

Enabling Efficient Random Access to Hierarchically-Compressed Data

Feng Zhang †, Jidong Zhai ♦, Xipeng Shen #, Onur Mutlu ★, Xiaoyong Du †

†Renmin University of China

♦Tsinghua University

#North Carolina State University

★ETH Zürich



ETH Zürich

Outline

1. Background
2. Motivation
3. Operations to Support
4. Challenges
5. Our Solution
6. Evaluation
7. Conclusion

1. Background

- TADOC: Text Analytics Directly on Compression

Input:

file0: w1 w2 w3 w1 w2 w4
w1 w2 w3 w1 w2 w4

file1: w1 w2 w1

(a) Original data

Rules:

R0 → **R1** **R1** SPT1 **R2** w1
R1 → **R2** w3 **R2** w4
R2 → w1 w2

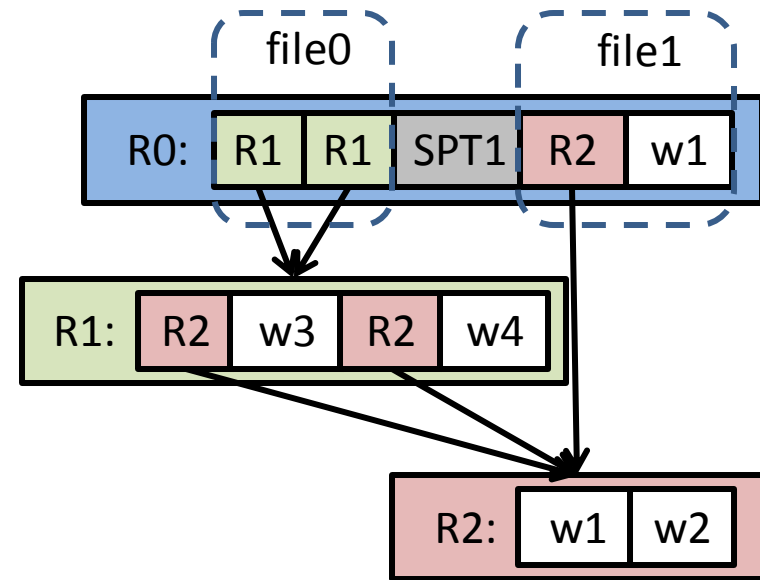
(b) TADOC compressed data

w1: 0 w2: 1 w3: 2
w4: 3 **R0**: 4 **R1**: 5
R2: 6 SPT1: 7

(d) Numerical representation

4 → 5 5 7 6 0
5 → 6 2 6 3
6 → 0 1

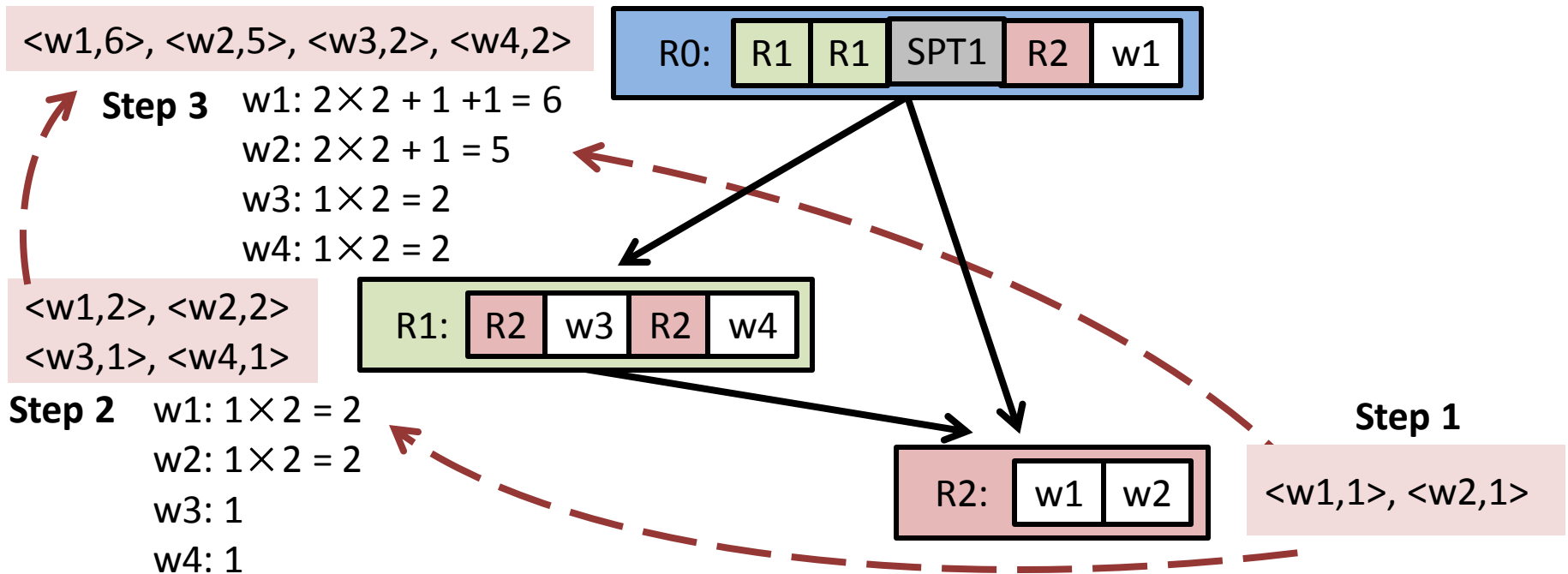
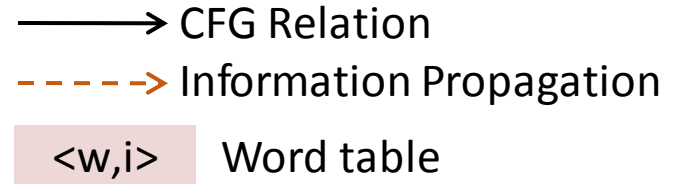
(e) Compressed data in numerical form



(c) DAG Representation

1. Background

- Example: word count



2. Motivation



News



Legal files

```

05:52:19 info AboutLog 170 0 "Changing App Context: (0: 1) -> (0: 4097)"
05:52:19 info Status 163 0 "INSTALL SPACE REQUIRED: (0: 108896) KB; SPACE AVAILABLE: (1: 12361979) KB; LOC: (2: C:)\"
05:52:19 info Step 25 56502 "User advanced from '0: Connecting' page to '1: Summary' page."
05:52:19 info Selection 99 0 "The following lines show the default selections for the Summary page when it first appears:"
05:52:19 info Selection 99 0 "This is a new installation of 2016 SP 6."
05:52:19 info Selection 102 0 "Download Options: (Operation=install only (do not download) [Enabled: 1]), (Background downloader=Do not use [Enabled: 1]) [Enabled]
05:52:19 info Selection 103 0 "Installation Location: (Installation location=C:\Program Files\SOLIDWORKS Corp [Enabled: 1]), (Install from=C:\SOLIDWORKS\DOWNLO
05:52:19 info Selection 104 0 "ToolboxHide Wizard Options: (Toolbox installation location=C:\SOLIDWORKS Data [Enabled: 1]), (Toolbox installation method=New Tool
05:52:19 info Selection 105 0 "Estimated installation size: 9.7 GB"
05:52:20 info Step 25 56502 "User advanced from '0: Summary' page to '1: Products to install' page."
05:52:24 info Selection 37 56511 "The user clicked the button representing the (0: OK) operation on the (1: Products to install) page."
05:52:24 info Selection 38 56510 "The user changed the installation action of (0: ProductView 360 Network Render Client) to: (1: Install ProductView 360 Network Render Client 2
05:52:24 info Selection 38 56510 "The user changed the installation action of (0: SOLIDWORKS Electrical) to: (1: Do not install SOLIDWORKS Electrical 2016 SP10)"
05:52:24 info Selection 38 56510 "The user changed the installation action of (0: APY Tools) to: (1: APY Tools 2016 SP6) (install manually)"
05:52:24 info Selection 38 56510 "The user changed the installation action of (0: SOLIDWORKS Workgroup PDM x64 Client Add-in) to: (1: Do not install SOLIDWORKS Workg
05:52:24 info Selection 38 56510 "The user changed the installation action of (0: SOLIDWORKS Workgroup PDM x64 Add-in) to: (1: Do not install SOLIDWORKS Workgroup PDM
05:52:24 info Status 163 0 "INSTALL SPACE REQUIRED: (0: 841178) KB; SPACE AVAILABLE: (1: 123606692) KB; LOC: (2: C:)\"
05:52:25 info Status 163 0 "INSTALL SPACE REQUIRED: (0: 841178) KB; SPACE AVAILABLE: (1: 123606692) KB; LOC: (2: C:)\"
05:52:27 info Step 25 56502 "User advanced from '0: Summary' page to '1: Summary' page."
05:52:33 info Selection 37 56511 "The user clicked the button representing the (0: OK) operation on the (1: Download Options) page."
05:52:33 info Step 25 56502 "User advanced from '0: Download Options' page to '1: Summary' page."
05:52:37 info Step 25 56502 "User advanced from '0: Summary' page to '1: Installation Location' page."
05:52:48 info Message 9 56500 "Dialog shown: (0: C:\Program Files\SOLIDWORKS 2016 The path does not exist. Do you want to create it now?). User selected: (1: Yes)"
05:52:48 info Status 163 0 "INSTALL SPACE REQUIRED: (0: 841178) KB; SPACE AVAILABLE: (1: 123606692) KB; LOC: (2: C:)\"
05:52:48 info Step 23 56500 "Page '0: Installation Location' is being removed or refreshed."
05:52:50 info Selection 37 56511 "The user clicked the button representing the (0: OK) operation on the (1: Installation Location) page."
05:52:50 info Status 163 0 "INSTALL SPACE REQUIRED: (0: 841178) KB; SPACE AVAILABLE: (1: 123606692) KB; LOC: (2: C:)\"
05:52:50 info Status 163 0 "INSTALL SPACE REQUIRED: (0: 841178) KB; SPACE AVAILABLE: (1: 123606692) KB; LOC: (2: C:)\"
05:52:50 info Step 25 56502 "User advanced from '0: Installation Location' page to '1: Summary' page."
05:52:56 info Step 25 56502 "User advanced from '0: Summary' page to '1: Toolbox Options' page."
05:52:57 info Step 23 56500 "Page '0: Toolbox Options' is being removed or refreshed."
05:54:43 info Selection 37 56511 "The user clicked the button representing the (0: OK) operation on the (1: Toolbox Options) page."
05:54:44 info Status 163 0 "INSTALL SPACE REQUIRED: (0: 841178) KB; SPACE AVAILABLE: (1: 123606698) KB; LOC: (2: C:)\"
05:54:44 info Step 25 56502 "User advanced from '0: Toolbox Options' page to '1: Summary' page."
05:54:48 info Selection 37 56511 "The user clicked the button representing the (0: <Back> operation on the (1: Summary) page."
05:54:48 info Selection 38 56513 "The following lines show the selections for the Summary page when the user selected to begin a download or installation."
05:54:48 info Selection 99 0 "This is a new installation of 2016 SP 6."
05:54:48 info Selection 102 0 "Download Options: (Operation=install only (do not download) [Enabled: 1]), (Background downloader=Do not use [Enabled: 1]) [Enabled]
05:54:48 info Selection 103 0 "Installation Location: (Installation location=C:\Program Files\SOLIDWORKS 2016 [Enabled: 1]), (Install from=C:\SOLIDWORKS\DOWNLOA
05:54:48 info Selection 104 0 "ToolboxHide Wizard Options: (Toolbox installation location=C:\Program Files\SOLIDWORKS 2016 [Enabled: 1]), (Toolbox installation method=New Tool
05:54:48 info Selection 44 56517 "The estimated installation size is: 9.7 GB"
05:54:48 info Selection 44 56517 "The source folder that will be used is: (0: G:\SOLIDWORKS\DOWNLOADED\SOLIDWORKS 2016 x64 SP6)"
05:54:48 info Selection 45 56517 "The installation Manager Source Folder that will be used is: (0: C:\SOLIDWORKS\DOWNLOADED\SOLIDWORKS 2016 x64 SP6\10488)"
05:54:48 info Step 25 56502 "User advanced from '0: Summary' page to '1: Install Progress' page."
05:54:48 info Selection 107 0 "Writing serial numbers to the registry."
    
```

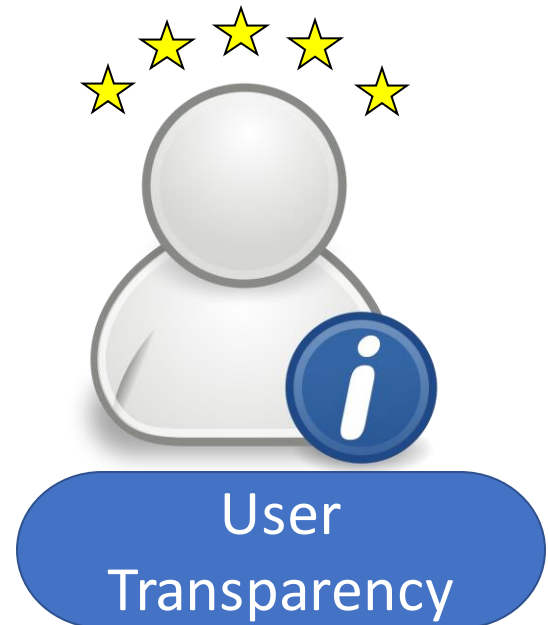
Logs

Random Access

- search
- extract
- count
- insert
- append

3. Operations to Support

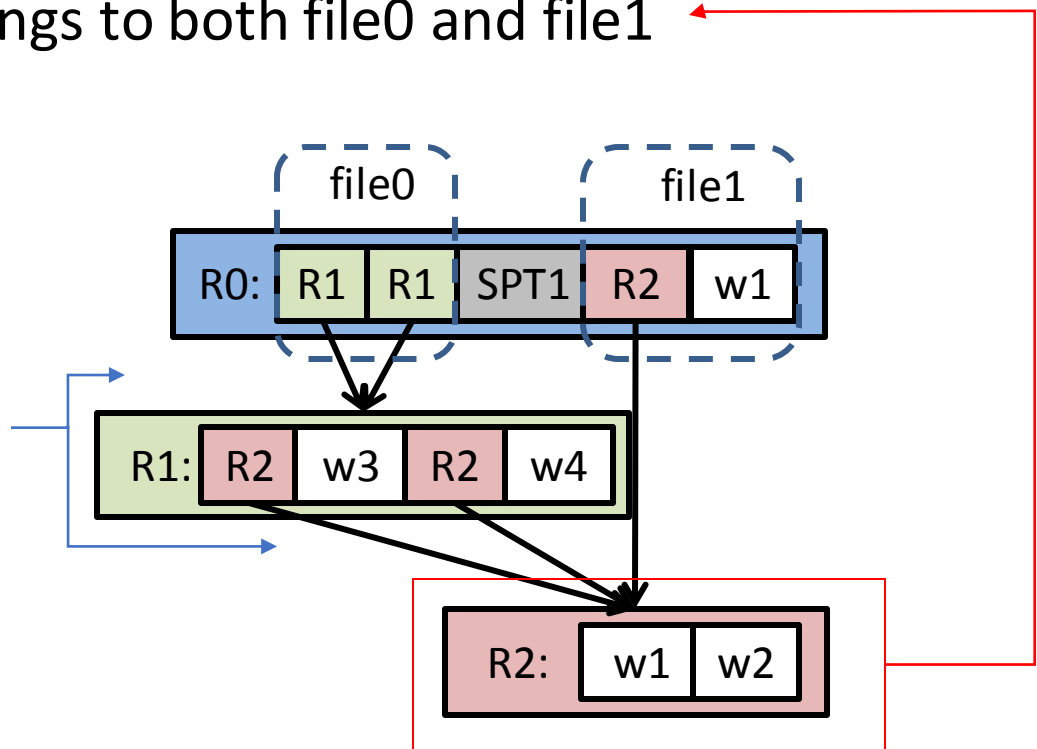
- Random Access
 - **extract(file,offset,length)**
 - **insert(file,offset,string)**
 - **search(file,word)**
 - **append(file,string)**
 - **count(file,word)**



4. Challenges

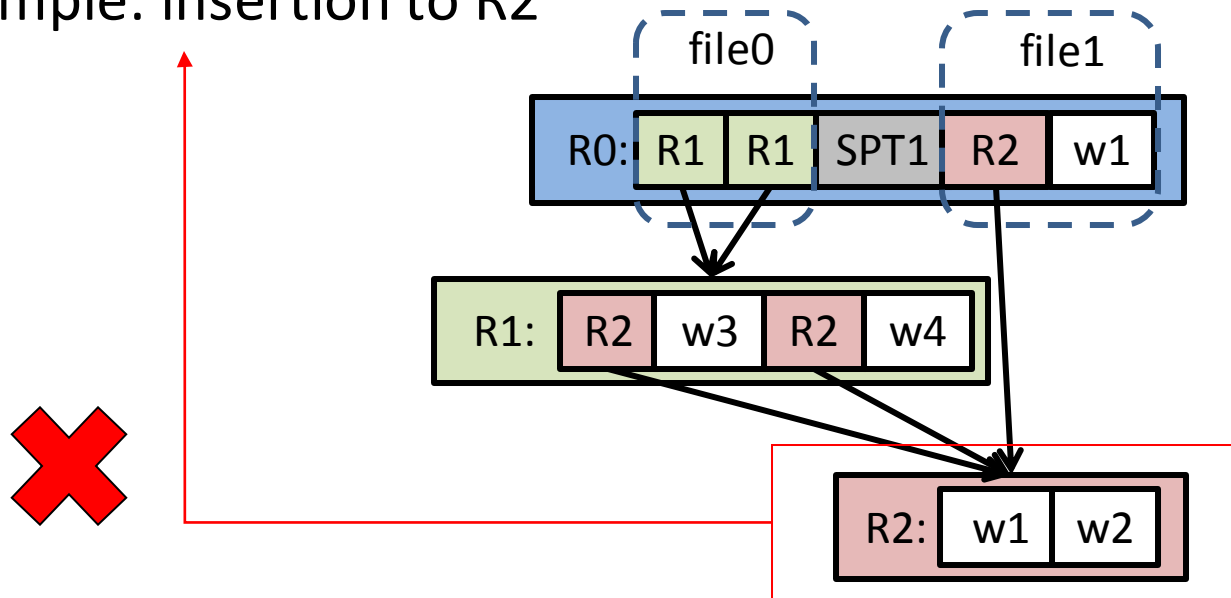
- Hierarchical Structure of the DAG
 - Example: R2 belongs to both file0 and file1

- Uni-Directionality
 - Edges



4. Challenges

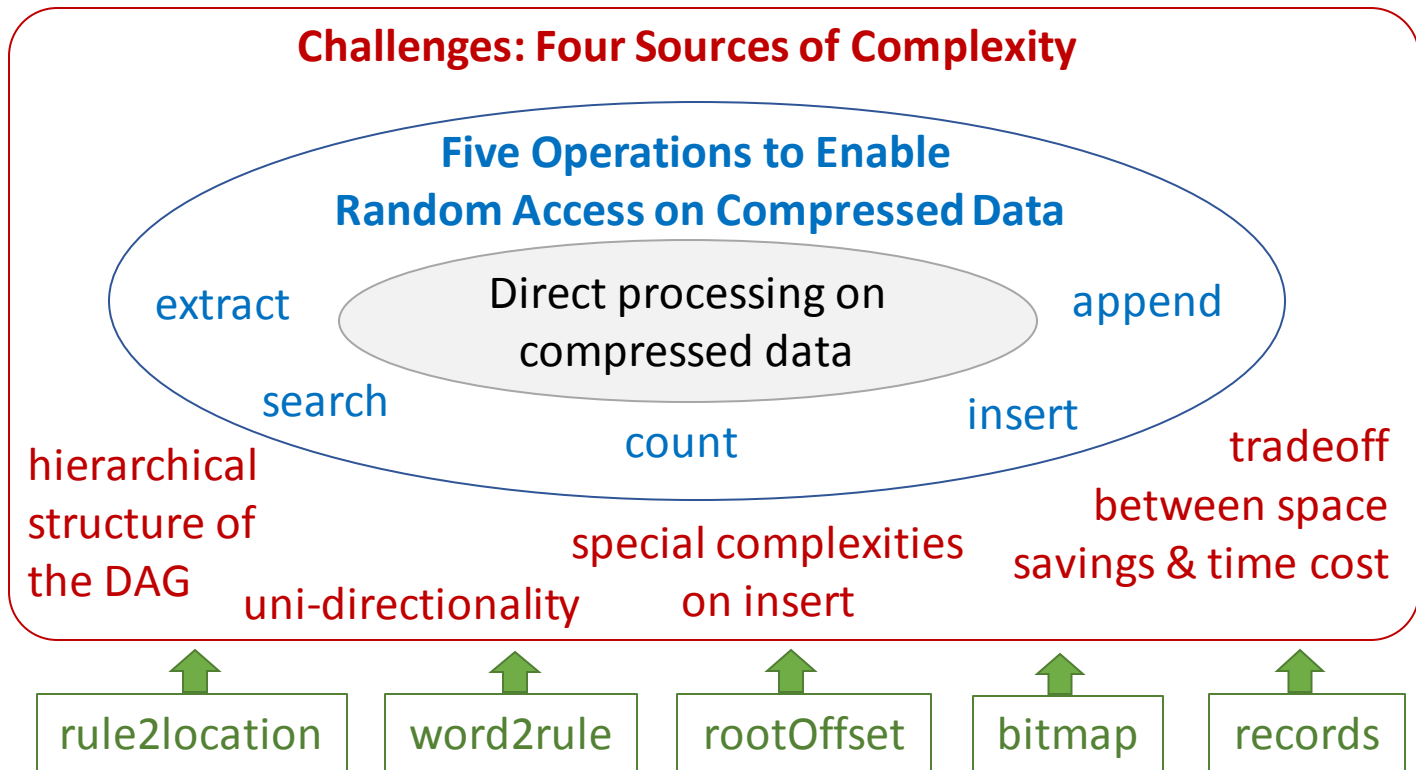
- Special Complexities on Insert
 - Example: insertion to R2



- Tradeoff between Space Savings and Time Cost
 - Index space cost

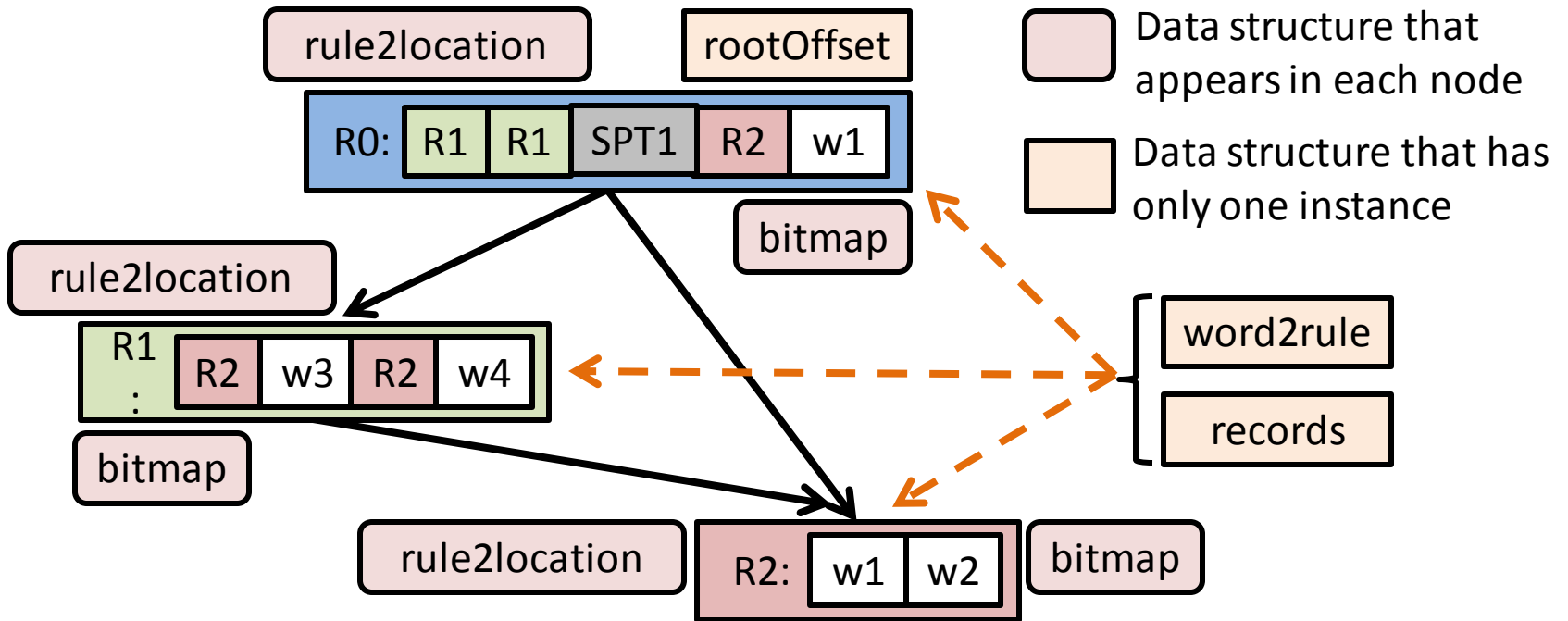
5. Our Solution

Solution Techniques: Five Data Structures



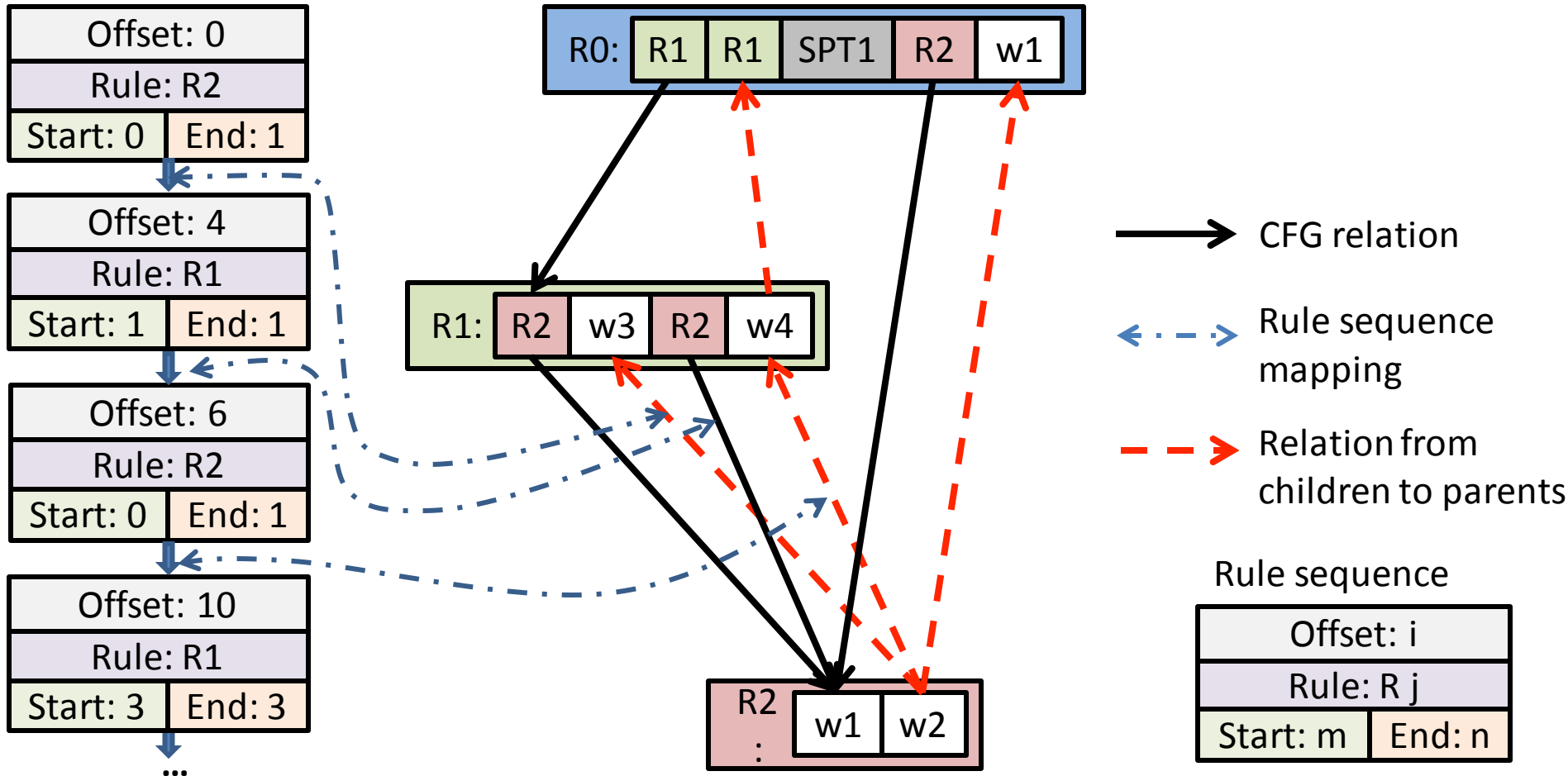
5. Our Solution

- Relations for data structures



5. Our Solution

- `extract(file,offset,length)`



5. Our Solution

- `search(file,word)`
 - `word2rule`
 - `rule2location`

- `count(file,word)`
 - `word2rule`
 - `rule2location`

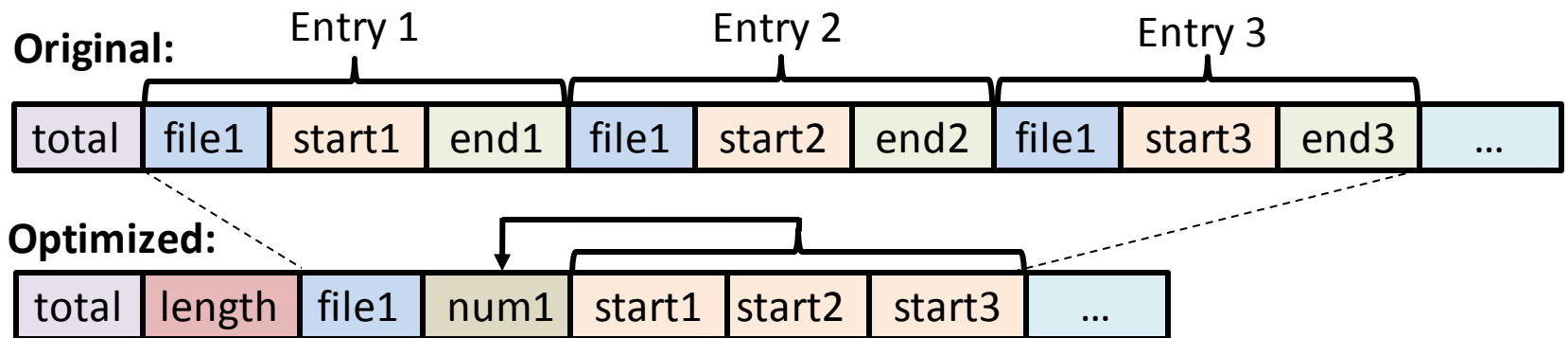
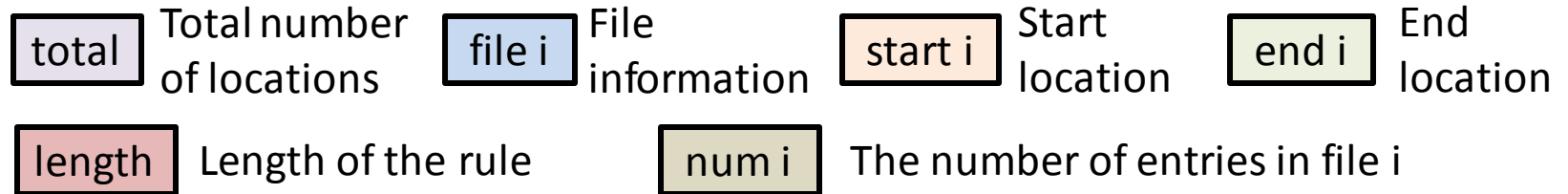


Illustration of *rule2location* optimization.

5. Our Solution

- insert(file,offset,string)
- append(file,string)

The Record Data Structure

```
struct Record{
  int fileID; // file, such as file1
  int fileOffset; //file offset to insert, such as 100
  int ruleID; // the rule ID to insert, such as 0
  int ruleLocation; //the inserted location, such as 2
  int replaceWord; //the replaced word, such as w2
  string content; //content string
  int ptr; //the recordID inserted at the same place. Default is -1
  int ruleStartOffset; //the starting offset of the rule to insert, such as 0
};
```

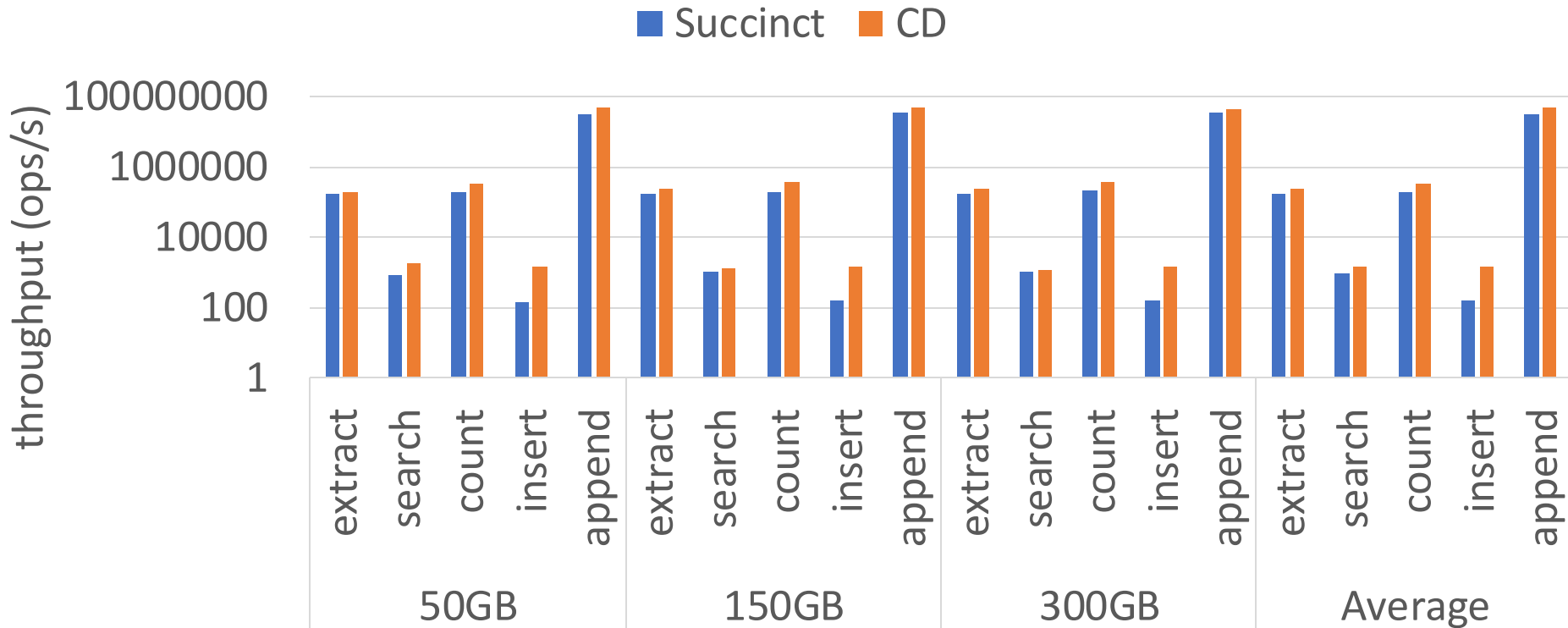
6. Evaluation

- Five operations
 - search, extract, count, insert, append
- Five datasets
 - 580 MB ~ 300 GB
- Two platforms
 - Single node
 - Spark cluster (10 nodes on Amazon EC2)



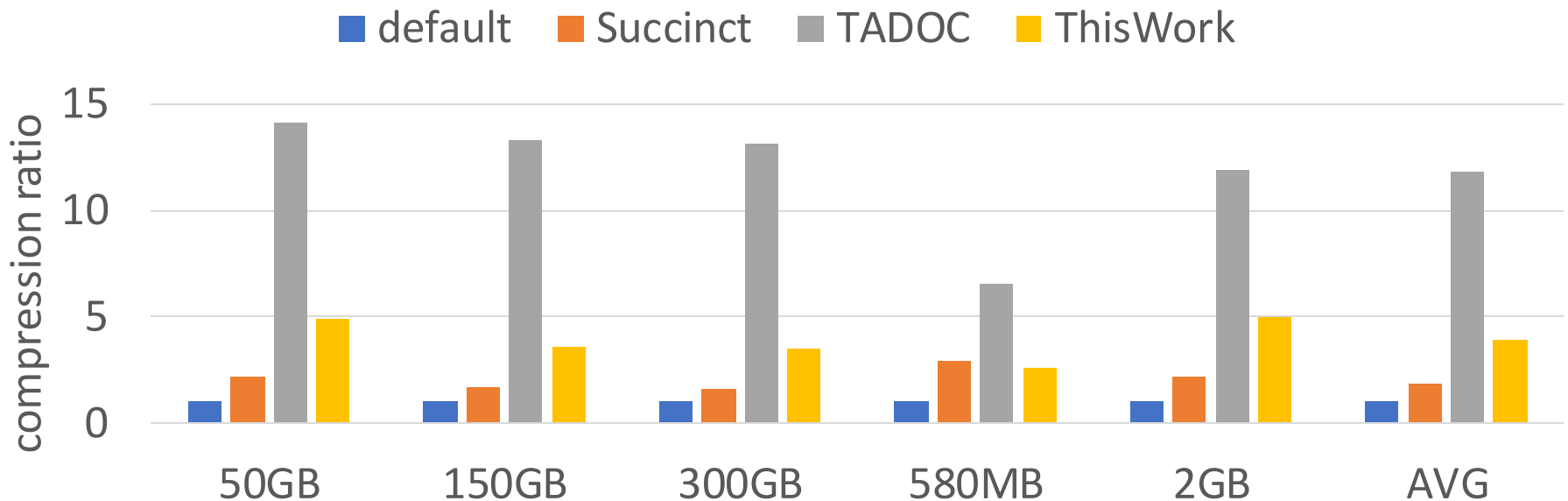
6. Evaluation

- On average, the overall throughput of our proposed techniques (CD) is **3.1x** of Succinct's throughput in a distributed environment.



6. Evaluation

- The average compression ratio we observe is **3.9**, which is still much more compact than the **1.8** compression ratio of Succinct.
- compression ratio = original size / compressed data size



7. Conclusion

- We provide a set of new techniques that enable efficient random access operations on hierarchically-compressed data
- Compatible with TADOC: data traversal operations
- We remove a major barrier against practical adoption of direct text analytics on compressed data.

Thanks!

- Any questions?

Feng Zhang †, Jidong Zhai ◊, Xipeng Shen #, Onur Mutlu ★, Xiaoyong Du †

†Renmin University of China

◊Tsinghua University

#North Carolina State University

★ETH Zürich



ETH Zürich