The Fountain of Knowledge

Using Digital Fountains to provide reliable storage in the Data Centre

Toby Moncaster, Andrew W. Moore, Jon Crowcroft

Data Centre Storage Requirements

Speed: Needs to keep up with application speed. Often read speed is more important than write. This leads to "move the application to the data" approaches.

Reliability: Data should not be corrupted, also need to ensure synchronisation between copies of the same data.

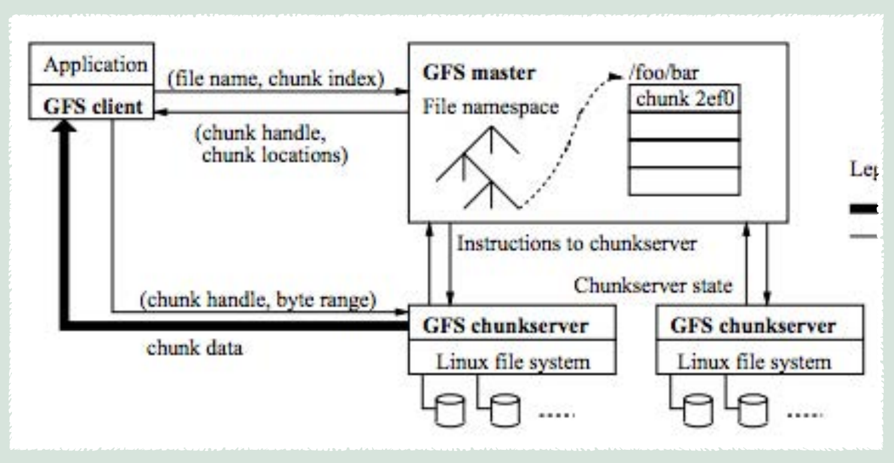
Resilience: Storage should be resilient against local and regional failures. Multiple copies of data may be required. Failures may come singly or in sequence.

Existing Approaches

Storage Area Networks (e.g. NetApp)

Provides dedicated specialist network to provide fast read/write access to centralised disk arrays. Storage appears as local disk.

Central metadata (e.g. GFS)



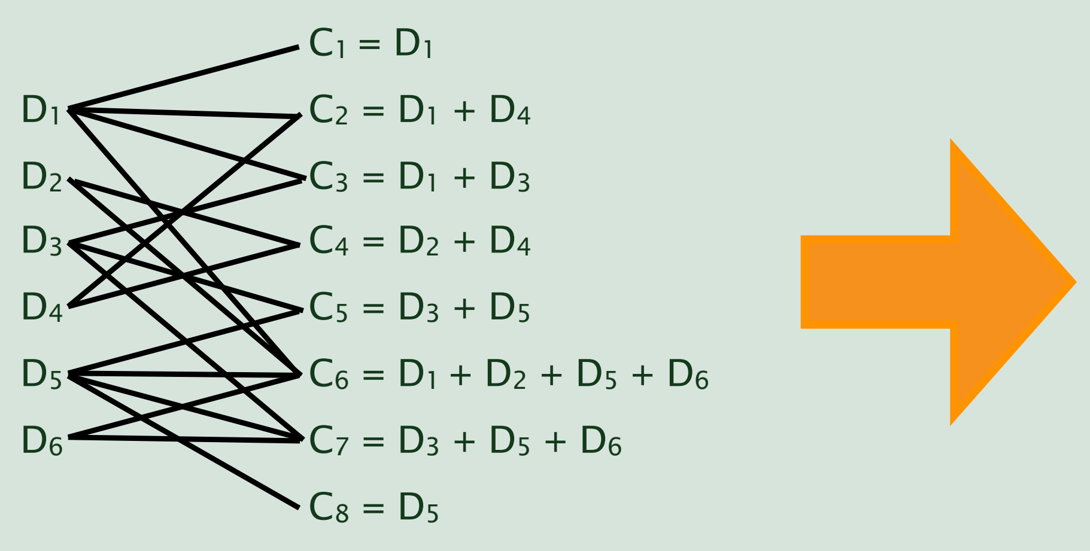
Storage divided into chunks. A single central metadata server controls how data are distributed and replicated between these. Ideal for oversubscribed networks.

Distributed (e.g. Flat Datacenter Storage)

Storage split into blobs. Data split into tracts. Each blob has a tract server which allocates space to each of the tracts. Central metadata server lists where the tract servers are. Uses hash function to split data among tracts. Only works in full bisection bandwidth networks.

Digital Fountains

Data is encoded using sparse erasure codes such as Tornado, Luby Transform or Raptor codes. Receiver needs to obtain slightly more encoded data than the original raw source. Can receive code chunks from multiple sources.



Receive C_1, C_2, C_7, C_3, C_4, C_5
Use C_1 to recover D_1
Use C_2 and D_1 to recover D_2
Then wait till you receive C_3
Use C_3 and D_1 to recover D_3
Use C_4 and D_2 to recover D_4
Use C_5 and D_3 to recover D_5
Use C_7 and D_3 to recover $(D_5 + D_6)$
Use $(D_5 + D_6)$ and D_5 to recover D_6

Data is divided into segments. Code chunks are created by XORing sets of data segments. To decode, repeatedly XOR the coded blocks until original data is recovered.

Data Centre Knowledge Fountains

Data is split into chunks and encoded into a digital fountain. Encoded chunks are stored at locations across data centre. Resilience is achieved by creating multiple encodings of each tract. These are stored at separate physical locations.

To access the data simply download sufficient encoded chunks to decode the original data. Overhead is 5-10% depending on the encoding used.

Choosing Where to Store the Data

Use a hash-based indexing approach similar to Flat Datacenter Storage. Each storage location (or blob) has a tractserver. These work just as in FDS. Central metadata server stores simple index.

Index key is generated using simple hash of file descriptor and blob ID. Index returns list of tract servers where code blocks for that file exist.

Detailed Operation

Writing Data: Split the data into blocks. Decide on the degree of redundancy. Encode multiple versions of the data using a suitable encoding scheme. Ask the metadata server to choose which tract servers to store the code chunks.

Accessing data: Use the metadata server to identify all the tract servers with code chunks for the file. Select tract servers based on suitable heuristics (location, speed, etc). Download sufficient chunks to recover data.

Recovering from failures: If a storage blob fails the metadata server will choose alternative tract servers for the data on that blob. Because of redundant encoding many blobs need to fail before data is unrecoverable.

Is it Suitable for Data Centres?

Speed: System is fast since code chunks are downloaded in parallel. Digital Fountain encoding is efficient and fast.

Reliability: Data is encoded across many blocks. If a block is corrupted or lost there is plenty of redundancy to recover from this.

Resilience: The system has a high degree of resilience because the encoding only requires sufficient code blocks to be downloaded which can come from any storage blob.

UNIVERSITY OF CAMBRIDGE

Systems Research Group
srg@cambridge:~$

Computer Laboratory, 15 JJ Thomson Avenue,
Cambridge, CB3 0FD, UK

EPSRC
Pioneering research and skills

INTERNET
INTelligent Energy awaRe NETworks