Caching less for better performance: Balancing cache size and update cost of flash memory cache in hybrid storage systems
(Presented at USENIX FAST’12)

Yongseok Oh (ysoh@uos.ac.kr)
PhD Candidate
University of Seoul
Hybrid Storage Systems

• Combine emerging SSDs and conventional HDDs
  ▪ Provide SSD-like performance at HDD-like price
• Our focus: issues managing SSD cache
  ▪ SSD cache solution is key to performance
How do we achieve optimal performance?

Optimal Partitioning Flash Cache Layer (OP-FCL) [FAST’12]
Cheetah Storage (Fictitious Company)

New Project: Hybrid Storage System

How to design an SSD cache solution?

SSD characteristics:
- Asymmetric Read/Write Cost
- GC Cost

Workload patterns:
- Read Request Ratio
- Hit Ratio

Okay, let me explain the critical issues in SSDs

Bob (SSD Designer)

Alice (Cache Designer)
Limitations
1. Asymmetric read-write cost
2. Erase-before-write (no overwrite)

[1] Samsung Large-block SLC NAND (K9WAG08U1A)
[2] Seagate 7200rpm Hard Disk (ST380011A)
FTL as Main Component of SSD

- FTL (Flash Translation Layer): emulating HDD interface
  - Logical address to physical address
FTL: Initial State

- **Mapping Table**
- **NAND Flash Memory**
- **Data Space**
- **Logical address to Physical address**
- **Over-provisioned Space (OPS) (Reserved for GC)**

- **Read/Write Sectors**
FTL: Remapping Operation

Write requests: 1, 3

Read/Write Sectors

Mapping Table

FTL

NAND Flash Memory

Log-structured Write

Invalidated

Remapped
FTL: Garbage Collection

1. Selecting a victim
   - Least Utilized Victim Block

2. Copying pages
   - GC requires copying pages and erasing block

3. Erasing a block
   - Erase

4. Finishing GC
   - Free
Implication of OPS on Performance

• Initial OPS size influences GC performance
  ▪ E.g., Write Amplification Factor (WAF) [SYSTOR09]

SSD Logical Layout

Data

OPS

More OPS

Victim block

1 copy

Low GC cost

SSD Logical Layout

Data

OPS

Less OPS

Victim block

3 copy

High GC cost
Amplified Write Cost of SSD

- Average write cost varies due to GC
  - Depending on OPS size [LFS92, HyLog04, Janus-FTL10]
- System administrator increases OPS size
  - Write performance is optimized

![Graph showing amplified write cost of SSD]

Utilization (%) SSD

- Calculated by our cost model
- More OPS

SSD

Data

OPS

Utilization = \frac{Data \ Space \ Size}{Total \ Space \ Size}

Key to performance

IBM Watson Research Center
OPS is Key Factor to Performance of SSD

Table 2. HP IO Accelerator overprovisioning

<table>
<thead>
<tr>
<th>Overprovisioning Setting</th>
<th>Percentage of Overprovisioning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>28%</td>
</tr>
<tr>
<td>Capacity</td>
<td>10%</td>
</tr>
<tr>
<td>Performance</td>
<td>50%</td>
</tr>
</tbody>
</table>

Source: https://www.fusionio.com/load/-media-/2kmwsc/docsLibrary/white_paper_Understanding_HP_IO_Accelerators.pdf
New Project: Hybrid Storage System
How to design an SSD cache solution?

SSD Characteristics
1. Asymmetric Read/Write cost
   - Distinguishing read and write is necessary
2. GC cost
   - Increasing OPS is required to reduce GC cost

Let me tell you why we have to know…
### Workload Characterization of Server Traces

- Read request ratio is different
  - Separation is required

<table>
<thead>
<tr>
<th>Workload</th>
<th>Avg. Req. Size (KB)</th>
<th>Request Amount (GB)</th>
<th>Read Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Read</td>
<td>Write</td>
<td>Read</td>
</tr>
<tr>
<td>Financial [UMASS]</td>
<td>5.7</td>
<td>7.2</td>
<td>6.9</td>
</tr>
<tr>
<td>Home [FIU]</td>
<td>22.2</td>
<td>3.9</td>
<td>15.3</td>
</tr>
<tr>
<td>Search Engine [UMASS]</td>
<td>15.1</td>
<td>8.6</td>
<td>15.6</td>
</tr>
<tr>
<td>Exchange [MSR]</td>
<td>9.89</td>
<td>12.4</td>
<td>114.36</td>
</tr>
<tr>
<td>MSN [MSR]</td>
<td>11.48</td>
<td>11.12</td>
<td>107.23</td>
</tr>
</tbody>
</table>

* [UMASS] UMASS trace repository, http://traces.cs.umass.edu
Hit ratio is key metric.
Many algorithms have been studied:
- LRU, LRFU, LRU-K, 2Q, ARC
Hit ratio increases as utilization goes up.

High hit rate using Less OPS

---

**Cache Hit Ratio vs Utilization**

<table>
<thead>
<tr>
<th>Financial</th>
<th>Home</th>
<th>MSN</th>
<th>Exchange</th>
<th>Search Engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>80</td>
<td>60</td>
<td>40</td>
<td>20</td>
</tr>
</tbody>
</table>

LRU Algorithm is used
New Project: Hybrid Storage System
How to design an SSD cache solution?

SSD Characteristics
1. Asymmetric Read/Write cost
   - Distinguishing read and write is necessary
2. GC cost
   - Increasing OPS is required to reduce GC cost

Workload patterns
1. Read request ratios
   - Differentiating read and write is necessary
2. Hit ratio
   - Decreasing OPS is a must to improve hit ratio of SSD caches
New Project: Hybrid Storage System
How to design an SSD cache solution?

<table>
<thead>
<tr>
<th>1. Different Read/Write cost</th>
<th>1. Read request ratios</th>
<th>Solved</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. GC cost decreases</td>
<td>2. Hit ratio increase</td>
<td>Unsolved</td>
</tr>
</tbody>
</table>

Bob (SSD Designer)

Read and write are separated

Alice (Cache Designer)
Still fighting!!!
Tradeoff: Caching Benefit vs GC Efficiency

Bob (SSD Designer)

I need more OPS to mitigate GC cost!

Hit Ratio

SSD cache

OPS

Caching Space (read + write)

GC Cost

Alice (Cache Designer)

I need more caching space to improve cache hit ratio!

Yongseok Oh (SSD Cache Designer)

I have a solution. Optimal Partitioning Algorithm!
Outline

• Introduction
• Our Approach: Optimal Partitioning Algorithm
• Evaluation
• Conclusion
Our Contributions

• OP-FCL (Optimal Partitioning Flash Cache Layer)
  ▪ Dynamically split caching space into read, write, and OPS
  ▪ Based on workload dependent cost model

• DiskSim based implementation

• Extensive evaluation using real workload traces
Our Idea: Optimal Partitioning Scheme

Storage Model
SSD cost
HDD cost

Workload Pattern
Hit Ratio
I/O Ratio

Optimal Partitioning with Hybrid Cost Model

\( C_{HY}(u, r) \)

Our goal: Find optimal \( u \) and \( r \)

SSD Cache Partitioning
Read
Write
OPS

u
1-u

u: utilization of caching space (e.g., \( 0 \leq u \leq 1.0 \))

r
1-r

r: read ratio in caching space (e.g., \( 0 \leq r \leq 1.0 \))

HDD

IBM Watson Research Center
Overview of Hybrid Cost Model

Hybrid Cost Model: $C_{HY}(u, r)$

Hybrid Read Cost: $C_{HR}(u, r)$

Hybrid Write Cost: $C_{HW}(u, r)$

Expected I/O cost based on $u$ and $r$
Hybrid Storage Model

Expected I/O cost of hybrid storage

Hybrid Cost:
\[ C_{HY}(u, r) = C_{HR}(u, r) \cdot IO_R + C_{HW}(u, r) \cdot IO_W \]

Hybrid Read Cost

\[ C_{HR}(u, r) = H_R(u, r) \cdot C_{PR} + (1 - H_R(u, r)) \cdot (C_{PR} + C_{PW}(u)) \]

Hybrid Write Cost

\[ C_{HW}(u, r) = H_W(u, r) \cdot C_{WH} + (1 - H_W(u, r)) \cdot (C_{PR} + C_{DW} + C_{PW}(u)) \]

Please, refer to the paper (11 equations included!!)

Hybrid Read Cost

Write Rate

Hybrid Write Cost

Write Rate

Please, refer to the paper (11 equations included!!)

Hybrid Read Cost

Write Rate

Hybrid Write Cost

Write Rate
1. Observe Hit Ratio
   (a) Read hit ratio
   (b) Write hit ratio

2. Calculate Cost based on \( u \) and \( r \)
   Hybrid Cost Model: \( C_{HY}(u,r) \)

3. Find \( u \) and \( r \) resulting in Optimal I/O Cost
   Optimal Point at \( op_u = 0.64 \), \( op_r = 0.25 \)

4. Adjust SSD Cache partition
   SSD Cache: e.g., 4GB
   \( op_u = 0.64 \), \( op_r = 0.25 \)
   Read 0.64GB, Write 1.92GB
   OPS 1.44GB

Optimal Partitioning Algorithm with Hybrid Cost Model
Adapt to Workload Pattern

Initial Partition

Resized

Optimal partitioning

Hit ratio curves

1st Period

Workload Tracking

2nd Period

Workload Tracking

Optimal u and r

Nth Period

Workload Tracking

Optimal partitioning

Hit ratio curves

Optimal u and r

Resized

I/O Rate

Time (min)

Write rate

Read rate

IBM Watson Research Center
Optimal Partitioning Flash Cache Layer (OP-FCL)

I/O request arrives

Seq. I/O Detector

If identify, go to HDD

Cache Miss

Page Replacer

Read LRU

Write LRU

Mapping Manager

Translation Table

Logical to Physical

SSD Cache

HDD

If non-seq. I/O, go to SSD Cache

Workload Tracker

(ghost buffer)

Hit Curves

Periodically Execute

Partition Resizer

Workload Dependent Optimal Partitioning

Shrink

Enlarge

Cache Hit

IBM Watson Research Center
Outline

- Introduction
- Hybrid Cost Model
- Evaluation
- Conclusion
Evaluation Setup

• Hybrid Storage Simulator
  ▪ CMU DiskSim 4.0 and MSR SSD extension

• Configurations
  ▪ Config. 1: 4GB SSD cache + 10K RPM HDD
  ▪ Config. 2: 16GB SSD cache + three 10K RPM HDDs

• Workload traces
  ▪ Financial [UMass] with Config. 1
    ◆ Random write dominant, OLTP application running at a financial institution
  ▪ Search Engine [UMass] with Config. 1
    ◆ Random read dominant, Web search engine
  ▪ Exchange [SNIA] with Config. 2
    ◆ Random read/write mixed, Microsoft employee e-mail server
Partitioning Schemes

1. **Fixed Partitioning (FP-FCL)**
   - Fixed size OPS
   - Typical SSD product
   - Single LRU list

2. **Read Write (RW-FCL)**
   - Fixed size OPS
   - Distinguish read and write
   - Two LRU lists

3. **Optimal Partitioning (OP-FCL)**
   - *Dynamically adjusted*
   - *Based on our model*
   - Two LRU lists
Response Time Results

- **OP-FCL shows near-optimal performance**
- **Optimal performance depends on workload characteristics**
Dynamic Adjustment

- OP-FCL dynamically adjusts cache space according to workload
- Financial and Exchange
  - Considerable OPS to lower garbage collection cost
- Search Engine
  - Mostly caching space to maintain read data
Effect on Lifetime of SSD Cache

- **Lifetime of SSD cache is an important issue**
- Optimal point of lifetime differs from that of performance
- Our focus is to improve the performance of SSD cache
- Optimizing lifetime of SSD cache left as future work
Summary of Results

- OP-FCL shows near optimal performance
- Optimal performance depends workload pattern
  - 50% OPS for Financial
  - 5% OPS for Search Engine
  - 30% OPS for Exchange