Scaling 30TB+ SMR

SAS4 Performance Gains and HAMR-SMR vs. ePMR-SMR

David Gerstein
CTO
david@leil.io

The International Conference on Massive Storage Systems and Technology

Santa Clara University

June 1, 2026

MSST

leil

About Leil

Single Software Stack. Two Paths.

HDD-Native™ Distributed File System which implements concepts introduced by Google File System.

LeilFS™ — Open Source

leil FS

- Included in Ubuntu, Debian, Fedora and major Linux distributions
- Core file system capabilities – reduced feature set
- Free to deploy, no license cost
- Community support

LeilOS™ — Enterprise

leil OS

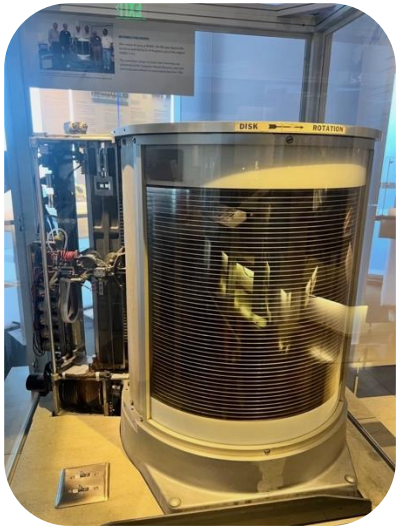
- Full feature set: SMRT Engine, advanced EC, variable capacity, management console
- ICE and head depopulation (roadmap)
- License cost can be zero — support is the standard engagement model
- Reseller Program: Margins + Deal Registration + MDF

HQ Tallinn, Estonia, EU

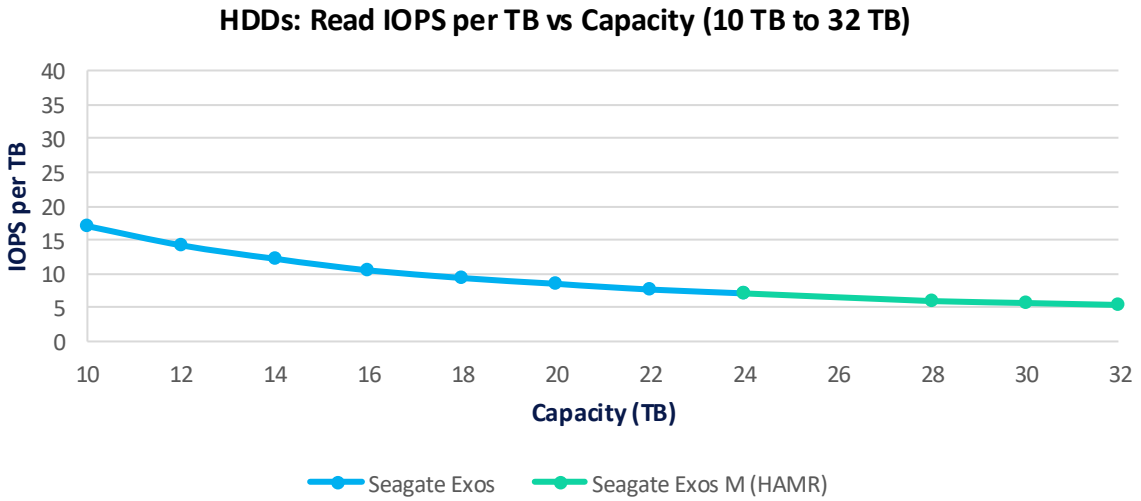
MSST

HDDs Through Time: 5MB to 36TB

Shifting Workloads: Random Access to Sequential for HDD Efficiency



1956



2026

**The IBM 305 RAMAC, the world's first commercial hard disk drive, was introduced on September 14, 1956*

The Misconception

~~*“SMR is too slow for serious workloads.”*~~

We measured. It isn't.

+50%

Multistream Writes — HAMR over ePMR

EC6+2, 8× 30TB

+14–16%

Reads & Writes — SAS4 over SAS3

60 drives, identical FIO sweep

~1.5×

Writes — Experimental zone allocation

102× WD Data102, EC6+2

What We'll Prove

Three independent variables — each one moves the number

PILLAR 01

Vendor

HAMR (Seagate) vs. ePMR (WD)

Single-drive bandwidth and EC6+2 multi-drive throughput diverge by up to 1.7× at saturation.

PILLAR 02

Interface

SAS4 24G 2-port vs. SAS3 12G 4-port

Same drives, same FIO sweep.
24G adds ~15% top-line at 60 jobs.

PILLAR 03

HBA

1× ATTO ESAH-12F0 vs. 2× Broadcom 9500-8e

One slot, one ATTO card matches or beats two Broadcom cards on raw reads at the SAS3 ceiling.

Test Setup

Two reference platforms, two HDD vendors, two HBA vendors, EC6+2 and raw throughout

Platforms

Reference A

WD Ultrastar Data102 · 102 × drives · SAS3 12G

Reference B

AIC XP1-S407VA014U · 60 × drives · SAS4 24G

Network

Dual-port 100 GbE QSFP28

Software

LeilFS 5.10 (Stable + Experimental)

Drives under test

HAMR

Seagate 30 TB SMR · ST30000NM011F

ePMR

WD 30 TB SMR · WSH723200ALN604

Excluded

Toshiba MG series — CMR only, SMR are not yet qualified

Storage controllers

SAS3 12G

1 × ATTO ESAH-12F0-GT0 (16 ports) — primary

SAS3 12G

2 × Broadcom 9500-8e — comparison

SAS4 24G

1 × ATTO H240F

Vendor: STX vs. WDC

At 30TB the two vendor technologies diverge. We measured how, and where it matters.

- Raw single-drive write — bandwidth per FIO job
- EC6+2 multi-drive reads (8 × drives, LeilFS 5.10 Stable)
- EC6+2 multi-drive writes (8 × drives, LeilFS 5.10 Stable)

EC6+2 Reads: 8 × HAMR vs. 8 ePMR

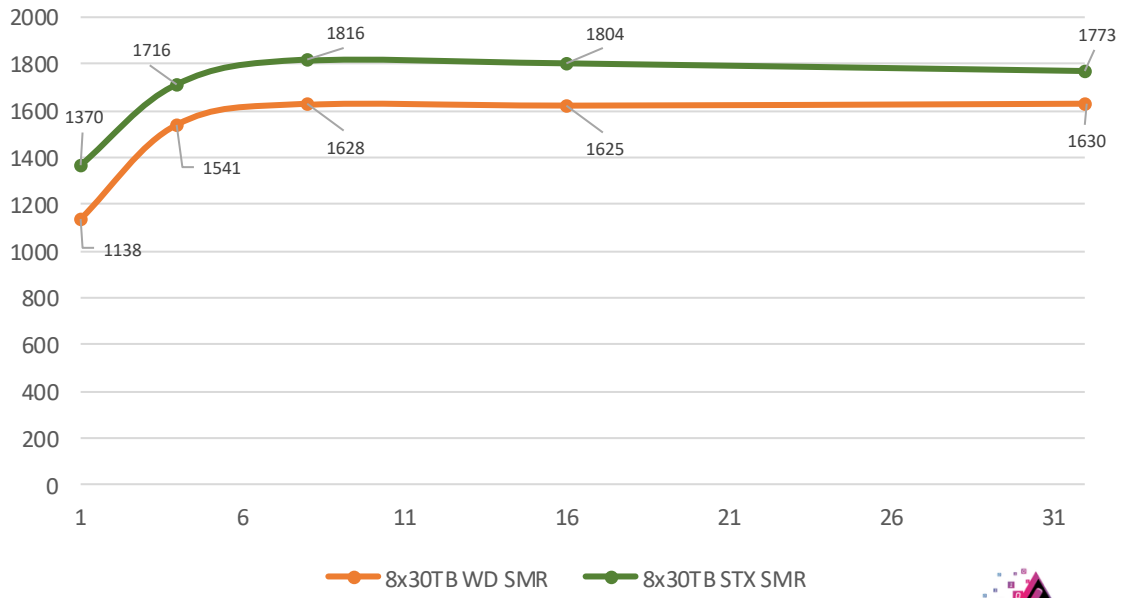
LeilFS 5.10, sequential reads across the chunkserver pool

+9%

STX sustains 1.77–1.82 GB/s reads vs. ePMR's 1.62–1.63 GB/s once the FIO sweep saturates.

Saturated by 8 jobs

Read scaling flattens beyond 8 concurrent jobs — nearly drive ceiling.



EC6+2 Writes: 8 × HAMR vs. 8 ePMR

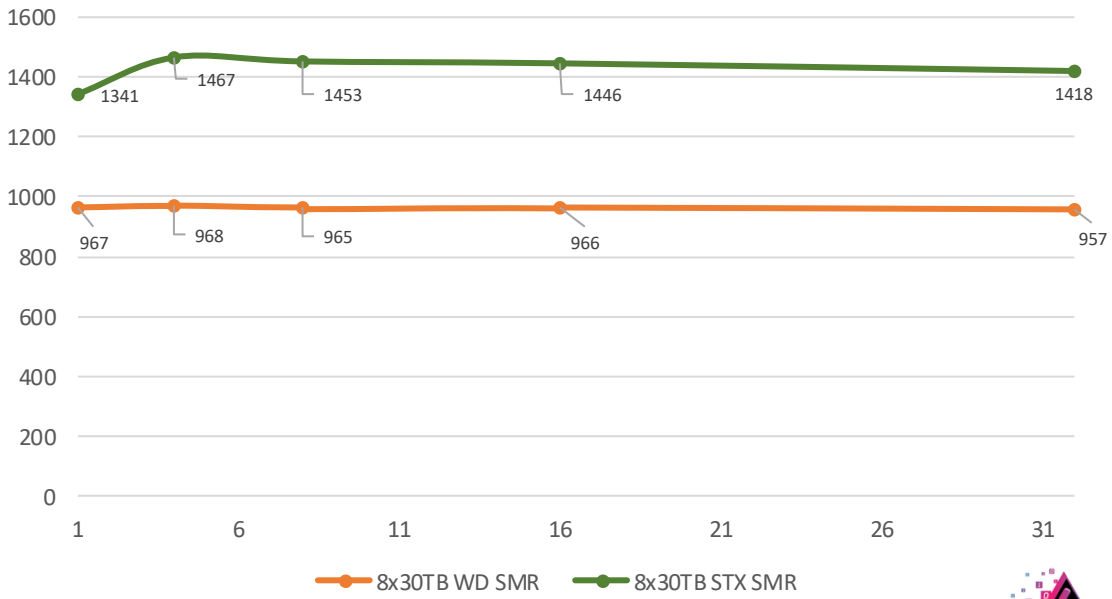
LeilFS 5.10, sequential reads across the chunkserver pool

+50%

STX sustains ~1.45 GB/s writes vs. WDC's ~960 MB/s across the full job sweep.

Flat = saturated

WDC plateaus immediately — drive's limit, not the file system.



Methodology: LeilFS FIO Sweep SMR

Parameterized job-count sweep across the chunkserver pool

```
# Low job counts — large per-job size
for jobs in 1 2 4 8 10; do
  sudo fio --name=test${jobs} --directory=$MOUNT \
    --size=64G --rw=write --numjobs=${jobs} \
    --ioengine=libaio --group_reporting \
    --bs=1M --direct=1 --iodepth=1
  sleep 5
done

# High job counts — smaller per-job size
for jobs in 20 30 40 60 80 102; do
  sudo fio --name=test${jobs} --directory=$MOUNT \
    --size=16G --rw=write --numjobs=${jobs} \
    --ioengine=libaio --group_reporting \
    --bs=1M --direct=1 --iodepth=1
  sleep 5
done   # repeat with --rw=read for the read sweep
```

Methodology: Client/Chunkserver Tuning

What's different from default — for reproducibility

Client mount (8x test)

```
sfsmount /mnt/mount211 \
  -H 192.168.50.211 -o \
  sfswritecachesize=4096,\
  sfschunkserverwriteto=10000,\
  sfschunkserverwavereadto=2000,\
  sfschunkservertotalreadto=8000,\
  maxreadaheadrequests=2,\
  cacheexpirationtime=5000,\
  sfscacheperinodepercentage=100,\
  readbuffersexpirationtime=3000,\
  readworkers=16,\
  sfswriteworkers=16,\
  readcachemaxsizepercentage=80,\
  readaheadmaxwindowsize=65536,\
  sfswritewindowsize=256,\
  sfscachemode=NEVER
```

Chunkserver tuning

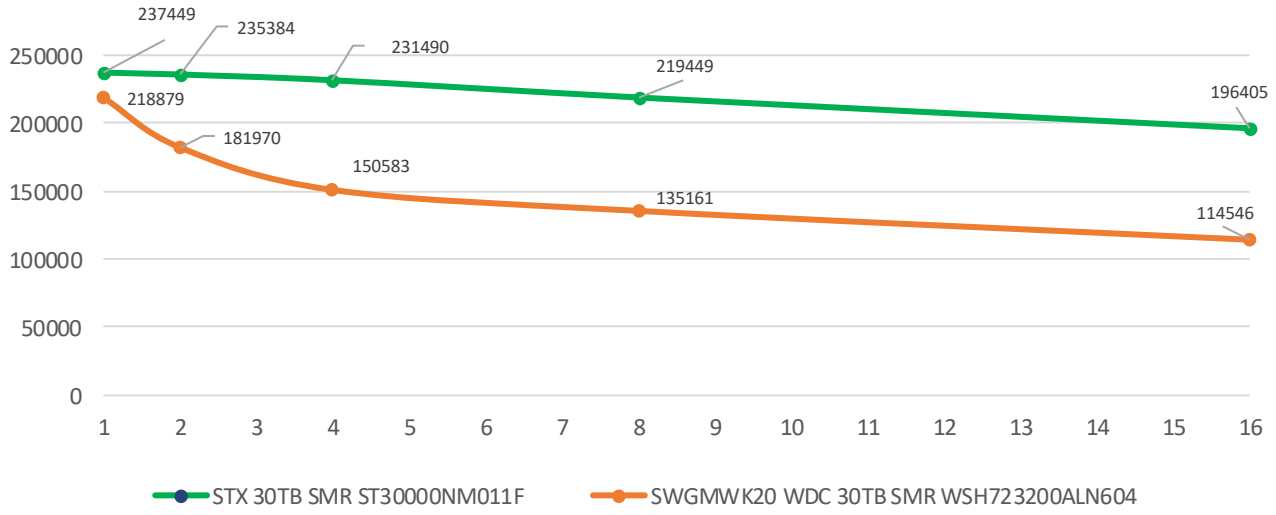
```
HDD_TEST_FREQ = 10000
GARBAGE_COLLECTION_FREQ_MS = 0
PERFORM_FSYNC = 0
HDD_CHECK_CRC_WHEN_WRITING = 0
HDD_CHECK_CRC_WHEN_READING = 0
NR_OF_NETWORK_WORKERS = 1
NR_OF_HDD_WORKERS_PER_NETWORK_WORKER = 4
MAX_BLOCKS_PER_HDD_WRITE_JOB = 32
MAX_BLOCKS_PER_HDD_READ_JOB = 1024
MAX_PARALLEL_HDD_READ_JOBS_PER_CS_ENTRY = 1
WRITE_BUFFERING_SIZE_MB = 64
```

Raw Write Bandwidth per Drive

Single-drive sequential write, MB/s vs FIO job

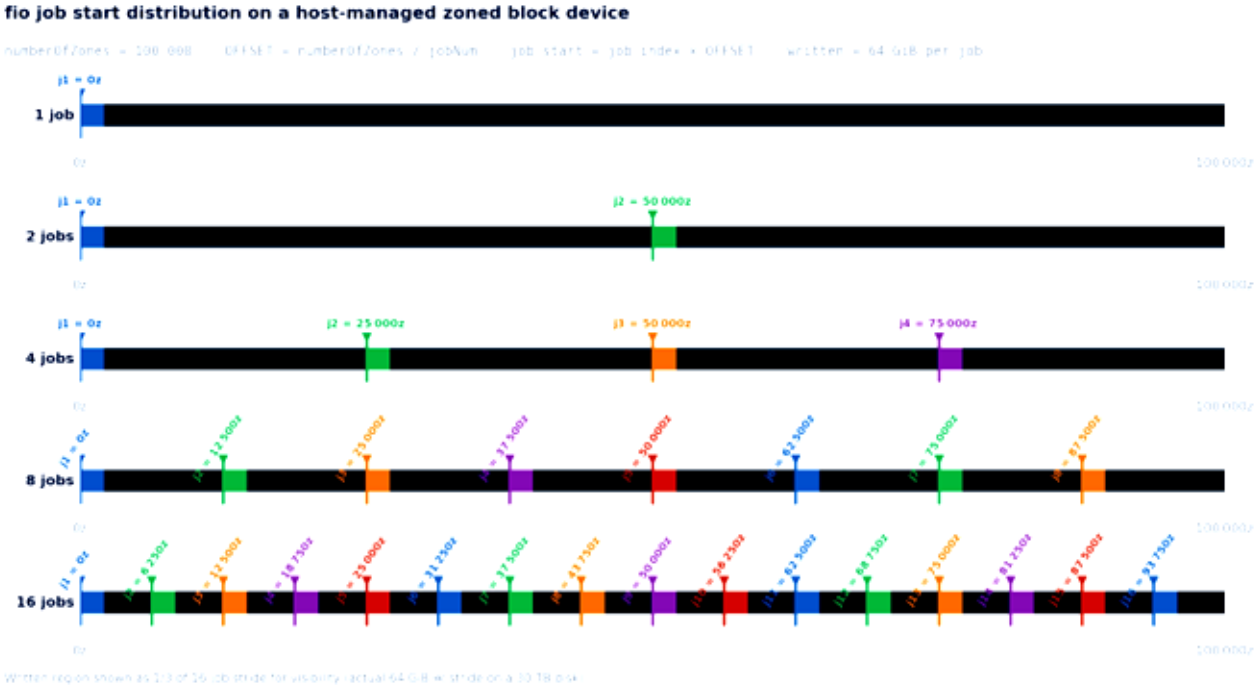
58% delta

At 16 concurrent jobs, STX holds 192 MB/s/drive while WD collapses to 112 MB/s/drive. STX loses only 17% across 1→16 jobs. WD loses 48%.



30TB STX SMR vs 30TB WD SMR Raw

Methodology and Parameters




```
# Each job writes up to 64 GiB
bs=64K
iodepth=1
direct=1
runtime=180s
zonemode=zbd

# Distribute jobs across the disk by zone position
OFFSET = numberOfZones / jobNum
job_start = job_index * OFFSET

# Goal: throughput vs. multiple write streams to
# different disk regions
```

Why this matters

The raw single-drive comparison stresses the drive's ability to handle multiple concurrent write streams landing in different zones — the load pattern an EC chunkserver actually produces.

*The picture shows an illustrative representation of the distribution.

PILLAR 02 - INTERFACE

Interface: SAS4 vs. SAS3

24G isn't just a faster cable — it changes the ceiling for high-density SMR.

- 60-drive sequential reads — same drives
- 60-drive sequential writes — same drives
- Why 2-port @ 24G outperforms 4-port @ 12G

SAS4 24G vs. SAS3 12G — Seq Reads

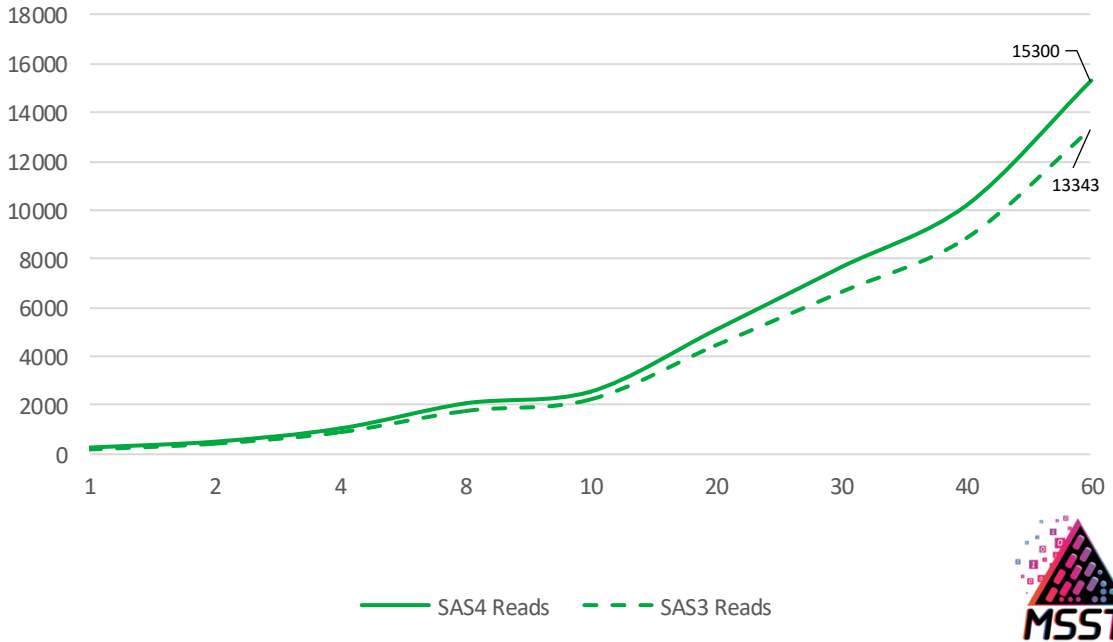
60-drive AIC platform vs WD JBOD102 with 60 drives, raw read bandwidth across FIO job sweep

+14.7%

At 60 jobs, SAS4 delivers 15.3 GB/s reads vs. SAS3's 13.3 GB/s.

Steady delta

Lead opens past 20 jobs — bandwidth headroom, not just peak.



SAS4 24G vs. SAS3 12G — Seq Writes

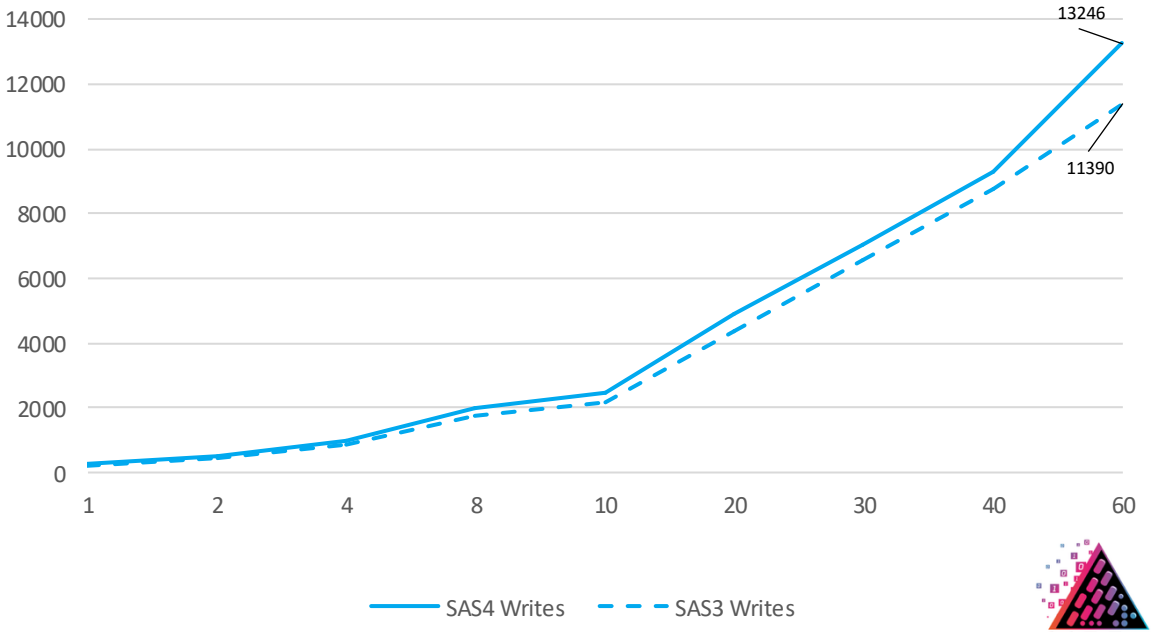
60-drive AIC platform vs WD JBOD102 with 60 drives, raw read bandwidth across FIO job sweep

+16.3%

At 60 jobs, SAS4 sustains 13.25 GB/s writes vs. SAS3's 11.39 GB/s.

Linear to 40 jobs

Both interfaces scale linearly until the 12G fabric saturates.



Methodology — Raw SMR FIO Profile

Drives are accessed in ZBD mode, offsets skip the random conventional zones

```
# FIO profile parameters
[global]
rw=write        # also rw=read for the read sweep
bs=1M
iodepth=1
direct=1
ioengine=libaio
group_reporting
size=10G
zonemode=zbd       # SMR-only; remove for CMR
offset=260113956864 # skip conventional random zones
```

```
# One fio per drive, two NUMA nodes in parallel, two repetitions per job count
numactl --cpunodebind=0 --membind=0 fio --alloc-size=$((256*1024*1024*1024)) "$profile1" &
numactl --cpunodebind=1 --membind=1 fio --alloc-size=$((256*1024*1024*1024)) "$profile2" &
```

HBA:
ATTO vs. Broadcom

Two cards in two slots, versus one card in one slot. What you actually get from each.

SAS3 12G 4-Port HBA ATTO vs Broadcom

1x ATTO ESAH-12F0-GT0 vs 2x Broadcom 9500-8e for WD Data102 (102 drives)

+5.4%

READ DELTA

14,705 vs. 13,951 MB/s

+0.4%

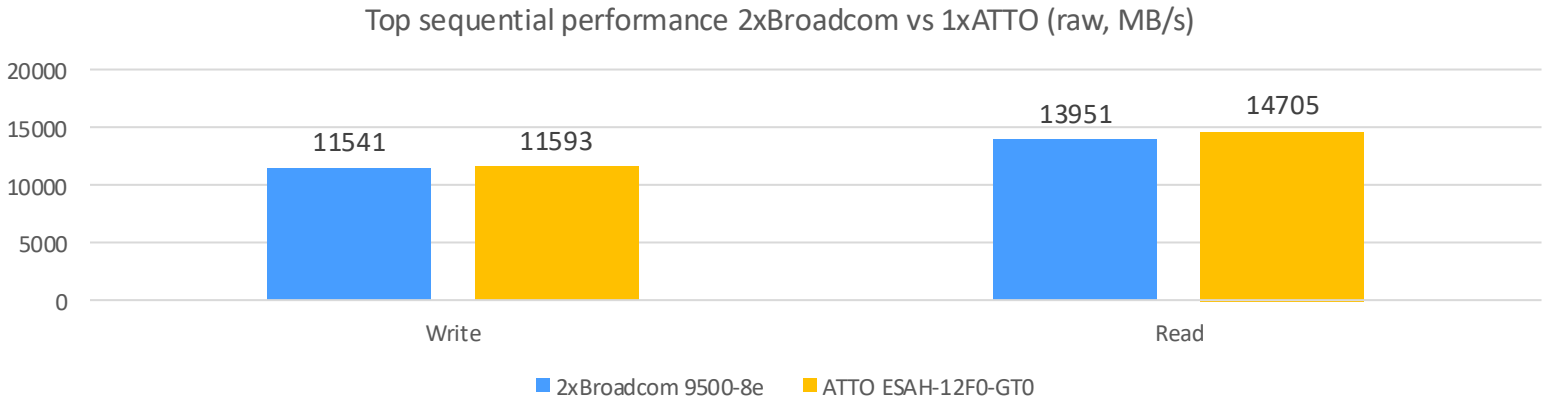
WRITE DELTA

11,593 vs. 11,541 MB/s

1 / 2

SLOTS USED

ATTO replaces two Broadcom cards



Software matters: Zone Assignment

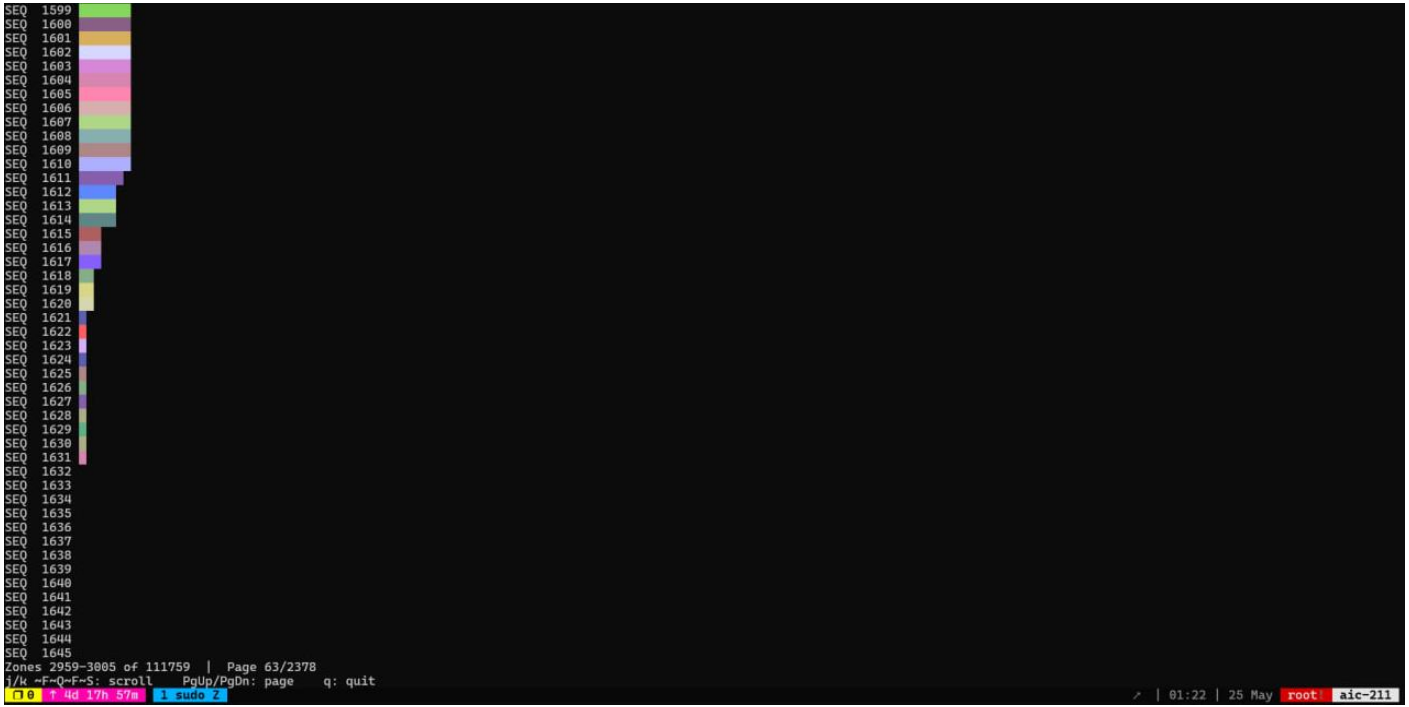
The biggest single-day performance gain we measured this year wasn't a new drive — it was how LeilFS decides which zone to write next.
Two zone-assignment strategies, one trade-off: read locality vs. write throughput. The next slides show both — and a +57% write gain.

Zone Assignment: Stable Branch

LeilFS places new chunks into different zones than the previous ones.

“Spread”

- Spreads each new chunk into a different zone, so appends to one chunk aren’t sequential at the drive.
- Win: fragmentation stays near zero, so sequential reads are excellent.
- Cost: writes at the drive level jump if the next operation is non-appending.



Zone Assignment: Experimental Branch

LeilFS fills zones one-by-one; most writes are strictly sequential at the drive.

“Fill One Zone”

- Fills one zone at a time; a client flush hint (pwritev) enables larger batched writes.
- Win: single-writer ordering keeps writes perfectly sequential — +57% throughput.
- Trade-off: slightly more fragmentation, yet reads hold near baseline.

Writes: Stable vs Experimental

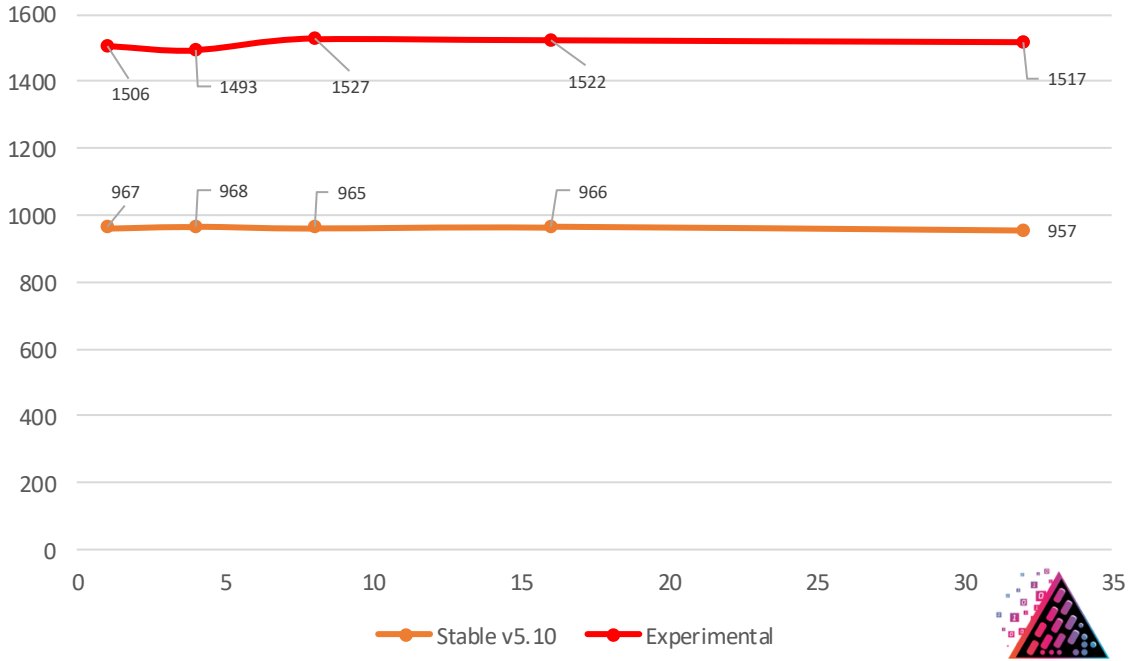
LeilFS 8 × 30TB SMR WD ePMR, EC6+2, sequential writes.

+57%

Experimental holds ~1.52 GB/s writes vs. Stable's ~965 MB/s across the sweep.

Limit-line behavior

Experimental flatlines at the drive's physical write ceiling — same shape, higher line.



Same WD 30TB ePMR drives · EC6+2 · WRITE_BUFFERING_SIZE_MB=64 in Experimental

Reads: Stable vs Experimental

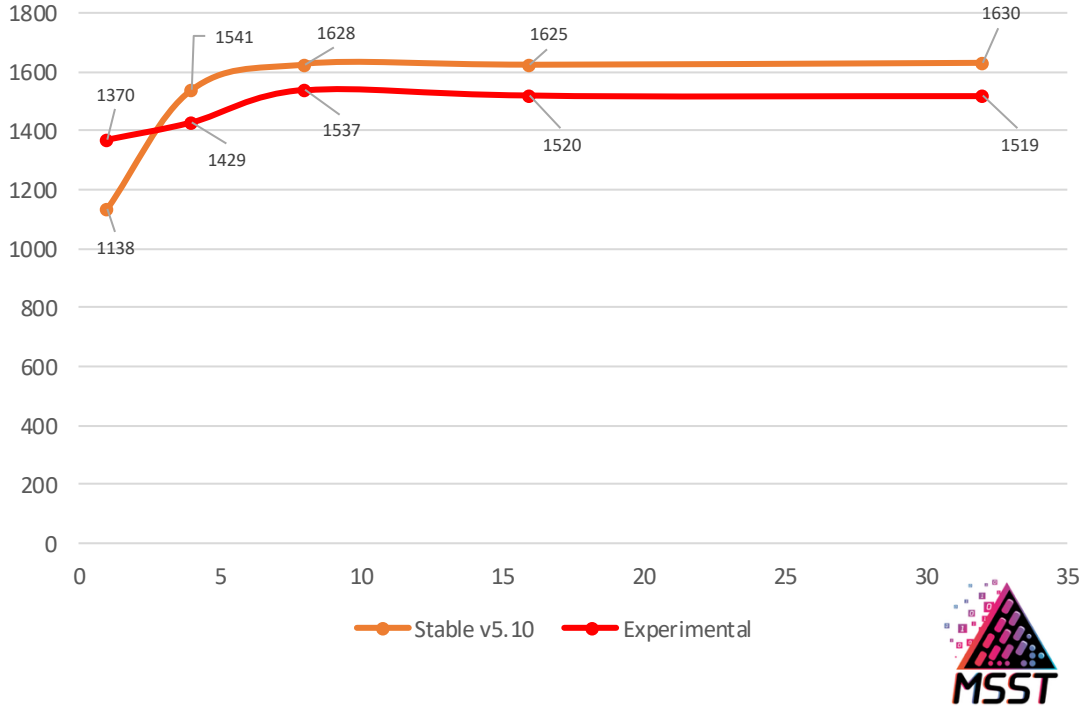
LeilFS 8 × 30TB SMR WD ePMR, EC6+2, sequential reads.

−7%

Reads drop slightly — 1.52 GB/s Experimental vs. 1.63 GB/s Stable at 32 jobs.

Tradeoff is favorable

+57% on writes for -7% on reads. Worth it for write-heavy archive ingest.



CUSTOMER VOICE

Real Numbers from PoC, Q1-2026

Benchmarked and tested in cooperation with

Poznańskie Centrum Superkomputerowo-Sieciowe
Poznan Supercomputing and Networking Center

WD Reference Architecture

Leil SMR Hardware SAS3 12G Blueprint. Extra Performance with Zoning Setup

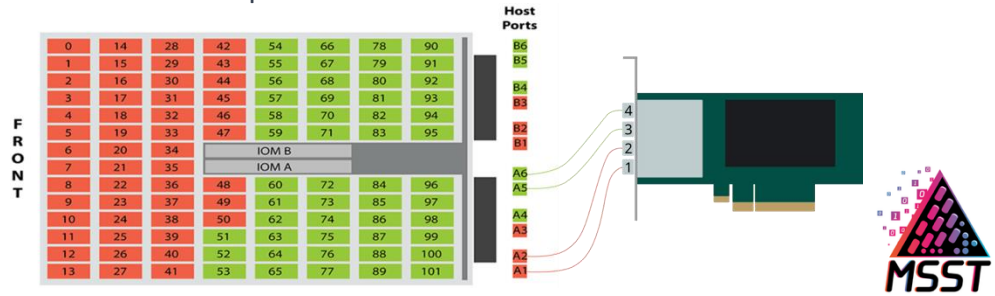
Ports 1-2 (wide port) of ATTO HBA are connected to A1 & A2 (red zone)
Ports 3-4 (wide port) are connected to A5 & A6 (green zone).

We do not use the second IOM because we are using SATA drives (same will be for new gen JBODs 3000 Series with 24Gb HBAs).

HM-SMR Drives
Single-port 6 Gb/s SATA

HBA
1x ATTO ESAH-12F0-GT0: 16-port 12Gb HBA

To Host
2x Mini-SAS HD ports per zone
4x Mini-SAS HD ports total



SAS3 12G 4-Port HBA Raw Performance

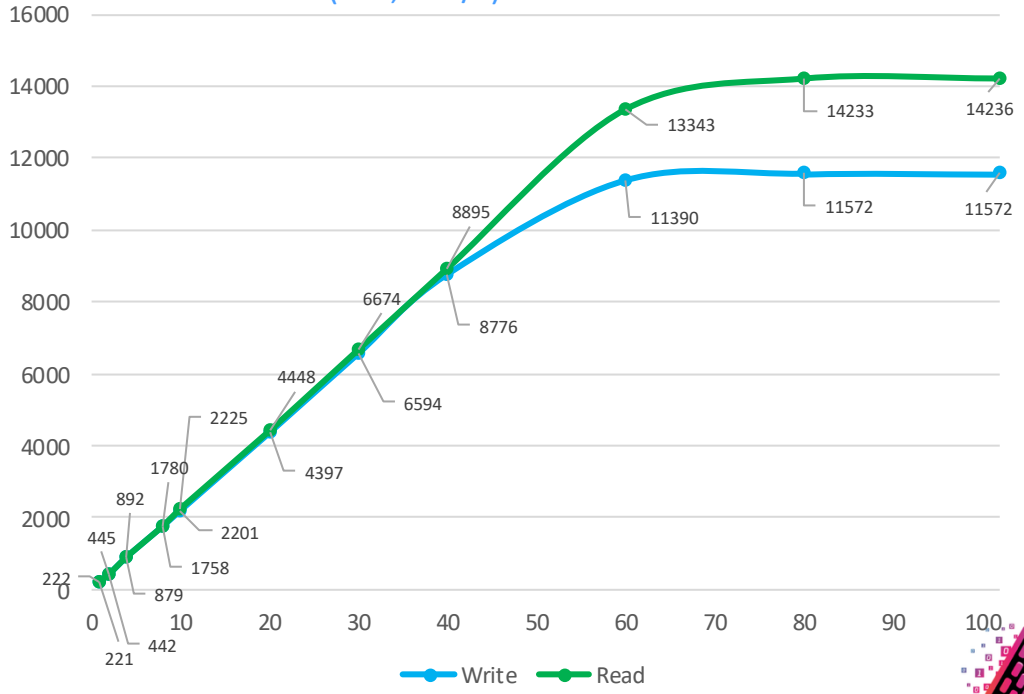
Sequential Read and Write Performance vs. Number of FIO Jobs (raw, MB/s)

14GB/s per Data102

With a single ATTO HBA WD Data102 sustains ~14,236GB/s reads and 11.572Gb/s writes.

Flat = saturated

WDC plateaus immediately — HBA/interface is the limit, not the drive.



LeilFS 5.x SMR Performance

Benchmarked and tested in cooperation with

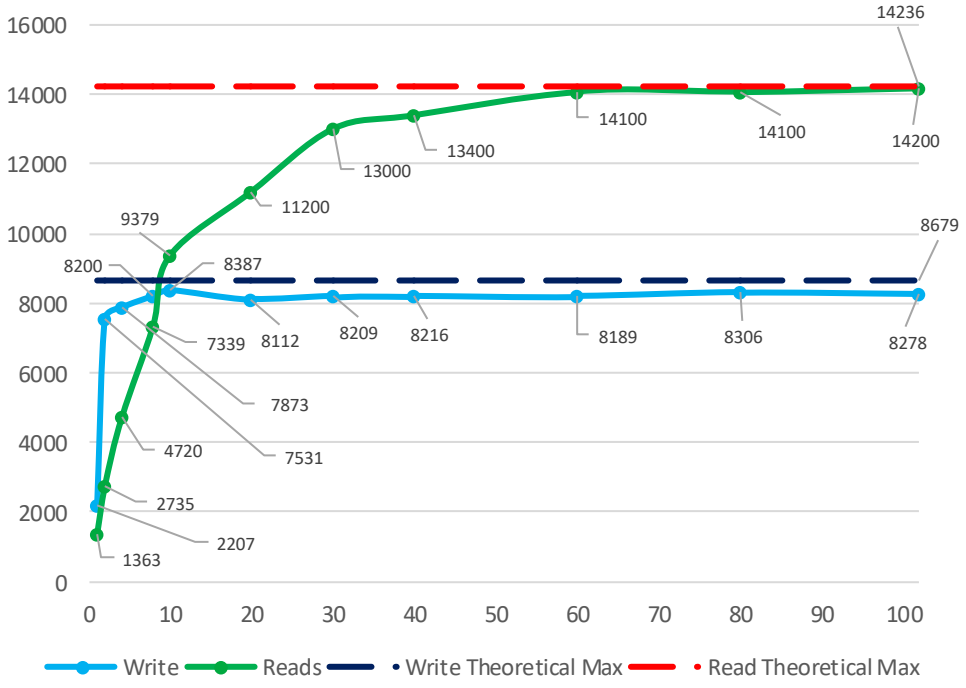
Poznańskie Centrum Superkomputerowo-Sieciowe
Poznan Supercomputing and Networking Center

99.7 % Read Efficiency

At 102 jobs, LeilFS sustains 14.2 GB/s Reads vs. 14.236 GB/s Raw performance.

95.4% Write Efficiency

At 102 jobs, LeilFS sustains 8.278 GB/s Writes vs. 8.679 GB/s Raw performance.



Takeaways: The Numbers

SMR isn't the bottleneck — paired with the right stack, it nearly matches raw HDD throughput.

PILLAR 01

Choose SMR tech wisely

+50% writes / +9% reads — HAMR beats ePMR at 30TB (EC6+2, 8 drives). Only a zone-aware stack converts that capacity into near-native HDD performance.

PILLAR 02

Standardize on SAS4

+14–16% throughput — SAS4 24G 2-port over SAS3 12G 4-port at 60 drives. Free headroom on the same sizing.

PILLAR 03

The file system is your biggest lever — don't stay hardware-agnostic

+57% write throughput — from zone-assignment tuning alone on identical hardware, for just a 7% read trade-off.

HDD-Native vs Hardware Agnostic

We Speak the Language of Physics

+20%

Capacity

Higher areal density from SMR, plus mixed CMR/SMR and variable-capacity support — about +1 TB usable per drive.

99%

Performance

Near-raw HDD throughput that scales linearly with every drive added — plus command duration limits for predictable latency.

+40%

Reliability

Longer drive life via flexible erasure coding, hot-spot avoidance, and head depopulation.

20%

Sustainability

Lower power per TB from SMR, ICE idle power-off (PIN3), and a lean code path that trims compute overhead.

MSST

Questions?

Thank you.

And Let’s Talk.

The International Conference on Massive Storage Systems and Technology

Santa Clara University

June 1, 2026