Datasets for Benchmarking Floating-Point Compressors

Fabian Knorr, Peter Thoman and Thomas Fahringer
Distributed and Parallel Systems Group
University of Innsbruck, Austria
{fabian,petert,tf}@dps.uibk.ac.at

Abstract

Compression of floating-point data, both lossy and lossless, is a topic of increasing interest in scientific computing. Developing and evaluating suitable compression algorithms requires representative samples of data from real-world applications. We present a collection of publicly accessible sources for volume and time series data as well as a list of concrete datasets that form an adequate basis for compressor benchmarking.

Introduction

Efficient compression of floating-point stream [1][2][3] and volume data [4][5][6] has seen significant advances in the past years. Finding competitive trade-offs between the compression ratios achieved and the computational power invested is the primary objective of this research domain. Actually benchmarking those quantities relies on suitable input data, which must be similar to the real-world use case of each algorithm.

Research groups usually rely on their own collection of test data, often without sufficient references to their sources. This severely limits the comparability of results from different authors, sometimes even rendering precise reproduction impossible.

In this paper, we present a selection of publicly available datasets with appropriate references to aid future publications on the matter.

Public Data Sources

Infrared Science Archive The NASA/IPAC Infrared Science Archive (IRSA)[1] is an image data archive for astronomical infrared and submillimeter missions. Among others, it serves images from the Spitzer Space Telescope, which saw prior use in compression research [1]. Data is available in the FITS format [7], an image format common in astronomy that supports floating-point data. The IRSA is funded by the National Aeronautics and Space Administration and operated by the California Institute of Technology.

Spitzer Space Telescope's First Look Survey (FLS) [8] and Frontier Fields [9] data archives contain two-dimensional image files of various sizes in the FITS single-precision floating-point format[2][3].

[1] https://irsa.ipac.caltech.edu

[2] https://irsa.ipac.caltech.edu/data/SPITZER/FLS/images

[3] https://irsa.ipac.caltech.edu/data/SPITZER/Frontier/images

Radio Telescope Data Center The Radio Telescope Data Center (RTDC)[4] of the Smithsonian Astrophysical Observatory serves data from various US radio telescope sites including the Submillimeter Array and the CfA Millimeter-Wave Telescope.

The Submillimeter Array (SMA) is a radio telescope interferometer for submillimeter wavelength observations located in Hawaii [10]. It is a joint project between the Smithsonian Astrophysical Observatory and the Academia Sinica Institute of Astronomy and Astrophysics and is funded by the Smithsonian Institution and the Academia Sinica.

The CfA Millimeter-Wave Telescope at the Harvard–Smithsonian Center for Astrophysics is an observatory for interstellar molecular clouds [11].

Observational data for both is available as three-dimensional, single-precision FITS floating-point images[5][6].

Hubble Legacy Archive The Hubble Legacy Archive (HLA)[7] provides a large selection of imagery from the Hubble space telescope in the FITS image format. Is a joint project of the Space Telescope Science Institute (STScI), the Space Telescope European Coordinating Facility (ST-ECF), and the Canadian Astronomy Data Centre (CADC).

IEEE SciVis Contest The IEEE Scientific Visualization Contest[8] is a yearly installment of the IEEE Visualization Conference (VIS). Simulation datasets from various domains are provided to contestants who submit their own approaches to visualizing them. Some of these datasets are multi-dimensional, binary single-precision floating-point data:

- SciVis 2004: 3D time steps from simulation of a hurricane from the National Center for Atmospheric Research in the United States[9].
- SciVis 2006: 3D time steps from an earthquake simulation[10]
- SciVis 2018: 3D time steps from a of deep water asteroid impact simulation[11].
- SciVis 2020: 4D time-step tiles from a simulation of complex eddy transport mechanisms eddies in the Red Sea [12][12].

HDRI Haven HDRI Haven[13] is a source for high-resolution, high-dynamic range (HDR) photographs. Images are available in the Radiance HDR format (also known

[4] `https://www.cfa.harvard.edu/rtdc`
[5] `https://www.cfa.harvard.edu/rtdc/SMAimages`
[6] `https://www.cfa.harvard.edu/rtdc/CO/CompositeSurveys`
[7] `http://hla.stsci.edu/`
[8] `http://sciviscontest.ieeevis.org`
[9] `http://sciviscontest.ieeevis.org/2004/data.html`
[10] `http://sciviscontest.ieeevis.org/2006/download.html`
[11] `https://sciviscontest2018.org/`
[12] `https://kaust-vislab.github.io/SciVis2020/data.html`
[13] `https://hdrihaven.com`

as RGBE), a shared-exponent floating-point image format. These can be converted to univariate full-width floating point grids by re-mapping the color space, e.g. extracting the luminance component.

Open Scientific Visualization Datasets Pavol Klacansky provides an online repository of multidimensional simulation data from various domains[14]. Both single- and double precision datasets exist.

Scientific IEEE 754 Floating-Point Datasets Martin Burtscher provides test data used in his compression research online. There exist separate pages for single-precision floating-point datasets[15] used for evaluating the SPDP compressor [2] and double-precision floating-point datasets[16] used, among others, for evaluating the FPC algorithm [1].

UCI Machine Learning Repository The UCI Machine Learning Repository is a collection of databases, domain theories and data generators for the empirical analysis of machine learning algorithms. [13] [17]. Some datasets contain time series data, useful for benchmarking one-dimensional compressors.

Freesound Freesound[18] is a collaborative database of audio snippets, samples and recordings. Some data is offered as 32-bit floating-point PCM audio, which corresponds to a single-precision time series.

University of Innsbruck Simulation data collected at the University of Innsbruck (UIBK) for the purpose of compressor benchmarking can be found on the DPS group website[19].

Test Datasets

We propose the collection of data samples from Figure 1 as a baseline for compressor benchmarking. Where no single-precision equivalent exists, double-precision datasets can be truncated to single-precision if the source data range allows it.

- `msg_sppm` and `msg_sweep3d`, taken from Martin Burtscher's repository, are numeric messages sent by a node in a parallel system running ASCI Purple solvers.
- `snd_thunder` is a 32-bit float PCM audio recording of thunder, obtained from Freesound user "Guialgarve"[20].

[14] `https://klacansky.com/open-scivis-datasets`
[15] `https://userweb.cs.txstate.edu/~burtscher/research/datasets/FPsingle`
[16] `https://userweb.cs.txstate.edu/~burtscher/research/datasets/FPdouble`
[17] `https://archive.ics.uci.edu/ml/index.php`
[18] `https://freesound.org`
[19] `https://dps.uibk.ac.at/~fabian/datasets`
[20] `https://freesound.org/people/Guialgarve/sounds/523100`

dataset	source	data type	dimensions	extent
`msg_sppm`	Burtscher	single, double	1	34,874,483
`msg_sweep3d`	Burtscher	single, double	1	15,716,403
`snd_thunder`	Freesound	single	1	7,898,672
`ts_gas`	UCI MLR	single	1	4,208,261
`ts_wesad`	UCI MLR	single	1	4,588,553
`hdr_night`	HDRI Haven	single	2	$8{,}192 \times 16{,}384$
`hdr_palermo`	HDRI Haven	single	2	$10{,}268 \times 20{,}536$
`hubble`	HLA	single	2	$6{,}036 \times 6{,}014$
`rsim`	UIBK	single, double	2	$2{,}048 \times 11{,}509$
`spitzer_fls_irac`	IRSA	single	2	$6{,}456 \times 6{,}389$
`spitzer_fls_vla`	IRSA	single	2	$8{,}192 \times 8{,}192$
`spitzer_frontier`	IRSA	single	2	$3{,}874 \times 2{,}694$
`asteroid`	SciVis 2018	single	3	$500 \times 500 \times 500$
`astro_mhd`	UIBK	single	3	$128 \times 512 \times 1024$
`astro_mhd`	UIBK	double	3	$130 \times 514 \times 1026$
`astro_pt`	UIBK	single, double	3	$512 \times 256 \times 640$
`flow`	Klacansky	double	3	$16 \times 7{,}680 \times 1{,}0240$
`hurricane`	SciVis 2004	single	3	$100 \times 500 \times 500$
`magrecon`	Klacansky	single	3	$512 \times 512 \times 512$
`miranda`	Klacansky	single	3	$1{,}024 \times 1{,}024 \times 1{,}024$
`redsea`	SciVis 2020	double	3	$50 \times 500 \times 500$
`sma_disk`	RTDC	single	3	$301 \times 369 \times 369$
`turbulence`	Klacansky	single	3	$256 \times 256 \times 256$
`wave`	UIBK	single, double	3	$512 \times 512 \times 512$

Figure 1: Proposed test datasets, types and grid sizes

- `ts_gas` is a time series of average temperature-modulated metal oxide gas sensor readings [14] obtained from the UCI Machine Learning Repository[21]. The file contains readings as truncated floating-point values in text form. The readings from all sensors were averaged to obtain a time series with usable precision.

- `ts_wesad` is a time series of average readings from physiological and motion sensors during a stress-affect lab study [15]. The data was obtained from the UCI Machine Learning Repository[22]. The file contains readings as truncated floating-point values in text form. The readings from all sensors were averaged to obtain a time series with usable precision.

- `hdr_night` and `hdr_palermo` are the luminance components of two HDR photographs from HDRI Haven, "Preller Drive"[23] and "Palermo Sidewalk"[24]. The images in the largest available resolution were decoded into single-precision

[21] `https://archive.ics.uci.edu/ml/datasets/Gas+sensor+array+under+dynamic+gas+mixtures`
[22] `https://archive.ics.uci.edu/ml/datasets/Gas+sensor+array+under+dynamic+gas+mixtures`
[23] `https://hdrihaven.com/hdri/?c=night&h=preller_drive`
[24] `https://hdrihaven.com/hdri/?c=outdoor&h=palermo_sidewalk`

floating-point RGB bitmaps and converted to the HSL color space to extract the luminance component.

- `hubble` is an image of the Tadpole Galaxy (UGC10214) from the Hubble Legacy Archive. The dataset id is `hst_8992_03_acs_wfc_f475w`.
- `rsim` is a radiosity field from room response simulation for time-of-flight imaging [16]. After extending the implementation to double-precision arithmetic, we simulated the "Medium" scene to obtain 2048 time steps.
- `spitzer_fls_irac`, `spitzer_fls_vla` and `spitzer_frontier` are images from the Spitzer First Look Survey and Frontier Fields observations. All datasets were obtained from the Infrared Science Archive.
  - `spitzer_fls_irac` was taken by the Spitzer Infrared Array Camera Component (IRAC) [17] as part of the First Look Survey mission. We chose the `chain1_main_mosaic.fits` image.
  - `spitzer_fls_vla` was taken by the Very Large Array (VLA) radio telescope in preparation of the Spitzer First Look Survey Mission [18], for which a single image is available.
  - `spitzer_frontier` is part of the frontier fields observation. Our source is the `MACS0717.IRAC.1.mosaic.fits` file from the MACS0717 dataset.
- `asteroid` is the last time-step of the SciVis 2018 asteroid impact simulation.
- `astro_mhd` is the temperature component of a magnetohydrodynamic simulation of solar wind interactions in the colliding-wind binary system Eta-Carinae [19]. The simulation was performed separately in single and double precision with a slight variation in size.
- `astro_pt` is one velocity vector component of a particle transport simulation in the LS 5039 system. This simulation was performed separately in single and double precision as well.
- `flow` are the last 16 timesteps of the pressure field of a direct numerical simulation of fully developed flow at different Reynolds numbers in a plane channel [20].
- `hurricane` is the precipitation component of time step 35 of the SciVis 2004 hurricane simulation.
- `magrecon` is a single time step from a computational simulation of magnetic reconnection [21].
- `miranda` is one time step of a density field in a simulation of the mixing transition in Rayleigh-Taylor instability [22].
- `redsea` is salt content component from the last time step in the SciVis 2020 contest Red Sea eddy simulation.

- `sma_disk` is observational data of a circumstellar disk from the Submillimeter Array radio inferometer[25].
- `wave` are time steps from a wave propagation simulation on a two-dimensional surface. We modified the `wave_sim` simulation code from the Celerity distributed memory runtime [23] to support double-precision arithmetic and computed 512 time steps.

References

[1] Martin Burtscher and Paruj Ratanaworabhan, "FPC: A high-speed compressor for double-precision floating-point data," *IEEE Transactions on Computers*, vol. 58, no. 1, pp. 18–31, 2008.

[2] Steven Claggett, Sahar Azimi, and Martin Burtscher, "SPDP: An automatically synthesized lossless compression algorithm for floating-point data," in *2018 Data Compression Conference*. IEEE, 2018, pp. 335–344.

[3] Annie Yang, Hari Mukka, Farbod Hesaaraki, and Martin Burtscher, "MPC: a massively parallel compression algorithm for scientific data," in *2015 IEEE International Conference on Cluster Computing*. IEEE, 2015, pp. 381–389.

[4] Lawrence Ibarria, Peter Lindstrom, Jarek Rossignac, and Andrzej Szymczak, "Out-of-core compression and decompression of large n-dimensional scalar fields," in *Computer Graphics Forum*. Wiley Online Library, 2003, vol. 22, pp. 343–348.

[5] Peter Lindstrom and Martin Isenburg, "Fast and efficient compression of floating-point data," *IEEE Transactions on Visualization and Computer graphics*, vol. 12, no. 5, pp. 1245–1250, 2006.

[6] Nathaniel Fout and Kwan-Liu Ma, "An adaptive prediction-based approach to lossless compression of floating-point volume data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2295–2304, 2012.

[7] Donald Carson Wells and Eric W Greisen, "FITS: A flexible image transport system," in *Image Processing in Astronomy*, 1979, p. 445.

[8] Martin J Burgdorf, Martin Cohen, James G Ingalls, S Ramirez, Jeonghee Rho, SR Stolovy, Sean J Carey, SB Fajardo-Acosta, WJ Glaccum, G Helou, et al., "The galactic first-look survey with the Spitzer space telescope," *Advances in Space Research*, vol. 36, no. 6, pp. 1050–1056, 2005.

[9] JM Lotz, A Koekemoer, D Coe, N Grogin, P Capak, J Mack, J Anderson, R Avila, EA Barker, D Borncamp, et al., "The Frontier Fields: Survey design and initial results," *The Astrophysical Journal*, vol. 837, no. 1, pp. 97, 2017.

[10] Paul TP Ho, James M Moran, and Kwok Yung Lo, "The submillimeter array," *The Astrophysical Journal Letters*, vol. 616, no. 1, pp. L1, 2004.

[11] Thomas M Dame, Dap Hartmann, and P Thaddeus, "The milky way in molecular clouds: a new complete co survey," *The Astrophysical Journal*, vol. 547, no. 2, pp. 792, 2001.

[12] Habib Toye, Peng Zhan, Sabique Langodan, George Krokos, Omar Knio, Ibrahim Hoteit, et al., "Impact of atmospheric and model physics perturbations on a high-resolution ensemble data assimilation system of the red sea," in *EGU General Assembly Conference Abstracts*, 2020, p. 6000.

[13] Dheeru Dua and Casey Graff, "UCI machine learning repository," 2017.

[25] `https://www.cfa.harvard.edu/rtdc/SMAimages/080403_034904_hd98800.html`

[14] Jordi Fonollosa, Sadique Sheik, Ramón Huerta, and Santiago Marco, "Reservoir computing compensates slow response of chemosensor arrays exposed to fast varying gas concentrations in continuous monitoring," *Sensors and Actuators B: Chemical*, vol. 215, pp. 618–629, 2015.

[15] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 400–408.

[16] Peter Thoman, Markus Wippler, Robert Hranitzky, and Thomas Fahringer, "RTX-RSim: Accelerated Vulkan room response simulation for time-of-flight imaging," in *Proceedings of the International Workshop on OpenCL*, 2020, pp. 1–11.

[17] M Lacy, G Wilson, F Masci, LJ Storrie-Lombardi, PN Appleton, L Armus, SC Chapman, PI Choi, D Fadda, F Fang, et al., "The infrared array camera component of the spitzer space telescope extragalactic first look survey," *The Astrophysical Journal Supplement Series*, vol. 161, no. 1, pp. 41, 2005.

[18] JJ Condon, WD Cotton, QF Yin, DL Shupe, LJ Storrie-Lombardi, G Helou, BT Soifer, and MW Werner, "The SIRTF first-look survey. i. VLA image and source catalog," *The Astronomical Journal*, vol. 125, no. 5, pp. 2411, 2003.

[19] R Kissmann, K Reitberger, O Reimer, A Reimer, and E Grimaldo, "Colliding-wind binaries with strong magnetic fields," *The Astrophysical Journal*, vol. 831, no. 2, pp. 121, 2016.

[20] Myoungkyu Lee and Robert D Moser, "Direct numerical simulation of turbulent channel flow up to $Re_\tau \approx 5200$," *Journal of Fluid Mechanics*, vol. 774, pp. 395–415, 2015.

[21] Fan Guo, Hui Li, William Daughton, and Yi-Hsin Liu, "Formation of hard power laws in the energetic particle spectra resulting from relativistic magnetic reconnection," *Physical Review Letters*, vol. 113, no. 15, pp. 155005, 2014.

[22] Andrew W. Cook, William Cabot, and Paul L. Miller, "The mixing transition in Rayleigh-Taylor instability," *Journal of Fluid Mechanics*, vol. 511, pp. 333–362, 2004.

[23] Peter Thoman, Philip Salzmann, Biagio Cosenza, and Thomas Fahringer, "Celerity: High-level C++ for accelerator clusters," in *European Conference on Parallel Processing*. Springer, 2019, pp. 291–303.