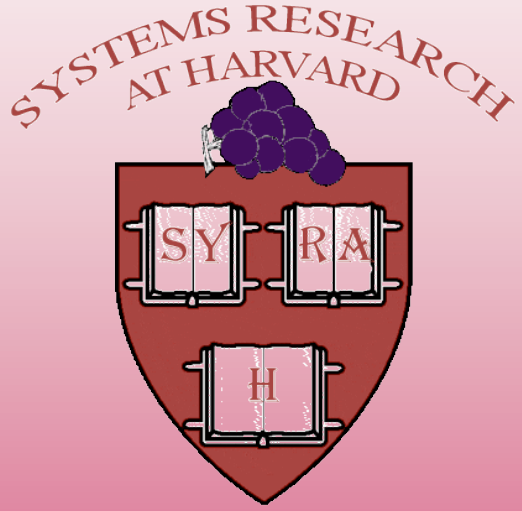
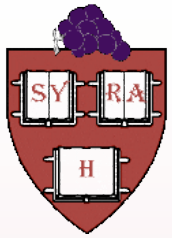
Provenance-Aware Sensor Data Storage Systems



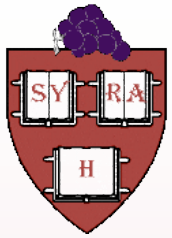
Jonathan Ledlie, Chaki Ng, David Holland,
Kiran-Kumar Muniswamy-Redy, Uri Braun, Margo Seltzer

April 9, 2005



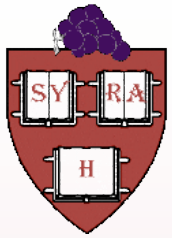
Outline

- Provenance: It's not just about Science
- Provenance-Aware Storage Systems
- Approaches and Research Challenges in PASS
- Conclusions



More than Science

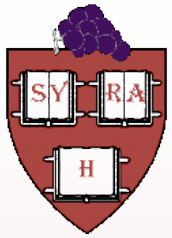
- Homeland security: from what did I derive this conclusion?
- Archival: what is the record of ownership of this document?
- Business: will this document stand up in a court of law?
- Science: how did I (they) get this result?



A Technical Definition of Provenance

- Attribute-value pairs.
  - Some attributes are standard (OS, CPU, process, parameters)
  - Some attributes are application-specific (BLAST query)
  - Some attributes may be domain-specific (astronomy coordinate precision)

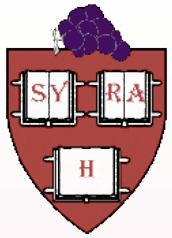
This is an instance of a general problem: some data has meta-data that is as important as the data itself.



The State of Provenance Today

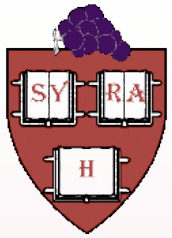
- Most provenance is entered manually.
- Provenance is a parallel, but separate data set from the actual data.
- In many fields, provenance is simply lacking.

There must be a better way!



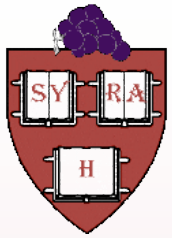
Outline

- Provenance: It's not just about Science
- Provenance-Aware Storage Systems
- Approaches and Research Challenges in PASS
- Conclusions



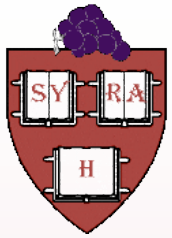
Provenance-Aware Storage Systems (PASS)

- Storage systems (e.g., file systems) in which provenance is a first class object (meta-data).
  - Maintained by the file system.
  - Kept consistent with the data itself.
  - Maintained in the presence of deletion of the data.
- *Provenance is generated and maintained as automatically as possible.*
- Support for rich indexing of provenance.



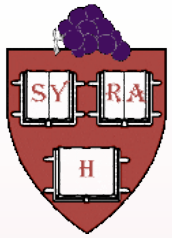
Automatic Provenance Generation

- There are four types of data:
  1. New data: provenance is inside a user's head.
  2. Data from a device: sensor network, microarray data, images, etc.
  3. Derived data: results from a transformation of existing data.
  4. Databases
- Type 1: requires manual intervention.
- Type 2: requires semi-automatic translation.
- Type 3: fully automated maintenance.
- Type 4: need specific DB-style solution.
- Operating system tracks and generates provenance for all transformations.



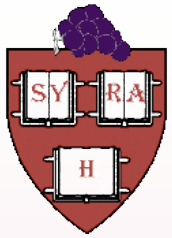
Index and Query

- Users will want to query on provenance
  - Show me everything derived from my file
  - Show me everything upon which I depend
  - How did I get here?
- Provenance schema is not fixed
  - My experiment will have different parameters from yours; parameters are part of the provenance of the result.
- This is the intersection of databases and file systems.



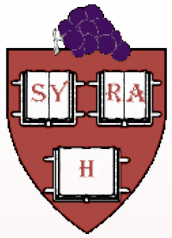
Outline

- Provenance: It's not just about Science
- Provenance-Aware Storage Systems
- Approaches and Research Challenges in PASS
- Conclusions



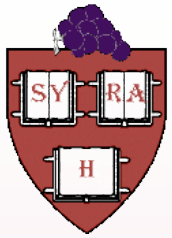
The PASS Agenda

- PASS-I: Integrate provenance with the file system.
- PASS-II: Automatically generate and maintain provenance on a local system.
- PASS-III: Automatically generate and maintain provenance in a network file system or other distributed environment.
- PASS-IV: Support distributed query across a collection of PASS devices.



Research Challenges

- Provenance Issues
- Systems Issues
- Data Management Issues

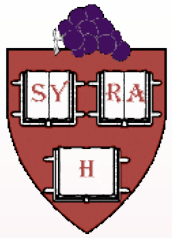


Provenance Issues

- Integrity
  - Trusting OS vs app-generated provenance?
- Security
- Cycles

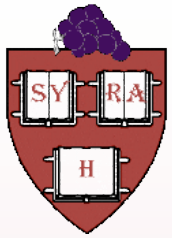
P1	P2
W(a)	
	R(a)
	W(b)
R(b)	

- Pruning



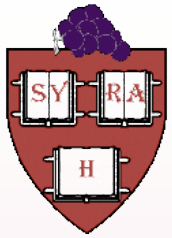
Systems Issues

- When is provenance created?
- When does it become queriable?
- How do you enforce provenance across a wire?
- Do we need a new network file system protocol?
- What do you do about distributed provenance?



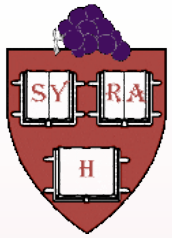
Data Management Issues

- Efficient ancestor/descendant queries in the face of multiple parents, and potentially long ancestry chains.
- Rapid queries on schema-less data.
- Attribute names mean different things to different people.



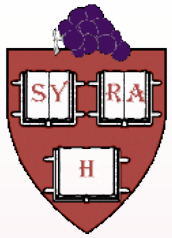
Outline

- Provenance: It's not just about Science
- Provenance-Aware Storage Systems
- Approaches and Research Challenges in PASS
- Conclusions



Status

- Focusing on scientific users.
  - Willing users in biology, physics, astronomy.
  - First PASS: command-line programs.
  - Second PASS: interface with application tools (e.g., Matlab, packaged software)
- We have a solution for transformations, device interfaces need to be customized.
- Put the system in the hands of users in April.



Conclusions

- Provenance is vital for research reproducibility.
- It is also vital in a number of other fields.
- The storage system is the *right* place to manage provenance.
- I believe that provenance is the next “big thing” in storage systems.
- Ten years from now, PASS will be as ubiquitous as RAID is today.