Holistic Flash Management for Next Generation All-Flash Arrays

Roman Pletka, Nikolas Ioannou, Ioannis Koltsidas, Nikolaos Papandreou, Thomas Parnell, Haris Pozidis, Sasa Tomic
IBM Research – Zurich

Aaron Fry, Tim Fisher
IBM Flash Systems Development

Agenda

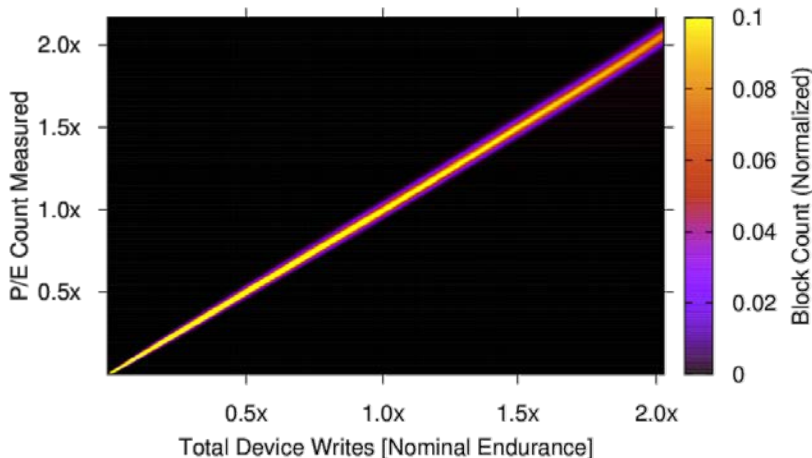
- Motivation
  - Flash characteristics
- Holistic Flash management
  - Overview Flash management functions
  - Design goals
- Evaluation
  - Write amplification and endurance results
- Conclusion & Questions

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

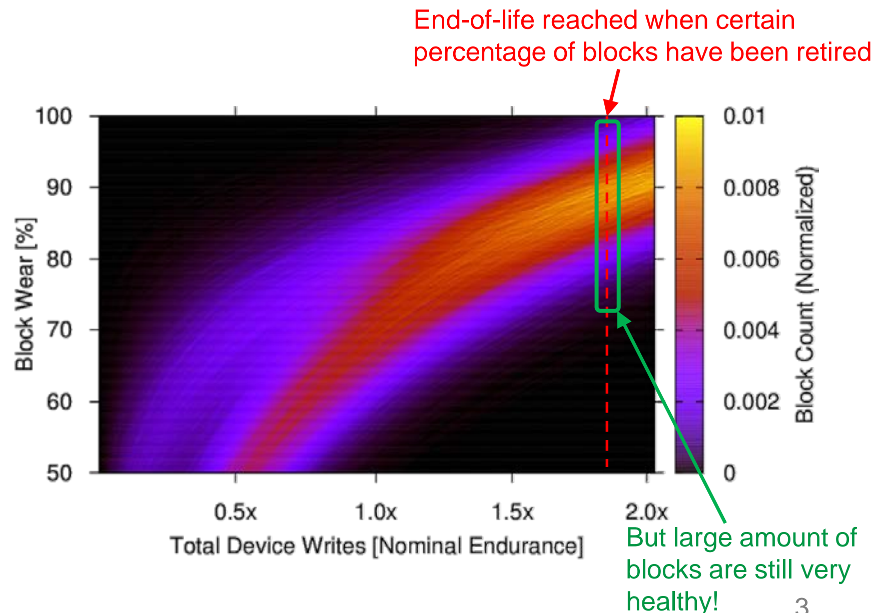
Flash Characteristics: Block Wear

Block wear : A quantity defining the quality of a block. Typically related to the raw bit error rate (RBER) of the worst page in a block and the error correction capability of the ECC used.

Conventional wear leveling can efficiently equalize P/E counts even under heavily skewed workloads such as Zipfian 95/20

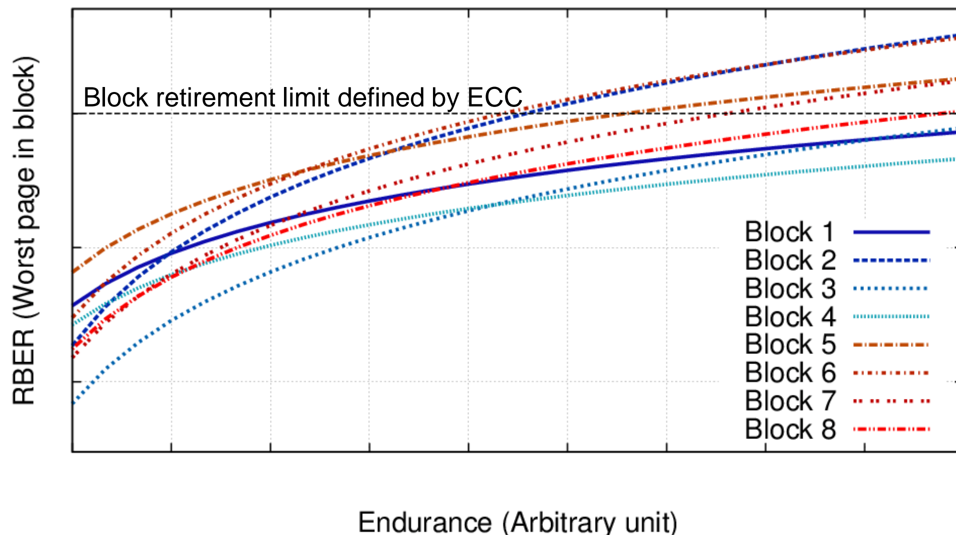


Block wear varies significantly across blocks



Flash Characteristics: Block Wear (cont.)

- Block wear of 8 selected blocks of standard MLC Flash:



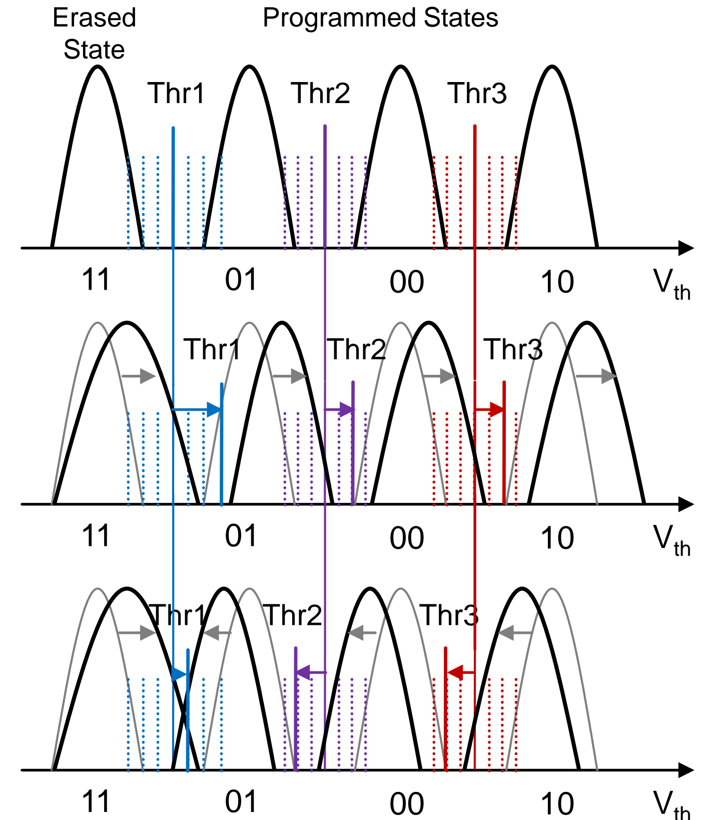
Observations:

- Some blocks look unhealthy in the beginning, but become significantly healthier than others (e.g., block 5 compared to 2 and 6)
- Other blocks are initially very healthy but age quickly (e.g., block 2).

- RBER in early life cannot be used to estimate block endurance.
- Continuously monitoring each block's health is required.

Flash Characteristics: Threshold Voltages

- Characterization results show that nominal threshold voltages are often suboptimal
- Appropriate shifting of threshold voltages can reduce the error rates significantly
- Threshold voltages shift depending on:
  - Number of P/E cycles
  - Number of reads a page has seen since programmed
  - Retention time
  - Individual block/page characteristics



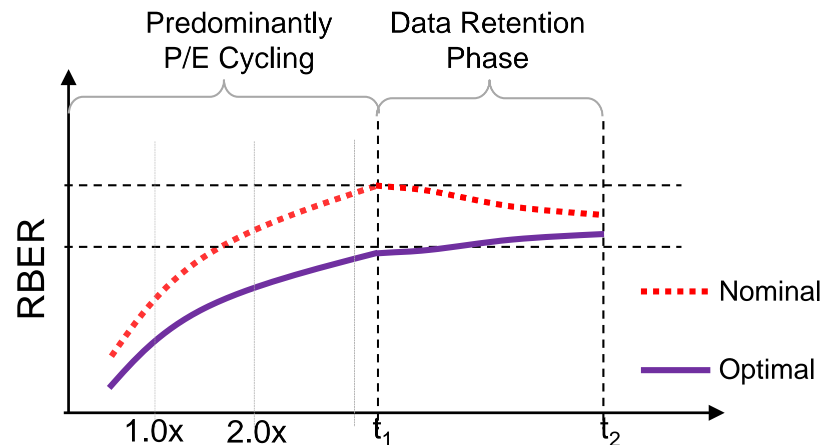
Holistic Flash Management

Function	Description / Observations	Design Goals
Garbage Collection	Reclaim invalidated space due to out-of-place writes. Relocation of valid data leads to write amplification (WA).	- Smarter data placement using heat segregation reduces write amplification.
Wear Leveling	Traditional approaches equalize usage of Flash cells by balancing P/E cycles of blocks. Wear leveling moves further increase WA.	- Equalize block health instead of P/E cycles. - Dynamic wear leveling: Smarter data placement using health binning. - Static wear leveling: Reduce to strict minimum to ensure retention targets.
Health Management	Blocks that reach the error correction capability of the ECC must be retired. Retired blocks eat up over-provisioning and ultimately limit device endurance even if there are still many good blocks available.	- Continuously monitor block health and shift threshold voltages accordingly - Actively narrow health distribution of all blocks with health binning. At end-of-life remaining good blocks only have little P/E cycles left.
Error Detection & Correction	Consumer SSDs perform read-retry and/or rely on ECC schemes using soft information. Read latency deteriorates with age of the device.	- Stronger ECC that does not require read-retry - Variable Stripe RAID™ - Array-level RAID

Dynamic Read Level Shifting

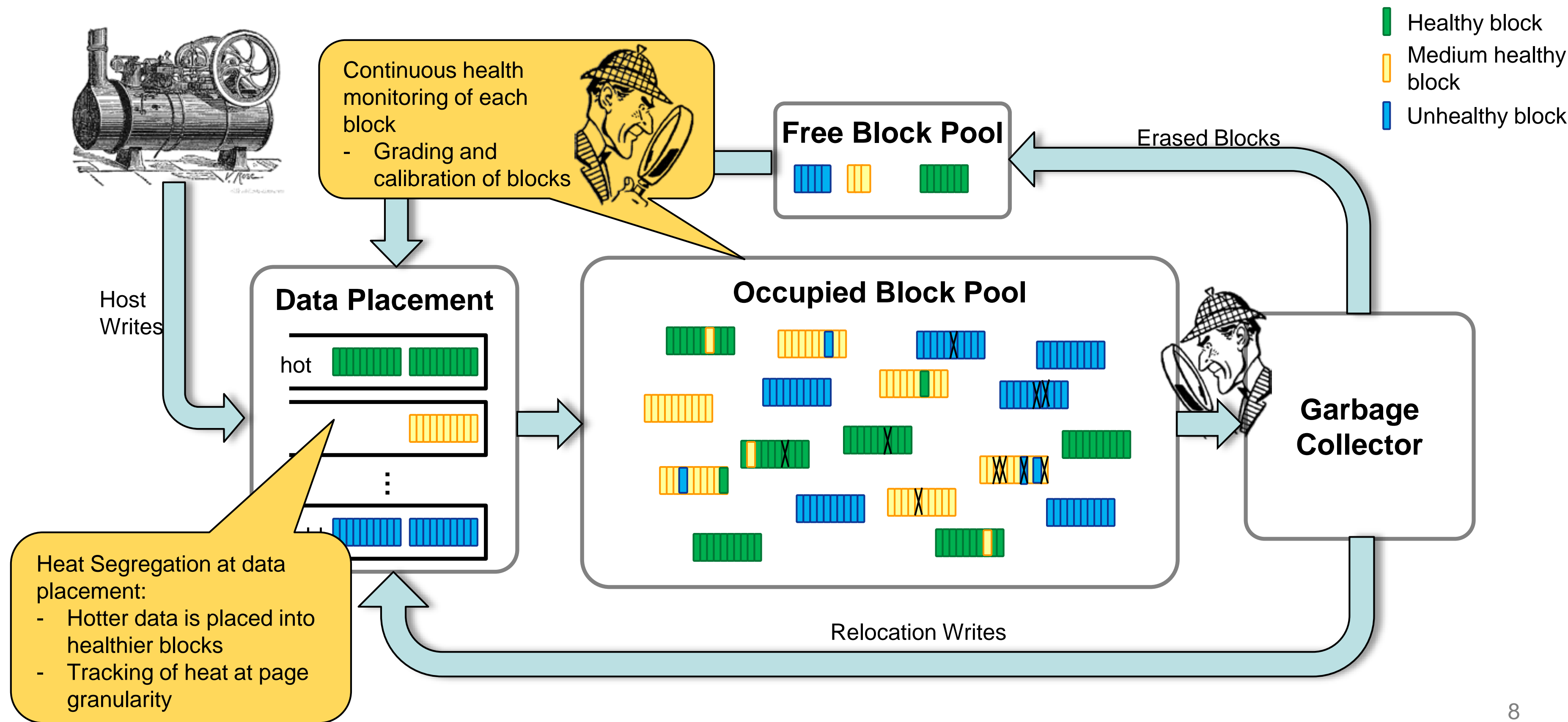
Dynamic Read Level Shifting:

- Extensive characterization is required to determine behavior of read levels under different conditions.
- Dynamic Read Level Shifting requires special access modes to the Flash.
- Optimal read levels must be continuously updated (calibrated) in the background.
- Use special techniques to reduce meta-data overhead.
- Dynamic read level shifting significantly contributes to maximizes flash endurance.

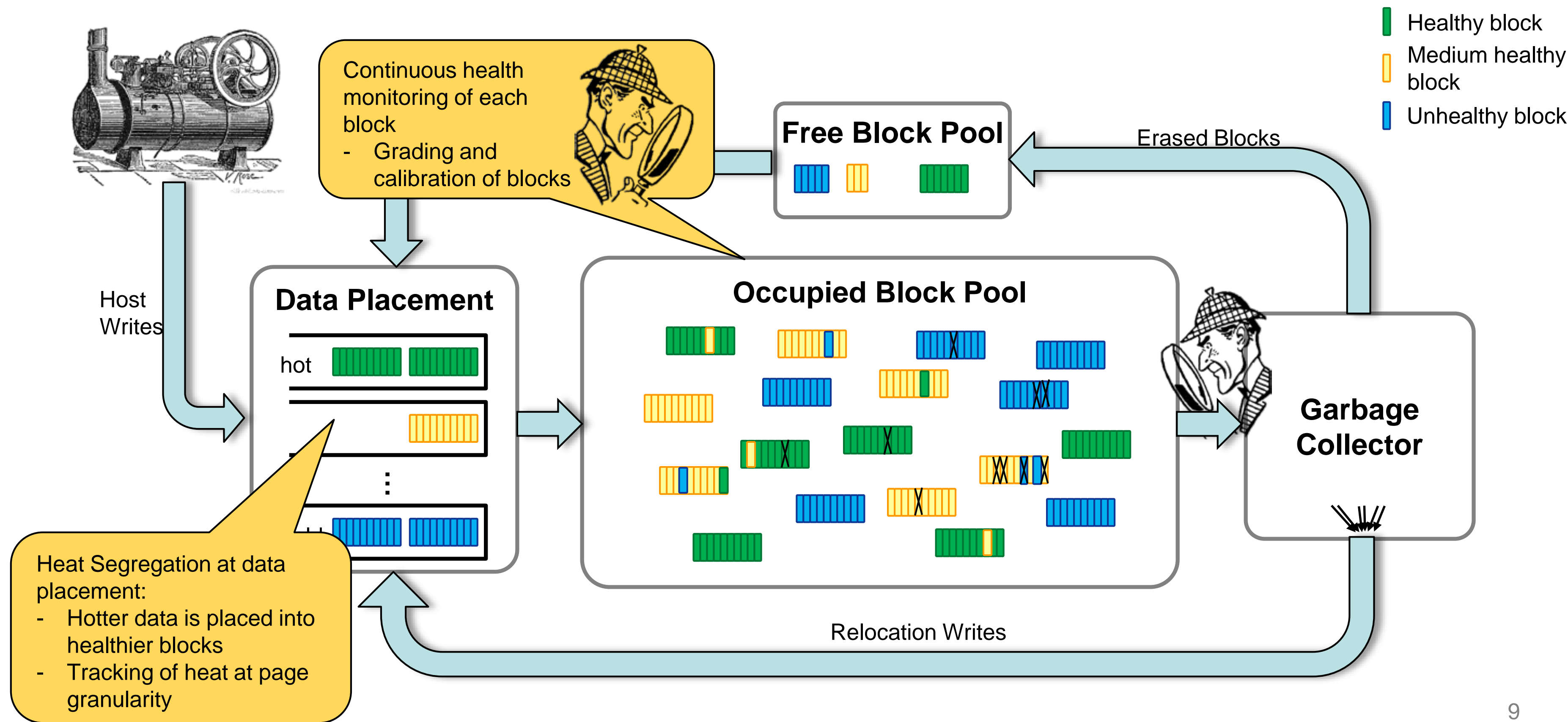


[1] Using Adaptive Read Voltage Thresholds to Enhance the Reliability of MLC NAND Flash Memory Systems,
N. Papandreou, T. Parnell, H. Pozidis, T. Mittelholzer, E. Eleftheriou, C. Camp, T. Griffin, G. Tressler, A. Walls, GLSVLSI 2014

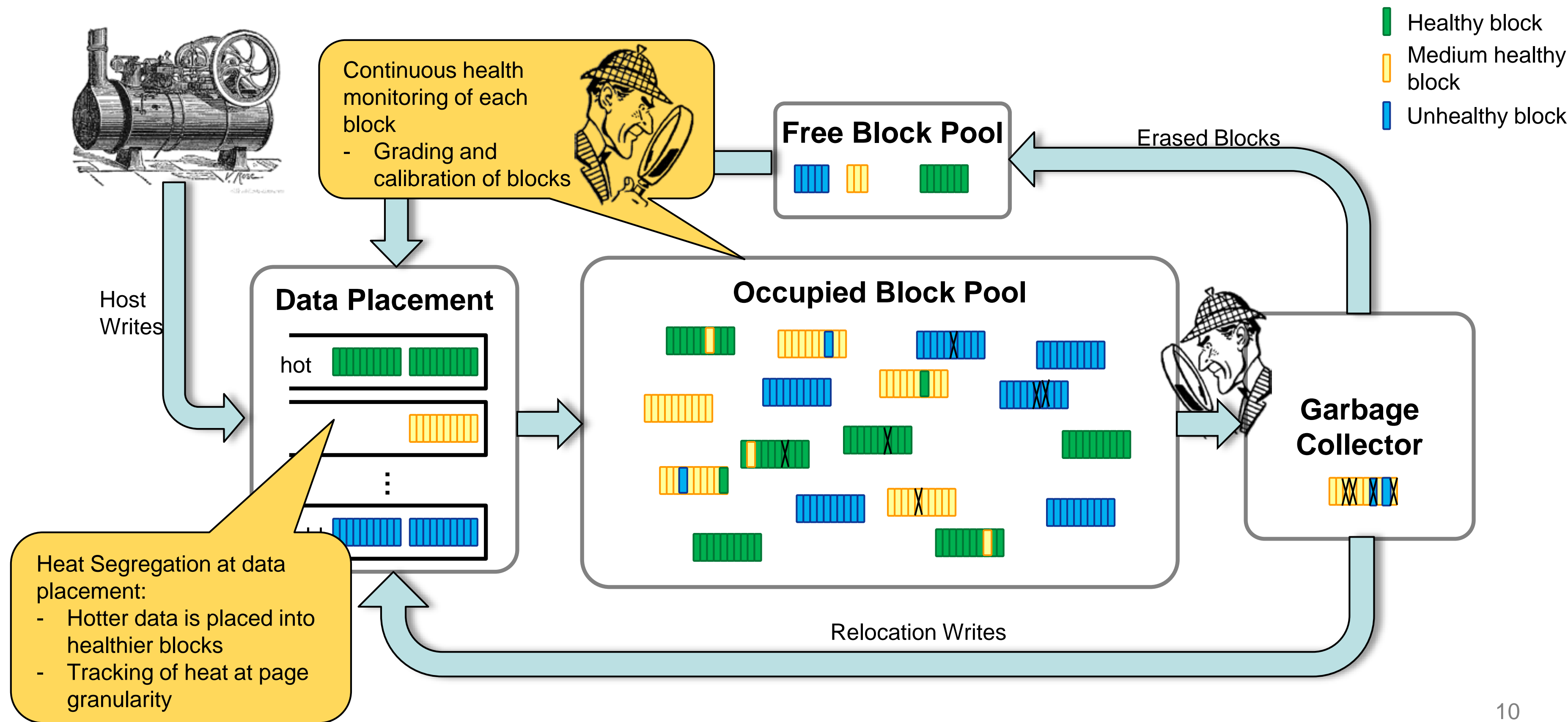
Data Placement with heat segregation and health binning



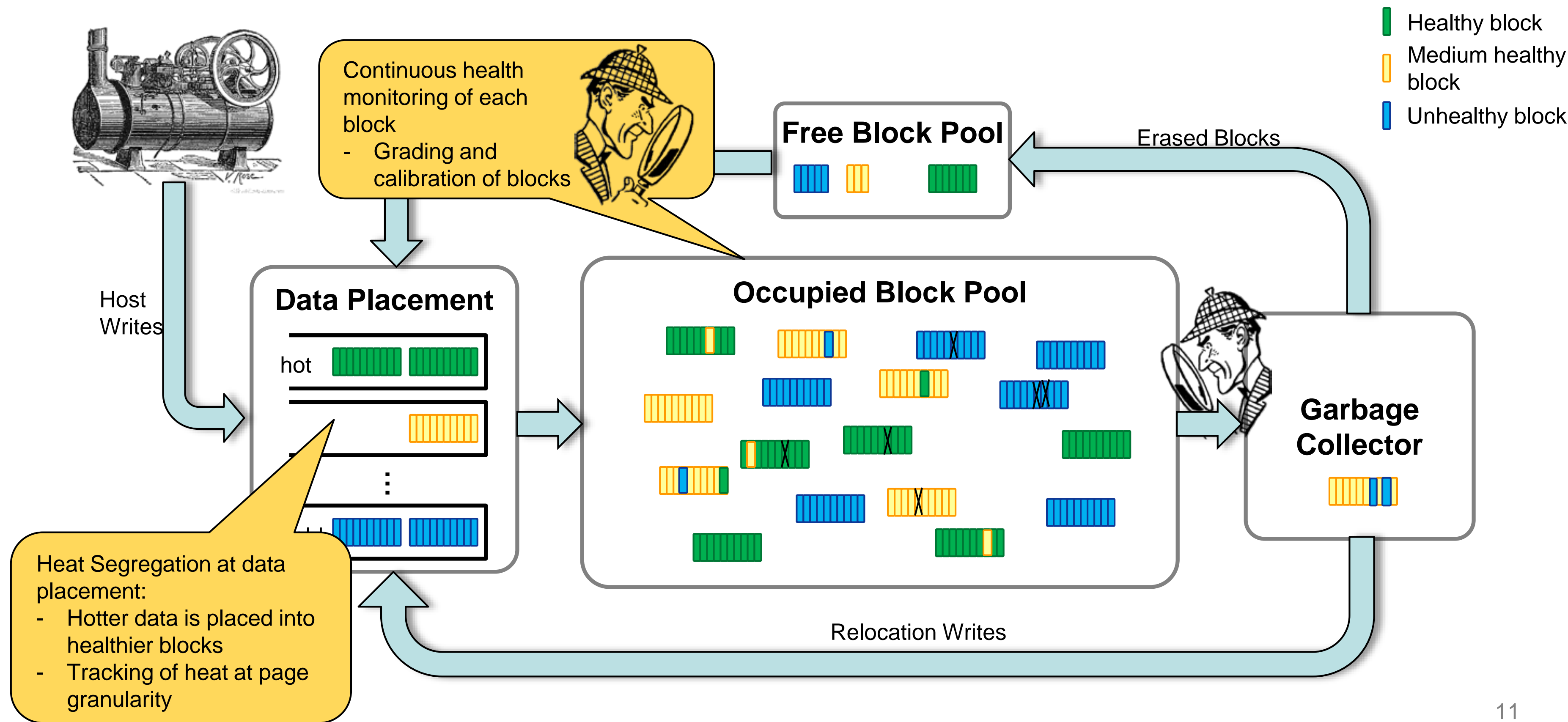
Data Placement with heat segregation and health binning



Data Placement with heat segregation and health binning

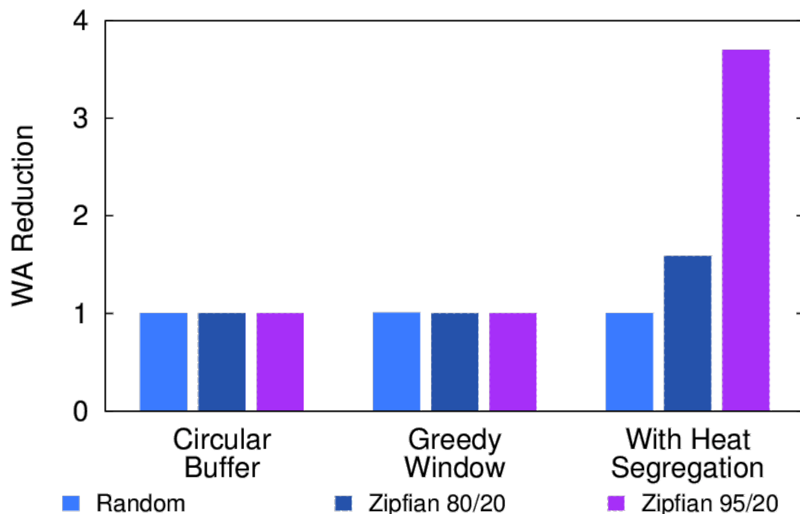


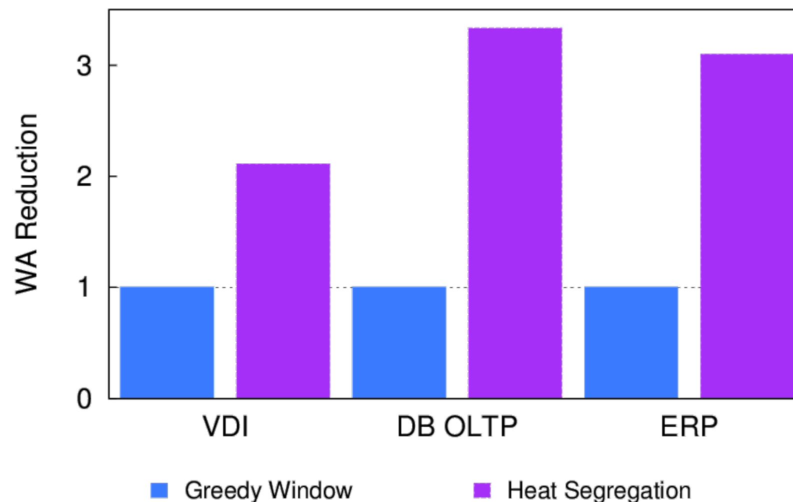
Data Placement with heat segregation and health binning



Write Amplification Results

- Write amplification results obtained from a simulation environment [▸ C.31] and real flash cards.
- Write amplification reduction of more than 3x in skewed workloads compared to baseline Circular Buffer and Greedy Window garbage collectors





Flash Health Evaluation Methodology

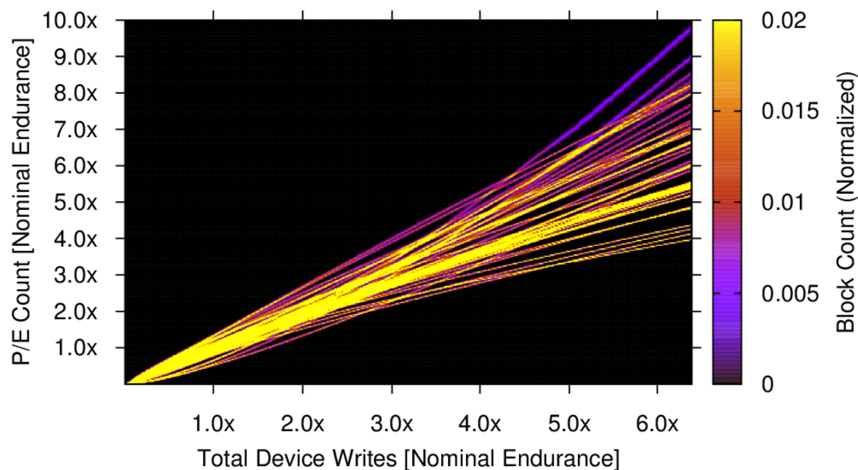
- We cannot wear out a real flash card within reasonable amount of time. Therefore design and analysis of our flash management algorithms w.r.t. endurance has been done on a simulator before integration and verification in an enterprise-level all-flash array. See presentation **[▸ C-31]** for more details on the simulation environment.
- Large-scale characterization data from different Flash generations and manufacturers (20-15nm MLC Flash) are used to create an accurate block wear model (e.g., Gaussian mixture model).
- Skewed write workloads (Zipfian) and real world traces used to determine write amplification and endurance characteristics
  - Note, Zipfian-type skewed write workloads have no static data. This is reasonable as our design point only adds minimal WA due to static wear leveling.

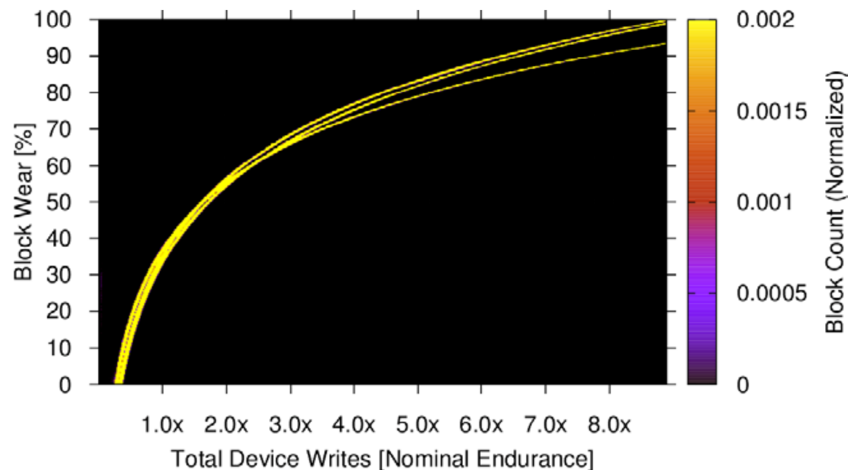
Flash Memory SUMMIT

Health Evaluation Results

IBM

- P/E cycle and wear distributions using heat segregation and health binning using a highly skewed workload (Zipfian 95/20):

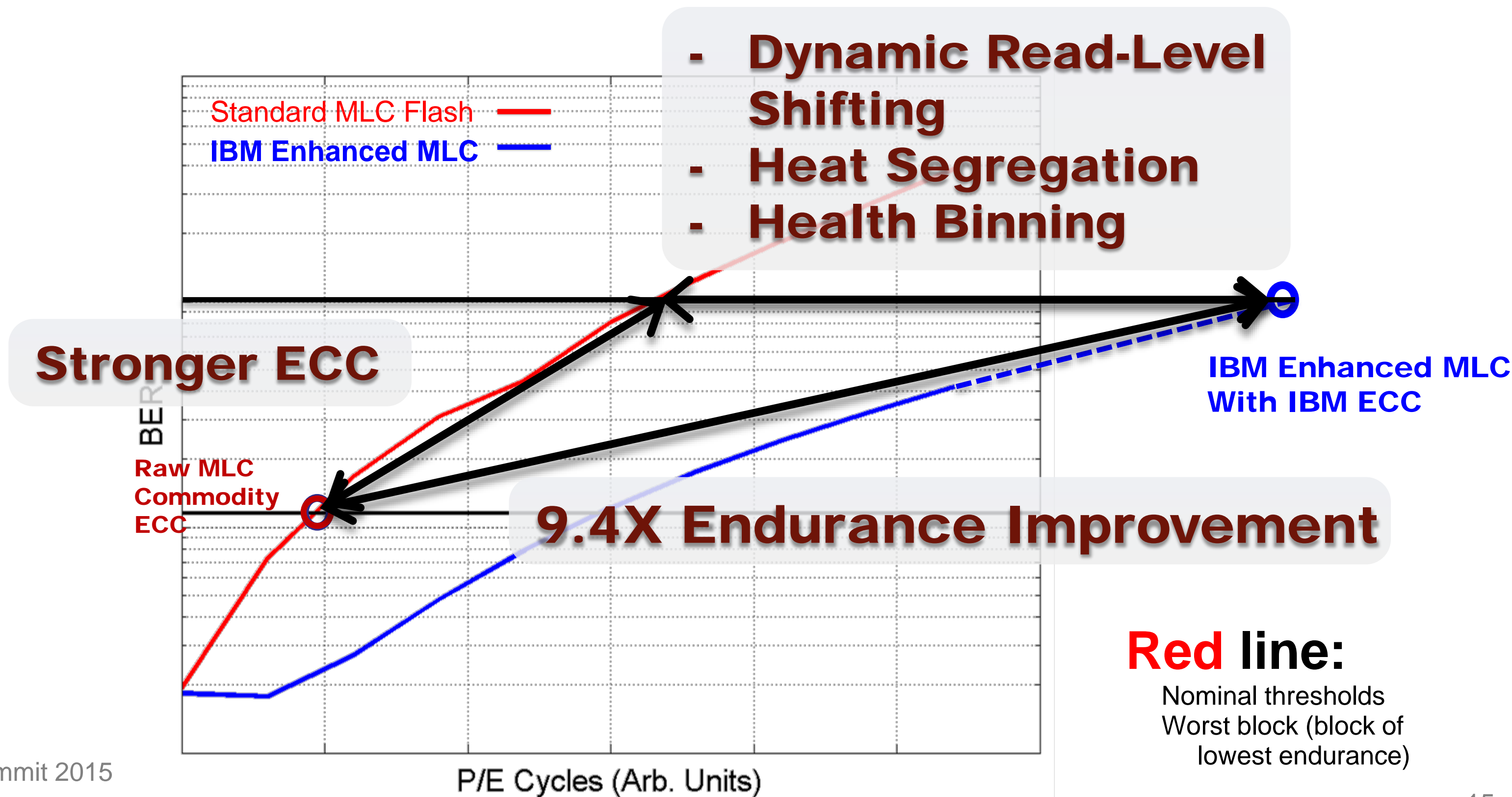




- P/E count distribution is narrow at beginning and widens with increasing number of writes
- At end-of-life, best block exhibits roughly 2.5x P/E cycles than the worst one

- Generally very narrow block wear distribution which marginally widens with increasing writes
- Best blocks consumed 93% of their endurance at end-of-life
- Up to 57% improvement in endurance

IBM Enhanced MLC



Red line:
Nominal thresholds
Worst block (block of lowest endurance)

Flash Memory SUMMIT

Conclusion

IBM

To achieve same endurance targets for enterprise all-flash arrays with new generations of NAND Flash:

- Stronger error correction codes (ECC) are required that maintain consistent latency characteristics for the entire lifetime.
- Threshold voltages must be continuously adapted in the background.
- Increasing differences in wear characteristics of blocks must be balanced based on health characteristics of blocks, not P/E cycles.
- Smart data placement combines heat segregation and health binning to reduce write amplification and increase endurance

Questions…

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