New York University
A private university in the public service

Prof. Dennis Shasha
Courant Institute of Mathematical Sciences
Department of Computer Science
New York University
251 Mercer Street
New York, N.Y. 10012, USA
Telephone: (212) 998-3086
Electronic mail: shasha@cs.nyu.edu
Web: http://cs.nyu.edu/cs/faculty/shasha/index.html

July 23, 2005

Dr. X (whom should this go to?)
SMIS Project
INRIA Rocquencourt
Domaine de Voluceau
BP 105
78153 Le Chesnay Cedex
France

Motivation of Research Project

The principal goal of this research project is to unify leading edge research in cryptographic file systems with databases, especially mobile databases in the context of the SMIS project and peer-to-peer databases in the context of the ATLAS project (Nantes). A significant secondary goal includes beginning collaboration with bioinformatics efforts at INRIA (Mireille Regnier) and Marne-la-Vallee (Maxime Crochemore) where I already know the researchers, but also possibly elsewhere in the biological and bioinformatics groups in Paris (LORIA, Ecole des Mines, IRISA, Paris 13, LIRIS, LRI, LAMI, and Institut Pasteur).

This would be my third year at INRIA. The first one was with projet Rodin (Patrick Valduriez) in 1991-1992. The second was with projet Caravel (Eric Simon) in 1998-1999. Since my first visit, the track record of collaborations between INRIA researchers and me has been strong, not only in papers (three ACM SIGMOD, three VLDB, and one ACM TODS) but also in informal mentoring. While at INRIA, I helped establish a few research efforts mostly by writing the first versions of the core algorithms (most recently, for Le Subscribe and Ajax) and helping to direct the efforts thereafter. Also, I have helped several young researchers from those groups obtain positions at research and development organizations in the United States. Most have since returned to France. Since 2001, I have hosted four engineering students from ISMIA and ESIEE (Adrien Chamming's, Raphael Pouzet, Frederick Gie, and Marc-Olivier Caillot) and one mathematics student (Olivier Bernardi) from ENS for short-term internships at New York University. Thus, my roles include those of collaborator, instigator, and conduit.

Having reviewed the fruits of previous efforts, let me now turn to new projects.

Database Accesses in an Untrusted Environment

Mobile devices have become increasingly powerful, but there will always be a gap between what one can do on a mobile device and what one can do with the full resources of the computational grid offered by one's enterprise or in a peer-to-peer community resource.

While moving, one may be content with what the mobile device can do. Once one has arrived at a destination by contrast, one may want to "plug in" and use all resources. Doing so securely is a challenge that is partly met by secureid cards and other such access mechanisms. In this distributed world, "perimeter security" makes less and less sense. There is no physical perimeter. Even electronic perimeters, provided for example by SSL, are extremely porous in this day of outsourced functionality. The open nature of P2P systems makes security, in particlar data access control, a central challenge as we cannot rely on trusted servers.

Even within organizations, trust cannot be complete. For example, the New York financial community is less worried about frontal attacks on their firewalls by would-be hackers than by rogue maintenance organizations that tap internal cables in order to steal information. In such an environment, each agent in an organization should attempt to ensure the integrity of information as much as possible, even in the face of untrusted support staff, industrial spies, and untrusted networks. At the same time, the full resources of the internet should continue to be brought to bear on problems. That is, one wants various degrees of integrity (a lot for corporate information but perhaps less for sports news) from different information sources, while using the maximum computational resources available.

In work begun at NYU and MIT, David Mazieres and I have designed SUNDR (Principles of Distributed Computing 2002 and Operating Systems Design and Implementation 2004), a system that protects information from snooping through cryptography (previously known), makes modifications tamper-evident through digital signatures (previously known), and makes system administrator subversions evident eventually (novel). A particularly pernicious attack that remains possible even with cryptography and digital signatures is that a system administrator could set up two versions of a file system and let agent A access one and agent B access the other, so as to destroy consistency and therefore sew confusion. (In French, this might be called a *zizanie* attack, but we called it, rather more prosaically, a *forking* attack.)

SUNDR detects forking attacks between agents A and B as soon as one agent attempts to access the same file as the other. As this could take some time after the attack, it is of interest to see whether out-of-band communication could detect such attacks. Mobile devices are one way to provide an out-of-band feature. In addition, the SUNDR design requires private information (such as key and signature information), which would ideally be held in mobile devices.

The major disadvantage of SUNDR is that it applies only to UNIX-style file systems. This makes it very inefficient for fine granularity access as would be required in a database application. Thus the two goals of this project are to:

- Extend the SUNDR guarantees to databases at various levels of granularity.
- Allocate the SUNDR computational requirements to the device that best can handle them. This may require using one or several untrusted servers to perform certain crytographic protocols.

Bioinformatics/Computaional Biology

My work in bioinformatics has been driven by biologists. Weekly meetings with plant and worm biologists in an informal setting have led to novel uses of combinatorial design for iterative learning, an interactive visualization system for multiple experiments, and a specialized analysis tool for phylogeny and hybridization. The reason that this work has not been limited to database work is that, in my work with biologists, I take on the role of a computer science generalist. This has led to papers in journals such as Plant Physiology, Systems Biology, Genome Research, and Science.

As of now, I have no plans for a specific new project in this area, but I plan to interact with researchers in Paris and elsewhere in Europe to start a collaboration. There are many opportunities to make progress in this field, thanks to (i) the high quality work going on in France in biology and machine learning and (ii) the fact that the cost of sequencing and other assays is poised to go down by a factor of 100 in a few years. The second fact in particular will make the role of data management and related machine learning efforts ever more significant, particularly as pertains to the problem of "omic" (i.e., genome, proteome, metabalome) database curation.

Summary

Besides learning a lot during my years at Inria, I have enjoyed them immensely. My language skills are good enough to discuss technical issues and to lecture in French, though I write in English. Three of my books have been published in France by Odile Jacob. At the same time, I have made contributions to the groups I have visited and hope to continue to do so.

Sincerely,

Dennis Shasha

Pertinent bibliography

SUNDR papers

1. Jinyuan Li, Maxwell Krohn, David Mazires, and Dennis Shasha “Secure Untrusted Data Repository (SUNDR)” Proceedings of the 6th Symposium On Operating Systems Design and Implementation (OSDI ’04) San Francisco, CA. December, 2004.
2. ”Building secure file systems out of Byzantine storage”, David Mazieres and Dennis Shasha, Principles of Distributed Computing, 2002. pp. 108-117.

Some biology or biology-inspired publications

1. “Adaptive Combinatorial Design to explore Large Experimental Spaces: approach and validation” Laurence V. Lejay, Dennis E. Shasha, Peter M. Palenchar, Andrei Y. Kouranov, Alexis A. Cruikshank, Michael F. Chou, Gloria M. Coruzzi *Systems Biology*, volume 1, issue 2, December 2004, pp. 206-212.
2. ”A gene expression map of the Arabidopsis root” Kenneth Birnbaum, Dennis E. Shasha, Jean Y. Wang, Jee W. Jung, Georgina M. Lambert, David W. Galbraith, and Philip N. Benfey *Science*, Dec 12 2003: 1956-1960
3. Mitchell Levesque, Dennis Shasha, Wook Kim, Michael G. Surette, and Philip N. Benfey “Trait-To-Gene: A Computational Method for Predicting the Function of Uncharacterized Genes” *Current Biology*, vol. 13, 129-133, January 21, 2003.
4. “Using Combinatorial Design to Study Regulation by Multiple Input Signals. A Tool for Parsimony in the Post-Genomics Era” Dennis Shasha, Andrei Kouranov, Laurence Lejay, Michael Chou, and Gloria Coruzzi, *Plant Physiology*, Dec. 2001 127(4):1590-1594.
5. “cis Element/Transcription Factor Analysis (cis/TF): A Method for Discovering Transcription Factor/cis Element Relationships” Kenneth Birnbaum, Philip N. Benfey, and Dennis E. Shasha *Genome Research* 2001 11: 1567-1573.
6. “Finding Patterns in Three Dimensional Graphs: Algorithms and Ap-

plications to Scientific Data Mining” Xiong Wang, Jason T-L Wang, Dennis Shasha, Bruce Shapiro, Isidore Rigoutsos, and Kaizhong Zhang *IEEE Transactions on Knowledge and Data Engineering*, pp. 731-749, 2002.

Collaborations with INRIA researchers

1. “Transaction Chopping: Algorithms and Performance Studies” Dennis Shasha, F. Llirbat, E. Simon, P. Valduriez *ACM Transactions on Database Systems*, October 1995, pp. 325-363.
2. ”Declarative Data Cleaning: Language, Model, and Algorithms” Dana Florescu, Helena Galhardas, Cristian Saita, Dennis Shasha, and Eric Simon *VLDB, 2001*, pp. 371-380
3. ”WebFilter: A High-throughput XML-based Publish and Subscribe System” Francoise Fabret, Francois Llirbat, Joao Pereira, Arno Jacobsen and Dennis Shasha *VLDB, 2001.* pp. 511-520.
4. “Filtering Algorithms and Implementation for Very Fast Publish/Subscribe” Francoise Fabret, Francois Llirbat, Joao Pereira, Ken Ross, Dennis Shasha *SIGMOD 2001*, pp. 115-126.
5. “AJAX: An Extensible Data Cleaning Tool” Dana Florescu, Helena Galhardas, Dennis Shasha and Eric Simon *SIGMOD 2000*
6. “Publish/Subscribe on the Web at Extreme Speed” Francoise Fabret, Francois Llirbat, Joao Pereira, Dennis Shasha *VLDB 2000*
7. “Simple Rational Guidance for Chopping Up Transactions” Dennis Shasha, E. Simon and P. Valduriez *SIGMOD 1992*, pp. 298-307