STORAGE DEVELOPER CONFERENCE

SDC22 | Fremont, CA
September 12-15, 2022

BY Developers FOR Developers

A SNIA Event

Optimal Performance Parameters for NVMe SSDs

Presented by Dan Helmick

Agenda

- Example Drive with Terminology
- Optimal Write Parameters
  - NOWS, NPWG, NPWA
- Optimal Deallocate Parameters
  - NPDG, NPDGL, NPDA, and NPDAL
- Optimal Read Parameters
  - NPRG, NPRA

Motivation

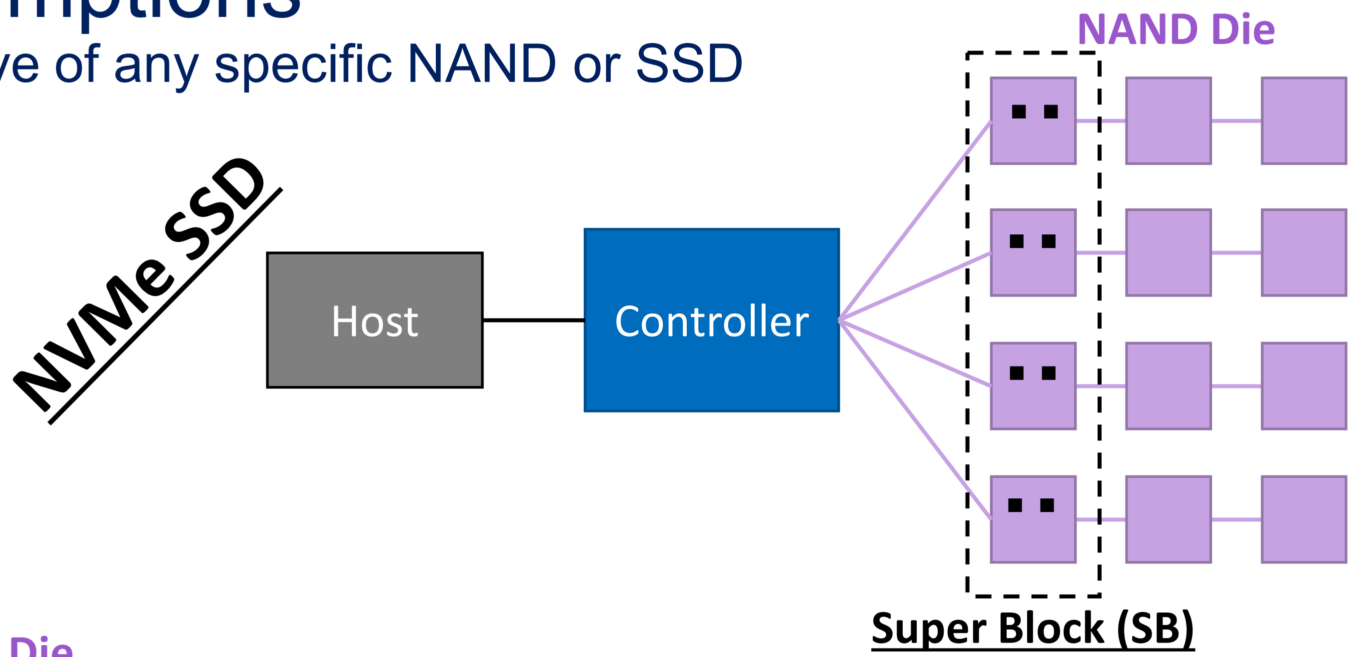
- Presentation intended as initial guidance for Enterprise class SSDs setting parameters and Hosts' expectations of those Optimal Performance parameters for NVMe SSDs

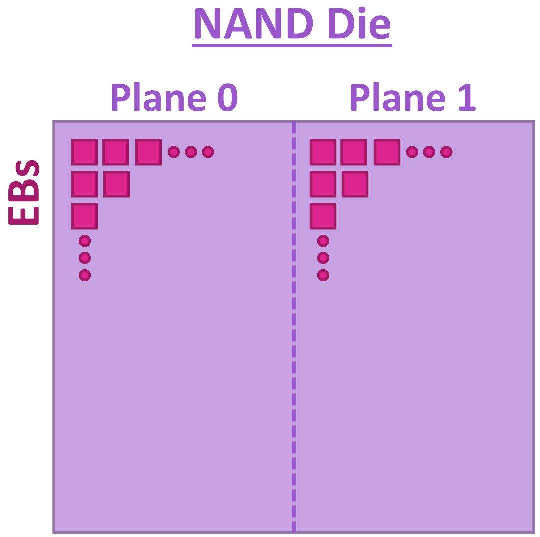
- Topics
  - OPTPERF enabled NPWG, NPWA, NOWS, NPDG, NPDGL, NPDA, NPDAL
  - OPTRPERF enabled NPRG, NPRA, and NORS
  - NVMe Spec References
    - NVM Express Command Set Specification 1.0b
    - TP4090 Enhanced Deallocation Granularity
    - TP4116 Optimal Read Size and Granularity
    - TP4148 Enhanced Namespace Preferred Deallocation Alignment

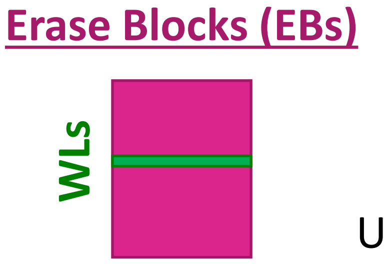
Example SSD with Size Assumptions

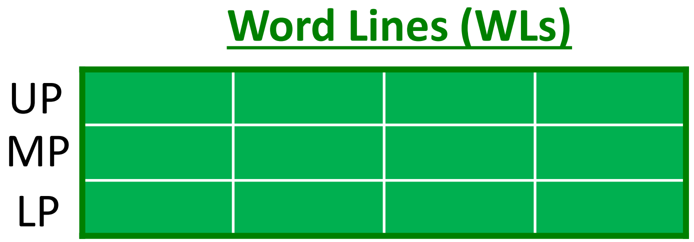
For example purposes only. Not Representative of any specific NAND or SSD

- 4 Channels
- 12 NAND Die
  - 3 Die per channel
- 2 Planes per Die
- 1,024 EBs per Plane
- SB = Logical Abstraction
  - 1 EB per Plane from 1 Die per Channel
  - Sometimes called: Garbage Collection Unit
- 256 WLs per EB
  - EB = 12MiB
- 3 Pages per WL in TLC NAND
  - Upper Page = UP
  - Middle Page = MP
  - Lower Page = LP
- 4 mapping units per Page
- 4KiB user data per mapping unit



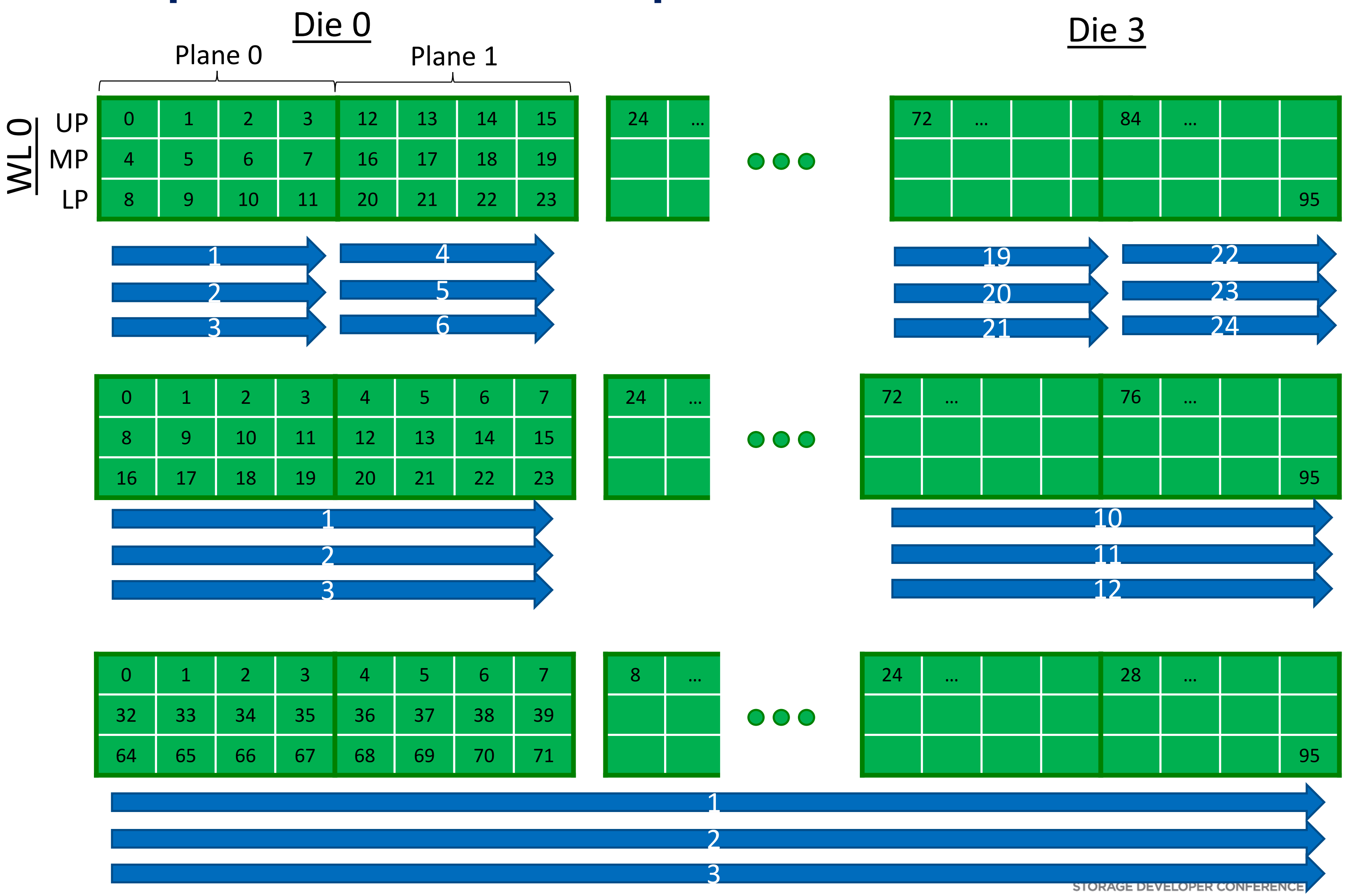






Some Example Fill Sequences for Super Blocks

- Full WL Fill
  - Fill each WL
  - Fill each Die
- Die Stripe Fill
  - Fill each Page on both Planes
  - Fill each Die
- SB Stripe Fill
  - Fill each Page on all Die



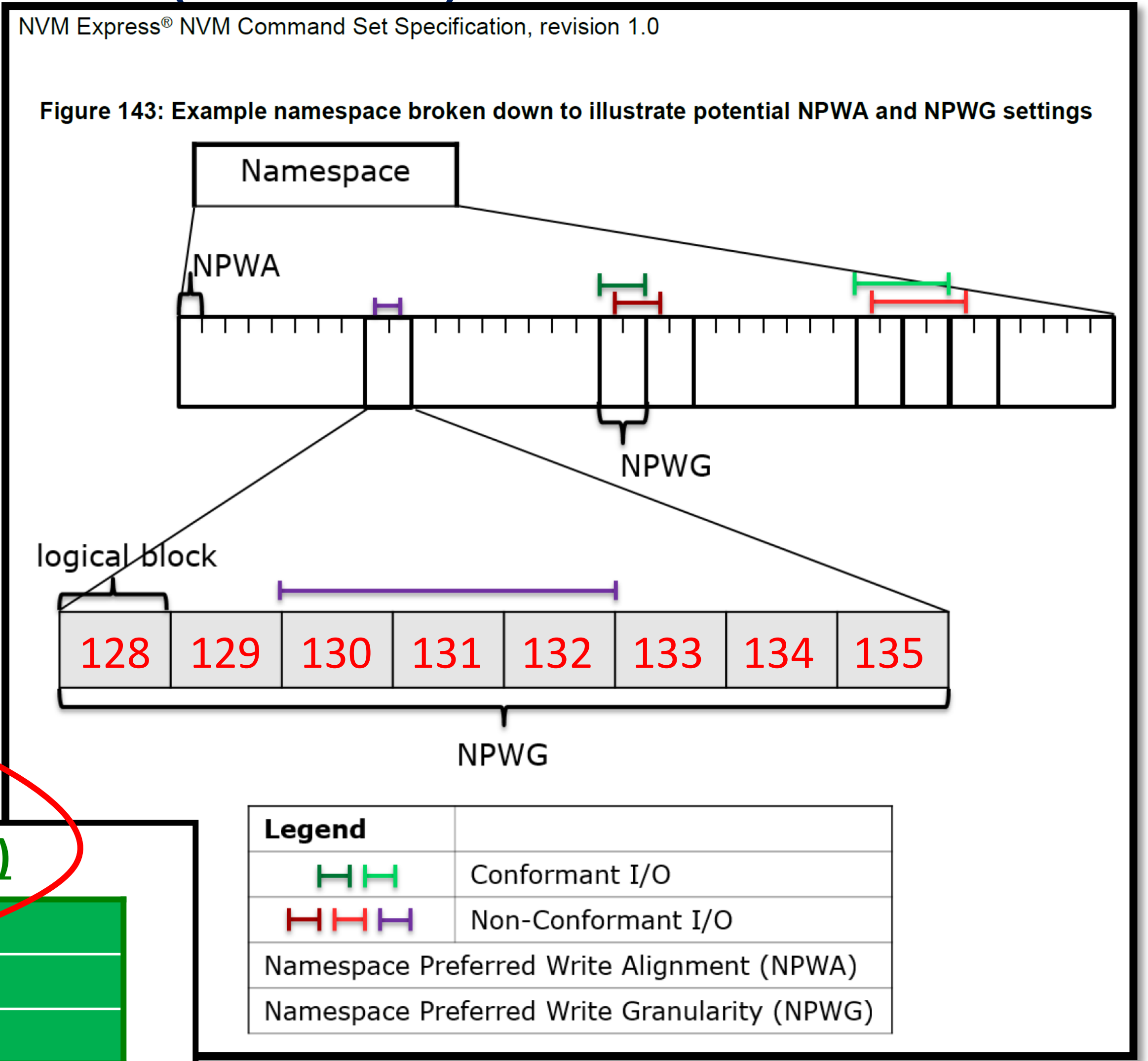
 |

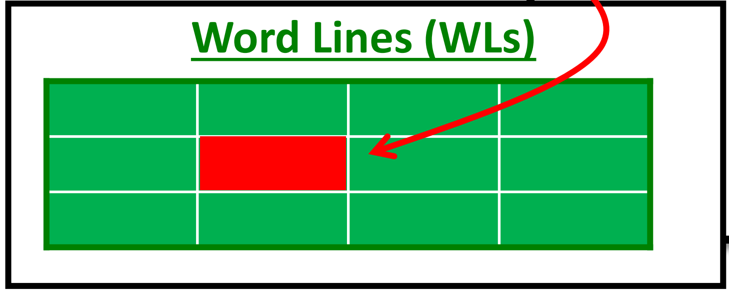
STORAGE DEVELOPER CONFERENCE SDC 22

Namespace Preferred Write Granularity (NPWG) and Namespace Preferred Write Alignment (NPWA)

- Example Image
  - Reminder: Zeros based count of Logical Blocks
  - NPWA = 3
  - NPWG = 7
  - Breakout images for IOs follows
- Added Information for Clarity of Example
  - Sample LBA numbering
  - Logical Block Size = 512B
  - Mapping Unit = 4KiB
  - WLs = 3 pages of 16KiB each
- Conclusions
  - Writes by SSD will be at the Mapping Unit size
  - NPWA = 3 is a poor choice for this example NAND
  - NPWA = NPWG = 7 is a better choice

Continuing with these assumptions





NPWG and NPWA Continued

- Conformant IO
  - NPWG = NPWA = 7
  - Every 4KiB write perfectly matches one Mapping Unit in the SSD
  - Aligned Writes of N * 4KiB length perfectly modify N Mapping Units
- Non-Conformant IO
  - Cause various Read-Modify-Write (RMW) behaviors
    - ***Write Amplification***
  - Offset starting LBAs cause a read of the **Head Runt**
  - Offset ending LBAs cause a read of the **Tail Runt**
  - Larger writes will efficiently overwrite the contiguous central data
    - Misalignment impacts reduce with larger IOs
  - Writes < NPWG might be able to read **Head and Tail runt data** from one Mapping Unit

Figure 145: Host write I/O command following NPWA and NPWG

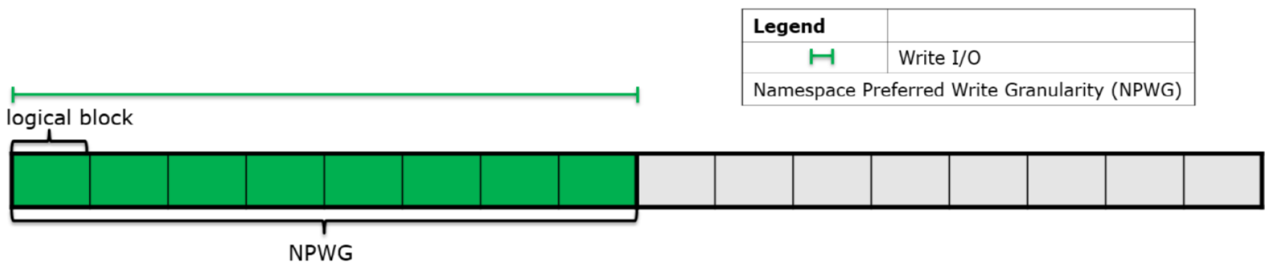


Figure 146: Host write I/O command following NPWG but not NPWA attributes

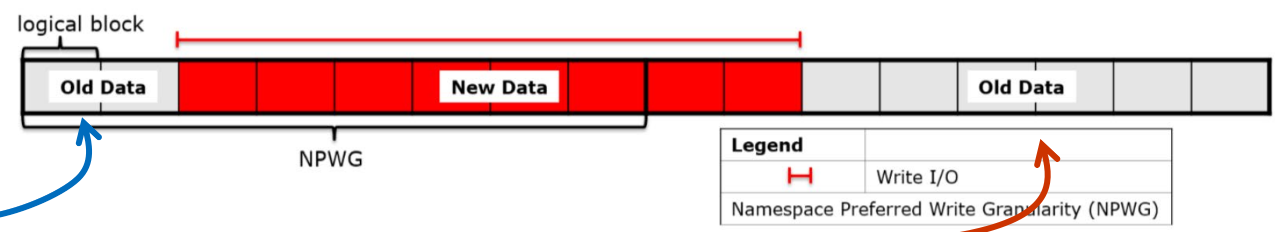
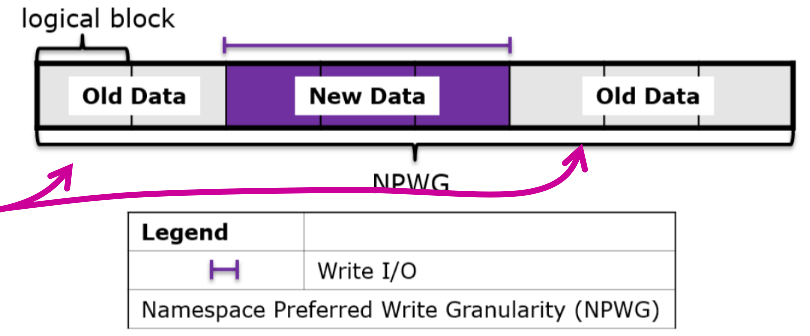
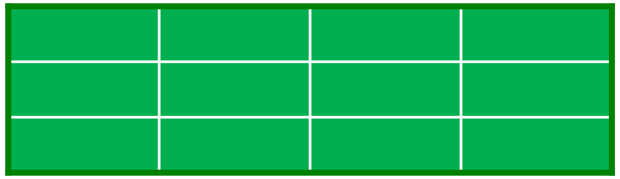


Figure 144: Non-conformant Write Impact



Word Lines (WLs)



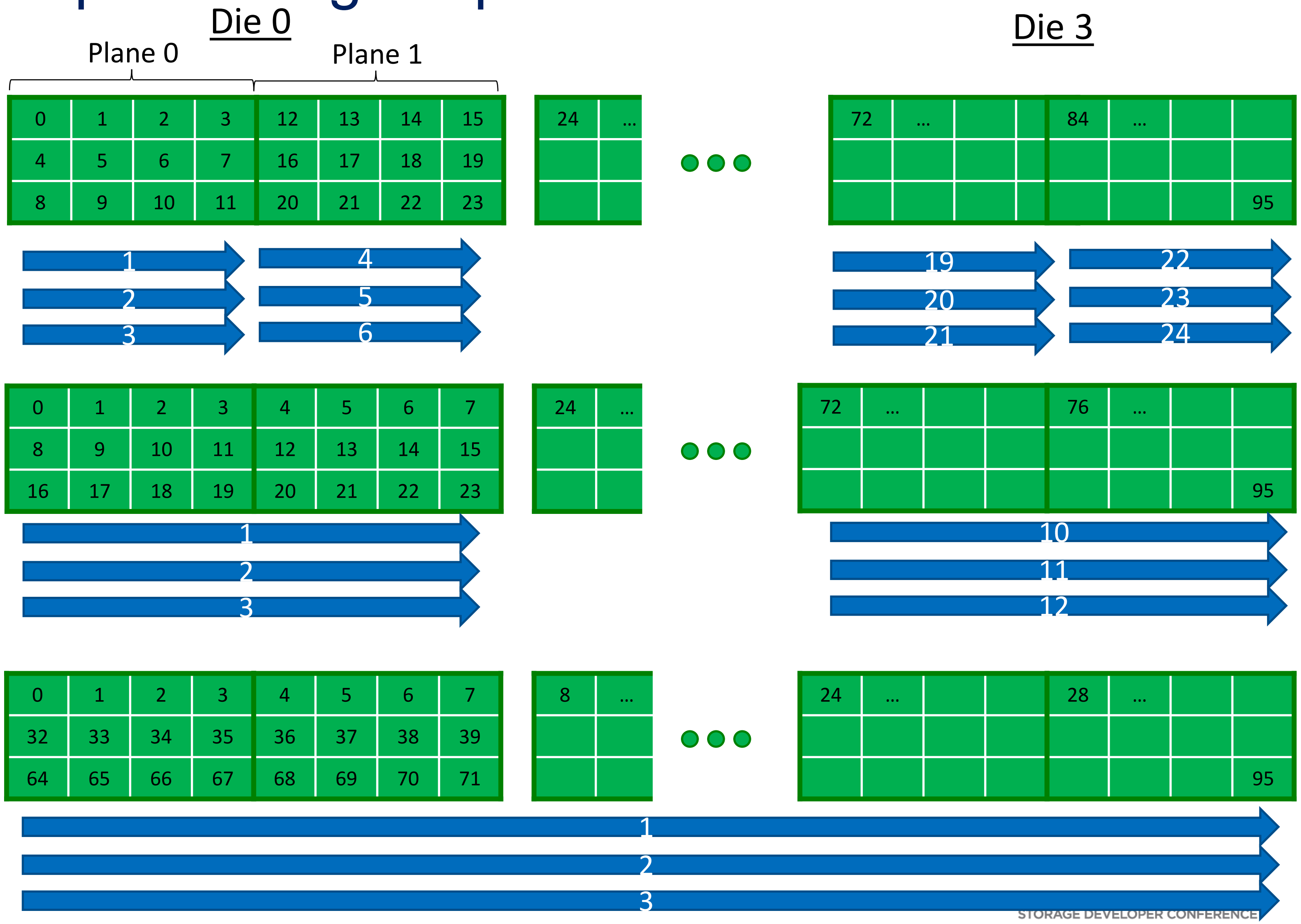
STORAGE DEVELOPER CONFERENCE SDC 22

NPWG and NPWA Discussion

- NPWG
  - Equal to Mapping Unit can be a great choice
- NPWA
  - Normally going to be set equal to NPWG
- Other values can be used
  - Unusual controller features or Metadata placements
  - New NAND access possibilities
- Summary
  - Generally set and defined by the drive and NAND characteristics
  - Avoids most harmful write performance penalties
    - Example focus today: Read-Modify-Write penalties
  - **Does not necessarily fully optimize write performance**
    - But may get you close to optimal

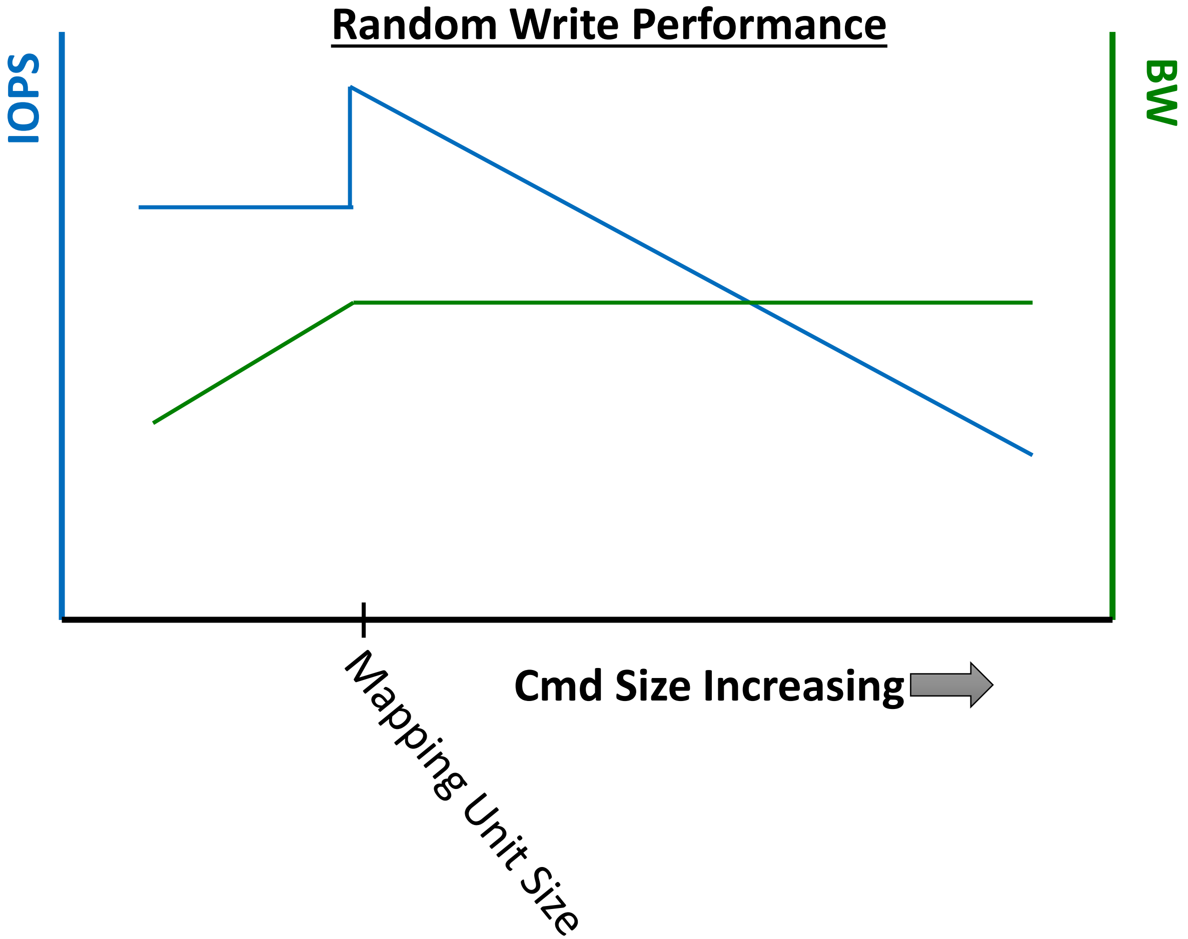
Namespace Optimal Write Size (NOWS)
in the Context of Some Example Writing Sequences

- Full WL Fill
  - NOWS = 16KiB
  - GC would be triggered with Full Page Reads
- Die Stripe Fill
  - NOWS = 32KiB
  - Perhaps 16KiB
    - SQ/CQ overheads are sufficiently small
    - More than QD1
    - Host cannot accumulate 32KiB reliably
- SB Stripe Fill
  - NOWS = 128KiB
- Conclusions
  - Focus of this slide is an architectural discussion on NAND access optimizations for NOWS
    - Other optimizations possible
  - NOWS = 4KiB is a valid setting
    - But it doesn't provide the Host any guidance above and beyond NPWG
  - Command lengths of M * (NOWS) expected to be equal or better performance than NOWS
  - Customer conversations have great value for setting an NOWS for their deployments
    - Feasibility of accumulating write sizes from end Applications
    - Expectations on Host outstanding QD or QoS



Write Performance and Optimal Performance Settings

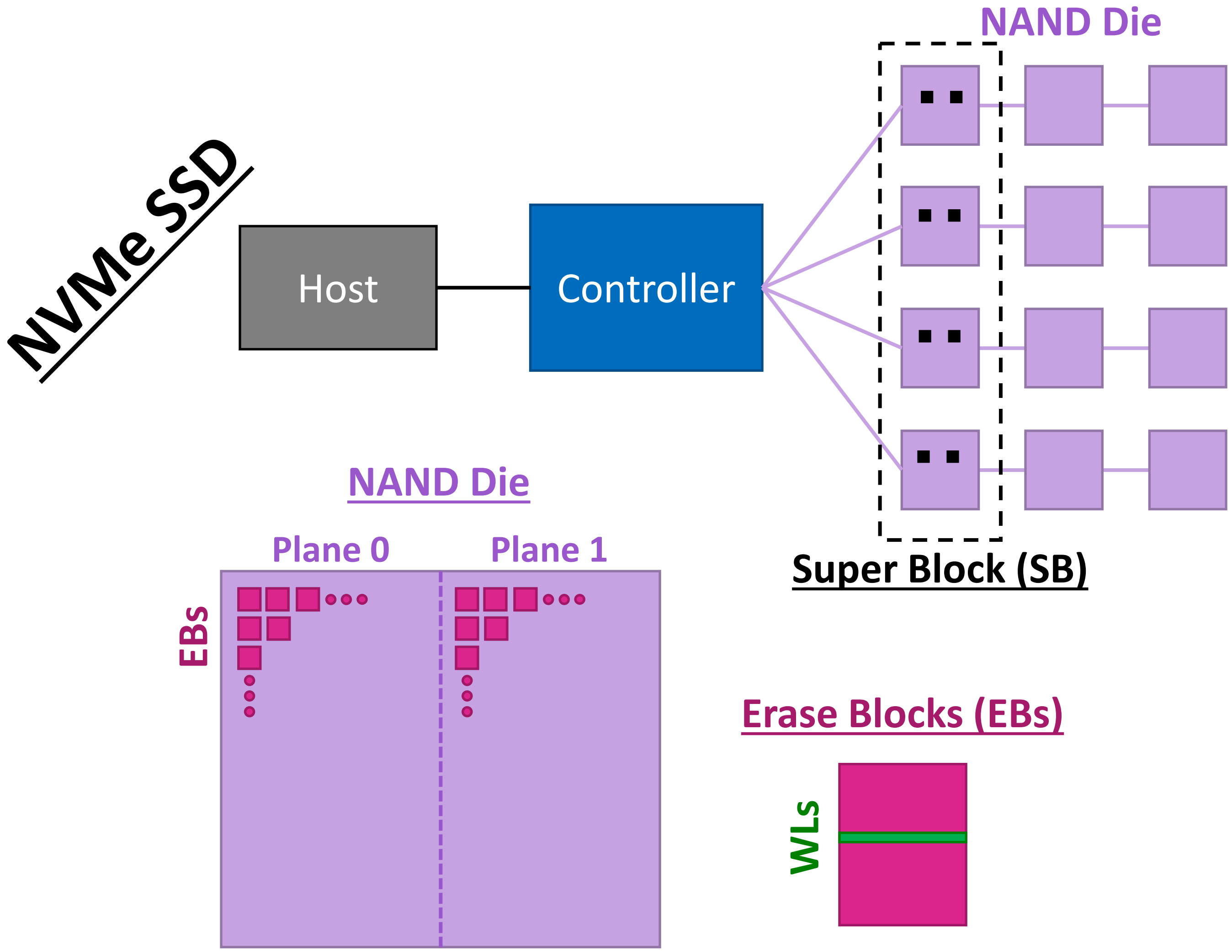
- Performance Plot sketched for Context
  - Ignoring Vendor specific behavior changes
- RMW impacts are clear at less than Mapping Unit Size
  - NPWG and NPWA is easily distinguished
- For Large Command Sizes
  - Write Buffer Obscures internal impacts
  - NOWS Settings can be hidden
- NOWS Selection could focus on Secondary Aspects
  - Mixed Performances (BW or QoS)
  - Read Performances that are measured later
  - Controller QoS impacts
  - Endurance or Media aspects
- Customer–Vendor discussions can be valuable for setting NOWS



Optimal Deallocate Parameters

Namespace Preferred Deallocate Granularity and Alignment (NPDG and NPDA)

- Optimal NPDG and NPDA provide Write Amplification Factor (WAF) = 1
  - Achieved if EBs are erased without any GC
- Example for NAND managed in SB groupings
  - Sequential Writes fill a SB
  - SB = (1EB per Plane) * (2 Plane per Die) * (1 Die per Channel) * (4 Channels)
  - SB ~= 100MiB for example numbers in this presentation
- NPDGL and NPDAL
  - Added to the spec to enable very large SB erases



Some Recommendable Host Uses of NPDG and NPDA
*Assuming NPDG = NPDA = SB Size

- Similarities to Streams Granularity Size (SGS)
  - Circular logs sized to NPDG=NPDA=SGS
  - Sequential writes deallocated in NPDG multiples
    - SSD Cache Manangement
  - Log Structured File Systems
- Similarities to Zoned Namespaces (ZNS)
  - Writing too much to a Zone provides hard errors while overwriting NPDG and NPDA will receive no information from the drive
    - NPDG and NPDA are most relevant as guidance for single tenant Conventional NVMe usage
  - Sequential writes deallocated in NPDG multiples
  - Log Structured File Systems
    - RocksDB has been used in many presentations on ZNS

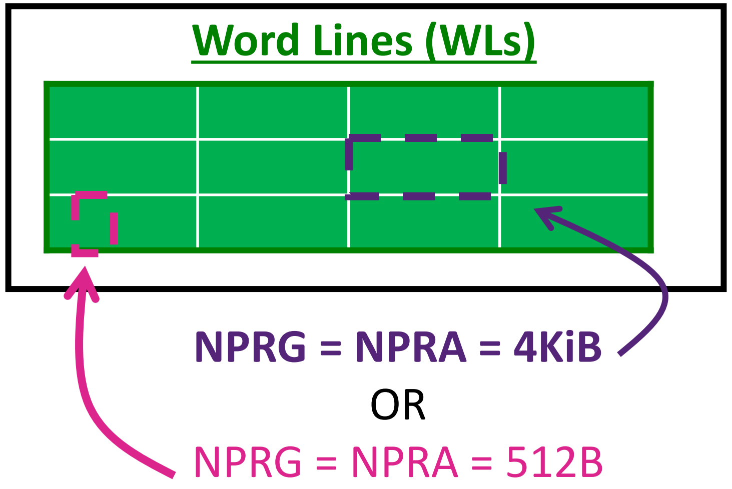
Why would NPDG and NPDA Differ from SB Size?

- Difficult for Hosts to aggregate SB Size of data
  - Today's SSDs can have 16+ Channels and EBs grow in capacity every generation
- Internal controller HW capabilities and optimizations
- NAND Features
- Behavior alignments of Write Fill and/or SSD Algorithms

Vendor-Customer discussions can provide more specific guidance and alignment to customer use-cases

Namespace Preferred Read Granularity (NPRG) and Namespace Preferred Read Alignment (NPRA)

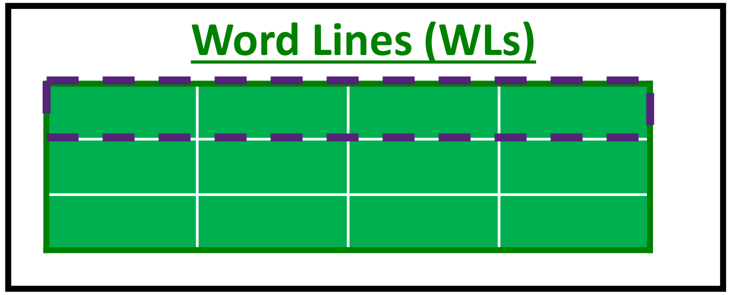
- Read Granularities and Alignments follow similar Host rules as Optimal Write parameters
- Reasons NPRG/NPRA may be Different from NPWG/NPWA
  - NAND often has the ability to be more precise on Reads with fewer physical restrictions
  - ECC capabilities may be optimized for common accesses
  - Controller Metadata or User Data layouts on the NAND
  - No RMW penalty for small Reads
- NPRG/NPRA selections can be a minor impactor in many SSDs



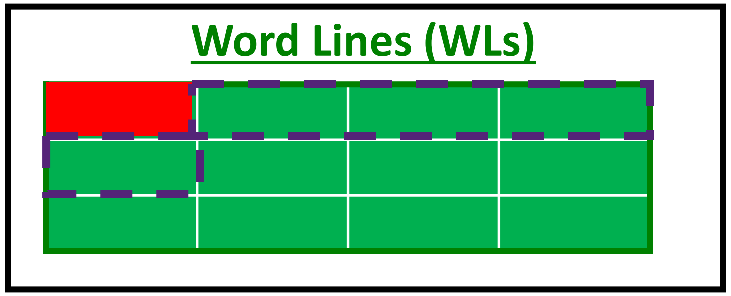
 |

Namespace Optimal Read Size (NORS)

- Similar to NOWS; NORS can have several best answers
  - NOWS
  - Mapping Unit
  - Page or Page Stripe
- Setting NORS benefits from a Vendor-Customer conversation
  - Optimizing SSD Architecture and Host SW should be an ongoing bi-directional conversation
- One Caveat for NORS
  - NORS values larger than a Mapping Unit may not be valid if an errant Write throws off alignments.



- NORS of 16KiB successfully gives full page reads



Non-adherence to write-related performance attributes (i.e. NPWG, NPWA, NPDG, NPDGL, NPDA, and NOWS), across all the namespaces in:

a) the same NVM Set;
b) the same Endurance Group when NVM Sets are not supported; or
c) the NVM subsystem when Endurance Groups are not supported,

may affect the level of read optimization achievable through the usage of NORS as described in this section.

- One write **fails** to follow a NOWS of 16KB.
- Therefore, NORS of 16KiB is now reading from 2 pages
- Format likely required to fix

Dan's Summarized Guidelines on Optimal Performance Parameters

No strict rules in this list

- NPWG = NPWA = Mapping Unit
- NOWS = NAND or Controller Optimized Size
  - NOWS > NPWG
    - Provide additional information. Don't report NOWS=NPWG.
  - NOWS = Page is a fair start
- NPDG = NPDA = SB groupings
  - But often this may be too large for Hosts to aggregate
  - Second best values should be discussed
  - NPDG/NPDA can be used as additional guidance for writes (SGS and Zone like behavior)
- NPRG = NPRA = Logical Block Size
  - Or perhaps Mapping Unit
- NORS = NOWS
  - Or perhaps Mapping Unit
  - NORS > NPRG
    - Provide additional information. Don't report NORS=NPRG

Host Programing

In order of Importance

- Writes
  1. NPWG = Hard lower bound
  2. NOWS = Goal write accumulation size
  3. Multiples of NPWG and NOWS
- Deallocates
  1. Deallocate large spans ASAP
  2. NPDG or multiples = Goal
- Reads
  1. Request the data needed
     - Don't read extra to reach NPRG or NORS
     - Optimize SW stack for Writes and Deallocates first
  2. NPRG = Goal read size
  3. NORS = Nice to have
  4. Multiples of NPRG and NORS = Bonus points

STORAGE DEVELOPER CONFERENCE SDC 22

Please take a moment to rate this session.

Your feedback is important to us.

STORAGE DEVELOPER CONFERENCE