Hardware-based Malware Detection using Low-level Architectural Features

Meltem Ozsoy*, *Member, IEEE,* Khaled N. Khasawneh†, *Student Member, IEEE,* Caleb Donovick‡, *Student Member, IEEE,* Iakov Gorelik‡, *Student Member, IEEE,* Nael Abu-Ghazaleh†, *Senior Member, IEEE* Dmitry Ponomarev‡, *Senior Member, IEEE*

Abstract—Security exploits and ensuant malware pose an increasing challenge to computing systems as the variety and complexity of attacks continue to increase. In response, software-based malware detection tools have grown in complexity, thus making it computationally difficult to use them to protect systems in real-time. Therefore, software detectors are applied selectively and at a low frequency, creating opportunities for malware to remain undetected. In this paper, we propose Malware-Aware Processors (MAP) - processors augmented with a hardware-based online malware detector to serve as the first line of defense to differentiate malware from legitimate programs. The output of this detector helps the system prioritize how to apply more expensive software-based solutions. The always-on nature of MAP detector helps protect against intermittently operating malware. We explore the use of different features for classification and study both logistic regression and neural networks. We show that the detectors can achieve excellent performance, with little hardware overhead. We integrate the MAP implementation with an open-source x86-compatible core, synthesizing the resulting design to run on an FPGA.

Index Terms—malware detection, architecture, security, low-level features

1 Introduction

Computing systems are under continuous attacks by increasingly motivated and sophisticated adversaries. These attackers use vulnerabilities to compromise systems and deploy malware. Malware is a general term for malicious software, which can be defined as any software system that can damage non-damaging software systems [38] . A recent IT industry survey shows that the number of security incidents in 2014 rose by 48% from 2013, to an astounding 42.8 million incidents [49]. An estimated 11% of these incidents cost $10 million or more each. The US Director of National Intelligence has ranked cybercrime as the top national security threat, higher than that of terrorism, espionage, and weapons of mass destruction [15].

Although significant effort continues to be directed at making systems more difficult to attack, the number of exploitable vulnerabilities is overwhelming. Attackers obtain privileged access to systems in a variety of ways, such as drive-by-downloads with websites exploiting browser vulnerabilities [8], network-accessible vulnerabilities [56] or even social engineering attacks [5]. Many vulnerabilities are not publicly known [65], and the slow update cycles make vulnerabilities exploitable long after they are discovered. Attackers only need to succeed in exploiting a single vulnerability to completely compromise a system. Thus, it is essential to invest in approaches to detect malware so that infections can be stopped and damage contained.

Increasing sophistication of malware makes its detection more difficult. A significant challenge faced by malware detection is related to constrained resources — the resource requirements needed for detection make it prohibitive to monitor every application all the time. Typical techniques proposed for online malware detection include VM introspection [25], dynamic binary instrumentation [18], information flow tracking [64], [47], and software anomaly detection [27]. These solutions each have coverage limitations and introduce substantial overhead (e.g., 10x slowdown for information flow tracking is typical in software [63]). The problem is especially critical for mobile environments where memory limitations and the energy cost of detection impose substantial limits on the resources that a system can dedicate to online malware detection. For these reasons, dynamic analysis techniques are typically conducted only on the cloud (for example, Google's Bouncer [44]), using automated inputs and for a limited time. On the user side, these difficulties limit malware detection to static signature-based scanning tools [20] which have known limitations [42] that allow attackers to bypass them and remain undetected.

In this paper, we motivate and present MAP (Malware-Aware Processor) — a hardware-based malware detector that uses low-level features to classify malware from normal programs as they execute. Because it is implemented using low-complexity hardware, malware monitoring can be always on with negligible overhead. We use the term low-level to mean architectural information about an executing program that does not

*Security and Privacy Lab., Intel Corp., Hillsboro, OR Email:{meltem.ozsoy}@intel.com

†CSE and ECE Departments, University of California, Riverside, CA 92521, Email:{kkhas001,naelag}@ucr.edu

‡CS Department, Binghamton University, Binghamton, NY 13902–6000. Email:{cdonovi1,igoreli1,dima}@cs.binghamton.edu

require modeling or detecting program semantics. Low-level information includes architectural events such as cache miss rates, branch prediction outcomes, dynamic instruction mixes, and data reference patterns.

Successful offline classification based on low-level features has recently been shown by Demme et al. [17]. MAP advances the state of the art relative to this work in the following ways:

- **Real-time malware detection:** real-time detection includes a new time-series component where successive decisions from the classifier are evaluated to detect anomalous behavior. We explore simple Exponentially Weighted Moving Average (EWMA) approach for detecting malware. In contrast, the offline problem uses after-the-fact analysis with the benefit of the complete data for the process lifetime. Thus, the online detection results demonstrate (for the first time) that classification over windows of execution can also separate malware from normal programs.
- **Hardware implementation using simpler classifiers:** a hardware implementation has significant benefits over software detection for this problem. First, direct access to hardware features is possible at low cost. Hardware detection can be always on, for all programs, with low complexity and power overhead. In contrast, software implementations require additional resources, are limited by the available performance counters, and incur significant costs. On the other hand, hardware implementations necessitate simpler classifiers than those available in software. This paper demonstrates that such simple classifiers can be effectively used to detect malware.
- **Exploration of complexity/detection tradeoffs:** we investigate both linear classifiers as well as neural network based classifiers. We explore the tradeoff between complexity and classification effectiveness. We also study a number of optimizations to the hardware implementation of both the base classifier and the time-series detector.
- **Two level detection framework:** False positives are likely to occur due to simple classification algorithms and the low-level features used. Thus, hardware detection is not sufficient on its own. We propose a two-level detection framework with MAP being the first line of defense. The goal of MAP is to prioritize running processes such that a heavy-weight software solution can be guided to protect or scan more suspicious processes first, reducing the effort and time to detection as compared to using the second level for all processes. To avoid building complex and stateful low-level models in hardware, the first-level hardware detector is based on the low-level features that are easily collectable in hardware. In contrast, the slow second-level software detector can be an IDS that is using full semantic information.

A major advantage of MAP is that it can react to a malware quickly, acting as a low-level alert system for further software protection. The hardware detector of MAP is always on, without affecting the available resources and with minimal energy consumption. At the same time, it can be built to use architectural events that are expensive and difficult to obtain at the software level (e.g., through performance counters). MAP is envisioned to operate synergistically with existing virus scanning utilities and other one-time analysis tools; it continues to monitor the system to detect any malware that evades such tools.

We developed a fully functional hardware description of MAP hardware detector using Verilog, and integrated it within an open source x86-compatible core implementation. Our evaluations show that MAP data collection delay fits within a single cycle of the processor. Moreover, for features related to instructions, the logic is located at the commit stage of the processor pipeline, therefore avoiding any negative impact on the cycle time, instruction throughput and execution time of the program. At a time where CPU manufacturers are showing increasing willingness to invest in hardware support for security [33], [58], [61], [55], [24], MAP offers an attractive mixture of significant impact on security and low complexity.

We did not consider how the detector should evolve to the changing nature of malware: a practical deployment will require a secure channel to update the detector configuration. Our contribution is to study the use of online hardware detection of existing malware. In particular, we did not explore how attackers will react to the presence of such a detector to attempt to hide the behavior of malware. Adversarial classification is a branch of machine learning that can assist with the evolution of attackers over time as commonly occurs in a security context [16]. Techniques from this space (such as feature randomization [60]) can be integrated into our design to make it more resilient to attacker evolution.

The remainder of the paper is organized as follows. Section 2 and Section 3 overview the malware detection approaches and examine a number of candidate low-level features. Section 4 presents the proposed online detectors. Section 5 presents the implementations of the proposed detectors, and evaluates their timing and complexity. Section 6 presents an evaluation of the real-time detection system based on MAP. In Section 8 we present the related work. Finally, Section 9 offers our concluding remarks.

2 Background and Preliminaries: Low-Level Malware Detection

Malware detectors typically use high-level information such as behavior models of programs based on system calls, accessed/created files and thread creation events [20] to capture common features of malware. In contrast, MAP uses low-level information that can be collected during the execution of programs such as architectural events, instructions and memory addresses, and the mix of executed instruction types.We refer to these features as *low-level* features.

In this section, we show that low-level information collected and processed in hardware can effectively distinguish malware from normal programs using simple classifiers. The classification in this section is done after-the-fact, similar to prior work [17], but differs in that the classifiers are simpler and more suitable for hardware implementation. Moreover, the section introduces the set of features that we use as representatives of the different available classes of low-level information.

We study two different classification algorithms for MAP: (1) Logistic Regression (LR), which is a simple linear classification algorithm. LR attempts to linearly separate malware from normal programs in the feature space. In general, the programs are not linearly separable so LR provides a probability between 0 to 1 for the likelihood of a program being malware. To convert this likelihood to a binary decision, we pick a threshold above which programs are considered malware; and (2) Neural Network (NN) which consist of a network of perceptrons that when trained, approximates a classification function that most likely could have generated the training data. LR is equivalent to a single perceptron in an NN [7]; thus, we expect NNs to perform better than LR but also to have higher implementation complexity.

For this experiment, the classifiers are trained based on the chosen low-level features collected using the PIN toolset [14]. In a hardware implementation these features would be collected directly from the hardware; for example, opcode frequencies can be collected directly at the commit stage of the processor pipeline.

2.1 Data Set & Data Collection

We used the University of Mannheim malware dataset for this study [3]. We downloaded the corresponding samples of 1,087 malware programs from the Offensive Computing website [45]. Using the VirusTotal [59] malware classification interface, we identified different types and families of these programs. We followed Microsoft's classification [4] which identified 9 malware families in total which are shown in Table 1. For normal program samples, we used a variety of programs including system programs, browsers, text editing applications and the SPEC2006 benchmarks. Overall, we analyzed 467 regular programs in our evaluations.

In order to collect the data, we used a virtual machine running a 32-bit Windows 7 operating system. We disabled the firewall and Windows Security Services on this machine and connected it to the network to support malware operations.

The collected data was divided into training, testing, and validation sets as shown in Table 1. We used a balanced training set (roughly equal number of malware and normal programs).

TABLE 1: Malware Dataset

Family	Train	Test-1	Val	Test-2	Total
Vundo	14	2	5	21	42
Emerleox	10	5	4	33	52
Virut	8	3	7	46	64
Sality	12	2	4	46	64
Ejik	7	6	4	101	118
Looper	10	3	6	145	164
AdRotator	14	1	2	119	136
PornDialer	11	6	4	196	217
Boaxxe	13	6	0	211	230

3 Feature Selection

There is a large number of different candidate low-level features available at the microarchitecture level. We explore this space by evaluating three types of features: (1) features based on executed instructions; (2) features based on memory address patterns; (3) features based on architectural events. We selected candidates from each category driven by both ease of collection through binary instrumentation as well as estimated implementation complexity. We introduce these selected features in the remainder of this section. We also evaluate their off-line detection performance using our candidate classifiers to allow comparison to prior work [17] which used more complex classifiers and in some cases different features.

3.1 Features Related to Architectural Events

One group of features is based on microarchitectural events which are not directly visible to the program. Demme et al. [17] used performance counters on the ARM chip to capture architectural features including the number of memory reads, memory writes, software updates to the program counter, unaligned memory accesses, immediate branches and taken branches. We explore these same features for the x86 instruction set with the exception of software updates to the PC which are not possible on x86. We call these features ARCH.

The value of the architectural features is collected once every 10,000 committed instructions following the value used by Demme et al. [17]; we later study the imact of the instruction window size on detection performance. At the end of each period, the detection algorithm classifies whether this execution period is representative of malware or of a normal program based on the collected feature data. These architectural features attempt to capture the similarity of the architectural events between malware.

TABLE 2: Features based on Architectural Events

Feature	Description
ARCH	Frequency of memory read/writes, taken & immediate branches and unaligned memory accesses

3.2 Features Related to Memory Addresses

The typical operations of malware include accessing files and updating/reading windows registry entries. This type of behavior results in similar access patterns to memory addresses during program execution. In order to capture this behavior, we examined the use of memory addresses as a detection feature. Specifically, we calculated the distance between the memory address of the current load/store instruction and the memory address of the first load/store operation in the group of 10K instructions. We used two different approaches for memory address features: (i) We created a histogram of read distances and write distances separately quantized into bins. At every period, we store the frequency of each bin to create the feature vector (MEM1 in Table 3); and (ii) this feature is similar to MEM1, but in this case we only use a binary existence vector for the read/write histogram features. The feature bits are set to one if a distance that falls into that bin is encountered during the execution (MEM2).

TABLE 3: Features based on Memory Addresses

Feature	Description
MEM1	Frequency of memory address distance histogram
MEM2	Memory address distance histogram mix

3.3 Features Related to Instructions

The distribution of executed instructions are another promising feature for classification. Instruction opcode is one of the features previously used for offline malware detection [51], [52], [9], [63].

We constructed opcode features in two ways. First, we created a list of most frequently used opcodes from malware and regular programs, we combined the top 35 opcodes that showed the largest difference in frequency between malware and regular programs (INS2 in Table 4).We also used the same opcode features in the form of a binary vector, where each element indicates if an instruction with that opcode has been executed (INS4).

TABLE 4: Features based on Instructions

Feature	Description
INS1	Frequency of instruction categories
INS2	Frequency of opcodes with largest difference
INS3	Existence of categories
INS4	Existence of opcodes

The instruction category features are based on Intel XED2[12] instruction category classes. Instead of tracking individual opcodes, we track frequencies of the instruction categories. There are 58 different instruction categories and the feature vector has one entry for each category. For example, all arithmetic instructions are in the $BINARY$ category, all bit manipulation instructions are in $LOGICAL$ category and data movement instructions are in $DATAXFER$ category. We use frequency of categories (INS1) and existence of categories (INS3) as separate feature vectors. Using categories as features generalizes the instruction types such that many similar instructions are counted only with one feature. In contrast, INS2 tracks frequency of opcodes that are commonly encountered either in malware or regular programs, while INS4 tracks the existence of these opcodes in the period.

3.4 Features Related to Branches

Another low-level indicator of activity of the program is its control flow behavior, which we capture through the branch instruction characteristics such as frequencies of branch opcodes and distribution of branch target distances. We selected 33 branch opcodes for testing branch frequencies, and we created 20 different distance groups for branches: 10 groups have positive distances and the other 10 groups have negative distances. We have two versions of each feature, one based on frequency and one on existence resulting in the four features in Table 5.

TABLE 5: Features based on Branches

Feature	Description
BRNCH1	Existence of direction categories
BRNCH2	Existence of branch categories
BRNCH3	Frequency of direction categories
BRNCH4	Frequency of branch categories

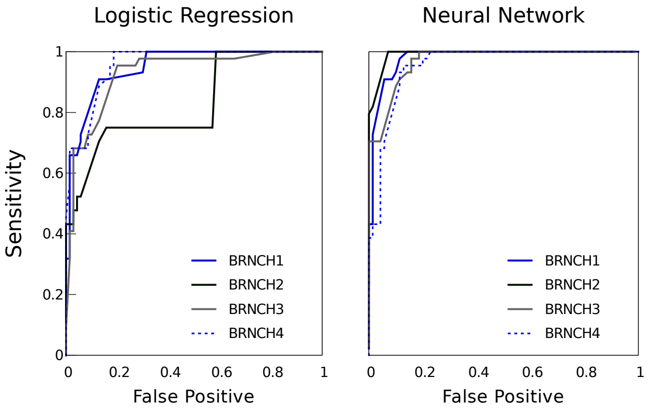


Fig. 1: Detection Performance of Branch-related Features

3.5 Offline detection evaluation

Evaluation of classification performance is based on the sensitivity and specificity of the model. *Sensitivity (S)* is the fraction of malware that are classified correctly and *Specificity (C)* is the fraction of normal programs classified correctly (1-C is the fraction of false positives). To evaluate classification performance and to select the best performing thresholds and features, Receiver Operating Characteristics (ROC) graphs[6] are used. We present the ROC graph for each feature in Figure 2.

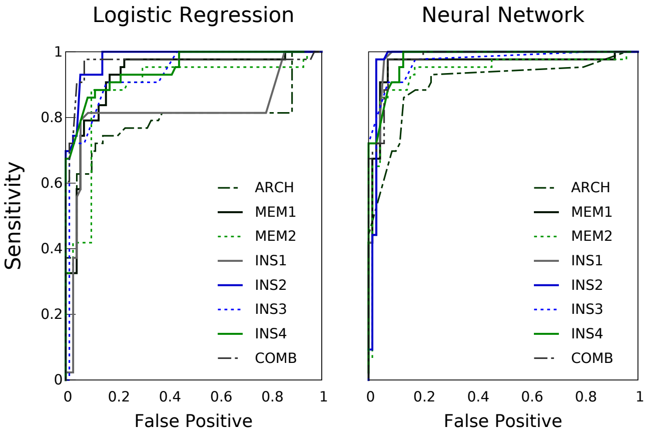


Fig. 2: Detection Performance of all Features

In order to evaluate the features, we use *after-the-fact* detection performance: simply, if the majority of classifier decisions show malicious behavior then the program is labeled as malware, otherwise it is labelled as regular. The threshold for each feature selected at the point where (S+C) sum is maximized.

Figure 2 shows the Receiver-Operating Characteristics (ROC) graph for the two classifiers across the different features we studied. In an ROC graph, S is plotted as a function of *FP rate*. FP rate is calculated by dividing the number of false positives by the number of actual negative instances ($FPrate = FP/(FP + TN)$, where TN is the number of True Negatives). The upper left corner of an ROC graph (0,1) provides the best classification performance with no false positives and 100% Sensitivity. We discuss the performance of the different features in more detail below.

Architectural Features. ARCH feature can correctly identify 70% of malware with only 10% false positives with the basic LR model. For the more complex NN model, the classification rate increases to 88%; however, the false positives also increase to 20%. Architectural features have already been shown to be effective for Android malware[17] using complex machine learning classifiers; they are also somewhat effective for detecting malware on x86 using simpler classifiers. Because of their modest classification performance, we did not pursue these features further.

Memory Address Features. Detection performance of both MEM1 and MEM2 features significantly outperforms the ARCH feature. Both the NN and the LR models can detect 90% of malware with the NN model having only 4% false positives for MEM1. The frequency based feature (MEM1) not only classifies better than the histogram mix feature (MEM2), but also achieves the best false positive rate among all features using the NN. However, the mix features (MEM2) are easier to collect and are simpler to classify (they do not require multiplication), allowing low complexity hardware implementations.

Instruction Mix Features. Instruction traces provide significant information about program execution. These features provide the highest accuracy among the set we considered: Figure 2 shows that most of the instruction based features achieve nearly 100% sensitivity with around 10% false positive rate using the NN model. The LR model is less effective than NN model for all features. Our hardware implementation is based on the INS2 feature which can detect all malware in our test set with only 9% and 16% false positive rates for NN and LR respectively.

Branches Features. We performed simulation of those features on both linear regression model and neural network model, using selected interval of 10K instructions. The ROC graph for branch features is presented in Figure 1. The figure shows that BRNCH4 performance using linear regression outperform the performance of the rest of features vectors with 100% sensitivity and 18.6% false positives. On the other hand, the BRNCH2 performance using neural network gave the best performance since it can detect 100% of the malware with only 7.2% false positives.

Combining Features. In addition, we evaluated the use of combinations of features to attempt to combine their strengths. All features can be combined together to create a powerful detection. This design point is marked as COMB in Figure 2. As expected, both models perform best when all features are used together. However, this significantly increases the implementation complexity of MAP.

4 Online Malware Detection

In this section, we introduce the online detection component of MAP. Detecting malware execution during runtime is a time-series analysis problem where the time-series consists of the successive decisions of the classifier. To be effective, the detection algorithm must filter out occasional false positives and quickly detect true malicious behavior.

To make a decision that considers past behavior of programs, but is not dominated by them, we use Exponentially Weighted Moving Average (EWMA) [30]. EWMA is a form of a low-pass filter commonly used to smooth out transients in a time-series signal, giving more weight to more recent inputs. EWMA computation requires floating point operations and is not suitable for efficient hardware implementation. Instead, we use a fixed-point implementation by first considering binary decisions from the base classifier (making the time-series consist of 1's for malware and 0's for normal decisions). We then use a window of these decisions with integer weights that best correspond to the chosen smoothing factor (α), which determines how much weight to give to the latest classification compared to the weight given to prior samples.

In Figure 3 we show the precise EWMA result (for $\alpha = 0.2$) and a fixed point hardware implementation for an arbitrary binary input stream. For the results in Figure 3, the input stream is assumed to have 20 bits and the window size for fixed point implementation is 8. As seen from the graph, the approximate hardware implementation closely tracks the precise EWMA estimate. The hardware implementation has a weight for each input in a window: the weight of an input in k^{th} order (W_k) is calculated by $W_k = 2^{\lfloor n/2 \rfloor} + \sum_{i=1}^{\lfloor k/2 \rfloor} 2^i$ where n is the window size and $0 \leq k < n$. There are two accumulators, one for regular labels and one for malware labels. The last step performs a subtraction operation and obtains the absolute difference between the summations.

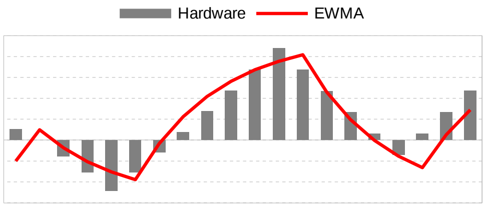


Fig. 3: EWMA vs. Fixed-point Approximation

Figure 4 shows the impact of the window size on the detection performance for the LR-based model with a trained threshold. While small windows cause around 100% false positive rate, the number of false positives decreases significantly with larger windows. As the window size continues to increase, false negatives also increase because malware behavior is more likely to be missed with larger windows. We use a window size of 16 to balance these two effects.

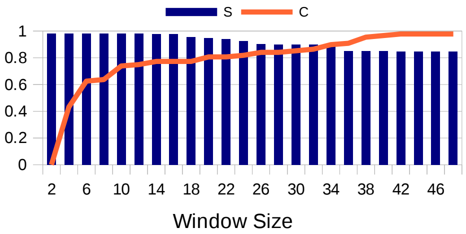


Fig. 4: Effect of Window Size on Detection Performance

5 Implementation

In this section, we describe the design and implementation of MAP using both the Logistic Regression (LR) and the Neural Network (NN) classifiers. In addition, we introduce some optimizations to simplify the implementation and evaluate their effect. The MAP logic is located at the end of the processor pipeline after the instruction commit stage; for instruction-based features, we only consider committed instructions. For the NN classifier, we consider the trade-offs between performance and complexity: increasing the number of neurons improves detection at the cost of more complex hardware implementation.

5.1 The MAP Microarchitecture

The general MAP microarchitecture is depicted in Figure 5. The Feature Collection (FC) component collects and prepares the feature being used for classification and provides it as an input to the Prediction Unit (PU). The PU implements the classifier (the LR or the NN) that provides a binary decision on one feature vector with 1 indicating malware, and 0 indicating normal program. The output of the PU is therefore a time-series consisting of the sequence of the PU decisions over time. This time-series is the input to the Online Detection (OD) module that carries out the time-series moving average analysis to provide a real-time decision on the currently executing program as explained in Section 4.

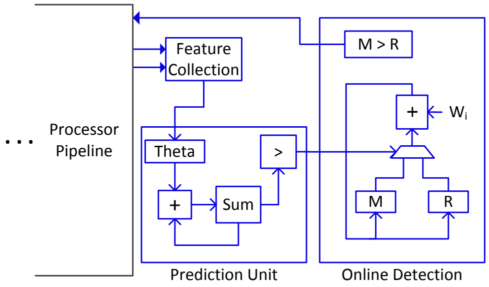


Fig. 5: MAP Microarchitecture with LR

For the implementation analyzed in this paper, we use the INS2 feature. Thus, the FC unit collects the committed instruction trace from the commit stage of the core pipeline. Other features require collection from the appropriate source of the feature events, such as the branch prediction unit, the memory management unit, or the fetch logic.

The MAP logic operates as follows. The FC unit collects and sends the features to the PU. The PU classifies the collected feature vector every classification period (we used 10K instruction period as with prior work [17]). The predictions are sent to the online detection module which applies the time-series algorithm as described in Section 4 to make a decision about the process. The counters in the OD module are treated as part of the process state; they are stored, restored and reset along with the process state on a context switch. A more secure option would be to store these counters in hardware. Since there are only two 32-bit registers in the OD module, it can synchronize with running processes without creating extra complexity.

5.1.1 *Logistic Regression Prediction Unit*

We implemented the logistic regression prediction unit using INS2 feature. The feature vector has 50 elements to represent selected opcodes. The Θ vector represents the weight of each feature as a floating point number based on the detector training. In the future, we envision a secure process that allows the update of Θ to allow the detector to evolve with evolving malware.

In a standard implementation of logistic regression [29], the features are multiplied with their weights (Θ) and accumulated to calculate the hypothesis. As a final step, the hypothesis is translated to a value between 0 and 1 by *sigmoid* function and the input is labeled according to the threshold. In theory, updating the feature vector for every commit and calculating the result at the checking granularity (10K instructions) is sufficient. However, in our implementation it is not necessary to wait for the end of the period. For every new committed instruction, we set the corresponding element of the feature vector to 1 and add its weight to the total value. However, we only send the detection signal to the OD unit when 10K instructions have committed. Therefore, in our implementation, the multiplication operation is not required. The feature weights (Θ), created after training, are all floating point numbers, but they are converted to 16-bit fixed point numbers with 3 integer and 13 fractional bits. The use of fixed-point arithmetic instead of floating point significantly reduces the complexity of our design [11]. For our studies, we used scalar pipeline. For a superscalar pipeline, there will be multiple bits set for each committed instruction and multiple adders will be required.

The final step of logistic regression is the sigmoid function and prediction. Sigmoid is an asymptotic function that creates values between 0 and 1. We discretize the prediction to produce a boolean classification using simple thresholding: if the classification threshold is 0.5, then all hypothesis values larger than 0 ($sigmoid(0) = 0.5$) will be classified as class 1 (malicious programs). The implementation of actual sigmoid function is not necessary since the threshold can be compared to the sum, instead of the sigmoid of the sum. In the last step of our LR implementation, we only compare this value with the predetermined threshold and send the result to the OD module.

5.1.2 Neural Network Prediction Unit

We implemented the neural network classifier as a multilayer perceptron (MLP) with 50 input features and a single hidden layer with 19 neurons. This configuration provides the best detection performance in the feature space we explored. In parallel to our machine learning model [53], we use $tanh$ as an activation function. An MLP with a single hidden layer operates by training a set of weights for each hidden neuron and the output neuron. Each hidden neuron calculates the dot product of their weights and feature vector, this value is then passed to a sigmoid function (in our case $tanh$). The output neuron operates like the hidden neurons except the output neuron uses the outputs of the hidden neurons as inputs, instead of using the feature vector.

We evaluated two designs with the same functionality. Our base design was implemented with performance constraints so that the neural network calculations are done in parallel. We then optimized this design for space constraints by serializing the operation of the neural network, which significantly reduced the number of operational units.

Similar to our LR implementation, both NN designs accumulate feature weights as feature data becomes available. Next, we calculate $\sum_{i=1}^{L} \tanh(a_i) \cdot w_i$ where L is the number of hidden neurons, a_i are the accumulated neuron values and w_i are the weights for each neuron in the output layer. Notice that we could not emit the actual implementation of the $tanh$ function while implementing the NN logic, because this time the output neuron requires the actual $tanh$ of the values calculated in the hidden layer. To reduce the complexity of both designs $tanh$ is approximated by a Look-up Table (LUT) [41]. In particular, the lookup table based implementation of $tanh$ function has a total absolute error of 0.062425 (error integrated over all input values of $tanh$). To further reduce complexity, we used fixed-point operations instead of floating point ones. To prevent the loss of precision and to reduce overflows, we use 16-bit values (3 integer plus 13 fractional bits) prior to multiplication and 32-bit values (6 integer plus 26 fractional bits) post multiplication. Finally we do not perform the final $sigmoid$ operations, opting instead to simply compare the resulting sum to a precalculated threshold.

Base Design. The base neural network design operates by calculating $tanh(a_i)$ for each a_i in parallel. Next, each $tanh(a_i)$ is multiplied by w_i (the corresponding weight) to generate the inputs to the ouput neuron in parallel. Finally, the products are summed using a reduction tree of adders to compute the sum in $\log_2(L)$ cycles. The final sum is compared with the threshold to produce the prediction. This design allows the classifier to be activated every cycle and produce a prediction in $T(tanh) + T(mul) + T(add) \cdot \log_2(L) + T(compare)$ cycles, where $T(x)$ is the number of cycles needed to perform x. However, the design requires L 16 bit accumulators, $tanh$ LUTs and multipliers, along with $\lceil \frac{L}{2} \rceil$ 32 bit adders.

Optimized Serial Design. The serial design operates by storing the accumulated values in a buffer, then multiplexing the values through a pipeline consisting of $tanh$, multiply, and accumulate. The final sum is compared to the threshold to produce the prediction. This design requires $T(setup) + T(tanh) + T(mul) + T(add) + L + T(compare)$ cycles to complete. While this unit is active, the accumulation of the feature data continues. However, another classification cannot be initiated until the previous feature set has been fully processed. Similar to the base parallel design, the serial design requires L 16-bit accumulators. However, as shown in Figure 6, the serial design requires only 1 $tanh$ LUT, 1 multiplier and 1 32 bit accumulator.

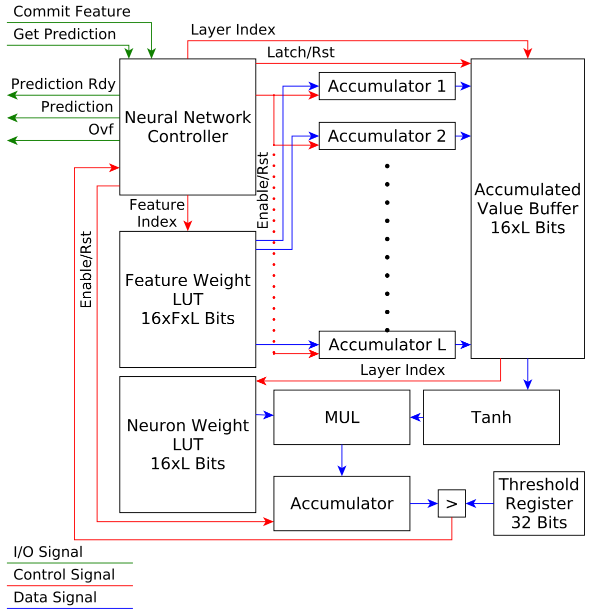


Fig. 6: Neural Network Serial Design

5.2 FPGA Implementation and Cycle Time Impact

We implemented MAP on an open source x86 processor (AO486) [2] using Verilog. The processor is a 32-bit in-order pipelined implementation of the Intel 80486 ISA. We synthesized the core with the MAP logic at the end of the pipeline on an Altera DE2-115 FPGA board [1] using Quartus II 13.1 software. We evaluated three different prediction unit options for MAP and summarized their time, area and power impact in Table 6. The MAP design with the LR prediction unit is extremely light-weight in terms of complexity and its impact on the core power and area is under 1%. The increase of the cycle time is caused by the exception transfer to the processor pipeline. However, it can be easily eliminated if the MAP exception transfer is performed over two cycles. For the NN prediction units, the base design requires substantial area and consumes significant power; in contrast, the optimized design uses only 5.67% of the core area. The cycle time impact of the NN designs could be reduced by deepening their pipelines. The processor area breakdown is shown in Figure 7 and MAP takes up 0.28-5% of the logic cells depending on the prediction unit choice.

We note that this overhead is relative to the small Intel 80486 class open core we extended. When considering modern cores which have two orders of magnitude more transistors, the overhead is likely to constitute a much lower percentage of the total core area.

TABLE 6: MAP's effect on core

	LR	NN Base	NN Serial
Logic Cells	+0.28%	+13.12%	+5.67%
Frequency	-1.93%	-2.28%	-5.53%
Power Usage	+0.08%	+5.23%	+1.66%

Our goal of implementing MAP on an FPGA was to show that it has minimal impact on the processor cycle time, power and area for a realistic system implemented within an x86 processor.

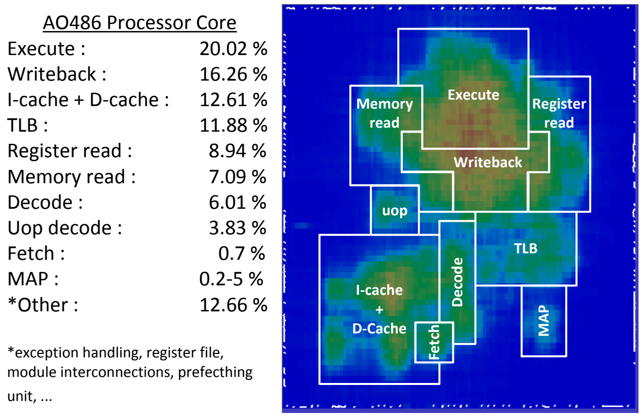


Fig. 7: MAP integrated into AO486 processor core

5.3 Two-level Framework and Integration Issues

We envision MAP to be part of a two-level framework where the hardware unit alerts the Operating System to invoke a more sophisticated analysis or isolation tool to monitor processes identified by the hardware component to be suspicious. Although our goal in this paper is primarily to explore the design space of the hardware component, it is important to understand how the integration between the hardware and the system can be carried out securely. For instance, malware may corrupt the operating system handlers to disable the communication between MAP and the software infrastructure. One possibility is to leverage recent hardware support for isolation, such as the ARM Trustzone [62] or the Intel SGX [40] to ensure that the communication with the second level of protection is secure. In a similar vein, the hardware component must be able to adapt to evolving malware. This requires a secure update mechanism (e.g., via attestation [54]) that allows the weights and thresholds of the classifier to be adapted by a trusted authority (for example, security provider); this is a problem similar to secure firmware update.

For simplicity, we assumed that the output of the first level classifier is a discrete malware/no-malware decision. However, the output of the classifier is a confidence value which we discretized using a threshold. One could then pass this confidence value to the second level allowing it to more finely prioritize the scanning of the processes based on the confidence in the decision and the availability of resources. Alternatively, MAP may even provide richer information to the second level, for example providing a summary of the feature vector exhibited by the suspcious application.

6 Evaluation MAP in Online Detection

In this section, we present the online detection results showing both conventional detection effectiveness (such as the ROC graph), as well as the translation from prediction unit outputs to online detection signals at runtime. We also present a sensitivity study of the impact of the classification instruction window size, which we have been fixing at 10K instructions, on the effectiveness of the detection.

6.1 Detection Effectiveness

Our hardware implementation of online detection is based on INS2 feature, because of the ease of collecting the feature vector in hardware as well as its excellent performance during offline analysis. In Figure 8, we show the detection success using the ROC graphs. The first graph shows the sensitivity of the detector that is based on an LR prediction unit. As seen from the results, it can detect almost 90% of the malware with 6% false positive rate at its most optimal configuration. The same feature can detect 93% of malware with the same false positive rate, if after-the-fact detection was possible. The second ROC graph in Figure 8 shows detection performance of the detector with an NN-based prediction unit. While the INS2 feature can detect all malware with 7% false positive rate with after-the-fact detection, it can still detect 94% of malware at runtime with the same false positive rate.

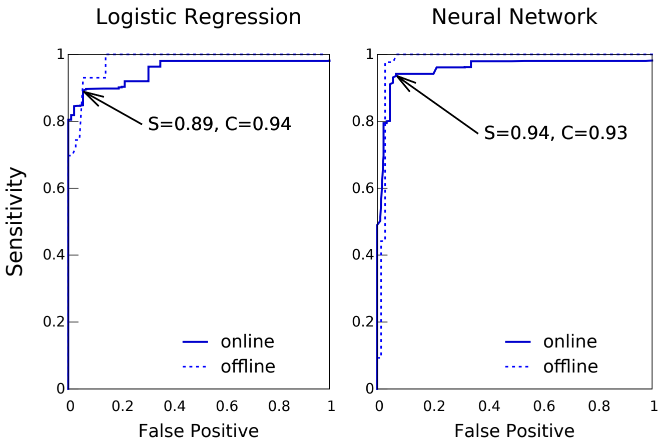


Fig. 8: Online Detection Performance

Next, we show how periodic signals from Prediction Unit (PU) are translated into a detection signal at runtime by the Online Detection (OD) counters. In Figures 9 and 10, we show the first 200 instances of 10K instruction periods for a malware sample from *Virut* family and one of the Spec2K6 benchmarks (*mcf*). In Figure 9, the prediction unit is implemented using the LR model. For *Virut* sample, the PU output shows that the executed program is a malware in the beginning. However, after some period of time the output becomes indicative of a regular program, causing PU to output zeros. The online detection logic smooths these infrequent signals and correctly predicts that the executed program is malware. Similarly, for *mcf*, the ”malicious program” output signals are smoothened by the OD unit.

In Figure 10, we show the generation of the detection signal by the OD unit from the periodic outputs of the PU that implements the NN model. As seen from the figure, the NN prediction is more sensitive to the behavior of the program compared to the LR. For *Virut*, NN generates some ”regular program” outputs even in the first phase of *Virut*. Again, smoothing these discrete signals from the PU output successfully creates a continuous correct detection result at runtime. For *mcf*, the NN model generates less ones than LR, because of the sensitivity of the model is higher.

A MAP configuration must fit within the desired hardware budget. With a neural network, it is possible to get better sensitivity than with logistic regression. However, the hardware requirements for the LR implementation are almost negligible and therefore, subject to hardware budget constraints, it may prove to be a more attractive candidate.

6.2 Impact of Classification Window

Thusfar, we have been using a classification window of 10K instructions: the features are accumulated for the duration of this window, and a classification decision on the feature is taken using LR or NN. In this study, we evaluate the sensitivity of the detection to the size of this window for INS2 feature using LR.

We collected the feature data for different window sizes and carried out the classification. The results of these experiments are presented in Figure 11. The accuracy of detection was poor for low classification windows. However, as the window size is increased, the detection accuracy increases, but it is not stable until the window size approaches 10K instructions. These results support the choice of 10K classification windows.

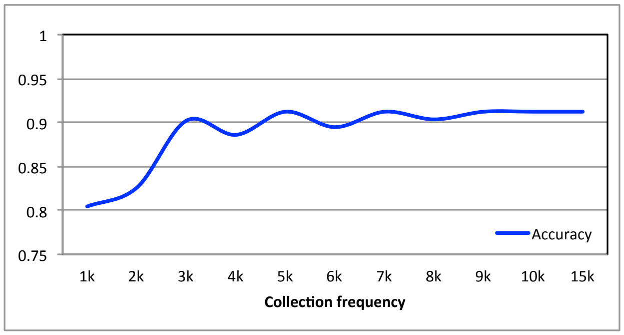


Fig. 11: Sensitivity to Classification Window Size

7 Feature Selection to Reduce Feature Sizes

The features used by MAP's classifiers use long vectors that measure the frequency or existence of different low-level events. For example, the INS2 feature vector consists of 50 bits corresponding to the constituent opcodes. However, it is possible that some of these opcodes do not contribute significantly to the classification success. If we

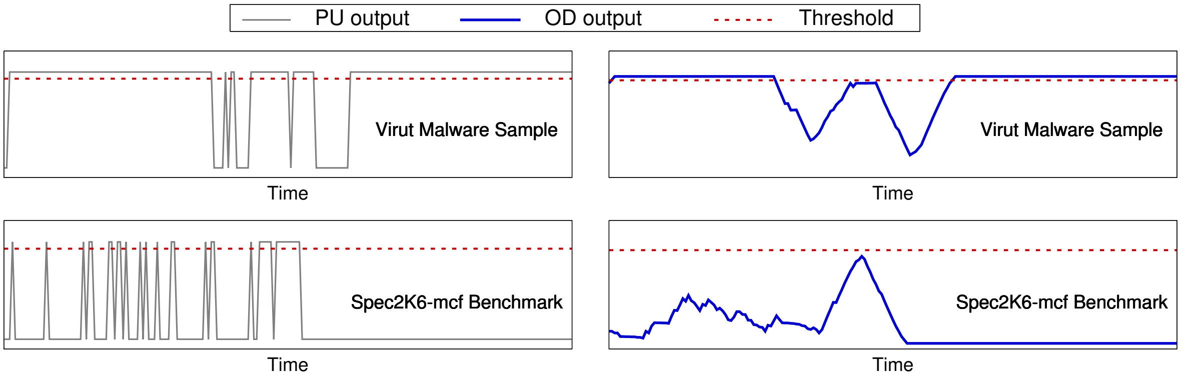


Fig. 9: Translation of Prediction Unit Output to Online Detection Signal at Runtime with LR-Based Detector

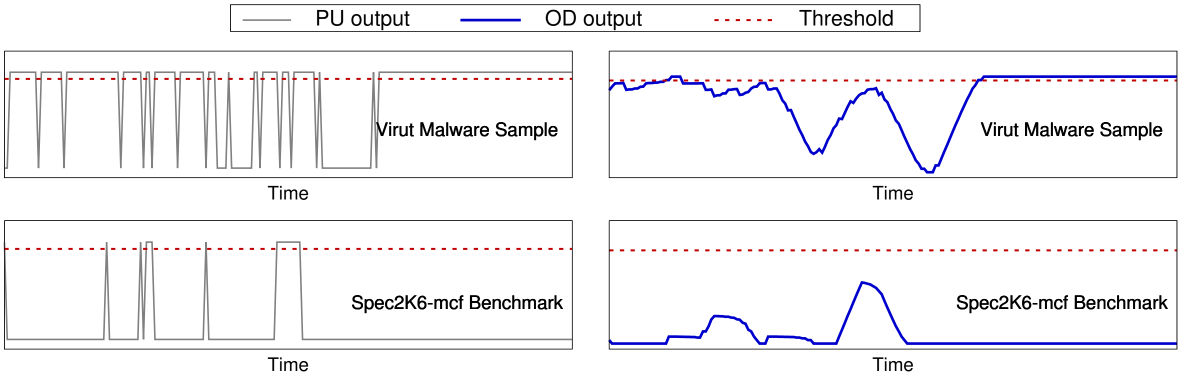


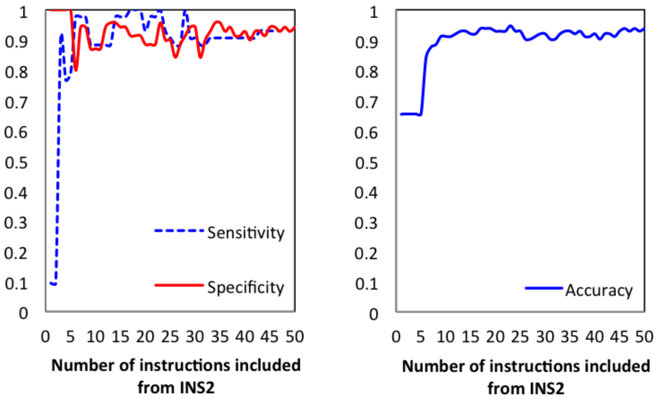
Fig. 10: Translation of Prediction Unit Output to Online Detection Signal at Runtime with NN-Based Detector

recognize and remove these features, we end up with smaller feature vectors, which simplifies the hardware implementation.

We first looked at the Θ vector resulting from using logistic regression on all 50 opcodes. Since Θ has a weight for each opcode, the opcodes with low weights are unlikely to be contributing meaningfully to the classification result. Thus, we sorted the opcodes in a descending order based on the absolute value of their weights. Furthermore, we trained 50 detectors in the following way. The first detector uses only the opcode with the highest weight. The second detector adds the next highest opcode, and so on. The final detector is the original INS2 detector with all 50 opcodes.

The detectors performance is shown in Figure 12. The figure shows that the performance with only 6 opcodes is only slightly worse than the one that uses all the opcodes. A detector built with just six highest-weighted opcodes can detect 90% of the malware with 20% false positives.

The feature selection approach above is simple but *ad hoc* in nature. Therefore, we explored two formal feature selection methods. The first method, Stepwise Regression [34], is an iterative method where at each iteration, all the features are tested using F-test and only the best feature among them is added to the model. The F-test is used to check if the means between two features are significantly different. The best feature is the feature that has the least p-value. The p-value represents the probability that the results could have happened by chance. The method terminates when all the features (opcodes) have been added or when the p-value for the remaining features is below a predefined threshold. In our experiment, we selected a high threshold so that we make sure to include all the opcodes in the model. We sorted the opcodes based on the order they were selected to be included in the model and 50 detectors were built iteratively and evaluated, as with the previous experi-

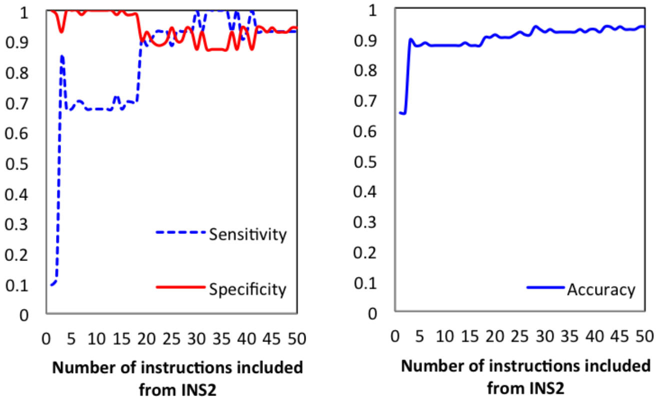


(a) Sensitivity and specificity

(b) Accuracy

Fig. 12: INS2 feature size impact using sorted weights

ment. The performance of the detectors is presented in Figure 13. The graph shows a notable increase in the accuracy when using the first three opcodes that were included to the model using stepwise regression. The detector can detect 83% of the malware with 8% false positives.



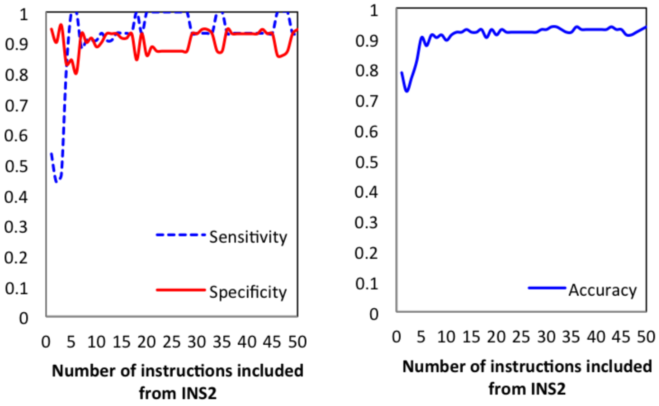
(a) Sensitivity and specificity

(b) Accuracy

Fig. 13: INS2 feature size impact using stepwise regression

The second feature selection method that we used was the sequential method, which is also an iterative method that tries to sequentially add features (opcodes) to the model until it reaches the point where no further improvement in prediction is achieved when adding more features. The sequential method calculates the mean criterion (sum of squared errors) values for all candidate features subsets at each step. Then the subset that minimizes the mean criterion value is chosen. As before, we sorted the opcodes based on the order they were selected to construct 50 detectors. Figure 14 shows the performance of the detectors. The sequential feature selection method outperformed the other two methods (sorted weights, and stepwise regression) in terms of decreasing the feature vector size while keeping the performance high. The detector that was built using the first five opcodes that were included in the model by the sequential method can detect all malware programs with less than 16% false positives.

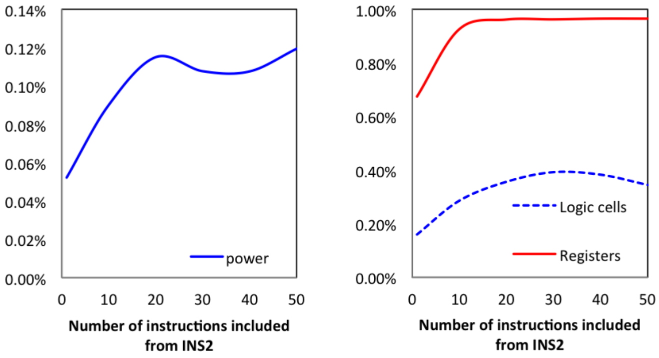
To show the optimization that can be achieved when reducing the features vector size, we estimated the area and power of the MAP hardware implementation of various feature vectors. Figure 15 shows the INS2 feature size impact on the area and power of the MAP logic. The results show that a detector built by the sequential method using just first five highest-weight opcodes reduces the power of the MAP logic by 50%, the number of logic cells by 33%, and the number of registers by 17% compared to an implementation that uses all of the opcodes.



(a) Sensitivity and specificity

(b) Accuracy

Fig. 14: INS2 feature size impact using sequential method



(a) Power

(b) Logic cells and registers

Fig. 15: INS2 feature size impact on MAP hardware complexity

8 Related Work

Malware Detection. Malware detection is an area that has attracted extensive research and commercial interest over the past decade. In general, malware detection techniques are either static (focusing on the structure of a program or system) or dynamic (analyzing the behavior during execution) [31]. Detection approaches are also classified as signature-based (looking for signatures of known malware) or anomaly-based (modeling the normal structure/behavior of programs or systems and detecting deviations from this model).

Static approaches including virus and spyware scanners are the first line of defense in malware detection. Originally, these scanners are operated using pattern matching to look for signatures of known malware. However, these approaches can be easily evaded using program obfuscation or simple code transformations that preserve the function of the malware but make it not match the patterns known to the scanner [43]. More advanced detectors based on semantic signatures have been proposed, and significantly improved the performance of static scanners [13]. Static approaches are limited and can be bypassed by sophisticated attackers [42].

In particular, code obfuscation techniques (polymorphic malware), and malware encryption (packing or metamorphic malware) are both sufficient to hide even from these more advanced detectors [42].

Dynamic detection observes the behavior of the program (or the system) as it runs and interacts with the environment. Dynamic behavior-based detection attempts to detect deviations from normal behavior of a program as it operates. It detects anomalies in the observed behavior compared to its model of normal behavior, which is often program-specific, to identify malware. A large number of software malware detectors have been investigated that vary in terms of the monitored events, the normal behavior model, and the detection algorithm [28], [50], [32], [31], [37]. The advantage of dynamic detection is that it is resilient to metamorphic and polymorphic malware [42], [39]; it can even detect previously unknown malware. However, disadvantages include a typically high false positive rate, and the high cost of monitoring during run-time. Moreover, since detection is a one time (or periodic) process, malware can evade detection either probabilistically or by recognizing that it is being observed and acting normally for that period.

Most similar to our work, RiskRanker uses a rule-based lightweight detection pass to rank the risk posed by different Android based Apps [26]. The analysis requires around 4 days of processing time, to identify a high risk set (comprising about 3% of the scanned 118,000 Apps). About one fourth of this set was found to actually have malware, including 322 zero-day exploits. MAP uses the same premise of a two-level monitoring; however, we do so in real-time for live systems.

Use of Low-level Features. A number of earlier works explored low-level features for malware detection. Bilar et al [9] examine the frequency of opcode use in malware. Santos et al and Yan et al evaluate opcode sequence signatures [52], [63], while in particular, opcode sequence signatures were found to effectively classify metamorphic malware. Runwal et al [51] study opcode sequence similarity graphs. These techniques obtain this information from running programs and malware inside heavyweight profiling tools such as Pin [14]. Moreover, all of these works consider offline analysis, rather than online detection.

Demme et al [17] collect performance counter statistics for programs and malware under execution. They show that offline machine learning tools can effectively classify malware. They conjecture that an online detector can therefore be built but do not explore this idea further. Our work builds on this evidence to develop a lightweight online hardware-supported malware detector. Tang et al [57] demonstrated that unsupervised learning on low-level feature can also successfully classify malware offline; unsupervised learning may be more amenable to detecting novel malware and attacker evolution. However, unsupervised learning also requires more sophisticated analysis implying more complex hardware implementations.

In general, we expect our proposed solution to operate effectively with other solutions by monitoring the processes to detect any malware that escapes detection using these other techniques. An orthogonal line of research pursues protection of application secrets even in the presence of compromised system software layers and malware [21], [22], [40].

9 Concluding Remarks

This paper contributes an always-on hardware malware detection engine called MAP. MAP is integrated at the commit stage of a conventional processor, which enables it to collect low-level features with low power consumption, and without software interference. MAP builds on recent important work that showed that hardware counters can be used to classify malware from normal programs off-line [17]. We explore the use of different low-level features for online detection, and show that these features using logistic regression can achieve excellent sensitivity and reasonable false positive rates.

Because of the false positives which are common in anomaly-based malware detection approaches, we propose to use MAP in combination with a heavier-weight software-based detector. In particular, MAP prioritizes the scanning order of processes such that those processes that are most anomalous are scanned first. There are a number of interesting integration issues when interfacing the two levels that form part of our future research. Moreover, the always-on nature of MAP makes it difficult for malware to avoid detection. We developed the hardware design for MAP and showed that its delay, complexity and energy consumption are small.

Our future work considers a number of follow-up directions. First, we would like to understand the phenomena that causes malware to behave differently from normal programs in the low-level feature space, to be able to better select features and anticipate attacker evolution. For example, most modern malware uses one of a relatively few packer utilities to encrypt the code and avoid signature based detection; perhaps the signature of these packers are contributing to the detection efficiency. Second, with every use of anomaly detection in an adversarial setting, one must expect the attackers to attempt to adapt. Thus, the behavior is not static and the detectors must evolve in reaction to attacker evolution. This problem of adversarial learning is well-studied in the machine learning community and we hope to integrate suitable approaches from that community to address this important issue. Finally, we would like to explore improvements to the detection including the use of alternative machine learning algorithms, the use of ensemble learning to build detectors specific to different malware categories to enhance detection success, and performance and power optimizations to the detection implementation.

We also expect our technique to be esepcially sensitive to Code Reuse Attacks, including both return-oriented [56] and jump-oriented [10] which remain dangerous vulnerabilities despite some promising solutions [46], [66], [35], [36]. In particular, these attacks have a unique computational footprint which will naturally allow our low level classifiers to identify it as malware. Similarly, certain classes of side channel and covert channel attacks [48], [19], [23] are also extremely dangerous and difficult to detect, but have a distinctive computational footprint that results from the need to cause contention on shared resources.

10 Acknowledgement

This material is based on research sponsored by the National Science Foundation grant CNS-1018496. Caleb Donovick was partially supported through the REU supplement award CNS-1338672. Iakov Gorelik was partially supported by the REU Site Award CCF-1005153.

References

[1] "De2-115 development and education board," 2010, http://www.altera.com/education/univ/materials/boards/de2-115/unv-de2-115-board.html.

[2] "The ao486 project," 2014, accessed May 2014 at http://opencores.org/project,ao486.

[3] "Laboratory for dependable distributed systems university of mannheim," 2014, accessed Feb. 2014 at http://pi1.informatik.uni-mannheim.de/malheur/.

[4] "Malware protection center," 2014, accessed May 2014 at http://www.microsoft.com/security/portal/mmpc/shared/malwarenaming.aspx.

[5] S. Abraham and I. Chengalur-Smith, "An overview of social engineering malware: Trends, tactics, and implications," *Technology in Society*, vol. 32, no. 3, pp. 183–196, 2010.

[6] Y. Abu-Mostafa, M. Magdon-Ismail, and H. Lin, *Learning from Data: A short course*. AMLBook, 2012.

[7] C. Aldrich and L. Auret, *Unsupervised process monitoring and fault diagnosis with machine learning methods*. Springer, 2013.

[8] S. Bandhakavi, S. King, P. Madhusudan, and M. Winslett, "Vex: Vetting browser extensions for security vulnerabilities." in *Proc. USENIX Security Symposium*, 2010.

[9] D. Bilar, "Opcode as predictor for malware," 2007.

[10] T. Bletsch, X. Jiang, V. W. Freeh, and Z. Liang, "Jump-oriented programming: a new class of code-reuse attack," in *Proceedings of ASIACCS*. ACM, 2011, pp. 30–40. [Online]. Available: http://doi.acm.org/10.1145/1966913.1966919

[11] J. Cavanagh, *Computer Arithmetic and Verilog HDL Fundamentals*. CRC Press, 2009.

[12] M. Charney, "Xed2 user guide," 2011, http://software.intel.com/sites/landingpage/pintool/docs/56759/Xed/html/main.html.

[13] M. Christodorescu, S. Jha, S. Seshia, D. Song, and R. Bryant, "Semantics-aware malware detection," in *Proc. IEEE Symposium on Security and Privacy*, 2005, pp. 32–46.

[14] C.Luk, R. Cohn, R. Muth, H. Patil, A. Klauser, G. Lowney, S. Wallace, V. Reddi, and K. Hazelwood, "Pin: building customized program analysis tools with dynamic instrumentation," in *Proc. PLDI*, 2005.

[15] C. Cooper, "Intelligence chief offers dire warning on cyberattacks," 2013, an article on CNET retrieved from http://www.cnet.com/news/intelligence-chief-offers-dire-warning-on-cyberattacks/.

[16] N. Dalvi, P. Domingos, M. Sumit Sanghai, and D. Verma, "Adversarial classification," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004, pp. 99–108.

[17] J. Demme, M. Maycock, J. Schmitz, A. Tang, A. Waksman, S. Sethumadhavan, and S. Stolfo, "On the feasibility of online malware detection with performance counters," in *Proceedings of the 40th Annual International Symposium on Computer Architecture*, ser. ISCA '13. New York, NY, USA: ACM, 2013, pp. 559–570. [Online]. Available: http://doi.acm.org/10.1145/2485922.2485970

[18] A. Dinaburg, P. Royal, M. Sharif, and W. Lee, "Ether: malware analysis via hardware virtualization extensions," in *Proceedings of the 15th ACM conference on Computer and communications security (CCS)*, 2008, pp. 51–62.

[19] L. Domnitser, A. Jaleel, J. Loew, N. Abu-Ghazaleh, and D. Ponomarev, "Non-monopolizable caches: Low-complexity mitigation of cache side channel attacks," *ACM Trans. Architecture and Code Optimization*, vol. 8, no. 4, pp. 1–35, Jan. 2012.

[20] M. Egele, T. Scholte, E. Kirda, and C. Kruegel, "A survey on automated dynamic malware-analysis techniques and tools," *ACM Computing Surveys (CSUR)*, vol. 44, no. 2, 2012.

[21] J. Elwell, R. Riley, N. Abu-Ghazaleh, and D. Ponomarev, "A non-inclusive memory permissions architecture for protecting against cross-layer attacks," in *Proc. International Symposium on High Performamce Computer Architecture (HPCA)*, Feb. 2014.

[22] D. Evtyushkin, J. Elwell, M. Ozsoy, D. Ponomarev, N. Abu-Ghazaleh, and R. Riley, "Iso-x: A flexible architecture for hardware-managed isolated execution," in *Proc. International Symposium on Microarchitecture (MICRO)*, Dec. 2014.

[23] D. Evtyushkin, D. Ponomarev, and N. Abu-Ghazaleh, "Covert channels through branch predictors," in *Proc. of the Workshop on Hardware and Architecture Security and Privacy (with ISCA)*, 2015.

[24] "Intel architecture instruction set extensions programming reference," 2014, accessed Feb. 2014 at http://download-software.intel.com/sites/default/files/319433-015.pdf.

[25] T. Garfinkel and M. Rosenblum, "A virtual machine introspection based architecture for intrusion detection," in *Proc. Usenix Symposium on Network and Distributed System Security (NDSS)*, 2003.

[26] M. Grace, Y. Zhou, Q. Zhang, S. Zou, and X. Jiang, "Riskranker: scalable and accurate zero-day android malware detection," in *Proceedings of the 10th international conference on Mobile systems, applications, and services (MobiSys)*, 2012, pp. 281–294.

[27] G. Gu, P. Porras, V. Yegneswaran, M. Fong, and W. Lee, "Bothunter: Detecting malware infection through ids-driven dialog correlation," in *Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium*, 2007.

[28] S. Hofmeyr, S. Forrest, and A. Somayaji, "Intrusion detection using sequences of system calls," *Journal of computer security*, vol. 6, no. 3, pp. 151–180, 1998.

[29] D. W. Hosmer Jr. and S. Lemeshow, *Applied Logistic Regression*. John Wiley & Sons, 2004.

[30] R. J. Hyndman, A. B. Koehler, J. K. Ord, and R. D. Snyder, *Forecasting with exponential smoothing*. Springer, 2008.

[31] N. Idika and A. Mathur, "A survey of malware detection techniques," technical Report, Departemnt of Computer Science, Purdue University. Accessed Feb. 2014 at: http://cyberunited.com/wp-content/uploads/2013/03/A-Survey-of-Malware-Detection-Techniques.pdf.

[32] G. Jacob, H. Debar, and E. Filiol, "Behavioral detection of malware: from a survey towards an established taxonomy," *Journal in computer Virology*, vol. 4, no. 3, pp. 251–266, 2008.

[33] V. G. Jim Guilford, Kirk Yap, "Fast SHA-256 Implementations on Intel Architecture Processors," Intel Corporation, Tech. Rep., May 2012.

[34] J. B. Kadane and N. A. Lazar, "Methods and criteria for model selection," *Journal of the American statistical Association*, vol. 99, no. 465, pp. 279–290, 2004.

[35] M. Kayaalp, M. Ozsoy, N. Abu-Ghazaleh, and D. Ponomarev, "Branch regulation: Low overhead mitigation of code reuse attacks," in *Proceedings of ISCA*, 2012.

[36] M. Kayaalp, T. Schmitt, J. Nomani, D. Ponomarev, and N. Abu-Ghazaleh, "Scrap: Architecture for signature-based protection from code reuse attacks," *2013 IEEE 19th International Symposium on High Performance Computer Architecture (HPCA)*, vol. 0, pp. 258–269, 2013.

[37] C. Kolbitsch, P. M. Comparetti, C. Kruegel, E. Kirda, X.-y. Zhou, and X. Wang, "Effective and efficient malware detection at the end host." in *USENIX Security Symposium*, 2009, pp. 351–366.

[38] S. Kramer and J. Bradfield, "A general definition of malware," *Journal in Computer Virology*, vol. 6, no. 2, 2010.

[39] L. Martignoni, M. Christodorescu, and S. Jha, "Omniunpack: Fast, generic, and safe unpacking of malware," in *IEEE Annual Computer Security Applications Conference (ACSAC)*, 2007, pp. 431–441.

[40] F. McKeen, I. Alexandrovich, A. Berenzon, C.Rozas, H. Shafi, V. Shanbhogue, and U. Savagaonkar, "Innovative instructions and software model for isolated execution," in *Wkshp. on Hardware and Architectural Support for Security and Privacy, with ISCA'13*, 2013.

[41] P. Meher, "An optimized lookup-table for the evaluation of sigmoid function for artificial neural networks," in *VLSI System on Chip Conference (VLSI-SoC), 2010 18th IEEE/IFIP*, Sept 2010, pp. 91–95.

[42] A. Moser, c. Kruegel, and E. Kirda, "Limits of static analysis of malware detection," in *IEEE Annual Computer Security Applications Conference (ACSAC)*, 2007, pp. 421–430.

[43] C. Nachenberg, "Computer virus-antivirus coevolution," *Communications of the ACM*, vol. 40, no. 1, pp. 46–51, Jan. 1997.

[44] J. Oberheide and C. Miller, "Dissecting the android bouncer," 2012, presentation at SummerCon, accessed online in October 2015 from http://diyhpl.us/~bryan/papers2/security/android/summercon12-bouncer.pdf.

[45] "Open Malware," accessed Feb. 2014 at: http://www.offensivecomputing.net/.

[46] K. Onarlioglu, L. Bilge, A. Lanzi, D. Balzarotti, and E. Kirda, "Gfree: Defeating return-oriented programming through gadget-less binaries," in *Proc. of Annual Computer Security Applications Conference (ACSAC)*, 2010, pp. 49–58.

[47] M. Ozsoy, D. Ponomarev, N. Abu-Ghazaleh, and T. Suri, "SIFT: A low-overhead dynamic information flow tracking architecture for smt processors," in *Proceedings of the ACM International Conference on Computing Frontiers*, May 2011.

[48] C. Percival, "Cache missing for fun and profit," 2005, http://www.daemonology.net/papers/htt.pdf.

[49] PWC CIO and CSO Offices, "The global state of information security survey," 2015.

[50] M. Roesch, "Snort: Lightweight intrusion detection for networks." in *Proc. Usenix System Adminsitration Conference (LISA)*, 1999, pp. 229–238.

[51] N. Runwal, R. M. Low, and M. Stamp, "Opcode graph similarity and metamorphic detection," *J. Comput. Virol.*, vol. 8, no. 1-2, pp. 37–52, May 2012. [Online]. Available: http://dx.doi.org/10.1007/s11416-012-0160-5

[52] I. Santos, F. Brezo, J. Nieves, Y. K. Penya, B. Sanz, C. Laorden, and P. G. Bringas, "Idea: Opcode-sequence-based malware detection," in *Engineering Secure Software and Systems*. Springer, 2010, pp. 35–43.

[53] M. Schmid, "A feed forward multi-layer neural network," 2010.

[54] A. Seshadri, M. Luk, A. Perrig, L. van Doorn, and P. Khosla, "Scuba: Secure code update by attestation in sensor networks," in *Proceedings of the 5th ACM workshop on Wireless security*, 2006, pp. 85–94.

[55] "Software Guard Extensions Programming Reference," 2014, accessed Feb. 2014 at http://download-software.intel.com/sites/default/files/319433-015.pdf.

[56] H. Shacham, "The geometry of innocent flesh on the bone: Return-into-libc without function calls (on the x86)," in *Proceedings of CCS*. ACM Press, Oct. 2007, pp. 552–61.

[57] A. Tang, S. Sethumadhavan, and S. Stolfo, "Unsupervised anomaly-based malware detection using hardware features," in *Research in Attacks, Intrusions and Defenses*, ser. Lecture Notes in Computer Science, 2014, vol. 8688, pp. 109–129.

[58] P. Team, "Pax non-executable pages design & implementation," http://pax.grsecurity.net/docs/noexec.txt.

[59] "VirusTotal," accessed Feb. 2014 at: https://www.virustotal.com/en/.

[60] Y. Vorobeychik and B. Li, "Optimal randomized classification in adversarial settings," in *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014)*, 2014.

[61] "Crimeware protection: 3rd generation intel core vpro processors," 2014, accessed Feb. 2014 at http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/3rd-gen-core-vpro-security-paper.pdf.

[62] J. Winter, "Trusted computing building blocks for embedded linux-based arm trustzone platforms," in *Proceedings of the 3rd ACM workshop on Scalable trusted computing*, 2008, pp. 21–30.

[63] G. Yan, N. Brown, and D. Kong, "Exploring discriminatory features for automated malware classification," in *Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2013, pp. 41–61.

[64] H. Yin, D. Song, M. Egele, C. Kruegel, and E. Kirda, "Panorama: capturing system-wide information flow for malware detection and analysis," in *Proceedings of the 14th ACM conference on Computer and communications security (CCS)*, 2007, pp. 116–127.

[65] H. Zhang, D. She, and Z. Qian, "Android root and its providers: A double-edged sword," in *Proceedings of ACM Conference on Computer and Communications Security (CCS)*, 2015.

[66] M. Zhang and R. Sekar, "Control flow integrity for cots binaries," in *Proc. 22nd Usenix Security Symposium*, 2013.

Meltem Ozsoy is a Research Scientist at Intel Labs, Hillsboro, OR. She received her PhD in Computer Science from SUNY Binghamton. Her research interests are in the areas of computer architecture and secure system design.

Khaled N. Khasawneh is a PhD student in the Department of Computer Science and Engineering at the University of California at Riverside. He received his MS degree in Computer Science from SUNY Binghamton in 2014. His research insterests are in architecture support for security.

Iakov Gorelik is currently a software engineer at CitiBank. He received his Bachelors Degree in Computer Science from SUNY Binghamton.

Caleb Donovick is an undergraduate student in the Department of Computer Science at SUNY Binghamton.

Nael Abu-Ghazaleh is a Professor in the Computer Science and Engineering department and the Electrical and Computer Engineering department at the University of California at Riverside. His research interests are in the areas of secure system design, parallel discrete event simulation, networking and mobile computing. He received his PhD from the University of Cincinnati in 1997.

Dmitry Ponomarev is a Professor in the Department of Computer Science at SUNY Binghamton. His research interests are in the areas of computer architecture, secure and power-aware systems and high performance computing. He received his PhD from SUNY Binghamton in 2003.