



# SPADE: A retrospective

Ashish Gehani, SRI

# Background

- Pre-20<sup>th</sup> century:
  - Experimental science
    - Hypotheses derived from experience
    - Physical phenomena measured
    - Steps and data recorded by hand
  - Theoretical science
    - Mathematical models
    - Conjectures based on analysis
    - Results derived by hand
- Late 20<sup>th</sup> century:
  - Computational science
    - Commoditization of sensors
    - Large volumes of data
    - Analyses involve significant computation
    - Hypotheses emerge from data exploration



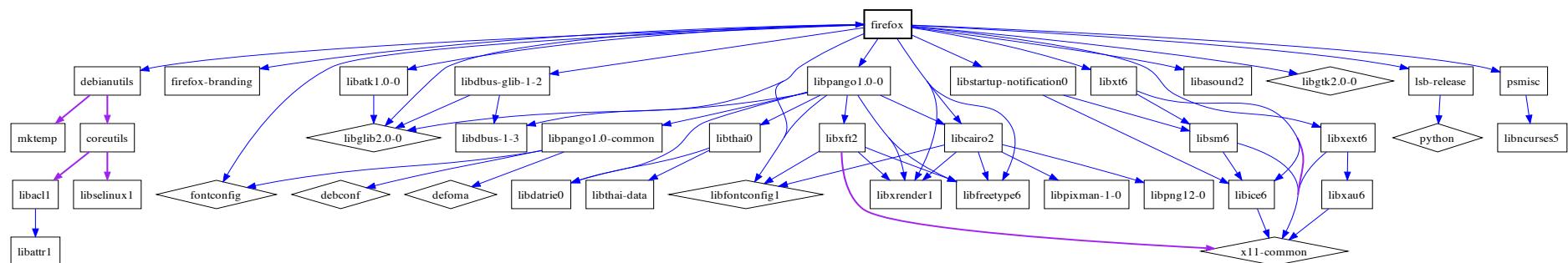
Credit: commons.wikimedia.org



Credit: commons.wikimedia.org

# Motivation

- Application context is complex
- Code dependencies
  - Linked libraries
  - System services
  - Utility programs

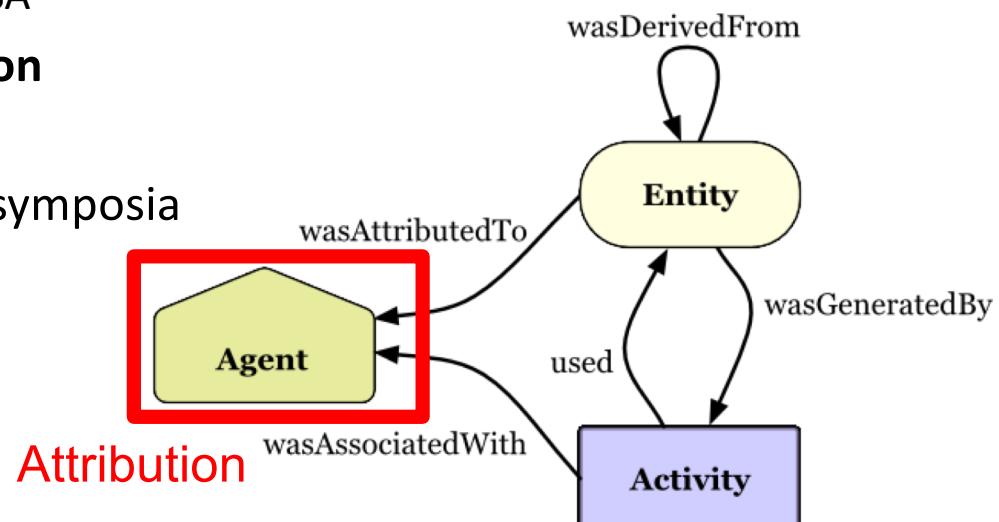


- Environmental dependencies
  - Shell variables
  - Shared memory contents
- Changes in any can affect output

# Data Annotation and Provenance

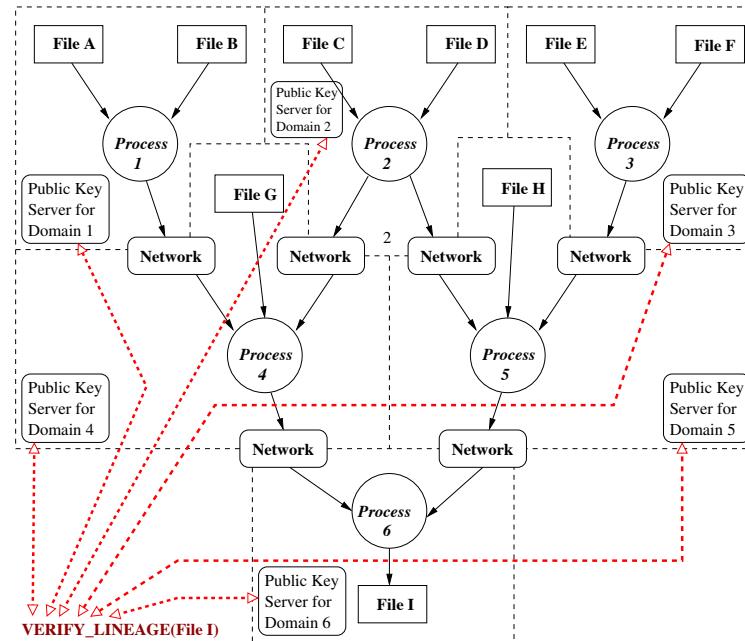
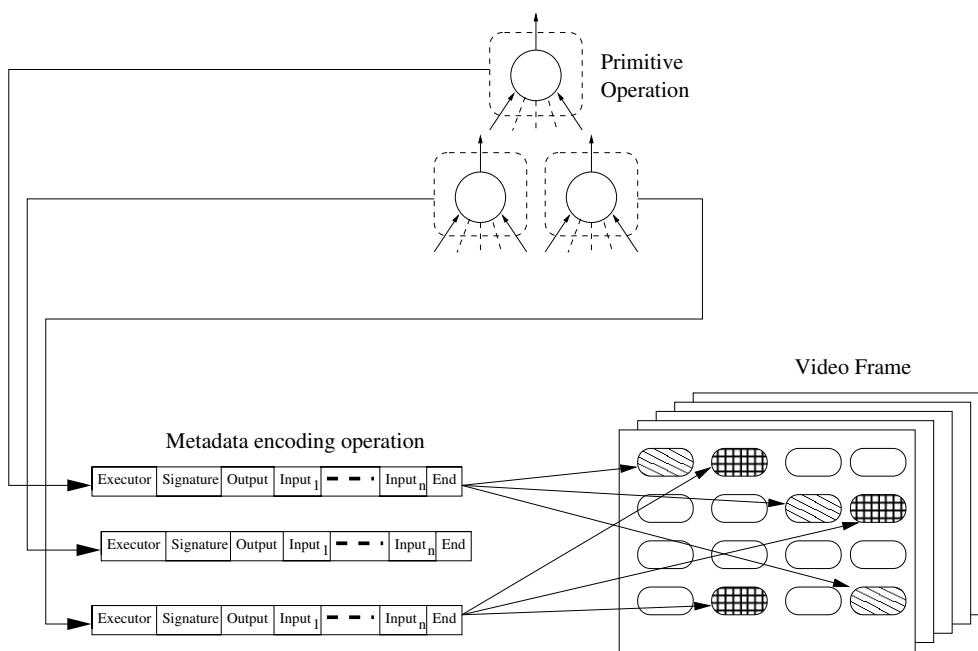
- Initial meetings:
  - 2002 : **Data Derivation and Provenance**
    - Argonne National Laboratory, Chicago, USA
  - 2003 : **Data Provenance and Annotation**
    - e-Science Institute, Edinburgh, UK
  - 2008-9 : **Principles of Provenance** – 6 symposia
    - e-Science Institute, Edinburgh, UK
  - 2012 : **Principles of Provenance**
    - Dagstuhl, Germany
- Emerging specifications:
  - 2007, 2011 : **Open Provenance Model** (versions 1.0, 1.1)
  - 2013 : **W3C PROV** standard
  - 2015-2019 : DARPA Transparent Computing **Common Data Model** (versions 1-20)
- Ongoing event series:
  - 2006- : Biennial **International Provenance and Annotation Workshop**
  - 2009- : Annual **USENIX Theory and Practice of Provenance**
  - 2014- : Biennial **ProvenanceWeek** co-located events

Credit: [www.w3.org/TR/prov-primer/](http://www.w3.org/TR/prov-primer/)



# Precursors (1/2)

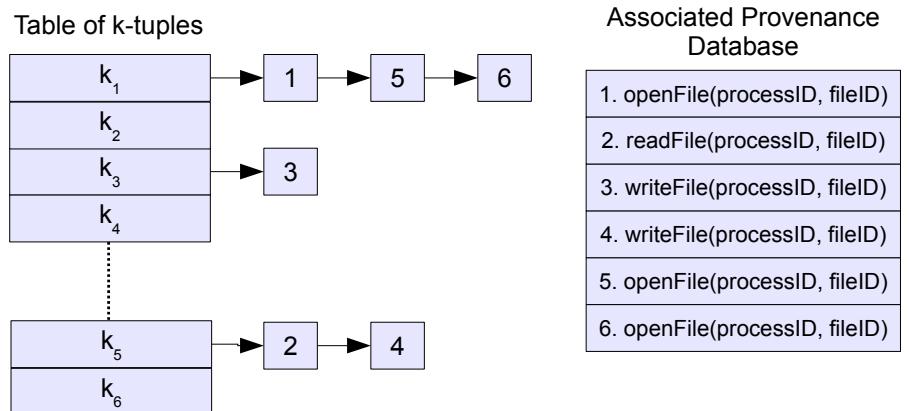
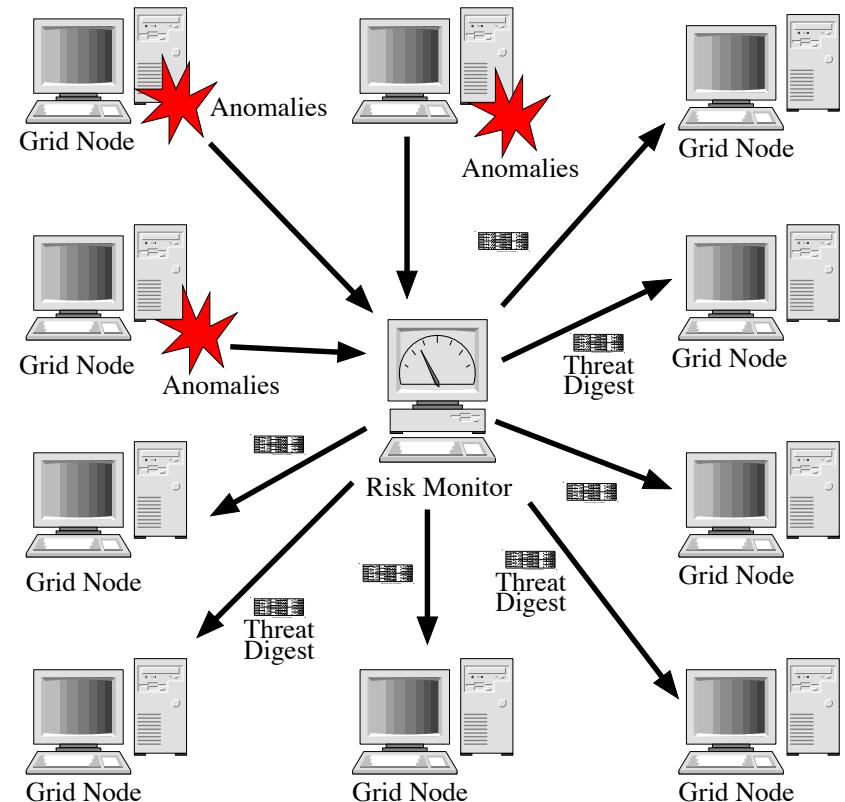
- Application-specific provenance
- Tracking authorship of video mashups
- Custom data model, schema
- *In-band encoding of metadata*
- **VEIL: A System for Certifying Video Provenance, IEEE Symposium on Multimedia, 2007**



- Initial distributed provenance effort
- *Decoupled metadata from source*
- **Bonsai: Balanced Lineage Authentication, Annual Computer Security Applications Conference, 2007**
- **Tracking and Sketching Distributed Data Provenance, IEEE Conference on e-Science, 2010**
- **Mendel: Efficiently Verifying the Lineage of Data Modified in Multiple Trust Domains, ACM Symposium on High Performance Distributed Computing, 2010**

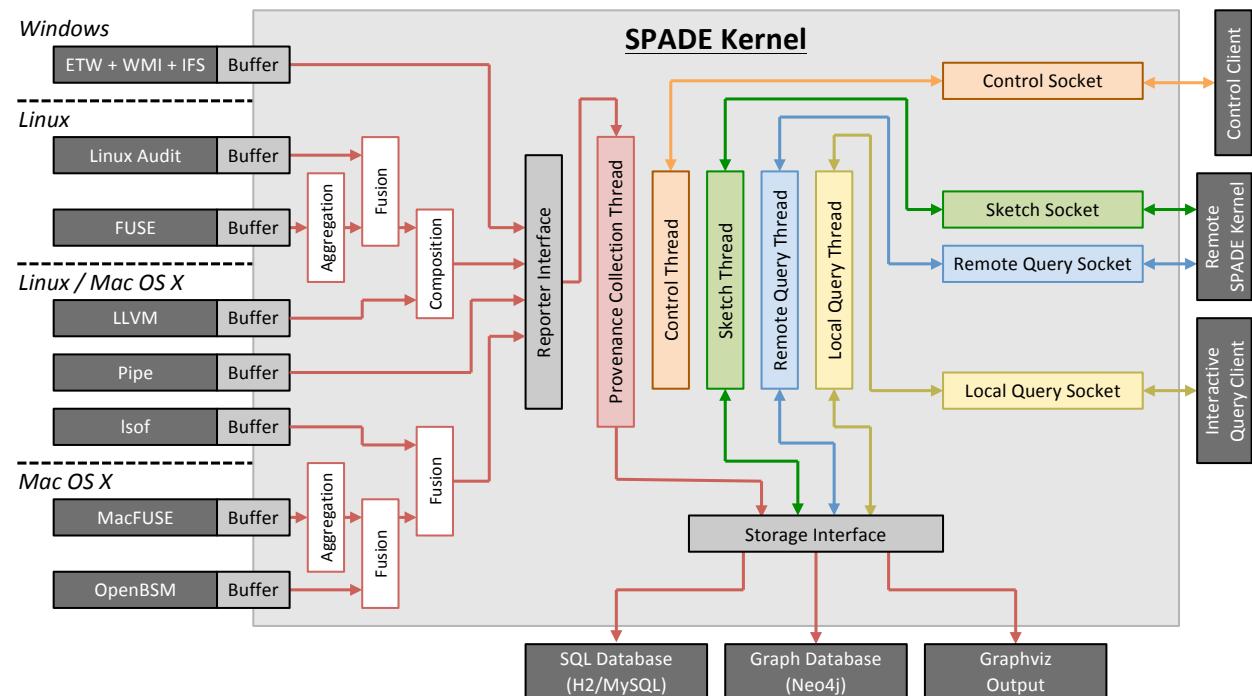
## Precursors (2/2)

- Early focus on cluster / Grid environments
- Influenced by DARPA Application Communities program
- *Relating anomalies to provenance*
- **Steps Toward Managing Lineage Metadata in Grid Clusters, USENIX Theory and Practice of Provenance, 2009**
- **Fine-Grained Tracking of Grid Infections, ACM/IEEE Conference on Grid Computing, 2010**
- **Identifying the Provenance of Correlated Anomalies, ACM Symposium on Applied Computing, 2011**



# SPADE (version 2)

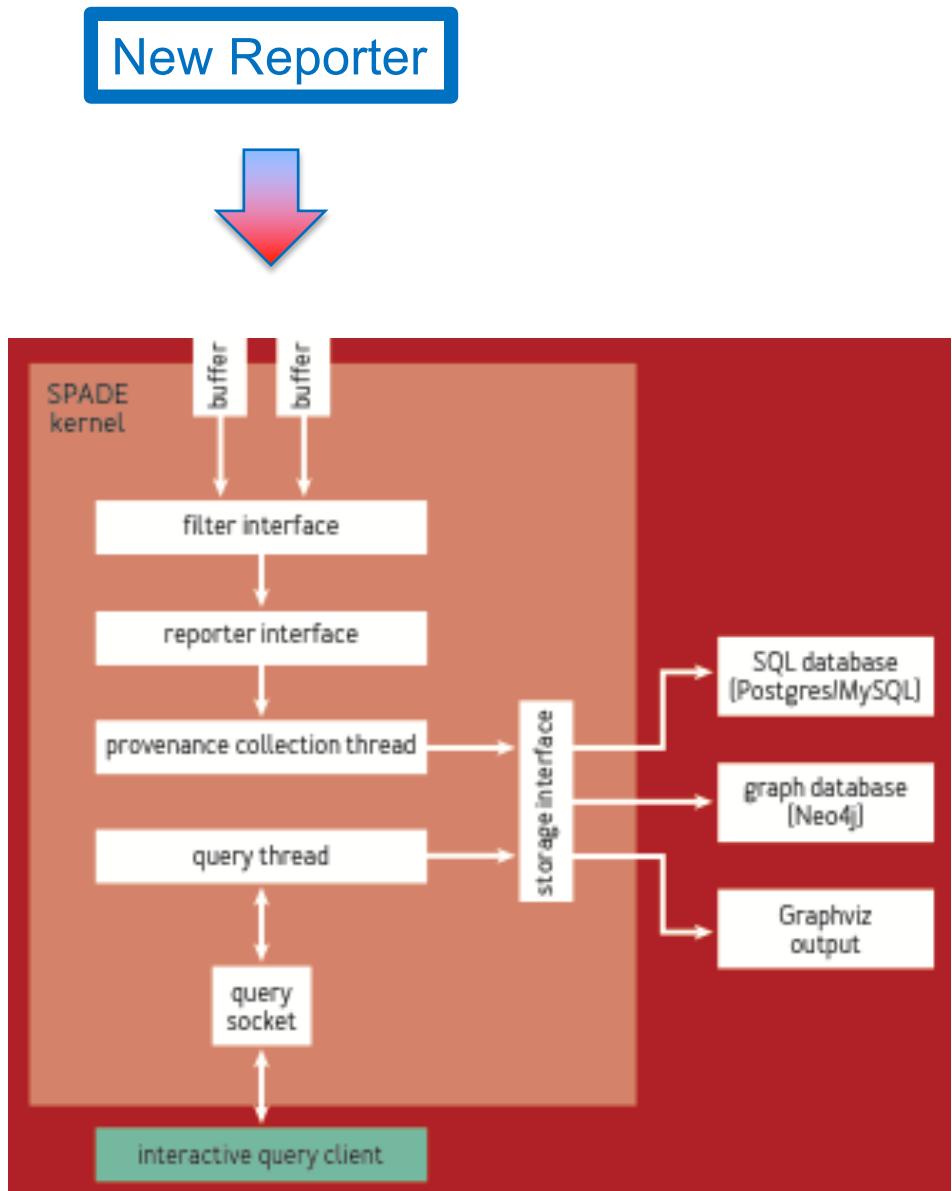
- Motivated by development, deployment experiences
- Re-architected, re-implemented to accommodate:
  - Diverse domains
  - Evolving attributes
  - Variable granularity
  - *Component decoupling*



- **SPADE: Support for Provenance Auditing in Distributed Environments, ACM/IFIP/USENIX Conference on Middleware, 2012**

# New Domain Workflow

- Study application
- Identify significant agents, activities, entities
- Build *causal model* that relates elements
- Create / configure *instrumentation*
- Develop a *SPADE Reporter* to:
  - Ingest event stream
  - Infer provenance
  - Emit *property graph* elements

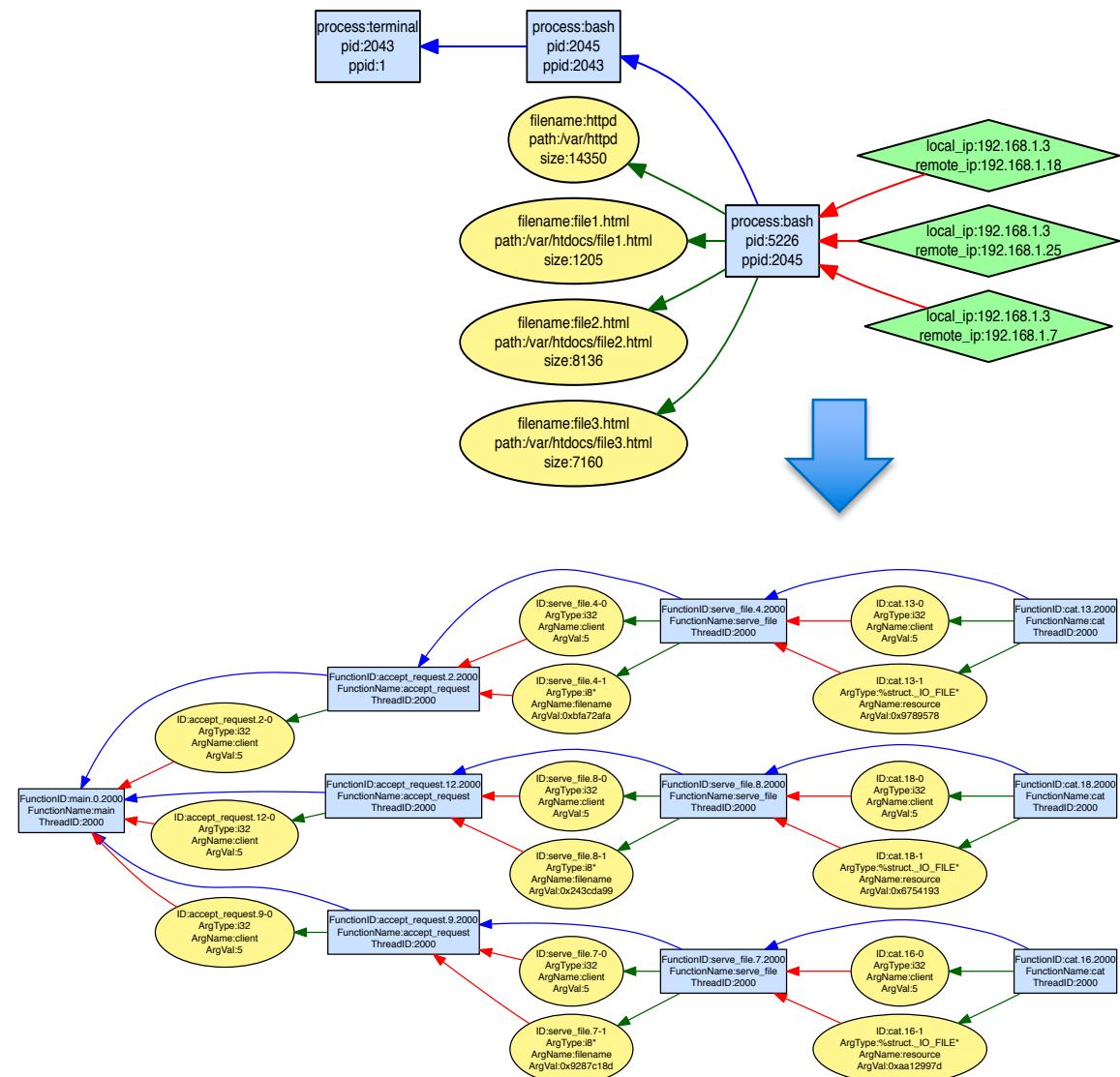


# Looking Inside

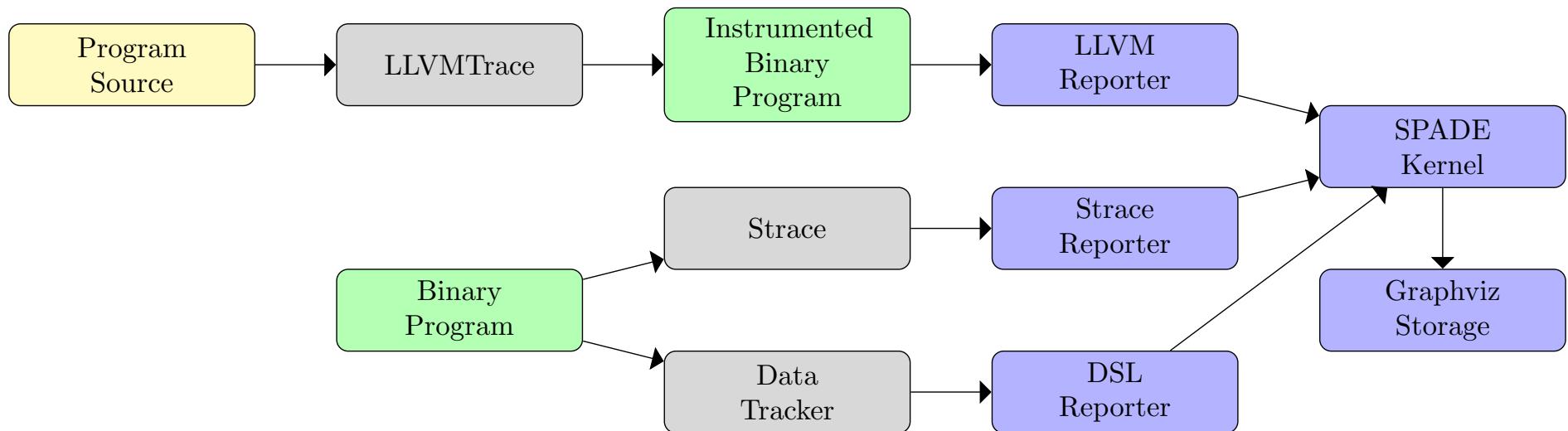
- *Dependency conflation* arises when:
  - Instrumentation is at coarser level of abstraction
  - Causality manifests at finer granularity
- Compiler instrumentation supports intra-process observation

Multiple abstraction levels

Towards Automated Collection of Application-Level Data Provenance, USENIX Theory and Practice of Provenance, 2012



# Comparing Approaches

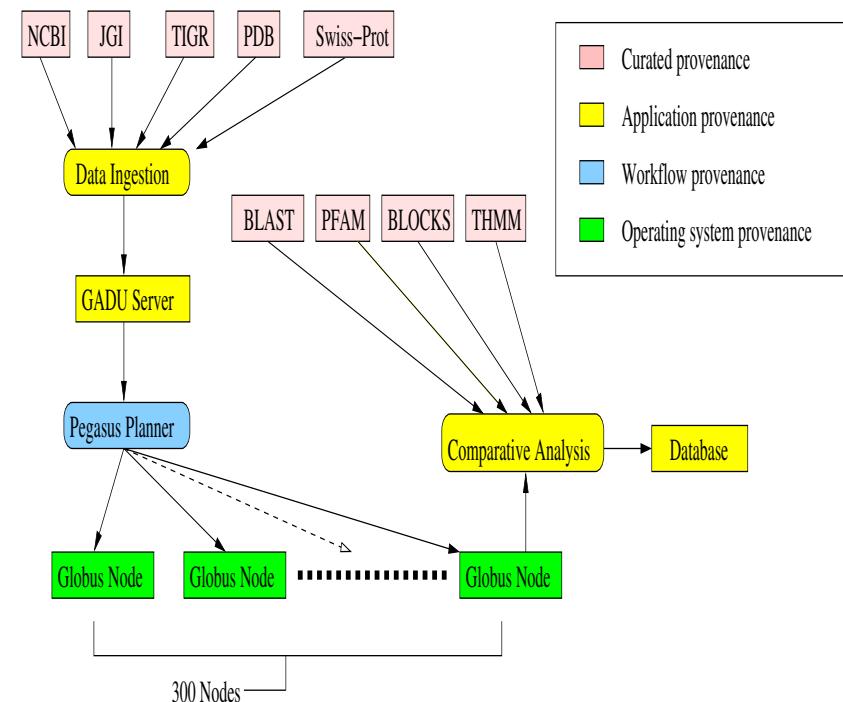


- **Tradeoffs in Automatic Provenance Capture, International Provenance and Annotation Workshop, 2016**

	system call analysis	static, compile-time instrumentation	dynamic, instruction-level instrumentation
integration effort	easy	medium	easy
prov. granularity <sup>6</sup>	file-level	function-level	byte-level
analysis scope	process and children	process, no dyn. lib.	process and children
false positives	many	depends on configured scope	negligible, tracks use of individual bytes
execution overhead	depends on the size of program I/O	depends on the number of function calls	high, depends on the taint tag type used
Reporter	strace reporter	LLVMTrace	DataTracker

# Integrating Provenance

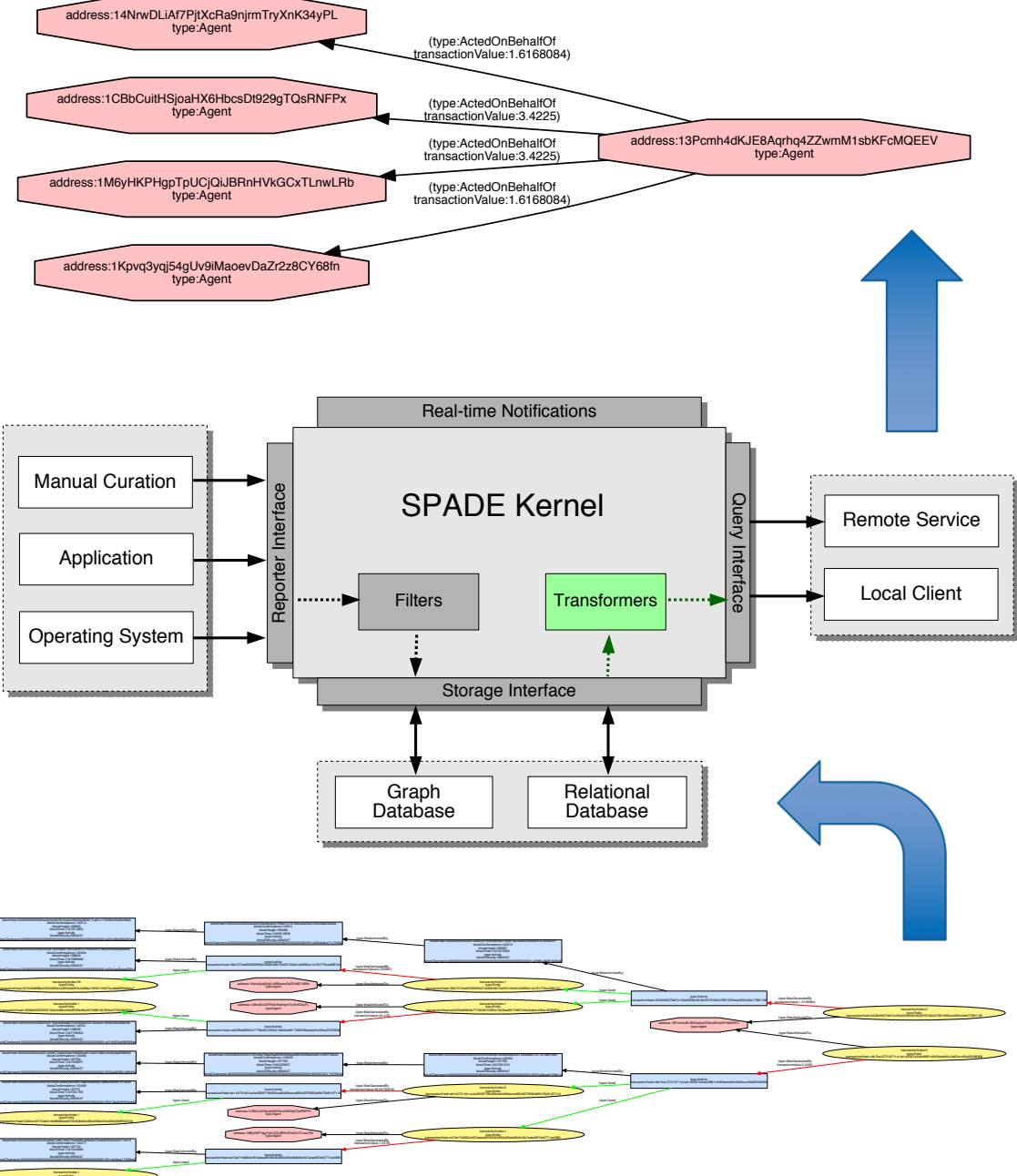
- Merging streams with *filters*
  - Aggregation (in time)
  - Fusion (of complementary sources)
  - Composition (from different layers)
- Policy-based integration
  - Facilitates *what-if* analysis
- For graph abstraction
  - Integration constraints
    - Account for influence of agents on activities, entities
  - Attribution fidelity controlled by:
    - Threshold of matching
    - Trust tolerance



- **Policy-Based Integration of Provenance Metadata, IEEE Symposium on Policies for Distributed Systems and Networks, 2011**
- **Provenance-Only Integration, USENIX Theory and Practice of Provenance, 2014**

# Scaling

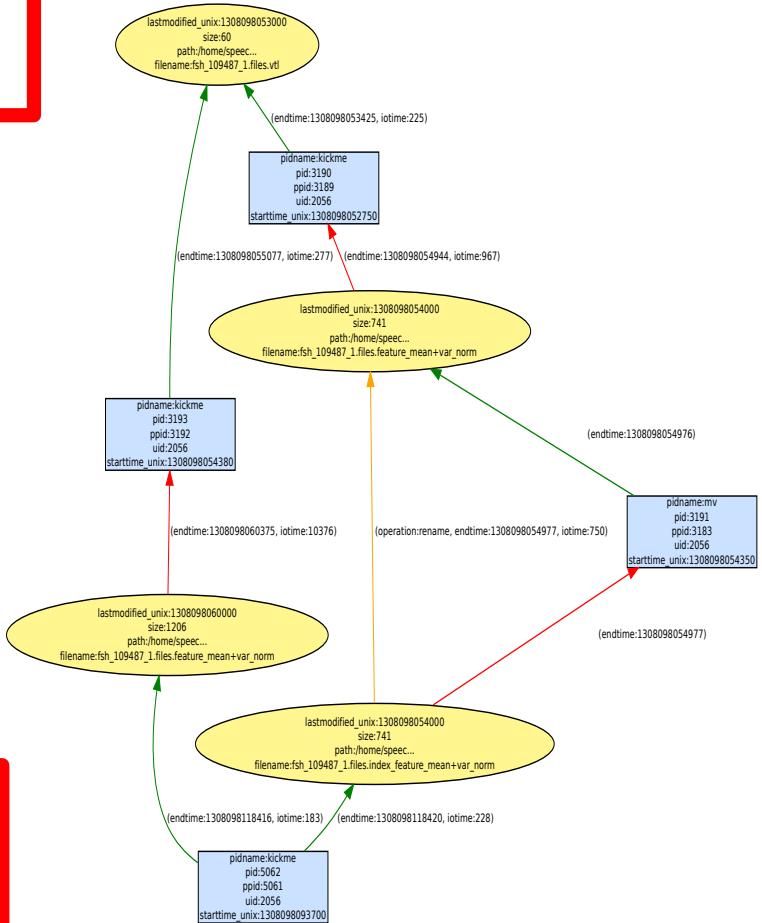
- ``Big Provenance'':
  - Bitcoin blockchain
  - Audit logs
- *Transformers*
  - Limit abstraction scope
  - Operate at query time
  - Dynamic graph rewrite
- **Scaling SPADE to "Big Provenance"**, *USENIX Theory and Practice of Provenance*, 2016
- **Streaming Provenance Compression**, *Lecture Notes in Computer Science*, Vol. 11017, Springer, 2018



# Querying

Intuitionistic logic

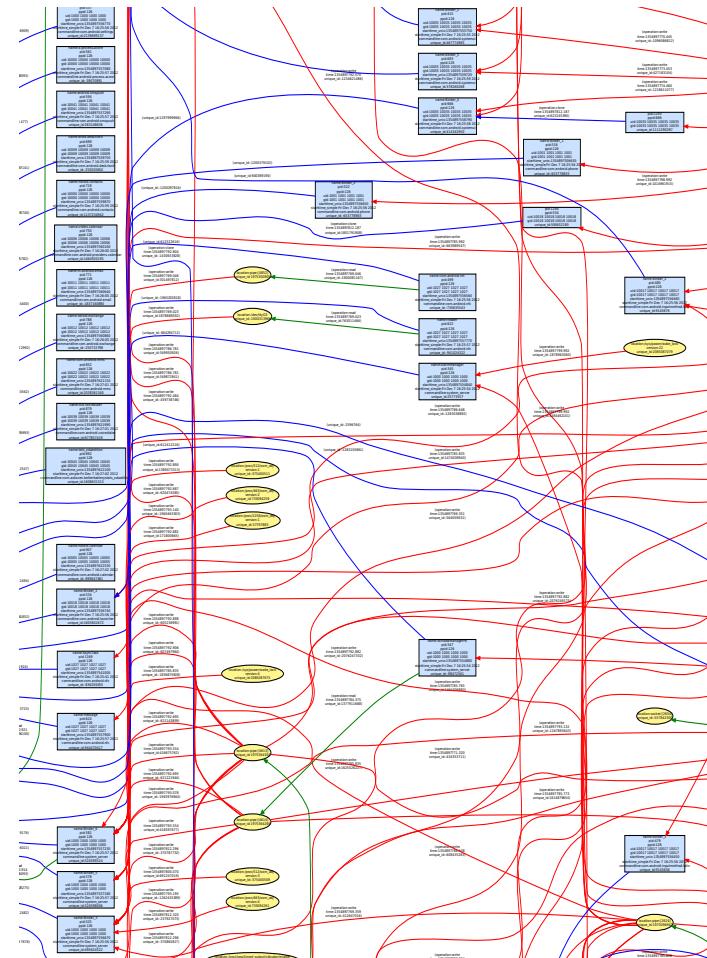
- **System Support for Forensic Inference, Advances in Digital Forensics V, 2009**
- **Efficient Querying of Distributed Provenance Stores, ACM Challenges of Large Applications in Distributed Environments, 2010**
- **Declaratively Processing Provenance Metadata, USENIX Theory and Practice of Provenance, 2013**
- **ProvMark: A Provenance Expressiveness Benchmarking System, ACM/IFIP Middleware Conference, 2019**
- **Digging Into "Big Provenance" (With SPADE), Communications of the ACM, Vol. 64(12), 2021**



Rich query surface  
(supports faceted search, set operations,  
aggregate statistics on big data)

# Diagnostics

- **Android Provenance: Diagnosing Device Disorders, USENIX Theory and Practice of Provenance, 2013**
  - **Discrepancy Detection in Whole Network Provenance, USENIX Theory and Practice of Provenance, 2020**
  - **Clarion: Sound and Clear Provenance Tracking for Microservice Deployments, USENIX Security Symposium, 2021**

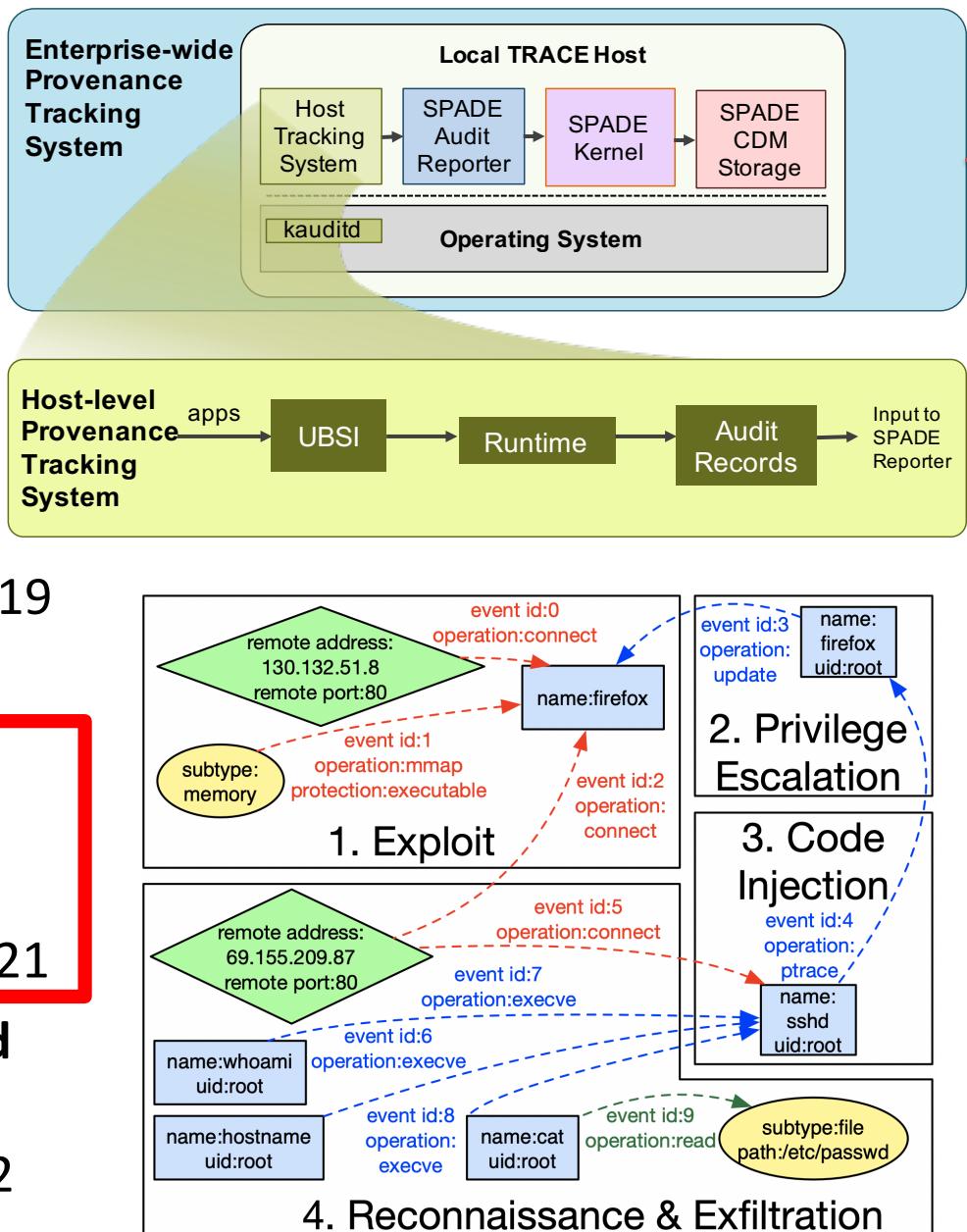


# Security

- **Using Provenance Patterns to Vet Sensitive Behaviors in Android Apps,** *Conference on Security and Privacy in Communication Networks*, 2015
- **Mining Data Provenance to Detect Advanced Persistent Threats**, *USENIX Theory and Practice of Provenance*, 2019

Partial observability (facilitates scaling)

- **TRACE: Enterprise-Wide Provenance Tracking For Real-Time APT Detection**, *IEEE Transactions on Information Forensics and Security*, 2021
- **PACED: Provenance-based Automated Container Escape Detection**, *IEEE Conference on Cloud Engineering*, 2022



# Impact

- Research Infrastructure
  - Competing concerns (community use / design iteration)
  - 100+ GitHub stars / 60+ forks
  - Anecdotal: Used in software build / staging
- Academic
  - 250+ citations
  - Anecdotal: Used to create other systems
- Datasets
  - Provenance Benchmark Challenge
  - DARPA Transparent Computing Adversarial Engagements (3 & 5)
- Industry
  - Streamlined + extended version licensed to AccuKnox (container security company)