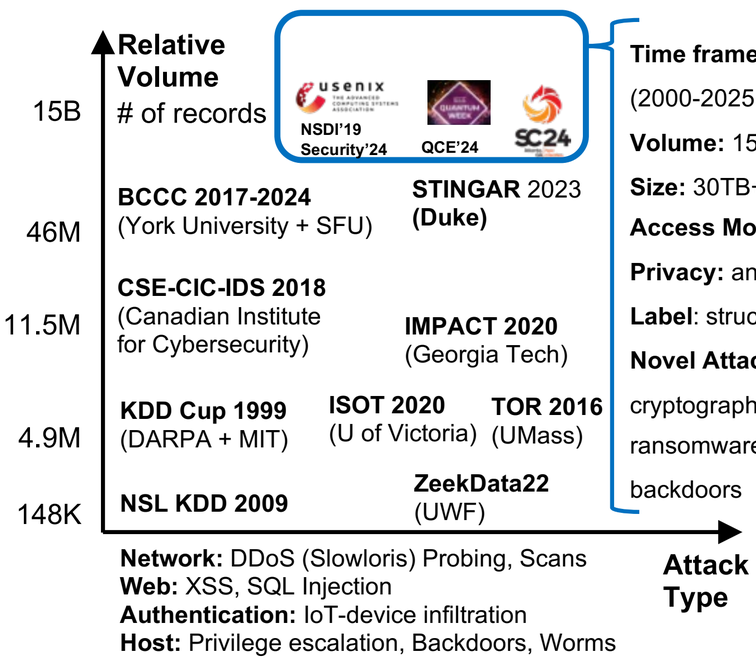
Live Evaluations of Real-World Security Data Lake from National Cyberinfrastructure

Phuong Cao, National Center for Supercomputing Applications
Ravishankar Iyer, University of Illinois Urbana Champaign


Problem: *Supercomputing infrastructure for AI innovations are prime targets for disruption of cyberinfrastructure.*

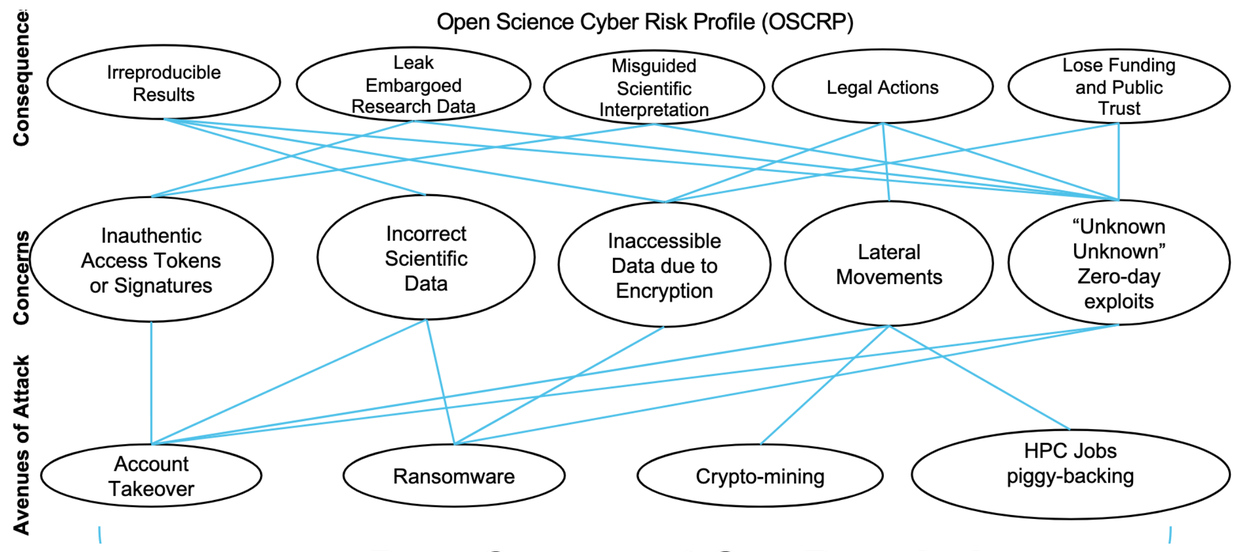
Gap: *Lacking cross-CI longitudinal attack data, making evaluation of detection models detached from real attacks.*

Existing Work



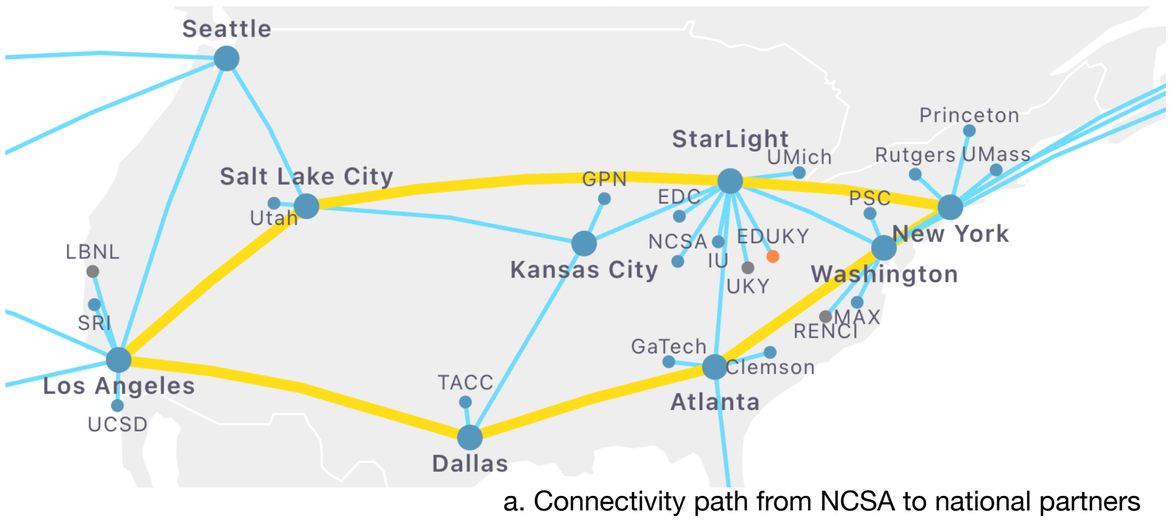
Time frame: longitudinal (2000-2025 onwards)
Volume: 15B+ records
Size: 30TB+ on disk (syslogs, pcaps)
Access Mode: Dynamic streams
Privacy: anonymization on the fly
Label: structured (SQL) and ground truth
Novel Attacks: quantum-resistant cryptography, AI-driven malware, ransomware, supply-chain SSH backdoors

Network: DDoS (Slowloris) Probing, Scans
Web: XSS, SQL Injection
Authentication: IoT-device infiltration
Host: Privilege escalation, Backdoors, Worms

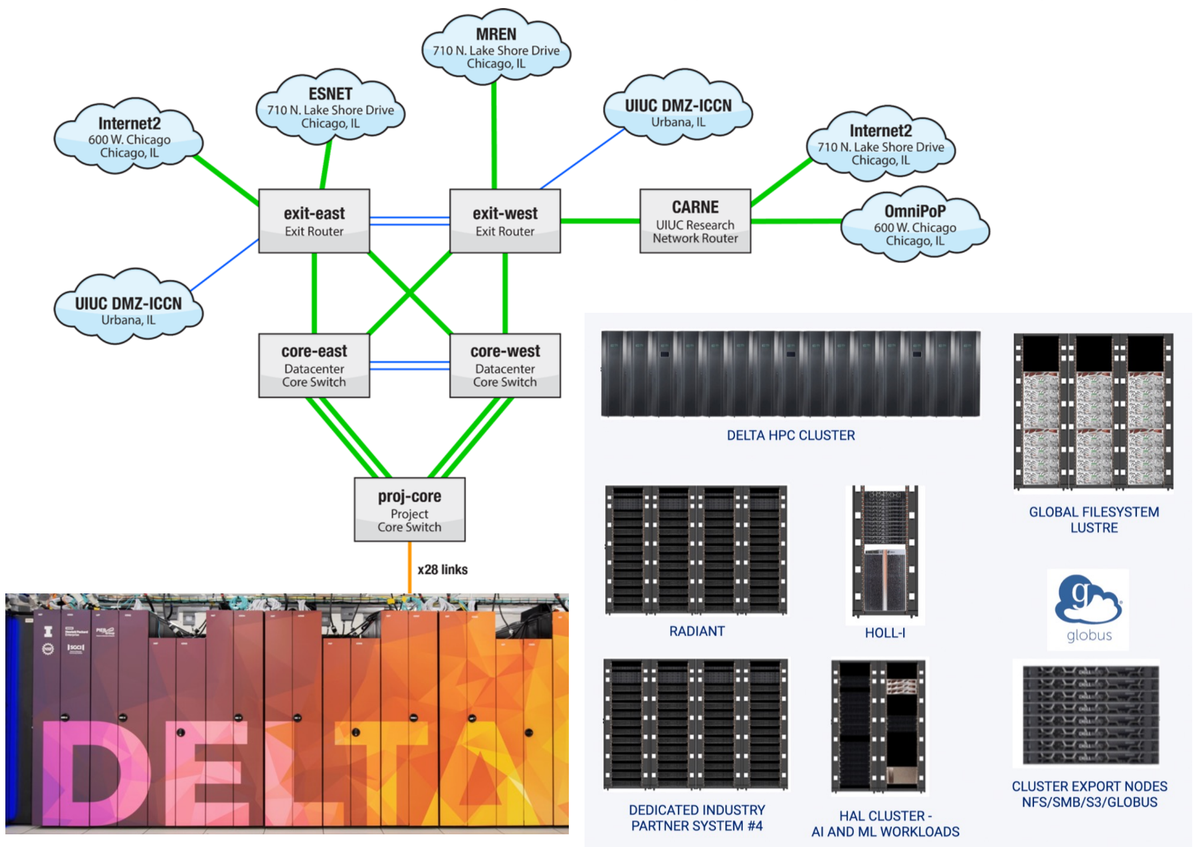


Data Source of Our Data Lake

Zeek border tap + East-West traffic



a. Connectivity path from NCSA to national partners

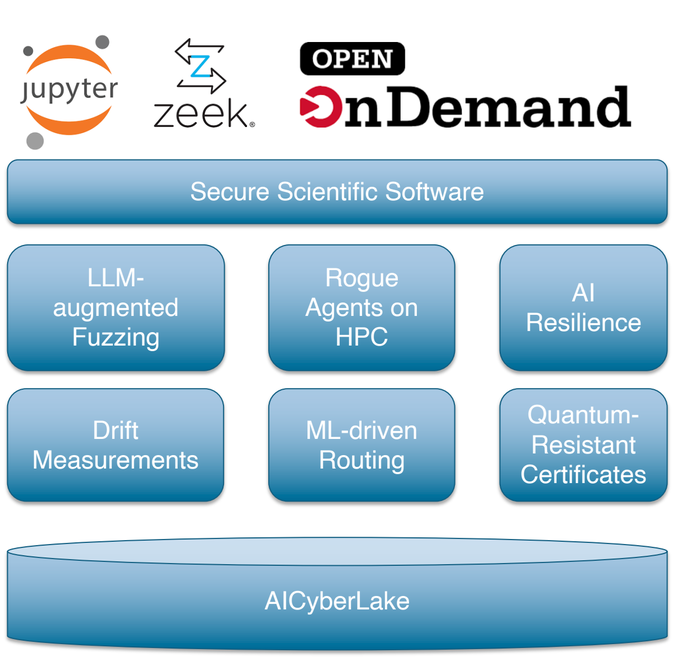


b. System and Network Diagram of Delta and related storage systems at NCSA

Data Schema

Monitoring Tool	Security Scope
Zeek (formerly Bro)	Used for deep packet inspection and network protocol analysis. Sufficient for capturing initial handshakes (SSH/TLS) and identifying anomalous flows (e.g., DNS tunneling).
Central Syslog	Collects application and system-level events from thousands of hosts, providing visibility into local actions like privilege escalation.
Honeypots (CAUDIT)	Specifically designed to attract and log brute-force attempts at scale, providing detailed traces of attack attempts, malformed token manipulations, and credential guesses.
NetFlow	Provides high-level traffic volume data, useful for detecting large-scale data exfiltration that might bypass deeper inspection because of volume.
Osquery	Enables SQL-like querying of host-level state, such as running processes and kernel module modifications.
LDMS	Lightweight Distributed Metric Service (from Sandia) for monitoring Blue Waters nodes.
Audit Logs	System/application logs from NCSA supercomputers (Blue Waters, Delta, and peers such as the FABRIC testbed).

Enabling and Securing Applications



Standardization Efforts

Organization/Document	Project Description
6GQ	Postquantum Cryptography for 6G Networks [1]
IETF draft-miller-sshm-00	ML-DSA 65/Ed25519 Composite Signatures in SSH [2]
NIST SP 800-53 AI Security	Securing AI Systems [3]
NIST SP 800-224 HPC Security	High-Performance Computing (HPC) Security Overlay [4]
ISO/IEC JTC 1 SC27 WG2	PQC ALGORITHMS: Amending ISO/IEC 18033-2 [5]

Example CVEs found in AICyberLake

Package	Vulnerability ID	Reported by our Team and Impact Summary
Zeek (Bro)	Our Reported Vulnerability	Rapid reset leading to DoS, LDAP evasion [6].
SciTokens	CVE-2026-32716/14	Improper prefix matching; unauthorized access [7].
Jupyter Lab	CVE-2026-42266; Pending	Malicious extensions installation [8].
CILogon	Our Reported Vulnerability	Path traversal via OAuth implementation [9]

Pending Publications

Cui, Shengkun, Archit Patke, Ziheng Chen, Aditya Ranjan, Hung Nguyen, Phuong Cao, Brett Bode et al. "Characterizing Modern GPU Resilience and Impact in HPC Systems: A Case Study of A100 GPUs." In *2025 55th Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, pp. 1-6. IEEE, 2025.

Gupta, Ragini, Shinan Liu, Ruixiao Zhang, Xinyue Hu, Xiaoyang Wang, Hadjer Benkraouda, Pranav Kommaraju, Phuong Cao, Nick Feamster, and Klara Nahrstedt
"Generative active adaptation for drifting and imbalanced network intrusion detection."
arXiv preprint arXiv:2503.03022 (2025).

Results + Data Disseminations

Data namespace on OSDF

- Public namespace on released data
- Private namespace on embargoed data
- Background jobs on non-urgent, longitudinal attack data on OSG HTC

NIST Secure HPC working group

Published *Characterizing GPU Resilience at Supercomputing'25 and AI agents DSN'26*

https://pmcao.github.io/projects/aicyberlake/