様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成２５年　６月　３日現在

機関番号：２４５０６
研究種目：挑戦的萌芽研究
研究期間：2011～2012
課題番号：２３６５００３１
研究課題名（和文）堅牢生と安全性を備えたユビキタスデータアクセスの研究

研究課題名（英文）A Study of Ubiquitous Access with Confidentiality and Availability

研究代表者
申　吉浩（SHIN YOSHIHIRO）
兵庫県立大学・応用情報科学研究科・教授
研究者番号：６０５２３５８７

研究成果の概要（和文）：データを断片に分割し、断片毎にアーカイブ先を複数ランダムに選択する、冗長なアーカイブ方式を提案している。この方式において、断片の長さもランダムに決定するようにすることで、アーカイブデータとランダムに生成したデータの区別ができないことを、理論的にも統計的にも証明した。また、プロトタイプ実装により、実用上十分に高速に処理を行えることを確認した。クラウドにおける安価なバックアップサービスとして実用化できるものと期待する。

研究成果の概要（英文）：We propose a data archive scheme that divides the input data into many tiny fragments and distributes each fragment across multiple storages chosen at random.　In this research, we have proved that the archived data generated by our scheme can not be distinguished from randomly generated data from the probability and statistical point of view.　Also, we have developed a prototype of the scheme, and have verified that the scheme is efficient sufficiently for the practical use.　We expect that this scheme is useful to realize inexpensive back-up services in the cloud computing.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	2,700,000	810,000	3,510,000

研究分野：暗号理論・機械学習理論
科研費の分科・細目：情報学・計算機システム・ネットワーク
キーワード：クラウドコンピューティング、セキュリティ、データバックアップ、アーカイブ、可用性、秘匿性

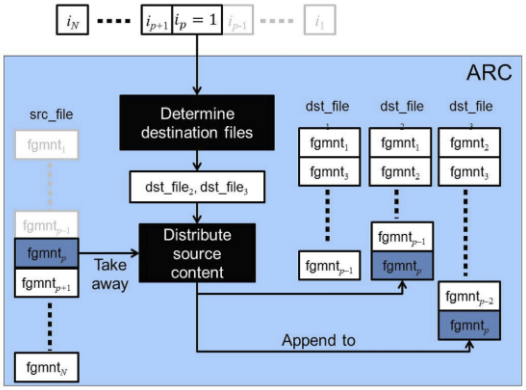
１．研究開始当初の背景

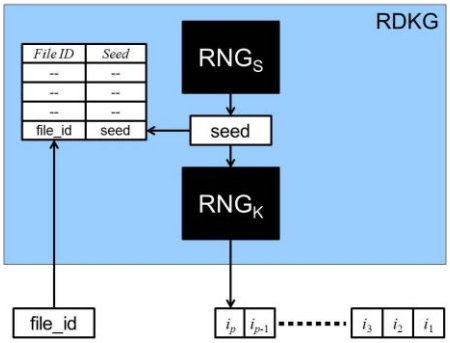
近年のモバイルコンピューティング及びインターネットの目覚ましい進歩、特にクラウドコンピューティングの実現に向けた産学官の熱意ある取り組みにより、ユビキタスデータアクセスが俄に現実味を帯びてきた。ユビキタスデータアクセスとは、いつでも、どこからでも、必要なデータに任意にアクセスすることを可能とする環境を指す。特に、クラウドコンピューティングを利用したユビキタスデータアクセスでは、ユーザは、ノート PC・スマートフォンなどのモバイル端末を用いて、インターネット上に遍在するストレージにアクセスし、いわば、巨大容量のネットワークドライブに常時接続している感覚で、写真や動画などの大容量データをストレスなくアップロード・ダウンロードする。一方、インターネットは管理された環境ではなく、可用性のレベルは不安定で、安全でもないことを認識しなければならない。ネットワークの帯域はまちまちであり、通信の集中による遅延の発生はしばしばで、サーバやルータの故障でサービスがダウンする危険も現実的である。また、WIMAX 等、 高速な

WiFi ネットワークの地域カバー率が上がってきているが、データ転送速度は、ハードディスクや USB メモリほどには高速ではなく、「ユビキタス感」を実感できるほどではない。一方、安全の 観点からは、通信経路での盗聴や改竄の脅威を原理的に免れ得ず、クラウドコンピューティングでは サービスは不特定に提供されるので、サービス（の管理者）を信用できる根拠もない。
則ち、セキュリティマネジメントの言葉で言うならば、CIA（Confidentiality・Integrity・Availability = 秘匿性・完全性・可用性）のいずれの要素も保証されていないということになる。本研究 では、分散アーカイブを基礎的なアイデアとして、CIA のうち C 及び A の要件を満足するユビキタスデータアクセスのための基礎技術を研究する（CIA の I については MIC、デジタル署名等有効な技術が存在する）。
分散アーカイブとはデータを複数のストレー ジに分散してアーカイブすることである。実は、インターネット上のストレージにアーカイブするデータの可用性を、現在のインターネットを前提として 向上させるためには、冗長性を持たせた分散アーカイブが唯一の有効な方策である。例えば、(n, k) 閾堅牢性という考え方では、データを n 箇所の異なるストレージに重複を許して分散アーカイブし、 n 箇所のうち k 箇所のストレージへのアクセスが得られればもとのデータが復元できる。言い換える と、n－k 箇所のストレージが失われても、残り k 箇所のストレージからデータの可用性を保証できる。因みに、RAID-5 は (n, n－1) 閾堅牢性の具体的な実現例である。

２．研究の目的
クラウドコンピューティングを利用して、ユビキタスデータアクセスを実現する為の技術を研究する。ユビキタスデータアクセスでは、ユーザは、モバイル端末からインターネットに遍在するストレージにアクセスし、ネットワークドライブのようにストレージを利用する。しかしながら、ユビキタスデータアクセスの実現にインターネットを利用する為には、インターネットが内包する可用性及び安全性の問題を解決しなければならない。本研究では、分散アーカイブに関する筆者の基礎的な発見に基づいて、上記問題に対する学術的解決を提案し、結果をプロトタイプとして実現する。
本研究の成果の活用方法として、クラウドを利用した安価なバックアップサービスの提供がある。阪神淡路・東日本大震災は、中小零細企業における業務データの喪失、それに起因する事業継続の阻害が重要な問題であることを明らかにした。大企業は、データセンターが運用するバックアップサービスを利用する資力があるが、中小零細企業では経費節減のためローカルバックアップが主流であった。コンピュータの故障等に対しては、ローカルバックアップは確かに有効な対策であるが、建物ごと破壊してしまう地震等の天災では、全く効果がない。クラウドの利用は、中小零細企業でも利用できる安価なリモートバックアップを提供する可能性があり、現在、社会の注目を集めている。しかしながら、前述の理由により、機密性と可用性の問題を解決しない限り、信頼できるリモートバックアップは実現できない。本研究は、この問題を解決することを目的としている。

３．研究の方法
本研究では、(n, k)閾堅牢性と (n, k)閾秘匿性の二つの考え方が基礎となる。(n, k)閾堅牢性は n 箇所のストレージのうち k 箇所にアクセスできれば、データを完全にリトリーブできる性質を指す。他方、(n, k)閾秘匿性は、データが分散アーカイブされている n 箇所のストレージのうち k － 1 箇所からデータが漏洩したとしても、データを完全には復元することができないという性質を 指す(効率性の観点から、Shamir 等の秘密分散法における厳密性は仮定しない)。例えば、データを nCn−k+1 個の 断片に分割し、それぞれの断片を n － k ＋ 1 箇所のストレージにコピーして記録する。異なる断片は、異なる n － k ＋ 1 個のストレージの組に記録されるとすると、このアーカイブ方式は (n, k)閾堅牢性と(n, k)閾秘匿性の両方を満足する。
さて、上記のアーカイブ方式を改良したものが、本研究で提案する方式である。即ち、データをもっと多くの小さな断片に分割し、それぞれの 断片を記録する n － k ＋ 1 個のストレージの組みを、断片毎に独立に、かつ、一様ランダムに選ぶようにする。
下図は、この分散の仕組みを示したものである。入力となるデータ（src_file）は、複数の断片(fgmnt$_p$)に分割される。一方、擬似乱数生成器(RNG)により生成される乱数列(i_p)中の乱数を順に取得し、その乱数をインデックスとして予め定められたテーブルを参照して､対応する断片をアーカイブする先の n－k ＋ 1 個のストレージを特定する(Determine destination files)。最後に、対応する断片を、先に選択されたストレージ（dst_file）の末尾に当該断片を追記する。断片のアーカイブ先が、擬似乱数生成器により生成される乱数によって決定される点が重要であり、本研究における工夫である。

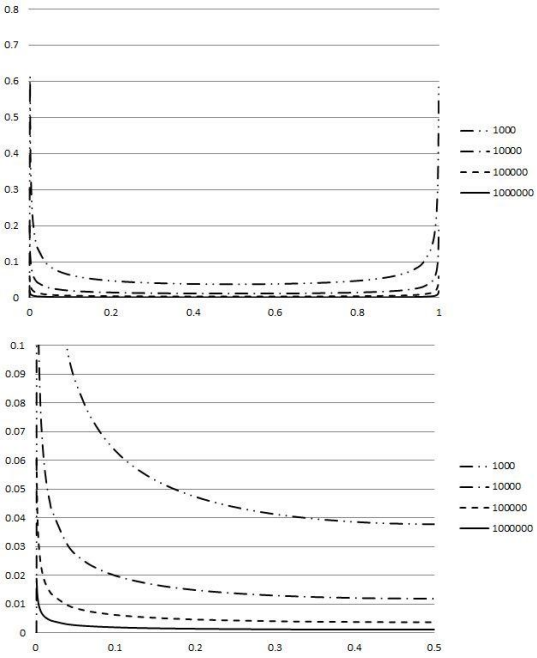




この工夫により、オリジナルデータの一部が本質的に欠損することを保証する(n, k)閾秘匿性に加えて、漏洩したアーカイブデータからもデータが漏洩しないという更に進化した秘匿性が得られる。
本研究では、提案方式の秘匿性を、以下の二つの方法で検証する。
（あ）オリジナルデータに関するあるモデルを仮定し、当該モデルの上で、アーカイブデータとランダムデータの差異を確率的に評価する。
（い）ウェブ上から実際のファイルを取得し、（あ）で求めた結果が現実世界で成り立つかを統計的手法を用いて検証する。
更に、提案方式の実現性と速度等の性能を評価することを目的として以下を実施する。
（う）プロトタイプを開発し、処理速度等の評価を行う。

４．研究成果
前記の実施項目（あ)、(い)、(う）それぞれについて、研究の成果を述べる。

（あ）モデルに基づく確率評価
まず、確率評価の出発点は、アーカイブファイルに現れる断片のオリジナルデータにおける位置を推定するという攻撃の確率的評価である。攻撃者による推定が正しい確率を計算し、その近似式を求めてグラフ化すると、右図上のグラフとなる。横軸は、オリジナルデータ中のビット位置を表す。このグラフから分かるように、データの先頭と最後尾では、推測が的中する確率は高いが、先頭と最後尾の近傍を除いた広い範囲で、低い値にとどまることが分かる。所謂、バスタブ型のグラフとなる。
更に、断片への分割数を大きくすると、バスタブの底部の確率は急速に小さくなることが分かる。右図下のグラフは、右図上のグラフの半分を拡大したものである。断片数を1000 から 1000000 まで増やした時、バスタブの底部は x 軸にはりついていく様子が分かる。提案方式を実用化した場合は、断片長は数ビット程度と想定されるため、断片数は非常に大きくなる。



上記のことから、データの先頭と最後尾の極く狭い近傍をのぞけば、断片のオリジナルデータ中での位置を特定することは、確率的に不可能であることがわかる。
更に、提案アーカイブ方式では、オリジナルデータに含まれるビット０とビット１の割合は、アーカイブデータに継承され、漏洩してしまう。ここでは、ビット０とビット１の割合は既知であるとして、その割合を保つように生成されるランダムビット列とアーカイブデータとの差異を確率的に評価する。即ち、あまり長くない有限長のビットパターンがアーカイブデータ中の特定の位置に出現する確率を求め、その確率が位置に依存せずに概ね一定であれば、少なくとも有限ビットパターンに関してランダムビット列と差がないと結論する。
もう少し詳しく説明する。テキストは、１バイト及び２バイト単位で符号化される。従って、アーカイブの単位である断片の長さをバ

イト単位とすると、文字情報が漏洩してしまう。従って、断片長はバイト単位としてはならないが、それでも情報の漏洩が心配される。例えば、ASCII 符号化は、実質的に 7 ビット符号化であり、各バイトの先頭ビット（Most Significant Bit, MSB）は必ず 0 となる。即ち、アーカイブデータ中に現れるビット 0 をバイトの境界（delimiter）と推測して、文字を特定することが出来てしまうかもしれない。
本研究の成果として、断片長を 8 と素となるように取り、かつ、ASCII 文字がランダムに選択されるテキストを入力とすると、ファイルの先頭と最後尾の近傍を除いて、アーカイブファイルは近似的にランダムデータとなることを示した。より正確には、比較的短い任意のビットパターンに対して、そのビットパターンの後に０または１が現れる条件付き確率が、近似的に、パターンの出現位置に依存しないことを、数学的に証明した。この事実は、提案方式により生成されるアーカイブデータが、上記条件付き確率に基づいて生成されるランダムビット列と近似的に一致することを意味する。
同じ事実が、S-JIS、UTF-8、JPEG など、ASCII 以外の符号化方式にも成立することを証明することもできる。

（い）実際のドキュメントは、上記のランダムモデルに厳密に従うことはない。文字毎に出現頻度は異なり、また、文字間の出現には相関がある。例えば、英文では、アルファベット E の出現頻度と Z の出現頻度の間には大きな隔たりがある。単語としては、THE の出現頻度が非常に高いことも知られている。
モデルを考える場合には、こうした自然言語の知識を導入すると、非常に難しくなり、かつ、数学的に扱うことが困難になるため、統計的にはランダムモデルで近似できると考える。この仮定に基づいた検証が、（あ）の後半で述べたものである。
この理論的な解析に基づいて、提案方式が実用でも有効であることを検証するために、ウェブから種々の符号化方式で作成された文書データを数多く取得し、上記のビットパターンに関する確率が成立するかを検証した。検証の手法としては、分散分析（ANOVA）を用い、大きな危険率に対して帰無仮説（観測される確率が理論値と等しい）が棄却できるかを検証した。
その結果は大変興味深いものであった。
まず、アーカイブする際の断片長を固定すると、帰無仮説を棄却できてしまう。即ち、理論値と観測値の間に有意に差があり、アーカイブデータとランダムビット列を有意に区別することが可能で、秘匿性に関して本研究で提唱している安全性の基準（ランダムビット列と区別できない）を満たさない。
それに対し、断片長をもランダムに選ぶという工夫を加えることで、帰無仮説を棄却できなくなることを発見した。ただ単に棄却できないだけでなく、算出された p 値は 1.0 に近く、高い信頼性をもって、観測値と理論値とが一致すると結論することができる（統計的には、差が検出できないというのが正確な表現になる）。
即ち、分散先をランダムに選ぶだけではなく、断片長をもランダムに選ぶことにより、実際の文書（画像なども含む）をアーカイブしたファイルは、（与えられた確率分布に従って）ランダムに生成されるビット列と統計的に区別がつかないことが確かめられた。

（う）提案手法の実現性及び性能の評価のために、提案手法を C++によりプロトタイプとして実装した。実装作業は、科学的アルゴリズムの実装に実績を有するソフトウェア開発会社に委託した。
実装したプロトタイプを評価した所、ビット操作があるにもかかわらず、非常に高速であり、使用した場合の体感的にも、データの転送時間のほうが遥かに大きいことが確かめられた。速度上のボトルネックが通信にあることがわかり、アーカイブアルゴリズムとして実用的であることを検証することができた。
現在は、開発したプロトタイプに基づいて、バックアップサービスのプロトタイプを作成し、実ユーザを用いた実証実験を計画している。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）
[1] A. Tallat, H. Yasuda and K. Shin. Technology of secure file archiving in the uniformly random distributed archive scheme. Journal of Information Security, Scientific Research, Vol.4, No.1: pp.42–53, 2013.
[2] A. Tallat, K. Shin, H. Lee and H. Yasuda. Some remarkable properties of the uniformly random distributed archive scheme. Advances in Information Science and Service Science, Vol.4, No.11: pp.114–124 2012.

〔学会発表〕（計１件）
[1] A. Tallat, K. Shin and H. Yasuda. Some remarkable properties of uniformly random distributed archive scheme In The 6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT 2011), pp.586–591, IEEE Explore, 29

November 2011, Jeju, Korea

６．研究組織
(1)研究代表者
申　吉浩（SHIN YOSHIHIRO）
兵庫県立大学・応用情報科学研究科・教授
研究者番号：６０５２３５８７