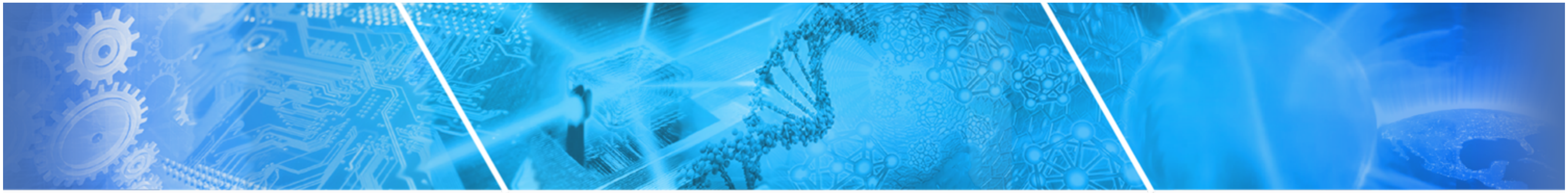
SAND2021-10880PE



LDRD

Laboratory Directed Research and Development

SECURE Overview

And now, the end is near
And so we face the final curtain

Ali Pinar, PI

Why Do We Need Cyber Experimentation?

To study complex cyber systems (e.g., resilience)

- answer “what if questions” with high-confidence

Emulytics

- assess confidence in our results under uncertainty

Uncertainty Quantification

- make robust decisions under uncertainty in an adversarial environment
- **Adversarial Optimization**

  with rigor

- *<u>Challenge:</u> Can we trust this approach for high consequence systems?*

- *<u>Inspiration:</u>* Sandia’s know-how and capabilities from our nuclear stockpile stewardship

> Rigorous cyber experimentation should be a pillar of science of cyber security, just as computational Science and Engineering (CSE) is a pillar of science.

An Overview of the Process

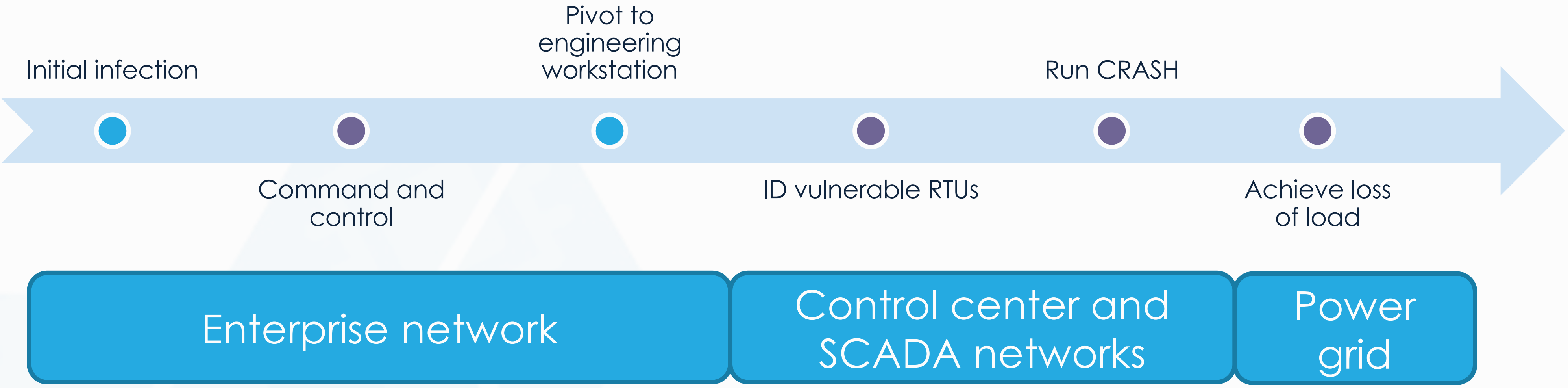
Exemplar problem: Is our power grid resilient against an attack as in Ukraine?

- Ukraine attack was based on Crash Override Malware
- The attacker gains remote access to power grid components to turn them on and off.

- Previous presentations organized around basic research elements
  - Integration was not clear
- Today's presentations organized around integration tasks
  - Each effort utilizes all research elements

Inspired by the EAB feedback

Exemplar goal and approach



Initial infection

Command and control

Pivot to engineering workstation

ID vulnerable RTUs

Run CRASH

Achieve loss of load

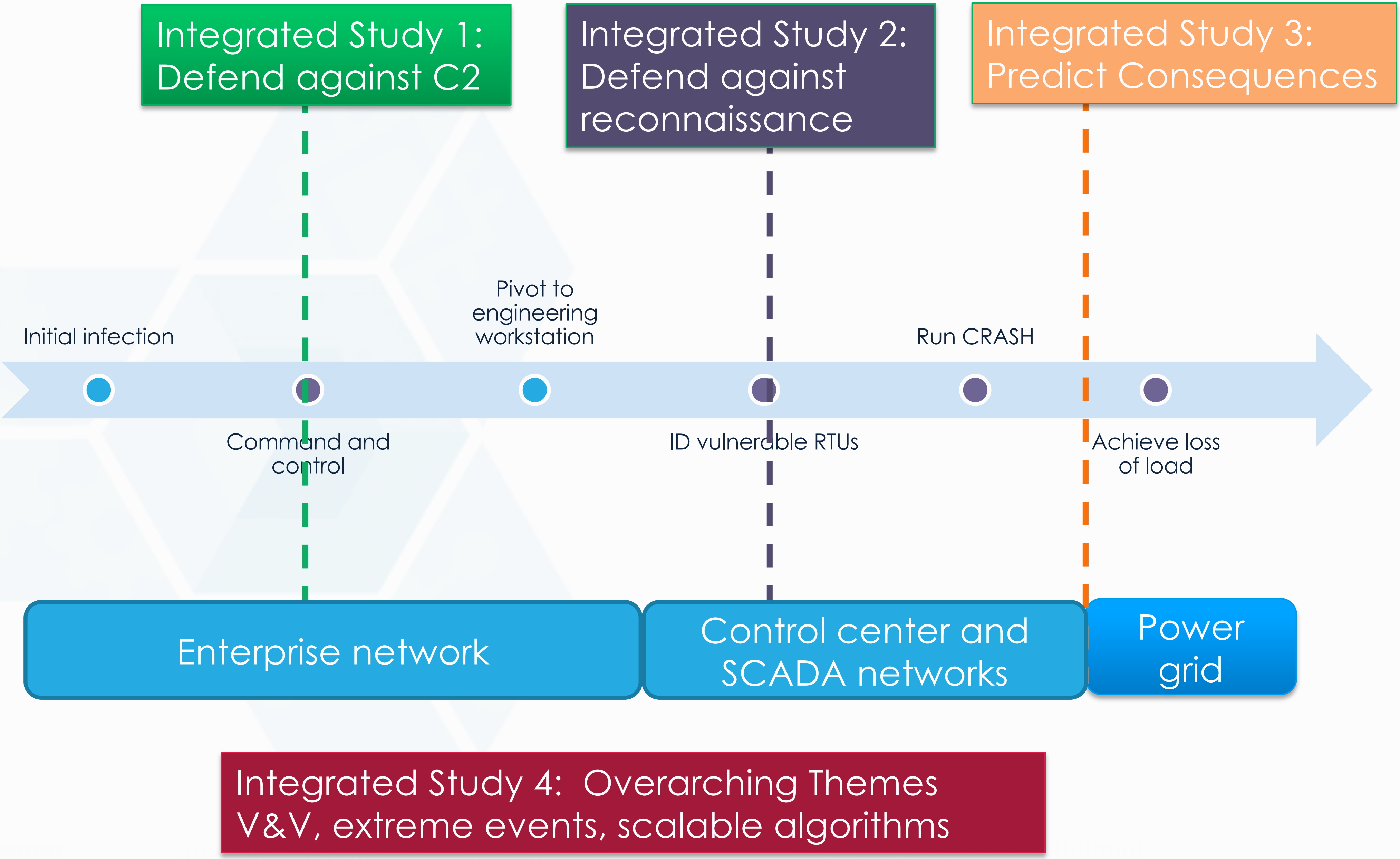
Enterprise network

Control center and SCADA networks

Power grid

- Goal: characterize loss of load resulting from malware infection in enterprise network
  - Account for uncertainties in threat, network conditions
- Approach
  - Piecewise studies to inform Markov transition probabilities and uncertainties
  - End-to-end SCEPTRE demonstration

Exemplar studies



Integrated Study 1: Defend against C2

Integrated Study 2: Defend against reconnaissance

Integrated Study 3: Predict Consequences

Initial infection

Command and control

Pivot to engineering workstation

ID vulnerable RTUs

Run CRASH

Achieve loss of load

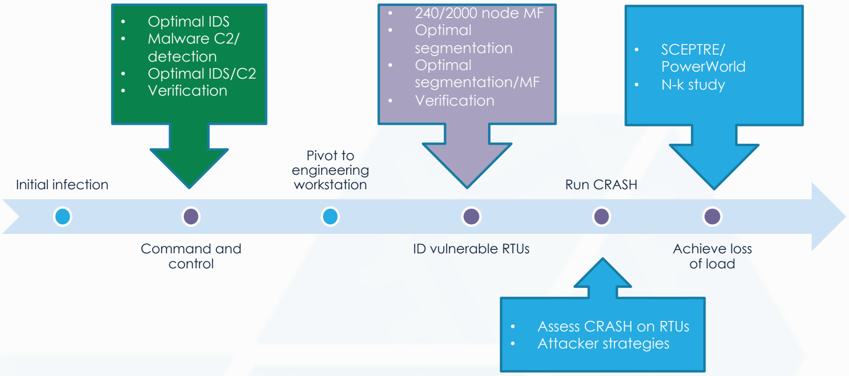
Enterprise network

Control center and SCADA networks

Power grid

Integrated Study 4: Overarching Themes V&V, extreme events, scalable algorithms

Connecting experiments to end-to-end threat analysis



- Optimal IDS
- Malware C2/ detection
- Optimal IDS/C2
- Verification

- 240/2000 node MF
- Optimal segmentation
- Optimal segmentation/MF
- Verification

- SCEPTRE/ PowerWorld
- N-k study

Initial infection

Command and control

Pivot to engineering workstation

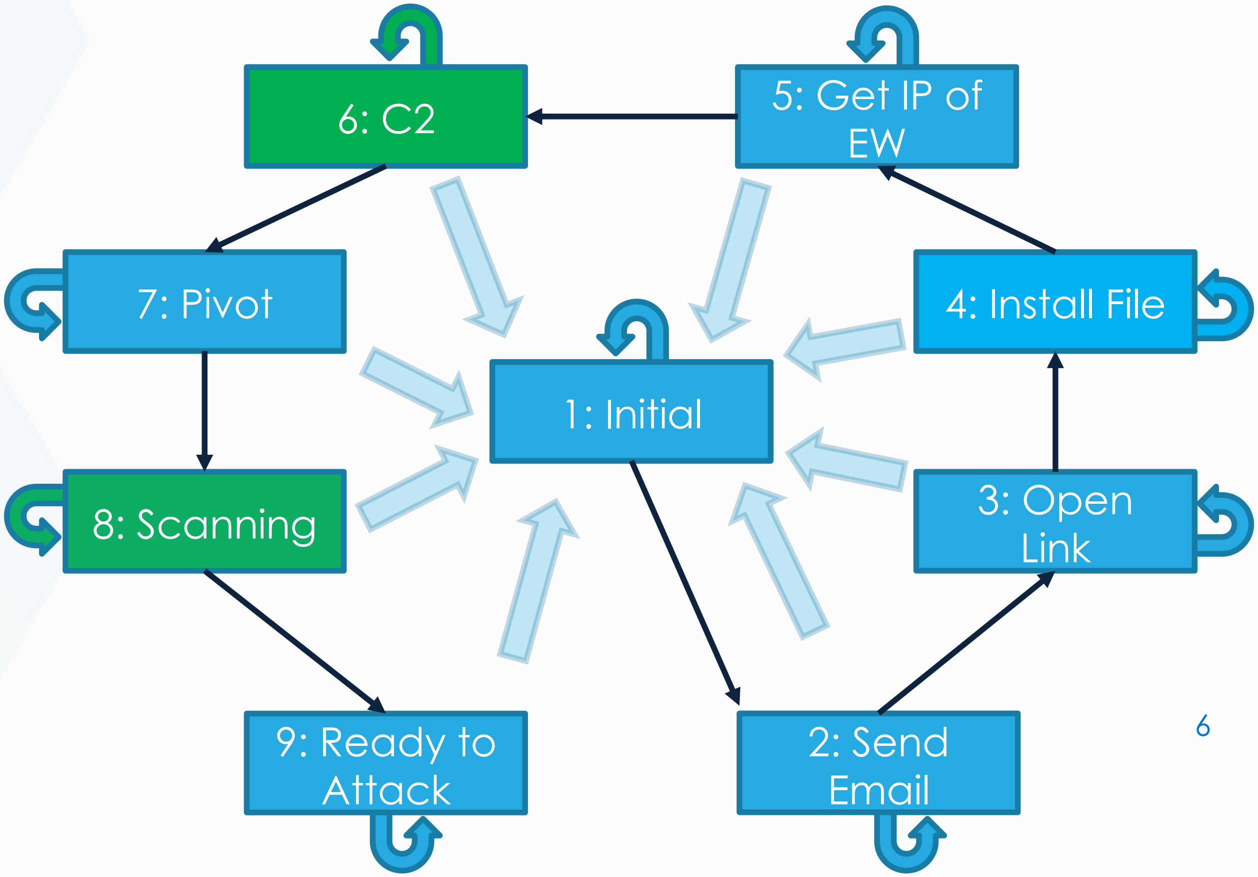
ID vulnerable RTUs

Run CRASH

Achieve loss of load

- Assess CRASH on RTUs
- Attacker strategies

Markov Model



Last year: we used MITRE ATT&CK to populate transition probabilities

NOW: we incorporate results from our emulation experiments to populate steps 6 and 8 in green

Connecting experiments to end-to-end threat analysis

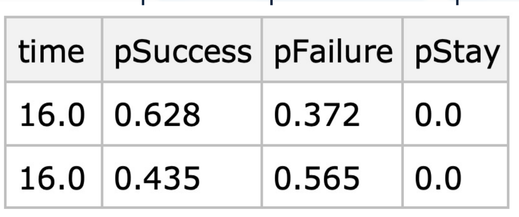
- Experimental data processed to represent:
  - Success (attacker goes to next step)
  - Failure (have to start anew)
  - Stay in place

- In addition to the various combinations of state 6 and state 8 transition probabilities shown below, we analyzed various attack/defender strategies

- **We are performing uncertainty analysis on the Markov model itself**

State 6 (C2):

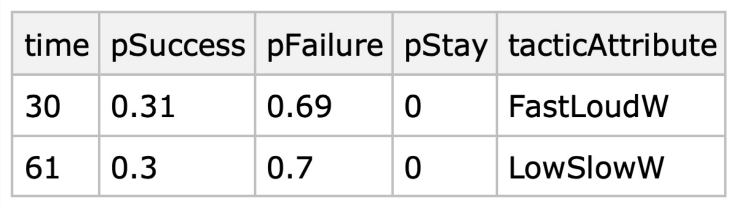
- >= 10 detections for alert
- Exp 1: Snort stressed/drops packets
- Exp 2: No packet drop



time	pSuccess	pFailure	pStay
16.0	0.628	0.372	0.0
16.0	0.435	0.565	0.0

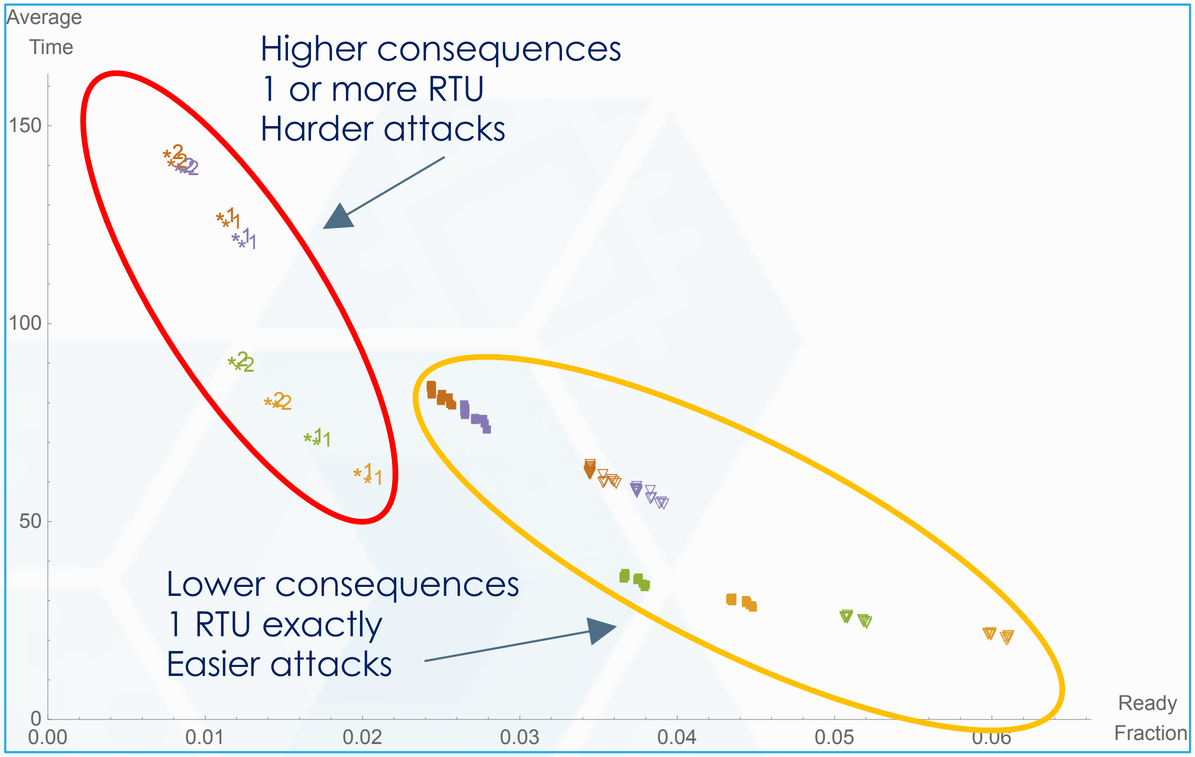
State 8 (Scanning/Detection):

- Fast scan: ts=30s
- Slow scan: ts=61s



time	pSuccess	pFailure	pStay	tacticAttribute
30	0.31	0.69	0	FastLoudW
61	0.3	0.7	0	LowSlowW

Results – Accounting for attacker/defender uncertainties



Higher consequences
1 or more RTU
Harder attacks

Lower consequences
1 RTU exactly
Easier attacks

Defender goal: push attacker toward top-left of the plot (e.g. through better IDS)

Each dot on the chart above represents a combination of C2 data, scanning/detection data, and attacker/defender strategy

- Plotting attack success metrics from Markov analysis: mean time to attack success vs. fraction of time in the "READY" state.
- Extended our analysis framework to support **UQ in transition probabilities**, and **variations in each step's inherent timestep.**
  - Default timestep = 1 hour/state, but states 6 and 8 use different values
- Experiments provide range of transition probabilities (depending on scenario, attacker strategy, etc.)

Markov analysis allows:

- Estimates of how secure the system is under attack
- Ranking of attacker/defender strategies

So What?

- What changed?
  - We produced an ***objective*** process that can ***quantify*** security.
    - All assumptions are listed; all processes are repeatable;
    - All experiments are verified; all models are validated;
  - We have a scientific processes that can, and will be improved.
    - No more disagreeing with expert opinions.
    - Instead challenge assumptions; propose better algorithms/metrics.

- What can we do now? ***Quantifiable Security***
  - Quantify return on investment for cyber security
  - Rigorously compare two proposed remediations
  - Identify critical components both for improving security and model fidelity
  - Quantify attack consequences and enable mission-driven cyber security
    - Zoom in on extreme events

SECURE's Legacy

- Rigorous cyber experimentation is essential and achievable.
  - Long way in front of us, but long way behind us.
  - The path forward is more clear. No more admiring the problem, we can line out specific tasks.
- We are building a community around us that will continue to work with us.

Lasting Impact:

- Cyber experimentation can be a pillar of science of security, when complemented with rigorous mathematical techniques.

- Rigorous cyber experimentation can provide to integrated cyber systems what CSE has provided to nuclear weapon's programs.

SCIRE Institute to carry on SECURE's Legacy

- **SCIRE** Institute: **S**andia's **C**yber security **I**nnovation and **R**igorous **E**xperimentation Institute
- SCIRE Goals:
  - Promote interdisciplinary research to support rigorous cyber experimentation
  - Serve as a bridge between basic and applied research
  - Raise awareness in the national security community
  - Keep the SECURE team together
    - and reach out for broader partnerships
- First workshop is scheduled for Nov 9-10, 2021

Inspired by the EAB feedback

What was good?

- EAB Feedback
- Outstanding team culture
  - Essential to keep the team together
  - Young teams members
- Strong support and convergence on the SECURE's goals
- Many strong stand-alone technical results
- Verification and Validation results will be landmarks
- Connections with academia and exporting complicated challenges as well-defined problems

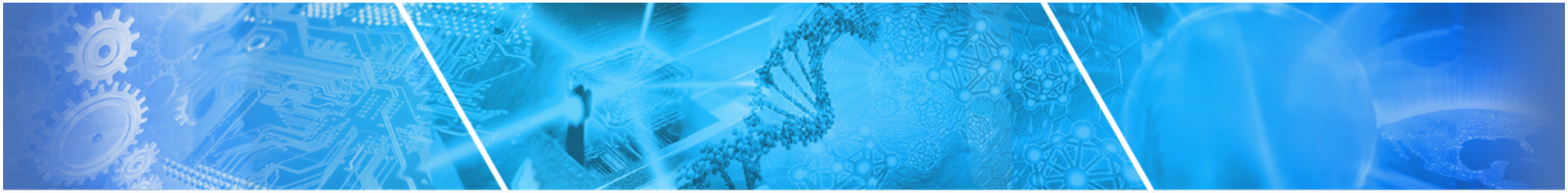
What could have been better?

- Pandemic was not part of the risk management plan.
  - Affected individual performances as well as lessening the advantage of being a well-functioning team
  - Affected out-reach activities
- Staff turnover slowed down progress
  - Recovered each time
  - Can be more bigger issue as we make more progress
- Some tasks were much harder than anticipated
  - e.g., Verification of computer experiments
- Scalability remains a challenge.
- External visibility needs to be improved.
  - Both a Lab-level problem and a project-level problem

SECURE Products

Inspired by the EAB feedback

- First version of the manual (both theory and user) is ready as an online document
- Software:
  - Python Adversarial Optimization (PAO) is released
  - Scenario Orchestration (ScOrch) will be released
  - Many others transferred to applied projects
- 29 Publications (accepted, submitted and in preparation)
- 15 technical presentations



LDRD

Laboratory Directed Research and Development

Command and Control (C2): Integrated Cyber Experimentation Exemplar

Jared Gearhart

August 26th, 2021

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Since last EAB, formed cross-cutting C2 team

Jared Gearhart,
C2 Integration Lead

Casey Glatter,
C2 Emulation

Jamie Thorpe,
Verification

Seth Hanson,
C2 Emulation

Eric Vugrin,
Math Modeling

Bert Debusschere,
UQ, PCE

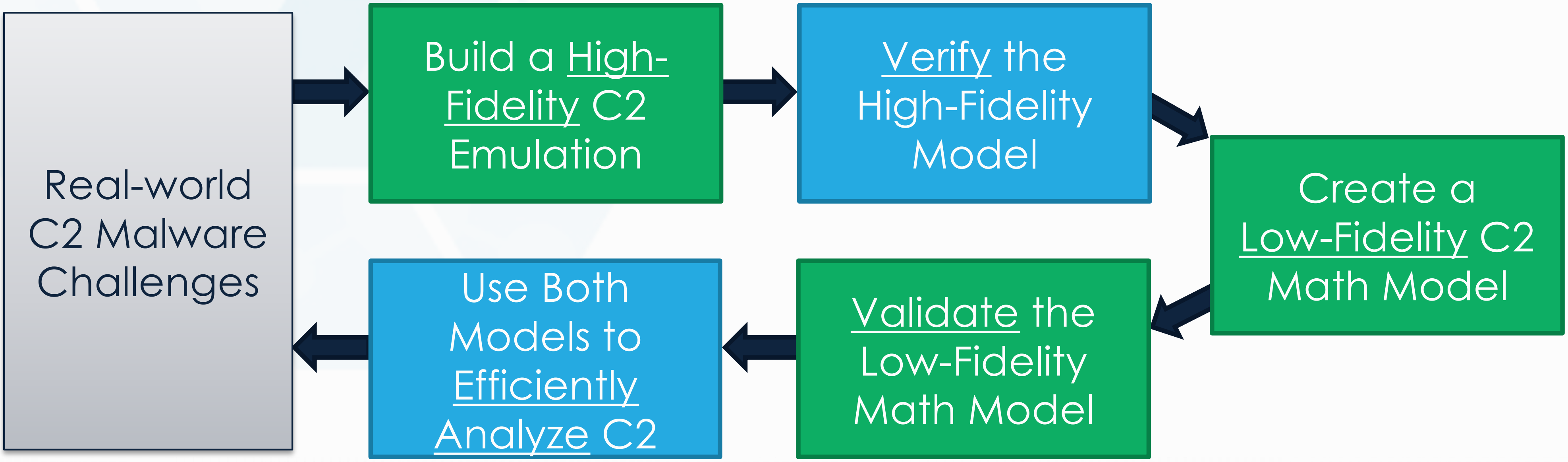
Gianluca Geraci,
UQ, MFUQ

Former Team Members: Jerry Cruz (Emulation), Trevor Rollins (Statistical Tests)

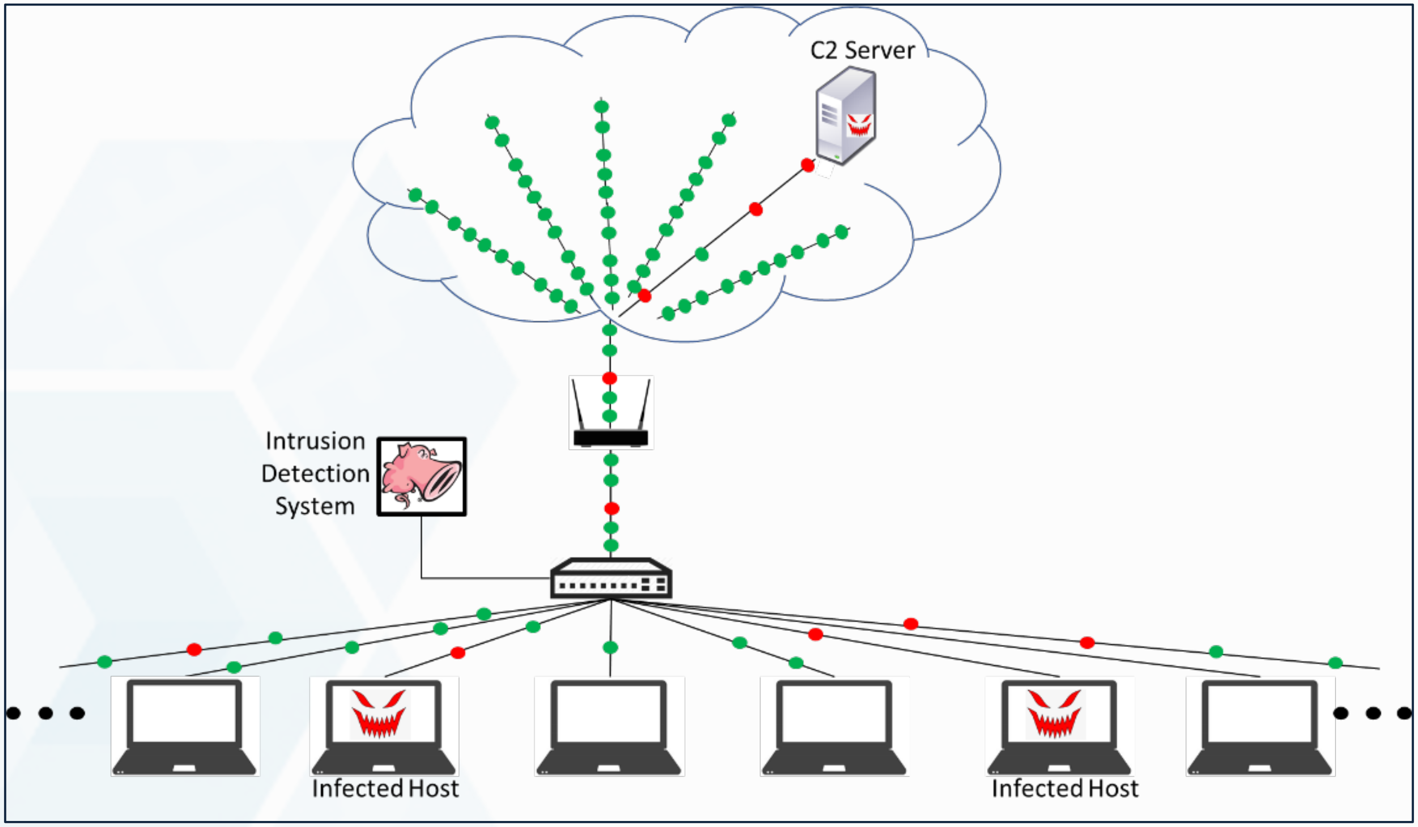
C2 Exemplar Goal: Answer both specific C2 and general cyber experimentation questions

- C2 specific goals:
  - How long does it take to detect a C2 channel?
  - Which factors have the largest impact on the IDS system?
- General research goals:
  - What emulation capabilities are required?
  - Can we develop a math model to approximate emulation?
  - Can the emulation and math model be used in conjunction?

C2 Exemplar Analysis Process


Real-world C2 Malware Challenges
Build a High-Fidelity C2 Emulation
Verify the High-Fidelity Model
Create a Low-Fidelity C2 Math Model
Validate the Low-Fidelity Math Model
Use Both Models to Efficiently Analyze C2

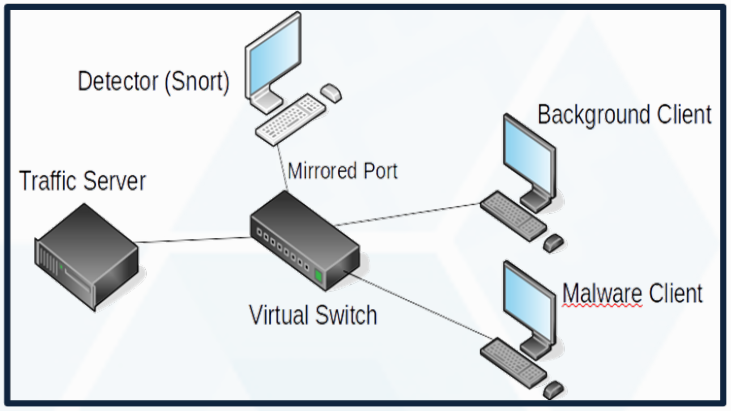
C2 Scenario: Detect malicious (Emotet) traffic between infected host(s) and a C2 server



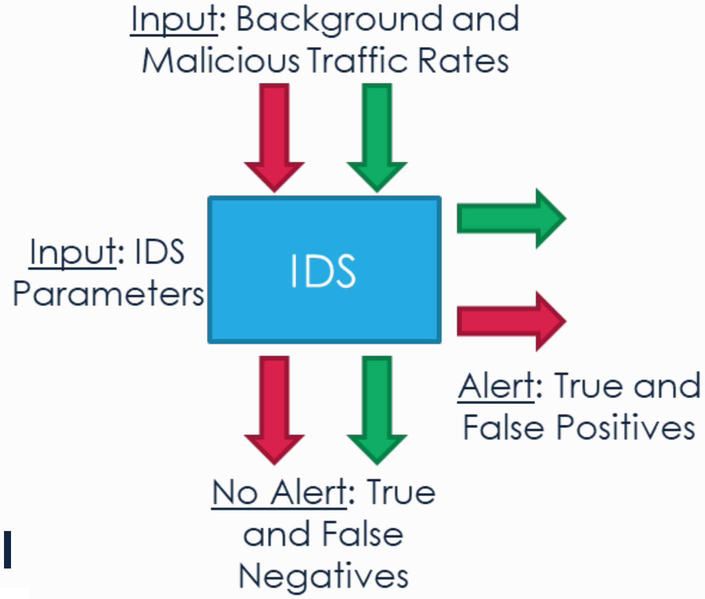
- Three main drivers:
  - IDS configuration
  - Volume of malicious traffic
  - Volume of benign traffic

Summary of Previous C2 Work

Created a High-fidelity Emulation Model – "Expensive"



Created a Low-fidelity Math Model – "Cheap"



Input: Background and Malicious Traffic Rates

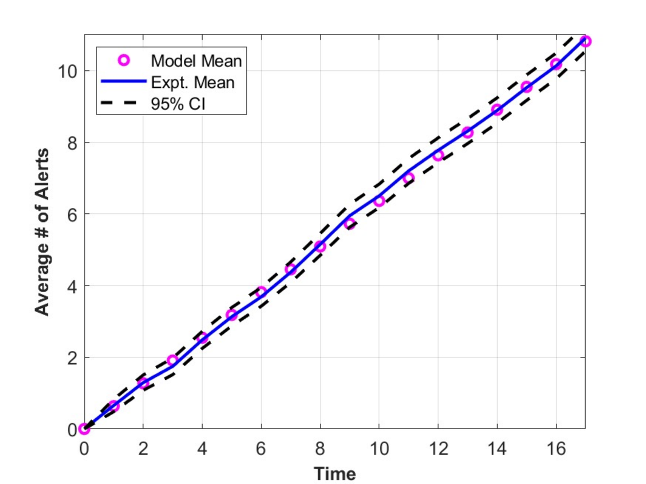
Input: IDS Parameters

IDS

Alert: True and False Positives

No Alert: True and False Negatives

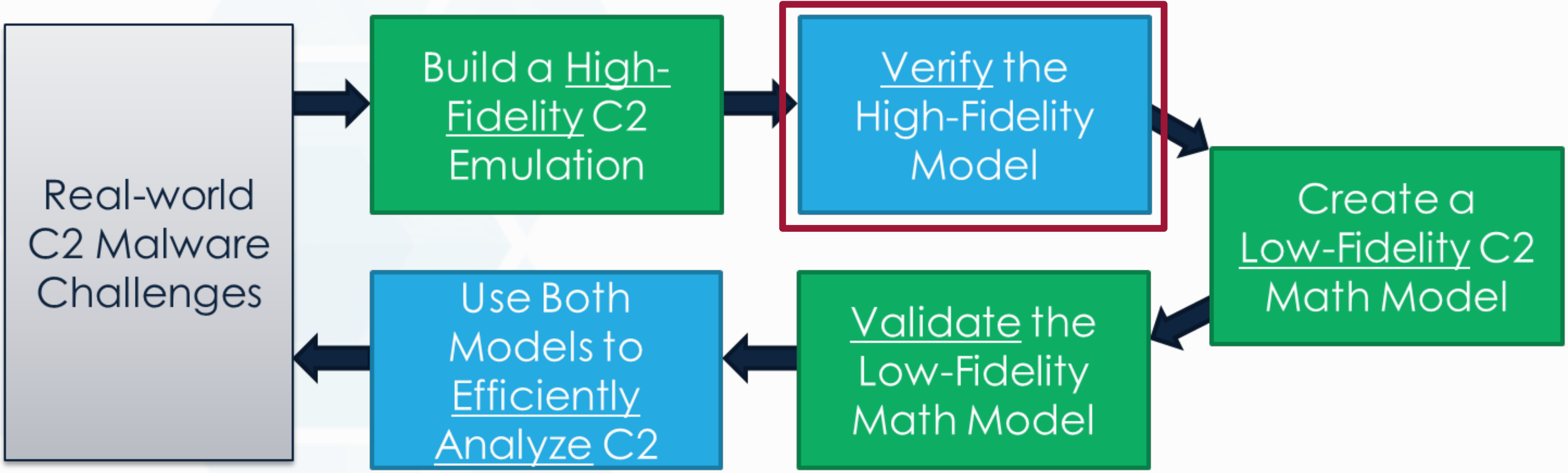
Developed Methods to Validate the Math Model



Eric D. Vugrin, Seth Hanson, Jerry Cruz, Casey Glatter, Thomas Tarman, and Ali Pinar, "Detection of command and control traffic: model development and experimental validation," submitted to *43rd IEEE Symposium on Security and Privacy*.

Focus Since Last EAB

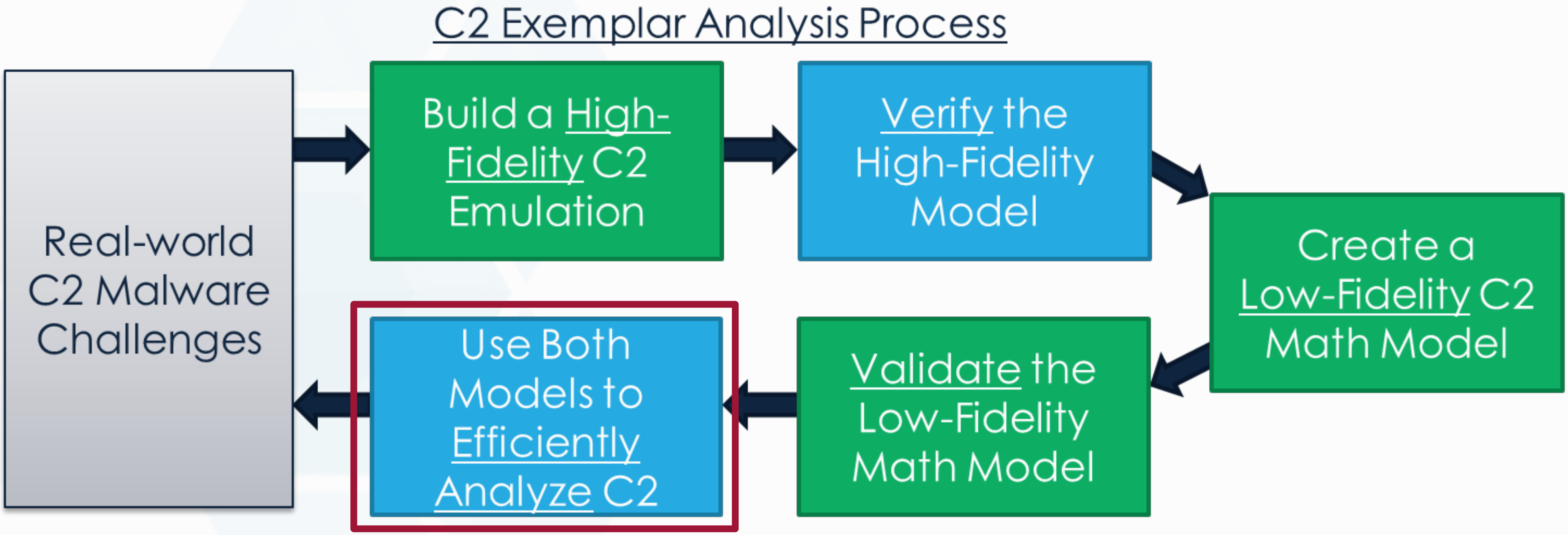
C2 Exemplar Analysis Process


Real-world C2 Malware Challenges
Build a High-Fidelity C2 Emulation
Verify the High-Fidelity Model
Create a Low-Fidelity C2 Math Model
Validate the Low-Fidelity Math Model
Use Both Models to Efficiently Analyze C2

Verification: How do we assess the trustworthiness of emulation experiments?

- Focus on over-subscription of emulation resources
- Verification strategy:
  - Establish baseline: Run experiments in series with sufficient resources
  - Intentionally over-subscribe: Iteratively increase parallel experiments
  - Monitor metrics to identify deviations/indicators:
    - C2 metrics: Number of alerts over time
    - Telemetry: CPU usage, stolen cycles, etc.
- Key findings:
  - C2 results have been less conclusive than the scanning study
    - Not currently able to over-subscribe the C2 emulation
    - Using remaining time on LDRD to understand issue (potentially related to I/O)
  - However, only identified this because we were doing verification

Focus Since Last EAB



C2 Exemplar Analysis Process

Real-world C2 Malware Challenges

Build a High-Fidelity C2 Emulation

Verify the High-Fidelity Model

Create a Low-Fidelity C2 Math Model

Validate the Low-Fidelity Math Model

Use Both Models to Efficiently Analyze C2

Analysis Goal: Efficiently identify key parameters

- 12 uncertain parameters
  - 4 discrete
  - 8 continuous

Benign Traffic Parameters

Parameters	Units	Value	Distribution
Benign traffic per host	Packets per sec	5-100	Continuous log-uniform
Fraction of benign packets with Emotet signatures	No Units	1e-5-1e-3	Continuous log-uniform
Detection rate for signatures in benign traffic	No units	0.9-0.99	Continuous uniform

IDS and Environment Parameters

Parameters	Units	Value	Distribution
Total number of workstations	No units	10	Fixed
Average packet size	Bytes	150-250	Continuous uniform
Snort capacity	Bytes per second	1e5, 2e5, 5e5, or 1e6	Discrete with equal probability
Number of CPUs	No units	8	Fixed
Number of CPUs to maximize snort	No units	1-8	Discrete with equal probability
Other CPU Processes	No units	0-7	Discrete with equal probability
Drop rate multiplier	No units	0.9-1.1	Symmetric continuous triangular distribution

Malware Traffic Parameters

Parameters	Units	Value	Distribution
Number of infected workstations	No units	0-10	Discrete with equal probability
Malware traffic per infected host	Packets per sec	4-10	Continuous uniform
Fraction of malware packets with emotet signatures	No units	0.1-0.2	Continuous uniform
Detection rate of signatures for malware traffic	No Units	0.9-0.99	Continuous uniform

Polynomial Chaos Expansion (PCE): Screen parameters using the low-fidelity model

- PCE surrogates represent the Quantity of Interest (QoI) as a linear combination of orthogonal polynomials in the input variables
  - Efficient
  - Offers Global Sensitivity Analysis (GSA) information for free
- Screening study:
  - Random sampling of parameters over joint discrete-continuous space
    - PyApprox (https://sandialabs.github.io/pyapprox/index.html)
  - Analyze low-fidelity model results from 3,000 samples

Bert J. Debusschere, Gianluca Geraci, John D. Jakeman, Cosmin Safta, and Laura Swiler, "Polynomial Chaos Expansions for Discrete Random Variables in Cyber Security Emulytics Experiments", SIAM CSE 2021 (virtual), March 1, 2021 (Talk presented by Bert Debusschere)

PCE: Identify the main effects for key metrics

Key Findings

- Total alerts primarily affected by volume of emotet traffic
- False positives primarily affected by volume of benign "Emotet" traffic

Utilize this information to focus subsequent multi-fidelity analysis

Main Effects for Select Results

	Parameters	Total Alerts, t = 5 sec.	False Positives, t = 5 sec.
Significant	Number of infected workstations	**0.87**	0.00
	Fraction of benign packets with Emotet signatures	0.00	**0.51**
	Benign traffic per host	**0.01**	**0.20**
Moderate	Malware traffic per infected host	**0.05**	0.00
	Fraction of malware packets with emotet signatures	**0.03**	0.00
	Snort capacity	**0.01**	**0.01**
	Other CPU Processes	**0.01**	0.00
Minimal	Number of CPUs to maximize snort	0.00	0.00
	Average packet size	0.00	0.00
	Detection rate for signatures in benign traffic	0.00	0.00
	Detection rate of signatures for malware traffic	0.00	0.00
	Drop rate multiplier	0.00	0.00

Multi-fidelity Uncertainty Quantification (MFUQ): Exploiting both models for efficient UQ

$$\hat{Q}^{MF} = \underbrace{\frac{1}{N}\sum_{i=1}^{40} Q^{(i)}_{minimega}}_{} + \underbrace{\alpha \left(\frac{1}{N}\sum_{i=1}^{40} Q^{(i)}_{math} - \frac{1}{r \times 40}\sum_{j=1}^{r\times 40} Q^{(j)}_{math} \right)}_{}$$

High-Fidelity Term

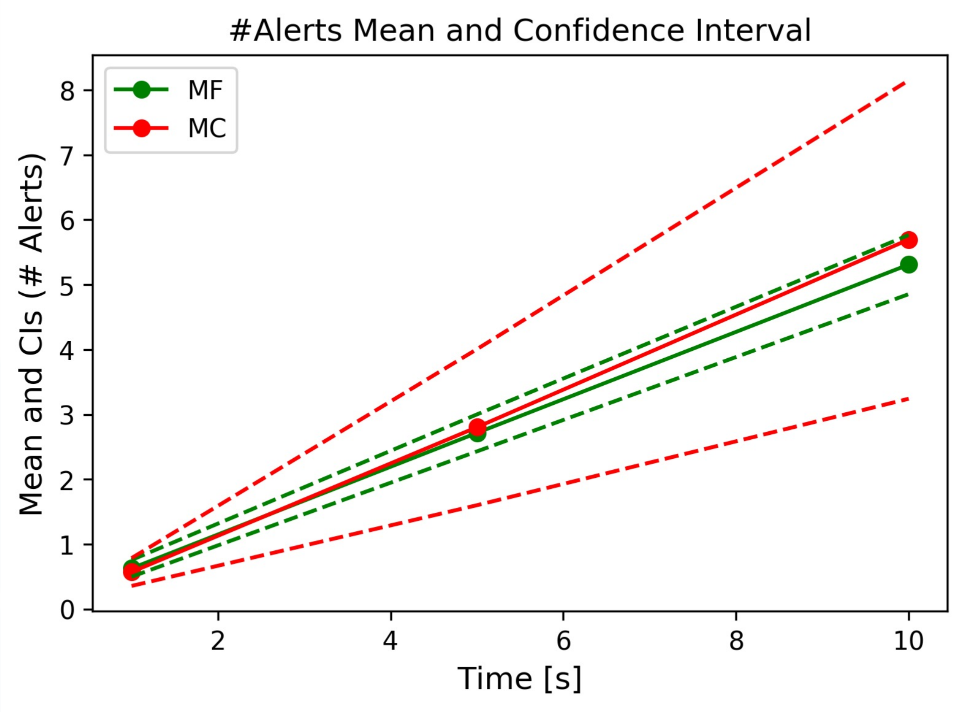
Monte Carlo Estimator
"Expensive"
(~18 hours for 40x10 runs)

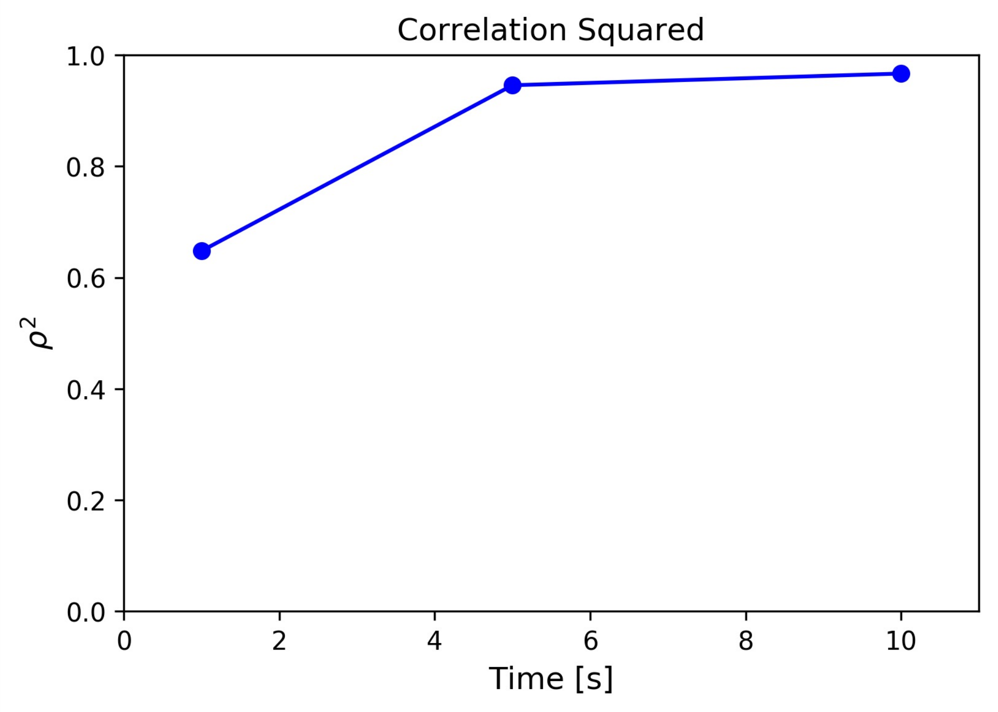
Low-Fidelity Term

Unbiased
Reduces Variance
"Cheap" (0.4 secs/run)

Geraci, G., Crussell, J., Swiler, L.P. and Debusschere, B. J. "Exploration of Multifidelity UQ Sampling Strategies for Computer Network Applications." *International Journal of Uncertainty Quantification,* 2021. Pp. 93-118. DOI: 10.1615/Int.J.UncertaintyQuantification.2021033774

MFUQ estimates mean number of alerts with less variability than Monte Carlo





- MC: 40 settings (10 iterations each)
- MF: Equal to 40.53 MC runs
  - 86,840 low-fidelity runs
  - Low-fidelity can be run on a PC

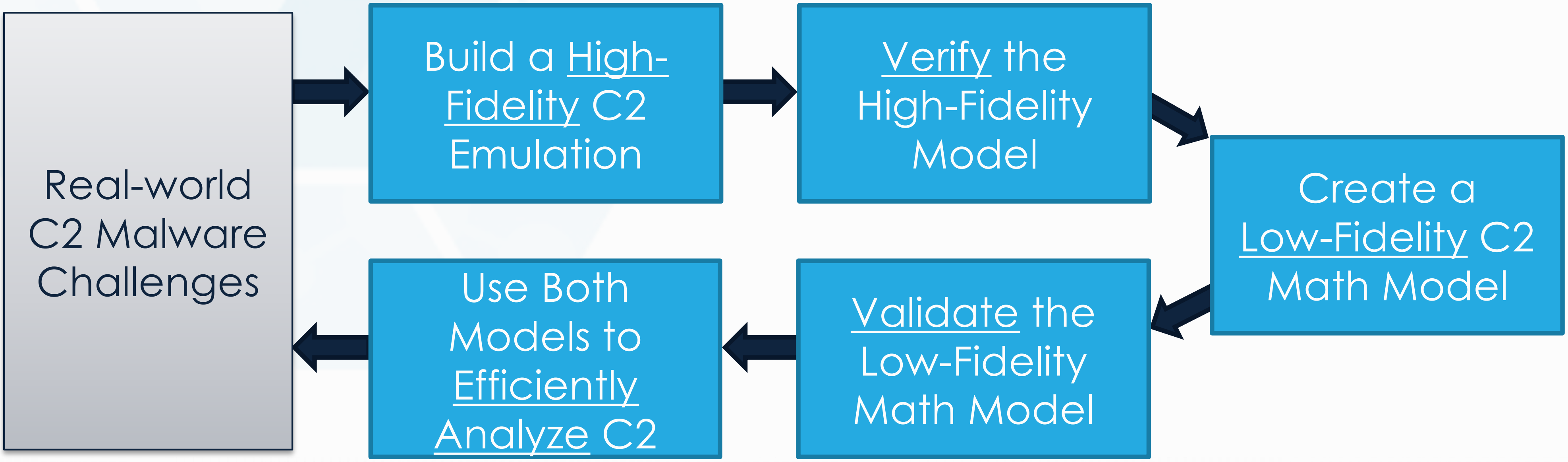
Great news, but only possible because:

1) Low- and high-fidelity models are correlated
2) Low-fidelity model is efficient

Outcomes

- Integrated multiple capabilities to create a rigorous experimentation process for the C2 problem
  - Rigorous experimentation is an improvement on existing work
  - More detailed write-up will be available on SCIRE website
- Publications on model and specific capabilities

C2 Exemplar Analysis Process


Real-world C2 Malware Challenges
Build a High-Fidelity C2 Emulation
Verify the High-Fidelity Model
Create a Low-Fidelity C2 Math Model
Validate the Low-Fidelity Math Model
Use Both Models to Efficiently Analyze C2

Lessons Learned and Future Work

- Lessons learned:
  - Emulations have large inherent variability; highlights need for UQ
  - Starting with the questions before emulating is important
  - Scenario orchestration (ScOrch) is a game-changer for experimentation
- What needs to be done:
  - Additional applications to answer questions related to generalization
    - Do we start to converge on a core set of methods?
  - Reducing development time for math modeling
    - Math models can be useful, but only if accurate; require expertise to build
    - Time and budget required to build math models must be accounted for
  - Just scratched the surface on verification
    - Which indicators matter?
    - How do we say two things are "equal"?

LDRD

Laboratory Directed Research and Development

SCADA network/grid effects studies

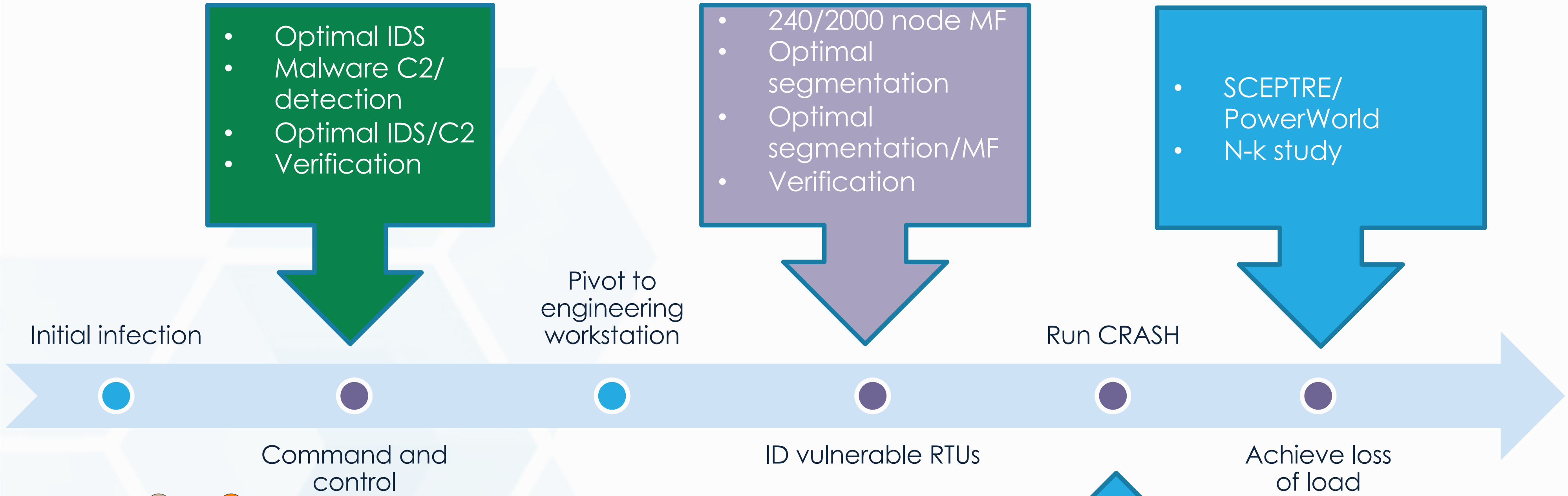
T. Tarman, A. Pinar, L. Swiler, T. Rollins, J. Gearhart, M. Sahakian, E. Vugrin, S. Hanson, C. Glatter, J. Cruz, J. Thorpe, B. Arguello, E. Johnson, C. Phillips, A. Outkin, T. Schulz, collaborators at Texas A&M

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Exemplar studies



- Optimal IDS
- Malware C2/ detection
- Optimal IDS/C2
- Verification

- 240/2000 node MF
- Optimal segmentation
- Optimal segmentation/MF
- Verification

- SCEPTRE/ PowerWorld
- N-k study

Initial infection

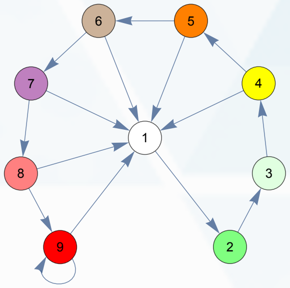
Command and control

Pivot to engineering workstation

ID vulnerable RTUs

Run CRASH

Achieve loss of load



- Assess CRASH on RTUs
- Attacker strategies

- Markov model – distributions, alternative graph structures
- Threat emulation
- Generalization to other APTs
- Validation

Threat modeling

Exemplar studies

- Optimal IDS
- Malware C2/ detection
- Optimal IDS/C2
- Verification

- 240/2000 node MF
- Optimal segmentation
- Optimal segmentation/MF
- Verification

- SCEPTRE/ PowerWorld
- N-k study

Initial infection

Command and control

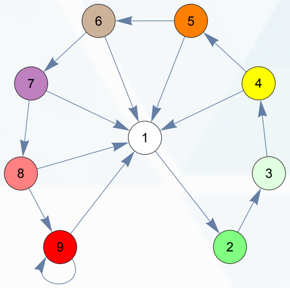
Pivot to engineering workstation

ID vulnerable RTUs

Run CRASH

Achieve loss of load

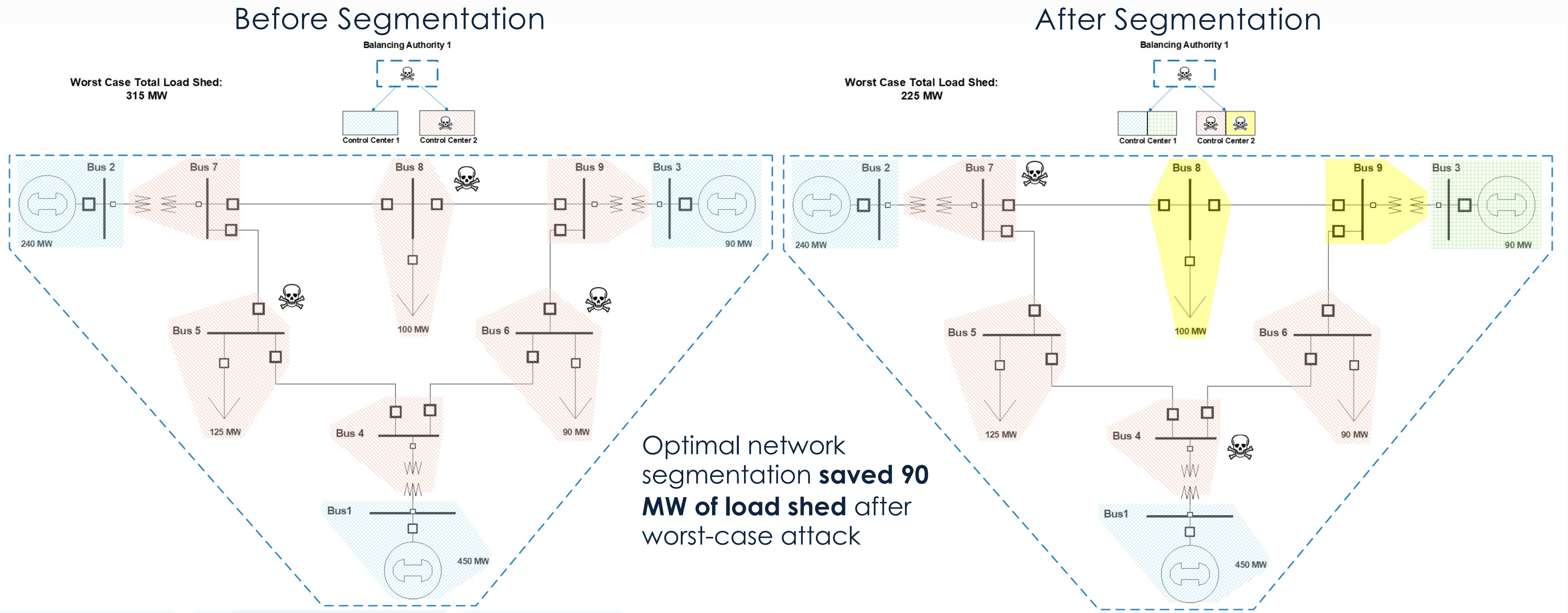
- Assess CRASH on RTUs
- Attacker strategies



- Markov model – distributions, alternative graph structures
- Threat emulation
- Generalization to other APTs
- Validation

Threat modeling

Power system cyber-physical network segmentation

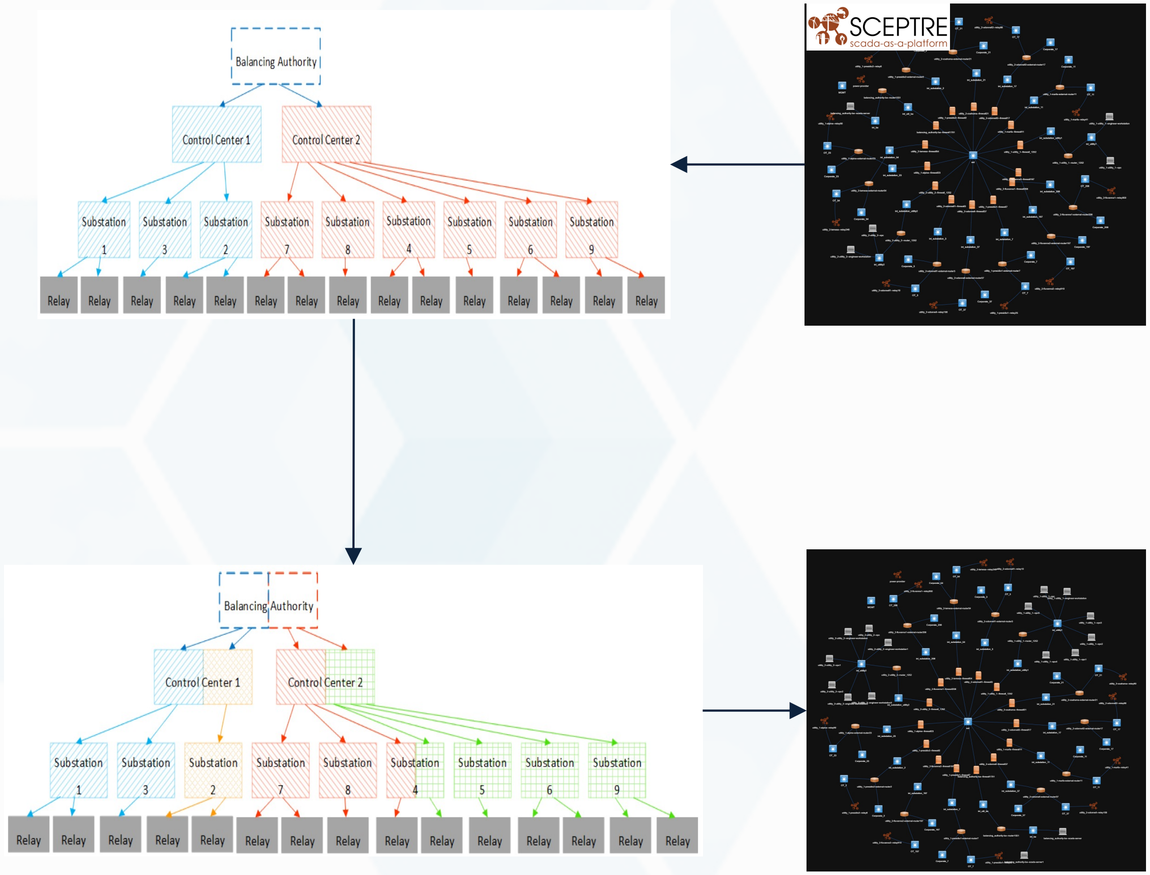


Optimal network segmentation **saved 90 MW of load shed** after worst-case attack

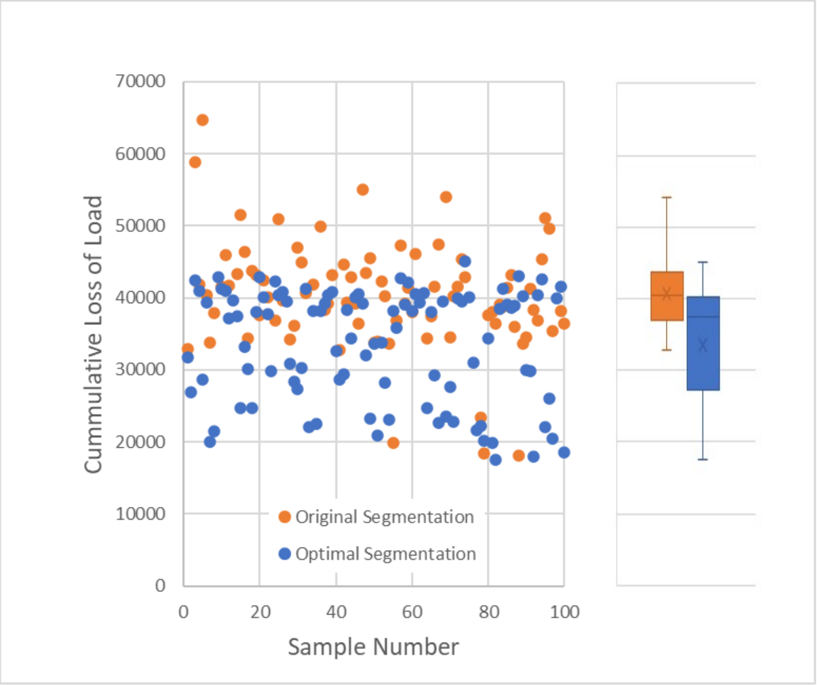
- **Trilevel network segmentation interdiction model** includes:
  - **IT administrator** – how to segment SCADA system against worst-case attack
  - **Attacker** – how to attack the grid after segmentation
  - **Grid operator** – Reoptimizes power flow to serve as much load as possible after worst-case attack
- Trilevel models are **strongly NP-hard**
  - Only 9-bus and 30-bus systems can be solved using new academic bilevel branch-and-bound solvers
- Our academic partners developed a **specialized algorithm** for solving a slightly simplified version of this model
  - Obtained results for the **2000-bus** synthetic system (small SCADA system)

Optimal Segmentation - Can mathematical optimization help identify what experiments to run in emulation?

Optimization/Emulation Workflow



Example Results

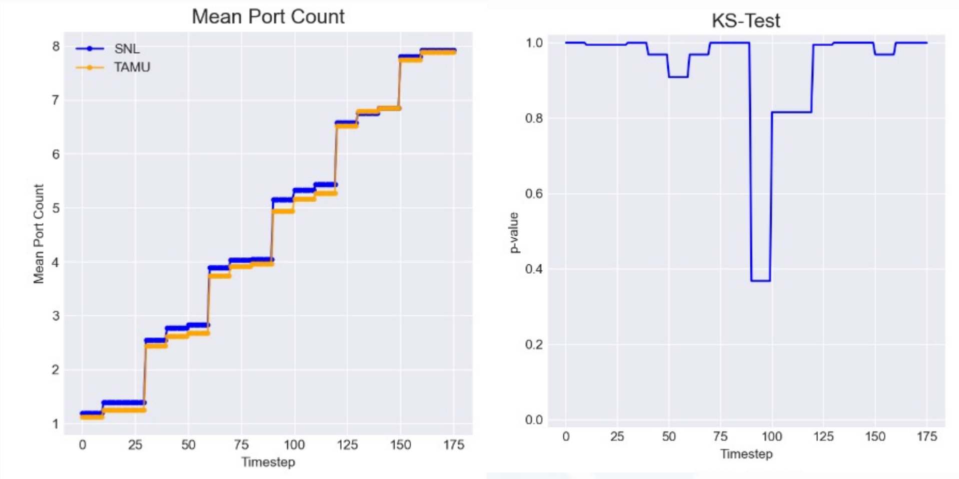


Takeaway: Designed a workflow that interfaces emulation with mathematical optimization to investigate network segmentation

Takeaway: Mathematical optimization identifies a segmentation policy that is more robust under a CrashOverride attack

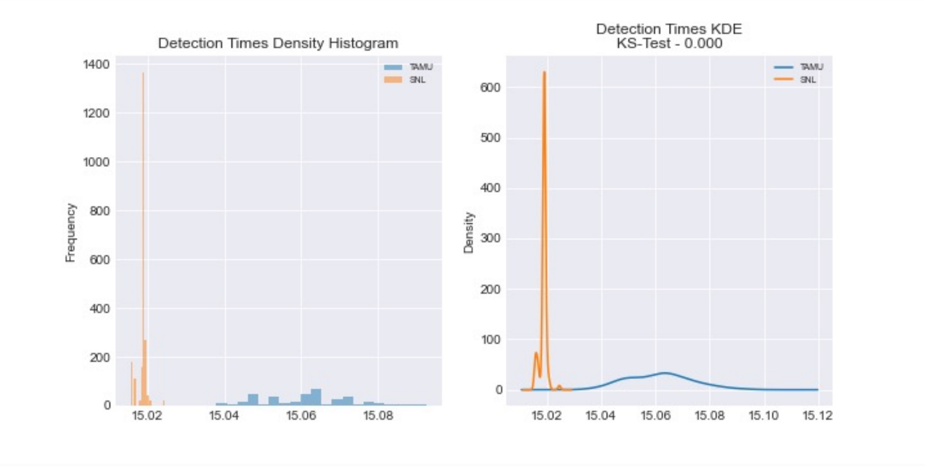
Experiment reproducibility and validation: KS test provides a good metric for comparison

Slow



Port discovery

Fast



Detection

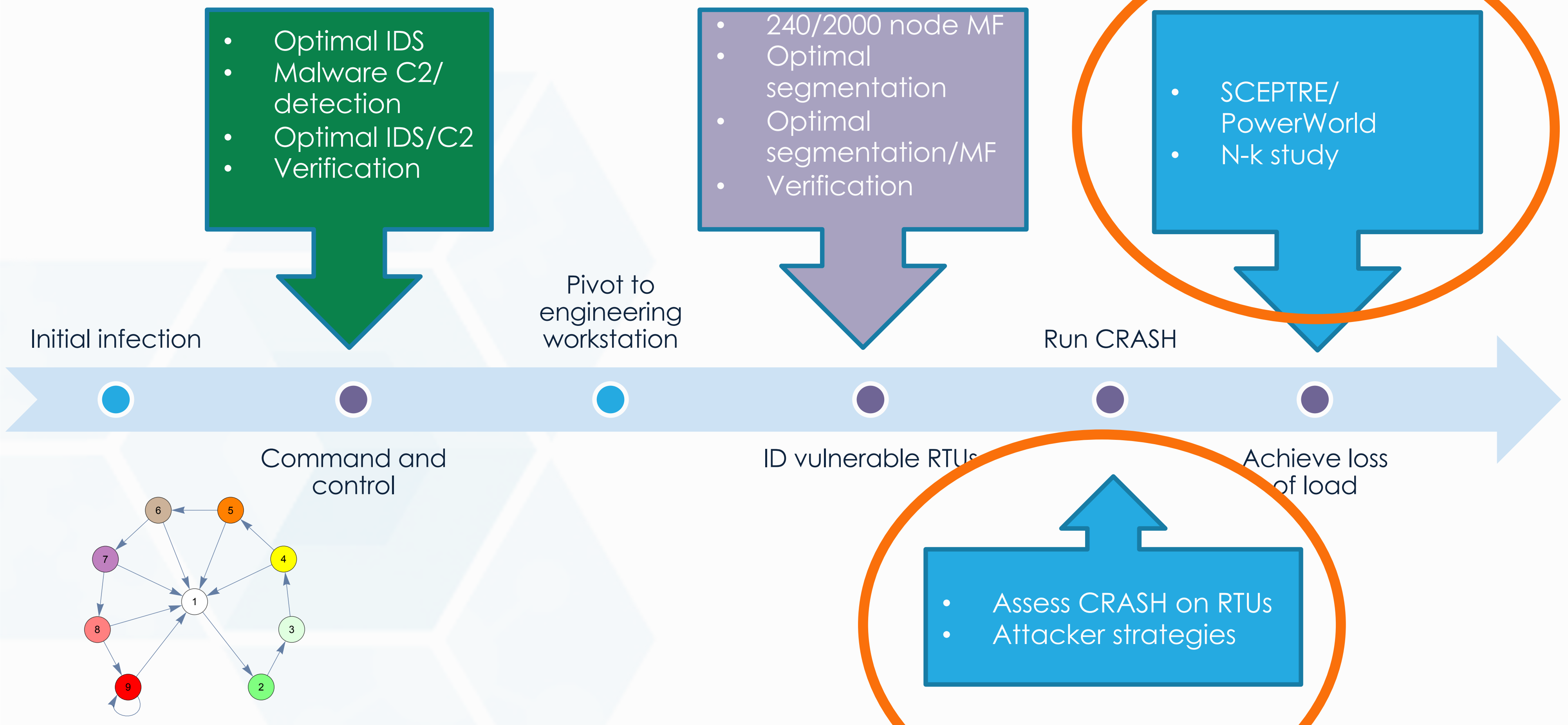
- KS test uncovers subtle differences, and provides statistical test to accept/reject whether CDF differences are statistically significant
- “Agreement” must be considered w.r.t. the question that is being answered
- Other metrics are described in the paper

T. D. Tarman, T. Rollins, L.P. Swiler, J. Cruz, E. Vugrin, H. Huang, A. Sahu, P. Wlazlo, A. Goulart, and K. Davis. Comparing reproduced cyber experimentation studies across different emulation testbeds. USENIX 14th Cyber Security Experimentation and Test (CSET) Workshop. Aug. 9, 2021. SAND2021-5696C.

What did we learn about reproducing emulation experiments?

- Even after providing a comprehensive writeup and details of the experiment, both teams still required significant coordination to reproduce the experiment.
- It can be challenging to determine if small differences are due to differences in the hardware/emulation platform OR due to an implementation detail that is not correctly reproduced.
    - Subject matter expertise is critical
- Statistical tests and ensembles of replicate results can help in this comparison as they provide some estimate of the uncertainty inherent in the results on one platform.
- Recommendations
    - Public repositories for experimental artifacts
    - Need consensus in artifacts and how testbed technologies use them
    - Understand differences between common cyber experimentation platforms
    - Appropriate metrics, depending on experiment question/objective
        - Distance measures between experimental results
        - Metrics to determine effects of platform differences on results

Exemplar studies



Initial infection

- Optimal IDS
- Malware C2/ detection
- Optimal IDS/C2
- Verification

Command and control

Pivot to engineering workstation

- 240/2000 node MF
- Optimal segmentation
- Optimal segmentation/MF
- Verification

ID vulnerable RTUs

Run CRASH

- Assess CRASH on RTUs
- Attacker strategies

- SCEPTRE/ PowerWorld
- N-k study

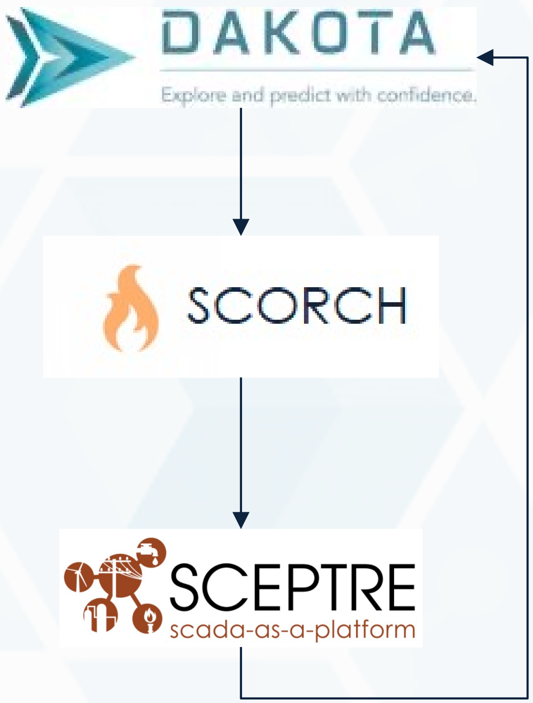
Achieve loss of load

- Markov model – distributions, alternative graph structures
- Threat emulation
- Generalization to other APTs
- Validation

Threat modeling

Can we apply UQ methods to evaluate the power grid impacts of the CrashOverride malware?

UQ/Emulation Workflow



DAKOTA

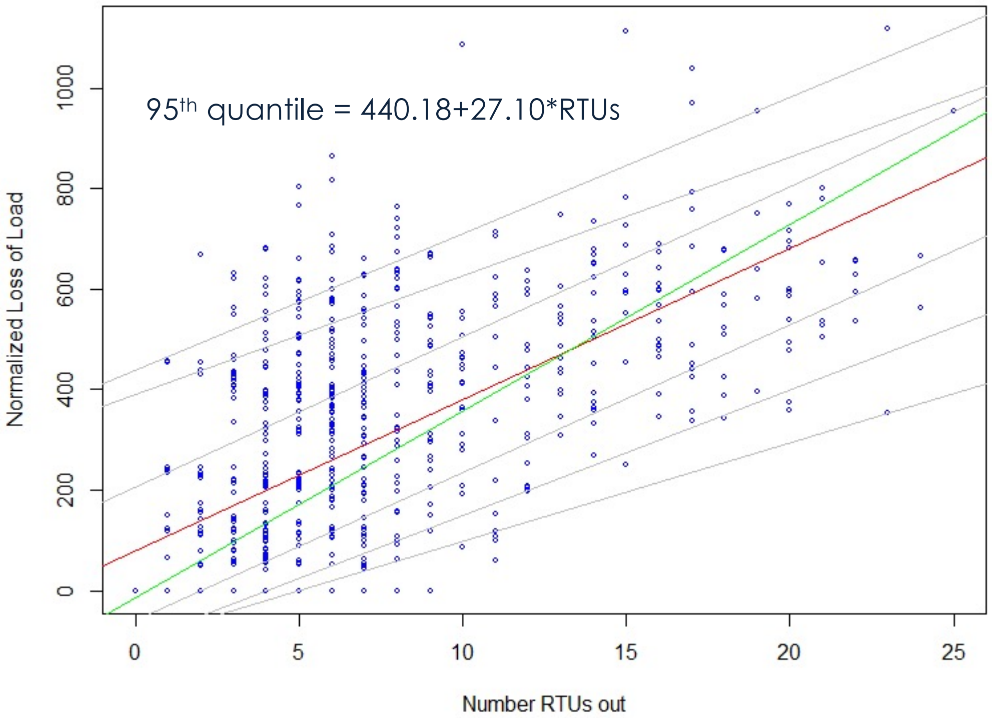
Explore and predict with confidence.

SCORCH

SCEPTRE

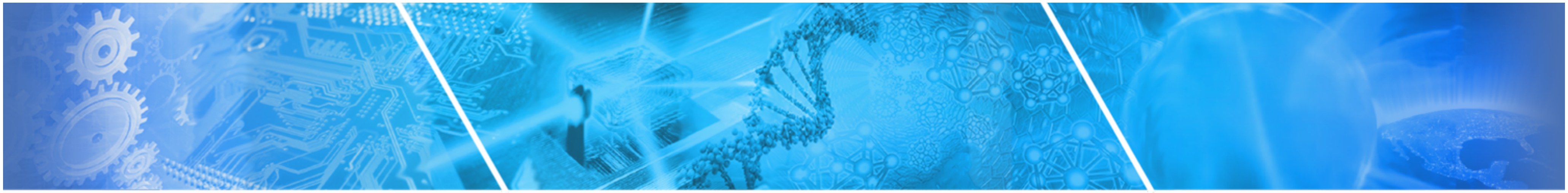
scada-as-a-platform

Example Results



Takeaway: Traditional UQ tools can be coupled with emulation capabilities to enable better characterization of uncertainty

Takeaway: Strength of analysis depends on approach. Quantile regression does better at finding a linear fit



LDRD

Laboratory Directed Research and Development

SECURE Verification and Validation

Laura Swiler

Tom Tarman

Jamie Thorpe

Bert Debusschere

Kasimir Gabert

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

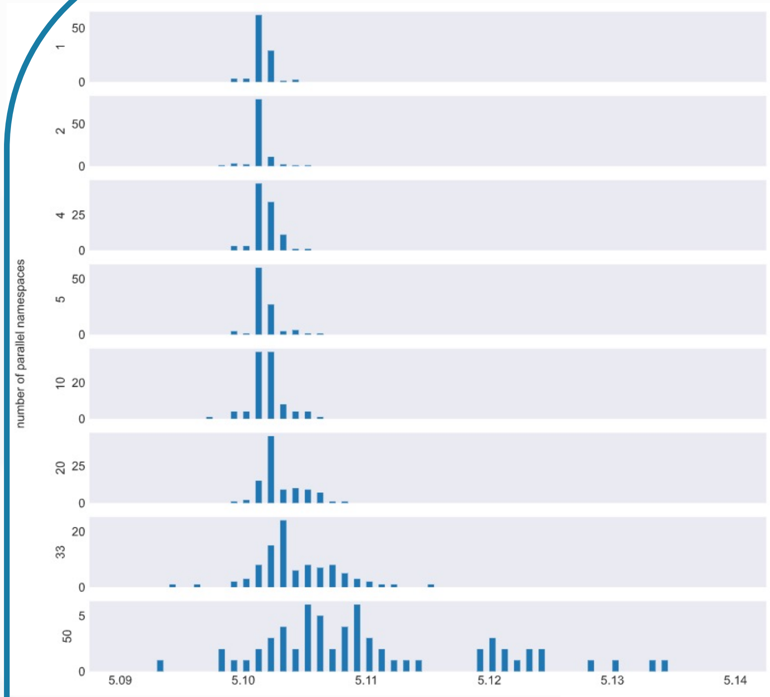
Verification

Detecting Over-subscription of Resources in Emulation-Based Experiments

- Scanning detection scenario case-study used for verification
- Repeated studies launched using increasing number of namespaces
- A namespace is **an experiment that is isolated in its own VLAN or set of VLANs.**
  - Each namespace has its own copy of each machine in the scenario and they are networked through a unique set of VLANs.
  - We can run multiple namespaces in parallel while the experiments remain isolated in their own namespace.
- Results presented by Jerry and Trevor in November

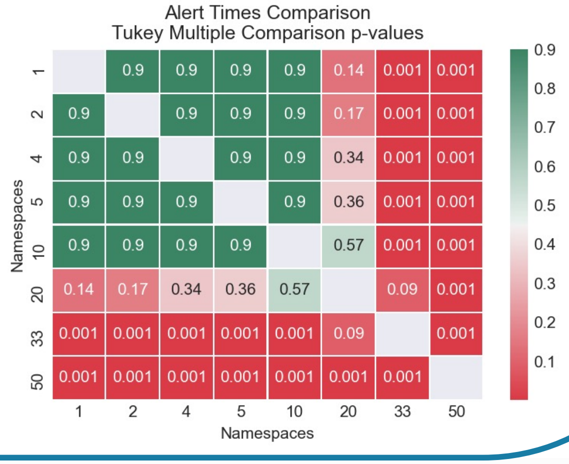
Alert Times - All

- Distribution of alert times shift as namespaces are added
- Quantified similarity with Tukey Multiple Comparison Test
  - Shows clear drop in similarity after 10 namespaces
- Large p-value indicates that the null hypothesis can't be rejected
  - $H_0: \mu_1 = \mu_2$
  - **Larger p-value -> similar results**



Alert Times Distribution

Tukey Multiple Comparison

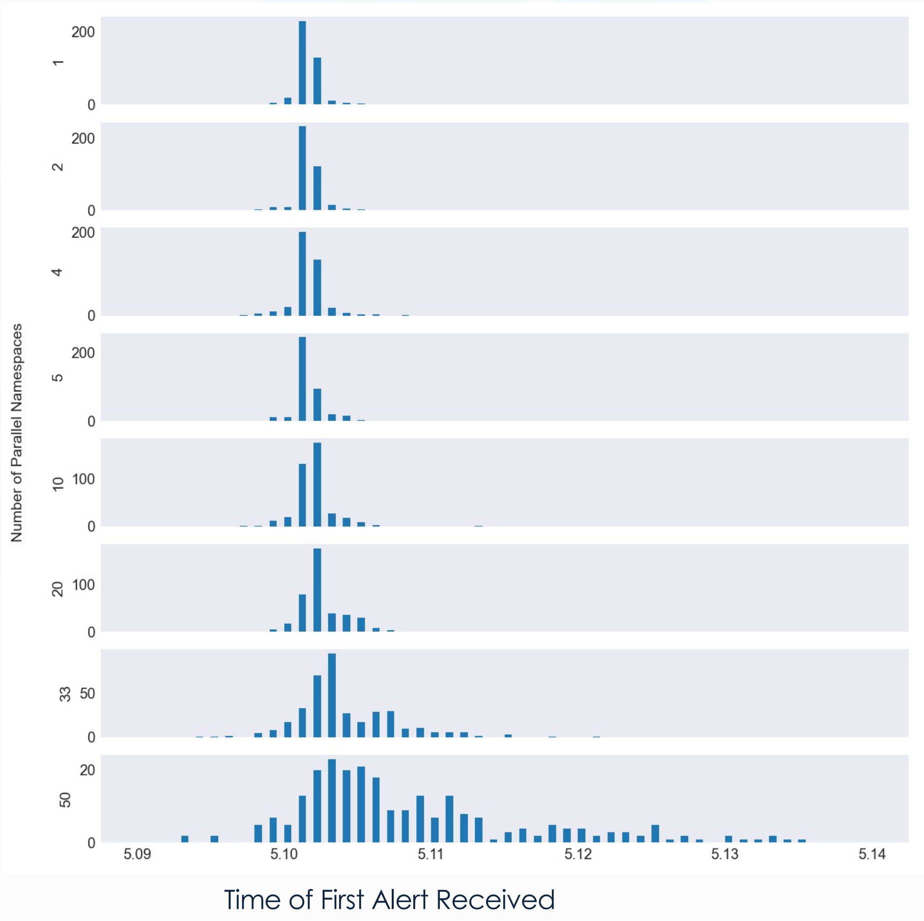

Alert Times Comparison
Tukey Multiple Comparison p-values

Namespaces	1	2	4	5	10	20	33	50
1		0.9	0.9	0.9	0.9	0.14	0.001	0.001
2	0.9		0.9	0.9	0.9	0.17	0.001	0.001
4	0.9	0.9		0.9	0.9	0.34	0.001	0.001
5	0.9	0.9	0.9		0.9	0.36	0.001	0.001
10	0.9	0.9	0.9	0.9		0.57	0.001	0.001
20	0.14	0.17	0.34	0.36	0.57		0.09	0.001
33	0.001	0.001	0.001	0.001	0.001	0.09		0.001
50	0.001	0.001	0.001	0.001	0.001	0.001	0.001	

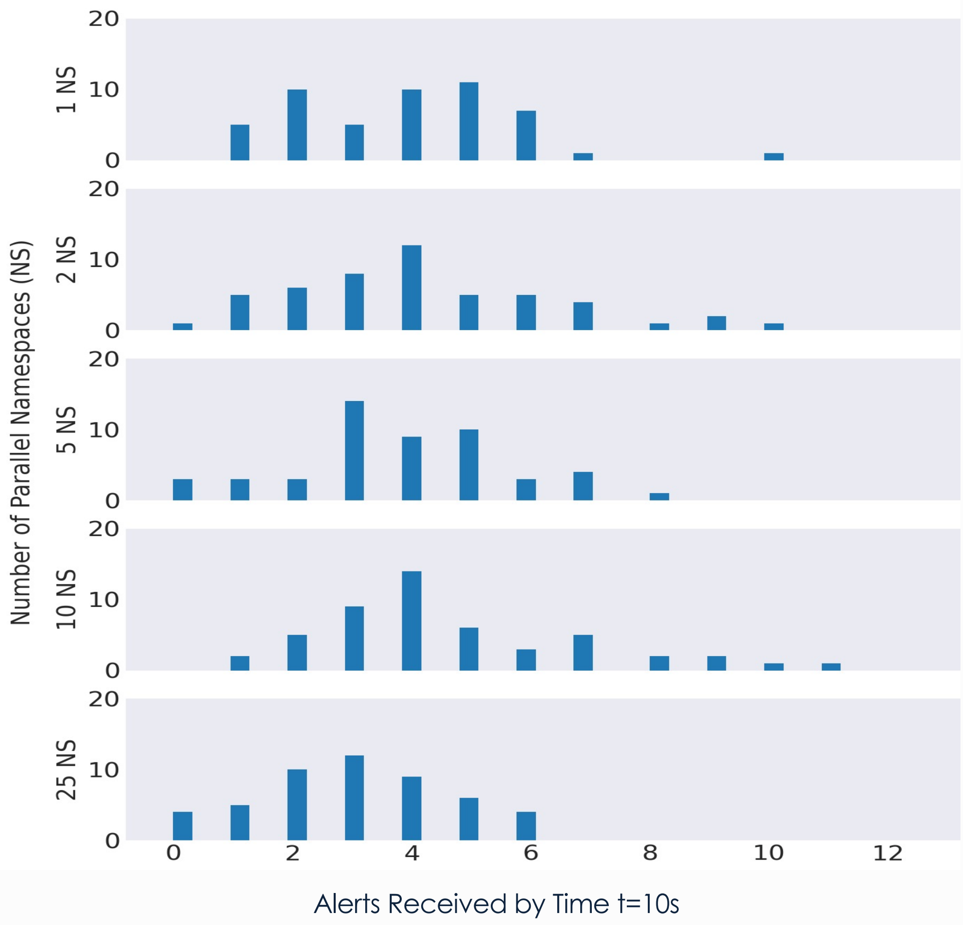
Scanning/Detection vs Command & Control

- Hypothesis: Verification Process from Scanning/Detection is **Generalizable**
- Command & Control Results are inconclusive: we don’t see data we should reject as we push the experiment to more namespaces. Why?

Scanning/Detection, All Replicates

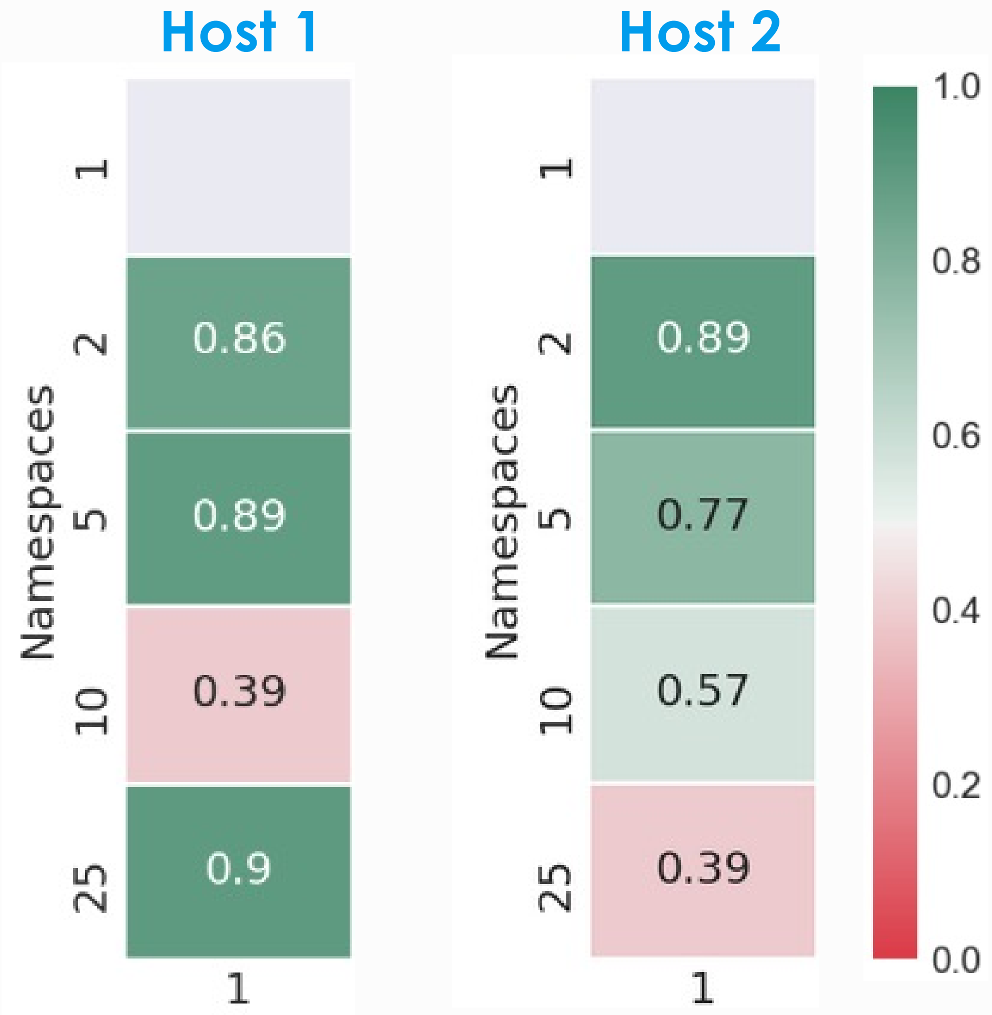


Command & Control, All Replicates



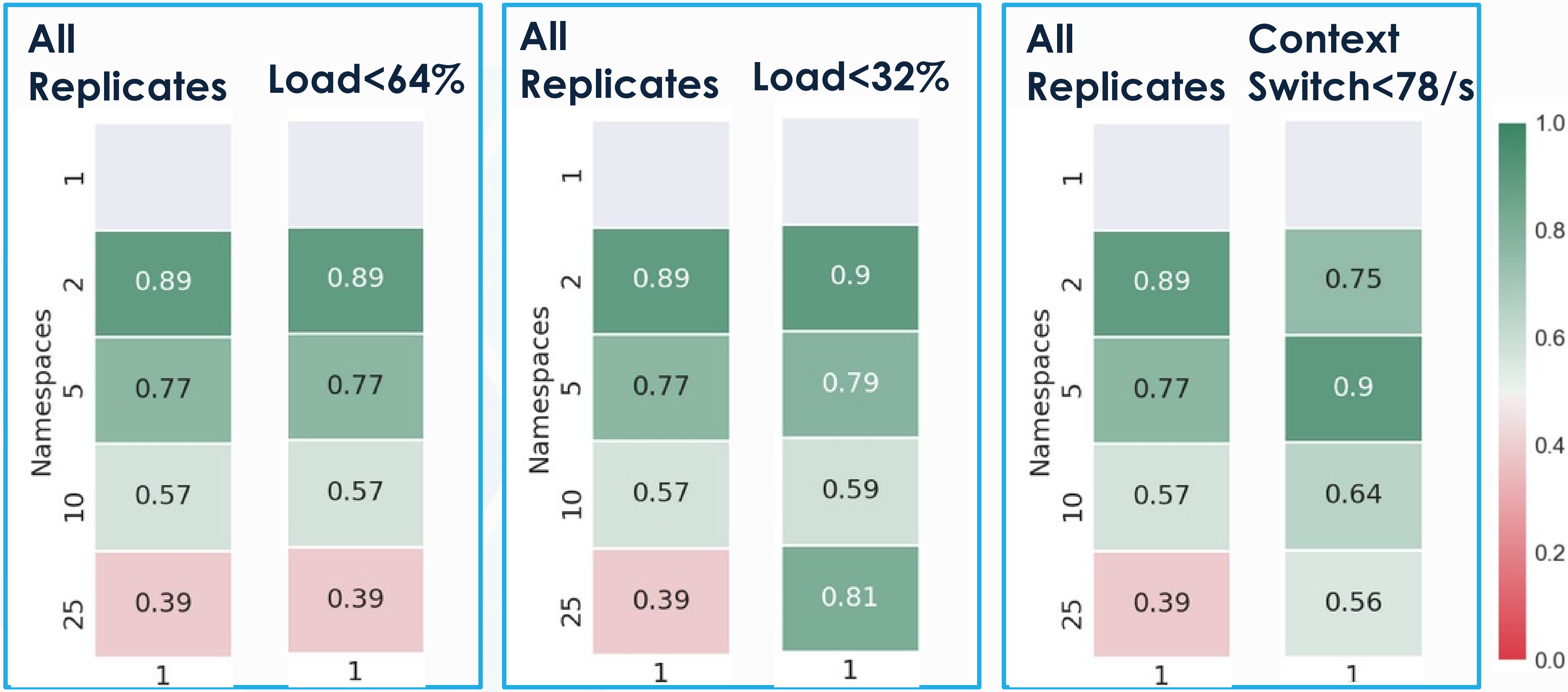
Command & Control Results: Host 1 vs Host 2

- Baseline: 1 Namespace
- Tukey Analysis - Compare Quantity of Interest Distribution for Increasing Parallel Namespaces
  - All Replicates Included
- Two Different Physical Hosts, Supposedly Identical Configuration
- Difference in Tukey for Different Hosts, even with Apparently Identical Resources



The statistical results indicate we would not reject any of these runs.

Command & Control Results:
Use of Telemetry to Filter Unreliable Replicates



Replicates Remaining: 250/250

Replicates Remaining: 211/250

Replicates Remaining: 179/250

Bottom Line: Balance Desired Similarity to Baseline with Number of Replicates being Removed

Can we use Telemetry to Identify Unreliable Experiment Results?

- **Repeated the Approach** from Scanning/Detection Scenario
  - Telemetry: Stolen Cycles, Load, Throughput, Context Switches, User Time
  - Identify **Threshold for Telemetry** that Experiments Should Not Cross
- **Command & Control Data Challenges**
  - Addressing Experimentation Bugs
  - Collecting, Processing, and Storing **Large Volumes** of Data
    - 17 Mb PCAP file -> 359 Mb JSON -> 320,000 entries -> **several hours** to store **per replicate**
    - Solution reduces to 1 stored entry per replicate, taking **~1 hour** to process **in total**
  - Adapting Analysis to **New Quantities of Interest, new Thresholds**
- **Differences in the two experiments**
  - The Command & Control scenario seemed more complex, yet resources were not pushed to oversubscription
  - **Scanning/Detection scenario involved 27 VMs, the C2 scenario involves 4.**
  - Hard to tease out what is the critical factor: the number of VMs vs. packets/sec vs. overall traffic
  - Scanning/Detection had 27VMs x 25 namespaces = 675 TCP handshakes tracked through one host kernel; vs. C2 had 4VMs x 25 namespaces =100 TCP handshakes.
  - Additionally, Scanning/Detection used tc (traffic control) to implement packet drops; this is high overhead and a stateful process.

Summary and Takeaways

1. We were interested in repeatability of our experiments under different platform configurations
2. We instrumented the platform to collect telemetry from the VMs in the experiments and from the physical host running the experiments. This telemetry can be used to develop metrics (or indicators).
3. We tested the efficacy of the telemetry metrics by removing replicates that violated the thresholds and testing whether different numbers of namespaces gave statistically different results

- **Takeaways:**
  - Both the ***system load*** and ***throughput*** help filter out replicates that generate statistically different results for scanning/detection
  - These same metrics do not indicate a problem for the C2 scenario. We currently think that we are not stressing the C2 scenario as much.
  - Currently we are investigating how we can push the C2 scenario further:
    - More namespaces: issues with Python threading in SCORCH
    - More traffic: PCAP size involves significant postprocessing times (hours to days)
  - The instrumentation of these verification experiments and statistical analysis of the telemetry metrics to find clear indicators of anomalies has been challenging.

Validation

Validation

- **Validation Metrics**
  - **Ultimate goal is to compare physical testbeds with emulation**
  - **Reproducibility Study: SNL minimega vs. TAMU CORE testbed (Tom's talk at CSET)**
  - **Latest study: SNL minimega vs. TAMU physical testbed.**
    - This is what I am presenting today. Note the TAMU "physical" testbed is only partly physical: has real RTUs for the open ports, but the closed ports are emulated and the filtered ports are handled by the switch.

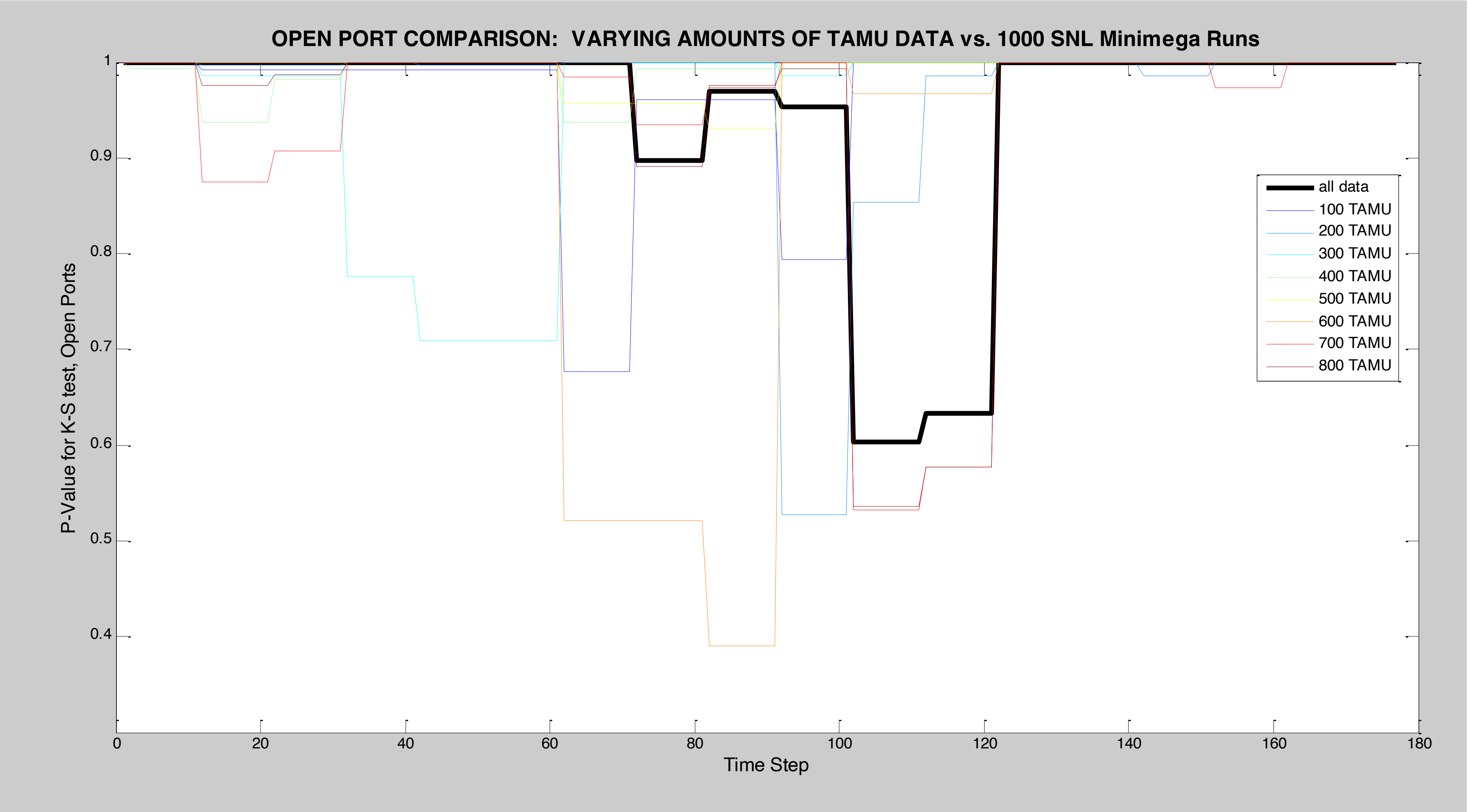
- **Scaling**
  - **How much validation at small scale can we use to build up a validation case at large scale?**
  - Kasimir Gabert's Ph.D. work at Georgia Tech.
    - Graph analysis approach: instead of topology graph, create communications graphs

Validation with physical test data

- This is closer to a true validation
- The TAMU testbed is still a mix of emulated/virtualized components and actual physical components
- For the network scanning/detection, the four Remote Terminal Units that have the "open" ports are physical units.
- We used the same experimental orchestration
- 1000 minimega runs and 823 TAMU physical results

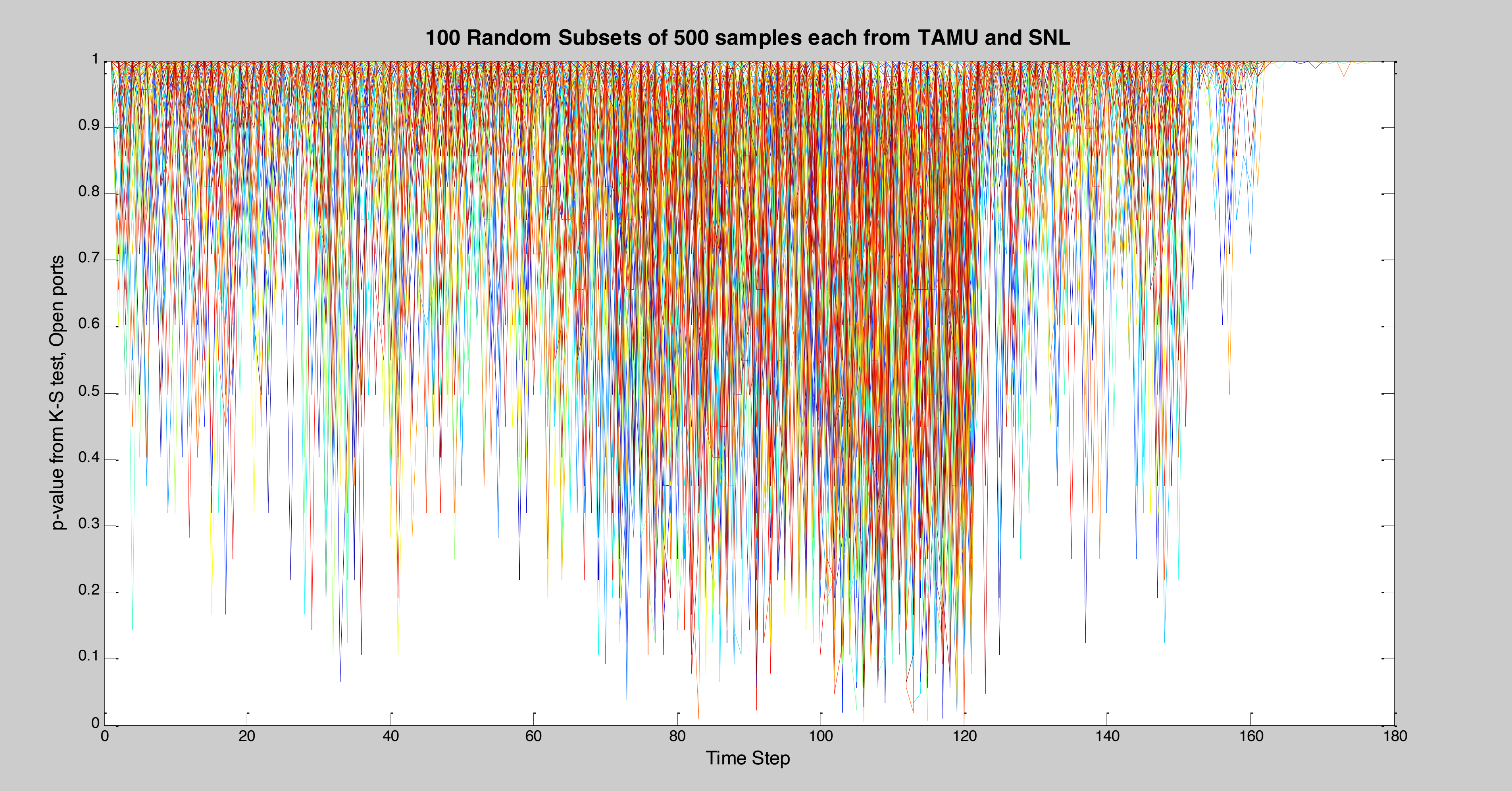
Validation test comparison: 1000 SNL minimega runs vs. 823 Physical tests at TAMU

- We don’t see monotonic improvement in K-S test statistics



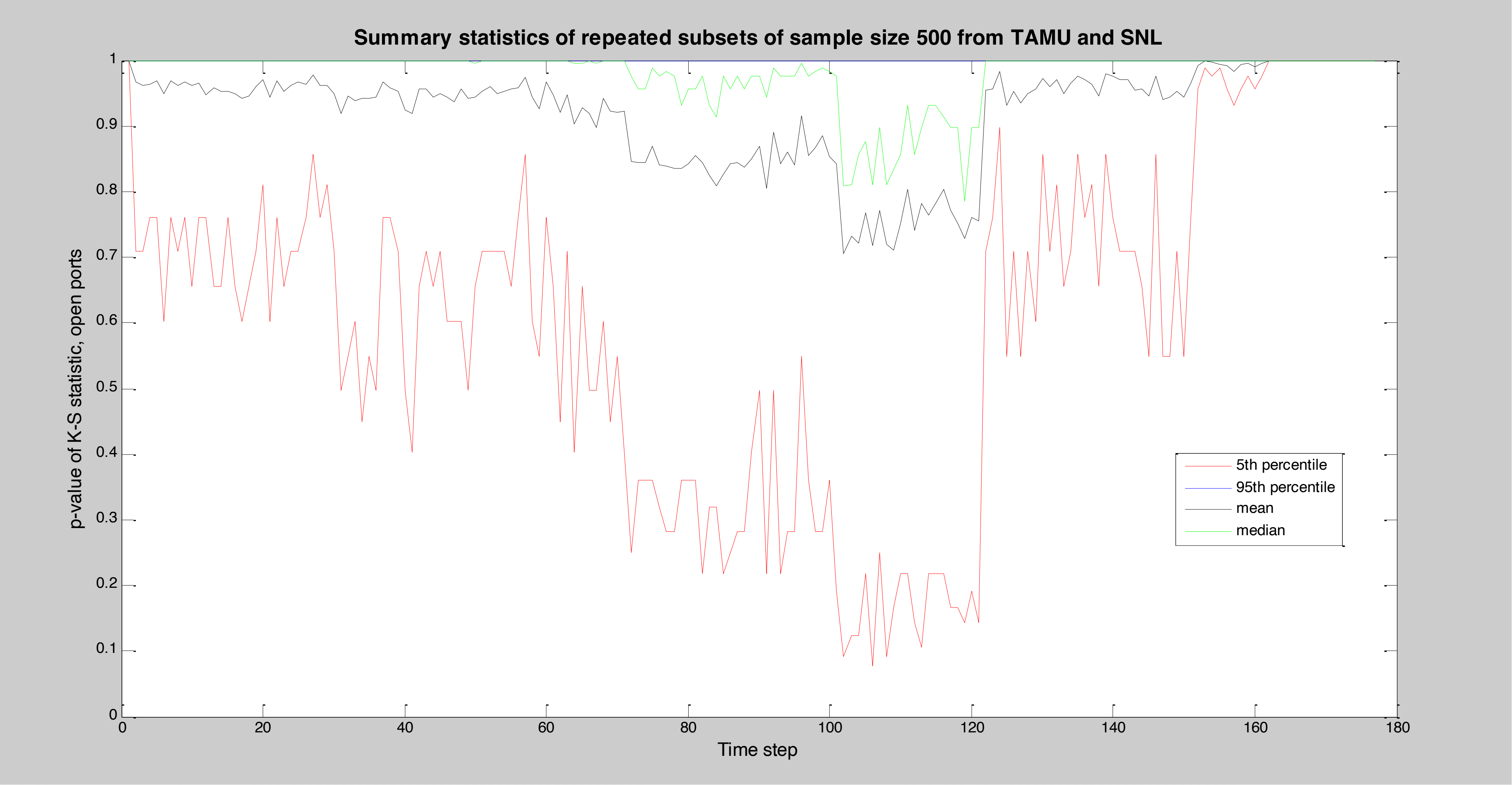
Take random subsets of 500 samples each from SNL and TAMU

- LARGE VARIABILITY



Statistics on the p-values

- Mean and median indicate good agreement. The low values of the 5th percentile between 100-120 seconds help identify times which have some realizations with less agreement.



Lessons learned: Validation

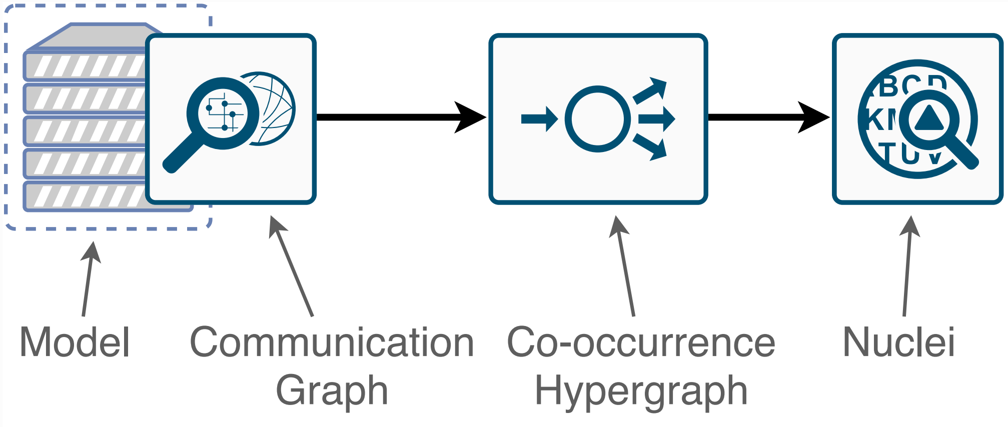
- Need consensus in artifacts and how testbed technologies use them
  - Understand differences between common cyber experimentation platforms
  - Virtualization technologies (CPUs, network interfaces, switching, etc.)
  - Public repositories for experimental artifacts
- It can be challenging to determine if small differences are due to differences in the hardware/emulation platform OR due to an implementation detail that is not correctly reproduced.
  - Subject matter expertise is critical
- Statistical tests and ensembles of replicate results can help in this comparison as they provide some estimate of the uncertainty inherent in the results on one platform.
  - Appropriate metrics, depending on experiment question/objective
- Repeatability + Reproducibility can strengthen Validity
  - A result that holds under repetition and various modeling environments is more likely to be true on the real system
  - Key to bringing rigor to cybersecurity modeling

Prioritizing Components to Validate

- Hierarchical validation: validate components, sets of components
- How to prioritize components given a finite validation budget?
  - Previously, only SME guidance is available
- High-level idea: monitor all communication within the model, find what is important to the model output programmatically
- Critical assumptions: (1) components are low-latency (2) key model output events occur due to network traffic

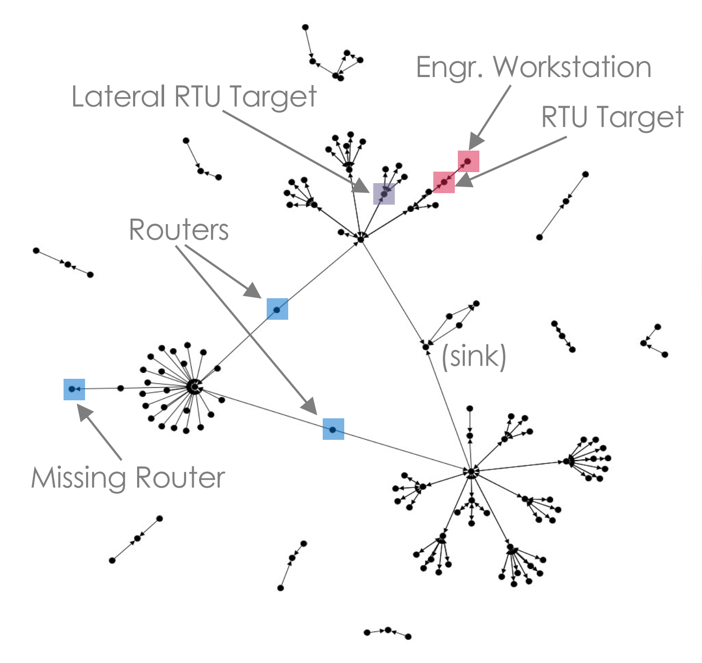
Approach

1. Run the model
2. Build communication graph
3. Build co-occurrence hypergraph around outputs
4. Track cores/nuclei, return dense regions that occur with outputs

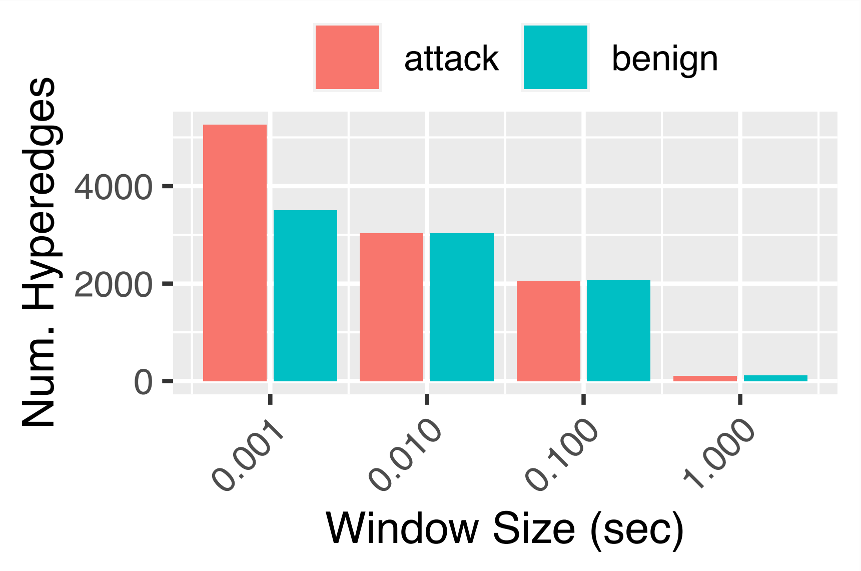


CrashOverride Experiment Results

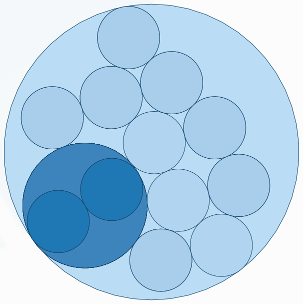
Communication graph



Window Size (Δ) Effect



Hypergraph cores with $\Delta = 0.01$



Resulting Ordered Components

Engr. Workstation	10.53.0.18
RTU Target	10.52.2.3
Lateral RTU Target	10.52.11.2
Impacted Routers	10.53.0.17 10.54.0.17
Missing Router	172.30.0.18

Tracking Dense Regions of Changing Graphs

- *A Unifying Framework to Identify Dense Subgraphs on Streams: Graph Nuclei to Hypergraph Cores*, WSDM (18.6% acceptance)
  - Problem: nuclei are expensive, cannot recompute when graph changes
  - Approach: convert to a hypergraph and maintain cores instead
- *Shared-Memory Scalable k-Core Maintenance on Dynamic Graphs and Hypergraphs*, IPDPSW ParSocial
  - Problem: prior approaches do not use multiple processors effectively
  - Approach: use the connection between h-indices and k-cores
- *ElGA: Elastic and Scalable Dynamic Graph Analysis, SC (26.8%)*
  - Problem: large, changing graphs become too big to fit in memory
  - Approach: use consistent hashing to load balance, sketches for state
- Coreness to Cores: Batch Dynamic Algorithm to Efficiently Find k-Cores, prepared
  - Problem: prior work performs decompositions but not core hierarchies
  - Approach: use a query-efficient index to maintain full hierarchies

Closing thoughts

Summary

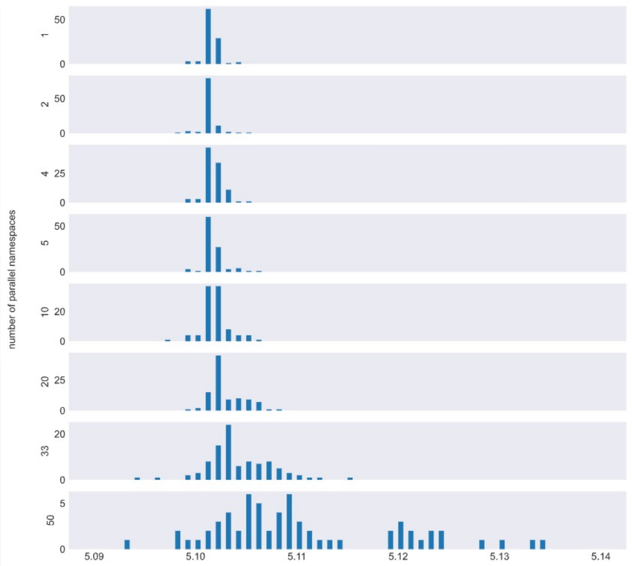
- Verification
  - Focus on oversubscription of resources
  - Instrumented minimega, results sent back to ElasticSearch for postprocessing analysis
  - Scanning/detection showed more variance in runs that were spread over many namespaces, use of telemetry metrics such as system load, stolen cycles, and throughput
  - C2 results inconclusive
- Validation
  - Validation metrics: use of K-S test, area metric, comparison of means
  - Scaling: Physical experiments at TAMU, Kasimir's research
  - Statistical distribution tests showed more variability than expected
  - Slow convergence of distribution test statistics
- Uncertainty quantification
  - Efficient UQ methods exist (structured designs, multifidelity UQ, PCE)
  - In some cases, we may be able to run enough emulations to formalized a tail probability relationship (e.g. quantile regression)
  - CHALLENGE: Identify effects of stochasticity AND parameter uncertainty, use of both information in tail probability estimation with MF methods

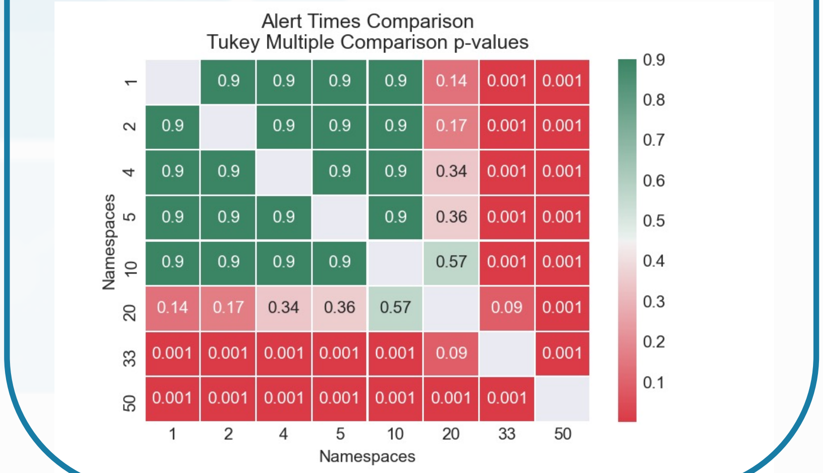
Backup

Alert Times – System Load Cutoff

- Replicates with **System Load** above a threshold were filtered out
- 244 / 760 replicates removed
  - Filtered NS 33 & 50 only

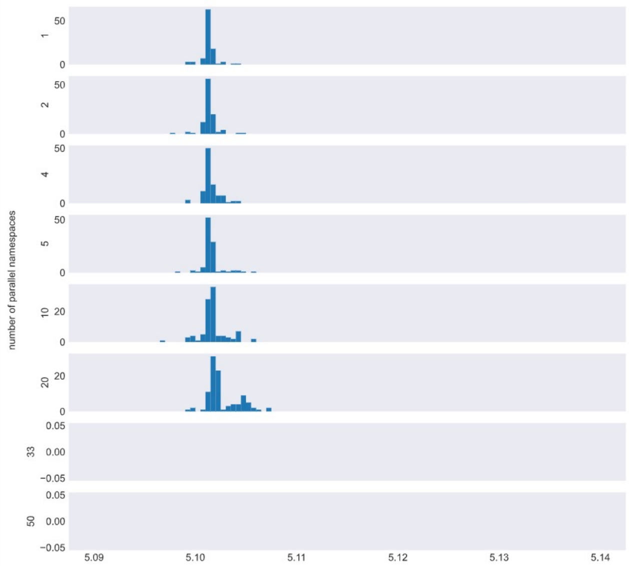
All Replicates

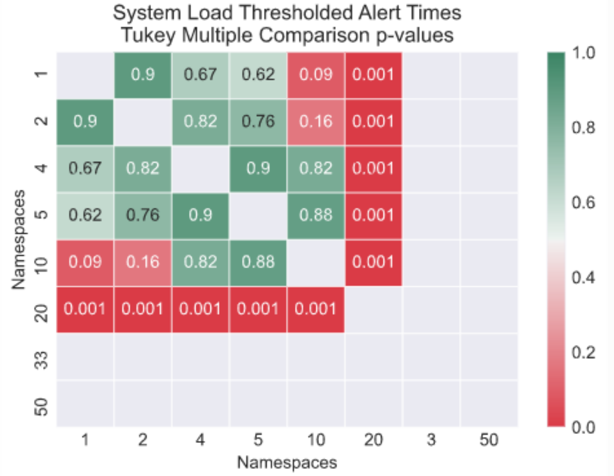



Alert Times Comparison
Tukey Multiple Comparison p-values

	1	2	4	5	10	20	33	50
1		0.9	0.9	0.9	0.9	0.14	0.001	0.001
2	0.9		0.9	0.9	0.9	0.17	0.001	0.001
4	0.9	0.9		0.9	0.9	0.34	0.001	0.001
5	0.9	0.9	0.9		0.9	0.36	0.001	0.001
10	0.9	0.9	0.9	0.9		0.57	0.001	0.001
20	0.14	0.17	0.34	0.36	0.57		0.09	0.001
33	0.001	0.001	0.001	0.001	0.001	0.09		0.001
50	0.001	0.001	0.001	0.001	0.001	0.001	0.001	

Filtered Replicates




System Load Thresholded Alert Times
Tukey Multiple Comparison p-values

	1	2	4	5	10	20	3	50
1		0.9	0.67	0.62	0.09	0.001		
2	0.9		0.82	0.76	0.16	0.001		
4	0.67	0.82		0.9	0.82	0.001		
5	0.62	0.76	0.9		0.88	0.001		
10	0.09	0.16	0.82	0.88		0.001		
20	0.001	0.001	0.001	0.001	0.001			
33								
50								

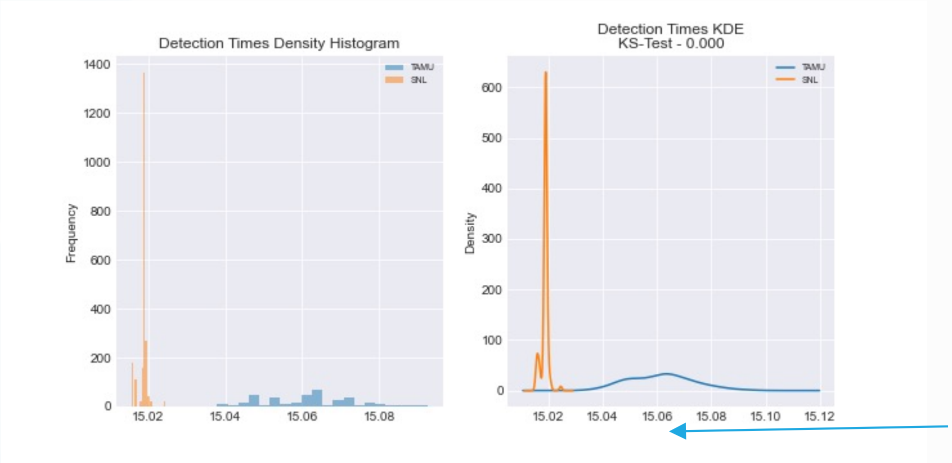
How well was the experiment reproduced?

- We may care about the differences in magnitude and not care about distributional differences.

Alert detection times

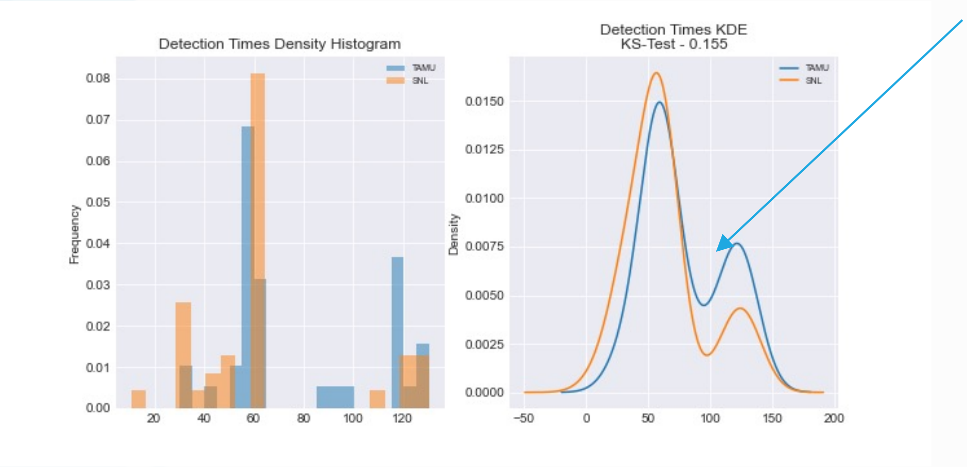
Fast – No Drop – Fixed Nmap order

- KS-test: 0.000
- Least variable experiment



Slow – Drop – Random Nmap order

- KS-test: 0.155
- Most variable experiment



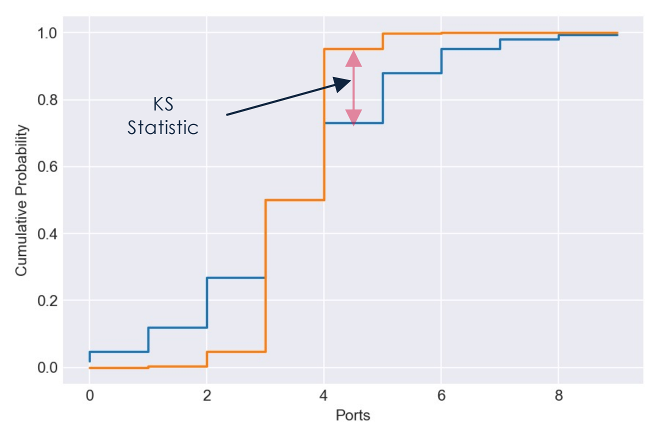
The validation metrics depend on the question being asked: Are these differences significant due to differing hypervisors, time synchronization, and experiment orchestration? Are they acceptable to be used in a larger attack model?

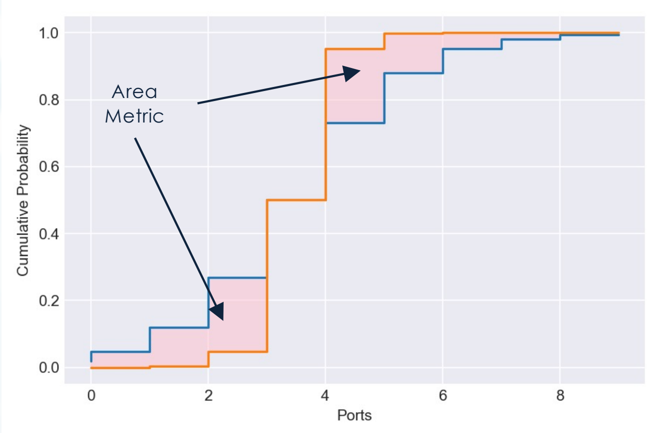
Metrics Used in this Study

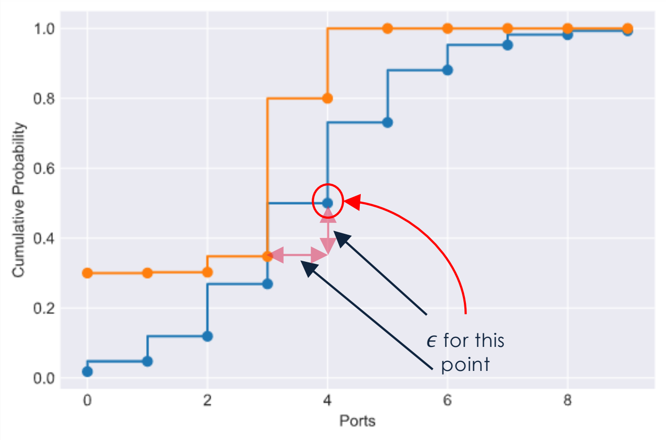
- **KS-Test**
  - Maximum value of the distance in the CDFs of two samples
  - P-value for this statistic is used
    - $H_0\colon CDF_1 = CDF_2$
    - **Large values imply similarity**

- **Area Metric**
  - Sum of the differences in area between the CDFs of two samples [1]
  - This is not a p-value, **small values imply similarity**

- **Relative Hausdorff Distance**
  - Developed as a metric to quantify graph similarity
  - For every point (p) on CDF_1 there is a point (p') on CDF_2 such that:
    - $|p - p'| \leq \epsilon p$
    - $|CDF_1(p) - CDF_2(p')| \leq \epsilon CDF_1(p)$
    - The minimum ϵ that satisfies this is the RH distance [2]
  - Allows for "play" in vertical and horizontal direction
  - This is not a p-value, **small values imply similarity**







[1] K.A. Maupin, L.P. Swiler, N.W. Porter, "Validation Metrics for Deterministic and Probabilistic Data," *Journal of Verification, Validation and Uncertainty Quantification*, Vol. 3, September 2018.

[2] O. Simpson, C. Seshadhri, and A. McGregor, Catching the head, tail, and everything in between: A streaming algorithm for the degree distribution, in 2015 IEEE International Conference on Data Mining, IEEE, nov 2015.

Metrics

- **Kolmogorov-Smirnov Test Statistic**
  - Well known non-parametric statistical test for equality of distributions

$$\mathrm{D}_{n,m} = \sup_{x} | F_{1,n}(x) - F_{2,m}(x) |$$

  - Test statistic converges to Kolmogorov distribution: there are formulas for rejection of the null hypothesis $\mathrm{CDF}_1 = \mathrm{CDF}_2$ at various confidence levels.

- **Area metric**
  - Does not just depend on the largest discrepancy between CDFs: accounts for the entire difference
  - No formal acceptance metric or statistical test

$$\mathrm{A}_{n,m} = \sum_{x} | F_{1,n}(x) - F_{2,m}(x) |$$

  - Units are in same units as the measurement: Ferson et al. argue it is best not to normalize but use some judgement about acceptable tolerances

Scott Ferson, William L. Oberkampf, Lev Ginzburg. "Model validation and predictive capability for the thermal challenge problem," *Computer Methods in Applied Mechanics and Engineering,* Volume 197, Issues 29–32, 2008

Metrics

- **Relative Hausdorff**
  - Originally developed for graph analysis, for quantities like the complementary cumulative degree distribution of large graphs
  - The distributions F_1 and F_2 are (ε, δ) close by the Relative Hausdorff distance if

$$\forall x, \exists x' \in [(1-\varepsilon)x, (1+\varepsilon)x'] \quad such\ that\ |F_1(x) - F_2(x')| \le \delta F_1(x)$$

  - Note that although the degree distributions involve discrete variables, the examples used have a much larger support and are smoother than our empirical distribution functions for port counts, for example

- **Other ideas: perform data alignment before applying metrics**
  - Edit distance: if we know the timing is different for some fundamental reason (e.g., a 2 second offset), can we shift the times of the second CDF and use the above metrics
  - There are also functional analysis tools which attempt to overlay signals in x- and y- coordinates (e.g. align peaks of signals as well as phase).

J. Derek Tucker, Wei Wu, Anuj Srivastava, "Generative models for functional data using phase and amplitude separation" *Computational Statistics & Data Analysis,* Volume 61, 2013, Pages 50-66.

Backup: Experimental Setup

Analysis ID	1	2	3	4	5
# Parallel Namespaces	1	2	5	10	25
# Replicates per Namespace	50	25	10	5	2

- Analyses run in succession on single Carnac nodes with NUMA nodes 2 and 3 disabled in order to limit scheduling capability
- ScOrch tool used for experiment orchestration

# CPUs on IDS VM	1
Benign Traffic Rate/s	1000
Benign Probability of Emotet Signature	0.001
Malicious Traffic Rate/s	20
Malicious Probability of Emotet Signature	0.125

LDRD

Laboratory Directed Research and Development

"Life After LDRD"

Derek Hart, PM

25 August 2021

U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Goals

- Create a sustained internal team
  - Members of the current GC and others around the lab
- Ensure sufficient follow-on funding to keep momentum moving
- Build on EAB and other external interactions to build a collaborative community
- Motivate internal community to accept a more scientific approach to cybersecurity

Opportunities

- Good news: Government is waking up to the seriousness of cyber, especially when it comes to defending critical infrastructure
- Bad news: Much of the discussion is around hygiene and doing the basics. Less focused on how to measure the degree of cybersecurity of a system
- NNSA is launching a cyber survey of the nuclear deterrent enterprise
- DHS CISA and DOE are in line to receive large increases in fund to confront cyber issues

Challenges

- Slow recognition of cyber as a science within the government
  - Engineering enterprise vs. a scientifically-informed enterprise
- Changing hearts and minds is taking time
- Funding for lower TRL, strategic cyber work is difficult to secure when the current problems are legion
- Classic research vs. tools/capabilities
- Team lost a little momentum mid-year when 3 staff members left the project (one was a key contributor).

Concrete Funding

- **COVID!!!**
- **Submitted 3 SECURE-aligned follow-on LDRDs**
  - Only 1 was funded
  - Call came at inopportune time where we lost 3 team members in short order
  - **A continuing LDRD is developing SECURE tools to make them more usable**
- **Re-aligned a currently funded DoD effort to apply SECURE-like approaches to a network of interest to the sponsor**
  - Slightly limited funding and sponsor's interest in a short term win vs long-term capability
- **New DoD work focused on creating high fidelity networks and results**
  - Small initial funding and advocate for the work retired
- **DHS funded effort out of S&T for NRMC to continue pursuing SECURE work**
  - Smaller funding than anticipated with reduced scope

Legacy - Internal

- Re-directed internal “Emulytics Community” toward the broader “Cyber Experimentation” field, which is a huge win for SECURE
  - Ensures internal funds to support planning, roadmapping, and program development efforts
  - Broadens engagement beyond the typical emulation/simulation staff to other disciplines around the lab
- Inspired research thrusts in corporate computing and cyber-physical mission areas
- Reframed how we talk about cybersecurity within the Lab

Legacy – External

- Workshop
  - Hosted in partnership with UC-Davis
  - Scheduled for November 2021
- Sandia Cybersecurity Institute on Rigorous Experimentation (SCIRE)
  - External-facing Website
    - communicate key work performed on SECURE
    - Acknowledge partners
    - Share links to current research being published by key partners in academia and national lab
    - Host externally viable versions of the Handbook
- Under Emulytics Community, continue hosting collaborative discussions with academic partners: USC-ISI, Purdue CERIAS, UIUC