(Integrity Justified) Experimental Provenance

Patrick McDaniel, Pennsylvania State University

Workshop on GENI and Security
Davis, CA -- January 22, 2009

Provenance

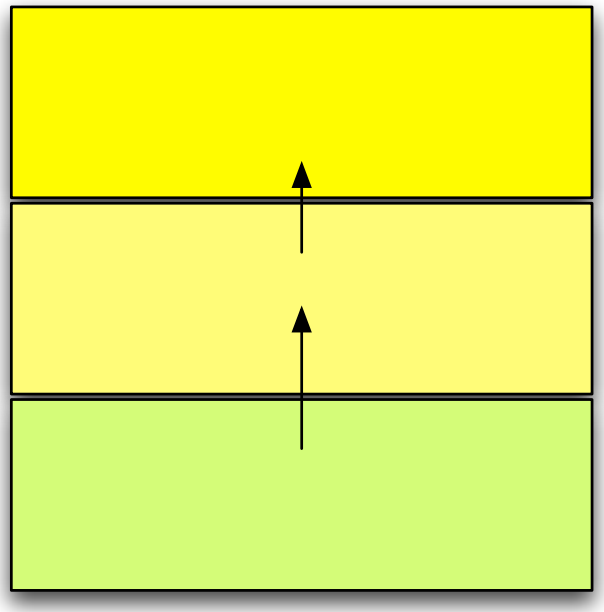
PENNSTATE 1855

- A human scale problem:
  - Data often comes from many sources ...
  - ... is synthesized/influenced by complex/hidden processes ...
  - ... thus, how do you really know what the data means?
- *Data provenance* immutably identifies how data came to be in the state it is.
  - Who/what contributed to it?
  - What was it based on?
  - When was it generated?
  - Why was it generated?
  - How was it generated?

Why GENI provenance?

PENNSTATE 1855

- Error handling
  - Detection, isolation, and recovery
- Source attribution
  - Forensics, consistency, believability
- Experimental Reproducability
  - Extension, instrumentation
- Data revision
  - Updates, correction, extension, refinement
- Evidentiary
  - Evidence that data is legitimate/legal (certification, verification)
- *Experimental data can only be judged in light of how, when and where it comes from*



GENI System Provenance

PENNSTATE 1855

- Assessing system provenance is key to understanding achieving the goals of GENI
  - What software was a component (slice/aggregate) running?
  - What inputs and configuration were used?
  - What security policy was being enforced?
    - e.g., isolation, data protection, privacy
- Stated as experimental *criteria* during the setup/acceptance
  - Think about sensitive experiments: *NCR*-esque, proprietary algoritms, opt-in with personal information
  - Determines apparatus acceptability of validation

GENI adoption requires answers to these questions

Integrity Justified Provenance

- Integrity measurement techniques provide information about the instantaneous state of a system, but *not* its data, or over time, or for other computational elements (VMs)
- What if you could build an aggregate of mutually attesting components that uses that apparatus to attest to the system state, protection state, data, and environment.
  - ... and tie a proof of that aggregate to experimental results.
- Building on the shared reference monitor (Shamon)