Regulatory-Compliant Data Management

Radu Sion
NSAC Laboratory
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794
sion@cs.stonybrook.edu
Phone: (631) 632 - 1672
Fax: (631) 632 - 1690

Marianne Winslett
DAIS Laboratory
Department of Computer Science
University of Illinois
Urbana, IL 61801
winslett@cs.uiuc.edu
Phone: (217) 333-3536
Fax: (217) 244-6500

Summary. Digital societies and markets increasingly mandate consistent procedures for the access, processing and storage of information. In the United States alone, over 10,000 such regulations can be found in financial, life sciences, healthcare and government sectors, including the Gramm-Leach-Bliley Act, Health Insurance Portability and Accountability Act, and Sarbanes-Oxley Act. A recurrent theme in these regulations is the need for regulatory-compliant data management as an underpinning to ensure data confidentiality, access integrity and authentication; provide audit trails, guaranteed deletion, and data migration; and deliver Write Once Read Many (WORM) assurances, essential for enforcing long-term data retention and life-cycle policies.

Unfortunately, current compliance data management WORM mechanisms are vulnerable to faulty behavior or insiders with incentives to alter stored data because they rely on simple enforcement primitives such as software and/or hardware device-hosted on/off switches, ill-suited to their target threat model. In practice, these first-generation mechanisms allow an insider using off-the-shelf resources to replicate illicitly modified data onto seemingly-identical units without detection.

More generally, the design of compliance data management is extremely challenging due to the conflict between security, cost-effectiveness, and efficiency. For example, the requirement to find requested information quickly means in practice data must be indexed. But trustworthy indexing of compliance data is a challenging problem, as it is easy to tamper with traditional indexes stored on WORM. Further, trustworthy indexes will make it very hard to delete all traces of documents that are past their retention periods, as required by certain regulations. Yet another complicating factor is the decades-long retention periods required by many regulations; it is unrealistic to expect data to reside on the same device for so long.

In this tutorial, we will discuss achieving strongly compliant data management in realistic adversarial settings. Specifically, we will explore designs for compliant data management systems that offer guaranteed document retention and deletion, quick lookup, and compliant migration, together with support for litigation holds and several key aspects of data confidentiality. Moreover, we will discuss the benefits of the recent advent of tamper-resistant, general-purpose trustworthy hardware which opens the door to fundamentally new assurance paradigms, e.g., by deploying this new hardware running certified code at the data management server. As heat-dissipation concerns greatly limit the performance of tamper-resistant processors, our goal is to investigate and evaluate software architectures for leveraging a secure processor in the server stack with minimal impact on cost and efficiency.

One of the main challenges of such an endeavor lies in securing the system against attack by insiders with superuser powers, and in balancing the conflicting requirements for trustworthiness, high performance, and low cost. For example, direct implementations of the server data management stack inside secure hardware will fail in practice due to poor performance. Efficient protocols making use of such hardware will need to do so sparsely, asyn-

chronously from the main data flow. We will discuss mechanisms that minimize the SCPU overhead for expected transaction loads, and use amortized commitment schemes to enforce write-once-read-many (WORM) semantics at the throughput rate of the servers ordinary processors. We will explore how to deploy existing work on compliance indexing to create efficient indexes secured by novel cryptographic constructs such as fast short-lived signatures. We will discuss achieving compliant data migration through backwards-compatible, inter-device trust chains and fast re-encryption.

Audience: extremely broad, non-CS. The intended audience is extremely broad, to include not only researchers but **also practitioners** in all areas of information processing that involve data-centric information regulation. Recent compliance regulations are intended to foster and restore humans trust in digital information records and, more broadly, in our businesses, hospitals, and educational enterprises. As increasing amounts of information are created and live digitally, compliance data management will be a vital tool in restoring this trust and ferreting out corruption and data abuse at all levels of society.

Proposed Length: 3 hours. This duration will allow the discussion of the three main assurance levels, namely WORM semantics and data migration, secure indexing, and secure deletion. The tutorial is structured accordingly into specific sections. The audience is expected to gain a solid understanding of the main challenges involved in designing and implementing regulatory-compliant data management systems (see below for a detailed structural outline).

Pre-requisites: none. The tutorial is designed to only require extremely broad knowledge of computer science. Moreover, we will accommodate a non-CS audience (e.g., practitioners in financial or health-care domains) by including intuitive, easy to follow analogies. A general introduction to data security will be included as part of the tutorial.

Biography of Speakers.
Radu Sion is an assistant professor of Computer Sciences in Stony Brook University and the director of the Network Security and Applied Cryptography Laboratory. His research focuses on data security and information assurance mechanisms. Collaborators and funding partners include Motorola Labs, IBM Research, the Center of Excellence in Wireless and Information Technology CEWIT, the Stony Brook Office for the Vice-President for Research and the National Science Foundation. Dr. Sion is serving on the organizing committee of numerous data management and information security conferences, such as SIGMOD, ICDE, ICDCS, CCS, Financial Cryptography, USENIX Security a.o.
Marianne Winslett received her PhD in Computer Science from Stanford University in 1987. She has been an assistant, associate, full, and adjunct professor in the Department of Computer Science at the University of Illinois. Her research interests are in databases and related areas, especially security in open systems and parallel I/O for high-performance scientific computation. She received a Presidential Young Investigator Award from the National Science Foundation in 1989 and Xerox Awards for Faculty Research in 1990 and 1997. She is currently on the editorial board of ACM Transactions on Database Systems and is a former editor for IEEE Transactions on Knowledge and Data Engineering and the vice-chair of ACM SIGMOD.

Overview. Over 10,000 regulations govern the management of documents in the US alone [27], in financial, life sciences, health-care industries and the government. These regulations impose a wide range of regulatory policies, ranging from information life-cycle management (e.g., mandatory data retention and deletion) to audit trails and storage confidentiality. Examples include the Gramm-Leach-Biley Act [20], Health Insurance Portability and Accountability Act [32], Federal Information Security Management Act [33], Sarbanes-Oxley Act [34], Securities and Exchange Commission rule 17a-4 [31], Department of Defense Records Management Program under directive 5015.2 [28], Food and Drug Administration 21 CFR Part 11 [30], and the Family Educational Rights and Privacy Act [29]. While each regulation has its own unique characteristics, certain assurance features are broadly mandated:

- **Guaranteed Data Retention.** To address this requirement, the goal of compliant data management is to support WORM semantics: once written, data cannot be undetectably altered or deleted before the end of their regulation-mandated life span, even with physical access to its hosting server.

- **Quick Lookup and Queries.** In light of the massive amounts of data subject to compliance regulations, the regulatory requirement for quick data retrieval can only be met by accessing the data through indexing structures. Such indexes must be efficient enough to support a target throughput, and must be secured against insiders who wish to remove or alter compromising information before the end of its mandated lifespan.

- **Secure Deletion.** Once data has reached the end of its lifespan, it can (and in some cases must) be deleted. Deleted records should not be recoverable even with unrestricted access to the underlying medium; moreover, after data is deleted, no hints of its existence should remain on the server, even in the indexes. We use the term *secure deletion* to describe this combination of features.

- **Compliant Data Migration.** Retention periods are measured in years. For example, national intelligence information, educational records, and certain health records have retention periods of over 20 years. To address this requirement, compliant data management needs *data migration* mechanisms that allow information to be transferred from obsolete to new storage media while preserving its associated security guarantees.

- **Litigation Holds.** Even if a data record has reached the end of its lifespan, it should remain fully accessible if it is the subject of current litigation.

- **Data Confidentiality.** Only authorized parties should have access to compliance data. To meet this requirement, access should be restricted even if the storage media are stolen, and access to meta-data such as indexes should also be limited.

In addition to these features, a common thread running through many of these regulations is the perception of powerful insiders as the primary adversary. These adversaries have superuser powers coupled with full access to the storage system hardware. This corresponds to the perception that much recent corporate malfeasance has been at the behest of CEOs and CFOs, who also have the power to order the destruction or alteration of incriminating records. Since the visible alteration or destruction of records is tantamount to an admission of guilt in the context of litigation, a successful adversary must perform their misdeeds *undetectably*.

Major vendors have responded by offering compliance storage and WORM products (which we will survey in this tutorial), including IBM [14], HP [10], EMC [6], Hitachi Data Systems [9], Zantaz [35], StorageTek [24], Sun Microsystem [25], Network Appliance [21] and Quantum Inc. [22].

Unfortunately, these products and research prototypes do not fully satisfy *any* of the requirements listed above. Most importantly, they are fundamentally vulnerable to faulty behavior or insider attack, because they rely on enforcement primitives such as software and/or simple hardware device-hosted on/off switches. As just one example, consider a recent IBM US patent for a disk-based WORM system whose drives selectively and permanently disable their write mode by using programmable read only memory (PROM) circuitry. "One method of use employs selectively blowing a PROM fuse in the arm electronics of the hard disk drive to prevent further writing to a corresponding disk surface in the hard disk drive. A second method of use employs selectively blowing a PROM fuse in processor-accessible memory, to prevent further writing to a section of logical block addresses (LBAs) corresponding to a respective set of data sectors in the hard disk drive" [15].

This method does not provide strong WORM guarantees. Using off-the-shelf resources, an insider adversary can penetrate storage medium enclosures to access the underlying data (and any flash-based checksum storage). She can then surreptitiously replace a device by copying an illicitly modified version of the stored data onto a identical replacement unit. Maintaining integrity-authenticating checksums at device or software level does not prevent this attack, due to the lack of tamper-resistant storage for keying material. By accessing integrity checksum keys, the adversary can construct a new matching checksum for the modified data on the replacement device, thus remaining undetected. Even if we add tamper-resistant storage for keying material [1], a superuser is likely to have access to keys

while they are in active use: achieving reasonable data throughputs will require integrity keys to be available in main memory for the main (untrusted) run-time data processing components.

Achieving a secure, cost-effective, and efficient design when the insider is the adversary is extremely challenging. To defend against insiders, we need processing components that are both tamper-resistant and *active*, such as general-purpose trust-worthy hardware. By offering the ability to run logic within a secured enclosure, such devices allow fundamentally new paradigms of trust. Trust chains spanning untrusted and possibly hostile environments can now be built by deploying secure tamper-resistant hardware at the storage components' site. The trusted hardware can run certified logic; close proximity to data coupled with tamper-resistance guarantees allow an optimal balancing and partial decoupling of the efficiency/security trade-off. Assurances can now be both efficient and secure.

However, trusted hardware devices are not a panacea. Their practical limitations pose a set of significant challenges in achieving sound regulatory-compliance assurances. Specifically, heat dissipation concerns under tamper-resistant requirements limit the maximum allowable spatial gate-density. As a result, general-purpose secure coprocessors (SCPUs) are often significantly constrained in both computation ability and memory capacity, being up to one order of magnitude slower than host CPUs.

Such constraints mandate careful consideration in achieving efficient protocols. Direct implementations of the full processing logic *inside* the SCPU are bound to fail in practice due to lack of performance. The server's main CPUs will remain starkly under-utilized and the entire cost-proposition of having fast untrusted main CPUs and expensive slower secured CPUs will be defeated. Efficient protocols need to access the secure hardware sparsely, asynchronously from the main data flow.

We will explore these challenges and show how to leverage this new paradigm to achieve strong regulatory compliance for storage systems in realistic adversarial settings. Specifically, we will discuss achieving secure designs offering guaranteed record retention and deletion, quick lookup, and compliant migration, together with support for litigation holds and several key aspects of data confidentiality.

Structure. The tutorial will discuss three distinct compliance assurances, namely: (1) record-level Write-Once Read-Many (WORM) assurances, (2) trust-worthy indexing, and (3) secure deletion. Specifically, in (1) we discuss existing tape-, optical-, and disk-based WORM mechanisms. We then explore the main drawbacks and vulnerabilities of such solutions and discuss achieving secure designs [12]. In (2) we analyze existing mechanisms for secure indexing for various media [2, 5, 23, 26] and their vulnerabilities in the considered adversarial model [17]. We then discuss indexing mechanisms impervious to such attacks, specifically Generalized Hash Trees (GHT) [11, 36], supporting exact-match lookups of records based on attribute values, most suitable for use with structured data. We will then explore also unstructured domains and discuss thrust-worthy keyword search [17], while surveying main results in the established area of thrust-worthy indexing in outsourced data [8], including authenticated dictionaries [3,7], and query correctness mechanisms [4,13,16,19]. We will discuss indexing mechanisms with secure hardware awareness and explore how such mechanisms can interact with the underlying WORM layer, discussed in (1). In (3), we will analyze the fact that, upon data record disposal (as mandated by numerous regulations [28,32]) just erasing records from WORM is insufficient, as their contents (or artifacts thereof) may be recoverable from indexes. We will then explore the main challenges associated with this task as well as a set of emerging solutions [18], including logical deletion methods, history independent data structures and trusted hardware - aware mechanisms. Finally the tutorial also includes a brief general-audience introduction to main data security primitives.

References

[1] Trusted Platform Module (TPM) Specifications. Online at `https://www.trustedcomputinggroup.org/specs/TPM`, 2006.

[2] Bruno Becker, Stephan Gschwind, Thomas Ohler, Bernhard Seeger, and Peter Widmayer. An Asymptotically Optimal Multiversion B-tree. *The VLDB Journal*, 5:264–275, 1996.

[3] Kaouthar Blibech and Alban Gabillon. Chronos: an authenticated dictionary based on skip lists for timestamping systems. In *SWS '05: Proceedings of the 2005 workshop on Secure web services*, pages 84–90, New York, NY, USA, 2005. ACM Press.

[4] Premkumar T. Devanbu, Michael Gertz, Chip Martel, and Stuart G. Stubblebine. Authentic third-party data publication. In *IFIP Workshop on Database Security*, pages 101–112, 2000.

[5] Malcolm C. Easton. Key-Sequence Data Sets on Indelible Storage. *IBM Journal of Research and Development*, May 1986.

[6] EMC. Centera Compliance Edition Plus. Online at `http://www.emc.com/centera/` and `http://www.mosaictech.com/pdf_docs/emc/centera.pdf`, 2007.

[7] M. Goodrich, R. Tamassia, and A. Schwerin. Implementation of an authenticated dictionary with skip lists and commutative hashing, 2001.

[8] Hakan Hacigumus, Balakrishna R. Iyer, and Sharad Mehrotra. Providing database as a service. In *ICDE*, 2002.

[9] Hitachi Data Systems. The Message Archive for Compliance Solution, Data Retention Software Utility. Online at `http://www.hds.com/solutions/data_life_cycle_archiving/achievingregcompliance.html`, 2007.

[10] HP. WORM Data Protection Solutions. Online at `http://h18006.www1.hp.com/products/storageworks/wormdps/index.html`, 2007.

[11] W. Hsu and S. Ong. Fossilization: A Process for Establishing Truly Trustworthy Records. *IBM Research Report*, (10331), 2004.

[12] Windsor Hsu, Lan Huang, and Shauchi Ong. Content Immutable Storage: Truly Trustworthy and Cost-Effective Storage for Electronic Records. Research Report RJ 10332. Technical report, 2004.

[13] HweeHwa Pang and Arpit Jain and Krithi Ramamritham and Kian-Lee Tan. Verifying Completeness of Relational Query Results in Data Publishing. In *Proceedings of ACM SIGMOD*, 2005.

[14] IBM Corp. IBM TotalStorage Enterprise. Online at `http://www-03.ibm.com/servers/storage/`, 2007.

[15] IBM Corporation and Daniel James Winarski and Kamal Emile Dimitri. United States Patent 6879454: Write-Once Read-Many Hard Disk Drive, 2005.

[16] Maithili Narasimha and Gene Tsudik. Authentication of Outsourced Databases using Signature Aggregation and Chaining. In *Proceedings of DASFAA*, 2006.

[17] Soumyadeb Mitra, Windsor W. Hsu, and Marianne Winslett. Trustworthy Keyword Search for Regulatory-Compliant Records Retention. In *Proceedings of VLDB*, 2006.

[18] Soumyadeb Mitra and Marianne Winslett. Secure Deletion from Inverted Indexes on Compliance Storage. In *Proceedings of the StorageSS Workshop*, 2006.

[19] E. Mykletun, M. Narasimha, and G. Tsudik. Authentication and integrity in outsourced databases. In *ISOC Symposium on Network and Distributed Systems Security NDSS*, 2004.

[20] National Association of Insurance Commissioners. Graham-Leach-Bliley Act, 1999. `www.naic.org/GLBA`.

[21] Network Appliance Inc. SnapLock Compliance and SnapLock Enterprise Software. Online at `http://www.netapp.com/products/software/snaplock.html`, 2007.

[22] Quantum Inc. DLTSage Write Once Read Many Solution. Online at `http://www.quantum.com/Products/TapeDrives/DLT/SDLT600/DLTIce/Index.aspx` and `http://www.quantum.com/pdf/DS00232.pdf`, 2007.

[23] Peter Rathmann. Dynamic Data Structures on Optical Disks. In *1st International Conference on Data Engineering*, 1984.

[24] StorageTek Inc. VolSafe secure tape-based write once read many (WORM) storage solution. Online at `http://www.storagetek.com/`, 2007.

[25] Sun Microsystems. Sun StorageTek Compliance Archiving Software. Online at `http://www.sun.com/storagetek/management_software/data_protection/compliance_archiving/`, 2007.

[26] T. Krijnen and L. G. L. T. Meertens. Making B-Trees Work for B.IW 219/83. The Mathematical Centre, Amsterdam, The Netherlands, 1983.

[27] The Enterprise Storage Group. Compliance: The effect on information management and the storage industry. Online at `http://www.enterprisestoragegroup.com/`, 2003.

[28] The U.S. Department of Defense. Directive 5015.2: DOD Records Management Program. Online at `http://www.dtic.mil/whs/directives/corres/pdf/50152std_061902/p50152s.pdf`, 2002.

[29] The U.S. Department of Education. 20 U.S.C. 1232g; 34 CFR Part 99:The Family Educational Rights and Privacy Act (FERPA). Online at `http://www.ed.gov/policy/gen/guid/fpco/ferpa`, 1974.

[30] The U.S. Department of Health and Human Services Food and Drug Administration. 21 CFR Part 11: Electronic Records and Signature Regulations. Online at `http://www.fda.gov/ora/compliance_ref/part11/FRs/background/pt11finr.pdf`, 1997.

[31] The U.S. Securities and Exchange Commission. Rule 17a-3&4, 17 CFR Part 240: Electronic Storage of Broker-Dealer Records. Online at `http://edocket.access.gpo.gov/cfr_2002/aprqtr/17cfr240.17a-4.htm`, 2003.

[32] U.S. Dept. of Health & Human Services. The Health Insurance Portability and Accountability Act (HIPAA), 1996. `www.cms.gov/hipaa`.

[33] U.S. Public Law 107-347. The E-Government Act, 2002.

[34] U.S. Public Law No. 107-204, 116 Stat. 745. The Public Company Accounting Reform and Investor Protection Act, 2002.

[35] Zantaz Inc. The ZANTAZ Digital Safe Product Family. Online at `http://www.zantaz.com/`, 2007.

[36] Qingbo Zhu and Windsor W. Hsu. Fossilized index: the linchpin of trustworthy non-alterable electronic records. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 395–406, New York, NY, USA, 2005. ACM Press.