Application level load balancing in a web switch or security oriented filtering based on the content signatures for deep packet inspection and delayed input DFA

K.Madhuri, M.Suman, K.Rajasekhar Rao

Abstract— There is a growing demand for network devices capable of examining the content of the Data packet in order to improve the network security and provide application specific services. Most high performance systems that perform deep packet inspection implement simple string matching algorithms to match packets against a large but finite set of strings .However there is a growing interest in using regular expressions based pattern matching. Since regular expressions offer Superior expressive power and flexibility. DFA representations are typically used to represent the regular expressions. However the DFA representation of regular expressions sets arising in network applications requires large amount of memory, limiting their practical application. In this paper I introduce a new representation for regular expressions called delayed input DFA (D2FA) which substantially reduces the space requirement as compared to a DFA. A D2FA is constructed by transforming a DFA via incrementally replacing several transmissions of the automata with a single default transmission. This approach dramatically reduces the number of transmissions in between the states. Given the substantially reduced space requirements, we can describe an efficient architecture that can perform deep packet inspection at multi gigabit rate.

Index Terms—DFA, Regular expressions, deep packet inspection, Network Intrusion Detection, Security, NFA and Delayed input DFA called D2FA.

1 Introduction

MANY critical network services handle packets based on payload content, In addition to the structure information found in the packet headers. Forwarding packet based on content (either for the purpose of application lever load balancing in a web switch or security oriented filtering based on the content signatures.) requires new level of support in networking equipment. To see why, we must consider how regular expressions are implemented. A regular expression is typically represented by a deterministic finite automaton. For any regular expression it is possible to construct a DFA with minimum number of states. The memory required to represent the DFA is intern determined by the product of the number of states and the number of transmissions from each state for ASCII alpha bet each state will have 256 out going transmissions. In this paper I introduce a highly compact DFA Representation , Our approach reduces the number of transmissions associated with each state the main observation is that groups of states in DFA often have identical outgoing transmissions and we can use this duplicate information to reduce memory requirement. For example suppose there are two states s1 and s2 that make transmissions to the same set of states { S } , for some set of input characters , { C } We can eliminate these transmissions from one state say s1 by introducing a default transmission from s1 to s2 that is followed for all the characters in{ C } Essentially s1 now only maintains unique next states for those transmissions not common to s1 and s2. And uses the default transmission to s2 for common transmissions. We refer to a DFA with such default transmissions as delayed input DFA (D2FA).

In practice the proper and effective construction of default transmission leads to the trade of between the size of the DFA representation and the memory bandwidth required to traverse it [1]. In a standard DFA the input character leads to a single transmission between the states. In a D2FA an input character lead to multiple default transmissions before it is consumed along a normal transmission.

2 Background and Related Work

Deep packet inspection has recently gained popularity as it provides the capability to accurately classify and control traffic in terms of content, applications, and individual subscribers. Some important applications require the

- *K.Madhuri is currently working as a Assistant Prof. in K L University, India, PH-+91-9885928333. E-mail: madhuriecm@kluniversity.in*
- *M.Suman is currently working as a Associate Prof. in K L University, India, PH-+91-9848187437. E-mail: suman.maloji@gmail.com*
- *K.Rajasekhar Rao is currently working as a Principal & Dean-Student Welfare in K L University, India, E-mail:krr_it@yahoo.co.in*

deep packet inspection are listed below.

- Network intrusion detection and system and prevention (NIDS/NIPS) generally scan the packet header and payload in order to identify a given set of signatures of well known security threats.
- Layer7 switches and firewalls provide content based filtering load balancing, authentication and monitoring.Application-aware web switches, for example provide scalable and transparent load balancing in data centers.
- Content based traffic management and routing can be used to differentiate traffic classes based on type of data in a packet.

Deep packet inspection often involves scanning every byte of the packet payload and identifying a set of matching predefined patterns. Traditionally rules have been represented as exact match strings consisting of known patterns of interest. Naturally due to their wide adoption and importance several high speed and efficient string matching algorithms have been proposed recently. Many researchers have been proposed high speed pattern matching hard ware architectures.

3 Delayed input DFA's

It is well known that for any regular expression set there will be a DFA with minimum number of states the memory needed to represent a DFA is determined by the number of transmissions from one state to another or equivalently the number of edges in the graph representation We introduce the modification to the DFA that can be represented much more compactly. Our modification is based on the techniques used in the string matching algorithm [2]. I extend their technique and apply it to DFA s obtained from regular expressions, rather than a simple string tests. As demonstrated in this document, the numbering for sections upper case Arabic numerals, then upper case Arabic numerals, separated by periods. Initial paragraphs after the section title are not indented. Only the initial, introductory paragraph has a drop cap.

4 Bounding default paths

If our only objective was minimizing the space requirement of the D2FA, it would suffice to find maximum weight spanning tree in the space reduction graph. The tree edges corresponding to the state pairs between which we create default transmissions.

Unfortunately when this procedure is applied to the DFA's arising in typical network applications, the resulting default transmission tree has many long paths implying that the D2FA may need to make many transmissions for each input character consumed[3]. We can improve the performance some what by selecting a tree root that is centrally located with in the spanning tree. However this still leaves us with many long default paths, the natural way to avoid long default paths is to construct a maximum weight spanning tree with a specified bounded diameter. Fortunately we have found that fairly simple methods based on classical maximum spanning tree algorithms, yield good result for D2FA construction.

5 Results on some regular expression sets

In order to evaluate the space reduction achieved by the delayed input DFA. We can perform an experiment on the regular expression sets used in a variety of network applications. One most important dataset are the regular expression sets used in CISCO systems this set contains more than 750 moderately complex expressions which are used to detect the anomalies in the traffic. It is widely used across several Cisco security appliances and Cisco commonly employs general purpose processor with gigabyte.

6 Conclusion

In this paper I introduce a new representation of the regular expressions called the delayed input DFA which significantly reduces the space requirement of the DFA by replacing its multiple transmissions with a singly default transmission. By reduction we can show that construction of d2fa from the DFA is NP hard .We can therefore present heuristics for D2FA construction that provide deterministic performance guarantees. Some of the results suggest that a D2FA constructed from the DFA can reduce memory space requirements by more than 95 percent.

Acknowledgment

I would like to thank Sailesh Kumar, Department of CSE, Washington University for hosting this research and making the available data.

References

J.E hopcraft and J.D Ullman introduction to Automata theory and computation.

S. Antonatos Generating realistic work load for network intrusion detection systems.

Fang Yu et all Fast and memory efficient regular expressions

S. Antonatos, et. al, "Generating realistic workloads for network intrusion detection systems," In ACM Workshop on Software and Performance, 2004.

A. V. Aho and M. J. Corasick, "Efficient string matching: An aid to bibliographic search," Comm. of the ACM, 18(6):333–340, 1975.

B. Commentz-Walter, "A string matching algorithm fast on the average," Proc. of ICALP, pages 118–132, July 1979.

S. Wu, U. Manber," A fast algorithm for multi-pattern searching," Tech. R. TR-94-17, Dept. of Comp. Science, Univ of Arizona, 1994.

Fang Yu, et al., "Fast and Memory-Efficient Regular Expression Matching for Deep Packet Inspection", UCB tech. report, EECS-2005-

8.

N. Tuck, T. Sherwood, B. Calder, and G. Varghese, “Deterministic memory-efficient string matching algorithms for intrusion detection,” IEEE Infocom 2004, pp. 333--340.