Data Loss Prevention Technologies

● Tomoyoshi Takebayashi ● Hiroshi Tsuda ● Takayuki Hasebe
● Ryusuke Masuoka

(Manuscript received April 14, 2009)

Although data loss has become an important problem that needs to be solved in many types of organizations, possible routes of data loss have become complicated and numerous, making countermeasures difficult to develop. Fujitsu Laboratories is developing data loss prevention technologies that integrate know-how related to mobile devices, data searching technologies, and security technologies like encryption that have been under development at Fujitsu for some time. These data loss prevention technologies are undergoing verification and testing within the company with an eye towards commercialization. This paper introduces three technologies that will enable information to be used in an expanded workplace with guaranteed safety, without placing a burden on the user. The expanded workplace will extend beyond the borders of a company to include offline work at a customer's office and at other organizations including collaborating companies. The paper begins by introducing a solution for moving data safely outside the company using a novel universal serial bus (USB) memory device. Next, it presents an E-mail filter for preventing the erroneous sending of E-mail containing sensitive information. Finally, it describes document management technology for protecting a document across its entire lifecycle including editing. It also touches upon data loss prevention technologies in the coming era of software as a service (SaaS) and Cloud computing.

1. Introduction

While the loss or leakage of data has become a major problem that needs to be solved in all kinds of organizations, the routes along which data can be lost have become complicated and numerous, making data loss countermeasures all the more difficult. According to the Japan Network Security Association, the number of publically reported data loss incidents in 2007 involving personal or confidential information was 864. These incidents affected more than 30 million people with total projected compensation for damages amounting to more than 2 trillion yen (about US$21 billion).[1)] The most common data loss medium or route was, as ever, paper (40.4%), followed by the Web/Net such as by Winny file-sharing software (15.4%), portable recordable media such as universal serial bus (USB) devices (12.5%), entire personal computers (10.9%), E-mail (9.8%), and other means including mobile phones (5.9%). These figures reflect the diverse routes over which digital data can flow.

In response to this problem, Fujitsu Laboratories and Fujitsu Laboratories of America have been developing data loss prevention technologies that integrate existing technologies and know-how related to mobile devices, data searching technologies, and security technologies such as encryption. They have demonstrated the effectiveness of these technologies through in-house use and are now working towards their commercialization.

2. Progress in data loss prevention

2.1 Trends in data loss prevention solutions

In the early stages, to defend against outside attacks such as viruses and unauthorized access that pose a threat to data, there was a need for network security technology that could set up a protective wall around an organization, as exemplified by the well-known firewall technique. Next, as the mobile work style began to spread and individual devices such as personal computers (PCs) that store data needed to be protected, security technology came to include the encryption of contents stored on hard disk drives, Trusted Computing Group[2] technology, and end-point security technology such as thin clients.

However, although data loss as a threat from inside the organization has become increasingly complicated, individual end-point measures in themselves have become limited and there is a need for information-centric security technologies. The aim here is to protect an organization's critical information and access to it wherever that information may exist.[3] Specific examples of information-centric security technologies are data loss prevention (DLP) and enterprise rights management (ERM).

DLP is a mechanism that identifies sensitive information by content—regardless of whether it is data in the process of being transmitted over the network (data in motion [DIM]), data on a server (data at rest [DAR]), or data at an end-point like a PC (data in use [DIU])—and prevents it from leaking outside the company.[4]

ERM, meanwhile, is a technology that applies digital rights management (DRM) to internal corporate documents to permanently control access rights to sensitive information whether it is located inside or outside the company. With ERM, existing office documents can be extended as a subject of access control, and uniform management of usage restrictions (viewing, printing, etc.) can be achieved via a server.

Looking forward to an era in which corporate information is stored and used via software as a service (SaaS) and Cloud computing, we can expect the methods of data loss to evolve as well. Nevertheless, we can expect the trend towards protecting information itself in DLP- and ERM-like ways to become increasingly important.

2.2 Achieving a secure work place across multiple organizations

In response to social demands and legal obligations, corporations are beginning to control sensitive information in a server-based centralized manner using thin clients, in contrast to end-point control using a PC or other terminal. However, the use of conventional thin clients cannot adequately provide safe sharing of sensitive information across multiple organizations such as in collaborative projects with affiliated companies or other firms.

The secure work space across multiple organizations that we envision is shown in **Figure 1**. This concept expands the existing intranet and teleworking environment to one that includes even offline work at a customer's or collaborator's office. The idea here is to guarantee a secure work space throughout this expanded environment and to enable information to be accessed and used without placing a burden on the user. Achieving this secure work space requires the following three technologies, which are described in more detail in Section 3.

1) Secure information environment

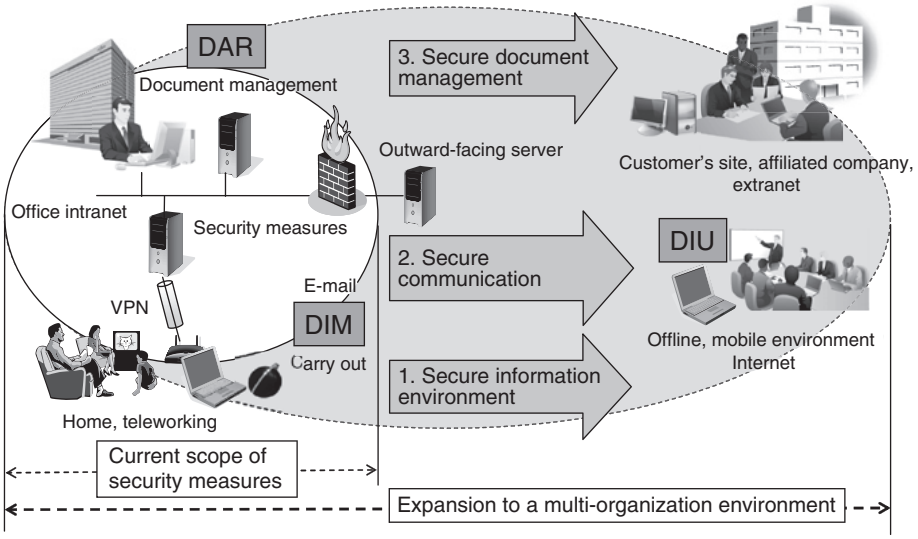
Applies end-point security technology to enable data to be carried around safely even in the case of offline work

2) Secure communication

Applies DLP technology to prevent the loss of data via E-mail

3) Secure document management

Applies ERM technology to protect documents across their lifecycles and to prevent



VPN: Virtual private network

Figure 1
Secure work space.

the outflow of documents via Winny or other file-sharing software

3. Necessary technologies for secure work space

3.1 Secure information environment: making it safe to carry data

3.1.1 Safe data movement

There have been many incidents of USB memory devices containing company data being lost outside the company or of data copied from a USB memory to a PC subsequently being leaked by some means. In response, some companies have prohibited the use of USB memories, but this provides only a superficial solution to the problem. It is natural for employees to want to use USB memories, which today are the simplest and most convenient way to carry files to a customer's office that may lack network access. For this reason, a countermeasure that takes into account actual business conditions is needed.

With this in mind, we consider that sensitive information belonging to either the company or the customer can be safely carried around on a USB memory and used as needed if (1) data can be automatically erased should the USB memory become lost and (2) sensitive information cannot be copied to devices other than prescribed USB memories or servers.

3.1.2 USB memory device with automatic data erase

To make it safe to carry data outside the company on USB memories, we have developed a USB memory device with an automatic data erase function and file redirect technology. It incorporates a microprocessor, battery (automatically recharged via USB), and clock. It performs authentication on the basis of various conditions such as user, elapsed time, and connected PC. If certain conditions are not satisfied, the data is erased by the device itself or alternatively the device is simply rendered unusable. This scheme prevents data leakage through loss or theft of this USB memory device.

File redirect technology is implemented as driver software that enables data in this device to be copied to only company-prescribed USB memory devices or servers and that prevents that data from being copied to a local hard disk drive. This technology can also be linked with

the secure document management technology described later to perform various types of data control such as preventing sensitive information from being printed or transferred as E-mail attachments.

If the abovementioned techniques are combined, data protection can be correctly applied throughout the process of moving sensitive information on a server (DAR) to a USB memory device (DIM), working on the data offline outside the company (DIU), and returning the data to a DAR state without the user being particularly aware of such protection. An example of such data protection is shown in **Figure 2**. Confidential data within the dotted lines (trust boundary) is prevented from being leaked to the outside.

The built-in clock enables the device to automatically erase stored data after a user-selectable preset period of time (default: 24 hours). In a typical use case, sensitive information from a customer can be obtained at the customer's office and carried back safely to the company, where it can then be copied to a prescribed server and analyzed as needed.

3.1.3 Safe data movement in PC environment

One implementation of a "safe PC environment", a natural next step from the "safe data movement" described above, is called "Your PC Anywhere". The idea here is to safely reproduce the user's personal work environment anywhere and to prevent that environment from being maliciously altered. To this end, we have developed a "Your PC Anywhere" prototype that combines virtual machine technology with a microprocessor, memory, Trusted Platform Module (TPM), fingerprint sensor, USB memory, and an organic electroluminescent display. When the prototype USB memory device is connected to a PC, it can perform fingerprint authentication and check device integrity and then allow the PC to reproduce the user's personal work environment using a virtual machine. With "Your PC Anywhere" technology, we aim to satisfy governance demands that

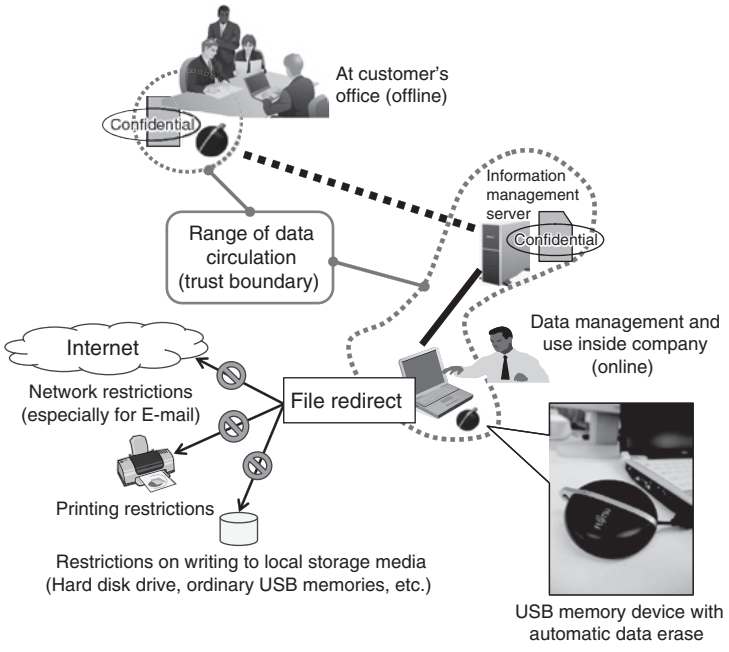


Figure 2
Data loss prevention solution based on USB memory device.

are becoming increasingly strict and to achieve countermeasures against data loss in response to an increase in internal crime.

3.2 Secure communication: countermeasures against erroneous E-mail sending

3.2.1 Difficulty of implementing countermeasures

It has been reported that 66.2% of business users have sent E-mails erroneously.[5] Most incidents of this type are caused by human error or negligence, which makes countermeasures somewhat difficult to develop. We have analyzed such incidents both inside and outside Fujitsu and have come up with eight original countermeasure levels. These cover in increasingly broader range as more levels are implemented from the top down, but the implementation burden on the user also increases.

L1: Automatically encrypts attached file

L2: Allows access to attached file only via Web server

L3: Rechecks before sending

L4: Prevents the sending of E-mail that conflicts with common rules

L5: Allows the sending of only E-mail that observes business rules

L6: Prevents erroneous attachment of confidential documents by interlinking with document management (such as secure information lifecycle management [ILM], described below)

L7: Checks E-mails manually by work review and an approval system that includes superiors

L8: Analyzes E-mail-trail logs and checks for similarities with confidential documents for malicious information leakage.

Encryption and delivery via a server acts as a fail-safe countermeasure only in the event that an E-mail is actually sent out in error. Alerts and rule-based checks, which are considered to be effective for preventing most incidents of erroneous E-mail sending, may lose their effectiveness if alerts are generated too frequently, which would cause users to become oblivious to them. Moreover, protecting confidential documents in conjunction with document management might be insufficiently thorough if users are required to take the time to register such documents manually.

3.2.2 Outbound E-mail filtering and its effectiveness

We have developed an outbound E-mail filtering tool that implements highly effective countermeasures L3–L5. This tool displays an alert window if the user attempts to send out E-mail that does not conform to established rules or that has a destination address containing a typographical mistake, for example (**Figure 3**). The E-mail will not be sent out unless a recheck is performed. This filtering tool has the following features.

1) Can be implemented without altering existing E-mail software or servers

Operating as a simple mail transfer protocol (SMTP) proxy, the filter performs a policy check and issues an alert if necessary before sending out E-mails. It prevents the transmission of the E-mail if the user fails to check a suspicious address or conform to the rules.

2) Highly uniform security level

The policy can be described by an administrator in extensible markup language (XML) in accordance with the business in question and can be delivered and enforced through RDF Site Summary (RSS).

3) Learning-based white list

The filter makes use of log analysis to learn the destination addresses that each individual user frequently uses so that rechecks of such destinations can be suppressed.

Outbound E-mail filtering is now being used as an internal tool in Fujitsu by over 70 000 workers. The policy focuses on the erroneous sending of E-mail to destinations outside the

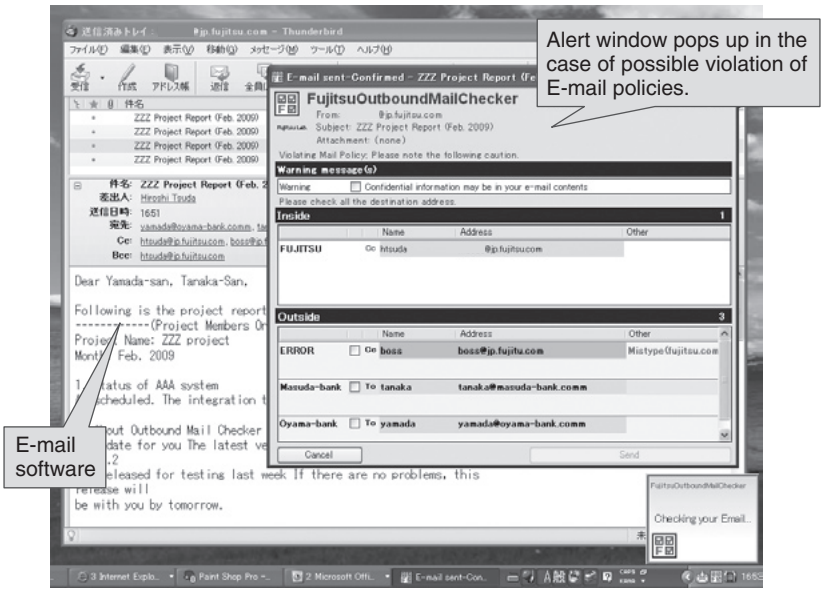


Figure 3
Outbound E-mail filtering interface.

company and is set so as not to generate an excessive number of alerts. For example, all E-mails to be sent outside the company might have to be checked, and there might be more than 20 types of alerts such as one for external E-mails that include specific words or phrases such as "confidential". Favorable comments from users of this system have been reported, such as "When replying to an E-mail, I can determine if there are people on the reply list who do not need to receive it", "I can immediately determine if I have mistyped an address", and "I can avoid mistakenly sending an E-mail to a competitor in the same industry". In addition, an analysis of the E-mail history log can reveal times when many E-mail mistakes are made, such as during the lunch break and at the end of the working day.

3.2.3 *Prevention of text-level outflow of sensitive information*

Sensitive information can be disclosed in text in the message body of an E-mail. A confidentiality check on text can detect general information like people's names and addresses by a pattern-matching process, but preparing individual pattern-match rules for company-confidential information is impractical. In addition, document hash functions and fingerprint technology that establish the identity of certain text suffer from a drop in detection performance if there are many locations where text has been changed.

As an implementation of countermeasures L6 and L8, we have developed technology that extracts text features (content signature) of sensitive information such as confidential E-mail and uses those features to compare sets of text for similarities.[6] A content signature, which is an extension of text-detection technology, is data obtained by calculations on the basis of the word positions in the text and on hash values. If an arbitrary section of E-mail content were to be edited, such as by insertion or deletion, and if that edited section were then to be embedded somewhere in another E-mail, the relative positions of words would be unchanged. Thus, a comparison of two signatures can reveal similarities between two sets of text.

Specifically, the content signature of E-mail related to confidential information that

is distributed via an internal mailing list can be automatically recorded by a server. Such a content signature can then be used to prevent data-leakage accidents, such as when a user involved with confidential E-mail accidently pastes or uses a portion of it in E-mail destined for customers.

We point out here that a content signature does not include word-based information such as persons' names. Thus, the leakage of a content signature itself, if that should occur, would not present a problem. This suggests that the use of content-signature technology could be expanded beyond the control of E-mail outflow. For example, it could be assigned as metadata of a protected document and used to check for text similarities in an environment like Cloud computing, where the content of a document cannot be directly read.

3.3 Secure document management

3.3.1 Protection across entire document lifecycle

An internal document must be protected across its entire lifecycle from its creation (or acquisition) to its distribution and modification and its eventual disposal. To this end, there is a need for document-management technology that can protect data in each of its states, from DAR to DIM and DIU. For this purpose, encryption is an important technology.

Many document-protection technologies based on the encryption of information achieve that protection in the distribution of read-only documents. However, when documents need to be modified (edited), a plain-text version must be temporarily made. Thus, the use (or reuse, as in editing) of an internal document deemed important by a company generates a risk of data loss.

3.3.2 Developed technology

To achieve document management and protection in a company, we have developed technology that can protect documents while they are being edited and that can be used in a variety of applications.

1) Protection during editing

To protect a document across its entire lifecycle including editing, we provide 1) storage of edited documents in a protected format (creation of a protected document), 2) inheritance of protection when copy and paste functions are being used, and 3) protection of even temporary files created by document-editing applications.

In the past, protection during copying and pasting was especially difficult to achieve. The technology that we have developed encrypts data stored on the clipboard when it is copied from a protected document. Then, if the encrypted data on the clipboard is pasted into an ordinary document, that entire document will thereafter be treated as a protected document as well (**Figure 4**).

2) Use in various applications

Most conventional products implement protection functions as an application plug-in, so the application in question must provide a plug-in function, and a plug-in must be developed for each application. To provide an application-independent solution, we have developed technology that integrates a protection mechanism through a common interface layer so that protection can be provided to a variety of applications.

The above two technologies make possible

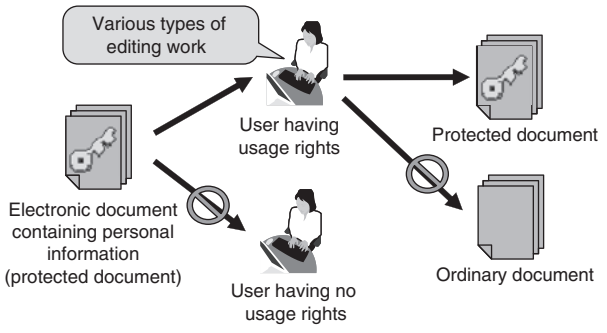


Figure 4
Safe editing of protected document.

a secure document management system that can prevent data loss even when a document is undergoing various types of editing.

4. Future developments: countermeasures against data loss in Cloud computing era

From here on, the mechanism for procuring information technology (IT) resources will undergo dramatic changes as the era of Cloud computing arrives. For the corporate world, however, a major obstacle to adopting Cloud computing is security. Countermeasures against data loss in the Cloud (data in Cloud [DIC]) are the next big security issue. However, this does not mean to say that corporate security requirements and governance and legal obligations are going to change. In the same way as when countermeasures against internal data loss are applied today in the DAR, DIM, and DIU scenarios, it will be necessary to determine where data is located (or from where and to where it is being moved) and to determine whether appropriate IT control is being applied to it. Countermeasures against data loss also need to be applied to the DIC scenario.

There is, however, a key difference between the problem of data protection as it is performed today and that in the Cloud. While the former usually involves two parties, (i.e., the company and the employee), the latter adds a third party in the form of the Cloud provider. This new problem is similar in nature to the problem of sharing information with another company. Once data is passed to another company, transparency and control of it is mostly lost. In short, a company simply has to "believe" that the other company is handling the data appropriately. Recognizing this, existing Cloud providers are having themselves audited by a third-party institution as in the SAS-70 Type II and ISO 27001 certification systems much as ordinary companies do with the aim of gaining the "trust" of their customers.

To provide a system that will enable companies to satisfy diverse security requirements and meet governance and legal obligations from here on, we must achieve transparency for the way data is handled in the Cloud (visualization within the Cloud) and make an environment for handling data in the Cloud.

From the macro viewpoint, internal data that is now confined to one company, as in the DAR, DIM, and DIU scenarios, will enter a world that includes the DIC scenario having multiple Clouds of different types. It will therefore be necessary to have a policy that provides metadata that describes how information is to be handled and a mechanism that enables information itself to ensure that the policy is applied to it. In this way, information itself will be able to ensure that it is handled according to a consistent policy regardless of whether it is located within a company or in the Cloud. We are entering an era that calls for a new form of information-centric security that will enable a consistent policy that includes the DIC scenario to be applied to information.

5. Conclusion

This paper described current activities toward the prevention of data loss in a multi-organization environment with a focus on three key technologies. Data loss can occur in diverse scenarios along various routes as a result of many factors, which means that a single technology or solution—even if capable of dealing with some threats to data—is incapable of dealing with all threats. It is therefore important that a variety of solutions be used in combination. The three technologies introduced here have a complementary relationship that supports a practical, composite solution to data threats. Furthermore, the risk of data loss via paper, which is known to be a major data leakage route, must also be tackled. To this end, the data-prevention technologies presented here must be linked with other security-related technologies developed by Fujitsu Laboratories such as paper encryption and biometric authentication.

Looking to the future, we plan to study countermeasures against data loss in the Cloud and to further expand our multi-organization framework for a secure workplace.

References

1) Japan Network Security Association: Fiscal 2007 Information Security Incident Survey Report. 2008.
http://www.jnsa.org/en/reports/incident.html
2) Trusted Computing Group.
http://www.trustedcomputinggroup.org/
3) Enterprise Strategy Group: Information-Centric Security and Data Erasure (White Paper), 2006.
4) G. Lawton: New Technology Prevents Data Leakage. *IEEE Computer*, Vol. 41, No. 9, pp. 14–17 (2008).
5) HDE: Survey of Erroneous E-Mail Transmission. (in Japanese), 2008
http://www.hde.co.jp/reports/20080423/
6) H. Tsuda: Toward Secure Use of Corporate Information. (in Japanese), Semantic Web Conference 2009, Keio University SFC, 2009.

Tomoyoshi Takebayashi
Fujitsu Laboratories Ltd.

Mr. Takebayashi received B.S. and M.S. degrees in Electrical Engineering from Yokohama National University, Japan in 1979 and 1981, respectively. He joined Fujitsu Laboratories Ltd. in 1981. He worked primarily on research and development of personal communication systems and services. He is currently a senior research fellow in the Software and Solution Laboratories researching and developing secure systems and DLP solutions. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan. He received the Young Engineers Award from IEICE in 1986.

Hiroshi Tsuda
Fujitsu Laboratories Ltd.

Dr. Tsuda received B.S., M.S., and Ph.D. degrees in Information Science from the University of Tokyo, Japan in 1987, 1989, and 1997, respectively. Since joining Fujitsu Laboratories in 1989, he has been engaged in research on the Semantic Web, Web mining and visualization, Web search engines, and knowledge processing. He was a visiting scientist at MIT as a W3C fellow in 2005. He is a research fellow of the Software and Solution Laboratories and currently leading a project on E-mail analysis and DLP.

Takayuki Hasebe
Fujitsu Laboratories Ltd.

Mr. Hasebe received B.S. and M.S. degrees in Electrical and Electronics Engineering from Tokyo Institute of Technology, Japan in 1983 and 1985, respectively. He joined Fujitsu Laboratories Ltd. in 1985 and has been engaged in research on communication systems, security hardware, and security systems. His current research is on a document leakage prevention system. He is a research fellow of the Software and Solution Laboratories.

Ryusuke Masuoka
Fujitsu Laboratories of America, Inc.

Dr. Masuoka received a Ph.D. degree in Mathematical Sciences from the University of Tokyo, Japan in 2000. After joining Fujitsu Laboratories Ltd. in 1988, he conducted research into neural networks, simulated annealing, and agent systems, which led to several Fujitsu products. Since moving to Fujitsu Laboratories of America, Inc. in March 2001, he has engaged in research on pervasive/ubiquitous computing and the Semantic Web, which led to task computing. He has now extended his research interests into trusted computing, the smart grid, software/security validation, and system-level design. He is the director of the Trusted Systems Innovation Group at Fujitsu Laboratories of America, Inc. and an adjunct professor of UMIACS, University of Maryland, USA. He is an IEEE member and an ACM senior member.