LSU

College of

Engineering

School of Electrical
Engineering & Computer Science

College of Engineering
School of Electrical Engineering & Computer Science

Accelerating Ransomware Defenses with Computational Storage Drive-Based API Call Sequence Classification

Louisiana State University

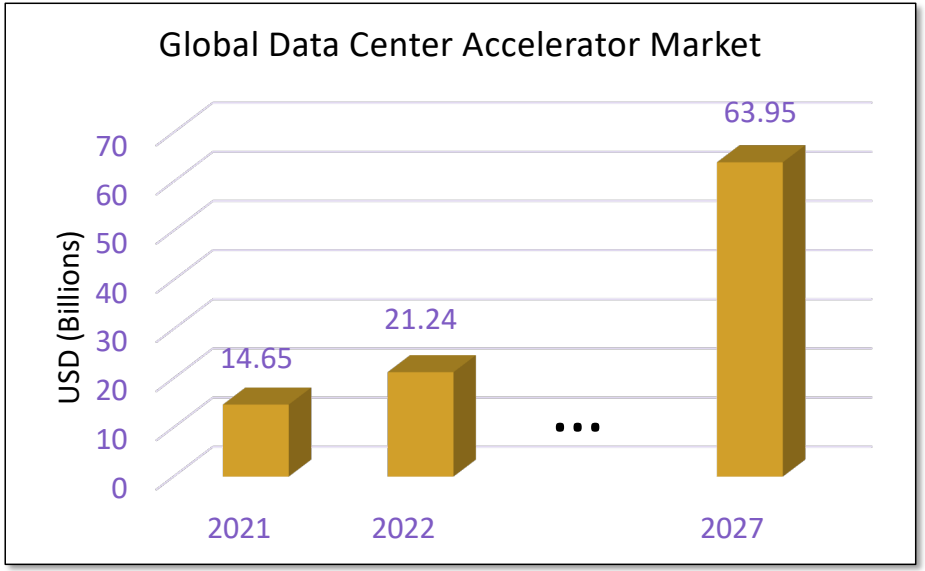
Division of Computer Science and Engineering (CSE)

Kurt Friday

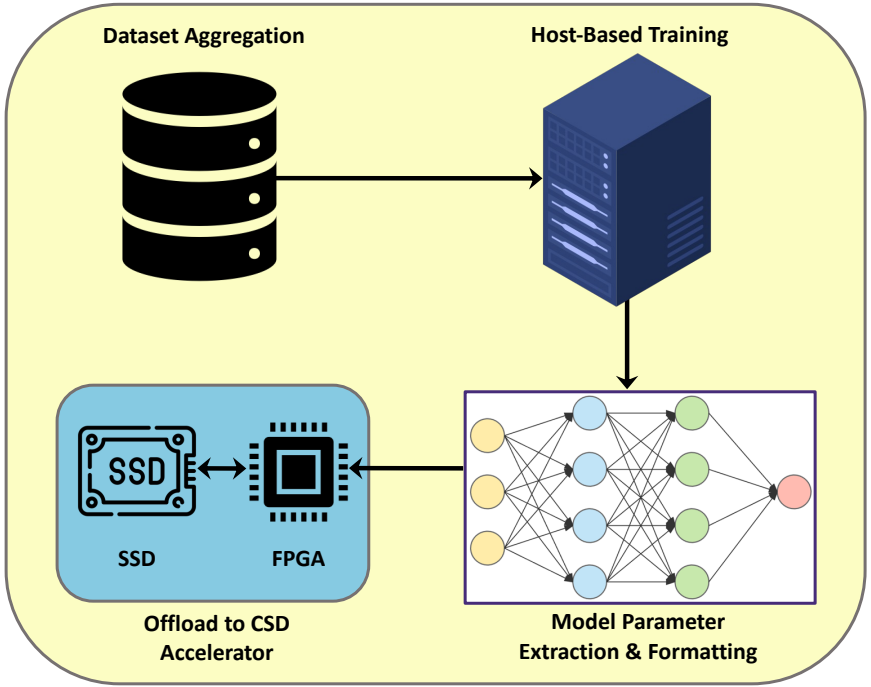
Elias Bou-Harb

Motivation

- Deep learning provides vital mechanisms for managing the exponential growth of data within data centers
- Data consumed in data centers has increased from 1.2 trillion GB to 59 trillion GB in the last decade
- CAGR (Compound Annual Growth Rate) of 24.7% in the global data center accelerator market

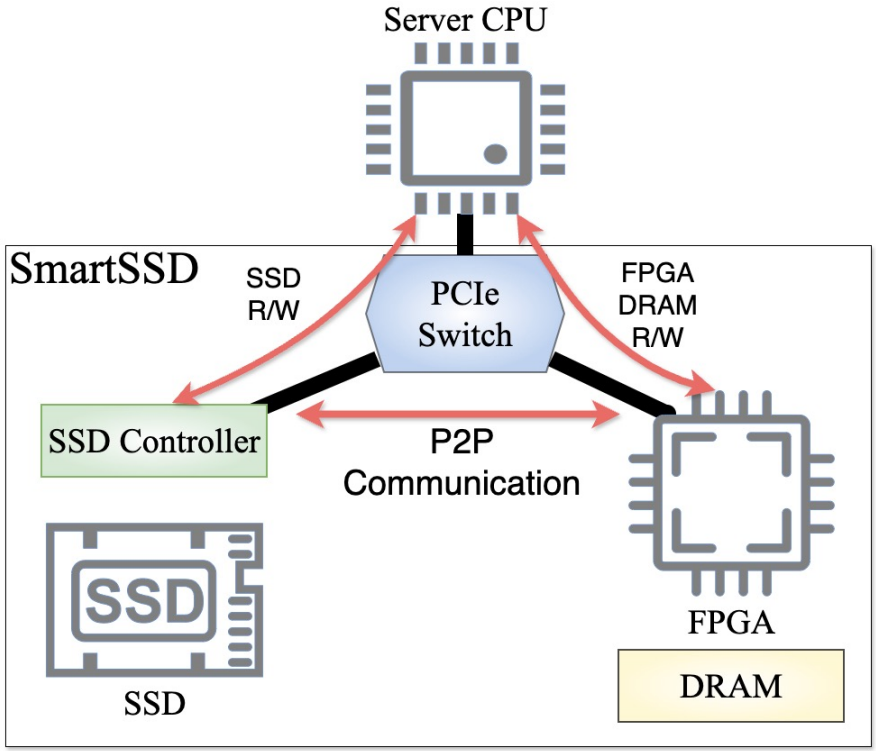


Overview & Contributions



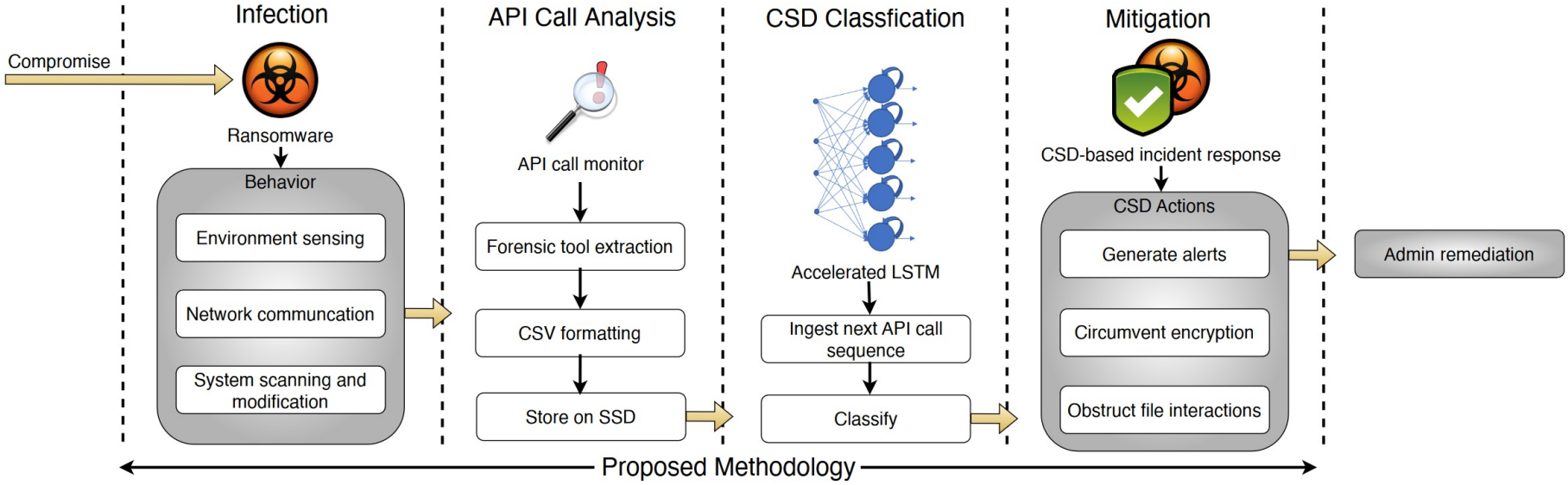
- Unique method for Accelerating deep learning classification via Computational Storage Drives (CSDs)
- Employ several enhancements to improve the model's inference speed, realizing an increase of 344.6x over NVIDIA's A100 GPU
- Showcase its capability of promptly and reliably detecting ransomware

CSD Primer

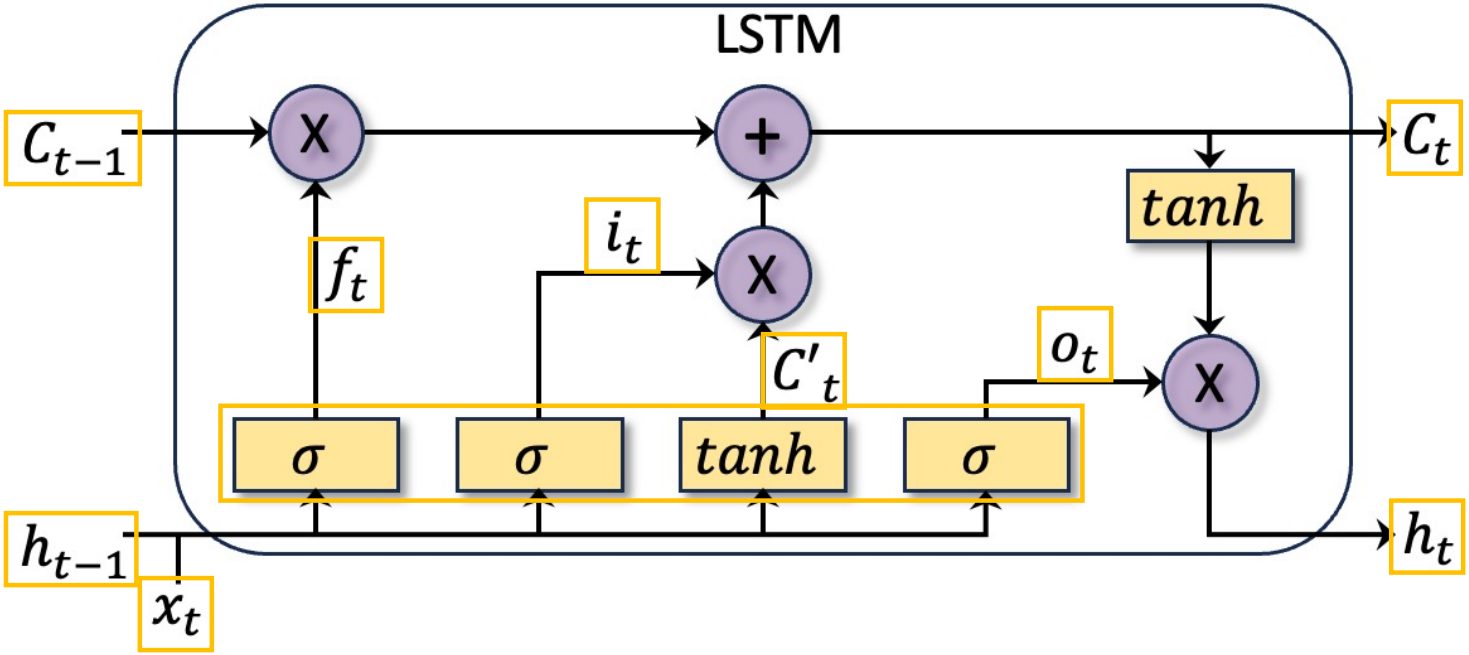


- Samsung 4 TB SSD with a Xilinx Kintex UltraScale FPGA
- CPU can issue FPGA computation and DRAM read/write requests
- Supports Peer-to-Peer (P2P) data movement over the internal data path between its NVMe SSD and FPGA
- P2P near-data computation can reduce or eliminate Host-to-SSD and Host-to-FPGA PCIe traffic, as well as related roundtrip latencies and performance degradations
- Enable appreciable energy consumption reduction

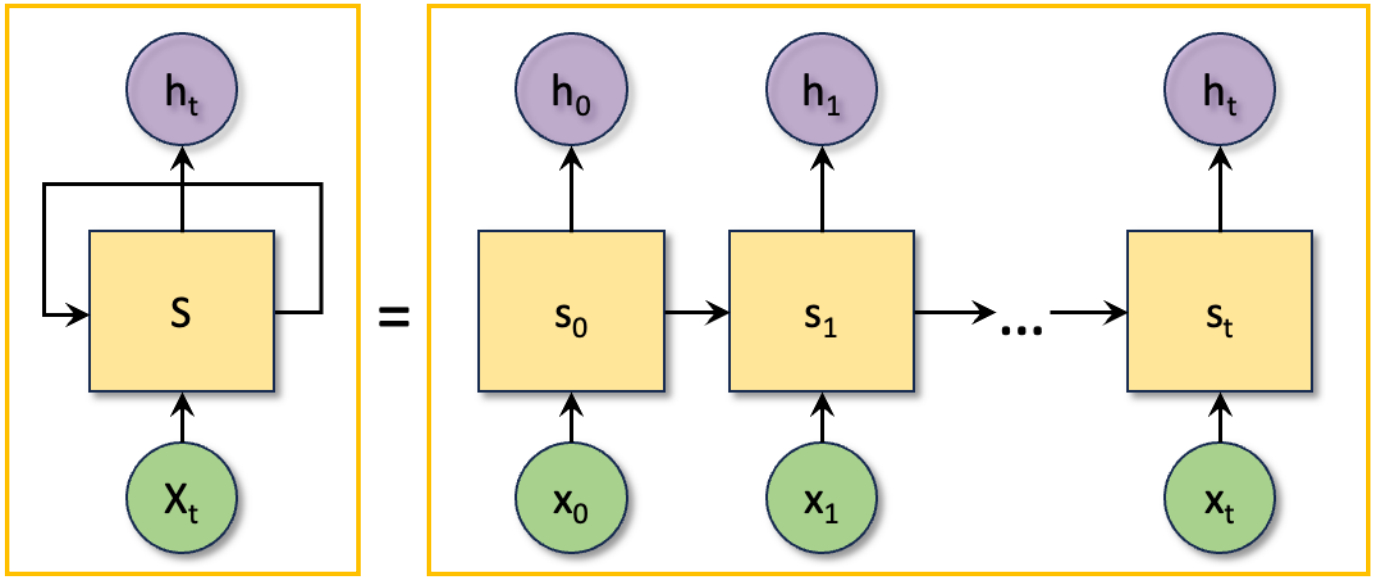
Proposed Methodology



Model Selection

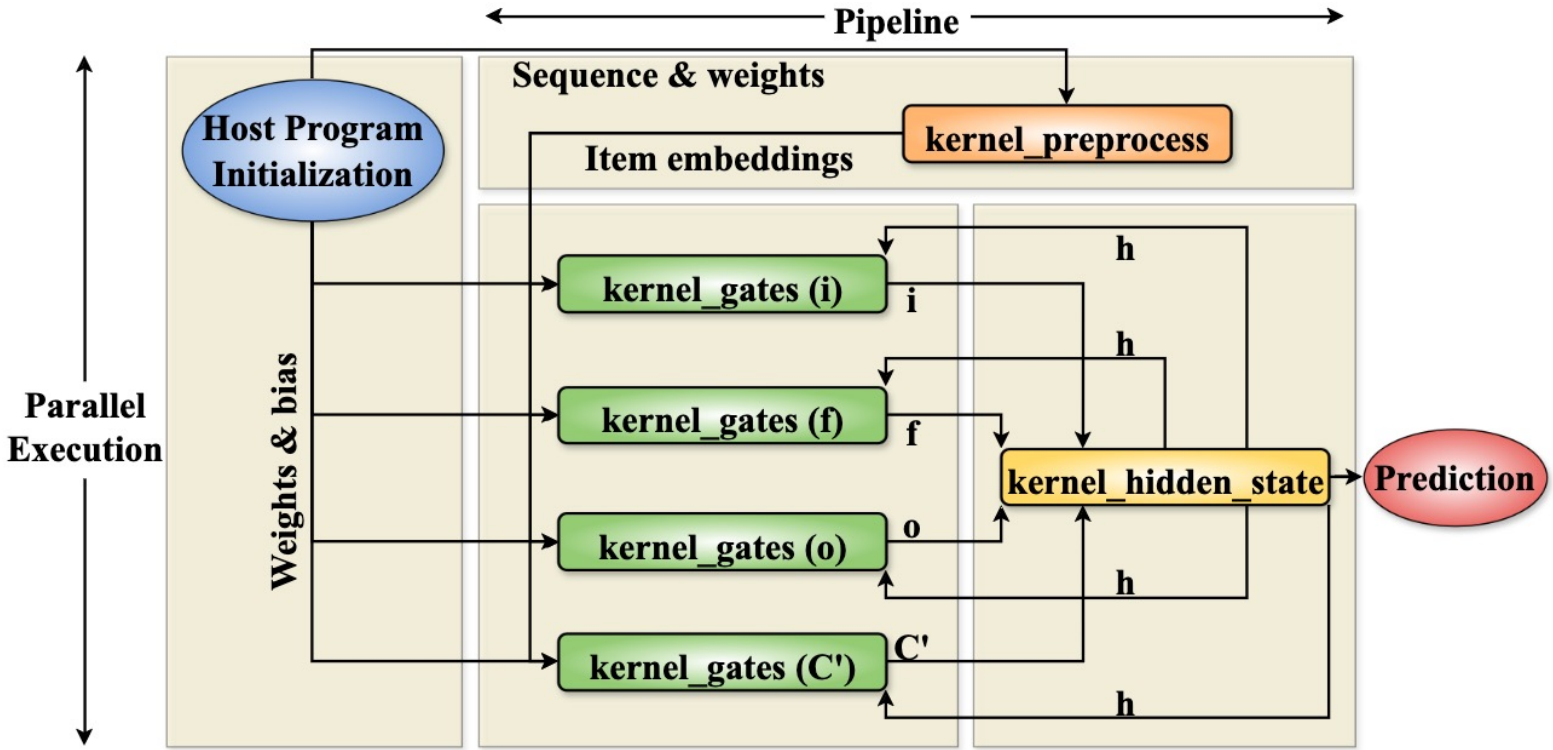


Advantages of LSTMs



- $i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$
- $f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$
- $o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$
- $C'_t = tanh(W_{C'}[h_{t-1}, x_t] + b_{C'})$

Kernel Implementation



Optimizations

Activation functions

- $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
- $softsign(x) = \frac{x}{abs(x) + 1}$

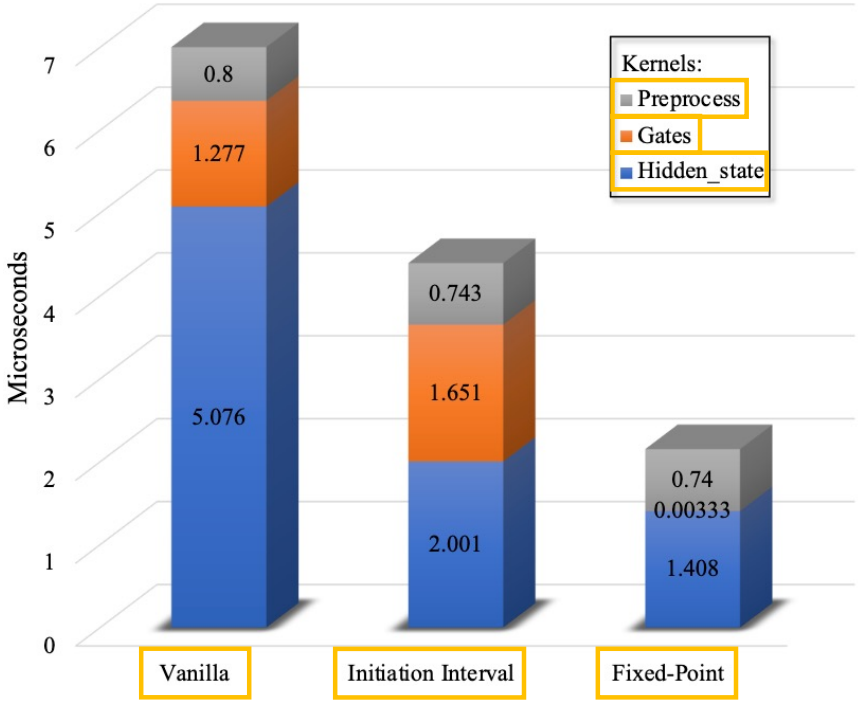
Pragmas

- HLS DATAFLOW
- HLS UNROLL
- HLS ARRAY PARTITION
- HLS PIPELINE II=1

Fixed-point arithmetic

- 10^6 scaling
  - e.g, 0.6533132 * 10^6 = 653313
- 10^{12} correction upon single multiplication
- 10^6 correction between chained multiplications

FPGA Inference Time

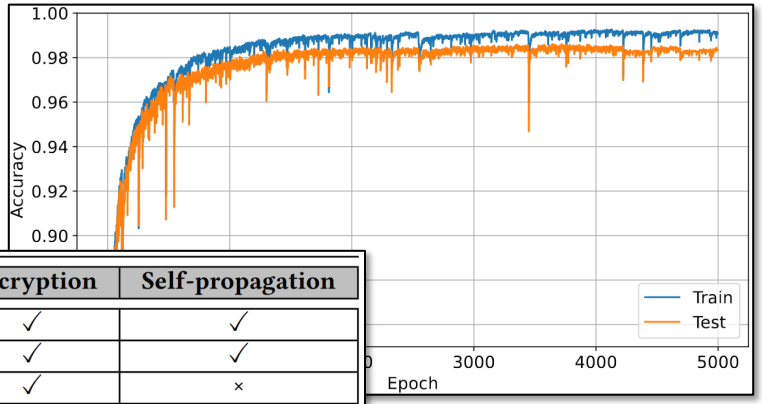


- Optimizations reduced forward pass time of first iteration from roughly 7.153 µs to 2.15133 µs
- Optimizations compared against an Intel Xeon CPU with 13 GB of RAM and an NVIDIA A100 GPU with 40 GB of video RAM
- Mean forward pass time of proposed approach surpassed the GPU by over 344×

	Execution time	95% CI
FPGA	2.15133 μs	N/A
CPU	991.57750 μs	217.46576 μs - 1765.68923 μs
GPU	741.35336 μs	394.45317 μs - 1088.25355 μs

Ransomware Detection

- Application Programming Interface (API) call sequences are ingested by the LSTM
  - Sequence lengths of 100 were utilized
- LSTM comprised of 7,472 parameters was employed
- Training for 3K epochs gave an accuracy, precision, recall, and F1 of 0.9833, 0.9789, 0.9890, and 0.9840, respectively



Family	Instances	Encryption	Self-propagation
Ryuk	5 variants	✓	✓
Lockbit	6 variants	✓	✓
Teslacrypt	10 variants	✓	×
Virlock	11 variants	✓	✓
Cryptowall	8 variants	✓	×
Cerber	9 variants	✓	×
Wannacry	7 variants	✓	✓
Locky	6 variants	✓	×
Chimera	9 variants	✓	×
Razy	12 variants	✓	×
TorrentLocker	3 variants	✓	×
Bitman	6 variants	✓	×
BadRabbit	5 variants	✓	✓

Questions