PYRAMID CODES: FLEXIBLE SCHEMES TO TRADE SPACE FOR ACCESS EFFICIENCY IN RELIABLE DATA STORAGE SYSTEMS

Cheng Huang, Minghua Chen, and Jin Li
Microsoft Research, Redmond, US
IEEE NCA, Boston, July 2007
networked storage on the rise …

- rapidly growing demands on storage systems
  - consumers, enterprises …
  - web services …
- using commodity components to build large scale storage systems is becoming a common practice
  - reliability is a must (five 9’s)
  - failure is norm and dealt with by redundancy
replication vs. erasure codes
- the fundamental trade-off

Pyramid Codes and recoverability theorem
- not YAC (yet another code)
- basic Pyramid Codes
- generalized Pyramid Codes
3-replication

- storage overhead: 3x
  - 12 data nodes + 24 replica nodes
- access/recovery cost (one data failure): 1x
(16, 12) erasure code

- storage overhead: 1.33x
  - 12 data nodes + 4 redundant nodes
- access/recovery cost (one data failure): 12x
replication vs. erasure codes (3)

<table>
<thead>
<tr>
<th></th>
<th>replication scheme</th>
<th>erasure codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>storage overhead</td>
<td>high (3x)</td>
<td>low (1.33x)</td>
</tr>
<tr>
<td>access/recovery cost</td>
<td>low (1x)</td>
<td>high (12x)</td>
</tr>
</tbody>
</table>

- in the end, storage is not that cheap
  - more storage → more machine, more space, more maintenance personal, etc. → 55% of data centers’ operating costs (Windows Live service data)

- network traffic is not free either
  - network in data centers can become bottleneck (Lian et al. ICDCS’05)

- same concerns for P2P storage …
the fundamental trade-offs in replication vs. erasure codes

- Erasure codes
- Pyramid Codes
- Replication
I. Basic Pyramid Codes (1)

- **Data Nodes**: d1, d2, …, d6, …, d7, …, d12
- **Redundant Nodes**: C1,1, C1,2, C1,3, C1,4

- C1,1: \{d1, d2, …, d6\}
- C1,2: \{d7, d8, …, d12\}

Diagram:
- **Data Nodes**: d1, d2, …, d6, …, d7, …, d12
- **Redundant Nodes**: C1,1, C1,2, C1,3, C1,4

Diagram has circular references.
1. basic Pyramid Codes (2)

- storage overhead: 1.42x
- access/recovery cost (one data failure): 6x
- recovery any 4 failures
I. basic Pyramid Codes (3)

- recover $d_5$ and $d_6$
- combine $C_{1,1}$ and $C_{1,2} \rightarrow C_1$; $C_{2,1}$ and $C_{2,2} \rightarrow C_2$
- recover $d_7$, $d_8$, $d_{11}$ and $d_{12}$
decoding is analogous to climbing up a Pyramid!
another erasure pattern

- is this erasure pattern recoverable at all?
  - no small group recovery
  - $C_{2,1}$ and $C_{2,2} \rightarrow C_2$, so only 3 redundant nodes at the global level

- counting failures/parities: 5 failures and 5 parities

- now what?
recoverability theorem (1)

- an erasure pattern is recoverable only if the corresponding Tanner graph contains a full-size matching.

- Tanner graph

failed data  survival redundancy
recoverability theorem (2)

an unrecoverable example

Tanner graph
no full-size matching!
the recoverability theorem is a necessary condition for all erasure codes

it is not sufficient for all known storage codes

including basic Pyramid Codes

generalized Pyramid Codes makes the condition sufficient

able to recover any erasure pattern ever possible to recover – optimal recoverably property
II. generalized Pyramid Codes (1)

- A generalized Pyramid Code can be constructed given any configuration (data/parity association).
  - Details in paper ...
- Any generalized Pyramid Code satisfies optimal recoverable property.
II. generalized Pyramid Codes (2)

- why is this a big deal?
  - **MDS codes** are optimal when redundant nodes and data nodes are fully associated
  - **Pyramid Codes** are optimal when redundant nodes and data nodes are partially associated

- contributions recap
  - a necessary condition theorem for recoverability
  - a construction algorithm for generalized Pyramid Codes, which achieve optimal recoverability
optimal decoding of generalized Pyramid Codes

- how to access/recover with minimum cost?
  - all failed nodes
  - or simply one failed node (say $d_{12}$)
- optimal decoding path
  - details in paper ...
summary

- the fundamental trade-off between storage overhead and access/recovery efficiency
- two classes of Pyramid Codes
- recoverability theorem
  - generalized Pyramid Codes are optimal