Something is changing in the Storage Panorama

Paolo Bianco
Systems Engineer, Dell

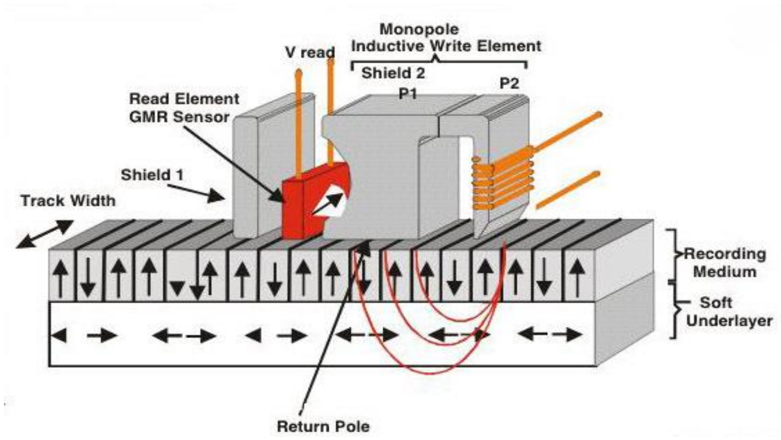
Agenda

- Nearline Capacity Drives Update
- SSDs and Performance Optimized Drives update
- Designing with SSDs
- Questions

Nearline Capacity Drives Update

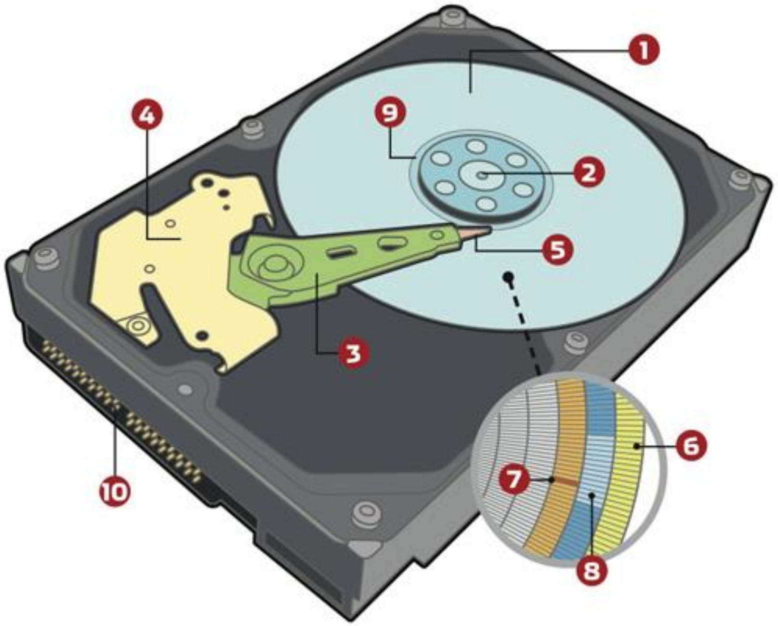
Introducing PMR

- Data on conventional HDDs platters is written in circular, concentrical tracks (about 75nm wide), separed by guard spaces.
- Total track width is larger than necessary because write head poles needs to be large enough to generate sufficient coercitive force magnetization swap
- Effective read track width could be (and it is) smaller



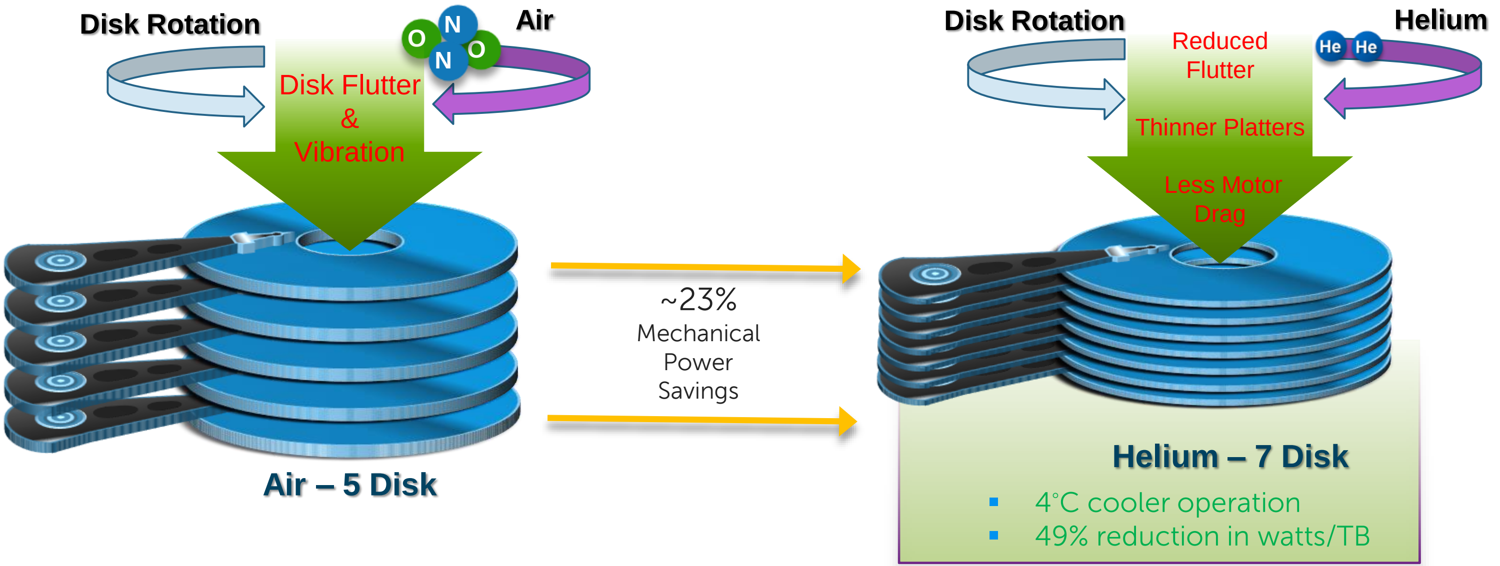
PMR Technology has reached its own limits

- Diameter of platters: 3,5"
  - Total Surface 11 sq.in
- Useful Data Surface: about 5.5 sq.in
- Max Capacity per platter side: 0,68TB
  - @ 1Tb/sq.in
- **Max Total Capacity (5 platters): 6,8TB**

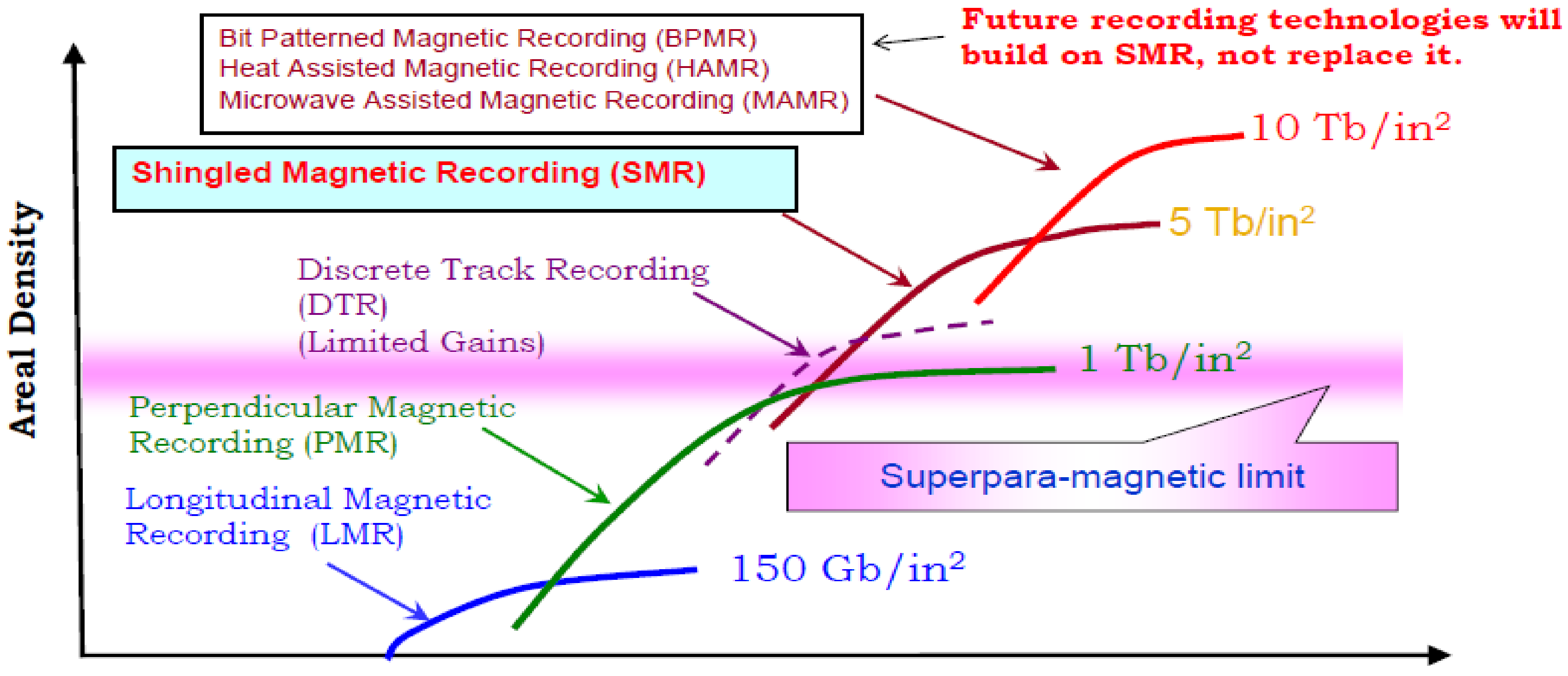


Helium-Filled: The last line of PMR Capacitive HDD

- Helium reduces mechanical power dissipated in air shear
- Allows platters to be placed closer together enabling more density
- 8TB He-Filled will probably be the last PMR-based cap.HDD generation on the market



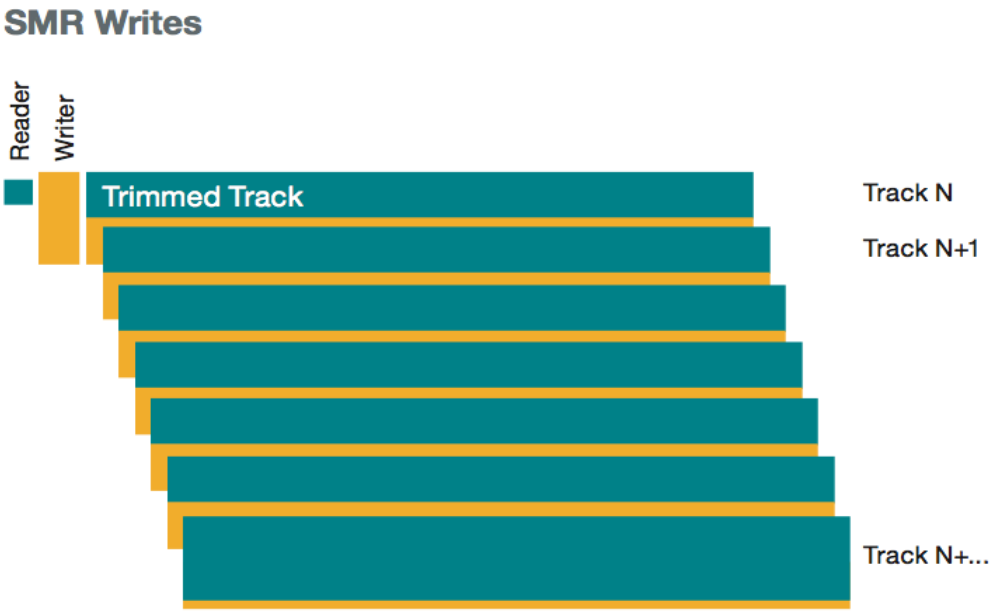
Magnetic Recording Technologies



Y. Shiroishi, Intermag 2009, FA-01

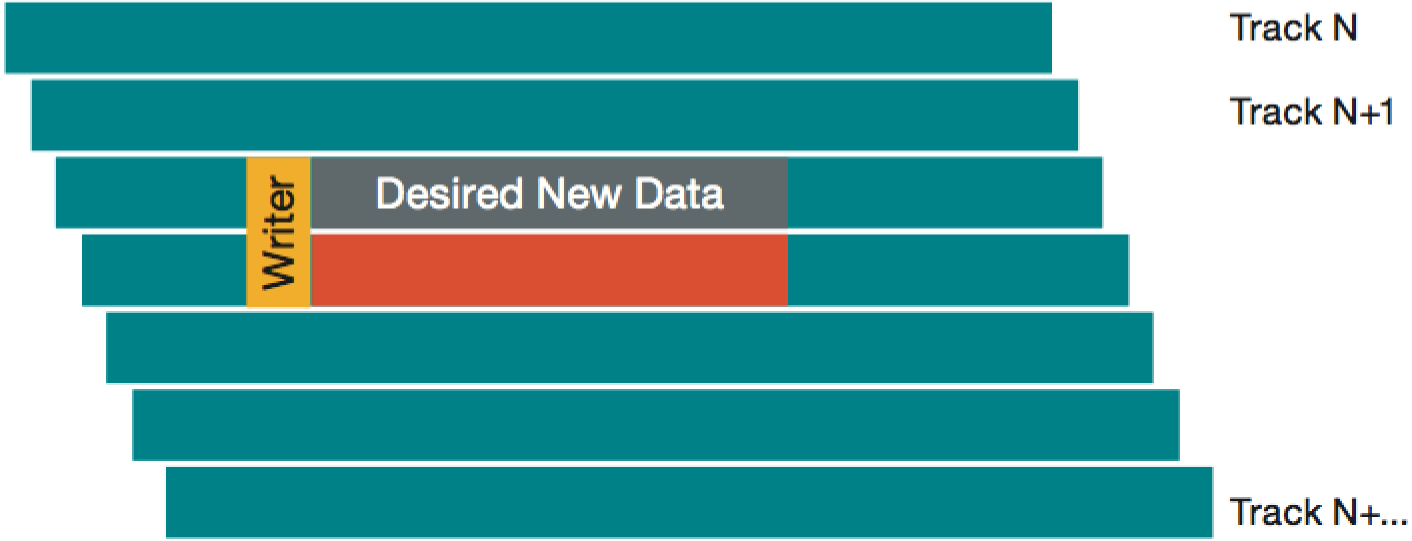
Introducing SMR

- With Shingled Magnetical Recording (SMR), clusters of tracks are superposed (just like «Roof Shingles») so that unnecessary track width space is recovered.



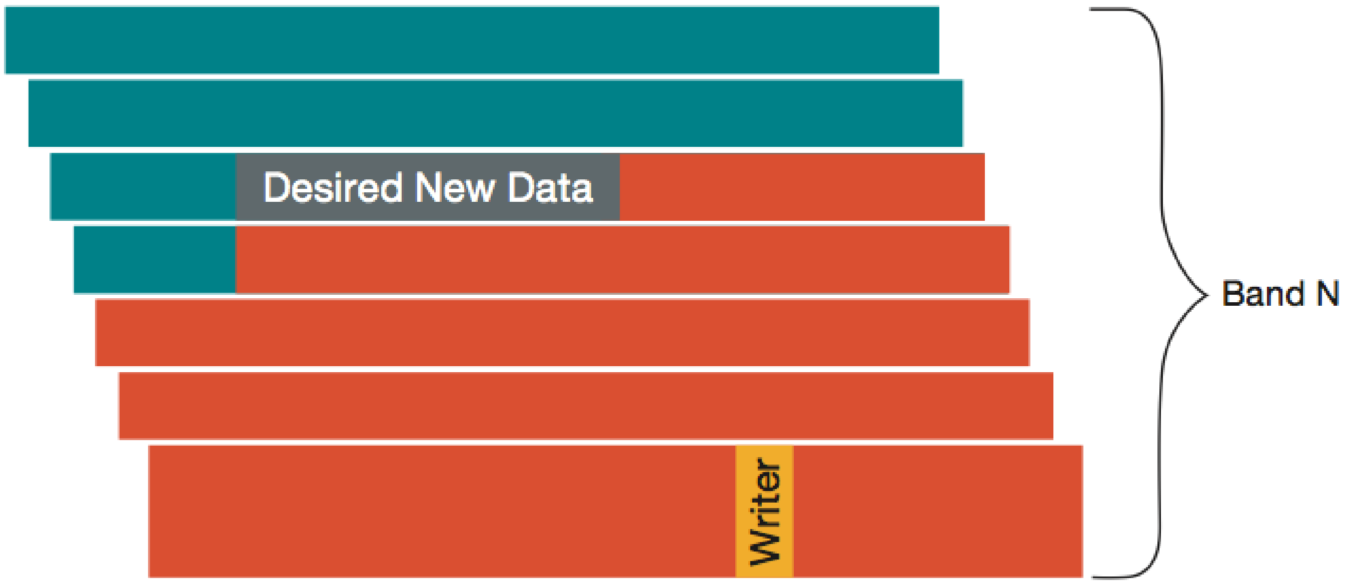
SMR Disadvantages

- When a new data is written, new data track overwrites subsequent tracks…



SMR Disadvantages

- We need then to Load in a buffer all data following the new track in a cluster…



- … and write down back again the cluster tracks starting from new data (aka R/M/W Penalty)

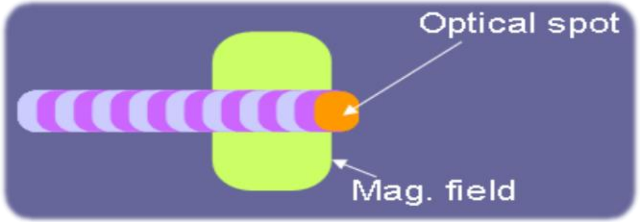
SMR Challenges

- How to avoid performance loss (sustained data rate) due to Read-Modify-Write?
  - Onboard controller defragmentation (just like SSDs)
  - OS-aware SMR media management (just like SSDs!)
  - Move BTL in Filesystem as exploring with FTL
  - T10 standards group working on this
- Short-term media capacity growing technology:
  - Move to HAMR in (probably) 3/5y
- Will all storage arrays manufacturers work on SMR-awareness?
  - Or will they try to just mitigate performance gaps with SSD caching?

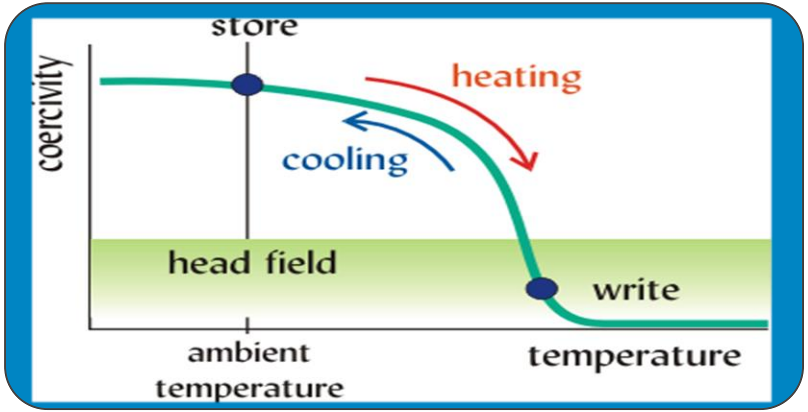
Near Future:Heat-Assisted Magnetic Recording

HAMR : A Whole New Recording System

- Density growth limited by ability to make smaller bits thermally stable
- HAMR combines laser and magnetic field to write the media
- Allows for use of much higher coercivity media and hence enables higher densities



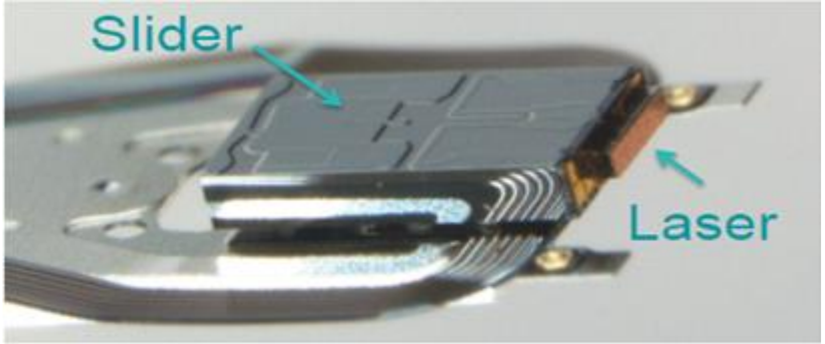
$$\frac{dH_{eff}}{dx} = \frac{dH_k}{dT} \cdot \frac{dT}{dx}$$



Industry projecting the introduction of HAMR technology around 2018

HAMR is not too far….

- The right is a photo of an actual HAMR drive. You can tell is a HAMR drive because it has the Laser Warning Stick stuck in front of it
- Below is a picture of an integrated HAMR head including the laser (not the same head used in the drive)
- First fully-functional public HAMR Drive demo run in Sept. 2012 by Seagate.

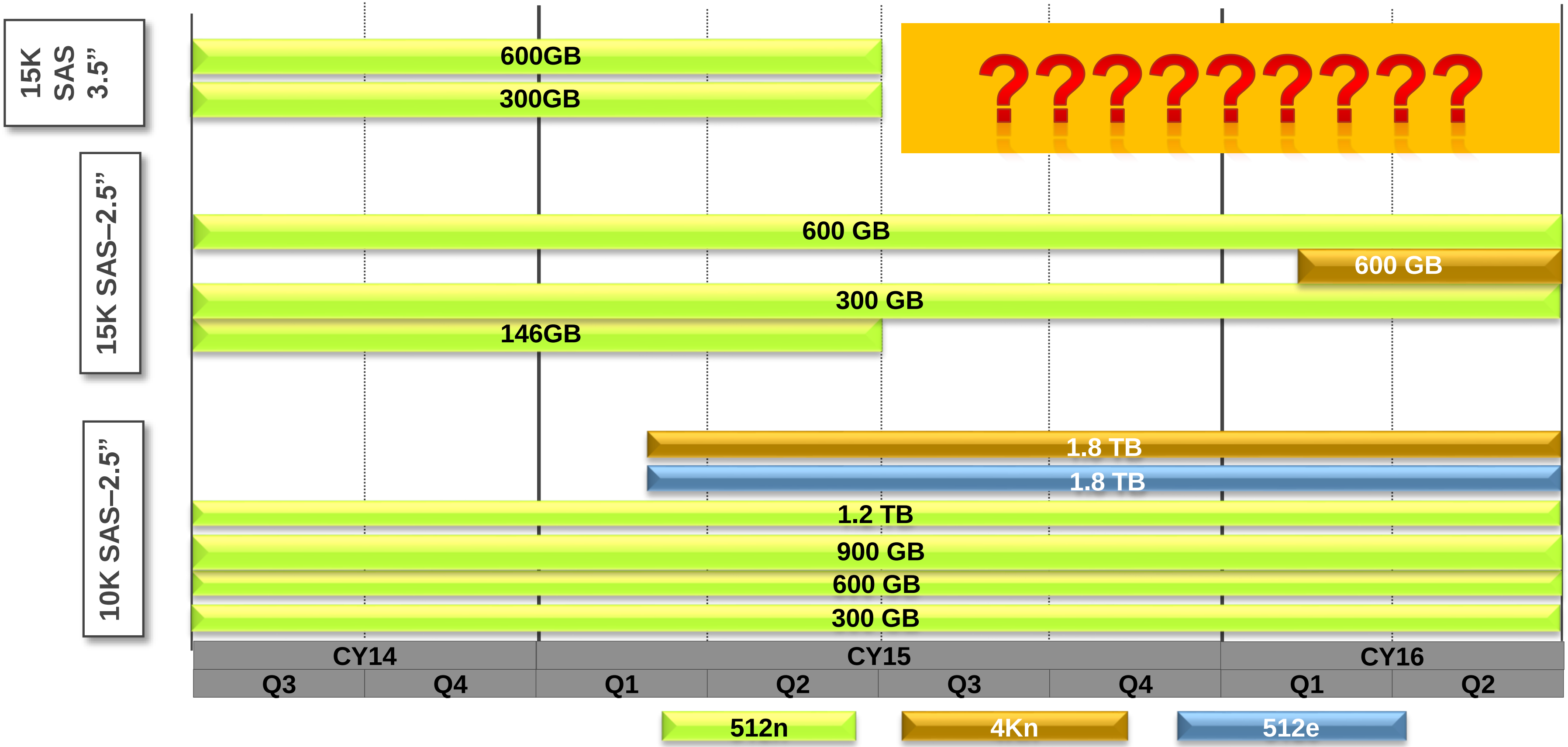


Key Takeaways

- Conventional Perpendicular Magnetic Recording technology has reached its maximum areal density limit
- 8TB NL drives will be the latest capacity drives based on standard Magnetic Recording technology (PMR)
- The drive industry is introducing a new areal density enabling technology called Shingled Magnetic Recording (SMR).
- This technology will partially alter the throughput and response time behavior of IO, especially for random writes.
- SMR is a transition technology toward HAMR, which is expected to appear in the next 3-4years (if nothing changes in Solid State memory market…)

SSDs and Performance Optimized Drives Update

Performance Optimized Enterprise HDDs



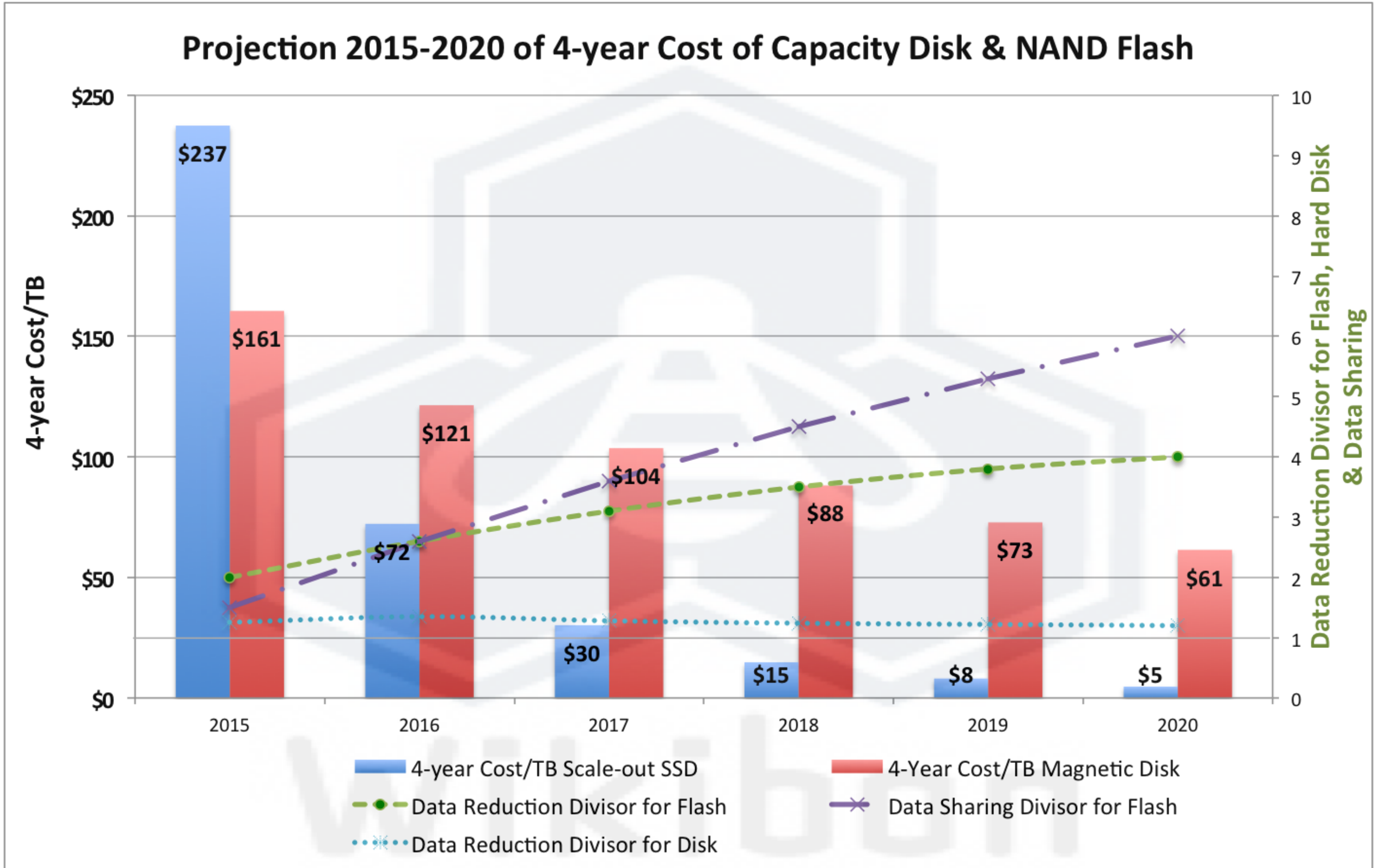
SSD predictions in 2008

Prediction 1: SSD will be bigger in capacity in 2015

SSD predictions in 2008

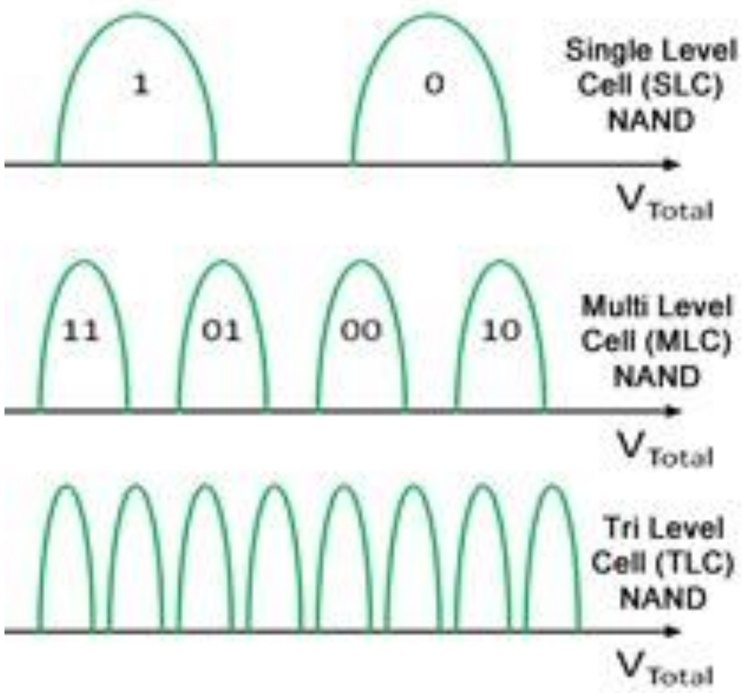
Prediction 2: SSD will be more cost effecti… 2009

Performance HDDs replaced by SSDs



NAND Memory Technologies

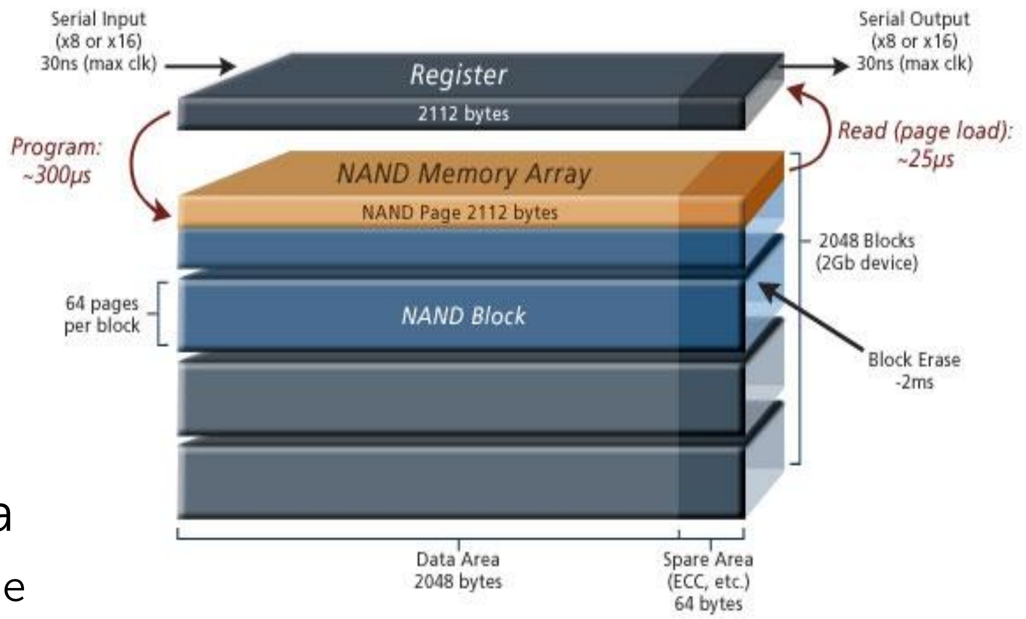
- Single Level Cell (SLC)
  - 1 bit/cell
  - Fastest
  - 100k P/E cycles
- e/HET Multi Level Cell (eMLC)
  - 2 bits/cell
  - Slightly Slower Writes
  - 30/40k P/E cycles
- Multi Level Cell (MLC)
  - 2 bits/cell
  - Slow
  - 10/20k P/E cycles
- Triple Level Cell (TLC)
  - 3 bits/cell
  - Slower
  - 3/5k P/E cycles



SSDs: anatomy of a NAND Chip

Asymmetrical access Storage Media

- Asymmetrical Read/Write
  - Read per page
  - Write per block (64 pages)
  - Block needs to be erased before a new write can occur
- Read/Modify/Write Penalty
  - Unavoidable!
- P/E cycles wears out the media
  - Electrical charges get trapped in the dielectric



Dealing with Wear Out

- Spare Capacity (2002)
- Wear Leveling
  - Distribute data to even out the use of cells. (2003)
  - Background Read Data Refresh (HET, 2012)
- Error Correction Coding
  - BCH (2005), LDPC(2010), Polar (2012)
- Compression (2011)
  - Lempel-Ziv or derivative
  - Reduces the effective amount of stored data
  - Transparent to the host!!
- De-Duplication (2012)
  - Reduces the effective amount of stored data
  - Transparent to the host!!
- Endurance Coding , aka Data Shaping (2013)
  - Transform input data into shaped data having less “0”
  - Minimize the number of programmed cells per P/E cycle
- Increase in die/chipset capacity

Steering away from SLC vs MLC discussion

Focusing on use profiles

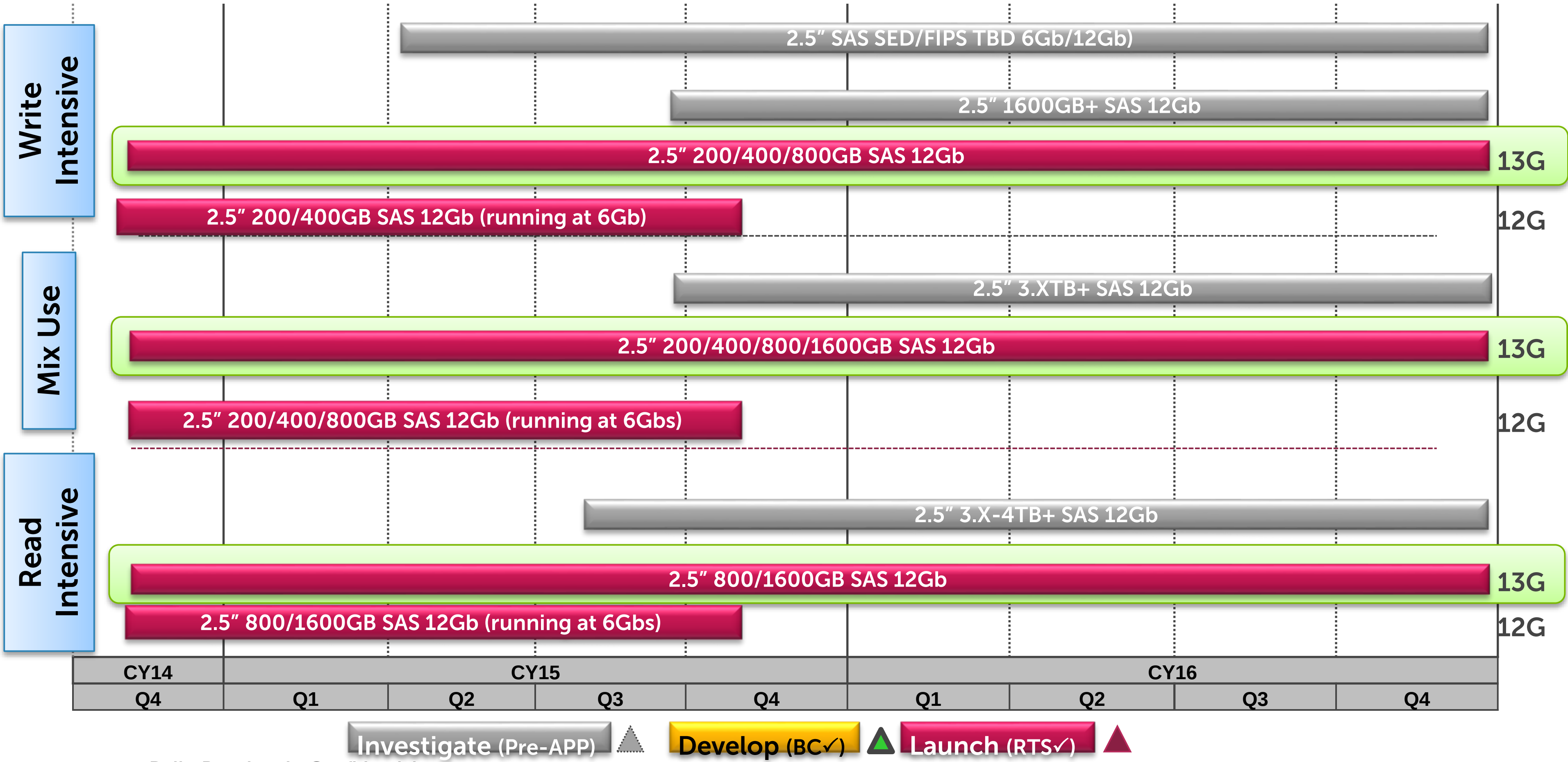
- Write Intensive SSDs
  - Mainly SLC
  - Highest longevity
  - Highest cost
- Read Intensive
  - MLC/eMLC
  - Lower longevity (but not affected by reads)
  - Lowest cost
- Multi-Use
  - Mainly eMLC
  - High performance and longevity
  - Medium cost

Flash-Optimized SSD Comparison

Storage Use	Write Intensive		Read Intensive
Market Terminology	Write Intensive (WI)	Mixed Use (MU)	Read Intensive (RI)
Workload	Mainstream Applications Any usage		Mostly Read 90/10 R/W Mix
Capacity	200 / 400 GB	800 GB	480 / 1600 GB
Endurance (Full writes / day)	30-10		3
Endurance (written PBs)	Up to 10 / 20 PB		Up to 8 PB
Random Read IOPS (*)	Up to 20K+		14K+
Random Write IOPS (**)	11K+	8K+	4K+
Sustained Write Bandwidth (***)	200 – 250 MB/s	150-225 MB/s	50 – 100 MB/s
List $/GB	Up to $20	$11	$4

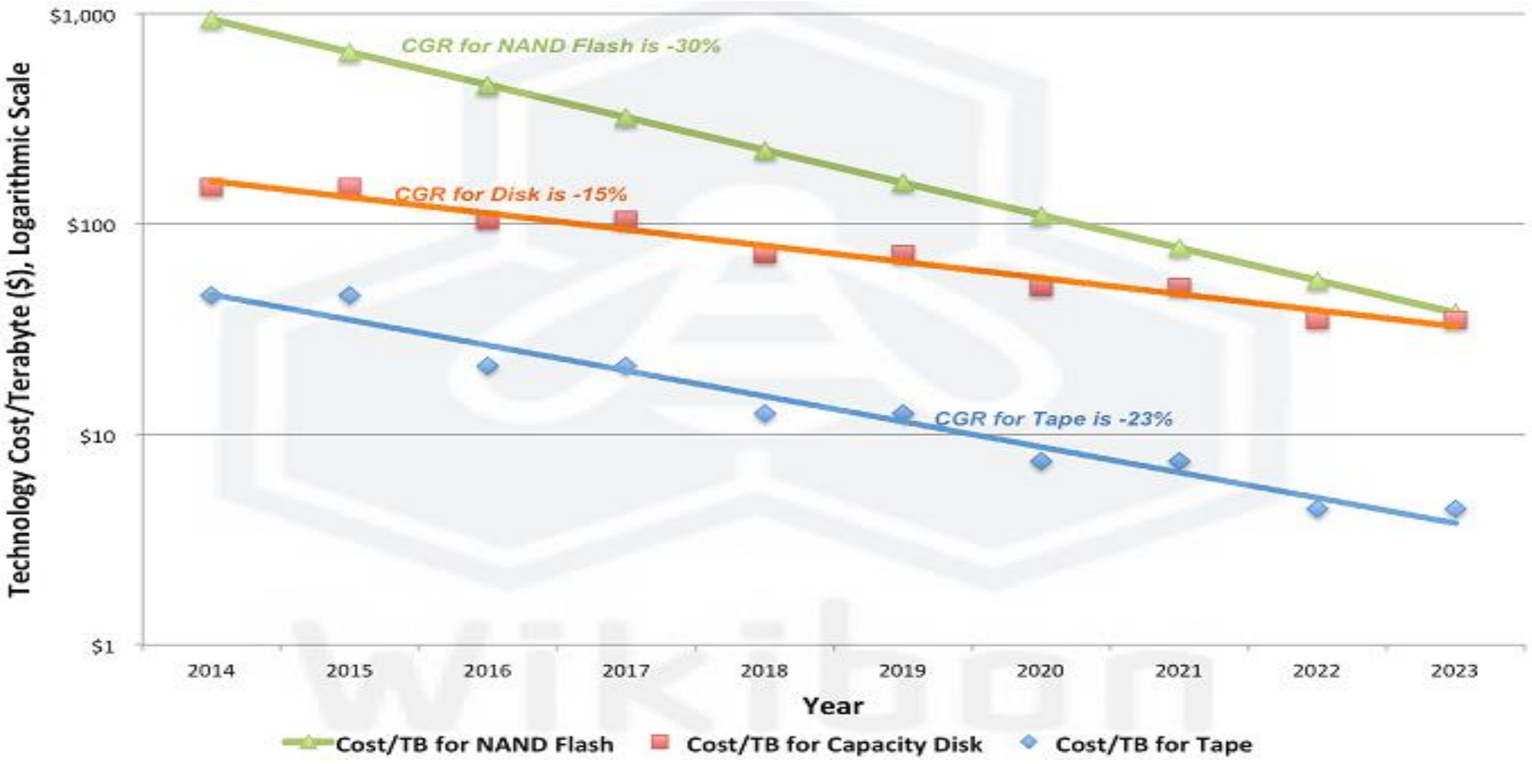
Managed NAND

Dell 2.5” SAS SSD Roadmap



Dell - Restricted - Confidential

SSDs and HDDs will still convive for a long time



Designing with SSDs

SSD as Cache

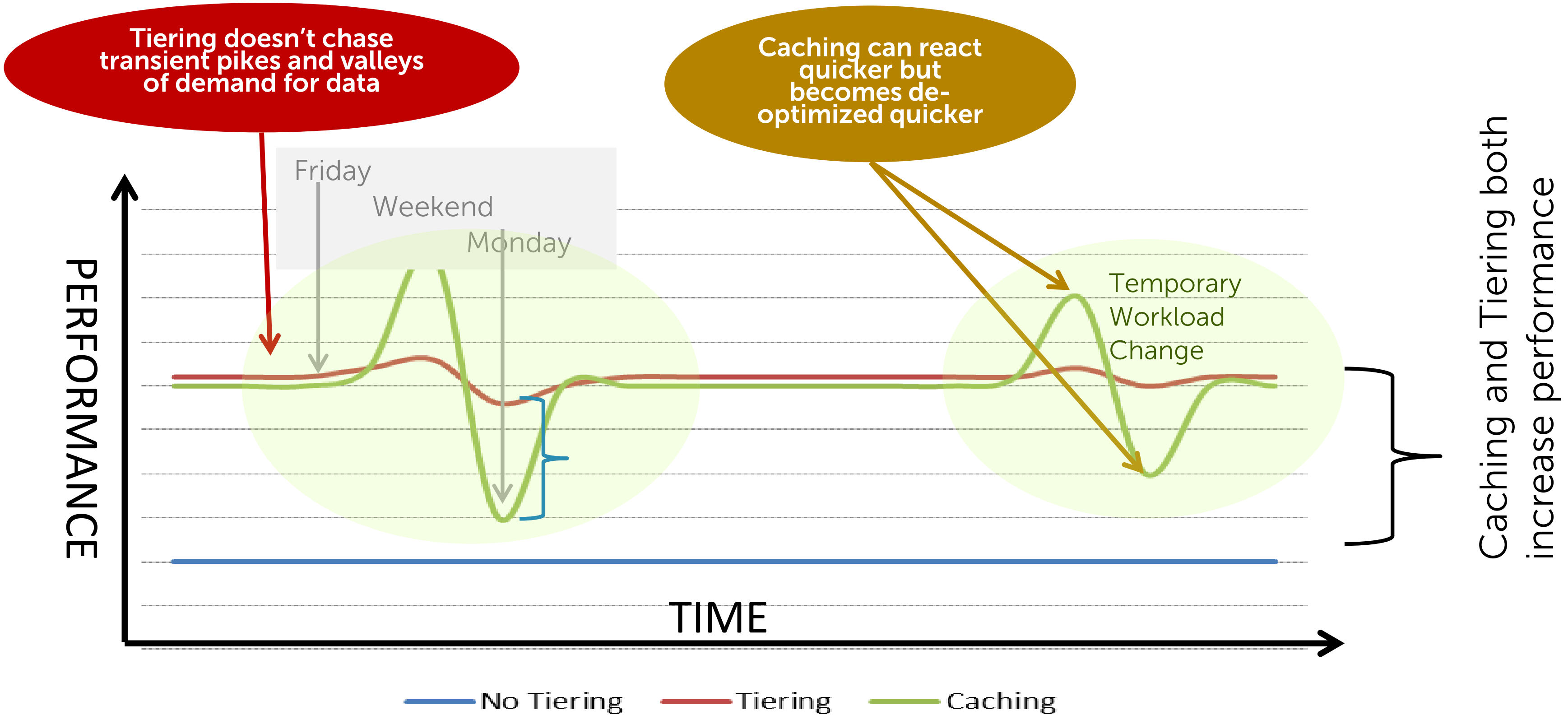
What is it?

- SSD cache moves data from an HDD virtual disk to the SSDs following a host read or write.
- Subsequent host read of the same LBAs can be read directly from the SSDs with a much lower response time than re-reading the data from the HDD virtual disk.
- **All PV MD36XX/38XX acquired by INFN can do this**

Workload characteristics that benefit from SSD Cache

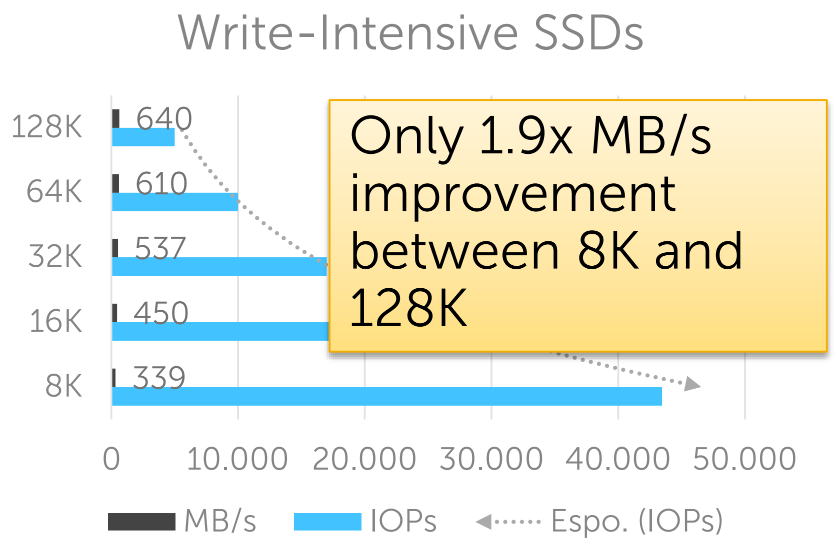
- Performance limited by HDD IOPs
- High percentage of Reads vs Writes / Large number of reads with intrinsic localty (repeated reads to the same or adjacent logical area of the LUN)
- The working size set that is repeatedly accessed is smaller than the SSD cache capacity.

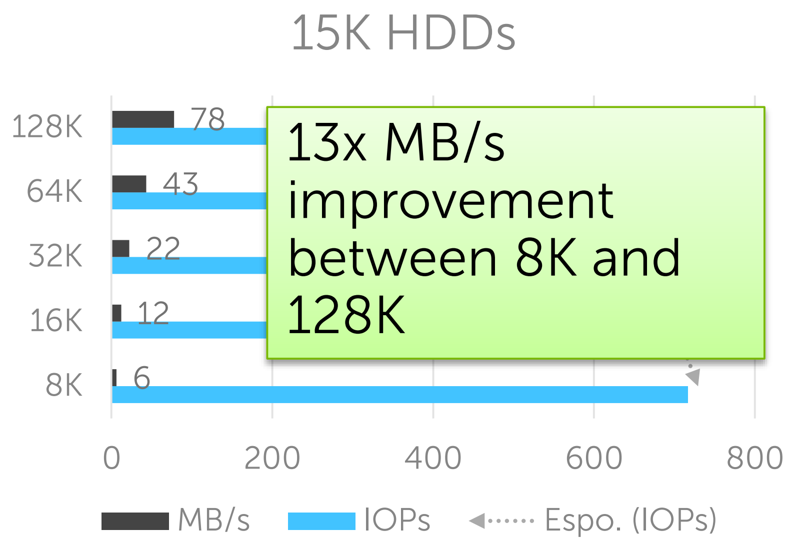
Caching is NOT Tiering (and vice-versa)…



SSDs have very high Write potential throttled by MB/s

100% Random Writes Raid 10



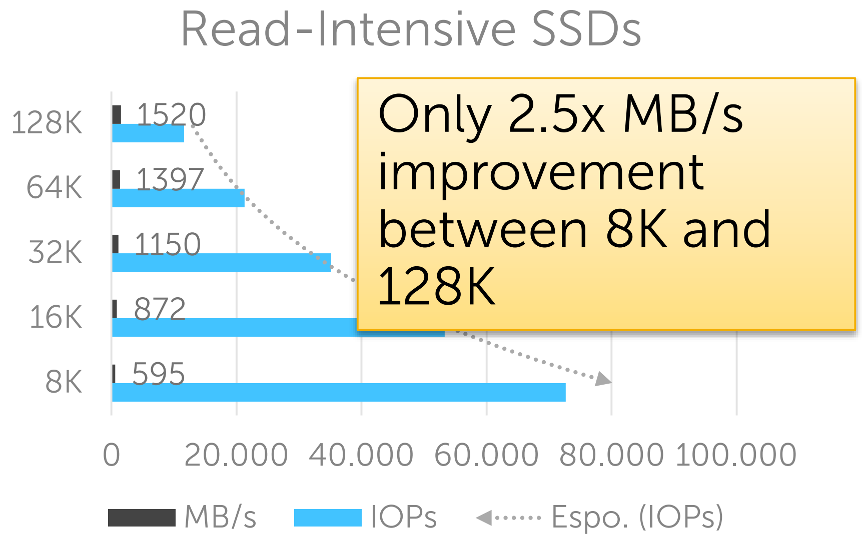


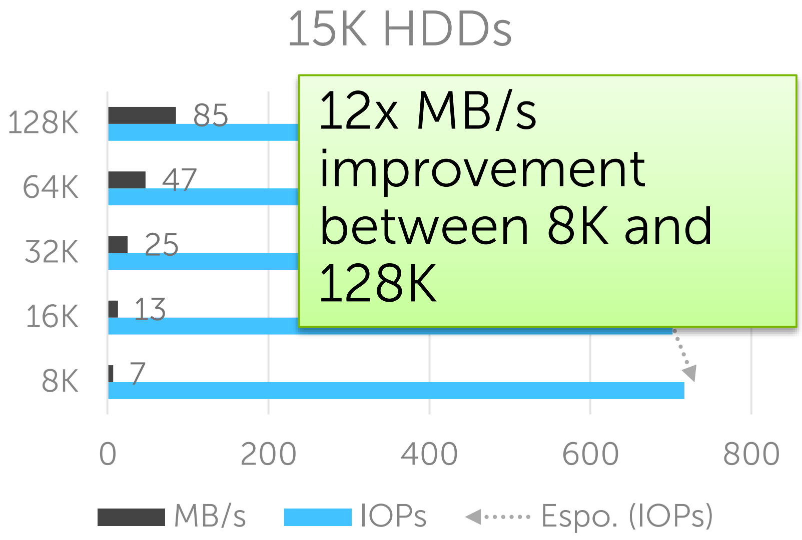
Use caution when sizing to IOP/s per disk method with SSDs*

Expect about 16x MB/s improvement between 8K and 128K

SSDs have very high Read potential throttled by MB/s

100% Random Reads Raid 5





Sizing may be more appropriately based on MB/s for SSDs *

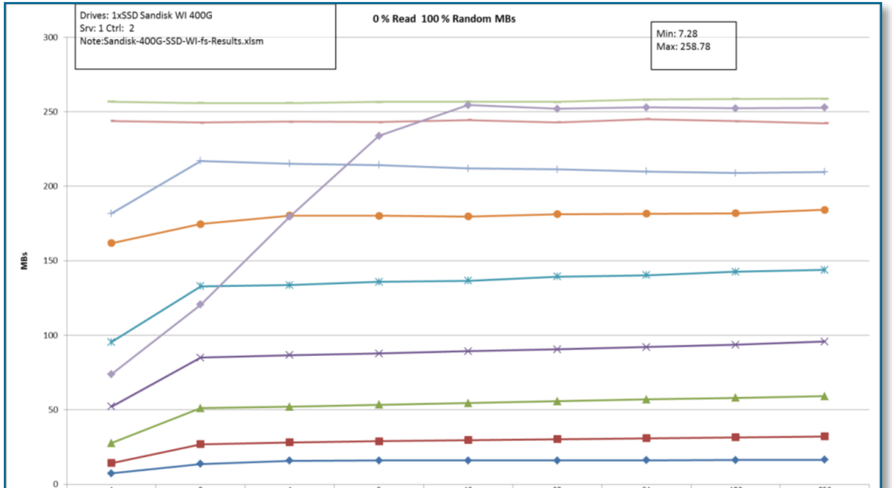
Designing with SSD and Spinning Disk Drives

Spinning Drives

- Need enough to provide needed IOPs

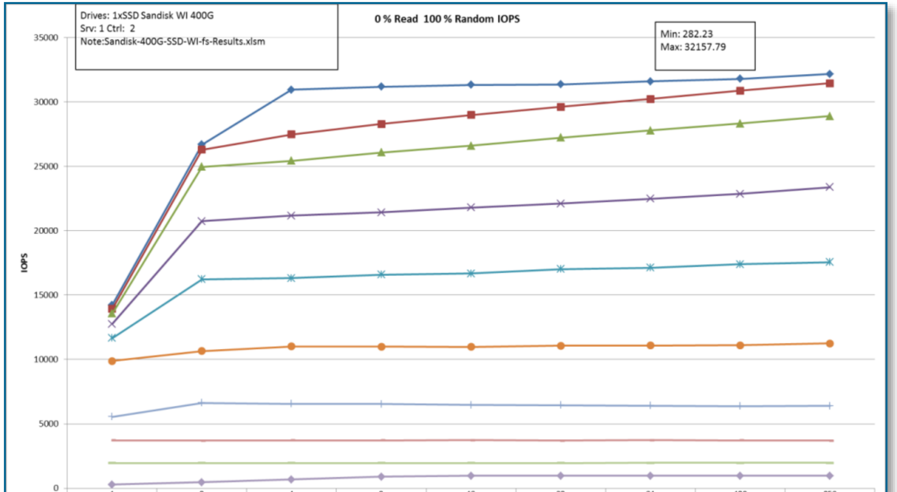
SSD

- Need to provide enough throughput



	1	2	4	8	16	32	64	128	256
512	7.283769	13.660804	15.847015	15.965485	16.040126	16.052258	16.174372	16.277342	16.46479
1k	14.251602	26.922944	28.139386	28.95379	29.681349	30.321949	30.958198	31.613901	32.200414
2k	27.768315	51.114551	52.060904	53.415607	54.478095	55.762433	56.889504	58.013675	59.170683
4k	52.230781	84.952892	86.693139	87.713989	89.294077	90.544059	92.088689	93.637902	95.773083
8k	95.502274	132.854913	133.701175	135.8515	136.6464	139.337544	140.342988	142.538584	143.871623
16k	161.881474	174.577718	180.296649	180.175541	179.796993	181.283156	181.554761	181.89878	184.2712
32k	181.663927	217.036077	215.097007	214.250748	211.864465	211.329611	209.983715	208.821523	209.636994
64k	243.83746	242.739205	243.282582	243.120427	244.468545	242.840227	245.11198	243.733901	242.294854
128k	256.728383	255.783869	255.842994	256.554[illegible]	256.70178	256.53437	258.175115	258.680581	258.776776
256k	73.984366	120.507684	179.65849	233.7770[illegible]	254.399452	252.002017	252.846414	252.313326	252.729826

System Queue Depth



	1	2	4	8	16	32	64	128	256
512	14226.11083	26681.25719	30951.2008	31182.58756	31328.37109	31352.06605	31590.56973	31791.68403	32157.79224
1k	13917.58022	26291.93772	27479.86934	28275.18567	28985.69264	29611.27846	30232.61481	30872.94973	31445.71692
2k	13558.74773	24958.27706	25420.36321	26081.83952	26600.63235	27227.75067	27778.07802	28326.98959	28891.93524
4k	12751.65558	20740.45222	21165.31708	21414.54799	21800.31168	22105.48304	22482.59015	22860.81586	23382.10043
8k	11657.9924	16217.64073	16320.94423	16583.43508	16680.46875	17008.97752	17131.71238	17399.72952	17562.45393
16k	9880.461059	10655.37827	11004.43413	10997.04232	10973.93754	11064.64577	11081.22319	11102.22045	11247.02148
32k	5543.94308	6623.415438	6564.239709	6538.413936	6465.590374	6449.267928	6408.194427	6372.727139	6397.613352
64k	3720.664375	3703.906325	3712.197601	3709.723318	3730.293954	3705.447802	3740.111999	3719.084187	3697.126064
128k	1958.682122	1951.476053	1951.927142	1957.357039	1958.479153	1957.20192	1969.719812	1973.576212	1974.310116
256k	282.22796	459.700333	685.342751	891.788704	970.456895	961.311406	964.532525	962.498953	964.087775

System Queue Depth

BS=32KB ⇨ 215MB/s ⇨ 1075 MB/s (6x Pack) ⇨ /2 = 537.5 MB/s (Raid10)

Maximum IOPS = 537.5 * 1024 = 550.400 KB / 32 KB = 17.200 IOPS (~6538*5/2)

Key Takeaways

- New Error Correction, Data Reduction and Cell Endurance algorythms can make Performance Drives replacement with SSDs a reality today.
- From now on, the drive industry will focus on SSDs more than 15k drives.
- 3,5” 15krpm drives will disappear shortly. Partial development will follow on 2.5” drives (lower seek time)
- Possible use of SSDs as Read Cache within INFN PV installed base. Read Cache, not Tiering!
- MB/s performance vs IO size scaling on SSDs not the same as HDDs. When dealing with SSD design, focus on MB/s performance and not on IOPS (If MB/s is a concern…)

Questions?