Giving Value to "Past Information" with NILFS (New Implementation of the Log-structured File System)

Dr. Satoshi Moriai
Senior Research Engineer, Supervisor
Group Leader, Kernel Group
Open Source Software Computing Project
NTT Cyber Space Laboratories

Open source software (OSS), which can be freely used, redistributed, and distributed as source code, has become popular as an effective means of reducing the cost of software development, system introduction, and total cost of ownership (TCO). The Kernel Group in NTT Cyber Space Laboratories has been researching and developing a new file system called NILFS that makes a hard disk into a type of "time machine". We asked Dr. Satoshi Moriai, the leader of the Kernel Group, to explain this fascinating file system to us.

Creating a landmark file system that can recover past data

—Dr. Moriai, please outline your current R&D theme for us.

In general, we are researching operating systems with the goal of constructing high-reliability systems. We are particularly focused on why data loss in computer systems happens and how we can prevent it. In this regard, we are working on a file system in which no data will be lost if at all possible.

When data is recorded to a medium to update the content, existing data will usually be overwritten unless special measures are taken. If the changed parts were continuously appended, however, data would not be corrupted even if power were lost during the writing process. This scheme, called a log-structured file system (LFS), would also provide a convenient history of all data changes. Research on LFS began in the 1990s, but these days, the capacity of recording media is much higher while the cost per megabyte is much lower. To date, however, a practical LFS has yet to be developed. Against this background, we have added an original algorithm to LFS and have successfully raised it to the commercial level as NILFS (new implementation of the log-structured file system), as shown in **Fig. 1**.

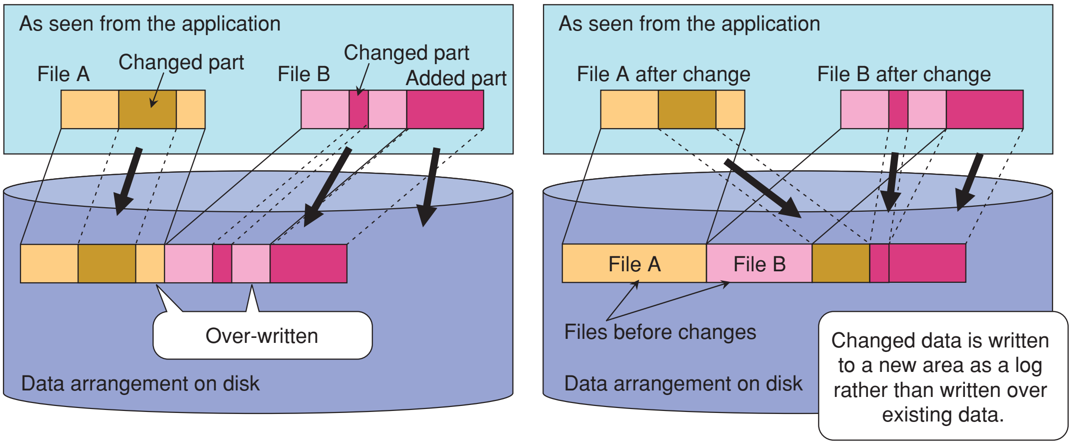
The principle of NILFS can be understood by imagining that the virtual space of the file system consists of semitransparent films (**Fig. 2**). When it is recording data, the system covers previously recorded data with a film and marks that film only with data that differs from past data. Consequently, while only the latest differences are marked on the top film, the entire file can be understood by looking down on the recorded data through all the overlaid layers. Moreover, data from a certain point in the past can be viewed by simply removing all of the layers after that point in time. This is an extremely simple principle. NILFS enables past data to be recovered while maintaining consistency in all files.

—How will the practical application of this research be useful to society?

The greatest benefit that NILFS can provide is improved reliability by preventing data corruption. In addition, since the source code is open, we can maintain and extend it ourselves, which saves on running costs. There is also no need to consult with vendors when we want to add new functions, which will

Uses a log-structured disk format.

- File data and metadata are not written to a fixed location; only the changes are logged sequentially on the disk.
  →Prevents data corruption; enables fast data writing



Principle of an ordinary file system

Principle of the log-structured file system

Fig. 1. Features of NILFS (1).

- Snapshot with time stamp and checksum is generated at a consistent point.
- Instant restoration of the snapshot is possible; recovery of past data is easy.
- Snapshots are generated automatically, without stopping service.
- B-tree used in data management provides efficient processing of large files and large numbers of files.
- Ordinary personal computers can be used: desktop Linux systems as well as enterprise servers.

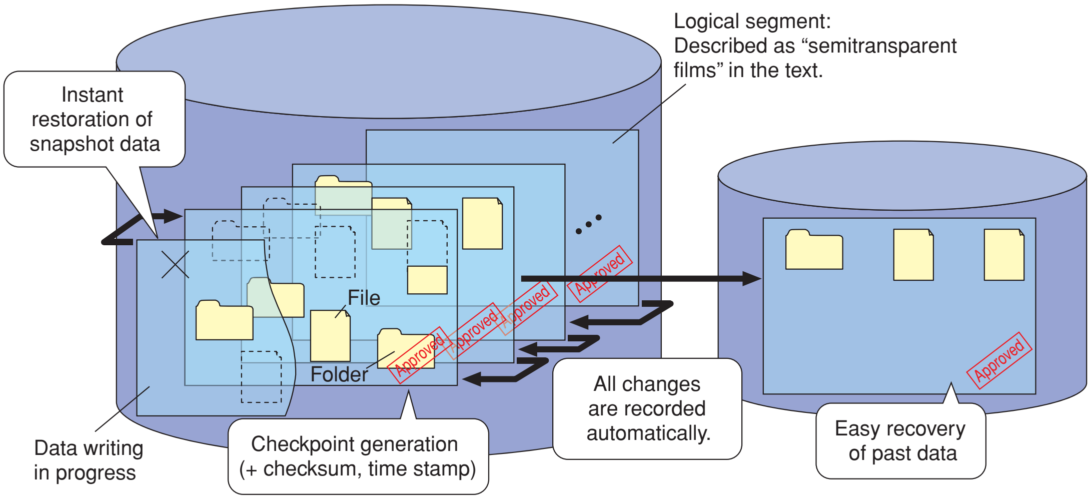

Logical segment:
Described as "semitransparent films" in the text.
Instant restoration of snapshot data
File
Folder
Approved
Approved
Approved
Approved
Approved
Data writing in progress
Checkpoint generation (+ checksum, time stamp)
All changes are recorded automatically.
Easy recovery of past data

Fig. 2. Features of NILFS (2).

enable us to add new features quickly. In the world of the Internet, speed is the key to survival. I want to reduce development time even to one day if possible. In such an environment, OSS can be a powerful tool.

—What are the key technical issues surrounding NILFS?

The conventional LFS only manages the current state of the file system and cannot provide the user with past data even though it may still exist on the disk. After trying various different ideas, we finally came to manage all the data on the disk in the form of a tree having multiple roots. We chose to make each tree a B-tree, which is a data structure widely used in databases. By doing so, we have improved index searching.

—What problems are you currently dealing with?

One problem is how to eliminate the drawbacks of saving the past. We have to admit that a system that can completely recover data from anytime in the past can also pose a security risk. One way of solving this problem is to mark the unnecessary snapshots and delete/retrieve them as trash. However, this mechanism is not implemented in the currently released version of NILFS. The reason is that retrieving trash while maintaining multiple snapshots is very difficult. So we are now focused on developing this.

—How do you think this research will develop in the coming years?

One direction it might take concerns the use of NILFS in combination with sensor networks. In this regard, our group is also researching technology separate from NILFS for gathering specific information from various types of sensors, and this technology can be combined with NILFS snapshots to improve the retrieval of past information. Each snapshot in the NILFS is specified by its creation time, but this level of detail far exceeds the memory ability of most people. With this in mind, we are experimenting with the use of “sketches” as hints for remembering the past (**Fig. 3**). For example, consider the operations that take place in an office where installed sensors record on a personal computer all of the movements occurring in the room. On viewing photos obtained by those sensors, an operator might remark: “I made that revision while I was drinking tea.” In short, information picked up by sensors can be used to help us return to a desired point in the past. We are also looking at ways of automatically selecting news stories from news recorded from the Internet and using those stories as hints for recalling the past.

Another direction our research may take involves the use of NILFS in conjunction with virtual machines. Virtual machine technology enables a single computer to create several computers virtually. Today, as hardware performance continues to increase by leaps and bounds, it often happens that the execution of certain services and tasks uses only a fraction of the available hardware resources. Virtual machine technology collects services and tasks that individually generate a light load and processes them simultaneously on a single computer. From the outside, it appears that multiple virtual computers are running although there is only one computer in reality. We might now ask “what can be done by combin-

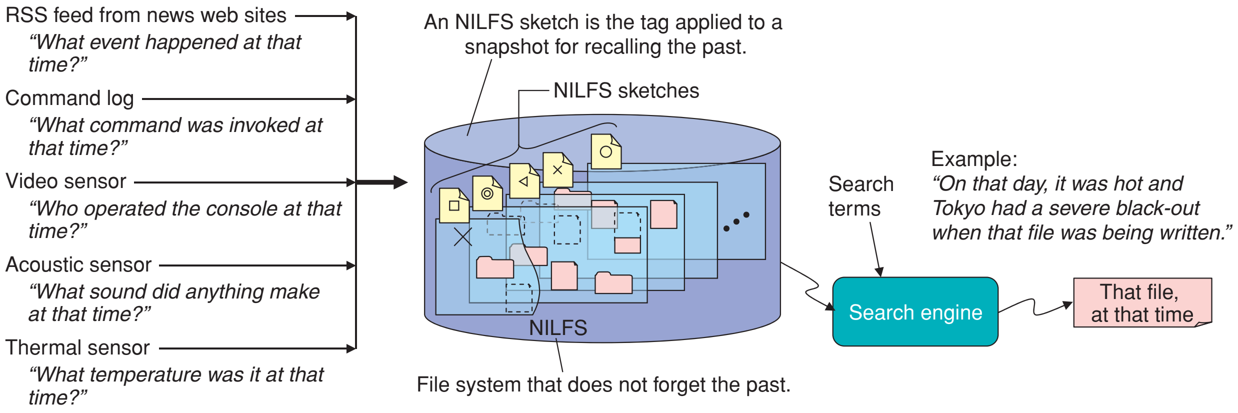

RSS feed from news web sites
“What event happened at that time?”
Command log
“What command was invoked at that time?”
Video sensor
“Who operated the console at that time?”
Acoustic sensor
“What sound did anything make at that time?”
Thermal sensor
“What temperature was it at that time?”
An NILFS sketch is the tag applied to a snapshot for recalling the past.
NILFS sketches
NILFS
File system that does not forget the past.
Search terms
Example:
“On that day, it was hot and Tokyo had a severe black-out when that file was being written.”
Search engine
That file, at that time

Fig. 3. NILFS desktop search.

ing virtual machines with NILFS?" One thing would be to create a "time machine" of sorts. Given a suitable piece of hardware, we can imagine one virtual machine running normally but being monitored and recorded to a log file on an NILFS volume, while another virtual machine reads the log in order to restore operations if necessary. The second virtual machine would constitute a backup system for the monitored system, with any desired interval selected for the time lag of the monitored system. Incorporating this time machine into a network could improve service reliability greatly.

Promoting NILFS in the Linux community

—Dr. Moriai, what is the worldwide trend in operating system research?

Operating systems (OSs) are being heavily researched in the USA and Europe. Moreover, India is currently providing the USA with considerable human resources in this field. On the other hand, I cannot say that OS research is flourishing among Japanese enterprises and researchers, which, as far as I'm concerned, is very unfortunate. In Japan, there is currently not a strong program for cultivating system software programmers. Though some of this is due to language and cultural differences, Japan appears to be content to stand on the sidelines in the world of OS research. The government considers this situation to be a problem and is setting up a number of national projects with the aim of developing system software programmers. We too realize that we need to make our presence felt.

—What kinds of international activities are now taking place with regard to NILFS?

We are just beginning to introduce our activities to OSS-related organizations. Actually, we've only been researching NILFS since the autumn of 2004. Once garbage collection techniques (for discarding obsolete data) for segment cleaning mature, I think we will have a practical, working system that will attract much attention.

As far as international activities are concerned, we are particularly focused on the Linux community. As you probably know, this is a spontaneous and virtual volunteer group that has much influence in the OSS world. Our goal is to provide NILFS for the people who require the NILFS functions and to improve NILFS in collaboration with its users.

—How have your activities been received?

The response to NILFS has been fairly good. Favorable opinions have appeared in various blogs (web logs).

—Are you collaborating with any other companies or research institutions?

In our research of sensor networks and ubiquitous computing systems, we are now involved in several collaborative efforts with universities and other groups at NTT Laboratories. For example, we are working with one university to develop a new software platform for ubiquitous computing. Teaming up with a party having hardware expertise while we take care of the system side is a form of collaboration that we often engage in. Looking forward, we plan to actively seek out tie-ups both inside and outside the company to promote our expertise and obtain know-how that we are lacking in.

Researching Internet protocol processing and realtime operating systems

—What gave you your technical foundation?

Well, I majored in information engineering at university and researched automatic speech recognition. At that time, heuristic approaches were the main currents of this research. As this was hardly acceptable for me, I devised a mechanism that could be learned by a computer using statistical techniques. In this way, by applying human learning methods to a computer, my system may have been a forerunner to artificial neural networks. This research gave me the opportunity to operate computers and write programs, and it was through this work that I gradually awakened to the fascinating aspects of computers.

—What motivated you to enter NTT?

I obtained my doctorate while researching speech recognition. But I had also become interested in computers as I just mentioned, and as it turned out, I became heavily involved in the research on computer systems after entering the company. As a research environment, I had heard that NTT gives researchers a high degree of freedom even in selecting research themes, and I thought that I would be able to pursue future technology R&D with long-range prospects even though NTT Laboratories was a corporate research institution.

—What specific research themes have you been involved with up to now?

After entering NTT in 1988, I first researched network protocol processing models, and then, as the Internet began to expand, I extended the themes to Internet protocol processing and distributed operating systems. In that research, the problem presented to me was how to achieve smooth playback of multimedia content like audio and video on the Internet. Here, I eventually came to the conclusion that the answer lay not only in protocols but also in operating systems, and I expanded my attention to developing a realtime OS. Therefore, in 1994, I joined the joint research project of Keio University, Carnegie Mellon University, NTT, and other companies, in order to develop a new realtime OS called “Real-Time Mach” and its application family. Next, from 2001 to 2004, I was placed in charge of system operations at an Internet service provider (ISP) in the NTT Group. Finally, on completing this assignment, I returned to NTT Laboratories in 2004.

—How did your R&D of NILFS first come about?

It came about in large part through my experiences at the ISP, where I tried various ways to improve system reliability. But no matter how advanced the systems that I prepared were, it was difficult to guarantee complete preservation of information by only hardware means. Thankfully, data-loss accidents were prevented beforehand by the efforts of the excellent engineers on my team, but I did encounter a number of potentially damaging problems. In the end, I realized that a primitive and reliable countermeasure in the form of continuous round-the-clock backing up was essential. I wondered whether an efficient but inexpensive method for this purpose existed, and it was then that I began my NILFS research.

—What do you aspire to in your R&D activities?

In truth, I want to create systems that people enjoy. Though there are bound to be some aspects of computers that probably aggravate all of us, machines were originally created to make things easier for people. With that as a starting point, I have always had the desire in my R&D work to create a computer that could guide people even if they should operate it in a somewhat erroneous manner. Also, despite the remarkable evolution of hardware, it is not being sufficiently utilized, and that goes for hard disks as well. My goal is to make the surplus resources of the computer useful by using them to give the user a sense of security, but not to simply dress up the user interface.

In recent years, moreover, I have developed a deep desire to educate many engineers on the interesting aspects of OS technology. With the OS field viewed as a less than spectacular one, the number of Japanese engineers involved in it is decreasing year by year. I would also like to convey the interesting aspects of this layer of R&D if only to prevent the decline in software in Japan, a country that tends to import too much software.

NTT Laboratories: The best environment for interdisciplinary R&D

—What kind of R&D would you like to pursue in the years ahead?

Well, I still don’t know what my ultimate goal in R&D might be. But I can say that, despite the fact that the Internet has become so huge and popular, I have always thought that the amount of information that you can get your hands on is only a small part of what’s out there, and it’s that phenomenon that I would like to focus on. Today, the Internet makes it easy for us to retrieve the latest information about something, but there is a huge amount of past information that provides the background to current information. For example, there is an enormous amount of data printed on paper sitting on library shelves in addition to the documents that have been digitized to be accessible via the Web. I would like to make information of this type, which can be called a cultural treasure from mankind’s past, open to the public online so that people can be free to examine and use it as desired. Developing technology for this purpose is a theme that I would like to become involved in.

—What is it like working at NTT Laboratories for you personally?

It’s a unique place where you can conduct research in your own way. It has a great pool of talented people and superb facilities, and funds can be obtained if your theme matches the business needs or R&D strategies of the NTT Group. Experts from a wide variety of fields—starting with information and communications and including human interfaces, recognition systems, and devices—are all gathered here under one roof, so to speak. The atmosphere is also quite open and broadminded. NTT Laboratories is the best environment for giving concrete form to a novel idea and for creating a large-scale interdisciplinary system. This kind of research institution is rare in Japan and also on the international level.

—Dr. Moriai, please leave us with a message for young researchers.

More than anything else, please have a very clear idea of what you want to do. Although doing well what you have been assigned is very worthy work, you cannot grow simply on that. As a researcher, you should always be aware of your objectives. I would also like you to be very good in some language. Whether that be English, Japanese, or even a programming language, I don't care. Having expressive power to convey exactly what you're thinking to people or machines is a vitally important skill for researchers. With clear-cut objectives and language ability, I have no doubt that you will fulfill your dreams.

Interviewee profile

■ Career highlights

Satoshi Moriai received the B.E. degree in electrical engineering and the M.E and Ph.D. degrees in information engineering from Tohoku University, Miyagi, in 1983, 1985, and 1988, respectively. Since joining NTT Laboratories in 1988, he has been engaged in research on network protocol processing, internet systems management, enterprise systems architecture, and distributed/realtime/secure operating systems. From 1994 to 1998, he was a visiting researcher in the multimedia platform project and the micro-kernel next-generation project at Keio University. In 1999, he and his co-developers released the Real-Time Mach operating system NTT version. From 2001 to 2004, he was General Manager in the Network Engineering Department of Plala Networks, Inc. Since re-joining NTT Laboratories in 2004, he has been working on dependable Linux kernels, virtual machines, and ubiquitous computing architectures. He is a member of USENIX, the Japan Society for Software Science and Technology, and the Information Processing Society of Japan.