Exploring The Milky Way

(Or Any Other NDim Dataset)

Francesc Alted / @FrancescAlted

The Blosc Development Team / @Blosc2

CEO ironArray / @ironArray

Python User Group Castelló

June 27th 2023

Agenda

The Gaia Dataset

Blosc2 NDim and NDArray Objects

Exploring the Milky Way with Blosc2

Automatic Compression Tuning with Btune

Conclusions

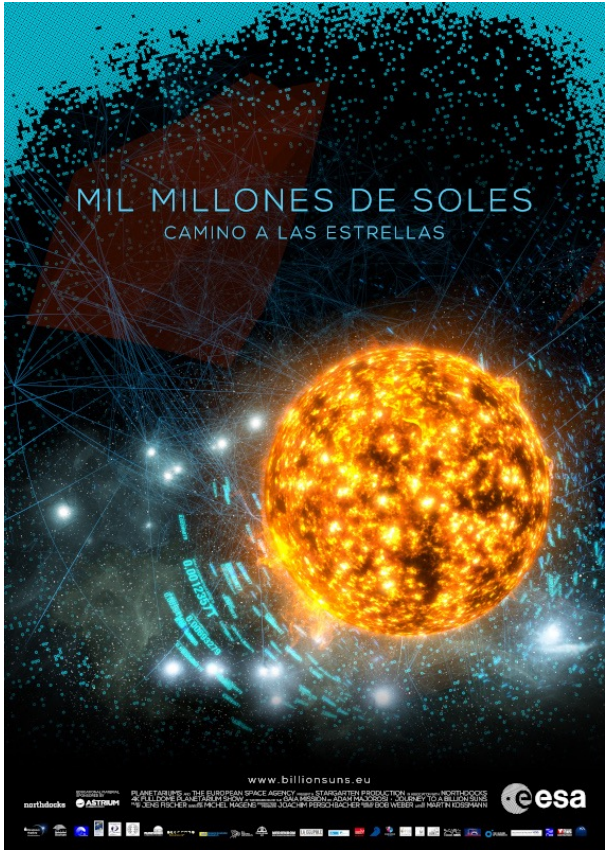
Disclaimer

I am not attached to the Gaia collaboration at all.

All statements said here about scientific facts on Gaia data might be plain wrong.

Idea agafada de documental vist al planetari de Castelló

Per desgràcia, sembla que ja no està disponible...



En compte d’això...

Cinema

ELS INVASSORS DE MART

Descripció:

Cicle de cinema de ciència ficció al Saló d'actes del Planetari de Castelló.

El fiil argumental són les diferents visions que el cinema de Hollywood ha tingut dels alienígenes al llarg dels darrers 70 anys.

Els invasors de Mart. Des de la finestra de la seua habitació, el xicotet Jimmy albira un platet volant que aterra prop de la seua casa.

L'estrany comportament que a partir de llavors mostren les persones que l'envolten, el porta a demanar ajuda.

Després d'investigar el cas, arriben a la conclusió que tot el que està ocorrent forma part d'un pla d'invasió de la Terra des del planeta Mart.

Divendres 07 de juliol 18.00

En compte d’això...

Cinema

2001 UNA ODISEA DE L'ESPAI

Descripció:

Cicle de cinema de ciència ficció al Saló d'actes del Planetari de Castelló.

El fil argumental són les diferents visions que el cinema de Hollywood ha

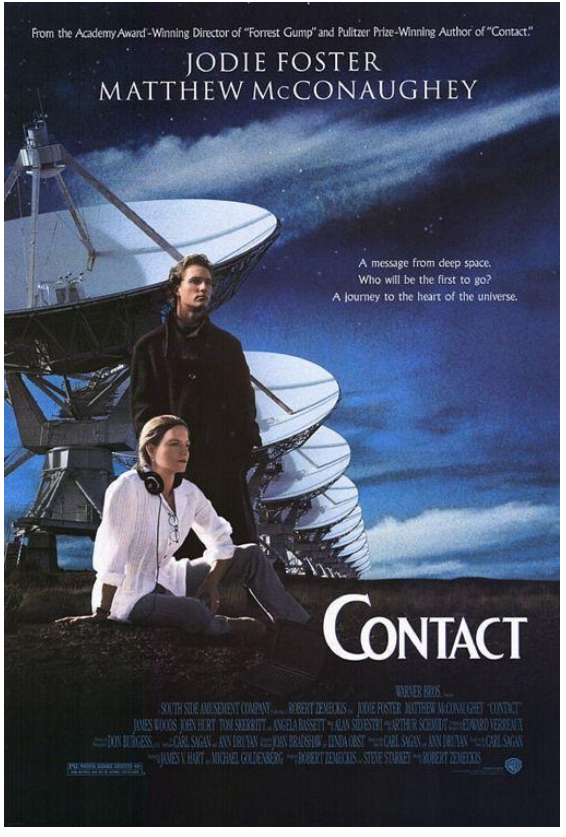
tingut dels alienígenes al llarg dels darrers 70 anys.

2001 Una odissea en l'espai.Narra els diversos períodes de la història de la humanitat, no sols del passat, sinó també del futur.

Fa milions d'anys, abans de l'aparició del "homo sapiens", uns primats descobreixen un monòlit que els condueix a un estadi d'intel·ligència superior. Milions d'anys després, un altre monòlit, enterrat en una lluna, desperta l'interés dels científics. Finalment, durant una missió de la NASA, *HAL 9000, una màquina dotada d'intel·ligència artificial, s'encarrega de controlar tots els sistemes d'una nau espacial tripulada.

Dijous 20 de juliol 18.00

En compte d’això...



Cinema

CONTACT

Descripció:

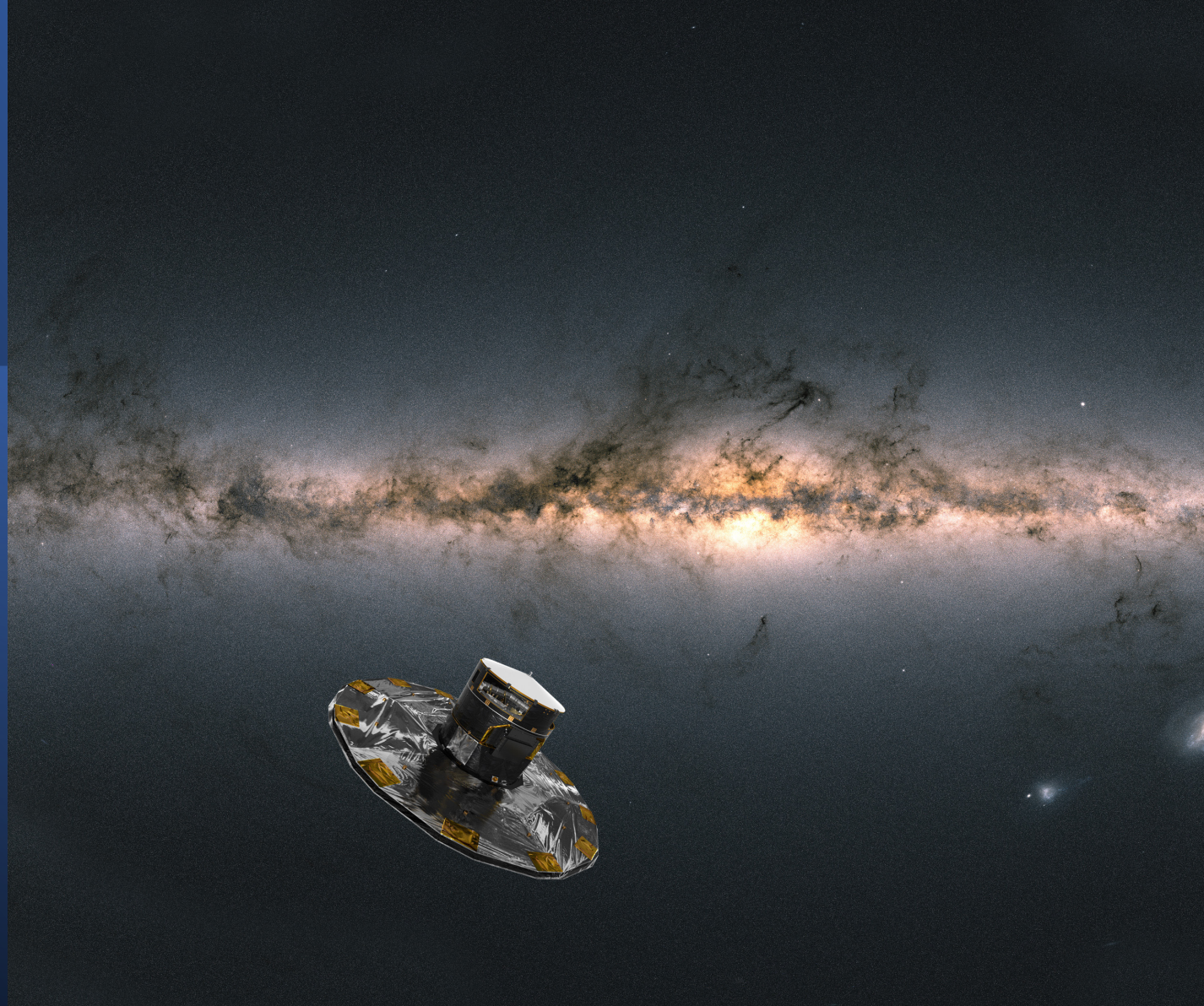
Cicle de cinema de ciència ficció al Saló d'actes del Planetari de Castelló.

El fil argumental són les diferents visions que el cinema de Hollywood ha tingut dels alienígenes al llarg dels darrers 70 anys.

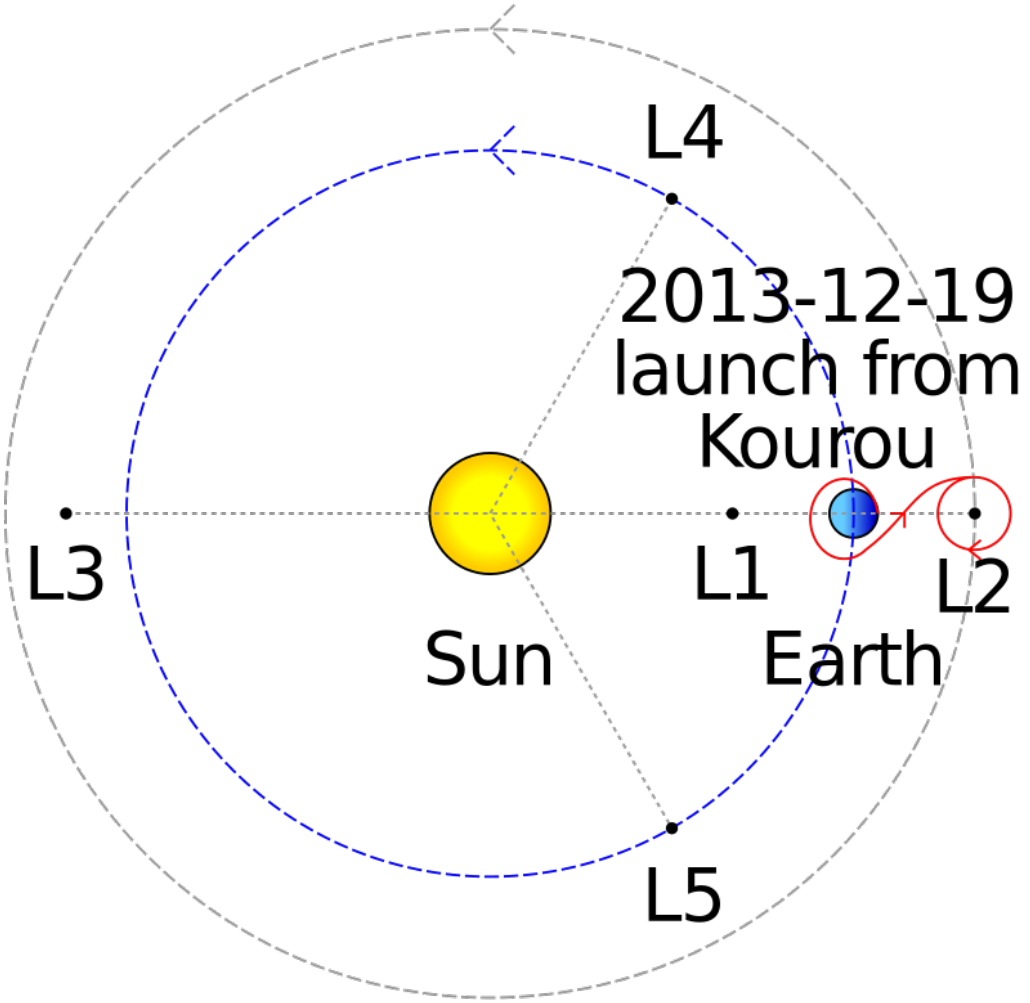
Contact. Després de la prematura mort dels seus pares sent una xiqueta, Eleanor Arroway va perdre la fe en Déu. Com a contrapartida, ha concentrat tota la seua fe en la investigació: treballa amb un grup de científics que analitzen ones de ràdio procedents de l'espai exterior amb la finalitat de trobar senyals d'intel·ligència extraterrestre. El seu treball es veu recompensat quan detecta un senyal desconegut que sembla contindre les instruccions de fabricació d’una màquina que permetria reunir-se amb els autors del missatge.

Dijous 03 d’agost 18.00

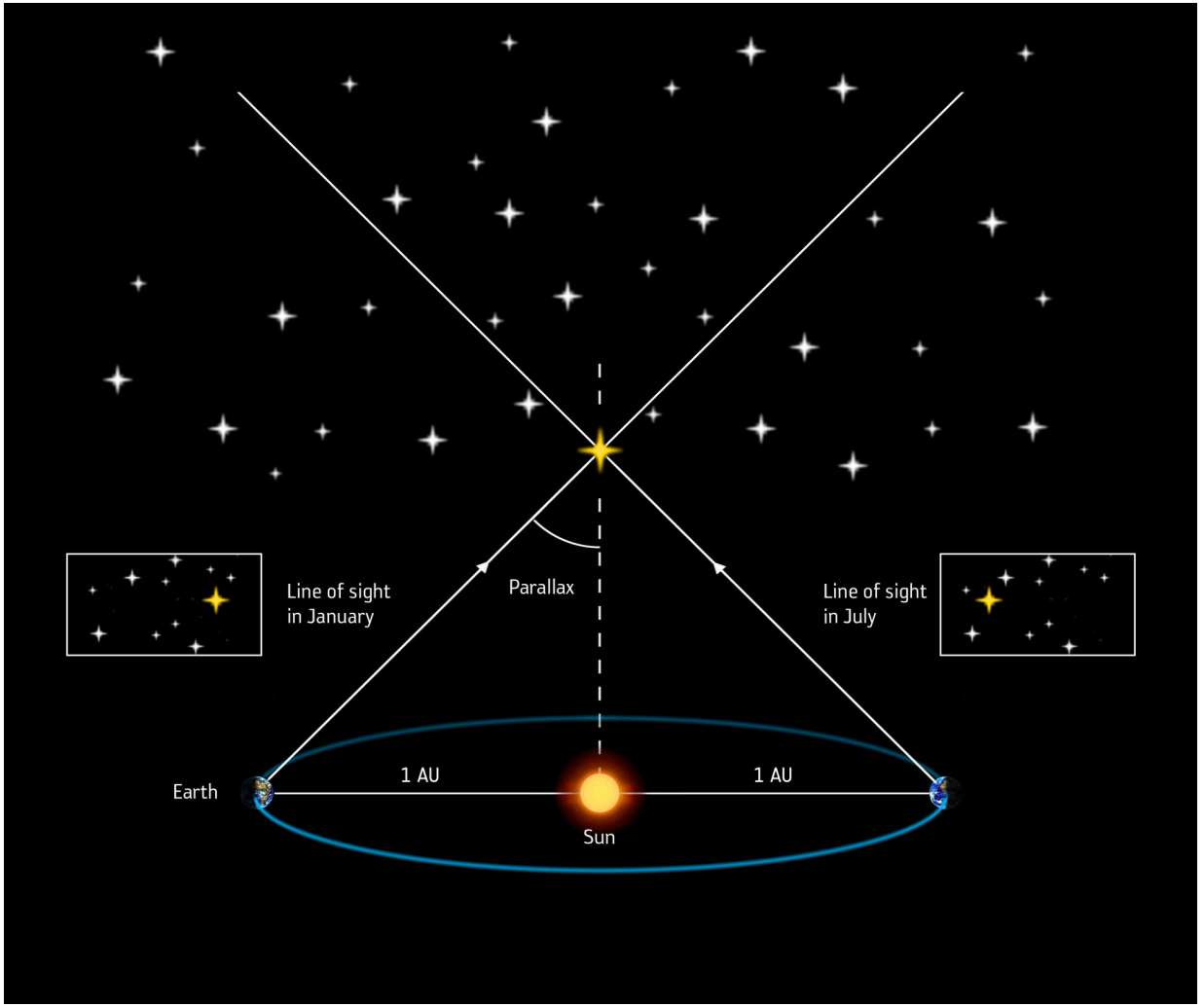
Gaia mission: measure the position of stars in the Milky Way (and other info too)



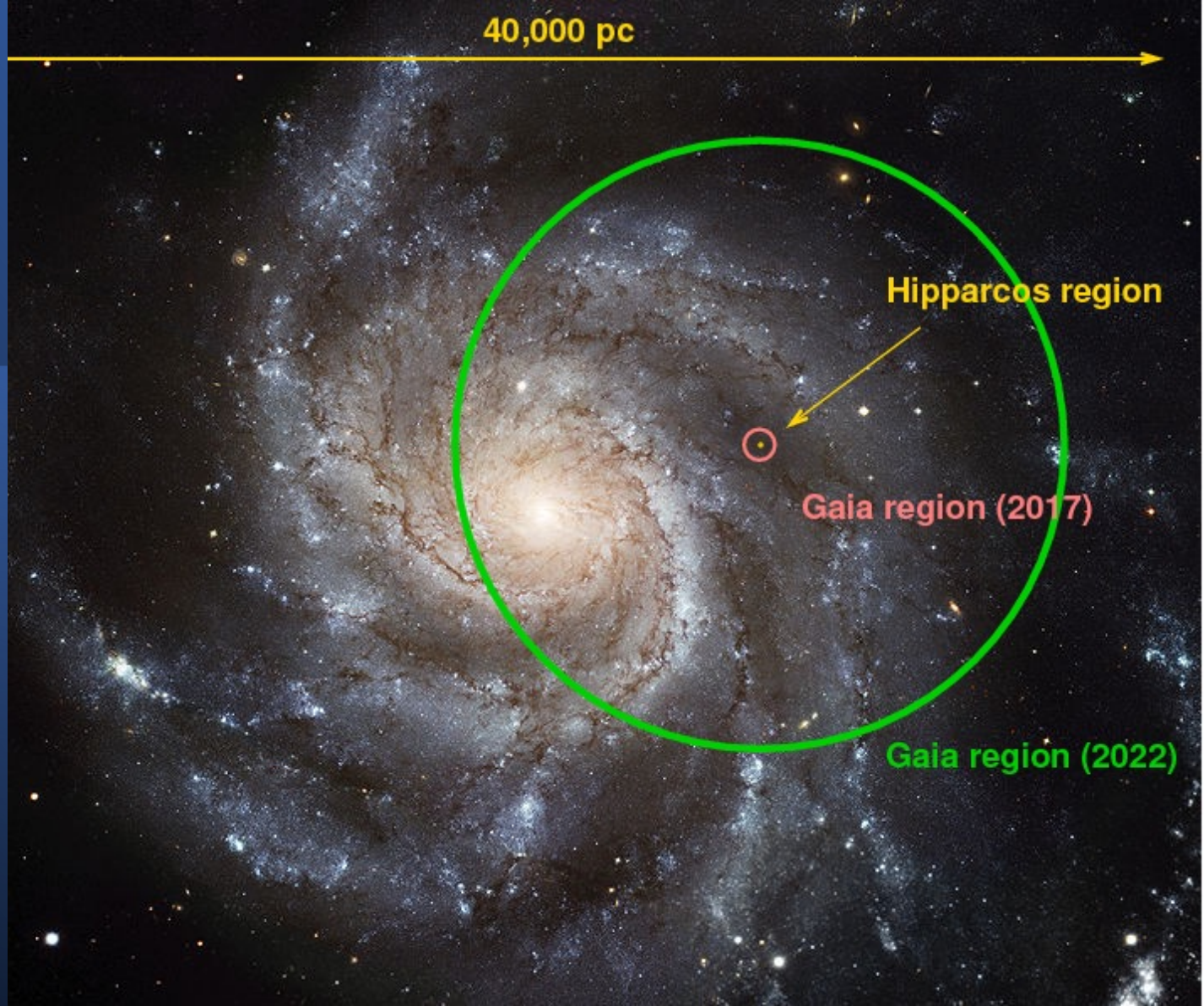
Gaia telescope orbits around L2 (recently James Webb joined too)



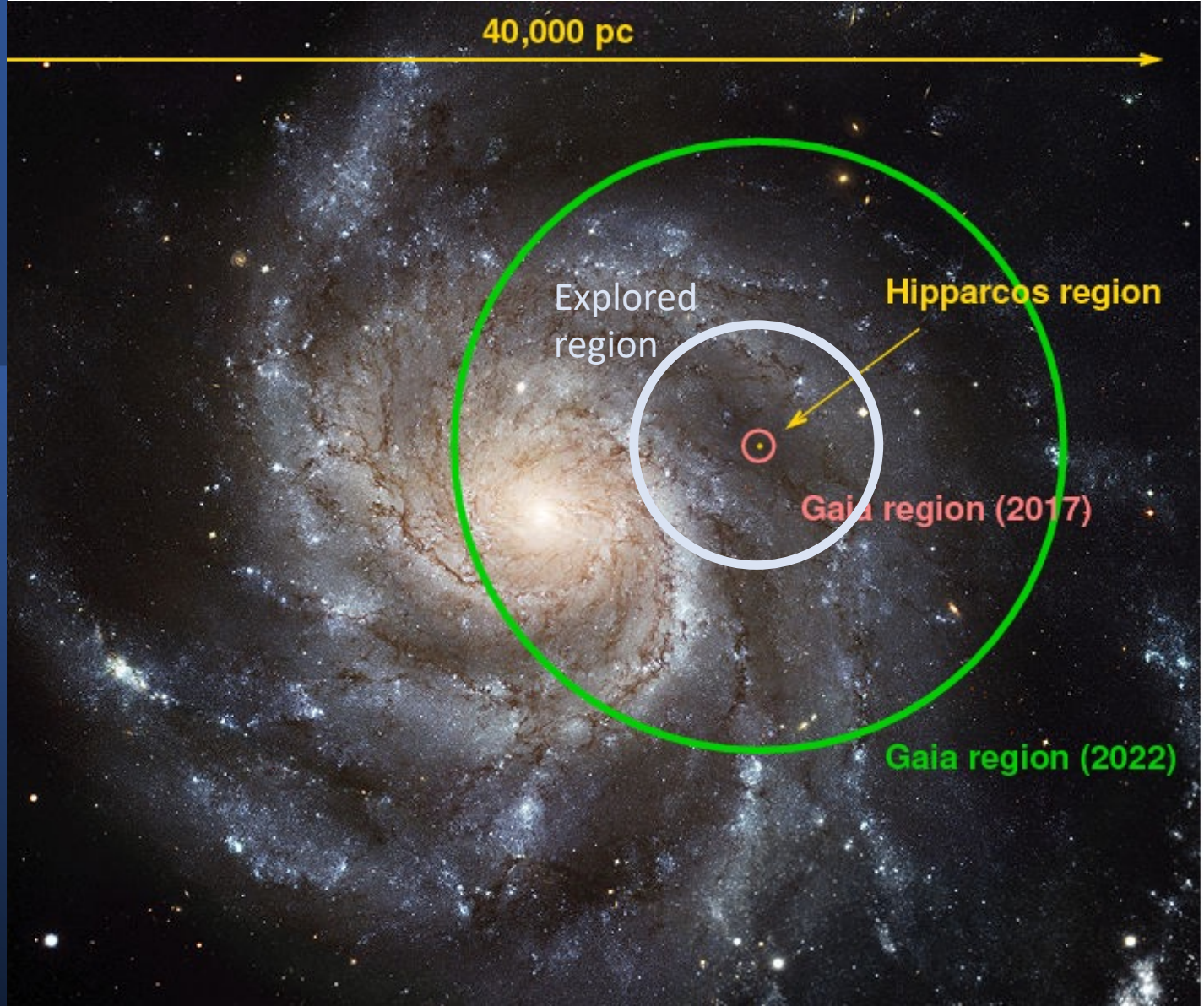
Gaia uses Parallax to measure distance to nearby stars



Gaia can observe larger regions as more observations cummulated



We will restrict ourselves to a region which is 10,000 light years (3,000 parsec) away from Gaia



The Gaia dataset

The Main Source Catalog

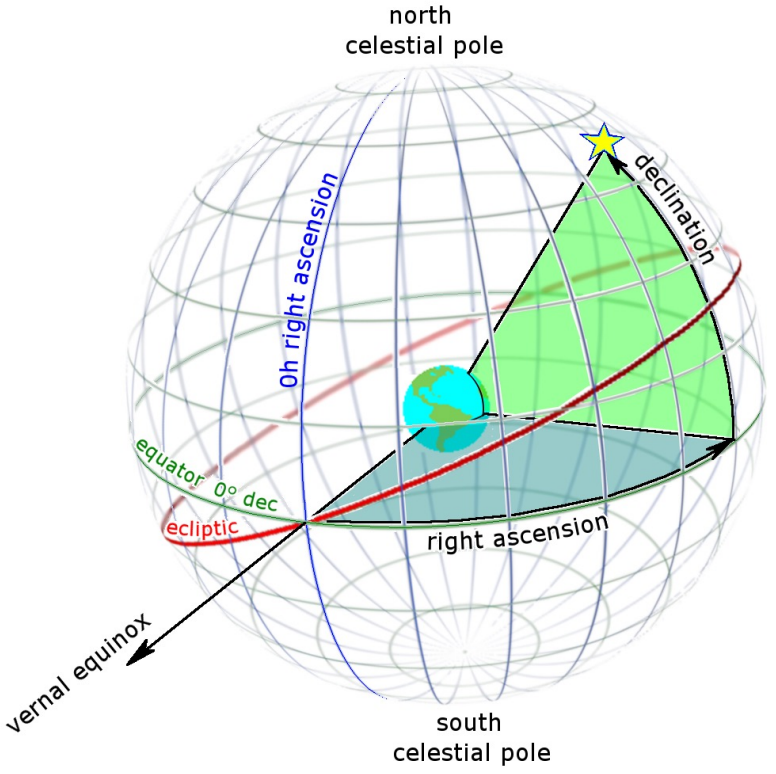
- A set of public CSV files with Gaia observations of many star parameters
- We are interested mainly in:
    - **ra** : Right ascension (double, Angle[deg])
    - **dec** : Declination (double, Angle[deg])
    - **parallax** : Parallax (double, Angle[mas])
- From these, and some elemental spherical geometry, we can read and filter the stars in a radius of 10,000 light years.

https://gea.esac.esa.int/archive/documentation/GDR3/Gaia_archive/chap_datamodel/sec_dm_main_source_catalogue/ssec_dm_gaia_source.html

Spherical to Cartesian Coordinates

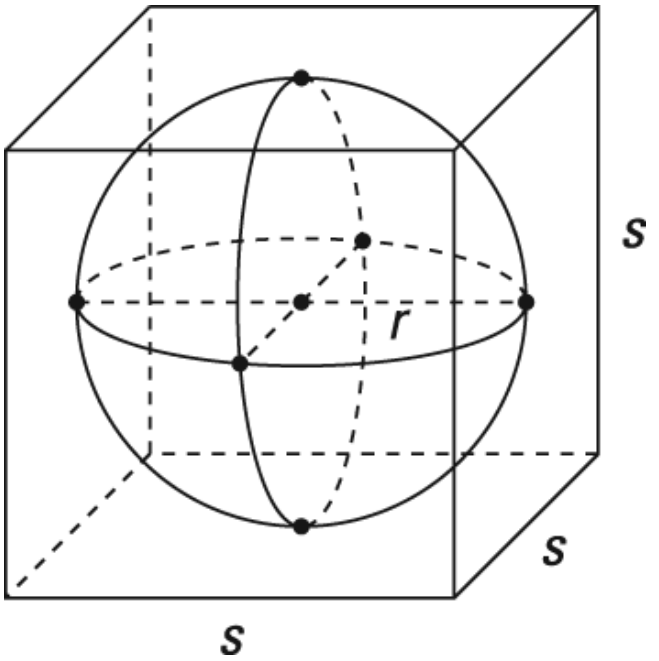
$$x = \rho \sin\theta \cos\varphi$$
$$y = \rho \sin\theta \sin\varphi$$
$$z = \rho \cos\theta$$

Easy to convert. Beware, angles must be in radians, whereas Gaia raw data provides degrees!



The Exploration Cube

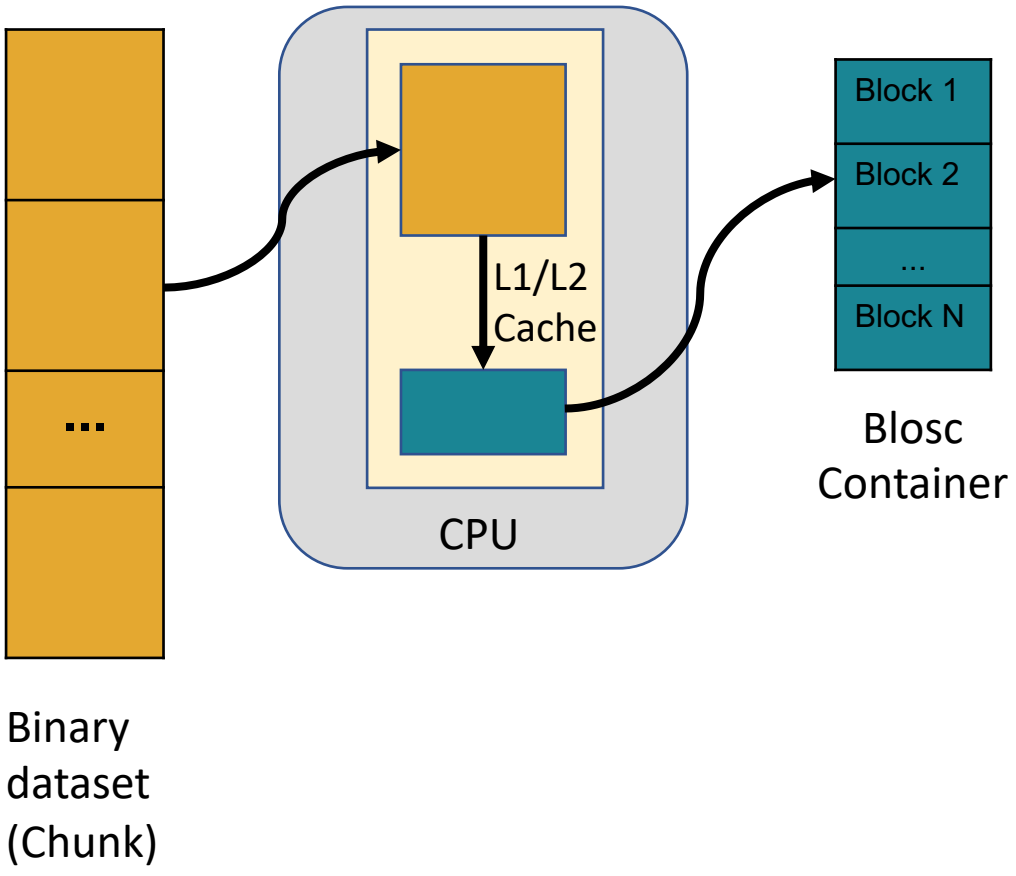
- **Geospatial data**
- Radius of the inscribed sphere:
  r = 10,000 light years
- Length of the cube side:
  S = 20,000 light years
- Every cell in the cube is 1 cubic light year.
- That means 8 trillions cells in total (8 TB!).



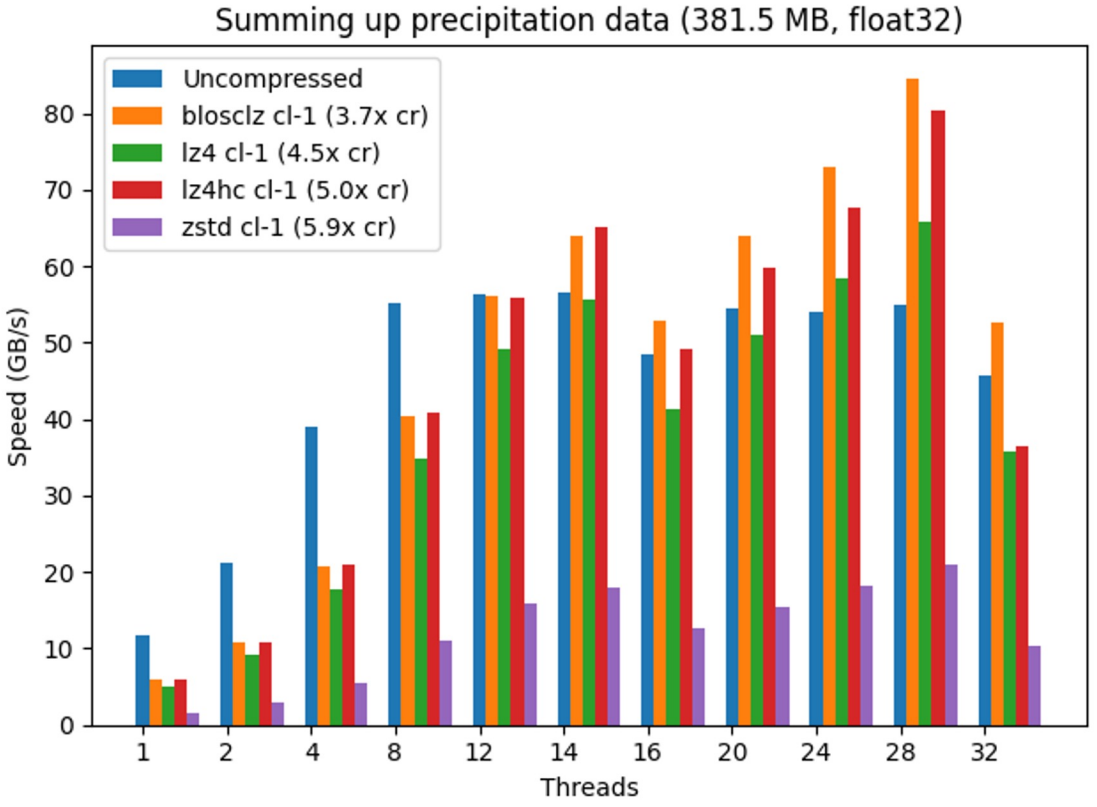
Enter Blosc2 Ndim for a solution.

What is Blosc?

- Sending data from CPU to memory (and back) faster than *memcpy()*.
- Split in blocks for better cache use: divide and conquer.
- It can use different filters (e.g. shuffle, bitsuffle) and codecs (e.g. LZ4, Zlib, Zstd, BloscLZ).



Breaking Memory Walls



https://www.blosc.org/posts/breaking-memory-walls/

Where is Blosc used?

Blosc is used in many places in the PyData ecosystem:

- HDF5 / h5py (via hdf5plugin)

- HDF5 / PyTables (native)

PyTables

- Zarr (via numcodecs)

Zarr

- ironArray (Blosc2)

ironArray

Lots of terrabytes compressed (and decompressed) on a daily basis!

Blosc (Francesc Alted) Winner of Google's Open Source Peer Bonus in 2017

"To recognize and celebrate external contributors to the open source ecosystem Google depends on."

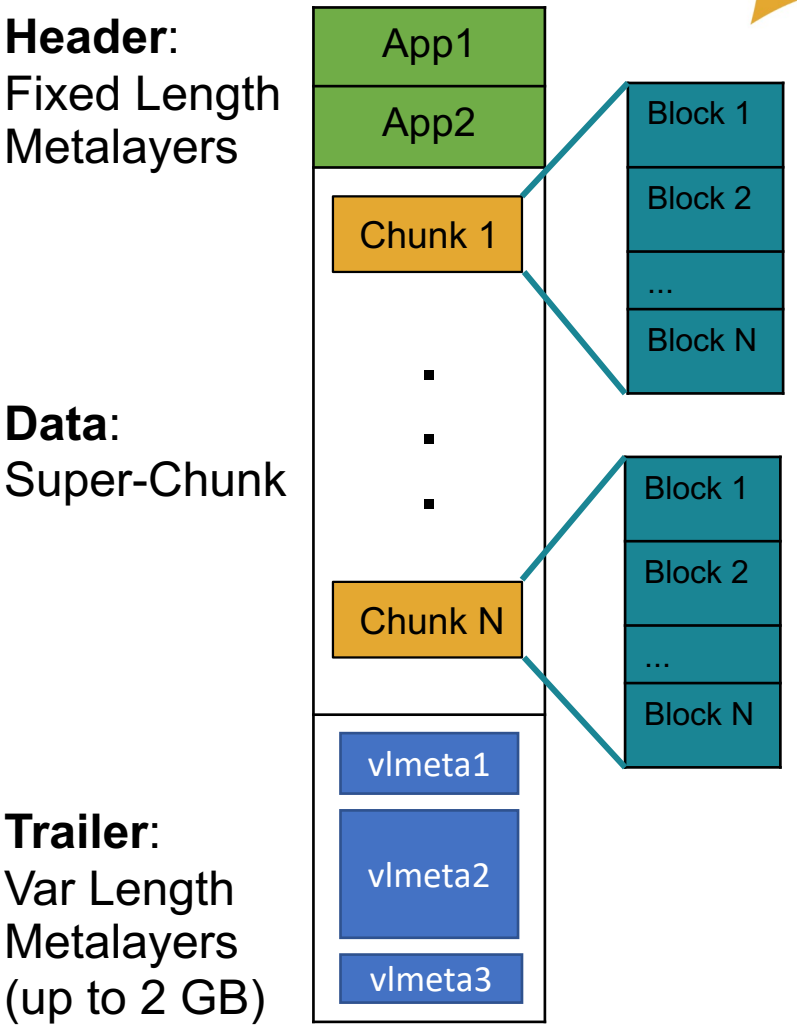
Some of the projects that won the award the same year:

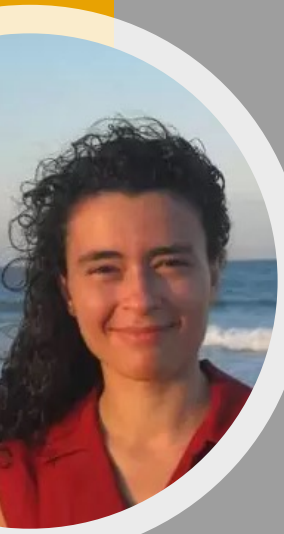
- SQLite (Dan Kennedy, Joe Mistachkin, Richard Hipp)
- NumPy (Sebastian Berg)
- Ffmpeg (Michael Niedermayer)
- Flask (Armin Ronacher)

Google Open Source

What is Blosc2?

- Next generation of Blosc(1), a high performance compressor.
- Blosc2 adds 63-bit containers that expand over the existing 31-bit containers (chunks) in Blosc1.
- Metalayers for adding info for apps and users.






The Blosc Development Team

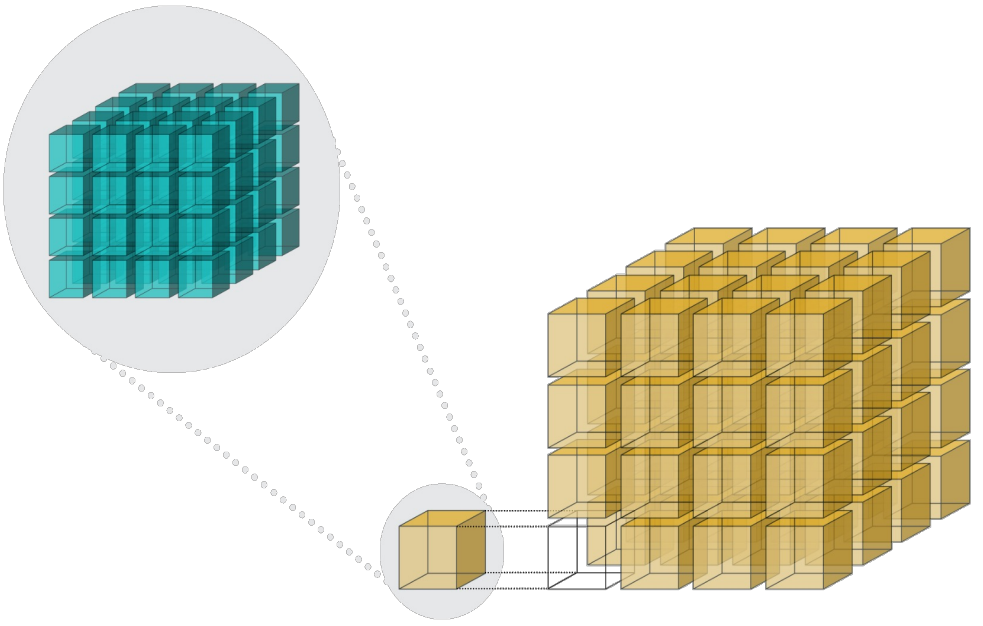

- Aleix Alcacer
- Oscar Guiñón
- Marta Iborra
- Alberto Sabater
- J. David Ibáñez
- Francesc Alted (BDFL)

NDim And NDArray

Blosc2 Goes Multidimensional

C-Blosc2 NDim: Multidimensions for C

- Each NDim array is split in chunks
- Each chunk is split in blocks
- All the partitions are multidimensional!
- Metalayer representing both multidimensionality and **data types** (new!)



https://www.blosc.org/c-blosc2/reference/b2nd.html

NDArray: Blosc2 NDim for Python

```python
import blosc2

a = blosc2.full((4, 4), fill_value=9)
a.resize((5, 7))
a[3:5, 2:7] = 8
print(a[:])
```

Output:

```
[[9 9 9 9 0 0 0]
 [9 9 9 9 0 0 0]
 [9 9 9 9 0 0 0]
 [9 9 8 8 8 8 8]
 [0 0 8 8 8 8 8]]
```

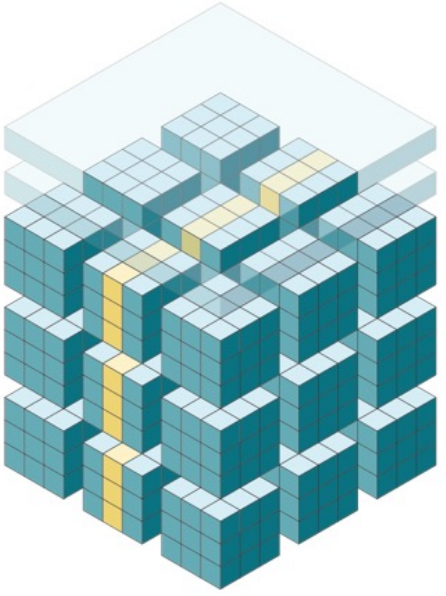
Features:

- Create arrays in memory or on disk
- Flexible resize (including shrinking)
- Support for all NumPy data types
- Efficient conversion from/to NumPy
- Mimic NumPy API
- Version 2.1 out; meant for production

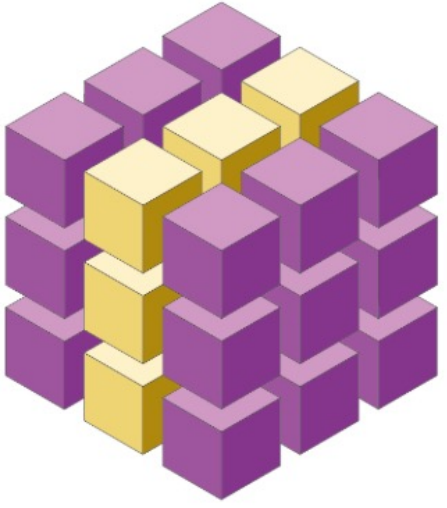
https://www.blosc.org/python-blosc2/reference/ndarray_api.html

Leveraging the second partition in Blosc2 NDim

Much more selective and faster queries!

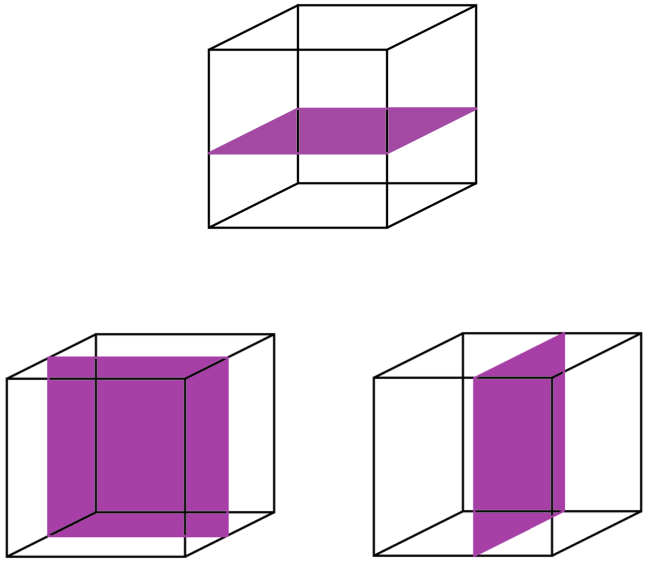


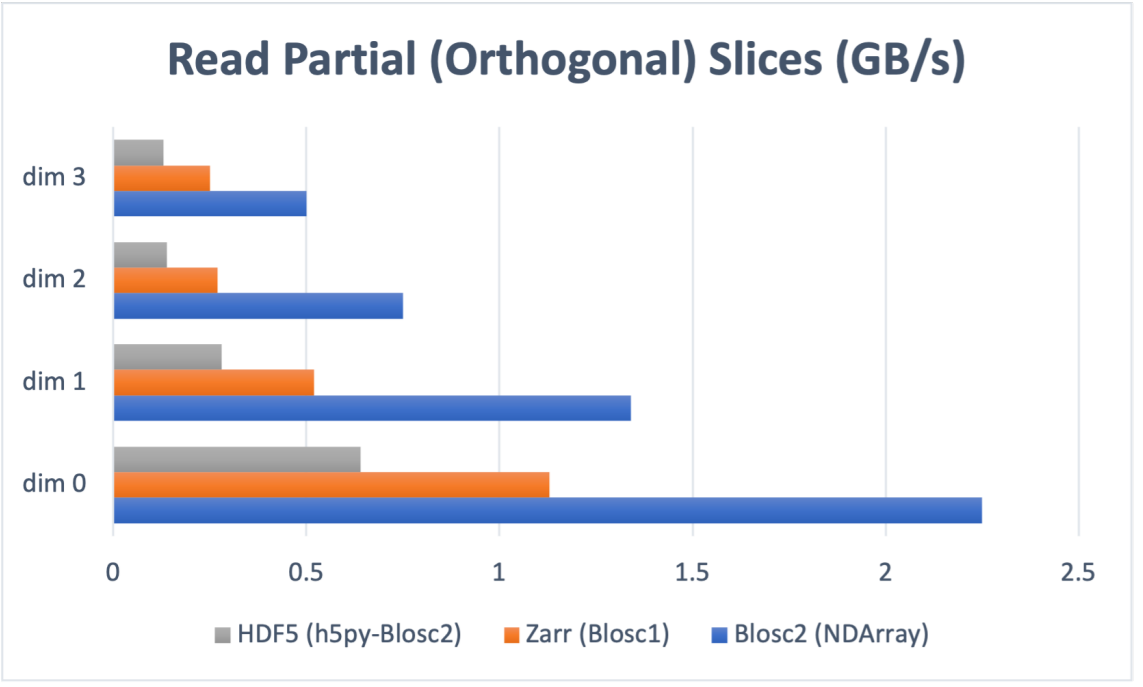
Blosc2 NDim



HDF5 / Zarr / others

Blosc2 NDim partial read performance





Faster slicing due to higher data selectivity in double partitioning

Walking Over Metadata

Latest version of Python-Blosc2 comes with iterchunks_info, which allows to iterate over meta-info from chunks.

This is much faster than regular iterchunks iterator because it does not decompress anything.

Example:

https://github.com/FrancescAlted/Gaia/blob/main/traverse-3d.py

Interlude

Visualize a 3D datagrid with 8 trillion cells and 0.7 billion of stars

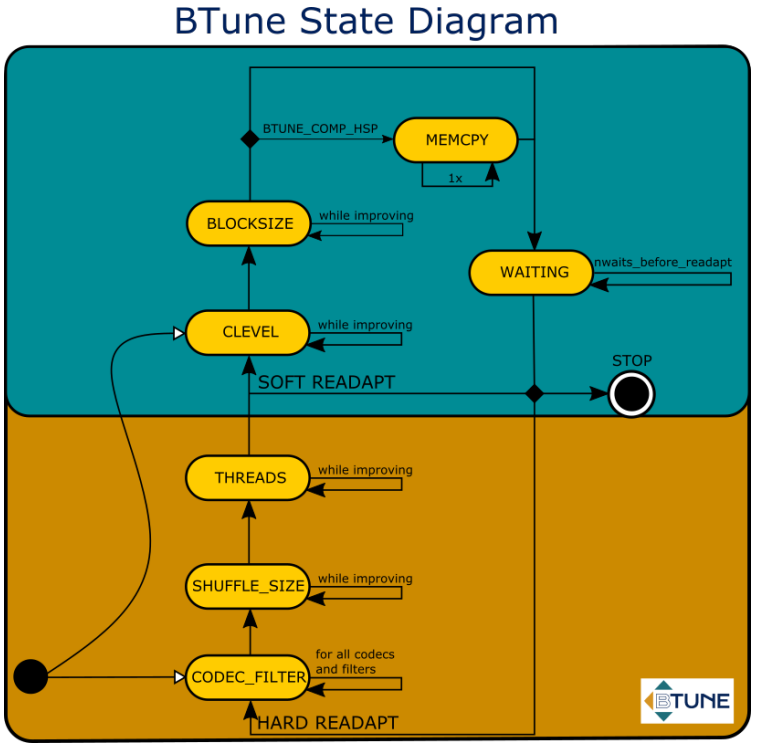
Btune

Optimize Compression To User’s Needs

Fine Tuning Performance with BTune

https://btune.blosc.org

- BTune can fine tune the different parameters of the underlying Blosc2 storage to perform as best as possible.
- Can be trained to find the best codec & filter with deep learning.
- **Looking for beta testers!**



Btune Operation

Three ways:

- **Plain Blosc2_Btune plugin**: Use the dynamic plugin directly (can be slow)
- **Btune Models**: Ask the Blosc Development Team for a Neural Net model adapted to your datasets for faster operation
- **Btune Studio**: Use the training package locally for generate your own models for your datasets by yourself

Plain Blosc2_Btune plugin

```
$ pip install blosc2-btune
```

After installing, Blosc2 will load Btune whenever there is a BTUNE environment variable. For example:

```
$ BTUNE_BALANCE=0.5 BTUNE_TRACE=1 python examples/schunk.py
```

Plain Blosc2_Btune plugin

```
faltet@Mac-mini-de-Francesc ~/b/python-blosc2 (main)> BTUNE_BALANCE=0.9 BTUNE_TRACE=1  python examples/schunk.py
-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
BTune version: 1.0.0
Performance Mode: COMP, Compression balance: 0.900000, Bandwidth: 20 GB/s
Behaviour: Waits - 0, Softs - 5, Hards - 10, Repeat Mode - STOP
TRACE: Environment variable BTUNE_MODELS_DIR is not defined
-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
BTune version: 1.0.0
Performance Mode: COMP, Compression balance: 0.900000, Bandwidth: 20 GB/s
Behaviour: Waits - 0, Softs - 5, Hards - 11, Repeat Mode - STOP
TRACE: Environment variable BTUNE_MODELS_DIR is not defined
WARNING: Empty metadata, no inference performed
|    Codec   | Filter | Split | C.Level | Blocksize | Shufflesize | C.Threads | D.Threads |   Score    |  C.Ratio   |   BTune State   | Readapt | Winner
|       zstd |      0 |     1 |       3 |         0 |           4 |         6 |         6 |   0.000195 |  2.5e+04x |   CODEC_FILTER |    HARD | S
|       zstd |      0 |     0 |       3 |         0 |           4 |         6 |         6 |    0.00152 |     1.53x |   CODEC_FILTER |    HARD | W
|       zstd |      1 |     1 |       3 |         0 |           4 |         6 |         6 |   0.000211 |      205x |   CODEC_FILTER |    HARD | W
|       zstd |      1 |     0 |       3 |         0 |           4 |         6 |         6 |   0.000106 |      138x |   CODEC_FILTER |    HARD | -
|       zstd |      2 |     1 |       3 |         0 |           4 |         6 |         6 |   0.000238 |      210x |   CODEC_FILTER |    HARD | W
|       zstd |      2 |     0 |       3 |         0 |           4 |         6 |         6 |   0.000151 |      279x |   CODEC_FILTER |    HARD | W
|       zlib |      0 |     1 |       3 |         0 |           4 |         6 |         6 |     0.0032 |     1.53x |   CODEC_FILTER |    HARD | -
|       zlib |      0 |     0 |       3 |         0 |           4 |         6 |         6 |    0.00285 |     1.47x |   CODEC_FILTER |    HARD | -
|       zlib |      1 |     1 |       3 |         0 |           4 |         6 |         6 |   0.000478 |      113x |   CODEC_FILTER |    HARD | -
|       zlib |      1 |     0 |       3 |         0 |           4 |         6 |         6 |   0.000395 |     57.1x |   CODEC_FILTER |    HARD | -
|       zlib |      2 |     1 |       3 |         0 |           4 |         6 |         6 |   0.000421 |      145x |   CODEC_FILTER |    HARD | -
|       zlib |      2 |     0 |       3 |         0 |           4 |         6 |         6 |   0.000495 |      149x |   CODEC_FILTER |    HARD | -
|       zstd |      2 |     0 |       3 |         0 |           4 |         6 |         6 |   0.000216 |     83.6x |   THREADS_COMP |    HARD | -
|       zstd |      2 |     0 |       3 |         0 |           4 |         5 |         6 |    0.00065 |      205x |   THREADS_COMP |    HARD | -
|       zstd |      2 |     0 |       6 |         0 |           4 |         6 |         6 |    0.00173 |      282x |         CLEVEL |    HARD | W
faltet@Mac-mini-de-Francesc ~/b/python-blosc2 (main)>
```

Btune Models

- Instead of trying out different codecs and filters for each chunk, Btune can be trained for your datasets and infer, in real time, when to use a combination of codec and filter that suits some requirements (favor speed, favor cratio, or a balance).
- The Blosc Development team is in charge to train a neural network model for the typical datasets. Then, the user drop the model in some directory and inform Btune about the location. E.g.:

```
BTUNE_MODELS_DIR=./models_sample BTUNE_USE_INFERENCE=3
python examples/schunk_roundtrip.py
```

Btune Models

Using 3 inferences. Note how fine-tuning is still carried out (tweaking):

```
PYTHONPATH=. BTUNE_BALANCE=0.5 BTUNE_PERF_MODE=COMP BTUNE_TRACE=1  BTUNE_MODELS_DIR=./models_sample/ BTUNE_USE_INFERENCE=3 python examples/schunk_roundtrip.py
-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
BTune version: 1.0.0
Performance Mode: COMP, Compression balance: 0.500000, Bandwidth: 20 GB/s
Behaviour: Waits - 0, Softs - 5, Hards - 11, Repeat Mode - STOP
INFO: Model files found in the '../blosc2_btune/models_sample//' directory
TRACE: time load model: 0.000063
TRACE: Inference category=4 codec=1 filter=0 clevel=5 splitmode=2 time entropy=0.000267 inference=0.000010
|    Codec    | Filter | Split | C.Level | Blocksize | Shufflesize | C.Threads | D.Threads |   Score   |  C.Ratio   |   BTune State   | Readapt | Winner
|        lz4 |       0 |     1 |       5 |         0 |           4 |        16 |        16 |  0.000627 |         2x |    CODEC_FILTER |    HARD | W
TRACE: Inference category=4 codec=1 filter=0 clevel=5 splitmode=2 time entropy=0.000029 inference=0.000002
|        lz4 |       0 |     1 |       5 |         0 |           4 |        16 |        16 |  0.000478 |         2x |    CODEC_FILTER |    HARD | -
TRACE: Inference category=4 codec=1 filter=0 clevel=5 splitmode=2 time entropy=0.000023 inference=0.000002
|        lz4 |       0 |     1 |       5 |         0 |           4 |        16 |        16 |  0.000273 |         2x |    CODEC_FILTER |    HARD | -
|        lz4 |       0 |     1 |       5 |         0 |           4 |        16 |        16 |  0.000272 |         2x |    CODEC_FILTER |    HARD | -
|        lz4 |       0 |     0 |       5 |         0 |           4 |        16 |        16 |  0.000225 |         2x |    CODEC_FILTER |    HARD | W
|        lz4 |       0 |     0 |       5 |         0 |           4 |        16 |        16 |  0.000239 |         2x |    THREADS_COMP |    HARD | -
|        lz4 |       0 |     0 |       5 |         0 |           4 |        15 |        16 |  0.000674 |         2x |    THREADS_COMP |    HARD | -
|        lz4 |       0 |     0 |       5 |         0 |           4 |        16 |        16 |  0.000497 |         2x |          CLEVEL |    HARD | W
|        lz4 |       0 |     0 |       4 |         0 |           4 |        16 |        16 |  0.000217 |         2x |          CLEVEL |    SOFT | W
|        lz4 |       0 |     0 |       5 |         0 |           4 |        16 |        16 |  0.000217 |         2x |          CLEVEL |    SOFT | W
|        lz4 |       0 |     0 |       6 |         0 |           4 |        16 |        16 |  0.000209 |         2x |          CLEVEL |    SOFT | -
|        lz4 |       0 |     0 |       5 |         0 |           4 |        16 |        16 |  0.000224 |         2x |          CLEVEL |    SOFT | -
|        lz4 |       0 |     0 |       6 |         0 |           4 |        16 |        16 |   0.00024 |         2x |          CLEVEL |    SOFT | -
|        lz4 |       0 |     0 |       5 |         0 |           4 |        16 |        16 |  0.000227 |         2x |          CLEVEL |    SOFT | -
|        lz4 |       0 |     0 |       6 |         0 |           4 |        16 |        16 |   0.00024 |         2x |          CLEVEL |    SOFT | -
|        lz4 |       0 |     0 |       5 |         0 |           4 |        16 |        16 |  0.000189 |         2x |          CLEVEL |    SOFT | -
|        lz4 |       0 |     0 |       6 |         0 |           4 |        16 |        16 |  0.000194 |         2x |          CLEVEL |    SOFT | -
|        lz4 |       0 |     1 |       5 |         0 |           4 |        16 |        16 |  0.000239 |         2x |    CODEC_FILTER |    HARD | -
|        lz4 |       0 |     0 |       5 |         0 |           4 |        16 |        16 |  0.000196 |         2x |    CODEC_FILTER |    HARD | -
|        lz4 |       0 |     0 |       5 |         0 |           4 |        15 |        16 |  0.000525 |         2x |    THREADS_COMP |    HARD | -
```

Btune Studio

Sometimes, people handling a large number of datasets want to have the possibility to generate neural network models by themselves.

Btune Studio provides all the necessary tooling for doing that.

Conclusion

Blosc2 and the Multidim Milky Way

We have shown:

- The Gaia dataset offers great info on 1.7 billion of stars freely and readily accessible via CSV files.
- Blosc2 Ndim and NDArray can be used to easily create huge sparse matrices representing large spatial volumes (in this case, 8 trillions cells)
- These arrays can be explored easily via e.g. Plotly
- **BTUNE,** a tool for automatically select best Blosc2 parameters

Blosc2: a highly efficient and flexible tool for
compressing your data, your way

Thanks to donors & contracts!

HUAWEI

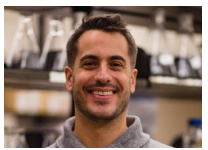
ironArray

NUMFOCUS
OPEN CODE = BETTER SCIENCE

python SOFTWARE FOUNDATION

Google

ESRF
The European Synchrotron



Jeff Hammerbacher

Without them, we could not have possibly put Blosc2 into production status: Blosc2 2.0.0 came out in June 2021; now at 2.8.0.

Thank you! Questions?

https://blosc.org/

We make compression better