Synopsis Report

On

A Semi Custom ASIC for Circumventing SPAM

By

Mr. Santosh A. Shinde

Under the Guidance of
Dr. R.K. Kamat

Department of Electronics
Shivaji University,
Kolhapur

July 2007

Synopsis Report

1. Subject	: Electronics
2. Name of the Student	: Mr. Santosh A. Shinde
Qualification	: M.Sc.
Address	: Department of Electronics, Shivaji University, Kolhapur
3. Research Guide	: Dr. R. K. Kamat
Qualification	: M.Sc. Ph.D.
Address	: Reader in Electronics, Department of Electronics, Shivaji University, Kolhapur – 416 004
4. Title of Proposed Thesis	: A Semi Custom ASIC for Circumventing SPAM

Introduction:

In today's networked world where email and internet access has become the main means of communication, the word spam is encountered quite often. Spam is referred to as flooding the Internet with many copies of the same message, in an attempt to force the message on people who generally would not otherwise choose to receive it. There is some debate about the source of the term, but the generally accepted version is that it comes from the Monty Python song, "Spam spam spam spam, spam spam spam spam, lovely spam, wonderful spam…" Like the song, spam is an endless repetition of worthless text. Another school of thought maintains that it comes from the computer group lab at the University of Southern California who gave it the name because it has many of the same characteristics as the lunchmeat Spam as given below:

- Nobody wants it or ever asks for it.
- No one ever eats it; it is the first item to be pushed to the side when eating the entree.
- Sometimes it is actually tasty, like 1% of junk mail that is really useful to some people.

In 1998, the New Oxford Dictionary of English, which had previously only defined "spam" in relation to the trademarked food product, added a second definition to its entry for "spam": "Irrelevant or inappropriate messages sent on the Internet to a large number of newsgroups or users."[4]

Generally the word SPAM refers to the more well-known and common form i.e. e-mail spam. However, there exists other types of spam in a variety of Internet communication mediums such as instant messaging, discussion boards, mobile phones with text messaging, newsgroups, Internet telephony, blogs etc. — wherein basically any device or client that provides a means for communications. In the present work the focus is to circumvent email SPAM that emanates from particular IPs.

Historical Aspects:

It is worthwhile here to review the historical aspects of the SPAM. Einar Stefferud, a longtime net hand, reports that DEC announced a new DEC-20 machine in 1978 by sending an invite to all ARPANET addresses on the west coast, using the ARPANET directory, inviting people to receptions in California. They were chastised for breaking the ARPANET appropriate use policy, and a notice was sent out reminding others of the rule. This is regarded as the first SPAM message sent over the internet.[1] Tom Van Vleck, co-author of the CTSS MAIL command, reports an even earlier spam sent on MIT's Compatible Time Sharing System (CTSS) as far back as 1971. A system administrator named Peter Bos used CTSS MAIL to send everybody the anti-war message that: "THERE IS NO WAY TO PEACE. PEACE IS THE WAY." He reports the spammer defended it by saying, "but this is important." He was also an authorized admin, so this one is somewhat harder to classify. A great history of early mail systems provides more details in some of the references [2].

Even much before 1978, the first known instance of unsolicited commercial communication being via telegram on September 13, 1904, however the term "spam" for this practice had not yet been applied. Later in the 1980s the term was adopted to

describe certain abusive users who frequented BBSs and MUDs, who would repeat "SPAM" a huge number of times to scroll other users' text off the screen.[3]

Commercial spamming started in force on March 5, 1994, when a pair of lawyers, Laurence Canter and Martha Siegel, began using bulk Usenet posting to advertise immigration law services. The incident was commonly termed the "Green Card spam", after the subject line of the postings. The two went on to widely promote spamming of both Usenet and e-mail as a new means of advertisement—over the objections of Internet users they labeled "anti-commerce radicals." Within a few years, the focus of spamming (and antispam efforts) moved chiefly to e-mail, where it remains today.[1]

Spam, Need of investigation:

The huge increase in email spam in recent years is beginning to take its toll on the online world. Some email users say they are using electronic mail less now because of spam. More people are reporting they trust the online environment less[5]. Increasing numbers are saying that they fear they cannot retrieve the emails they need because of the flood of spam. They also worry that their important emails to others are not being read or received because the recipients' filters might screen them out or the emails might get lost in the rising tide of junk filling people's inboxes.
Data from a national survey[5] suggests that spam is beginning to undermine the integrity of email and to degrade the online experience. In large numbers, Internet users report that they trust email less and some even use email less because of spam. Users now worry that the growing volume of spam is getting in the way of their ability to reliably send and receive email. It is uncontrollably clutters their inboxes and imposes uninvited, deceptive, and often disgustingly offensive messages. Some of the key figures in this regard are as follows:

- 25% of email users say the ever-increasing volume of spam has reduced their overall
- use of email; 60% of that group says spam has reduced their email use in a big way.
- 52% of email users say spam has made them less trusting of email in general.
- 70% of email users say spam has made being online unpleasant or annoying.
- 30% of email users are concerned that their filtering devices may block incoming
- email.
- 23% of email users are concerned that their emails to others may be blocked by
- filtering devices.
- 75% of email users are bothered that they can't stop the flow of spam.
- 80% of email users are bothered by deceptive or dishonest content of spam.
- 76% of email users are bothered by offensive or obscene content of spam.

SPAM not only troubles the users with unsolicited emails, but also chocks the internet bandwidth considerably making the life miserable for the system administrators.

The above facts and figures justify the need to take up the investigation of the present research problem.

Research Problem:

- **Hypothesis:**

"The semicustom ASIC to be developed in the course of research work will effectively filter the SPAM owing to the dynamic reconfiguration attribute of the state of the art FPGAs on fly."

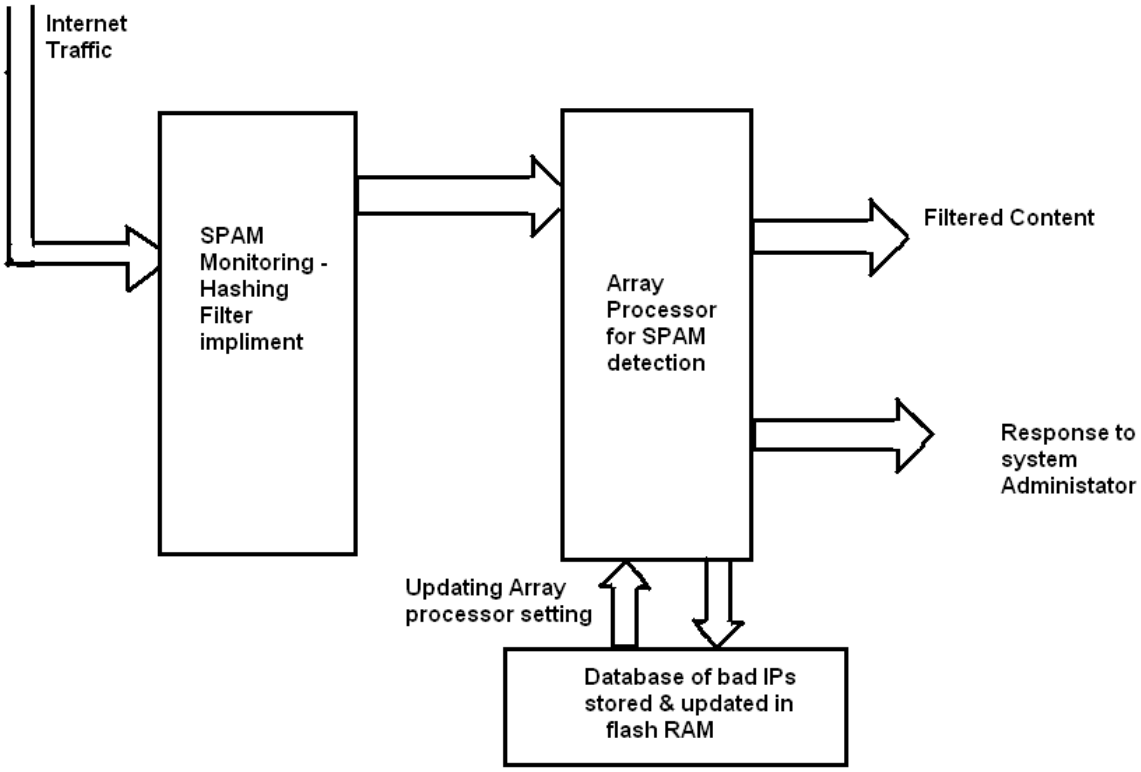
- **Problem Statement:**

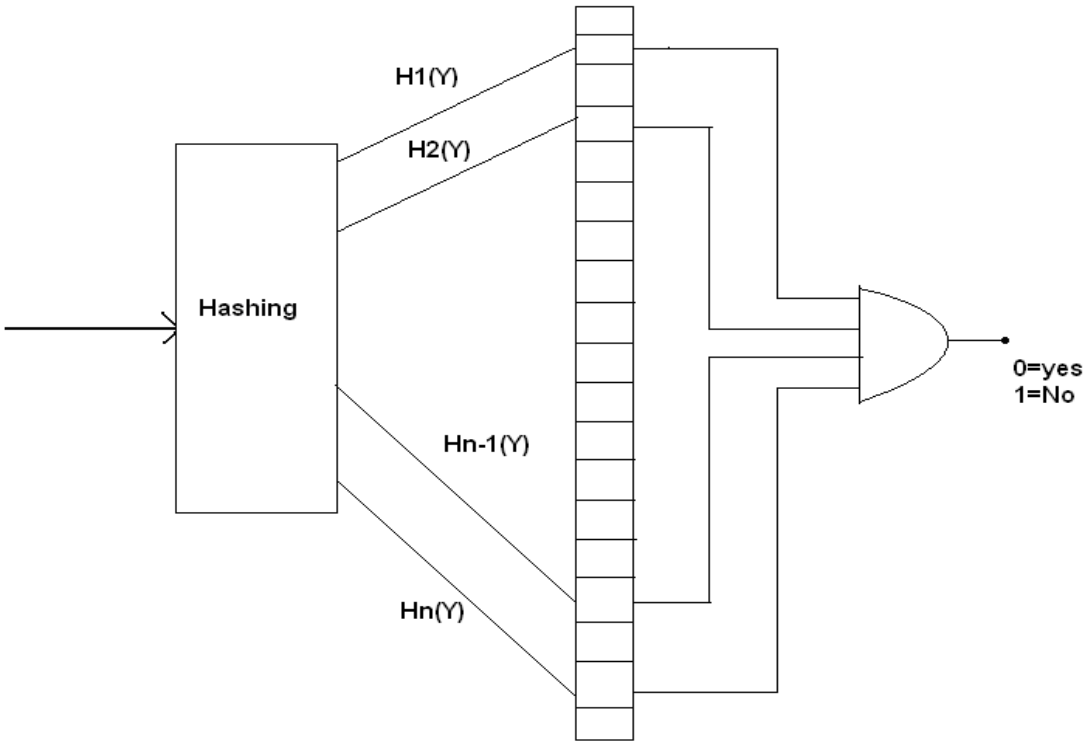
The research work comprises of development of FPGA based ASIC with dynamic reconfiguration on fly to filter out the SPAM on the basis of the detected IP address of the spammer.

- **Elaboration of problem statement:**

In the proposed work, it is planned to design a FPGA based inbound mail processor module which will handle incoming mail from the Internet to perform most of the anti-spam processing. The proposed processor will be based on an distributed architecture with each element having memory and few registers to pass the parameters to the neighboring element. The array of processing elements will work on broadcast methodology to shift the values in the main network registers to the left, right, up or down to the neighboring elements. The local ALU implemented at each element will take care of the execution of the instruction passed on by the central processor. The inherent reconfiguration feature of the FPGAs will ensure the effective anti spam filter implementation based on a multi grid level implementation. The learning algorithm will be implemented using hashing filter technique supported by the flash ROM based database to categorize the incoming IPs into accepted and rejected category.

The block diagram of the proposed implementation and the filtering technique is shown in figures 1 and 2.


Internet
Traffic
SPAM
Monitoring -
Hashing
Filter
impliment
Array
Processor
for SPAM
detection
Filtered Content
Response to
system
Administator
Updating Array
processor setting
Database of bad IPs
stored & updated in
flash RAM


H1(Y)
H2(Y)
Hashing
Hn-1(Y)
Hn(Y)
0=yes
1=No

Significance of research work:

The research work attempts to tackle the spam problem by design and implementation of “Soft Hardware”. The need for such a FPGA based soft IP core implementation is summarized in the following paragraph.

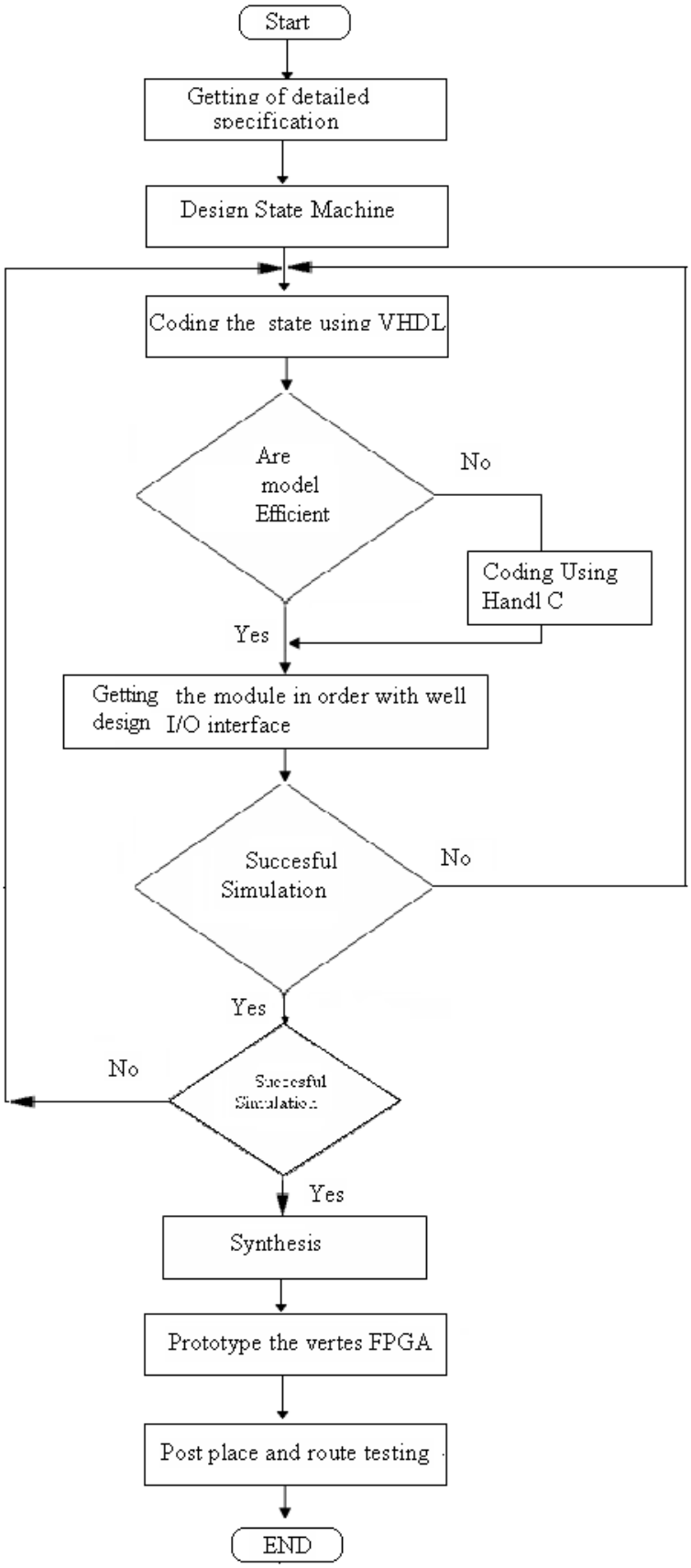
The computer networks today are looking for two major improvements which are interrelated to each other viz. speed and security. It is a common observation that if the later is tried to strengthen, the former degrades. Therefore, these days the networks use a layered approach with scanning at both the desktop, and the gateway using a security appliance. The problem of SPAM adds one more dimension to the above mentioned specifications. With more SPAM, the network bandwidth gets chocked inturn decreasing the speed. While introducing the spam filters and follwing the pattern matching techniques on the server or client itself leads to a bottleneck in speed on the client or server itself. A wayout is development of a standalone device at the gateway with specialized content processors to which the scanning task be offloaded to remove the speed bottleneck at the server or client end.

The dynamic reconfiguration property of the FPGAs, makes it ideal for content security with constant updation of the bad IP or contect of the SPAM which regularly alters its signature. Even as new content types, new attacks and new protocols become critical, embedded processors based on FPGAs can download new firmware to remain relevant. It is this ability, combined with the high-performance available in the latest generation of FPGAs that make them the best choice for content processing. Furthermore, being able to re-flash the firmware and hardware over the Internet gives appliance vendors additional revenue streams for the same product, making for a very compelling business case.

Methodology:

The hardware used for the proposed research work will be based on Virtex FPGA of Xilinix Inc. The software will be developed in webpack as well as some modules in Handel C. The prototyping of the system will be done by using Picoblaze an8-bit microcontroller developed by Xilinx to be used on their devices. The main advantage of using Picoblaze is its highly optimized for their products and requires only a small amount of resources. It is easy to use and the documentation is extensive and well written. It is powerful in comparison to the resources that it uses on the FPGA, it uses 85 slices and a BlockRAM and is capable of 40-70 MIPS.

The research work will commence with development of the state space design with detailed listing of the activities and updates in each state. Then it proceeds on the lines as shown in the following flowchart. The implementation will explore the possibility of the configuring PicoBlaze as a small web server with built in protocols for filtering the IP.


Start
Getting of detailed specification
Design State Machine
Coding the state using VHDL
Are model Efficient
No
Coding Using Handl C
Yes
Getting the module in order with well design I/O interface
Succesful Simulation
No
Yes
No
Succesful Simulation
Yes
Synthesis
Prototype the vertes FPGA
Post place and route testing
END

Literature Survey:

The fight against spam is being waged on two fronts, legal and technological. We hear from time to time about small claims and spectacular victories in the courtroom, but we believe--as do a majority of our antispam poll respondents--that legislative efforts alone will not eliminate spam[6].

As far as technologicy is concerned, the most popular way to circumvent the spam so far is “antispam filters”. Following types of spam filters are being used:

- Integrated, internet-based spam filters
- Integrated, algorithmic spam filters
- Proxy spam filters
- Server-side spam filters

Internet-based spam filters[7] are services provided by a third party. They store the virus definitions on their services, and the client that the user runs checks incoming mail against this database. Commonly, this database is built and refined by the users themselves. If a spam email is missed by the plugin, the user can tell the plugin that it is spam, and this will be reported back, and the spam definitions will be further refined. There advantages and disadvantages are as follows:

Pros:

- Works out of the box
- Challenges for senders
- Internal spam definitions as well as server-side definitions

Cons:

- Not free
- Some users report agitation from mail senders who don't want to deal with the challenges.

Algorithmic spam filtration is filtration that does not check email against a remote server. Rather, programs have to be "trained" to recognize what is spam and what is not. The volume of email that one receives is the primary factor in determining how quickly the plugin "learns." Some plugins learn faster than others, but it is primarily the amount of email that the plugin has to build a database of good and bad with that determines this.

The database that the plugin checks email against is the most important part. An advantage to these filters is that the user defines precisely what is spam and what is not. What one labels as spam is only spam to you, not someone else: there is no danger in marking that email from that listserve as spam, because it's not going to affect anyone else. The obvious disadvantage is that one has to build a database to begin with, and that means that there are going to be false positives, and lots of false negatives at first. With time and effort, the numbers for both will gradually decline. The most common sort of spam filtering is Bayesian sorting. Bayesian sorting is useful for more than just spam filtering. It can be used to sort mail into specific categories like work, personal, and spam. Most Bayesian software, however, only uses it to sort what is spam and what is

not. Roughly speaking, Bayesian sorting uses statistical analysis to see which words appear as spam, and which words do not, and it uses this information to build a composite score for a particular email. In theory, the larger the database gets, the more accurate it should be. Many people consider algorithmic sorting that is not based on the Bayesian method to be inferior.

Proxy spam filters are set up on a local machine, though others, such as POPFile can be set up on a remote machine, making them somewhat more manageable to deal with if you're often using e-mail from different locations. Some people dislike having two programs running when one would suffice. The flexibility allowed by proxy setups is one of the biggest draws, however, so it boils down to a matter of preference. One other draw is that users of Outlook Express can also play, whereas with many of the other plugin methods, they cannot. Server-side spam filters is obviously that software which resides on the actual mail server. What is does with incoming spam is dependent upon whether the server simply trashes incoming spam, or if it simply tags that mail as potential spam, and then allows the user to deal with it as he or she sees fit.

Spammers use questionable search engine optimization (SEO) techniques to promote their spam links into top search results. In this paper, we focus on one prevalent type of spam – redirection spam – where one can identify spam pages by the third-party domains that these pages redirect traffic to. We propose a fivelayer, double-funnel model for describing end-to-end redirection spam, present a methodology for analyzing the layers, and identify prominent domains on each layer using two sets of commercial keywords – one targeting spammers and the other targeting advertisers. The methodology and findings are useful for search engines to strengthen their ranking algorithms against spam, for legitimate website owners to locate and remove spam doorway pages, and for legitimate advertisers to identify unscrupulous syndicators who serve ads on spam pages.

Today's crucial information networks are vulnerable to fast moving attacks by Internet worms and computer viruses. These attacks have the potential to cripple the Internet and compromise the integrity of the data on the end-user machines. Without new types of protection, the Internet remains susceptible to the assault of increasingly aggressive attacks. A platform has been implemented that actively detects and blocks worms and viruses at multi-Gigabit/second rates. It uses the Field-programmable Port Extender (FPX) to scan for signatures of malicious software (malware) carried in packet payloads. Dynamically reconfigurable Field Programmable Gate Array (FPGA) logic tracks the state of Internet flows and searches for regular expressions and fixedstrings that appear in the content of packets. Protection is achieved by the incremental deployment of systems throughout the Internet.[20]

The security of the Internet can be improved using Programmable Logic Devices (PLDs). A platform has been implemented that actively scans and filters Internet traffic for Internet worms and viruses at multi-Gigabit/second rates using the Field-programmable Port Extender (FPX). Modular components implemented with Field Programmable Gate Array (FPGA) logic on the FPX process packet headers and scan for signatures of malicious software (malware) carried in packet payloads. FPGA logic is used to implement circuits that track the state of Internet flows and search for regular expressions and fixed-strings that appear in the content of packets. The FPX contains logic that allows modules to be dynamically reconfigured to scan for new signatures.

Network-wide protection is achieved by the deployment of multiple systems throughout the Internet[21].

An extensible firewall has been implemented that performs packet filtering, content scanning, and queuing of Internet packets at Gigabit/second rates. The firewall uses layered protocol wrappers to parse the content of Internet data. Packet payloads are scanned for keywords using parallel regular expression matching circuits. Packet headers are compared to rules specialized in Ternary Content Addressable Memories (TCAMs). All packet processing operations were implemented with reconfigurable hardware and .t within a single Xilinx Virtex XCV2000E Field Programmable Gate Array (FPGA). The singlechip firewall has been used to filter Internet SPAM and to guard against several types of network intrusion. Additional features were implemented in extensible hardware modules deployed using run-time reconfiguration[22].

A module has been implemented in Field Programmable Gate Array (FPGA) hardware that is able to perform regular expression search-and-replace operations on the content of Internet packets at Gigabit/ second rates. All of the packet processing operations are performed using reconfigurable hardware within a single Xilinx Virtex XCV2000E FPGA. A set of layered protocol wrappers is used to parse the headers and payloads of packets for Internet protocol data. A content matching server automatically generates, compiles, synthesizes, and programs the module into the Field-programmable Port Extender (FPX) platform[23].

Sr.No	Technique	Principle	Remark
1	Content Based filter	Matching the content to some words previously stored. It is installed at the receiver end . There are two types filter tools such as SpamAssassin, latest Bayesian algorithms	They falsely identify real mail as spam, and block it
2.	Blacklists	There are many competing blacklists of IPs , some of strong ethics, others more dubious. The mail from that IP is blocked	
3.	Collaborative filters	These filters, such as Vipul's Razor (now via CloudMark) rely on the first poor souls who get a spam reporting it to a central server. As the reports come in, the spam can be identified and rules can be written to block it	These are reasonably effective, and go after bulk, which is good. They have fewer false positives if done well
4	Spamtrap Filters	These are primarily used by BrightMail Inc., which is probably the largest commercial anti-spam operation. Brightmail maintains huge numbers of addresses seeded onto spammer lists. When messages arrive, they are almost surely spam, and human beings look at them to derive rules to filter out and retroactively delete the messages	Very few false positives, but unfortunately reportedly only about 60-70% effective.
5.	Hiding your address	The most common technique today seems to be hiding your E-mail address so that it can't be harvested by spammers.	Unfortunately, by using dictionary attacks, they are managing to spam people who have never exposed their E-mail in public.This is not a good solution.

References:

1. “Origin of the term "spam" to mean net abuse,” by http://www.templetons.com/brad/spamterm.html Retrieved on May 28, 2007.

2. “The History of Electronic Mail” By Tom Van Vleck: http://www.multicians.org/thvv/mail-history.html Retrieved on May 28, 2007.

3. “Spam (electronic) From Wikipedia, the free encyclopedia”, http://en.wikipedia.org/wiki/Spam_(electronic) Retrieved on May 31, 2007

4. “Oxford dictionary adds Net terms: The latest New Oxford Dictionary of English includes definitions for a slew of technology words and phrases”, By Rose Aguilar, Staff Writer, CNET News.com, http://news.com.com/2100-1023-214535.html Retrieved on May 31, 2007.

5. “Spam: How It Is Hurting Email and Degrading Life on the Internet”, Deborah Fallows, Senior Research Fellow, October 22 2003, http://www.pewinternet.org/pdfs/PIP_Spam_Report.pdf Retrieved on June 1, 2007

6. “AntiSpam Techniques Sick of SPAM”, May 13, 2004 - By Ron Anderson, http://www.networkcomputing.com/channels/security/showArticle.jhtml?articleID=20000059 Retrieved on June 1, 2007

7. Ask Ars: The best anti-spam solutions for Windows, by Rian "hanser" Stockbower, http://arstechnica.com/ask-ars/2003/anti-spam/page1.html

8. “Inside the SPAM Cartel, Trade Secrets from the Dark Side”, Jeffrey Posluns, Syngress Publications

9. “Filtering SPAM for Dummies, By Margaret Levine Young, Ray Everett-Church, John R. Levine, Published 2004, Wiley Inc.

10. “Degunking Your Email, Spam, and Viruses”, By Jeff Duntemann, Published 2004, Paraglyph Press

11. “Spam Wars: our last best chance to defeat spammers, scammers, and hackers”, By Danny Goodman, Published 2004, SelectBooks, Inc.

12. “Stopping Spam” By Simson L. Garfinkel, Alan Schwartz, Published 1998, O'Reilly

13. B. Leiba and N. Borenstein, A Multifaceted Approach to Spam Reduction, In Proceedings of the First Conference on Email and Anti-Spam, July, 2004.

14. R. Segal, J. Crawford, J. Kephart and B. Leiba, SpamGuru: An Enterprise Anti-Spam Filtering System. In Proceedings of the First Conference on Email and Anti-Spam, July, 2004.

15. I. Rigoutsos and T. Huynh, Chung-Kwei: a Pattern-discovery-based System for the Automatic Identification of Unsolicited E-mail Messages (SPAM). In Proceedings of the First Conference on Email and Anti-Spam, July, 2004.

16. M. Leonard, M. Rodriguez, R. Segal and R. Shoop, Managing Customer Opt-Outs in a Complex Global Environment, In Proceedings of the First Conference on Email and Anti-Spam, July, 2004.

17. P. Capek, B. Leiba, and M. N. Wegman. Charity Begins at your Mail Server, http://www.research.ibm.com/people/w/wegman/charity.htm, 2004.

18. S. Fahlman, Selling interrupt rights: A way to control unwanted e-mail and telephone calls. IBM Systems Journal, Volume 41, Number 4, pp. 759-766, 2002.

19. "Spam Double-Funnel: Connecting Web Spammers with Advertisers", http://research.microsoft.com/SearchRanger, by Yi-Min Wang, Ming Ma Yuan Niu, Hao Chen

20. Application of Hardware Accelerated Extensible Network Nodes for Internet Worm and Virus Protection, by John W. Lockwood, James Moscola, David Reddick, Matthew Kulig, and Tim Brooks, International Working Conference on Active Networks (IWAN), Kyoto, Japan, December, 2003.

21. Internet Worm and Virus Protection in Dynamically Reconfigurable Hardware; by John W. Lockwood, James Moscola, Matthew Kulig, David Reddick, Tim Brooks, Military and Aerospace Programmable Logic Device (MAPLD), Washington DC, 2003, Paper E10, Sep 9-11, 2003.

22. An Extensible, System-On-Programmable-Chip, Content-Aware Internet Firewall, by John W. Lockwood, Christopher Neely, Christopher Zuver, James Moscola, Sarang Dharmapurikar, and David Lim; Field Programmable Logic and Applications (FPL), Lisbon, Portugal, Paper 14B, Sep 1-3, 2003.

23. Implementation of a Streaming Content Search-and-Replace Module for an Internet Firewall, by James Moscola, Michael Pachos, John W. Lockwood, Ron P. Loui; Hot Interconnects 11 (HotI), Stanford, CA, USA, pp. 122-129, Aug. 2003.