

## **High Performance Apache SpamAssassin**

- **What do I mean by High Performance**

*I'm sure that when you first saw the title for this presentation you had in your mind, how do I make Apache SpamAssassin run faster. In deed that is a small piece of the puzzle and probably what I will spend most of our time today talking about. However, there are other aspects of making your SpamAssassin install into a high performance spam fighting engine and we will be covering those as well.*

- **Make it go faster**

*As I said, one of the tenents of having a high performance SpamAssassin install is pure speed, how fast can you filter your incoming message stream. For some this is key, with some installs 90+ percent of their incoming mail is spam, that means that the remaining 10 percent of good mail must wait in line. The faster we can be the sooner the good stuff can make it into your customers inboxes.*

- **Make it perform better**

*Once we have things running along at full speed we can spend a little time working on maing SpamAssassin catch more spam. In some cases we'll need the tips and tricks in the Making it Faster topic in order to run more things. We'll talk a little bit about running additional rules, making those rules work better and maybe even touch a little bit on generating your own scores.*

- **Make it easier to maintain**

*I hesitate to say final, because its no where near, so maybe the last piece of the puzzle we'll talk about today is making your SpamAssassin install easier to maintain. There are lots of different ways to accomplish this, some good, some bad. I imagine there are nearly as many ways as there are current installs of SpamAssassin. The key here is to find something that works well and use it, hopefully its just tweakable enough to do what you want with minimal hassle.*

- **Making Apache SpamAssassin Run Faster**

- **spamassassin vs. spamd/spamc**

*Bottom line, if you are attempting to setup a high performance SpamAssassin installation, don't use the spamassassin script. Take time, do it right. The process of integrating spamd into just about any setup can easily be found on the web.*

- **Hardware/Resources**

*The lynch pin in any SpamAssassin insallation is going to be the underlying hardware. SpamAssassin takes a lot of resources. What sort of hardware do you need? That really isn't an easy question to answer. The more power you have the more options you have.*

- **Turning off Network Checks**

*A fast way to gain some cycles for processing mail is turning off network checks. Compared to everything else, network checks take forever. Recent releases have done great things to help resolve these issues, timeouts and what not, but still it can take time. The downside is that recent releases have come to lean on these network checks in such a way that folks not running network checks will see more spam than*

## High Performance Apache SpamAssassin

*those who do not.*

*One thing you can do is push the network checks out of SpamAssassin and into the MTA. This lets you drop a lot of mail on the floor and make it so SpamAssassin never sees it.*

*DCC/Razor/Pyzor - These are systems that when they work they can really help, in some cases it might provide just enough to push a message over the threshold.*

*However, you pay the price, these systems are expensive time wise.*

- **Bayes, use it? If so what is the best way?**

*One question you'll want to answer is if you want to use Bayesian filtering or not. I love bayesian filtering, it does wonders for my mail stream, but its not for everyone. I encourage you to investigate and figure out if it will work well for you in your environment.*

- **Individual vs Site-Wide**

*Part of answering if it will work well for you is determining if you want to use individual or site-wide bayes filtering. Here is the difference, individual filtering means that each person who receives mail on your system has their own bayes database. That means that it is customized for their mail stream. Similarly site-wide bayes filtering means that the bayes database is shared by everyone. There are multiple schools of thought on this and we could probably go round and round all day trying to figure out what is best. Here is a simple formula. In order to have the maximum effectiveness, the bayes algorithm needs to act on a representative message stream. Now, while there is some "custom" spam in the world for the most part spam that gets sent to one person is most likely to be seen by many many others, so remove that from the equation. Non-spam, or ham messages are then the key. If you have a group of people that receive similar types of ham, for instance a corporate mail server, then site-wide filtering will probably work well for you. On the other hand, if your user base gets a vastly different types of ham, for instance a regional ISP that serves EVERYONE, then individual bayes filtering will work better. Like I said, there are other aspects and what works well for some doesn't always work well for others, do not be afraid to experiment and find what works best for you.*

- **DBD/File vs SQL**

*Ok, now that you've done the right thing and decided that you for sure want to use bayes filtering there is another consideration. Do you do a local DBD or SDBM file based database or go whole hog and put that sucker on a nicely tuned SQL server.*

*There are benefits and tradeoffs for each.*

- 1) Speed
- 2) Locking Issues
- 3) Ability to span multiple instances

## **High Performance Apache SpamAssassin**

- □ **Turning off what you don't need**

*Here is a quick way to help performance. If you don't need a feature, just turn it off. AutoWhitelist is a good example of this. Due to history, AutoWhitelist is on by default for most installations. It can be very handy, but for some it either a) doesn't work quite right or b) doesn't provide as much bang for buck as other solutions. Turn it off.*

*Same can be said for other pieces of SpamAssassin.*

*Bayes, Language Detection, HashCash, DCC/Razor/Pyzor*

- □ **Making Apache SpamAssassin Perform Better**

- □ **FN vs FP Rates**

*What is an FN or FP Rate?*

*FN is False Negative, or basically a spam message that was not caught by SpamAssassin.*

*FP is False Positive, or basically a non-spam/ham message that WAS caught by SpamAssassin.*

*The key here is that everyone views their acceptable FN/FP rate differently. For instance, in a business or corporate environment your allowed FP rate might be very low, after all who wants to miss that critical business opportunity. The converse might be someone like an ISP who is serving a mass of mom and pop users who don't really care if the latest chain letter or the email with the top ten lists of ways to annoy your neighbor get dropped on occasion.*

*The key point is that knowing this information you can then tune your installation the right way to service your users properly.*

- □ **Third-Party Rulesets**

*SpamAssassin has a long release cycle, that means that large batches of new rules do not get added or updated very often. That doesn't mean you can't or shouldn't add rules on your own.*

*Recent versions of SpamAssassin and for sure the next major version, 3.2, there will be a new mechanism for updating rules between releases.*

*Appropriately named sa-update, this script will query a central server, download new rules if they exist and install them for you. You can then restart your server and be up and running on new rules.*

*There are a couple of different projects that generate sets of rules. Check on the wiki for more information on various third party rulesets.*

- □ **Custom Rules/Plugins**

*Something else you can do, is write your own rules or plugins to tackle your spam.*

*These are the types of topics that can be made into a talk all their own. Again, there is*

## **High Performance Apache SpamAssassin**

*lots of information on the wiki, if you feel that you need to go down this route then you should definately check it out.*

- **Custom Scoring - possible beyond scope but lets be bold for now**

*This is for the extreme folks in the audience. Not necessarily recommended for the faint of heart. However, in some cases this option can really work well.*

*Talk about turning off network tests and getting rid of them.*

*That all said, this is a process that takes work. You need to*

*1) build up a large corpus, when we score a general release we use millions of messages and spend multiple weeks checking those messages and generating the scores.*

*2) The actual score generation piece is largely an artform. There is some information on the wiki. I'll admit right up front that this is something that will take some work, but can work very well once you put*

- **Making Apache SpamAssassin Easier to Maintain**
  - **SQL User Preferences**
  - **Maia/Amavis/etc**
  - **Custom Solutions**
  - **Third-Party/Commercial Solutions**