High-Performance SSD-Based RAID Storage

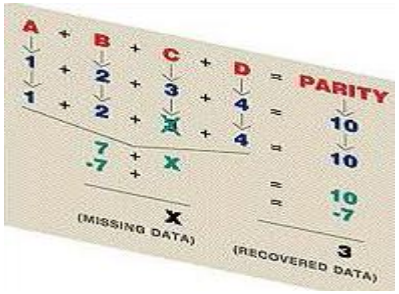
Madhukar Gunjan Chakhaiyar
Product Test Architect

Agenda

- HDD based RAID
- Performance-HDD based RAID Storage
- Dynamics driving to SSD based RAID Storage
- Evolution :Next gen Solution
- Turning Point - NAND SSD
- Performance Evaluation against Disk Specification
- Is adding SSDs worth investment over hard drive?
- SSD Internal Schema
- SSD based RAID Storage with Parity
- Performance Enhancement
- SSD based RAID Storage Solution

HDD based RAID

NO,, NOT THIS KIND OF RAID!!

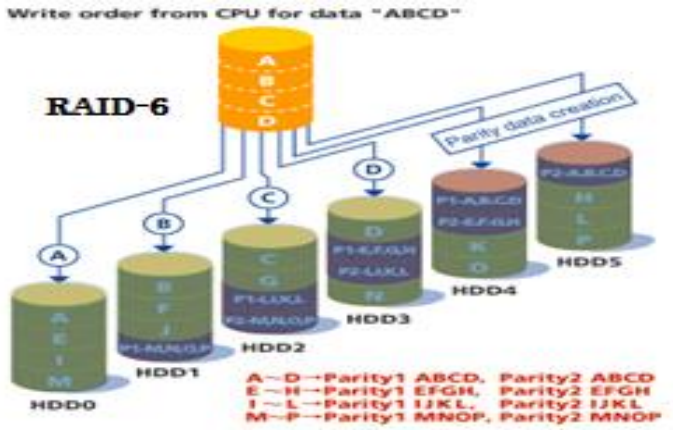


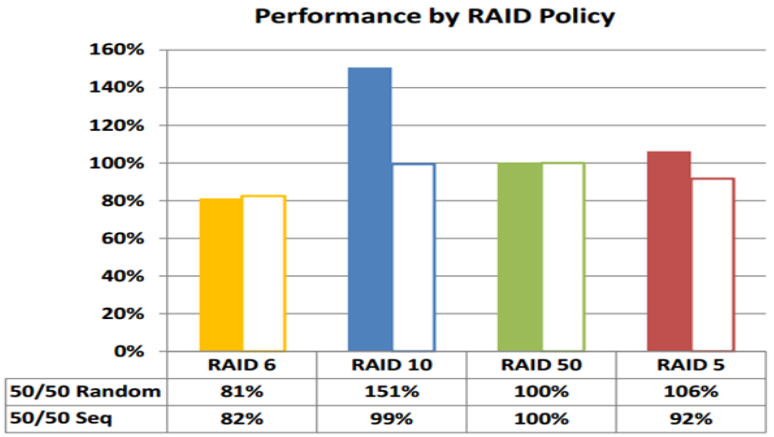
LBA Mapping for a 5-drive RAID 5 Array

DRIVE 1		DRIVE 2		DRIVE 3		DRIVE 4		DRIVE 5	
Host LBA	Drive LBA	Host LBA	Drive LBA	Host LBA	Drive LBA	Host LBA	Drive LBA	Host LBA	Drive LBA
0	0	8	0	0x10	0	0x18	0	P S	0
1	1	9	1	0x11	1	0x19	1	A E	1
2	2	0x0a	2	0x12	2	0x1a	2	R G	2
3	3	0x0b	3	0x13	3	0x1b	3	I M	3
4	4	0x0c	4	0x14	4	0x1c	4	T E	4
5	5	0x0d	5	0x15	5	0x1d	5	Y N	5
6	6	0x0e	6	0x16	6	0x1e	6	T	6
7	7	0x0f	7	0x17	7	0x1f	7	*	7
0x20	8	0x28	8	0x30	8	P S	8	0x38	8
	9		9		9	A E	9		9
	0x0a		0x0a		0x0a	R G	0x0a		0x0a
	0x0b		0x0b		0x0b	I M	0x0b		0x0b
	0x0c		0x0c		0x0c	T E	0x0c		0x0c
	0x0d		0x0d		0x0d	Y N	0x0d		0x0d
	0x0e		0x0e		0x0e	T	0x0e		0x0e
0x27	0x0f	0x2f	0x0f	0x37	0x0f	*	0x0f	0x3f	0x0f
0x40	0x10	0x48	0x10	P S	0x10	0x50	0x10	0x58	0x10
	0x11		0x11	A E	0x11		0x11		0x11
	0x12		0x12	R G	0x12		0x12		0x12
	0x13		0x13	I M	0x13		0x13		0x13
	0x14		0x14	T E	0x14		0x14		0x14
	0x15		0x15	Y N	0x15		0x15		0x15
	0x16		0x16	T	0x16		0x16		0x16
0x47	0x17	0x4f	0x17	*	0x17	0x57	0x17	0x5f	0x17

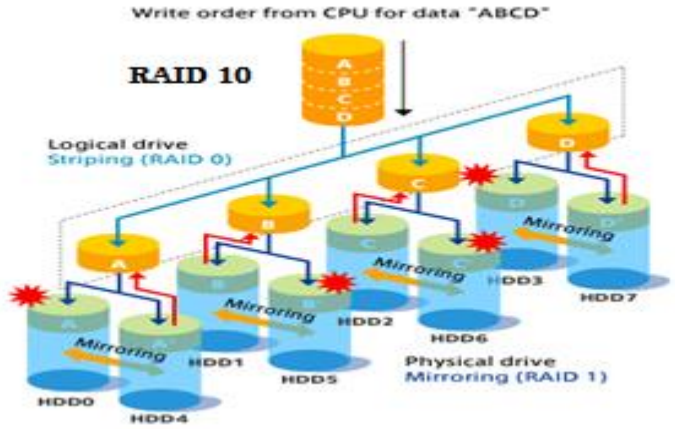
- RAID is a disk clustering technology- A Redundant array of Independent disks cabled together, so they appear as a single large device to the host systems.
- Split I/O operation into equal sized blocks and spread them evenly across the disks
- Fault Tolerance
- Data Protection
- Recovery and rebuilding of Data
- Blocks/Segment - A unit of storage made up of some number of blocks. In the diagram, a segment is shown to be made up of eight blocks. The size of a segment can vary from a minimum of eight blocks to a maximum of 0xfff8 blocks.
- Stripe - In the diagram, the storage area bounded by double lines is a “stripe.” The stripe size can be thought of in two ways, one not including the parity segment, and one including the parity segment.
- Parity - This segment contains data which is the “exclusive or” of the data contained in all the other segments on that stripe. This segment provides redundant information which allows for the regeneration of data which could be lost by a drive failure.

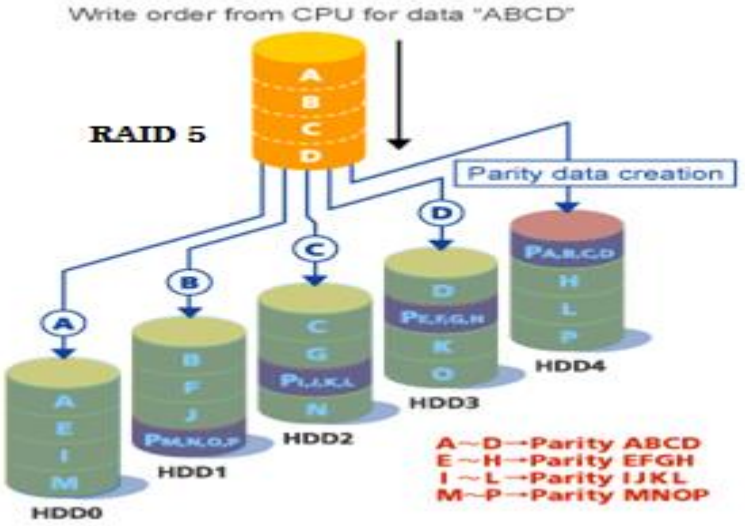
Performance-HDD based RAID Storage





	RAID 6	RAID 10	RAID 50	RAID 5
50/50 Random	81%	151%	100%	106%
50/50 Seq	82%	99%	100%	92%

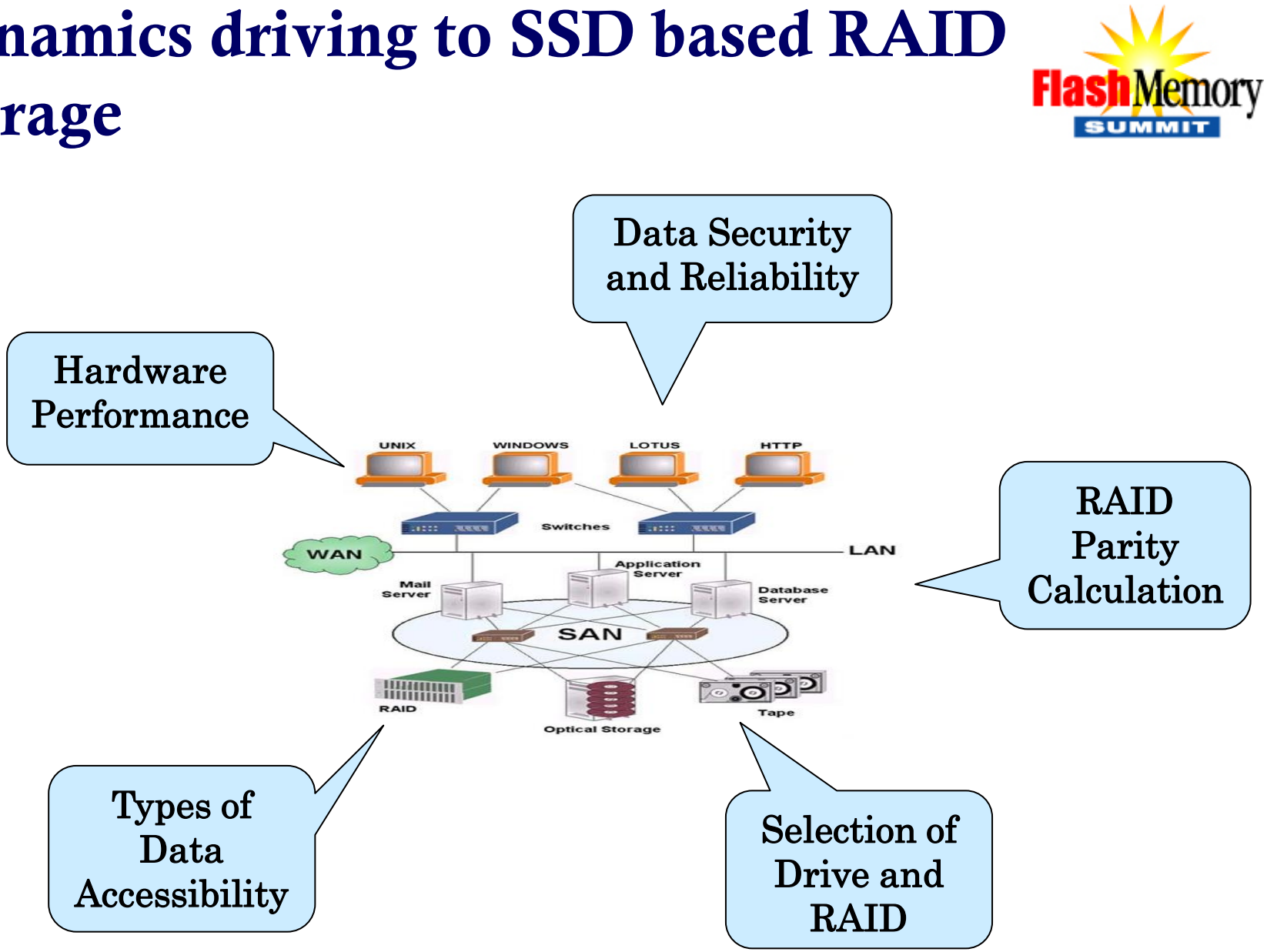




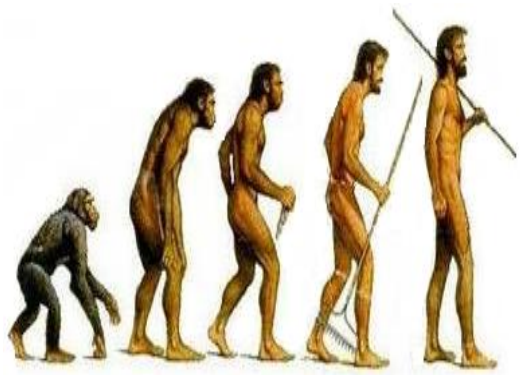
Source :: www.social.technet.microsoft.com/wiki/contents/articles/17634.storage-concepts-for-it-pros-raid-types-and-iops.aspx

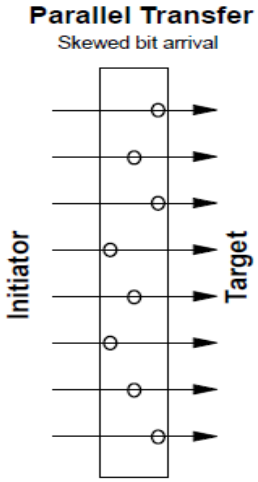
Source :: www.digitalmedix.com/raid_types.php

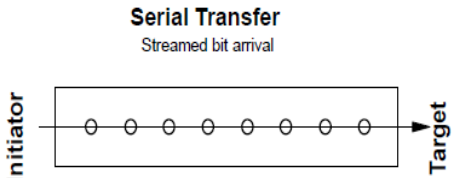
Dynamics driving to SSD based RAID Storage



Evolution :Next gen Solution







Turning Point - NAND SSD

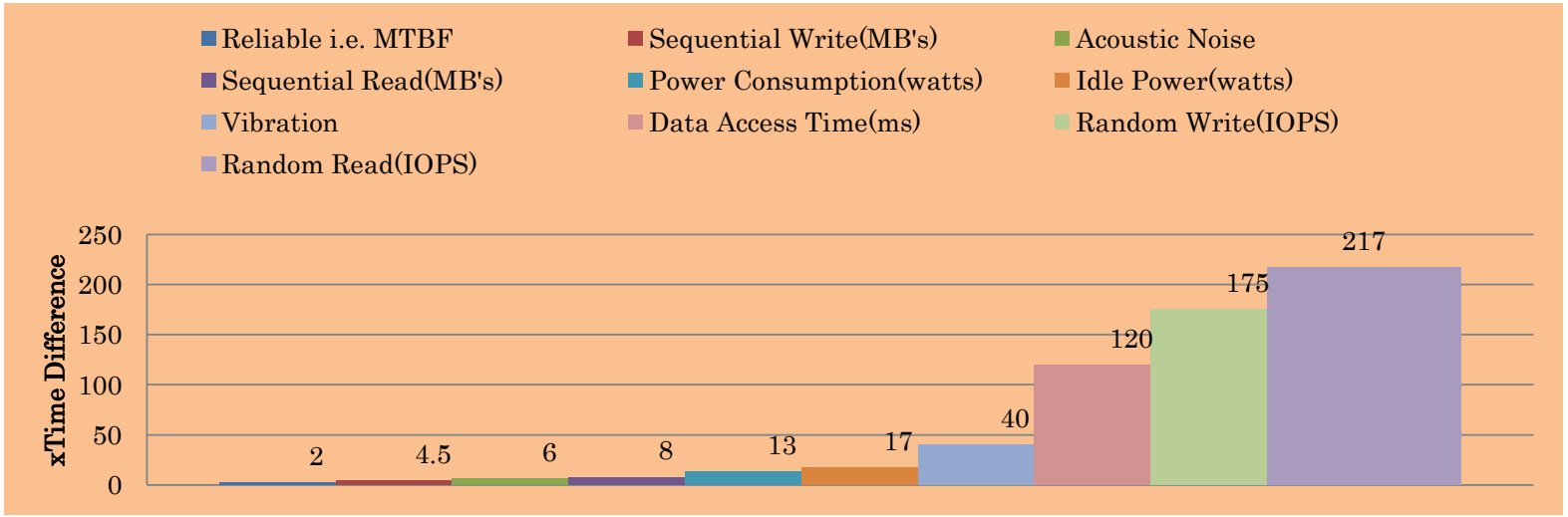
- Block I/O access
- Flash parallelism is a major factor in NAND SSD. Storage devices integrate NAND flash memory chips in parallel with IC to create SSD storage device.
- File System Suitability - Typically the same file systems used on hard disk drives can also be used on solid state disks. File system to support the TRIM command which helps the SSD to recycle discarded data.
- Wear leveling - Wear leveling is a process that is designed to extend the life of solid state storage devices. SSD controllers use a technique called wear leveling to distribute writes as evenly as possible across all the flash blocks in the SSD. Wear leveling also means that the physical address of the data and the address exposed to the operating system are different.
- Less Power Consumption

Turning Point - NAND SSD (contd.)

- High reliability and endurance - Mean time between failure is reduced as No moving parts deployed in Solid state drive. Enhanced write endurance for greater durability and 2 million hours MTBF.
- Compatibility with PCI - SSD contains host interface to support some form of physical host interface connection like USB, Fiber Channel, PCI-x Express, PCI-e Express, & SATA.
- IO Handling – SSD drive can handle high thread IO load.
- Compatible Form Factor - Typically comes in a 2.5" HDD form factor, or a custom form factor.
- Data Recovery - Ability to recover completely lost sectors, pages, blocks.
- SSDs offer a new level of RAID performance never seen before with traditional hard drives. Sequential read and write speeds indicate how a drive works with large contiguous files.

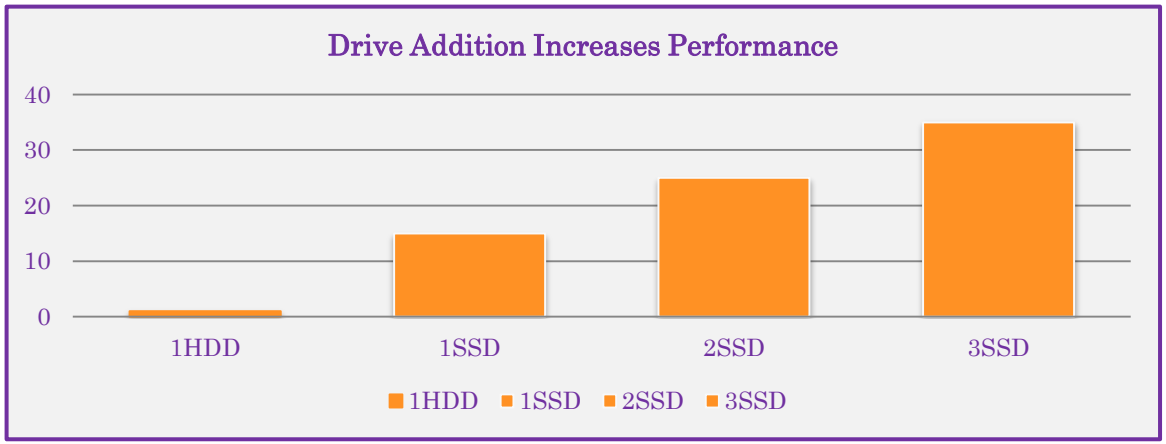
Performance Evaluation against Disk Specification

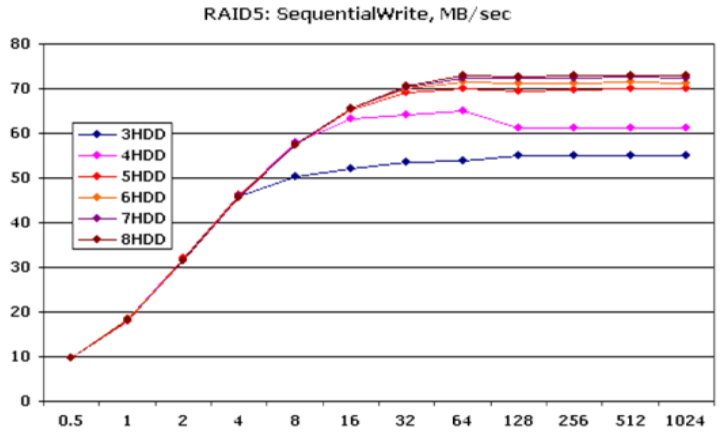
Performance factor	NAND Based SSD (500GB)	SATA HDD(7200rpm, 500GB)
Reliable i.e. MTBF	2M Hours	700K Hours
Sequential Write(MB's)	330	70
Acoustic Noise	None	0.6db
Sequential Read(MB's)	540	70
Power Consumption(watts)	0.12	1.75
Idle Power(watts)	0.04	0.8W
Vibration	20 G	0.5G
Data Access Time(ms)	0.1	~12
Random Write(IOPS)	70000	400
Random Read(IOPS)	98000	450

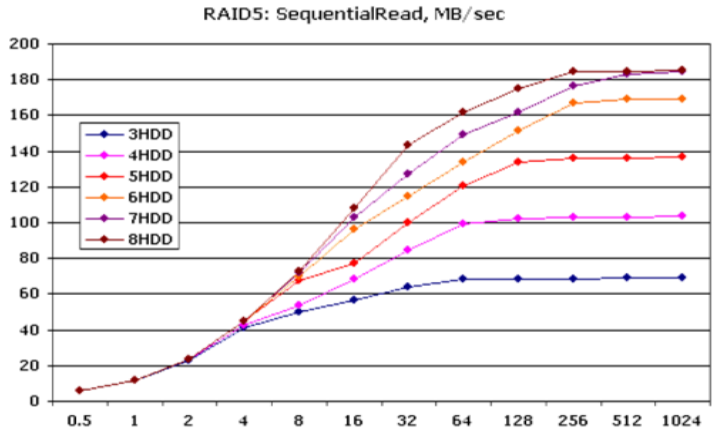


Is adding SSDs worth investment over hard drive?

- Yes, IO Performance increases with Drive addition.







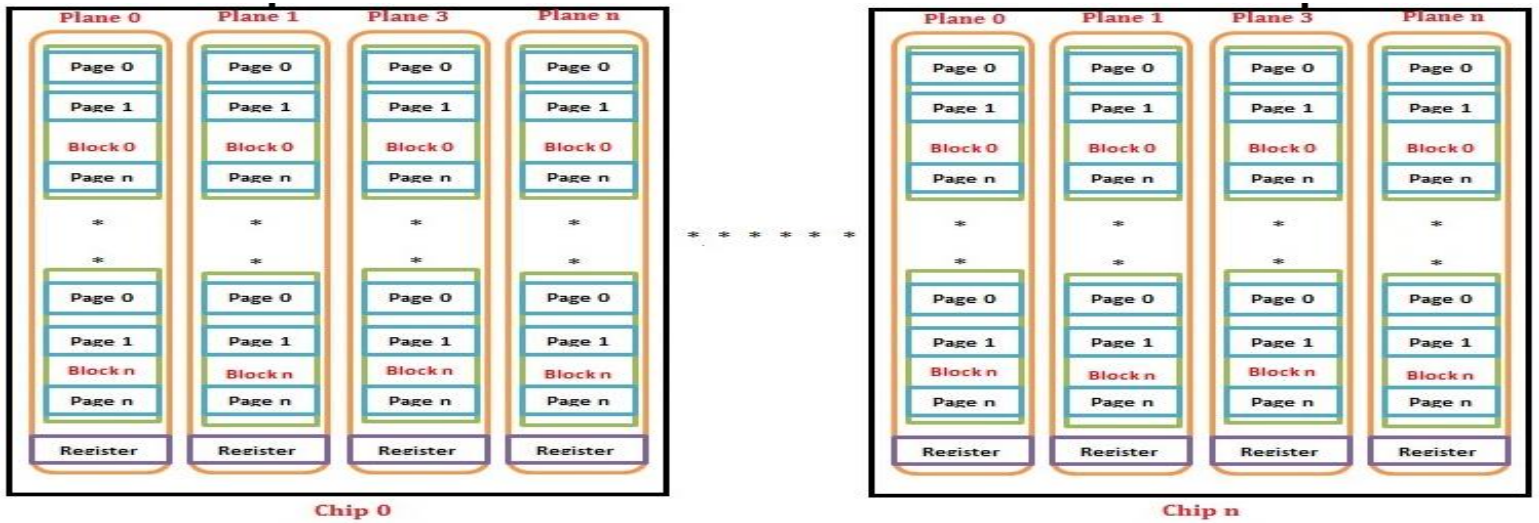
Source :: www.xbitlabs.com/articles/storage/display/3ware-7850.html

Is SSDs worth the investment over enterprise hard drives? (Contd.)

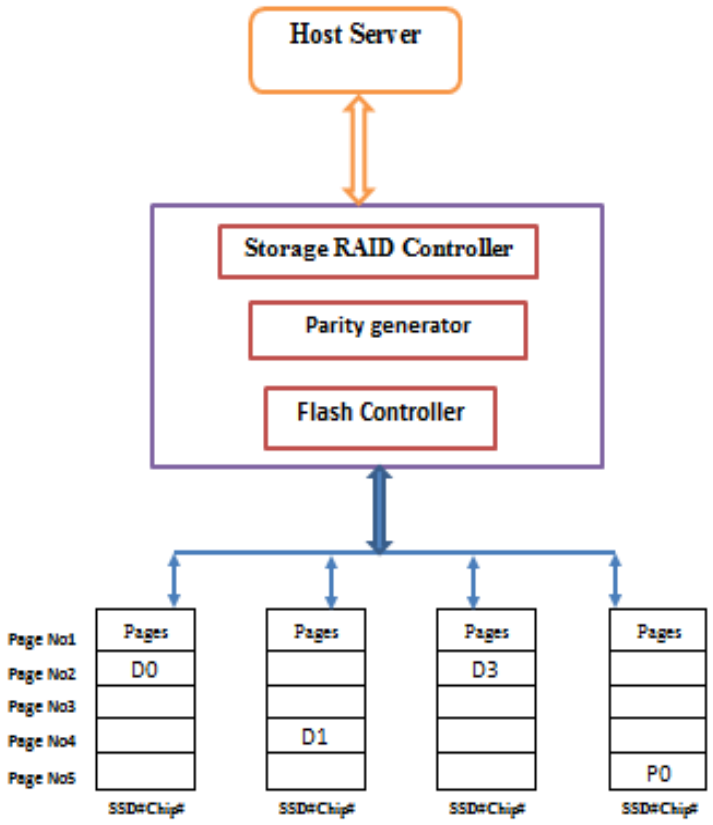
- A Hard Disk Drive are deeply affected by fragmentation, SSDs deliver consistent performance because of integrated circuits instead of physical spinning platters. Over time, as the outer sectors fill with data, the drive must write to progressively smaller sectors, which naturally store less data. Thus, additional movement is required to switch to the next available sector if more space is required. Additionally, data becomes fragmented with extended use, forcing the mechanical drive head to jump among inner and outer sections of the platter, negatively affecting performance even further.

SSD Internal Schema

- SSDs is an array of NAND Flash memory-based modules that are accessed in Parallel.
- First-gen SSDs used SLC (Single-Level Cell) flash, where each flash cell stores a single bit value.
- Emergence of MLC (Multi-Level Cell) technology stores multiple bits within each flash cell to expand capacity.
- NAND based flash package is composed of multiple chips. The chip is usually organized in planes which compose blocks with in.
- NAND based flash chips are organized in blocks consisting of multiple pages. The storage capacity per flash page ranges between 512 bytes and multiple kilobytes, while one block includes usually 64 or 128 pages.
- Both read and write operations are performed in unit of pages, and each page is of size 4KB
- A page is the smallest data unit for read and writes operations, while a block is the smallest data unit for erasing.



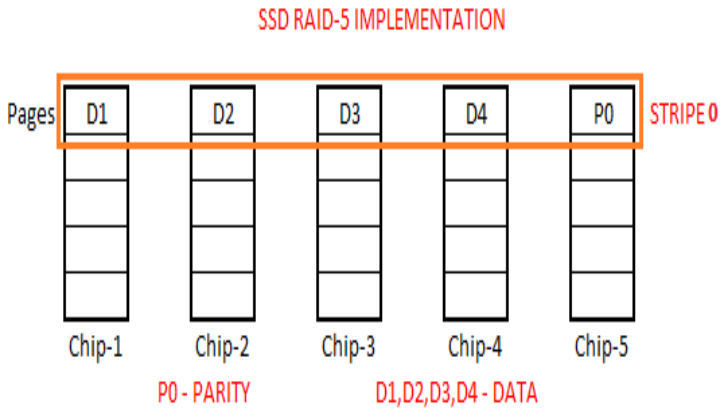
SSD based RAID Storage with Parity

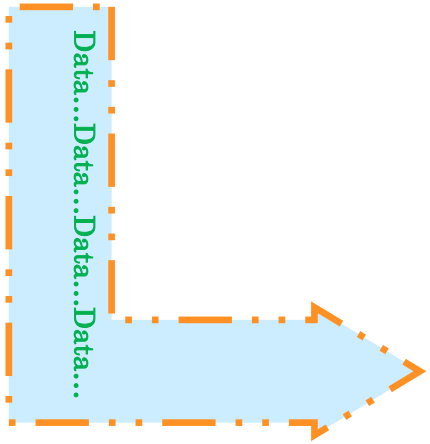


SSD Internal Provisioning at Enterprise Storage

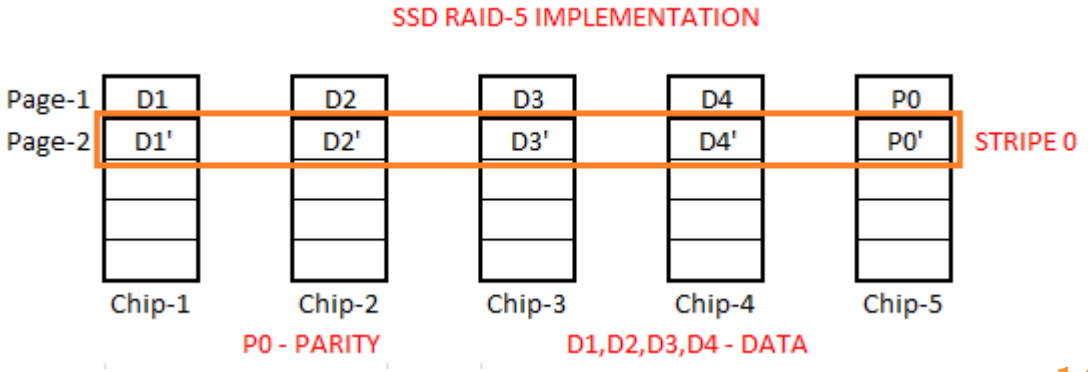
- SSD drives are kept at the backend of Storage controller and are connected to Storage RAID controller via Flash IO channel of the controller.
- To implement RAID with parity solution across SSD disk or chip the Storage RAID controller takes the assistance of Parity generator.
- Parity generated by parity generator is distributed across SSD drive and chip by Flash controller.
- A flash controller is the part of solid state flash memory that communicates with the host device and manages the flash file system directory.
- Even parity distribution - As an example, RAID-5, represented by (20; 20; 20; 20; 20) for 5 devices, where parity is distributed evenly across all devices.

Performance Enhancement : Read Modify Write

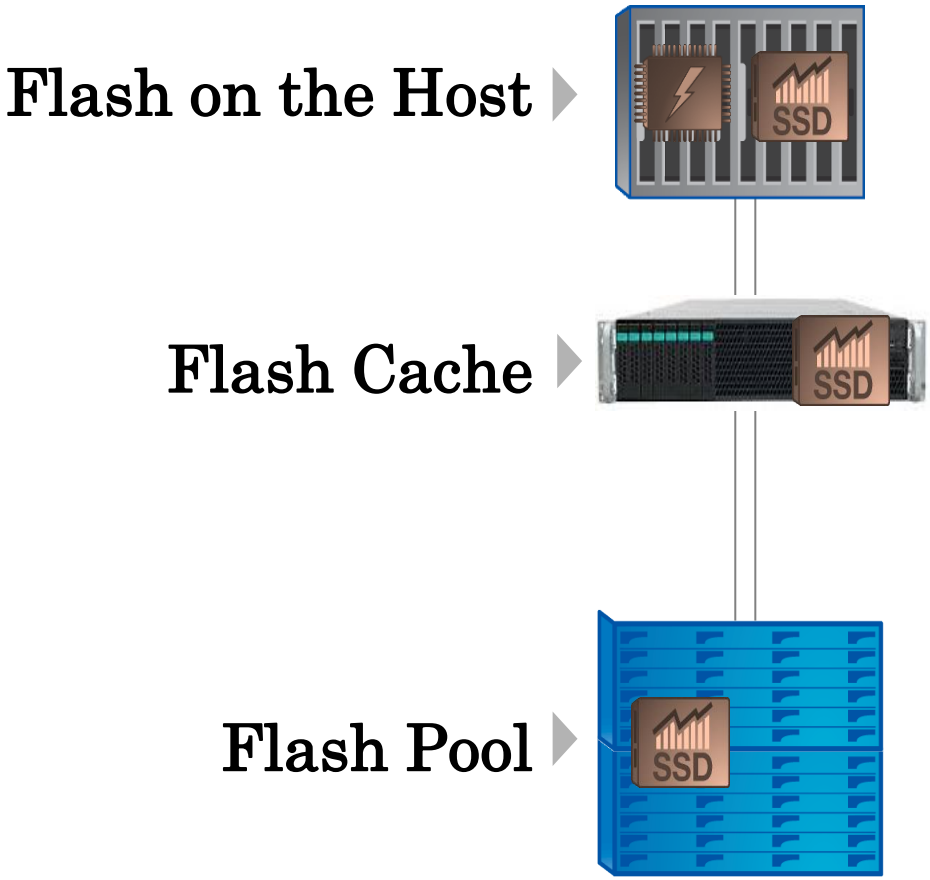


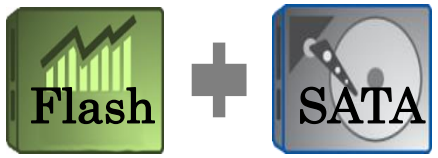


- Stripe 0 consists of pages D1 to D4 and parity P0. Stripe 0 metadata contains the location as (Pg1-Chp1, Pg1-Chp2,Pg1-Chp3,Pg1-Chp4,Pg1-Chp5)
- Assume data pages D1 through D4 are updated.
- Note that unlike disk-based RAID-5 where each old strip would be overwritten,
- Using read modify write, existing data must be read to calculate the new parity so that the new data and parity can be written to the chips.
- New updated data and updated parity must be written to the same chips.
- Metadata for stripe get updated with new Page No for each data i.e. as (Pg2-Chp1, Pg2-Chp2,Pg2-Chp3,Pg2-Chp4,Pg2-Chp5)

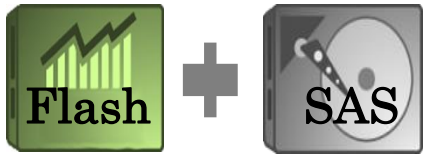


SSD based RAID Storage Solution





- 40% lower cost
- 55% lower $/GB



- 50% lower cost
- 50% lower $/IOPs


Thank
You!