HTTP retries, when, why, how?

Problem

- conflict between HTTP requirements and practices in field, with pressure from customer asking to explicitly violate the spec, but till what point ?

- failures are more and more common with instant-on/off services and users really expect the LB to retry because they can't guarantee anymore that they won't be breaking traffic!

- some LBs alledgedly violate the spec by retrying anything by default and present this as a modern behaviour, further increasing end-user demand.

What the standard says

Currently, RFC7230#6.3.1 is pretty clear :

- A proxy MUST NOT automatically retry non-idempotent requests.

- A user agent MUST NOT automatically retry a request with a non-idempotent method unless it has some means to know that the request semantics are actually idempotent, regardless of the method, or some means to detect that the original request was never applied.

Idea #1

- we could reuse the new 425 code ("too early") developed for TLS Early Data and extend it to any type of browser retry.

Note: I predict that many users will still not like to have this one alone as it provides a visibility of their failures.

Idea #2

- if the user agent has some means to know a request is idempotent and wishes that intermediaries retry it, it could possibly pass this information in a header field down the chain. This would not significantly inflate the traffic as it would only be on POST requests in practice, with a constant name+value couple (i.e. 1 byte in HPACK)

- we could then discuss with whatwg to add an application-to-client signal in HTML forms to explicitly mention the request is idempotent thus retryable.

Proposal: mix of this

- an application should signal known retriable requests in the forms

- when the user agent *knows* an intermediary may retry a request using an apparently non-idempotent method, it should add the header field

- when an intermediary faces a failure on a non-idempotent request, it may use 425 to let the client retry instead of returning a proxy error

- when an intermediary faces a condition forcing it to retry such a request tagged by the client, it should mention it in a response header field. This *could* help automated clients slow down on certain types of requests (e.g. auto-completion, file submissions, etc).