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‡

†‡
Flash Has Changed Over the Last Decade

- 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 $0.2/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 $0.2/GB

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

- bing
- MongoDB
- MapReduce

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

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

Flash-based SSD: A Black Box

Read/write interferences cause long (3x) tail latency!
Performance Interference in Shared SSD
Performance Interference in Shared SSD

Flash Translation Layer

Channel

Chip

Chip

Chip

Chip

Channel

Chip

Chip

Chip

Chip

Write

Read
FlashBlox: Hardware Isolation in Cloud Storage
FlashBlox: Hardware Isolation in Cloud Storage

Leveraging parallel chips for hardware isolation

Flash Translation Layer

Channel

Chip

Chip

Chip

Channel

Chip

Chip

Chip

Channel

Chip

Chip

Chip
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

Channel

Chip

Plane

Channel

Chip

Plane

Channel

Chip

Plane
New Abstractions for Hardware Isolation

Virtual SSD (Channel Level)

Virtual SSD (Chip Level)

Virtual SSD (Plane Level)

Software-based

High

Medium

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

<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>

- **Throughput**
- **Single Partition Size**
- **Price**

---

Azure DocumentDB

vSSD (Channel) ➔ vSSD (Chip) ➔ vSSD (Software)

Chip ➔ Channel ➔ Chip

vSSD (Channel)

Chip ➔ Channel ➔ Chip

vSSD (Software)
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

Flash Translation Layer

Channel

Chip

Chip

Channel

Chip

Chip

Channel

Chip

Chip

……..
Impact of Hardware Isolation on SSD Lifetime

The average rate at which flash blocks are erased

<table>
<thead>
<tr>
<th>#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>Average</td>
<td>1.5</td>
<td>0.5</td>
<td>1</td>
<td>1.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4.5</td>
<td>3.5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Impact of Hardware Isolation on SSD Lifetime

The average rate at which flash blocks are erased

![Bar chart showing the average rate of flash block erasure for various workloads.](chart.png)
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
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
How Frequently Should We Swap?

**Imbalance = 4**

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

Imbalance = 2

Channel 1

Channel 2

Channel 3

Channel 4
How Frequently Should We Swap?

Imbalance = \( \frac{4}{3} \)

<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></td>
</tr>
</tbody>
</table>

Used Erase Cycles

App
How Frequently Should We Swap?

Imbalance = 1

Channel 1: M
Channel 2: M
Channel 3: M
Channel 4: M

App
How Frequently Should We Swap?

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

Channel 1
\[ M \quad M \]

Channel 2
\[ M \quad M \]

Channel 3
\[ M \quad M \]

Channel 4
\[ M \quad M \]

App
How Frequently Should We Swap?

Imbalance = 1

Channel 1

Channel 2

Channel 3

Channel 4

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

after $K$ rounds of cycling:
Quantifying the Swapping Frequency

after $K$ rounds of cycling:

\[
\text{Wear Imbalance} = \frac{MK + M}{MK + M/N} = \frac{K + 1}{K + 1/N} \leq (1 + x)
\]
after $K$ rounds of cycling:

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

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

After $K$ rounds of cycling:

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

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

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
App

Channel 2 
Used Erase Cycles: M/3
App

Channel 3 
Used Erase Cycles: M/2
App

Channel 4 
Used Erase Cycles: 0
App
Adaptive Wear Leveling in Practice

Using erase rate as the trigger condition for swapping
Intra Channel Wear Leveling

- Channel 1
- Channel 2
- Channel 3
- Channel 4

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)
FlashBlox Architecture

- App
- Virtual SSD

Resource Manager

Channel-Level Wear Leveling

Chip-Level Wear Leveling

Flash

Pay-As-You-Go Model in Cloud
FlashBlox Architecture

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

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

- Intra Channel Swapping
- Inter Channel Swapping
- Channel-Level Wear Leveling
- Chip-Level Wear Leveling
- Resource Manager

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

- Intra Channel Swapping
- Other FTL Algorithms
- Intra Channel Swapping

FlashBlox Architecture

- Application
- Virtual SSD

16
FlashBlox
Experimental Setup

14 data center workloads

Yahoo Cloud Service Benchmark
Bing Search / Index / PageRank
Transactional Database
Azure Storage

16 channels
4 chips
4 planes
16 KB page size
Tail Latency Reduction with FlashBlox

App1  App2
Tail Latency Reduction with FlashBlox

App1  App2

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

App1-Software Isolation
App1-FlashBlox
App2-Software Isolation
App2-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

0 5 10 15 20 25 30 35 40 45 50 55
Impact of Channel Migration on Application Performance

Bing Search’s Performance During Channel Migration

Without Migration

With Migration

34%
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

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

2.6x reduction on tail latency

Near-ideal SSD lifetime

Swap once per 19 days
Thanks!

Jian Huang†
jianh@illinois.edu

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

Q&A