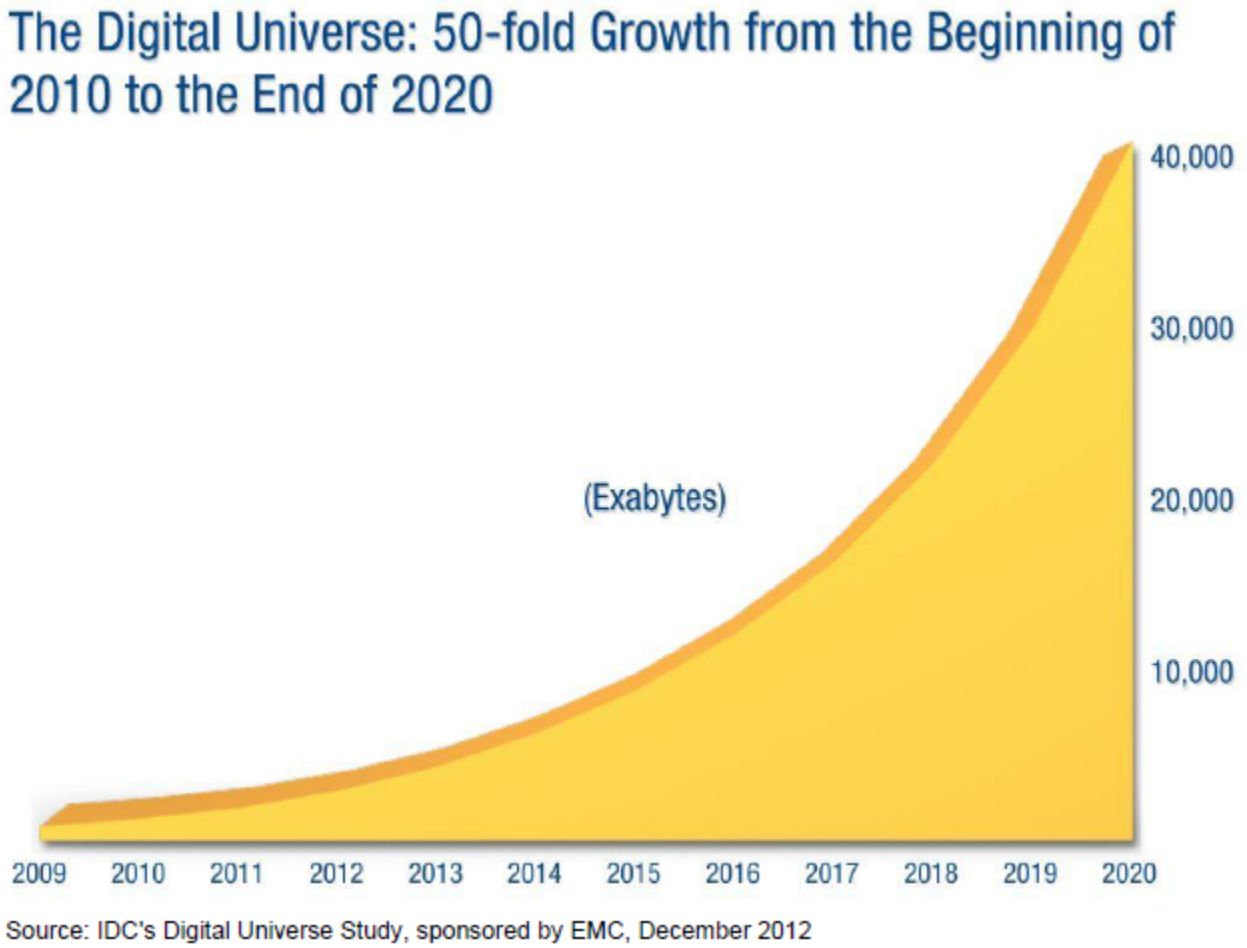
Deduplicating Compressed Contents in Cloud Storage Environment

Zhichao Yan[1], Hong Jiang[1], Yujuan Tan[2] and Hao Luo[3]
University of Texas Arlington[1]
Chongqing University[2]
University of Nebraska Lincoln[3]

Reality and Trend

- **Information explosion**



Reality and Trend

- **Pain point: how to manage storage growth**

2013-14	2012-13	IT / Storage Managers and Professionals
79%	77%	Managing storage growth
43%	45%	Designing, deploying, and managing Backup, Recovery, and Archive solutions
39%	36%	Making informed strategic/big-picture decisions (+8%)
38%	39%	Designing, deploying, and managing disaster recovery solutions
37%	31%	Designing, deploying, and managing storage in a virtualized server environment (+19%)
29%	27%	Lack of skilled storage professionals (+7%)
18%	16%	Designing, deploying, and managing storage in cloud computing environment. (+13%)
15%	15%	Lack of skilled cloud technology professionals
11%	10%	Convincing higher management to adopt cloud (+10%)
10%	7%	Infrastructure for Big Data analytics (+43%)
8%	4%	Managing external cloud service providers (+100%)

Table Source: IDC

Reality and Trend

- **Data Migrate to Cloud**

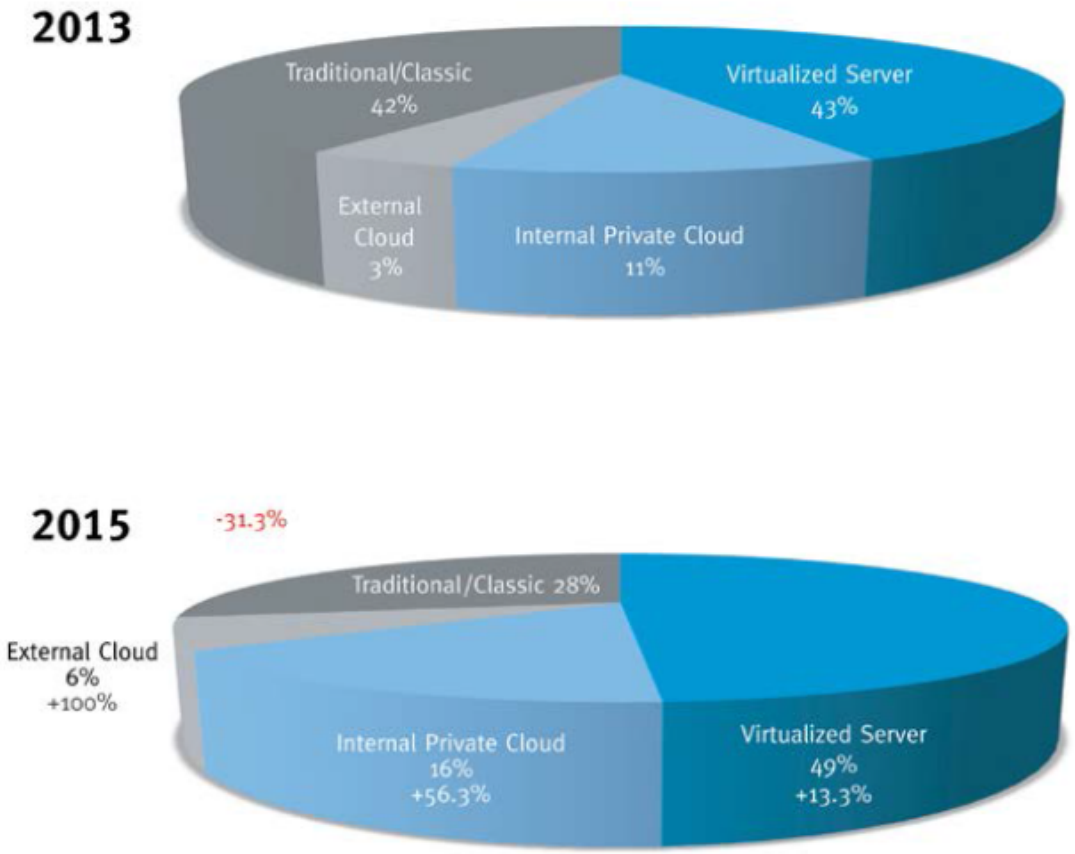


Figure Source: IDC

Reality and Trend

- **Network bandwidth (Low and Asymmetric)**

Broadband Speed Greater Than 10 Mbps(2014-2019) from Cisco

Region	>10 Mbps		>25 Mbps		>100 Mbps	
	2014	2019	2014	2019	2014	2019
Global	48%	68%	29%	33%	3%	7%
Asia Pacific	46%	73%	26%	37%	3%	8%
Latin America	27%	33%	9%	12%	1%	3%
North America	58%	74%	33%	45%	2%	8%
Western Europe	51%	62%	28%	37%	4%	10%
Central and Eastern Europe	53%	76%	34%	41%	2%	6%
Middle East and Africa	16%	20%	6%	8%	0.3%	1%

Summary of Existing Internet Plans of Time Warner Cable

TWC Plan	Ulti200	Ulti100	Extr	Basic	EvyDay
D/L Speeds(Mbps)	200	100	50	10	3
U/L Speeds(Mbps)	20	10	5	1	1
Price($/month)	60	50	40	30	15

Why data reduction

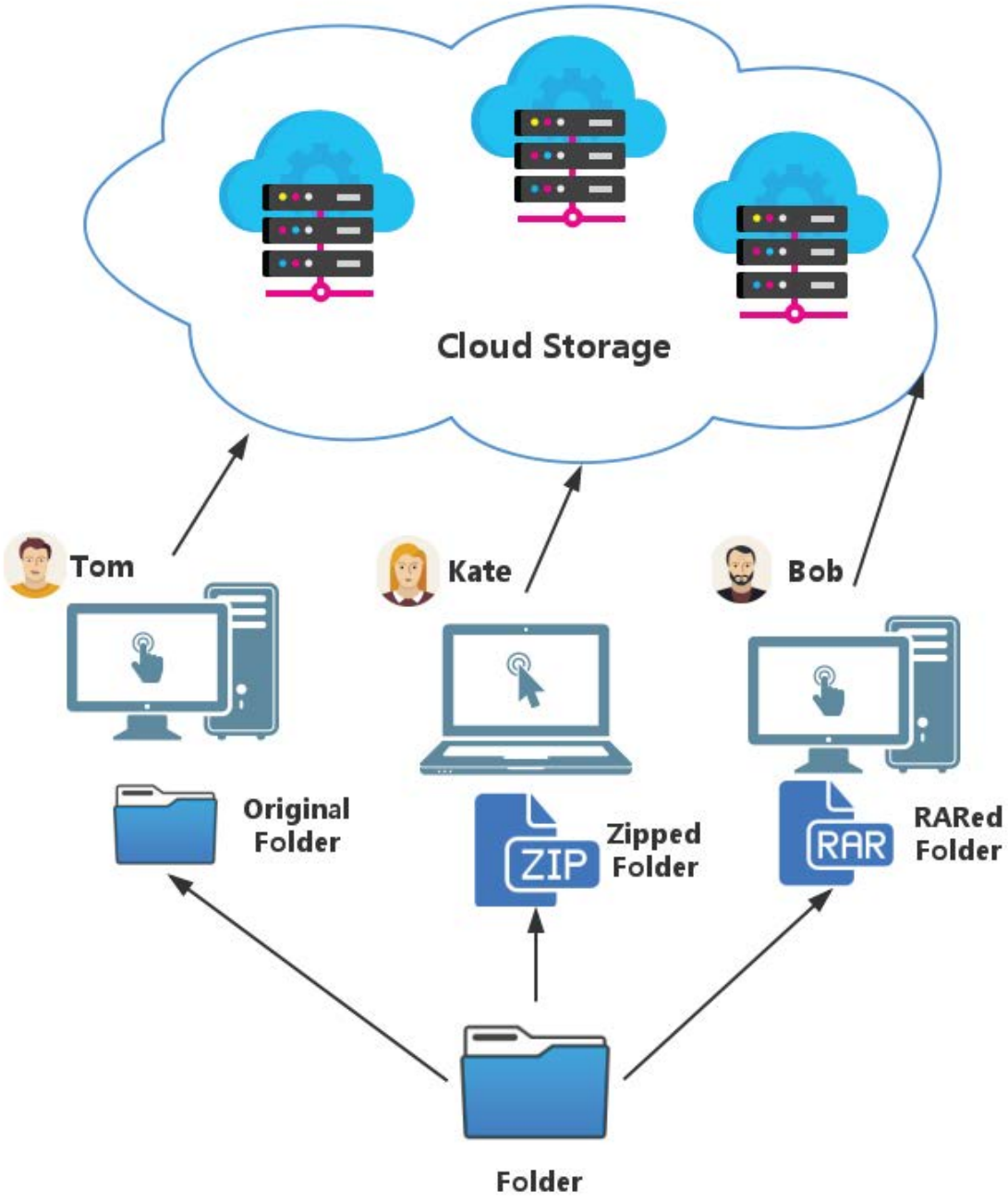
- **Information explosion → Huge amount of digital contents → How to store these data? →Cloud storage → Lower storage cost → Data reduction technology**
- **Network bandwidth→ Low and Asymmetric → How to transfer a large amount of data to cloud? → Data reduction technology**

Two common data reduction technologies

- **Data lossless compression**
  finds repeated strings within the specific range of the individual files and replaces them with a more compact coding scheme (Compression dictionary)
- **Data deduplication**
  identifies and removes the redundant files/chunks across all the files (maintaining pointers information to assemble the data from files/chunks for future access)

What will happen

- **Both end users and cloud service providers have performance (data transferring time) and economic (data storage cost) incentives to deploy data reduction technologies**
- **Cloud will become the digital content aggregating point in the digital universe, containing a lot of compressed packages from different end users**



A common scenario: Compression at the client side Dedup at the cloud side

Different users will use various compression tools to compress their data before sending to the cloud

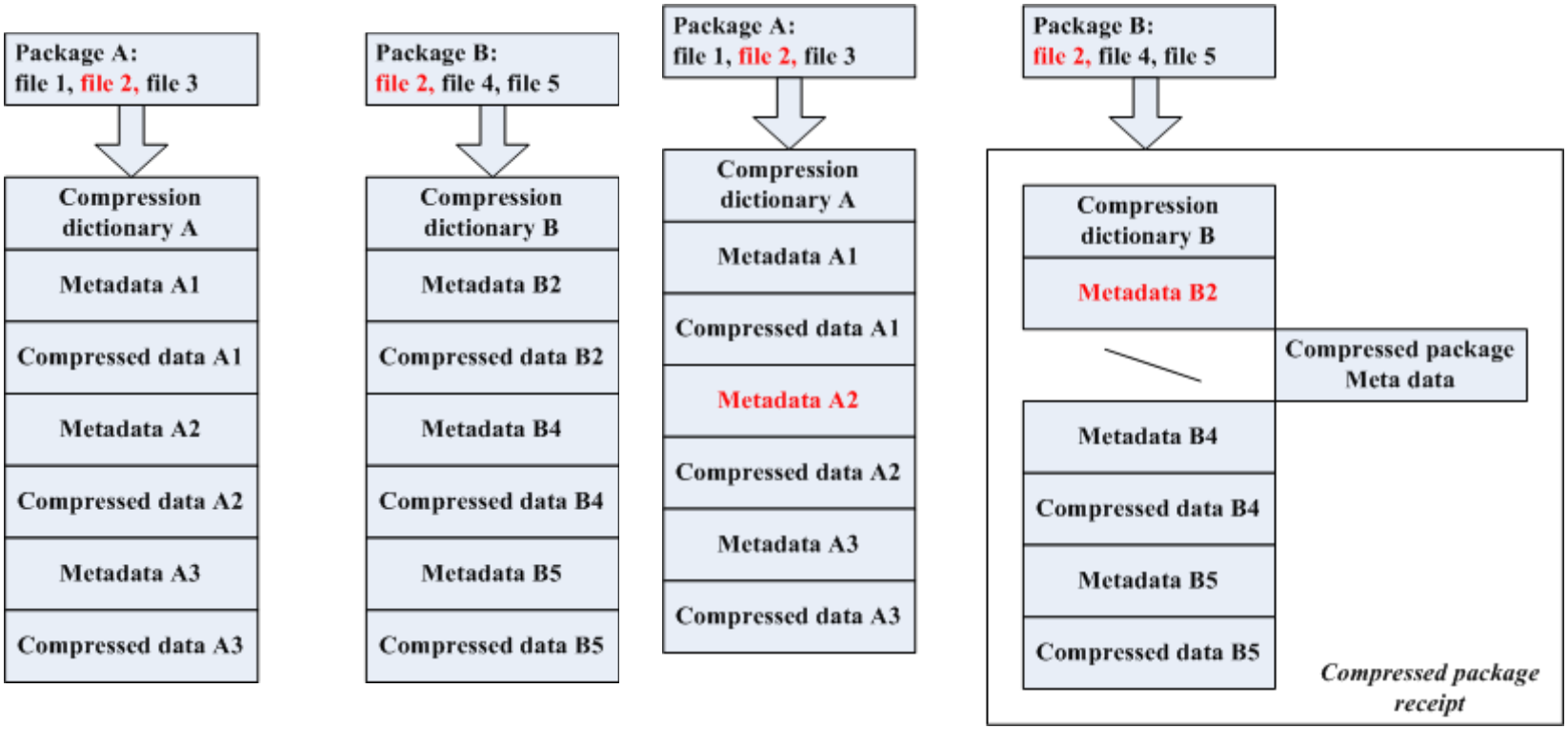
Problem

- Old dedup works well with plain data, but not with shuffled data in compressed packages
- Redundancy hidden within compressed contents might widely exist in cloud storage environment and will increase with the time
- Efficient cloud requires an approach to dedup such kind of redundant data within the compressed contents

X-Ray Dedup

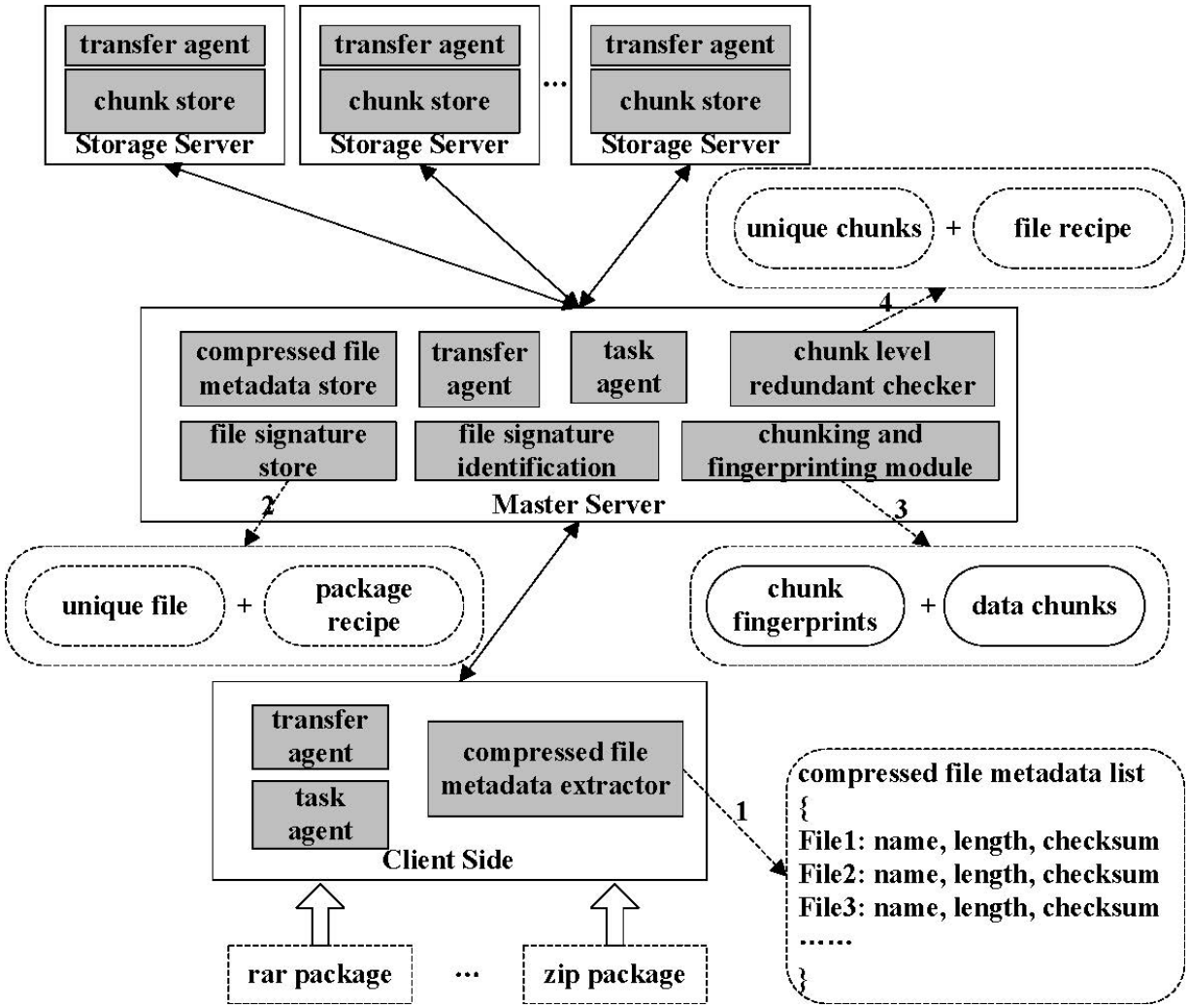
- Compression tools usually have some data integrity mechanism to avoid compressed data corruption
- Same checksum algorithm will generate the same checksum value for the same file no matter which compress algorithm it works with
- One most popular checksum mechanism is CRC32 (collision can be studied in future)
- Combined with original file length as ID for dedup

X-Ray Notion

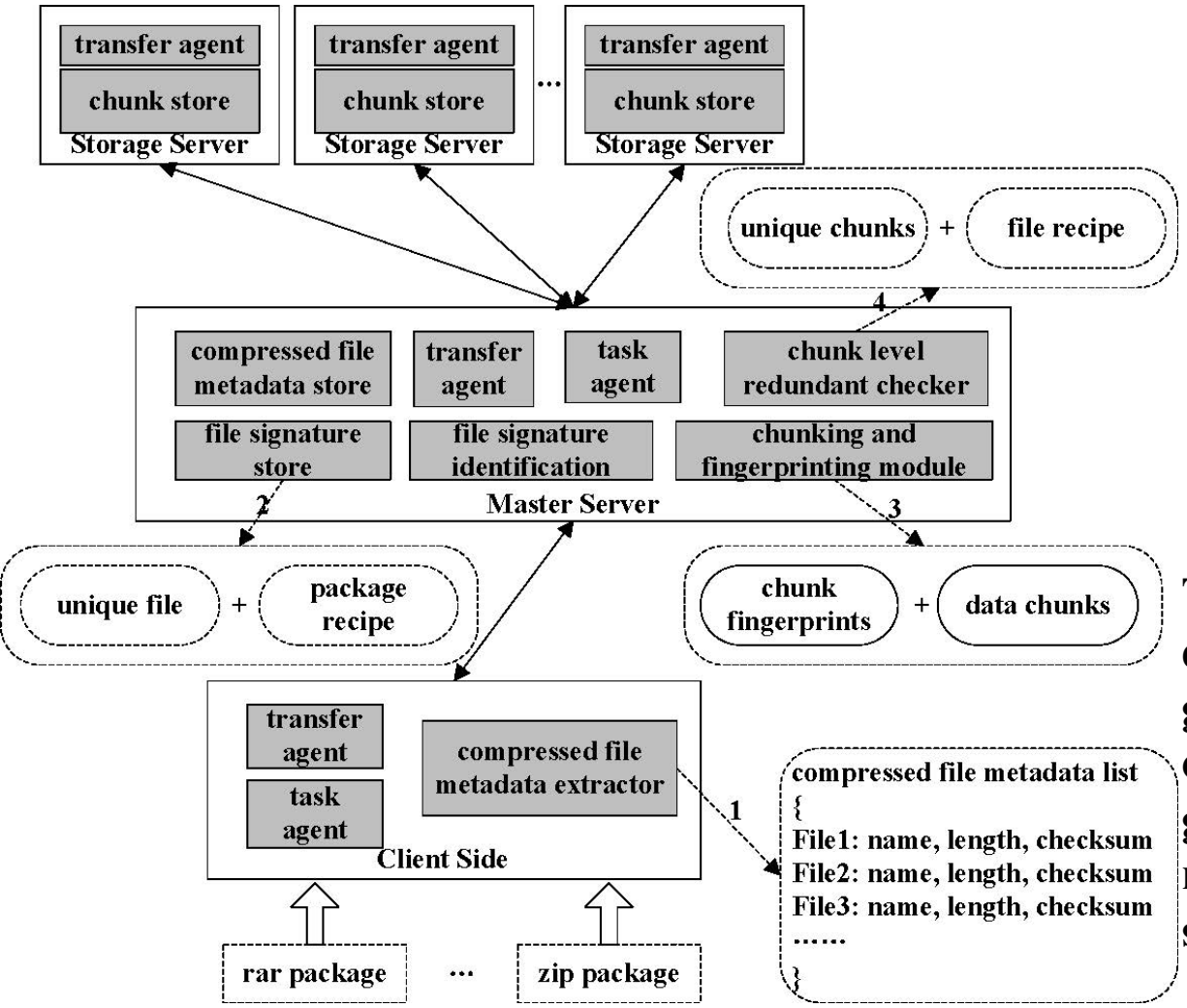


Checksums used in different compression tools: CRC32 , MD5, SHA1, RIPEMD-160, SHA256, SHA512, BLAKE2

System Overview



System Workflow



A file metadata extractor module on the client side extracts the metadata of compressed package by parsing through the compressed package (i.e., name, uncompressed length and checksum).

A file signatures store is used to help the file signature identification module identify and remove file-level data redundancy by its recorded files' metadata entries.

The unique (non-redundant) files are chunked to generate data chunks and their individual fingerprints.

The conventional chunk-level deduplication will be executed to generate file recipes and unique chunks. Finally, the previously generated package recipes, file recipes and unique chunks are stored to the storage servers.

Evaluation

- Based on chunk level dedup system (destor)
- Add an extra file level dedup for compressed contents
- Tar only checksum the header, we need to extend it to whole file content checksum for such kind of compressed tools like tar.gz and tar.xz, it can be translated at the sever side by adding some extra checksum information

Evaluation

Table 1: Compression tools

	tar	gz	xz	7z	rar
ubuntu	1.27.1	1.6	5.1.0α	9.20	4.20
windows	1.28-1	1.6	5.2.2	15.09β	5.31

Table 2: Sizes (KiB) of different compression formats under the Ubuntu / Windows platforms

	coreutils-8.25	linux-4.5-rc5
tar	49990 / 49990	642550 / 642550
xz	5591 / 5591	86287 / 86287
gz	12784 / 12784	132608 / 132609
7z	6169 / 5723	93561 / 89437
rar	12402 / 12401	156310 / 155135

20 versions of coreutils and 11 versions of Linux kernel

One version of coreutils or linux has about 2K~3K / 30K ~50K files

Some Results

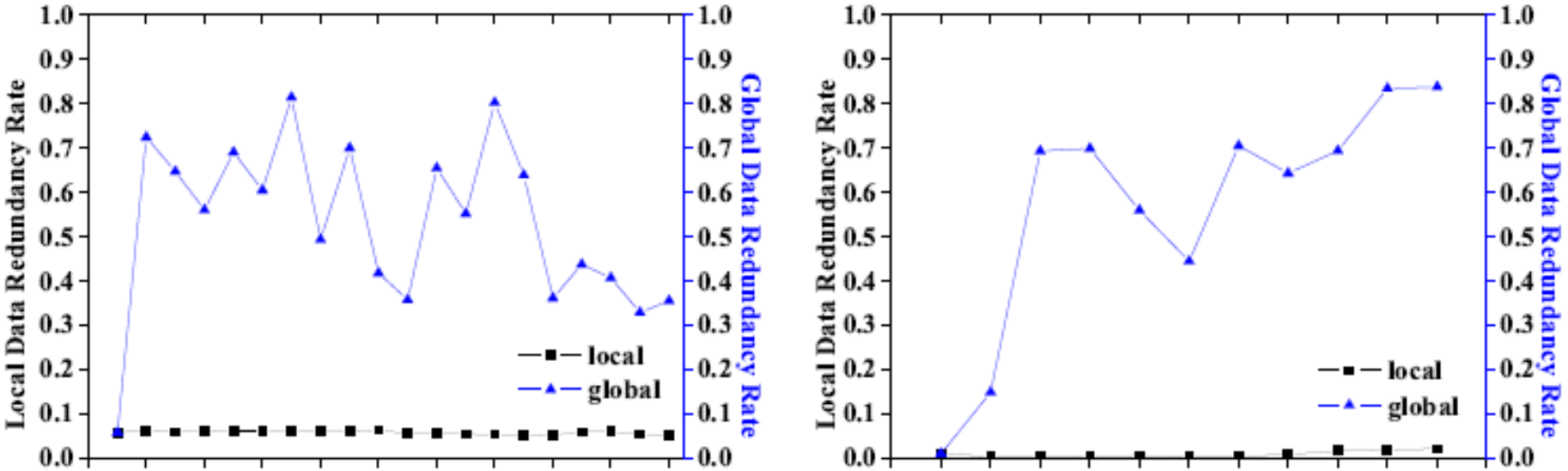
- What if we only deploy chunk level dedup on compressed packages (of the same content)

Table 3: Comparison of redundancy ratio (in percentage) between different compressed packages between the same content, whose row is Ubuntu and column is Windows

		coreutils				linux			
		xz	gz	7z	rar	xz	gz	7z	rar
coreutils	xz	100	0	0	0	0	0	0	0.05
	gz	0	7.6	0	0	0	0	0	0.05
	7z	0	0	0	0	0	0	0	0.05
	rar	0	0	0	1.0	0	0	0	0.05
linux	xz	0	0	0	0	100	0	0	0.05
	gz	0	0	0	0	0	5.6	0	0.05
	7z	0	0	0	0	0	0	0	0.05
	rar	0	0	0	0	0	0	0	0.24

Some Results

- **What about the hidden data redundancy? (local and global)**



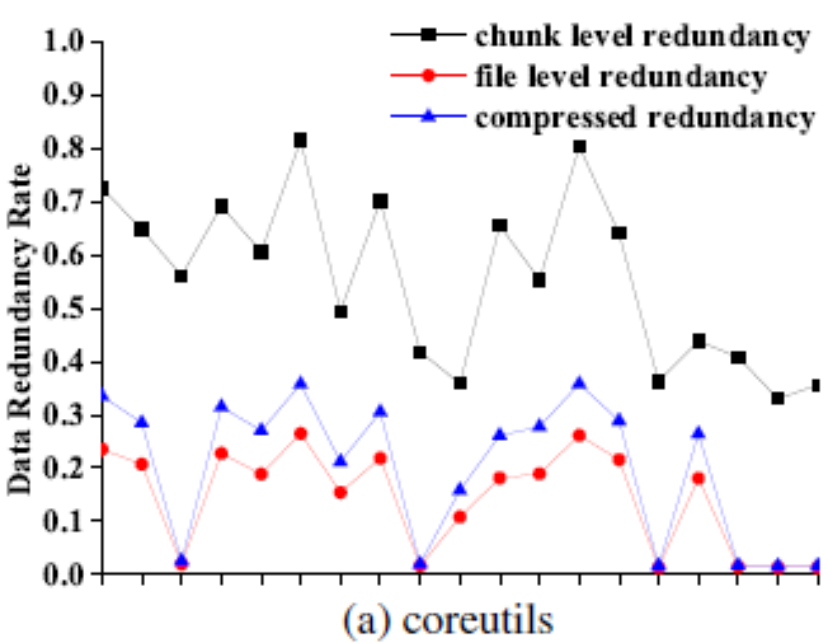
(a) coreutils (20 versions)

(b) linux (11 versions)

Figure 4: Real data redundancy throughout different versions of decompressed packages

Some Results

- **How much redundant data X-Ray dedup can reduce**



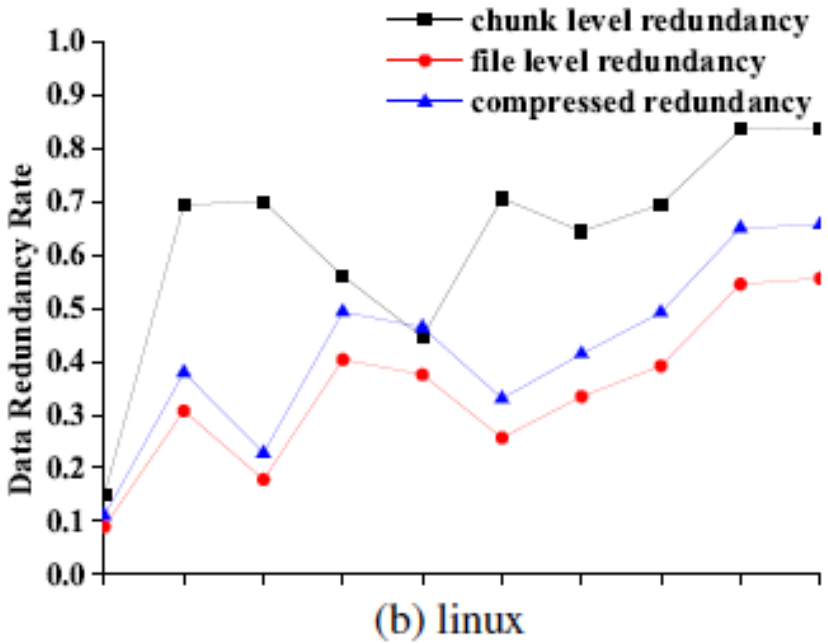


Figure 5: Compressed redundancy information of the X-Ray Dedup approach throughout all compressed packages

Compressed redundancy = compressed intact files' size / size of compressed package

Summary

- Find new ID (checksum + file length)to detect redundant file across the compressed packages
- An extra file level dedup designed for compressed files
- Significant reduce the capacity requirement for an efficient cloud storage environment

Thank you!