Compression techniques for BioFilm Optical Coherence Tomography (OCT) Data

Author: Max H. Faykus III
Email: mfaykus@clemson.edu
Advisors: Dr. Melissa C. Smith, Dr. Jon C. Calhoun

ABSTRACT

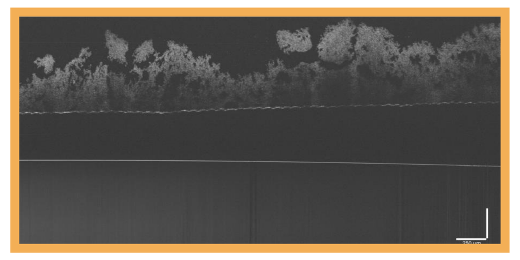
•Biofilms on ship hulls create massive drag penalties, but are smoothed and stabilized with natural marine microbes[7]
•Optical Coherence Tomography (OCT) is a fast and non-destructive technology to generate image data (>100 TB) of the biofilm [8,12,10]
•OCT data causes difficulties in storage and data transmissions[2]
•HPC is needed for data reduction to improve long term storage and data transfers
•The 3D structure of the OCT is leveraged in pre-processing to improve data reduction
•Lossy and lossless data reduction methods are utilized to compress OCT data
•We analyze lossy and lossless compressors in conjunction with preprocessing methods to determine largest compression ratio and maintain a high bandwidth

ARCHITECTURE

OCT data structure

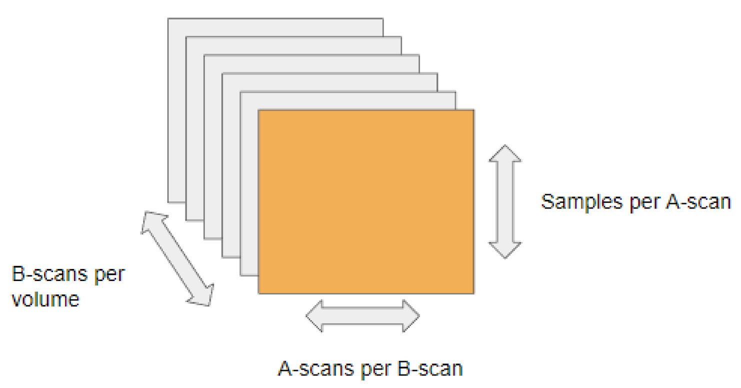
- OCT sends a beam of light at the biofilm, and backscatter echos are visualized.
- Time gating the series of echos to a depth profile line is a A-scan
- Be repeating this step over the sample on the xy-plane generates a 2D depth image which is the B-scan
- The series of B-scans creates a 3D volume scan

OCT Biofilm scan



A-scan

OCT File Structure

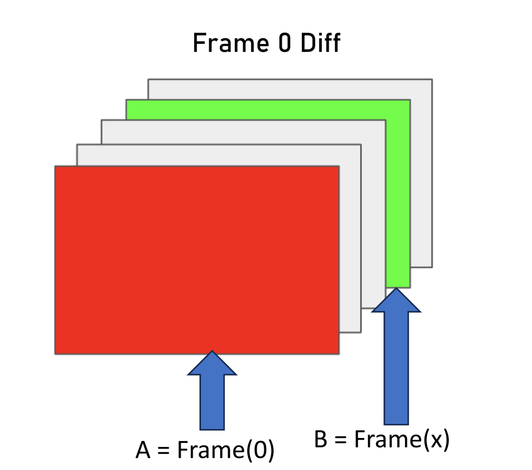


The data is saved as a series of tiff images that are joined together in this format to create the volume scan

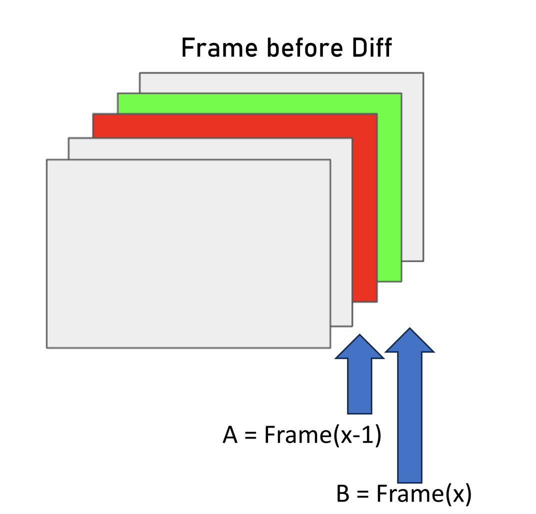
Data Pre-Processing

In order to leverage the 3D nature of OCT images, we explored two difference methods to leverage OCT data to improve the compression ratio and maintain compression and decompression bandwidth.

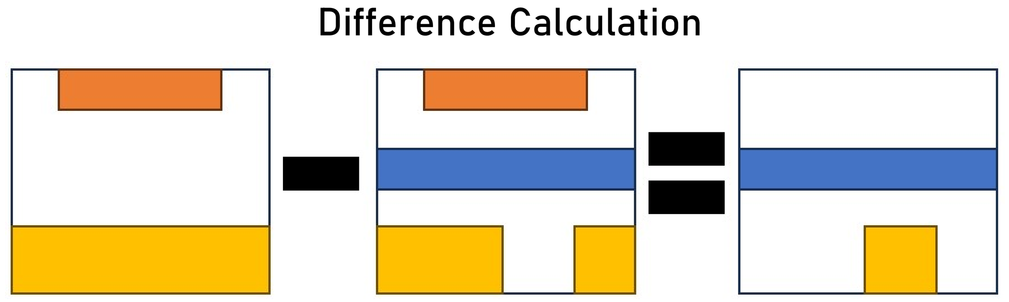
Frame 0 Diff (0Diff): The difference is calculated between the first frame of the OCT image and every subsequent image frame



Frame before Diff (Ldiff): The difference of each frame is taken from the preceding frame in the image



Difference Calculation



Removes similar features between the OCT frames to increase the number of zeros in the data

Methods

- The evaluations are performed on 13 different Biofilm experiments
- Each experiment is stored with each B-scan as a separate tiff file
- For lossy compression, the 0-255 uint8 image data is normalized to 0-1 float32 leveraging SZ and ZFP bounds
- Ssim is used over PSNR because the A-scan has a single structure across the image and the rest is background noise, The structural integrity of the biofilm is important

Accuracy metric

$$ssim(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

The structural similarity index measure (ssim) is a method of predicting similarity/quality between two different images[11]

Testing Methodology

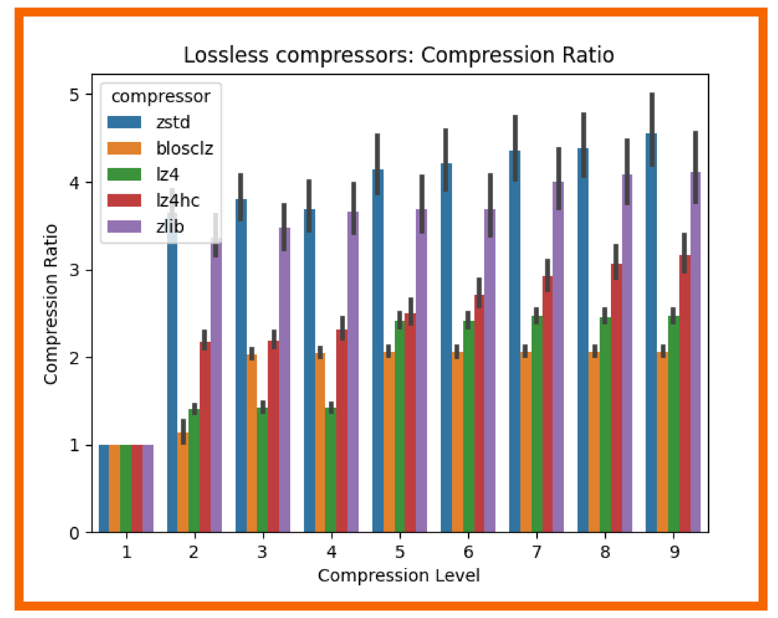
All Experiments run on the Clemson University Palmetto Cluster

Biofilm	Format
Volume	1000x505x50
1000	Samples per A scan
505	A-scans per B-scan
50	# of B-scans

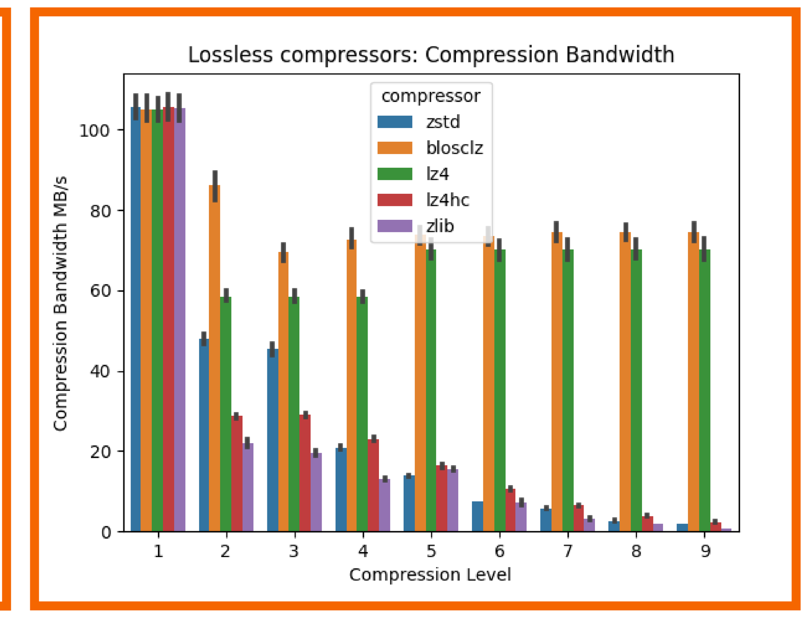
File System	Indigo
CPU	Intel® Xeon® Gold 6258R CPU @ 2.70GHz
Architecture	X86_64
GCC	12.1.0
SZ Version[6]	2.1.12
ZFP Version[9]	1.0.0
ZSTD Version[4]	1.5.5
LZ4 Version[3]	1.9.4
ZLIB Version[5]	1.2.13
BLOSC Version[1]	1.21.2
LIBPRESSIO Version[13]	0.94.0

RESULTS

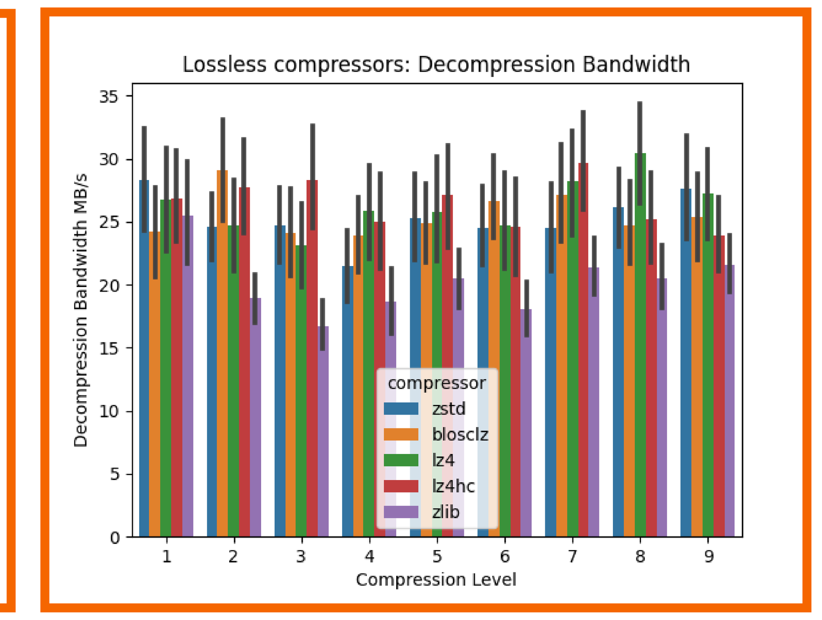
Lossless Compression Evaluation



Tested and compared multiple lossless compressors and compared the compression ratio CR, Zstd achieves CR ~4.5x



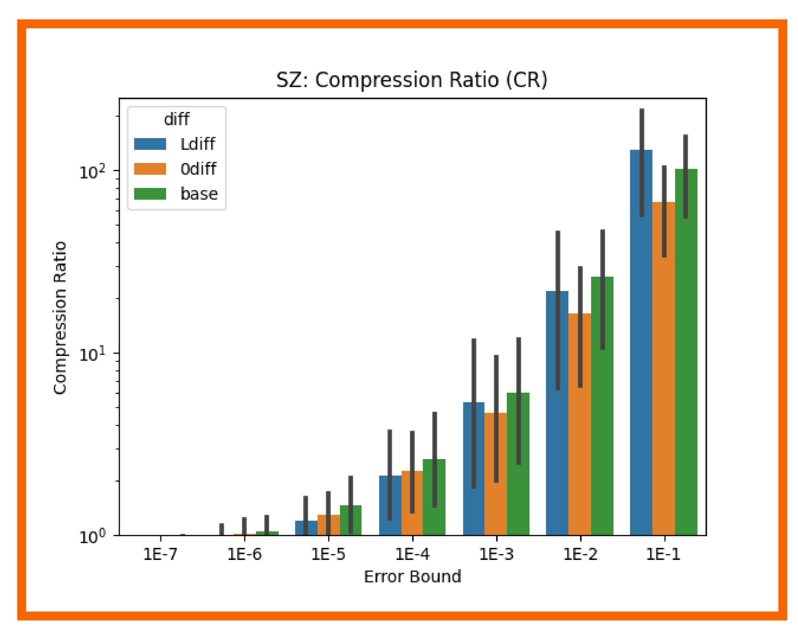
Compression bandwidth. blosclz achieves the highest compression bandwidth

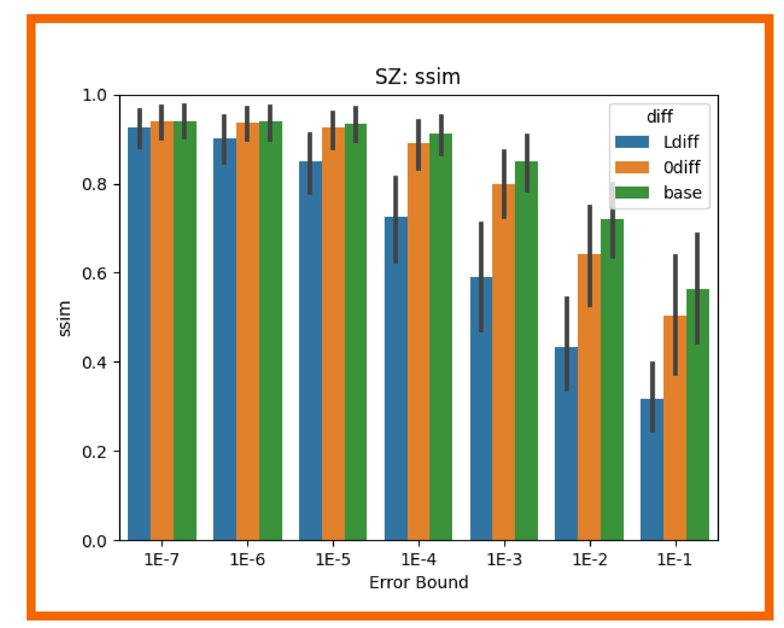


Decompression bandwidth lz4/lz4hc achieves the fastest decompression bandwidth

Pre-processing Evaluation

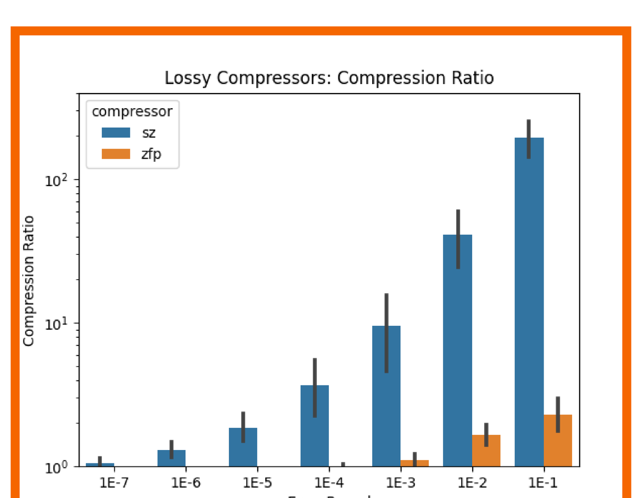
SZ is the only compressor that leverages pre-processing improving CR on higher error bounds; Ldiff achieved the highest CR at 927.8x



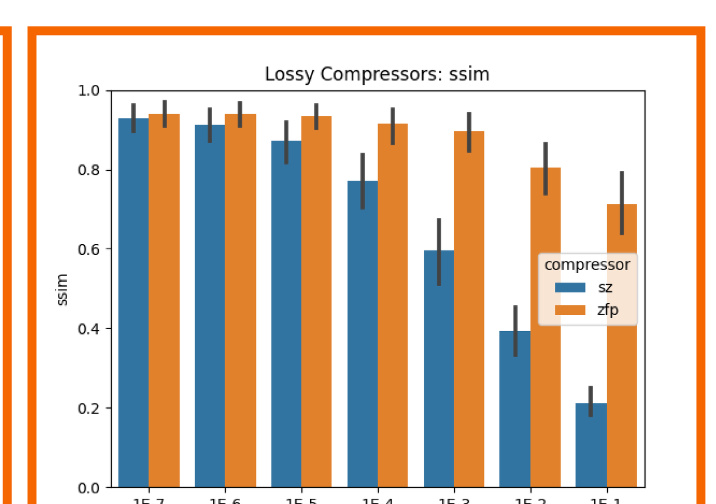


Ssim drops as the error bound increases, on higher bounds the pre-processing reduces ssim

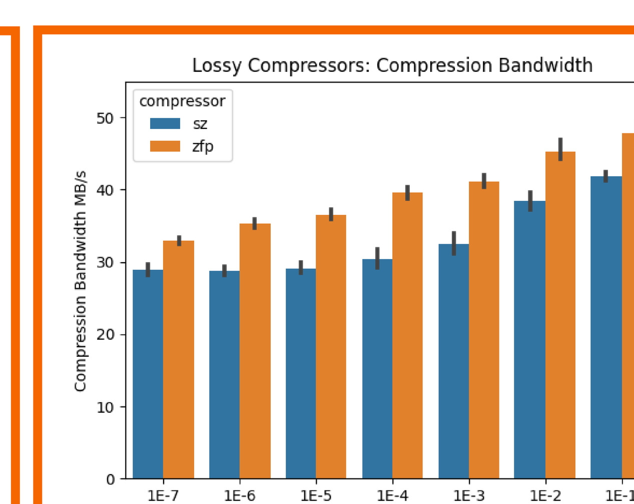
Lossy compression evaluation

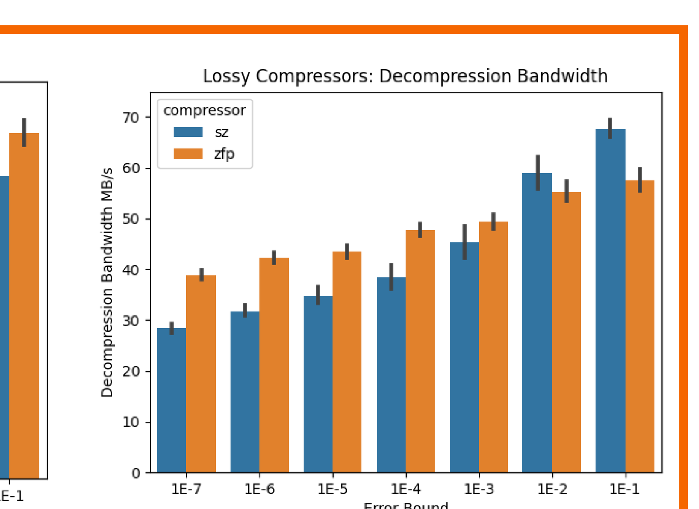


SZ achieves a higher CR when compared to ZFP



SZ ssim degrades drastically on higher error bounds ZFP





ZFP achieves a higher compression bandwidth and decompression bandwidth compared to SZ

CONCLUSIONS

- For lossless compressors zstd achieves the largest CR ~4.5x while performing average in compression bandwidth and decompression bandwidth
- Lossy compression trades data distortion for smaller file size
- Lossless trades compression time for smaller file size
- Pre-processing was able to improve the overall achievable compression ratio
- Overall, OCT biofilm data is highly compressible utilizing pre-processing and lossy compression; with Ldiff preprocessing on SZ achieves the highest compression ratio
- Pre-processing in conjunction with lossy compression improves the ability to analyze large OCT datasets

References

[1] Francesc Alted. 2017. Blosc, an extremely fast, multi-threaded, meta-compressor library.
[2] Josef F Bille. 2019. High resolution imaging in microscopy and ophthalmology: new frontiers in biomedical optics. (2019).
[3] Yann Collet. 2011. LZ4 lossless compression algorithm. (2011).
[4] Yann Collet and Murray Kucherawy. 2018. Zstandard Compression and the application/zstd
[5] Peter Deutsch and Jean-Loup Gailly. 1996. Zlib compressed data format specification version 3.3. Technical Report.
[6] Sheng Di and Franck Cappello. 2016. Fast error-bounded lossy HPC data compression with SZ. In 2016 ieee international parallel and distributed processing symposium (ipdps). IEEE, 730–739
[7] Andrea Farkas, Soonseok Song, Nastia Degiuli, Ivana Martić, and Yigit Kemal Demirel. 2020. Impact of biofilm on the ship propulsion characteristics
and the speed reduction. Ocean Engineering 199 (2020), 107033.
[8] David Huang, Eric A Swanson, Charles P Lin, Joel S Schuman, William G Stinson, Warren Chang, Michael R Hee, Thomas Flotte, Kenton Gregory, Carmen A Puliafito, et al. 1991. Optical coherence tomography. science 254, 5035 (1991), 1178–1181.
[9] Peter Lindstrom. 2014. Fixed-rate compressed floating-point arrays. IEEE transactions on visualization and computer graphics 20, 12 (2014), 2674–2683.
[10] Michael Wagner and Harald Horn. 2017. Optical coherence tomography in biofilm research: a comprehensive review. Biotechnology and Bioengineering114, 7 (2017), 1386–1402.
[11] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13, 4 (2004), 600–612.
[12] Zahid Yaqoob, Jigang Wu, and Changhuei Yang. 2005. Spectral domain optical coherence tomography: a better OCT imaging strategy. Biotechniques
39, 6 (2005), S6–S13.
[13] R. Underwood, V. Malvoso, J. C. Calhoun, S. Di, and F. Cappello, "Productive and performant generic lossy data compression with libpressio "in 2021 7th International Workshop on Data Analysis and Reduction for Big Scientific Data (DRBSD-7), pp. 1–10, IEEE, 202

Clemson University is acknowledged for generous allotment of compute time on the Palmetto cluster.
This material is based on work supported by the National Science Foundation under Grant Nos. MRI# 2024205, MRI# 1725573, MRI# 1725573, and CRI# 2010270..