SpZip: Architectural Support for Effective Data Compression In Irregular Applications

Yifan Yang, Joel S. Emer, Daniel Sanchez
{yifany, emer, sanchez}@csail.mit.edu

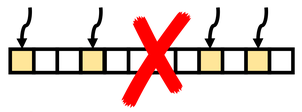
1. Motivation

- Irregular applications, such as graph analytics and sparse linear algebra, are important workload domains
- Irregular applications are often **memory bound**
- **Data compression** is an attractive approach to accelerate irregular applications

Prior Work

Hardware compression units for **sequentially accessed long streams**
e.g., IBM z15 [ISCA'20]

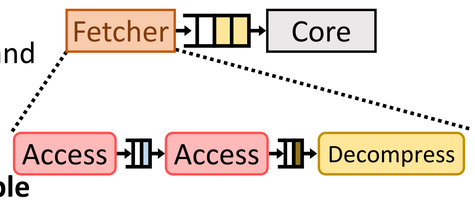
Compressed memory hierarchies support **random accesses**
e.g., VSC [ISCA'04]



2. Overview

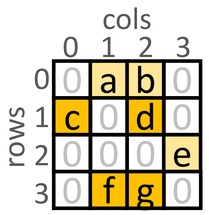
This work is optimized for **indirect, data-dependent accesses to short streams**

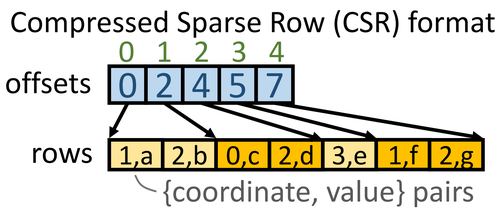
- Specialized hw to accelerate data access and decompression
- hw can be configured using a set of **composable** operators expressing various traversal patterns

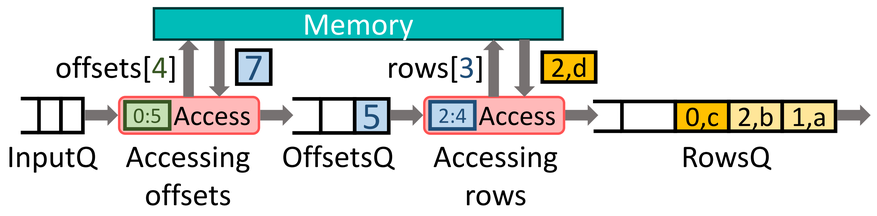


3. Dataflow Configuration

Sparse matrix traversal configuration

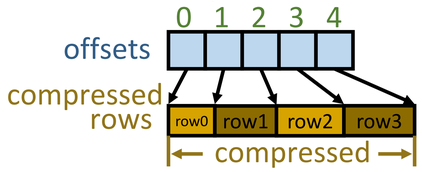






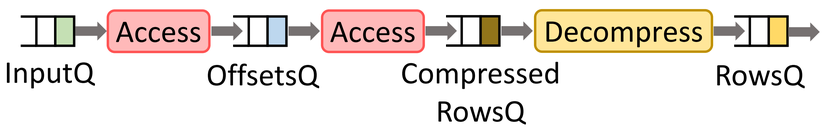
Row-compressed sparse matrix traversal configuration

CSR with individually compressed rows



Traversal Pseudocode

```
for i in range(numRows):                    in Fetcher
  for {col, val} in decompress(
                  rows[offsets[i]:
                       offsets[i+1]]):
    visit({col, val})                       in Core
```

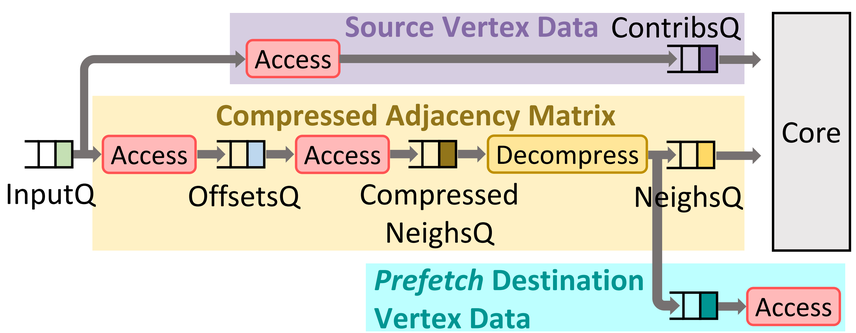


Multiple data structure traversal configuration in PageRank

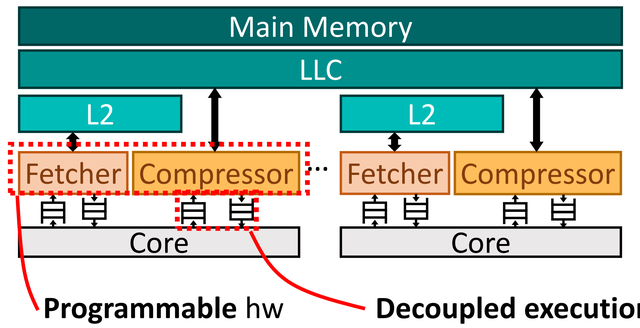
PageRank (PR) Pseudocode

```
def PRIter(Graph g, Array scores, Array contribs)
  for src in range(g.numVertices):
    for dst in decompress(g.neighs[g.offsets[src]:
                                   g.offsets[src+1]]):
      scores[dst] += contribs[src]
```

in Core — in Fetcher



4. SpZip Design

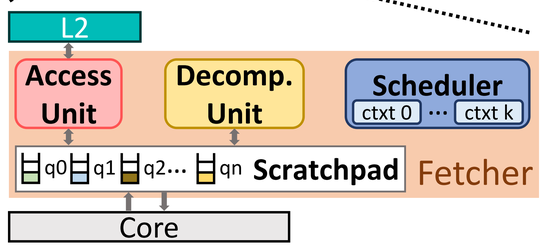


Programmable hw handling various traversal patterns

Decoupled execution hiding access and decompression latency

Compressor compresses newly generated data before storing it off-chip

Fetcher accelerates data structure traversal and decompression



Operators are **time-multiplexed** on the same physical unit

5. Evaluation

SpZip improves performance and reduces traffic