Title : "PASS: Provenance-Aware Storage System"

Authors: Margo Seltzer, David A. Holland, Kiran-Kumar Muniswamy-Reddy, Uri Braun, and Jonathan Ledlie

Provenance (also known as pedigree or lineage) means the complete history of a document. The provenance of a file, if available, has a variety of applications. The scientific community can use provenance to reproduce and validate results; the archival community can use provenance to keep track of the chain of ownership and transformations a document has undergone; business compliance software can use provenance to build more informed information lifecycle management (ILM) policies.

Unfortunately, in most computer systems today, provenance is an afterthought and is maintained by hand or is implemented as an auxiliary indexing structure parallel to the actual data. We believe that as all information flows through the operating system, the operating system in cooperation with the storage system should be responsible for the collection and management of provenance. To this end, we are designing a provenance-aware storage system (PASS): a system that transparently collects and maintains provenance.

Building a PASS presents various interesting research questions that we are currently exploring. In order to be useful, a PASS needs to support queries on provenance in addition to collecting and maintaining it. This raises the questions: How should we store provenance? What is the appropriate query interface? Also, the conventional file access control model is inadequate for provenance. For example, a user may have read access to a file, but may not have access to the provenance of the file.

This raises the question: what is the right security model for provenance? Another important issue is network awareness: what should be done with the provenance when data is transmitted across the wire? And last, what criteria do we use to evaluate a PASS?