LCIO: Large Scale Filesystem Aging

Matthew Bachstein
University of Tennessee, Knoxville
mbachste@vols.utk.edu

Feiyi Wang, Sarp Oral
Oak Ridge National Laboratory

Motivation

Summit Procurement

“Wouldn’t it be great if we could test what they just said.”

LifeCycle I/O = LCIO

Why Filesystem Aging?

Benchmarks answer:

> “How does the system perform in the current state?”

Filesystem performance => evolution of state over time.

“Benchmarking empty file systems cannot provide an accurate assessment of the real-world behavior of a file-system architecture”

Why Aging?

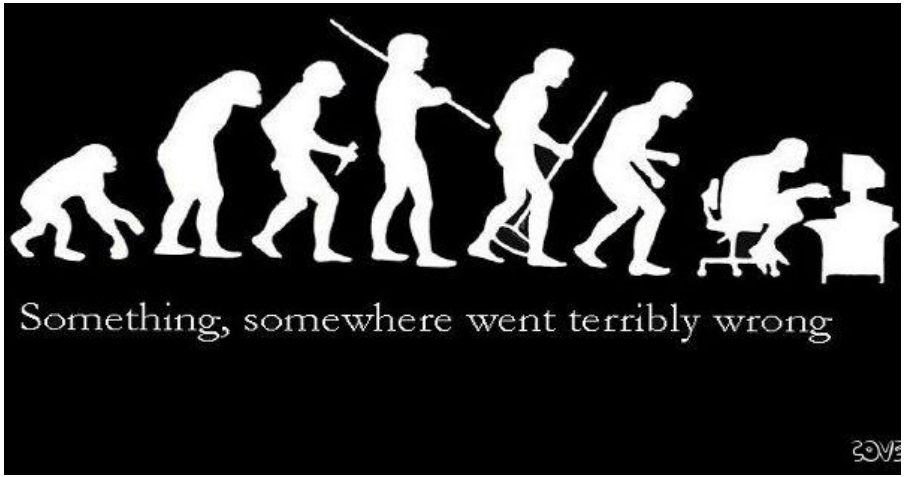
However, aging is quite difficult.

Have to touch the disk.

Two primary approaches exist:

Trace Replays

Probabilistic Convergence



Trace Replays

Smith & Seltzer used a trace replay.

Pros:

- Very accurate, guaranteed to generate the same evolution of states

Cons:

- Slow, effectively serial, relies on a tracefile (which may be huge)
- I.e. probably fine for a desktop system, probably not good for a PFS

Probabilistic Convergence

Probabilistically generated synthetic traces

Approach taken by Impressions, Geriatrix

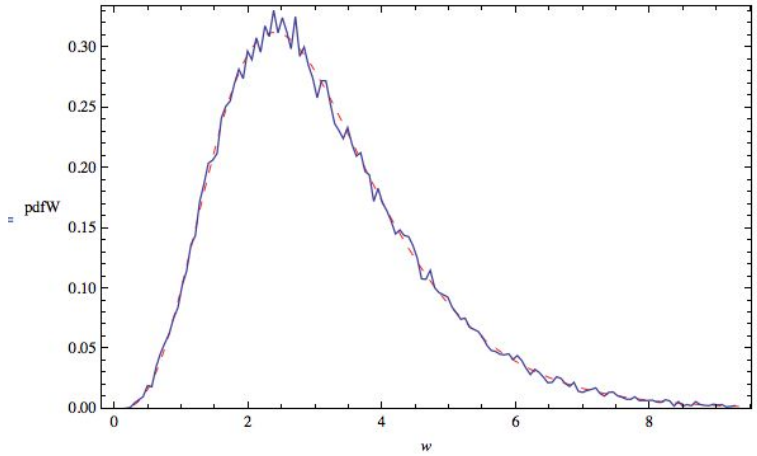
Pros:

Faster, parallelizable, less input data

Amenable to future projections

Cons:

Garbage in -> Garbage out



Previous Work

Impressions

- Statistically realistic metadata (correct tree depth, file extensions and contents, etc)
- ~ 30 mins for 12 GB of data (52,000 files, 4000 dirs)

Geriatrix

- Unique in that it focuses on converging to a file age distribution.
- ~7 hours (420 mins) to age a 90 GB file system

Limitations of Previous Work

Not designed with parallel file systems in mind

These tools are great for desktop-like systems

- Their utility in HPC is limited.

Need to respect the standard HPC guidelines

- Scalability and performance

LCIO Approach

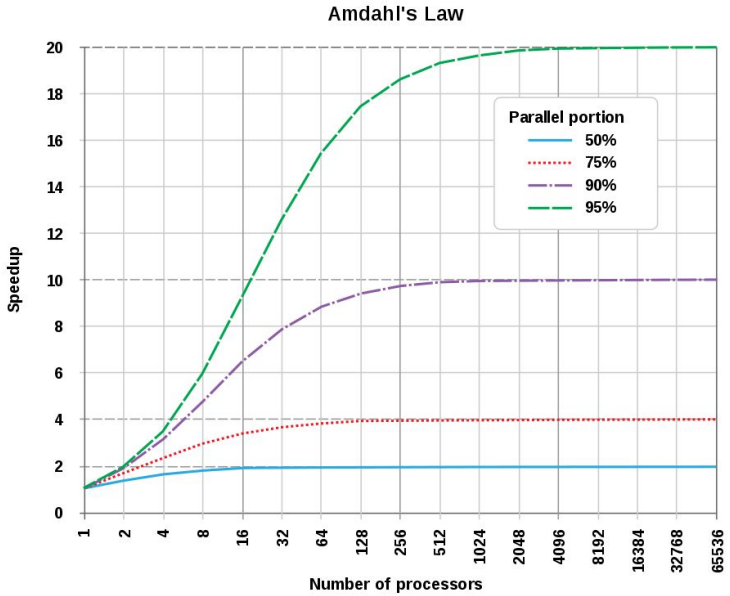
Focus on file size distribution

Pros:

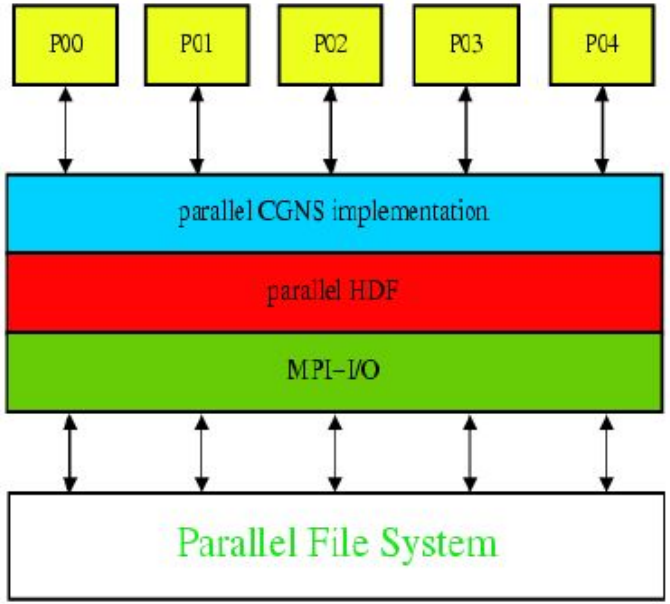
Easily paralllelizeable

Cons:

We trade off realism



LCIO Approach



Other minutia:

Epochs and ops

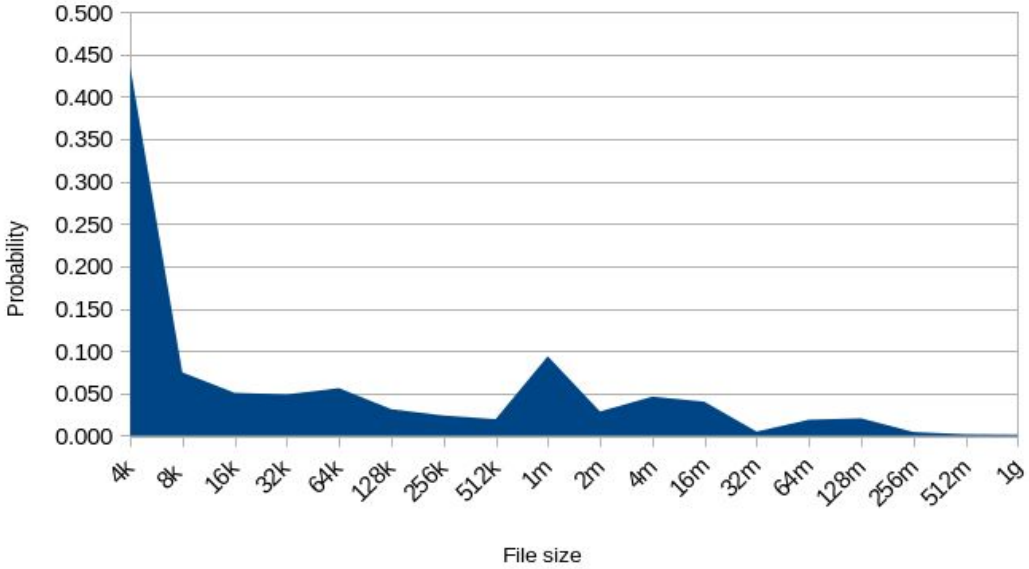
MPIIO & POSIX

Fallocate, fsync, etc….

LCIO Evaluation

This distribution is from a large scale (32 PB) Lustre filesystem at OLCF

The distribution was truncated to 1GB files from 4TB, miniscule tail probability



LCIO Evaluation

Evaluated on a Test and Development, IBM Spectrum Scale, 3.2 PB capacity

Exponential increases, 10k - 1 billion files, POSIX only

128 processes / run, 256 for the 1 billion run

After each run, IOR and MDtest were run to evaluate how the state changed

The 100m and 1b scale runs timed out (8 hour job limit on TDS)

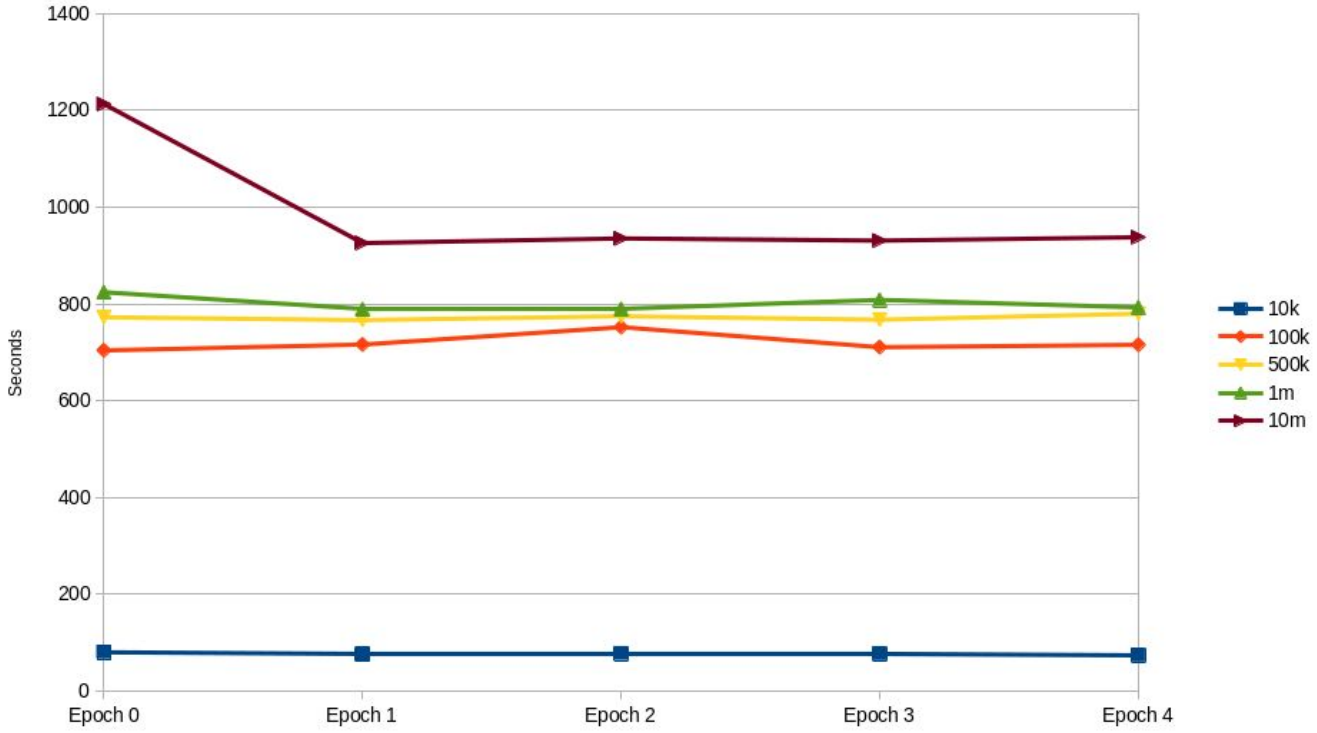
- Time to completion can be extrapolated from amount completed

Each rank had to write a minimum of 400 GiB of traffic

Data Per Rank

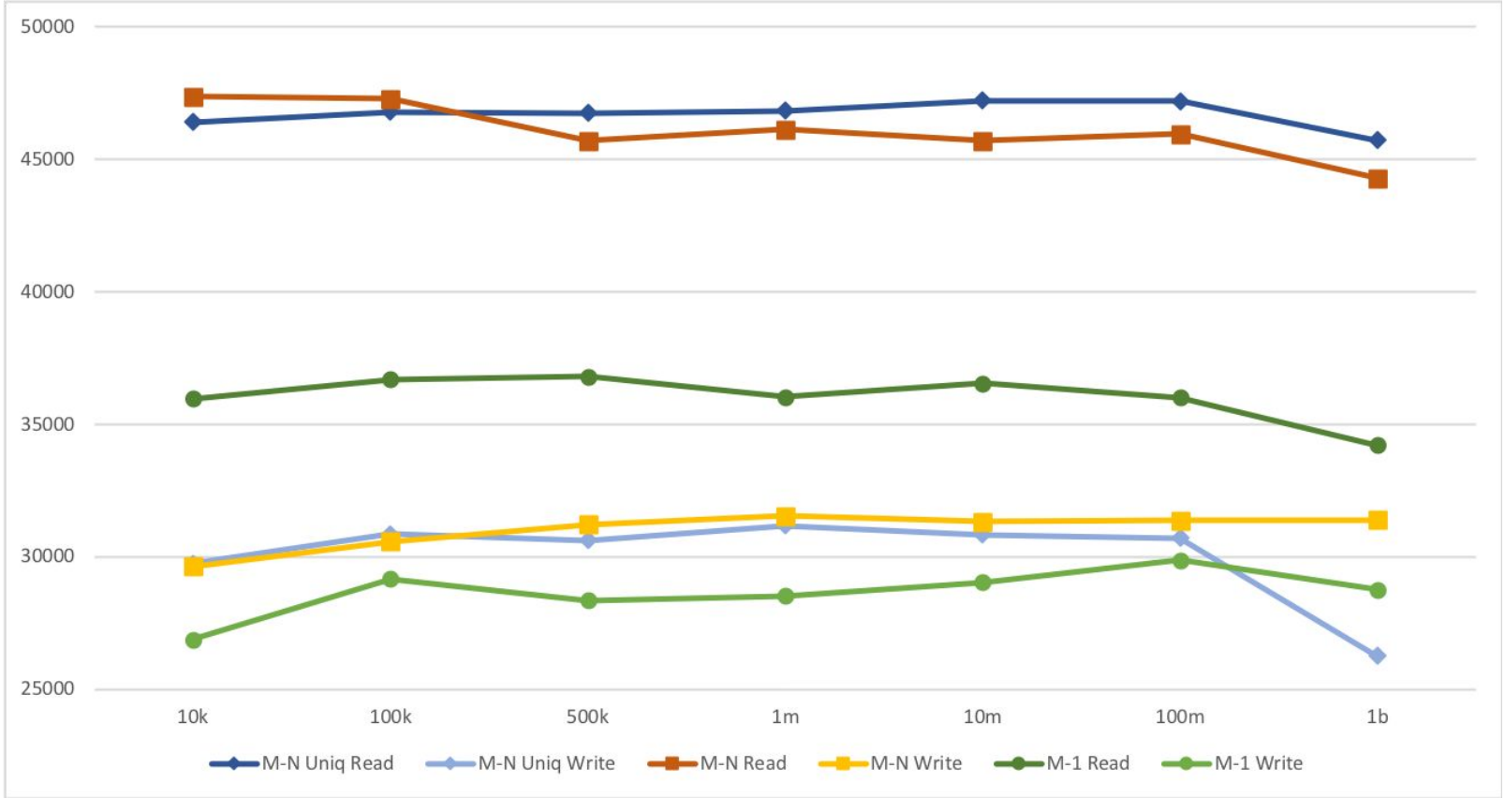
Number of Files	Avg Data written per rank	Total Execution Time (s)	Time for Initial Image (s)
10k	442 GiB	3,712	7.03
100k	453 GiB	3,678.6	80.4
500k	480 GiB	4,247	386.5
1m	530 GiB	4,710.3	706.9
10m	1.14 TiB	12,515.9	7,573.2
100m	Timeout	Timeout	[72,040]
1b	Timeout	Timeout	[288,160]

Time Per Epoch

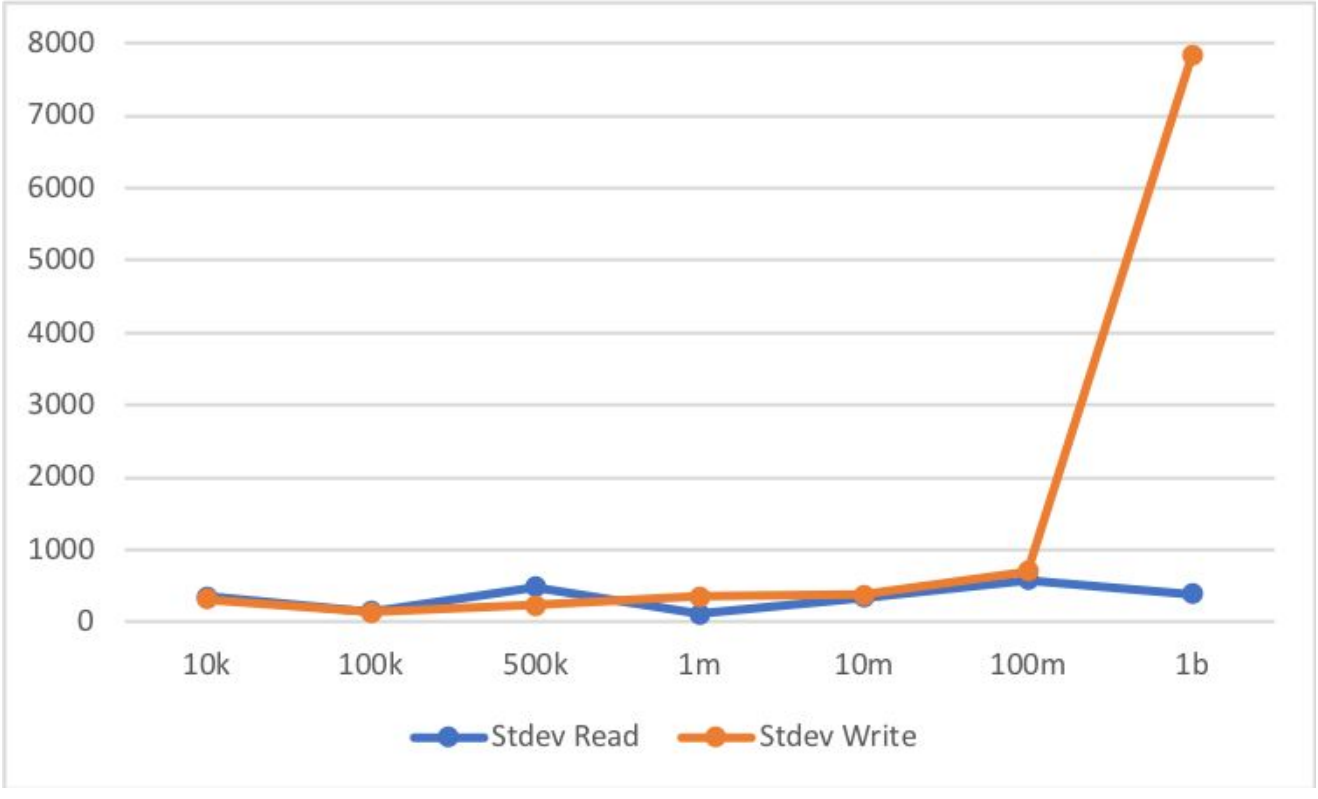


Utilization after Aging (TDS 3.2 PB system)

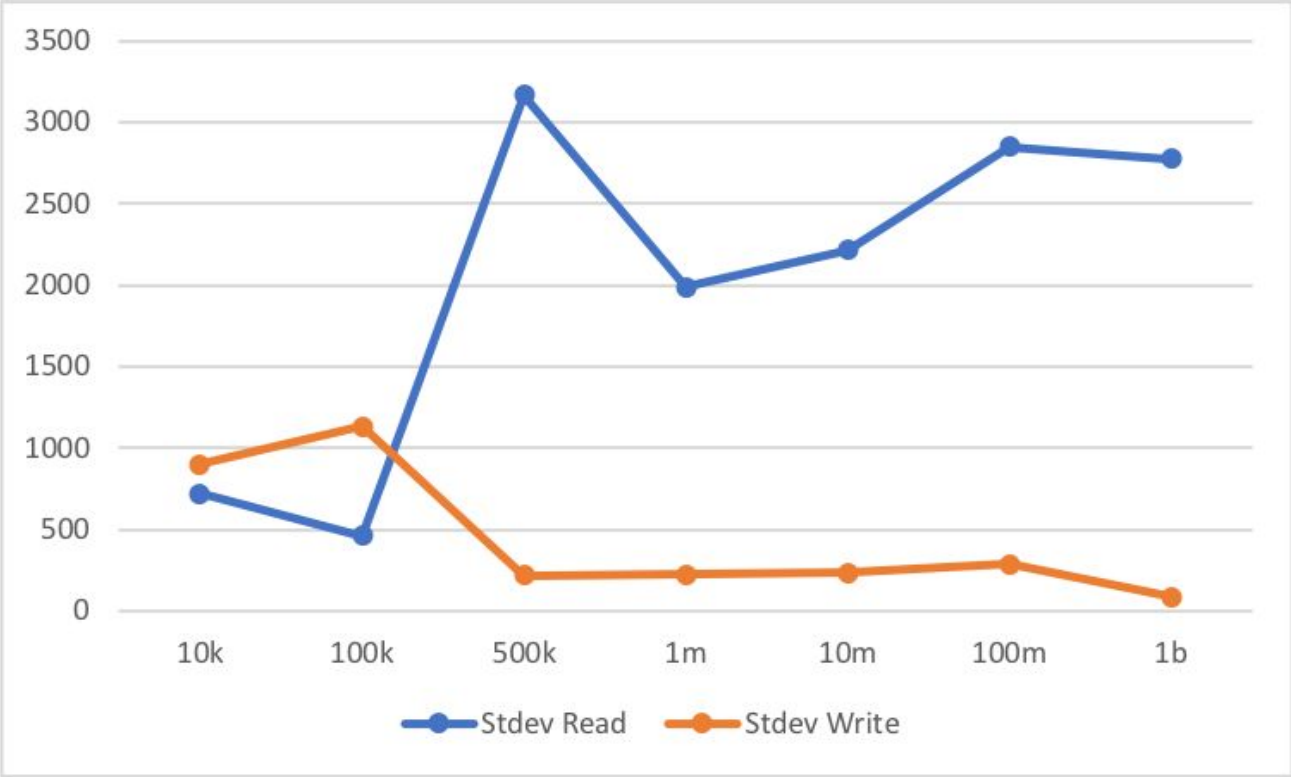
Number of Files	Used FS capacity (%)	Used inodes (%)
Base	53.3 (1.71 PB)	68.0
10k	53.3	68.0
100k	53.3	68.0
500k	53.4	68.1
1m	53.7 (1.72 PB)	68.3
10m	56.5 (1.81 PB)	70.0
100m	63.9 (2.05 PB)	75.2
1b	81.7 (2.61 PB)	87.3


50000
45000
40000
35000
30000
25000
10k
100k
500k
1m
10m
100m
1b
M-N Uniq Read
M-N Uniq Write
M-N Read
M-N Write
M-1 Read
M-1 Write

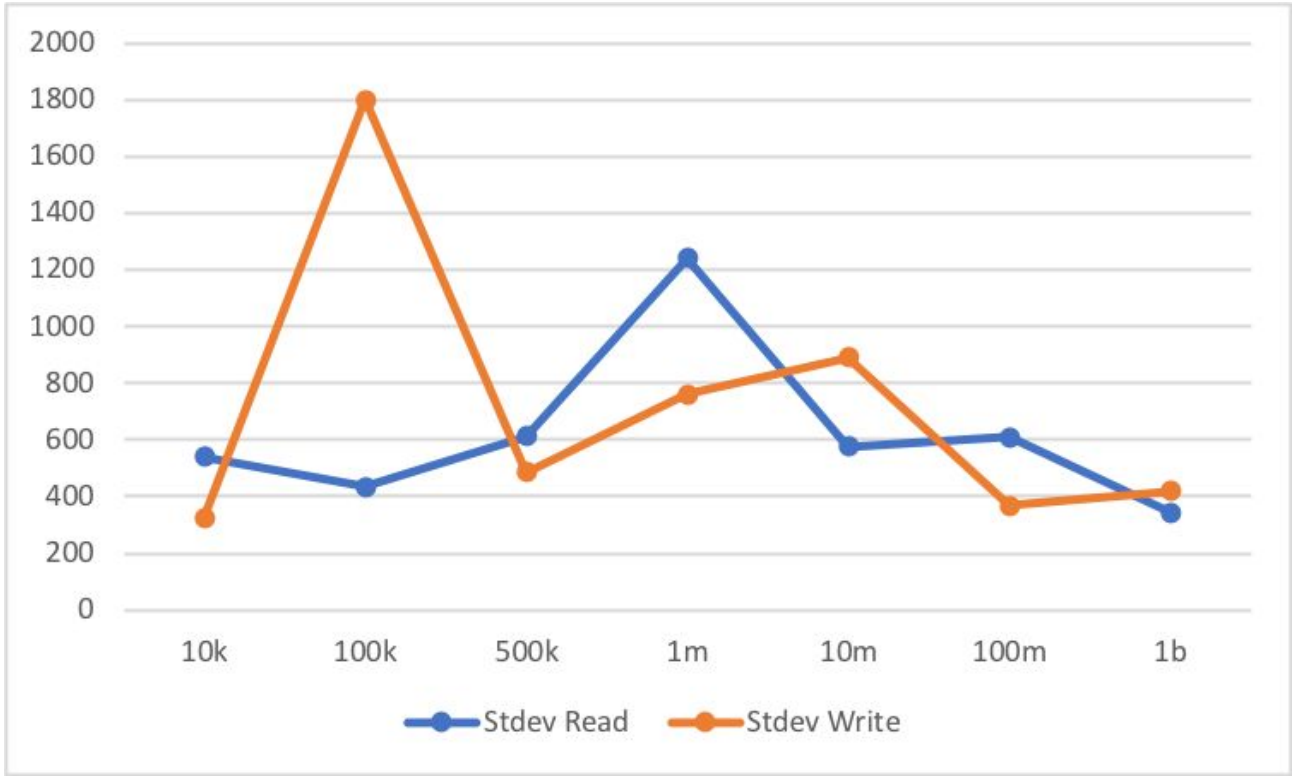
What does IOR tell us now? (M-N unique dir)



What does IOR tell us now? (M-N)



What does IOR tell us now? (M-1)



Conclusions

File system benchmarking is complicated

- Have to consider aging
- Be careful with means -> consider variance

LCIO ages large parallel file systems, in a flexible, scalable manner

Designed to be used alongside other tools (MDtest, IOR, etc)

References

Agrawal, N., Arpaci-Dusseau, A.C., Arpaci-Dusseau, R.H.: Generating realistic impressions for file-system benchmarking. Trans. Storage, 5(4), 16:1–16:30 (Dec 2009). https://doi.org/10.1145/1629080.1629086, http://doi.acm.org/10.1145/1629080.1629086

Axboe, J.: FIO: Flexible I/O Tester.https://github.com/axboe/fio

Conway, A., et al: File systems fated for senescence? nonsense, says science! In: 15th USENIX Conference on File and Storage Technologies (FAST 17). pp. 45–58. USENIX Association, Santa Clara,CA (2017), https://www.usenix.org/conference/fast17/technical-sessions/presentation/conway

IBM: IBM Spectrum Scale.https://en.wikipedia.org/wiki/IBM_Spectrum_Scale (2018)

Kadekodi, S., Nagarajan, V., Ganger, G.R.: Geriatrix: Aging what you see and what you don't see. a file system aging approach for modern storage systems. In: 2018 USENIX Annual Technical Conference (USENIX ATC 18). pp. 691–704.USENIX Association, Boston, MA (2018),https://www.usenix.org/conference/atc18/presentation/kadekodi

LLNL:IOR HPC Benchmark.https://www.nersc.gov/users/computational-systems/cori/nersc-8-procurement/trinity-nersc-8-rfp/nersc-8-trinity-benchmarks/ior/"(2017)

References (cont.)

(NERSC), N.E.R.S.C.C.:MDtest. https://www.nersc.gov/users/computational-systems/cori/nersc-8-procurement/trinity-nersc-8-rfp/nersc-8-trinity-benchmarks/mdtest (2013)

OLCF, O.R.L.C.F.: SPIDER Storage System.https://www.olcf.ornl.gov/olcf-resources/data-visualization-resources/spider/ (2018)

OpenSFS: Lustre. http://lustre.org/documentation/ (2018)

Smith, K.A., Seltzer, M.I.: File system aging - increasing the relevance of file system benchmarks. SIGMETRICS Perform. Eval. Rev. 25(1), 203–213 (Jun 1997). https://doi.org/10.1145/258623.258689, http://doi.acm.org/10.1145/258623.258689

Traeger, A., Zadok, E., Joukov, N., Wright, C.P.: A nine year study of file system and storage benchmarking. Trans. Storage 4(2), 5:1–5:56 (May2008). https://doi.org/10.1145/1367829.1367831,http://doi.acm.org/10.1145/1367829.1367831

12. Vazhkudai, S.S., et al: The design, deployment, and evaluation of the coral pre-exascale systems. In: Proceedings of the International Conference for High Per-formance Computing, Networking, Storage, and Analysis. pp. 52:1–52:12. SC '18,IEEE Press, Piscataway, NJ, USA (2018),http://dl.acm.org/citation.cfm?id=3291656.329172

Wang, F., Sim, H., Harr, C., Oral, S.: Diving into petascale production file systems through large scale profiling and analysis. In: Proceedings of the 2Nd Joint International Workshop on Parallel Data Storage & Data Intensive Scalable Computing Systems. pp. 37–42. PDSW-DISCS '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3149393.3149399, http://doi.acm.org/10.1145/3149393.3149399