



Systems and Internet Infrastructure Security

Network and Security Research Center
Department of Computer Science and Engineering
Pennsylvania State University, University Park PA

Secure Provenance in High-End Computing Systems

Patrick McDaniel, Radu Sion,
Marianne Winslett, Erez Zadok,
HEC FSIO 2009 - August 10, 2009

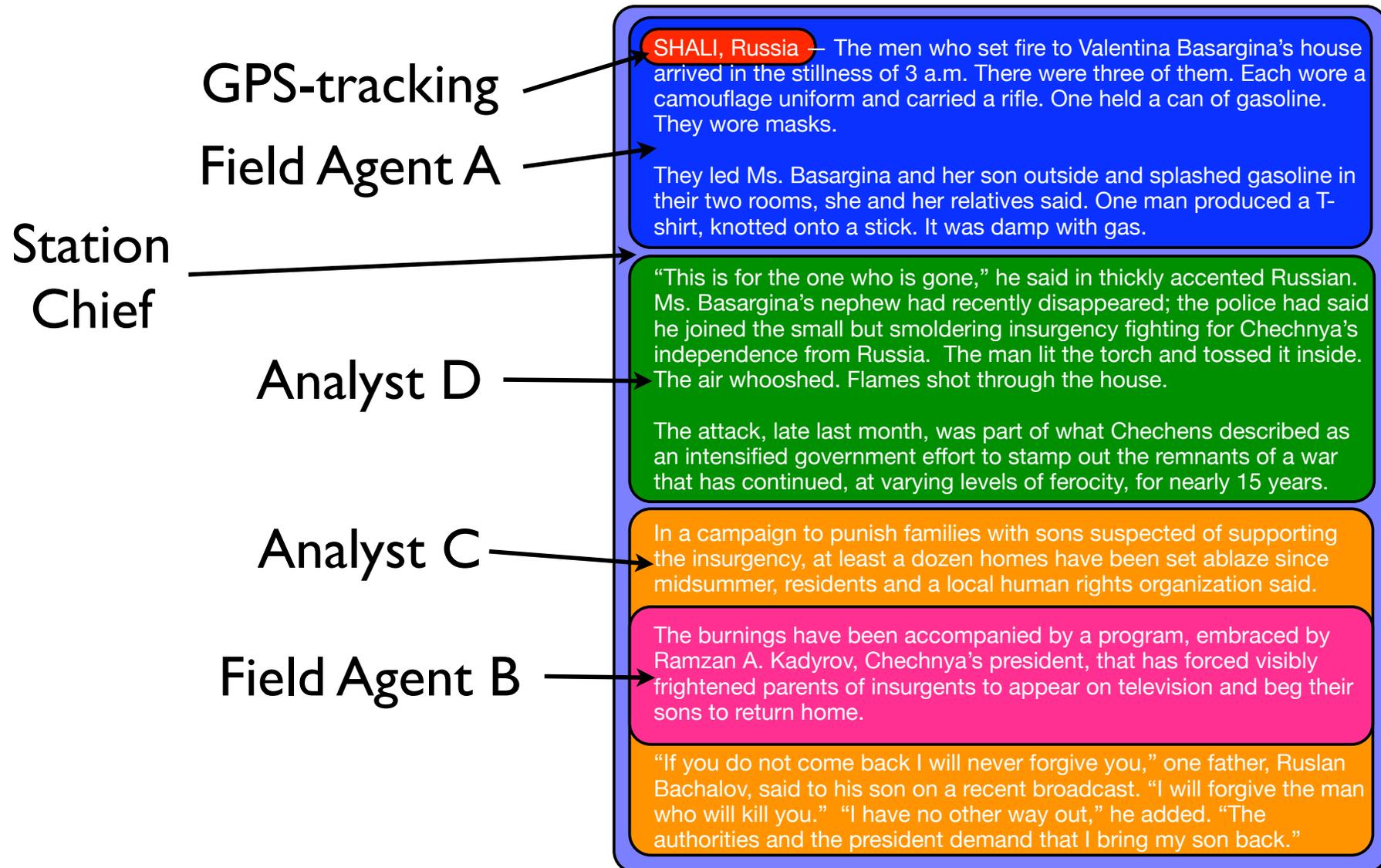
Provenance

HEC Data often comes from many sources and is synthesized by often distributed, complex, or hidden processes ..., how do you really know what the data means? Or how to interpret/filter/repair/reproduce it?

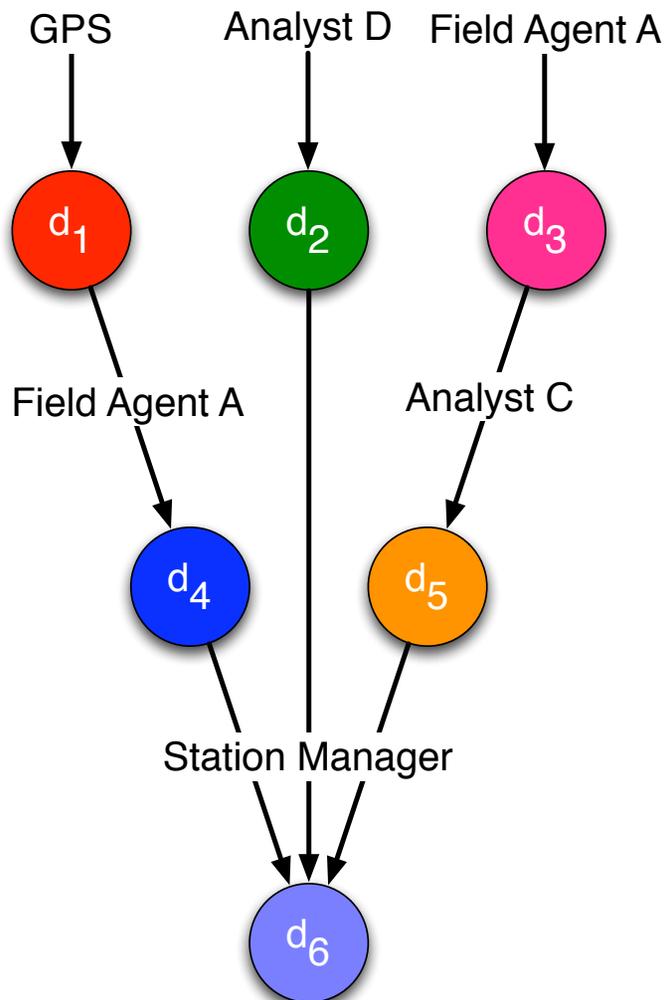
- *Data provenance* immutably identifies how data came to be.
 - ▶ **Who/what** contributed to it?
 - ▶ **What** was it based on?
 - ▶ **When** was it generated?
 - ▶ **Why** was it generated?
 - ▶ **How** was it generated?
- A *provenance chain* is a annotated history of the observed object.
- Standing calls from **government** (*intelligence, auditing and compliance certification*), the **scientific community** (*experimental data management, causal analysis*), and **industry** for systemic support for the collection and use of provenance data in HEC systems.



Tracking provenance



Tracking provenance



SHALI, Russia — The men who set fire to Valentina Basargina's house arrived in the stillness of 3 a.m. There were three of them. Each wore a camouflage uniform and carried a rifle. One held a can of gasoline. They wore masks.

They led Ms. Basargina and her son outside and splashed gasoline in their two rooms, she and her relatives said. One man produced a T-shirt, knotted onto a stick. It was damp with gas.

"This is for the one who is gone," he said in thickly accented Russian. Ms. Basargina's nephew had recently disappeared; the police had said he joined the small but smoldering insurgency fighting for Chechnya's independence from Russia. The man lit the torch and tossed it inside. The air whooshed. Flames shot through the house.

The attack, late last month, was part of what Chechens described as an intensified government effort to stamp out the remnants of a war that has continued, at varying levels of ferocity, for nearly 15 years.

In a campaign to punish families with sons suspected of supporting the insurgency, at least a dozen homes have been set ablaze since midsummer, residents and a local human rights organization said.

The burnings have been accompanied by a program, embraced by Ramzan A. Kadyrov, Chechnya's president, that has forced visibly frightened parents of insurgents to appear on television and beg their sons to return home.

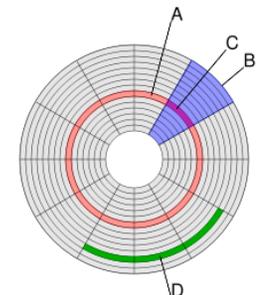
"If you do not come back I will never forgive you," one father, Ruslan Bachalov, said to his son on a recent broadcast. "I will forgive the man who will kill you." "I have no other way out," he added. "The authorities and the president demand that I bring my son back."

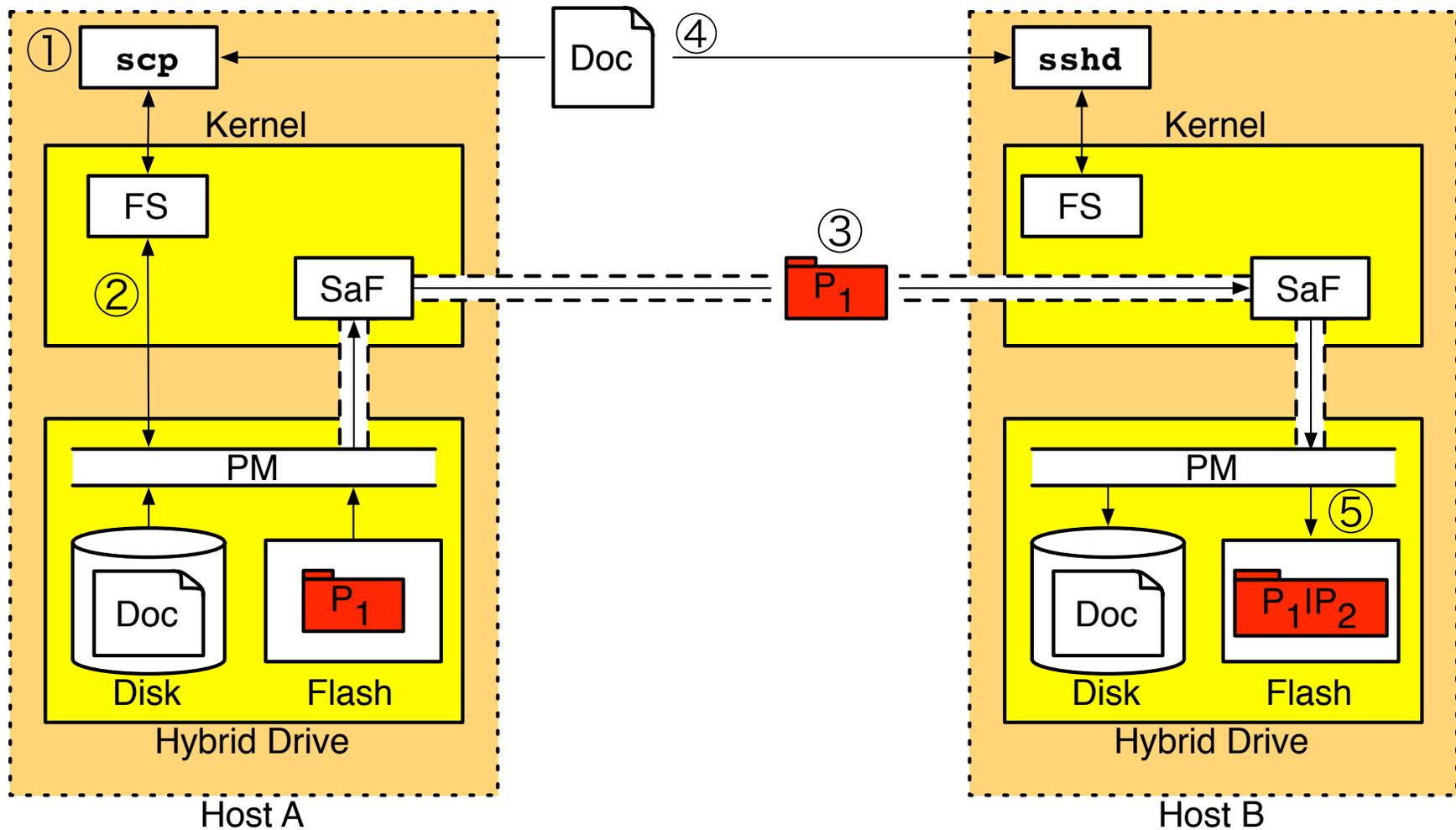
Future Vision

- Envision a world in which provenance was collected as a natural consequence of computation.
 - ▶ *End-to-end provenance* - all computational and storage elements of a HEC environment works in concert to collect and share provenance information.
 - ▶ Operating systems, storage,
- Core challenges:
 1. *Collection and storage of relevant provenance data within computing (host+storage).*
 2. *Coordinating/communicating provenance between systems/domains (host+storage+network).*
 3. *Modeling, calibrating, and mitigating costs (all).*
 4. *Retrieving and using provenance data.*

Provenance Monitors

- Tenets of secure system design holds that any kind of annotation should be implemented by external “observers” (monitors) [Anderson '72]
 - ▶ Such monitors must be (a) tamper-resistant, (b) completely mediate, and (c) verifiable.
 - ▶ Supports a threat model that preserves the correctness of the provenance chain even when the observed “application” to be compromised
 - ▶ Research: what are the appropriate TCB platforms for provenance in HEC systems
 - In operating system
 - In storage processor
 - Not just meta-data, securely bound meta-data (crypto)





Research Thrusts

- Host-focused Provenance
 - ▶ *In-kernel*: LSM “hook” based monitor (FS level provenance)
 - ▶ *In-storage*: on disk processor monitor (block level provenance)
- Distributed System Provenance
 - ▶ Low cost secure communication
 - ▶ Cryptographic constructions
- Cost/Performance Modeling and Mitigation
 - ▶ Develop calculus to understand cost and trade off granularity and strength of binding
- *Immediate goal*: reduce to practice (looking for apps)