FlashBlox: Achieving Both Performance Isolation and Uniform Lifetime for Virtualized SSDs

Jian Huang†  Anirudh Badam  Laura Caulfield
Suman Nath  Sudipta Sengupta  Bikash Sharma  Moinuddin K. Qureshi†

†Georgia Institute of Technology  Microsoft
Flash Has Changed Over the Last Decade

Performance Improvement

100x lower latency
5,000x higher throughput
Flash Has Changed Over the Last Decade

Performance Improvement

- 100x lower latency
- 5,000x higher throughput

Increased Parallelism

- Dozens of parallel chips
Flash Has Changed Over the Last Decade

Performance Improvement

- 100x lower latency
- 5,000x higher throughput

Increased Parallelism

- Dozens of parallel chips

Became Commodity

- Less than $0.3/GB
Flash Has Changed Over the Last Decade

**Performance Improvement**
- 100x lower latency
- 5,000x higher throughput

**Increased Parallelism**
- Dozens of parallel chips

**Became Commodity**
- Less than $0.3/GB

Significant improvements on Flash
Shared Flash-Based Solid State Disk (SSD) in the Cloud

- bing
- mongoDB
- MapReduce

Windows Azure
Shared Flash-Based Solid State Disk (SSD) in the Cloud

SSDs are virtualized and shared in data centers
Performance Interference in Shared SSD

Flash-based SSD: A Black Box
Performance Interference in Shared SSD

Read/write interferences cause long (3x) tail latency!

Flash-based SSD: A Black Box
Performance Interference in Shared SSD

Flash Translation Layer

Chip

Channel

…….

Write

Read
Performance Interference in Shared SSD

Flash Translation Layer

Channel

Chip

Chip

Channel

Chip

Chip

Channel

Chip

Chip

……

……

……

Write

Read

MongoDB

Hadoop

MapReduce

……
FlashBlox: Hardware Isolation in Cloud Storage
FlashBlox: Hardware Isolation in Cloud Storage

Leveraging parallel chips for hardware isolation
Internal Parallelism Enables Hardware Isolation
Internal Parallelism Enables Hardware Isolation

Channel-Level Parallelism
Internal Parallelism Enables Hardware Isolation

Channel-Level Parallelism

Chip-Level Parallelism
Internal Parallelism Enables Hardware Isolation

Plane-level parallelism is constrained as each chip contains only one address buffer.
Internal Parallelism Enables Hardware Isolation

Channel-Level Parallelism

Chip-Level Parallelism

Plane-Level Parallelism

Different parallelism level provides different isolation guarantee
New Abstractions for Hardware Isolation
New Abstractions for Hardware Isolation

Virtual SSD (Channel Level)

Virtual SSD (Chip Level)

Virtual SSD (Plane Level)

High

Medium

Low

Channel

Chip

plane
plane

Chip

plane
plane

Chip

plane
plane

Channel

Chip

plane
plane

Chip

plane
plane

Chip

plane
plane

Channel

Chip

plane
plane

Chip

plane
plane

Chip

plane
plane
New Abstractions for Hardware Isolation

Virtual SSD (Channel Level)

Virtual SSD (Chip Level)

Virtual SSD (Plane Level)

Software-based
Hardware Isolation Meets the Pay-As-You-Go Model in Cloud
Hardware Isolation Meets the Pay-As-You-Go Model in Cloud

Azure DocumentDB

Azure SQL Database

Amazon DynamoDB

vSSD (Channel)

vSSD (Chip)

vSSD (Software)
Hardware Isolation Meets the Pay-As-You-Go Model in Cloud

**Throughput**

**Single Partition Size**

**Price**

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
</tr>
</thead>
<tbody>
<tr>
<td>RU/s</td>
<td>250</td>
<td>1 K</td>
<td>2.5 K</td>
</tr>
<tr>
<td>GB</td>
<td>10 GB</td>
<td>10 GB</td>
<td>10 GB</td>
</tr>
<tr>
<td>Price</td>
<td>$25 USD</td>
<td>$50 USD</td>
<td>$100 USD</td>
</tr>
</tbody>
</table>

---

- **vSSD (Channel)**
- **vSSD (Chip)**
- **vSSD (Software)**

---

- **Azure DocumentDB**
- **Azure SQL Database**
- **Amazon DynamoDB**

---

**Chip**

**Plane**

---

**Channel**

---

**Throughput**

**Single Partition Size**

**Price**
Hardware Isolation Meets the Pay-As-You-Go Model in Cloud

Hundreds of vSSDs can be supported in a single server
Impact of Hardware Isolation on SSD Lifetime
Impact of Hardware Isolation on SSD Lifetime

The average rate at which flash blocks are erased

<table>
<thead>
<tr>
<th>Average #Blocks Erased/sec</th>
<th>YCSB-A</th>
<th>YCSB-B</th>
<th>YCSB-C</th>
<th>YCSB-D</th>
<th>YCSB-E</th>
<th>YCSB-F</th>
<th>Azure Storage</th>
<th>Bing Search</th>
<th>Bing PageRank</th>
<th>Bing Index</th>
<th>TPCC</th>
<th>TATP</th>
<th>TPCB</th>
<th>TPCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>1.5</td>
<td>1.5</td>
<td>3.5</td>
<td>4.0</td>
<td>4.5</td>
<td>4.5</td>
<td>4.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Impact of Hardware Isolation on SSD Lifetime

The average rate at which flash blocks are erased

Average #Blocks Erased/sec

YCSB-A  YCSB-B  YCSB-C  YCSB-D  YCSB-E  YCSB-F  Azure Storage  Bing Search  Bing PageRank  Bing Index  TPCC  TATP  TPCB  TPCE
Impact of Hardware Isolation on SSD Lifetime

The average rate at which flash blocks are erased

Flash blocks wear out at different rate with different workload
Impact of Hardware Isolation on SSD Lifetime

Flash Translation Layer

Channel

Write Intensive

Chip

Chip

Chip

Chip
FlashBlox Challenges

SSD Lifetime

Performance Isolation
FlashBlox Challenges
FlashBlox Challenges

SSD Lifetime | Performance Isolation
FlashBlox: Swapping Channels for Wear Balance

Adjusting the wear imbalance at a more coarse time granularity can achieve near-ideal SSD lifetime.
FlashBlox: Swapping Channels for Wear Balance

The channel that has incurred the maximum wearout

The channel that has the minimum rate of wearout
FlashBlox: Swapping Channels for Wear Balance

Channel migration takes 15 minutes, once per 19 days
Overall performance drops only for 0.04% of all the time
How Frequently Should We Swap?

Imbalance = MaxWear / AvgWear

Used Erase Cycles

Channel 1
Channel 2
Channel 3
Channel 4
How Frequently Should We Swap?

Imbalance = 4

Channel 1  Channel 2  Channel 3  Channel 4
How Frequently Should We Swap?

Imbalance = 2

Channel 1
M

Channel 2
M

Channel 3

Channel 4

Used Erase Cycles

App
How Frequently Should We Swap?

Imbalance = 4/3

Channel 1
Channel 2
Channel 3
Channel 4
How Frequently Should We Swap?

Imbalance = 1

<table>
<thead>
<tr>
<th>Channel 1</th>
<th>Channel 2</th>
<th>Channel 3</th>
<th>Channel 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
</tbody>
</table>

Used Erase Cycles

App
How Frequently Should We Swap?

**Imbalance** = \[ \frac{\text{MaxWear}}{\text{AvgWear}} \]

Channel 1: M M
Channel 2: M M
Channel 3: M M
Channel 4: M M
How Frequently Should We Swap?

<table>
<thead>
<tr>
<th>Channel 1</th>
<th>Channel 2</th>
<th>Channel 3</th>
<th>Channel 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
<tr>
<td>M</td>
<td>M</td>
<td>M</td>
<td>M</td>
</tr>
</tbody>
</table>

Imbalance = \( \frac{\text{MaxWear}}{\text{AvgWear}} \)

\( \begin{align*}
\text{Channel 1:} & \quad \frac{4}{3} \\
\text{Channel 2:} & \quad \frac{8}{5} \\
\text{Channel 3:} & \quad \frac{4}{3} \\
\text{Channel 4:} & \quad \frac{8}{7}
\end{align*} \)

How many times should we swap within SSD lifetime?
Quantifying the Swapping Frequency

Assume there are $N$ channels,

wear imbalance target: $1+x$

after $K$ rounds of cycling:

\[
\text{Wear Imbalance} = \frac{MK + M}{MK + M/N} = \frac{K + 1}{K + 1/N} \leq (1 + x)
\]
Quantifying the Swapping Frequency

Assume there are $N$ channels,

wear imbalance target: $1 + x$

after $K$ rounds of cycling:

$$\text{Wear Imbalance} = \frac{MK + M}{MK + M/N} = \frac{(K + 1)}{(K + 1/N)} \leq (1 + x)$$

$$K \geq \frac{(N - 1 - x)}{(Nx)}$$
Quantifying the Swapping Frequency

Assume there are $N$ channels, 

wear imbalance target: $1+x$

after $K$ rounds of cycling:

\[
\text{Wear Imbalance} = \frac{MK + M}{MK + M/N} = \frac{K + 1}{K + 1/N} \leq (1 + x)
\]

\[
K \geq \frac{(N - 1 - x)}{(Nx)}
\]

Example

If $N = 16$, $x = 0.1$, then $K = 9$, which means after swap $NK = 148$ times, we can guarantee the wear imbalance is bounded in 1.1
Quantifying the Swapping Frequency

Assume there are $N$ channels,

wear imbalance target: $1+x$

after $K$ rounds of cycling:

Wear Imbalance = $\frac{MK + M}{MK + M/N} = \frac{K + 1}{K + 1/N} \leq (1 + x)$

$K \geq \frac{(N - 1 - x)}{(Nx)}$

Example:

For an SSD with 5 years lifetime, swap once per 12 days can guarantee the channels are well balanced for worst case
Adaptive Wear Leveling in Practice

- Channel 1: Used Erase Cycles: $M$
- Channel 2: Used Erase Cycles: $M/3$
- Channel 3: Used Erase Cycles: $M/2$
- Channel 4: Used Erase Cycles: $0$

Legend:
- App
- Channel
Using erase rate as the trigger condition for swapping
Intra Channel Wear Leveling

Used Erase Cycles

Channel 1

Channel 2

Channel 3

Channel 4

[Diagram showing bar charts for Channel 1, Channel 2, Channel 3, and Channel 4 with varying heights representing used erase cycles.]
Intra Channel Wear Leveling

Chips will be swapped along with the channel migration
Intra Channel Wear Leveling

Chips will be swapped along with the channel migration

Intra-chip wear leveling mechanisms
FlashBlox Architecture

- App
- Resource Manager
- Channel-Level Wear Leveling
- Chip-Level Wear Leveling
- Flash
FlashBlox Architecture

App

Resource Manager

Channel-Level Wear Leveling

Chip-Level Wear Leveling

Flash

App

Virtual SSD

...
FlashBlox Architecture

Isolation, Bandwidth & Capacity Requirement (Virtual SSD to Parallel Chips Mappings)

- App
- Virtual SSD
- Resource Manager
- Channel-Level Wear Leveling
- Chip-Level Wear Leveling
- Flash
FlashBlox Architecture

App

Resource Manager

Channel-Level Wear Leveling

Chip-Level Wear Leveling

Flash

App

Virtual SSD

... Pay-As-You-Go Model in Cloud

App

Virtual SSD
FlashBlox Architecture

- Inter Channel Swapping
- Isolation, Bandwidth & Capacity Requirement (Virtual SSD to Parallel Chips Mappings)
- App
  - Virtual SSD
  - App
  - Virtual SSD
- Resource Manager
- Channel-Level Wear Leveling
- Chip-Level Wear Leveling
- Flash
FlashBlox Architecture

- **App**
- **Virtual SSD**
  - Isolation, Bandwidth & Capacity Requirement (Virtual SSD to Parallel Chips Mappings)
  - Channel
  - Inter Channel Swapping
  - Intra Channel Swapping
  - Other FTL Algorithms
  - Intra Channel Swapping

- **Resource Manager**
- **Channel-Level Wear Leveling**
- **Chip-Level Wear Leveling**
- **Flash**
FlashBlox
Experimental Setup

16 channels
4 chips
4 planes
16 KB page size

14 data center workloads

Yahoo Cloud Service Benchmark
Bing Search / Index / PageRank
Transactional Database
Azure Storage
Tail Latency Reduction with FlashBlox

Yahoo Cloud Service Benchmark (YCSB)

- A: Session store recording recent actions
- B: Photo tagging
- C: User profile cache
- D: User status update
- E: Threaded conversations
- F: User database
Tail Latency Reduction with FlashBlox

Tail latency reduction: 2.6x, average latency reduction: 1.4x
Impact of Channel Migration on Application Performance

Bing Search’s Performance During Channel Migration

Latency (milliseconds)

Time (Seconds)

Without Migration

With Migration
Impact of Channel Migration on Application Performance

Bing Search’s Performance During Channel Migration

![Graph showing the impact of channel migration on Bing Search's performance. The graph compares latency (milliseconds) over time (seconds) between 'without migration' and 'with migration.' It indicates a 34% increase in latency with migration.](image-url)
Impact of Channel Migration on Application Performance

Channel migration takes 15 minutes, once per 19 days
Overall performance drops only for 0.04% of all the time

Bing Search’s Performance During Channel Migration

Latency (milliseconds)

Time (Seconds)

Without Migration

With Migration

34%
FlashBlox
Summary

2.6x reduction on tail latency

Near-ideal SSD lifetime

Swap once per 19 days
Thanks!

Jian Huang†
jian.huang@gatech.edu

Anirudh Badam  Laura Caulfield  Suman Nath
Sudipta Sengupta  Bikash Sharma  Moinuddin K. Qureshi †

Georgia Tech

Q&A