RateLimit Headers

Communicate service status

HTTPAPI-WG @ IETF-114

ietf-httpapi-ratelimit-headers

[see the specifications]

RateLimit Fields - Goals

- communicate service limits, so clients can stop before being throttled out
- align all the *already existing* ratelimit headers and stop headers' proliferation
- express multiple RateLimit policies

STOP headers proliferation

X-RateLimit-UserLimit: 1231513

X-RateLimit-UserRemaining

X-Rate-Limit-Limit: name=rate-limit-1,1000

x-custom-retry-after-ms

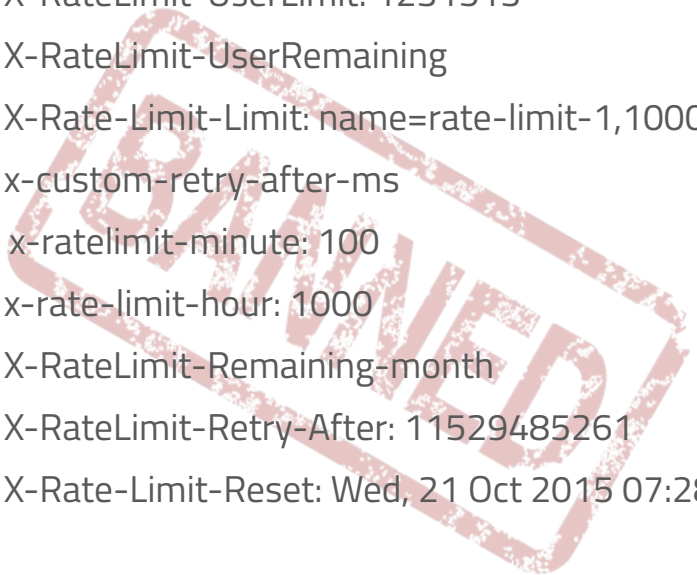
x-ratelimit-minute: 100

x-rate-limit-hour: 1000

X-RateLimit-Remaining-month

X-RateLimit-Retry-After: 11529485261

X-Rate-Limit-Reset: Wed, 21 Oct 2015 07:28:00 GMT



RateLimit-Limit: **SF-List** **#quota-units**

RateLimit-Remaining: **SF-Integer** **#quota-units**

RateLimit-Reset: **SF-Integer** **#delta-seconds**

RateLimit-Policy: **SF-List**

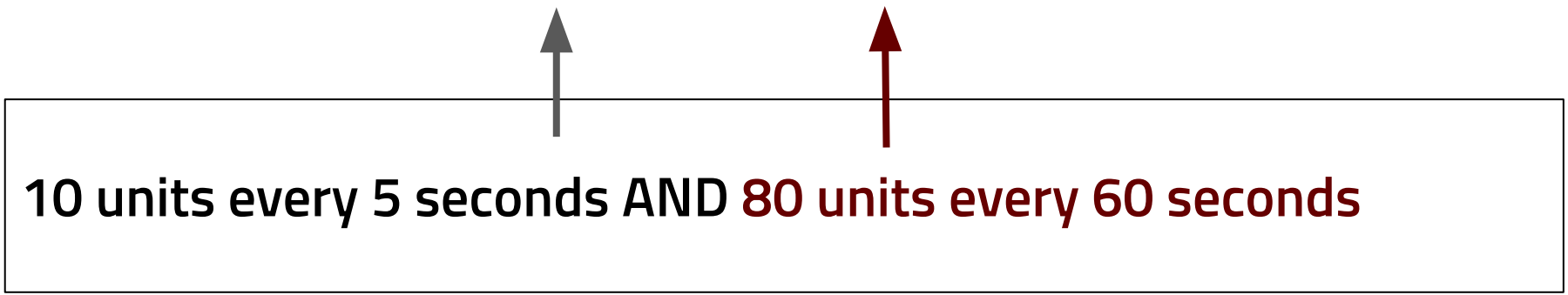
... and many more!

Extensibility via RateLimit-Policy

RateLimit-Limit: 10
RateLimit-Remaining: 6
RateLimit-Reset: 3

optional field with policy details and comments. Has Parameter registry.

RateLimit-Policy 10;w=5 , 80;w=60;comment="bar"



Technical choices

- #60 support **only delta-seconds** (no ntp skew & adjustment issues) like Retry-After
- #35 Use Structured-Fields

- flexible semantics to express dynamic policies, sliding windows and concurrency limits
- don't mention infrastructural concepts like connections

https://github.com/ietf-wg-httpapi/ratelimit-headers/pulls?q=is%3Apr+is%3Aclosed

Changes from -04

- #87 Added privacy considerations
- #103 No x-prefix in parameters
- #97,#98 Ignore malformed and cached fields
- #79 separate quota policies from limit
- Significant editorial refactoring, including referencing RFC9110

https://github.com/ietf-wg-httpapi/ratelimit-headers/pulls?q=is%3Apr+is%3Aclosed
https://www.ietf.org/rfcdiff?url2=draft-ietf-httpapi-ratelimit-headers-04

Open Issues before WGLC

- #105 No trailers: no one uses them for rate limiting. Defer to future specs.
- #65 Field names and combination proposes to replace the 3 basic fields to one field for conciseness

    RateLimit: limit=10, remaining=5, reset=60

#65 Field names and combination

Replace the 3 basic fields to one field for conciseness

RateLimit: limit=10, remaining=5, reset=60

Editors tried to accommodate this request (ML thread) but finally decided to not support this change. Motivations include:

- all platforms and middlewares decided to use separate, integer-valued fields for ratelimit, so there is no operational experience on providing/consuming this information in combined form;
- when in this I-D we combined different information in the same field, (e.g. in RateLimit-Limit) we were asked to separate them;
- SF parser are not available in current off-the-shelf tools like httpd/nginx or API gateways, and we do not advise processing non-bare-item SF with regexp

FAQ

Q: Are we inventing a new service management model?

A: No. We just standardize headers semantic for the many who *already* use this pattern.

Q: Why don't use timestamps for RateLimit-Reset?

A: Timestamps *require* NTP on both sides. NTP in the real world is hard (skew, adjust, IoT, ...). We like Retry-After too ;)

Thanks!

Roberto Polli - robipolli@gmail.com

Alex Martinez - alex@flawedcode.org

Backup slides