UZP: Representing sparse matrices as polyhedra

Alonso Rodríguez-Iglesias*, Gabriel Rodríguez*, Juan Touriño*, Louis-Noël Pouchet†

*Universidade da Coruña, GAC, CITIC, Spain. †Colorado State University, USA. E-mail: alonso.rodriguez@udc.es

Introduction

- There are many formats for representing sparse matrices, such as COO, CSR, CSC, DIA, and block or doubly compressed versions of them.
- Compression ratio of these formats varies strongly depending on the internal structure of the matrix to be compressed.
- A polyhedron describes a series of (nested) affine loops that can be used to represent a collection of points in space.
- An efficient file format for compressing, distributing and operating with matrices can be created by searching for polyhedra inside sparse matrices in a way that it represents the matrix in the most compressed way possible, as well as tuning it for efficient operation.
- Based on the principles of polyhedral compilation and looking for even better compression and operational efficiency, we present the UZP (Union of $\mathcal{Z}$-Polyhedra) file format.

Objectives

- To develop an efficient algorithm that performs pattern search.
- To dump those found patterns (polyhedra) into a UZP file.
- To make such UZP file more compressed than all previous file formats in all non best case scenarios.

Design Principles

- A 1-d polyhedron can be understood as a 1-layer for loop, such as:

```
for(int i = 0; i < 8; i++)
    <access to> A[1+1*i][1+2*i]
```

- The loop above traverses the points (1,1), (2,3), (3,5), (4,7), (5,9), (6,11), (7,13), (8,15), and thus can be represented by a polyhedron.
- This access pattern can be represented as $\{origin, polyhedron_{md}\}$, where the m-dimensional[a] polyhedron has $lattice$ and n elements: $\{origin, \{lattice, n\}\}$, which would translate into $\{(1,1), \{(1,2), 8\}_{1d}\}$.
- A matrix can be fully represented by a collection of such polyhedra.

Internal Representation

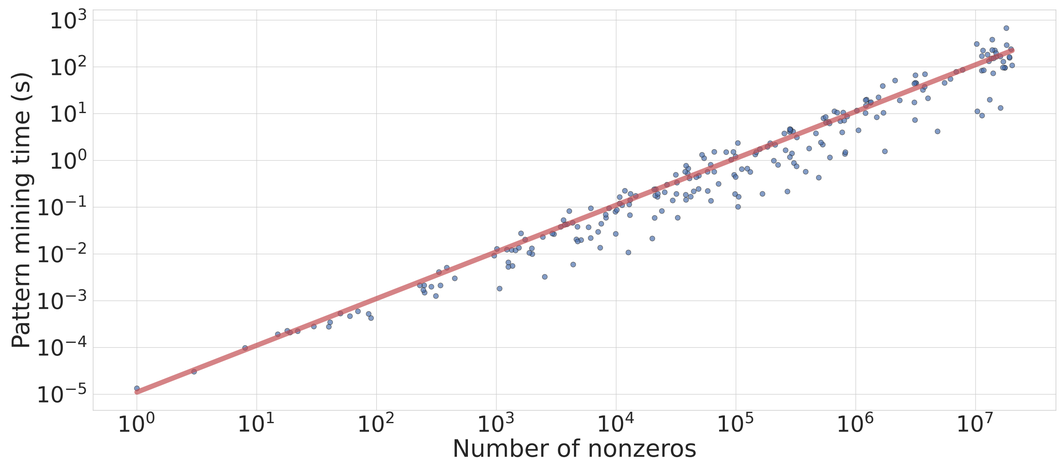
- The UZP File Format is divided into several differentiated sections:
  - Header: Describes the sparse matrix dimensions, nnz, etc.
  - Dictionary of Shapes (DoSh): Dictionay of different kinds of shapes.
  - List of Origins (LO): Where a shape of shape_id kind starts.
  - Vector of not-Included/Incorporated Data (ninc data): Indices of ninc data.
  - Vector of Included/Incorporated Data (inc data): Data values from LO and ninc.

Header	DoSh	LO	ninc data	inc data
nnz, inc data len, nrows, ncols, dimensions, shapes, data_ptr...	{ shape_id: _, lenghts: _, strides: _, lattice: _, } ...	total origins [(shape_id, row, col, data_offset), ...]	type of encoding: { 0: CSR; 1: CSC; 2: COO } [data] { CSR: row_ptr, col_idx CSC: col_ptr, row_idx, COO: row_idx, col_idx }	data in the order of traversal described by the LO, sorted by shape_id, followed by the values pointed by the ninc data coords

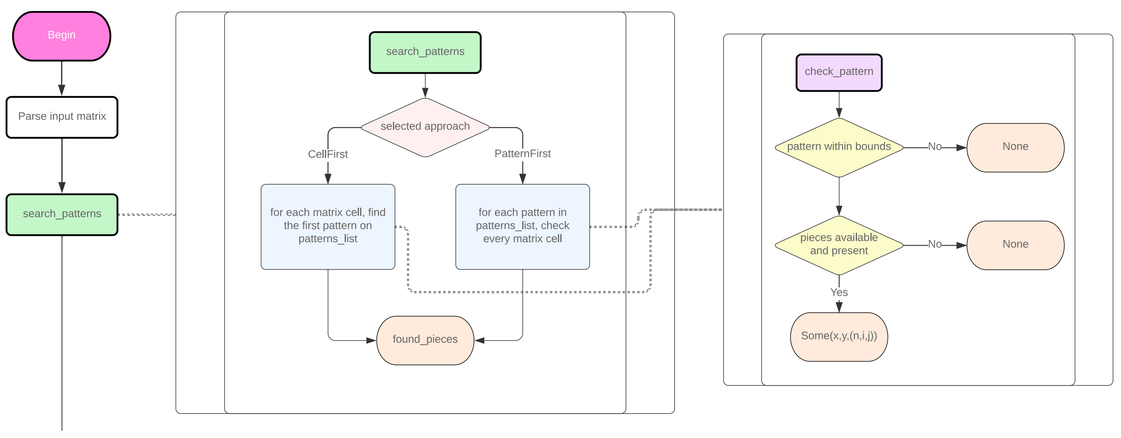
[a] Trailing md is for clarification purposes only. It is not necessary at all and can be easily inferred.

Pattern Search

- Very demanding time constraints, as pattern searching needs to be done in a reasonable time given the large sizes of many sparse matrices.
- Output compression and quality varies greatly depending on the algorithm and its hyper-parameters.

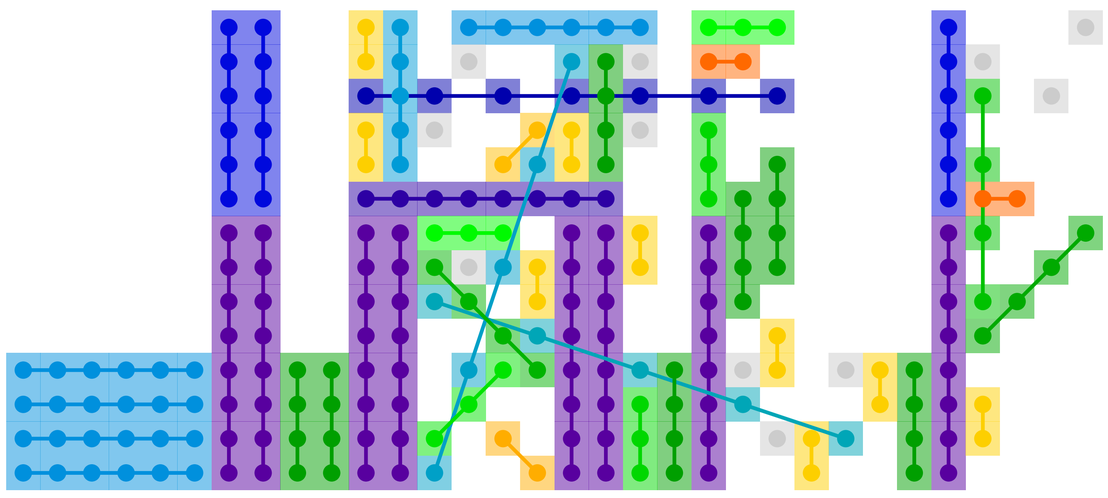


- Further compression can be achieved by simplifying two or more equispaced polyhedra together:
  - *Grouping*: Multiple (same lattice) consecutive polyhedra into a single one. These polyhedra must be consecutive and offset by $n * lattice$. e.g. $\{(0,0),\{(0,1),8\}_{1d}\} + \{(0,8),\{(0,1),8\}_{1d}\} \rightarrow \{(0,0),\{(0,1),16\}_{1d}\}$
  - *Augmentation*: Multiple same type (lattice and n) polyhedra with origins offset by the same amount one to the next. e.g. $\{(0,0),\{(0,1),8\}_{1d}\} + \{(3,2),\{(0,1),8\}_{1d}\} + \{(6,4),\{(0,1),8\}_{1d}\} \rightarrow \{(0,0), \{\{(3,2),3\}_{1d} \times \{(0,1),8\}_{1d}\}_{2d}\}$



Pattern Visualization

- As an example, we provide a polyhedral representation view of the sparse matrix `Maragal_1` from `sparse.tamu.edu`, rotated 90 degrees counterclockwise.



Acknowledgements

CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01).
Grant PID2022-136435NB-I00, funded by MCIN/AEI/ 10.13039/501100011033 and by "ERDF A way of making Europe", EU. Predoctoral grant of Alonso Rodríguez-Iglesias ref. FPU2022/01651, funded by the Ministry of Science, Innovation and Universities.