Detecting ICMP Rate Limiting in the Internet

Hang Guo and John Heidemann
USC/Information Sciences Institute
March 26, 2018 PAM’18
ICMP Probing is Widely Used In Network Research

Rate Limiting Could Distort Research Results!

University of Pennsylvania  Texas A&M University  University of Michigan
Delft University of Technology  UC San Diego  ICS
ETH Zurich
University of Maryland  Bielefeld  TU München  ICSI
CAIDA  Massachusetts Amherst  Univ. of Washington
Information Sciences Institute  Colorado State University
Swinburne University of Technology  Monash University
UC Berkeley  Internet Initiative Japan  USC  Inria
The Aerospace Corporation  University of Washington
Naval Postgraduate School  University of Waikato
Universite Nice Sophia Antipolis  MIT  Lastline
Motivations

• ICMP probing is the center of many network measurements.

• Undetected ICMP rate limiting could distort measurements, silently creating false conclusions
Contributions

- Develop a new, light-weight, detection algorithm: FADER
- Systematically look for ICMP rate limiting in the Internet.
- We showed
  - rate limiting is rare up to 0.38 ping/s per /24
  - probing up to 1 ping/s per /24 risk being rate limited.
Talk Structure

Methodology

Detection Results

Related Work and More
How Rate Limiting Affect Probing

Pinger knows probes and sees responses ==> how it knows if rate limited?

Probing: P pps per /24

Response: 0.5 P pps per /24

Internet

R1

R2

R3

0.5P pps per /24

Rate Limiter at 0.5P pps per /24

host1

host2

host256

Assuming host 1 to 256 responsive ==> response rate = request rate

probes dropped with 50% chance ==> 0.5P pps per /24 at egress
Insights for Detection

1. Rate limiting causes **probe loss**

2. This probe loss is randomized

---

**Slow Probing:** 0.01 Ping/s per /24

(< Rate Limit)

**Fast Probing to Same Blk:** 0.4 Ping/s per /24

(4*Rate Limit)
We probe target blocks at two rates: a slow ($P_s$) and a fast ($P_f$).

$P_s$, chosen carefully, can be under any rate limits.

$P_f$ sets upper bound for detection.

Modeled Availability for a Rate Limited /24 Block*

*More about our rate limiting models in paper
FADER: Rate Limiting Detection

We probe target blocks at two rates: a slow (Ps) and a fast (Pf).

Rate limit exists if

1. higher availability in Ps than Pf
2. random probe loss in Pf

if \( \frac{A(Ps) - A(Pf)}{A(Ps)} \geq 0.1 \)

Modeled Availability for a Rate Limited /24 Block

*See our paper for reasoning behind threshold 0.1
Rate limit exists if
(1) higher availability in Ps than Pf
(2) random probe loss in Pf

Q: How to detect randomness? A: signaled by response alternations: transition between responsiveness and un-responsiveness

Examples Of Response Alternations (9 Alternations Shown In White Squares )
Rate limit exists if
(1) higher availability in Ps than Pf
(2) random probe loss in Pf

More than threshold* response alternations ==> Random probe loss

*See our paper for details about this threshold
FADER: Rate Limit Estimation

We probe target blocks at two rates: a slow ($P_s$) and a fast ($P_f$)

Rate limit exists if
(1) higher availability in $P_s$ than $P_f$
(2) random probe loss in $P_f$

Rate limit estimated by block availabilities

Modeled Availability for a Rate Limited /24 Block

$$\text{Rate Limit} = \frac{A(P_f)}{A(P_s)} \times \text{Probing Rate}$$

*Details about this equation in paper*
Identify Where No Algorithm Can Work

- When Pf **right above** rate limits (not enough change in response)
- When Pf **far above** rate limits (block appears non-responsive, giving little info).

FADER Conservatively Output: **Un-Rate-Limited**

FADER Output: **Can-Not-Tell**

FADER is correct in the remaining areas

Theoretical Availability for a Rate Limited Block
Talk Structure

Methodology

Detection Results

Related Work and More
Question: How many /24 blocks are rate limited in the Internet?
Is Relatively Slow Probing Rate Limited?

Dataset: It71w Internet Census and Survey from USC/ISI

- **Ps**: Census (single-round, 0.0001 ping/s per block)
- **Pf**: Survey (multiple-round, 0.39 ping/s per block)
  - Covering ICMP rate limiting up to Pf rate
  - Covering 40,493 /24, ~2% of the responsive IPv4 space
  - Representing the whole Internet
Relatively Slow Rate Limit: Initial Results

blocks studied 40,493 (100%)
not-rate limited 24,414 (60%)
cannot tell 15,941 (39%)
rate limited 111 (0.27%)

Mostly barely responsive
(i.e. no one can tell without more info)
Relatively Slow Rate Limit: Re-Probing

- Validate 111 rate limited blocks
  - Re-probing them from 0.01 to 20 ping/s per block
  - If *experimental* values (like blk availability) of re-probing matches *rate limit model*, then block rate limited (true positive)
  - otherwise, block not rate limited (false positive)
Relatively Slow Rate Limit: Re-Probing Results

FADER Detection Results

<table>
<thead>
<tr>
<th>Description</th>
<th>Blocks</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>blocks studied</td>
<td>40,493</td>
<td>(100%)</td>
</tr>
<tr>
<td>not-rate limited</td>
<td>24,414</td>
<td>(60%)</td>
</tr>
<tr>
<td>cannot tell</td>
<td>15,941</td>
<td>(39%)</td>
</tr>
<tr>
<td>rate limited</td>
<td>111</td>
<td>(0.27%)</td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>false positives</td>
<td>105</td>
<td>(0.25%)</td>
</tr>
<tr>
<td>true positives</td>
<td>6</td>
<td>(0.015%)</td>
</tr>
</tbody>
</table>

High false positive (FP) by design:
1. FADER favor FP to avoid false negatives (by detecting with necessary conditions)
2. This trade-off is required to confirm the near-absence of rate limiting

Validate with Re-probing*
Relatively Slow Rate Limit: Conclusion

<table>
<thead>
<tr>
<th>blocks studied</th>
<th>40,493</th>
<th>(100%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>not-rate limited</td>
<td>24,414</td>
<td>(60%)</td>
</tr>
<tr>
<td>cannot tell</td>
<td>15,941</td>
<td>(39%)</td>
</tr>
<tr>
<td>rate limited</td>
<td>111</td>
<td>(0.27%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>false positives</th>
<th>105</th>
<th>(0.25%)</th>
<th>(95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>true positives</td>
<td>6</td>
<td>(0.015%)</td>
<td>(5%)</td>
</tr>
</tbody>
</table>

6 /24 from 2 ISPs:  
124.46.219.0  
124.46.239.0  
182.237.200.0  
182.237.212.0  
182.237.217.0  
202.120.61.0

FADER Detection Results

Validate with Re-probing*

ICMP probing up to 0.39 packet/s per block is not likely to be distorted

When we detect rate limit,  
Our rate limit estimation is generally correct*

*Details in paper
Relatively Slow Rate Limit: Does Results Hold Over Time

Detection to it70w census survey taken 2 months earlier

- Sharing 76% target blocks with it71w data
- Same amount of target blocks as it71w
Relatively Slow Rate Limit: Does Results Hold Over Time

The results are similar:
- 138 potential rate limited blocks (vs 111 in it71w)
- 5 actual limited (vs 6 in it71w)
  - 4 also detected in it71w*
  - the other 1 is not probed in it71w.

Consistency here
- Proves our conclusion about rate limiting holds over time
- Rules out possibility that the observed high FPs are caused by concurrent high-rate ICMP activities at target blocks.

*We saw inconsistency between it70w and it71w for two blocks. Details in paper.
Question: Is faster probing rate limited?

Knowing rate limiting is rare up to 0.38 ping/s per block
Is Faster Probing Rate Limited?

Dataset: Zippier ZMap’s 50-second TCP-SYN probing

- From 0.1M to 14M packet/s (0.007 to 0.97 packets/s per /24)
  - Covering rate limit up to 1 packet/s per /24
- Each scan targets a random sample of a 3.7 billion IP pool
  - Representing all possible IPv4 addresses (2^32 or 4.2 billion)

FADER applies to both ICMP and TCP-SYN probing
Is Faster Probing Rate Limited?

ZMap* reports a **unknown, linear availability drop** in probing results:

–“hit-rate begins to drop linearly after 4 Mpps… It is not immediately clear where probe packets are dropped

We believe **Rate Limiting** is the cause

Can Rate Limit Cause this Linear Drop?

We modeled ZMap’s measurement process:

- Modeled multiple rate limits from 4M pps to 14M pps ==> Aggregation of multiple hyperbolic losses ==> linear losses

Rate Limiting Model closely matches ZMap Experiments

Modeled Availability with Rate Limit
Zmap Experimental Availability

Rate Limiting could explain this linear response drop-off
Do ZMap’s Target Blocks Look Rate Limited?

We partially apply FADER to ZMap probing results

- **Ps**: slowest ZMap scan
- **Pf**: each of the other 15 faster scans

15 FADER tests for each ZMap block
- A block is **potentially rate limited** if classified as rate limited in at least one test

Cannot test random probe loss as ZMap do not repeat addresses
Do ZMap’s Target Blocks Look Rate Limited?

<table>
<thead>
<tr>
<th>blocks studied</th>
<th>56,550</th>
<th>(100%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 rate limited</td>
<td>53,460</td>
<td>(94.54%)</td>
</tr>
<tr>
<td>15 cannot tell</td>
<td>53,149</td>
<td>(93.99%)</td>
</tr>
</tbody>
</table>

Due to target block went dark in Ps
Do ZMap’s Target Blocks Look Rate Limited?

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>blocks studied</td>
<td>56,550</td>
<td>(100%)</td>
</tr>
<tr>
<td>0 rate limited</td>
<td>53,460</td>
<td>(94.54%)</td>
</tr>
<tr>
<td>15 cannot tell</td>
<td>53,149</td>
<td>(93.99%)</td>
</tr>
<tr>
<td>at least 1 rate limited</td>
<td>3,090</td>
<td>(5.46%)</td>
</tr>
</tbody>
</table>

~5% blocks are potential rate limited
Do ZMap’s Target Blocks Look Rate Limited?

<table>
<thead>
<tr>
<th></th>
<th>Values</th>
<th>95% CI</th>
<th>100% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>blocks studied</td>
<td>56,550</td>
<td>(100%)</td>
<td></td>
</tr>
<tr>
<td>0 rate limited</td>
<td>53,460</td>
<td>(94.54%)</td>
<td>(100%)</td>
</tr>
<tr>
<td>15 cannot tell</td>
<td>53,149</td>
<td>(93.99%)</td>
<td>(99.42%)</td>
</tr>
<tr>
<td>at least 1 rate</td>
<td>3,090</td>
<td>(5.46%)</td>
<td>(100%)</td>
</tr>
<tr>
<td>at least 13 rate</td>
<td>2,153</td>
<td>(3.81%)</td>
<td>(69.68%)</td>
</tr>
</tbody>
</table>

Further proof: ~70% of potential rate limit block consistently classified as rate-limited in most of 15 tests

Fast probing up to 1 packets/s per /24 risk being rate limited
Talk Structure

Methodology

Detection Results

Related Work and More
Related Work

Universite Nice Sophia Antipolis studies rate limiting for ICMP Time exceeded replies

- **Different Subject:** They reverse path while we forward path
- **Different Signals:** They constant response rate while we availability difference and random loss

Google examined policing to TCP video traffic

- **Different Subject:** They TCP Video while we ICMP
- **Different Coverage:** Theirs is greater (b/c Google is a major content provider)
- **Different Results:** They shows widespread rate limiting of TCP traffic
Conclusion

• We developed **FADER**, a light-weight rate limiting detector.

• We looked for ICMP rate limiting in the whole Internet.

• We conclude that
  – low-rate ICMP scan (up to 0.39 ping/s per block) are unlikely to be distorted
  – high-rate scan (up to 1 ping/s per block) risks being rate limited.

Check our paper for more about FADER (validation, more results, code available*) and apply it to more datasets!

*Requiring map-reduce