AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE

Managing data availability and integrity in federated cloud storage

Zarządzanie dostępnością i integralnością danych w sfederowanych zasobach chmury obliczeniowej.

Krzysztof Styrc

Promotor: dr inż. Marian Bubak (AGH)
Konsultant: Piotr Nowakowski (ACC Cyfronet)

Agenda

1. Introduction
   - Motivation
   - VPH-Share project background
   - Objectives of the thesis
2. State of the art
   - Standard methods for data integrity
   - Approaches to data integrity in the cloud
   - Drawbacks of the existing methods
3. Design and implementation
   - DRI service design
   - Validation algorithm
   - DRI service verification
4. Summary and future work

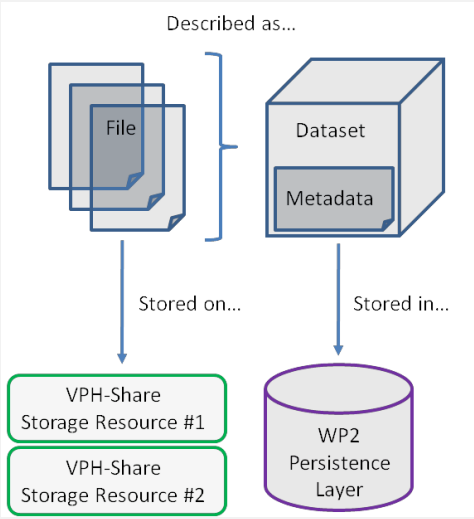
Motivation

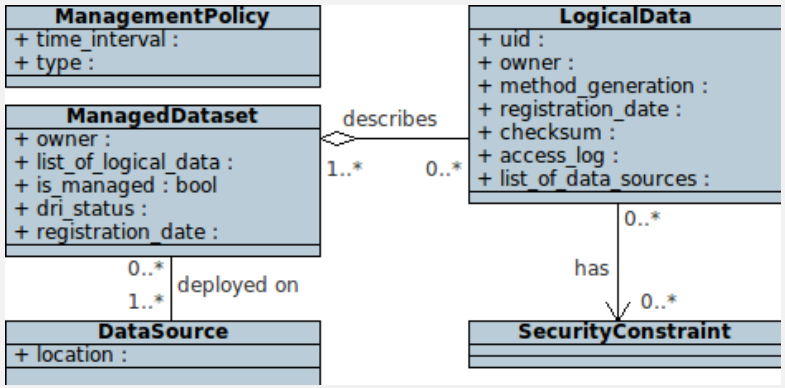
1. Technology shifts toward cloud computing paradigm
   - good quality–cost ratio and pay as you use
   - scalability, availability and SLAs
   - no IT infrastructure management
2. Cloud storage problems
   - data stored on external resources of cloud provider
   - SLA defined as best-effort, return of costs otherwise
   - cloud storage vendor lock-in
   - recent cloud storage failures and security breaches:
     - ★ deleted mails, blocked accounts in Gmail
     - ★ Amazon S3 downtimes
     - ★ unauthorized accessto files in GoogleDocs
3. Cloud storage data integrity challenges:
   - network latency and bandwidth limits
   - costs assosiated with data retrieval
   - only simple operations available, no possibility to execute code on stored data

Still it is required to ensure that the data is available and not corrupted

VPH-Share project background

- VPH-Share data overview
  - biomedical data stored in federation of cloud providers to avoid vendor lock-in
  - storage entity: dataset (set of files)





- Data integrity requirements
  - periodical and on-request monitoring of data availability and integrity
  - network efficient validation algorithm
  - data replication in federated cloud storage environment

Objectives of this thesis

The main objective of this thesis is to create a data reliability and integrity tool (DRI) in form of a web service that would monitor the availability and integrity of data stored in federated cloud storage.

Important aspects of this work

- design and implementation of validation web service prototype
- integration of DRI service with the rest of VPH-Share project
- periodical and on-request validation
- literature research on efficient cloud storage validation algorithm
- propose network efficient validation algorithm of data in the cloud

Standard methods for data integrity

Data integrity building blocks:

- hash functions (MD5, SHA-1, SHA-256),
- Message Authentication Code(MAC) – integrity and authenticity assurance
- Error Correcting Code(ECC) – corruption detection and correction

Popular approaches:

- MD5/SHA-1 checksum for software packages
- integrity checksums for network messages
- ECCs in hardware solution

Cloud storage data integrity challenges:

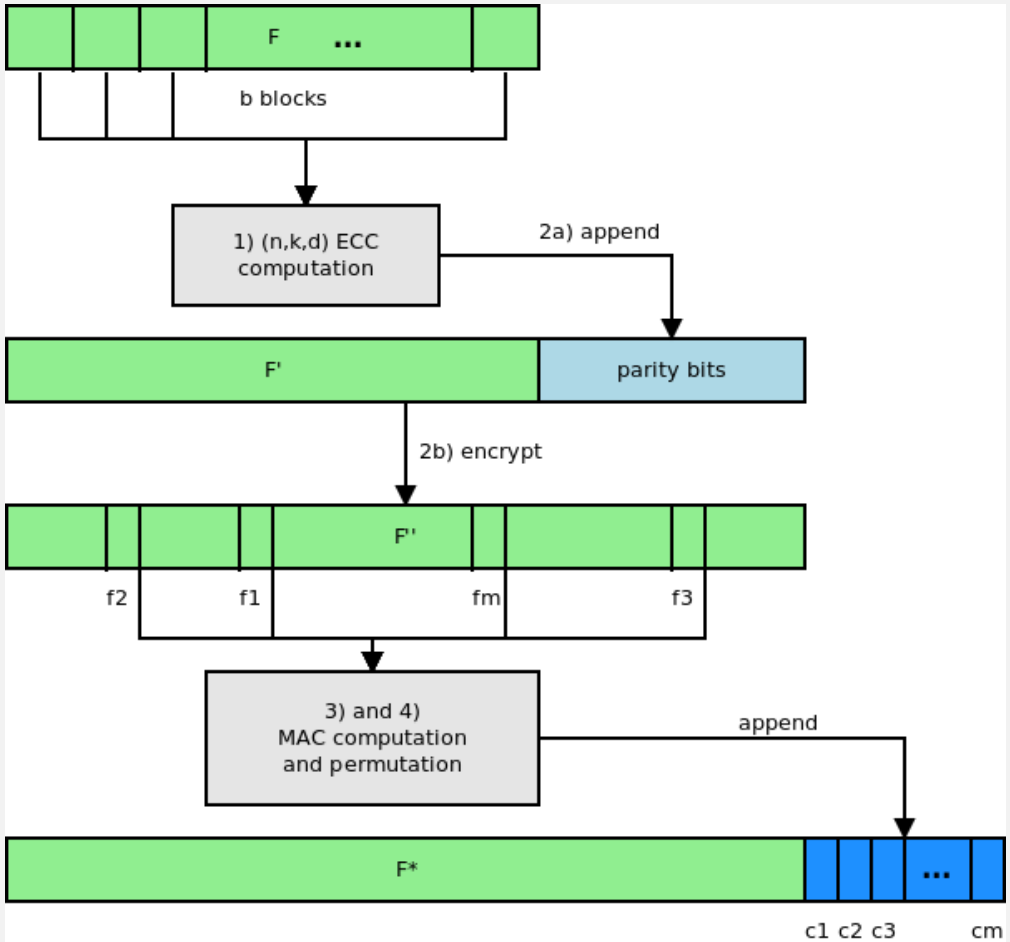
- huge amounts of data – inefficient remote validation
- externallly stored over broadband network – network limitations

If whole-file content validation is infeasible, then maybe we should try probabilistic validation

Approaches to data integrity in the cloud

Proof of Retrievability (POR) scheme:

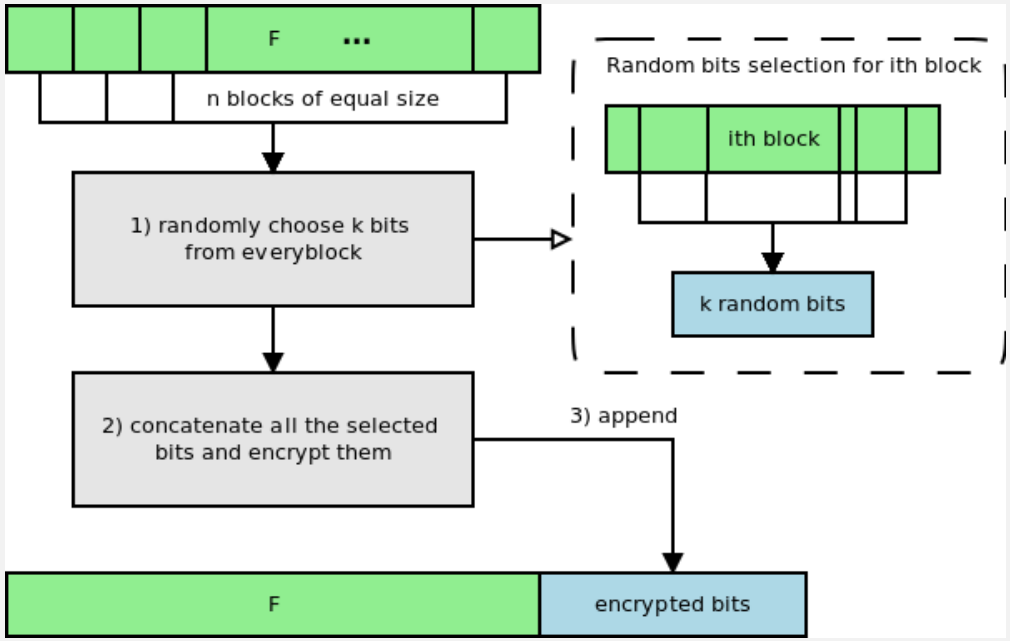
- divide a file F into b blocks and apply ECCs
- encrypt the file with appended parity bits
- select m blocks out of M, compute MACs and append them to file



Approaches to data integrity in the cloud

Data integrity proofs (DIP) scheme:

- divide a file F into n blocks and select randomly k bits from every block
- concatenate all selected bits and ecnrypt them
- append encrypted bits to the end of file

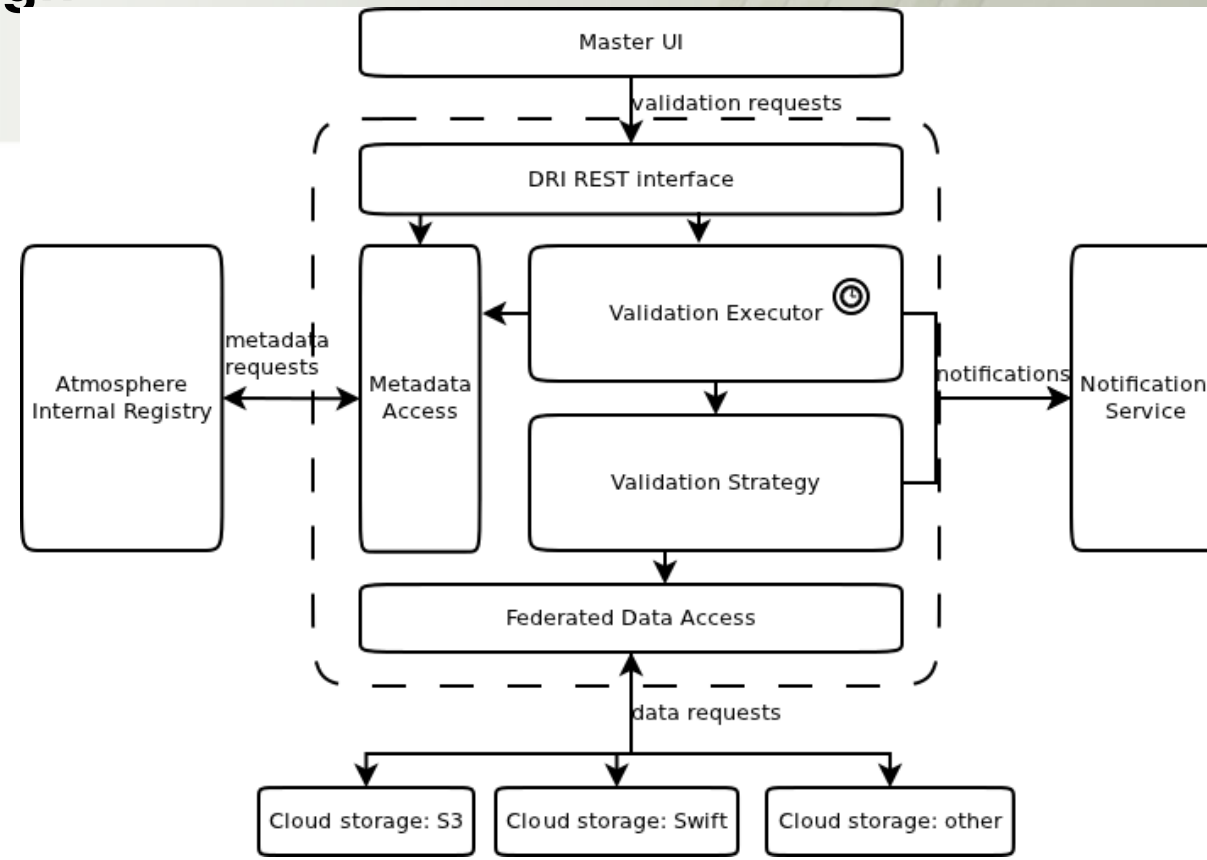


Drawbacks of the outlined approaches

- modifications applied to the original file (appending metadata, content encryption)
- assume computing capabilities on the prover side
- do not take into account cloud REST API limitations (no support for multiple HTTP Range requests)

DRI service design

- stateless REST web service in VPH-Share cloud environment
- periodical and on-request probabilistic validation of data in federated cloud storage
- data replication over cloud providers



Implementation technologies

- JClouds library – cloud storage abstraction
- Quartz – task scheduling
- JAX-RS – REST web service
- Java, Guice, Guava, Tomcat

DRIService

+ registerDataset(dataset : ManagedDatasetDescription) : ManagedDatasetI
+ unregisterDataset(id : ManagedDatasetID)
+ replicateDatasetToResource(id : ManagedDatasetID, source : DataSourceID
+ dereplicateDatasetFromResource(id : ManagedDatasetID, source : DataSou
+ datasetChanged(id : ManagedDatasetID, dataset : ManagedDatasetDescri
+ validateDataset(id : ManagedDatasetID) : Message
+ setManagementPolicy(policy : ManagementPolicy)
+ getManagementPolicy(id : ManagedDatasetID) : ManagementPolicy

DRI validation algorithm

Setup phase

- divide file F into n equal chunks
- compute MAC checksum for every chunk and store

Validation phase:

- randomly select k chunks
- compute MAC checksum of selected chunks and compare

Metric	our approach	whole-file approach
E_{det}	$\frac{k}{n}$	1
N_{over}	$\sim F \times \frac{k}{n}$	$\sim F$
T_{exec}	$\sim k \times (\frac{F}{n \times speed} + latency)$	$\sim \frac{F}{speed} + latency$

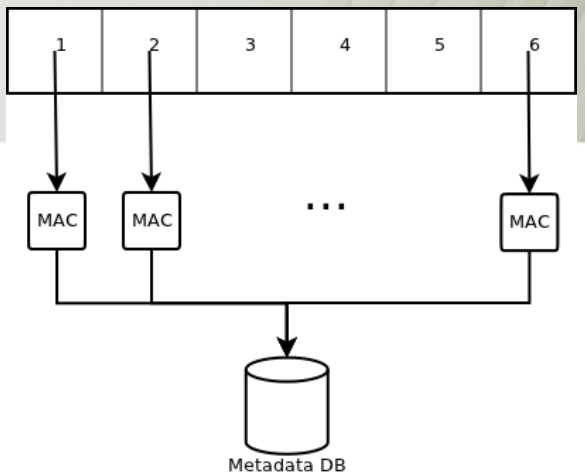


Figure: Setup phase

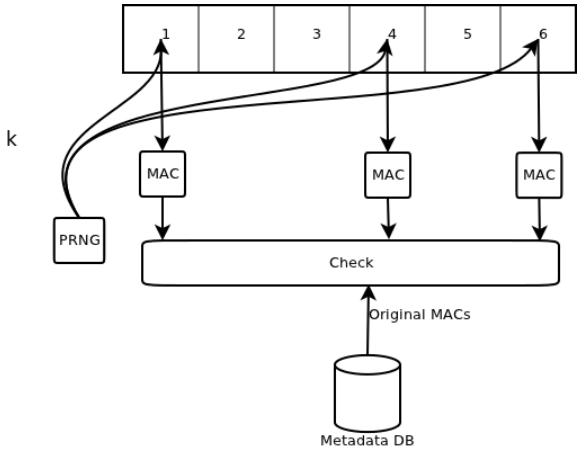


Figure: Validation phase

DRI service verification

DRI Notification Service

Dataset name	Notification status	Execution time	Time scheduled
test_dataset	Integrity errors detected	2s	8/10/13 12:52 PM

The dataset test_dataset is INVALID

Below is the detailed validation report:

Logical data identifier	Integrity status
moon.jpg	INVALID
earth.jpg	INVALID
time-machine.txt	UNAVAILABLE

test_dataset	Integrity errors detected	2s	8/10/13 12:51 PM

The dataset test_dataset is INVALID

Below is the detailed validation report:

Logical data identifier	Integrity status
moon.jpg	INVALID
time-machine.txt	UNAVAILABLE

test_dataset	Validation success	2s	8/10/13 12:47 PM
test_dataset	Tagged dataset as managed	5s	8/10/13 12:45 PM

Results and future work

Results:

- proposed an efficient algorithm for data validation in the cloud
- proposed methodology to monitor data reliability and integrity in the cloud
- enabled VPH-Share project users to monitor data integrity and notify in case of failures

Future work:

- Design how to combine DRI monitoring service with federated cloud storage data access layer
- Extract DRI component from VPH-Share context and share as open source
- Design better data validation algorithm in the cloud

Acknowledgement

More at `http://dice.cyfronet.pl/VPH-Share`

This thesis was realized partially in the framework of the following projects:

- Virtual Physiological Human: Sharing for Healthcare (VPH-Share) – partially funded by the European Commission under the Information Communication Technologies Programme (contract number 269978).
- Project UDA-POKL.04.01-01-00-367/08-00 "Improvement of didactic potential of computer science specialization at AGH", at the Department of Computer Science, AGH University of Science and Technology, Al. A. Mickiewicza 30, 30-059 Kraków

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY