Important Differences Between Consumer and Enterprise Flash Architectures

Robert Sykes
Director of Firmware

OCZ Technology

Introduction

- This presentation will describe key firmware design differences between consumer- and enterprise-class SSDs.
- Topics covered include end-to-end data protection, bad block recovery techniques for the latest NAND nodes such as Low-Density Parity Check (LDPC), advances in Flash Translation Layer techniques with a focus on firmware requirements for sustained performance, host interface buffering techniques, and a review of RAID and the flash interface.

Contents

- Consumer / Enterprise basics.
- End to End Data Protection
- Data Corruption / Correction (ECC, BCH, LDPC, RAID)
- Enterprise FTL architecture

Contents

- Consumer / Enterprise basics.
- End to End Data Protection
- Data Corruption / Correction (ECC, BCH, LDPC, RAID)
- Enterprise FTL architecture

Contents

- Consumer / Enterprise basics.
- End to End Data Protection
- Data Corruption / Correction (ECC, BCH, LDPC, RAID)
- Enterprise FTL architecture

Contents

- Consumer / Enterprise basics.
- End to End Data Protection
- Data Corruption / Correction (ECC, BCH, LDPC, RAID)
- Enterprise FTL architecture

Consumer vs Enterprise Basics

- There have been several discussions over the past few months on the forums asking; What is the difference between consumer and enterprise grade SSDs?
- Snapshot of some of the discussions on the forums describing the differences:
  - **Interface is the difference.**
    PCIe / SAS / NVMe are enterprise; SATA is consumer.
  - Consumer: Cost / Capacity / Performance / Data Integrity
  - Enterprise: Data Integrity / Performance / Capacity / Cost
  - Enterprise has data redundancy

Consumer / Enterprise Basics

- Enterprise has consistent performance
- Enterprise drives have greater endurance
- Enterprise drives have additional raw capacity
- Enterprise drives require custom applications for specific needs
- Client / consumer drives contain a single or only two SSDs; Enterprise could be comprised of several SSDs

Consumer / Enterprise Basics

- Turning to specifications for the answer: JEDEC helps to define the difference by specifying workload differences.
  - Data Usage Models
    - JEDEC 218 and 219 define client and enterprise usage models
      - Enterprise products will be used in heavy workload environments and are expected to perform.
        - » JESD 218A defines this as 24hrs / day at 55ºC with 3 months retention at 40ºC
      - Client drives generally have a lighter load.
        - » JESD 218A defines this as 8hrs / day at 40ºC with 1 year retention at 30ºC

Note: Data retention is defined here as the ability for the SSD to retain data in a power off state for a period of time.

Consumer / Enterprise Basics

- Enterprise SSDs should not really be defined by the interface, but by the application and system requirements.
  - However, many enterprise systems are not surprisingly built around fast hosts such as PCIe / NVMe as some enterprise systems require high-end performance (high IOPS) with sustained data rates.
- Questions to ask when designing your system:
  - Is the data WORM like (write once read many)?
  - Is the data volatile data (like swap data)?
  - Is the data redundant data (data backed up somewhere else)?

Consumer / Enterprise Basics

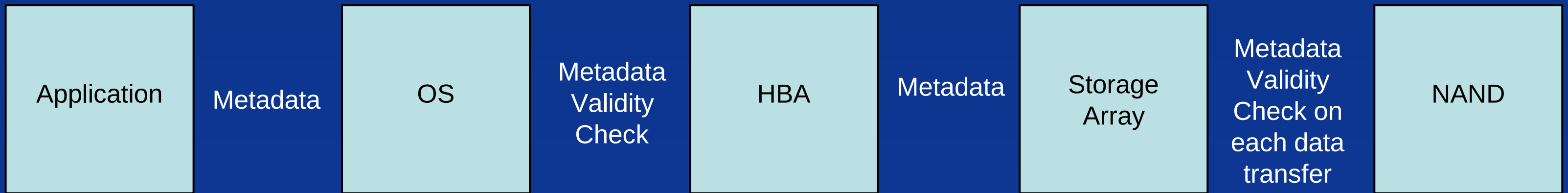
- Issues with sustained data read / writes:
  - During sustained activity in an enterprise system, the probability of a system level issues increases. This could be anything from a host memory issue, DRAM failure or uncorrectable flash page.

- The following methods can be used to prevent system issues:
  - End-to-End Data Protection
  - Read Retry
  - LDPC
  - RAID
  - Data Stirring
  - Temperature Throttling
  - Wear Leveling

End-to-End Data Protection

End-to-end Data Protection

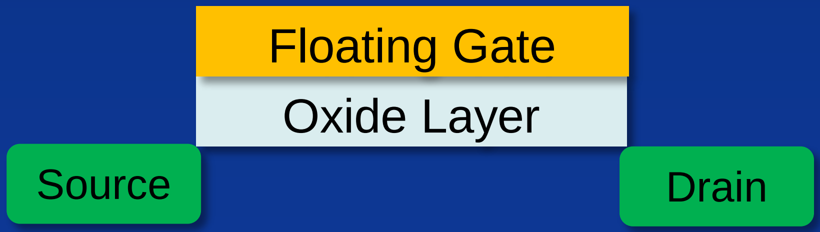
- End-to-End Data Protection
  - Why do we need it?
    - Protect against silent errors
      - » OS issues including device drivers problems
      - » Any issues within the HW or FW of the storage controller (i.e., bus issues, firmware writing the wrong data, etc.)

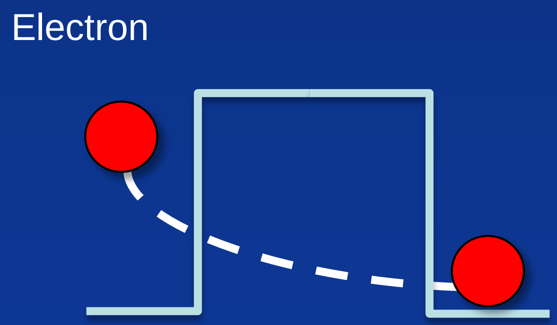


Data Corruption / Correction

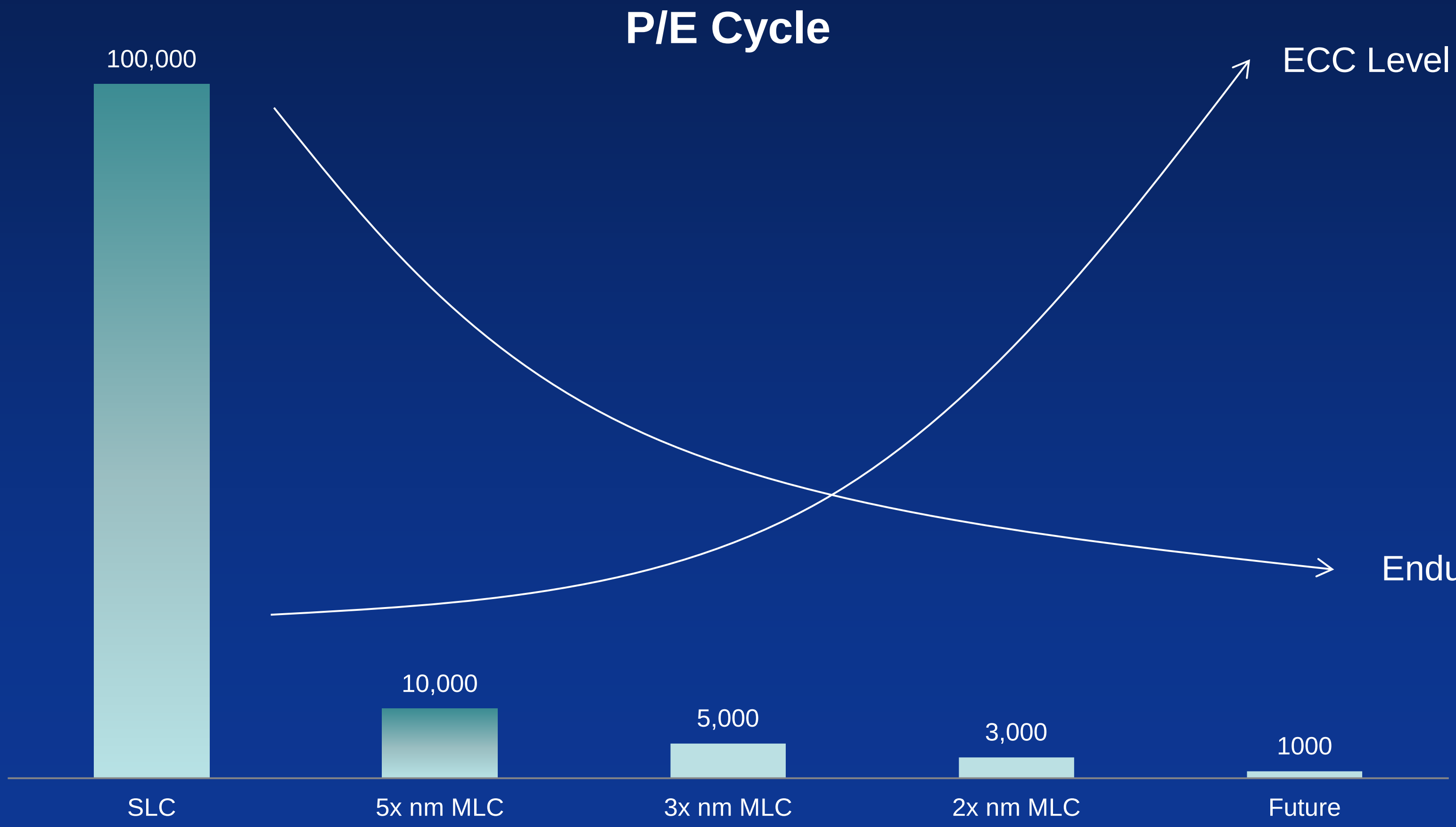
Data Corruption / Correction

- What are the possible causes?
  - NAND flash has an inherent weakness: (Quantum Mechanics!) as the NAND process shrinks so does the effectiveness of the oxide layer, in turn increasing the probability that a 0 becomes a 1 or vice versa when it shouldn't



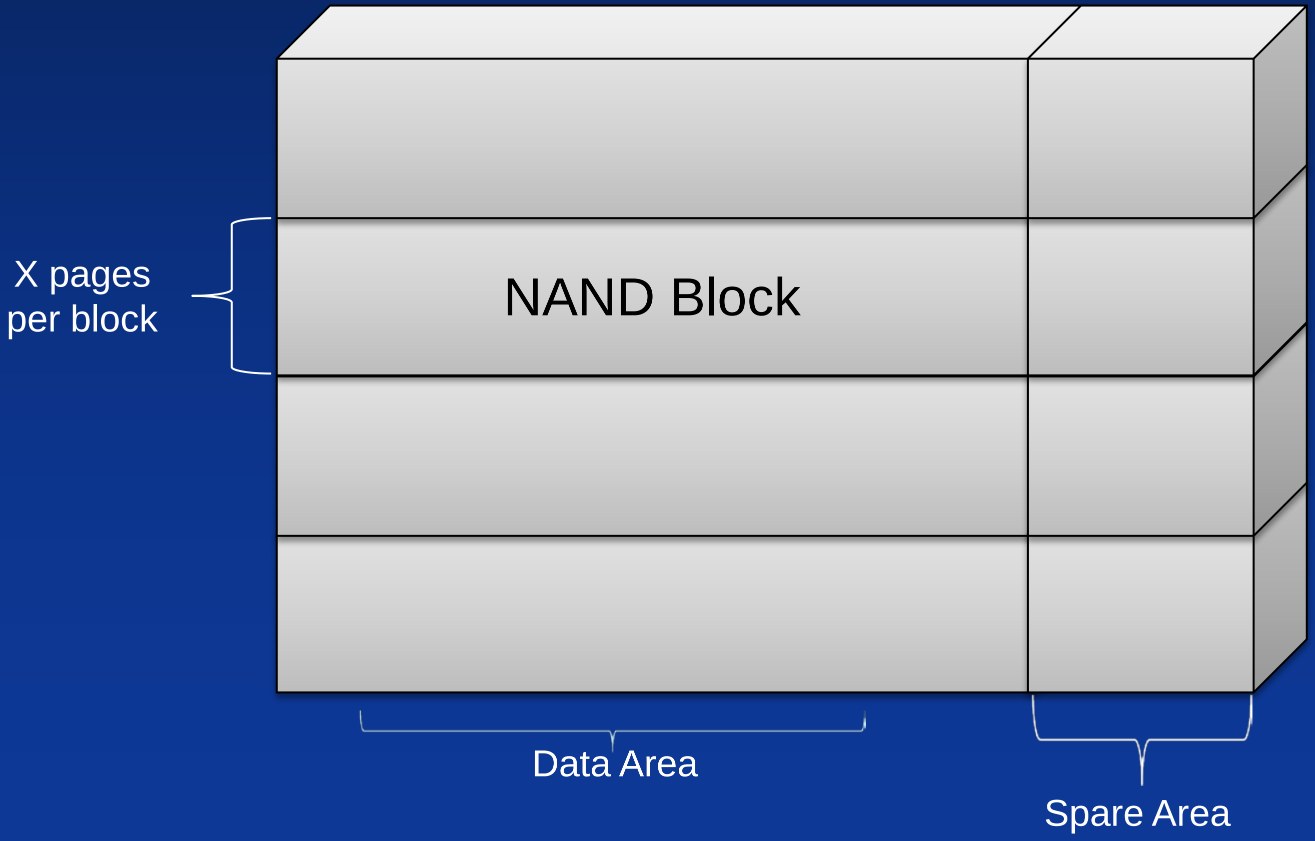


Data Corruption / Correction



Data Corruption / Correction

- How is ECC calculated?
  - Spare area is used for Bad Block markers (Mandatory), ECC and User Specific Data



Data Corruption / Correction

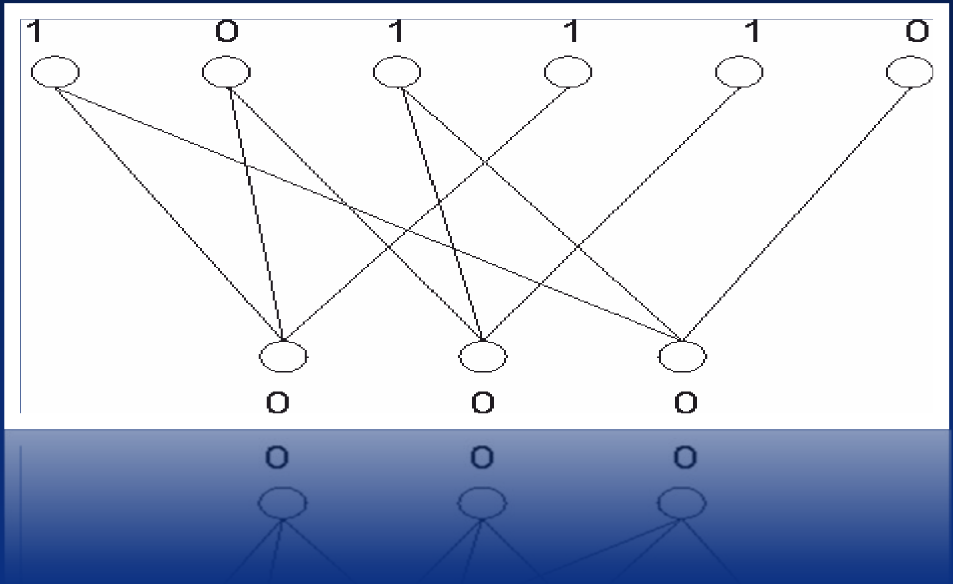
- The amount of spare area will depict the amount of ECC applied. For example:
  - 25nm MLC may have 450 bytes of spare area allowing 24bit ECC per 1K of data
  - 20nm MLC may have 750 bytes of spare area allowing 40bit plus ECC per 1K of data
- Note: The NAND vendors specify the level of ECC expected to meet the specified endurance of the drive
- Consumer drives could use BCH to achieve this level of ECC
  - BCH (Bose, Chaudhuri, Hocquenghem; invented 1959)

Data Corruption / Correction

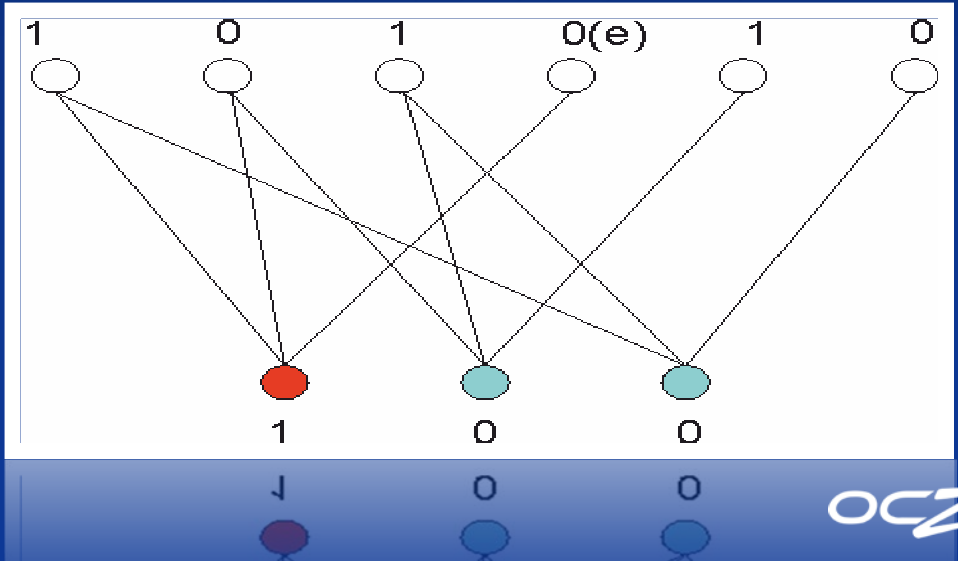
- Enterprise drives require greater endurance over consumer grade drives given the different requirements set out by the JEDEC standard; BCH may not be enough to meet these standards.
- One possible method to achieve this is to use LDPC (Low Density Parity Check)
  - LDPC can easily utilise soft information
  - LDPC can enhance endurance and retention
  - Requires new ratio(s) of user data to parity data
  - Compare with SNR/Frame Error Rate graphs, not a simple N-bit correction.
- How does LDPC work…..

Data Corruption / Correction

- LDPC code word
  - satisfied

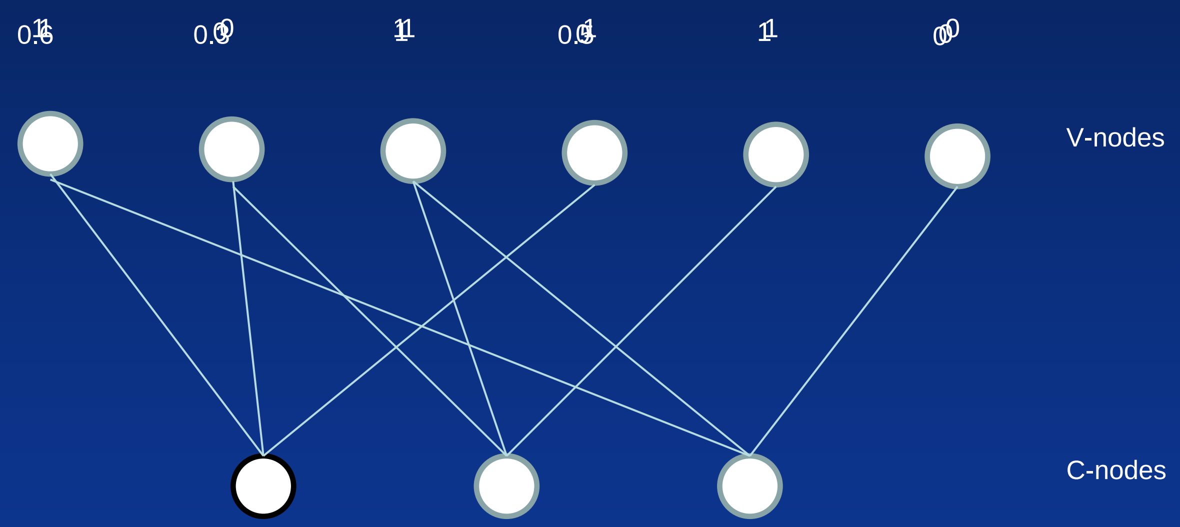


- LDPC code word
  – error d3



Data Corruption / Correction

- LDPC code – error v3 - decode iteration



v0-0	v1-0	v0-1
v1-1	v2-1	v2-1
v3-1	v4-1	v5-0

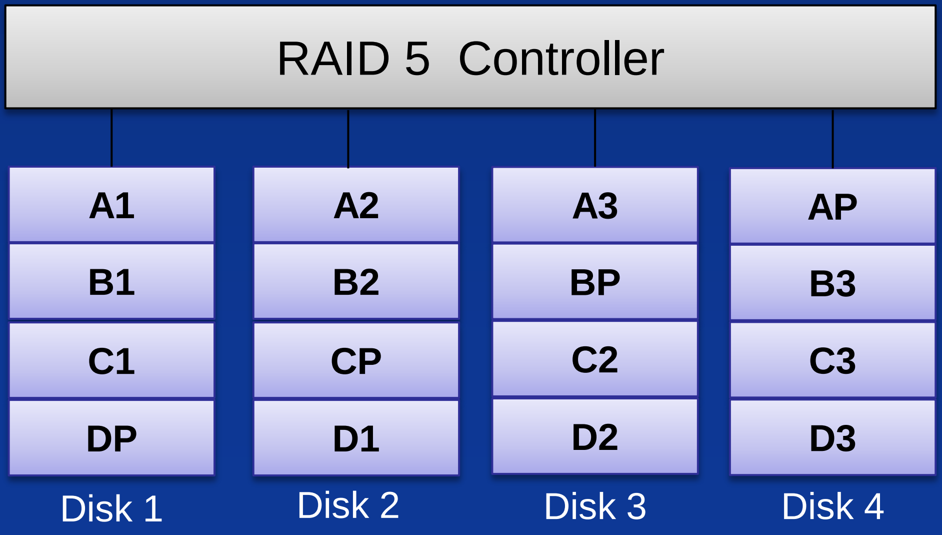
Data Corruption / Correction

- What happens when BCH and LDPC is not enough to recover the data?
- RAID could be used as a method to guarantee data but comes at the cost of additional NAND to store parity and data copies

RAID

RAID

- RAID protection
  - As the NAND process shrinks, NAND becomes more susceptible to error, sometimes ECC, BCH and LDPC may not be enough
- HDD vendors use RAID to protect data across physical drives

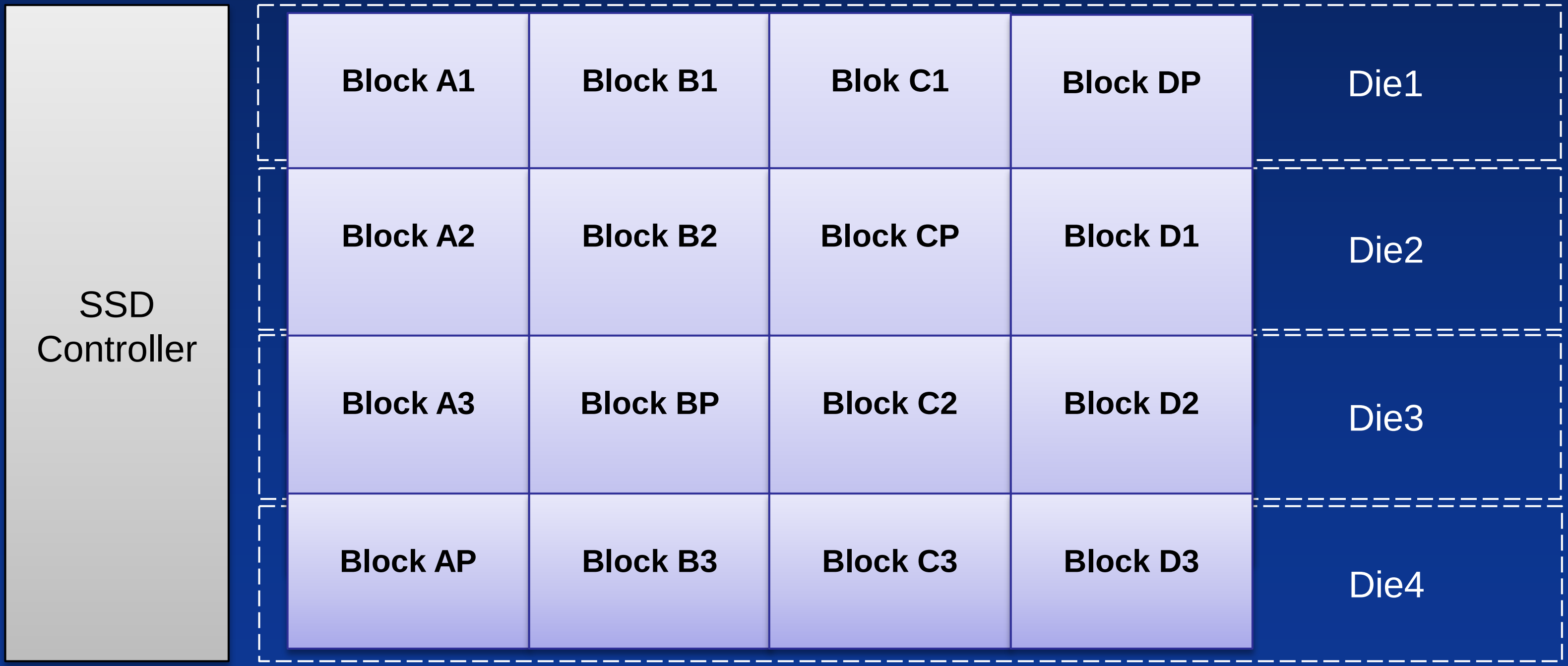


RAID 5 Controller

Disk 1	Disk 2	Disk 3	Disk 4
A1	A2	A3	AP
B1	B2	BP	B3
C1	CP	C2	C3
DP	D1	D2	D3

RAID

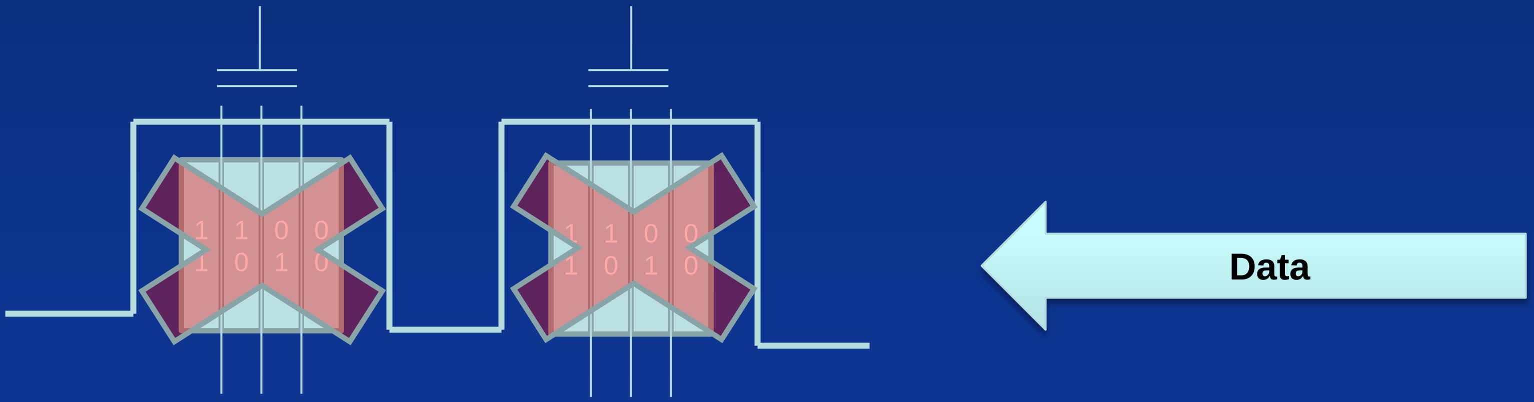
- A similar approach can be taken with SSDs



SSD Controller	Block A1	Block B1	Blok C1	Block DP	Die1
	Block A2	Block B2	Block CP	Block D1	Die2
	Block A3	Block BP	Block C2	Block D2	Die3
	Block AP	Block B3	Block C3	Block D3	Die4

RAID

- What is RAID protecting against?
  - Latest NAND nodes could be susceptible to mass data loss die to charge issues within the NAND
  - This can result in a whole world line data loss



Block marked as bad

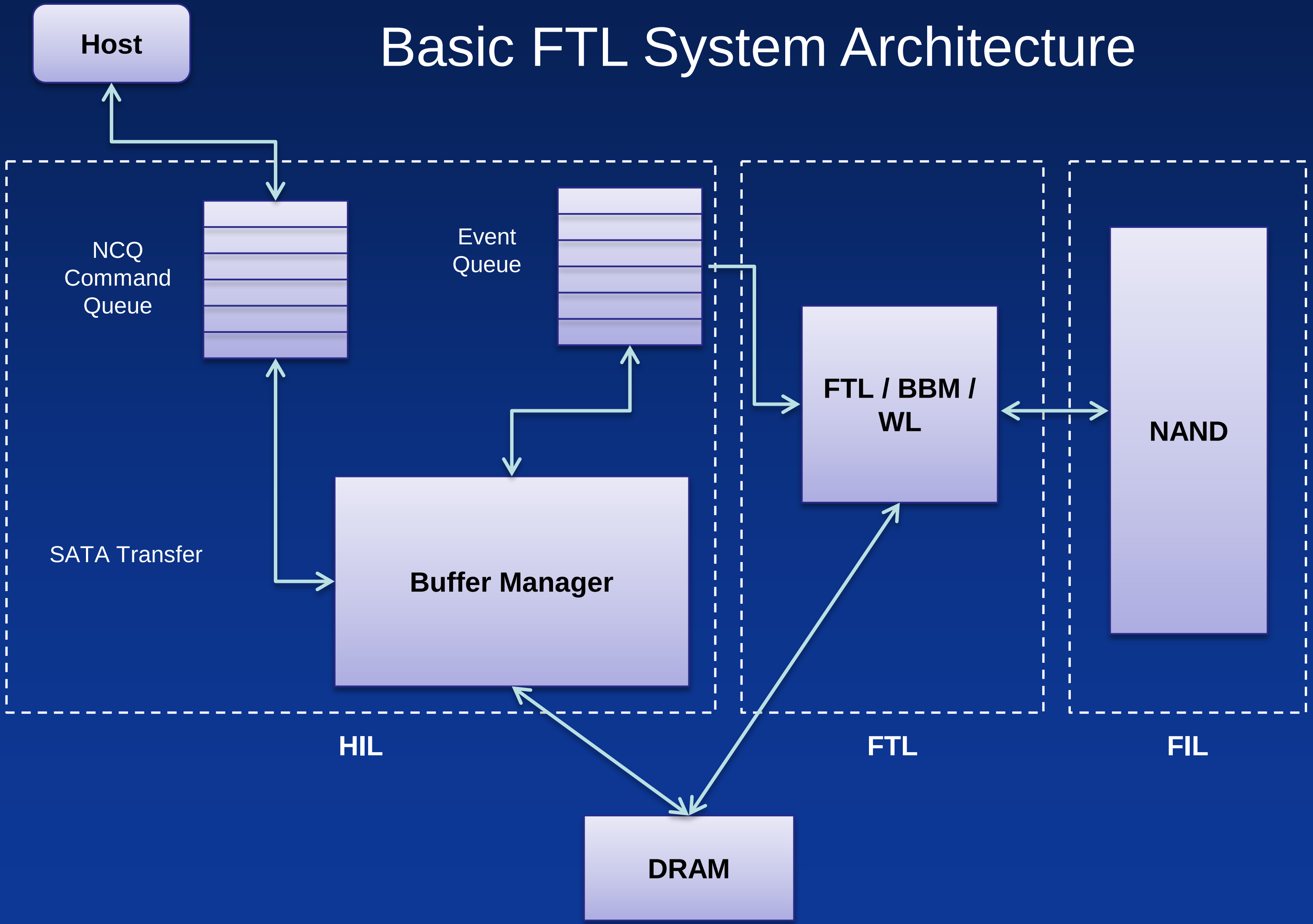
Enterprise FTL

Enterprise FTL

“You can put a price on performance,
but you cant buy low latency!”

Enterprise FTL

Basic FTL System Architecture



Enterprise FTL

- Enterprise drives may require sustained data rates that far exceed a general consumer grade NAND solution
  - As such, the system requirements are far more complex than a consumer grade drive
- In the previous example the system is defined for SATA only system
  - When considering the sustained data rates of NVMe, a different approach is required

Enterprise FTL

- In addition to coping with fast hosts, enterprise solutions also require reduced latency
- Updating the FTL mapping takes time:
  - For 16K page there would be one entry for each 16K
    - One entry is ~4B (i.e., for each 16K page written 4B of metadata must be stored in DDR)
    - For a 256GB drive with 16K page NAND:
      - » (256G / 16K) *4B = 64MB
    - For a 256GB drive with 8K page NAND:
      - » (256G / 8K) *4B = 128MB

Enterprise FTL

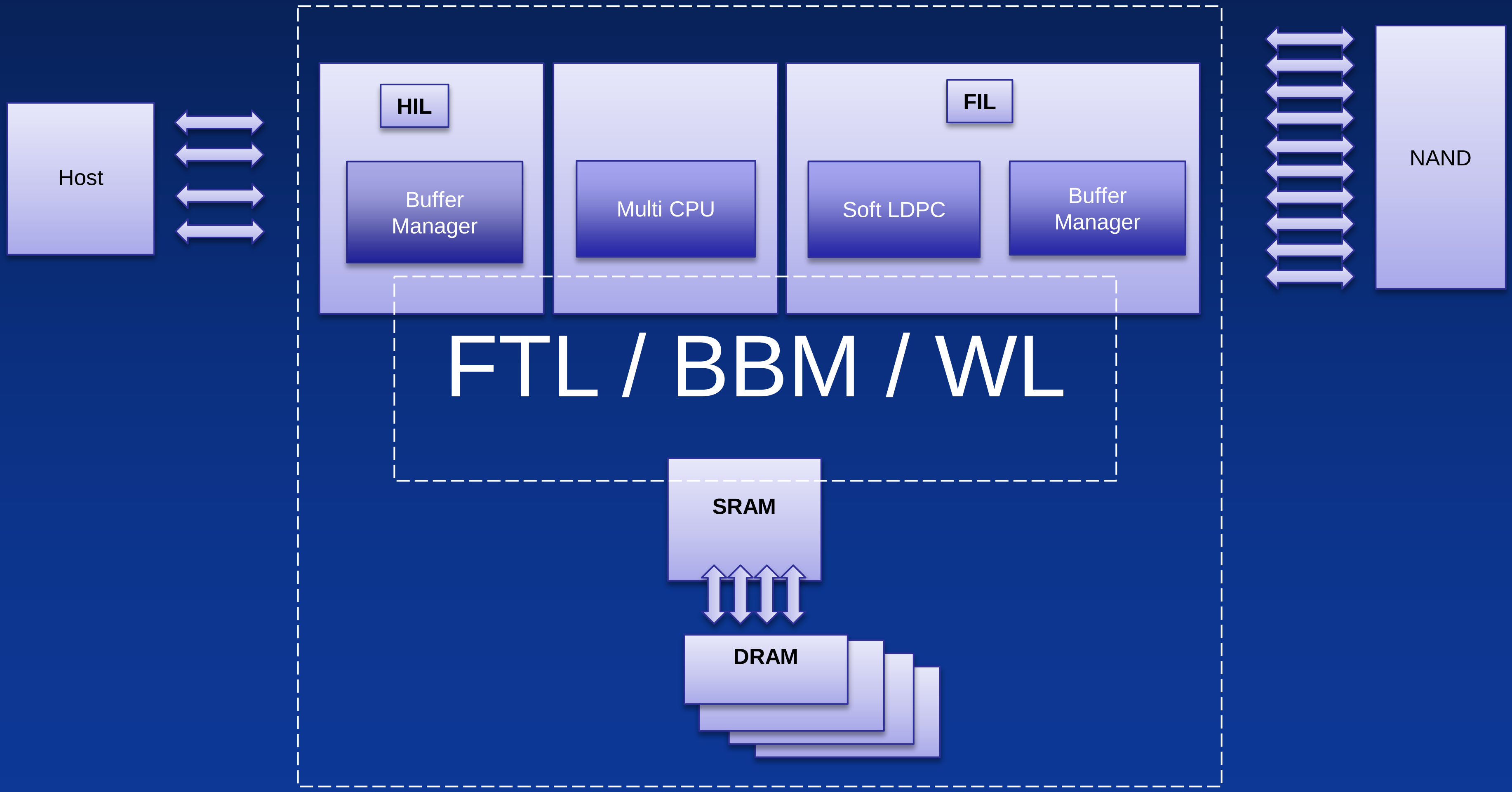
- Writing many mega bytes of mapping table data to the NAND on a regular basis could be a major contributing factor to latency issues in the system
- Instead of writing the complete table, latency can be reduced by splitting the table into smaller chunks of data and then creating a link between multiple maps

Enterprise FTL

- Now we have a new (very high level) FTL architecture we can address the performance bottlenecks

Enterprise FTL

Enterprise FTL System Architecture



Enterprise FTL

- Other factors to consider in an enterprise system:
  - Higher OP levels may be required to achieve good dirty performance
  - Inclusion of super caps or batteries to protect against power loss
  - Infield debugging, continuous logging of firmware behaviour accessible remotely for a speedy unobtrusive solution

Thank You!