Rethinking RAID for SSD based HPC

An Affordable and Reliable RAID system.

Yugendra Guvvala
Research Summary

07/28/2011
Agenda

• SSD vs. HDD
• Hybrid RAID
• Design Rationale of HRAID Architecture
• Solutions for HRAID Design Challenges
• Traditional HRAID
• NON-Traditional HRAID
  • Example Of Non-Traditional HRAID
• Optimization of HRAID
• Preliminary Results
  • Reliability
  • Performance
  • Cost Effectiveness
• Conclusion
SSD Vs. HDD

NAND-flash based solid-state drives (SSDs) provide superior performance over traditional hard disk drives (HDDs)

- SSDs provide a promising solution for I/O intensive HPC applications
- However, the cost and the capacity have often been cited as the major barrier of SSD deployment
- We evaluate SSDs and HDDs and design a Hybrid-RAID solution
SSD Vs. HDD

**Merits of SSDs:**
- Low-Latency
- High-Band width
- More reliability

**Barriers of SSDs:**
- High Cost
- Low Capacity
- Limited write cycles

**Merits of HDDs:**
- Low Price
- Huge Capacities
- Superior write cycles

**Barriers of HDDs:**
- Less Reliable
- Huge Latency

Considering listed merits and demerits pure SSD or HDD based RAID solutions are not optimal.
Hybrid-RAID (HRAID): Low-cost and High-latency HDDs as RAID drives for Low-latency and high-cost SSDs

Key Features:

- HDDs are used as RAID drives instead of SSDs.
- Every HDD is paired with a small flash Package.
- Usage of one large SSD instead of small flash package.

Design challenges:

- High Latency of HDDs.
- Huge random write requests to HDDs.
- Small size of flash package.
- SSD/Flash package becoming a bottle neck.
Design Rationale of HRAID Architecture

SSD

Flash Package / SSD

HDD
Solutions for HRAID Design Challenges

• Flash packages with about 1% the size of SSDs are coupled with the HDDs to bridge the huge gap between latencies of SSDs and HDDs.

• Addition of Flash package to SSDs would be able to handle small random write requests to the HDDs to update parity bits.

• For Large contiguous write requests HDDs perform on par or better compared to SSDs.

• Flash package can be replaced with SSD and multiple HDDs can share one SSD.

• Dynamically tuning number of HDDs to be coupled with one SSD/Flash package would improve the performance.
Block Diagram

SSD-2

SSD-1

SSD-3

SSD-4

Flash Package / SSD to buffer data

HDD

HDD

HDD
This designs are same as traditional RAID but by replacing SSDs with HDDs and tweaking them a little bit to improve reliability and performance.

- **HRAID-1**: Implemented by mirroring SSD data onto a HDD.
HRAID-4: Employs one HDD per array of SSDs to hold parity bits
Traditional HRAID

**HRAID-5**: Uses distributed parity, so HDDs belonging to an array hold parities of SSDs belonging to different arrays.
Traditional HRAID

**HRAID-6:** Uses distributed parity, employs 2 HDDS per array, so HDDs belonging to an array hold parities of SSDs belonging to different arrays.
Nitin Agrawal Design tradeoffs for SSD performance.
This design follows the idea of traditional RAID-6 but increases the reliability by as many factors as the user specifies.

- \( n \) SSDs and one HDD per array.

- \( m \) arrays per rack and \( k \) racks. Each rack has a SSD which is shared by all HDDs across all arrays in the rack for bridging latency.

- Size of HDD is \( p \) times the size of SSD.

- HDD partitioned into \( p \) equal partitions.

- \( p \) is number of dimensions of parity. Higher the value of \( p \) more reliable the system is.

- Higher the value of \( p \) more latency is involved to calculate the parity values.

- We can employ a GPU to speedup the parity calculations.
Non-Traditional HRAID Example

Design of a 3 dimensional \((p = 3)\) non-traditional HRAID.

- There are 3 racks \((k = 3)\), 3 arrays per rack \((m = 3)\), and 4 SSDs per array \((n = 4)\).
- There is one HDD per array and one bridging SSD per rack or one flash package per partition of HDD.
- The HDD is 3 times the size of SSD and is partitioned into 3 equal partitions say \(P_1\), \(P_2\), and \(P_3\).
- \(P_1\) acts as a parity bit holder for the SSDs in the array where HDD belongs to like HRAID-4.
- \(P_2\) holds parity bits of flash packages of SSDs in the arrays other than the array HDD belongs to but within the rack.
Non-Traditional HRAID Example

• $P_3$ holds parity bits of the dies of the SSDs belonging to different arrays in different racks.

• If there is one disk failure $P_1$ is used to reconstruct the disk as the parity is nearest.

• If there is an array failure $P_2$'s of different HDDs are used to reconstruct the array.

• If there is a rack failure then $P_3$'s of appropriate HDDs are used to reconstruct the data drives in the rack.

• Further these racks can be grouped and multiple UPSs can be used to power those groups and new dimension can be created to tolerate UPS failure.

• Thus we can extend the number of dimensions to create a better reliable system to tolerate multiple component failures.
Non-Traditional HRAID Example
Optimization: *Usage-Aware Optimization*

Considering the special features of SSDs and HRAID architecture, more optimizations can be performed:

- Track the request to SSDs and dynamically change redundancies of data copies when life time grows.

- The intelligence of the usage-aware optimization is to take the likelihood of disk failures into consideration for different redundancy in a redundant disk array system like HRAID, which has not been explored before.

- The reason is that the lifetime of SSDs is generally limited to certain number of write cycles and the possibility of disk failures increases as disk usage grows.
Primary Results: Reliability

- Traditional HRAID has comparable reliability as traditional RAID as their design features are same.
- Non-Traditional HRAID is more reliable as every bit of data on the disks has parities on $p$ different places.
- Partitions of HRAID hold parities of bits belonging to SSDs across racks, arrays, groups of racks, and powered by different UPSs thus providing more reliability.
- The optimization strategy like usage-aware optimization further improves the reliability.
Primary Results: \textit{Performance}

- Addition of flash package to HDD bridges the performance gap between HDDs and SSDs well, and make the matching between SSD and HDD possible.

- Performance of the Hybrid System is comparable and close to pure SSD based RAID system.
Primary Results: **Performance**

- **Experimental Results:**
  - Designs other than HRAID-1 the write requests on to HDDs is small random requests of size of few bytes as all are only parity bit updates.
  - Random write speed of SSDs is significantly higher
  - Random write speed of:
    - *Intel X25-E 40 GB MLC SSD is 35.4 MB/s*
    - *Western Digital VelociRaptor 600 GB HDD is 1.9 MB/s*
  - Using flash disks as bridging gap offsets this huge difference of latencies
  - For HRAID-1 the data request sizes are bigger than bridging flash packages.
  - The data request patterns are large contiguous, sequential writes.
  - Sequential write speeds of:
    - *Intel X25-E 40 GB MLC SSD is 41.9 MB/s*
    - *Western Digital VelociRaptor 600 GB HDD is 143.9 MB/s*
  - Thus HRAID performs well and is comparable to SSD based arrays.
Primary Results: Cost Effectiveness

- Considering cost of unit SSD and HDD in market.
  - 120 GB OCZ Vertex 3 VTX3-25SAT3 is $2.00/GB
  - 2 TB Western Digital WD20EARS Cavlar Green HDD is $0.037/GB.
- Cost savings for HRAID-1 is 1.97 times compared to pure SSD based RAID.
- Cost savings for HRAID-4 & HRAID-5 is 1.2 times compared to pure SSD based RAID.
- Cost savings for HRAID-6 is 1.4 times compared to pure SSD based RAID.
- Cost savings for Non-Traditional HRAID-6 is 1.5 times compared to pure SSD based RAID.
Primary Results: Cost Effectiveness
Future Work

- Employ GPUs for calculating parities in Non-Traditional HRAID.
- To reduce latencies of many levels of Parity bit.
Conclusion

• Taking advantage of Merits of both HDDs and SSDs, we propose HRAID.

• HRAID is capable of providing same reliable features and comparable performance as SSD-Based RAID.

• HRAID has an edge over SSD-RAID in cost effectiveness.

• HRAID provides a low-latency, low-cost and high-throughput solution.

• For Non-Traditional HRAID, use GPUs for faster calculation of parity values. (Research in progress)


Questions