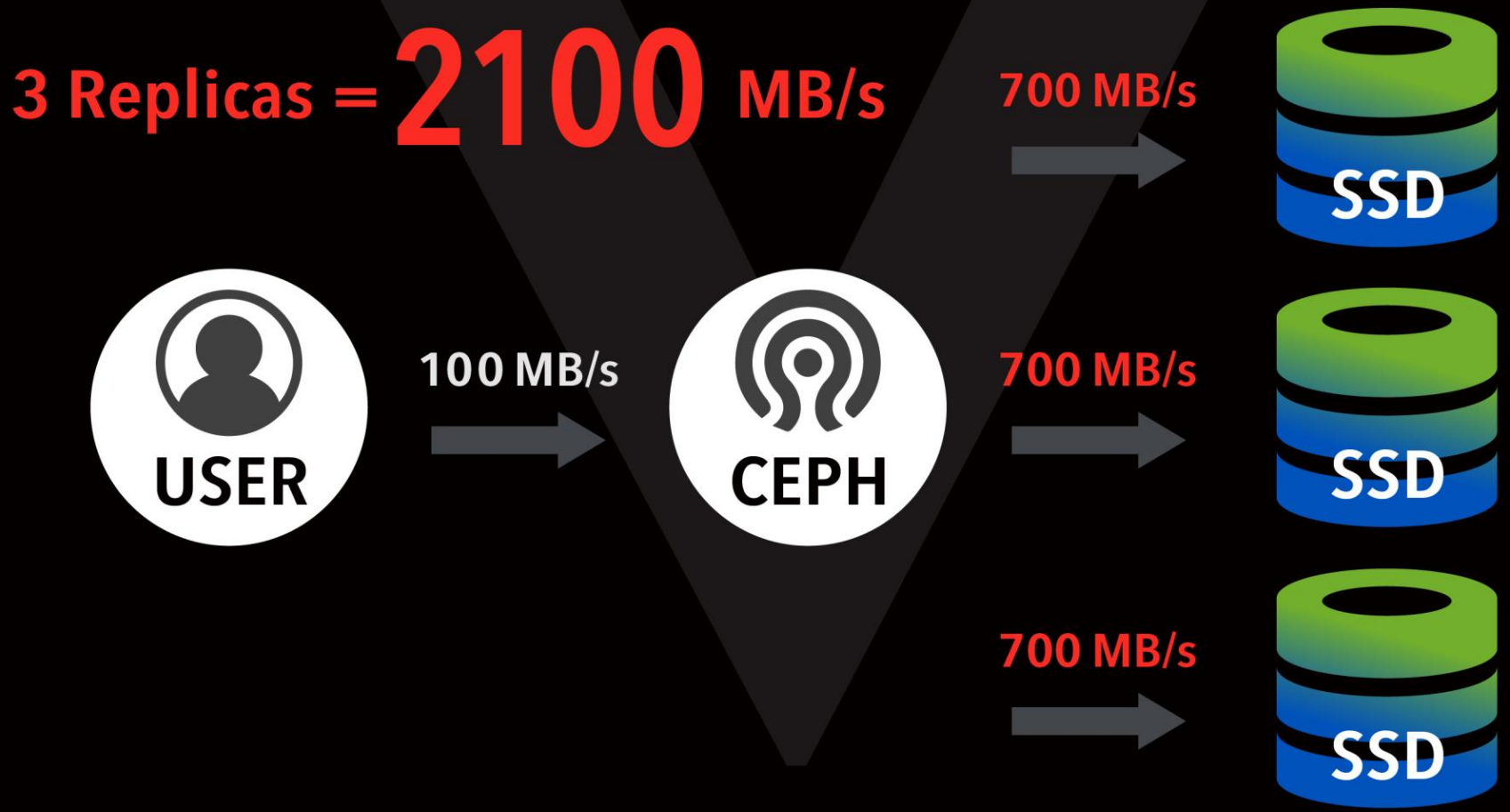
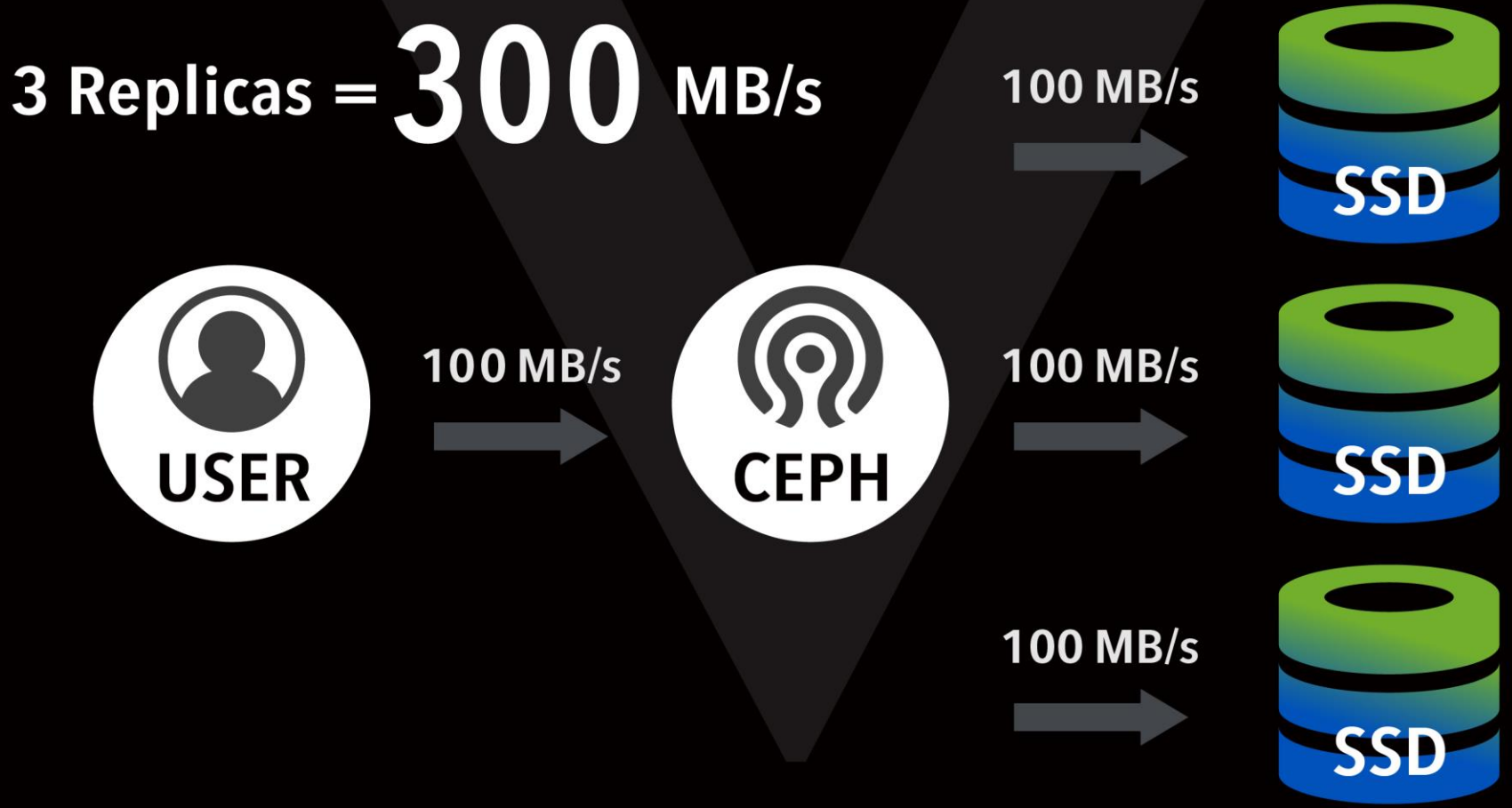
NVDIMM
The Savior of SSD Endurance in CEPH

David Tseng

Bigtera Inc. | Flash Memory Summit 2018


USER
100 MB/s
CEPH
700 MB/s
SSD


3 Replicas = 2100 MB/s
700 MB/s
SSD
USER
100 MB/s
CEPH
700 MB/s
SSD
700 MB/s
SSD


3 Replicas = 300 MB/s
USER
100 MB/s
CEPH
100 MB/s
SSD
100 MB/s
SSD
100 MB/s
SSD

Why CEPH ?

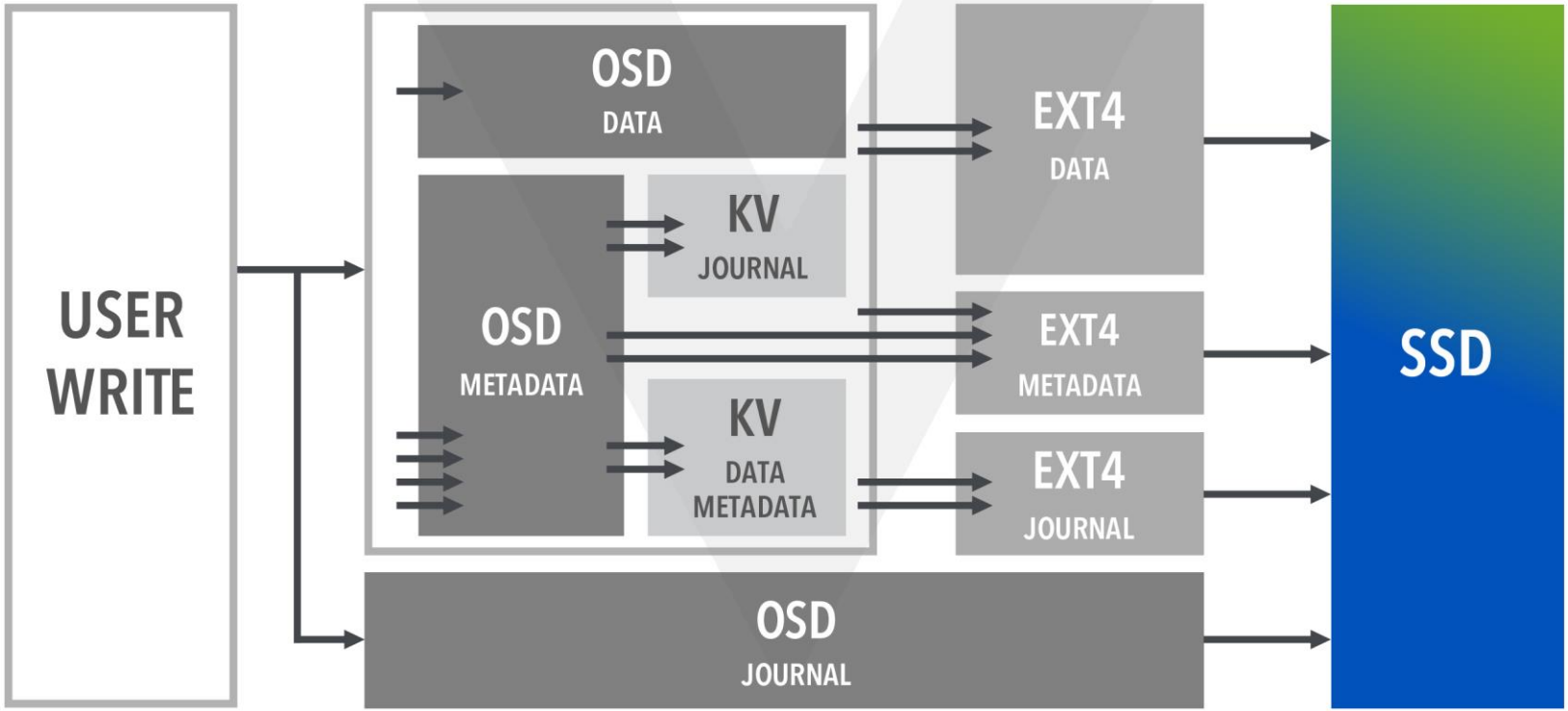
UNIFIED STORAGE

- Object
- Block
- Filesystem

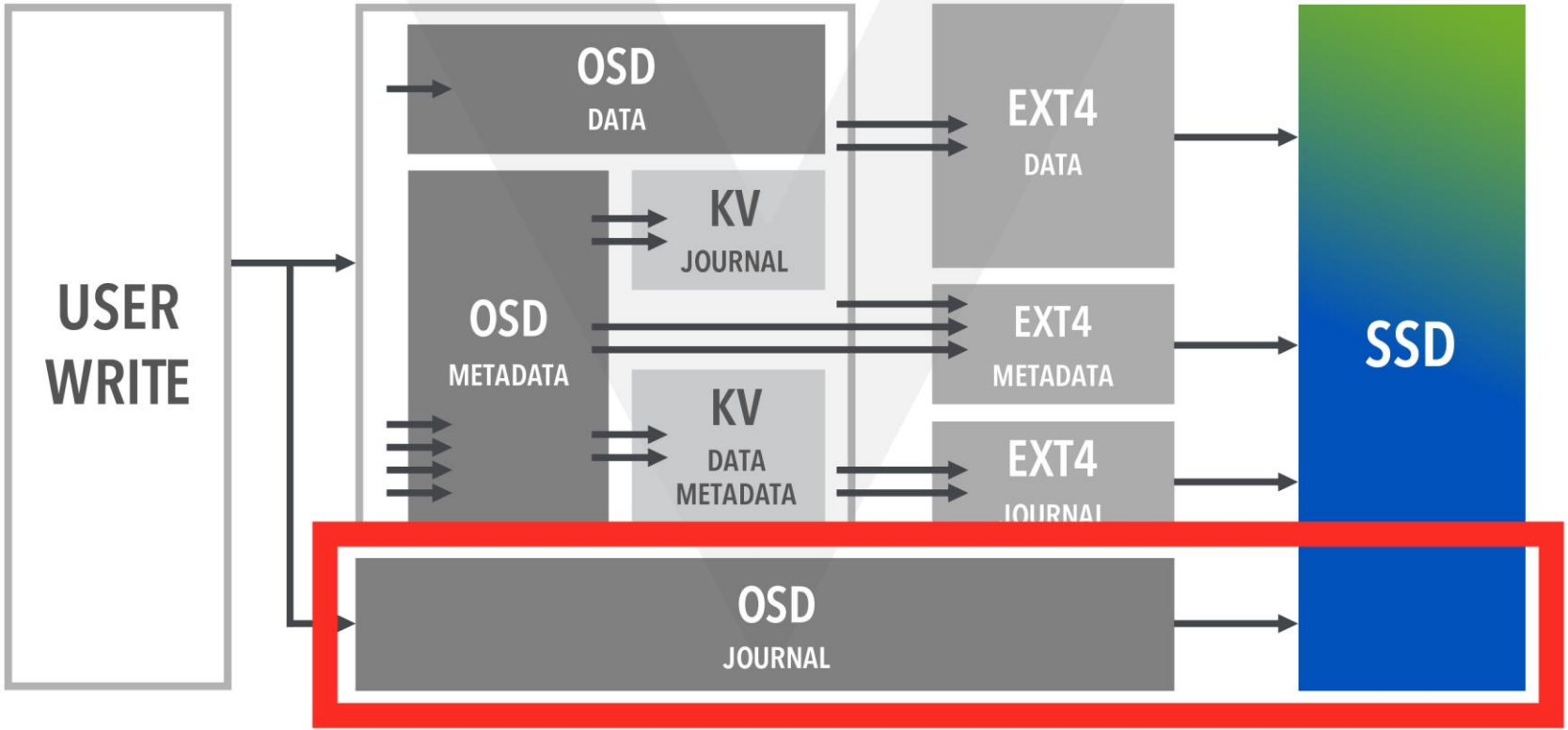
SCALE-OUT

- Pay as you grow
- No single point of failure

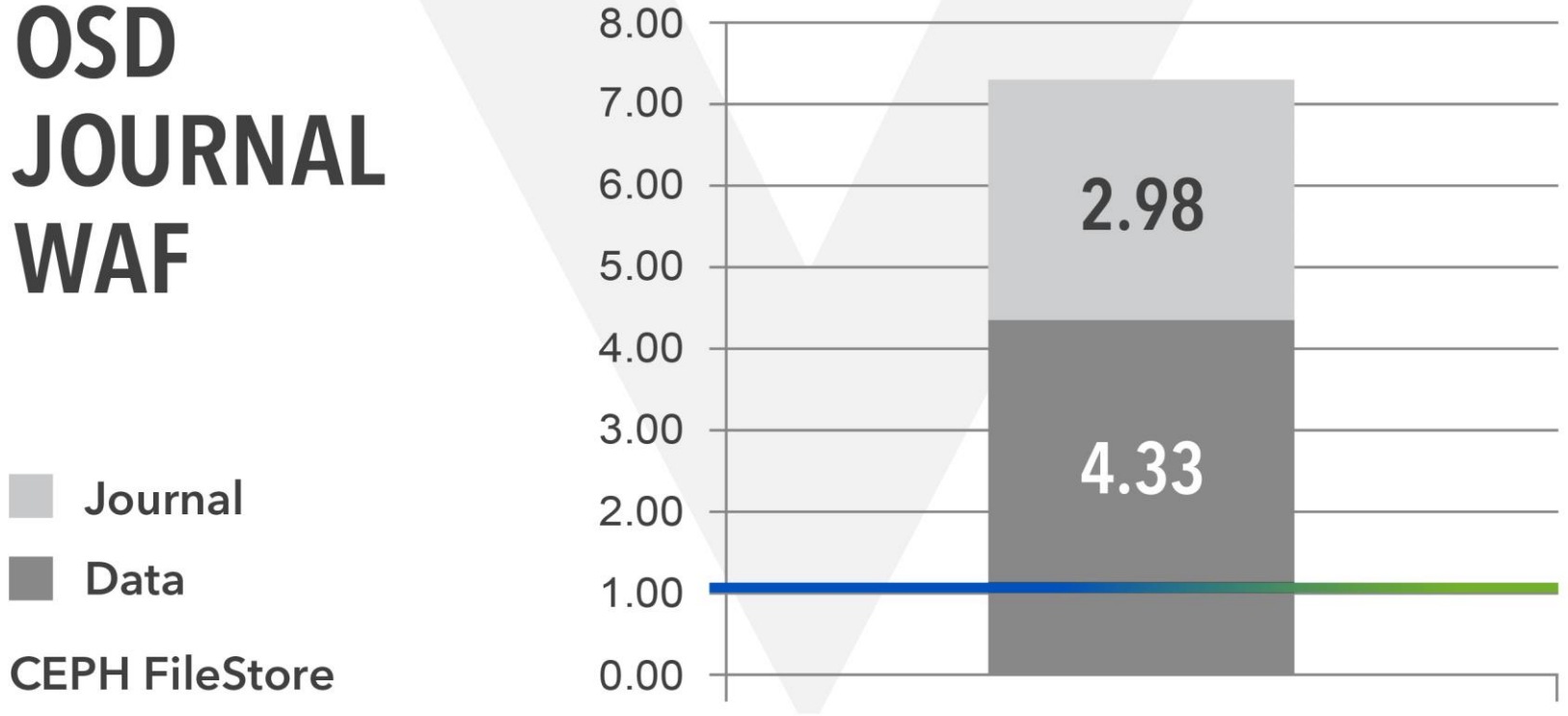
The Source of Write Amplifications



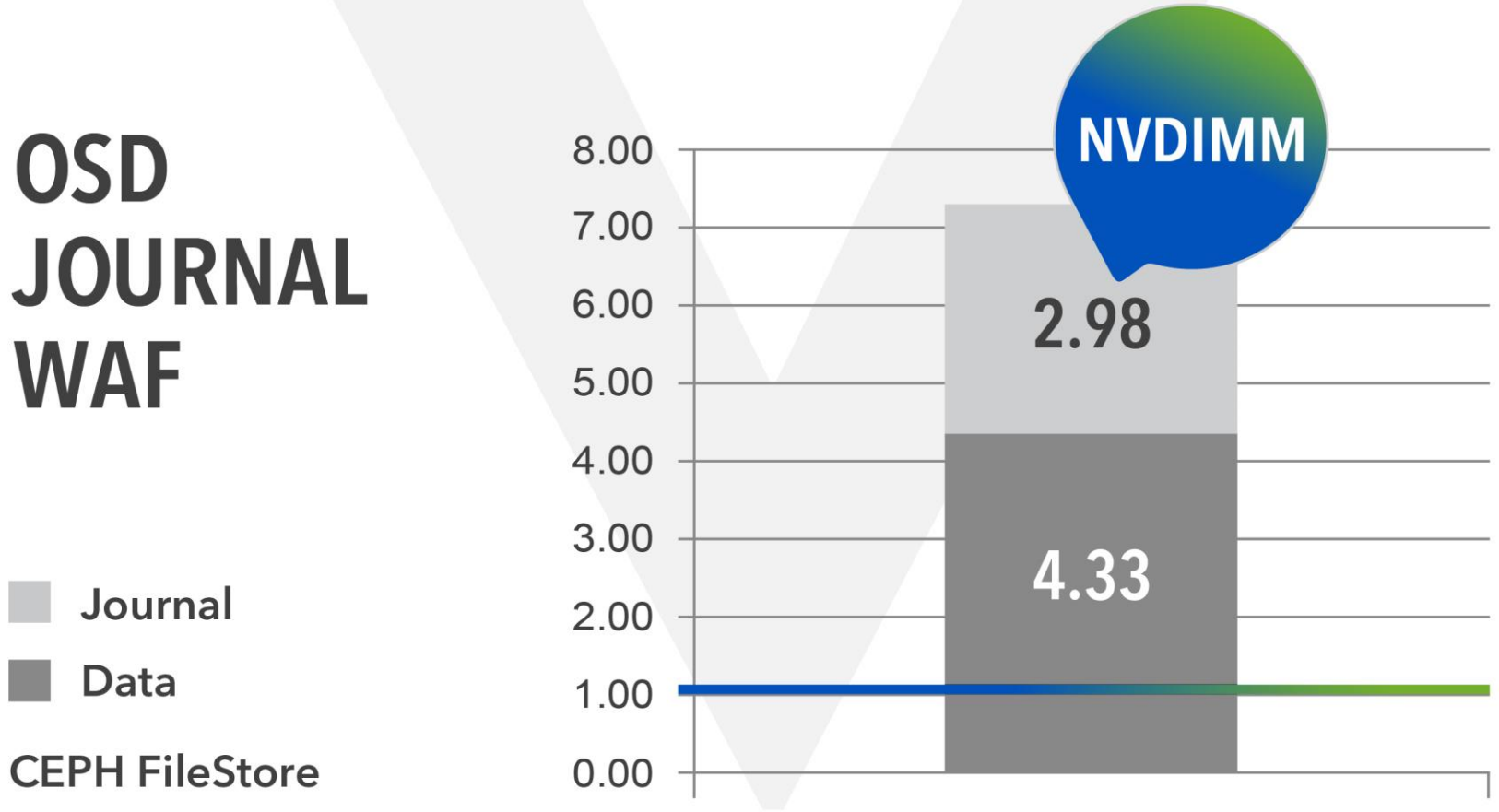
OSD JOURNAL

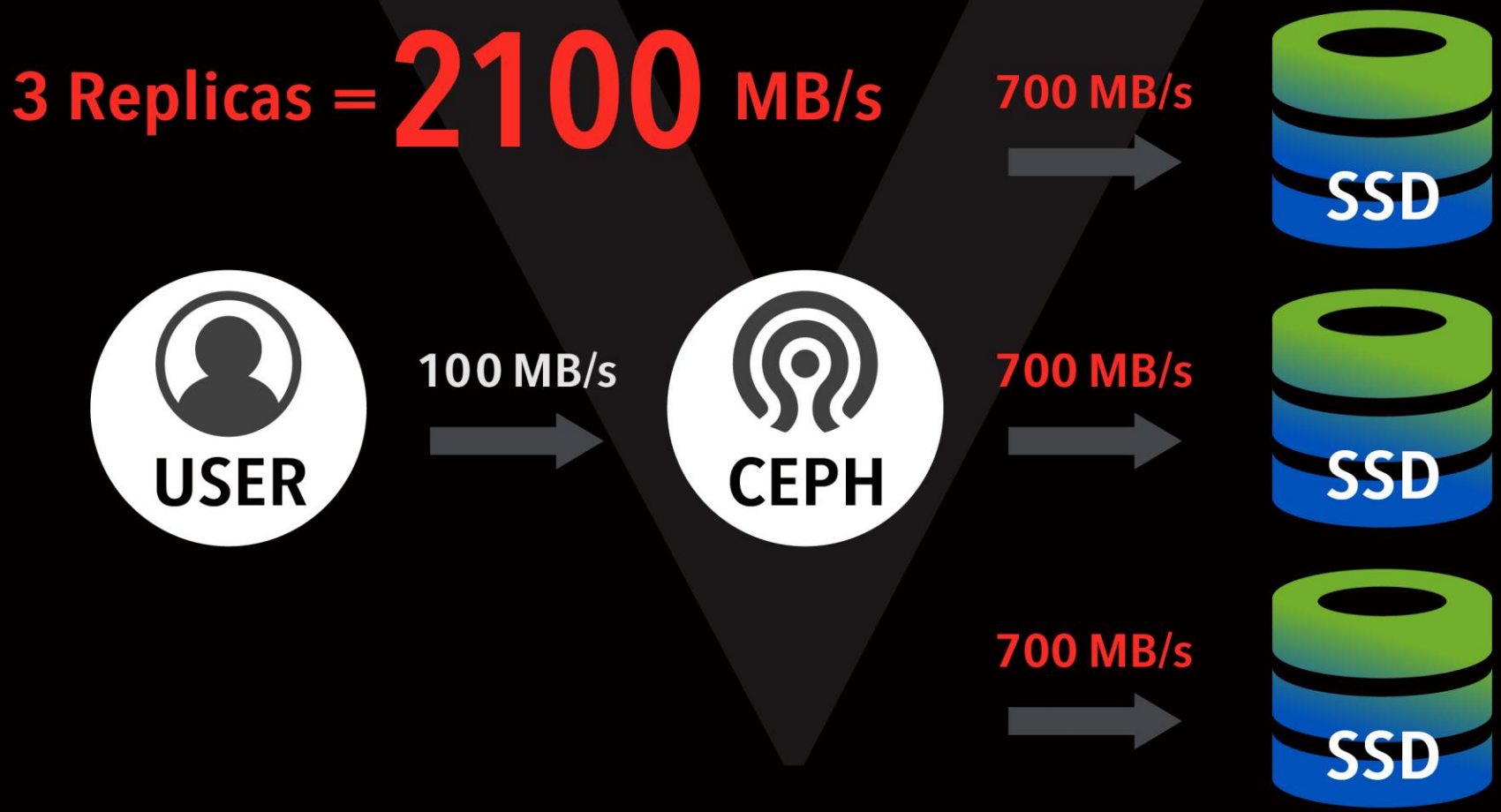

USER WRITE
OSD DATA
OSD METADATA
KV JOURNAL
KV DATA METADATA
OSD JOURNAL
EXT4 DATA
EXT4 METADATA
EXT4 JOURNAL
SSD

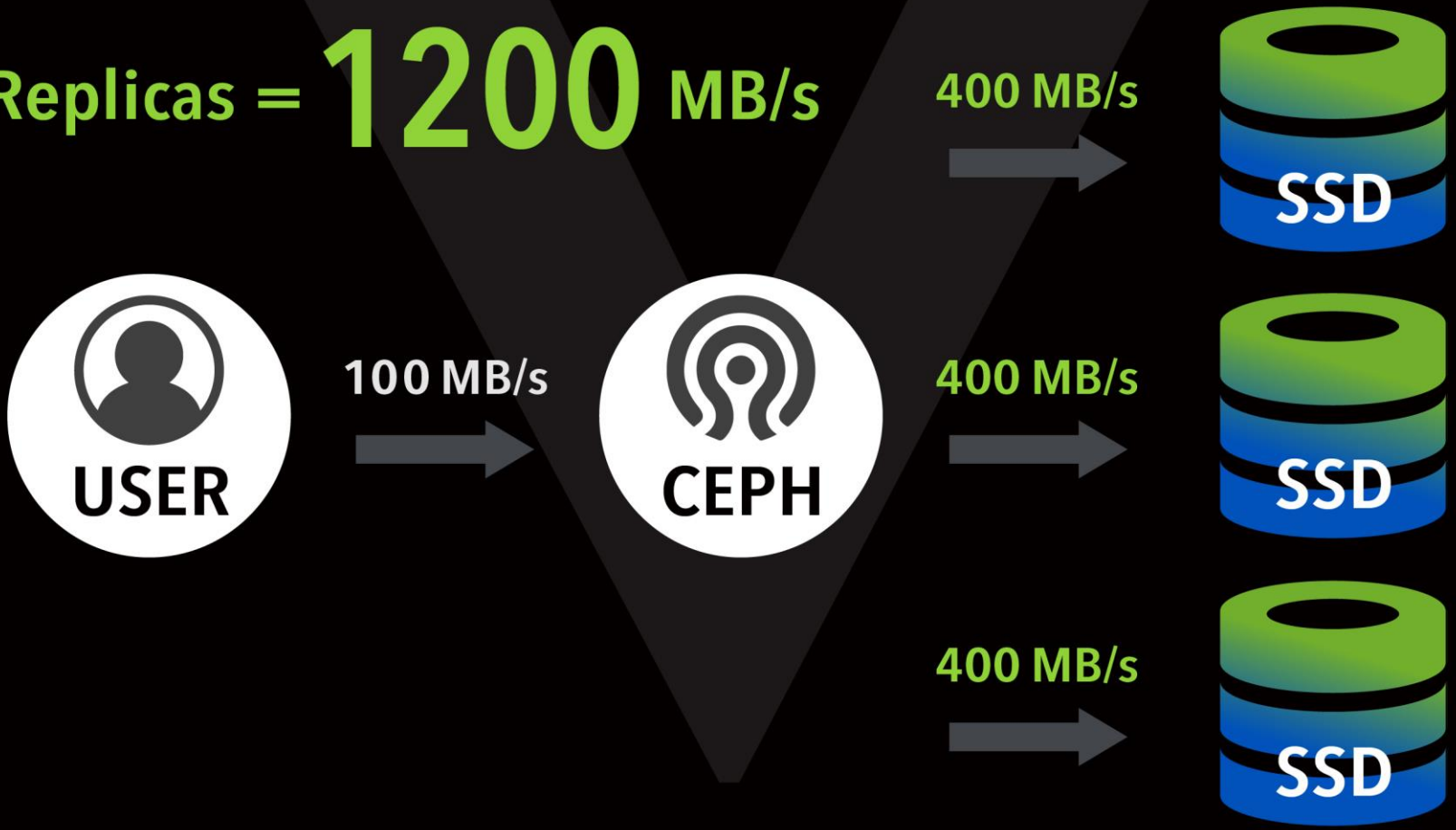
OSD JOURNAL WAF



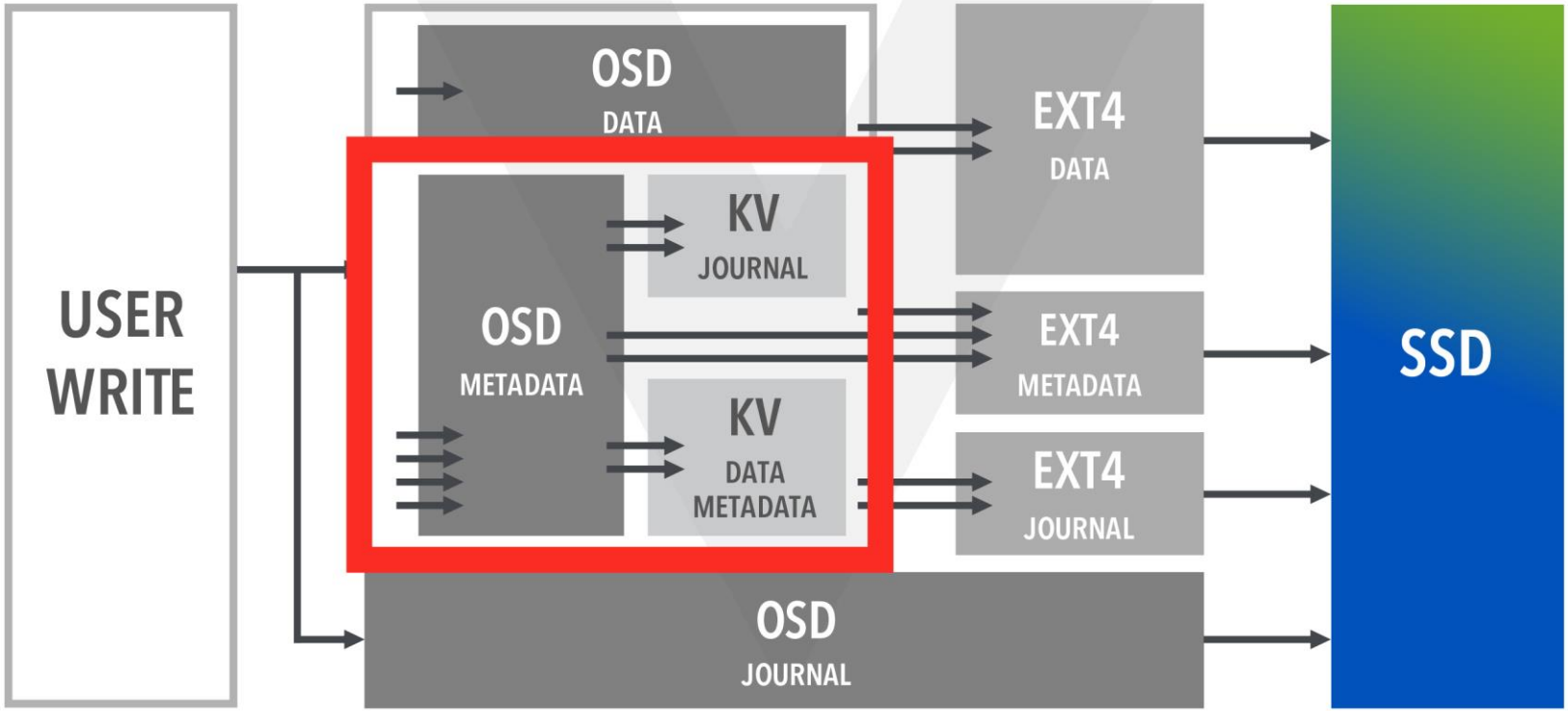
OSD JOURNAL WAF




3 Replicas = 2100 MB/s
700 MB/s
SSD
USER
100 MB/s
CEPH
700 MB/s
SSD
700 MB/s
SSD

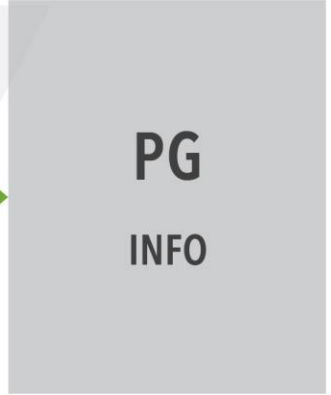

3 Replicas = 1200 MB/s
400 MB/s
SSD
USER
100 MB/s
CEPH
400 MB/s
SSD
400 MB/s
SSD

OSD METADATA



METADATA FOR EACH WRITE

Re-writing Same Key



METADATA STORAGE

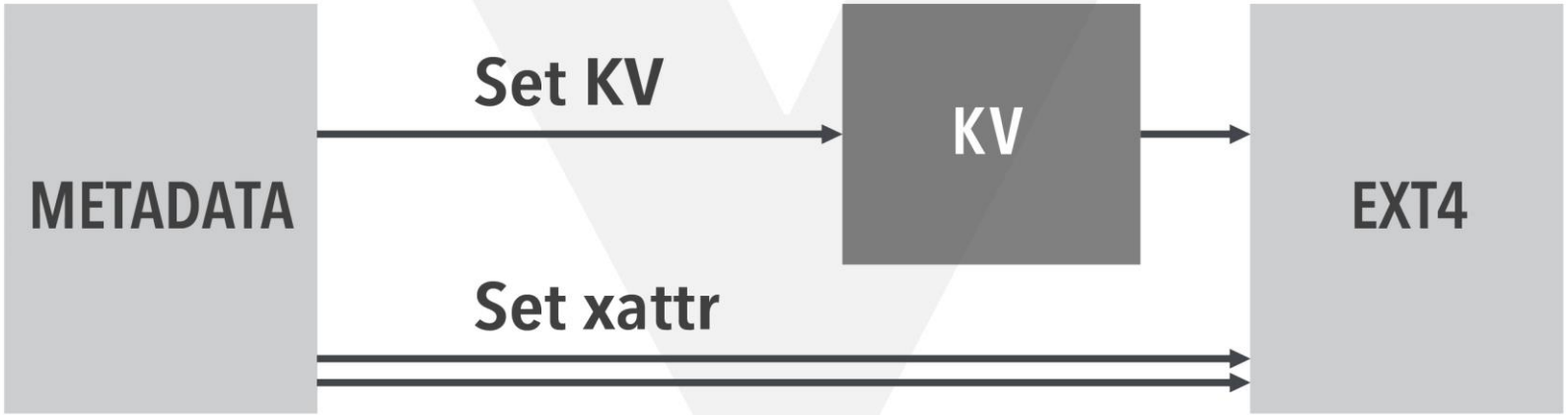
Limited length
Write amplification on Disk

Compaction overhead
Write amplification on Disk

Stores hot metadata
Minimal writes on Disk

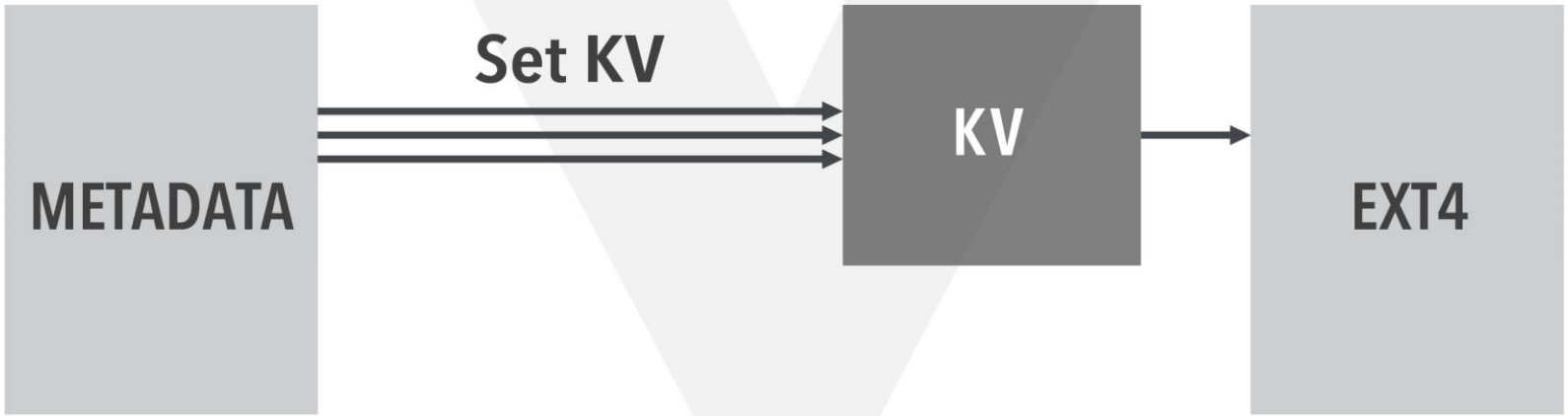
OSD METADATA

Re-writing same KV pair



OSD METADATA

Re-writing same KV pair

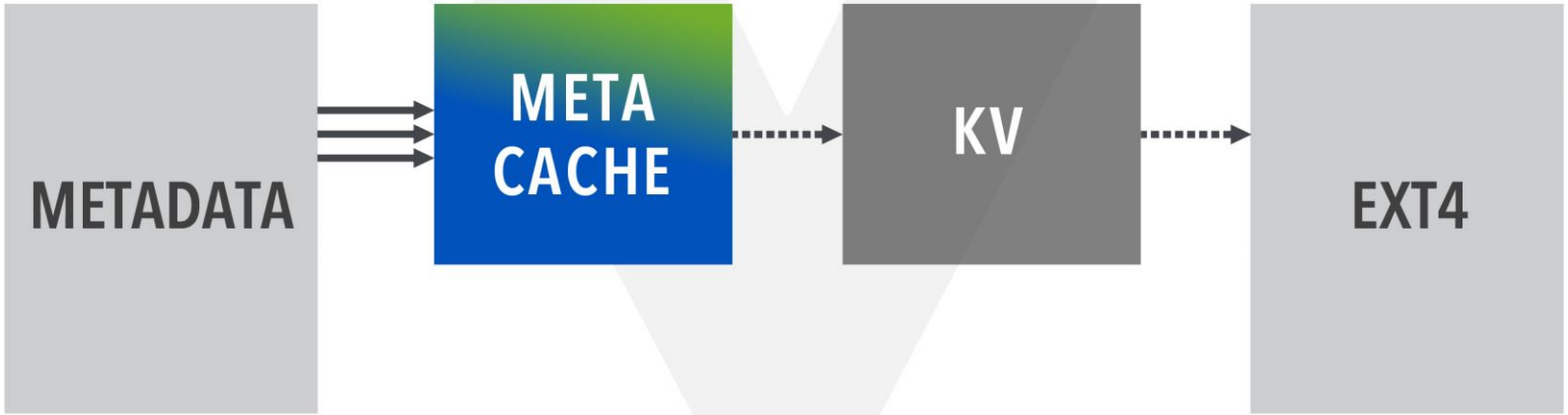


More Merge

Less EXT4-J Lock Contention

OSD METADATA

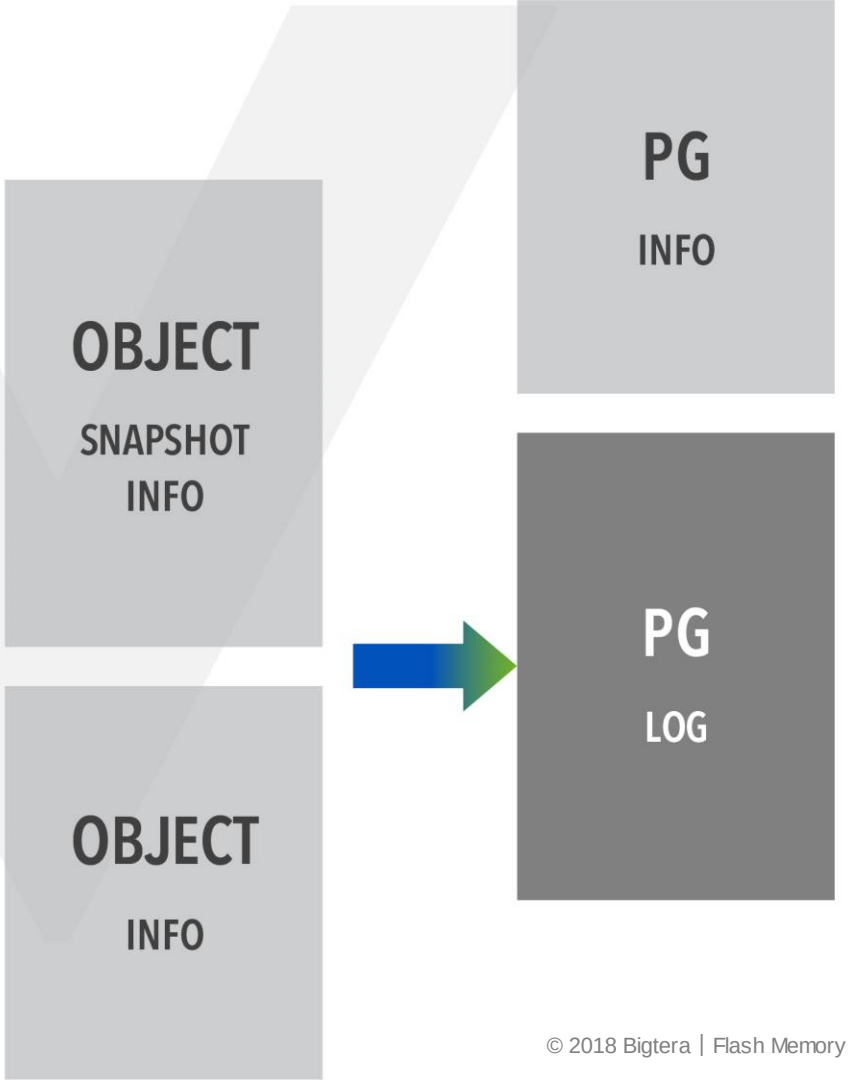
Re-writing same KV pair



Store Metadata in NVDIMM
Flush when
(Idle & High cache pressure)

METADATA FOR EACH WRITE

New Log for Each Write



OSD METADATA

PGLog Traditional behavior

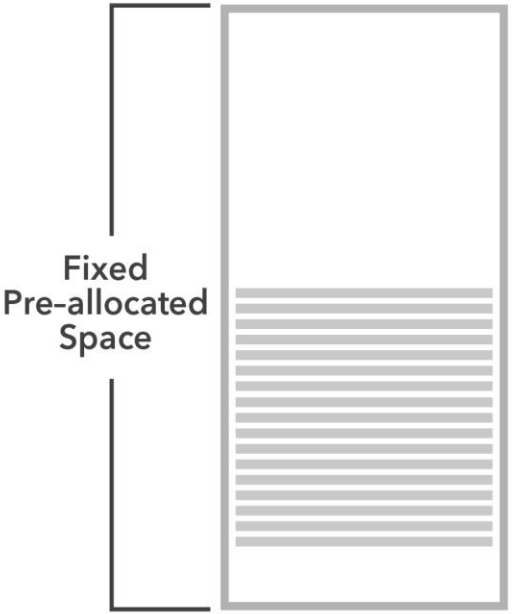
OSD METADATA

PGLog Traditional behavior

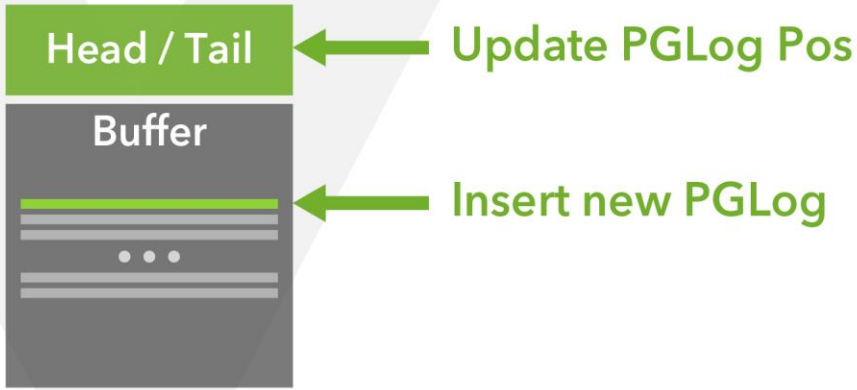
OSD METADATA

PGLog NVDIMM integrated

Rotational PGLog Object
Single file per PG



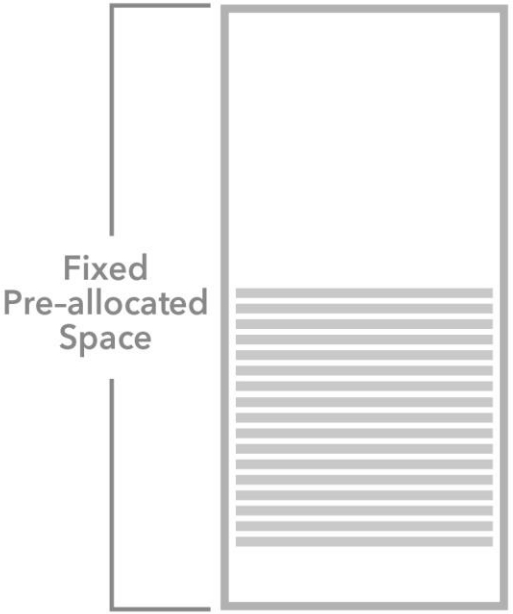
NVDIMM Cache Entry



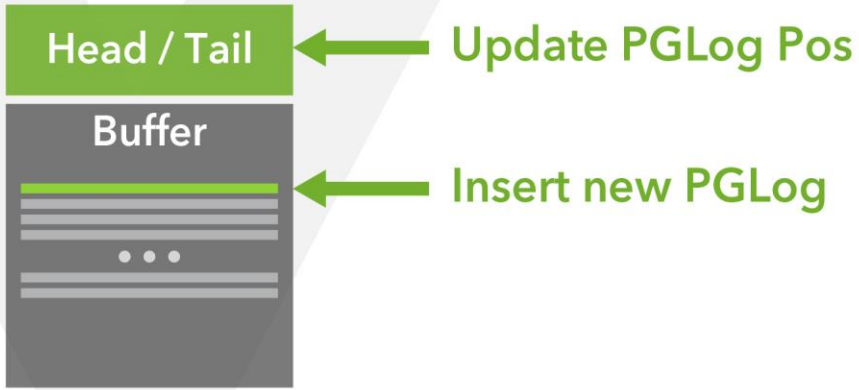
OSD METADATA

PGLog NVDIMM integrated

Rotational PGLog Object
Single file per PG

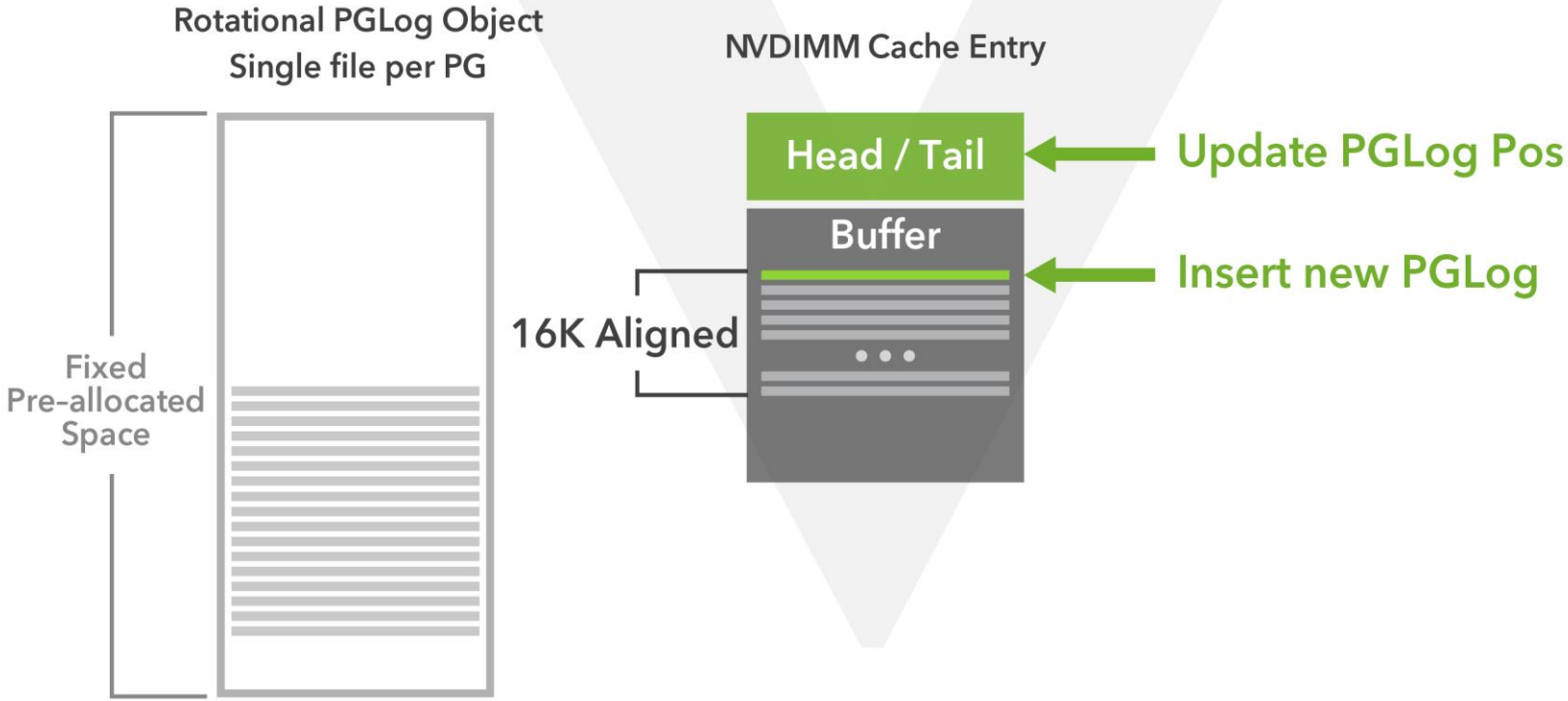


NVDIMM Cache Entry



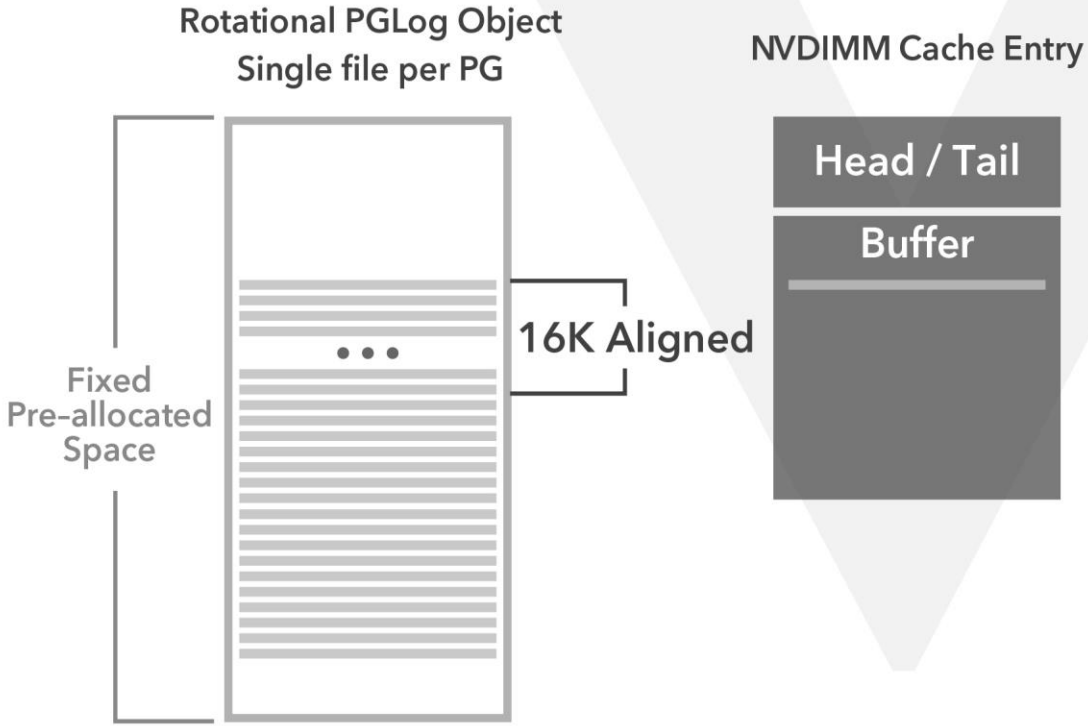
OSD METADATA

PGLog NVDIMM integrated



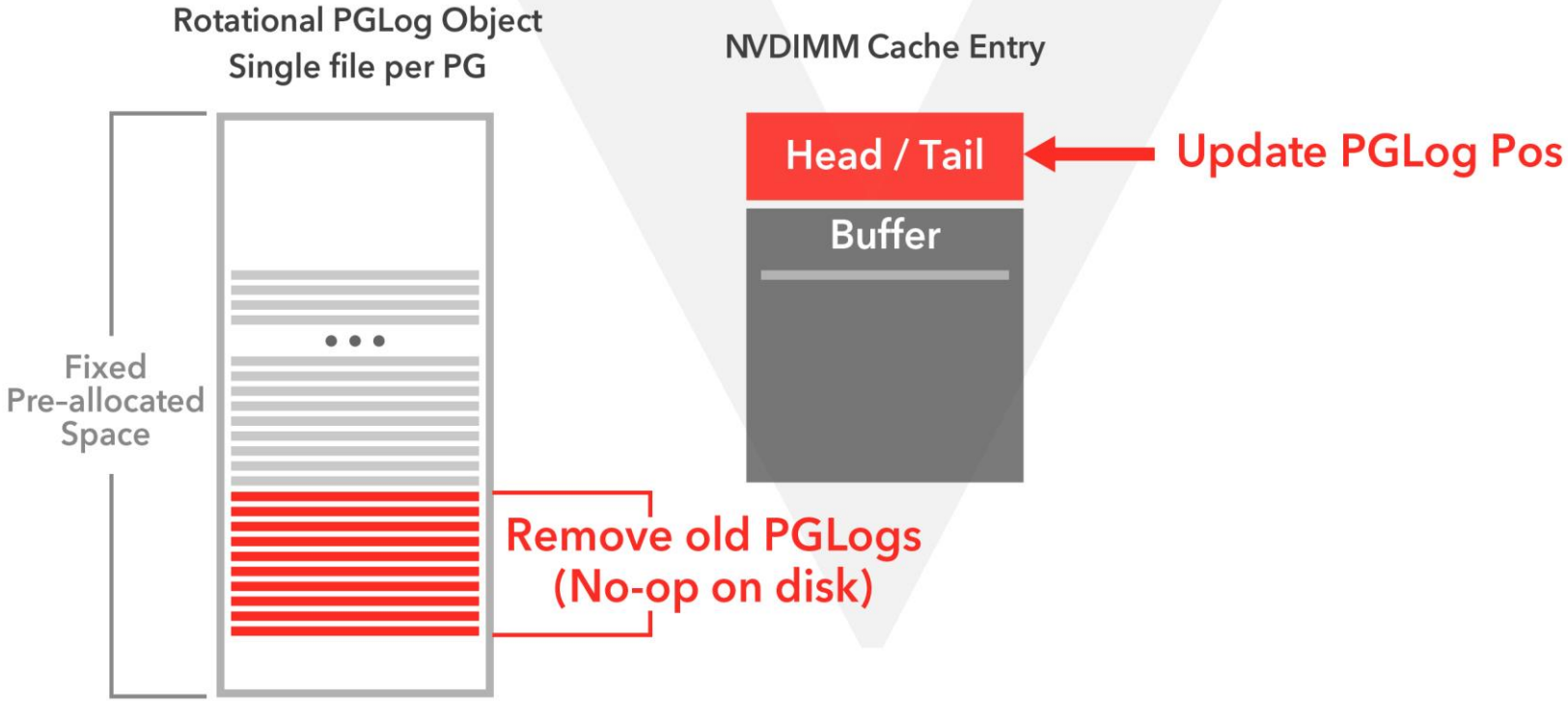
OSD METADATA

PGLog NVDIMM integrated



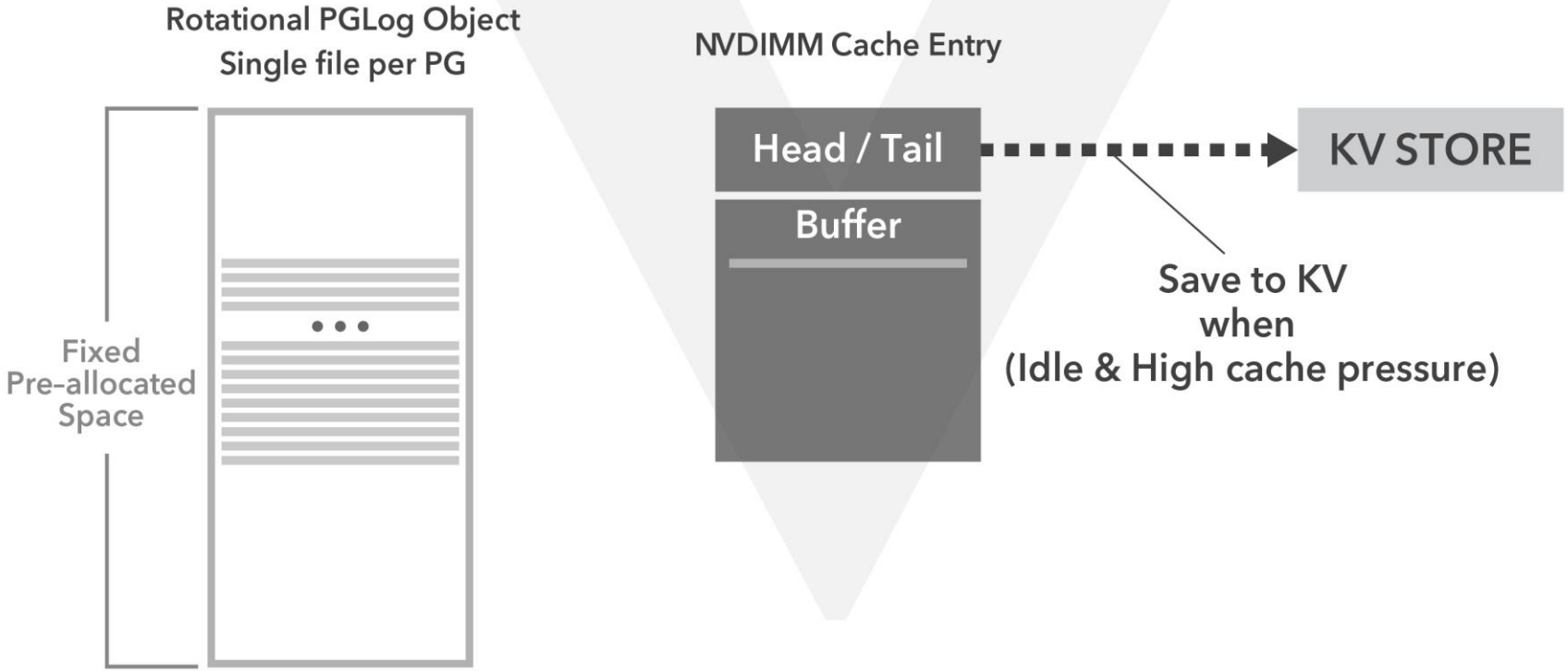
OSD METADATA

PGLog NVDIMM integrated

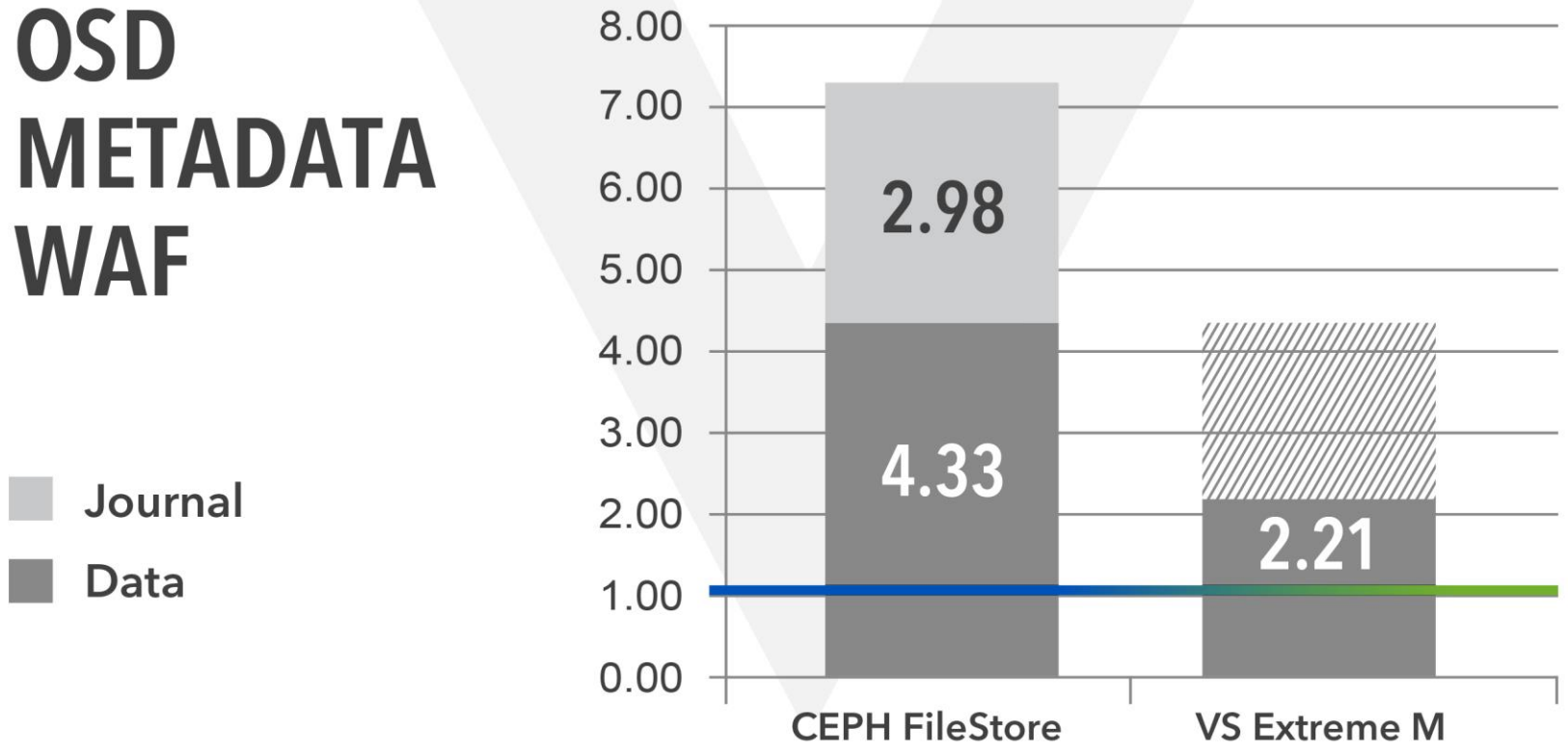


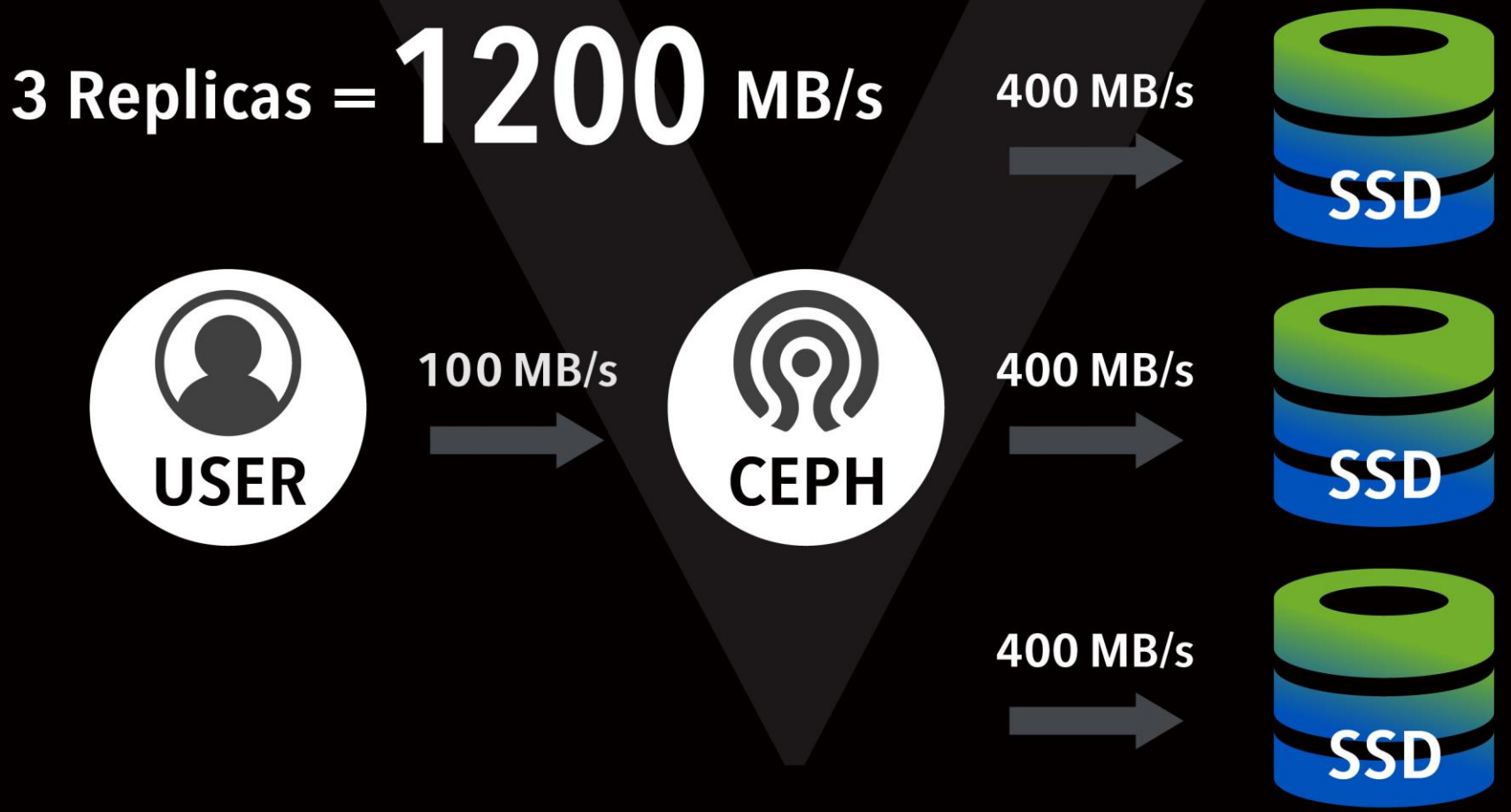
OSD METADATA

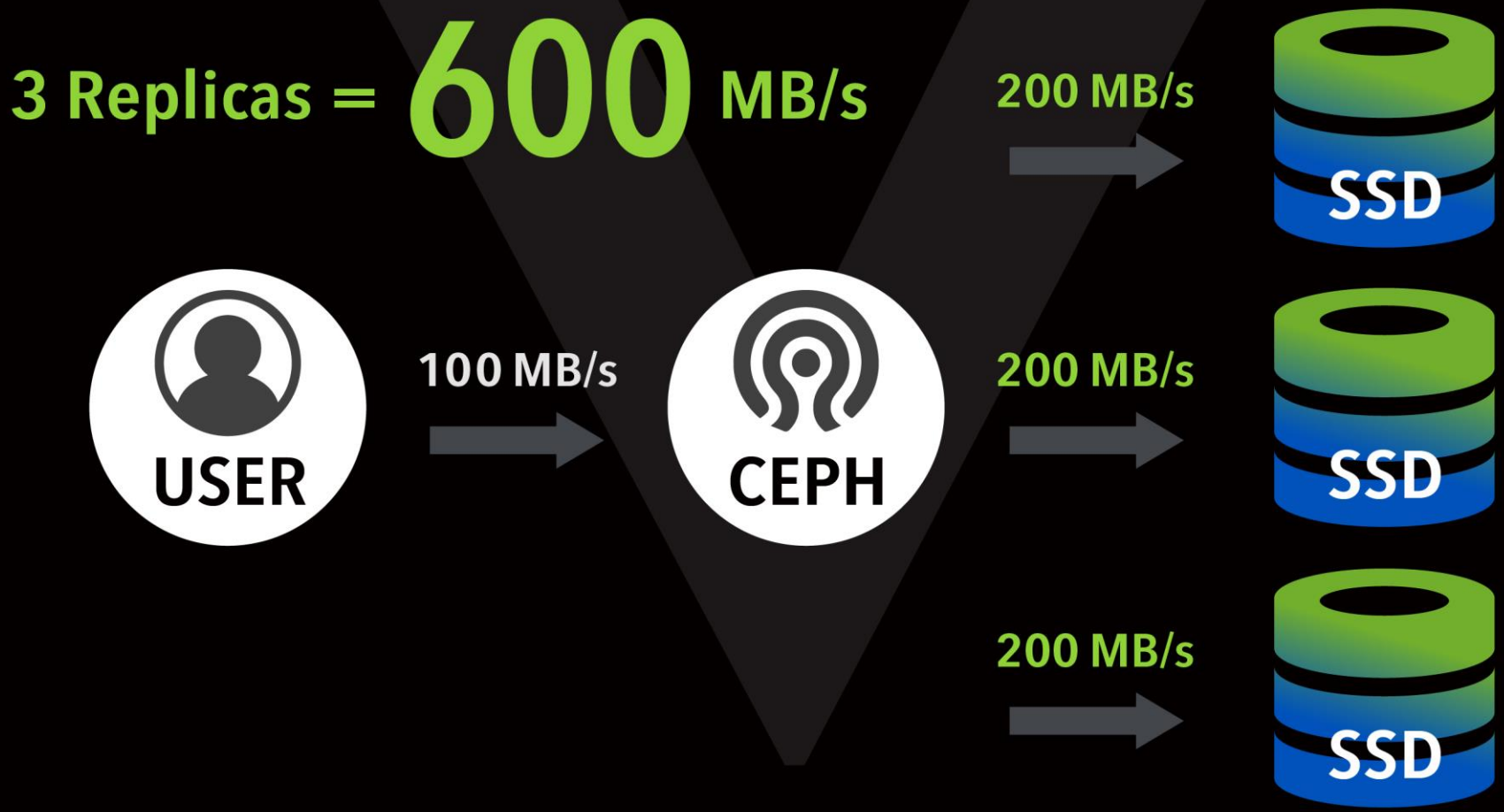
PGLog NVDIMM integrated



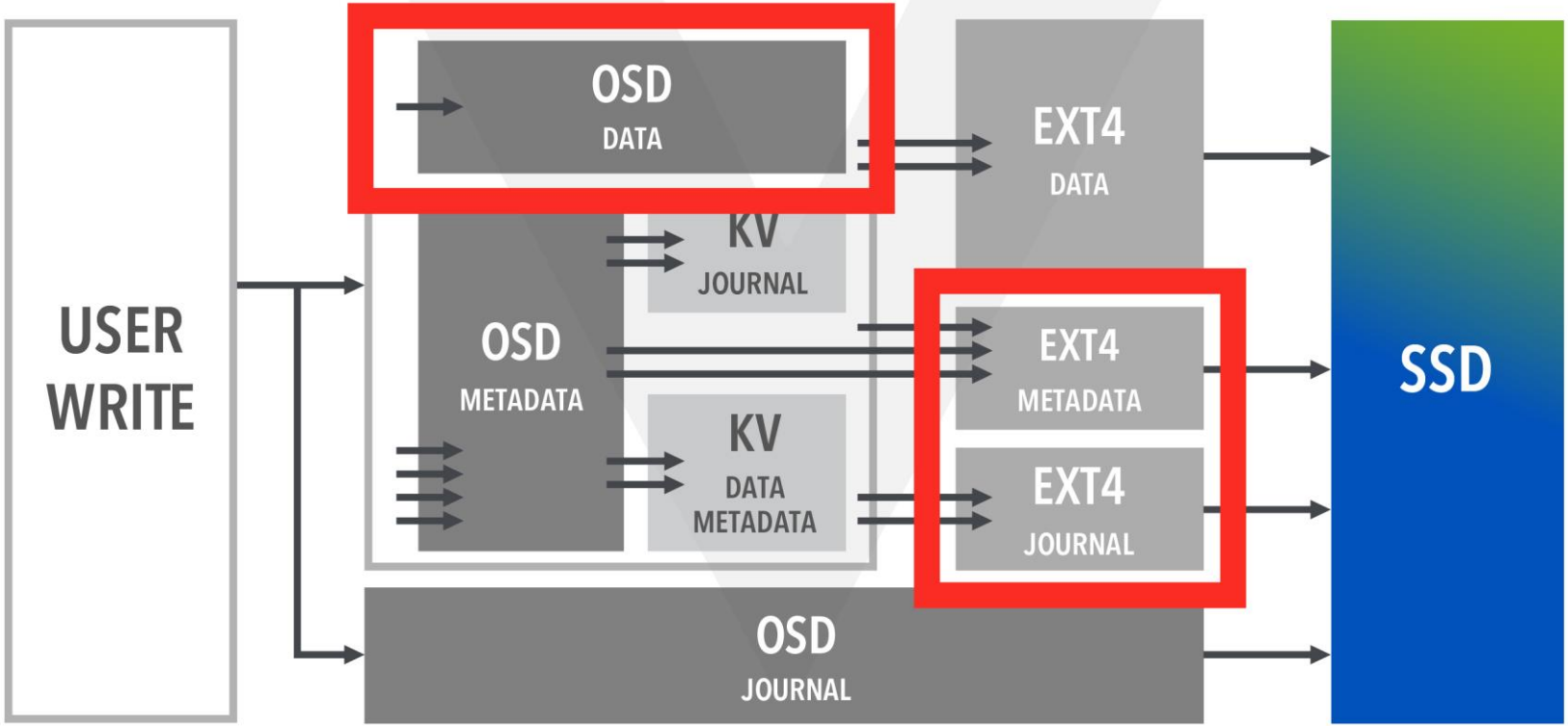
OSD METADATA WAF




3 Replicas = 1200 MB/s
400 MB/s
SSD
USER
100 MB/s
CEPH
400 MB/s
SSD
400 MB/s
SSD


3 Replicas = 600 MB/s
200 MB/s
SSD
USER
100 MB/s
CEPH
200 MB/s
SSD
200 MB/s
SSD

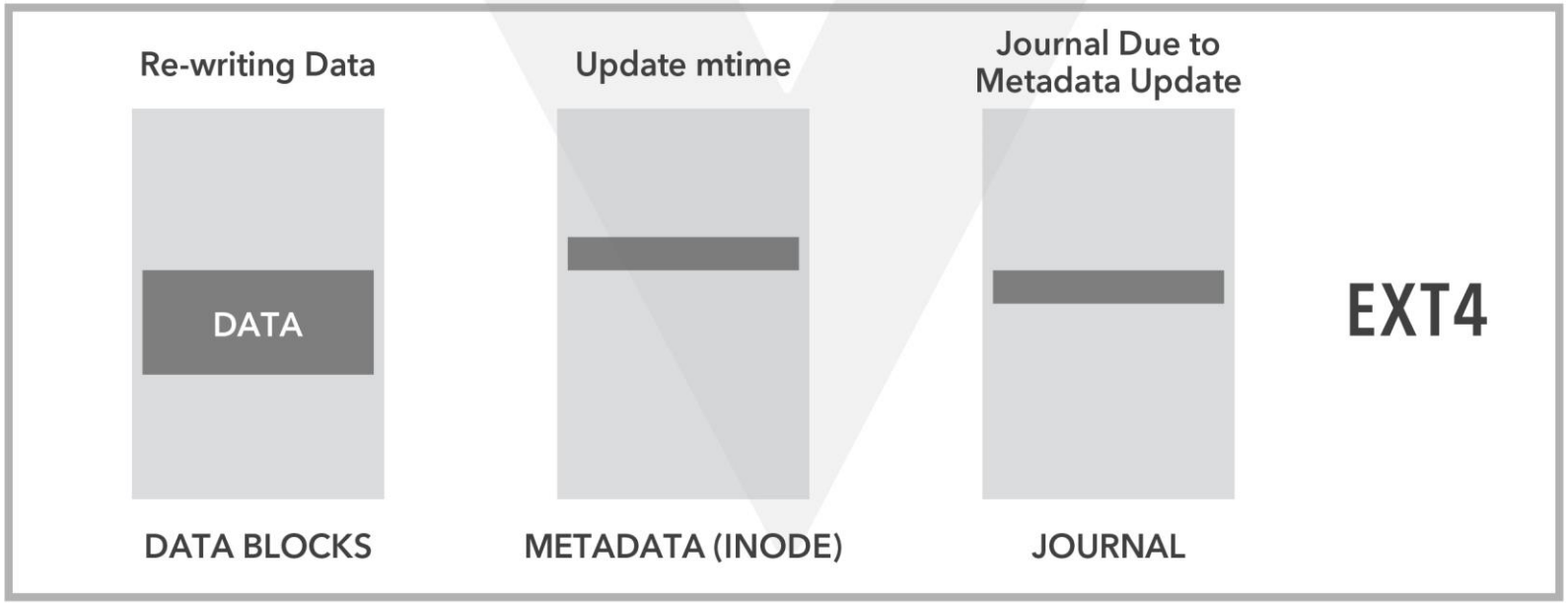
EXT4 JOURNAL & METADATA



EXT4 WRITE AMPLIFICATION

Keep track of mtime

Re-writing Data DATA



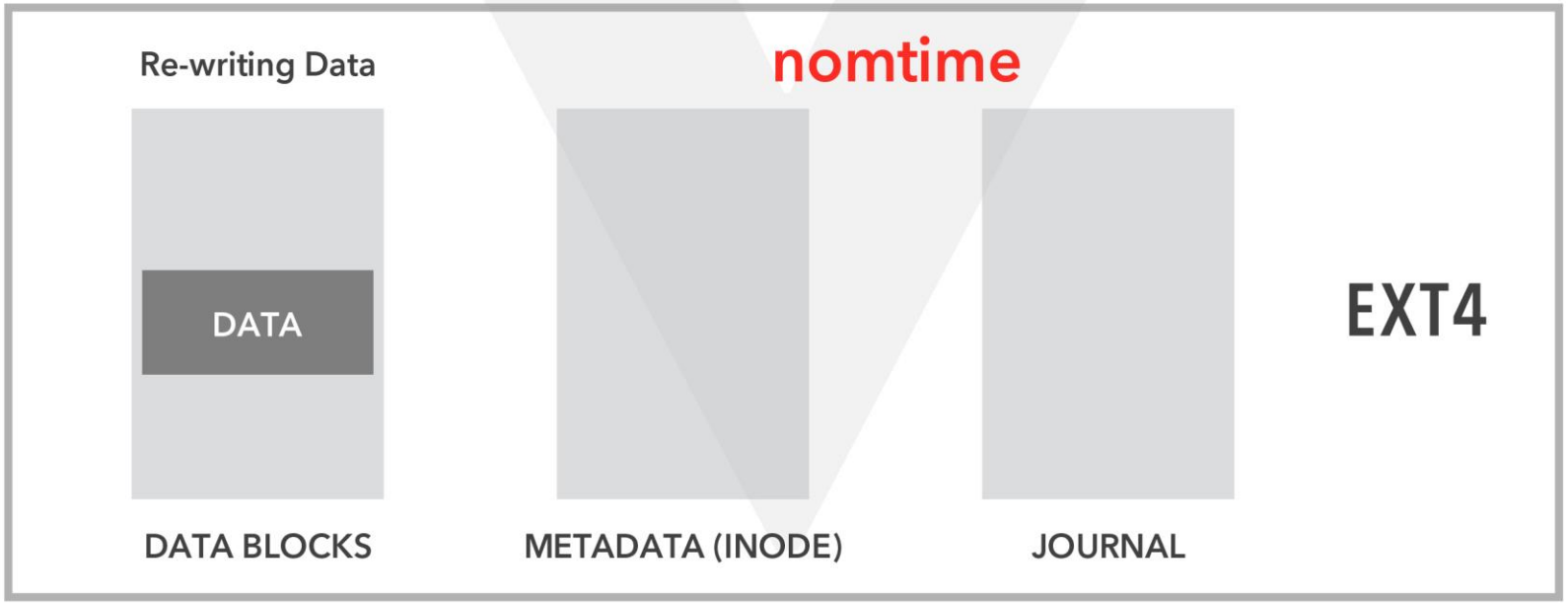
EXT4 WRITE AMPLIFICATION

Do not touch mtime

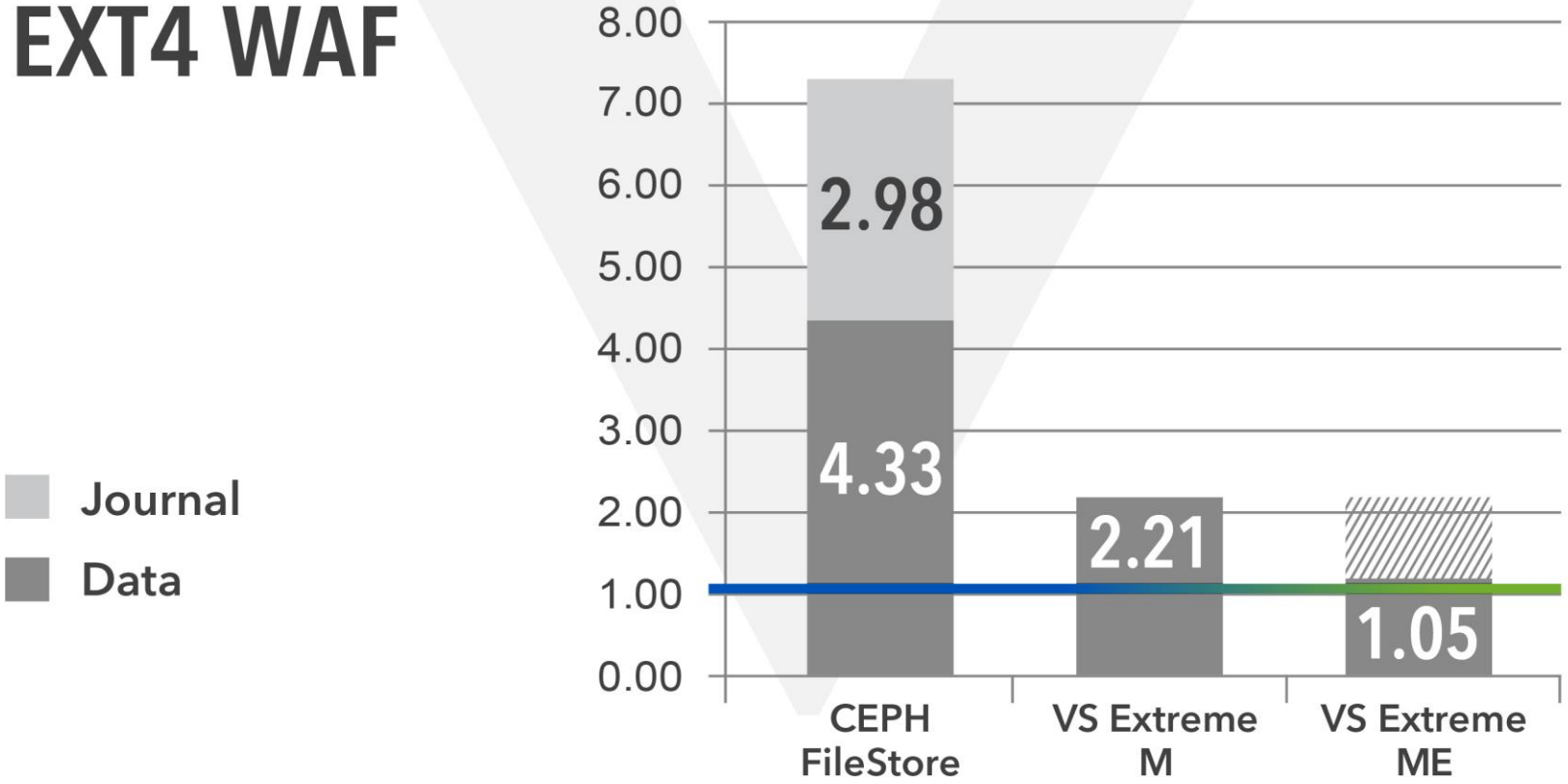
Re-writing Data

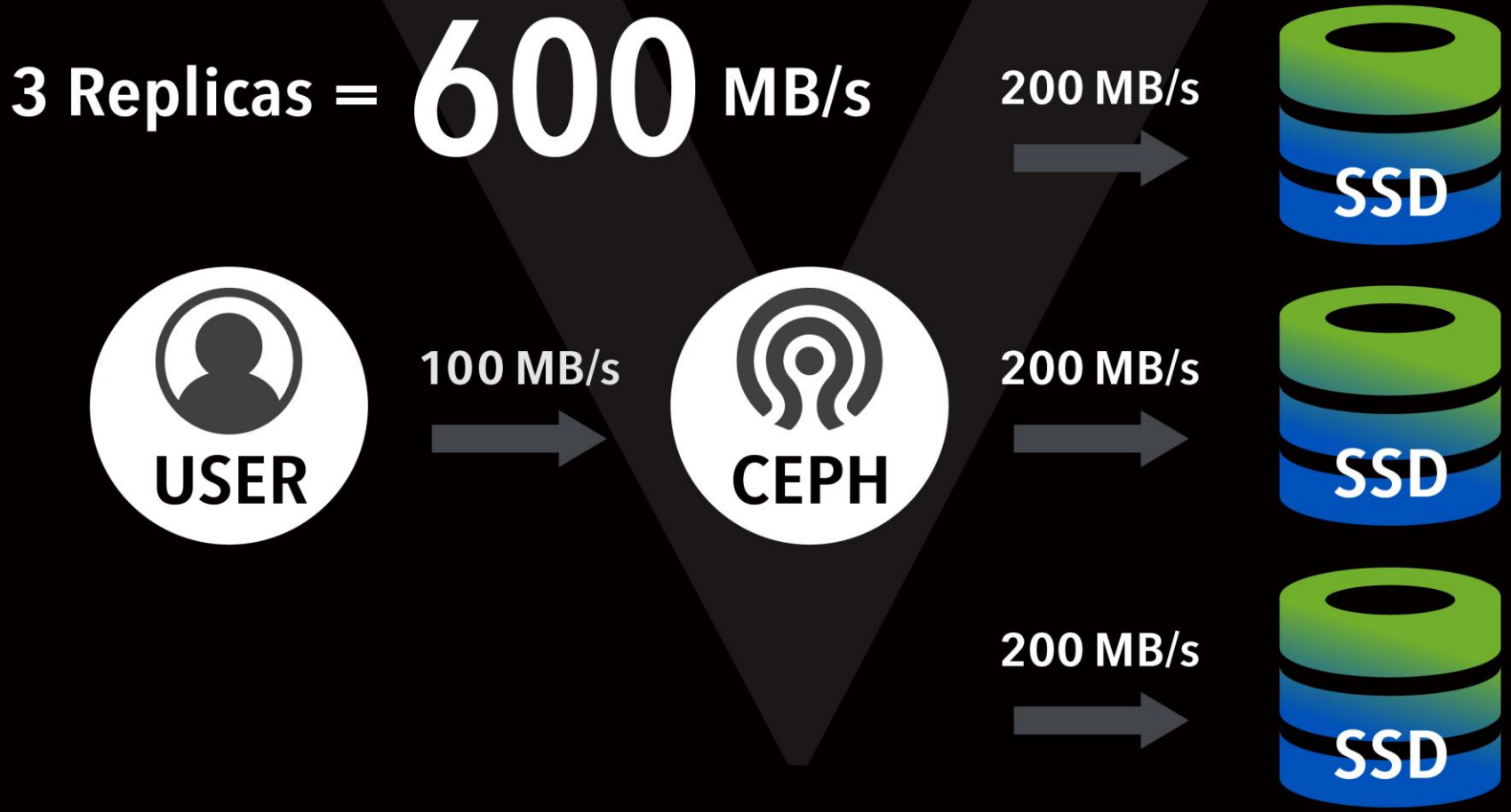
DATA

OSD does _not_ require mtime information on filesystem



EXT4 WAF


Journal
Data
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
2.98
4.33
2.21
1.05
CEPH FileStore
VS Extreme M
VS Extreme ME


3 Replicas = 600 MB/s
200 MB/s
SSD
USER
100 MB/s
CEPH
200 MB/s
SSD
200 MB/s
SSD

3 Replicas = 300 MB/s

100 MB/s

100 MB/s

100 MB/s

100 MB/s


LOCK FREE QUEUES
Messenger
OSD Workers
IO Workers
CPU OPTIMIZATIONS
Memory spatial / temporal locality
Huge pages
Avoid memory copy
Less context switches
CPU Affinity
NUMA consideration
VirtualStor™
EXTREME
AVOID LOCK CONTENTION
Per thread data
Smarter job scheduling
Atomic data structures
Parallel RW lock
OBJECT POOLS
Better reuse of frequently used objects
Better memory spatial locality
Lock free allocation / free
Huge page to have better TLB cache hit
NVFAStore
AIO supported backend
Specialized NVDIMM journal structure
No locks in all data path

VIRTUALSTOR EXTREME

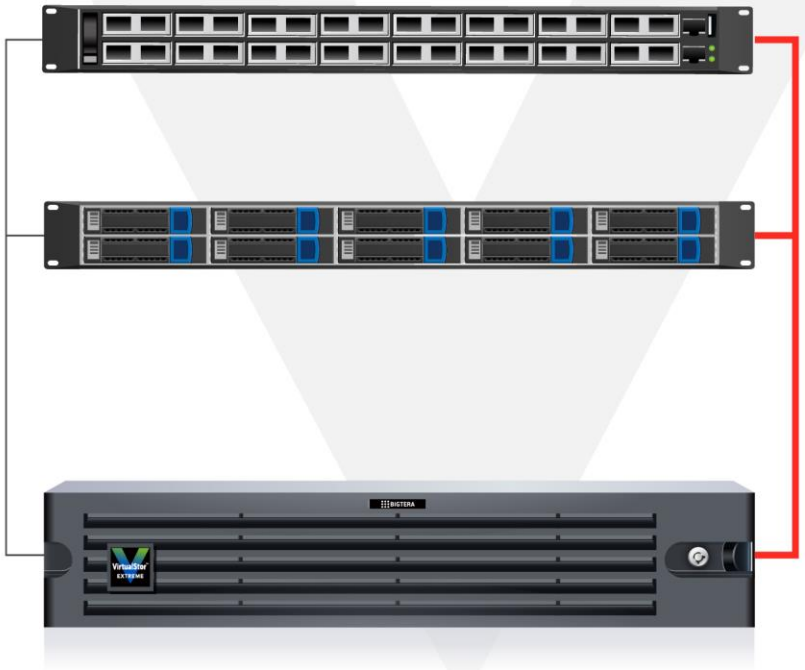
- Hardware Form Factor: 4 nodes in 2U chassis
- Spec Per Node:
  - 2x Intel Xeon Scalable Processor
  - 256 GB Memory
  - 2x 16 GB NVDIMM
  - 6x SATA SSD: [960 GB, 1.92 TB or 3.84 TB]
  - Network 2x 10GbE and 1x 1GbE IPMI
  - Add-on: Dual-Port 40GbE, Fiber Channel

BENCHMARK ENVIRONMENT

Brocade TurboIron 24X
10GbE Switch

Client Node* 4
Intel Gold 5118 x2
128GB RAM
Intel X710 10GbE
Ubuntu 16.04

Storage Nodes (2U4N)
SuperMicro 2029TP-HCOR
Intel Silver 4114 x2
192GB RAM
Viking NVDIMM 16GB x2
Intel 82599ES 10GbE
VirtualStor Extreme
Intel DC S4500 x6



Intel CeTune Benchmark

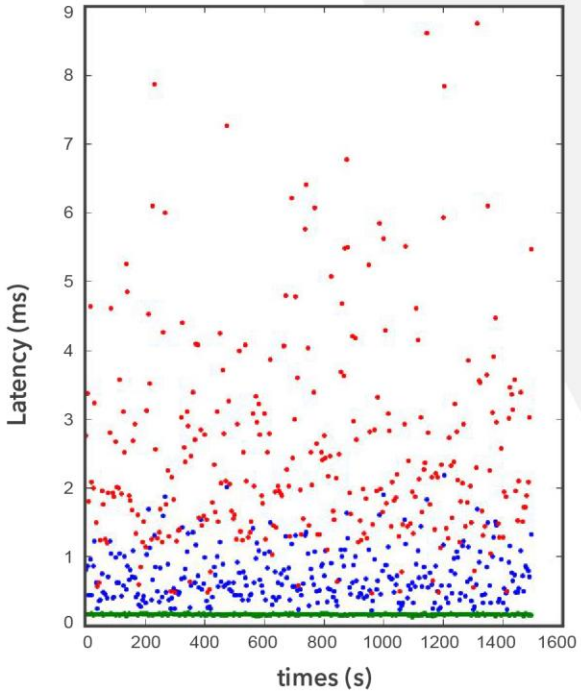
112GB Vdisk* 16 VM* 4 Nodes
7TB Dataset

1QD* 16 VM* 4 Nodes
64QD

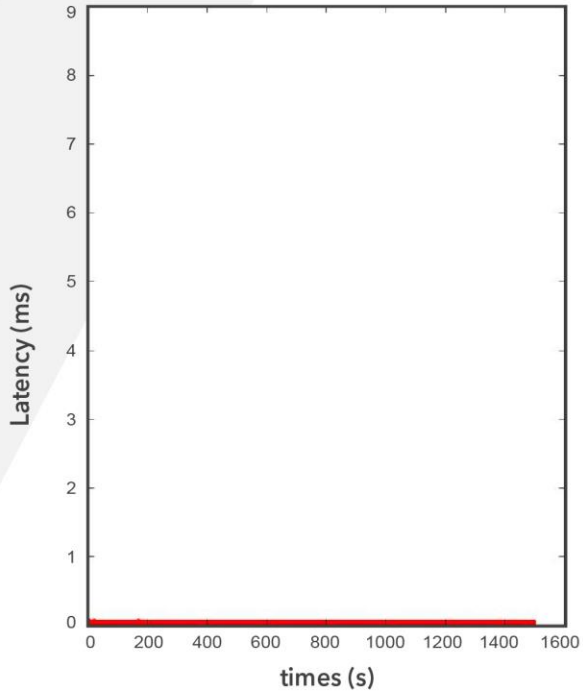
SSD LATENCY

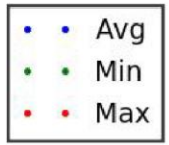
CEPH Mimic BlueStore VS VirtualStor Extreme

Mimic BlueStore



VirtualStor Extreme

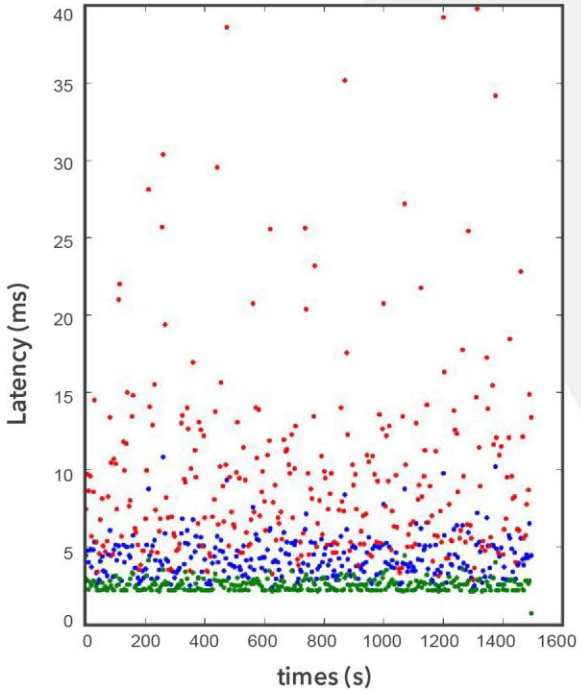




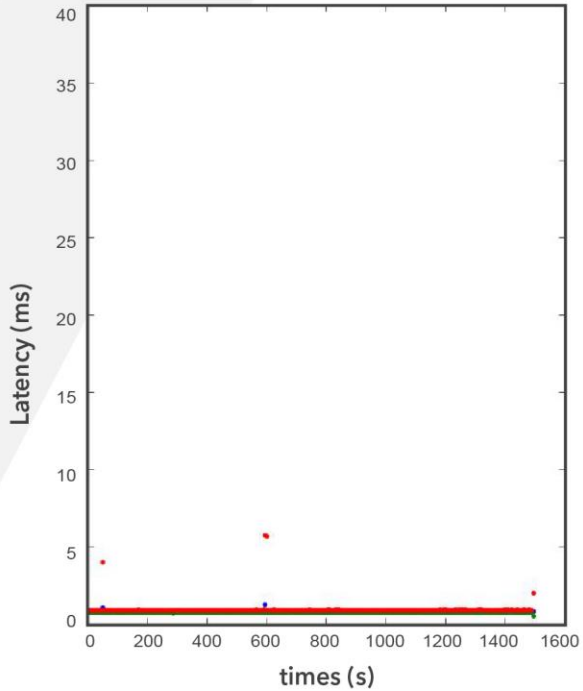
CLIENT LATENCY

CEPH Mimic BlueStore VS VirtualStor Extreme

Mimic BlueStore



VirtualStor Extreme



Thank you for your attention

Questions?