Achieving Latency and Reliability Targets with QLC in Enterprise Controllers

Roman Pletka, Nikolaos Papandreou, Radu Stoica, Haris Pozidis, Nikolas Ioannou
Tim Fisher, Aaron Fry, Kip Ingram, Andrew Walls

IBM Research Europe – Zurich Research Laboratory
IBM Systems – Houston TX

Flash Memory Summit

Outline

- Background
  - 3D NAND flash properties and challenges
  - Why is 3D QLC NAND flash so challenging?
- Mitigation techniques and strategies in a QLC controller
  - Data placement
  - Block calibration
  - Hybrid controller architecture
  - Read heat separation
- Conclusion

Background on 3D NAND Flash characteristics

- With increasing number of bits stored per cell, read, program, and erase latencies have been increasing in 3D NAND flash devices
- The manufacturer specified endurance for QLC flash has dropped to a level where traditional flash management technologies are not sufficient for enterprise storage applications
- Significant cost reductions with more bits per cell

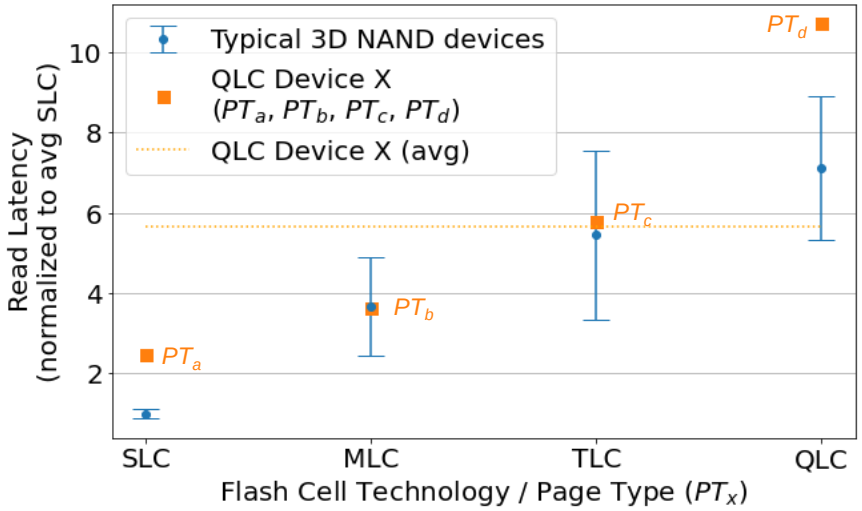
Operation	Flash cell technology			
	SLC	MLC	TLC	QLC
Page read	20 – 25 µs	55 – 110 µs	75 – 170 µs	120 – 200 µs
Page program	50 – 100 µs	0.4 – 1.5 ms	0.8 – 2 ms	2 – 3 ms
Block erase	2 – 5 ms	5 – 10 ms	10 – 15 ms	15 – 20 ms
Endurance	100'000	15'000	3'000 – 5'000	800 – 1'500
Cost per bit	–	+	++	+++
Density	1 – 256 Gb	16Gb – 2Tb	128Gb – 8Tb	1 – 16Tb

Source: [1] Understanding the design trade-offs of hybrid flash controllers R. Stoica, R. Pletka, N. Ioannou, N. Papandreou, S. Tomic, H. Pozidis, MASCOTS 2019

3D NAND Flash read latency characteristics

Read latency characteristics:

- Read latency varies as a function of the page type [2]
- Number of threshold voltage distributions increases exponentially with the number of bits stored per cell
- More read voltage levels must be tested upon a read operation affecting read latency



[2] P. Breen, N. Papandrou, G. Tressler, Component-Level Characterization of 3D TCL, QLC, and Low –Latency NAND, Flash Memory Summit 2019

Overview of mitigation techniques for QLC controllers

QLC challenges	Mitigation techniques
Erase latency increase	Erase suspend
Program latency increase	Program suspend
Read latency increase	- SLC cache/tier - Physical data layout (Page straddling) - Continuous background calibration of read voltage levels - Read heat separation - Compression

Endurance limitations	- Background calibration of read voltage levels
	- Write heat separation [4]
	- Hybrid controller architecture:
	  - Operate blocks in either a high-density multi-bit mode (e.g., QLC) or in a high-performance SLC mode
	  - dynamic SLC and QLC pools
	  - workload-optimized data placement
	- Wear leveling such as Health binning [5] within and between SLC and QLC tiers
	- Compression

[4] R. Pletka, I. Koltsidas, N. Ioannou, S. Tomić, N. Papandreou, T. Parnell, H. Pozidis, A. Fry, T. Fisher, "Management of next-generation NAND flash to achieve enterprise-level endurance and latency targets", ACM Trans. Storage, vol. 14, no. 4, Dec. 2018

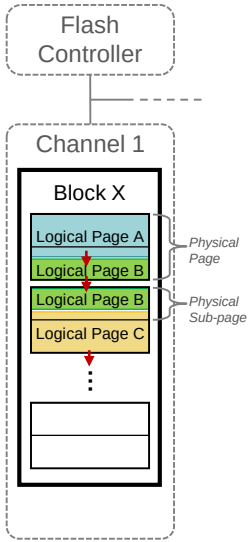
[5] R. Pletka, S. Tomic, "Health-Binning: Maximizing the Performance and the Endurance of consumer-level NAND flash", SYSTOR 2016

Page straddling

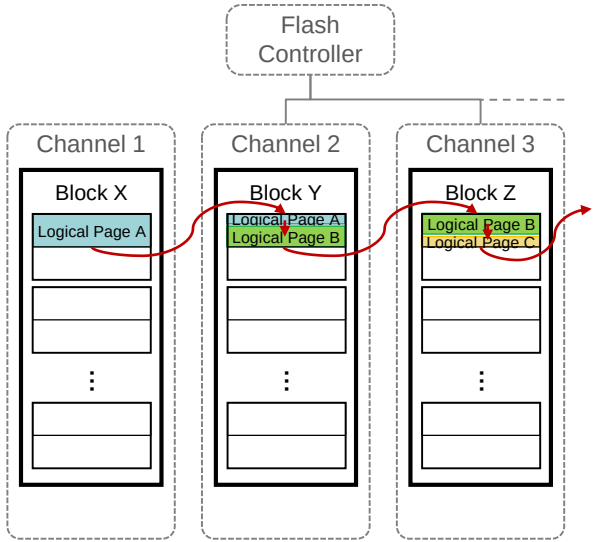
Intra-block vs. inter-channel data straddling

- Replace page read operation by sub-page reads on multiple channels executed concurrently
- Controller design considerations w.r.t the data layout:
  - Blocks on different channels must be organized into stripes
  - L2P mapping can be implicit for straddling logical pages.
  - Also works for compressed logical pages

Intra-block data straddling

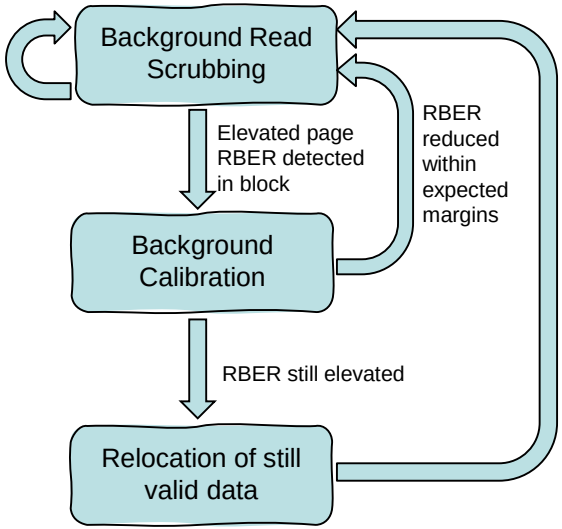


Inter-channel data straddling



Background Calibration

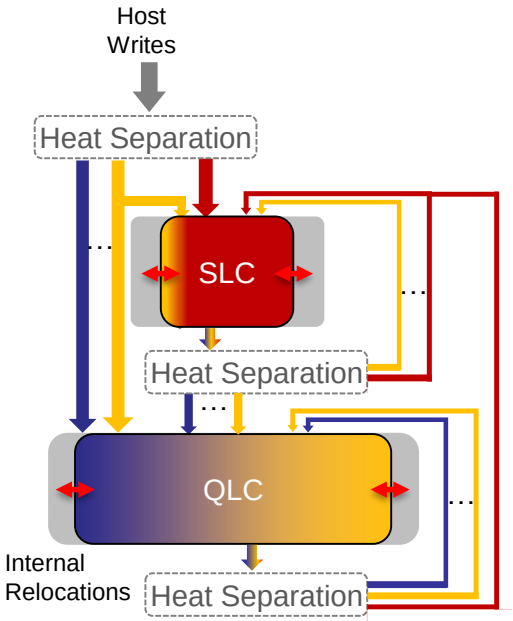
- Block calibration determines optimal read voltage levels continuously in the background
  - Calibration frequency is a function of:
    - Program-erase cycles of the block,
    - Number of reads the block has seen,
    - Retention time,
    - Page type, …
- Benefits:
  - Significant increase in endurance
  - Significant reduction of read retry operations
  - Due to the RBER reduction, ECC using hard-decision only can be used. As no soft-information is needed, read latency is reduced
  - Does not impact host read operations



Hybrid controller architecture for QLC Flash

Hybrid QLC controller design goals:

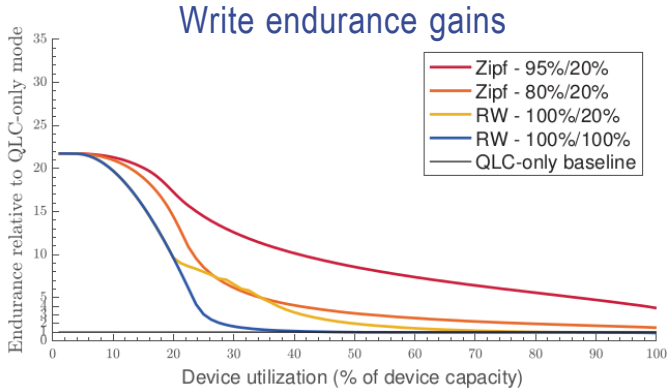
- Operate blocks in either a high-density multi-bit mode (e.g., QLC) or in a high-performance SLC mode
- Blocks can be moved between pools to adapt to workload changes (pool resizing) or for wear leveling purposes
- Data placement is done based on workload properties
  - Write heat separation: The co-location of data with similar update frequencies in the same block reduces internal write amplification [4]
  - Read heat separation re-arranges data in QLC blocks such that frequently read data is placed on faster pages (e.g., requiring fewer read thresholds to be tested)
  - Direct data placement to QLC avoids data being first written to SLC and later destaged to QLC

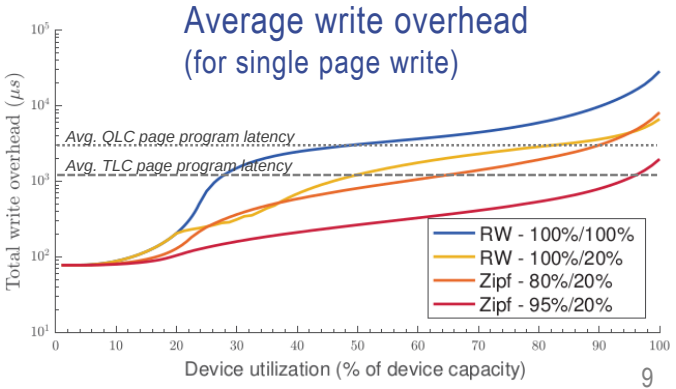


Benefits from a hybrid controller architecture

Modeling experiment (see [1] for details):

- Workloads:
  - Zipfian writes with skew factors 95/20 and 80/20
  - Uniform random writes to whole or partial data set
- Hybrid controller with full write heat separation
  - Controller determines optimal SLC/QLC ratio for given utilization
  - Controller determines optimal size of hot dataset to keep in SLC
- Results:
  - Significant improvements in both, endurance and write performance for skewed workloads
  - Skewed workloads outperform a TLC only controller even when device utilization is high





[1] R. Stoica, R. Pletka, N. Ioannou, N. Papandreou, S. Tomic, H. Pozidis, Understanding the design trade-offs of hybrid flash controllers, MASCOTS 2019

Information encoding in NAND flash cells

Information encoding in NAND flash cells uses so called Gray coding schemes (reflected binary coding):

- The binary values are ordered in such a way that two successive values only differ in a single bit
- Gray coding guarantees that only one physical page is affected by a single bit flip.
- There are many possible Gray coding schemes which differ in the number of read voltage levels that must be tested upon reading a page of a particular page type
- Today, NAND flash devices only implement a single coding scheme that cannot be changed.
- Example Gray coding scheme for QLC with 4 page types PT_a,…,PT_d:

Page Type	E	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L15	# Read voltage levels
PT_a (LSB)	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	1 → fastest page
PT_b	1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	1	2
PT_c	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	4
PT_d (MSB)	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	8 → slowest page

The potential of read heat separation

Read heat separation goals:

- Place frequently read data on faster pages
- Place data rarely read on slower pages

Achievable latency reductions depend on:

- The Gray coding scheme used
- Read skew of the workload
- Read and write workload mix

Maximum achievable average read latency reductions using {1, 2, 4, 8} Gray coding

(page types correspond to the read latencies of the QLC device X on slide 4)

Read Workload	Average Read Latency Reduction
Random	0.0%
Zipf 70/30	28.3%
Zipf 80/20	41.5%
Zipf 95/20	53.2%

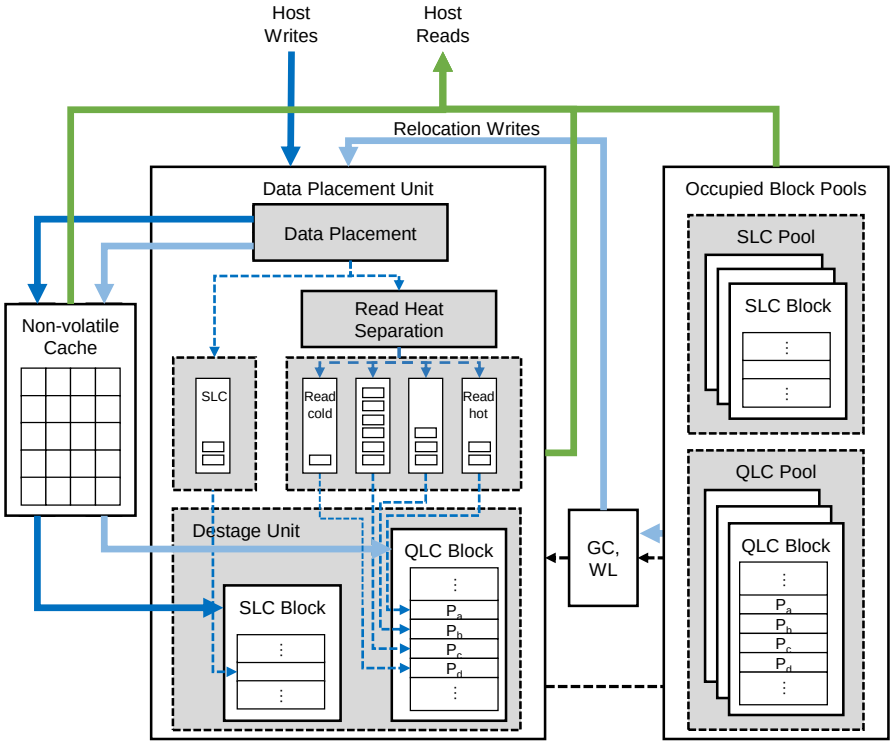
Faster than average TLC latency !

[3] R. Pletka, N. Papandreou, R. Stoica, H. Pozidis, N. Ioannou, T. Fisher, A. Fry, K. Ingram, A. Walls, Improving NAND flash performance with read heat separation, MASCOTS 2020

Controller architecture for read heat separation

General concept:

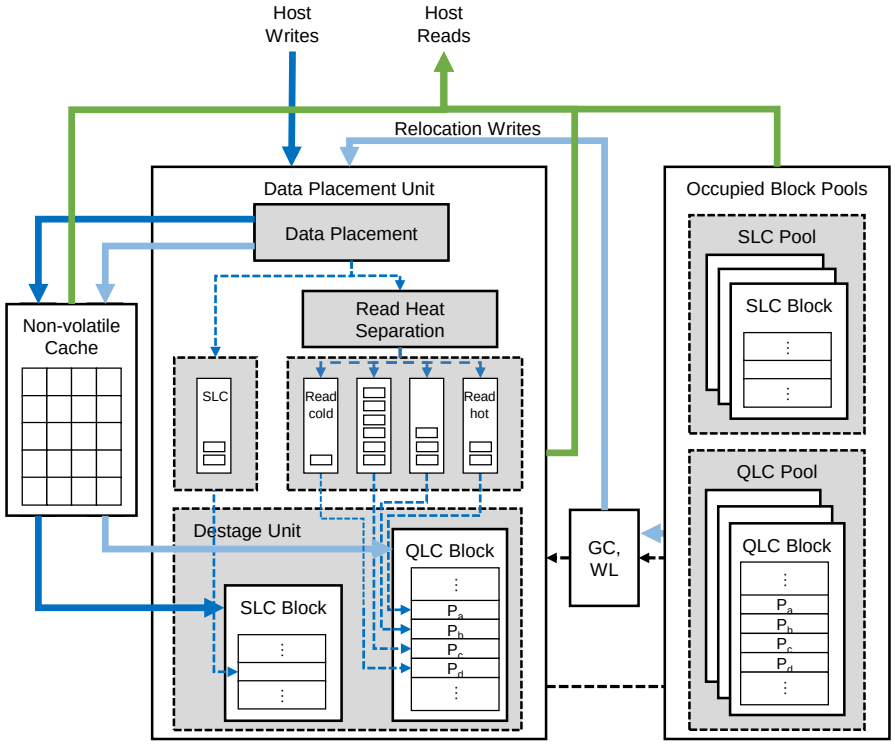
- Read heat information is tracked on every host read operation
- Data placement unit performs read heat separation based on read heat information for all write operations
  - 1 SLC command queue
  - 4 QLC command queues (one for each page type)
- Preferably, there are as many instances of data placement unit as there are write heat streams in the controller



Controller architecture for read heat separation

Write process:

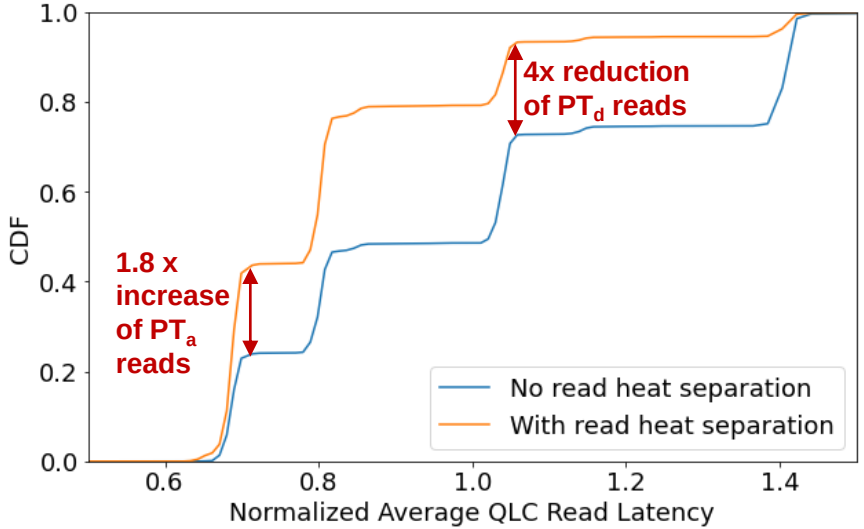
- Store write data in non-volatile destaged buffer (MRAM, battery-backed DRAM, SLC, …)
- Determine read heat (read heat information is tracked for each LBA)
- Perform read heat separation by placing the write request into the corresponding read heat queue
- Process writes by pulling from preferred read-heat queue. When preferred read-heat queue is empty, pull write request from the next best-matching non-empty read-heat queue.



Measurements from real SSDs

Read heat separation

- Experiment:
  - Preconditioned device with 80% of logical capacity used
  - 2 read workloads:
    - 95% of reads to only 5% of the written LBA space
    - 5% reads to 95% of the written LBA space
  - Uniform random write workload
- Results:
  - Reads from fastest PT_a pages increases by 1.8x
  - Reads from slowest PT_d pages reduced by 4x



17.1% **in average read latency reduction achieved**

Measurements from real SSDs

Workload:

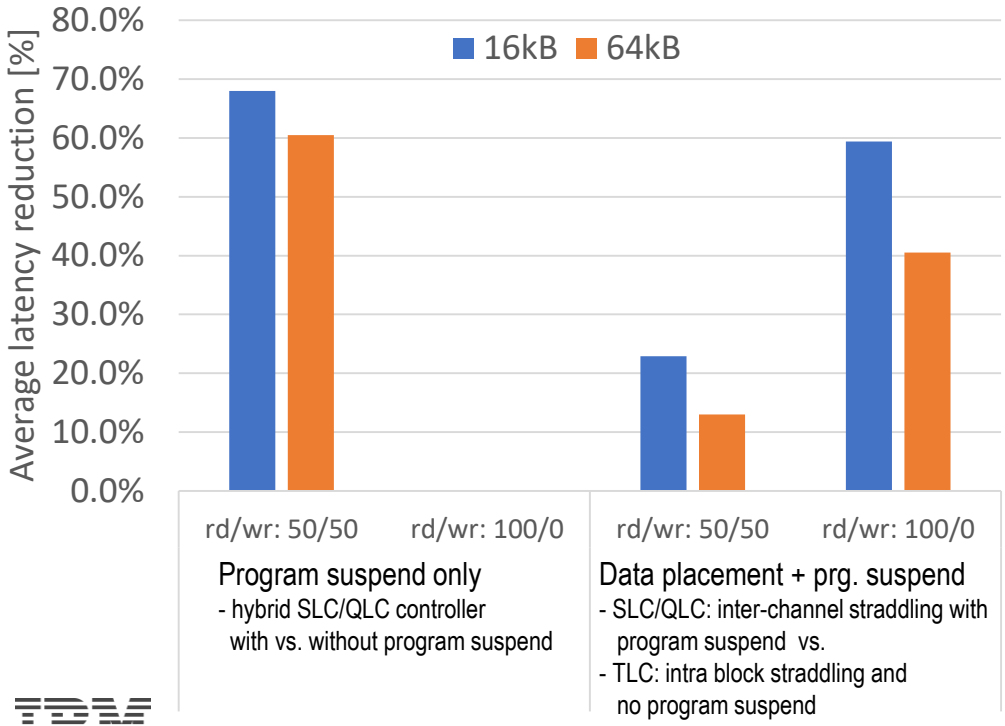
- Uniform random, high queue depth (constant IOPS)

Measurements:

- Average read latency reductions
  - From program suspend
  - From optimized data placement
  - 3D TLC only vs SLC/QLC NAND flash controller

Findings:

- Write suspend reduces the average read latency by up to 68% in mixed read/write workloads
- A hybrid SLC/QLC controller with intelligent data placement and program suspend can significantly outperform a traditional TLC controller



Conclusion

- Novel controller design approaches are essential to address the shortcomings of latest high-density 3D QLC NAND flash generations
- Key aspects to improve performance are
  - A dynamically resizable hybrid controller architecture
  - Data straddling on sub-page level
  - Workload dependent data placement using read and write heat separation
  - Continuous background block calibration
- Using these techniques, enterprise-level SSD controllers can be built using QLC flash that outperform previous NAND flash generations

Thank You!

Flash Memory Summit

Everything You Need To Know
For Success