Packed to the Brim: Investigating the Impact of Highly Responsive Prefixes on Internet-wide Measurement Campaigns

Patrick Sattler, Johannes Zirngibl, Mattijs Jonker, Oliver Gasser, Georg Carle, Ralph Holz

Wednesday 6th December, 2023

ACM Conference on emerging Networking EXperiments and Technologies 2023
Paris, France

Motivation

Port scans are an important building block for Internet research

- Rough overview of service deployments on IANA standardized ports
- Target acquisition for application layer scans
- Target selection in security use cases
  - e.g., Censys, Shodan, and Rapid7 use port scans as a baseline

Motivation

Port scans are an important building block for Internet research

- Rough overview of service deployments on IANA standardized ports
- Target acquisition for application layer scans
- Target selection in security use cases
    - e.g., Censys, Shodan, and Rapid7 use port scans as a baseline

Systematic distortions from port scans will affect results in all use cases

Motivation

Example

Step 1

- Scan routable IPv4 address space
- Tool: stateless port scanner (e.g., ZMap)

Motivation

Example

Step 1

- Scan routable IPv4 address space
- Tool: stateless port scanner (e.g., ZMap)

Step 2

- Perform application layer scans
- Targets: responsive hosts from step 1

Motivation

Example

Step 1

- Scan routable IPv4 address space
- Tool: stateless port scanner (e.g., ZMap)

Step 2

- Perform application layer scans
- Targets: responsive hosts from step 1

Step 3

- Evaluate application-layer results

Motivation

- RQ1: Are there distortions in port scans?
  - Distribution of responsive addresses is skewed due to highly responsive prefixes (HRPs)

Motivation

- RQ1: Are there distortions in port scans?
  - Distribution of responsive addresses is skewed due to highly responsive prefixes (HRPs)
- RQ2: To what extent does this impact port scans?
  - Between 20 %-75 % of responsive addresses are impacted
  - Different deployment strategies by ASes: HRPs only on specific ports or on all ports

Motivation

- RQ1: Are there distortions in port scans?
  - Distribution of responsive addresses is skewed due to highly responsive prefixes (HRPs)
- RQ2: To what extent does this impact port scans?
  - Between 20 %-75 % of responsive addresses are impacted
  - Different deployment strategies by ASes: HRPs only on specific ports or on all ports
- RQ3: What impact does this have on application-layer scans?
  - Lower success rate for targets within HRPs
  - Lower information gain per scanned target

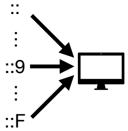
Related Work

- No in-depth analysis in IPv4; but indicators
  - All addresses within a prefix appear to be responsive; Izhikevich et al.[1]
    - Evaluation of application layer services on non-standard ports

[1] L. Izhikevich et al. 2021. LZR: Identifying Unexpected Internet Services. In Proc. USENIXSecuritySymposium [2]

Related Work

- No in-depth analysis in IPv4; but indicators
  - All addresses within a prefix appear to be responsive; Izhikevich et al.[1]
    - Evaluation of application layer services on non-standard ports
- Aliased prefixes in IPv6 hitlists; Gasser et al.[2]
  - All addresses handled by a single host
  - Induces bias in hitlist



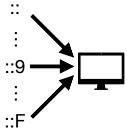
Aliased Prefix

[1] L. Izhikevich et al. 2021. LZR: Identifying Unexpected Internet Services. In Proc. USENIXSecuritySymposium [2]

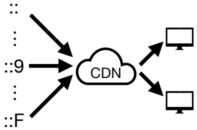
[2] O. Gasser et al. 2018. Clusters in the Expanse: Understanding and Unbiasing IPv6 Hitlists. In Proc. ACM Int. Measurement Conference [1]

Related Work

- No in-depth analysis in IPv4; but indicators
  - All addresses within a prefix appear to be responsive; Izhikevich et al.[1]
    - Evaluation of application layer services on non-standard ports
- Aliased prefixes in IPv6 hitlists; Gasser et al.[2]
  - All addresses handled by a single host
  - Induces bias in hitlist
- Fully responsive prefixes; Zirngibl et al.[3]
  - Need for a broader definition
  - CDN prefixes appear to be fully responsive in IPv6
  - Different considerations apply for aliased vs fully responsive prefixes



Aliased Prefix



Fully Responsive Prefix

[1] L. Izhikevich et al. 2021. LZR: Identifying Unexpected Internet Services. In Proc. USENIXSecuritySymposium [2]

[2] O. Gasser et al. 2018. Clusters in the Expanse: Understanding and Unbiasing IPv6 Hitlists. In Proc. ACM Int. Measurement Conference [1]

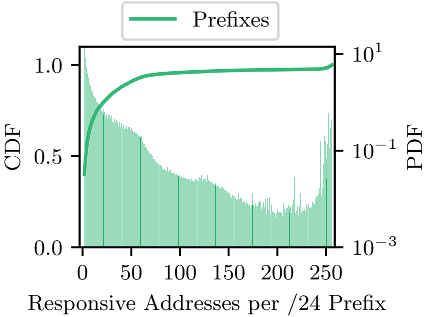
[3] J. Zirngibl et al. 2022. 2022. Rusty Clusters? Dusting an IPv6 Research Foundation. In Proc. ACM Int. Measurement Conference [3]

Research Questions

- RQ1: Are there distortions in port scans?
- RQ2: To what extent do highly responsive prefixes impact port scans?
- RQ3: What impact does this have on application-layer scans?

Are there distortions in port scans?

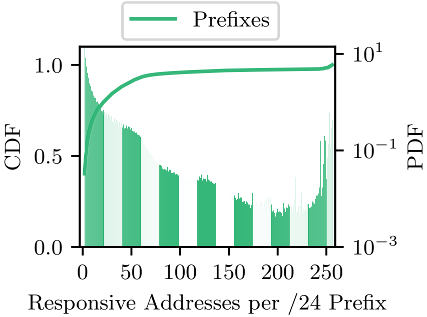
TCP/443 Port Scan Results

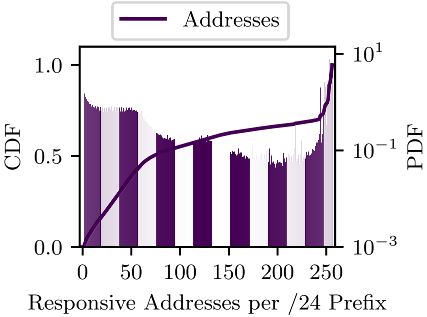


- 91 % of prefixes have less than 50 responsive addresses

Are there distortions in port scans?

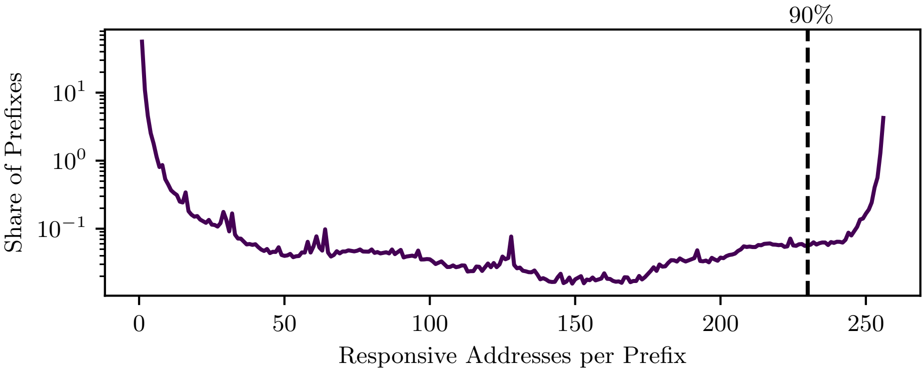
TCP/443 Port Scan Results





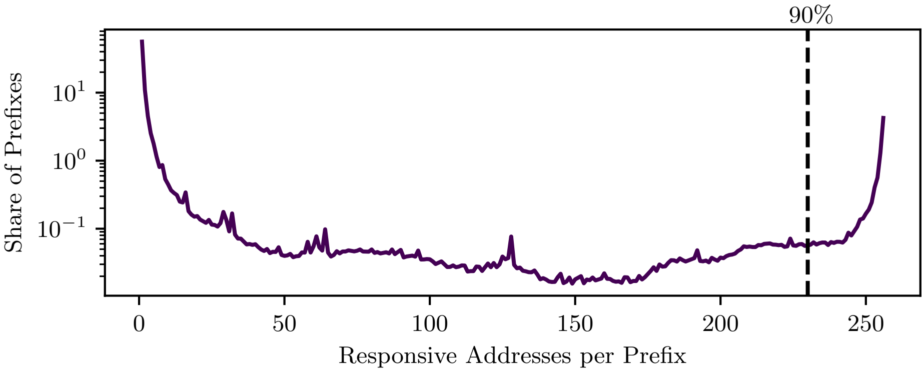
- 91 % of prefixes have less than 50 responsive addresses
- $>$ 30% of addresses are in highly responsive prefixes

HRP Definition



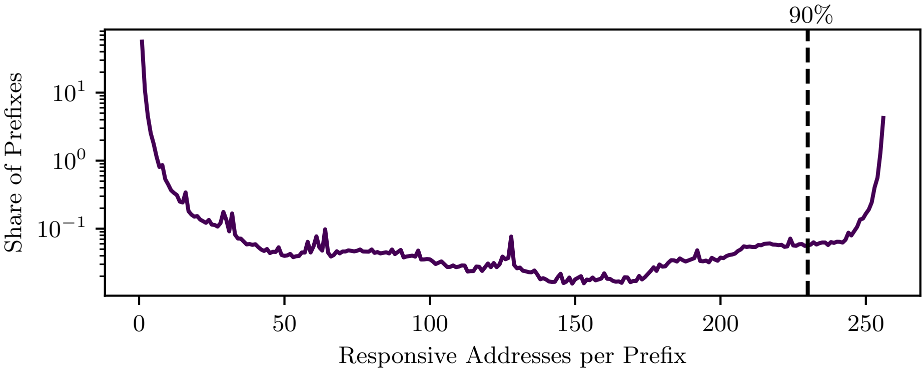
- Evaluation of reachable host per prefix across all analyzed ports

HRP Definition



- Evaluation of reachable host per prefix across all analyzed ports
- Extreme ends of distribution strongly influence overall result

HRP Definition



- Evaluation of reachable host per prefix across all analyzed ports
- Extreme ends of distribution strongly influence overall result
- HRPs are all prefixes with more than 90% responsive hosts

Datasets

Port Scan Datasets

- Weekly TCP/443 port scans from 2021 until end of 2022

Datasets

Port Scan Datasets

- Weekly TCP/443 port scans from 2021 until end of 2022
- Dedicated port scans from Munich and Saarbrücken for 36 ports

Datasets

Port Scan Datasets

- Weekly TCP/443 port scans from 2021 until end of 2022
- Dedicated port scans from Munich and Saarbrücken for 36 ports
- Port scans from Rapid7 Project Sonar (129 TCP ports; 19 UDP ports)

Datasets

Port Scan Datasets

- Weekly TCP/443 port scans from 2021 until end of 2022
- Dedicated port scans from Munich and Saarbrücken for 36 ports
- Port scans from Rapid7 Project Sonar (129 TCP ports; 19 UDP ports)

Application Layer Data

- TLS application layer results

Datasets

Port Scan Datasets

- Weekly TCP/443 port scans from 2021 until end of 2022
- Dedicated port scans from Munich and Saarbrücken for 36 ports
- Port scans from Rapid7 Project Sonar (129 TCP ports; 19 UDP ports)

Application Layer Data

- TLS application layer results
- Rapid7 TLS and HTTP results

Datasets

Port Scan Datasets

- Weekly TCP/443 port scans from 2021 until end of 2022
- Dedicated port scans from Munich and Saarbrücken for 36 ports
- Port scans from Rapid7 Project Sonar (129 TCP ports; 19 UDP ports)

Application Layer Data

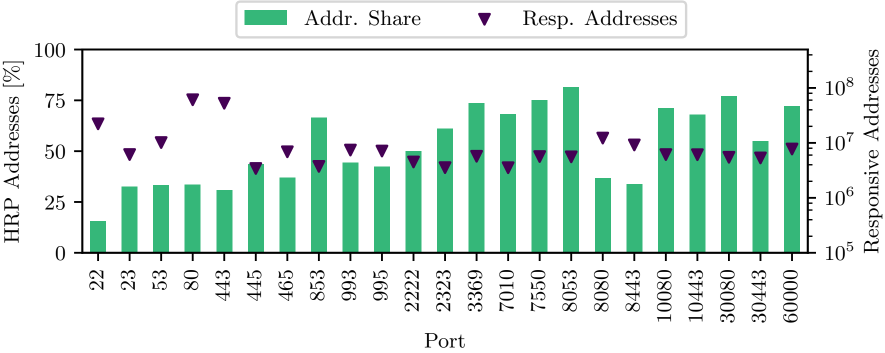
- TLS application layer results
- Rapid7 TLS and HTTP results
- OpenINTEL DNS data

Research Questions

- RQ1: Are there distortions in port scans?
- RQ2: To what extent do highly responsive prefixes impact port scans?
- RQ3: What impact does this have on application-layer scans?

To what extent does this impact port scans?

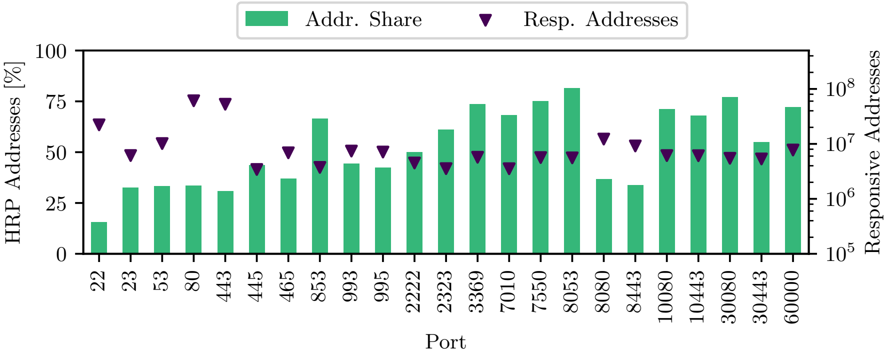
Comparison between TCP Ports



- 30% HRP address share for IANA standard ports, port 8080 and 8443
- Other services have up to 75% HRP share

To what extent does this impact port scans?

Comparison between TCP Ports



- 30% HRP address share for IANA standard ports, port 8080 and 8443
- Other services have up to 75% HRP share

→ Which ASes deploy HRPs?

To what extent does this impact port scans?

ASes Deploying HRPs

Top ASes by total number of HRPs

AS	Visible /24	HRP Share
AS16625 (Akamai)	22.9k	97.8%
AS20940 (Akamai)	24.7k	85.6%
AS7713 (Telin)	12.5k	52.5%
AS16509 (Amazon)	134.9k	4.4%
AS721 (DoD)	4.9k	91.3%
...		
AS13335 (Cloudflare)	3.1k	98.3%

- Four CDN/Cloud provider ASes, three ISPs, two DoD ASes, and one academic network
- Top five cover 64 % of all HRPs

To what extent does this impact port scans?

ASes Deploying HRPs

Top ASes by total number of HRPs

AS	Visible /24	HRP Share	Ports with HRPs	Visible Ports
AS16625 (Akamai)	22.9k	97.8%	3	5
AS20940 (Akamai)	24.7k	85.6%	5	136
AS7713 (Telin)	12.5k	52.5%	4	136
AS16509 (Amazon)	134.9k	4.4%	135	136
AS721 (DoD)	4.9k	91.3%	55	136
. . .				
AS13335 (Cloudflare)	3.1k	98.3%	136	136

- Four CDN/Cloud provider ASes, three ISPs, two DoD ASes, and one academic network
- Top five cover 64 % of all HRPs
- Some CDNs deploy HRPs on all visible ports

→ Different deployment reasons and strategies

Research Questions

- RQ1: Are there distortions in port scans?
- RQ2: To what extent do highly responsive prefixes impact port scans?
- RQ3: What impact does this have on application-layer scans?

What impact does this have on application-layer scans?

TLS Data

We use:

- TLS handshake results for port 443 from our local measurement campaign
- Certificate data by Rapid7
  - Only data for targets with certificates

What impact does this have on application-layer scans?

TLS Data

We use:

- TLS handshake results for port 443 from our local measurement campaign
- Certificate data by Rapid7
    - Only data for targets with certificates

- How many TLS services are active inside HRPs?
- What information gain can be expected when scanning HRPs?

What impact does this have on application-layer scans?

TLS Hosts in HRPs

- 84% of TCP/443 HRPs contain TLS responsive hosts
- Lower share of HRPs on other ports

		App. Layer Success
Port	# HRPs	# HRPs
443	64 435	54 203
8443	13 048	3287
25	33 294	3493
110	11 394	2553
853	8352	565

What impact does this have on application-layer scans?

TLS Hosts in HRPs

- 84% of TCP/443 HRPs contain TLS responsive hosts
- Lower share of HRPs on other ports
- Only half of these are highly responsive on the application layer

		App. Layer Success	
Port	# HRPs	# HRPs	>90 % Success
443	64 435	54 203	26715
8443	13 048	3287	809
25	33 294	3493	2210
110	11 394	2553	2379
853	8352	565	379

What impact does this have on application-layer scans?

TLS Hosts in HRPs

- 84% of TCP/443 HRPs contain TLS responsive hosts
- Lower share of HRPs on other ports
- Only half of these are highly responsive on the application layer
- Mail ports have large share of single identifier HRPs

		App. Layer Success		Same Identifier	
Port	# HRPs	# HRPs	>90 % Success	# HRPs	HRP [%]
443	64 435	54 203	26715	2718	10.2
8443	13 048	3287	809	384	47.5
25	33 294	3493	2210	2041	92.4
110	11 394	2553	2379	1944	81.7
853	8352	565	379	53	14.0

What impact does this have on application-layer scans?

- Few HRPs provide actual service on all addresses
- Responsive HRP hosts tend to have the same analyzed identifier (e.g., certificate)

What impact does this have on application-layer scans?

- Few HRPs provide actual service on all addresses
- Responsive HRP hosts tend to have the same analyzed identifier (e.g., certificate)
- Notable exceptions TCP/80 and TCP/443
    - These port scans are dominated by CDNs
    - We found different reasons for CDNs deploying HRPs:
        - IPv4 addresses are not easily available and CDNs use their available assets

What impact does this have on application-layer scans?

- Few HRPs provide actual service on all addresses
- Responsive HRP hosts tend to have the same analyzed identifier (e.g., certificate)
- Notable exceptions TCP/80 and TCP/443
  - These port scans are dominated by CDNs
  - We found different reasons for CDNs deploying HRPs:
    - IPv4 addresses are not easily available and CDNs use their available assets
    - Cloudflare deploys addressing agility techniques and TCP proxies on all ports

→ HRPs cause distortions in application layer scans (see single identifier prefixes)

What impact does this have on application-layer scans?

- Few HRPs provide actual service on all addresses
- Responsive HRP hosts tend to have the same analyzed identifier (e.g., certificate)
- Notable exceptions TCP/80 and TCP/443
    - These port scans are dominated by CDNs
    - We found different reasons for CDNs deploying HRPs:
        - IPv4 addresses are not easily available and CDNs use their available assets
        - Cloudflare deploys addressing agility techniques and TCP proxies on all ports

→ HRPs cause distortions in application layer scans (see single identifier prefixes)

New Application Layer Scanning Approach

- Filter HRPs from port scans before running the application layer scan
- Scan HRPs selectively (DNS and sample-based)

What impact does this have on application-layer scans?

- Few HRPs provide actual service on all addresses
- Responsive HRP hosts tend to have the same analyzed identifier (e.g., certificate)
- Notable exceptions TCP/80 and TCP/443
  - These port scans are dominated by CDNs
  - We found different reasons for CDNs deploying HRPs:
    - IPv4 addresses are not easily available and CDNs use their available assets
    - Cloudflare deploys addressing agility techniques and TCP proxies on all ports

→ HRPs cause distortions in application layer scans (see single identifier prefixes)

New Application Layer Scanning Approach

- Filter HRPs from port scans before running the application layer scan
- Scan HRPs selectively (DNS and sample-based)
- We applied this approach to our previous data:
  - 99 % of unique certificates are discovered
  - –75 % application layer probes

Conclusion

- RQ1: Are there distortions in port scans?
  - Defined and outlined HRP distortions in port scans

Conclusion

- RQ1: Are there distortions in port scans?
  - Defined and outlined HRP distortions in port scans
- RQ2: To what extent does this impact port scans?
  - Analyzed port scans across multiple ports
  - Showed that between 20 % and 75 % of responsive hosts are affected

Conclusion

- RQ1: Are there distortions in port scans?
  - Defined and outlined HRP distortions in port scans
- RQ2: To what extent does this impact port scans?
  - Analyzed port scans across multiple ports
  - Showed that between 20 % and 75 % of responsive hosts are affected
- RQ3: What impact does this have on application-layer scans?
  - Evaluated DNS and TLS data
  - Proposed a new more ethical scanning approach

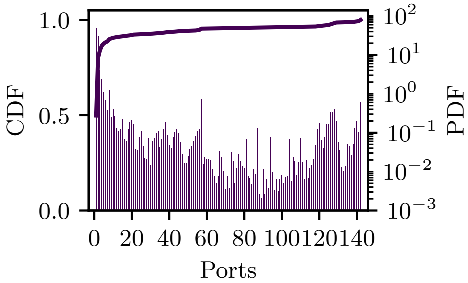
Conclusion

- RQ1: Are there distortions in port scans?
  - Defined and outlined HRP distortions in port scans
- RQ2: To what extent does this impact port scans?
  - Analyzed port scans across multiple ports
  - Showed that between 20 % and 75 % of responsive hosts are affected
- RQ3: What impact does this have on application-layer scans?
  - Evaluated DNS and TLS data
  - Proposed a new more ethical scanning approach
- *Tool and data openly available*
  - Tool to detect HRPs in port scans
  - Weekly new HRP data for ports 80 and 443

https://hrp-stats.github.io/

Backup Slides

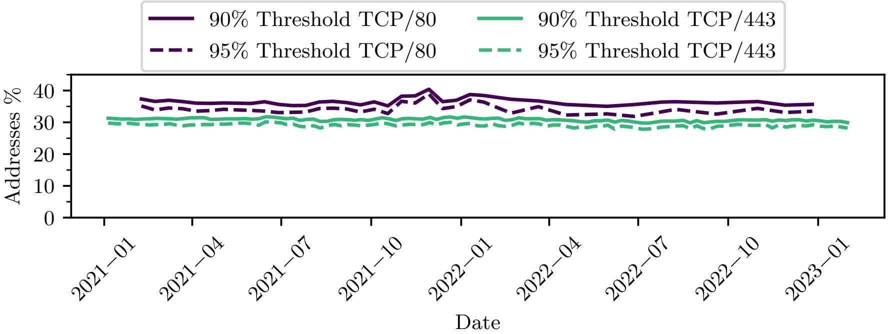
HRPs on multiple TCP ports



- 50% of HRPs appear only on a single port
- Spikes are due to our different data sources and logarithmic PDF axis
- Some prefixes are highly responsive on all analyzed ports

Backup Slides

HRP Stability



- Stable results over the period of two years
- We validated the stability of results between vantage points (see results in the paper)

Backup Slides

Domains in HRPs

- HTTP/S HRPs expose a larger number of DNS references
- Overall only a small fraction of responsive addresses is referenced by in DNS

	HRP	IP addresses	
DNS Ports (Using A records of NS names):			
TCP/53	12.0%	40.9k	1.4%
UDP/53	25.5%	29.0k	3.1%
Mail Ports (Using A records of MX names):			
TCP/25	18.5%	172.4k	2.0%
TCP/110	26.4%	126.0k	4.4%
TCP/143	26.3%	121.6k	4.3%
HTTP/S Ports (Using A records):			
TCP/80	34.4%	4.7M	11.0%
TCP/443	30.8%	2.0M	6.3%
TCP/8443	56.8%	517.3k	16.7%

Backup Slides

Domains in HRPs

- HTTP/S HRPs expose a larger number of DNS references
- Overall only a small fraction of responsive addresses is referenced by in DNS
- A large number of FQDNs and SLDs depend on services in HRPs

	HRP	IP addresses		FQDNs	SLDs
DNS Ports (Using A records of NS names):					
TCP/53	12.0%	40.9k	1.4%	161.6 k	115.6 M
UDP/53	25.5%	29.0k	3.1%	133.0 k	104.6 M
Mail Ports (Using A records of MX names):					
TCP/25	18.5%	172.4k	2.0%	3.0 M	3.7 M
TCP/110	26.4%	126.0k	4.4%	2.7 M	3.2 M
TCP/143	26.3%	121.6k	4.3%	2.7 M	3.2 M
HTTP/S Ports (Using A records):					
TCP/80	34.4%	4.7M	11.0%	171.4 M	-
TCP/443	30.8%	2.0M	6.3%	149.1 M	-
TCP/8443	56.8%	517.3k	16.7%	28.1 M	-

Bibliography

[1] O. Gasser, Q. Scheitle, P. Foremski, Q. Lone, M. Korczynski, S. D. Strowes, L. Hendriks, and G. Carle.
Clusters in the Expanse: Understanding and Unbiasing IPv6 Hitlists.
In Proc. ACM Int. Measurement Conference (IMC), 2018.

[2] L. Izhikevich, R. Teixeira, and Z. Durumeric.
LZR: Identifying unexpected internet services.
In Proc. USENIX Security Symposium, Aug. 2021.

[3] J. Zirngibl, L. Steger, P. Sattler, O. Gasser, and G. Carle.
Rusty Clusters? Dusting an IPv6 Research Foundation.
In Proc. ACM Int. Measurement Conference (IMC), 2022.