New Intrusion Classification System Used Parallel Clustering Algorithm

Bh.Dasaradha Ram[1], Dr.B.V.Subba Rao[2]
[1] Research Scholar, Dept of CSE
[2]Professor & Head, Dept of IT
[1]Rayalaseema University, Kurnool, (AP) – India
[2]pvpsiddhartha Engineering College (AP) – India

Abstract- *Intrusion Detection Technology is a research hotspot in the field of information security. This study introduces the types of traditional intrusion detection and data mining technology. The present article gives an overview of existing Intrusion Detection Systems (IDS) along with their main principles. Also this article argues whether data mining and its core feature which is knowledge discovery can help in creating Data mining based IDSs that can achieve higher accuracy to novel types of intrusion and demonstrate more robust behavior compared to traditional IDSs. In this research work new guidelines is proposed for an efficient GPU adaptation of Aho-corasick algorithm for regular expression matching. Also several techniques are introduced to optimization on GPU, including reducing global memory access, storage format for output table. In case of misuse detection, intrusion patterns are built automatically from a training data by the use of the random forest classification method. The adaptive immune system in our proposed architecture also takes advantage of the distributed structure, which has shown better self-improvement rate compare to centralized mode and provides primary and secondary immune response for unknown anomalies and zero-day attacks.*

Keywords- Artificial immune system, Innate immune system, Data mining, Random Forest and Weighted K-Means. Aho-Corasick, Graphics processing Unit, security, intrusion detection

I. INTRODUCTION

With the rapid development of the internet, the various attacks, which emerge endlessly in the network, have become a major threat to network and information security [1]. Traditionally, network users usually use firewall as the first line of defense for security. But with attacking tools and means becoming much more complicated, simple firewall is difficult to resist various attacks; therefore people put forward a kind of technology which can discover in time and report unauthorized or abnormal phenomena in the system, named intrusion detection technology [2]. Recent exploits also suggest that the more sensitive the information that is held is, the higher the probability of being a target. Several Retailers, banks, public utilities and organizations have lost millions of customer data to attackers, losing money and damaging their brand image [3]. In Multi pattern Matching algorithm we have to report all occurrence of pattern in given string. Multi pattern string matching use in number of application is network intrusion detection digital forensics, natural language processing [4]. For example Snort is open source network intrusion detection system which contained thousands of pattern that are match against packet in network for virus/worm signature detection [5]. They are correlation feature selection (CF) and minimal redundancy maximal relevance (mRMR). Another challenge of intrusion detection system is an imbalance between real and trained data [6]. Misuse detection has a key advantage is their high rate of accuracy in detecting known attacks. Their main drawback is the inability to detect novel attacks. Anomaly detection, built profiles based on normal behavior [7]. Anomaly detection, detect novel attacks hybrid intrusion detection system, combine the advantages of misuse and anomaly detection [8]. Machine learning methods can be organized based on the type of input available during training. There are three main categories of machine learning supervised semi-supervised and unsupervised algorithms [9]. Supervised machine learning algorithms need to be trained by labeled data to distinguish the normal and abnormal behavior of the network. Semi-supervised machine learning algorithms can be trained by attack-free unlabeled data [10].

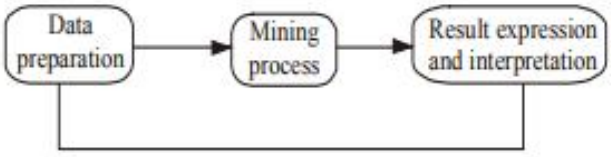

Fig. 1: The whole process of data mining

II. RELATED WORK

Many intrusion detection systems a set of rules are used to describe intrusions. The detection techniques are applied in misuse and anomaly detection [11]. The main drawback of these two techniques resides in the encoded models, which defines normal and malicious activities. In open source network security tool for intrusion detection in the campus network environment [12] . Two data mining techniques that are random forest and k-means algorithms are used in misuse, anomaly and hybrid detection [13]. The human immune system defends the human body against harmful and previously unseen foreign cells using lymphocyte cells. The foreign cells are called antigens, such as bacteria and viruses [14]. The artificial immune system is designed for the computational system and inspired by the it is applied to solving various problems in the field of information security, particularly intrusion detection systems [15]. The objective of IDS is to alert administrators of suspicious activities and in some cases even attempt to circumvent the attacks. The practices employed in IDSs do differ from other security techniques such as firewalls, access control or encryption which aim to secure the computer system [16]. With this being identified however it is strongly recommended that these security practices are used in conjunction with one another as this reinforces defense of a system and ensures that a much larger scope of a system is protected [17].

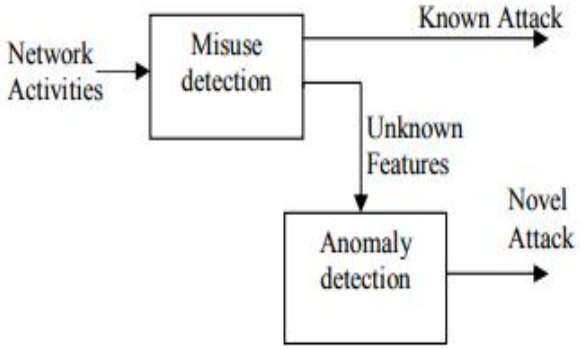

Fig.2. Overview of Hybrid Intrusion Detection

III. INTRUSION DETECTION SYSTEMS MODEL

Finding model is the core idea of data mining, which is applied into intrusion detection system through the data mining algorithm. The system can find the knowledge and rule which can be contributed to detect attack from the system log. Association rules mining algorithm [18]. Correspondingly it may be added that an intrusion detection system is the practical implementation of intrusion detection principles and mechanisms over a network [19]. Graphic Processing Units (GPUs) have been developed to cope with the high performance requirements of graphical and animation tasks. They have different architecture than Central Processing Units (CPUs)[20] stressing floating point operations, fine grained concurrency, and high data rate memories. They have usually taken the form of co-processors of the CPU.

- In the CPU, our first task is to move data to Memory of GPU from main Memory of CPU. In this, call does not return until all data has been fully transferred.
- Then we send GPU command that GPU will execute .In this operation control is return as soon as command are execute.
- The GPU execute the commands that have been received.
- After data is computed at GPU we retrieve that data in CPU by Coping from device memory to host memory. The CPU needs to know that the commands have been completed before retrieving the results.

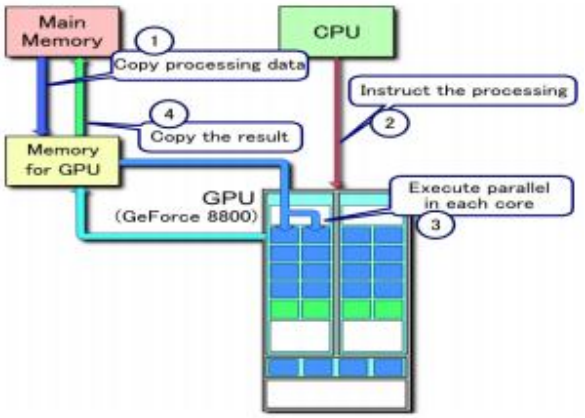

Fig 3: Processing Flow between GPU and CPU

IV. PROPOSED INTRUSION DETECTION SYSTEM

The clustering engine performs network traffic clustering into the self or non-self clusters through unsupervised learning techniques. The AIS engine consists of agents that cooperate for intrusion detection. The term agent originally comes from Artificial Intelligence (AI) and refers to anything that can view its environment through sensors and act upon that environment using actuators. The hybrid detection system has several key advantages [21]. In misuse detection, random forest classification algorithm is used to achieve high speed. Accuracy of misuse detection is also high. The performance level of anomaly detection tends to be reduced, due to a large number of connections.

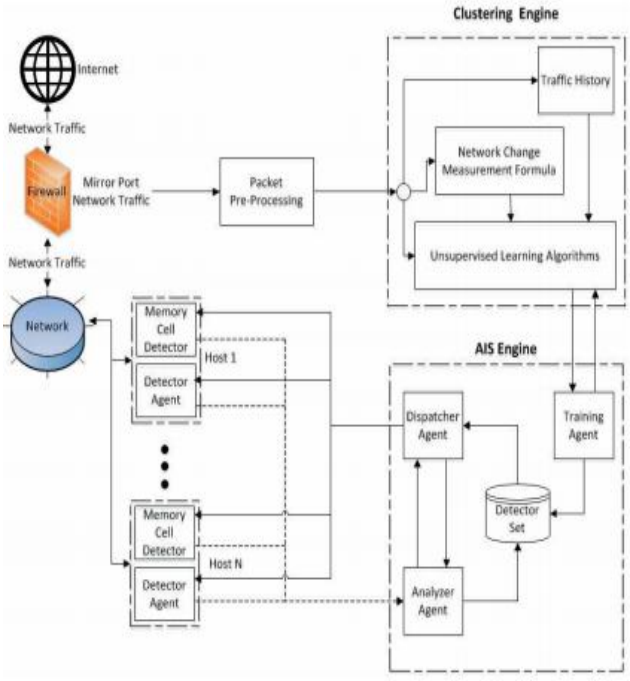

Figure 4. Proposed system architecture

A. Random Forest Algorithm

A random forest algorithm is a data mining, classification algorithm. It is an ensemble learning method for classification and regression. The random forest was developed by Leo Breiman and Adele Cutler. Random Forest grows many classification trees [22]. Random forest algorithm as follows:

1. If the number of training set is N, and the number of variable in the classifier be M.
2. The number of input variables m is selected randomly for each node of the tree. M should be much less than N.
3. Calculate the best split of this m in the training set.
4. During the forest growing, the value of m is held constant.
5. Each tree is grown to the largest extent and there is no pruning.

The offline phase uses the training dataset to build intrusion patterns used by the online phase. The pattern builder output the feature importance values used in the anomaly detection part.

B. Weighted k-means algorithm

Weighted k-means algorithm is a data mining clustering algorithm. It is a modified version of k-means algorithm by adding a weight of the data features. Algorithm as follows [16]:

1. Initially, K clusters are picked randomly as a "centroids".
2. Assigning each node to its closet centroids.
3. Calculating mean of assigned points. Relocating each centroid based on the mean value.
4. Repeat step 2and 3, until convergence will occur.

The weighted k-means algorithm requires minimum and maximum value of each data in our KDD'99 dataset for calculating the weight [22].

V. PARALLEL AHO-CORASICK ALGORITHM ON GPU

The Aho-Corasick algorithm was proposed in 1975 by Alfred V. Aho and Margaret J.Corasick [23] , and this is the most effective multi pattern matching algorithm. Aho-Corasick (AC) is a multi-string matching algorithm, meaning it matches the input against multiple strings at the same time. Multi-string matching algorithms generally pre-process the set of strings, and then search all of them together over the input text. Searching for a keyword is very efficient, because it only moves through the states in the state machine. If a character is match, goto() function is executed otherwise it follows fail() function [24].

Proposed work modifies the traditional Aho-corasick algorithm.

Algorithm:

Input: DFA state transition Table, Set of patterns {P1, P2, P3..Pn}, Input Text T.

Output: Locations where the patterns occur In T.

Begin

- Declare n thread one for each byte.
- Curent_state=0
- Pattern_length=0
- Number_of_ pattern=0
- For cursor=start_of_ string To end_of_ string
- If (DFAtable[current state][T[cursor]].next state ≠0) then
  1. If(DFAtable[current_state][T[cursor]].isFinal=0) then
  2. Current_state=DFAtable[current_state][T[Cursor]].next state
  3. Pattern_length=pattern_length +1
  4. Else

5. Match_Position=cursor-pattern length
6. Match_ state=current_state
7. Num_Pattern=num_Pattern +1
8. Else o pattern_length=0
9. Cursor_state=0

- End

Two important task in DFA matching is reading the input data and fetching next state from state table. This memory transfer can take lots of time in general memory latency is hide by using several threads in parallel .multiple thread can utilized memory by overlapping data with computation [25]. In traditional Aho-Corasick algorithm matrix is used to store state transition table. In parallel approach of Aho-Corasick algorithm, proposed algorithm use CSR representation of sparse matrix.

VI. RESULT

Performance Measure of Serial Algorithm We test the performance of serial implementation of Aho-corasick algorithm against Brute force attack. Here we use different number of pattern in length and size .Following result is tested on Intel 4 th generation coreI5 Machine time required to find the pattern in given packet. The number of samples of data set was 22545, which was sufficient for performing the evaluation and comparison between DBSCAN and K-means. Recall (REC) or True Positive Rate which estimated by dividing the correctly detected anomalies and the total number of anomalies, Precision of Anomalies (PREC) or positive predicted value which estimated by dividing the correctly classified positives by the total predicted positive count and finally the F1 score, which is the weighted average of the precision and recall. The self improvement rate in distributed mode is better than centralized mode and it reaches to its stable maximum amount after only 6 rounds while this happens after 10 rounds in centralized mode. This is because of dynamic distribution and synchronization of newly generated memory cells in each host to others.

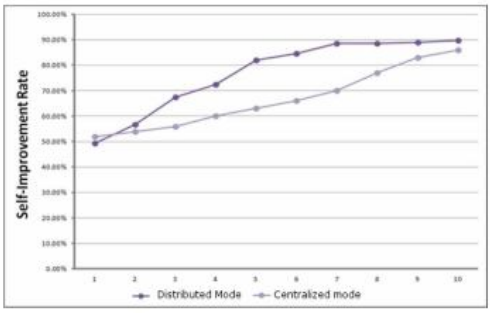

Figure 5. Comparison of self-improvement rate and centralized mode

VII. CONCLUSION AND FUTURE WORK

In misuse detection framework, intrusion patterns are built in the offline phase. The main characteristic of misuse detection techniques is in comparing network traffic against a predefined intrusion pattern in order to decide whether. Hybrid framework we used advantages of both misuse and anomaly detection, thus offering speed and accuracy to detect the intrusion. To improve the performance of clustering presented the experimental results of proposed innate immune mechanism using our network measurement formula. We also demonstrated that the distributed structure for this IDS is more efficient than the centralized mode. Aho-Corasick shows better performance as compare to other algorithm. We also tested parallel performance of Aho-Corasick algorithm using different language and we can say that Parallel implementation of Aho-Corasick is gives better performance as compare to OPEN-MP version of Aho-Corasick. Intrusion prevention system (IPS) is software that has all the capabilities of an intrusion detection system and can also attempt to stop possible incidents; this can be put forward as a Future work the current generations of IDS (HIDS and NIDS) are quite effective already; as they continue to improve they will become the backbone of the more flexible security systems. Improves the detection efficiency, and reduces the previous deviations caused by domain experts' hand writing mode.

REFERENCES

[1] Ulf lindvist, Phillip Brentano and Doug Mansur, "IDS Motivation, architecture, and An Early Prototype", Computer Security Laboratory, US Davis: 160-171
[2] L. Morgan (2014), List of Cyber Attacks and Data Breaches in 2014. IT Governance, 23 Dec.
[3] M. Watson, (2014). JP Morgan suffers data breach affecting 76 million customers. IT Governance, 23
[4] "Report and response regarding Leakage of Customers‟ personal Information." (10 September 2014). Last accessed on 17 February 2015,
[5] R.Bane, N.Shivsharan, "Network intrusion detection system (NIDS)", 2008, pp.1272-1277.
[6] S. T. Brugger, "Data mining methods for network intrusion detection", 2004, pp. 1-65.
[7] Snort Intrusion detection system.(2006). www.snort.org
[8] Reda M. Elbasiony, Elsayed A. Sallam, Tarek E. Eltobely, Mahmound M. Fahmy, "A hybrid network intrusion framework based on random forest and weighted k-means," Ain Shams Engineering Journal, 2013.
[9] S. Forrest, S. A. Hofmeyr, and A. Somayaji, "Computer Immunology," Commun. ACM, vol. 40, no. 10, pp. 88–96, Oct. 1997.

[10] S. A. Hofmeyr and S. A. Forrest, "Architecture for an Artificial Immune System," Evol. Comput., vol. 8, no. 4, pp. 443–473, Dec. 2000.
[11] S. and G. G. Feixian, "Research of Immunity-based Anomaly Intrusion Detection and Its Application for Security Evaluation of E-government Affair Systems.," JDCTA Int. J. Digit. Content Technol. its Appl., vol. 6, no. 20, pp. 429 – 437, 2012.
[12] Paul Dokas, Levent Ertoz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava and Pang-Nig Tan, "Data Mining for Network Intrusion Detection".
[13] D. Barbara, J. Couto, S. Jajodia, L. Popyack, and N. Wu, "ADAM: Detecting intrusions by data mining," in Proc. 2nd Annu. IEEE Workshop Inf. Assur.Secur., New York, Jun. 2001, pp. 11-16.
[14] A, Peterson. (5 December 2014). The Washington Post. "Why it˝s so hard to calculate the cost of the Sony Pictures hack." Last accessed on 29 January 2015,
[15] Trend Micro Incorporated, (22 December 2014). Simply Security. "The Reality of the Sony Pictures Breach." Last accessed on 29 January 2015, http://blog.trendmicro.com/reality-sony-pictures-breach/.
[16] R. Heady, G.F. Luger, A. Maccabe and M. Servilla, "The architecture of a Network Level Intrusion Detection System," Department of Computer Science, College of Engineering, University of New Mexico, 1990, pp. 1-17.
[17] R. Bace and P. Mell, "NIST Special Publication on Intrusion Detection Systems," Booz-Allen and Hamilton inc, Mclean VA, 2001, pp. 5-22.
[18] R.A. Kemmerer and G. Vigna, "Intrusion Detection : A brief History and Overview," Computer, 2002 [supplement to security and privacy magazine], pp. 27-30.
[19] Z. Zhou, Y. Xue, J. Liu, W. Zhang, and J. Li. MDH: A High Speed Multi-phase Dynamic Hash String Matching Algorithm for Large-Scale Pattern Set. Information and Communications Security,4861:201– 215, 2007 .
[20] Snort. Webpage containing information on the snort intrusion prevention and detection system.
[21] S. Dori and G.M. Landau. Construction of Aho Corasick Automaton in Linear Time for Integer Alphabets. Information Processing Letters, 98(2):66–72, 2006.
[22] CUDA Zone. Official webpage of the nvidia cuda api.
[23] Ken Thompson," Programming Techniques: Regular expression search algorithm", June 1968
[24] John E. Hopcroft and Jeffrey D. Ullman, Introduction to Automata Theory, Languages, and Computation, Addison-Wesley Publishing, Reading Massachusetts, 1979.
[25] M. O. Rabin and D. Scott. Finite automata and their decision problems IBM Journal of Research and Development, 3(2):114–125, 1959.