12 Days in the life of the IETF FTP service

(August 2 - 13, 2020)

This is additional data in response to discussion on the Call for Community Feedback: Retiring IETF FTP Service thread on ietf@ietf.org. A big thanks to Glen Barney and Henrik Levkowetz on the Tools Team who provided the raw data and supporting analysis.

1. Summary

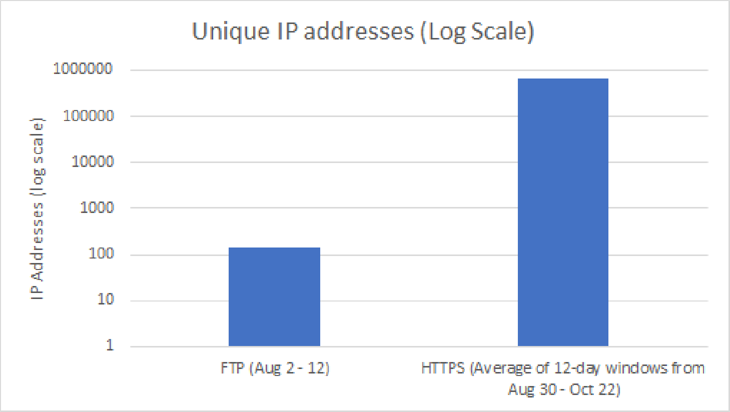
- <91 unique users of FTP (from 140 unique IPs)
- FTP unique IP counts are <0.02% of HTTP traffic
- 78% of all traffic is created by 5 users
- 40% of FTP users made 1 file request
- FTP requests are 0.2% of HTTP traffic (in the case of drafts/RFCs)
- >99% of all FTP traffic is users performing bulk downloads or search engine behavior

2. How many users?

140 unique IP addresses were observed. Analysis suggests that there appears to up to 91 “unique users” (this is an upper bound). Conversion of IP addresses to “users” was done by manual inspection:

- Combine IP addresses with same reverse DNS name look-ups (i.e., *.<searchengine>bot.com) (35 IP addresses combined into 1 notional user)
- Combine IP addresses from the seemingly same dynamic address pool making requests, that when combined, look like the same operation (18 IP addresses combined into 3 notional users)

For perspective, FTP is traffic is roughly 0.02% of the HTTPS traffic

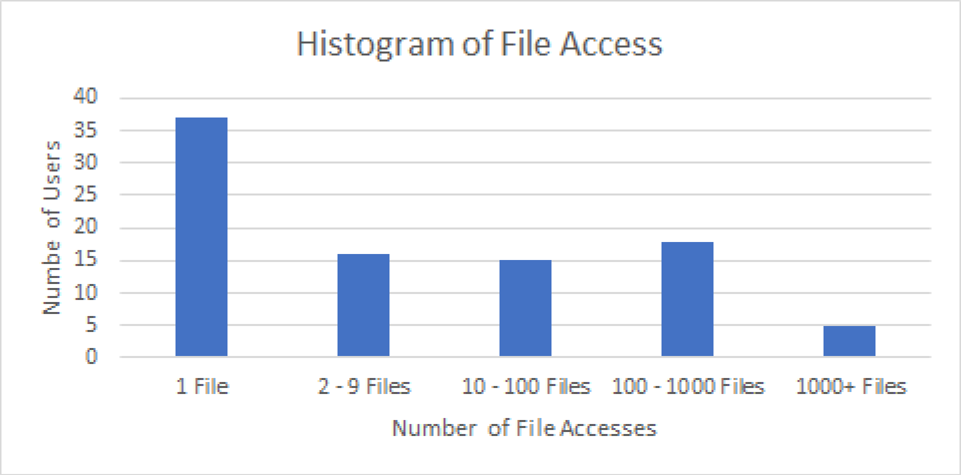


Note: HTTPS traffic is undercounted because only RFC/I-D access on tools.ietf was considered (log periods don’t match because only aggregate logs were available at the time of this analysis)

3. What was the request volume?

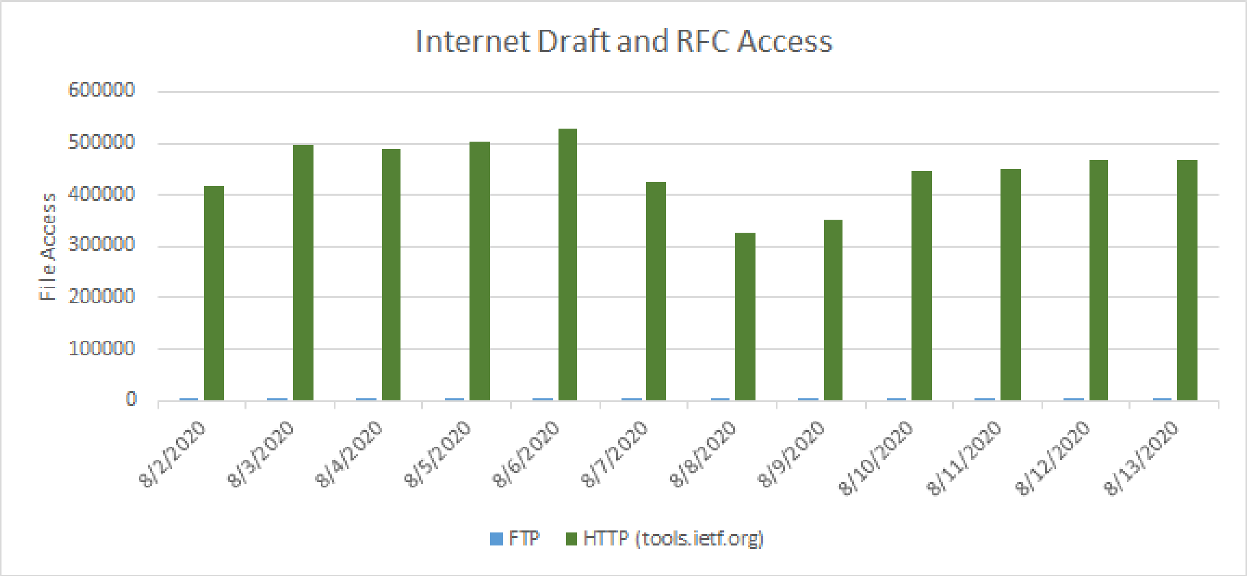
33673 unique requests were made.

Comparing request volume by "user"

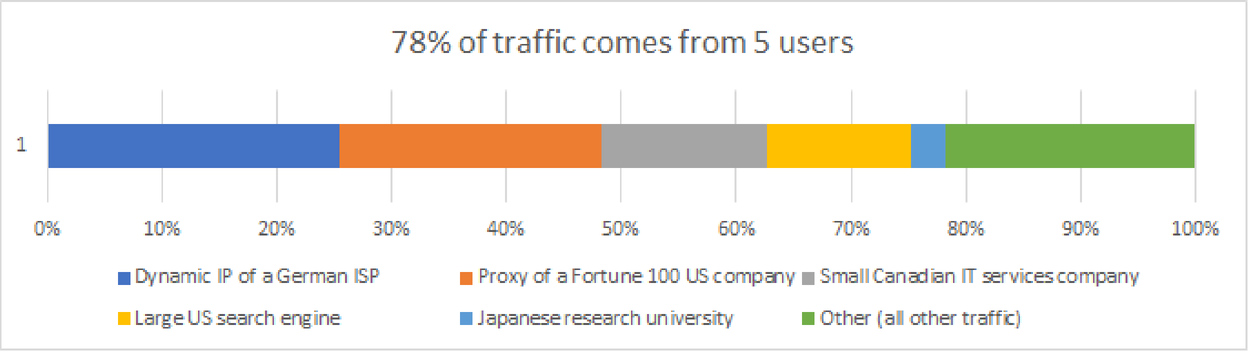


Comparing FTP vs. HTTP (linear scale)

FTP access is ~0.18% of comparable HTTP traffic (when comparing I-D and RFC access)

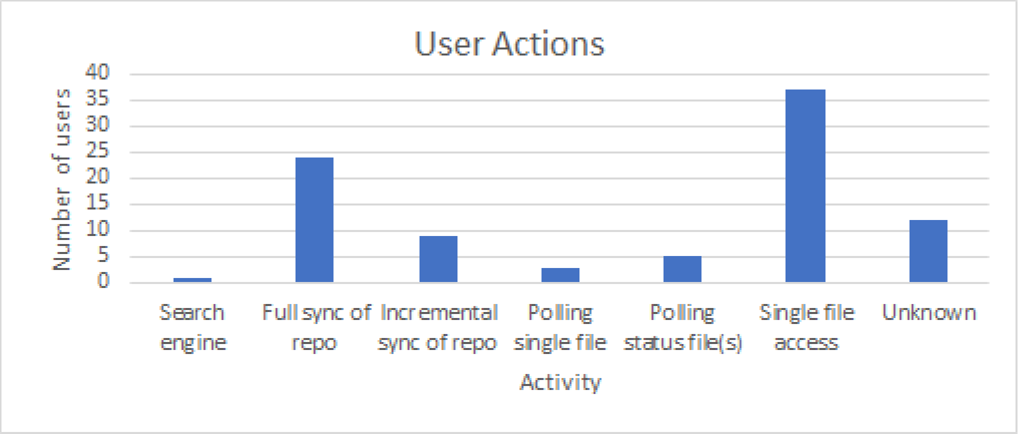


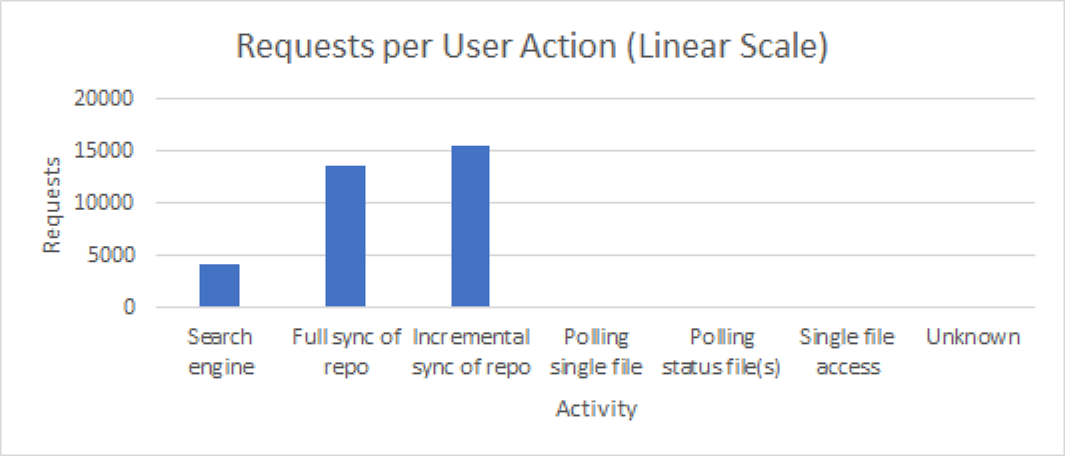
4. Who are the users?



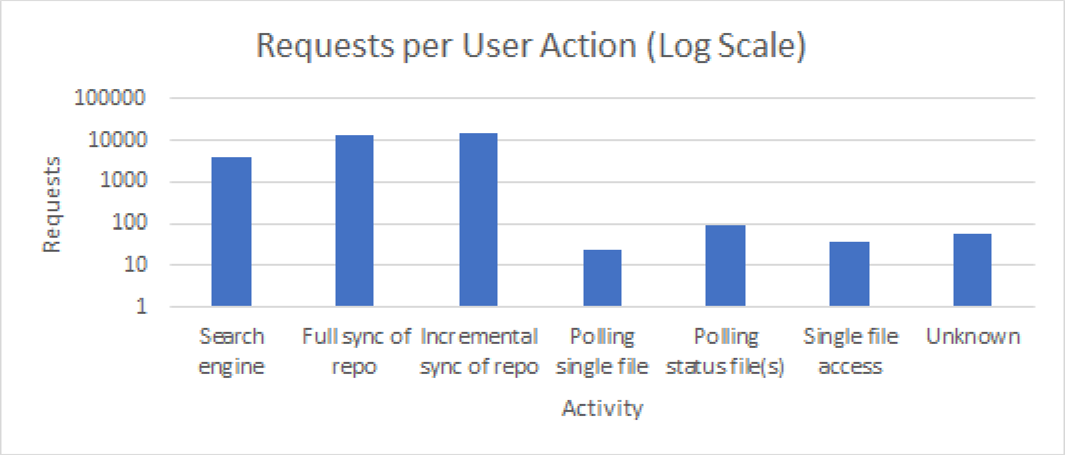
5. What are users doing?

Manual inspection of all requests categorized the behavior of the users. This is an imperfect analysis based on the IP/DNS name, the files being requested, and the periodicity of requests.





- Search engine = download of robots.txt or sync to CDN cache
- Sync of the repo = downloads an entire or parts of a repository (multiple requests where the names are algorithmically generated, e.g., alphabetical file order, sequential RFCs numbers)
- Incremental sync of the repo = downloads of new files since last update (downloading files that have changed or been added, may include repeated download of summary files)
- Polling a single file = repeated access to a single file
  - /ietf/ftpext/ftpext-charter.txt
  - /iana/ipv4-recovered-address-space/ipv4-recovered-address-space.txt
- Polling the status files = repeated access to one or some the summary files (e.g., rfc-index.txt, 1id-abstracts.txt)
- Single file access = on-time access to a single file (perhaps this is an inbound link?)
- Unknown = could not discern activity pattern; access between 2 - 11 files (sometimes drafts, RFCs, mailing lists)



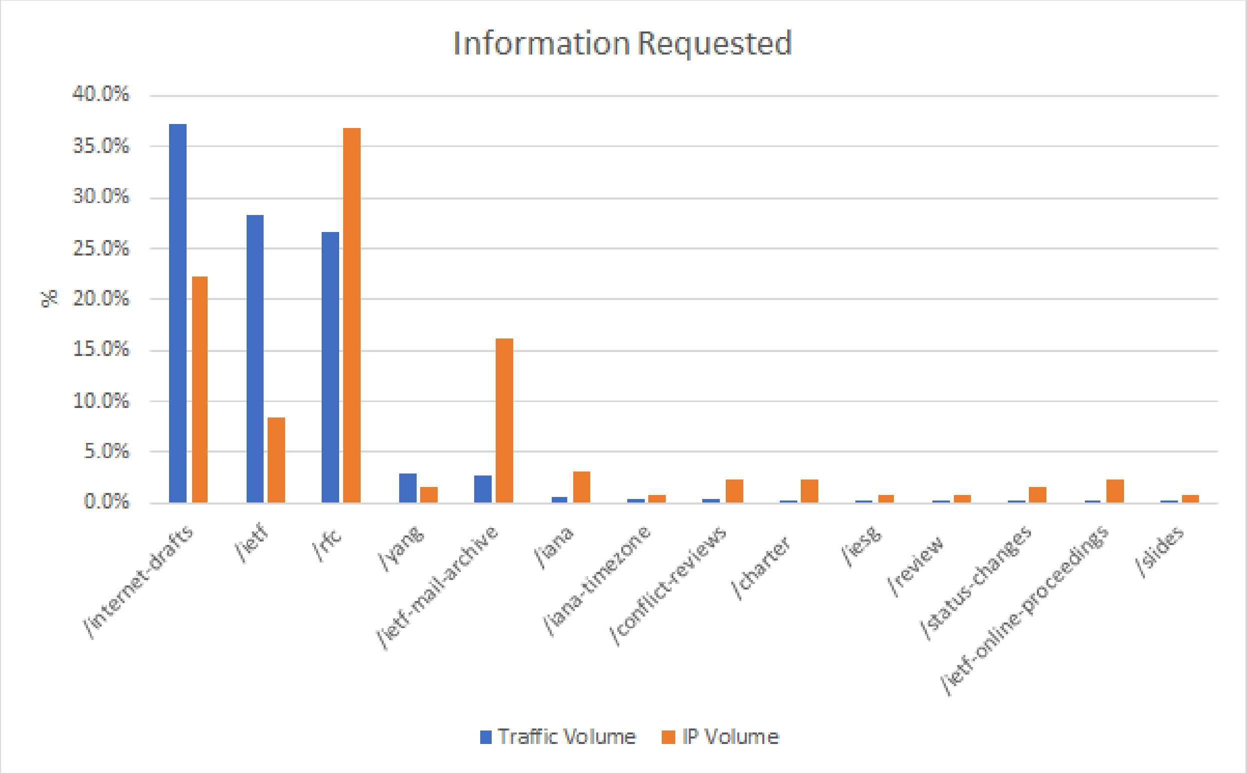
6. How much of this is users running scripts?

It is difficult to tell. “Search engine” (35 IP addresses) and “Single file access” (37 IP addresses) can safely be excluded as users running custom scripts. This leaves no more than 68 IP addresses (49%).

7. What information are the users requesting?

The chart below summarizes the proportion of total FTP usage by directory. The distribution of IP addresses accessing these directories is also provided as a rough sense of the size of the user population (search engine IPs were removed, and unlike the above, aggregating IP addresses likely to be the same users was not done).

- Access to documents (64%) is the most popular information
- Most of the access (69%) is to content where IETF is the authority (i.e., not RFCs, YANG or mirrored IANA content)



* If IP aggregation was applied as above, unique IP addresses for /ietf and /rfc would shrink