Detecting Anomalous Computation with RNNs on GPU-Accelerated HPC Machines

Pengfei Zou, **Rong Ge**

Clemson University

Ang Li, Kevin Barker

Pacific Northwest National Laboratory

Overview

- **The new threat in HPC**
  - Illicit workloads exploit powerful GPUs committed to HPC workloads

- **Our approach**
  - Leverage identifiable patterns of HPC workloads
  - Treat illicit workload detection as a classification problem
  - Devise RNN models to infer workloads from high-level profiles

- **Contribution**
  - An online illicit workload detection suitable for practical use
    - > 95% accuracy, with system level light weight profiling only
  - Techniques to handle data heterogeneity, irregularity and loss
  - Advanced RNN modeling for inference accuracy

Illicit Applications on HPC Systems

- **Illicit computations begin running on HPC systems**
  - Crypto mining
  - Password cracking
  - Denial-of-service (DoS) attacks
- **Common characteristics**
  - For-profit or malicious attacks instead of science
  - Resource intensive
    - Powerful GPU accelerators are ideal
  - Long execution time: days to weeks or longer
- **Risks and security issues to HPC**
  - Mission-critical applications deprived of computing cycles
  - data leaking, system damage, etc
  - Empowered hacks and attacks

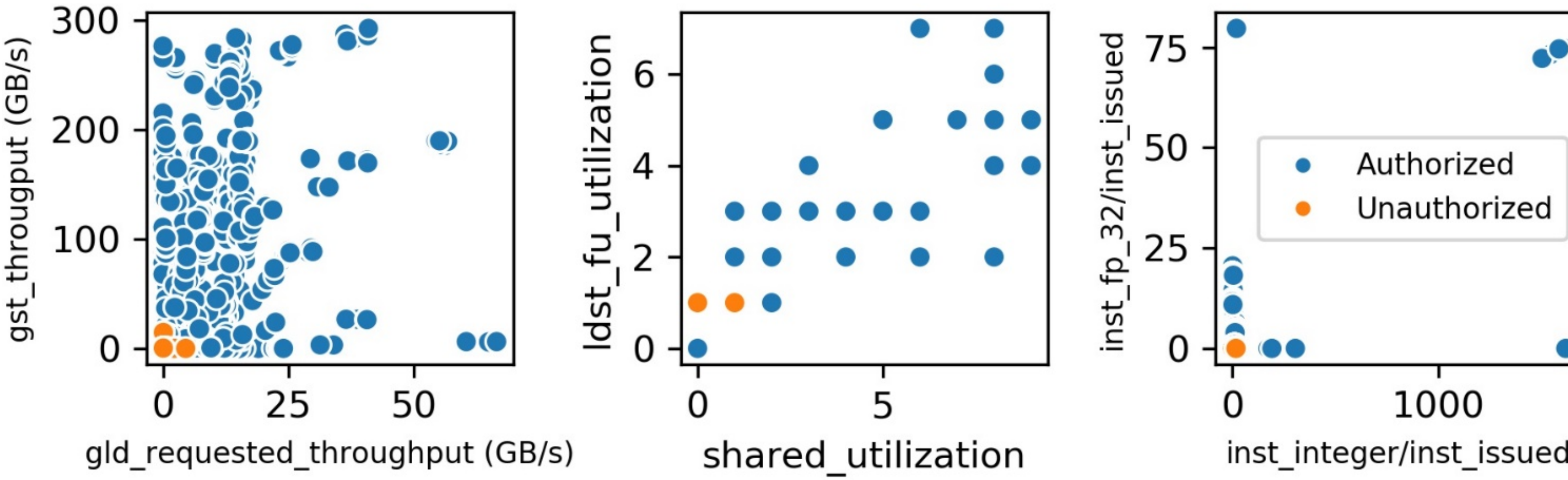
A Unique, New Thread

- **Penetrating login nodes imposes the risks**
  - HPC systems only protect login nodes
- **Authorized users can run illicit computations**
  - Authorization and authentication easily passed
- **Little barriers and guards exist**
  - Due to performance priority in HPC systems
  - Little or no network traffic monitoring and host auditing
- **Computations masked and offloaded to accelerators**
  - CPU-side monitoring and detection measures would fail

Novel security measures needed to detect illicit computation in HPC

Opportunities and Challenges

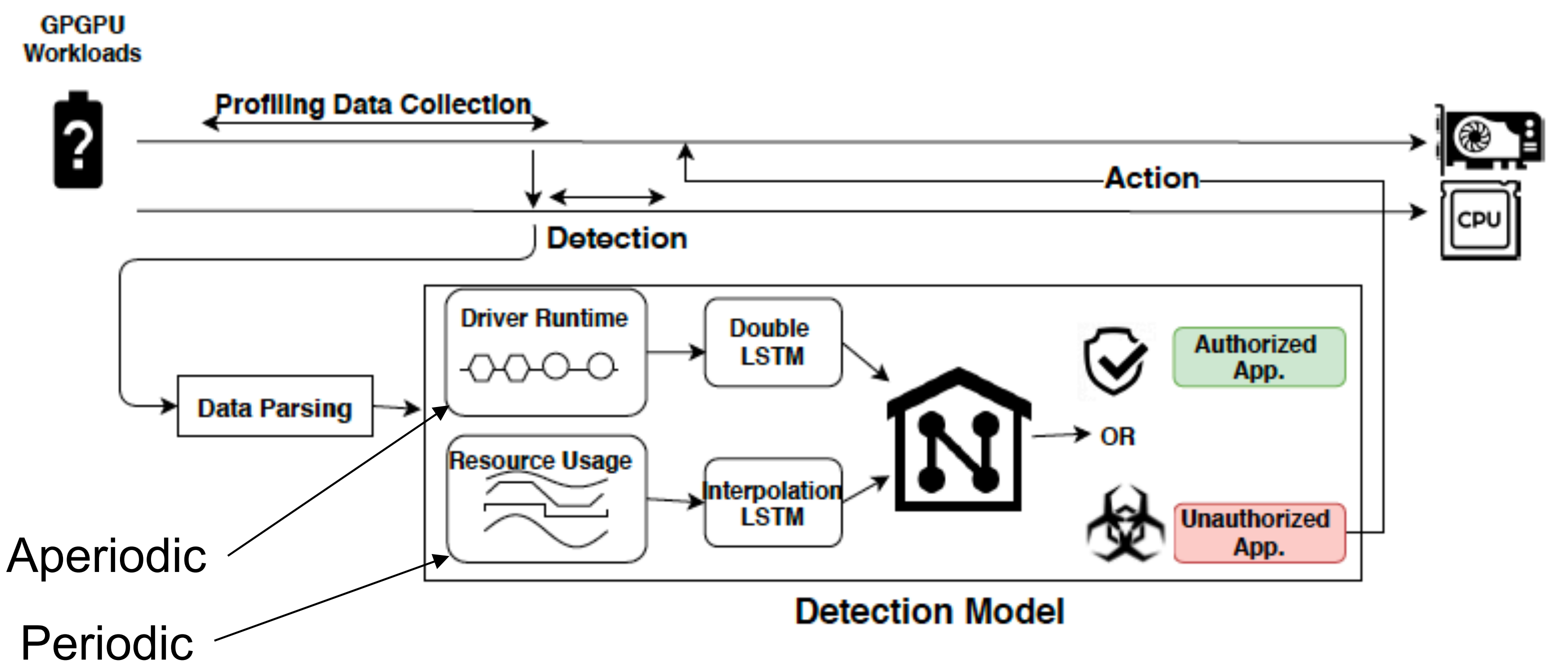
- **HPC workloads have unique patterns identifiable by ML**
  - A small set of programs with specific resource usage patterns
  - Certain kernels and functions, e.g., FFT, BLAS



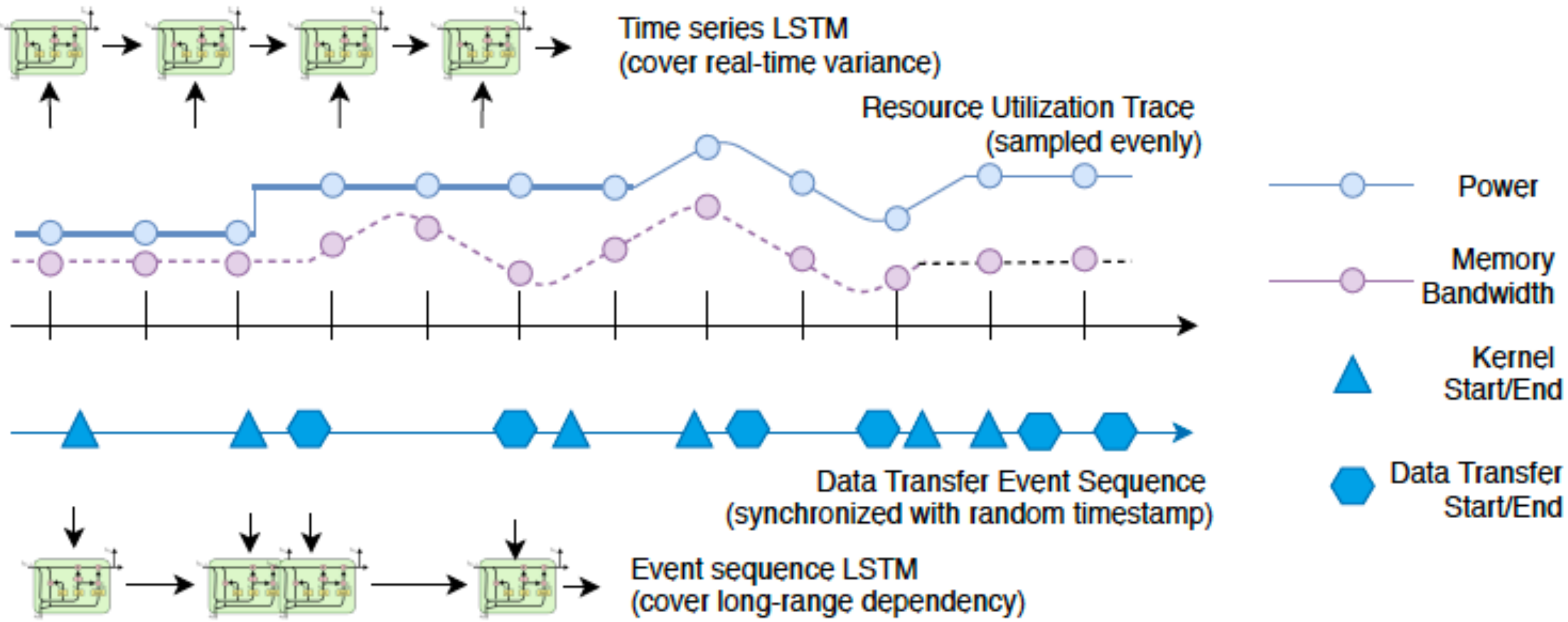
- **Accurate ML models use many HW counters as input**
  - Large overhead for online detection
  - Intrusive to user applications

Our Approach

- **Online illicit workload detection**
  - Illicit GPU computation detection as classification problems
  - Light-weight, common system level profiling for model input
  - Multiple input sequences for inference accuracy
  - Synergistic multi-RNNs to handle complex, heterogeneous inputs

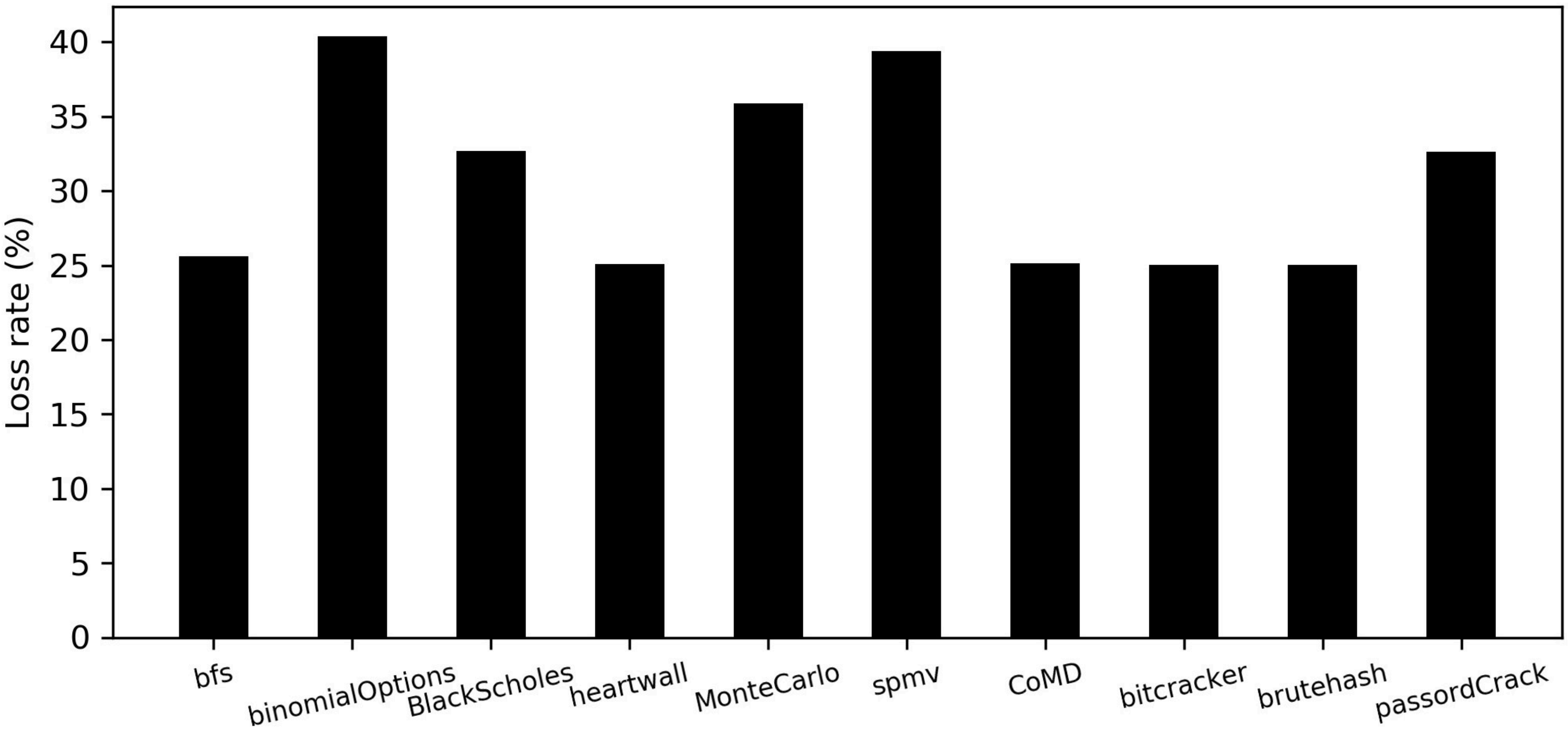


Data Heterogeneity



- **Heterogeneity in data sequences**
  - Varying sample losses in resource utilization sequences
  - Asynchronism between the types
- **Irregularity of event-based data sequence**

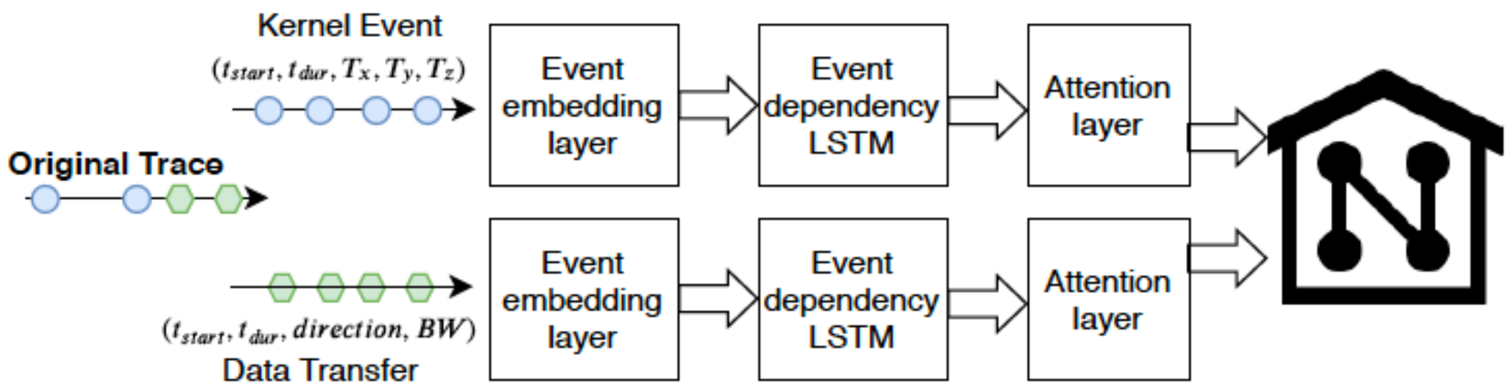
Sample Losses in Utilization Data



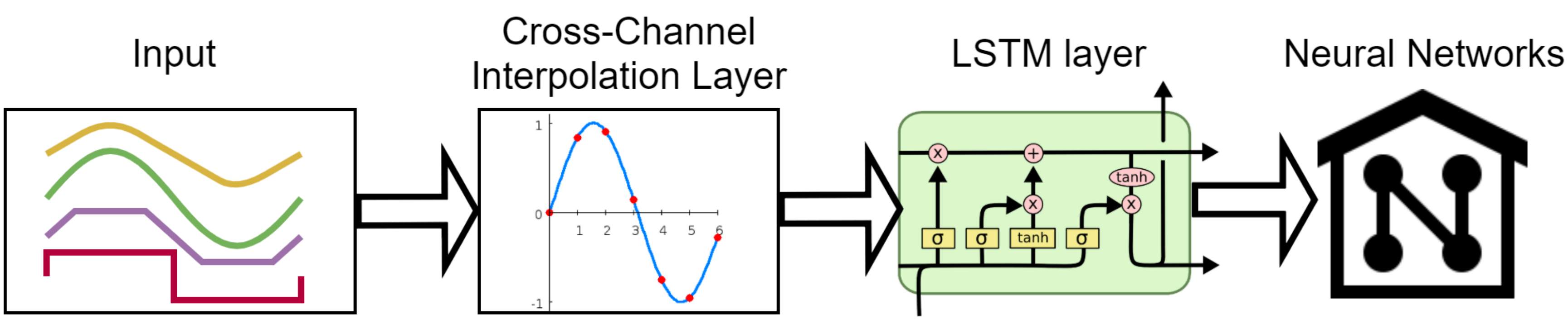
- **Nvidia-smi profiling loses samples**
  - E.g., 30% on average
- **Losses depend on application and sampling interval**
  - Different temporal information from different training apps

LSTM Layers for Advanced Training

- Split Layers for the event-based driver runtime



- Interpolation layer for the resource utilization sequences



Model Training and Validation

- **Workloads**
  - 83 authorized applications
    - Rodinia, Parboil, SHOC, PolyBench, exascale Proxy Apps, etc
  - 17 unauthorized applications from GitHub and BitBucket
    - Crypto mining, password cracking, brute force attacking…
- **Data collection**
  - Periodic resource utilization
    - Power, core utilization, memory footprint, memory bandwidth
  - Event based driver runtime
    - Kernel events: starting time, duration, configuration
    - Data transfer events: starting time, latency, direction, bandwidth
  - HW performance counters for counterpart comparison
- **Three generations of GPUs: K40, P100, and V100**

Selected Evaluation Results

Sequences	K40		P100		V100	
	seen	unseen	seen	unseen	seen	unseen
Events	98.2	78.1	96.7	81.2	97.0	77.8
Resource Util.	99.7	96.7	97.2	95.0	92.1	90.4
Combined	99.2	97.2	98.2	92.4	95.7	90.1

Accuracy

Sequences	K40		P100		V100	
	seen	unseen	seen	unseen	seen	unseen
Events	0.6	62.4	2.4	64.4	0.4	68.7
Resource Util.	1.3	12.6	1.5	8.2	2.8	8.5
Combined	3.1	11.4	1.4	4.1	1.9	7.2

False NR

Metrics / Data	Accuracy		FNR	
	seen	unseen	seen	unseen
Hardware metrics	98.5	91.2	1.3	59.5
Event & utilization sequences	98.2	92.4	1.4	4.1

vs. HMC based

Conclusion

- **A new thread in HPC**
  - Illicit computation takes execution cycles and empowers attacks
- **Our proposed online detection**
  - Lightweight profiling
  - Accurate detection with fused LSTMs using multiple data sequences
- **Our findings**
  - Illicit workloads have different patterns from HPC workloads
  - Multiple system-level profiling is sufficient for accurate detection
  - Fused RNNs are suitable for online detection