PASS: Provenance Aware Storage Systems

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

HARVARD UNIVERSITY
Division of Engineering and
Applied Sciences

Provenance refers to the complete history of a document. A Provenance Aware Storage System (PASS) treats the provenance of a file as a first class citizen: Provenance is generated and maintained by the system as transparently as possible.

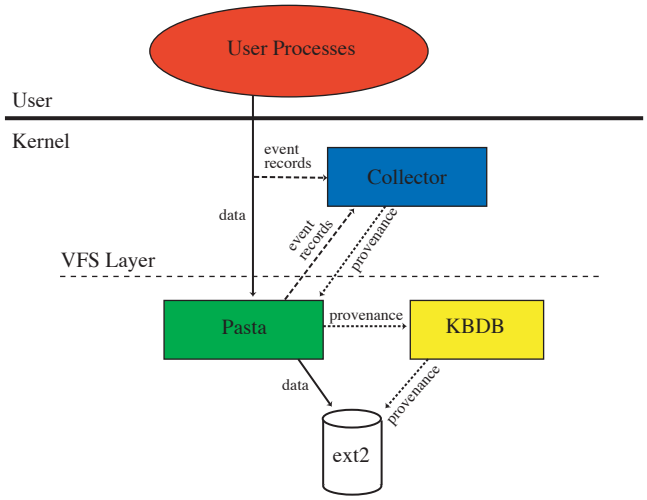
Motivation

- Existing solutions treat provenance as a parallel, but separate data set from actual data: provenance can go out of sync from data.
- Most provenance is Entered Manually: could lead to errors due to negligence.
- Many provenance systems are domain-specific.
- In Many fields, provenance is completely lacking

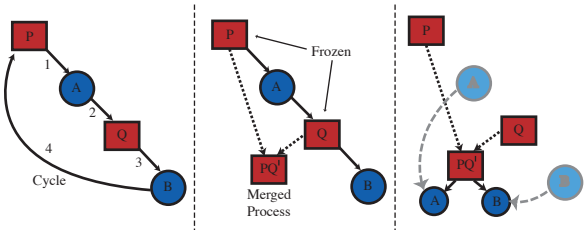
Applications of PASS

- Homeland security: from what sources did I derive this conclusion?
- Archival: What is the record of ownership/format conversions of this document
- Science: How did I (or they) get this result?
- Information life cycle management (ILM): tweak ILM policies for data belonging to a particular application
- Other domains: business compliance, software development, Electronic composition

PASS prototype Architecture

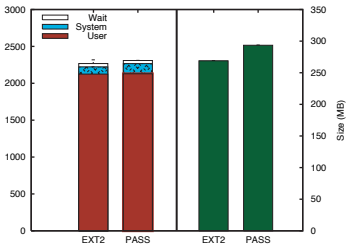


Dealing with cycles



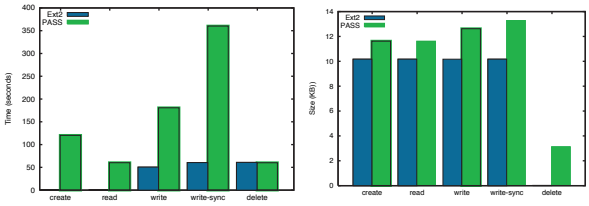
1. Cycle Detected
2. Freeze and Merge Processes
3. Re-parent Files

Linux Kernel Compile Benchmark



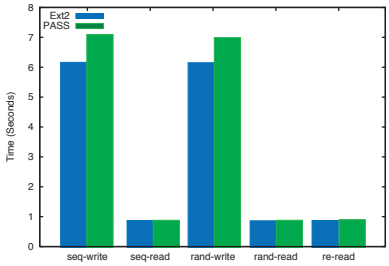
Small File Microbenchmarks

Experiment creates 10,000 files of 4KB each. In subsequent phases, all files are subjected to read, write, and write-sync operations and finally deleted.



Large File Microbenchmarks

Experiment creates 1 large file (100 MB) by sequentially writing to it. In subsequent phases, the file is sequentially read, randomly written, randomly re-written and re-read. The I/O size is 256KB



Research Challenges

- Security: What is the right security model of provenance?
- Search: can we do better than general-purpose search?
- APIs: how do we export provenance to support applications?
- The Wire: how do we implement a distributed PASS?
- Pruning: when can we delete provenance?
- Evaluation: how do we evaluate PASS? To what do we compare? what are the relevant metrics?

For more information: `http://www.eecs.harvard.edu/~syrah/pass`