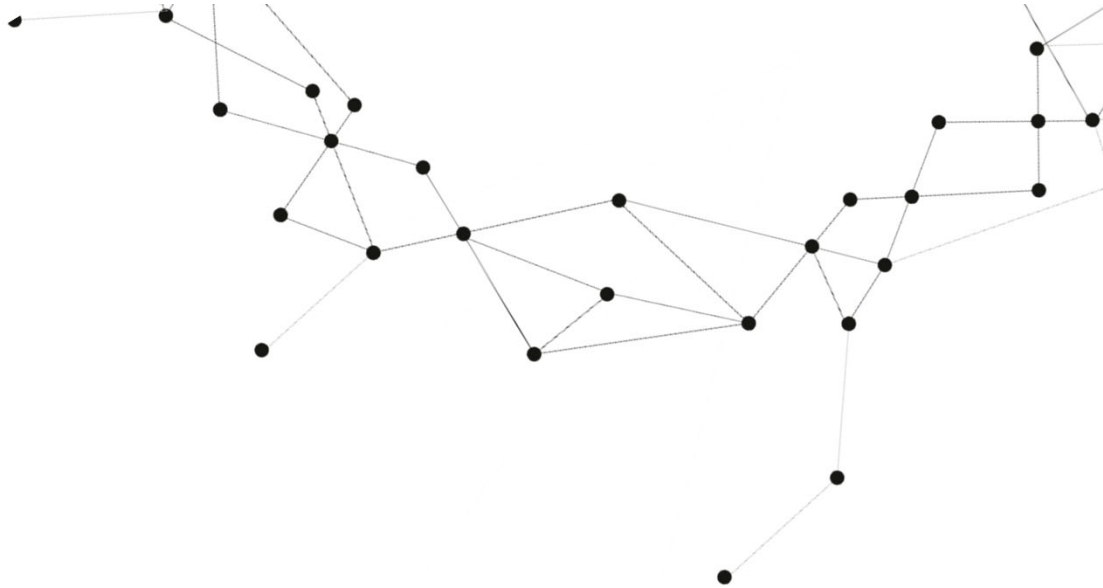


# CARD

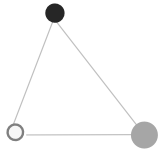
## Chunk Content Is Not Enough:

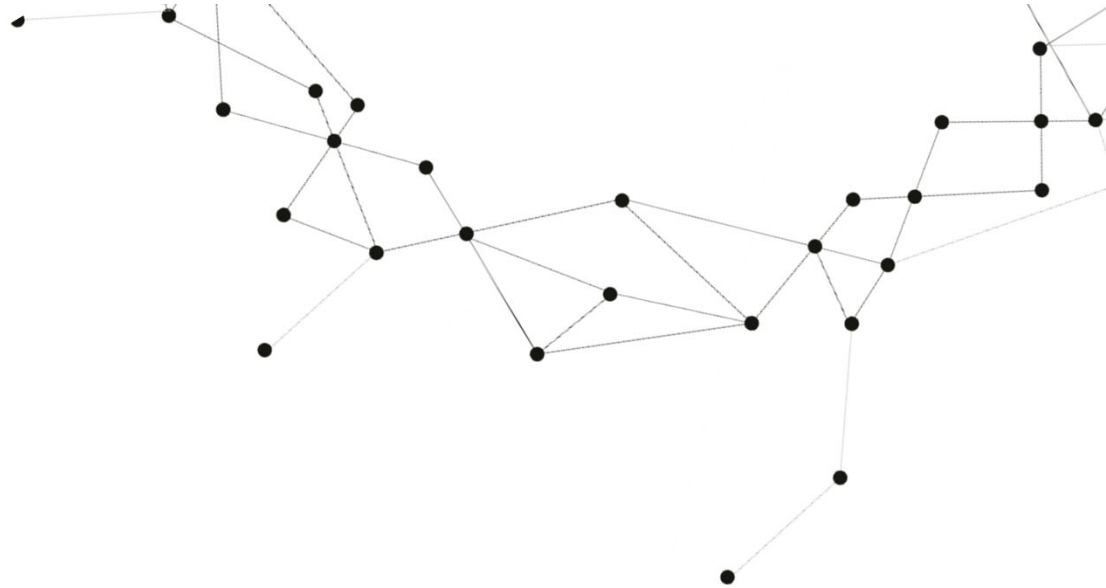
## Chunk-Context Aware Resemblance Detection For Deduplication Delta Compression

**Xuming Ye**, Xiaoye Xue, Wenlong Tian,
Zhiyong Xu, Weijun Xiao,
Ruixuan Li, Yaping Wan

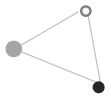
## Outline

- Background And Motivation
- Traditional solutions
- Problem & Solution & Design
- Evaluation
- Conclusion





# 1. Background And Motivation



## Background And Motivation

Nowadays, data deduplication is critical in the storage system. With the increase of the devices, such as IoT devices, mobile device, etc..., the ever-increasing demand of storage space is pressing.

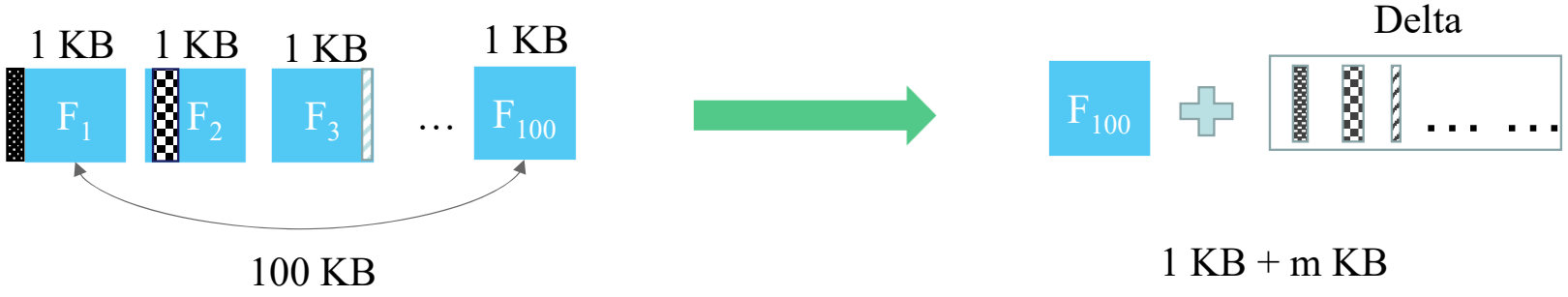

# Data deduplication

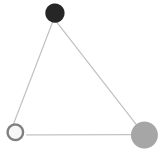
**Redundancy Deduplication**

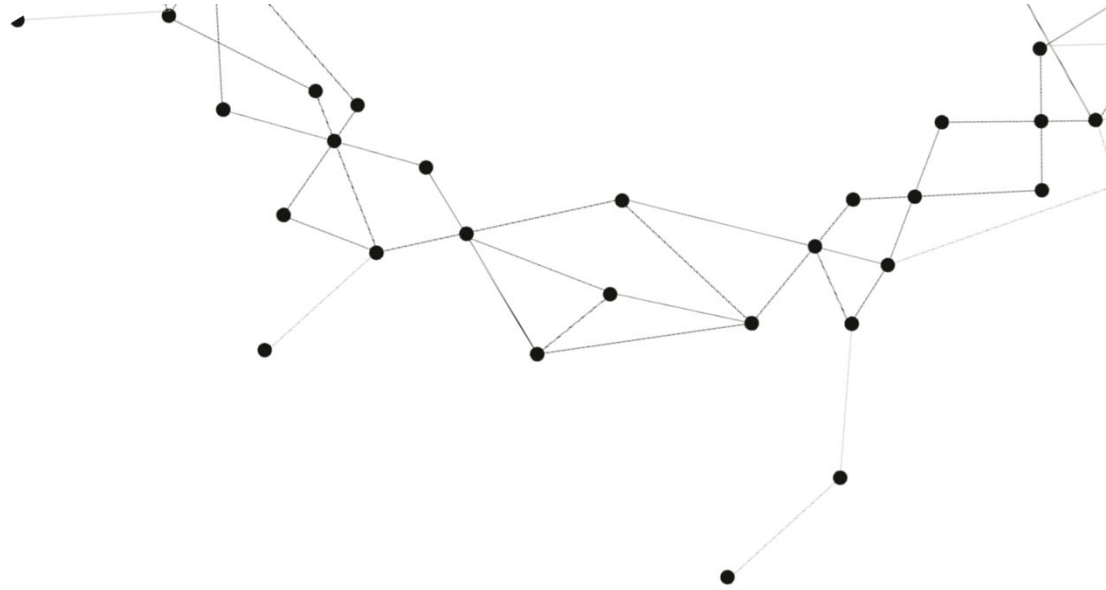
extra copies of the same data are deleted, leaving only one copy to be stored.


**Resemblance Deduplication**

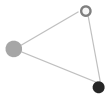
Although redundant data eliminate is efficient, but in the storage system, there are also much similar data.



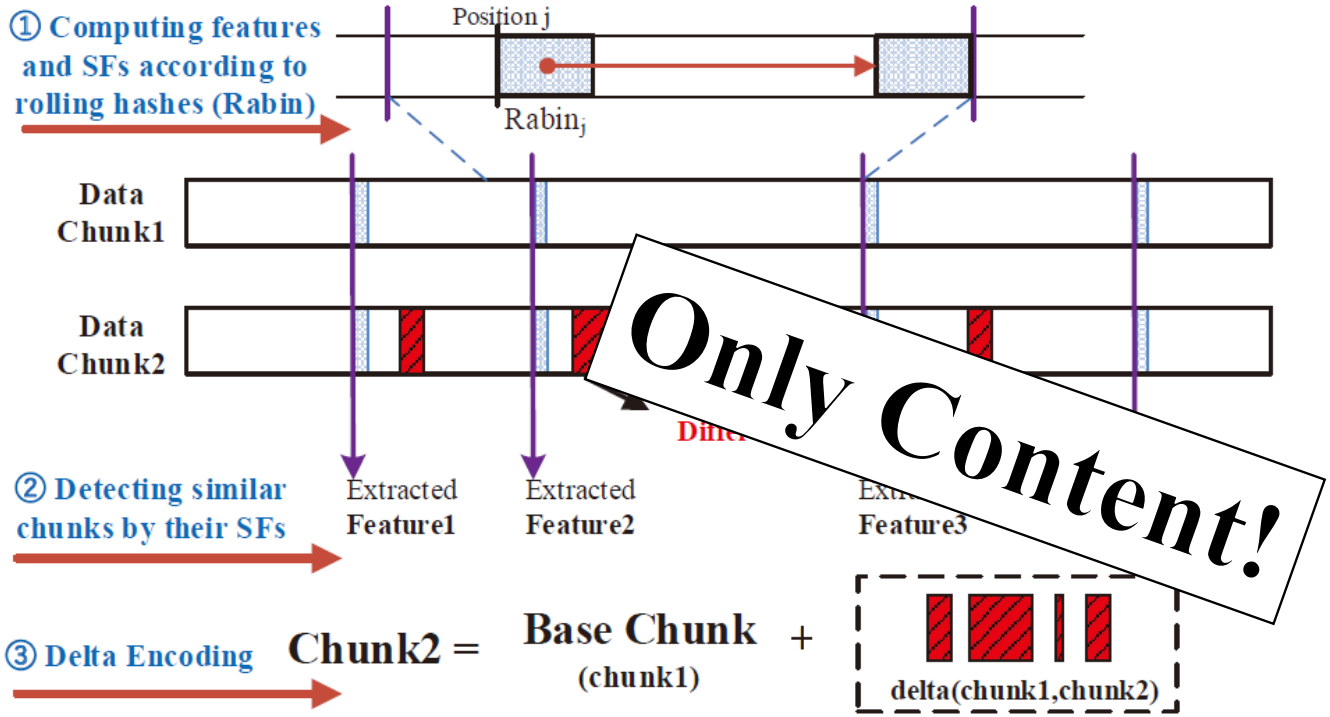




# 2. Traditional solutions



# Traditional Process



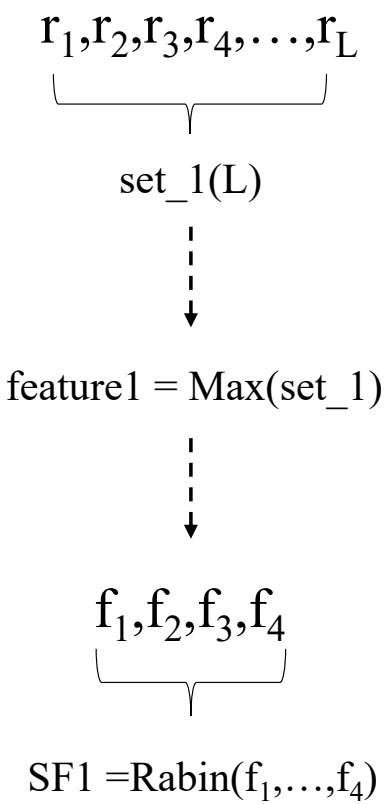
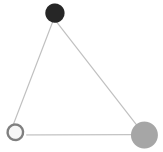
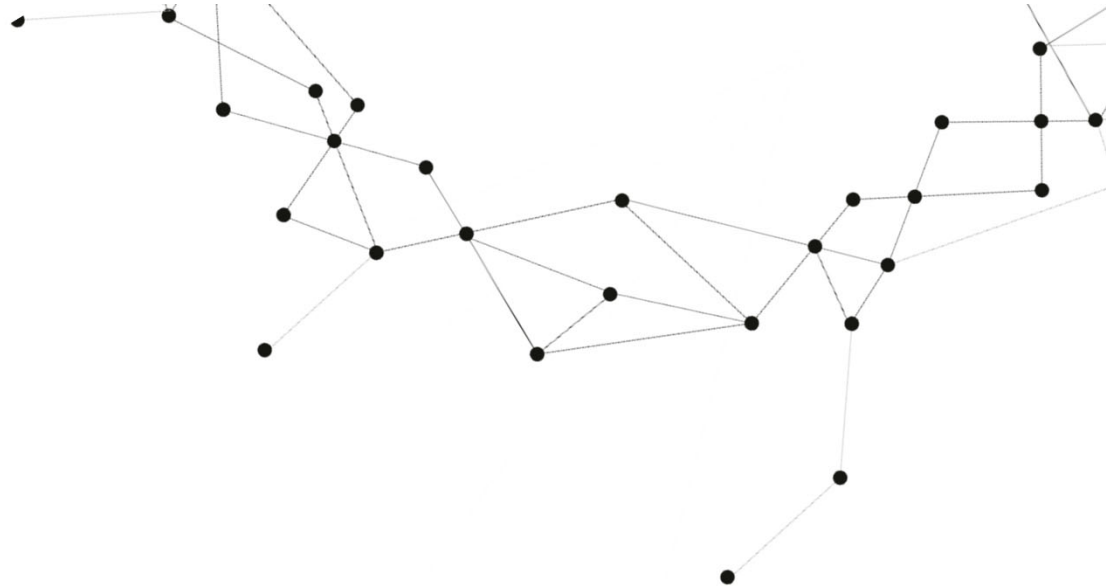
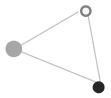




Figure 1: An example of delta compression on two similar chunks with the three typical steps: ① computing similarity, ② indexing, and ③ delta encoding.

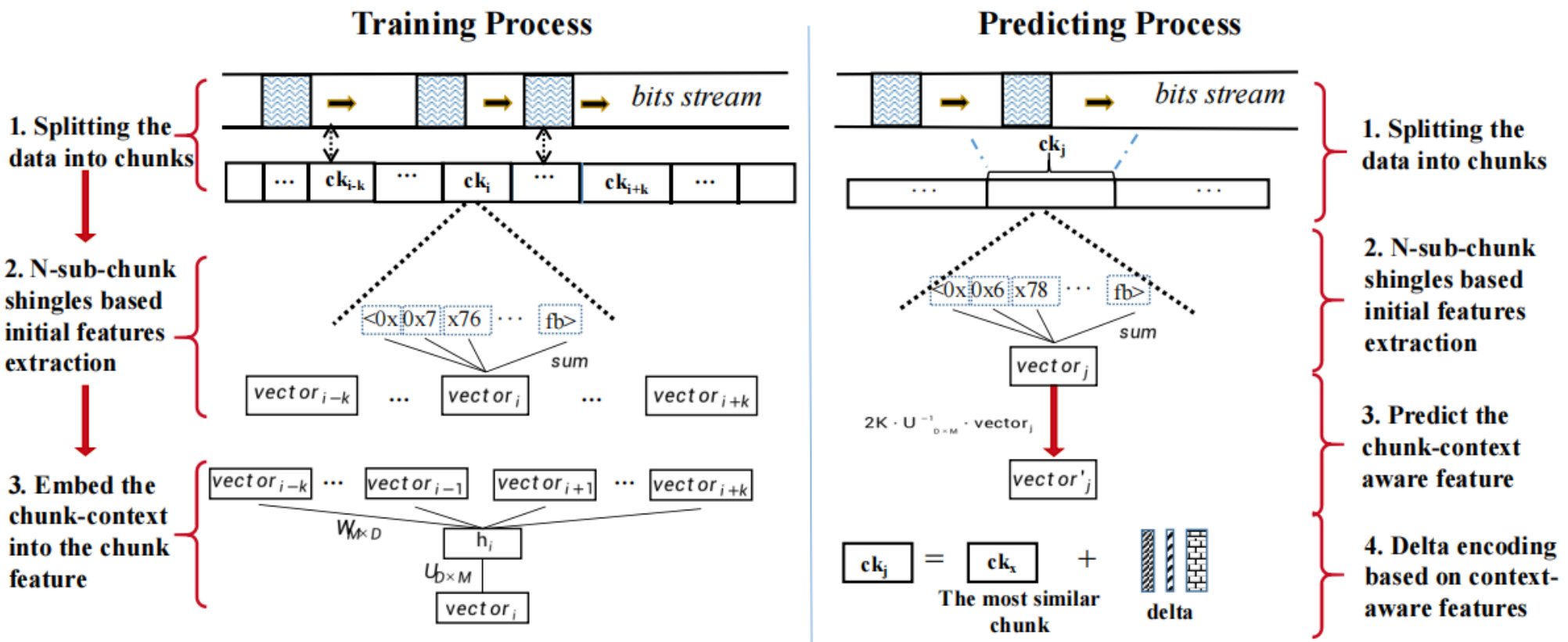




# 3. Problem & Solution & Design



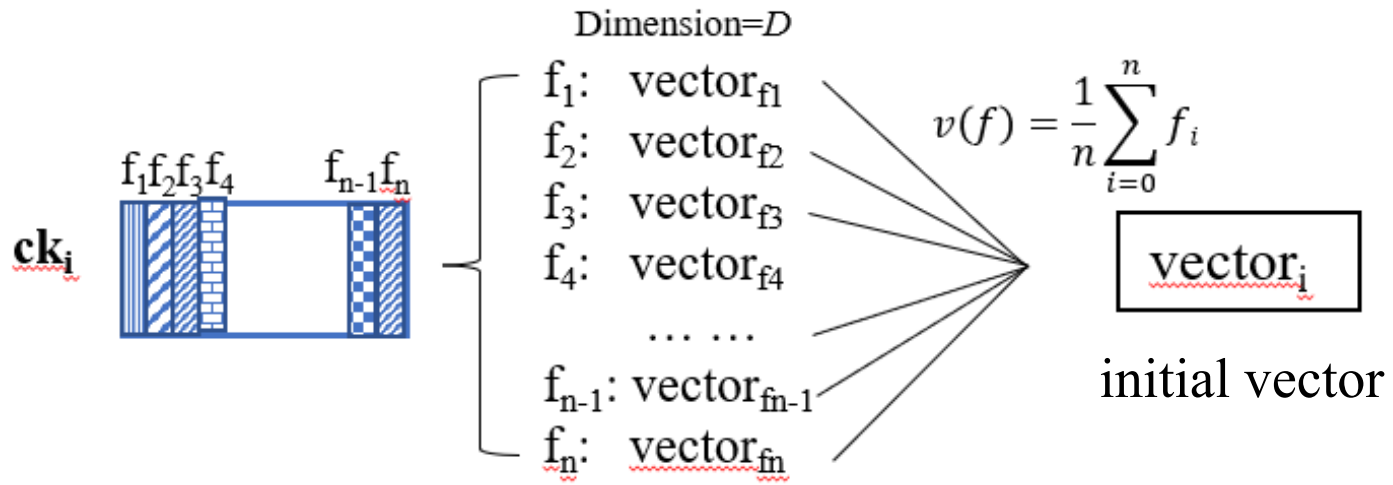
# Workflow


Training Process
Predicting Process
bits stream
1. Splitting the data into chunks
2. N-sub-chunk shingles based initial features extraction
3. Embed the chunk-context into the chunk feature
4. Delta encoding based on context-aware features
sum
The most similar chunk
delta

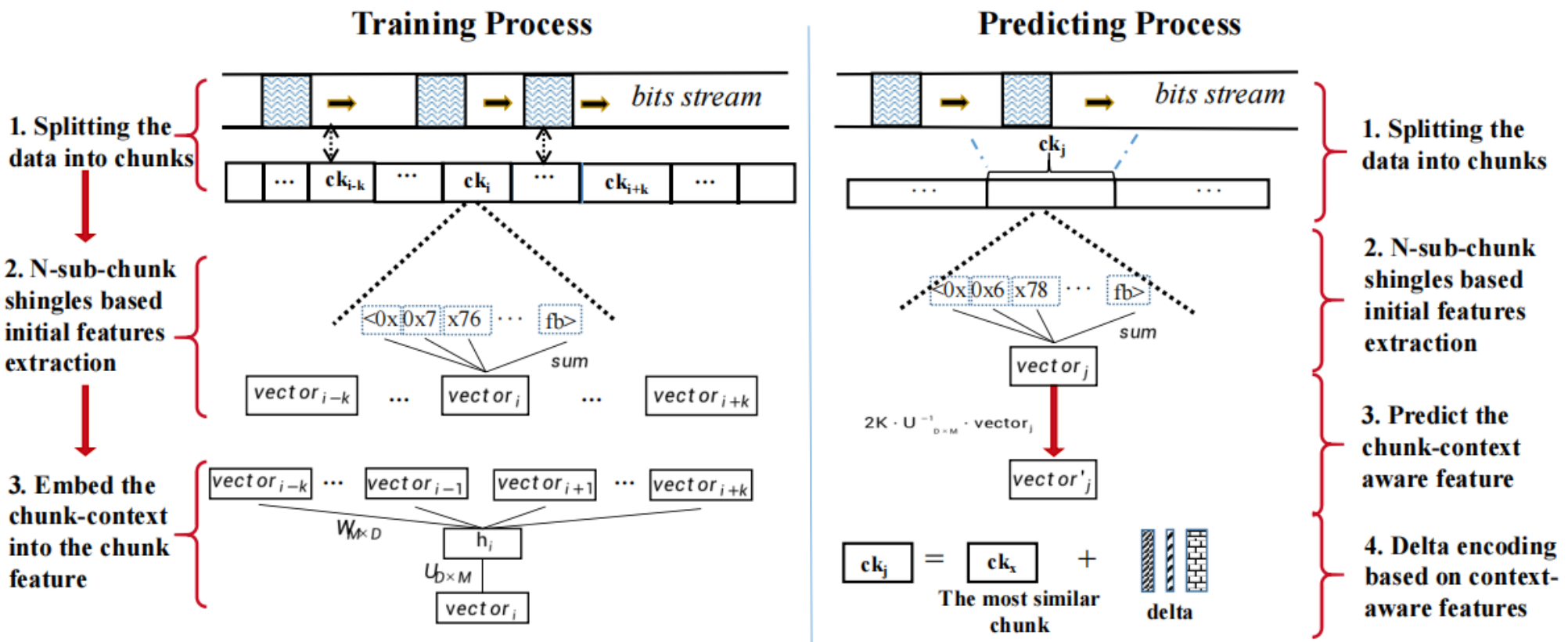
# CARD

- Extract Features

N-sub-chunk scheme



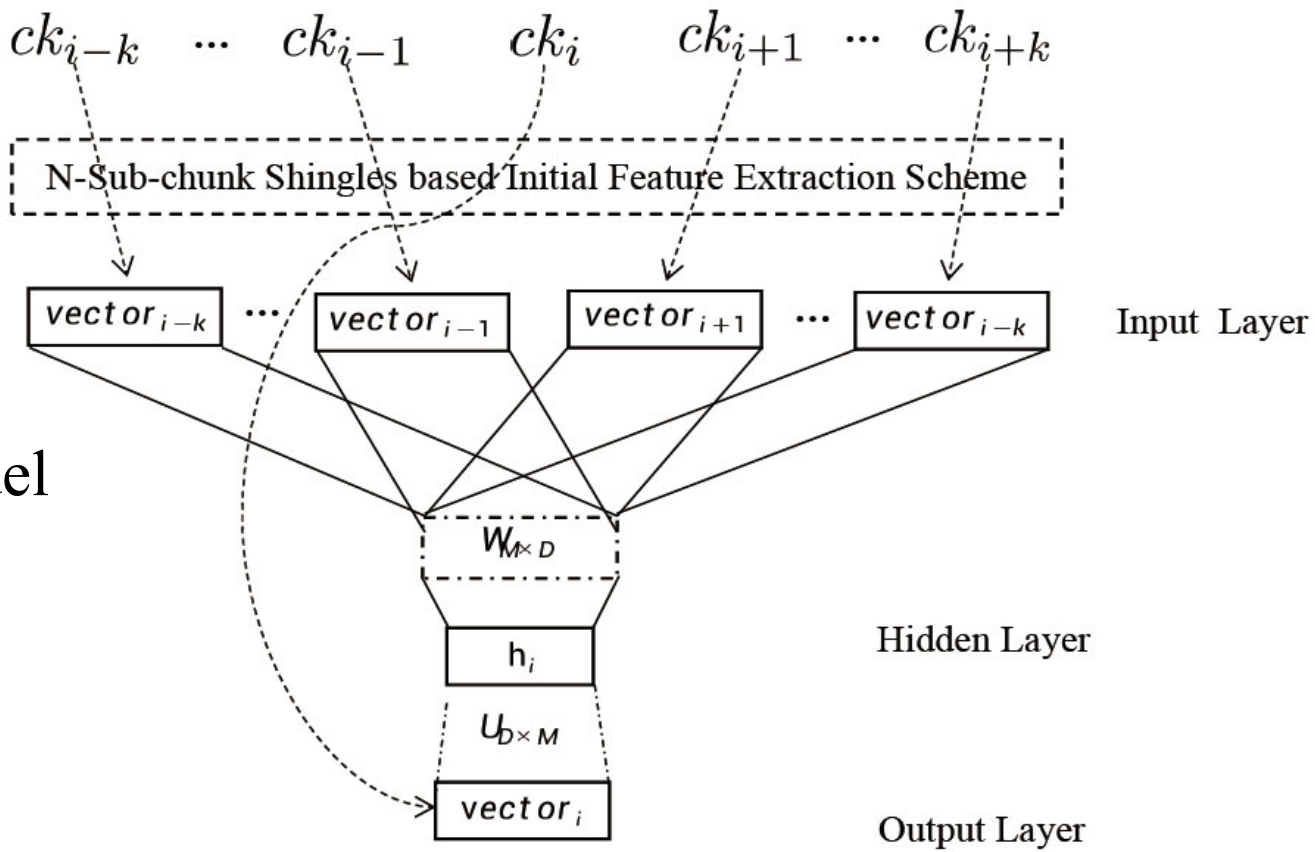
# Workflow


Training Process
Predicting Process
bits stream
1. Splitting the data into chunks
2. N-sub-chunk shingles based initial features extraction
3. Embed the chunk-context into the chunk feature
4. Delta encoding based on context-aware features
sum
The most similar chunk
delta

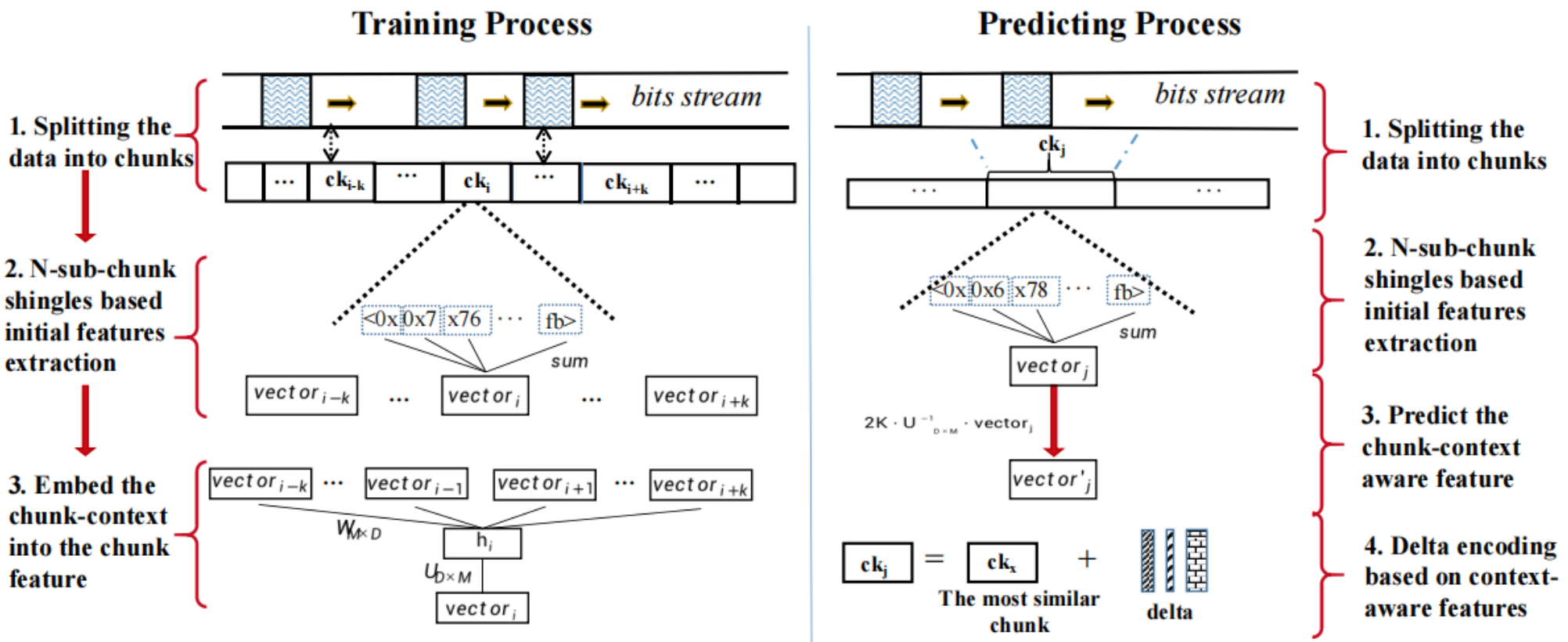
# CARD

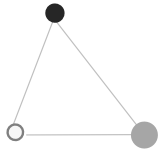
- Extract Features

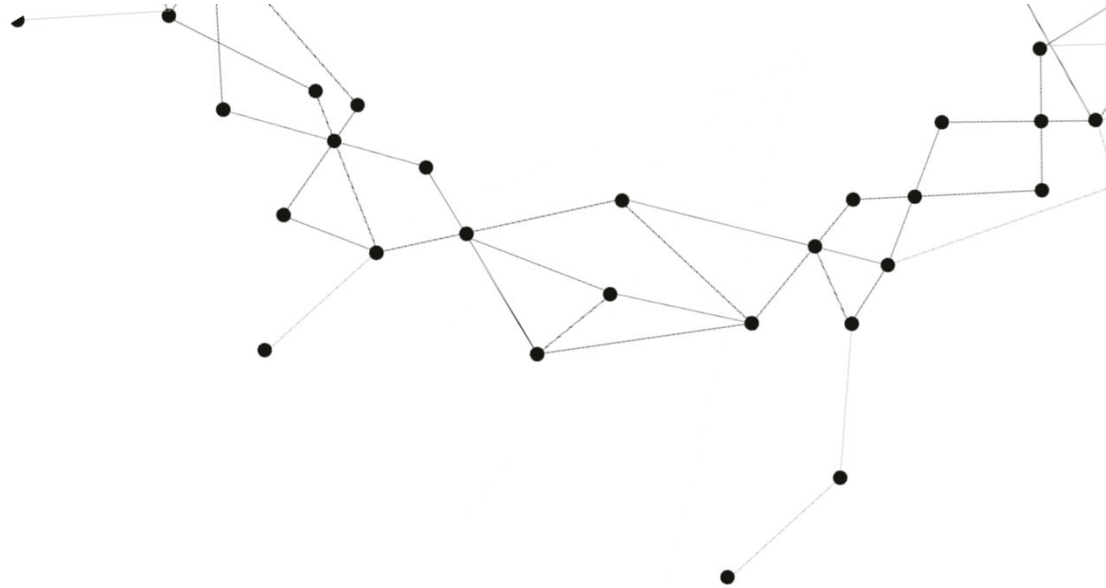
BP-Neural Network-based
Chunk-Context Aware Model



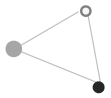
# Workflow


Training Process
Predicting Process
bits stream
1. Splitting the data into chunks
2. N-sub-chunk shingles based initial features extraction
3. Embed the chunk-context into the chunk feature
4. Delta encoding based on context-aware features
sum
The most similar chunk
delta

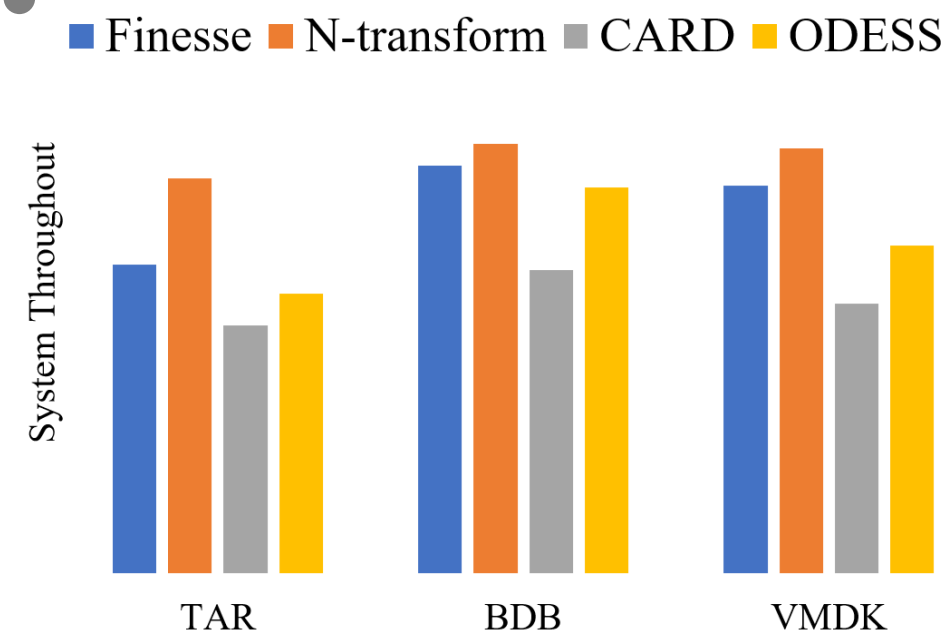


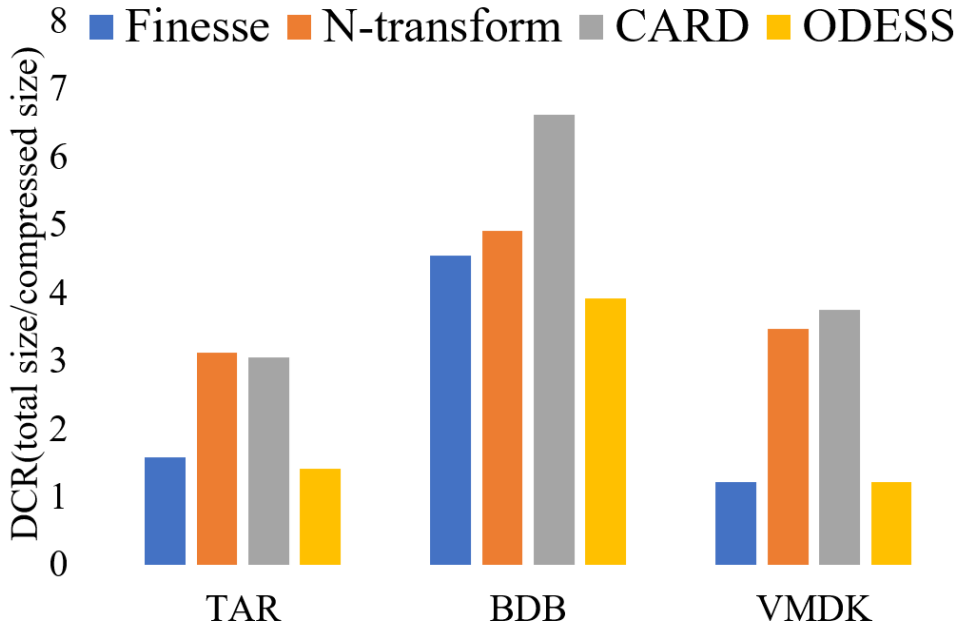


# 4. Evaluation

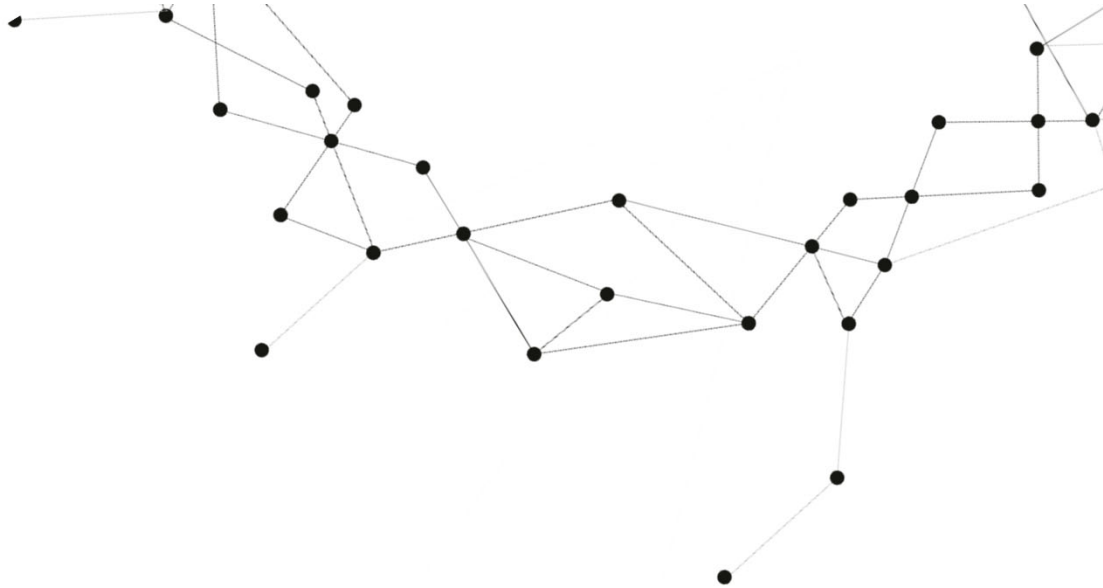


# Evaluation




| Name | Workload descriptions |
|---|---|
| TAR | Versions of Linux kernel source code. Each version is packaged as a tar file. |
| BDB | Backups of the real company database. |
| VMDK | Snapshots of an Ubuntu 18.04 VM image. |



# 5. Conclusion

## Conclusion

According to the presentation above, our advantages as the following.

- Features to vector
- N-sub-chunk shingle scheme
- Chunk-Context Model

# CARD

Thanks for watching

our arxiv: https://arxiv.org/abs/2106.01273