Chip-Level RAID with Flexible Stripe Size and Parity Placement for Enhanced SSD Reliability

Jaeho Kim (Postdoctoral researcher @ Hongik Univ.)
Donghee Lee (Professor @ Univ. of Seoul)
Sam H. Noh (Professor @ Hongik Univ.)
Introduction & Motivation

- Flash SSD products with RAID-5 like data protection

Fusion-io’s ioMemory
(Adaptive Flashback Tech.)

Baidu’s Software-Defined Flash

Micron’s P420m
(Redundant Array Independent NAND Tech.)

Shannon Systems’s Direct-IO
Introduction & Motivation

- Applying RAID-5 into SSD internal
  - Is RAID-5 suitable for flash chips?
  - Is RAID-5 really beneficial for SSD?

Quantitative analysis for reliability and lifetime?
Applying RAID-5 into SSD

- Apply **RAID-5** configuration to chips comprising the **SSD** device
Pros & Cons of RAID-5 in SSD

• Assumptions of quantitative analysis
  – Conventional SSD (denoted “ECC”): MLC 64GB with BCH code (4bit/512bytes) for ECC
  – RAID-5 SSD (denoted “RAID-5”): MLC 64GB applying RAID-5 without parity cache
  – Workload: Financial

• Pros of RAID-5: Improving reliability of SSD
• Cons of RAID-5: Decreasing lifetime of SSD

**Pros of RAID-5:**
- Improved reliability

**Cons of RAID-5:**
- Decreased lifetime

---

**Graphs:**
- **Line graph:**
  - UPER (Error Rate) vs. Total written bytes (1TB to 200TB).
  - Two lines, one for ECC and one for RAID-5, showing the error rate for each.
- **Line graph:**
  - P/E cycle limit (MLC) vs. Total written bytes (1TB to 200TB).
  - Two lines, one for ECC and one for RAID-5, showing the cycle limit for each.
How can we improve the reliability while prolong the lifespan of the SSD?
Outlines

• Introduction & Motivation

• **Challenges of RAID-5**

• Our Solution: eSAP-RAID

• Evaluations

• Analytic Models of RAID Schemes

• Conclusion
Applying RAID-5 into SSD: Challenges #1,2

- Out-of-place update property of flash memory
  - Parity writes increase write amplification (WA) in SSD
- LBN (Logical Block Number) based striping feature of RAID-5
  - Parity update overhead (Read-modify-write) for small write requests
  - Data are written to specific chip depending on the LBN of data
Applying RAID-5 into SSD: Challenge #3

- Open a window of vulnerability (for totally new data)
  - Small writes must wait until stripe fills up to write parity

raid 5 may loss incomplete stripe due to nonexistent parity

Totally new data
(Not updating existing data)

Write Req.

FlasMemory Summit
2015 Santa Clara, CA
Summary of the Challenges

• **Out-of-place update** property of flash memory
  – Parity update may decrease lifespan of flash memory

• **LBN based striping** feature of RAID-5
  – Must read old data or old parity for parity calculation
  – Data is written to specific chip depending on the LBN of data

• **Open a window of vulnerability**
  – Small writes must wait until stripe fills up to calculate parity
Outlines

• Introduction & Motivation
• Challenges of RAID-5

• **Our Solution: eSAP-RAID**
• Evaluations
• Analytic Models of RAID Schemes
• Conclusion
Our Solution: eSAP-RAID

RAID-5

Frequent parity update decrease lifespan of flash memory

* Must read data for new parity
* Skewed writes to particular chip lead to reduced lifespan

Open a window of vulnerability

Elastic Striping & Anywhere Parity (eSAP)

Dynamically construct a stripe based on arrival order of write requests regardless of LBN

Stripe size can be flexible with partial stripe parity

Solve
Write Cost: RAID-5 vs. eSAP

- Assumptions
  - Stripe 0 and 1 are already constructed in the SSD
    - Stripe 0: D0, D1, D2, and P0
    - Stripe 1: D3, D4, D5, and P1
  - Updated data separately arrive in D0’, D0’”, and D1’ order
**Write Cost: RAID-5 vs. eSAP**

- I/O cost: eSAP reduces the number of read & write operations.

<table>
<thead>
<tr>
<th></th>
<th>RAID-5</th>
<th>eSAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Write updated data</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Read for parity calc.</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Write parity</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Reduced cost of parity overhead
How can we protect new data ‘D8’ before parity write?

- RAID-5 may loss incomplete stripe due to without parity
- eSAP can protect the new data with partial stripe parity (flexible stripe size)
Outlines

• Introduction & Motivation
• Challenges of RAID-5
• Flash-aware New RAID Architecture

• Evaluations
• Analytic Models of RAID Schemes
• Conclusion
Evaluation Setup

- SSD extension with the DiskSim, which is a simulator for SSD
  - 8 flash memory chips, a stripe consists of 16 pages

- Evaluate three configurations
  - ECC: No parity (similar to RAID-0)
  - RAID-5: Conventional RAID-5 scheme
  - eSAP: Elastic Striping and Anywhere Parity-RAID (Proposed scheme)

- Characteristics of I/O workloads

<table>
<thead>
<tr>
<th>Workload</th>
<th>Total Data Req.(GB)</th>
<th>Write Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential</td>
<td>21.8</td>
<td>1.0</td>
</tr>
<tr>
<td>Random</td>
<td>30.2</td>
<td>1.0</td>
</tr>
<tr>
<td><strong>Financial</strong></td>
<td><strong>35.7</strong></td>
<td><strong>0.81</strong></td>
</tr>
<tr>
<td>Exchange</td>
<td>101.2</td>
<td>0.46</td>
</tr>
<tr>
<td>MSN</td>
<td>29.7</td>
<td>0.96</td>
</tr>
</tbody>
</table>
• eSAP reduces the response time over RAID-5
• eSAP prolongs the life span of SSD over RAID-5
• RAID-5 performs worst, especially for the financial workloads
  – Small writes incur heavy parity overhead
Analysis of Parity Overhead

- Parity overhead of RAID-5
  - Reads for parity calculations (PR)
  - Parity writes (PW)

- Parity overhead of eSAP
  - Parity writes (PW)
  - Partial stripe parity writes (PPW) for small write request
Outlines

• Introduction & Motivation
• Challenges of RAID-5
• Flash-aware New RAID Architecture
• Evaluations

• Analytic Models of RAID Schemes
• Conclusion
Analytic Models of RAID Schemes

• Goals of analytic models
  – Find expected lifespan (P/E cycles) of SSD with various I/O workloads
  – Project long-term reliability according to the lifespan of SSD

• Two factors affecting lifespan of SSD
  – Write Amplification Factor (WAF)
    • Garbage collection cost
  – Parity Write Overhead (PWO)
    • Term derived from this work (Mathematical model)
Parity write overhead are determined by

- 1) Size of RAID stripe
- 2) Size of write request
- 3) Starting position of write request within a stripe

From the PWO,

- We can estimate the number of page writes and erase operations
Analytic Models of RAID-5 and eSAP

• Expected lifespan of SSD with RAID-5

Analytic model of RAID-5

\[ N_{PW_{R5}}^{req} = N_P^{req} \left( 1 + N_{P_{R5}}^{PO} \right) \left( 1 + WAF(u) \right) \]

- P/E cycles & UPER
- WAF
- PWO of RAID-5

Parameters

Characteristics of I/O workload & SSD

• Expected lifespan of SSD with eSAP

Analytic model of eSAP

\[ N_{PW_{eSAP}}^{req} = N_P^{req} \cdot \left( 1 + N_{P_{eSAP}}^{PO} \right) + N_P^{req} \cdot \left( 1 + N_{P_{eSAP}}^{PO-WA} \right) \cdot WAF(u) \]

- P/E cycles & UPER
- WAF
- PWO of eSAP

Parameters

Characteristics of I/O workload & SSD

Please refer to our paper for details!!

Flash Memory Summit
2015 Santa Clara, CA
Projecting Long-term Reliability

- Procedure of projection
  1) Extract characteristics of I/O workloads and parameters of SSD
  2) Put extracted values into analytic models to expect lifespan of SSD
  3) Calculate reliability equations with expected lifespan of SSD
  4) Find Uncorrectable Page Error Rate (UPER) from the calculation
Analysis of Long-term Reliability

- Uncorrectable Page Error Rate (UPER) and life span of SSD
  - Financial workload
  - For 64GB MLC flash-SSD

**eSAP can improve reliability while limiting SSD’s wear**

**eSAP requires far less P/E cycles compared to RAID-5**

**eSAP has the lowest UPER among three schemes**

**eSAP can improve reliability while limiting SSD’s wear**

- Required error rate by HDD vendor
- P/E cycle limit (MLC)

Total written bytes

**Flash Memory Summit**
2015 Santa Clara, CA
Conclusion

• Reliability of flash based storage is getting more crucial

• A solution to improve reliability is to apply RAID configuration into SSD
  – Conventional RAID-5 is not suitable
  – eSAP: A novel flash-aware RAID scheme is proposed

• Derive the analytical model of RAID schemes in SSD
  – Derive performance and lifespan models of RAID schemes in SSDs
  – Project long-term reliability of SSDs
Thank you! & Questions?

Please refer to the paper for details,


Jaeho Kim (kjhnet@gmail.com)
Backup Slides
Accuracy of Model

- Accuracy ratio of the model compared to the experimentally obtained
  - Most of the cases, the difference between the model and the experimental results are within 10%

<table>
<thead>
<tr>
<th>Workloads</th>
<th>RAID-5</th>
<th>eSAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td>Random</td>
<td>0.99</td>
<td>0.91</td>
</tr>
<tr>
<td>Financial</td>
<td>0.99</td>
<td>0.93</td>
</tr>
<tr>
<td>Exchange</td>
<td>0.93</td>
<td>0.99</td>
</tr>
<tr>
<td>MSN</td>
<td>0.98</td>
<td>0.95</td>
</tr>
</tbody>
</table>
What is Determination factor for WAF & PWO?

- WAF and PWO are determined by characteristics of the I/O workloads

<table>
<thead>
<tr>
<th>Workload</th>
<th>Scheme</th>
<th># of Write req.</th>
<th>Avg. size of Write</th>
<th>Avg. $\mu$ of victim blocks for GC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential</td>
<td>RAID-5</td>
<td>368K</td>
<td>62K</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>eSAP</td>
<td>184K</td>
<td>124K</td>
<td>0</td>
</tr>
<tr>
<td>Financial</td>
<td>RAID-5</td>
<td>3617K</td>
<td>9K</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>eSAP</td>
<td>416K</td>
<td>78K</td>
<td>0.64</td>
</tr>
</tbody>
</table>

Flash Memory Summit 2015 Santa Clara, CA
ECC Bit Correction Requirements Trends

- Bit requirements for BCH

From: 1) ECC Options for Improving NAND Flash Memory Reliability – Micron, 2012
2) Signal processing and the evolution of NAND flash memory – Anobit, 2010
BER of MLC vs. TLC

BER of MLC Flash

BER of TLC Flash

Flash Memory Summit
2015 Santa Clara, CA