Trustworthy Whole-System Provenance for the Linux Kernel

Adam Bates, Dave (Jing) Tian,
Thomas Moyer, and Kevin R. B. Butler

In association with…

UF College of Engineering
UNIVERSITY *of* FLORIDA

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

USENIX Security Symposium, Washington D.C., USA
13 August, 2015

Provenance Matters!

Provenance Matters!

GIZMODO

How a Burnt Lady Gaga CD Helped Leak Thousands of Intelligence Files

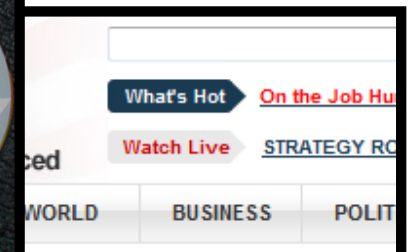
Kat Hannaford 107,554
Filed to: WIKILEAKS 11/29/10 7:20am

Provenance Matters!

GIZMODO

How a Burnt Lady Gaga CD Helped Leak Thousands of Intelligence Files

Kat Hannaford 107,554
Filed to: WIKILEAKS 11/29/10 7:20am



Office of the Inspector General
Commonwealth of Massachusetts

Glenn A. Cunha
Inspector General

Investigation of the Drug Laboratory at the William A. Hinton State Laboratory Institute 2002 – 2012

March 4, 2014

One Ashburton Place, Room 1311 | Boston, MA 02108 | (617) 727-9140 | www.mass.gov/ig

"40,000 Massachusetts defendants may be affected by chemists's alleged misdeeds"

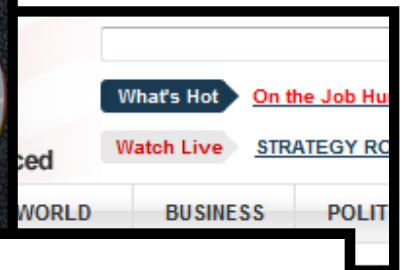
- Morgan Windsor, CNN, Aug 2013

Provenance Matters!

GIZMODO

How a Burnt Lady Gaga CD Helped Leak Thousands of Intelligence Files

Kat Hannaford 107,554
Filed to: WIKILEAKS 11/29/10 7:20am



But is our provenance secure?

Glenn A. Cunha
Inspector General

Investigation of the Drug Laboratory at the William A. Hinton State Laboratory Institute 2002 – 2012

March 4, 2014
One Ashburton Place, Room 1311 | Boston, MA 02108 | (617) 727-9140 | www.mass.gov/ig

defendants may be affected by chemists's alleged misdeeds"

- Morgan Windsor, CNN, Aug 2013

Overview

- **Linux Provenance Modules (LPM)**, for trustworthy provenance monitors in Linux.
- **Provenance-Based Data Loss Prevention**, to monitor and control the propagation of sensitive data in enterprise environments.
- **Evaluation**:
  - Collection agent imposes 3%-8% runtime overhead
  - Provenance queries return in under 3 milliseconds

Whole-System Provenance

Def: prov·en·ance \prä-və-nän(t)s\ *n*:

- A complete description of system *Agents*…
  - e.g., Users, Groups
- … controlling *Activities*…
  - e.g., Processes, Forks
- … and their interactions with *Controlled Data Types*.
  - e.g., Inodes, Sockets, IPC, Memory

Threat Model

- Provenance-Aware Adversary attempts to disable collection agent, tamper with logs, etc.
- Provenance-Aware Applications can be compromised, and may lie about system events.
- Kernel is trusted on install, but can later be attacked.
- PKI stores and distributes keys for Prov-Aware Hosts.

Design Goals

1. Completeness
   - Gapless descriptions of system activity
2. Tamperproof
   - Impervious to attacks launched in user space
3. Verifiable
   - Formal assurance of G1, G2

4. Authenticated Channel
   - Tamper-evident provenance transmission

5. Secure Disclosure
   - Validate annotations disclosed in user space

Rationale…

Reference Monitor Concept

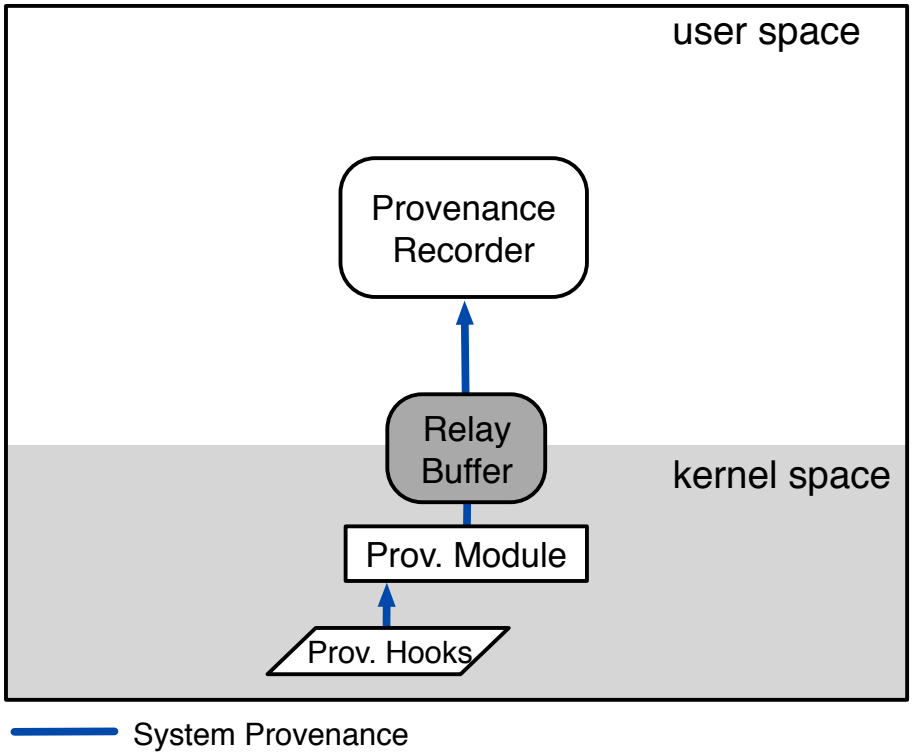
Networked Provenance

Layered Provenance

Design

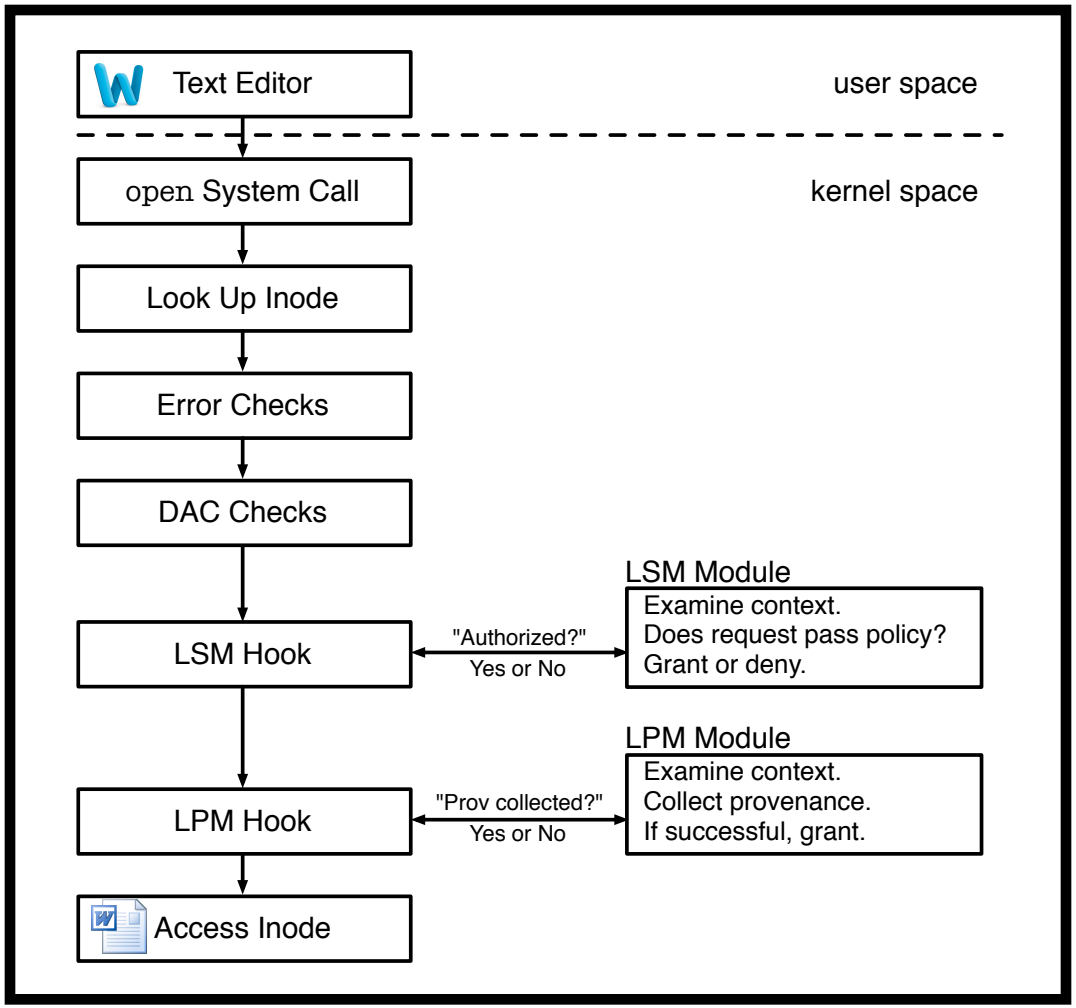
Kernel layer collection agent:

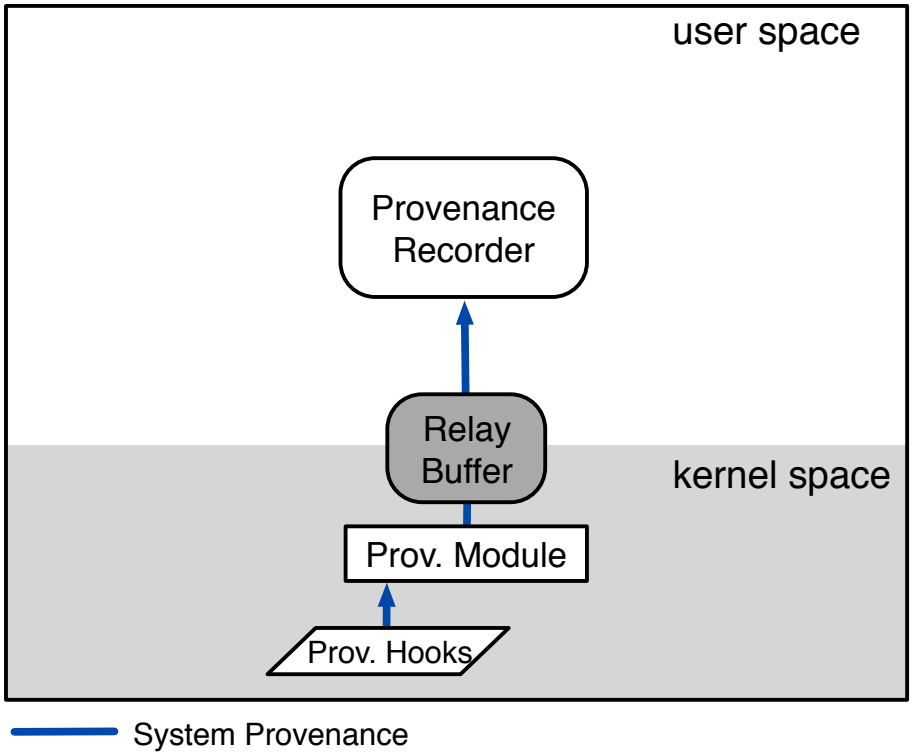
- LPM architecture mirrors Linux Security Modules.
- Kernel instrumented with 170 provenance hooks.
- Modules efficiently transmit provenance to user space with relay buffer.



Design

Kernel layer collection agent:



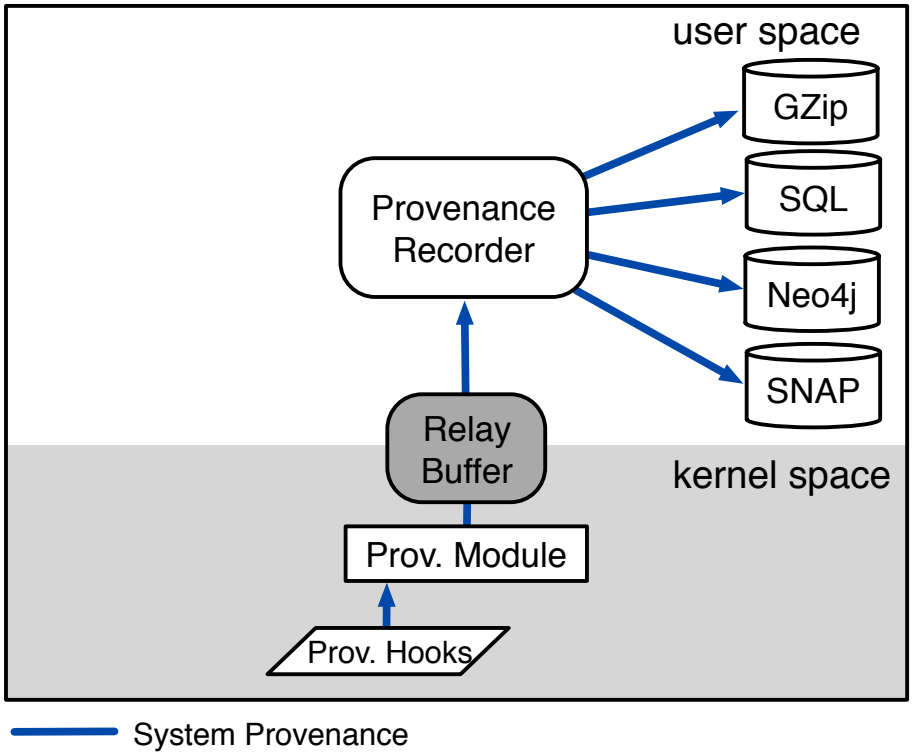


Example control flow through an LPM provenance hook.

Design

Support for provenance storage:

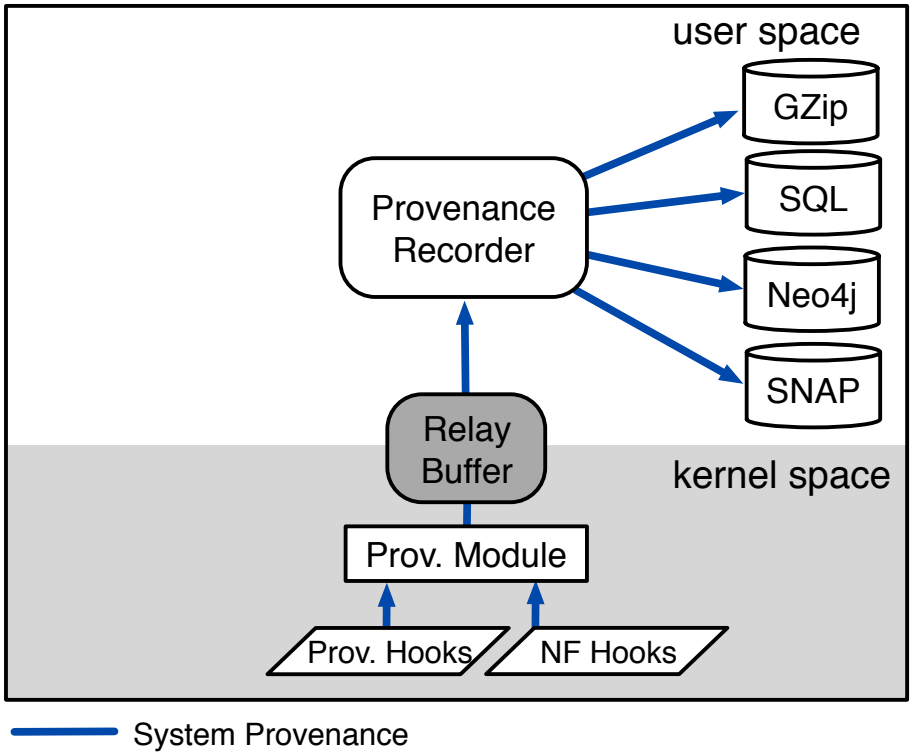
- Recorders translate provenance stream for various storage backends.
- Support recording to file, relational DBs, graph DBs.
- Upcoming: Accumulo.



Design

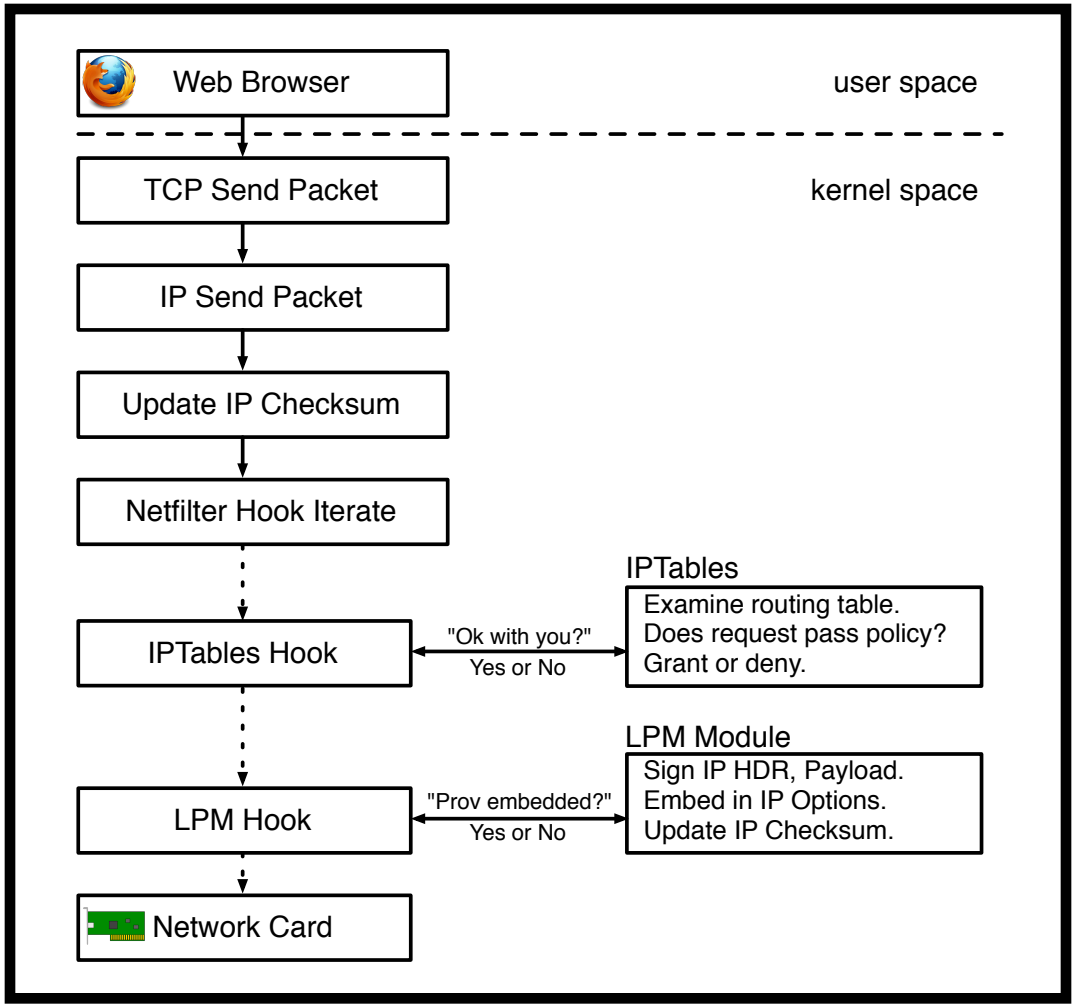
Support for networked provenance-aware systems:

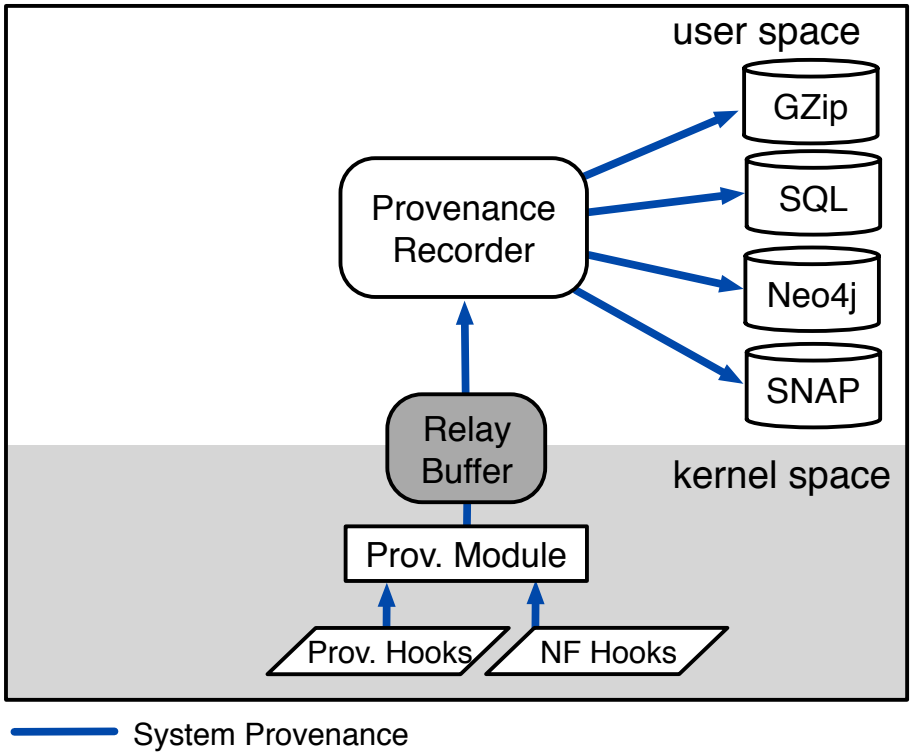
- Message Commitment Protocol enforced with Netfilter subsystem.
- LPM performs per-packet DSA signing and verification.
- Signatures are embedded in IP Options, ensuring (nearly) universal compatibility.



Design

Support for networked provenance-aware systems:



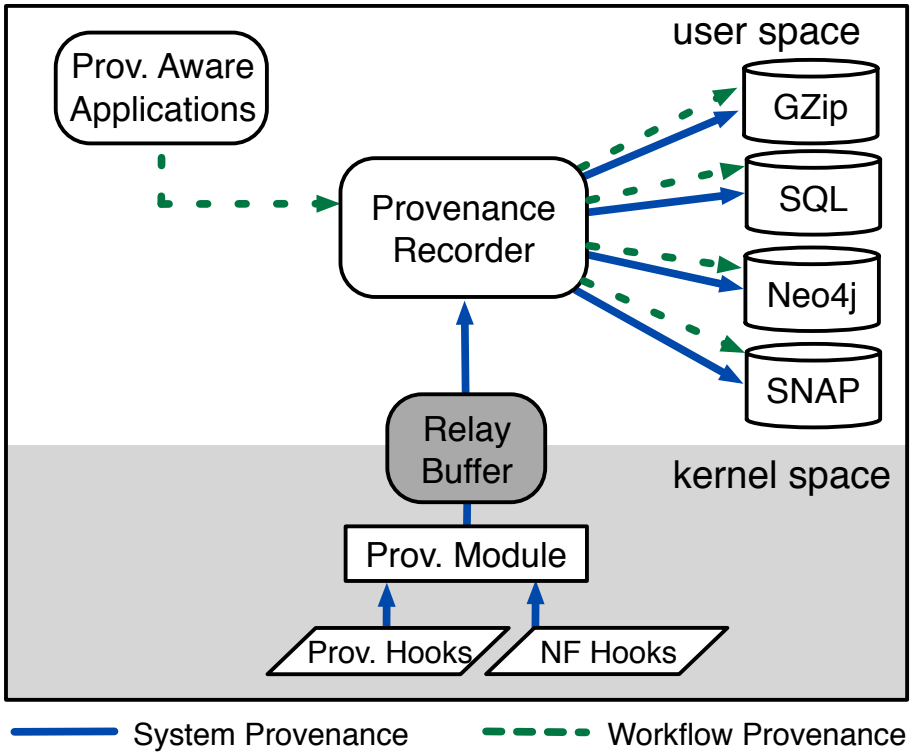


Example control flow for authenticated packet transmission.

Design

Support for layered provenance-aware systems:

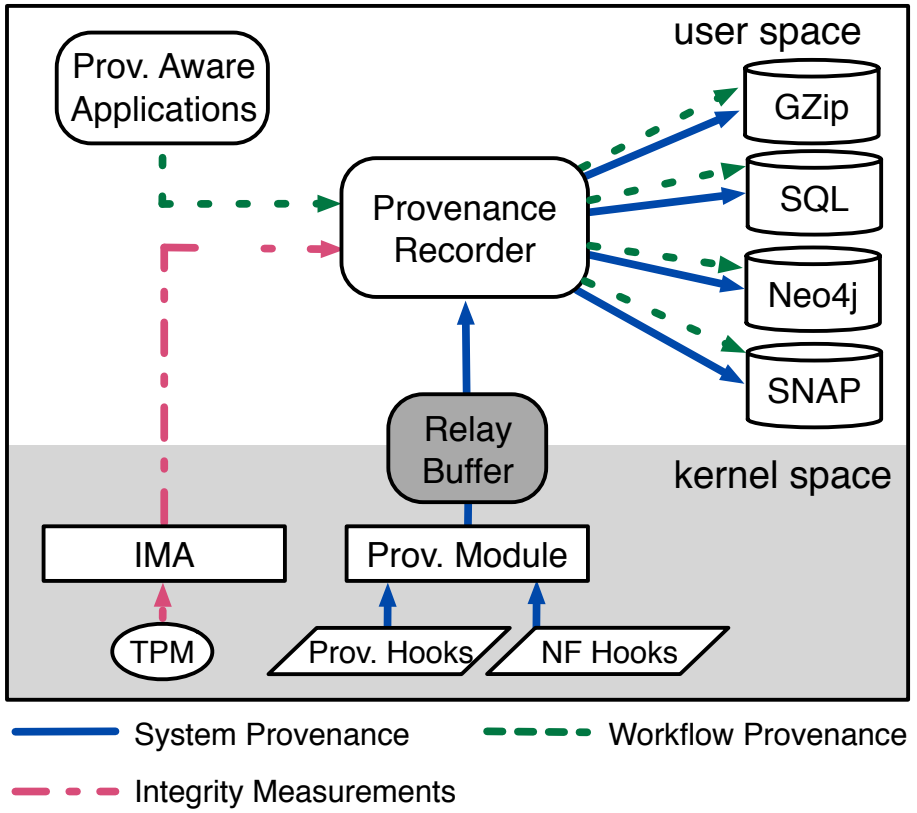
- Kernel provenance suffers from *semantic gap* problem.
- Layered provenance bridges the gap, but expands attack surface.
- Authenticity and integrity of workflow provenance must be validated, but how?



Design

Support for layered provenance-aware systems:

- LPM includes a gateway for upgrading low integrity provenance.
- Integrity Measurement Architecture (IMA) check verifies load time integrity of application.
- Only correctly validated provenance is recorded.



Analysis of Secure Deployment

1. Completeness

- Provenance hooks observe all *sensitive operations* performed on *controlled data types*.

2. Tamperproof

- SELinux preserves run-time kernel integrity
- Secure Boot techniques prevent booting into another kernel

3. Verifiable

- By mirroring LSM hooks, LPM inherits formal analysis that ensures complete mediation of controlled data types.

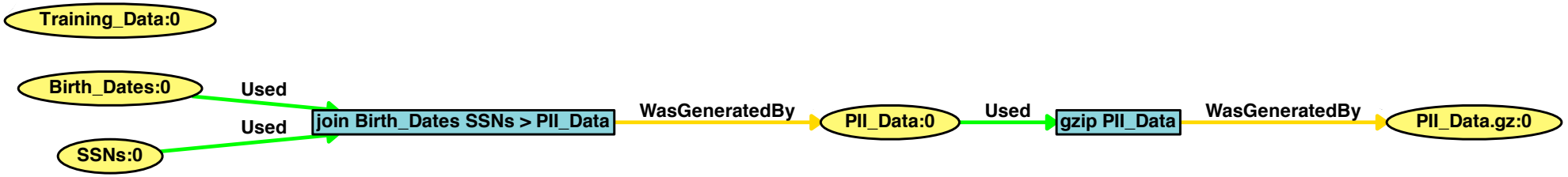
4. Authenticated Channel

- Message Commitment Protocol ensures integrity and identity.

5. Secure Disclosure

- IMA check verifies load time integrity of Prov-Aware Applications.

Data Loss Prevention



Provenance Graph: Two objects are fused together to create PII, then compressed.

Data Loss Prevention tools take the following forms:

- Regex-Based: Fails to recognize data transformations
- Manual Labelling: Not tamper proof, may fail to handle data fusions.
- ***Provenance-Based: All lineage information is recorded, any sensitive ancestry can be traced.***

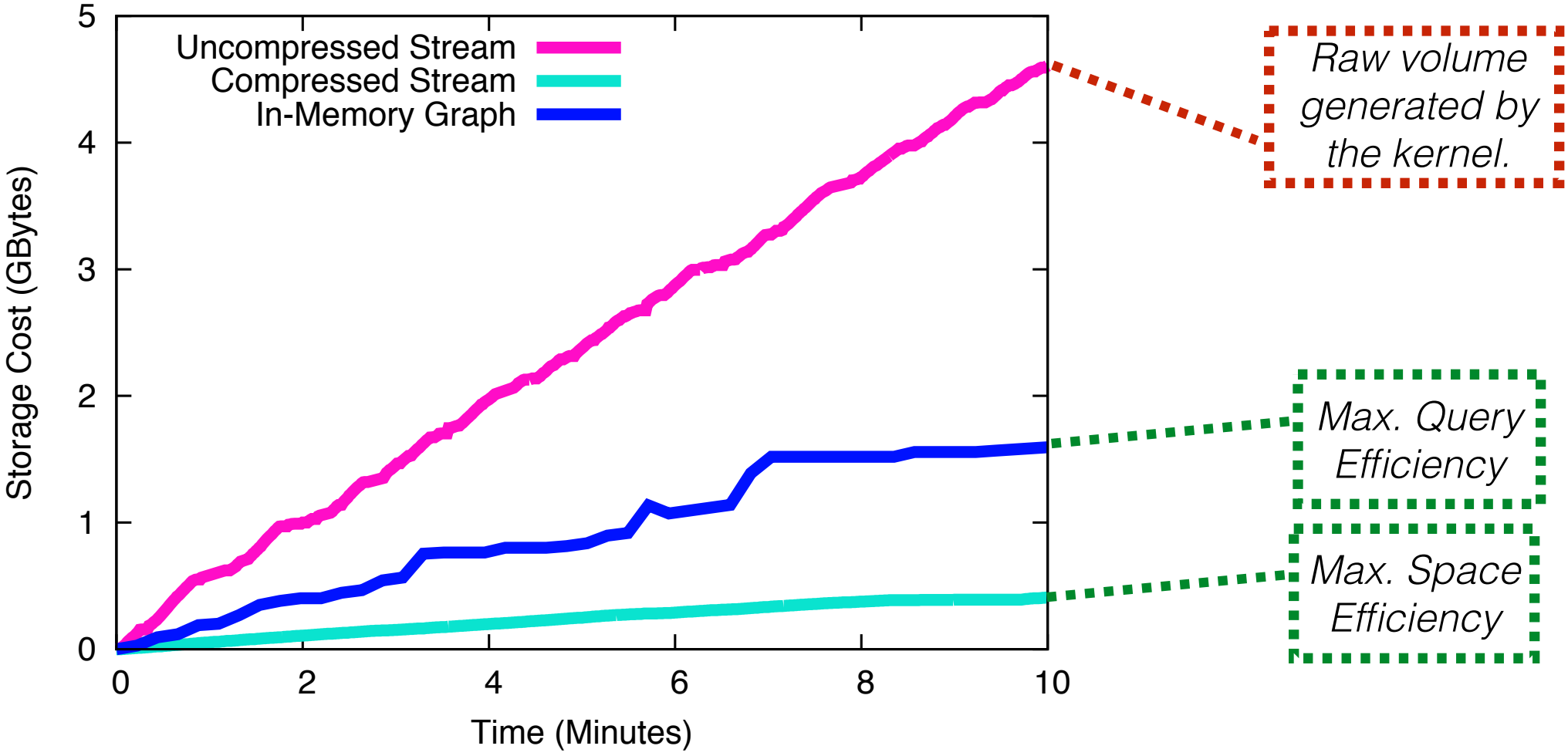
Evaluation: Collection Costs

Benchmark	Vanilla Kernel	LPM w/ Provmon	Overhead
Kernel Compilation	598 sec	612 sec	2.7%
Postmark	25 sec	27 sec	7.5%
Blast Sequencing	376 sec	390 sec	4.8%

Overhead is highest on I/O intensive tasks with frequent file creation, deletion, and open.

Costs amortize over reads and writes.

Evaluation: Collection Costs

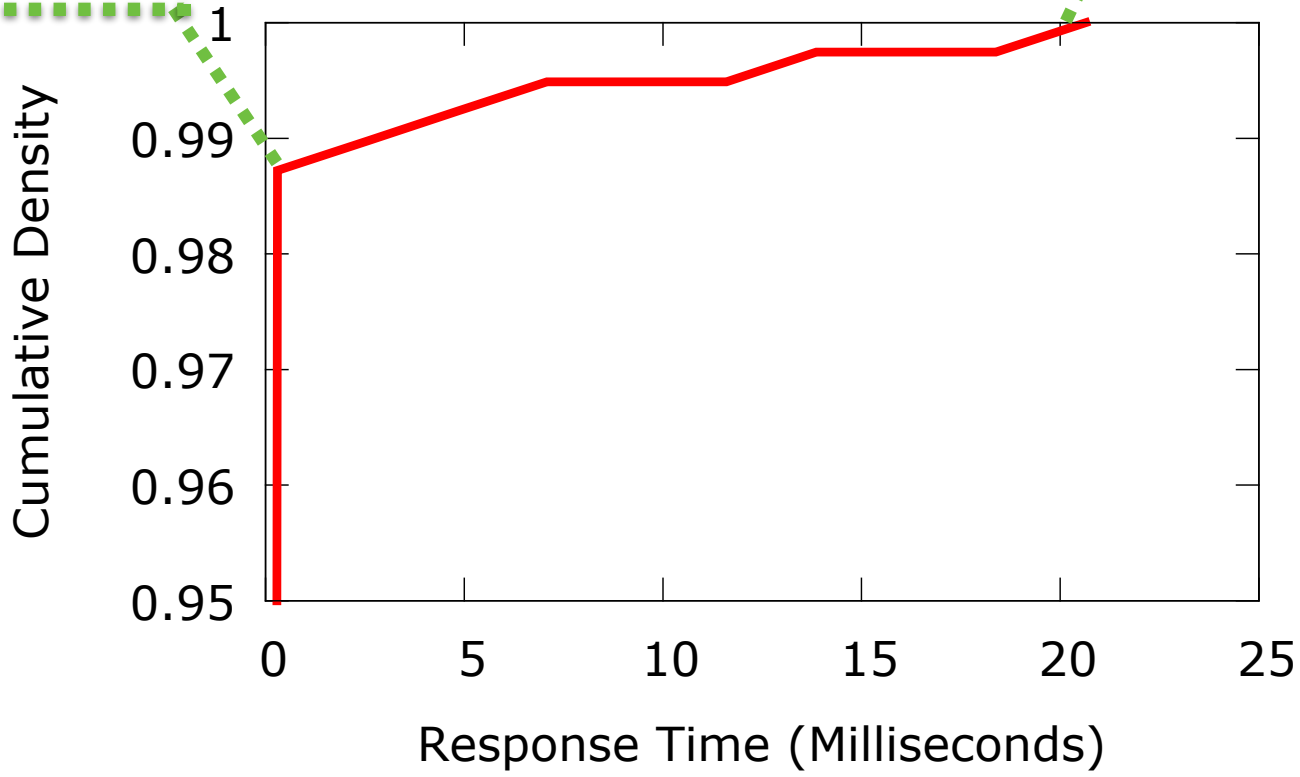


Storage overhead is high, but consistent with other system layer provenance/audit tools. Compression techniques can be used to reduce storage burden.

Evaluation: Query Costs

99% of queries return in 2.5 ms or less!

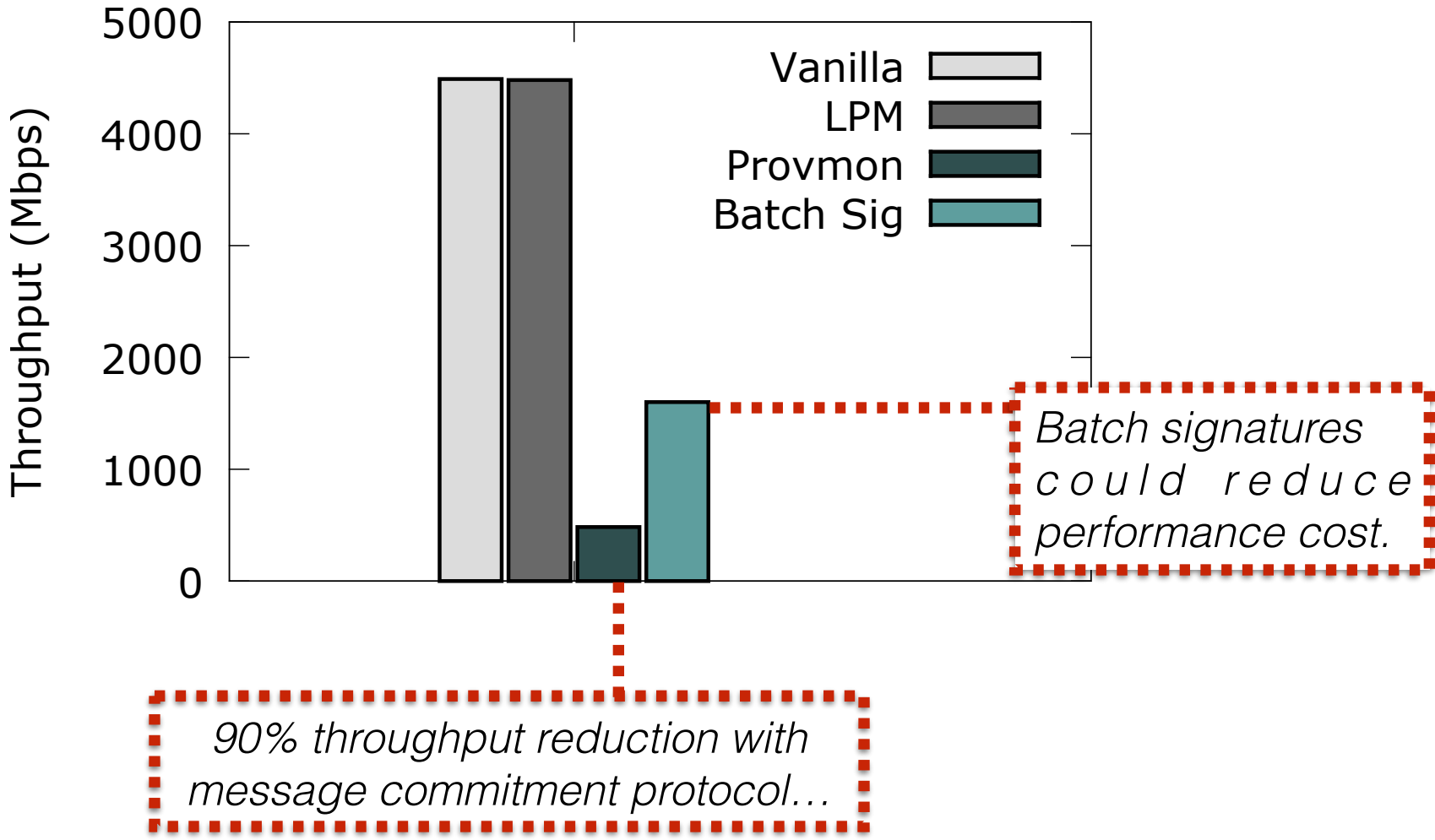
Worst case: 17,696 nodes Returns in 21 ms



PB-DLP Ancestry Queries for Inodes in a 6 million node graph.

(Only inodes with over 50 ancestors were considered)

Evaluation: Network Prov.



IPerf TCP benchmarks of Message Commitment Protocol.

Alternatives: SSL or IPSec, which require app rewriting.

Takeaways

In this work, we…

- identify the requirements for trustworthy provenance in distributed, heterogeneous environments.
- design, implement, and deploy the first fully-realized *provenance monitor*.
- propose a mechanism for provenance-based data loss prevention that offers improved capabilities over existing enterprise systems.

Questions?

Thank you for your time.
adammbates@ufl.edu

LPM is available at http://linuxprovenance.org