uniform random workloads (e.g., “always pick the best block”).
- Segregation of writes results in less good blocks being selected for data placement for data which supposedly seems to be colder, but in reality has the same update frequency due to the workload characteristics.

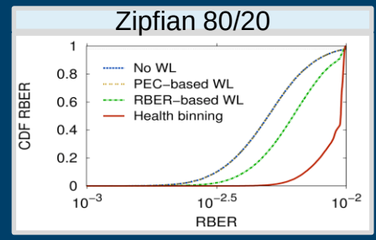


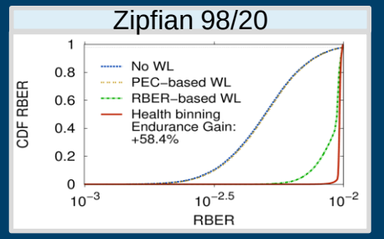
Skewed Workloads:

- Already with a very light workload skew, health binning dramatically increases endurance.
- RBER-based WL also sees some improvement, but significantly less than health binning
- Even with high workload skew, PEC-based WL is only marginally better than no WL.









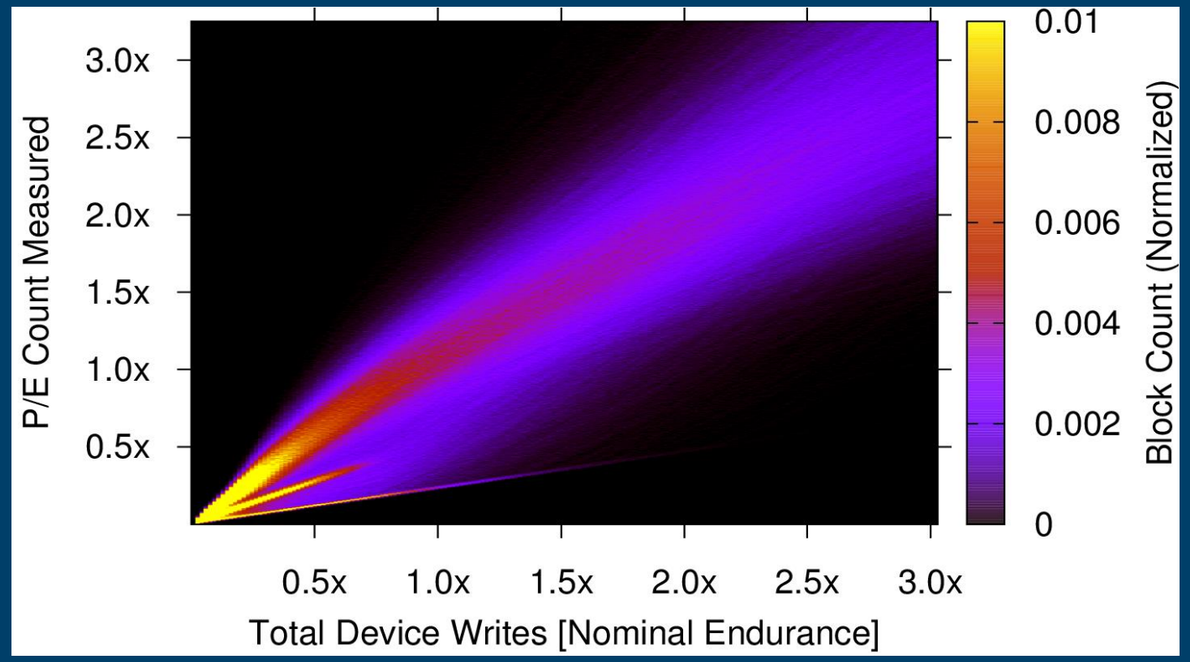
P/E Cycle and Wear Distribution using Health Binning

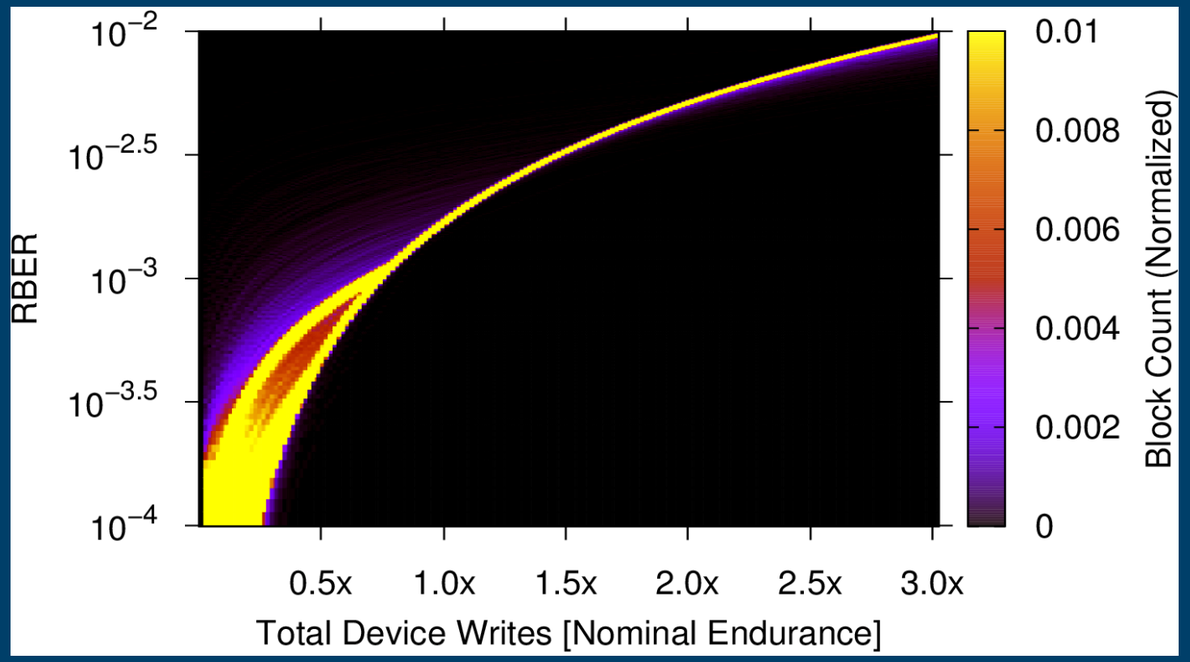
Simulation parameters :

- Workload: heavily skewed Zipfian 95/20, no static data
- Flash model: 19nm c-MLC, >10k different blocks
- Health Binning with 4 streams

Health binning results :

- Best blocks endure ~2.5x more P/E cycles than the worst ones
- Extremely narrow RBER distribution after 1.0x of total device writes





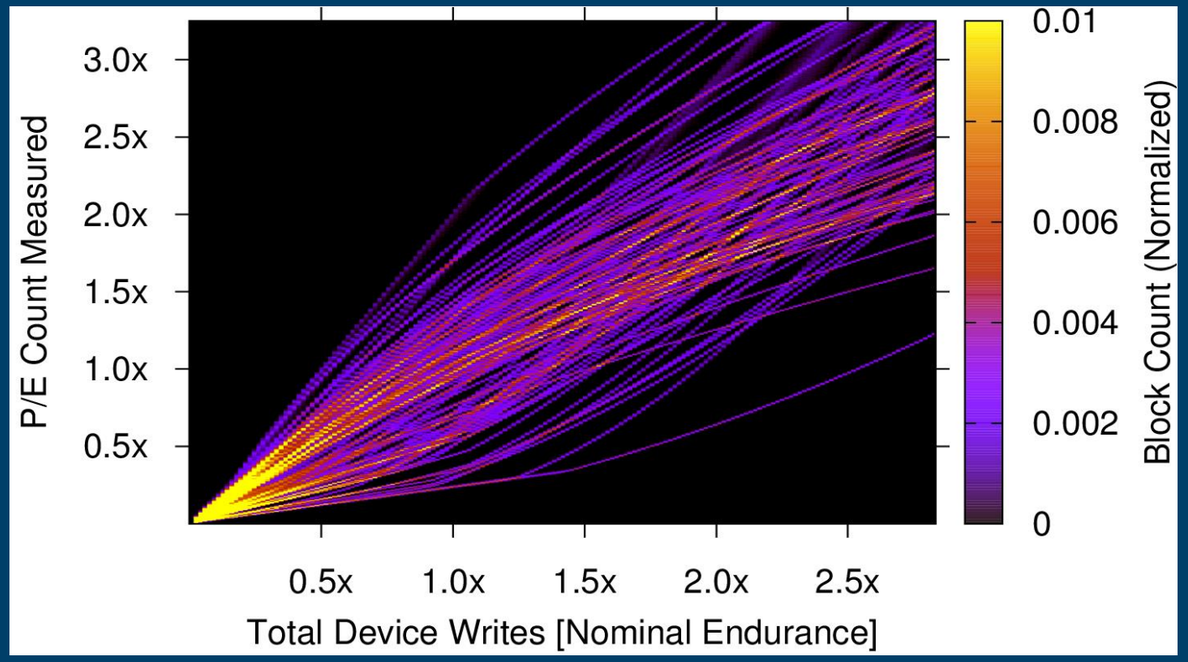
P/E Cycle and Wear Distribution using Health Binning (cont.)

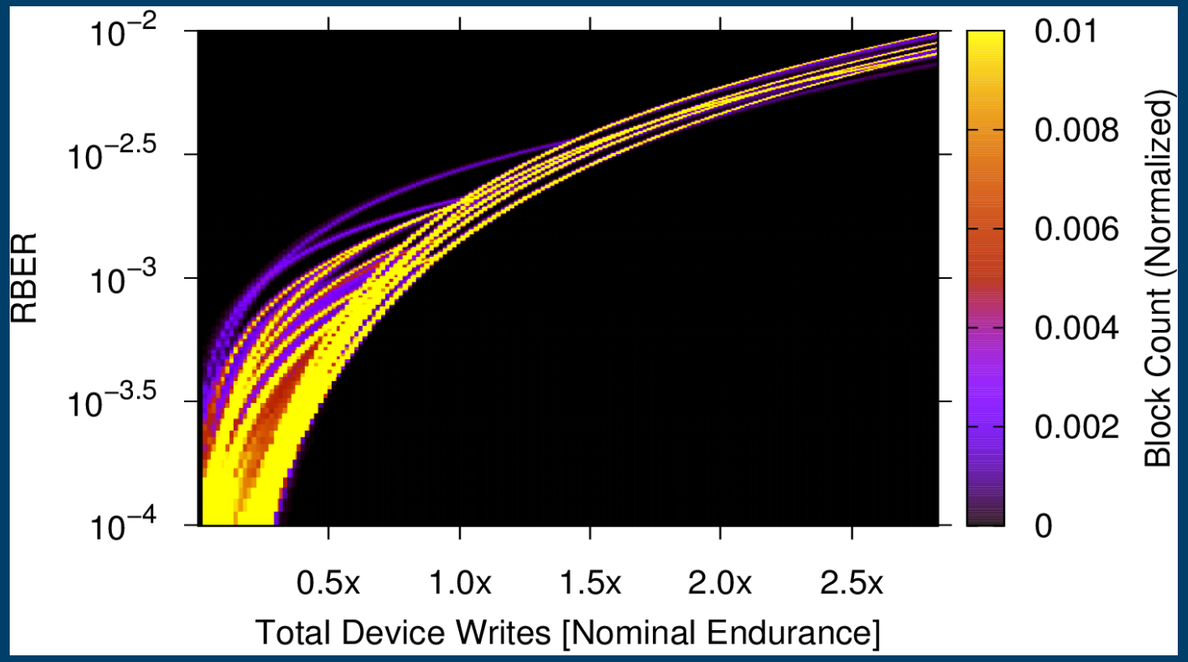
Simulation parameters :

- Workload: heavily skewed Zipfian 95/20, no static data
- Flash model: 19nm c-MLC, >10k blocks but only 100 different block types

Changes in wear characteristics for different block types now visible:

- Decreasing slopes indicate blocks that initially appeared better (lower RBER), but turned out to be less good blocks
- Increasing slopes indicate initially underestimated blocks
- RBER distribution at end-of-life slightly wider due to the small set of different block types

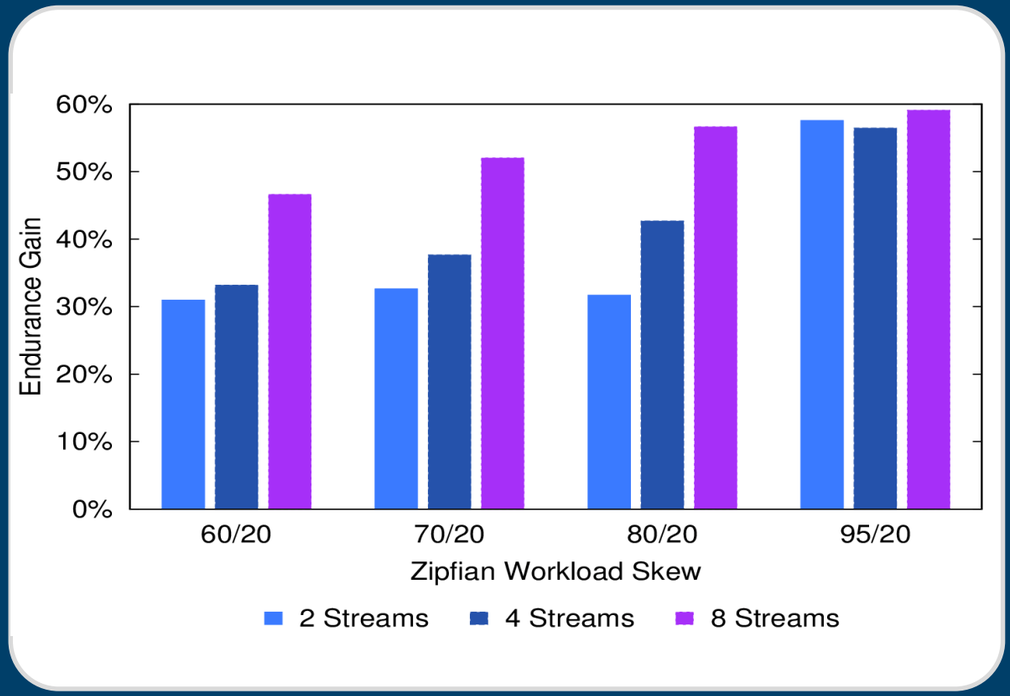




Segregation Granularity

Endurance gain is not only a function of workload skew but also depends on the number of streams at which segregation in data placement is performed :

- Analysis of 2, 4, and 8 streams and different workload skews using the 19nm cMLC flash model
- Generally, more streams result in higher endurance gains, even at lower workload skews.
- For highly skewed workloads (Zipfian 95/20), Health Binning achieves more than 95% of the available endurance irrespective of the number of streams
- Separate streams are maintained for relocations and host write requests
  - Allows for better capturing workload-dependent differences in temporal locality of the two types of write requests
- Limitations:
  - Each stream requires dedicated destage buffers (in DRAM) which limits the number of streams in a real implementation
  - More streams artificially reduce overprovisioning as blocks in the destage buffers are on average half filled.



Conclusions

- We observe significant differences in the endurance of flash blocks of the same device types with sub-20nm planar NAND flash devices. Traditional wear leveling approaches are not suitable to mitigate these deficiencies.
- Health Binning is a novel flash management technique that approaches WL from a device characteristics perspective:
  - Utilize RBER of the worst page in a block instead of P/E cycle count
  - Data placement segregates writes according to their update frequency and uses the best matching blocks available
  - Health Binning reduces Static WL to solely address retention and read disturb limitations of Flash
- Health Binning significantly outperforms traditional WL when workloads are skewed and achieves endurance gains of up to 80%
  - Health Binning further reduces the amount of read retry operations, hence achieves consistent latency
- Health Binning is a key enabler for consumer-level NAND Flash in enterprise storage systems
  - We integrated Health Binning into one of our commercially available product line
- Outlook:
  - Extend our Health Binning research to 3D-NAND flash devices

Questions ?

www.research.ibm.com/labs/zurich/cci/