

# **Hybrid Redundancy System**

## **New Approach to SSD Redundancy**

**Avraham (Poza) Meir**

**Anobit**

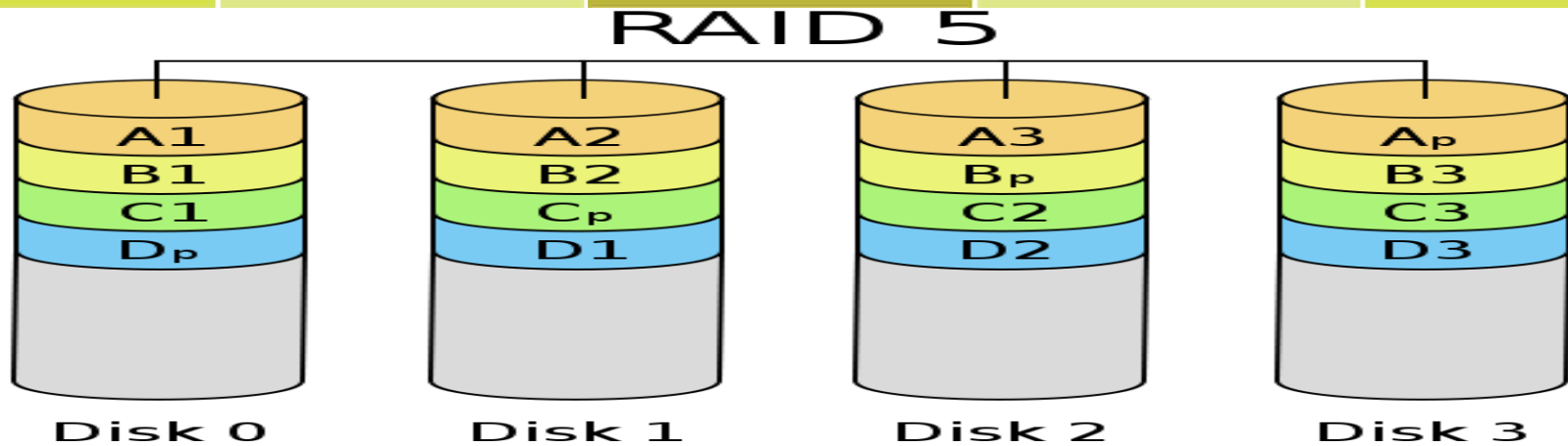
# Agenda

- ❑ Preface
- ❑ RAID Background
- ❑ SSD Reliability Myth
- ❑ SSD - Redundancy Implementation
- ❑ Hybrid Redundancy System
- ❑ RAID Systems - Comparison
- ❑ Summary

- ❑ SSD popularity is growing
- ❑ SSD is still considered similar to HDD
  - ❑ Same interfaces
  - ❑ Same RAID systems
- ❑ Basic differences
  - ❑ One NAND die is equivalent to plate-side + head
  - ❑ Each “head” is much faster than HDD head
- ❑ This presentation suggests an alternative approach to the reliability of multiple SSD systems
- ❑ In order to better understand it, let’s start with existing RAID systems

- ❑ RAID = Redundant Array of Independent Disks
- ❑ The purpose of RAID
  - ❑ Reliable Storage – create high reliability storage system using “regular reliability” elements
  - ❑ “No Single Point of Failure” – as enterprise systems need to provide very high data availability and reliability. The outcome: everything is duplicated (cables, ports, hosts est.)

# RAID Basics 2



## How RAID works

- Organization - data is organized in stripes (Stripe A, Stripe B etc). Each stripe has at least one parity check data
- Recovery – when drive fails its data is recovered by XORing the other data in the stripe

## Small Write Example (B1 Update)

- Process
  - Read B1
  - Read B<sub>p</sub>
  - Write New-B1
  - Subtract B1 from B<sub>p</sub> (XOR operation)
  - Add new-B1 to B<sub>p</sub> (XOR operation)
  - Write New-B<sub>p</sub>

Small single write is translated to 2 read and 2 write operations

# SSD Reliability Myth

- ❑ SSD Reliability
  - ❑ Number of NAND dies - typical SSD has 32 - 256 NAND dies
  - ❑ NAND die reliability: 10 - 100 FIT
  - ❑ SSD MTBF: 30K - 3M hours
- ❑ Reliability improvement
  - ❑ Improving Die Reliability - using better NAND screening
  - ❑ Redundancy - add redundant NAND dies and calculate parity in a similar manner to RAID.
  - ❑ Partial redundancy - enable recovery of failed page/block
- ❑ Redundancy implications (for small write operations)
  - ❑ Basic solution
    - ❑ Natural performance reduction as each operation is translated to 2 read + 2 write operations.
    - ❑ Write efficiency/Write amplification - write amplification is doubled which reduces performance to 50%
    - ❑ Logical endurance - reduced significantly due to write amplification increase
  - ❑ Complex RAID - complex techniques can reduce write amplification and recover most of the performance.

- ❑ SSD with internal RAID reliability
  - ❑ Good SSD MTBF: 3M - 19M hours
  - ❑ Lower Performance
  - ❑ Lower Endurance
- ❑ SAN RAID
  - ❑ RAID “on top of reliable SSD” will be implemented in any case as:
    - “No single point of failure”
  - is a major requirement

# Hybrid RAID Approach

- ❑ Principle
  - ❑ Use the critical SAN RAID system for recovery
  - ❑ SSD will have spare areas where recovered data can be written.
  - ❑ Note: spare areas are natural to SSD, it is the SSD overprovision.
- ❑ Structure
  - ❑ Overall System – regular RAID system
  - ❑ SSD – each drive has spare dies.
- ❑ How does it works
  - ❑ At the beginning, each SSD uses the spare dies as an additional over-provisioning area
  - ❑ At the point the SSD detects failed die, it reports to the RAID controller which LBAs were lost
  - ❑ RAID controller calculates the LBAs using other drives and writes them back to the SSD.
  - ❑ The SSD spreads the data on the other dies (reduces the original over-provisioning)
- ❑ The Benefits
  - ❑ Spare dies are added to over-provisioning and improve write efficiency
  - ❑ The whole over-provision can be used to support recovery of failed dies vs. a single die in the conventional approach.
  - ❑ No write amplification

# Comparison Criteria

- ❑ Reliability
  - ❑ Number of failed dies that can be corrected in one SSD
- ❑ IOPS / \$
  - ❑ The major benefits of SSD vs HDD = IOPS
  - ❑ This criterion represents the economical value of IOPS while all other parameters are the same (Density, Lifetime ....)
- ❑ Life GB/\$
  - ❑ Life GB represents the overall GB written to an SSD in its life time. The purpose of this definition is to neutralize the FLASH quality from cost structure. (Logical Endurance x Logical Density)
  - ❑ The criterion represents the life GB in economical values.

# Comparison Table

	External RAID+ Internal RAID	Hybrid RAID
Reliability <sup>1</sup>	Capable to correct <u>N</u> dies failure	Capable to correct <u>N+1</u> dies failure
IOPS / \$ <sup>2</sup>	50% - 90%	100%
Life GB / \$ <sup>3</sup>	50% - 90%	100%

## Comparison assumptions

- 200 GB user, "256" GB Gross Density
- NAND - 8GB NAND Dies
- SSD – 200GB + (N+1) spare dies for over provisioning and redundancy
- Over Provision – 25-27%

## Comments:

1. Hybrid system can handle many failed dies up to the over-provision (25-27%) with some performance degradation. It is assumed that after each die failure the stripe is re-built
2. As an outcome of performance improvement
3. As an outcome of endurance improvement

# Implementation Issues

- ❑ SAS Protocol – SAS protocol has the ingredients to support such operation
- ❑ SATA Protocol – doesn't have such support, needs updates

# SSD Reliability and Performance Definitions

## Reliability Definition

- Legacy – a Failure is consider as 1<sup>st</sup> read failure
- New – failure is considered till TBD read failures

## Performance and Performance Variation:

The new approach can lead to a product with inconsistent performance. Suggested below 2 definitions. The first is for applications that can benefit higher performance, when available, while the second is for applications that need consistency

### Large performance variance is allowed

- Minimal performance
- Average performance over many drives
- Average performance over long time

### Limited performance variance

- Average performance
- Performance Variance within relative short period

- ❑ New approach to SSD redundancy was proposed
- ❑ The Hybrid RAID approach is based on:
  - ❑ Hi-Level RAID will be implemented in any case (“no single point of failure”)
  - ❑ SSD by nature has many NAND dies, additional dies can be used as over-provision. Over-provision can be changed dynamically as dies fail.
  - ❑ Use the hi-level RAID data recovery to improve “SSD component” reliability with additional gain in performance and usability.
- ❑ Benefits
  - ❑ Better cost
  - ❑ Better reliability
- ❑ Application comment:
  - ❑ The Hybrid RAID approach is good for large systems with many SSDs. It is probably not efficient for DAS with a small amount of drives

- ❑ Our Expertise – FLASH enhancement
  - ❑ Better performance and significantly greater endurance.
- ❑ Invented new technology – Memory Signal Processing (MSP)
- ❑ Over 72 Patents (19 granted)
- ❑ 2 Business Lines
  - ❑ Consumer NAND controller for embedded mobile applications
    - ❑ Over 20M controllers shipped
  - ❑ Enterprise SSD – MLC SSD with SLC performance and reliability, leveraging MSP technology.
    - ❑ “Genesis” 2<sup>nd</sup> drive generation just announced
- ❑ Strategic relationship with NAND Vendors

**We Make Flash Better ...**

**Thank You**