Storage Efficiency Opportunities and Analysis for Video Repositories

HotStorage ‘15

Suganthis Dewakar, Sethuraman Subbiah, Gokul Soundararajan, Mike Wilson, Mark W. Storer, Kishore Udayashankar, Kaladhar Voruganti, Minglong Shao

NetApp Inc.
07/07/2015
Video files have deduplication

Video repository

No deduplication!
Video files have deduplication

45%

Video repository
But where..?
But where..?
But where..?
But where..?
Storage efficiency options

Video storage efficiency

- Compression
  - Lossy
  - Lossless
- Quality reduction
- On-demand transcoding
- Deduplication

System level
- Content-unaware

Application level
- Content-aware
Storage efficiency options

- **Video storage efficiency**
  - Compression
    - Lossy
    - Lossless
  - Quality reduction
  - On-demand transcoding
  - Deduplication
    - System level
      - Content-unaware
      - Content-aware
    - Application level
Outline

1) Introduction
2) Classification of storage efficiency techniques
3) Content-aware deduplication techniques
4) Evaluation
5) Conclusions and future work
Video file content

- Video: 87%
- Audio: <1%
- Header & Text: >12%
Video file layout
ISO Base Media Format

Header

Data
(87% video, 12% audio, <1% text)
Video file layout

ISO Base Media Format

Header

Data
(87% video, 12% audio, <1% text)

Audio Sequence

Video Sequence

Text Seq

Audio Sequence

Video Sequence

Text Seq
Video file layout
ISO Base Media Format

Header

Data
(87% video, 12% audio, <1% text)

Audio Sequence
Video Sequence
Text Seq
Audio Sequence
Video Sequence
Text Seq

Sample
Sample
Sample
Sample
Sample

© 2015 NetApp, Inc. All rights reserved.
Sample-based deduplication

Header

Data
(87% video, 12% audio, <1% text)

Audio Sequence | Video Sequence | Text Seq | Audio Sequence | Video Sequence | Text Seq

Sample | Sample | Sample | Sample | Sample | Sample
Sample-based deduplication technique

<table>
<thead>
<tr>
<th>Fingerprint</th>
<th>Link to data</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>40024</td>
<td>894</td>
</tr>
<tr>
<td>V2</td>
<td>54204</td>
<td>984</td>
</tr>
<tr>
<td>V3</td>
<td>62683</td>
<td>683</td>
</tr>
</tbody>
</table>
Sample-based deduplication technique

Video file

<table>
<thead>
<tr>
<th>Fingerprint</th>
<th>Link to data</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>40024</td>
<td>894</td>
</tr>
<tr>
<td>V2</td>
<td>54204</td>
<td>984</td>
</tr>
<tr>
<td>V3</td>
<td>62683</td>
<td>683</td>
</tr>
</tbody>
</table>
Sample-based deduplication technique

<table>
<thead>
<tr>
<th>Finger print</th>
<th>Link to data</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>40024</td>
<td>894</td>
</tr>
<tr>
<td>V2</td>
<td>54204</td>
<td>984</td>
</tr>
<tr>
<td>V3</td>
<td>62683</td>
<td>683</td>
</tr>
</tbody>
</table>

Video file

Signature table

Video File Extract samples
Sample-based deduplication technique

Video file

Signature table

<table>
<thead>
<tr>
<th>Fingerprint</th>
<th>Link to data</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>40024</td>
<td>894</td>
</tr>
<tr>
<td>V2</td>
<td>54204</td>
<td>984</td>
</tr>
<tr>
<td>V3</td>
<td>62683</td>
<td>683</td>
</tr>
</tbody>
</table>

© 2015 NetApp, Inc. All rights reserved.
Sample-based deduplication technique

Signature table

<table>
<thead>
<tr>
<th>Fingerprint</th>
<th>Link to data</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>40024</td>
<td>894</td>
</tr>
<tr>
<td>V2</td>
<td>54204</td>
<td>984</td>
</tr>
<tr>
<td>V3</td>
<td>62683</td>
<td>683</td>
</tr>
<tr>
<td>V4</td>
<td>74000</td>
<td>937</td>
</tr>
<tr>
<td>V5</td>
<td>80137</td>
<td>873</td>
</tr>
</tbody>
</table>

Video File → Extract samples → Compute signature → Deduplicate
Sequence-based deduplication

Header

Data (87% video, 12% audio, <1% text)

Audio Sequence

Video Sequence

Text Seq

Audio Sequence

Video Sequence

Text Seq

Sample

Sample

Sample

Sample

Sample
Sequence-based deduplication technique

<table>
<thead>
<tr>
<th>Signature</th>
<th>Link to data</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>30995</td>
<td>312</td>
</tr>
<tr>
<td>V2</td>
<td>12685</td>
<td>646</td>
</tr>
<tr>
<td>V3</td>
<td>19865</td>
<td>857</td>
</tr>
</tbody>
</table>

Audio signature table

<table>
<thead>
<tr>
<th>Signature</th>
<th>Link to data</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>40024</td>
<td>894</td>
</tr>
<tr>
<td>A2</td>
<td>62683</td>
<td>683</td>
</tr>
</tbody>
</table>

Video signature table
Sequence-based deduplication technique

Video file

<table>
<thead>
<tr>
<th>Signature</th>
<th>Link to data</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>40024</td>
<td>894</td>
</tr>
<tr>
<td>A2</td>
<td>62683</td>
<td>683</td>
</tr>
</tbody>
</table>

Audio signature table

Video signature table

<table>
<thead>
<tr>
<th>Signature</th>
<th>Link to data</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>30995</td>
<td>312</td>
</tr>
<tr>
<td>V2</td>
<td>12685</td>
<td>646</td>
</tr>
<tr>
<td>V3</td>
<td>19865</td>
<td>857</td>
</tr>
</tbody>
</table>
Sequence-based deduplication technique

Audio signature table

<table>
<thead>
<tr>
<th>Signature</th>
<th>Link to data</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>40024</td>
<td>894</td>
</tr>
<tr>
<td>A2</td>
<td>62683</td>
<td>683</td>
</tr>
</tbody>
</table>

Video signature table

<table>
<thead>
<tr>
<th>Signature</th>
<th>Link to data</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>30995</td>
<td>312</td>
</tr>
<tr>
<td>V2</td>
<td>12685</td>
<td>646</td>
</tr>
<tr>
<td>V3</td>
<td>19865</td>
<td>857</td>
</tr>
</tbody>
</table>
Sequence-based deduplication technique

Video file

Audio signature table

<table>
<thead>
<tr>
<th>Signature</th>
<th>Link to data</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>40024</td>
<td>894</td>
</tr>
<tr>
<td>A2</td>
<td>62683</td>
<td>683</td>
</tr>
</tbody>
</table>

Video signature table

<table>
<thead>
<tr>
<th>Signature</th>
<th>Link to data</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>30995</td>
<td>312</td>
</tr>
<tr>
<td>V2</td>
<td>12685</td>
<td>646</td>
</tr>
<tr>
<td>V3</td>
<td>19865</td>
<td>857</td>
</tr>
</tbody>
</table>
Sequence-based deduplication technique

Video file

Audio signature table

<table>
<thead>
<tr>
<th>Signature</th>
<th>Link to data</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>40024</td>
<td>894</td>
</tr>
<tr>
<td>A2</td>
<td>62683</td>
<td>683</td>
</tr>
<tr>
<td>A3</td>
<td>74000</td>
<td>937</td>
</tr>
<tr>
<td>A4</td>
<td>80137</td>
<td>873</td>
</tr>
</tbody>
</table>

Video signature table

<table>
<thead>
<tr>
<th>Signature</th>
<th>Link to data</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>30995</td>
<td>312</td>
</tr>
<tr>
<td>V2</td>
<td>12685</td>
<td>646</td>
</tr>
<tr>
<td>V3</td>
<td>19865</td>
<td>857</td>
</tr>
<tr>
<td>V4</td>
<td>50687</td>
<td>757</td>
</tr>
</tbody>
</table>
Outline

1) Introduction
2) Classification of storage efficiency techniques
3) Content-aware deduplication techniques
4) Evaluation
   - Experiments
   - Input dataset
   - Metrics
   - Results
5) Conclusions and future work
Existing deduplication techniques

- Content-unaware deduplication techniques
  - Fixed-sized deduplication
  - Variable-sized deduplication

- Fixed-sized
  - Block length – 4096 Bytes

- Variable-sized
  - Rabin fingerprint
  - Average block size: 4096 Bytes
  - Polynomial degree: 48
Experiments

Video 1

A1  V1  A2  V2  H
Experiments

Video 1

A1 | V1 | A2 | V2 | H

vs.

Video 2

A1' | V1 | A2' | V2 | H
Experiments

Video 1

A1 V1 A2 V2 H

VS.

A1’ V1 A2’ V2 H

Video 2
Experiments

After deduplication:
Experiments

After deduplication:

V1  V2
Experiments

After deduplication:
Experiments

Content variations

Base

A  V  A  V  H

VS.

Dubbed

A’  V  A’  V  H

© 2015 NetApp, Inc. All rights reserved.
Experiments

Content variations

Base

\[ \text{A} \quad \text{V} \quad \text{A} \quad \text{V} \quad \text{H} \]

vs.

Dubbed

\[ \text{A}' \quad \text{V} \quad \text{A}' \quad \text{V} \quad \text{H} \]

Closed caption

\[ \text{A} \quad \text{V} \quad \text{T} \quad \text{A} \quad \text{V} \quad \text{T} \quad \text{H} \]
Experiments

Content variations

- **Base**: AVAVH
- **Dubbed**: A'V'A'VH
- **Closed caption**: AVTH
- **Resolution**: AV'AVA'VH

vs.
Experiments

Content variations

Base

A

V

A

V

H

vs.

Dubbed

A'

V

A'

V

H

Closed caption

A

V

T

A

V

T

H

Resolution

A

V'

A

V'

H

Web optimized

H

A

V

A

V

© 2015 NetApp, Inc. All rights reserved.
Experiments

Video Segments

Base

vs.

Source: Madagascar
Experiments

Video Segments

Base vs. Different start time

Source: Madagascar
Experiments

Video Segments

Base

Different start time

Different end time

vs.

Source: Madagascar
Experiments

**Video Segments**

- **Base**
- **Different start time**
- **Different end time**
- **Subset of base**

Source: Madagascar
Input dataset

Video DVDs

- Movies and TV series
- Base, Dubbed, Closed Caption, Resolution
- Metadata in the beginning (Web Optimized)
- Different start time, different end time
- Subset of a video
Evaluation metrics

- **Storage saved**
  
  Deduplication ratio $\% = 100 - \left( \frac{\text{storage footprint after dedupe}}{\sum \text{size of files}} \right) \times 100$

  E.g. If a file is stored twice, storage saved is 50%

- **Deduplication overhead**

  Number of signatures maintained in the fingerprint table
Results

Content variation

Deduplication ratio

Deduplication overhead

© 2015 NetApp, Inc. All rights reserved.
Results

Content variation

![Graph showing deduplication ratio and overhead for different content types and resolution levels.]

- **Deduplication ratio**
  - Dubbed
  - CC
  - Resolution
  - Web-optimized

- **Deduplication overhead**
  - Fixed-sized (F)
  - Variable-sized (V)
  - Sample-based audio (S)
  - Sample-based video (S)
  - Sequence-based audio (Q)
  - Sequence-based video (Q)

© 2015 NetApp, Inc. All rights reserved.
Results

Content variation

Deduplication ratio

<table>
<thead>
<tr>
<th>Dubbed</th>
<th>CC</th>
<th>Resolution</th>
<th>Web-optimized</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>V</td>
<td>S</td>
<td>Q</td>
</tr>
<tr>
<td>F</td>
<td>V</td>
<td>S</td>
<td>Q</td>
</tr>
<tr>
<td>F</td>
<td>V</td>
<td>S</td>
<td>Q</td>
</tr>
<tr>
<td>F</td>
<td>V</td>
<td>S</td>
<td>Q</td>
</tr>
</tbody>
</table>

Deduplication overhead

<table>
<thead>
<tr>
<th>Dubbed</th>
<th>CC</th>
<th>Resolution</th>
<th>Web-optimized</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>V</td>
<td>S</td>
<td>Q</td>
</tr>
<tr>
<td>F</td>
<td>V</td>
<td>S</td>
<td>Q</td>
</tr>
<tr>
<td>F</td>
<td>V</td>
<td>S</td>
<td>Q</td>
</tr>
<tr>
<td>F</td>
<td>V</td>
<td>S</td>
<td>Q</td>
</tr>
</tbody>
</table>
Results

Content variation

Deduplication ratio

Deduplication overhead

© 2015 NetApp, Inc. All rights reserved.
Results

Content variation

Deduplication ratio

Deduplication overhead
Results

Content variation

Deduplication ratio

<table>
<thead>
<tr>
<th>Dubbed</th>
<th>CC</th>
<th>Resolution</th>
<th>Web-optimized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed-sized (F)</td>
<td>Variable-sized (V)</td>
<td>Sample-based audio (S)</td>
<td>Sample-based video (S)</td>
</tr>
</tbody>
</table>

Deduplication overhead

<table>
<thead>
<tr>
<th>Dubbed</th>
<th>CC</th>
<th>Resolution</th>
<th>Web-optimized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of signatures (in millions)</td>
<td>Fixed-sized (F)</td>
<td>Variable-sized (V)</td>
<td>Sample-based audio (S)</td>
</tr>
</tbody>
</table>
Results

Content variation

Deduplication ratio

<table>
<thead>
<tr>
<th>Content Type</th>
<th>Fixed-sized (F)</th>
<th>Variable-sized (V)</th>
<th>Sample-based audio (S)</th>
<th>Sample-based video (S)</th>
<th>Sequence-based audio (Q)</th>
<th>Sequence-based video (Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dubbed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web-optimized</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Deduplication overhead

<table>
<thead>
<tr>
<th>Content Type</th>
<th>Fixed-sized (F)</th>
<th>Variable-sized (V)</th>
<th>Sample-based audio (S)</th>
<th>Sample-based video (S)</th>
<th>Sequence-based audio (Q)</th>
<th>Sequence-based video (Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dubbed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web-optimized</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

Content variation

Deduplication ratio

Deduplication overhead

- Fixed-sized (F)
- Variable-sized (V)
- Sample-based audio (S)
- Sample-based video (S)
- Sequence-based audio (Q)
- Sequence-based video (Q)
Results

Content variation

Deduplication ratio

Deduplication overhead

- Fixed-sized (F)
- Variable-sized (V)
- Sample-based audio (S)
- Sample-based video (S)
- Sequence-based audio (Q)
- Sequence-based video (Q)
Results

Content variation

Deduplication ratio

Deduplication overhead

© 2015 NetApp, Inc. All rights reserved.
Results

Content variation

**Deduplication ratio**

<table>
<thead>
<tr>
<th></th>
<th>Dubbed</th>
<th>CC</th>
<th>Resolution</th>
<th>Web-optimized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed-sized (F)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variable-sized (V)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample-based audio (S)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample-based video (S)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence-based audio (Q)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence-based video (Q)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Deduplication overhead**

<table>
<thead>
<tr>
<th></th>
<th>Dubbed</th>
<th>CC</th>
<th>Resolution</th>
<th>Web-optimized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed-sized (F)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variable-sized (V)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample-based audio (S)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample-based video (S)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence-based audio (Q)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sequence-based video (Q)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Results

Content variation

Deduplication ratio

Deduplication overhead

© 2015 NetApp, Inc. All rights reserved.
Results

Video segments

Deduplication ratio

- Different start times
- Different end times
- Subset of another video

Deduplication overhead

- Different start times
- Different end times
- Subset of another video

Fixed-sized (F)
Variable-sized (V)
Sample-based audio (S)
Sample-based video (S)
Sequence-based audio (Q)
Sequence-based video (Q)
What did we learn?

- Different techniques work well for different environments
- Content-aware techniques work well for controlled environments
- Variable-sized works well for repositories with multiple unknown formats
- Sequence-based has 81% lower signature table size
- Sample-based provides maximum deduplication to other techniques
- Exclude audio when using sample-based technique
- Variable-sized fails when sample ordering within sequences are different
Conclusion

- Deduplication works for video files
- Different ways storage efficiency is achieved in video repositories
- Compared the different system-level deduplication techniques
- Obtained inferences from the results
- In future, study different video file formats like FLV, AVI, etc.
Thank you