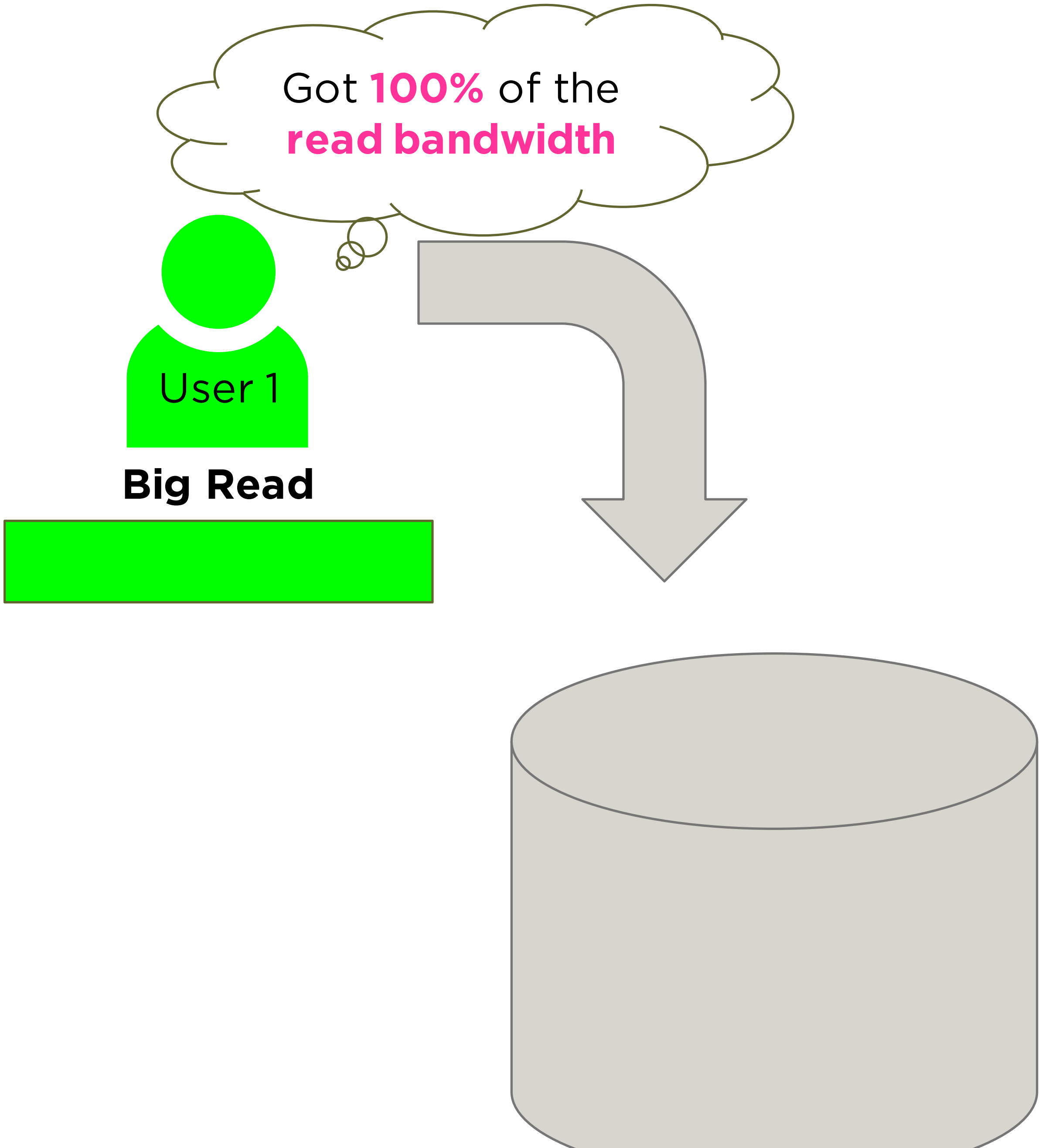
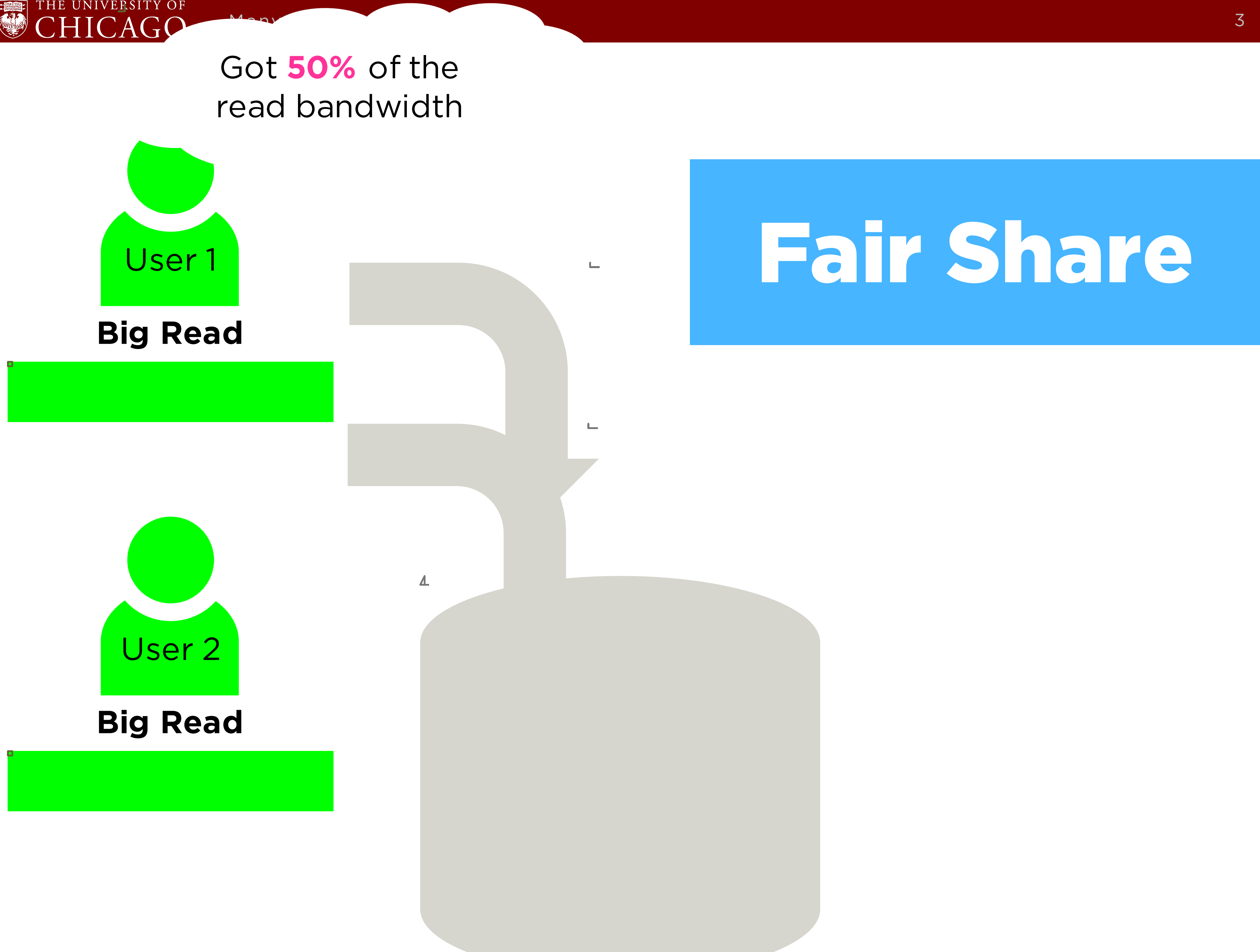
Manylogs
Improving CMR/SMR Disk Bandwidth & Latency

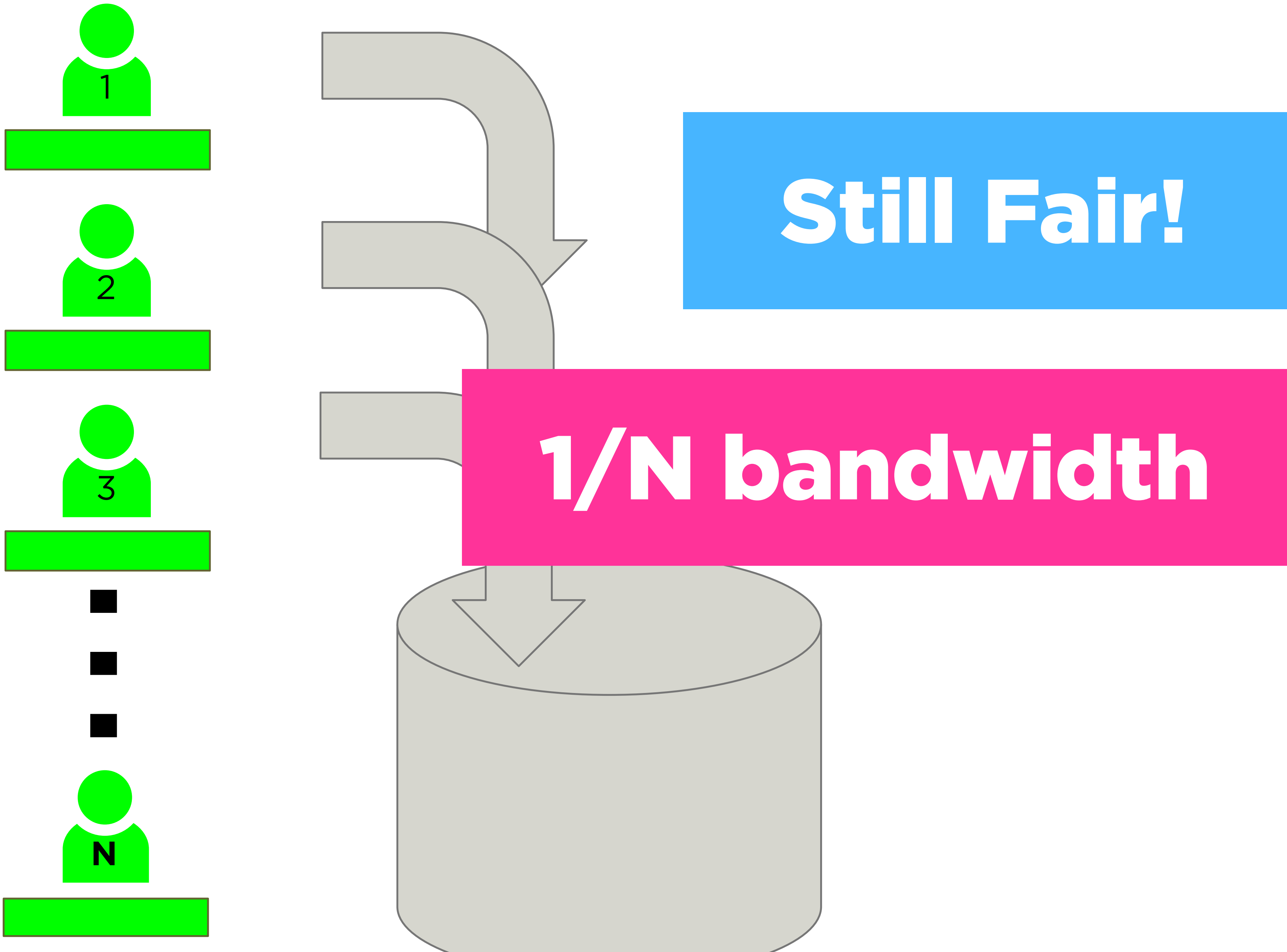
Tiratat Patana-anake, Vincentius Martin†,
Nora Sandler, Cheng Wu, and Haryadi S. Gunawi

THE UNIVERSITY OF CHICAGO

† SURYA UNIVERSITY


Got 100% of the read bandwidth
User 1
Big Read

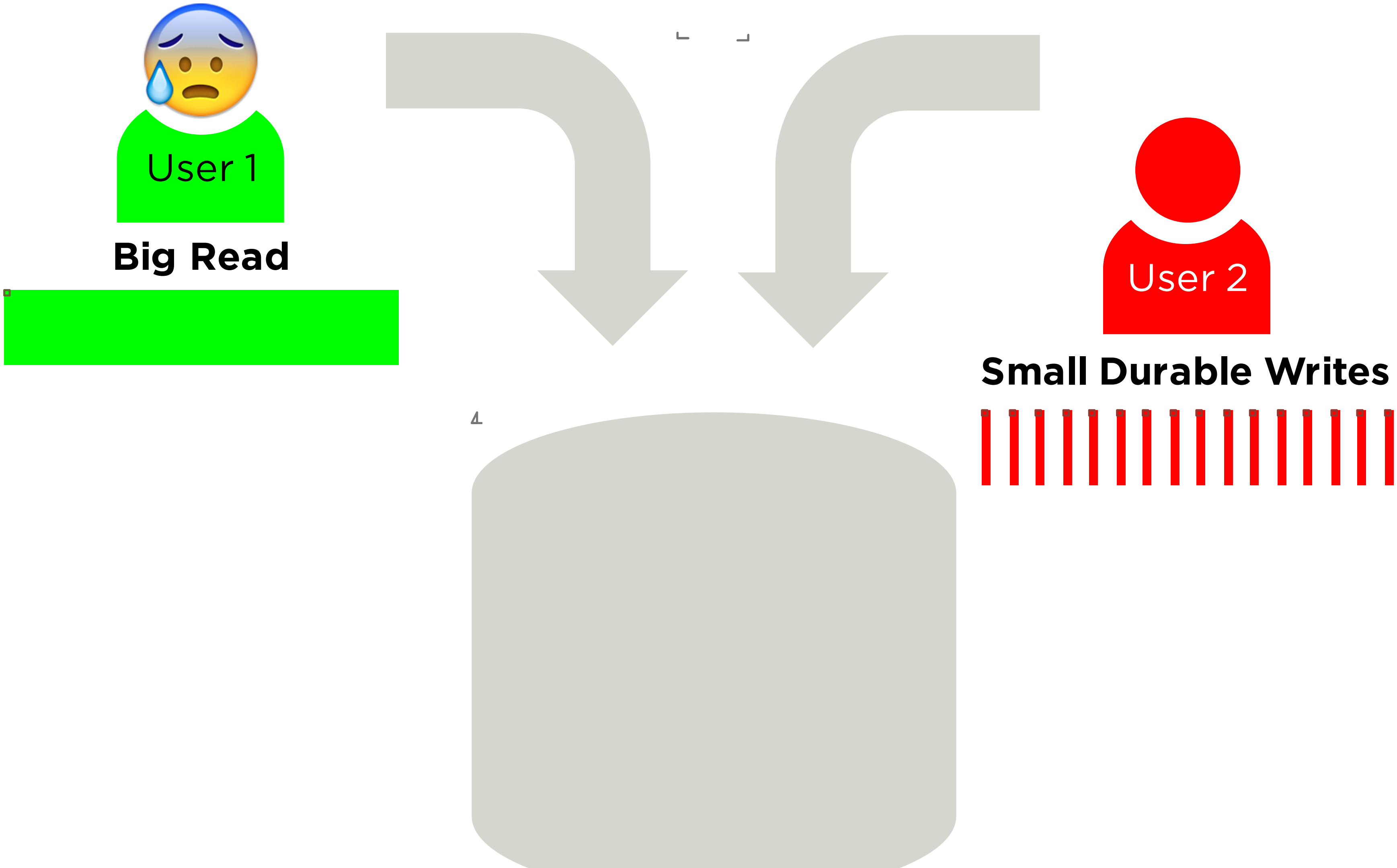

Got 50% of the
read bandwidth
User 1
Big Read
User 2
Big Read
Fair Share


1
2
3
N
Still Fair!
1/N bandwidth

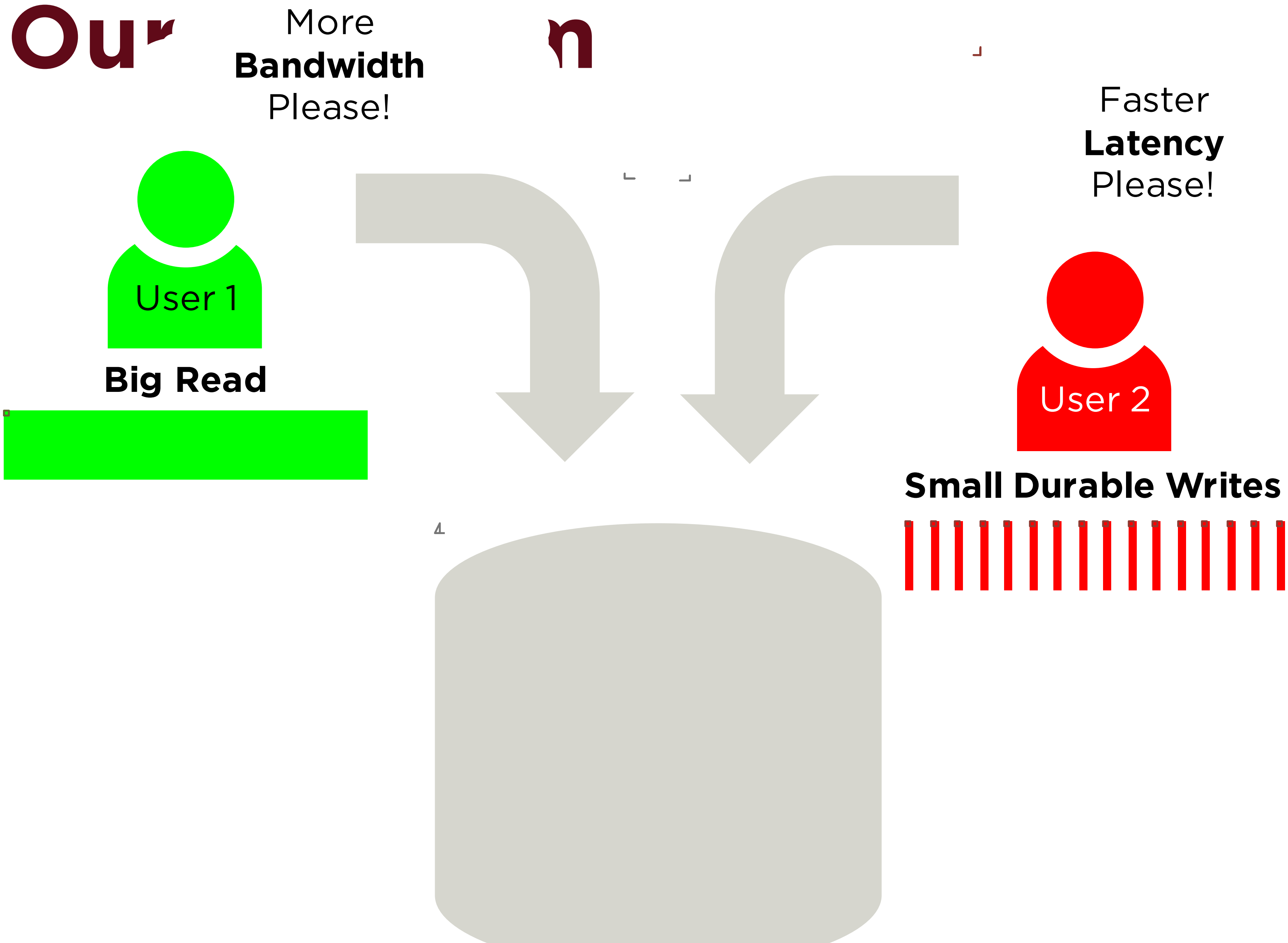
Our

Oh no!

5% bandwidth?!

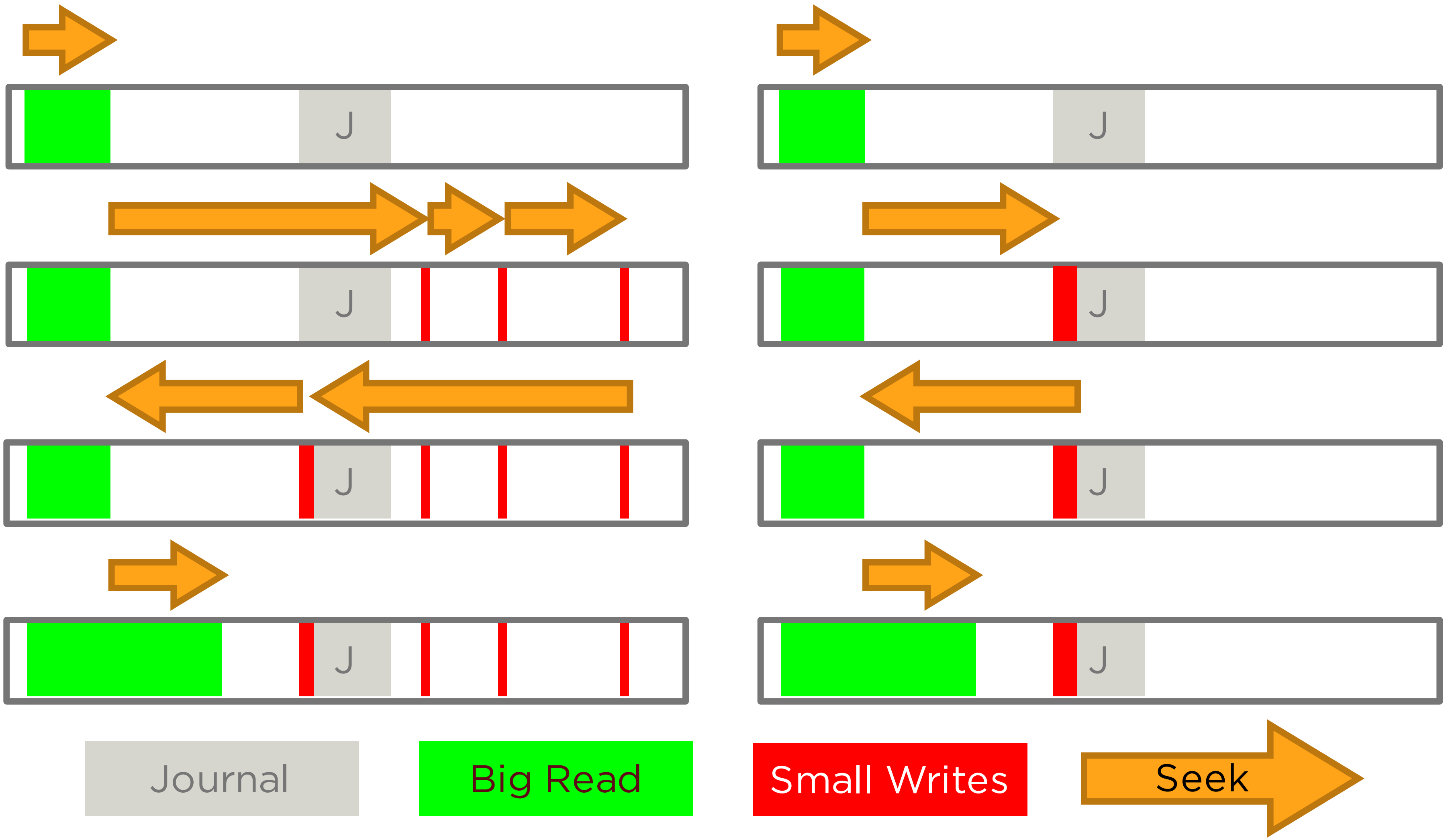


Ou n


More
Bandwidth
Please!
Faster
Latency
Please!
User 1
Big Read
User 2
Small Durable Writes

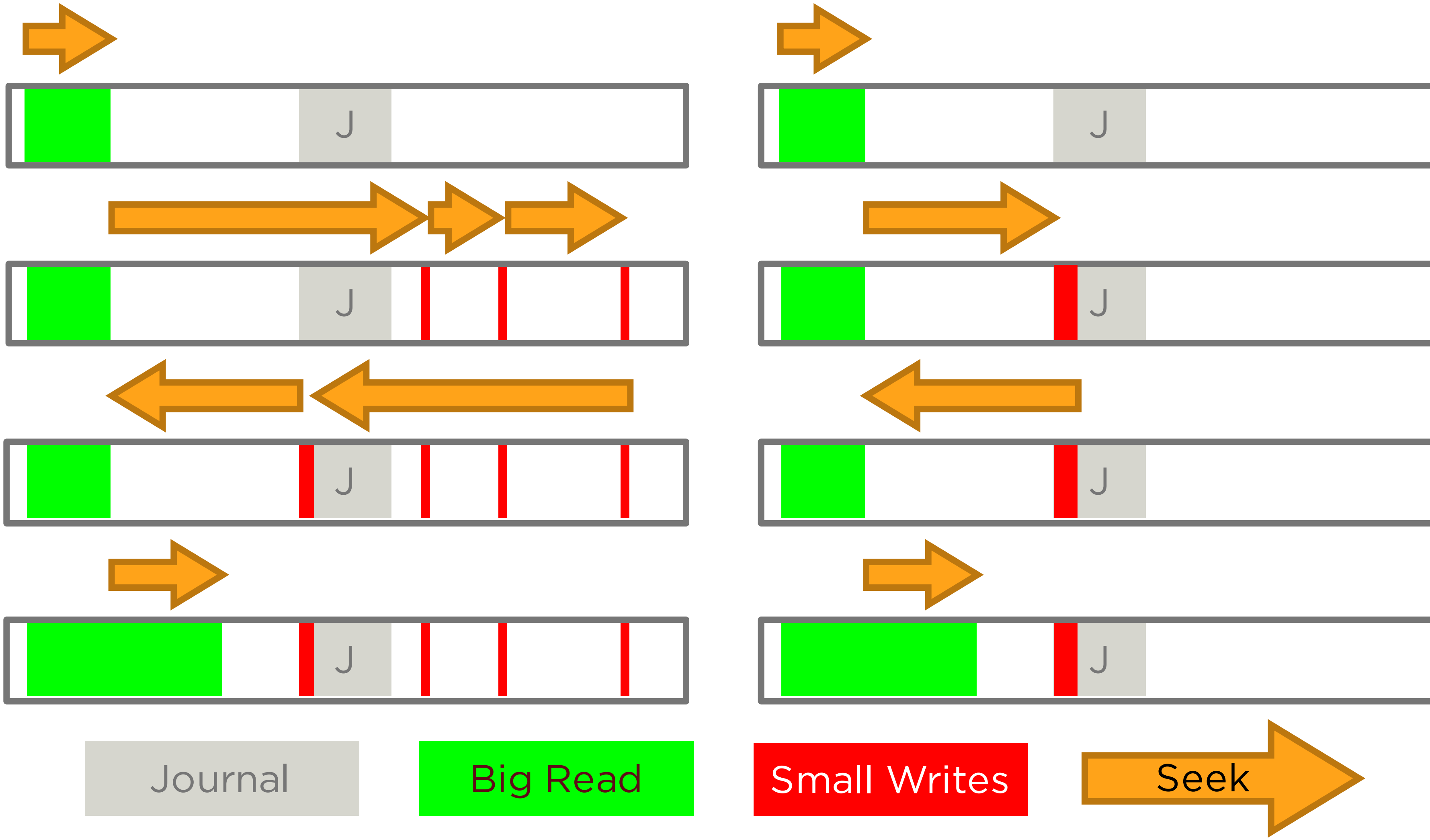
Ordered Journaling

Data Journaling

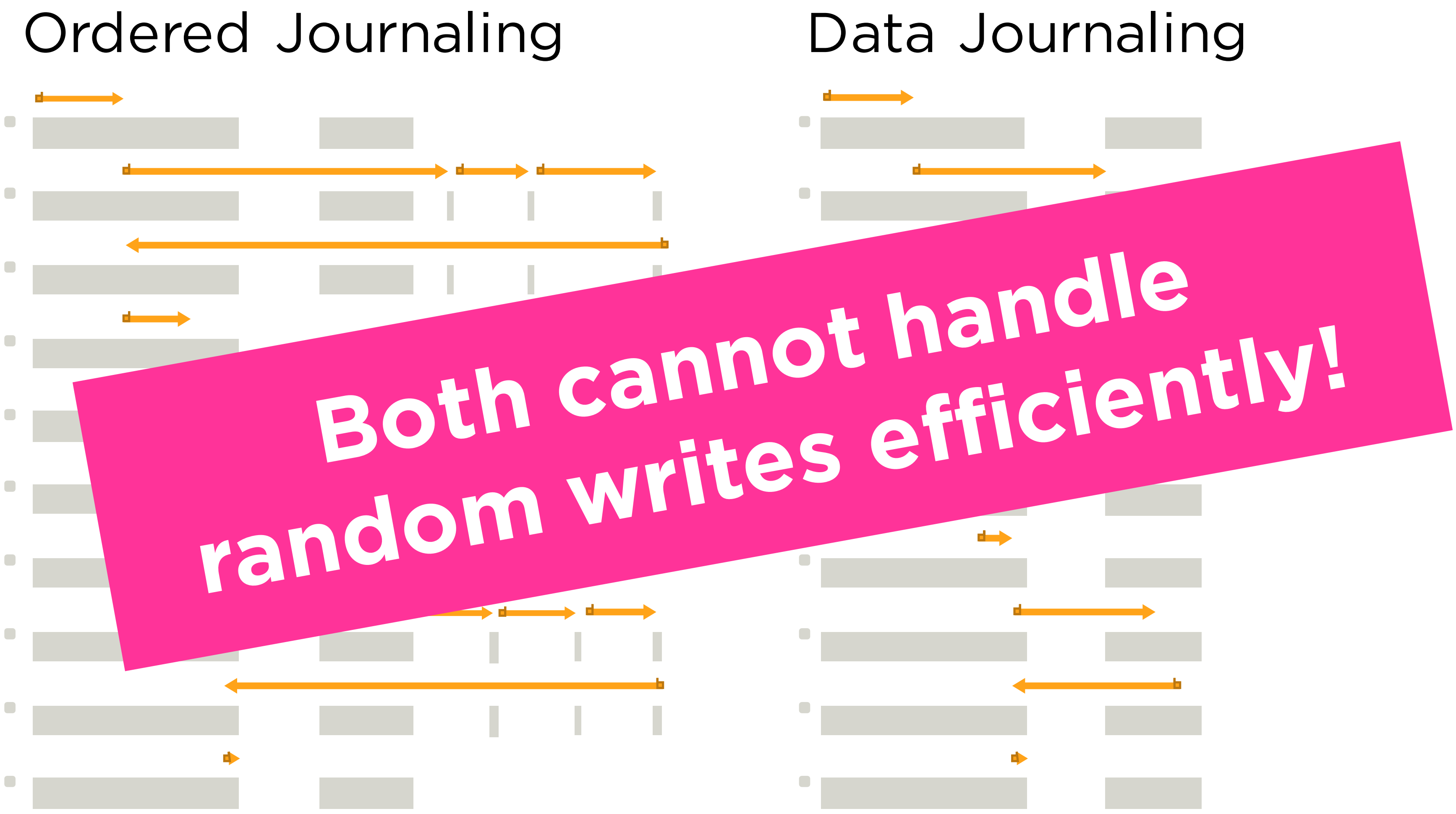


Ordered Journaling

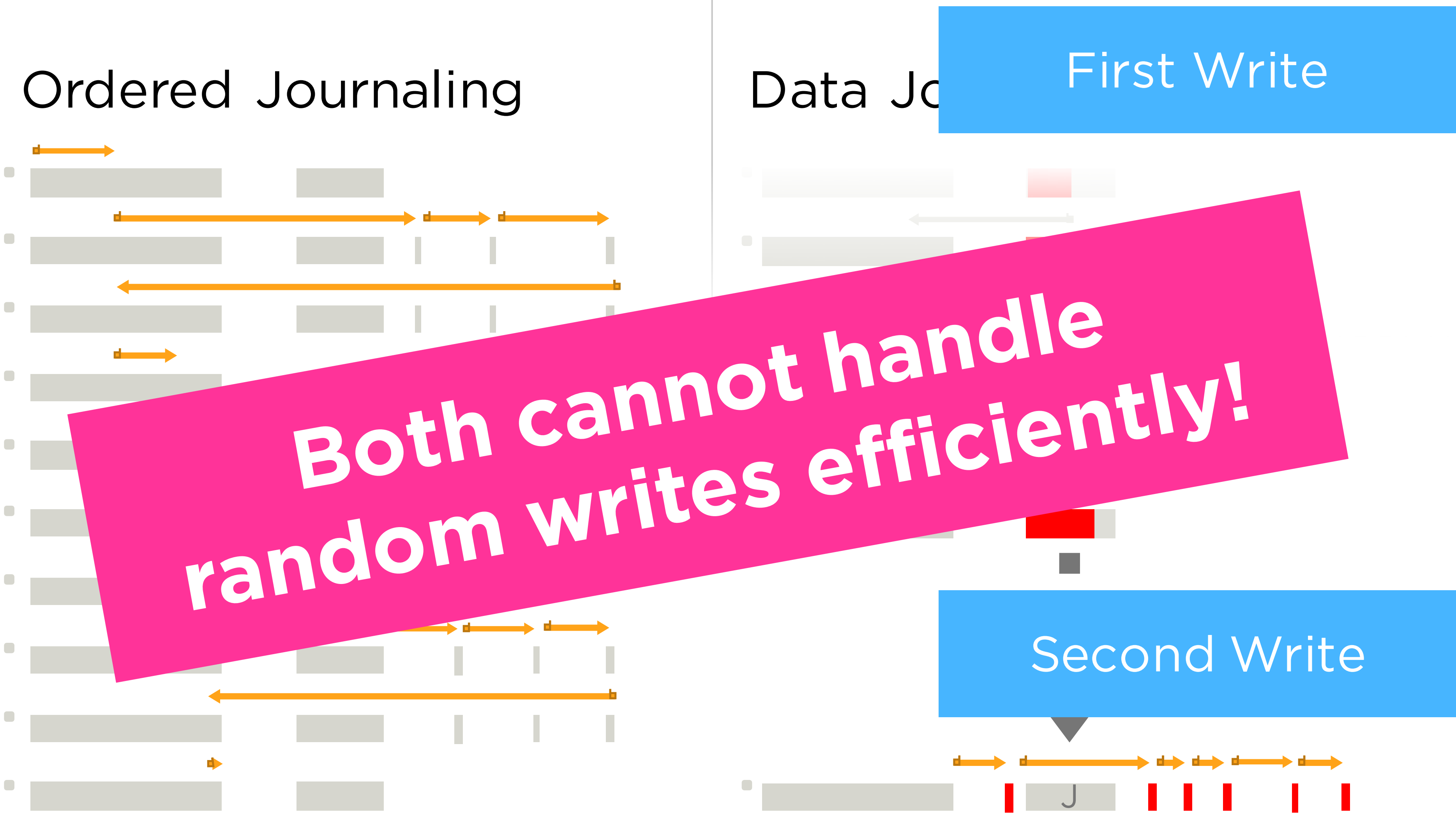
Data Journaling



Problems with Current Journaling



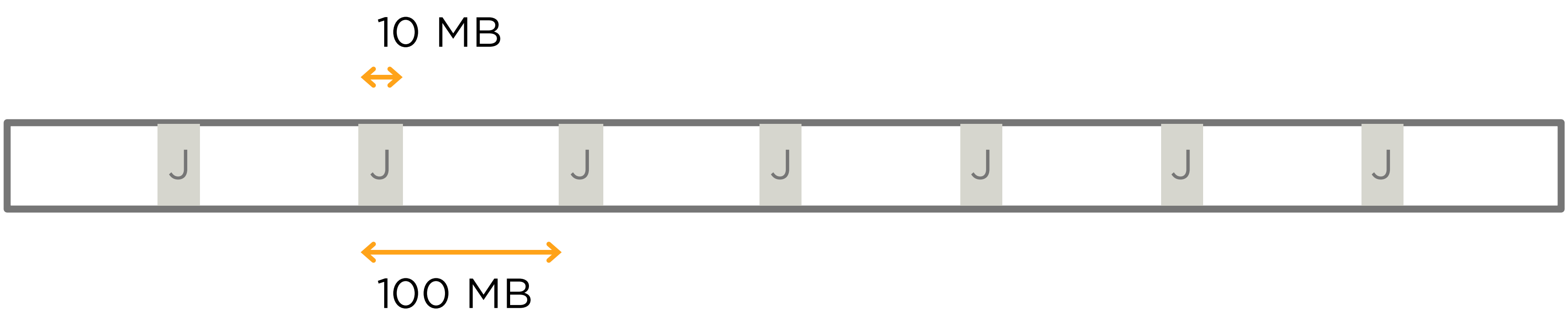
Problems with Current Journaling



Introducing Manylogs

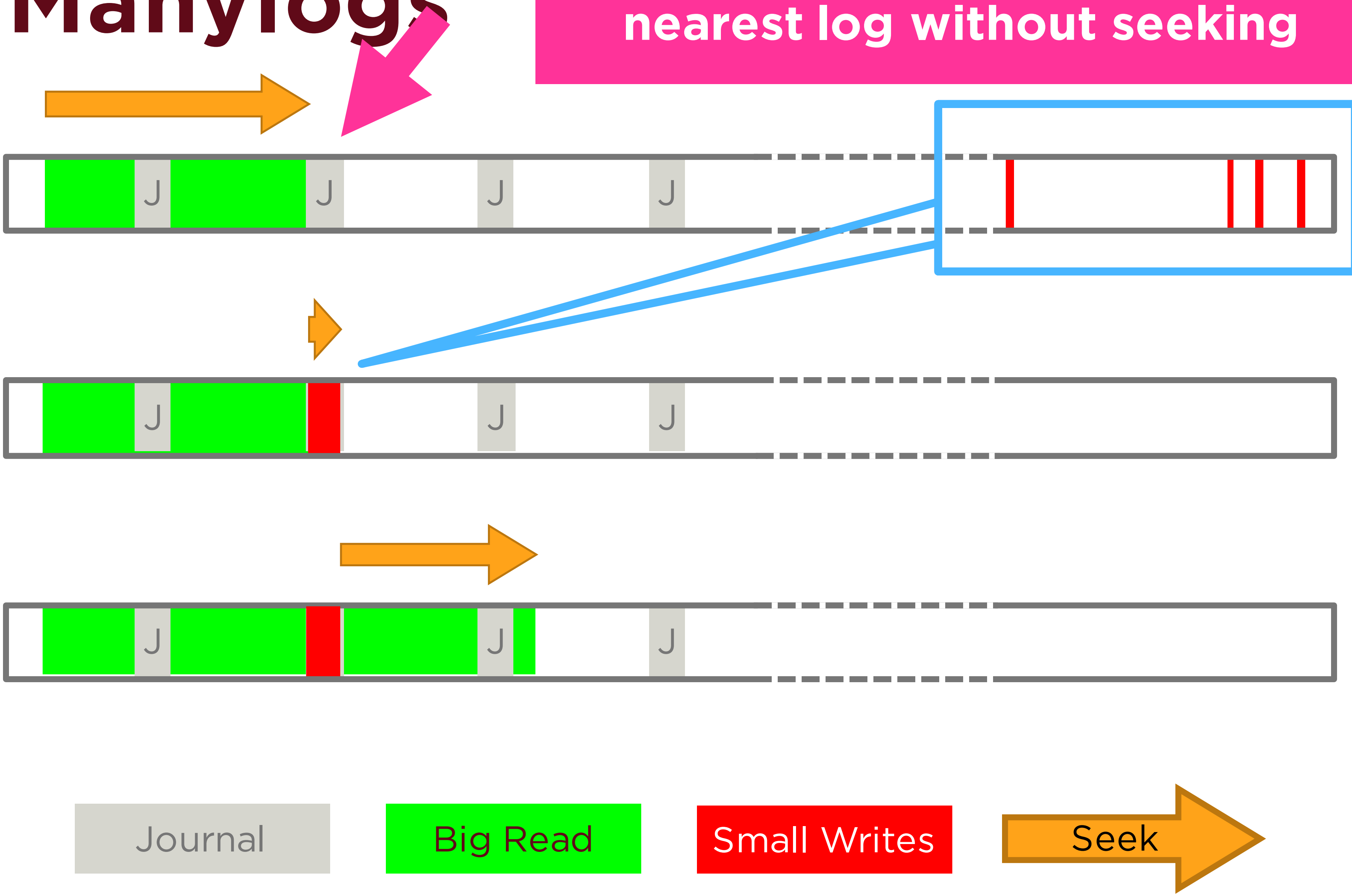
Single Log

Manylogs



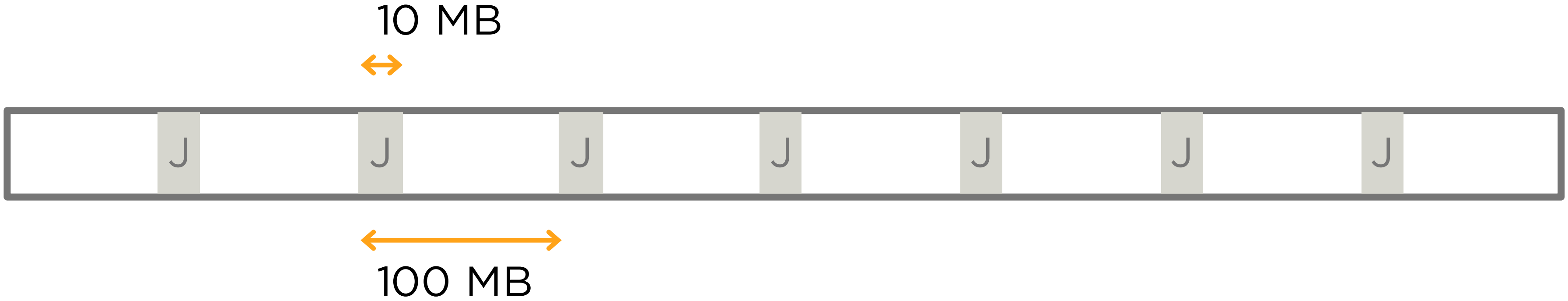
Manylogs

Small writes made durable to the **nearest log without seeking**

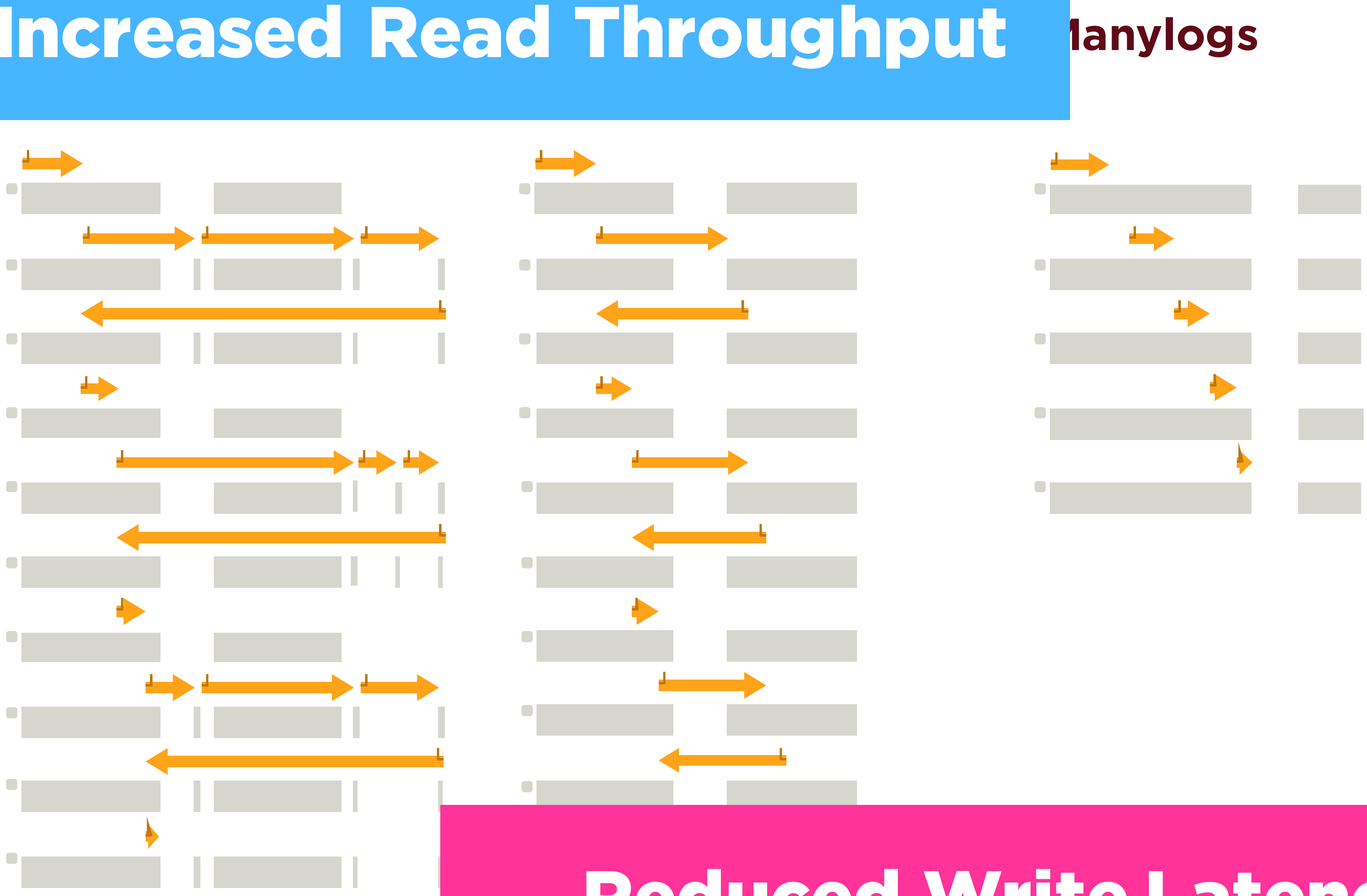


Manylogs

- 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

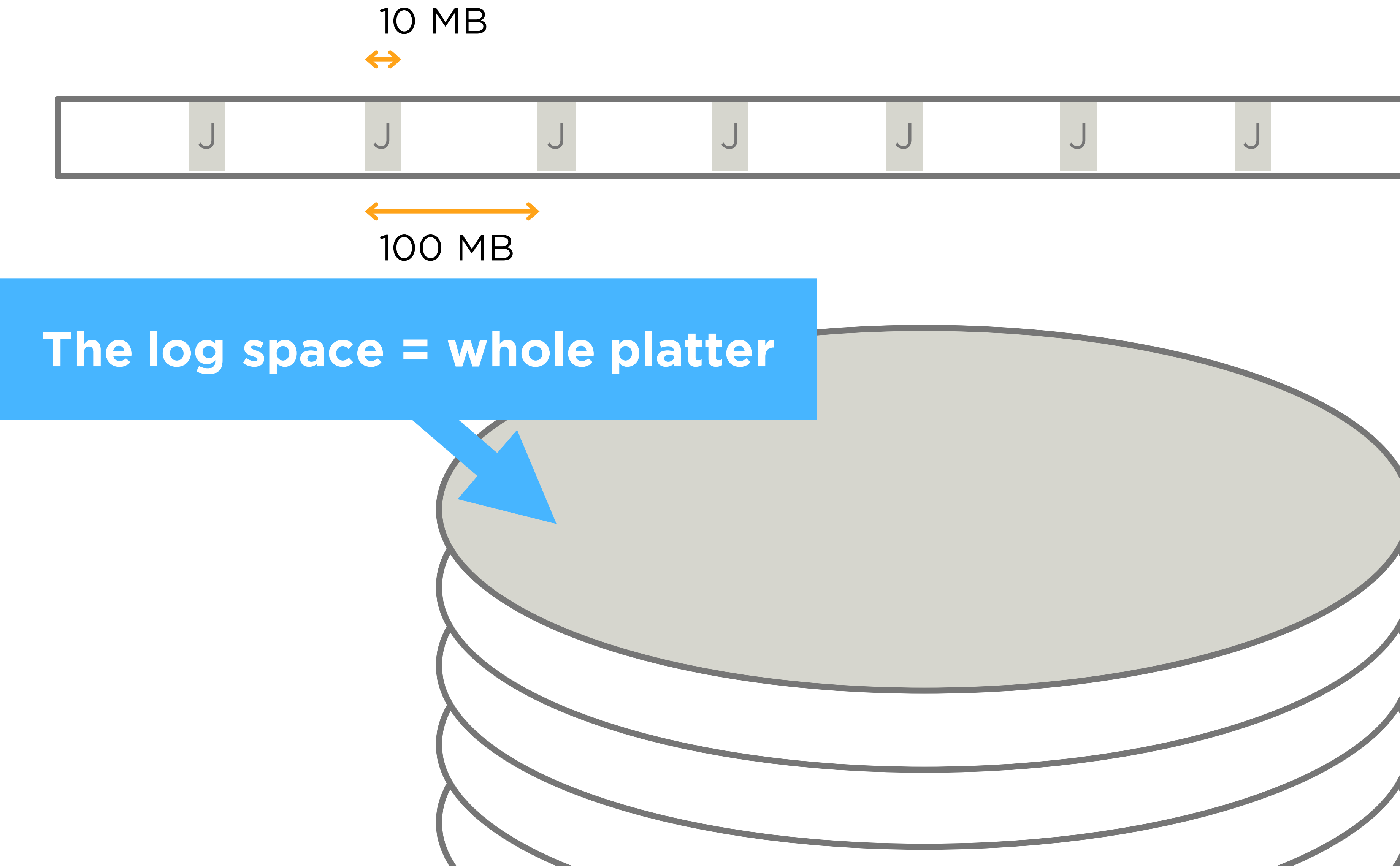


Increased Read Throughput

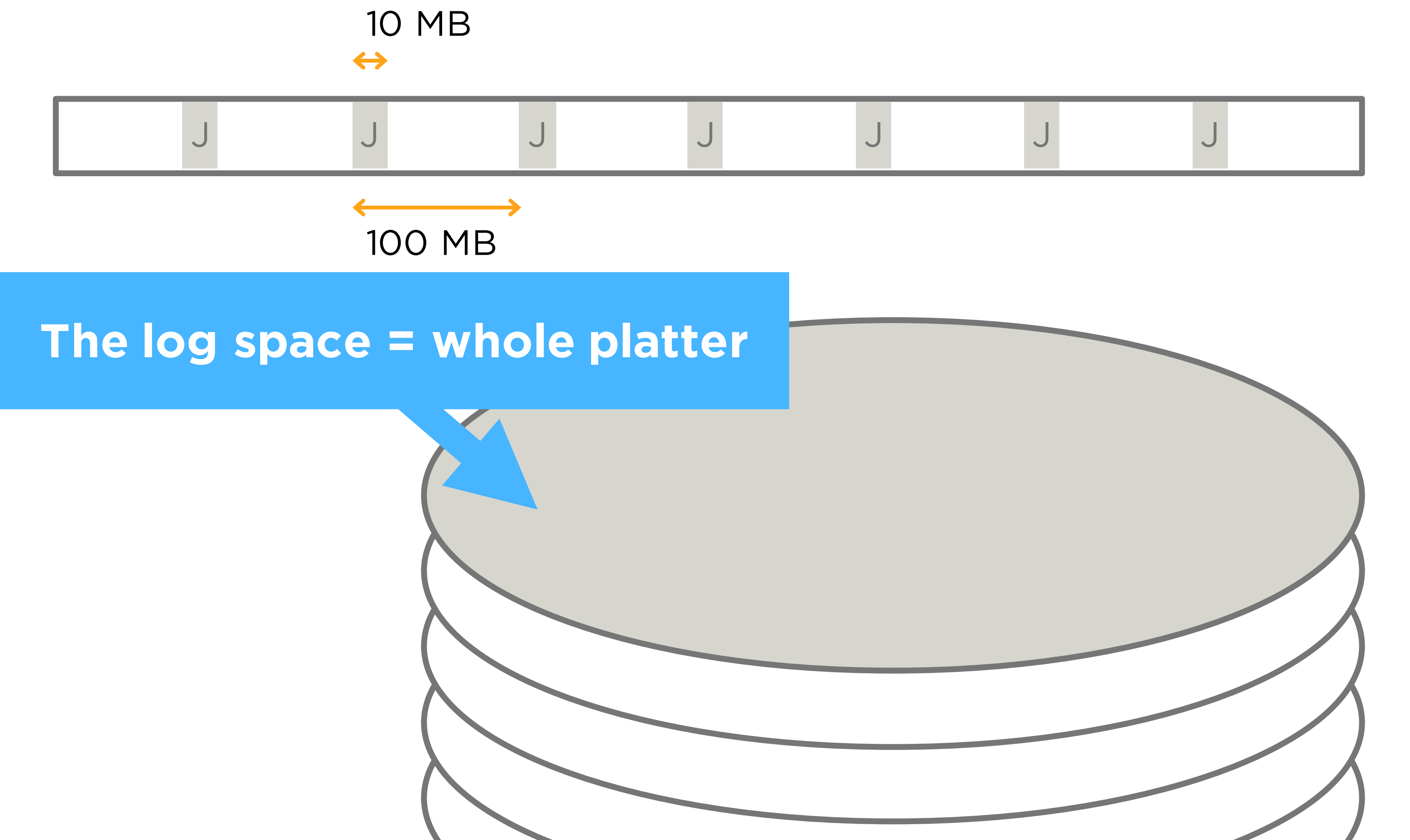


Reduced Write Latency

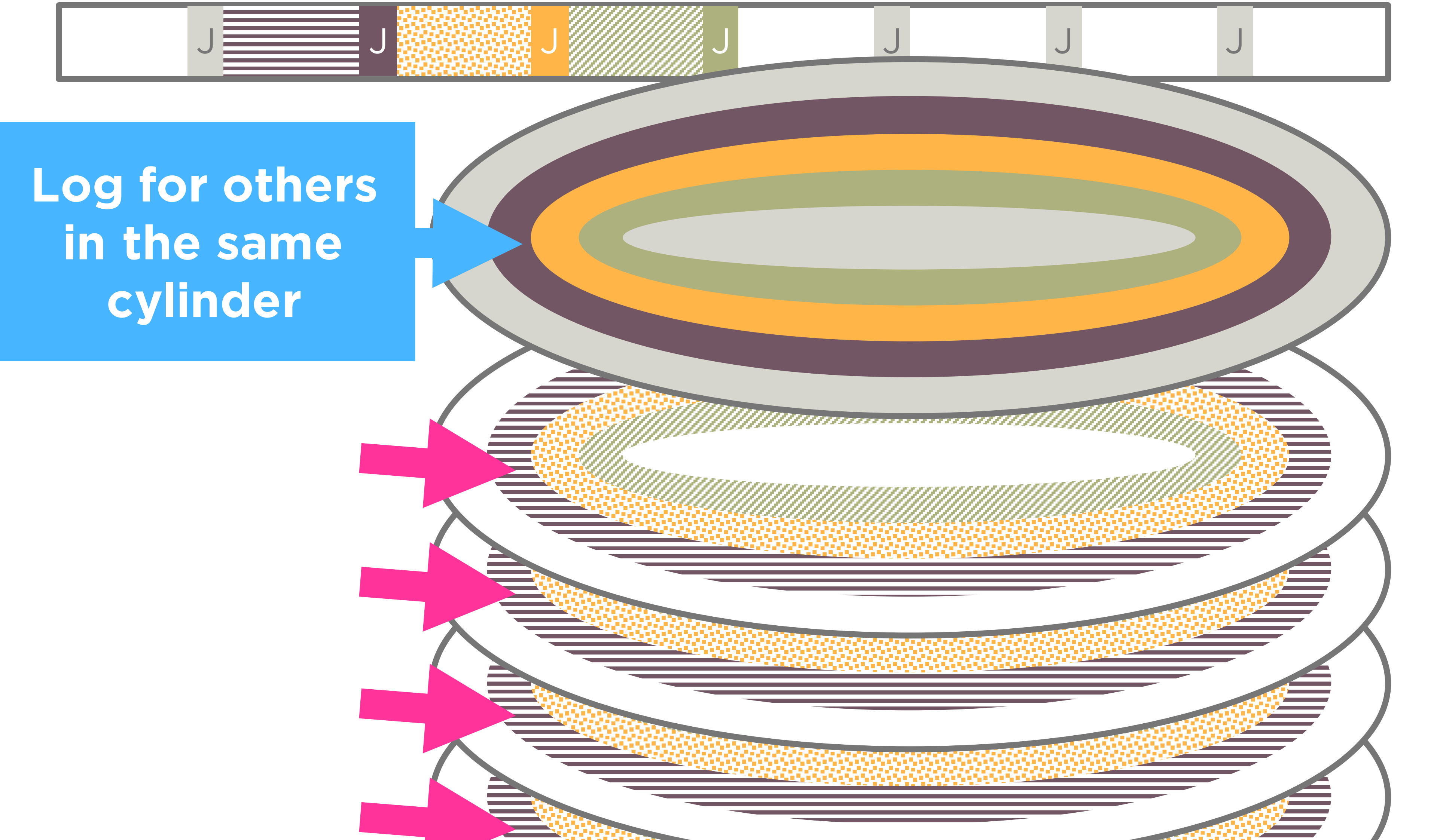
Where are logs on the disk?

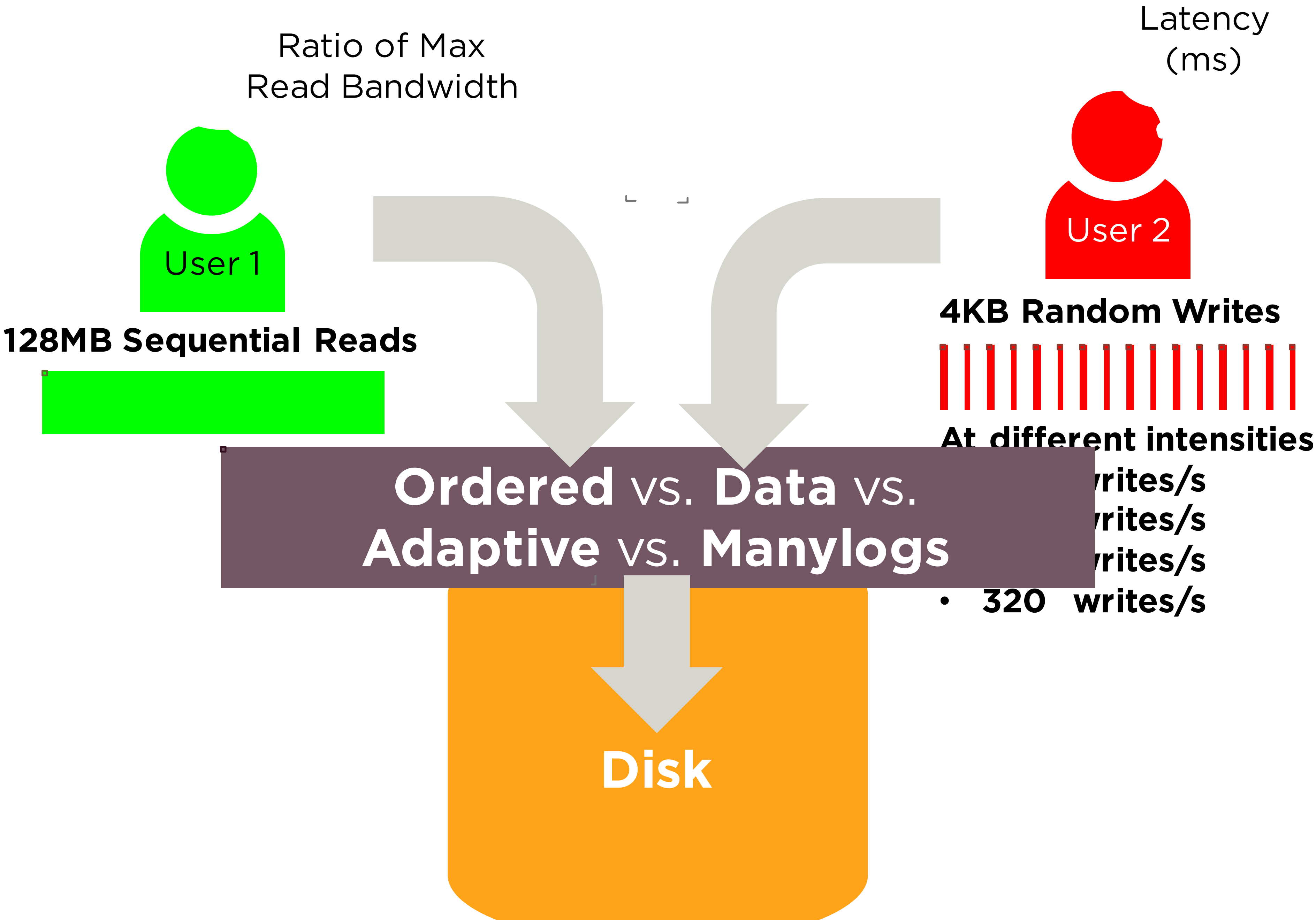


Where are logs on the disk?



Same cylinder = No seek!



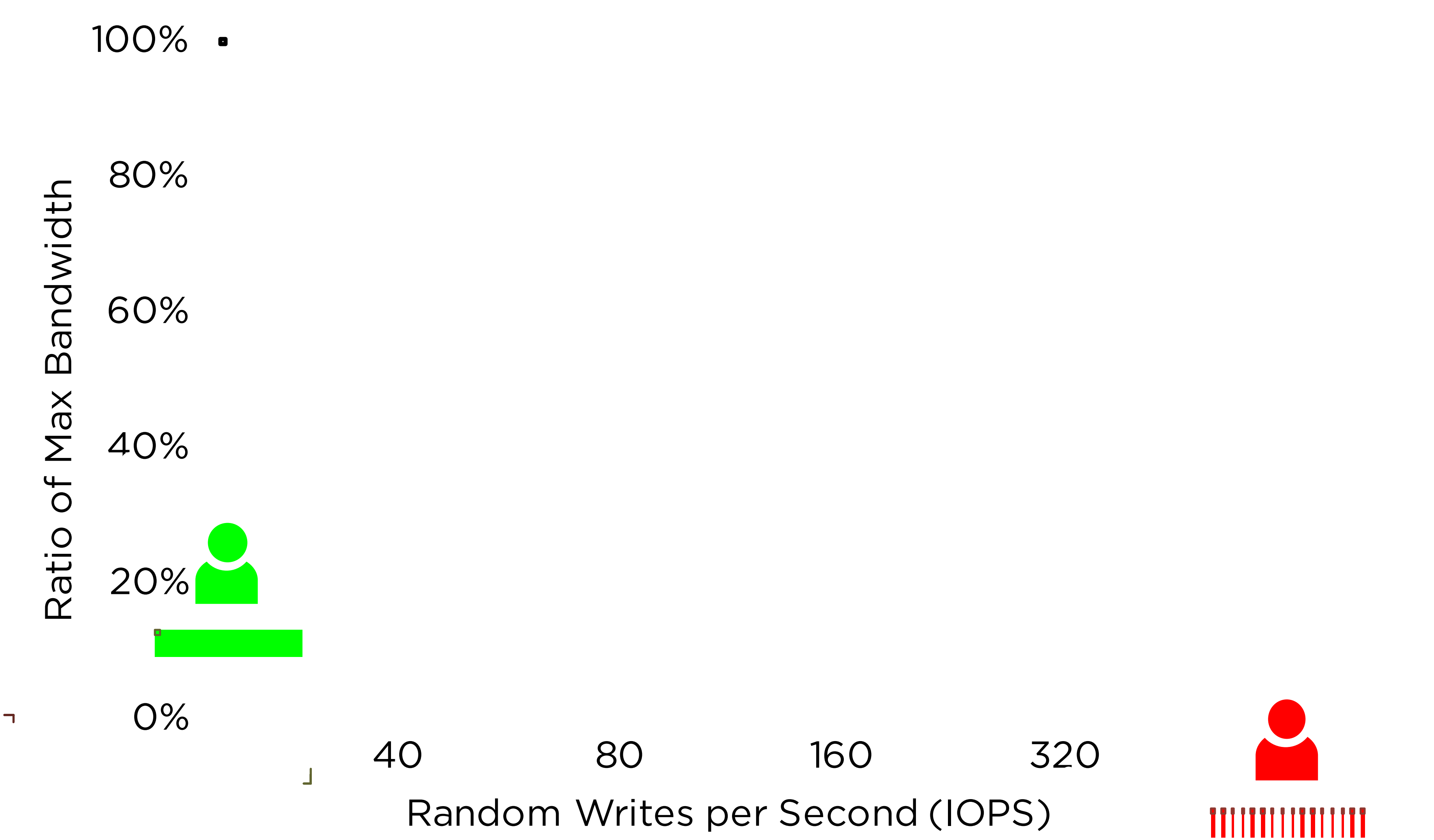

Ratio of Max
Read Bandwidth
Latency
(ms)
User 1
128MB Sequential Reads
User 2
4KB Random Writes
At different intensities
• 320 writes/s
Ordered vs. Data vs.
Adaptive vs. Manylogs
Disk

Adaptive Journaling

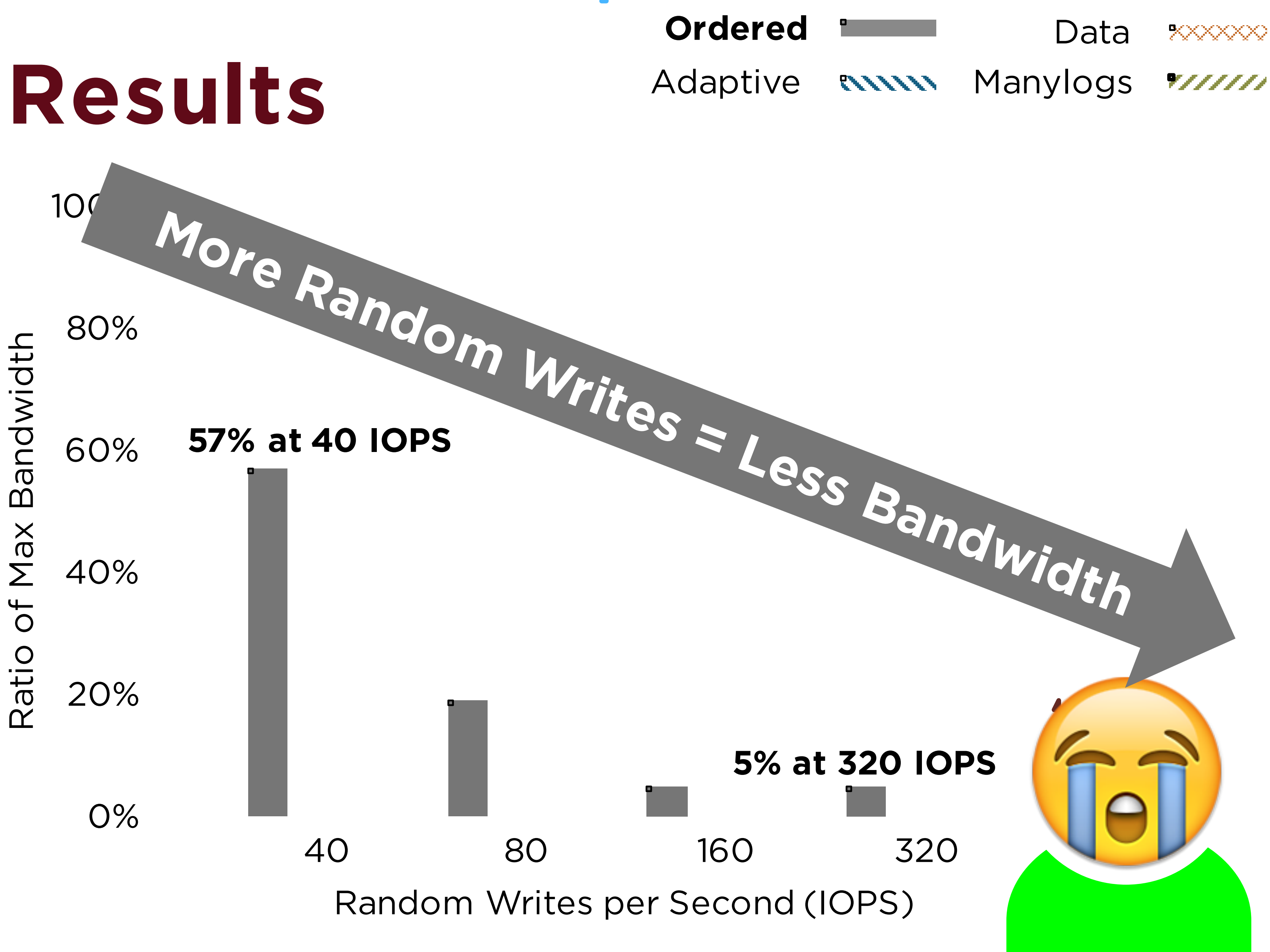
- Middle ground between ordered journaling and data journaling
- **Single-log** design
- Prabhakaran et al., ATC '05

Vijayan Prabhakaran, Andrea C. Arpaci-Dusseau, and Remzi H. Arpaci-Dusseau. "Analysis and Evolution of Journaling File Systems." In USENIX Annual Technical Conference, General Track, pp. 105-120. 2005.

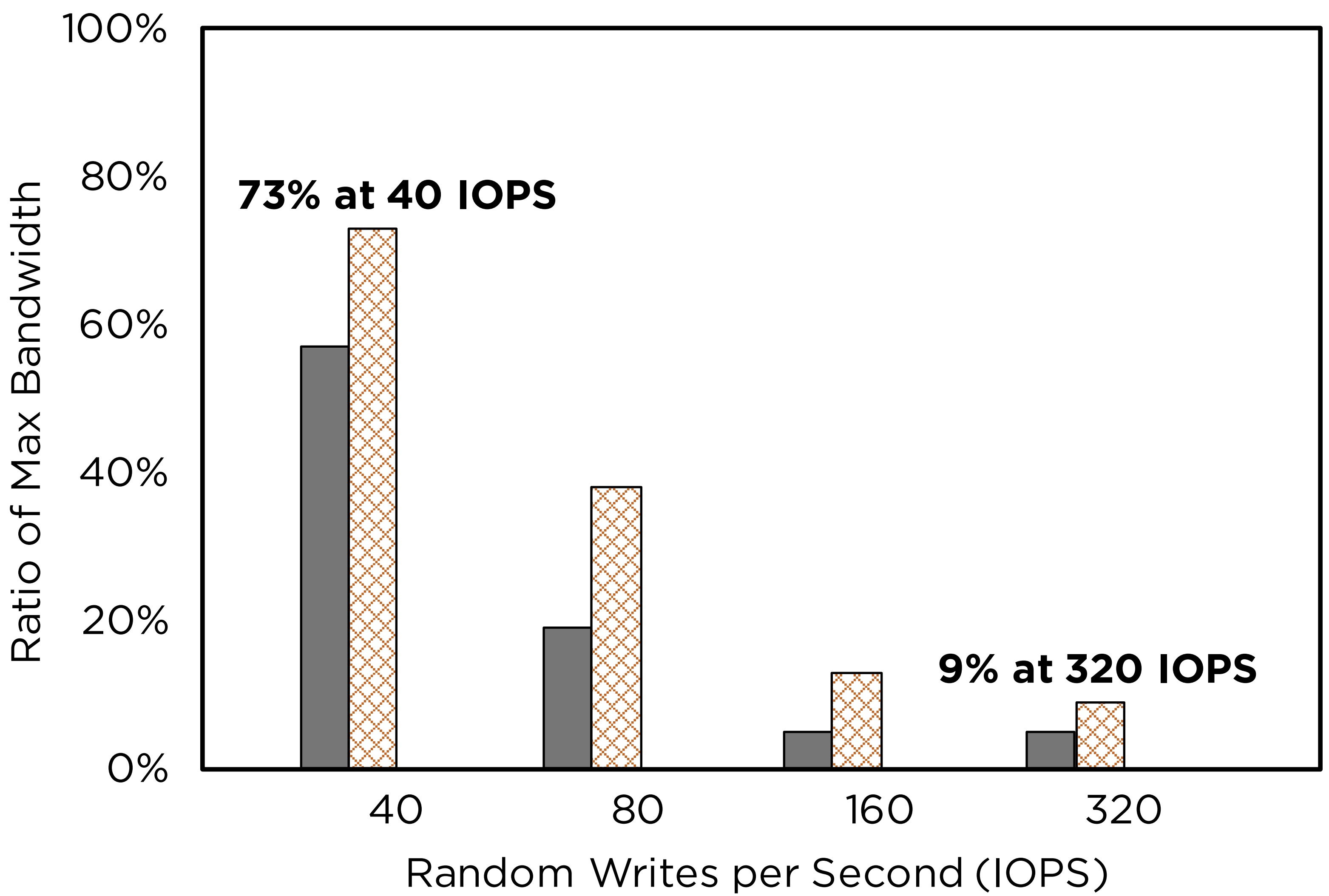
Results


Ordered
Data
Adaptive
Manylogs
100%
80%
60%
40%
20%
0%
Ratio of Max Bandwidth
40
80
160
320
Random Writes per Second (IOPS)

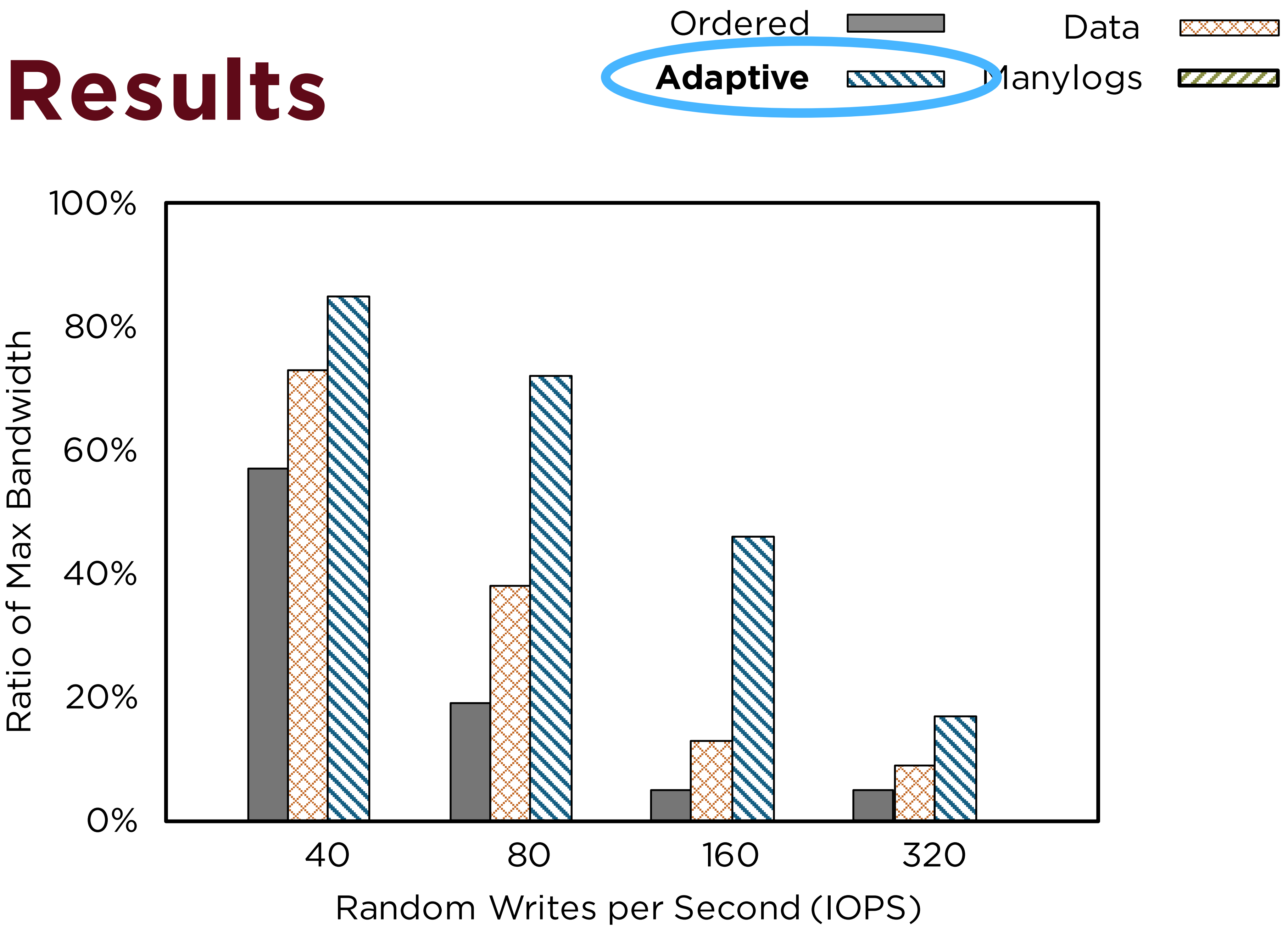
Results


Ordered
Data
Adaptive
Manylogs
More Random Writes = Less Bandwidth
57% at 40 IOPS
5% at 320 IOPS
Ratio of Max Bandwidth
100
80%
60%
40%
20%
0%
40
80
160
320
Random Writes per Second (IOPS)

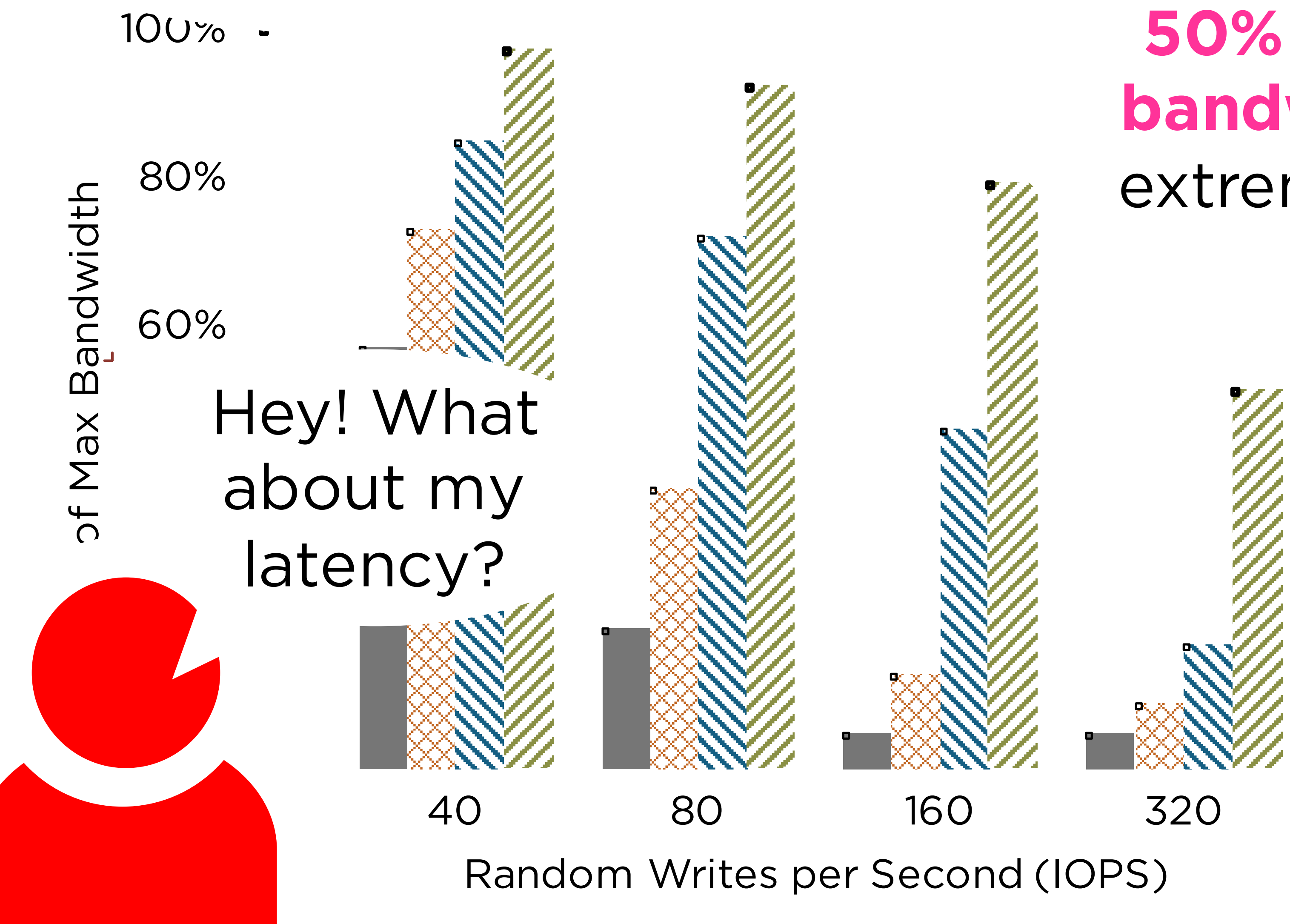
Results


Ordered
Adaptive
Data
Manylogs
100%
80%
60%
40%
20%
0%
Ratio of Max Bandwidth
73% at 40 IOPS
9% at 320 IOPS
40
80
160
320
Random Writes per Second (IOPS)

Results

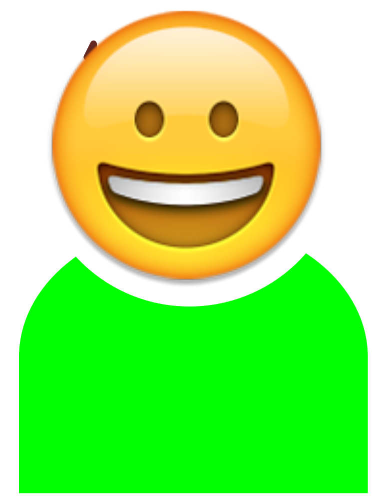

Ordered
Adaptive
Data
Manylogs
100%
80%
60%
40%
20%
0%
Ratio of Max Bandwidth
40
80
160
320
Random Writes per Second (IOPS)

Manylogs gives the **most bandwidth**

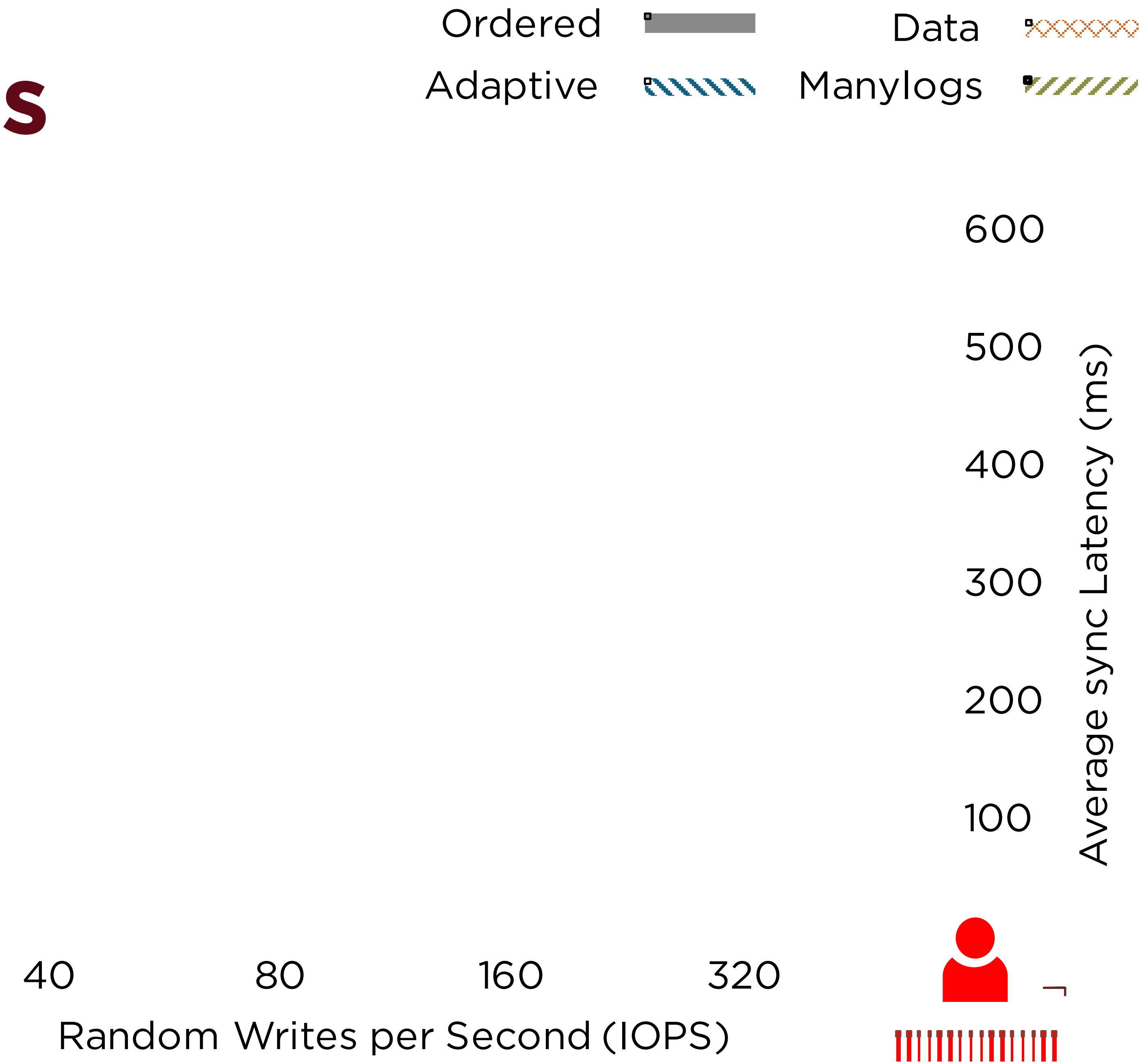


50% of max bandwidth at extreme IOPS

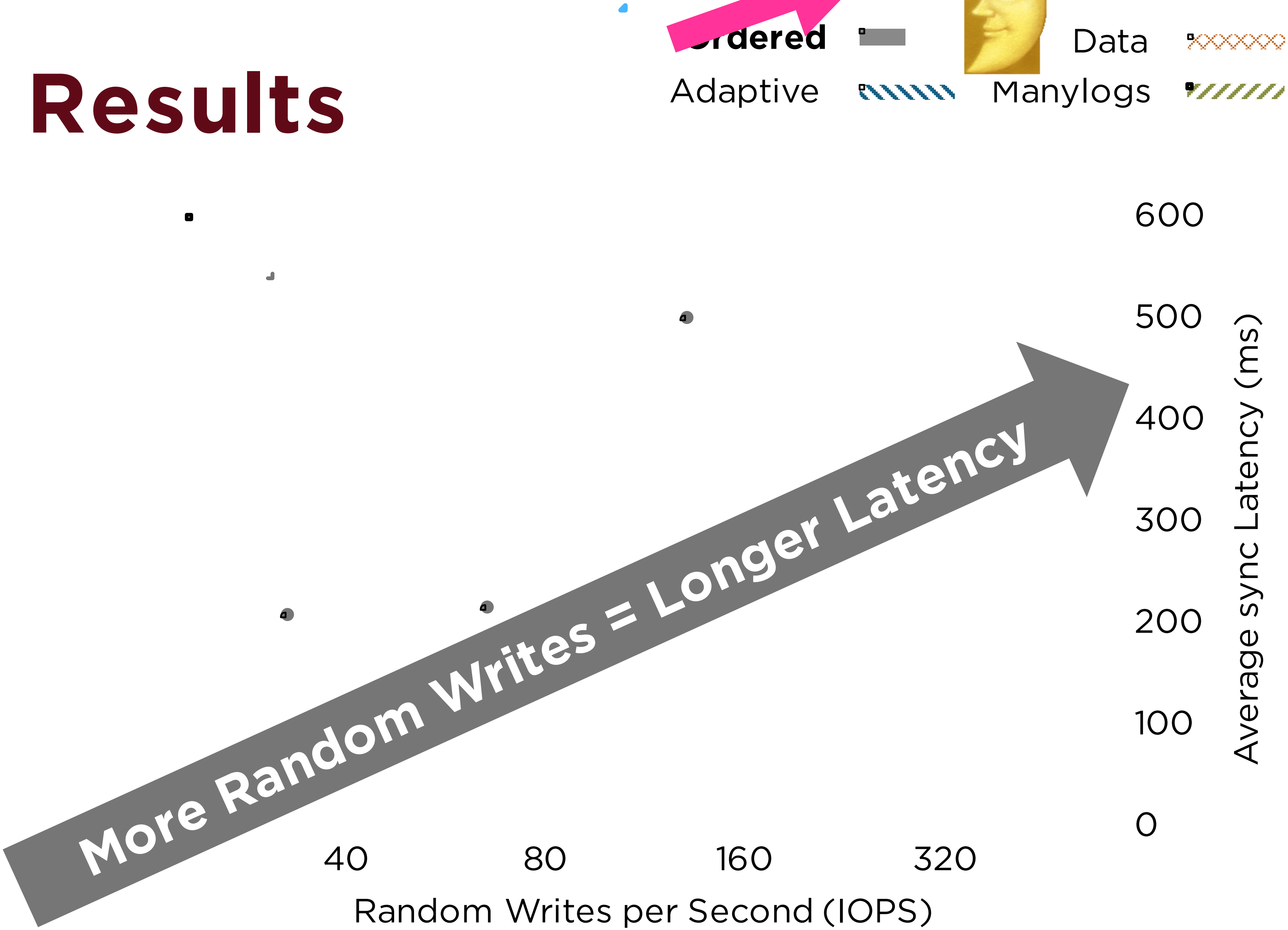
Hey! What about my latency?



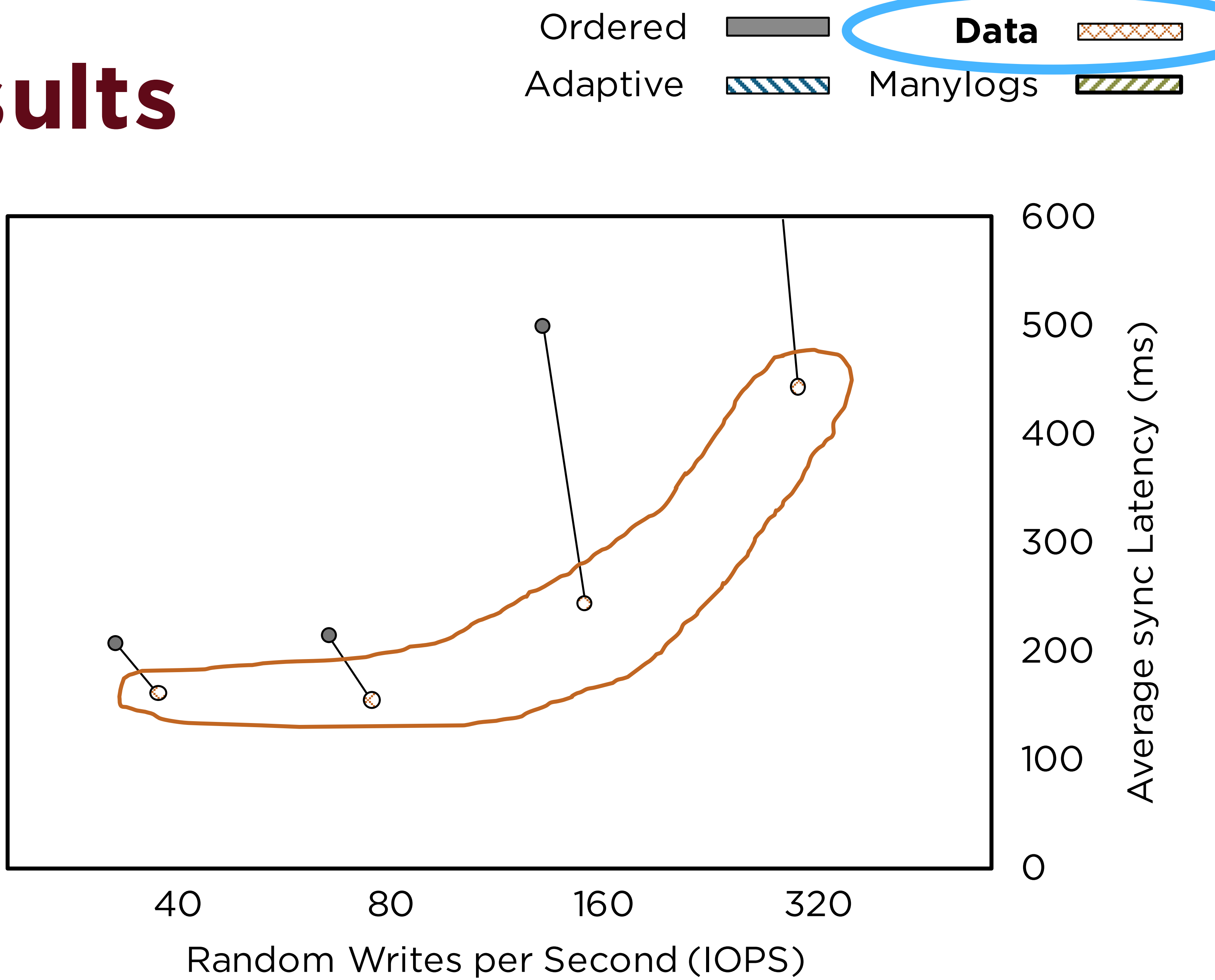
Results


Ordered
Data
Adaptive
Manylogs
600
500
400
300
200
100
Average sync Latency (ms)
40
80
160
320
Random Writes per Second (IOPS)

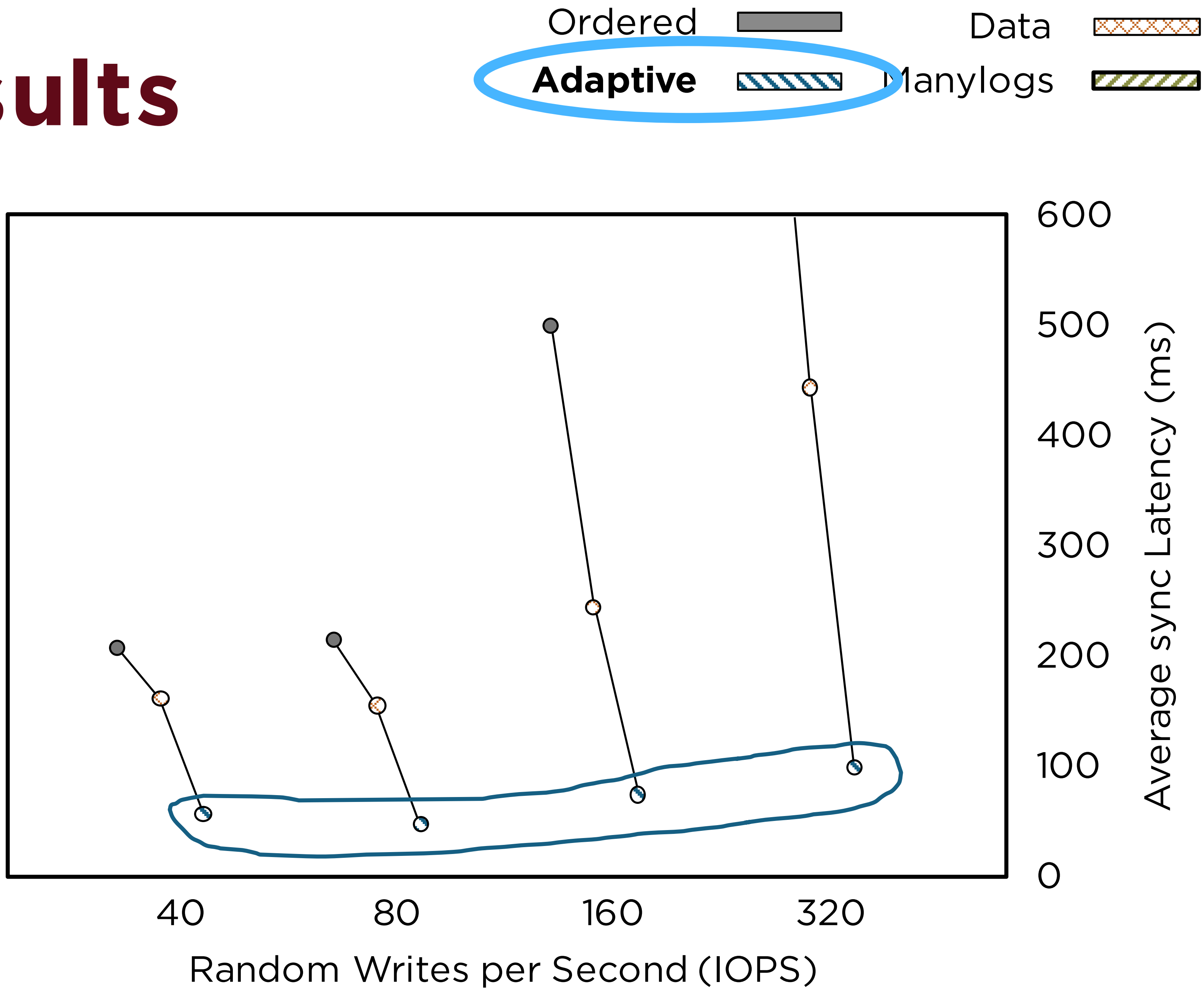
Results


Ordered
Data
Adaptive
Manylogs
600
500
400
300
200
100
0
Average sync Latency (ms)
40
80
160
320
Random Writes per Second (IOPS)
More Random Writes = Longer Latency

Results

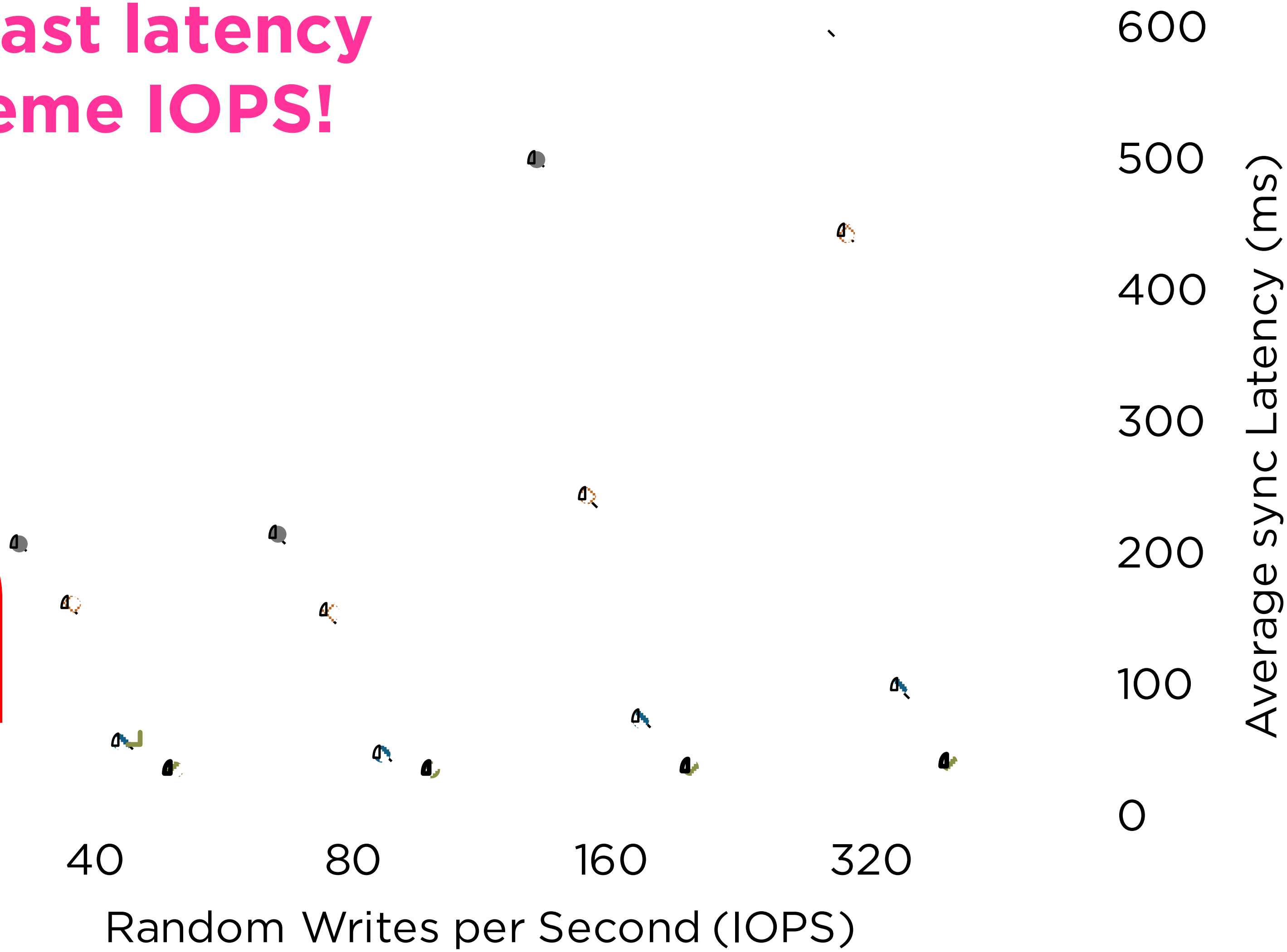

Ordered
Data
Adaptive
Manylogs
600
500
400
300
200
100
0
Average sync Latency (ms)
40
80
160
320
Random Writes per Second (IOPS)

Results

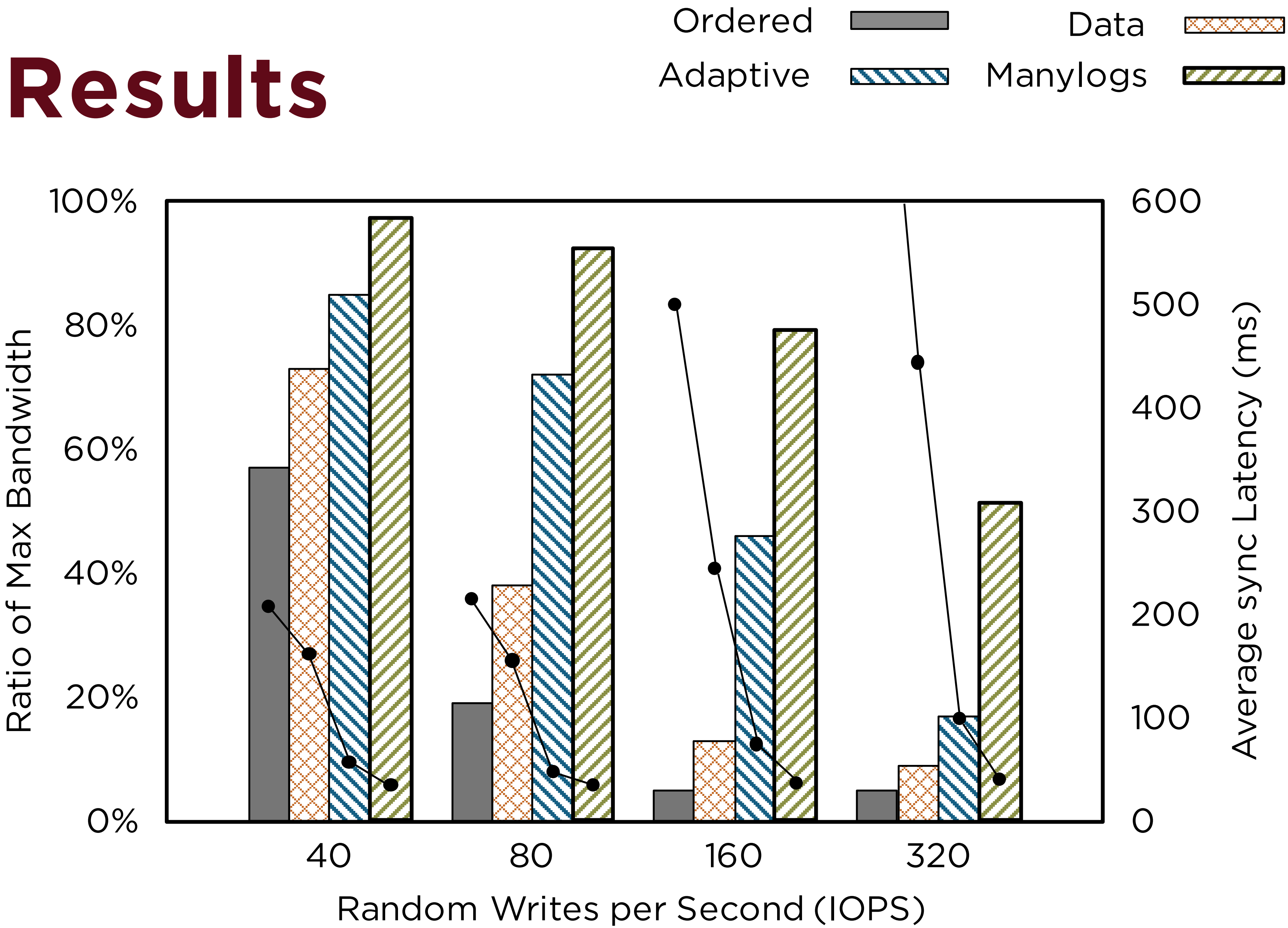

Ordered
Data
Adaptive
Manylogs
600
500
400
300
200
100
0
Average sync Latency (ms)
40
80
160
320
Random Writes per Second (IOPS)

Results

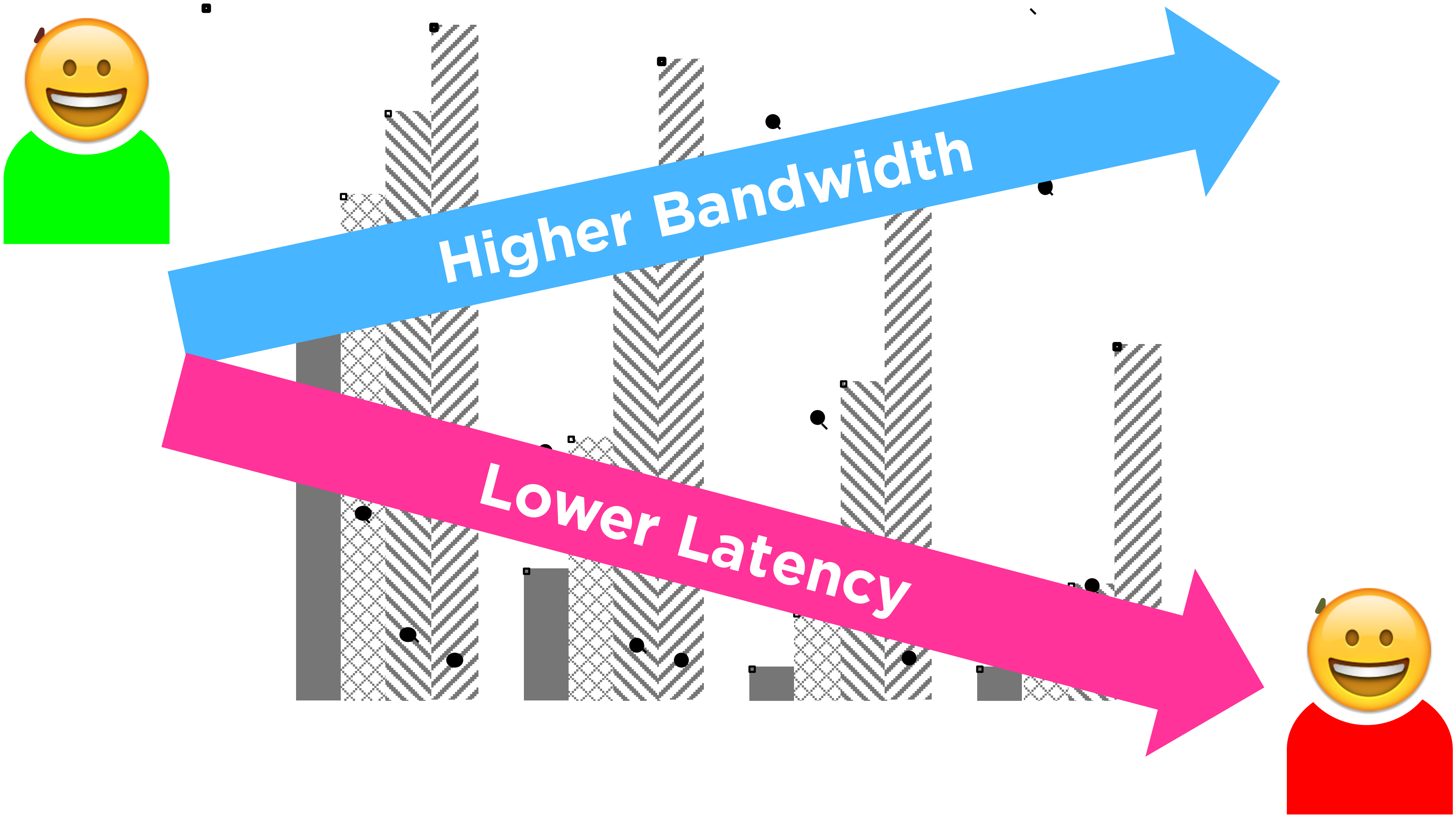
WOW! **Fast latency at extreme IOPS!**

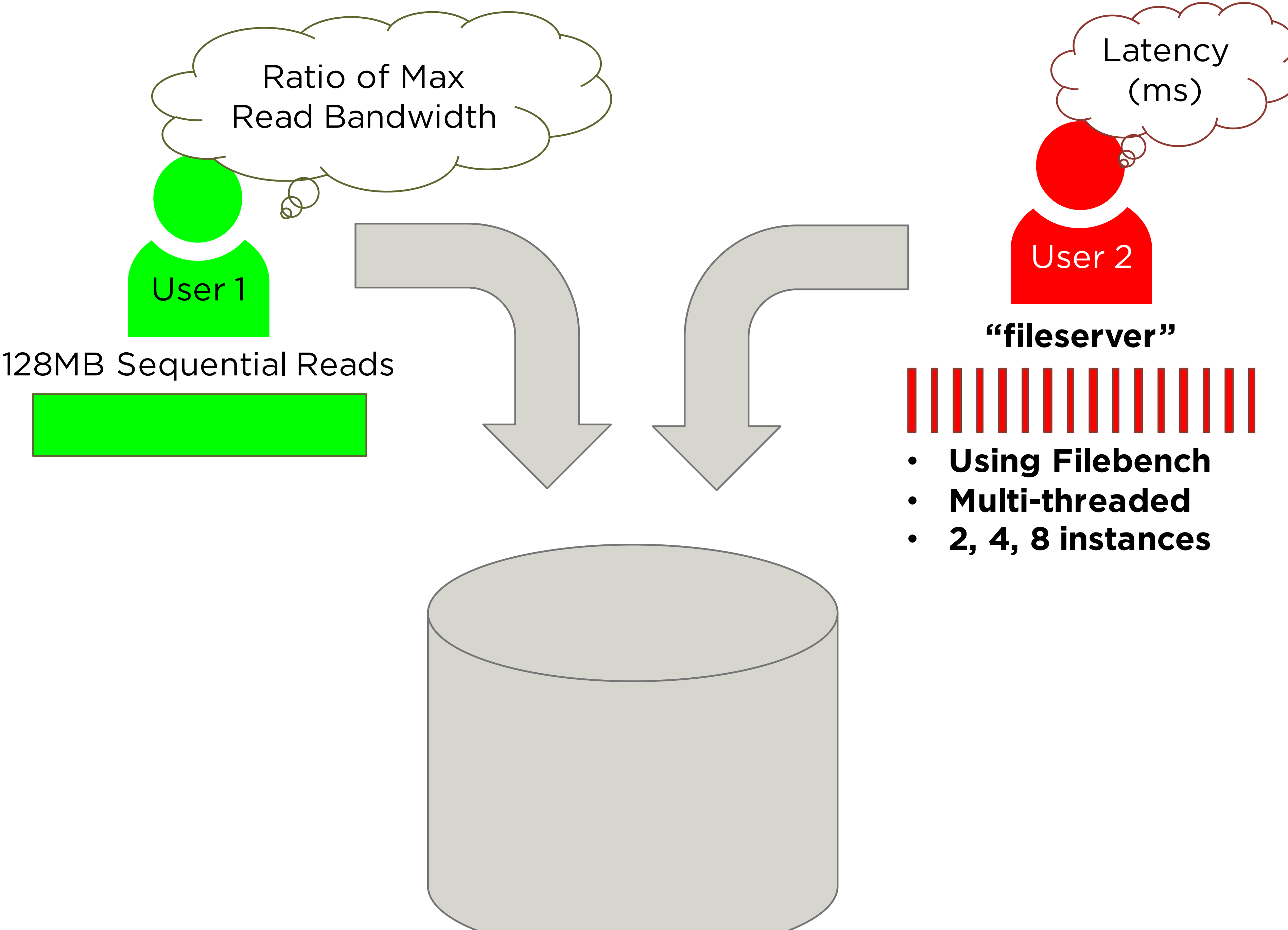


Results


Ordered
Data
Adaptive
Manylogs
100%
80%
60%
40%
20%
0%
Ratio of Max Bandwidth
600
500
400
300
200
100
0
Average sync Latency (ms)
40
80
160
320
Random Writes per Second (IOPS)

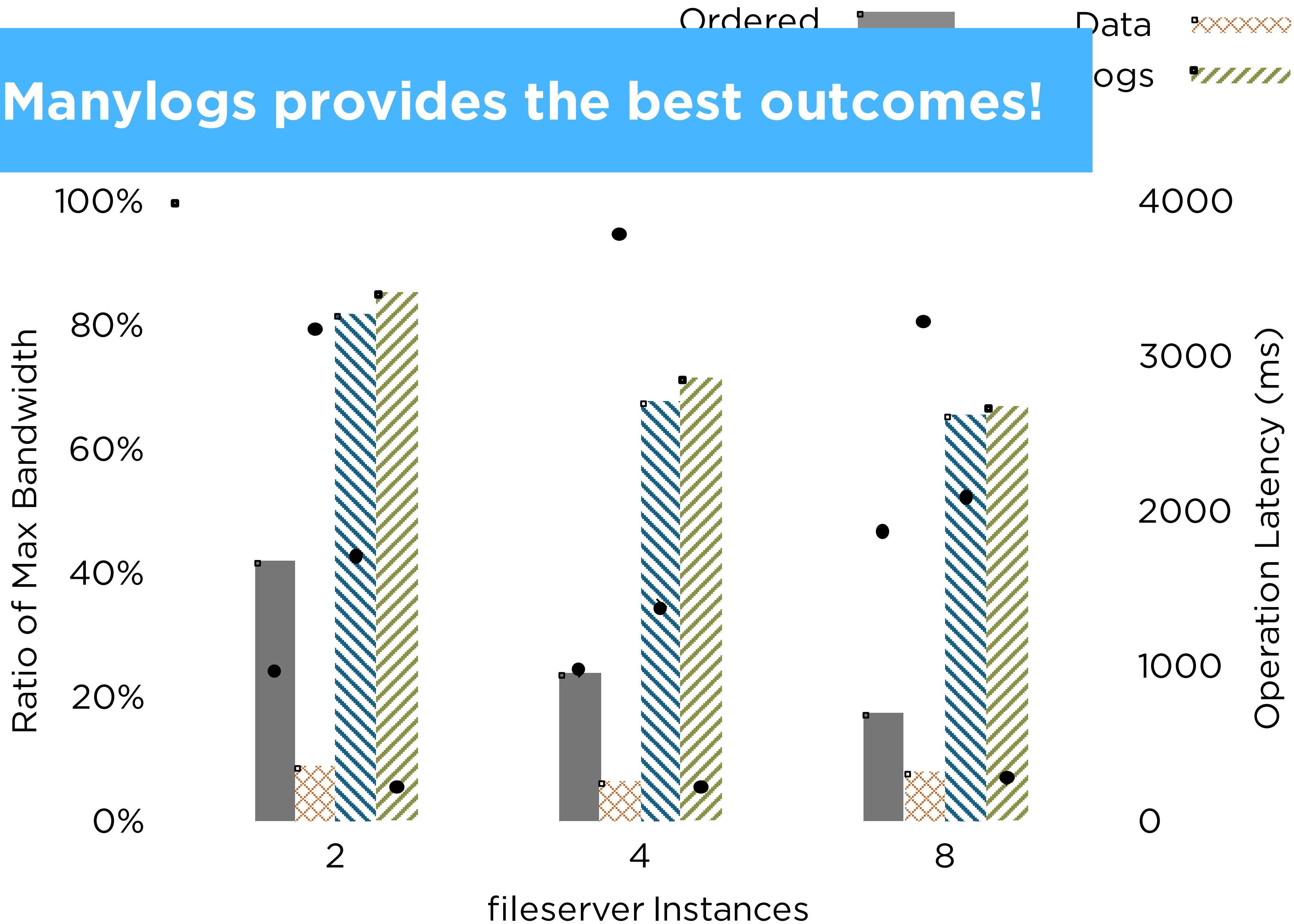
Results





- **Using Filebench**
- **Multi-threaded**
- **2, 4, 8 instances**

Manylogs provides the best outcomes!



Checkpointing

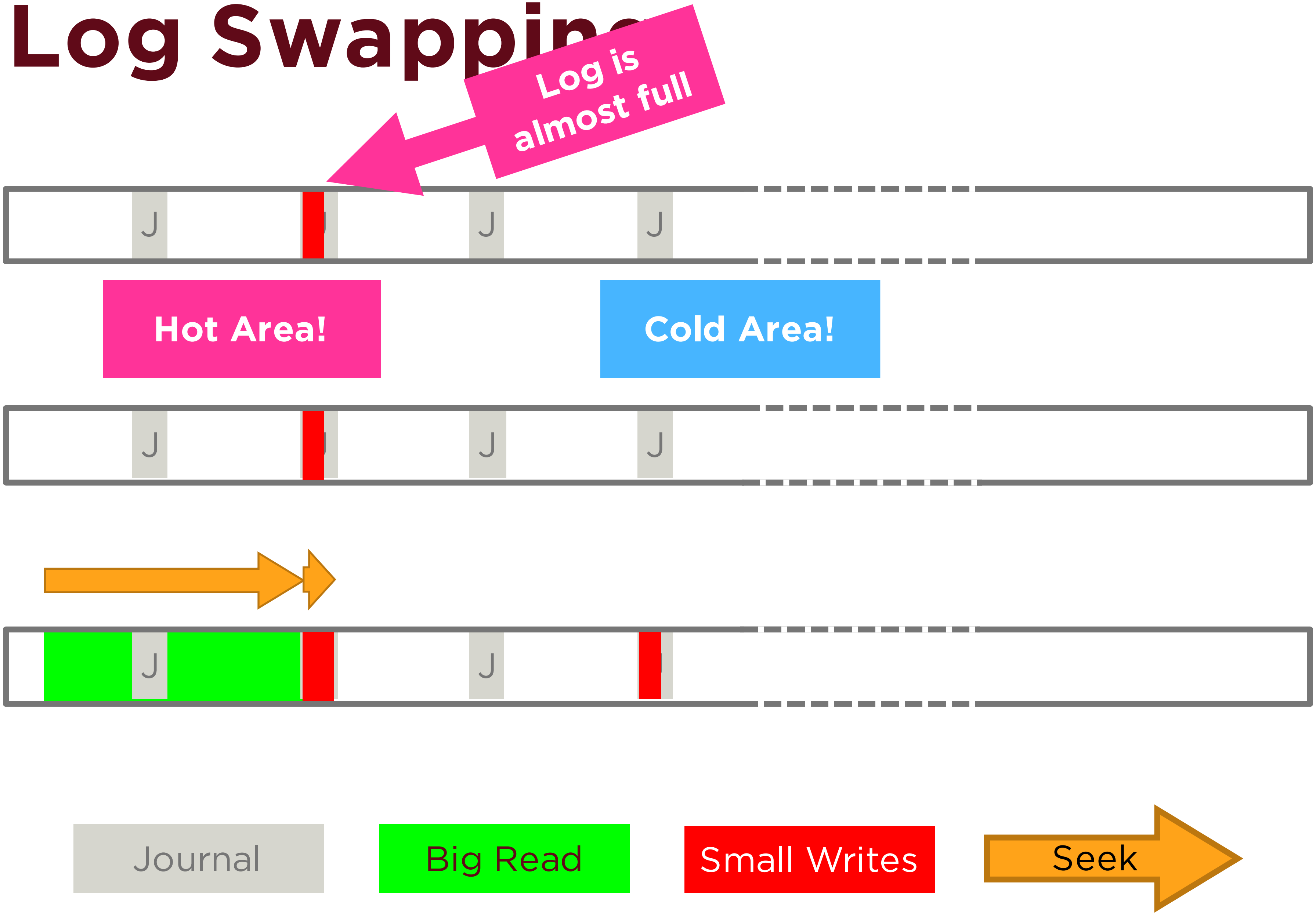
Data Journaling

- 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**

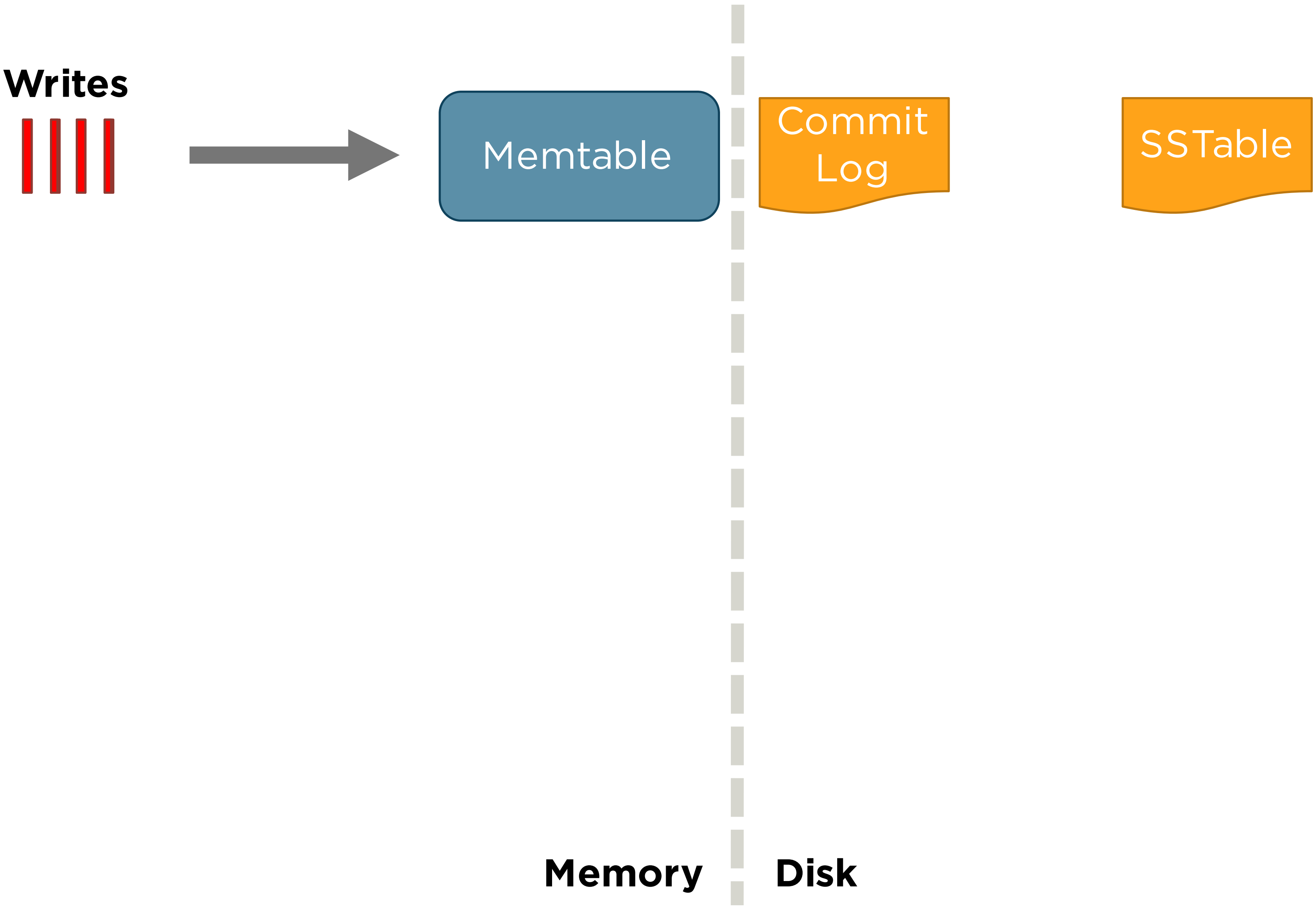
Log Swapping



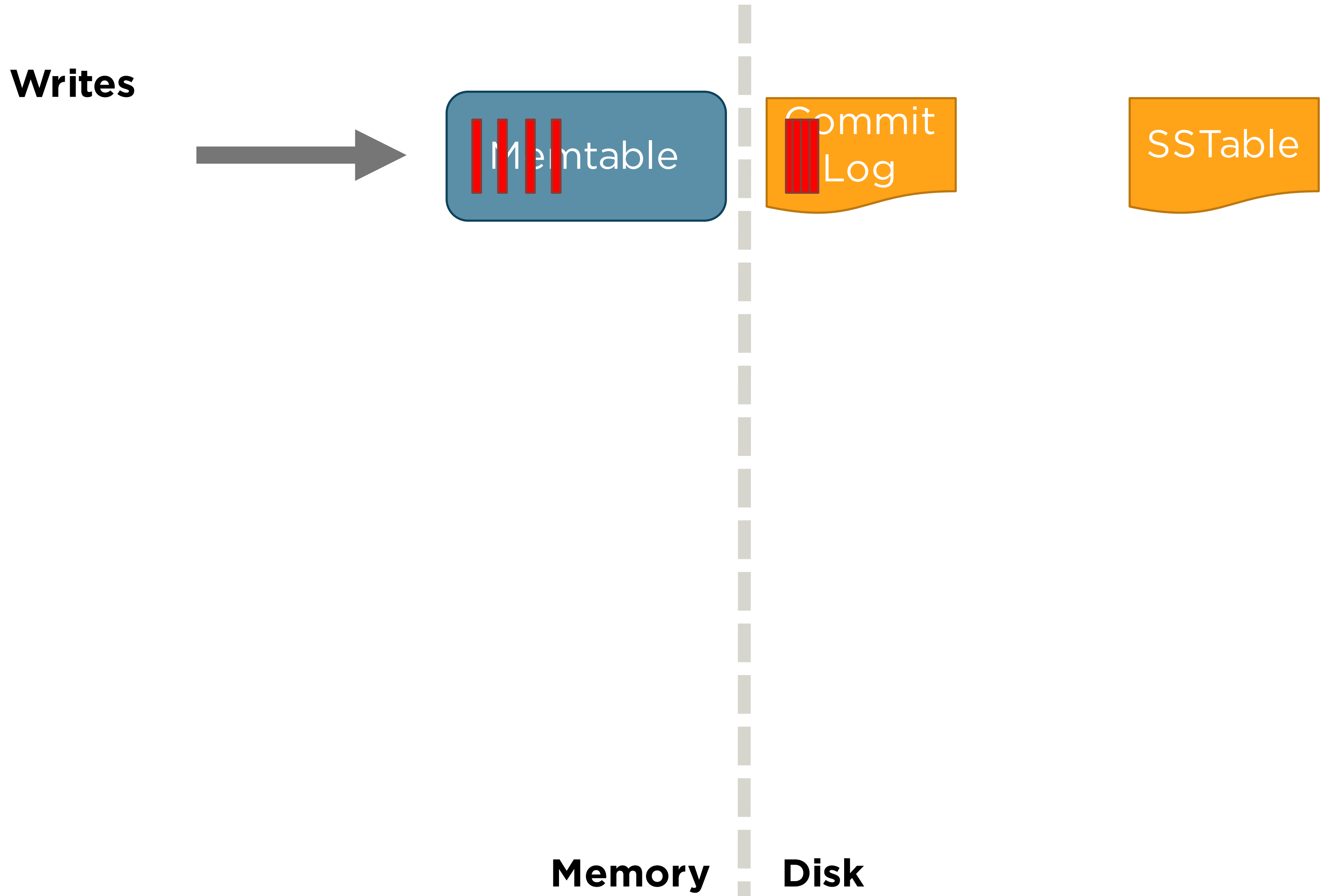
Integrations

- File System (MLFS)
  - Durability-Only Mode (O_DUR)
- SMR Disk (MLSMR)
- RAID

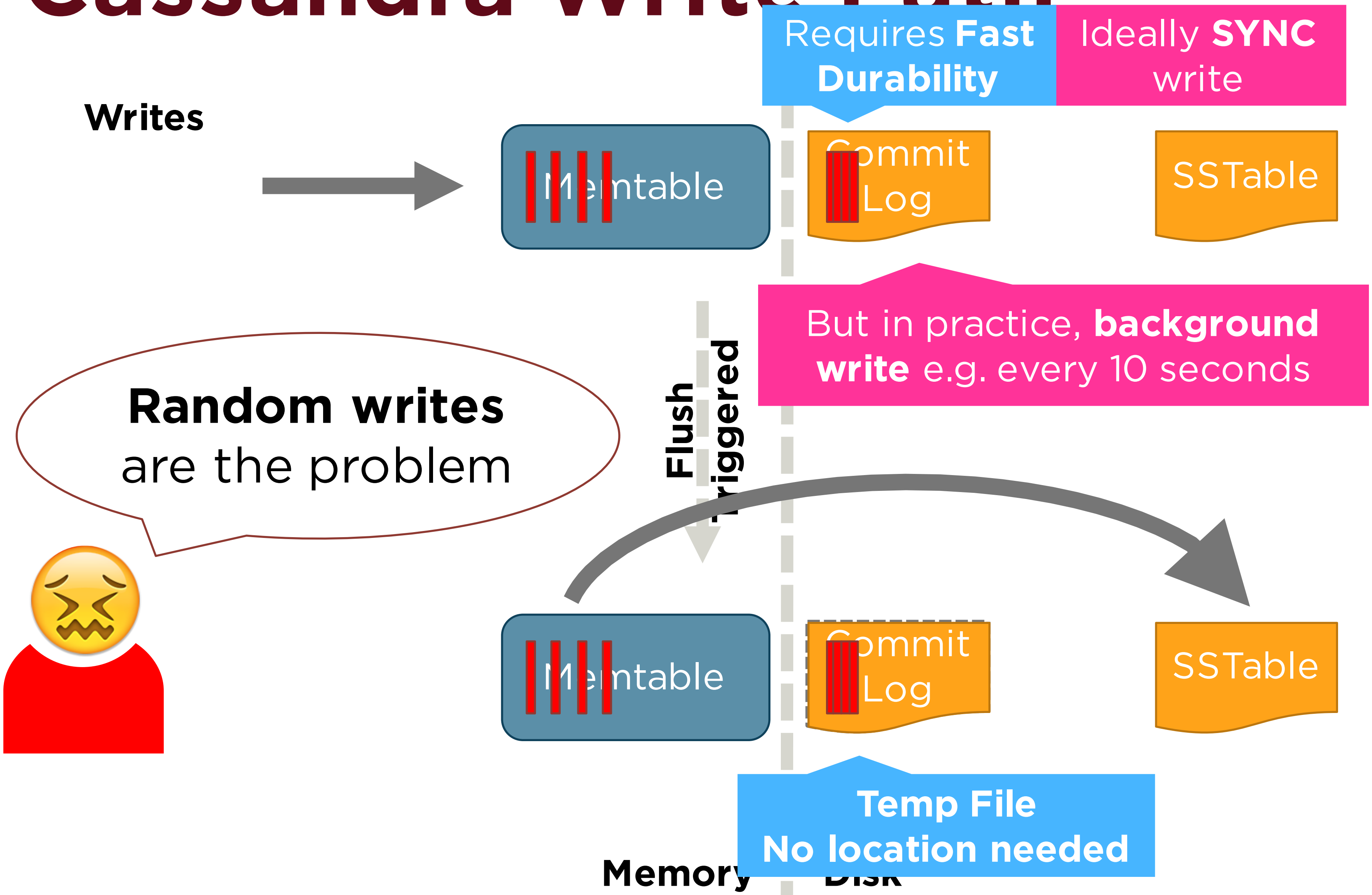
Cassandra Write Path



Cassandra Write Path

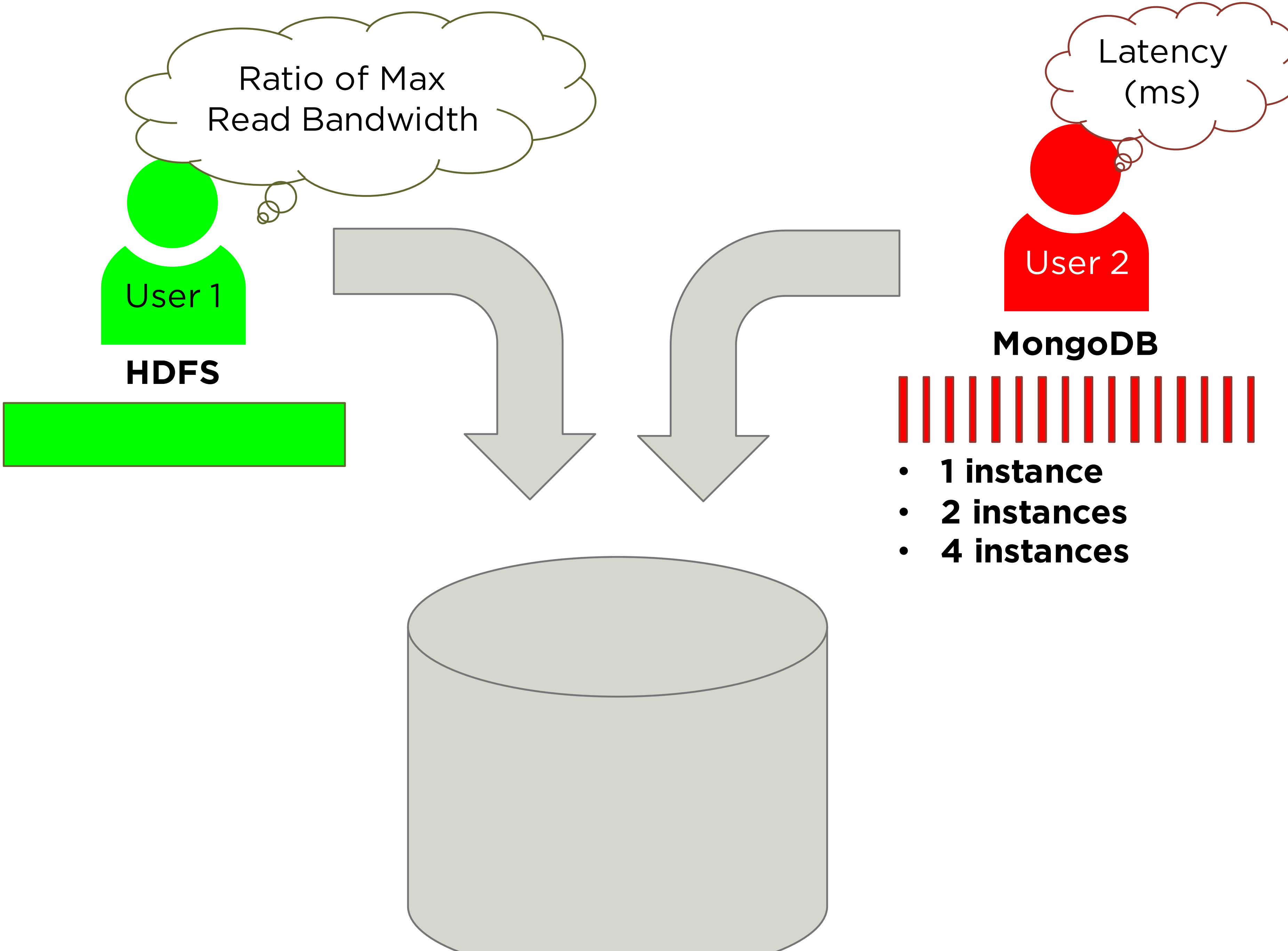


Cassandra Write Path

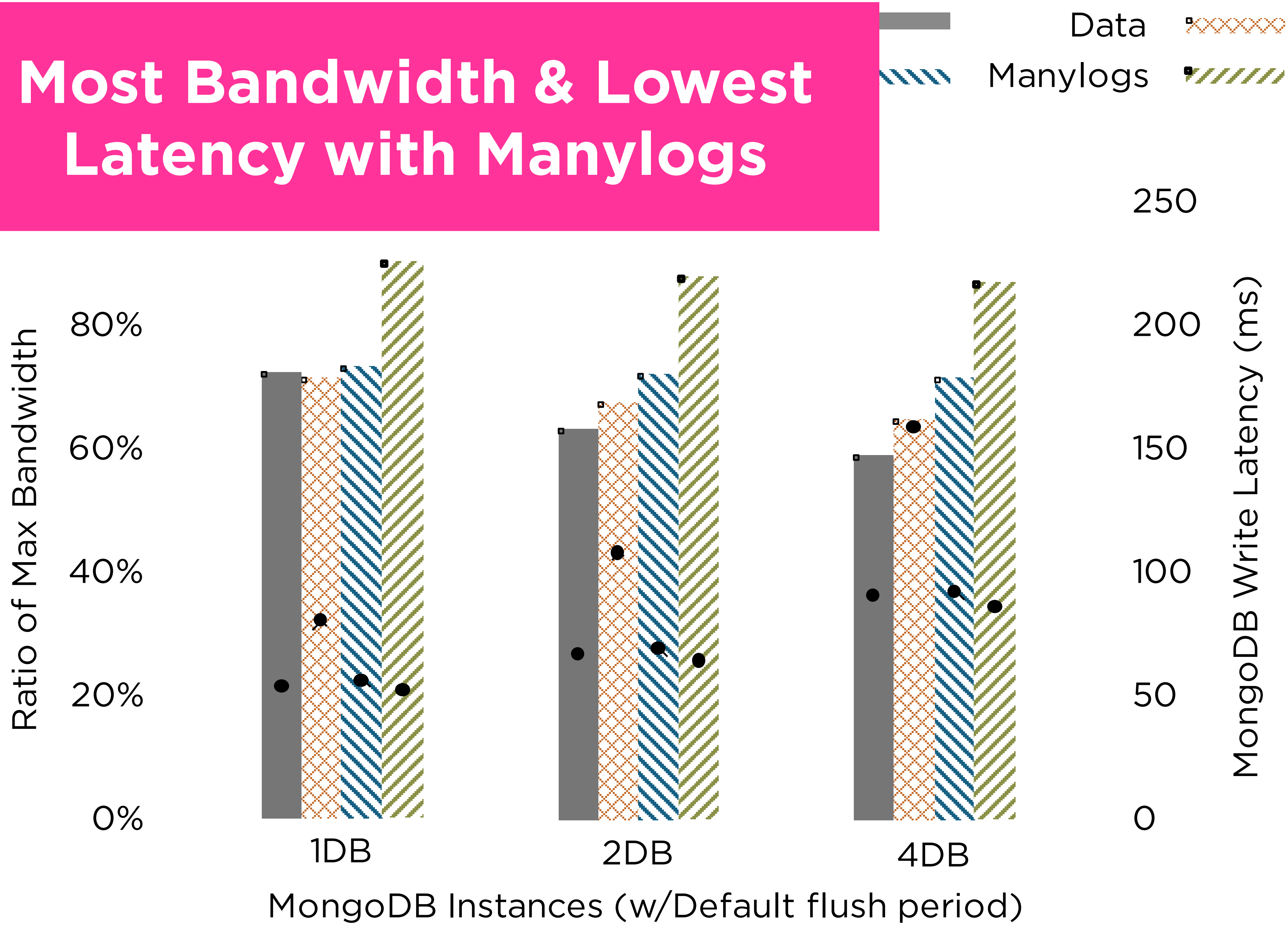


open(file, O_DUR);

- 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!**


Ratio of Max Read Bandwidth
User 1
HDFS
Latency (ms)
User 2
MongoDB
• 1 instance
• 2 instances
• 4 instances

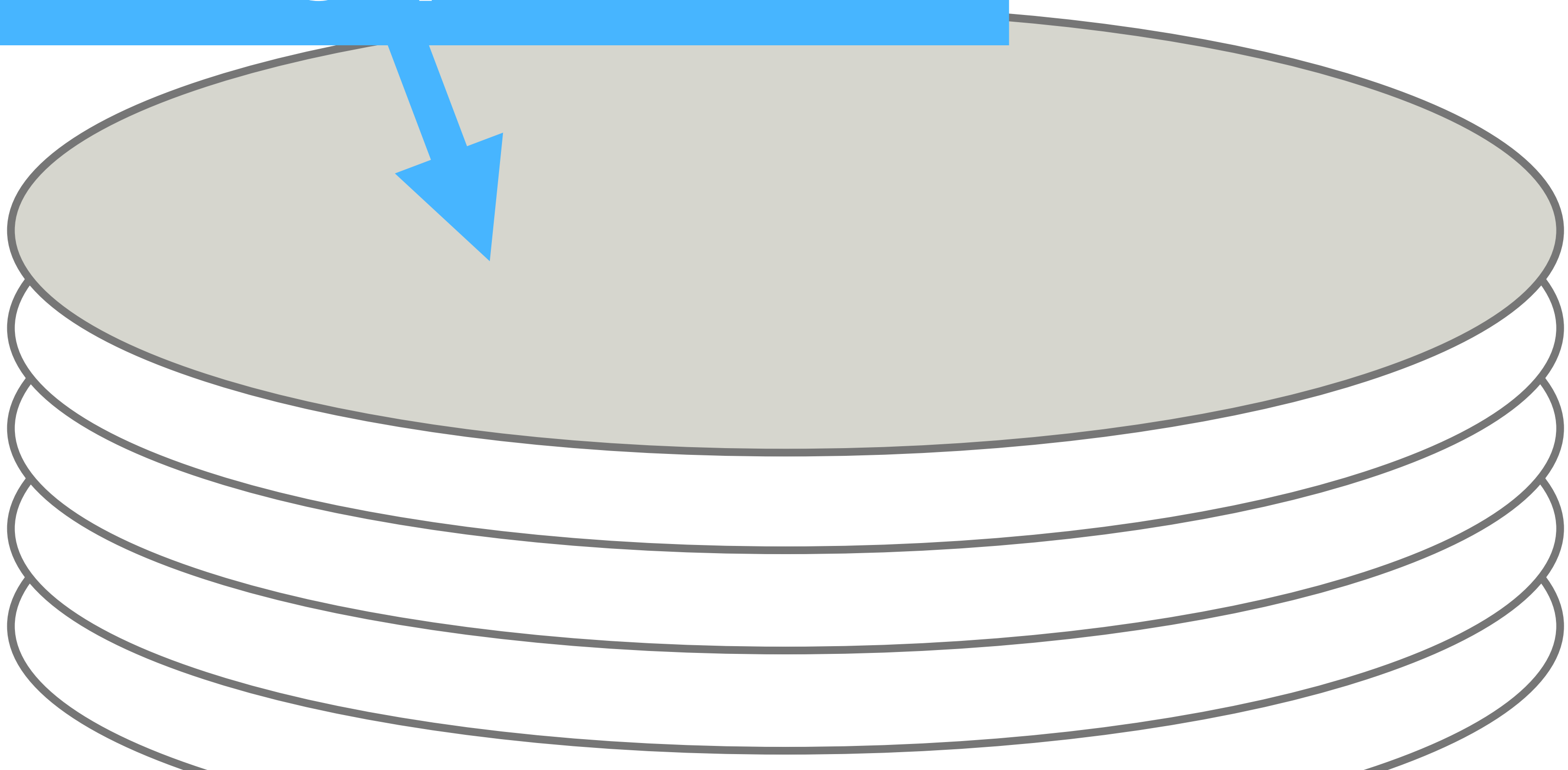
Most Bandwidth & Lowest Latency with Manylogs



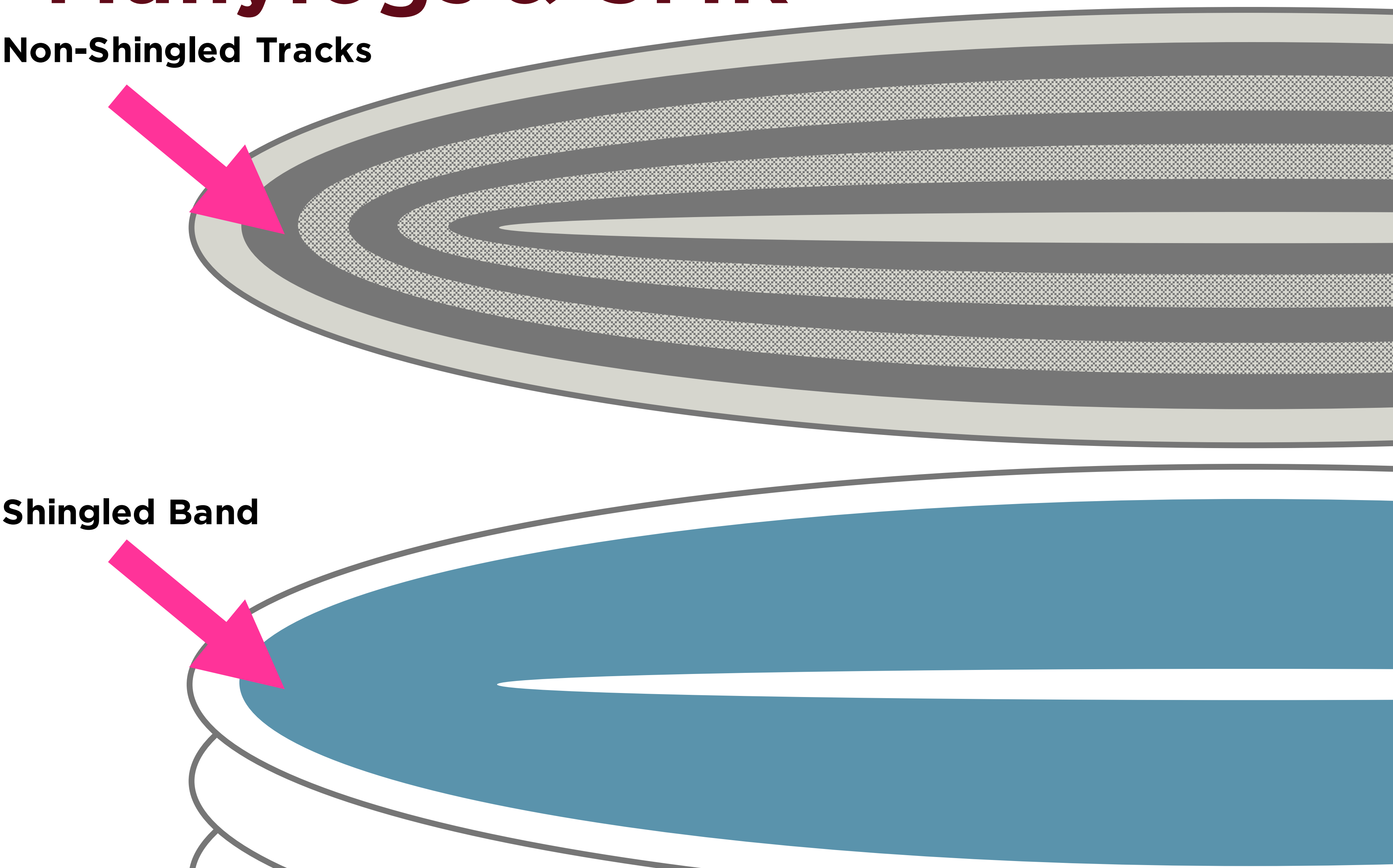
Manylogs & SMR

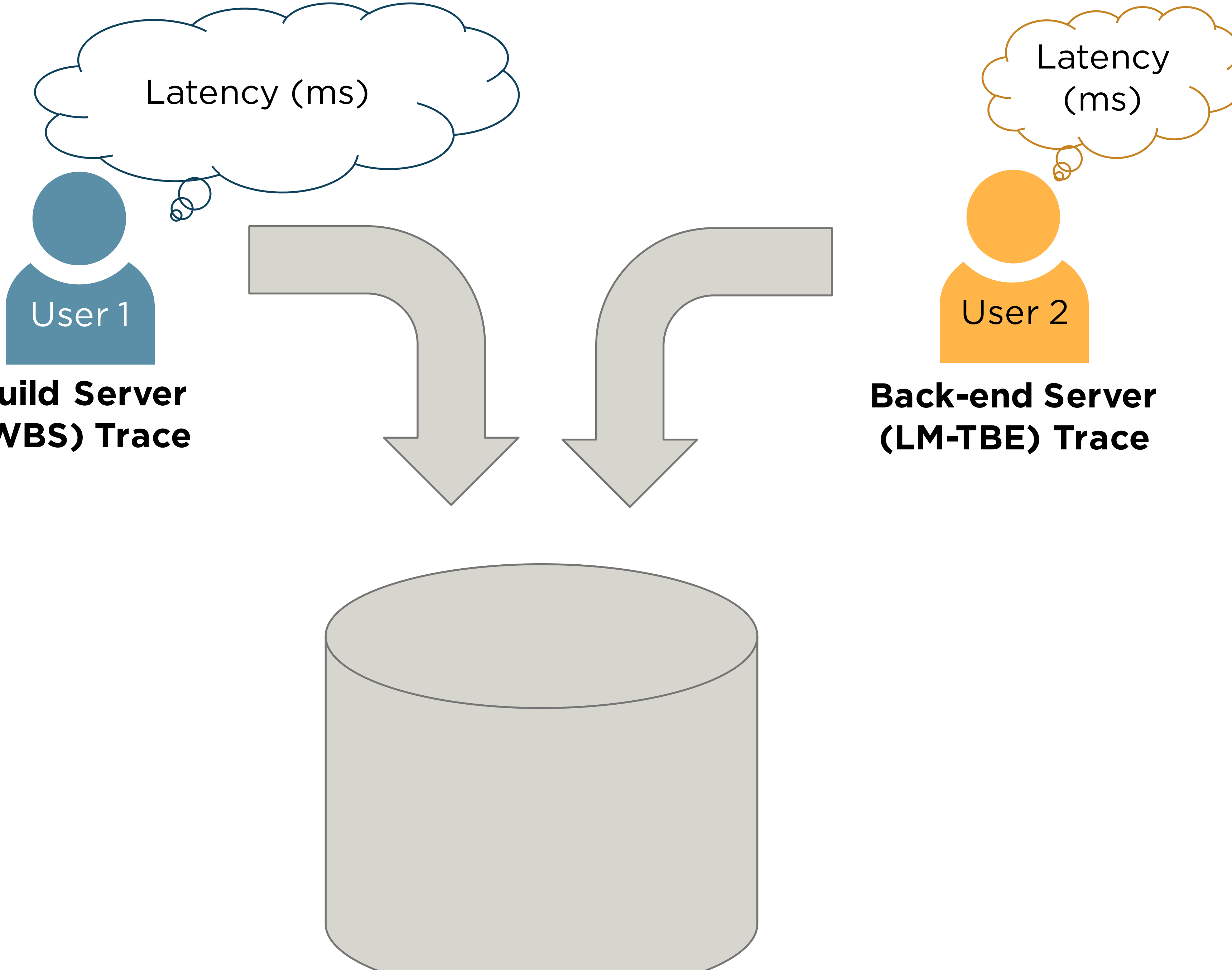
One non-shingled surface

= log space



Manylogs & SMR

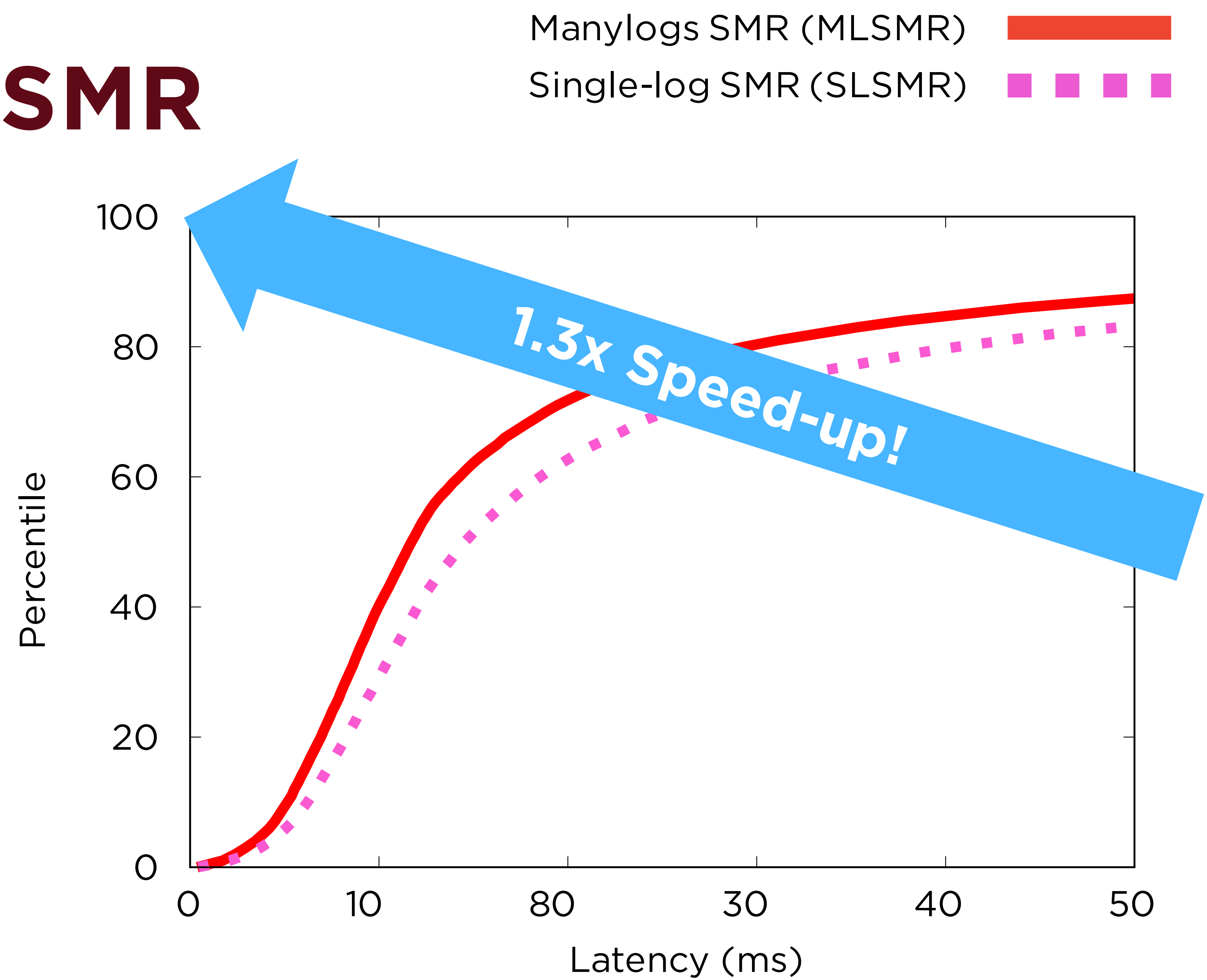




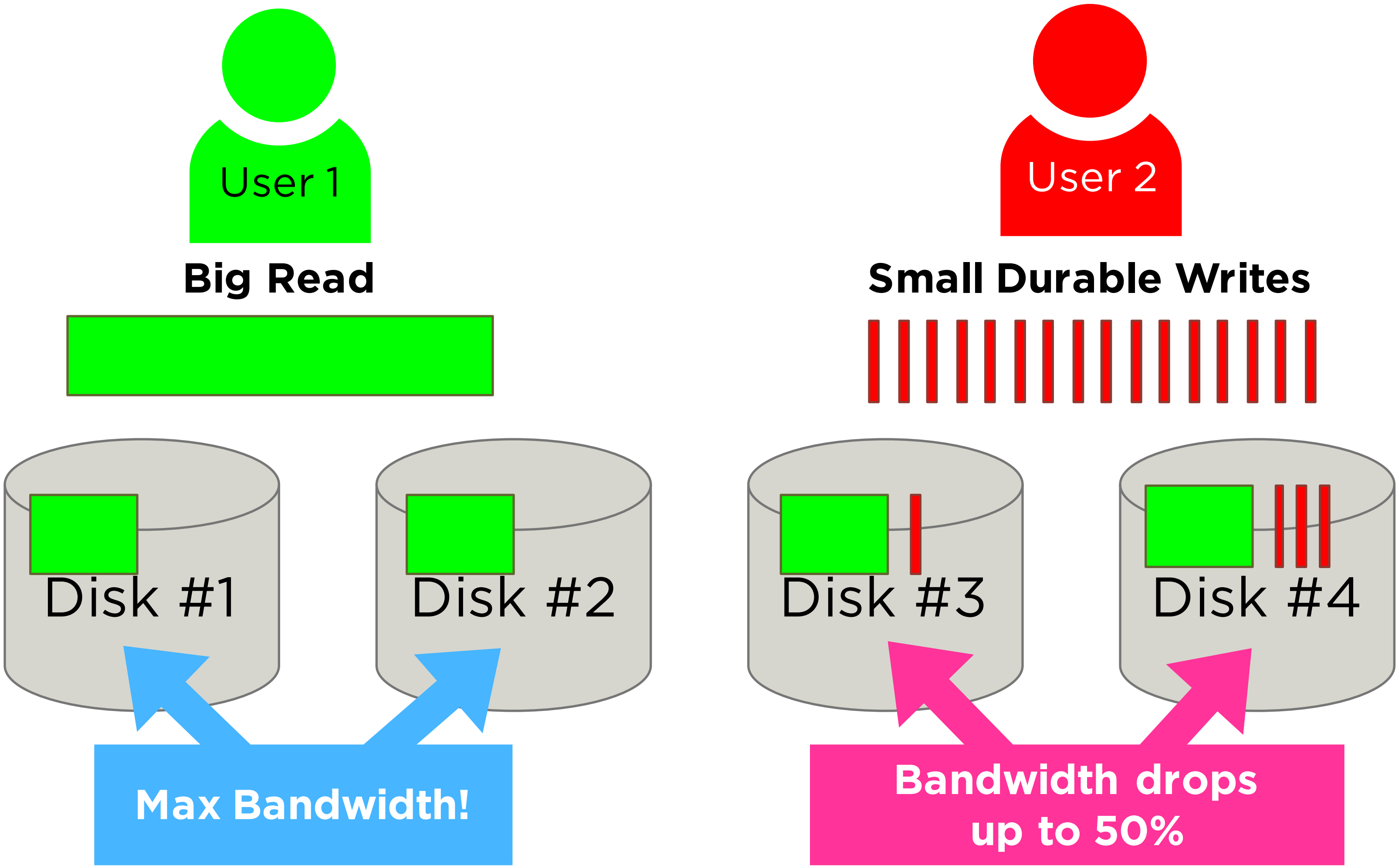
Build Server (WBS) Trace

Back-end Server (LM-TBE) Trace

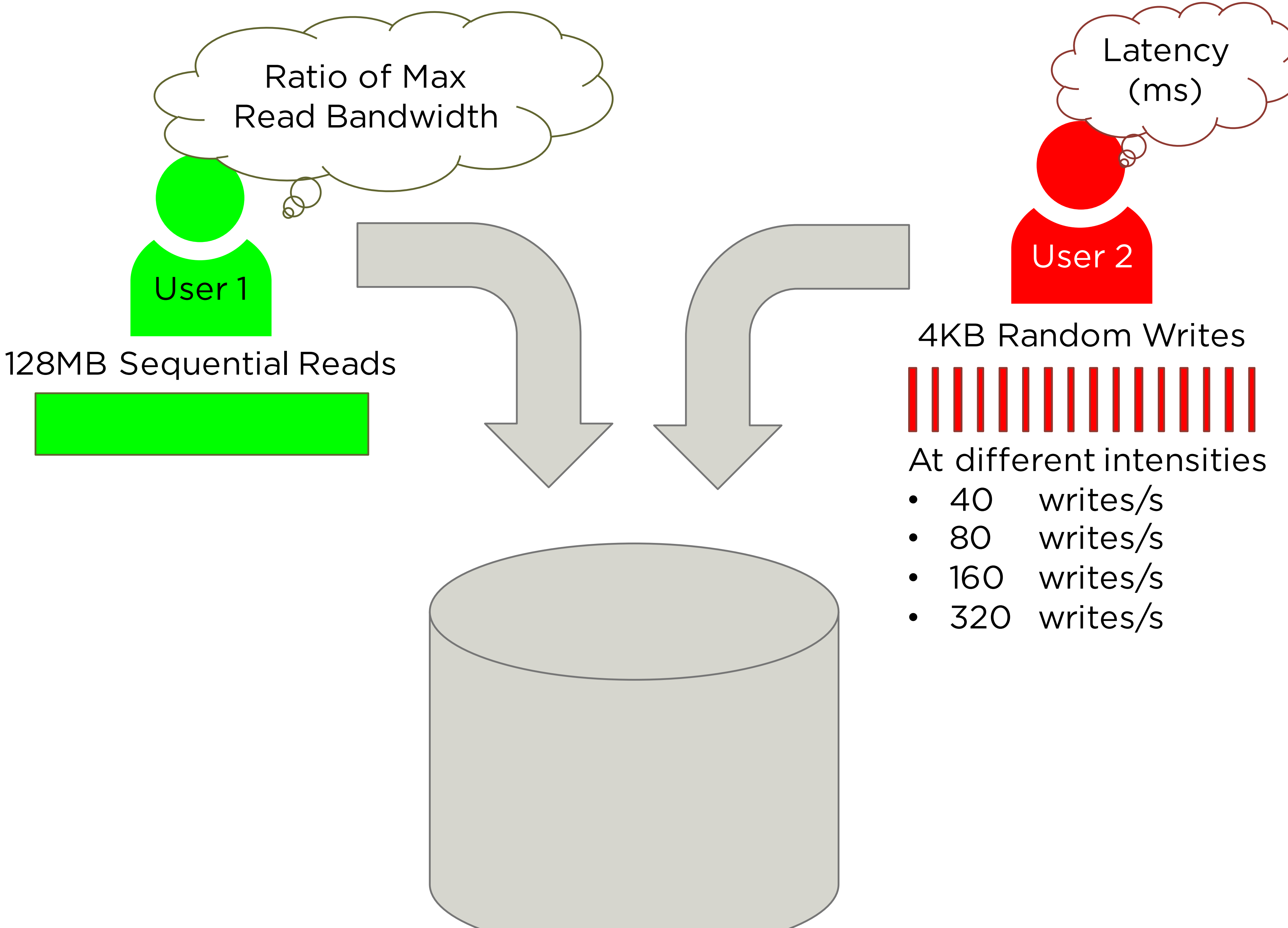
SMR


Manylogs SMR (MLSMR)
Single-log SMR (SLSMR)
1.3x Speed-up!
Percentile
100
80
60
40
20
0
0
10
80
30
40
50
Latency (ms)

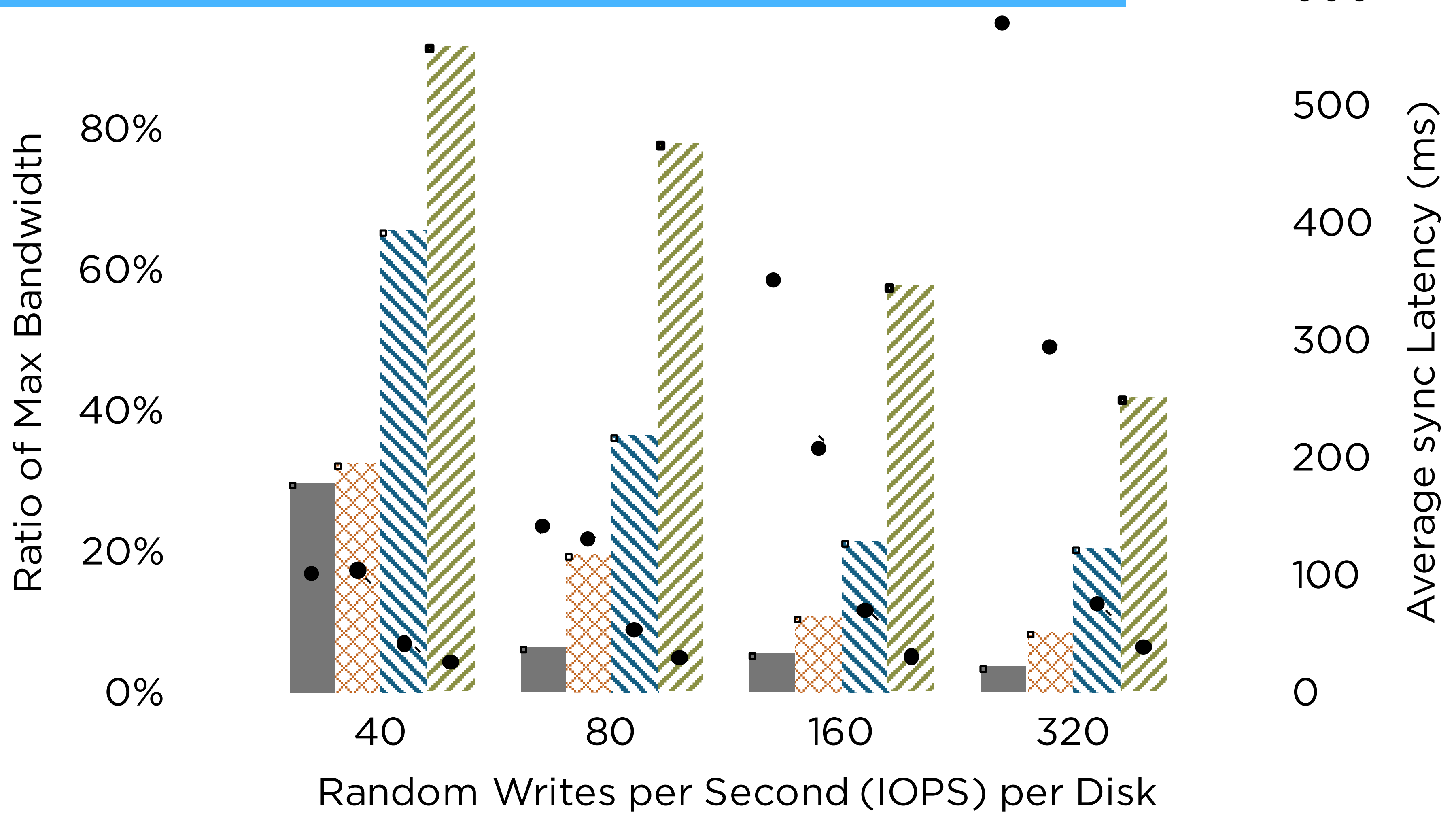
Manylogs & RAID



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.


Ratio of Max
Read Bandwidth
User 1
128MB Sequential Reads
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!

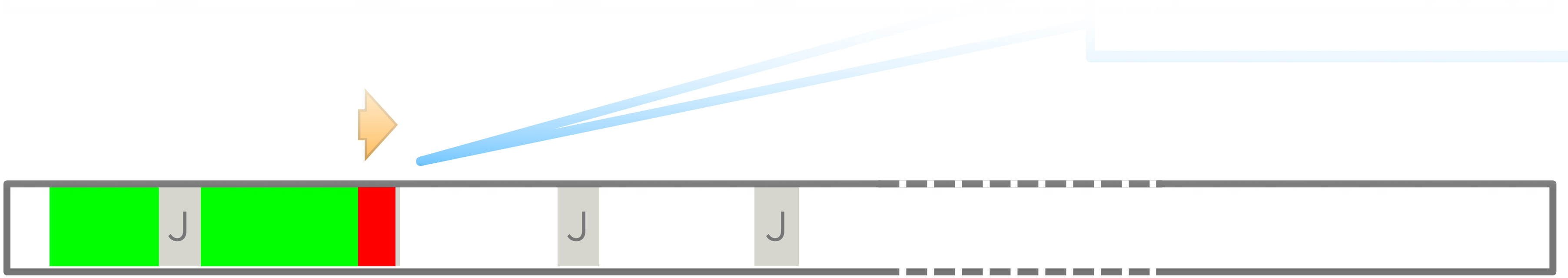


More in the paper

- 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

	Bandwidth Speed-up	Latency Speed-up
vs. Ordered	3.7x	5.7x
vs. Adaptive	2.7x	2.0x
vs. Single-log SMR		1.3x

Thank you!
Questions?

UCARE

THE UNIVERSITY OF CHICAGO

http://ucare.cs.uchicago.edu