Tracking Emigrant Data via Transient Provenance

Stephanie Jones, Christina Strong,
Darrell Long, Ethan Miller

June 21, 2011

NSF Center for Research in Intelligent Storage
Storage Systems Research Center
Center for Information Technology Research in the Interest of Society
Baskin Engineering
UC Santa Cruz

Introduction

- Data leaks are harmful for companies and government agencies
- Figuring out who leaked your data and how is hard
- If a leak is found, information gathering is critical
  - Who could have leaked the data?
  - When did the data leave the system?
  - What else did the user access at the time?
- Data provenance can be extended to track emigrant data

Assumptions

- All data is kept on a central storage system
- Provenance store uses the PASS framework
- High Performance Computing (HPC) environment
- Clients connect over the network
  - NFS, CIFS, SSH, FTP
- Adversary is a trusted user with malicious intent

Transient Provenance

- Data can leave central storage in two ways
  - Copied or moved to an directly connected external drive
  - Copied over a network connection
- Emigrant data is tracked via ***ghost objects***
- Ghost objects represent a period of time when data has left the central storage system
- Ghost objects differ from regular provenance
  - Do not track data ancestry
  - Are not meant to be immutable

Tracking Data Leaks

- Ghost objects can be used to identify suspect users
- Each ghost represents a period of time during which data were accessible from outside the provenance system's control
- Querying over the provenance graph for leaked data will return all relevant ghost objects
  - Users who accessed the data
  - How the data left the central storage system
  - Where the leaked data went (external drive or IP address)

Conclusion

- Data provenance can be extended to track emigrant data via ***ghost objects***
- Ghost objects are used to track when data emigrates from a storage system
- Querying ghost objects can identify sources of data leaks
  - Identify potential suspects
  - Provide timeframe for the leak and the set of data