On the Importance of Evaluating Storage Systems’ $Costs

Zhichao Li, Amanpreet Mukker, and Erez Zadok

http://www.fsl.cs.stonybrook.edu
Hybrid Drive (PCM/SSD/HDD/etc)

Tiering

Caching

6/17/2014

On the Importance of Evaluating Storage Systems' $Costs
Problem Statement

- Performance alone is not enough and dollar cost matters
- Lacking empirical TCO study when systems deploy SSD

Our work?

- We built a cost model to justify gained performance
- We built and evaluated two empirical systems: tiering vs. caching
Cost Model

● Upfront purchase
  ◆ Price per capacity ($/GB)

● Total cost of ownership (TCO)
  ◆ Energy and power costs
    ▪ Lookup local electricity authority
  ◆ Endurance cost
    ▪ $\Delta$Endurance(dev) x Cost(dev)
  ◆ Service and space cost

Scale TCO to predict long-term costs
# Energy and Power Costs

<table>
<thead>
<tr>
<th>Prices ($/)</th>
<th>Power &lt;= 7KW</th>
<th>Power &lt;= 145KW</th>
<th>Power &gt; 145 KW</th>
</tr>
</thead>
<tbody>
<tr>
<td>offpeak</td>
<td>0.0863 0</td>
<td>0.0191 0</td>
<td>0.0218 0</td>
</tr>
<tr>
<td>peak</td>
<td>0.1052 0</td>
<td>0.0340 48.78</td>
<td>0.0446 28.76</td>
</tr>
<tr>
<td>intermediate</td>
<td>0.0863 0</td>
<td>0.0317 5.94</td>
<td>0.0356 8.13</td>
</tr>
</tbody>
</table>

PSE & G

(Per KWh and per KW)
Device Endurance Model

Model:

- Track writes to the SSD *(convert reads)*
  \[ E_{SSD}(t) = 1 - \frac{\#\text{writes}(t)}{\text{Limit}_{SSD}} \]

- Count the HDD start-stop cycles
  \[ E_{HDD}(t) = 1 - \frac{\#\text{start-stop}(t)}{\text{Limit}_{HDD}} \]

- \[ \Delta E_{SSD} = E_{SSD}(t_2) - E_{SSD}(t_1) \]

- \[ \Delta E_{HDD} = E_{HDD}(t_2) - E_{HDD}(t_1) \]

- Endurance Unit: eu
  \[ [0, 1,000,000] = [\text{dead, new}] \]
Architecture
Tiering Architecture

Application I/Os

Virtual device (/dev/mapper/xx)

Device Mapper Framework

GreenDM

Mapping Table

linear

SSD

Migrate

HDD

Spin-down

User

Kernel
Caching Architecture

Application I/Os

Virtual device (/dev/mapper/xx)

Device Mapper Framework

- Caching
- Cache Entry
- linear

SSD

Copy

HDD

Spin-down

User

Kernel
Tiering Data Management

Virtual Extents (VEs)

Logical Extents (LEs)

LBA

VBA

Mapping Table

6/17/2014

On the Importance of Evaluating Storage Systems' $Costs
Caching Data Management

- Asynchronous
- Queue write-backs

Cache Entry Table

<table>
<thead>
<tr>
<th>CID</th>
<th>PID</th>
<th>Flags</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>n</td>
<td>3</td>
<td>...</td>
</tr>
</tbody>
</table>

Cache

0 1 2 3 ... n

Extents

Eviction

Write policy

Pre-fetch

Light-weight mapping

map_bio

Primary

0 1 ... k ... m-1 m

HDD
Evaluation
Experimental Setup

• Lenovo Think-Center
  ◆ Core 2 Quad 2.66GHz CPU and 4GB RAM
  ◆ Intel 300GB SSD ($529) and Seagate 2TB HDD ($200)
• Wattsup Pro ES for power/energy measurement
• Three baselines
  ◆ 1) Mylinear; 2) SSD-only; 3) HDD-only
• Various parameters
  ◆ PT, MCML, and ES, for example
Web-Search Trace Replay

- Block trace from UMass trace repository
- Drive size: 32GB
- Average read size: 16KB (num: 1,055,236)
- Average write size: 8KB (num: 212)
Web-search Cost Results

**Time-factor 1**

- **SSD-only**: large purchase, small TCO
- **HDD-only**: small purchase, small TCO
- **Hybrids**: medium purchase, large TCO
- **GreenDM costs more than Mylinear**
- **Larger PT, smaller TCO**

**Time-factor 100,000**

([min, avg, max]=[0.2, 2.1, 7.7] yrs)
On the Importance of Evaluating Storage Systems' Costs

Web-search Tiering vs. Caching

- Read-intensive workload
- Similar throughputs
- Similar short-term cost
- Tiering incurs larger TCO

Throughput

Time Factor 1

Time Factor 100,000

([min, avg, max]=[0.2, 2.1, 7.7] yrs)

Aggregated primary SSD I/Os
FIU Online Trace Replay

- Block trace from FIU trace repository
- Drive size: 8GB
- Average read size: 8KB (num: 655,526)
- Average write size: 4KB (num: 4,211,786)
FIU Online Cost Results

Time-factor 1

SSD-only: large purchase, large TCO
HDD-only: small purchase, small TCO
Hybrids: medium purchase, large TCO
GreenDM costs more than Mylinear
Smaller ES, smaller TCO

Time-factor 100,000
([min, avg, max] = [0.7, 3.3, 9.8] yrs)

On the Importance of Evaluating Storage Systems' $Costs
FIU Online Tiering vs. Caching

Evaluate storage systems across performance, long-term cost, and workloads.

- Caching write back
- Similar short-term cost
- Caching incurs larger TCO
- Write-back I/Os

Time Factor 100,000

([min, avg, max] = [0.7, 3.3, 9.8] yrs)
Observed Trends Summary

- **Parameters matter for TCO**
  - Read-intensive Web-search: larger PT
  - Write-intensive Online: smaller ES

- **HDD-only drive**
  - Least initial and long-term costs; but lowest performance

- **SSD-only drive**
  - Highest initial cost; various long-term costs; but highest performance

- **Hybrids (two modes)**
  - Medium initial costs; various long-term costs; and medium performance
  - Hybrids incurs more costs than Mylinear; but higher performance

- **Versatility effects**
  - Different system configurations lead to various long-term costs
Limitations

- Costs did not include several factors:
  - Computer hardware, air conditioning
  - Labor power, financing
- Simplified several conditions:
  - Data center hardware setup more complex
  - Data center workloads more complex
- Coarse-grained endurance model
- Journaling support
Related Work

- Simulation
  - [HybridStore 2011]
- No SSD endurance
  - [EDT 2011, LSM 1996]
- No TCO with SSD
  - [Floyer 2009]
Conclusion

- Cost model to justify performance gained
  - Upfront purchase cost
  - Total cost of ownership

Hope for more long-term cost models of not only hybrids, but also all future storage systems
Q&A

On the Importance of Evaluating Storage Systems’ $Costs

Zhichao Li, Amanpreet Mukker, and Erez Zadok

http://www.fsl.cs.stonybrook.edu
## SSD Replacement Statistics

<table>
<thead>
<tr>
<th></th>
<th>Web-search</th>
<th></th>
<th></th>
<th>Online</th>
<th></th>
<th></th>
<th>File-server</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSD-only</td>
<td>HDD-only</td>
<td>My-linear</td>
<td>GreenDM Cache</td>
<td>SSD-only</td>
<td>HDD-only</td>
<td>My-linear</td>
<td>GreenDM Cache</td>
<td>SSD-only</td>
</tr>
<tr>
<td>Time (yrs)</td>
<td>0.22</td>
<td>7.73</td>
<td>6.18</td>
<td>1.72</td>
<td>1.64</td>
<td>0.68</td>
<td>9.8</td>
<td>2.29</td>
<td>1.2</td>
</tr>
<tr>
<td># SSD</td>
<td>4.3</td>
<td>0</td>
<td>0.96</td>
<td>46.76</td>
<td>41.13</td>
<td>47.7</td>
<td>0</td>
<td>40.6</td>
<td>47.43</td>
</tr>
<tr>
<td>Days per SSD lasts</td>
<td>18.67</td>
<td>∞</td>
<td>2349.7</td>
<td>13.43</td>
<td>14.55</td>
<td>5.2</td>
<td>∞</td>
<td>20.59</td>
<td>9.23</td>
</tr>
</tbody>
</table>

Statistics with time factor 100,000

More failures as storage systems become complex

[Schroeder CFDR’07]
What if Unit Prices Change?

- Online with time factor 100,000:
  - HDD/SSD prices are reduced by half
  - Different absolute values, but similar trends
Comments on Flash-Array?

- Products: Dell Compellent, NetApp EF550, Pure Storage FA-400, etc
- Super fast performance, but with challenges:
  - Mis-matching read and write performance
  - Co-related failures
  - Parity Drive ages faster
  - The stripe size matters
  - The aging factor

[Jeremic SYSTOR’11]
Others?

- Less aggressive caching write-back?
  - Absolute value may change
  - Negative effects of write-back won’t change

- Counting metadata flush?
  - Fairly small to be ignored (<<1%)
  - Not closely related to the trade-off study
  - Presumably store in the SSD reserved space

- SSD/HDD ratio?
  - Example (1/4) and looking into (1/8)
  - Such SSD size is small compared with WL size
File-Server Workload

- Filebench with Gamma distribution
- Drive size: 8GB
- RAM size: 1GB
- Average I/O size: 128KB
File-server Cost Results

Time-factor 1

SSD-only: large purchase, medium TCO

HDD-only: small purchase, small TCO

Hybrids: medium purchase, large TCO

GreenDM costs more than Mylinear

Smaller Gamma, smaller TCO

Time-factor 100,000

([min,avg,max]=[0.2, 1.0, 2.4] yrs)
File-server Results

Throughput

Time Factor 100,000
([min,avg,max]=[0.2,1.0,2.4] yrs)

Time Factor 1

- Reads/writes workload
- Tiering higher throughputs
  - Caching write-back
- Similar short-term cost
- Tiering incurs larger TCO
  - Aggregated primary SSD I/Os
### All about Storage Devices

#### SSD: SLC, MLC

**Pick your combo by yourself!**

<table>
<thead>
<tr>
<th>Device</th>
<th>Performance</th>
<th>Price</th>
<th>Capacity</th>
<th>Endurance</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tape</td>
<td>Low+</td>
<td>Low++</td>
<td>Large+</td>
<td>Large</td>
<td>Low</td>
</tr>
<tr>
<td>HDD</td>
<td>Low</td>
<td>Low</td>
<td>Large</td>
<td>Large</td>
<td>High</td>
</tr>
<tr>
<td>SMR</td>
<td>Low</td>
<td>Low+</td>
<td>Large++</td>
<td>Large</td>
<td>High</td>
</tr>
<tr>
<td>SSD</td>
<td>High</td>
<td>High</td>
<td>Small</td>
<td>Small+</td>
<td>Low</td>
</tr>
<tr>
<td>PCM</td>
<td>High+</td>
<td>High+</td>
<td>Small+</td>
<td>Large</td>
<td>Low</td>
</tr>
<tr>
<td>DRAM</td>
<td>High+</td>
<td>High+</td>
<td>N/A</td>
<td>Large</td>
<td>Low</td>
</tr>
<tr>
<td>SCM</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

---

6/17/2014

On the Importance of Evaluating Storage Systems' $Costs
Why not Separating Reads/Writes?

- Reads/writes are equal to HDD
- Reads/writes are unequal to SSD
- Yes, it is possible to distinguish
- Migration model will be more complex
- Our trade-off study will be complex as well
- Future work to look even deeper
Workloads Mainly Reside in DRAM?

- What previous systems want?
  - Performance and transaction
  - But it costs more
  - As the WL grows, colder data goes to cheaper device

- Data filtered out of DRAM
  - Hybrid systems better trade-off performance and cost
To Live or To Die? That is a Question

- **DRAM?**
  - **Die**
    - SCM (e.g., PCM) blurs the boundary between DRAM and Disk and may replace DRAM (cheaper)
    - SCM is comparably fast. Complex DRAM cache may turn out to hurt the performance, instead
  - **Live**
    - OS needs a clean state sometimes

- **What if die?**
  - Then, all I/O goes directly to lower-level systems
On the Importance of Evaluating Storage Systems' $Costs

- **Read-intensive workload**
  - More data movement
  - Less SSD hit

- **Decreased throughput**
  - Less data movement
  - More SSD hit

- **Similar short-term cost**

- **Increased long-term cost**
  - More SSD accesses
On the Importance of Evaluating Storage Systems' $Costs

- **Write-intensive workload**
  - Less write-back I/Os

- **Increased caching throughput**
  - More Migrations and less SSD hit

- **Decreased tiering throughput**
  - More Migrations and less SSD hit

- **Similar short-term cost**
  - Net effect

- **Similar long-term cost**

---

Online (Capacity Ratio: 1/8)

---

![Graphs showing throughput and cost comparison between SSD-only, HDD-only, Mylinear, cache-4k, cache-16k, cache-64k, green-4k, green-16k, and green-64k configurations.](image-url)
File-server (Capacity Ratio: 1/8)

- Read-intensive workload
- Decreased throughput
  - More data movement
  - Less SSD hit
- Similar short-term cost
- Decreased long-term cost
  - Less SSD accesses