Marking HTTP Requests as Unimportant

draft-thomson-http-nice-02

Abstract

An HTTP "Nice" header field is defined. "Nice" marks a request as low priority. Gateways can choose to discard or delay the request, or provide a response from cache rather than forwarding it to an origin server. This enables constrained origin servers, such as those that rely on battery power, to avoid expending limited resources on serving requests.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on May 14, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of
1. Introduction

HTTP [RFC2616] servers are beginning to appear as the interface to a wide array of devices. Management interfaces in many devices have classically been provided as HTTP servers, but this trend now extends to HTTP APIs on a range of devices, including constrained devices. Constrained devices are those with limited processing power, network connectivity or battery capacity.

The Constrained Application Protocol (CoAP) [RFC7252] in particular is designed to provide devices with extremely limited capabilities a way to provide an HTTP-compatible interface to the information and services they provide. A CoAP-HTTP gateway [I-D.ietf-core-http-mapping] provides HTTP-capable clients a means of accessing these devices.

For a device that operates based on a battery, it is often crucial that the device remain dormant for extended periods. Radio communication in particular consumes a significant amount of power. Frequent communication limits the length of time that the device can operate. It is often the case that communication can be initiated, but this could require a significant expenditure of stored energy.

Many constrained devices rely on intermediaries such as the CoAP-HTTP gateway to terminate requests and mediate access. Clients that access the services provided by such limited devices can be unaware of the limited nature of the device serving the request, since they actually interact with the gateway. Even when the client is aware of these limitations, it is not always possible for clients to learn
whether any given request would cause significant expenditure of resources at the constrained device.

A push server [I-D.thomson-webpush-http2] provides similar functionality for devices with constrained resources. Providing a standard way to indicate that a request is not urgent allows a push server to discriminate between requests that the sender considers urgent and those that are unimportant.

This document defines an HTTP header field, "Nice" that can be used by clients to indicate that a request is not urgent or important enough to cause a constrained server to expend special effort to serve. An gateway that is aware that the origin server is unable to handle the request can instead terminate the request. The request is forwarded as normal to an origin server that is available.

An gateway can generate an error or 203 (Non-Authoritative Information) response in response to a nice request, avoiding the need to contact the constrained origin server. Alternatively, the gateway could delay the request until the origin server becomes available or serve a response from cache if that is possible.

No specific mechanism is defined for an origin server to inform gateways of absence or other indisposition.

1.1. Conventions and Terminology

At times, this document falls back on shorthands for establishing interoperability requirements on implementations: the capitalized words "MUST", "SHOULD" and "MAY". The meaning of these is described in [RFC2119].

The terms "intermediary" and "gateway" are defined in [RFC7230].

2. The Nice Header Field

The "Nice" header field indicates that a request is less important than a request that doesn't bear this header.

The value of the header field is a decimal number between 0 and 3 inclusive. Values greater than zero indicate increasing levels of unimportance. A lower value indicates greater urgency; for example, a value of 3 is less urgent or important than a value of 1. A value of 0 (or an absent "Nice" header field) indicates that the request is to be forwarded as normal.

Nice = "Nice" ":" ("3" / "2" / "1" / "0")
Multiple values for the header field MUST NOT be included. If multiple values are present, a gateway MAY choose to treat the request in any way it chooses.

For example, the following request indicates that it is not urgent:

GET /m HTTP/1.1
Host: device9710.example.net:11453
Nice: 2

An gateway might reject this request, indicating that the origin server is not available using a 503 status code.

HTTP/1.1 503 Service Unavailable
Date: Fri, 26 Jul 2013
Content-Type: text/plain;charset=utf8
Content-Length: 63

The server is asleep, don't disturb it unless it's urgent.

A key characteristic of this header field is that intermediaries and clients that do not understand its semantics treat requests so marked no different to any other requests. An intermediary that has no special information about the availability of the origin server will also forward the request. That means that requests from a client that does not include this header will always reach the origin server.

2.1. Policies for Treatment of Nice Requests

An origin server or gateway might use several inputs in determining the threshold at which a request is forwarded to the origin server. An origin server might either directly instruct the gateway about the threshold, or it might be provide specific information that can be used, in conjunction with knowledge the gateway has of the origin server, as input to an algorithm for determining the threshold. Potential inputs include:

- The relative cost of awakening a dormant server. Depending on the server, this cost may be assessed in different ways, including monetary, battery or time.

- The last time that the server was in active communication. Typically, wireless devices have a period of heightened availability just after sending or receiving data. During this period activation and communication can be significantly more efficient.
Application preferences or context. For example, a server might be configured to be more highly responsive to requests during certain times.

The following describes a potential set of policies regarding selection and treatment of "Nice" header field value:

nice: 1 = The client regards the request as relatively urgent, but not critically so. An gateway might use a heuristic with a moderate risk of false positives in determining whether the server is available. A gateway might also forward a request so marked to a dormant device that has a relatively low activation cost.

nice: 2 = The client regards the request as not urgent. A gateway might attempt to minimize the probability that it awakens a server, if it uses a heuristic in determining whether to forward requests.

nice: 3 = The client regards the request as being of trivial importance. A gateway might avoid forwarding requests unless there is strong indication that the origin server is available and willing to communicate.

Many different policies can be applied to the selection of a value for the "Nice" header field, as well as to the treatment of requests so marked. Specific applications might define a means for providing more specific policies.

2.2. Polling with Nice Requests

Marking a request as nice is quite useful for requests that do not require immediate action. Clients might wish to have the request fulfilled, but are willing to wait until the origin server is present. Such requests might be sent periodically until they succeed.

In some cases, origin server availability is predictable and known to the gateway. Some devices have predictable cycles of availability, which are used for brief bursts of communication. If the next time that the origin server is available is known, a gateway can include a "Retry-After" header field in a generated error response.

For example:

HTTP/1.1 503 Service Unavailable
Date: Fri, 26 Jul 2013 03:34:19 GMT
Retry-After: 4
2.3. Store and Forward

A gateway MAY hold requests for a limited amount of time, to be forwarded when the origin server becomes available (i.e., a "store and forward" mode of operation). Including a "Prefer" header field [RFC7240] with the "wait" tag provides the gateway information about how long the client is prepared to await a response. This could allow the gateway to reject the request immediately if the device is known to be unreachable for the entire duration.

Gateways MAY alternatively accept a request and return an immediate response, such as a 202 (Accepted) status code. Use of the "respond-async" token in the "Prefer" header field allows clients to expressly request this behavior.

3. Security Considerations

Lowering the priority with which a request is handled is unlikely to cause any special concern with respect to security.

Intermediaries that do not support the "Nice" header field might erroneously cache a response from an intermediary that handles the request without forwarding to the origin server. Intermediaries MUST NOT generate cacheable responses to requests containing an "Nice" header field. Intermediaries MAY however provide cached responses originally provided by the origin server.

4. IANA Considerations

The permanent message header field registry (see [RFC3864]) has been updated with the following registration:

Header field name: Nice
Applicable protocol: http
Status: standard
Author/Change controller: IETF
Specification document: this specification (Section 2)

5. Acknowledgements

The original idea for this header field was devised by Matthew Kaufman and Bruce Lowekamp, who realized the importance of making the header a negative rather than positive expression of priority.
6. References

6.1. Normative References


6.2. Informative References


Author's Address

Martin Thomson
Mozilla
331 E Evelyn Street
Mountain View, CA  94041
US

Email: martin.thomson@gmail.com