SSDs and RAID: What’s the right strategy

Paul Goodwin
VP Product Development
Avant Technology
SSDs and RAID: What’s the right strategy

- Flash Overview
- SSD Overview
- RAID overview
- Thoughts about Raid Strategies for SSDs
Flash Overview

• The smallest unit of NAND-based flash storage that can be written and subsequently read is a page.
  – Today these pages are typically 4 KB in size.
  – All bits in a blank page are set to 1s, and writing data to the page involves setting some of the bits within it to 0s.
  – Individual bits within a page cannot be reset to 1s; rewriting the page to a different bit pattern requires an intermediate erase operation that resets all bits back to 1.
• Erasures are performed over large blocks (e.g., of 128 KB) spanning multiple pages.
  – Blocks wear out as they are erased
  – Exhibiting increasing BERs
• The PE cycle limit is the point where the BER becomes unmanageably high.
SSD Overview

• An SSD is comprised of a specialized processor with an I/O interface, typically SATA, on one side and a number of flash interfaces on the other.
SSD Overview

• The Function of the controller is to look like a Disk Drive.
  – store and retrieve data to the flash devices connected to the back side of the controller.

• Additionally the controller has to manage the idiosyncrasies of the flash devices. In particular manage the PE cycle limitation of the flash devices. Wear leveling.
RAID Overview:

Increasing performance of CPUs and memories will be squandered if not matched by a similar performance increase in I/O. While the capacity of Single Large Expensive Disks (SLED) has grown rapidly, the performance improvement of SLED has been modest. Redundant Arrays of Inexpensive Disks (RAID), based on the magnetic disk technology developed for personal computers, offers an attractive alternative to SLED, promising improvements of an order of magnitude in performance, reliability, power consumption, and scalability. This paper introduces five levels of RAID.

- Patterson; A Case for Redundant Arrays of Inexpensive Disks (RAID) 1987
RAID Overview: Enter SSD

Increasing performance of CPUs and memories will be squandered if not matched by a similar performance increase in I/O. While the capacity of Single Large Expensive Disks (SLED) has grown rapidly, the performance improvement of SLED has been modest.

Redundant Arrays of Inexpensive Disks (RAID), based on the magnetic disk technology developed for personal computers, offers an attractive alternative to SLED, promising improvements of an order of magnitude in performance, reliability, power consumption, and scalability. This paper introduces five levels of RAIDS.

Redundant Arrays Inexpensive Disks
SSD are not really that really “inexpensive”
SSD are not Disks
New base to the acronym
Redundant Array of “Independent Drives”
Increasing performance of CPUs and memories will be squandered if not matched by a similar performance increase in I/O. While the capacity of Single Large Expensive Disks (SLED) has grown rapidly, the performance improvement of SLED has been modest. Redundant Arrays of Inexpensive Disks (RAID), based on the magnetic disk technology developed for personal computers, offers an attractive alternative to SLED, promising improvements of an order of magnitude in performance, reliability, power consumption, and scalability. This paper introduces five levels of RAID
RAID Overview:

The advantages of RAID are not without cost

• Performance
  – HD have only increased ~1.3X in speed over the past 10 years.
  – SSDs can, on their own, approach the wire rate.
  – Thru the RAID 0 component the SET can reach wire rate

• Reliability
  – HDs have random, unpredictable failures
  – SSDs have predictable and manageable failures
  – RAID is good for random and unpredictable failures

• Power consumption
  – HDs are power hungry
  – SSDs are lower in power
  – RAID has a power penalty for implementation

• Scalability
  – HDs have a power and vibration problem when configuring large arrays
  – SSDs scale well although are capacity challenged
  – RAID has mechanical and power penalties

Increasing performance of CPUs and memories will be squandered if not matched by a similar performance increase in I/O. While the capacity of Single Large Expensive Disks (SLED) has grown rapidly, the performance improvement of SLED has been modest. Redundant Arrays of Inexpensive Disks (RAID), based on the magnetic disk technology developed for personal computers, offers an attractive alternative to SLED, promising improvements of an order of magnitude in performance, reliability, power consumption, and scalability. This paper introduces five levels of RAID.
RAID Overview: RAID 0

*Not really RAID*

_Data is split evenly across two or more disks with no parity or redundancy._

- Fairly large performance benefits
  - For accesses that are larger than the stripe size the seek time of the array will be the same as that of a single drive.
  - For accesses that are smaller than the stripe size allow the drives to seek independently.
  - If the sectors accessed are spread evenly between the drives, the apparent seek time of the array will be half that of a single drive.
  - The transfer speed of the array will be the transfer speed of all the disks added together, the speed of the RAID controller becomes the limiting factor.
- If you are going to have more than one drive you should set them up in a RAID 0 set.
RAID 0: Implementations

*RAID 0 may significantly increase the performance of an SSD Storage subsystem*

- It’s all a matter of where you place it
- An external RAID enclosure may not provide any performance benefit for SSDs
  - SSDs running near SATA max won’t benefit if connected via eSATA
- A Raid controller directly connected to the I/O will realize the benefits of RAID 0.
RAID 0: Implementations

Storage subsystem on a card

<table>
<thead>
<tr>
<th>Density</th>
<th>40 GB to 4 TB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form Factor</td>
<td>PCIe, 8-lane</td>
</tr>
<tr>
<td>IOPS</td>
<td>Up to 240K Read/Write</td>
</tr>
<tr>
<td>Sequential or Random R/W</td>
<td>&gt;2 GB/s</td>
</tr>
<tr>
<td>RAID</td>
<td>0, 1, 5, 6, 10, 50, 60</td>
</tr>
</tbody>
</table>
RAID Overview: RAID 1

- Groups \( n + n \) disks as one virtual disk with the capacity of \( n \) disks.
- Data is replicated on the two disks.
- When a disk fails the data will be read from the failed mirror drive.
- Read and write performance equal to a single drive.
- Full Redundancy for protection of data.
- RAID 1 is more expensive in terms of disk space since twice the number of disks are used than required to store the data without redundancy.
RAID Overview: RAID 10

- Mirrored sets in a Striped configuration
- Provides fault tolerance and improved performance.
RAID Overview: RAID 5

provides data redundancy by using data striping in combination with parity across all physical drives in the set.

• Groups n disks as one large virtual disk with a capacity of (n-1) disks.

• Redundant information (parity) is alternately stored on all disks.

• When a disk fails, the virtual disk still works, but it is operating in a degraded state. The data is reconstructed from the surviving disks.

• Better read performance, but slower write performance.

• Redundancy for protection of data.
RAID Overview: RAID 6

provides data redundancy by using data striping in combination with double distributed parity across all physical drives in the set.

- RAID 6 groups n disks as one large virtual disk with a capacity of (n-2) disks.

- Parity information is alternately stored on all disks.

- The Raid 6 set remains functional with up to two disk failures. The data can be reconstructed from the surviving disks.

- Better read performance, but slower write performance.

- Increased redundancy for protection of data.

- Two disks per span are required for parity. RAID 6 is most expensive in terms of disk space.
RAID Overview: RAID 5 and RAID 6

RAID 5 and RAID 6 are expensive in terms of Flash operations

- Read the data block from drive
- Read the parity block from drive
- Compare the old data block with the write request.
  - For each bit that has flipped in the data block, flip the corresponding bit in the parity block
- Write the updated data block
- Write the updated parity block
RAID Overview: RAID 50

*Striping a RAID 5 drive set with a second RAID 5 Drive set.*

- Groups $n^s$ disks as one large virtual disk with a capacity of $s^*(n-1)$ disks, where $s$ is the number of spans and $n$ is the number of disks within each span.

- Redundant information (parity) is alternately stored on all disks of each RAID 5 span.

- Better read performance, but slower write performance.

- Requires proportionally as much parity information as standard RAID 5.

- Data is striped across all spans.

![Diagram of RAID 50 configuration with striping across multiple spans.](image-url)
RAID Overview: RAID 60

Striping a RAID 6 drive set with a second RAID 6 drive set.

- Groups N disks as one large virtual disk with a capacity of \( S \cdot (N-2) \) disks, where \( S \) is the number of sets and \( N \) is the number of disks within each span.

- Redundant information (parity) is alternately stored on all disks of each RAID 6 set.

- Better read performance, but slower write performance.

- Requires proportionally as much parity information as standard RAID 6.

- Data is striped across all spans.
RAID and SSD:

- Does RAID cause write amplification in SSDs?
  - In Parity applications (RAID 5 and RAID 6) can writing a small block result in parity that affects multiple blocks?
  - We are currently doing research on this to determine the effects of parity on SSD endurance.
  - Some evidence is leaning towards supporting this.

- Some of this investigation is blind since we don’t know the actual algorithms.
RAID and SSD:

- SSDs don’t eliminate the need for RAID
  - Although less likely they can still have random failures

- For RAID using SSDs we prefer RAID 10
  - Because of the potential for RAID write amplification.
  - Simple to implement
  - RAID 10 is the fastest to rebuild
  - Same cost as RAID 6
  - Faster than RAID 5 or RAID 6

- RAID may help survive increased errors as the life cycle of a SSD is reached but swapping out drives is still necessary.
  - Although they are likely to have been replaced by larger drives before that time is reached.
RAID and SSD:

- SSDs don’t eliminate the need for RAID
  - Although less likely they can still have random failures

- For RAID using SSDs we prefer RAID 10
  - Because of the potential for RAID write amplification.
  - Simple to implement
  - RAID 10 is the fastest to rebuild
  - Same cost as RAID 6
  - Faster than RAID 5 or RAID 6

- RAID may help survive increased errors as the life cycle of a SSD is reached but swapping out drives is still necessary.
  - Although they are likely to have been replaced by larger drives before that time is reached.
RAID 10 vs. RAID 5

- Tradeoff
  - RAID 5 can’t survive any 2 drive failures
  - RAID 5's need to recalculate and redistribute parity data on a per-write basis.
  - RAID 10 is faster to rebuild
RAID 10 vs. RAID 6

- Tradeoff
  - RAID 6 can replace any 2 drive failures
  - RAID 10 is faster to rebuild
RAID 10 vs. RAID 50

• Tradeoff
  - RAID 50 only increases the speed of a RAID 5 based subsystem.
  - SSDs are already faster.
  - RAID 10 is faster to rebuild and doesn’t suffer from any complexity
• Tradeoff
  – RAID 6 can replace any 2 drive failures
  – RAID 10 is faster to rebuild
RAID Strategies for SSDs: Next Steps

• More investigation
  – A lot more data needs to be collected

• SSD specific RAID controllers
  – Do we need new RAID controllers that work with SSDs?

• To be continued...