Designing Enterprise Controllers with QLC 3D NAND

Roman Pletka, Radu Stoica, Nikolas Ioannou, Sasa Tomic, Nikolaos Papandreou, Haralampos Pozidis

IBM Research – Zurich Research Laboratory

Outline

- Background
  - MLC and TLC NAND Flash controller design options for enterprise storage
  - Existing hybrid SSD controller architectures: advantages and limitations
- Motivations for a hybrid SLC/QLC controller
  - Can the endurance gain of SLC-mode outweigh the reduced capacity?
- SLC/QLC controller design options
  - Analyze potential endurance using a modeling approach
  - Comparison of different controller architectures

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

MLC/TLC NAND in enterprise storage

- How to achieve enterprise-level endurance with MLC/TLC NAND flash?
  - Strong error correction codes.
  - Dynamic threshold voltage shifting [1].
  - Heat segregation co-locates LBAs with similar update frequencies into the same block to reduce internal write amplification [2].
  - Health binning moves endurance limit from worst blocks to the average of all blocks by placing write hot data onto better blocks [3].

- But this is not sufficient for QLC NAND Flash! Why?
  - Manufacturer specified QLC endurance in the order of 800 – 1k program/erase cycles.
  - QLC blocks in SLC mode have ~ 40x more endurance specified, but 4x less capacity.
  - Number of levels to distinguish doubles from TLC => tighter margins increase RBER resulting in secret sauce being less effective…

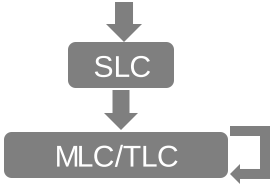
[1] Using Adaptive Read Voltage Thresholds to Enhance the Reliability of MLC NAND Flash Memory Systems, N. Papandreou et al., GLSVLSI 2014

[2] Holistic Flash Management for Next Generation All-Flash Arrays, R. Pletka et al., FMS 2015

[3] Health Binning: Maximizing the Performance and the Endurance of Consumer-level NAND Flash, R. Pletka and S. Tomic, SYSTOR 2016

Background on hybrid SSD controllers

1st Generation: Fixed-size SLC cache for MLC/TLC NAND Flash [4-7]



- Controller characteristics:
  - Use a small region of the Flash as a **static SLC cache**.
  - Data is first written to SLC, still valid data in SLC is later destaged to MLC/TLC when SLC cache is full
- Benefits:
  - Low write latency and write higher throughput for bursty write workloads with significant idle times
  - Read latency reduction for data read from SLC
  - Some manufacturers provide on-chip copy from SLC to MLC/TLC [8]
- Challenges:
  - Write speed drops significantly when SLC cache is full
  - Requires reasonably good endurance of MLC/TLC as many writes will eventually be destaged
  - Capacity reduction increases cost

How much can we improve endurance with such a design?

[4] A Hybrid Approach to NAND-Flash-Based Solid-State Disks, L. Chang et al., IEEE Trans. on Comp. 2010

[5] Improving performance and lifespan of MLC flash memory using SLC flash buffer, S. Im et al., Journal on Syst. Archit. 2010

[6] Samsung Solid State Drive TurboWrite Technology, 2013

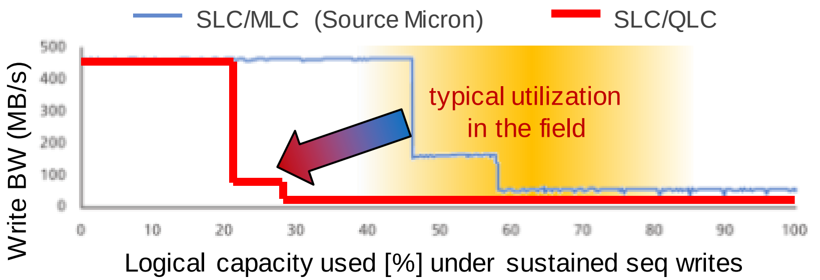
[7] Samsung SSD 850 Review , K. Vättö, AnandTech 2014, https://www.anandtech.com/show/8747/samsung-ssd-850-evo-review/2

[8] Leverage TLC technology to Advance your Corporate Environment, E. Bek, FMS 2014

Background on hybrid SSD controllers

2nd Generation: Adaptive SLC caching (i.e., Dynamic Write Acceleration DWA [9,10])

- Controller characteristics:
  - Dynamically switch flash programming modes SLC ⇔ MLC/TLC at the block level.
  - Number of blocks in SLC and MLC/TLC mode depends on logical capacity used.
  - Migration from SLC to MLC/TLC done in the background at idle time.
- Benefits:
  - Higher write throughput for sustained write workloads up to a certain limit (e.g., especially when utilization is low).
  - Read latency reduction for data read from SLC.
  - No user capacity reduction.
- Challenges:
  - Write speed drops when utilization reaches a certain level.
  - Requires reasonably good endurance of MLC/TLC as many writes will eventually be destaged.
  - Blocks remain statically assigned at a given utilization.
    => Endurance dictated by the minimum endurance of the pools
  - Requires idle times to cleanup SLC cache.
- Moving from SLC-MLC towards SLC-QLC caching:



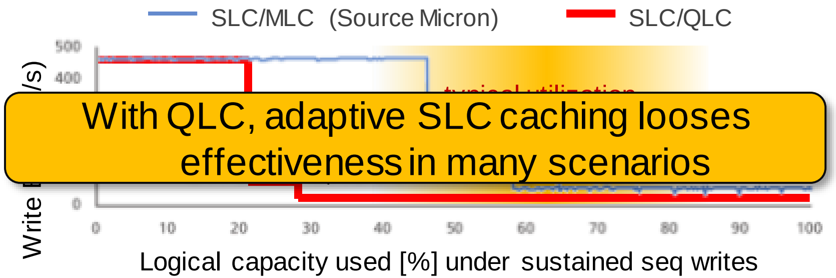
[9] Utilization-Aware Self-Tuning Design for TLC Flash Storage Devices, M. Yang et al., TVLSI 2010

[10] Optimized Client Computing With Dynamic Write Acceleration, D. Glen, Micron, 2014, https://www.micron.com/~/media/documents/.../brief_ssd_dynamic_write_accel.pdf

Background on hybrid SSD controllers

2nd Generation: Adaptive SLC caching (i.e., Dynamic Write Acceleration DWA [9,10])

- Controller characteristics:
  - Dynamically switch flash programming modes SLC ⇔ MLC/TLC at the block level.
  - Number of blocks in SLC and MLC/TLC mode depends on logical capacity used.
  - Migration from SLC to MLC/TLC done in the background at idle time.
- Benefits:
  - Higher write throughput for bursty write workloads, especially when utilization is low.
  - Read latency reduction for data read from SLC.
  - No user capacity reduction.
- Challenges:
  - Write speed drops when utilization reaches a certain level.
  - Requires reasonably good endurance of MLC/TLC as many writes will eventually be destaged.
  - Blocks remain statically assigned at a given utilization.
    => Endurance dictated by the minimum endurance of the pools
  - Requires idle times to cleanup SLC cache.
- Moving from SLC-MLC towards SLC-QLC caching:



With QLC, adaptive SLC caching looses effectiveness in many scenarios

[9] Utilization-Aware Self-Tuning Design for TLC Flash Storage Devices, M. Yang et al., TVLSI 2010

[10] Optimized Client Computing With Dynamic Write Acceleration, D. Glen, Micron, 2014, https://www.micron.com/~/media/documents/.../brief_ssd_dynamic_write_accel.pdf

Towards an SLC/QLC controller design

How can we do better ?

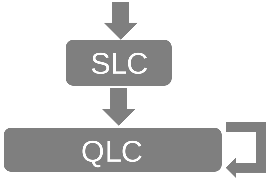
- Rethink data placement and pool sizing!

Modeling Approach:

- We assume an optimal controller design with the following properties:
  - Tracking of write hot data set (update frequencies) -> Optimal pool sizing
  - For the modeling we assume block swapping can be done at any point in time without penalty.
- Then determine endurance upper bounds for different workload types with optimal pool sizes:
  - We evaluate write amplification of each pool to evaluate device endurance for different drive utilization.

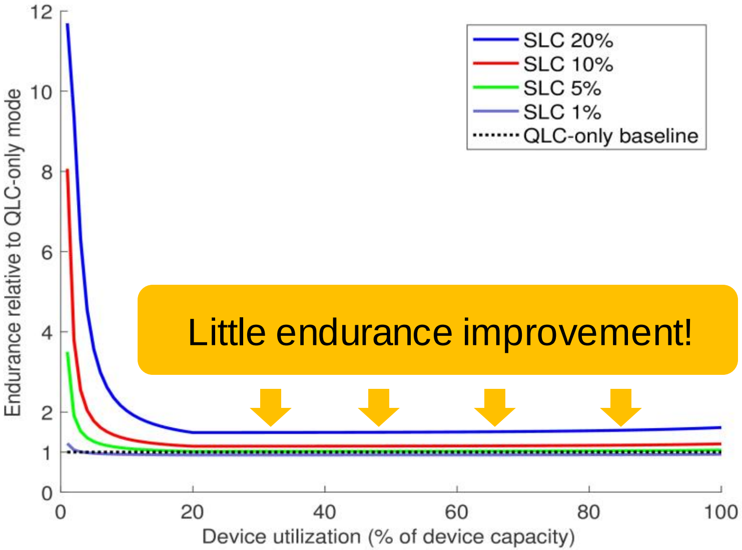
SLC/QLC controller with destage buffer

Fixed-size destage buffer (1st generation):



Experiment:

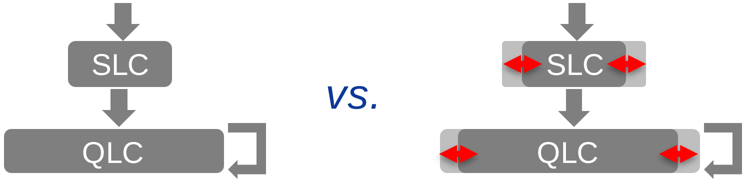
- Random write workload to 20% of the address space.
- Remaining utilized space holds static data.
- Controller parameters:
  - 1, 5, 10, 15% of physical blocks set to SLC mode.
  - 20% total over-provisioning irrespective of the SLC size.



Little endurance improvement!

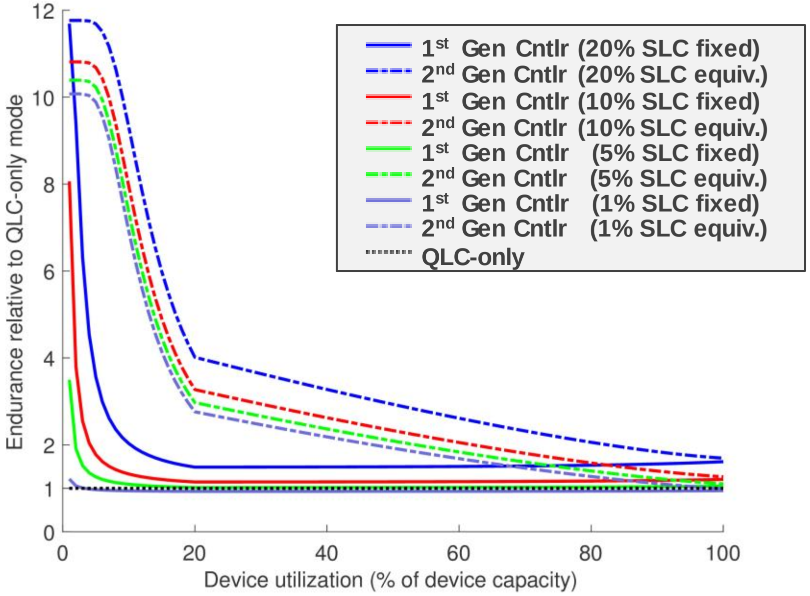
SLC/QLC controller with destage buffer

Fixed vs. optimally sized destage buffer:



Experiment:

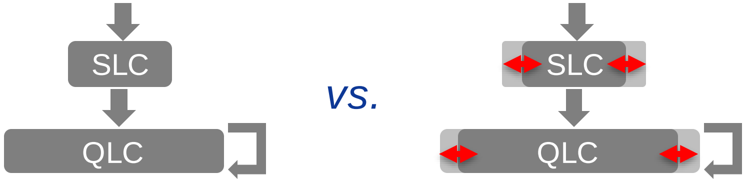
- Random write workload to 20% of the address space.
- Remaining utilized space holds static data.
- 1st Generation Controller with fixed SLC cache:
  - 1, 5, 10, 15% of physical blocks set to SLC mode.
  - 20% total over-provisioning irrespective of the SLC size.
- 2nd Generation Controller with adaptive SLC cache:
  - Assumes optimal SLC/QLC ratio for given utilization.



Significant endurance gain w.r.t. fixed SLC cache, however, endurance improvement are diminishing with higher device utilization.

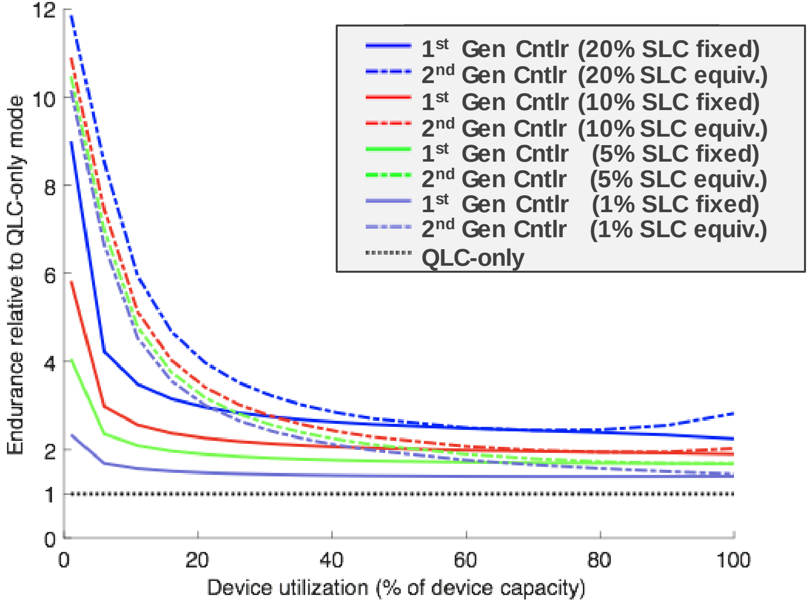
SLC/QLC controller with destage buffer

Fixed vs. optimally sized destage buffer:



Experiment:

- **Zipfian writes (80/20)** to utilized LBA space.
- Remaining utilized space holds static data.
- 1st Generation Controller with fixed SLC cache:
  - 1, 5, 10, 15% of physical blocks set to SLC mode.
  - 20% total over-provisioning irrespective of the SLC size.
- 2nd Generation Controller with adaptive SLC cache:
  - Assumes optimal SLC/QLC ratio for given utilization.



Significant endurance gain w.r.t. QLC only, additional endurance gains with adaptive SLC cache. However, endurance gains drop fast at low utilization.

Further related aspects

Other approaches:

- Phoenix: Reviving MLC blocks as SLC to extend NAND Flash devices lifetime, X. Jimenez et al., DATE 2013
  - Switch unreliable MLC blocks to SLC mode.
  - Endurance gains of 3-17% are achievable, but capacity of the device shrinks at the same time.

- Data compression:
  - Use data compression to reduce write amplification and get additional spare space. This will result in increased endurance.
  - The compression engine has to match the drive throughput which may be hard to achieve in a low-power controller.
  - No benefits for encrypted or already compressed data.

Conclusion

- An appropriate SLC/QLC controller design can achieve significant endurance gains even under high capacity utilization.
  - Fixed-size SLC destage buffers only achieve marginal endurance improvements.
  - Adaptive SLC caching is better than a fixed-size SLC destage buffer, but endurance improvements diminish with higher device utilization.
- Optimal cache sizing and data placement approaches are fundamental to enable QLC in enterprise storage systems.
  - We believe that, combined with existing Flash management technologies endurance targets for enterprise SSD controllers with QLC NAND Flash can indeed be achieved.
- Future work:
  - Study implications on the implementation complexity.

Thank You !

Questions ?

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