SPADE: A retrospective

Ashish Gehani, SRI

Background

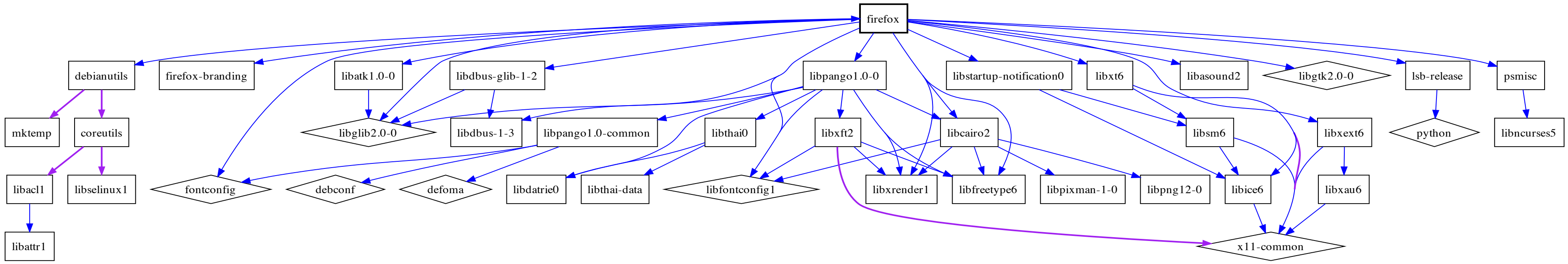
- Pre-20th century:
  - Experimental science
    - Hypotheses derived from experience
    - Physical phenomena measured
    - Steps and data recorded by hand
  - Theoretical science
    - Mathematical models
    - Conjectures based on analysis
    - Results derived by hand
- Late 20th century:
  - Computational 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

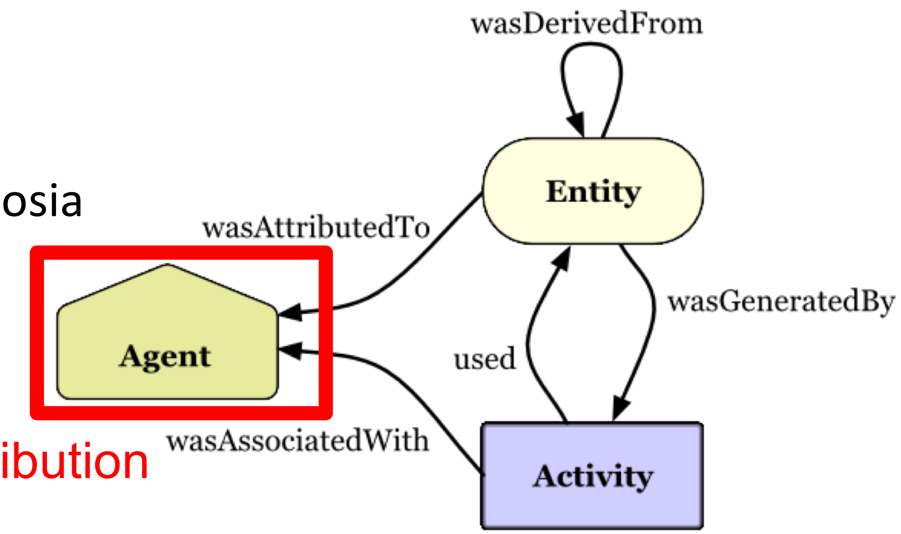


- Environmental dependencies
  - Shell variables
  - Shared memory contents
- Changes in any can affect output

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

Credit: www.w3.org/TR/prov-primer/

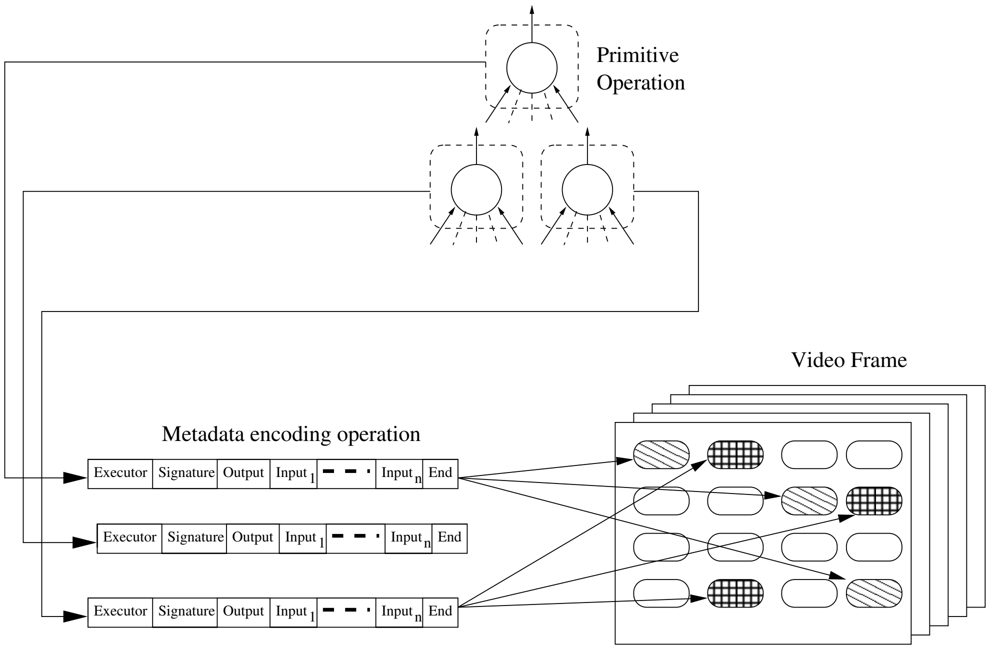


Attribution

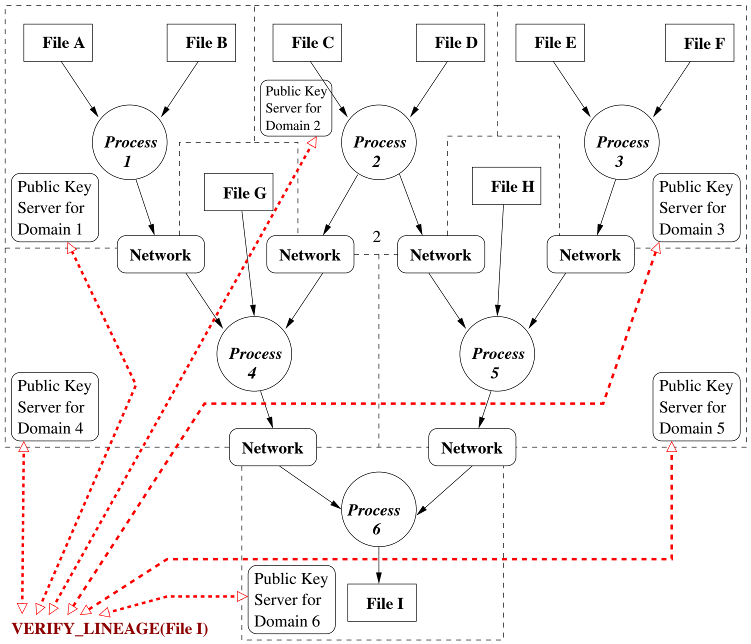
- Emerging specifications:
  - 2007, 2011 : **Open Provenance Model** (versions 1.0, 1.1)
  - 2013 : **W3C PROV** standard
  - 2015-2019 : DARPA Transparent Computing **Common Data Model** (versions 1-20)
- Ongoing event series:
  - 2006- : Biennial **International Provenance and Annotation Workshop**
  - 2009- : Annual **USENIX Theory and Practice of Provenance**
  - 2014- : Biennial **ProvenanceWeek** co-located events

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



Primitive Operation
Video Frame
Metadata encoding operation
Executor | Signature | Output | Input 1 | Input n | End
Executor | Signature | Output | Input 1 | Input n | End
Executor | Signature | Output | Input 1 | Input n | End



File A
File B
File C
File D
File E
File F
Public Key Server for Domain 2
Process 1
Process 2
Process 3
Public Key Server for Domain 1
File G
File H
Public Key Server for Domain 3
Network
Network
2
Network
Network
Process 4
Process 5
Public Key Server for Domain 4
Public Key Server for Domain 5
Network
Network
Process 6
Public Key Server for Domain 6
File I
VERIFY_LINEAGE(File I)

- 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

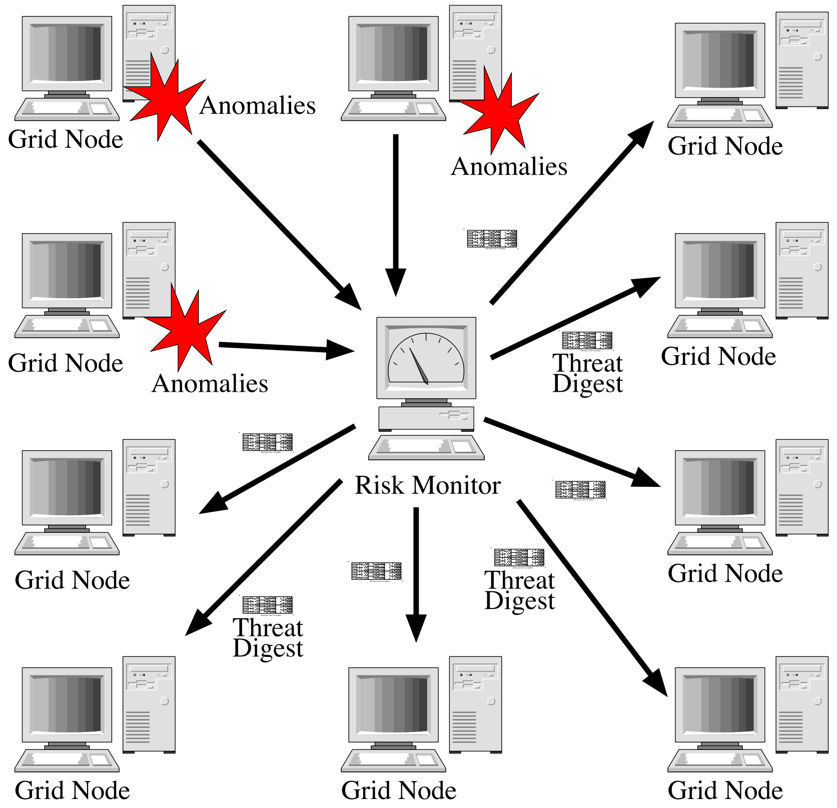
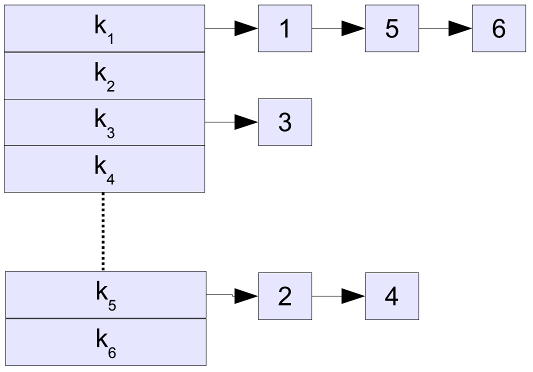


Table of k-tuples

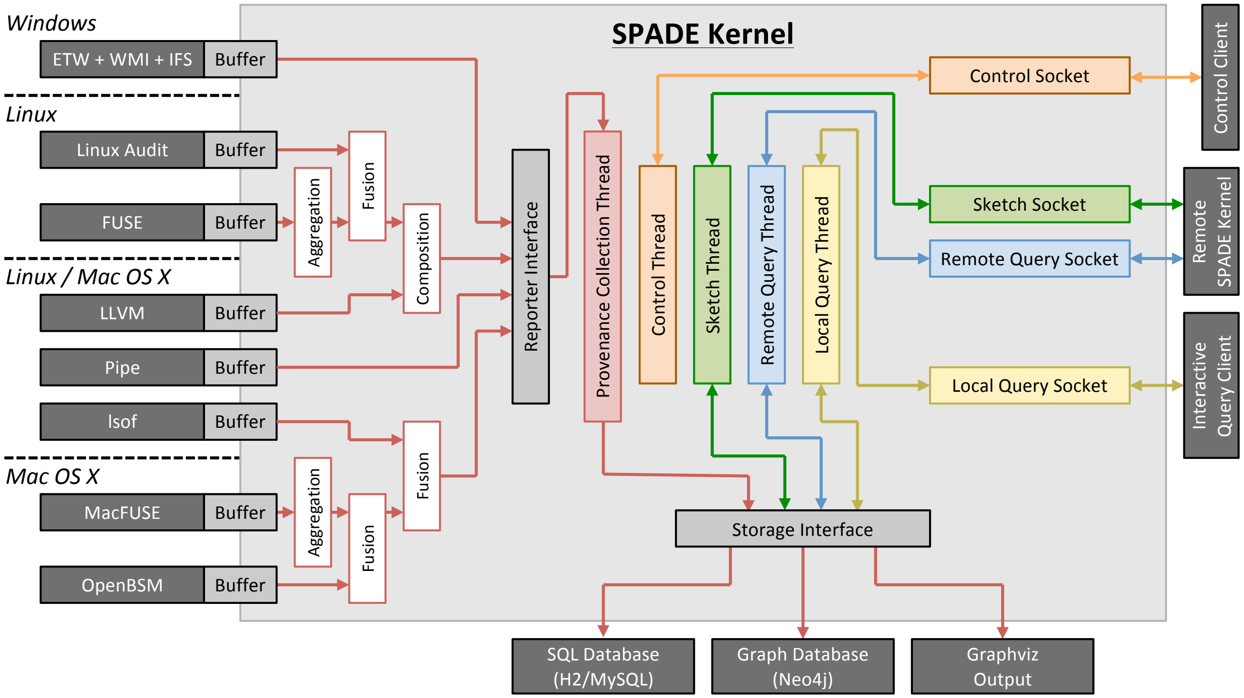


Associated Provenance Database

1. openFile(processID, fileID)
2. readFile(processID, fileID)
3. writeFile(processID, fileID)
4. writeFile(processID, fileID)
5. openFile(processID, fileID)
6. openFile(processID, fileID)

SPADE (version 2)

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

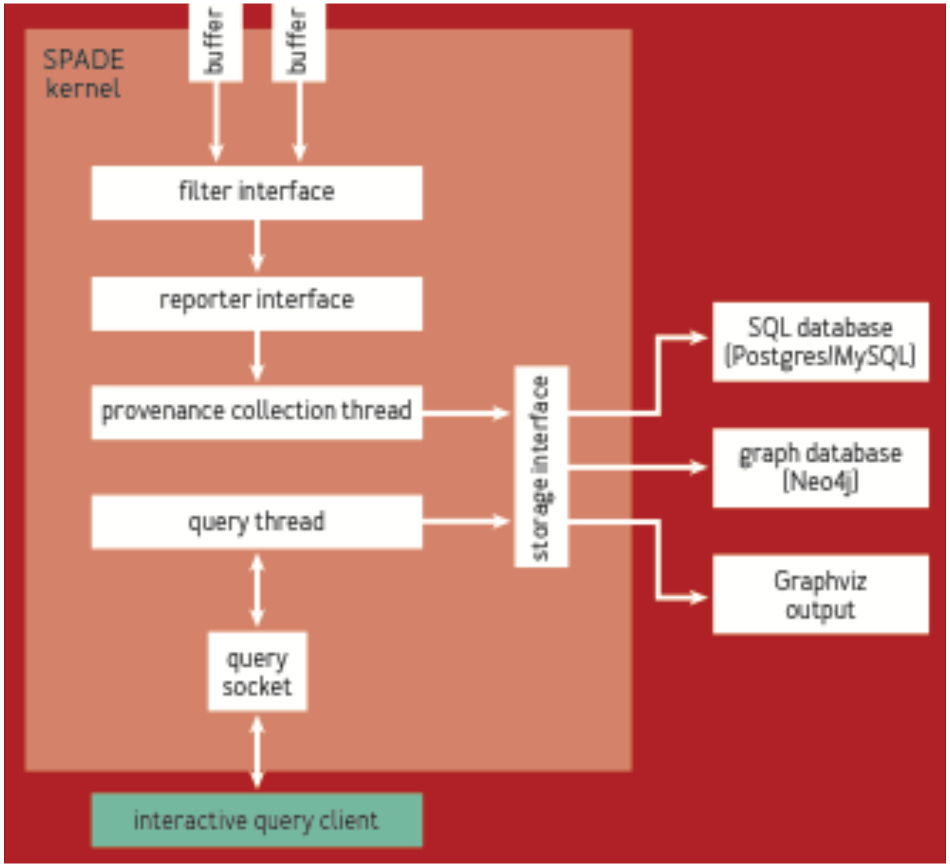


- **SPADE: Support for Provenance Auditing in Distributed Environments**, *ACM/IFIP/USENIX Conference on Middleware*, 2012

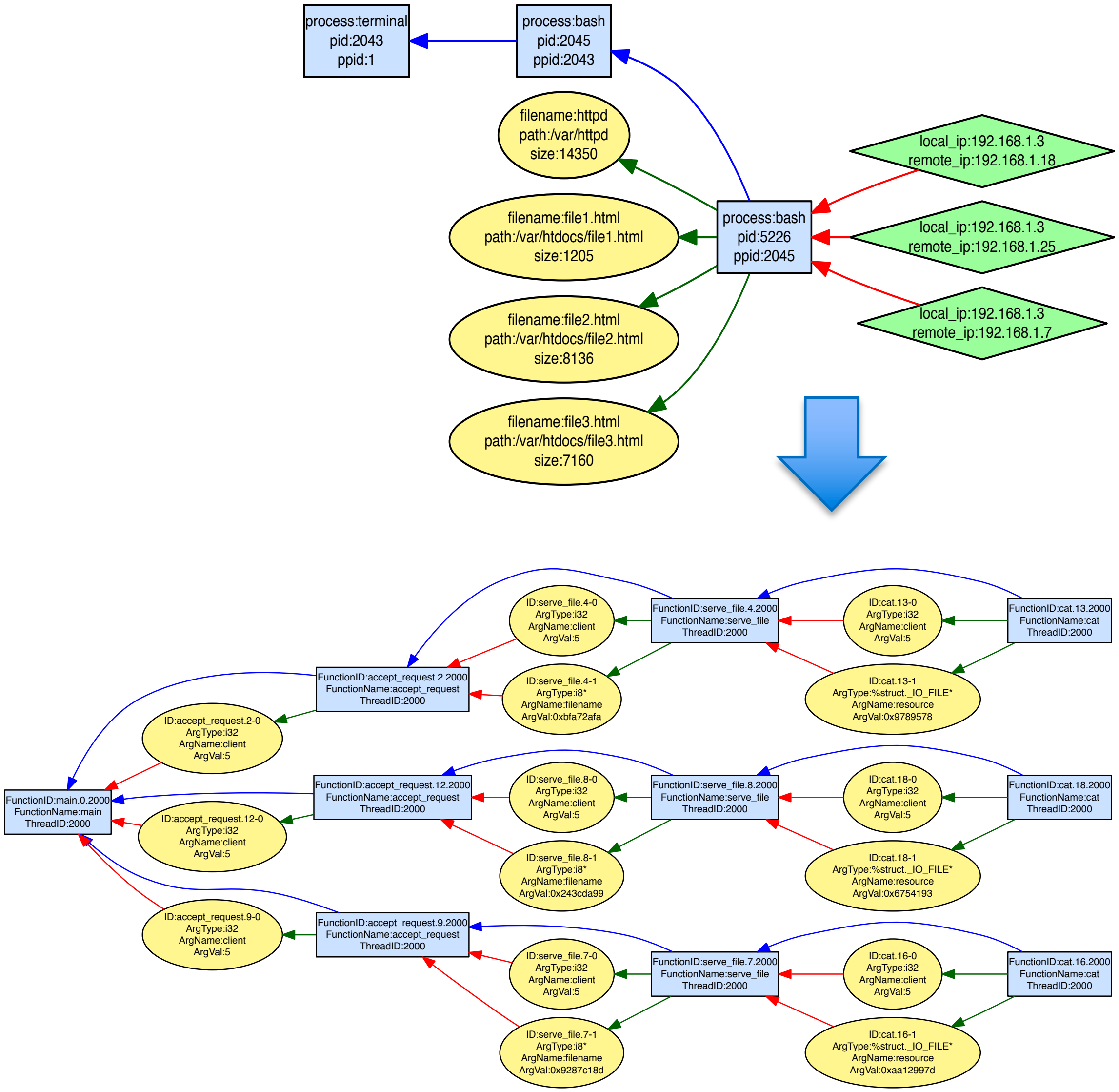
New Domain Workflow

- 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

New Reporter



Looking Inside



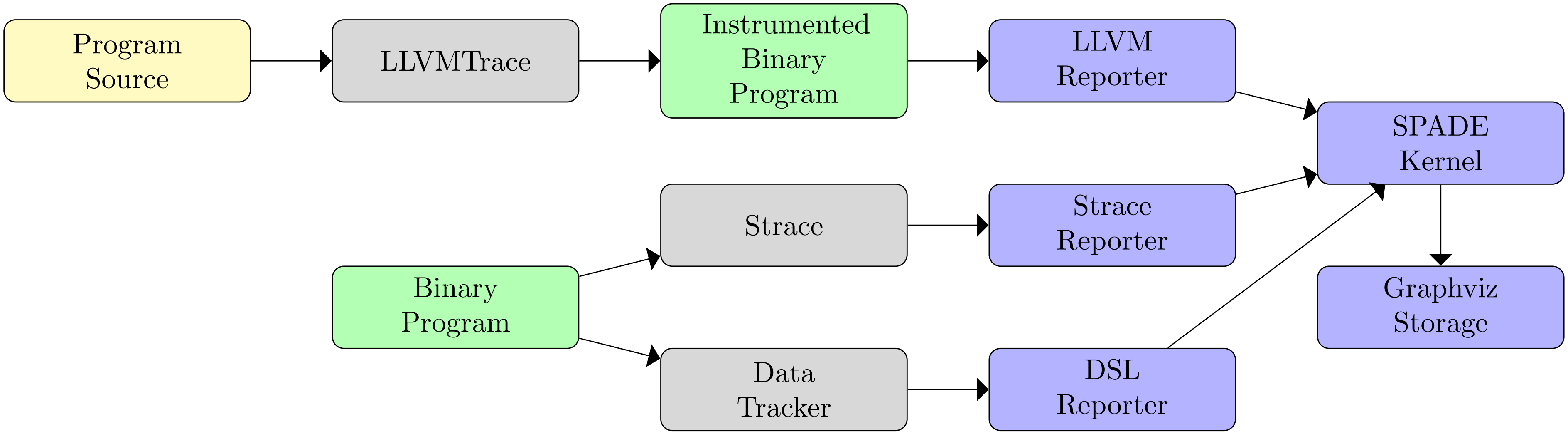
process:terminal pid:2043 ppid:1
process:bash pid:2045 ppid:2043
filename:httpd path:/var/httpd size:14350
filename:file1.html path:/var/htdocs/file1.html size:1205
filename:file2.html path:/var/htdocs/file2.html size:8136
filename:file3.html path:/var/htdocs/file3.html size:7160
process:bash pid:5226 ppid:2045
local_ip:192.168.1.3 remote_ip:192.168.1.18
local_ip:192.168.1.3 remote_ip:192.168.1.25
local_ip:192.168.1.3 remote_ip:192.168.1.7
FunctionID:main.0.2000 FunctionName:main ThreadID:2000
ID:accept_request.2-0 ArgType:i32 ArgName:client ArgVal:5
ID:accept_request.12-0 ArgType:i32 ArgName:client ArgVal:5
ID:accept_request.9-0 ArgType:i32 ArgName:client ArgVal:5
FunctionID:accept_request.2.2000 FunctionName:accept_request ThreadID:2000
FunctionID:accept_request.12.2000 FunctionName:accept_request ThreadID:2000
FunctionID:accept_request.9.2000 FunctionName:accept_request ThreadID:2000
ID:serve_file.4-0 ArgType:i32 ArgName:client ArgVal:5
ID:serve_file.4-1 ArgType:i8* ArgName:filename ArgVal:0xbfa72afa
ID:serve_file.8-0 ArgType:i32 ArgName:client ArgVal:5
ID:serve_file.8-1 ArgType:i8* ArgName:filename ArgVal:0x243cda99
ID:serve_file.7-0 ArgType:i32 ArgName:client ArgVal:5
ID:serve_file.7-1 ArgType:i8* ArgName:filename ArgVal:0x9287c18d
FunctionID:serve_file.4.2000 FunctionName:serve_file ThreadID:2000
FunctionID:serve_file.8.2000 FunctionName:serve_file ThreadID:2000
FunctionID:serve_file.7.2000 FunctionName:serve_file ThreadID:2000
ID:cat.13-0 ArgType:i32 ArgName:client ArgVal:5
ID:cat.13-1 ArgType:%struct._IO_FILE* ArgName:resource ArgVal:0x9789578
ID:cat.18-0 ArgType:i32 ArgName:client ArgVal:5
ID:cat.18-1 ArgType:%struct._IO_FILE* ArgName:resource ArgVal:0x6754193
ID:cat.16-0 ArgType:i32 ArgName:client ArgVal:5
ID:cat.16-1 ArgType:%struct._IO_FILE* ArgName:resource ArgVal:0xaa12997d
FunctionID:cat.13.2000 FunctionName:cat ThreadID:2000
FunctionID:cat.18.2000 FunctionName:cat ThreadID:2000
FunctionID:cat.16.2000 FunctionName:cat ThreadID:2000

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

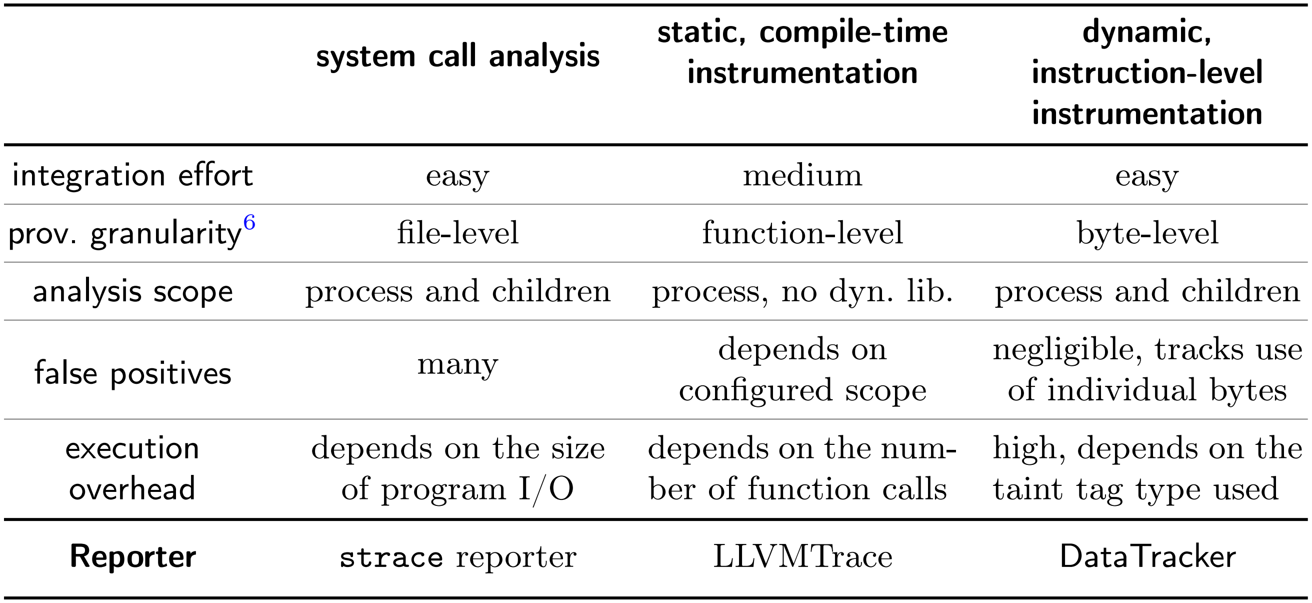
Multiple abstraction levels

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

Comparing Approaches



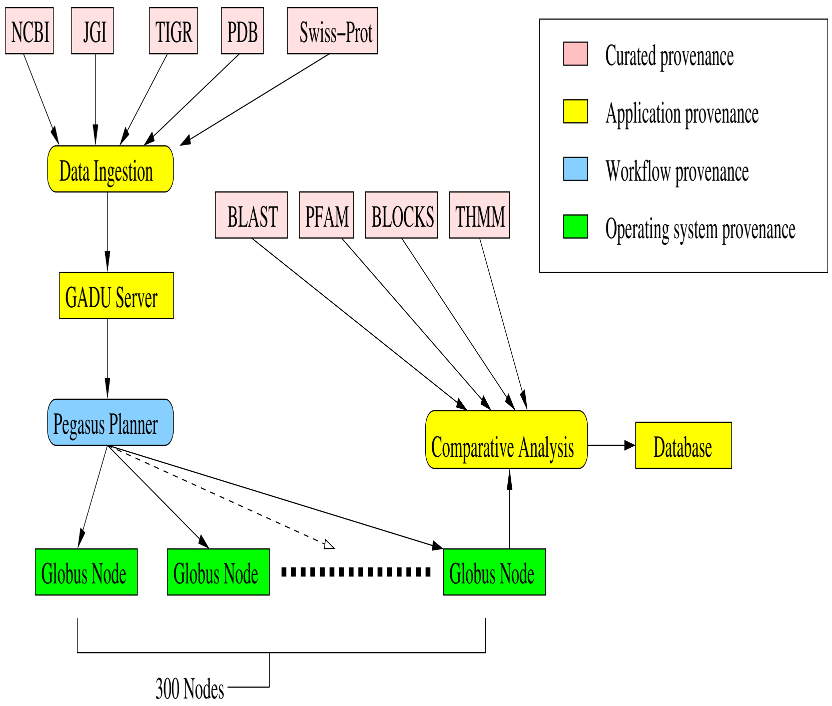
- **Tradeoffs in Automatic Provenance Capture,** *International Provenance and Annotation Workshop*, 2016



	system call analysis	**static, compile-time instrumentation**	**dynamic, instruction-level instrumentation**
integration effort	easy	medium	easy
prov. granularity[6]	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

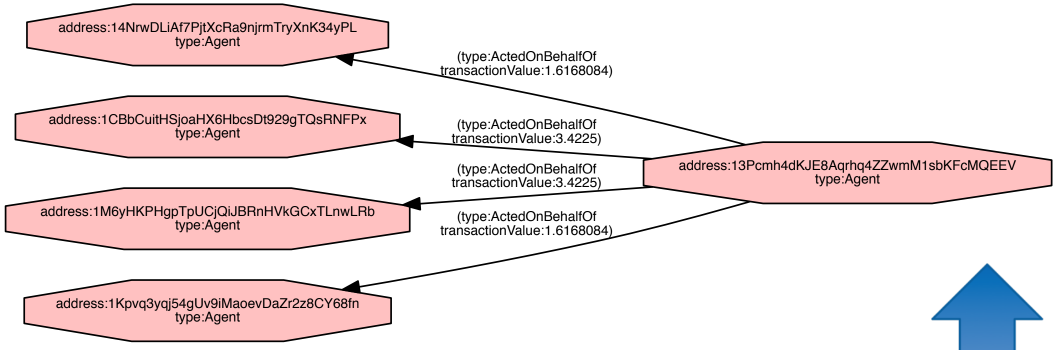
- Merging streams with *filters*
  - Aggregation (in time)
  - Fusion (of complementary sources)
  - Composition (from different layers)
- Policy-based integration
  - Facilitates *what-if* analysis
- For graph abstraction
  - Integration constraints
    - Account for influence of agents on activities,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

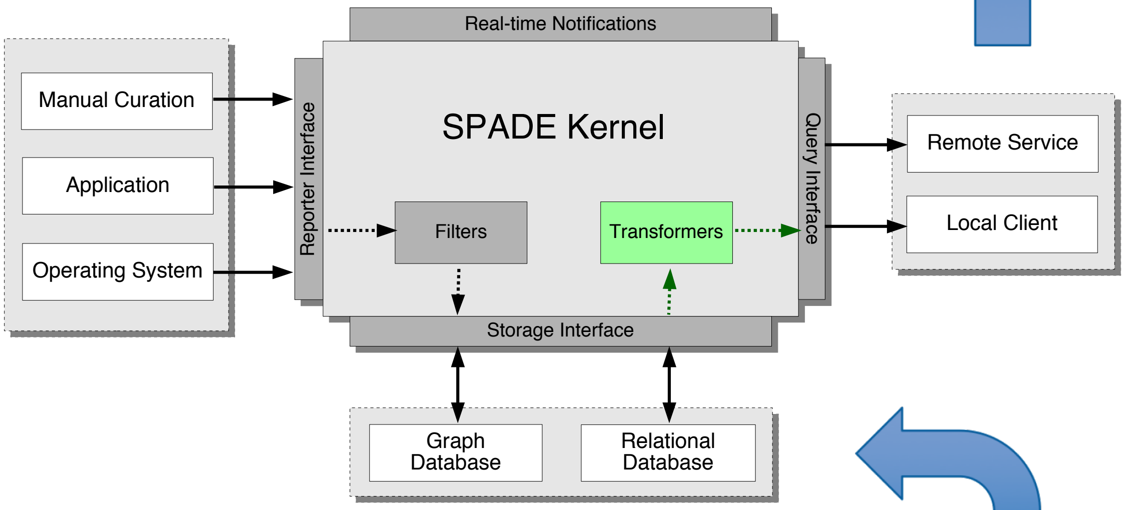


- **Policy-Based Integration of Provenance Metadata,** *IEEE Symposium on Policies for Distributed Systems and Networks*, 2011
- **Provenance-Only Integration**, *USENIX Theory and Practice of Provenance*, 2014

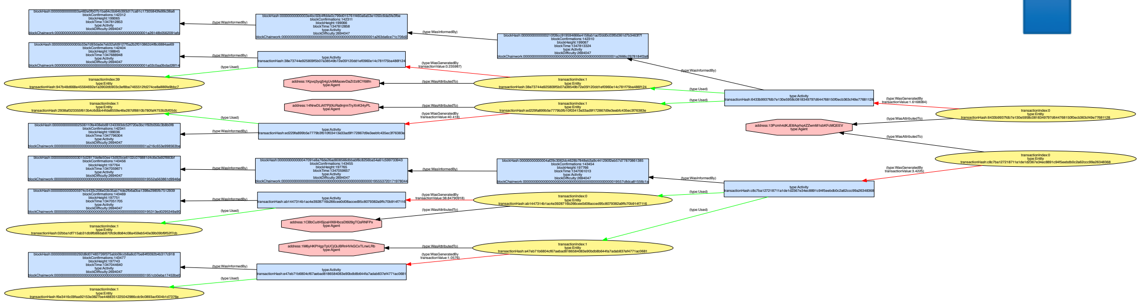
Scaling

- ``Big Provenance'':
  - Bitcoin blockchain
  - Audit logs
- *Transformers*
  - Limit abstraction scope
  - Operate at query time
  - Dynamic graph rewrite


address:14NrwDLiAf7PjtXcRa9njrmTryXnK34yPL
type:Agent
address:1CBbCuitHSjoaHX6HbcsDt929gTQsRNFPx
type:Agent
address:1M6yHKPHgpTpUCjQiJBRnHVkGCxTLnwLRb
type:Agent
address:1Kpvq3yqj54gUv9iMaoevDaZr2z8CY68fn
type:Agent
address:13Pcmh4dKJE8Aqrhq4ZZwmM1sbKFcMQEEV
type:Agent
(type:ActedOnBehalfOf
transactionValue:1.6168084)
(type:ActedOnBehalfOf
transactionValue:3.4225)
(type:ActedOnBehalfOf
transactionValue:3.4225)
(type:ActedOnBehalfOf
transactionValue:1.6168084)



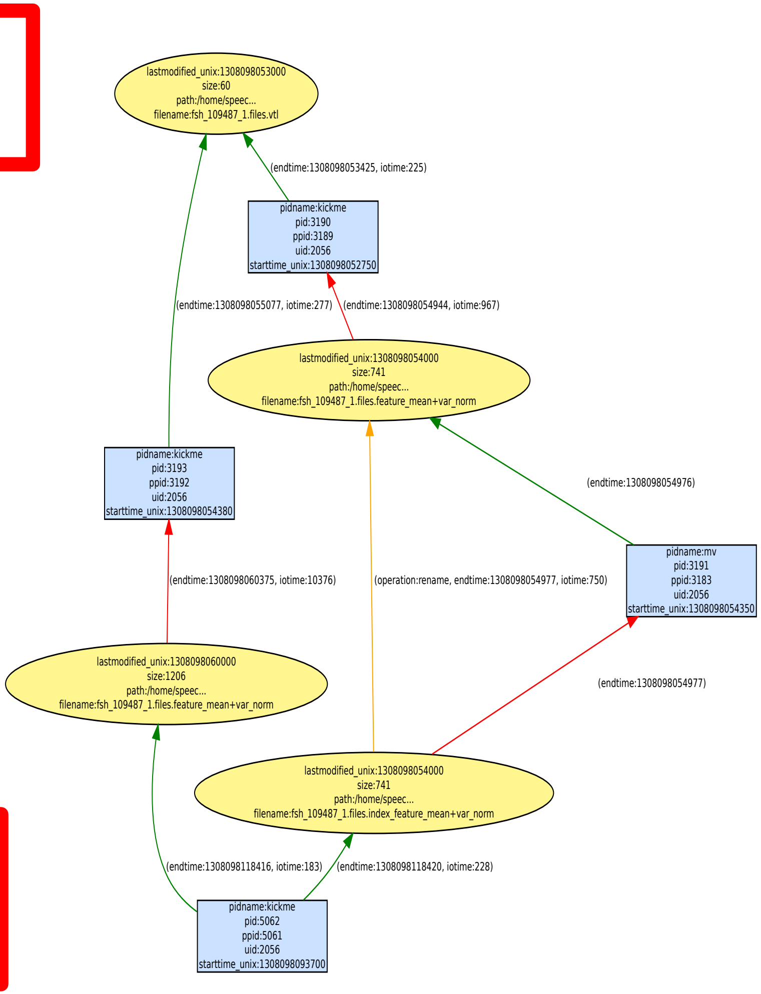
- **Scaling SPADE to "Big Provenance"**, *USENIX Theory an Practice of Provenance*, 2016
- **Streaming Provenance Compression**, *Lecture Notes in Computer Science*, Vol. 11017, Springer, 2018



Querying

Intuitionistic logic

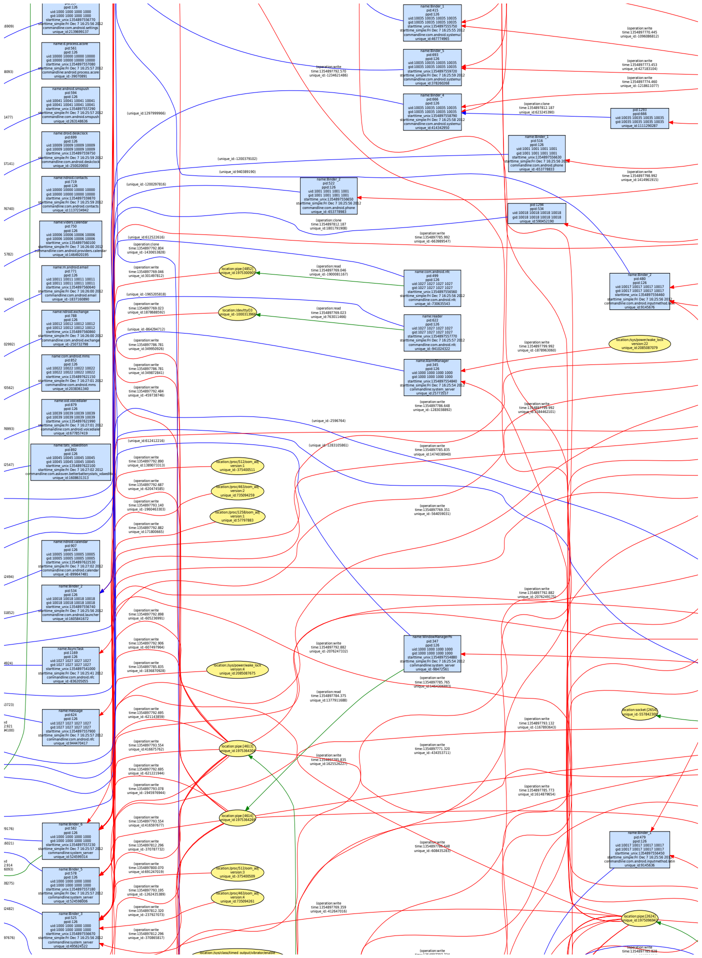
- **System Support for Forensic Inference**, *Advances in Digital Forensics V*, 2009
- **Efficient Querying of Distributed Provenance Stores**, *Challenges of Large Applications in Distributed Environments*, 2010
- **Declaratively Processing Provenance Metadata**, *USENIX Theory and Practice of Provenance*, 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


lastmodified_unix:1308098053000
size:60
path:/home/speec...
filename:fsh_109487_1.files.vtl
(endtime:1308098053425, iotime:225)
pidname:kickme
pid:3190
ppid:3189
uid:2056
starttime_unix:1308098052750
(endtime:1308098055077, iotime:277)
(endtime:1308098054944, iotime:967)
lastmodified_unix:1308098054000
size:741
path:/home/speec...
filename:fsh_109487_1.files.feature_mean+var_norm
pidname:kickme
pid:3193
ppid:3192
uid:2056
starttime_unix:1308098054380
(endtime:1308098054976)
pidname:mv
pid:3191
ppid:3183
uid:2056
starttime_unix:1308098054350
(endtime:1308098060375, iotime:10376)
(operation:rename, endtime:1308098054977, iotime:750)
lastmodified_unix:1308098060000
size:1206
path:/home/speec...
filename:fsh_109487_1.files.feature_mean+var_norm
(endtime:1308098054977)
lastmodified_unix:1308098054000
size:741
path:/home/speec...
filename:fsh_109487_1.files.index_feature_mean+var_norm
(endtime:1308098118416, iotime:183)
(endtime:1308098118420, iotime:228)
pidname:kickme
pid:5062
ppid:5061
uid:2056
starttime_unix:1308098093700

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

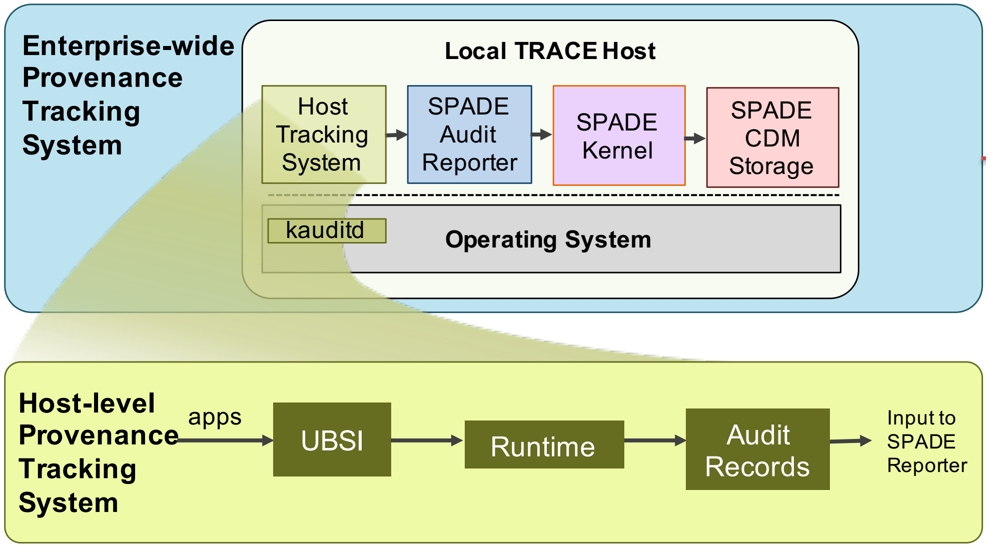


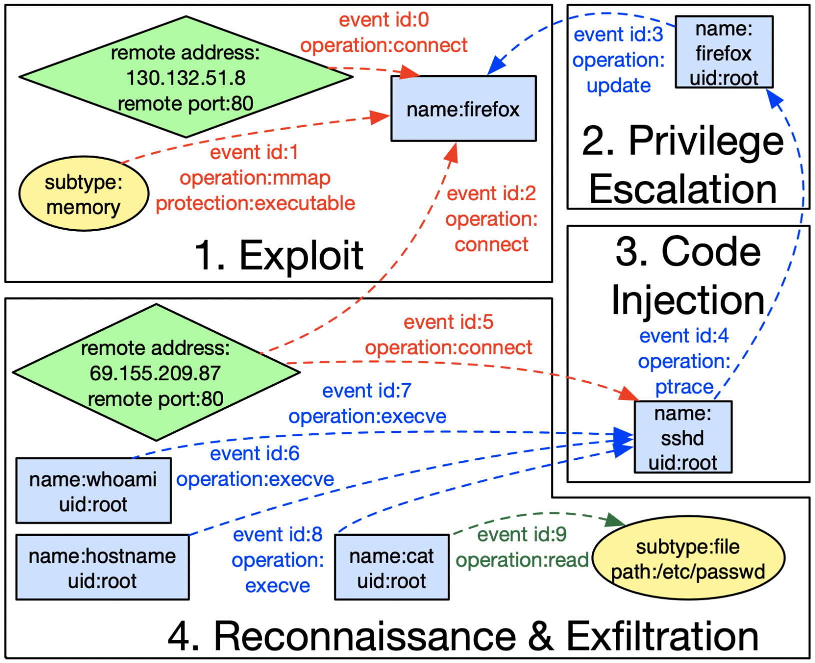
Security

- **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**, *10th IEEE International Conference on Cloud Engineering*, 2022




remote address: 130.132.51.8 remote port:80
event id:0 operation:connect
name:firefox
subtype: memory
event id:1 operation:mmap protection:executable
event id:2 operation: connect
1. Exploit
event id:3 operation: update
name: firefox uid:root
2. Privilege Escalation
3. Code Injection
event id:4 operation: ptrace
name: sshd uid:root
remote address: 69.155.209.87 remote port:80
event id:5 operation:connect
event id:7 operation:execve
event id:6 operation:execve
name:whoami uid:root
name:hostname uid:root
event id:8 operation: execve
name:cat uid:root
event id:9 operation:read
subtype:file path:/etc/passwd
4. Reconnaissance & Exfiltration

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)