Ensuring Data Storage Security in Cloud Computing

Cong Wang[1], Qian Wang[1], Kui Ren[1], and Wenjing Lou[2]

[1]ECE Department, Illinois Institute of Technology

[2]ECE Department, Worcester Polytechnic Institute

Outline

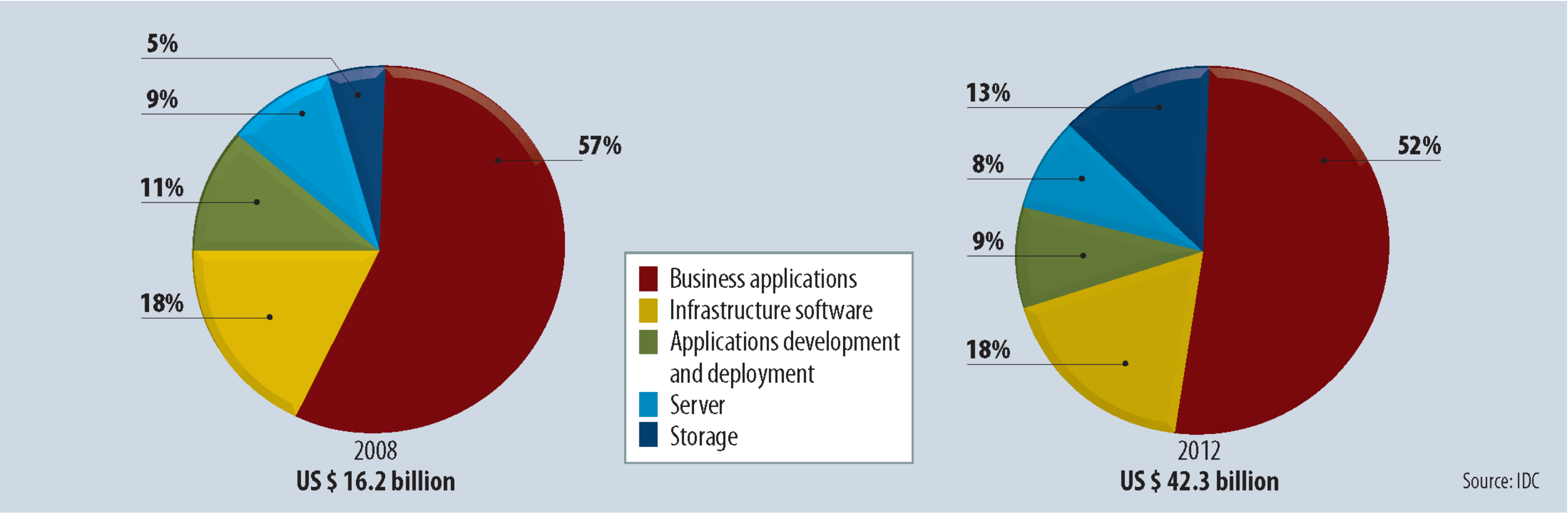
- Cloud Computing and Its Security Challenges
- Data Storage Security in Cloud Computing
- Our Approach
- Evaluation
- Concluding Remarks

Cloud Computing Background

Ubiquitous Security & PrivaCy Research Laboratory

- Cloud computing has been envisioned as the next-generation architecture of IT enterprise.

  on-demand self-service, ubiquitous network access, location independent resource pooling, rapid resource elasticity, usage-based pricing and transference of risk



Prediction from Market-research firm IDC, cloud-computing revenue will increase from US $16.2 billion to 42.3 billion during the next few years.

Image from: Neal Leavitt, "Is Cloud Computing Really Ready for Prime Time?," Computer, vol. 42, no. 1, pp. 15-20, January, 2009.

ILLINOIS INSTITUTE OF TECHNOLOGY

Cloud Computing Background

UbiSeC Lab
Ubiquitous Security & PrivaCy Research Laboratory

- Along with the coming of Cloud Computing is its untested deployment, correlated adversarial models and vulnerabilities:
  - Secure resource virtualization
  - Practical integrity mechanisms for data outsourcing
  - Secure computation outsourcing
  - Business and security risk models and clouds
  - Secure data management outsourcing
  - and many……

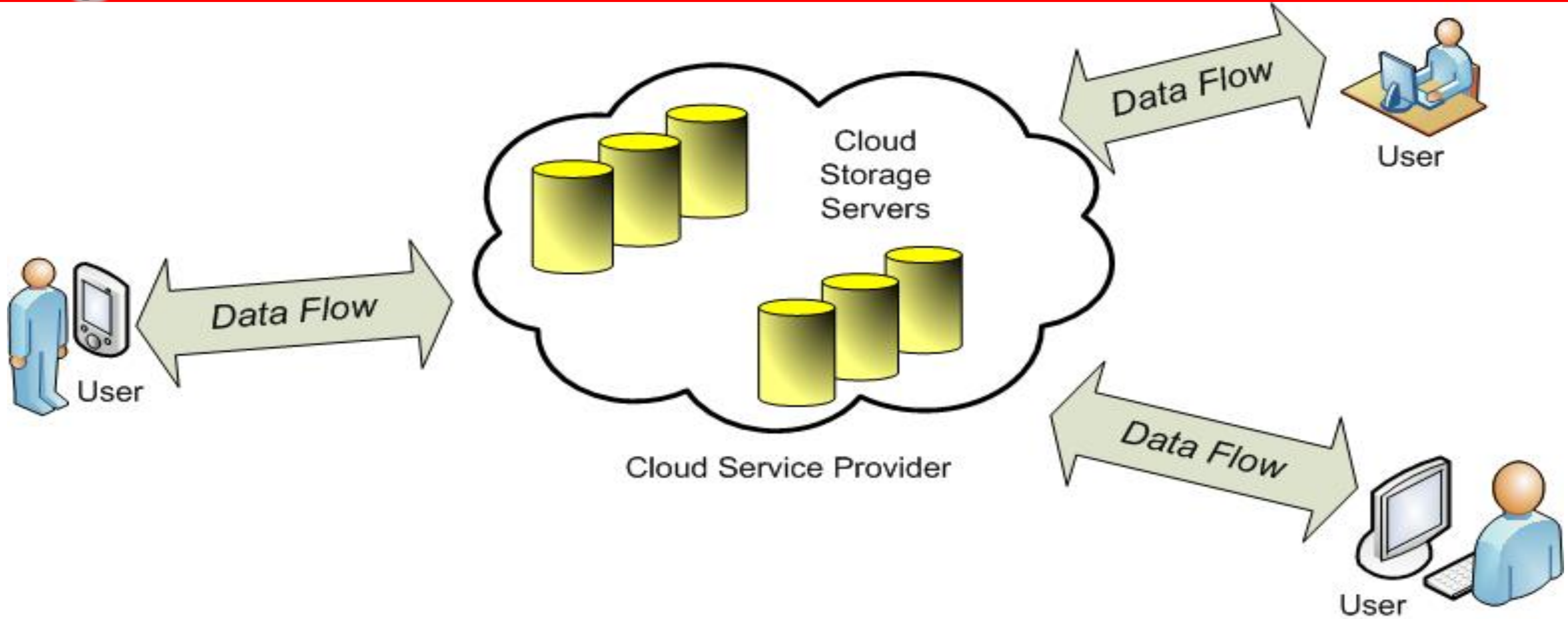
It is imperative that our community gets involved at this early stage and do it right for the first time!

ILLINOIS INSTITUTE OF TECHNOLOGY

Outline

- Cloud Computing Background
- Data Storage Security in Cloud Computing
- Our Approach
- Evaluation
- Concluding Remarks

Overview for Data Storage in Cloud



- From user′s perspective, data outsourcing brings:
  - Relief of the burden for storage management
  - universal access to data, independent of location
  - lower capital expenditure (CapEx) on hardware, software and services
- Data outsourcing *also eliminates users' ultimate control over the fate of their data.*

Shall We Trust the Cloud for data integrity?

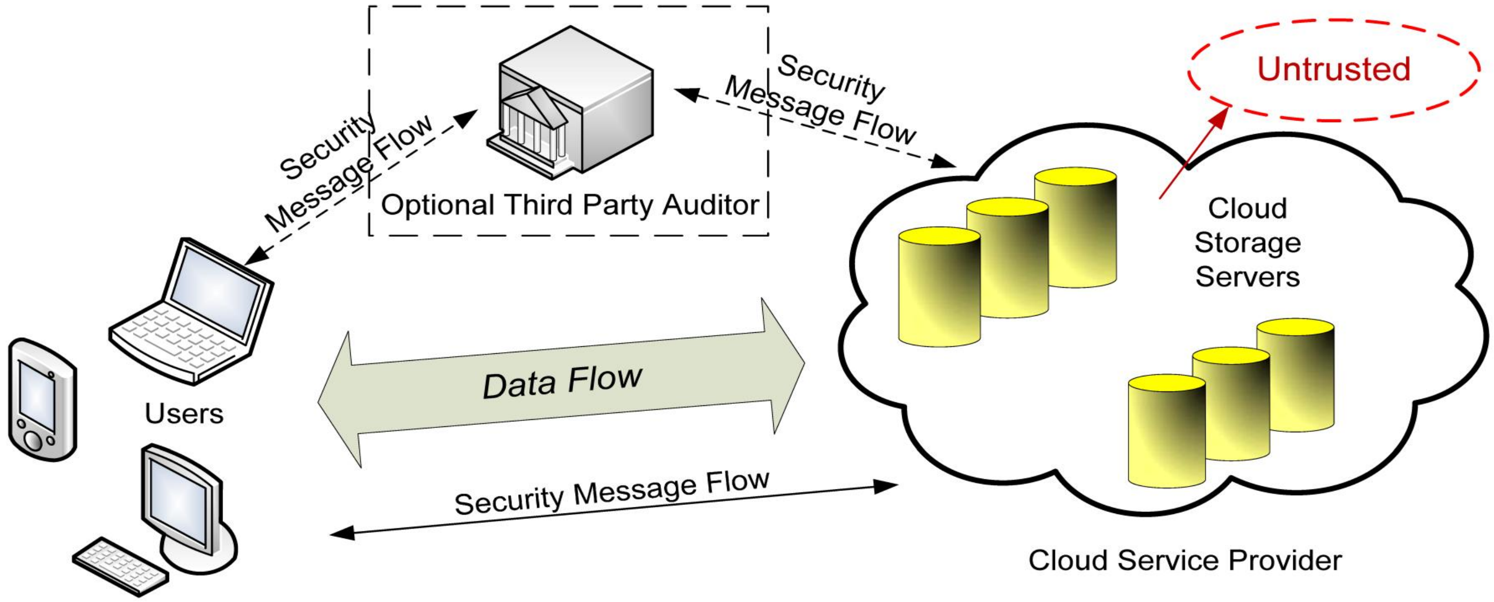
UbiSeC Lab
Ubiquitous Security & PrivaCy Research Laboratory

- Broad range of threats for data integrity still exist:
  - Internal: Byzantine failure, management errors, software bugs etc.
  - External: malicious malware, economically motivated attacks etc.
- Motivation for the Cloud service providers to cheat:
  - Discard rarely accessed data for monetary reason
  - Hide data loss incident for reputation.
- *While cloud data storage is economically attractive for the costs and complexity of long-term large-scale data storage, it doesn't offer guarantees on data integrity and availability.*

ILLINOIS INSTITUTE OF TECHNOLOGY

Problem Description

UbiSeC Lab
Ubiquitous Security & PrivaCy Research Laboratory



- Users should be equipped with security means so that they can make continuous correctness assurance of their stored data.

- Data integrity auditing tasks, if necessary, can be delegated to an optional Third Party Auditor (TPA).

ILLINOIS INSTITUTE OF TECHNOLOGY

Challenges for ensuring data integrity in Cloud

- Traditional crypto primitives can not be directly adopted.
  - No local copy of data at user side.
  - Retrieving large amount data for checking is unpractical.
    - *I/O burden on both servers and user, Huge network traffic, Expensive services charge, by byte of I/O and byte transferred*
- Data dynamics should be considered
  - Cloud is not just a data warehouse: data may be frequently updated.
    - *Most previous work on remote data integrity do not support data dynamics*
- Distributed protocols for storage correctness is demanded
  - Cloud is powered by data centers running in a simultaneous, cooperated and distributed manner
    - *Most previous work on distributed data storage only provide binary results for the storage correctness.*

Design Goals

- Storage Correctness Verification
  - Distributed protocol for storage correctness assurance
- Fast Data Error Localization (outperform the binary result)
  - Identifying misbehaving server(s)
- Explicit Dynamic Data Operation Support
  - Data modification, deletion and append are considered
- Dependability
  - Minimize the effect brought by data errors or server failures
- Efficiency

Outline

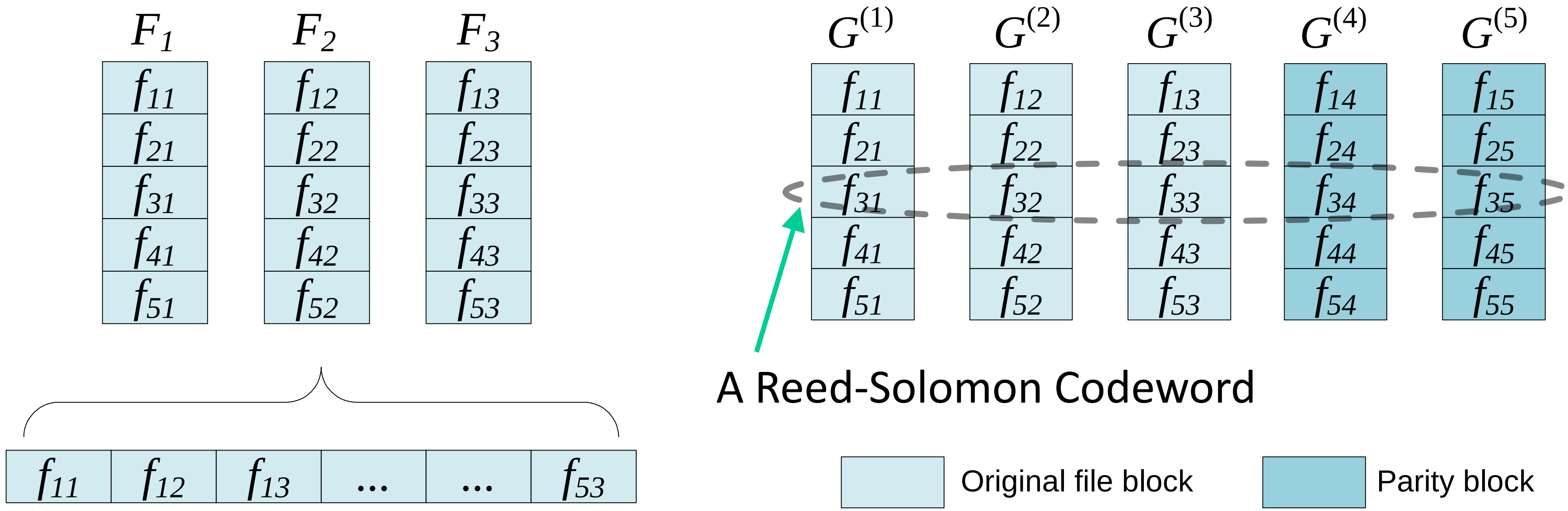
UbiSeC Lab
Ubiquitous Security & PrivaCy
Research Laboratory

- Cloud Computing Background
- Data Storage Security in Cloud Computing
- Our Approach
  - Ensuring Cloud Data Storage
  - Supporting Data Dynamics
- Evaluation
- Concluding Remarks

ILLINOIS INSTITUTE
OF TECHNOLOGY

Ensuring Cloud Data Storage

Ubiquitous Security & PrivaCy Research Laboratory



We rely on a $(m + k, k)$ Reed-Solomon erasure-correcting code to disperse the data file F redundantly across a set of $n = m + k$ distributed servers.

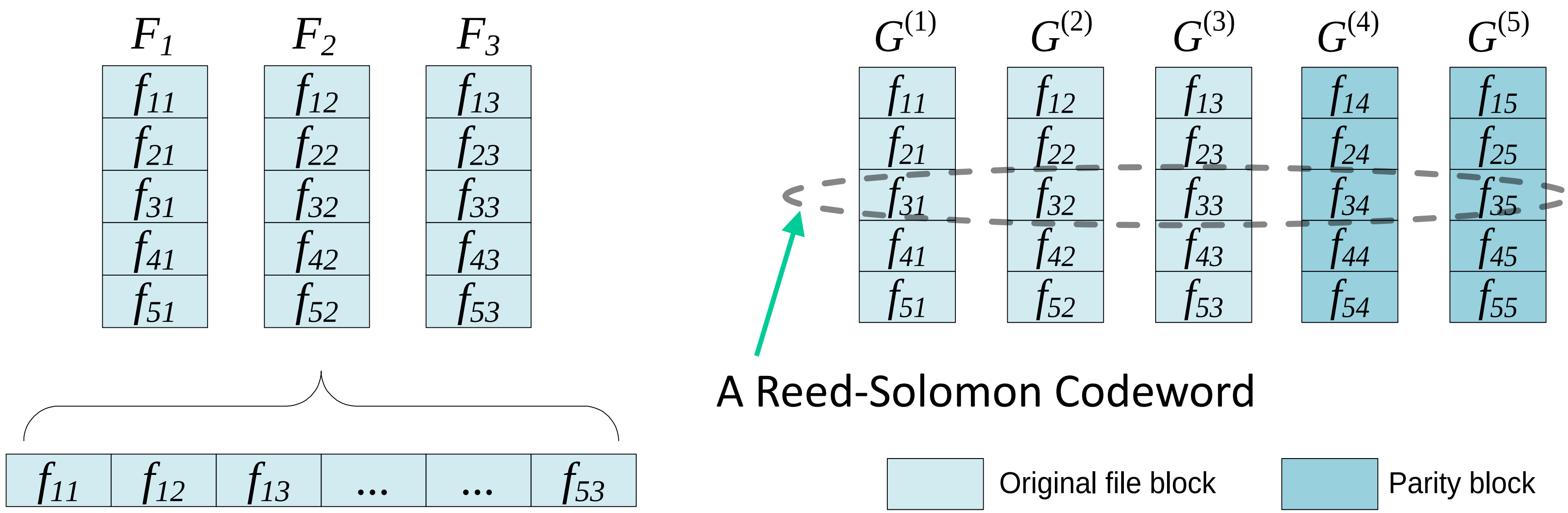
The systematic layout with parity vectors is achieved with the information dispersal matrix $\mathbf{A}$:

$$\mathbf{G} = \mathbf{F} \bullet \mathbf{A} = \mathbf{F} \bullet (\mathbf{I} \mid \mathbf{P}) = (F_1, F_2 \cdots F_m) \bullet (\mathbf{I} \mid \mathbf{P})$$

$$= (G^{(1)}, G^{(2)} \cdots G^{(m)}, G^{(m+1)}, \cdots G^{(m+k)})$$

ILLINOIS INSTITUTE OF TECHNOLOGY

Ensuring Cloud Data Storage



Based on the codeword relationship, we can verify the correctness of data block in each “row” via information dispersal matrix **A** (or **P**).

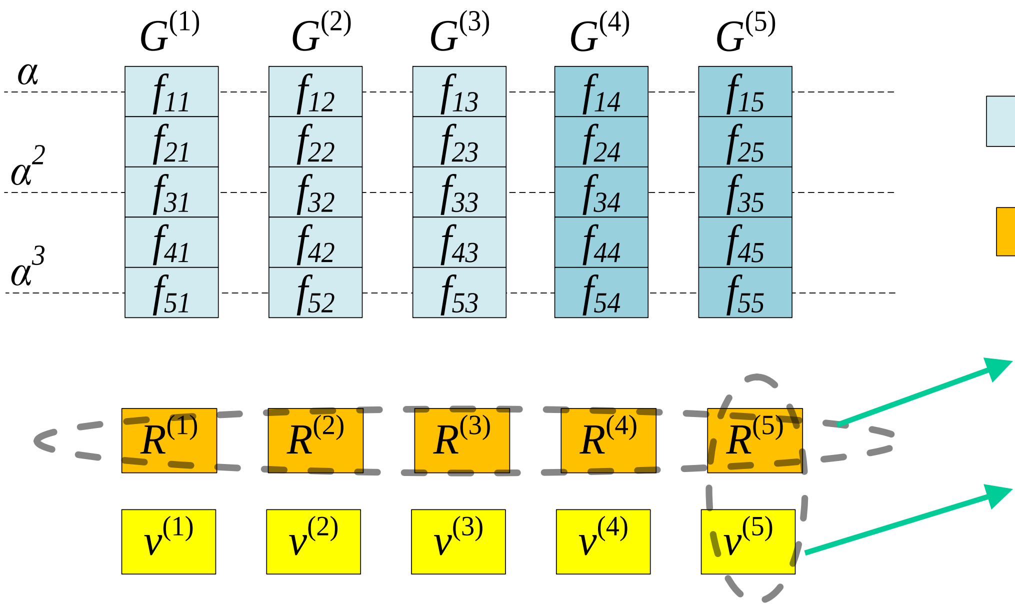
$$(f_{31}, f_{32}, f_{33})\cdot P = (f_{34}, f_{35})$$

Can we do better?

Drawbacks: 1. need block retrieval at first, which is proportional to vector length.
2. large communication overhead.
3. only binary result about the storage state.

Ensuring Cloud Data Storage

Ubiquitous Security & PrivaCy Research Laboratory



Original file block

Parity block

Server Response

Pre-computed token

1. Check if it is a valid a Reed-Solomon Codeword. If not, go to step 2.
2. Check if $R^{(j)}=v^{(j)}$ ($j = 1, \ldots, 5$) to identify the misbehaving server(s).

Random sampling + homomorphic token pre-computation(linear combination)

$$R^{(j)} = v^{(j)} = \sum_{q=1}^{3} \alpha^q * G^{(j)}[I_q], \{I_q = 1,3,5\} \; and \;\; j = \{1,\ldots,5\}$$

$$(R^{(1)}, R^{(2)}, R^{(3)}) \cdot P = (R^{(4)}, R^{(5)})$$

Advantages: 1. only small constant block retrieval is required

2. Finding misbehaving server(s)

3. Efficiency

ILLINOIS INSTITUTE OF TECHNOLOGY

Outline

UbiSeC Lab
Ubiquitous Security & PrivaCy Research Laboratory

- Cloud Computing Background
- Data Storage Security in Cloud Computing
- Our Approach
  - Ensuring Cloud Data Storage
  - Supporting Data Dynamics
- Evaluation
- Concluding Remarks

Supporting Data Dynamics

- Cloud data storage is not only for archive purpose
- General block-level operations: update, delete, append…
- Trivial way is to download all the data from the cloud servers and **re-compute** parity blocks and tokens
- Can we do better?

Supporting Data Dynamics

UbiSeC Lab
Ubiquitous Security & PrivaCy Research Laboratory

Logical representation of data dynamics,
including block update, append and delete

$$\begin{array}{ccc} F_{1*} & F_{2*} & F_{3*} \\ f_{11} & f_{12*} & f_{13} \\ f_{21*} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \\ f_{41} & f_{42} & f_{43*} \\ f_{51} & f_{52*} & f_{53} \\ f_{61} & f_{62} & f_{63} \end{array} = \begin{array}{ccc} F_{1} & F_{2} & F_{3} \\ f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \\ f_{41} & f_{42} & f_{43} \\ f_{51} & f_{52} & f_{53} \\ 0 & 0 & 0 \end{array} + \begin{array}{ccc} \Delta F_{1} & \Delta F_{2} & \Delta F_{3} \\ 0 & \Delta f_{12} & 0 \\ \Delta f_{21} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \Delta f_{43} \\ 0 & \Delta f_{52} & 0 \\ f_{61} & f_{62} & f_{63} \end{array}$$

Original file block   Modified block   Appended block

$$\mathbf{F}^{*} \cdot \mathbf{A} = (\mathbf{F} + \Delta\mathbf{F}) \cdot \mathbf{A} = \mathbf{F} \cdot \mathbf{A} + \Delta\mathbf{F} \cdot \mathbf{A}$$

Due to the linear property of Reed-Solomon code, we can “amend” the parity blocks, *without involving any* of *the unchanged blocks.*

ILLINOIS INSTITUTE OF TECHNOLOGY

Supporting Data Dynamics

Ubiquitous Security & PrivaCy Research Laboratory

- Similarly: we can "amend" the tokens, *without retrieving any of the unchanged blocks*.

$$v^{(j)} = \sum_{q=1}^{r} \alpha^q * G^{(j)}[I_q], \{I_q \in [1, \ldots, l] \mid 1 \le q \le r\}$$

- Suppose a block $G^{(j)}[I_s]$, which is covered by the specific token $v^{(j)}$, has been changed:

$$G^{(j)}[I_s] \to G^{(j)}[I_s] + \Delta G^{(j)}[I_s]$$

- The token $v^{(j)}$ can be updated:

$$v^{(j)} \leftarrow v^{(j)} + \alpha^s * \Delta G^{(j)}[I_s], s \in \{q\}.$$

Outline

- Cloud Computing Background
- Data Storage Security in Cloud Computing
- Our Approach
  - Ensuring Cloud Data Storage
  - Supporting Data Dynamics
- Evaluation
- Concluding Remarks

Security Analysis

UbiSeC Lab
Ubiquitous Security & PrivaCy Research Laboratory

- Detection Probability: Assume the adversary modifies or deletes the data blocks in z rows out of the total l rows in the encoded file matrix.

- Each time we samples r rows to check, the detection probability will be:

$$p_d = 1 - \prod_{i=0}^{r-1} \left(1 - \min\left\{\frac{z}{l-i}, 1\right\}\right) \geq 1 - \left(\frac{l-z}{l}\right)^r$$

If $z/l = 1\%$ and $r = 460$, the detection probability P_d is at least 99%.

ILLINOIS INSTITUTE OF TECHNOLOGY

Security Analysis

UbiSeC Lab
Ubiquitous Security & PrivaCy Research Laboratory

- Colluding Attack Resistance: Can we hide the secret encoding matrix **P** without affecting the validity of the checking results?

- Yes! *Make use of the linear property of Erasure Correcting Coding.*

- Adding random perturbations to the encoded file matrix and hence hide the secret matrix **P.**
  - The linear property of RS-code makes random perturbations easily stripped away for verification purposes.

ILLINOIS INSTITUTE OF TECHNOLOGY

Performance Evaluation

File Pre-distribution Cost

Set I	m=4	m=6	m=8	m=10
k = 2	567.45s	484.55s	437.22s	414.22s

Set II	k=1	k=2	k=3	k=4
m=8	358.90s	437.22s	584.55s	733.34s

The cost of parity generation in seconds for an 8GB data file on Intel Core 2 processor running at 1.86GHz. For set I, the number of parity servers k is fixed; for Set II, the number of data servers m is constant.

Token Pre-computation Cost

- To verify the data once per day for the next 5 years, the average token pre-computation cost according to our implementation is 51.97s per data vector, given r = 460.

Outline

UbiSeC Lab
Ubiquitous Security & PrivaCy Research Laboratory

- Cloud Computing Background
- Data Storage Security in Cloud Computing
- Our Approach
  - Ensuring Cloud Data Storage
  - Supporting Data Dynamics
- Evaluation
- Concluding Remarks

Concluding Remarks

- 1) Instead of giving only binary results about the storage state across the distributed servers, our work further provides the *localization* of data error.
- 2) In addition to ensuring cloud data integrity, the new scheme supports secure and efficient *dynamic* operations on data blocks, including: update, delete and append.
- 3) Extensive security and performance analysis shows that the proposed scheme is highly efficient and resilient under various untrusted server scenarios.

Thank You!