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 \(^2\), 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) \(^3\) and Frontier Fields \(^4\) data archives contain two-dimensional image files of various sizes in the FITS single-precision floating-point format\(^5\).

---

\(^1\)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 images⁵⁄⁶.

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 Contest⁸ 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 States⁹.
- SciVis 2006: 3D time steps from an earthquake simulation¹⁰.
- SciVis 2018: 3D time steps from a of deep water asteroid impact simulation¹¹.

HDRI Haven  HDRI Haven¹³ is a source for high-resolution, high-dynamic range (HDR) photographs. Images are available in the Radiance HDR format (also known

---

⁴https://www.cfa.harvard.edu/rtdc
⁵https://www.cfa.harvard.edu/rtdc/SMAimages
⁶https://www.cfa.harvard.edu/rtdc/CO/CompositeSurveys
⁷http://hla.stsci.edu/
⁸http://sciviscontest.ieeevis.org
⁹http://sciviscontest.ieeevis.org/2004/data.html
¹⁰http://sciviscontest.ieeevis.org/2006/download.html
¹¹https://sciviscontest2018.org/
¹²https://kaust-vislab.github.io/SciVis2020/data.html
¹³https://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
<table>
<thead>
<tr>
<th>dataset</th>
<th>source</th>
<th>data type</th>
<th>dimensions</th>
<th>extent</th>
</tr>
</thead>
<tbody>
<tr>
<td>msg_sppm</td>
<td>Burtscher</td>
<td>single, double</td>
<td>1</td>
<td>34,874,483</td>
</tr>
<tr>
<td>msg_sweep3d</td>
<td>Burtscher</td>
<td>single, double</td>
<td>1</td>
<td>15,716,403</td>
</tr>
<tr>
<td>snd_thunder</td>
<td>Freesound</td>
<td>single</td>
<td>1</td>
<td>7,898,672</td>
</tr>
<tr>
<td>ts_gas</td>
<td>UCI MLR</td>
<td>single</td>
<td>1</td>
<td>4,208,261</td>
</tr>
<tr>
<td>ts_wesad</td>
<td>UCI MLR</td>
<td>single</td>
<td>1</td>
<td>4,588,553</td>
</tr>
<tr>
<td>hdr_night</td>
<td>HDRI Haven</td>
<td>single</td>
<td>2</td>
<td>8,192 × 16,384</td>
</tr>
<tr>
<td>hdr_palermo</td>
<td>HDRI Haven</td>
<td>single</td>
<td>2</td>
<td>10,268 × 20,536</td>
</tr>
<tr>
<td>hubble</td>
<td>HLA</td>
<td>single</td>
<td>2</td>
<td>6,036 × 6,014</td>
</tr>
<tr>
<td>rsim</td>
<td>UIBK</td>
<td>single, double</td>
<td>2</td>
<td>2,048 × 11,509</td>
</tr>
<tr>
<td>spitzer_fls_irac</td>
<td>IRSA</td>
<td>single</td>
<td>2</td>
<td>6,456 × 6,389</td>
</tr>
<tr>
<td>spitzer_fls_vla</td>
<td>IRSA</td>
<td>single</td>
<td>2</td>
<td>8,192 × 8,192</td>
</tr>
<tr>
<td>spitzer_frontier</td>
<td>IRSA</td>
<td>single</td>
<td>2</td>
<td>3,874 × 2,694</td>
</tr>
<tr>
<td>asteroid</td>
<td>SciVis 2018</td>
<td>single</td>
<td>3</td>
<td>500 × 500 × 500</td>
</tr>
<tr>
<td>astro_mhd</td>
<td>UIBK</td>
<td>single</td>
<td>3</td>
<td>128 × 512 × 1024</td>
</tr>
<tr>
<td>astro_mhd</td>
<td>UIBK</td>
<td>double</td>
<td>3</td>
<td>130 × 514 × 1026</td>
</tr>
<tr>
<td>astro_pt</td>
<td>UIBK</td>
<td>single, double</td>
<td>3</td>
<td>512 × 256 × 640</td>
</tr>
<tr>
<td>flow</td>
<td>Klacansky</td>
<td>double</td>
<td>3</td>
<td>16 × 7,680 × 1,024</td>
</tr>
<tr>
<td>hurricane</td>
<td>SciVis 2004</td>
<td>single</td>
<td>3</td>
<td>100 × 500 × 500</td>
</tr>
<tr>
<td>magrecon</td>
<td>Klacansky</td>
<td>single</td>
<td>3</td>
<td>512 × 512 × 512</td>
</tr>
<tr>
<td>miranda</td>
<td>Klacansky</td>
<td>single</td>
<td>3</td>
<td>1,024 × 1,024 × 1,024</td>
</tr>
<tr>
<td>redsea</td>
<td>SciVis 2020</td>
<td>double</td>
<td>3</td>
<td>50 × 500 × 500</td>
</tr>
<tr>
<td>sma_disk</td>
<td>RTDC</td>
<td>single</td>
<td>3</td>
<td>301 × 369 × 369</td>
</tr>
<tr>
<td>turbulence</td>
<td>Klacansky</td>
<td>single</td>
<td>3</td>
<td>256 × 256 × 256</td>
</tr>
<tr>
<td>wave</td>
<td>UIBK</td>
<td>single, double</td>
<td>3</td>
<td>512 × 512 × 512</td>
</tr>
</tbody>
</table>

**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&amp;h=preller_drive
\(^{24}\)https://hdrihaven.com/hdri/?c=outdoor&amp;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.

- **magnecon** 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 interferometer\(^5\).

• **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


\(^{25}\)https://www.cfa.harvard.edu/rtdc/SMAimages/080403_034904_hd98800.html


