SplitScreen: Enabling Efficient, Distributed Malware Detection

Sang Kil Cha, Iulian Moraru, Jiyong Jang, John Truelove, David Brumley, David G. Andersen
Carnegie Mellon University
Malware Scanning

Malware Sigs (Strings, RegExps, etc.)

Files

ClamAV

Exact Signature Matching

9c5750139409a8b187015938ff0f…

Found Malware
Signature-based Scanning

- Currently fastest method
  - Emerging alternatives slower (e.g., behavior-based)
  - Signature scanning likely part of practical solutions

- Widely Deployed
  - $2 billion industry
  - Symantec, Trend Micro, ClamAV, etc. all use signature-based scanning

- Millions of existing signatures
Signature Count is exploding

- **2005:** 0K
- **2006:** 0K
- **2007:** 100K
- **2008:** 200K
- **2009:** 700K

#Sigs in ClamAV
More Signatures = More Memory!

![Graph showing the growth of signatures in ClamAV and memory usage from 2005 to 2009. The graph indicates a significant increase in both signatures and memory requirements over the years.]
More Signatures = Poor Cache Performance!

![Graph showing the correlation between the number of signatures in ClamAV and cache misses/misses in memory over the years 2005 to 2009.](image-url)
More Signatures = Slower!

Throughput

11 MB/s

5 MB/s

Memory

Cache Misses

Year

2005  2006  2007  2008  2009

#Sigs in ClamAV

0K  100K  200K  300K  400K  500K  600K  700K

Throughput

Memory

Cache Misses

More Signatures = Slower!
SplitScreen:

$\geq 2x \text{ the speed, }$

$\leq \frac{1}{2} \text{ the memory.}$
Opportunity: Few Signatures Matched

4 month study of CMU email malware

< 1% of signatures used by ClamAV for all malware

Detect all the malware using $\approx 1000$ signatures.

Detect 80% malware using 5 signatures.
SplitScreen Architecture

- **All Malware Sigs**
- **Cache-Efficient**
- **FFBF**
- **Identified Sigs**
- **Suspect Files**
- **Exact Signature Matching**
- **Reduce # of files (5~10%), but not miss malware**
- **Reduce # of sigs. (< 1%)**
- **Found Malware**
SplitScreen Design

- Feed-Forward Bloom Filter (FFBF)
  - FFBF Initialization
  - FFBF Scanning: File filtering
  - FFBF Scanning: Pattern filtering

- Cache-efficient Bloom filter design
Feed-Forward Bloom Filter (FFBF): Initialization

Patterns (signatures)

Hash Functions

All-patterns
FFBF Scanning: File Filtering

Patterns (signatures)

- d99094ef??c97864
- 70aabb(\texttt{cc}|\texttt{dd})bb

Target File: A.exe
... ffd99094f04 ...

All-patterns

100100010110

... Rolling Hash

Found Suspect

13
FFBF Scanning: Pattern Filtering

Patterns (signatures)

<table>
<thead>
<tr>
<th>d99094ef??c97864</th>
</tr>
</thead>
<tbody>
<tr>
<td>70aabb(cc</td>
</tr>
</tbody>
</table>

... ffd99094f04 ...

Target File: A.exe

All-patterns

1 0 0 1 0 0 0 1 0 1 1 0

Matched-patterns

0 0 0 1 0 0 0 0 0 0 1 0

Rolling Hash
FFBF Scanning Recap

... d99094ef??c97864 70aabb(cc|dd)bb ...

A.exe
B.exe

FFBF

< 1%

d99094ef??c97864

A.exe

5 ~ 10%

Exact Signature Matching
Cache-Efficient Bloom Filter

≈10MB for 500k sigs

Standard

Cache-Efficient

L2/L3 Cache Size

Cache-resident part: Use separate hash functions.

Non-cache-resident part: check only if the cache-resident part has hits for all the hashes.
The rest of the talk

- Evaluation of SplitScreen on Intel 2.4 GHz Core 2 Quad with 4 GB of RAM
  - Throughput
  - Cache performance
  - Memory use

- On-demand signature distribution
Throughput (1.6 GB Clean Files)

- **SplitScreen**
  - 500K sigs: \(2x\) throughput
  - 3M sigs: \(4x\) throughput

- **ClamAV**

Throughput (MB/s)

Number of signatures

- 500K
- 1M
- 2M
- 3M
Better Cache Performance

- Number of signatures
- Cache Misses
- ClamAV
- SplitScreen
Less Memory

Memory Use (MB)

Number of signatures (in log scale)

ClamAV

SplitScreen

Router Limit (Approx.)

iPhone 3GS Limit

1/2 Memory for 2009 sigs
Reducing Signature Distribution Cost

- Option 1: Private data in cloud
  - Cloud-based virus scanning, e.g., CloudAV
  - SplitScreen accelerates scanning

- Option 2: Private data not in cloud
  - On-demand signature distribution
On-demand signature distribution

Server

Client

Malware Sigs

Files

All-patterns Bit Vector

Matched-patterns Bit Vector

Identified Sigs

Exact Signature Matching
Lower Initial Signature Distribution Cost

10x less server bandwidth

10x less sigs update cost
More details in paper

- Handling short signatures
- Choosing which part of signature for FFBF
- FFBF tuning
  - Number of hash functions
  - Size of the sliding window
  - FFBF size
SplitScreen:

- Malware scanning at 2x - 4x higher throughput, 1/2 - 1/10 memory
- Enables malware scanning on weak devices
  - Embedded systems, ..., maybe iPads
- Enables on-demand signature distribution
  - Minimize overhead given millions of sigs
- Source code at http://security.ece.cmu.edu/