Manylogs
Improving CMR/SMR Disk Bandwidth & Latency

Tiratat Patana-anake, Vincentius Martin†, Nora Sandler, Cheng Wu, and Haryadi S. Gunawi

THE UNIVERSITY OF CHICAGO

SURYA UNIVERSITY
Got 100% of the read bandwidth
Got 50% of the read bandwidth

User 1

Big Read

User 2

Big Read

Fair Share
Still Fair!

1/N bandwidth
Our

Oh no!

5% bandwidth?!

User 1
Big Read

User 2
Small Durable Writes
Our

More
Bandwidth
Please!

User 1
Big Read

User 2
Faster
Latency
Please!

Small Durable Writes
Ordered Journaling

Data Journaling

Journal  Big Read  Small Writes  Seek
Ordered Journaling

Data Journaling

Journal

Big Read

Small Writes

Seek
Problems with Current Journaling

Ordered Journaling

Data Journaling

Both cannot handle random writes efficiently!
Problems with Current Journaling

Ordered Journaling

Data Journaling

First Write

Second Write

Both cannot handle random writes efficiently!
Introducing Manylogs

Single Log

Manylogs

10 MB

100 MB
Manylogs

Small writes made durable to the nearest log without seeking
Manylogs

- Reserved log spaces uniformly across the disk
  - 10 MB every 100 MB
- Follow the disk head (last big I/O)
- Redirect Small Writes (e.g. ≤ 256 KB)
  - Nearest log: log closest to last big I/O
- Sequential Writes are left untouched

Diagram:
- 10 MB
- 100 MB
- J J J J J J J J
Increased Read Throughput

Reduced Write Latency
Where are logs on the disk?

10 MB

100 MB

The log space = whole platter
Where are logs on the disk?

10 MB

[Diagram showing log space equal to the whole platter]

100 MB

The log space = whole platter
Same cylinder = No seek!

Log for others in the same cylinder
Ratio of Max Read Bandwidth

User 1

128MB Sequential Reads

User 2

Latency (ms)

4KB Random Writes

At different intensities writes/s

- Ordered vs. Data vs. Adaptive vs. Manylogs

Manylogs @ MSST '16 18
Adaptive Journaling

- Middle ground between ordered journaling and data journaling
- **Single-log** design
- Prabhakaran et al., ATC ‘05

Results

Ratio of Max Bandwidth

Ordered
Adaptive
Manylogs

Random Writes per Second (IOPS)
Results

More Random Writes = Less Bandwidth

57% at 40 IOPS
5% at 320 IOPS
Results

73% at 40 IOPS

9% at 320 IOPS
Results

Ordered

Adaptive

Data

Manylogs

Ratio of Max Bandwidth

0%

20%

40%

60%

80%

100%

40

80

160

320

Random Writes per Second (IOPS)
Manylogs gives the **most bandwidth**

Hey! What about my latency?

50% of max bandwidth at extreme IOPS
Results

Ordered Adaptive Manylogs

Average sync Latency (ms)

600
500
400
300
200
100

40 80 160 320
Random Writes per Second (IOPS)
Results

More Random Writes = Longer Latency
Results

![Graph showing performance metrics for Ordered, Adaptive, and Manylogs]

- **Random Writes per Second (IOPS)**:
  - Ordered: 40, 160, 320
  - Adaptive: 40, 160, 320
  - Manylogs: 40, 160, 320

- **Average sync Latency (ms)**:
  - Ordered: 0, 100, 200, 300, 400, 500
  - Adaptive: 0, 100, 200, 300, 400
  - Manylogs: 0, 100, 200, 300, 400

---

**Data**

Ordered

Adaptive

Manylogs
Results

The graph shows the relationship between Random Writes per Second (IOPS) and Average sync Latency (ms) for different log structures. The data is represented as follows:

- **Ordered**: Solid black line and markers
- **Adaptive**: Dashed blue line and markers
- **Manylogs**: Gray dotted line and markers

The x-axis represents Random Writes per Second (IOPS), ranging from 40 to 320. The y-axis represents Average sync Latency (ms), ranging from 0 to 600.

The graph indicates that Adaptive log structures generally have lower Average sync Latency compared to Ordered and Manylogs structures, especially at higher IOPS values.
Results

WOW! Fast latency at extreme IOPS!

Ordered Adaptive Manylogs
Results

The chart shows the ratio of max bandwidth and average sync latency for different random writes per second (IOPS). The X-axis represents the IOPS, ranging from 40 to 320. The Y-axis represents the ratio of max bandwidth ranging from 0% to 100%. The chart compares four categories: Ordered, Adaptive, Data, and Manylogs. The bars indicate the ratio of max bandwidth, while the lines show the average sync latency in milliseconds (ms).
Results

Higher Bandwidth

Lower Latency
User 1
128MB Sequential Reads

Ratio of Max Read Bandwidth

User 2
Latency (ms)

“fileserver”

- Using Filebench
- Multi-threaded
- 2, 4, 8 instances
Manylogs provides the best outcomes!
Checkpointing

Data Journaling

- Periodically
  - Usually every 5 secs

- Journal can get filled fast because all writes are in the journal!

Manylogs

- “Lazy” or “Off-hours”

- Rarely full because just small writes are redirected

Log Swapping
Log Swapping

Log is almost full

Hot Area!

Cold Area!

Journal
Big Read
Small Writes
Seek
Integrations

- File System (MLFS)
  - Durability-Only Mode (O_DUR)

- SMR Disk (MLSMR)

- RAID
Cassandra Write Path
Cassandra Write Path

Writes

Memtable

Commit Log

SSTable

Memory

Disk
Cassandra Write Path

Random writes are the problem

Writes

Memtable

Commit Log

SSTable

Memtable

Commit Log

SSTable

Temp File

No location needed

Flush Triggered

Requires Fast Durability

Ideally SYNC write

But in practice, background write e.g. every 10 seconds
open(file, O_DUR);

- Need **fast durability** but not location constraints
- Content of files will be put in Manylogs regardless of the write size
- Never checkpoint their content
- Random writes are not a problem anymore!
User 1
HDFS

Ratio of Max Read Bandwidth

User 2
MongoDB

Latency (ms)

- 1 instance
- 2 instances
- 4 instances
Most Bandwidth & Lowest Latency with Manylogs

Ratio of Max Bandwidth

MongoDB Instances (w/Default flush period)

MongoDB Write Latency (ms)
Manylogs & SMR

One non-shingled surface
= log space
Manylogs & SMR

Non-Shingled Tracks

Shingled Band
Manylogs SMR (MLSMR) vs. Single-log SMR (SLSMR)

1.3x Speed-up!
Manylogs & RAID

User 1

Big Read

Disk #1

Disk #2

Max Bandwidth!

User 2

Small Durable Writes

Disk #3

Disk #4

Bandwidth drops up to 50%

Mingzhe Hao, Gokul Soundararajan, Deepak Kenchammana-Hosekote, Andrew A. Chien, and Haryadi S. Gunawi. ”The Tail at Store: A Revelation from Millions of Hours of Disk and SSD Deployments.” FAST’16.
128MB Sequential Reads

Ratio of Max Read Bandwidth

User 1

Latency (ms)

User 2

4KB Random Writes

At different intensities
- 40 writes/s
- 80 writes/s
- 160 writes/s
- 320 writes/s
10x Bandwidth Speed-up
14x Latency Speed-up at 320 IOPS!
More in the paper

- Block-Level Manylogs
- Other workloads
  - Sequential Writes
  - “varmail”
  - More Traces
- Log Size
- Logged Write Size
- Mapping Table
Manylogs

- Reserved log spaces uniformly across the disk
- Redirect small writes to the nearest log
- Can help with **NoSQL, SMR, RAID**, and more!
- Provide up to **5x speed-up** on average
## Manylogs

<table>
<thead>
<tr>
<th></th>
<th>Bandwidth Speed-up</th>
<th>Latency Speed-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>vs. Ordered</td>
<td>3.7x</td>
<td>5.7x</td>
</tr>
<tr>
<td>vs. Adaptive</td>
<td>2.7x</td>
<td>2.0x</td>
</tr>
<tr>
<td>vs. Single-log SMR</td>
<td></td>
<td>1.3x</td>
</tr>
</tbody>
</table>
Thank you!
Questions?

http://ucare.cs.uchicago.edu