Fingerprinting Anomalous Computation with RNN for GPU-Accelerated HPC Machines

Pengfei Zou, Ang Li¶, Kevin Barker¶, Rong Ge**

*Clemson University, ¶Pacific Northwest National Laboratory

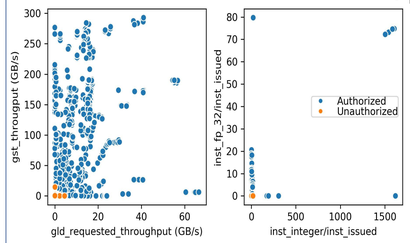
Motivation

- ✓ GPUs has become the main computational contributor for HPC systems
- ✓ It is a serious concern that GPUs are exploited by Illicit workloads (e.g. cryptocurrency, password cracking) for unauthorized computation [1]
- ✓ For HPC security and resource allocation, GPU accelerated HPC systems need fast, lightweight automated illicit workloads detection [2] [3]
- ✓ Our initial profiling suggests that illicit workloads can be discriminated from authorized HPC workloads

Different Profiles of HPC illicit workloads

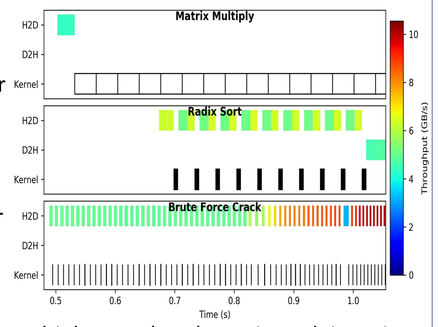
- **Floating point vs integer operations**:
  - Illicit workloads: high hash operations, low floating point operations,
  - HPC workloads: floating point linear algebra and FFT operations.

  Such difference reflects on microarchitectural events, data movement, and resource utilization.



- **Microarchitectural events**:
  - ✓ HPC workloads have different integer and floating point ratio
  - ✓ HPC workloads have higher L2, DRAM, SFU utilization
  - ✓ HPC workloads have higher load and store throughput

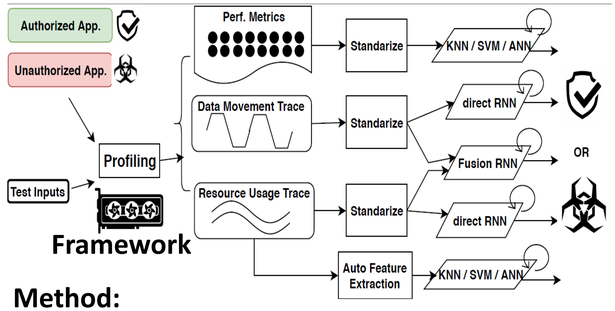
- **Data movement:**
  - ✓ Illicit workloads barely transfer data from device to host
  - ✓ Authorized kernels last longer
  - ✓ Optimized HPC app. overlap computation and data transfer



- **Resource utilization**:
  - ✓ HPC workloads typically have multiple complex phases in each iteration
  - ✓ HPC workloads generally have high resource utilization variation

These different profiles encourage illicit workload detection. However, simple classifiers would fail to accurately categorize the diverse HPC workloads

Machine learning based automatic detection



Method:

- intragroup variance In both authorized/unauthorized workloads is high, simply classifications methods fail
- Recurrent neural networks (RNN) is fit for online time-series data classification
- Three data resource are independent; the acquisition overhead and classification accuracy are different

Microarchitectural **events selection:**

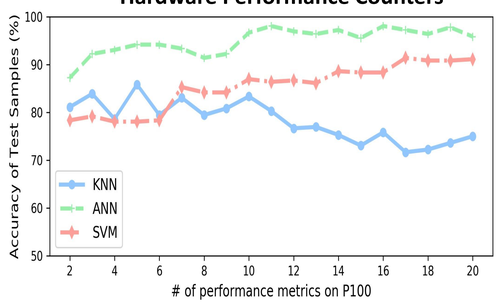
- Use performance counters to record as many events as possible
- Use Support Vector Classification (SVC) to select the 20 most related events for classification.
- Remove events that have high VIF

Data movement & resource utilization trace processing:

- Extract features from trace using *tsfresh* [4]
- Select features with SVC to filter coefficient features for ANN/SVM/KNN models
- Align the trace length by truncating or stretching the traces for RNN model

Results

Hardware Performance Counters



- ✓ DRAM, SFU, L2 Cache PMCs are most related PMCs for classification
- ✓ Within 20 counters, using more hardware counters leads to higher accuracy
- ✓ ANN models have the highest accuracy (98%)
- ✓ Comparing to other sources, PMC based detection is the most accurate but its data collection and storage incurs the highest overhead

- ✓ Fusion RNN models trained with both traces improve accuracy over those with either
- ✓ RNN trained with data movement trace achieves higher accuracy than that with resource utilization trace
- ✓ RNN classification method can be deployed online for illicit workloads detection
- ✓ RNN model achieves higher accuracy than ANN when trained with resource utilization trace

Accuracies of models trained with data movement & resource utilization traces

Data source	ML model	K40	P100	V100
Data movement	RNN	90%	93%	92%
Resource utilization	RNN	89%	90%	88%
	ANN	89%	88%	89%
Both traces	Fusion RNN	93%	97%	93%

Discussion & Future Plan

- RNN models trained with data movement and resource utilization traces provide accurate illicit workloads detection with a low overhead
- RNN models may fail to detect some illicit workloads if the workloads have a high resource utilization, but such failure can be corrected by ML models with PMCs.
- Authorized workloads that have low floating-point operation and resource utilization have a chance to be mis-classified as unauthorized workloads
- In the future, we plan to build an online hierarchical risk model and detection framework to analyze and combine results from multiple sources

Conclusion

- ✓ Authorized and unauthorized workloads show difference on microarchitectural activities, data movement and resource utilization
- ✓ It is possible to design data-driven ML models to identify illicit workloads from HPC workloads
- ✓ PMC-based detection is the most accurate if only one data source is used but consumes resource for data collection
- ✓ Online resource utilization trace could provide initial diagnose with very low overhead

Reference

[1] Yongdong Wu, Zhigang Zhao, Feng Bao, and Robert H Deng. Software puzzle: A counter measure to resource-inflated denial-of-service attacks. IEEE Transactionson Information Forensics and security, 10(1):168–177, 2014

[2] Davide Balzarotti, Roberto Di Pietro, and Antonio Villani. The impact of gpu-assisted malware on memory forensics: A case study. Digital Investigation, 14:S16–S24, 2015

[3] Hoda Naghibijouybari, Ajaya Neupane, Zhiyun Qian, and Nael Abu-Ghazaleh. Rendered insecure: Gpu side channel attacks are practical. In Proceedings of the2018 ACM SIGSAC Conference on Computer and Communications Security, pages2139–2153. ACM, 2018.

[4] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. Timeseries feature extraction on basis of scalable hypothesis tests (tsfresh – a pythonpackage). Neurocomputing, 307:72 – 77, 2018.

[5] John Demme, Matthew Maycock, Jared Schmitz, Adrian Tang, Adam Waksman, Simha Sethumadhavan, and Salvatore Stolfo. On the feasibility of online malware detection with performance counters. In ACM SIGARCH Computer Architecture News, volume 41, pages 559–570. ACM, 2013.

Acknowledgement

This work is supported in part by the U.S. National Science Foundation under Grants CCF-1551511, CNS-1551262 and U.S. DOE Office of Science, Office of Advanced Scientific Computing Research, under award 66150: "CENATE - Center for Advanced Architecture Evaluation

Demo

Evaluated Benchmarks