

Old Dominion University ODU Digital Commons

VMASC Publications

Virginia Modeling, Analysis & Simulation Center

2018

A Reliable Data Provenance and Privacy Preservation Architecture for Business-Driven Cyber-Physical Systems Using Blockchain

Xueping Liang

Sachin Shetty

Old Dominion University, sshetty@odu.edu


Deepak K. Tosh

Juan Zhao

Danyi Li

See next page for additional authors

Follow this and additional works at: https://digitalcommons.odu.edu/vmasc_pubs

 Part of the [Computer and Systems Architecture Commons](#), [Digital Communications and Networking Commons](#), and the [Information Security Commons](#)

Repository Citation

Liang, Xueping; Shetty, Sachin; Tosh, Deepak K.; Zhao, Juan; Li, Danyi; and Liu, Jihong, "A Reliable Data Provenance and Privacy Preservation Architecture for Business-Driven Cyber-Physical Systems Using Blockchain" (2018). *VMASC Publications*. 39.
https://digitalcommons.odu.edu/vmasc_pubs/39

Original Publication Citation

Liang, X., Shetty, S., Tosh, D., K., Zhao, J., Li, D., & Liu, J. (2018). A reliable data provenance and privacy preservation architecture for business-driven cyber-physical systems using blockchain. *International Journal of Information Security and Privacy*, 12(4), 68-81.
doi:10.4018/IJISP.2018100105

This Article is brought to you for free and open access by the Virginia Modeling, Analysis & Simulation Center at ODU Digital Commons. It has been accepted for inclusion in VMASC Publications by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

Authors

Xueping Liang, Sachin Shetty, Deepak K. Tosh, Juan Zhao, Danyi Li, and Jihong Liu

A Reliable Data Provenance and Privacy Preservation Architecture for Business-Driven Cyber-Physical Systems Using Blockchain

Xueping Liang, Institute of Information Engineering, Chinese Academy of Sciences, China & School of Cyber Security, University of Chinese Academy of Sciences, China & Old Dominion University, USA

Sachin Shetty, Old Dominion University, USA

Deepak K. Tosh, Department of Computer Science, University of Texas at El Paso, USA

Juan Zhao, Tennessee State University, USA

Danyi Li, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

Jihong Liu, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

ABSTRACT

Cyber-physical systems (CPS) including power systems, transportation, industrial control systems, etc. support both advanced control and communications among system components. Frequent data operations could introduce random failures and malicious attacks or even bring down the whole system. The dependency on a central authority increases the risk of single point of failure. To establish an immutable data provenance scheme for CPS, the authors adopt blockchain and propose a decentralized architecture to assure data integrity. In business-driven CPS, end users are required to share their personal information with multiple third parties. To prevent data leakage and preserve user privacy, the authors isolate and feed different information retrieval requests using tokens specifically generated for each type of request. Providing both traceability of data operations, and unlinkability of end user activities, a robust blockchain-based CPS is prototyped. Evaluation indicates the architecture is capable of assured data provenance validation and user privacy preservation at a low overhead.

KEYWORDS

Blockchain, Cyber Physical Systems, Privacy Preservation, Reliability, Survivability

1. INTRODUCTION

Typical Cyber-Physical Systems (CPS) connect physical infrastructure to integrated computing devices and data storage facilities, with a combination of computation, communication and control. CPS are increasingly deployed in military, electrical and medical systems, as well as logistics or industrial production processes. However, due to system instability and existing vulnerabilities in the heterogeneous subsystems, the control system may be faced with random system failures or

DOI: 10.4018/IJISP.2018100105

Copyright © 2018, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

even malicious cyber attacks. Meanwhile, end users of the CPS could be encountered with potential privacy concerns. Recent research (Han, Shah, Luk, & Perrig, 2007) indicates that the collected data of indoor humidity could reveal user activities, thus becoming a data leakage point, which could significantly raise privacy concerns for end users. It is also reported (Grid, 2010) that the smart meter can collect data from Home Area Network (HAN) to reveal home smart appliances, making end user privacy at high risks.

The distributed ledger which is being used by cryptocurrencies like Bitcoin (Nakamoto, 2008) and (Wood, 2014), is a decentralized architecture running among distributed and untrusted network nodes with cryptography algorithm and consensus mechanism, providing traceability and data protection for each transaction witnessed by participating nodes. Blockchain is one implementation of distributed ledger where a chain of blocks are generated from transactions between nodes. The adoption of blockchain in CPS is rarely studied but is quite promising. Due to the decentralized architecture of blockchain and the removal of trust among distributed nodes, the robustness of CPS can be improved with the redundancy capability achieved by the distributed copies maintained by every single node. Blockchain based data provenance is proposed (Liang et al., 2017) to preserve the integrity of data generated from communication and control procedures, with the capability to defend against deception attacks (Shirey, 2007).

According to the framework for CPS (Griffor, Greer, Wollman, & Burns, 2017) issued by the US National Institute of Standards and Technology (NIST), cybersecurity for CPS must address how a system can continue to function correctly when under attack, provide mechanisms that support fault-tolerance with mission- or business-driven priorities, and enable the system to fail-safe. Those requirements indicate the urgency of developing a survivable and reliable CPS. Modern power grid system, namely smart grid, is proposed in many countries to realize a reliable, scalable, manageable, extensible, secure, interoperable and cost-effective electric cyber-physical infrastructure (Khaitan & McCalley, 2013). A typical smart grid system consists of power generation, transmission, distribution and consumption domains and we aim to address the reliability and privacy preservation issues in these domains. Specifically, we focus on business-driven situations and adopt blockchain to design a reliable power delivery (in generation, transmission and distribution domain) provenance and privacy preserving user interface (in consumption domain especially in HAN), as a step towards a fully survivable architecture. However, to facilitate CPS with blockchain architecture, several critical issues need be solved. We identify the concerns regarding the integration of blockchain with CPS and then propose a solution to fulfill the objectives of reliability and privacy protection. In this paper, we use power system as a sample CPS to illustrate how blockchain can be utilized in such environment. Most importantly, we implement a blockchain based power supply chain data provenance architecture for power delivery, and privacy protection scheme to prevent sensitive personal data leakage. Performance evaluation indicates that the proposed architecture achieves the above objectives at a low overhead with security guarantee.

1.1. Contributions

We summarize our contributions as follows:

1. We initiate the research of identifying the challenges of blockchain integration into CPS, and designing a reliable data provenance architecture with preserved privacy for business-driven CPS based on blockchain. To the best of our knowledge, this is the first work towards designing a reliable and privacy preserving power system provenance architecture;
2. We design a reliable data provenance architecture to enable traceability for data processing and operations in CPS. In addition, we transform data provenance into immutable and permanent blockchain transactions efficiently using hash algorithms, to protect data integrity and resist tampering. The integrity of the data and the trusted timestamping is guaranteed by the consensus mechanism used in the block mining process;

3. We design a privacy preserving scheme with unlinkability between user activities within the business domain by leveraging attribute based verification to minimize the disclosure of personal information and verify anonymous identity. The evaluation indicates that the architecture brings low overhead.

The rest of the paper is organized as follows. Section 2 introduces the advantages blockchain can bring, challenges to integrate blockchain into CPS, as well as attribute based verification and the threat model. Section 3 introduces the overall system design and two system processes, including data provenance along the data transmission logic and the privacy preserving customer interface within HAN. We give a security analysis and performance evaluation in Section 4. Section 5 presents related work and concludes the paper.

2. BUILDING BLOCKS

The proposed system is based on two modules, one is the adoption of the blockchain technology for data provenance in CPS architecture, the other is the attribute-based token verification scheme for privacy-preserving user authentication. Blockchain is a new technology which uses a distributed public ledger to record transactions and facilitate trusted delivery of transactions across a distributed network without involvement of centralized authority or intermediaries. Blockchains have the following advantage over centralized databases: (1) Ability to directly share databases across diverse boundaries of trust in situations where it is difficult to identify a trusted centralized arbitrator to enforce constraints of proof of authorization and validity. Blockchain transactions leverage self-contained proofs of validity and authorization based on a verification process enforced by multiple validating nodes and a consensus mechanism that ensures synchronization. (2) Ability to provide robustness in an economical fashion without the need for expensive infrastructure for replication and disaster recovery.

Attribute-based verification is widely adopted for cloud services where service providers are not trusted, and users are concerned that their sensitive data could be accessed by unauthorized parties, because only the necessary set of attributes are captured for authentication purposes. This brings benefits of reduced communication cost, and most importantly, provides a privacy-friendly interface between the service provider and the requesting user. Moreover, the attribute-based verification is flexible regarding to the multiple tokens generated for different application scenarios and different level of sensitive data required. This section summarizes blockchain capabilities in securing CPS, and points out the challenges to integrate blockchain which are addressed in our system design in Section 3. A formal explanation of the attribute-based verification is presented for further adoption during the user authentication phase. Threat model specifies the system adversarial capabilities and security issues we intend to solve.

2.1. What Blockchain can do in CPS

Blockchains provide built-in technical mechanisms to handle tasks which would otherwise require complex institutional processes. Nodes in a blockchain automatically self-configure, connect and sync with each other in a peer-to-peer fashion. The feature of built-in redundancy avoids the need for closely monitoring and provides the ability to tolerate multiple communication link failures. External users can broadcast transactions to any node, and it is ensured that disconnected nodes will be caught up on missed transactions. The detailed capabilities that enable blockchain as a compelling technology to serve as solutions to some critical security issues in CPS are discussed in the following:

1. Each mined block in the blockchain contains a list of transactions which are then hashed as a Merkle root. This attribute can be used to handle sensitive information and maintain the integrity of data. The hash of the previous block is involved in the generation of the current block hash,

making the block capable to prevent and detect any form of tampering. Any manipulation of block data is impossible without being detected. This immutability of blockchain benefits CPS by way of providing the reliability of data provenance and implicit linkability between lists of transactions;

2. Based on the timestamping function and the chain based architecture, mathematical certainty over the provenance and tracking of every component could be assured in the system. Any record anchored on the blockchain cannot be modified and thus the recorded event sequencing is preserved. The validation of the data integrity is provable by traversing the blockchain state which is maintained by distributed nodes;
3. Computation and data operation process can be deployed among multiple untrusted parties, but the computation and data operation results are trusted. This can effectively reduce the risks of single point of failures and the DDoS attack (Douligeris & Mitrokotsa, 2004).

2.2. Challenges for Blockchain Integration in CPS

1. Data volume in CPS could be a heavy load for the processing unit. To deal with multiple concurrent data streams and operation records, it is required that the blockchain should manage frequent data records and handle large data sets during a limited time span. However, blockchain based architecture requires a specific time cost for both block mining and consensus scheme;
2. Data from different CPS domains are differently formatted during creation, exchange and storage. It is even more complex to process, understand and manage data that is transmitted across boundaries of several subsystems. The data interoperability poses a challenging task for blockchain integration, especially for the design logic of blockchain program such as smart contract in Ethereum (Wood, 2014) or chaincode by Hyperledger (Cachin, 2016);
3. The public nature of blockchain architecture is a huge challenge for dealing with sensitive data especially when it comes to the market and customer domain. Security measures are required to make sure that only authorized access is allowed, and user privacy is not at risk. Some blockchain implementations such as Hyperledger support channel scheme which provides an isolated communication method but still adds to the risks of data leakage during channel participation.

2.3. Attribute Based Verification

To protect user privacy within the business domain of the CPS, it is crucial that the user identity is not easily exposed and there are no linkability between normal user activities. For verification, real world identities such as a passport or a driver's license can expose sensitive information of a person, such as name, date of birth and address, resulting in drawbacks towards privacy risks. Not only the user's real-world activities could be tracked, but also online activities are exposed. To authenticate user in a more privacy friendly way, attribute-based data sharing (Yu, Wang, Ren, & Lou, 2010) is proposed so that user can disclose only necessary information to service providers without revealing the real identity and other personal information. For example, the customer would only reveal billing status and power load demand instead of billing statement or detailed power usage, to indicate personal information and request to a certain service provider.

Suppose a user has an attribute set, noted as $A_{user} = a_1, a_2, \dots, a_n$. There are service providers, noted as $SP = sp_1, sp_2, \dots, sp_n$, requiring user authentication so that only users with specific attribute(s) could be authenticated. For simplicity, we suppose there is one attribute requested by each service provider and the required attribute for sp_i is noted as $A_{sp_i} = f(a_i)$, where $f(a_i)$ represents an interpretation of the original attribute a_i . To generate a presentation token to the service provider that requests user attributes, a trusted issuer which is responsible for maintaining the user information database will first issue a credential by signing user attributes and user private key a . Proof of statement will then be used to generate and present a token containing the selected attributes to the service

provider and prove the possession of required attributes. Such proof of statement, such as Zero Knowledge Proof (Goldreich & Oren, 1994), will be adopted for selective attribute disclosure and minimizing the information a service provider can access or obtain.

2.4. Threat Model

We consider a power system where there are large quantities of distributed nodes which are utilized for computing, communication and controlling during the power supply delivery to end users. Those devices could be controlled by insiders to perform designed behaviors, but insiders are not always trusted. Consequently, the integrity and availability of data storage could be compromised due to unauthorized access and malicious manipulation, making data provenance a challenging task.

The power supply chain and market participants include billing, investigation, home sensor/device providers, supply chain coordinators and other power related service providers. In the power system close to the HAN, end users or customers are exposed to interactions with these participants. Different levels of personal information are retrieved during the whole process which indicates potential data leakage points and privacy threats for end users or customers. Related third parties like billing and insurance companies or even the service providers are curious about personal information such as power usage and home activities while they should only ask for the minimized necessary information they need. The verifier from the service provider may attempt to infer and link user activities. Also, the service provider and verifier could even collude with each other.

We assume that the participating blockchain nodes are not trusted but perform the assigned tasks of collecting and computing the hash of data records, thus reliably achieve agreement among the untrusted nodes. Moreover, the communication channels between users and the service provider is secure and each user credential and token as well as private keys are well protected.

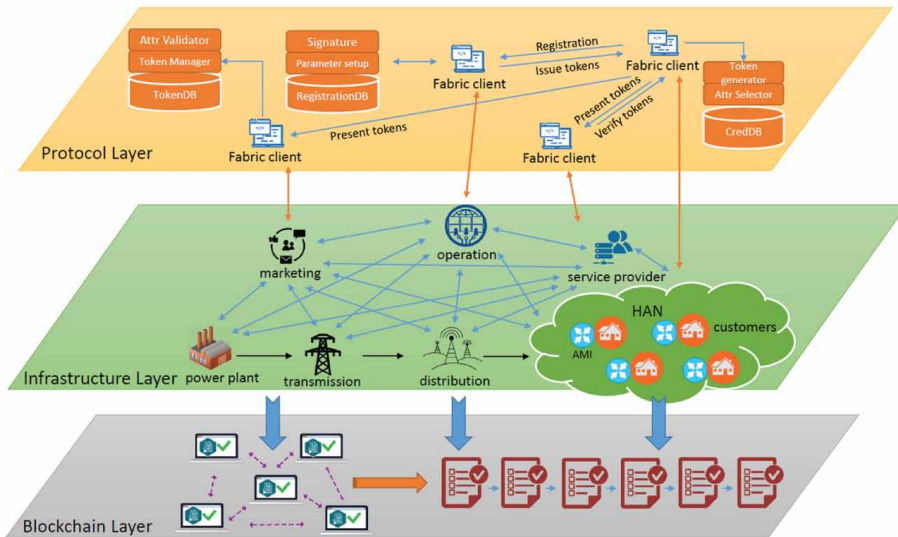
3. SYSTEM DESIGN

To design a blockchain facilitated smart grid system, we adopt Hyperledger Fabric (Cachin, 2016), which is an open source blockchain platform with the capability of chaincode deployment, to provide the tamper-resistant and traceable data provenance functionality. Especially, for the privacy preserving customer interface with operation components and service providers, we disclose customer attributes in a selective and unlinkable manner. The architecture is shown in Figure 1.

Three layers are designed in the blockchain-based smart grid architecture, namely protocol layer, infrastructure layer and blockchain layer. System entities include power plant, transmission, distribution sectors, and HAN, as well as marketing, operation, and service provider domains, as illustrated in the infrastructure layer in Figure 1. The power plant, transmission, distribution sectors, and marketing, operation, and service provider domains are responsible for power generation and delivery, while HAN serves as an interface between AMI (Advanced Metering Infrastructure) devices and customers. Each component in the infrastructure layer and its activities during power generation and delivery are captured and anchored to the blockchain layer for provenance purposes. The blockchain layer comprises distributed computing nodes and copies of all blocks of transactions. The decentralized architecture protects provenance data integrity in a reliable manner. To enhance the scalability of blockchain integration with the system, the tree-based data batching method is adopted for concurrent data input.

In protocol layer, there are Fabric clients, representing marketing, operation, service providers and customers, responsible for processing power transactions. For one thing, the Fabric client is responsible for retrieve all the provenance data across the entire infrastructure. For another, the Fabric client is also responsible for the verification and privacy protection of the customers. Different roles are assigned in the protocol layer for token-based verification, such as token issuer, prover and verifier. The customer (prover) tokens are generated using different attributes required by the service providers (verifier), by requesting to identity servers (issuer) in the operation domain. The prover only

Figure 1. A blockchain-based smart grid architecture with reliable data provenance and privacy preservation



selects necessary attributes to generate specific tokens for different services, preserving customer privacy with minimum exposure of sensitive information. Meanwhile, different verifiers cannot link tokens presented by the same prover. By properly attribute selection, even the same verifier cannot link different tokens presented by the same prover.

To illustrate the design of the architecture in detail, we consider two major tasks of the system in the following sections, including data provenance for power delivery and the customer interface in HAN. By integrating blockchain, the system can achieve the decentralization and reliability of the data provenance, and the privacy preservation of the customer interface with auditable activity logs.

3.1. Reliable Data Provenance for Power Delivery

Smart grid system components include power generation, transmission, distribution and customer, while there are three domains including market, operation and service provider, all of which host a Fabric client to support the provenance generation. Each Fabric client relies on a peer node to communicate with neighboring peers. The data provenance procedures include provenance data retrieval, provenance data uploading and integrity validation. Data operations along the power delivery process are monitored and retrieved by the data provenance generator and then uploaded to the Fabric platform. By consensus scheme and block mining, the provenance data will be patched and anchored in the ledger as a transaction. The ledger itself is a chained structure so each block is integrity protected and any modification would result in all the following blocks being affected. Detailed process is as follows.

3.1.1. Phase One: Provenance Data Retrieval

The data operator is monitored when performing control commands and communications. With increasing development of sensors, controllers and integrated tools, it is easier to monitor complex grid operations uninterruptedly. Transactions include maintaining and controlling of the grid, dynamic load, fault and outage management. An example operation record is retrieved in JSON format with metadata as follows:


```
{  
  "nodeID": "1001",  
  "operation type": "reclose",  
  "affecteddevice": "devicegroup1",  
  "timestamp": "1475679929",  
  "location": "39.858425, 116.287148",  
  "device type": "switch",  
  "load": "[]",  
  "priority": "30",  
  "status": "close"  
}
```

3.1.2. Phase Two: Provenance Data Uploading

To protect the operation data and improve the process efficiency of concurrent operation records, the JSON formatted data is hashed to a string with fixed length. This resolves the incompatibility issue between subsystems across the infrastructure. By adopting a Merkle Tree (Szydlo, 2004) structure, the data uploading process is transformed to a data batching and tree construction task, with a computation complexity of $O(n)$. Eventually, the tree root is uploaded to the blockchain network and recognized as a transaction, to form a block together with other transactions. The block is validated by distributed peers and will be appended to the existing blockchain after a consensus is reached. The uploaded data record will be updated when a proof of integrity is generated. The proof of integrity is generated with the metadata from the specific block where the tree root is anchored, following the Chainpoint 2.0 standard (Vaughan, 2016).

3.1.3. Phase Three: Provenance Data Integrity Validation

On receiving the proof of integrity, it can be validated by querying the blockchain network. If the proof is consistent with the block information it points to, the data integrity and the timestamping is then trusted. The validation process is designed to run periodically so that cyber incidents can be identified in a timely manner. By examining the proof of integrity, the provenance data is traceable throughout the entire life cycle of a single operation event or a piece of data operation unit. The provenance data can be used for further investigation to detect malicious behaviors and intrusions.

With all three phases above, we achieve a reliable data provenance architecture with scalability. The data anchored to the blockchain network is verified and integrity protected continuously by distributed nodes. The scalability is achieved by batching input data sets into tree-based data structure using the algorithm, where each tree leaf node represents a data record while the tree root node is anchored to the blockchain network as a transaction. The hash algorithm helps to format the data inputs and achieve a consistent interface between the blockchain nodes.

3.2. Privacy Preserving Customer Interface

In HAN, home appliances interact with service providers or other AMI devices for power supply and data exchange (Liu, Xiao, Li, Liang, & Chen, 2012). The following processes illustrate the proposed architecture for preserving consumer privacy using selective attribute disclosure. Three entities, such as issuer, customer and the service provider, are involved in the customer interface. The issuer belongs to the operation domain and is responsible for maintaining the customer database, generating signatures on customer information, and initializing system parameters for following data operations. The service provider will propose policies for each service, ask the customers for verification and provide services to authenticated and qualified customers. The attribute selector is used to check whether the policies are satisfied so that the customers are qualified. The token database is maintained by the marketing domain and service providers. The customer maintains a database for credentials which are received

from the issuer, chooses attributes needed based on the policies made by the service providers and generates multiple tokens accordingly. The detailed procedures include system setup, registration, credential issuance, token presentation and verification.

3.2.1. Phase One: System Setup

To issue credentials for the customer, the issuer from the operation domain first generates cryptographic parameters. Based on Strong RSA assumption (Cramer & Shoup, 2000), given a special RSA modulus $n = pq$, where p and q are two prime numbers and kept as the secret key of the issuer, SK_I . A set of random numbers are generated, $R_0, R_1, \dots, R_l, Z, S \in QR_n$, where QR_n is a quadratic residue. The other parameters such as Hash functions (Ramakrishna & Zobel, 1997) and attribute length are also predefined in this stage.

3.2.2. Phase Two: Registration

To subscribe to the power supply service, customers need to register first to obtain a user account by communicating with the issuer in a reliable manner. This unspecified manner is out of the scope of this paper. The customer provides a list of attributes $A_{customer} = a_0, a_1, \dots, a_l$ and generates a private key a as well. For credential requests, the attribute list and a nonce is sent to the issuer.

3.2.3. Phase Three: Credential Issuance

On receiving the credential request from the customer, the issuer signs on a commitment

$R = R_0^a \prod_{i=1}^l R_i^{a_i} \pmod{n}$, with the signature $A = (Z / S^v R)^{1/e}$, where e is chosen by the issuer. When

the customer receives the signature, the customer will then validate the signature using a new v which is only known to the customer. Finally, the signature will be updated and stored in the customer credential database. To preserve activity unlinkability, the customer can request a list of credentials with different attribute sets signed by the issuer.

3.2.4. Phase Four: Token Presentation

This step is for the customer to prove to the service provider sp_i the truth that the customer indeed owns the combination of the necessary attributes to satisfy a given service requirement, namely R_s . A token will be generated by the customer depending on the specific R_s , including the common value list *common*, a challenge c to the nonce n_1 from the service provider and a s indicating the proof of possession of a set of attributes with specific values.

3.2.5. Phase Five: Token Verification

To validate the proof of possession, the service provider sp_i computes a t value list using the s value list and then the challenge verification $c_1 = H(\text{context} | \text{common} | t | n_1)$ using the Fiat-Shamir challenge scheme (Fiat & Shamir, 1986). If the challenge verification c_1 and c matches, then the token verification is successful.

3.2.6. Phase Six: Channel Setup

For each service provider sp_i , there is a Fabric channel for them to set up and join. Each channel represents a business scenario and can host one or more service providers to conduct transactions. To set up a channel, the service provider first deploys the chaincode, which is responsible for transaction processing, to the channel. Customers can choose a channel to transact with the service providers for

an isolated and single-purpose communication. For each channel, a customer can present a different token to the service provider so that tokens presented in different channels are not linkable. To participate in the channel, necessary attributes should be met. After the service provider sp_i checks the required attributed provided by the customer, the customer will then be able to join the channel.

3.2.7. Phase Seven: Transaction Processing

For customers requesting services from service provider sp_i , there are three types of operations provided by Fabric that a customer can perform, including instantiate, invoke and query. For the first interaction between the customer and the service provider sp_i , the chaincode needs to be instantiated by setting the initial value. Then changes can be made using the invoke method. Query methods is used for looking up a specific value. Other methods can be defined to satisfy different business requirements. The transactions are captured when chaincode methods are called. In this way, the business process is recorded in the subledger generated from the current channel, which will become part of the system ledger.

With all seven phases above, we achieve a privacy preserving customer interface with minimum information exposure. The data transmitted during the verification process is also reduced for better communication efficiency. The service provider can only obtain the necessary sensitive information from customers without linking historical transactions performed by the same customer.

4. SYSTEM EVALUATION

4.1. Security Analysis

The proposed system integrates blockchain in a smart grid system and provides reliable provenance for data generated and exchanged during system operations. The challenges for blockchain integration summarized in Section 2.2 are addressed in the system architecture. First, large data volume can be handled during provenance data uploading in a tree-based data structure, making the system scalable while improving the throughput. Second, data interoperability is addressed by standardizing the format for data collection. In our design, we adopt the JSON format which is sufficient for describing an event captured in the system while still apply to different system components and equipment. Third, the decentralized architecture is secured by the token-based access control mechanism so the potential data leakage and privacy concerns are mitigated, ensuring the system reliability and accountability.

The reliability of the data provenance relies on two critical data processing techniques. For one thing, each node accommodating a Fabric client can generate provenance data, which is hashed to a binary string. In this case, the hash algorithm maintains the data consistency between the original provenance data and the uploaded data. The protected provenance data prevents the data tampering from untrusted insiders and outsiders. For another, the real-time provenance data hash uploading leads to the integrity protection from the ledger with the proof of integrity and trusted timestamping. This is the second shelter against data modification and provides non-repudiation. The hashed data ensures that both data operations and operator's identity is recorded and integrity-protected, preventing user impersonation attacks. The public blockchain also represents the traceability for all the uploaded data records. On the other hand, each distributed node maintains an individual copy of the ledger state which is resistant against denial of service attack and single point of failure. Meanwhile, the provenance data generated by distributed nodes naturally preserves a logical order of the operation events, adding to the difficulty of maliciously manipulating a certain node or a set of nodes.

The privacy preservation of customer interface is maintained by the credential mechanism in a way that customers can select a minimum set of attributes to satisfy the requests from the service providers, reducing potential sources of personal data leakage and avoiding linkability between activities which further enhances privacy by removing a side channel. The unlinkability is achieved in that each token presented to the service provider only contains the proof of possession of a specific

attribute sets s , and during token verification, only this proof is verified without exposure of the exact attribute values. This preserves the unlinkability between customer transactions. The unlinkability between the issuer and verifier is also preserved so that even the issuer and the verified collude with each other, the privacy of the customer is still protected.

4.2. Performance Evaluation

To evaluate the performance for data provenance and privacy protection, we test the architecture on top of Hyperledger Fabric on a desktop with Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz (Skylake with 8MB cache) and 32GB RAM. The Identity Mixer technology (Bichsel et al., 2009) by IBM is used to implement attribute-based tokens. With different numbers of provenance data records, we build provenance data on each entity, following the basis of subject-based provenance. We use one peer to represent each participating node or device. For redundancy, two peers for each device could be adopted. The open source cloud platform Owncloud (Steven J. & Vaughan-Nichols, 2012), is used to receive provenance data and batching data into blockchain transactions.

First, the response time for the blockchain based data provenance architecture is evaluated against the data provenance without the blockchain client. Figure 2 shows the average response time to upload provenance data with increasing number of peers while Figure 3 shows the average response time to validate provenance data with increasing number of peers, both of which indicates that the Fabric based data provenance architecture stays steady and brings low overhead for data uploading and validation.

Compared to the previous experiment evaluation results of a blockchain-based provenance architecture for the Internet-of-Things (Kaku et al., 2017), the proposed system architecture has better performances considering a lower response time cost with different numbers of peers during provenance data generation and validation. This makes our system a practical solution for cyber physical systems, since there are not as frequent data operations as in the cloud computing environment and most system components are fixed during system establishment and system operations during system running.

For the generation of key and other cryptographic materials, we test performance with different key lengths, as is shown in Table 1. For performance considerations, the key length of 1024 bit as the system key length is used.

There are different stages for the generation of attribute based credentials and anonymized tokens. The time cost and required data transmission size for each stage is shown in Table 2, from which

Figure 2. Average response time for data provenance uploading with increasing number of peers

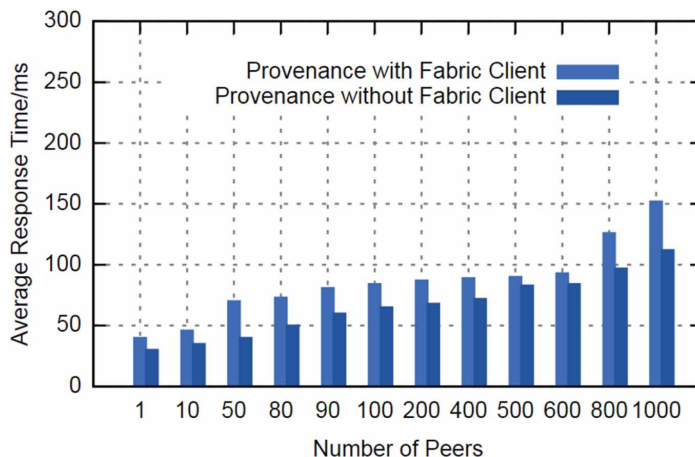


Figure 3. Average response time for data provenance validation with increasing number of peers

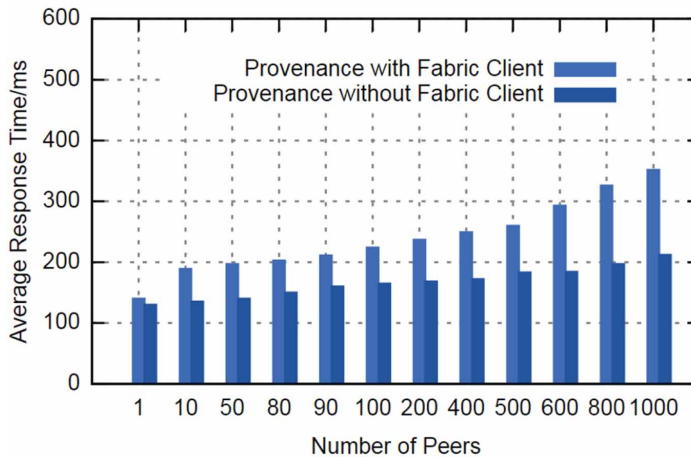


Table 1. Overhead in credential generation phase with different key lengths

	Key Length (bit)	Time Spent (ms)	Data Size (KB)
Experiment 1	512	134	3.04
Experiment 2	1024	334	3.49
Experiment 3	2048	29155	4.4

Table 2. On chain overhead in credential generation phase

	Phase	Time Spent (ms)	Data Size Received (KB)
Experiment 1	System Parameter Setup	290	3.49
Experiment 2	Issuer Parameter Setup	355	12.44
Experiment 3	Issuance Request from User	41	5.52
Experiment 4	Credential Issuance	475	5.64
Experiment 5	Token Presentation	260	5.21
Experiment 6	Token Verification	125	1.41

we can conclude that the average overhead for attributed based verification is acceptable, under the circumstances that the verification is launched only the first time when customers request services.

5. RELATED WORK AND CONCLUSION

Blockchain based security architecture for the communication of CPS is proposed in (Yin, Bao, Zhang, & Huang, 2017) and used in a cotton production process, for improvement of the security and timeliness of the equipment communications and data exchange. Privacy protection for end user energy usage can be deployed in two ways. Secure hardware and additional electrical components can be adopted, as proposed by (Kalogridis, Denic, Lewis, & Cepeda, 2011) where a power router and a

rechargeable battery is utilized to moderate appliance load and use patterns to avoid being recognized and tracked by malicious intruders. Existing work mostly focus on data aggregation or anonymization. Our work addresses the unlinkability among different interactions with different parties where some necessary personal data is requested, and we managed to minimize the data disclosure with attribute based verification. A user-centric distributed solution for privacy-preserving analytics using multi-party computation is also proposed (Bestavros, Lapets, & Varia, 2017).

Aside from blockchain adoption in CPS sector, the decentralization and security characteristics of blockchain have attracted researchers and developers to design and develop various applications in different aspects, such as smart contracts, distributed DNS, and identity management etc. There are also blockchain-aimed methods developed to provide security guarantees or enhancement. Tierion (Vaughan, Bukowski, & Rempe, 2017) provides a platform for uploading and publishing data records into the Blockchain network. With public APIs available, Tierion is convenient for integrating applications that demand need of blockchain. Developers can post metadata using HTTP request into Tierion data store and fetch record information. Each data record has a record ID which can be used to retrieve the blockchain receipt generated based on the blockchain transactions. The blockchain receipt contains the transaction ID which will be used to locate a transaction and the block that hosts the transaction. In this way, the data record posted on the blockchain cannot be tampered and the integrity is assured. Blockchain application in information-centric network for name based security of content distribution has also been proposed (Fotiou & Polyzos, 2016). Enigma is a decentralized computation platform with guaranteed privacy which uses blockchain network to control the network, manage access control and identity, and create tamper-proof log of events (Zyskind, Nathan, & Pentland, 2015). Guardtime provides industrial-scale blockchain services using Keyless Signature Infrastructure (KSI) and secure one-way hash function, which is quantum-immune in contrast to RSA (Buldas, Kroonmaa, & Laanoja, 2013). Guardtime also proposed a blockchain standard for digital identity and a protocol for authentication and digital signature which provides a simplified mechanism for revocation management and long-term validity (Buldas, Laanoja, & Truu, 2014). It is pointed out that the blockchain usage in the future Procurement 4.0, as a potential solution to substantiate Industry 4.0 (Nicoletti, 2018). Blockchain-based information sharing is emphasized to improve flexibility and to enable monitoring of risks as well as to establish preventive actions (Urciuoli, 2015). Design principles for application of Bitcoin data structure is explored in (English & Nezhadian, 2017). The application of semantic blockchains is proposed to solve the issue of data integration with flexibility (Brewster, 2015).

In this paper, we propose a reliable data provenance architecture for generalized CPS and utilize blockchain to protect data integrity and detect tampering. For privacy preservation, especially in some business-driven scenarios in CPS, we adopt the attribute based verification for minimum information exposure. We take power system as an example and illustrate the design on top of Hyperledger Fabric and address those challenges in terms of blockchain integration in CPS systems, including scalability, compatibility and privacy concerns. Evaluation shows that the proposed architecture is capable of integrity protection and privacy friendly. To the best of our knowledge, this is the first data provenance architecture towards building a survivable critical infrastructure in CPS using blockchain. In the future, we will deploy the architecture in an emulated smart grid environment and evaluate the performance.

ACKNOWLEDGMENT

This work was supported by Office of the Assistant Secretary of Defense for Research and Engineering (OASD (R& E)) agreement FA8750-15-2-0120. The work was also supported by a grant from the National Natural Science Foundation of China (No.61402470) and the research project of Trusted Internet Identity Management (2016YFB0800505 and 2016YFB0800501).

REFERENCES

- Bestavros, A., Lapets, A., & Varia, M. (2017, January). User-centric distributed solutions for privacy-preserving analytics. *Communications of the ACM*, 60(2), 37–39. doi:10.1145/3029603
- Bichsel, P., Binding, C., Camenisch, J., Groß, T., Heydt-Benjamin, T., Sommer, D., et al. (2009). *Cryptographic protocols of the identity mixer library*. Tech. Rep. RZ 3730, Tech. Rep.
- Brewster, C. (2015). *Semantic blockchains in the supply chain*. Academic Press.
- Buldas, A., Kroonmaa, A., & Laanoja, R. (2013). Keyless signatures' infrastructure: How to build global distributed hash-trees. In *Nordic Conference on Secure IT Systems* (pp. 313-320). Springer. doi:10.1007/978-3-642-41488-6_21
- Buldas, A., Laanoja, R., & Truu, A. (2014). Efficient Implementation of Keyless Signatures with Hash Sequence Authentication. *IACR Cryptology ePrint Archive, 2014*, 689.
- Cachin, C. (2016). Architecture of the hyperledger blockchain fabric. *Workshop on distributed cryptocurrencies and consensus ledgers*.
- Cramer, R., & Shoup, V. (2000). Signature schemes based on the strong rsa assumption. *ACM Transactions on Information and System Security*, 3(3), 161–185. doi:10.1145/357830.357847
- Douligieris, C., & Mitrokotsa, A. (2004). Ddos attacks and defense mechanisms: Classification and state-of-the-art. *Computer Networks*, 44(5), 643–666. doi:10.1016/j.comnet.2003.10.003
- English, S. M., & Nezhadian, E. (2017). *Application of Bitcoin Data-Structures & Design Principles to Supply Chain Management*. arXiv preprint arXiv:1703.04206.
- Fiat, A., & Shamir, A. (1986). How to prove yourself: Practical solutions to identification and signature problems. In *Conference on the theory and application of cryptographic techniques* (pp. 186–194). Academic Press.
- Fotiou, N., & Polyzos, G. C. (2016). Decentralized name-based security for content distribution using blockchains. In *Computer Communications Workshops (INFOCOM WKSHPS), 2016 IEEE Conference on* (pp. 415-420). IEEE.
- Goldreich, O., & Oren, Y. (1994). Definitions and properties of zero-knowledge proof systems. *Journal of Cryptology*, 7(1), 1–32. doi:10.1007/BF00195207
- Grid, N. S. (2010). Introduction to nistir 7628 guidelines for smart grid cyber security. Guideline, Sep.
- Griffor, E. R., Greer, C., Wollman, D. A., & Burns, M. J. (2017). *Framework for cyber-physical systems: Volume 2, working group reports*. Special Publication (NIST SP)-1500-202.
- Han, J., Shah, A., Luk, M., & Perrig, A. (2007). *Don't sweat your privacy using humidity to detect human presence*. Academic Press.
- Kaku, E. (2017). *Using blockchain to support provenance in the internet of things* (Doctoral dissertation).
- Kalogridis, G., Denic, S. Z., Lewis, T., & Cepeda, R. (2011). Privacy protection system and metrics for hiding electrical events. *International Journal of Security and Networks*, 6(1), 14–27. doi:10.1504/IJSN.2011.039630
- Khaitan, S. K., & McCalley, J. D. (2013). Cyber physical system approach for design of power grids: A survey. In *Power and energy society general meeting (pes), 2013 IEEE* (pp. 1–5). IEEE.
- Liang, X., Shetty, S., Tosh, D., Kamhoua, C., Kwiat, K., & Njilla, L. (2017). Prochain: A blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability. In *Proceedings of the 17th IEEE/ACM international symposium on cluster, cloud and grid computing* (pp. 468–477). IEEE.
- Liu, J., Xiao, Y., Li, S., Liang, W., & Chen, C. P. (2012). Cyber security and privacy issues in smart grids. *IEEE Communications Surveys and Tutorials*, 14(4), 981–997. doi:10.1109/SURV.2011.122111.00145
- Nakamoto, S. (2008). *Bitcoin: A peer-to-peer electronic cash system*. Academic Press.
- Nicoletti, B. (2018). The Future: Procurement 4.0. In *Agile Procurement* (pp. 189–230). Cham: Palgrave Macmillan. doi:10.1007/978-3-319-61085-6_8

- Ramakrishna, M. V., & Zobel, J. (1997). Performance in practice of string hashing functions. In *Database systems for advanced applications '97, proceedings of the fifth international conference on database systems for advanced applications (dasfaa), Melbourne, Australia, April 1-4, 1997* (pp. 215–224). Academic Press.
- Shirey, R. W. (2007). *Internet security glossary, version 2*. Academic Press.
- Szydło, M. (2004). Merkle tree traversal in log space and time. In Eurocrypt (Vol. 3027, pp. 541–554). Academic Press.
- Urciuoli, L. (2015). Cyber-resilience: A strategic approach for supply chain management. *Technology Innovation Management Review*, 5(4), 13–18. doi:10.22215/timreview/886
- Vaughan, W. (2016). *Chainpoint: a standard blockchain proof protocol*. Academic Press.
- Vaughan, W., Bukowski, J., & Rempe, G. (2017). *Chainpoint: a standard blockchain proof protocol*. Academic Press.
- Vaughan-Nichols. (2012). *OwnCloud: Build your own or manage your public cloud storage services*. ZDNet.
- Wood, G. (2014). *Ethereum: A secure decentralised generalised transaction ledger*. Ethereum Project Yellow Paper, 151.
- Yin, S.-y., Bao, J.-s., Zhang, Y.-m., & Huang, X.-d. (2017). *M2m security technology of cps based on blockchains*. Academic Press.
- Yu, S., Wang, C., Ren, K., & Lou, W. (2010). Attribute based data sharing with attribute revocation. In *Proceedings of the 5th ACM symposium on information, computer and communications security* (pp. 261–270). ACM.
- Zyskind, G., Nathan, O., & Pentland, A. (2015). *Enigma: Decentralized computation platform with guaranteed privacy*. arXiv preprint arXiv:1506.03471.

Xueping Liang is a Cybersecurity Researcher at Tennessee State University (TSU) and a PhD student at Institute of Information Engineering, Chinese Academy of Sciences. Her research interests include Cyber and Cloud security, Trusted Computing, and Blockchain technology. Sachin Shetty is an Associate Professor in Virginia Modeling, Analysis and Simulation Center and Center for Cyber Security Education and Research, Department of Modeling, Simulation and Visualization Engineering, Old Dominion University. His research interests include Cyber and Cloud security, Trusted Computing, and Blockchain technology.

Deepak K. Tosh is a Cybersecurity Researcher at Norfolk State University (NSU). Prior to that he was a postdoctoral researcher at Tennessee State University. He has received his Ph.D. in Computer Science and Engineering from University of Nevada, Reno in Summer 2016. He received his masters from University of Hyderabad, India in Spring 2012. His research interests include Cyber and Cloud security, Cyber-Threat Information Sharing, Incentivization Models, Cyber-Insurance, and Blockchain technology.

Juan Zhao is working as a Postdoc fellowship in Department of Biomedical Bioinformatics at Vanderbilt University. She worked as a Research Scientist and Adjunct Graduate Faculty at Tennessee State University from Sept. 2015 to Sept. 2017. She received her Ph.D. degree from University of Chinese Academy of Sciences in 2012, and B.E. from Shandong University in 2006, both degrees in Computer Science. Prior to joining Tennessee State University, she worked as an Associate Research Professor from 2012 to 2015 in Chinese Network Information Center, Chinese Academy of Sciences. Her research interests include machine learning in cyber security, biomedical informatics, transfer learning, anomaly detection, feature engineering, natural language processing and social network analysis.

Danyi Li is a Cybersecurity Researcher at Institute of Information Engineering, Chinese Academy of Sciences. Her research interests include Cyber and Cloud security, Trusted Computing, and Blockchain technology.

Jihong Liu is a PhD student at Institute of Information Engineering, Chinese Academy of Sciences. His research interests include Cyber and Cloud security, Trusted Computing, and Blockchain technology.