Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 29, 2010

S. Loreto
Ericsson
P. Saint-Andre
Cisco
G. Wilkins
Webtide
S. Salsano
Univ. of Rome "Tor Vergata"

Oct 26, 2009

Best Practices for the Use of Long Polling and Streaming in Bidirectional HTTP

draft-loreto-http-bidirectional-01

Status of this Memo

Copyright Notice

Abstract

There is widespread interest in using the Hypertext Transfer Protocol (HTTP) to enable asynchronous or server-initiated communication from a server to a client as well as from a client to a server. This document describes how to better use HTTP, as it exists today, to enable such "bidirectional HTTP" using "long polling" and "HTTP streaming" mechanisms.

Table of Contents

1. Introduction

The Hypertext Transfer Protocol [HTTP-1.1] is a request/response protocol. HTTP defines the following entities: clients, proxies, and servers. A client establishes connections to a server for the purpose of sending HTTP requests. A server accepts connections from clients in order to service HTTP requests by sending back responses. Proxies are intermediate entities that can be involved in the delivery of requests and responses from the client to the server and vice versa.

In the standard HTTP model, a server cannot initiate a connection with a client nor send an unrequested HTTP response to the client; thus the server cannot push asynchronous events to clients. Therefore, in order to receive asynchronous events as soon as possible, the client needs to poll the server periodically for new content. However, continual polling can consume significant bandwidth by forcing a request/response round trip when no data is available. It can also be inefficient because it reduces the responsiveness of the application since data is queued until the server receives the next poll request from the client.

To improve this situation, several server push programming mechanisms have been implemented in recent years. These mechanisms, which are often grouped under the common label "Comet" [COMET], enable a web server to send updates to clients without waiting for a poll request from the client. Such mechanisms can deliver updates to clients in a more timely manner while avoiding the latency experienced by client applications due to the frequent open and close connections necessary to periodically poll for data.

The two most common server push mechanisms are "Long Polling" and "HTTP Streaming":

Long Polling: The server attempts to "hold open" (not immediately reply to) each HTTP request, responding only when there are events to deliver. In this way, there is always a pending request available to use for delivering events as they occur, thereby minimizing the latency in message delivery.

HTTP Streaming: The server keeps a request open indefinitely; that is, it never terminates the request or closes the connection, even after it pushes data to the client.

It is possible to define other technologies for bidirectional HTTP, however such technologies typically require changes to HTTP itself (e.g., by defining new HTTP methods). This document focuses only on bidirectional HTTP technologies that work within the current scope of

HTTP as defined in [HTTP-1.1] and [HTTP-1.0].

The remainder of this document is organized as follows. Section 2 analyzes the "long polling" technique. Section 3 analyzes the "HTTP streaming" technique. Section 4 provides an overview of the specific technologies that use server-push technique. Section 5 lists best practices for bidirectional HTTP using existing technologies.

The preferred venue for discussion of this document is the hybi@ietf.org mailing list; visit <https://www.ietf.org/mailman/listinfo/hybi> for further information.

2. Long Polling

2.1. Definition

With the traditional or "short" polling technique, a client sends regular requests to the server and each request attempts to "pull" any available events or data. If there are no events or data available, the server returns an empty response and the client waits for a period before sending another poll request. The frequency depends on the latency which can be tolerated in retrieving updated information from the server. This mechanism has the drawback that the consumed resources (server processing and network) strongly depend on the acceptable latency in the delivery of updates from server to client. If the acceptable latency is low (e.g., on the order of seconds) then the frequency of the poll request can cause an unacceptable burden on the server, the network, or both.

By contrast with such "short polling", "long polling" attempts to minimize both latency in server-client message delivery and the processing/network resource as compared to normal polling techniques. The server achieves these efficiencies by responding to a request only when a particular event, status, or timeout has occurred. Once the server sends a long poll response, typically the client immediately sends a new long poll request. Effectively this means that at any given time the server will be holding open a long poll request, to which it replies when new information is available for the client. As a result, the server is able to asynchronously "initiate" communication.

The basic life cycle of an application using "long polling" is as follows:

1. The client makes an initial request and then waits for a response.

2. The server defers its response until an update is available, or a particular status or timeout has occurred.

3. When an update is available, the server sends a complete response to the client.

4. The client typically sends a new long poll request, either immediately or after a pause to allow an acceptable latency period.

The long polling mechanism can be applied to either persistent or non-persistent HTTP connections. The use of persistent HTTP connections will avoid the additional overhead of establishing a TCP/IP connection for every long poll.

2.2. Long Polling Issues

The long polling mechanism introduces the following issues.

Header Overhead: With the long polling technique, every long poll request and long poll response is a complete HTTP message and thus contains a full set of HTTP headers in the message framing. For small infrequent messages, the headers can represent a large percentage of the data transmitted. This does not introduce significant technical issues if the network MTU allows all the information (including the HTTP header) to fit within a single IP packet. On the other hand, it can introduce business issues related to data cost, as the amount of transferred data can be significantly larger than the real payload carried by HTTP.

Maximal Latency: After a long polling response is sent to a client, the server must wait for the next long polling request before another message can be sent to the client. This means that while the average latency of long polling is close to one network transit, the maximal latency is over three network transits (long poll response, next long poll request, long poll response). However, because HTTP is carried on TCP/IP, packet loss and retransmission can occur, so maximal latency for any TCP/IP protocol will be more than three network transits (lost packet, next packet, negative ack, retransmit).

Connection Establishment: A common criticism of both short polling and long polling is that these mechanisms frequently open TCP/IP connections and then close them. However, both polling mechanisms work well with persistent HTTP connections that can be reused for many poll requests. Specifically, the short duration of the pause between a long poll response and the next long poll request avoids the closing of idle connections.

Allocated Resources: Operating systems and network appliances will allocate resources to TCP/IP connections and to HTTP requests outstanding on those requests. The long polling mechanism requires that for each client, both a TCP/IP connection and an HTTP request are held open. Thus it is important to consider the resources related to both of these when sizing a long polling application. Typically the resources used per TCP/IP connection are minimal and can scale reasonably. Frequently the resources allocated to HTTP requests can be significant, and scaling the total number of requests outstanding can be limited on some gateways, proxies, and servers.

Graceful Degradation: A long polling client or server that is under load has a natural tendency to gracefully degrade in performance at a cost of message latency. If load causes either a client or server to run slowly, then events to be pushed to clients will queue (waiting either for a long poll request or for available CPU to use a held long poll request). If multiple messages are queued for a client, then they may be delivered in a batch within a single long poll response. This can significantly reduce the per-message overhead and thus ease the work load of the client or server for the given message load.

3. HTTP Streaming

3.1. Definition

The "HTTP streaming" mechanism keeps a request open indefinitely. It never terminates the request or closes the connection, even after the server pushes data to the client. This mechanism significantly reduces the network latency because the client and the server do not need to open and close the connection.

The basic life cycle of an application using "HTTP streaming" is as follows:

1. The client makes an initial request and then waits for a response.

2. The server defers the response to a poll request until an update is available, or a particular status or timeout has occurred.

3. Whenever an update is available, the server sends it back to the client as a part of the response.

4. The data sent by the server does not terminate the request or the connection. The server returns to step 3.

The HTTP streaming mechanism is based on the capability of the server to send several pieces of information on the same response, without terminating the request or the connection. This result can be achieved by both HTTP/1.1 and HTTP/1.0 servers.

A HTTP response content length can be defined using 3 options:

Content-Length header: This indicates the size of the entity body in the message, in bytes.

Transfer-Encoding header: The ’chunked’ valued in this header indicates the message will be break into chunks of known size.

End of File (EOF): This is actually the default approach for HTTP/1.0 where the connections are not persistent. Clients do not need to know the size of the body they are reading; instead they expect to read the body until the server closes the connection. Although with HTTP/1.1 the default is for persistent connections, it still possible to use EOF by setting the ’Connection:close’ header in either the request or the response, to indicate that the connection should not be considered ’persistent’ after the current request/response is complete. The client’s inclusion of the ’Connection: close’ header field in the request will also prevent pipelining.

The main issue with EOF is that it is difficult to tell the difference between a connection terminated by a fault and one that is correctly terminated.

An HTTP/1.0 server can use only EOF as a streaming mechanism. By contrast, both EOF and "chunked transfer" are available to an HTTP/1.1 server.

The "chunked transfer" mechanism is the one typically used by HTTP/1.1 servers for streaming. It does so by including the header "Transfer-Encoding: chunked" at the beginning of the response, which enables it to send the following parts of the response in different "chunks" over the same connection. Each chunk starts with the hexadecimal expression of the length of its data, followed by CR/LF (the end of the response is indicated with a chunk of size 0).

```
HTTP/1.1 200 OK
Content-Type: text/plain
Transfer-Encoding: chunked

25
This is the data in the first chunk

1C
and this is the second one

0
```

Figure 1: Transfer-Encoding response

A HTTP/1.0 server will omit the Content-Length header in the response to achieve the same result, so it will be able to send the following parts of the response on the same connection (in this case the different parts of the response are not explicitly separated by HTTP protocol, and the end of the response is achieved by closing the connection).

3.2. HTTP Streaming Issues

The HTTP streaming mechanism introduces the following issues.

Network Intermediaries: The HTTP protocol allows for intermediaries (proxies, transparent proxies, gateways, etc.) to be involved in the transmission of a response from server to the client. There is no requirement for an intermediary to immediately forward a partial response and it is legal for it to buffer the entire response before sending any data to the client (e.g., caching transparent proxies). HTTP streaming will not work with such intermediaries.

Maximal Latency: Theoretically, on a perfect network, an HTTP streaming protocol’s average and maximal latency is one network transit. However, in practice the maximal latency is higher due to network and browser limitations. The browser techniques used to terminate HTTP streaming connections are often associated with JavaScript and/or DOM elements that will grow in size for every message received. Thus in order to avoid unlimited memory growth in the client, an HTTP streaming implementation must occasionally terminate the streaming response and send a request to initiate a new streaming response (which is essentially equivalent to a long poll). Thus the maximal latency is at least three network transits. Also, because HTTP is carried on TCP/IP, packet loss and retransmission can occur, so maximal latency for any TCP/IP protocol will be more than three network transits (lost packet, next packet, negative ack, retransmit).

Client Buffering: There is no requirement in existing HTTP specifications for a client library to make the data from a partial HTTP response available to the client application. For example, if each response chunk contains a statement of JavaScript, there is no requirement in the browser to execute that JavaScript before the entire response is received. However, in practice most browsers do execute JavaScript received in partial responses, but some require a buffer overflow to trigger execution, so blocks of white space can be sent to achieve buffer overflow.

Framing Techniques: Using HTTP streaming, several application messages can be sent within a single HTTP response. The separation of the response stream into application messages needs to be perfomed at the application level and not at the HTTP level. In particular it is not possible to use the HTTP chunks as application message delimiters, since intermediate proxies might "re-chunk" the message stream (for example by combining different chunks into a longer one). This issue does not affect the long polling technique, which provides a canonical framing technique: each application message can be sent in a different HTTP response.

4. Overview of Technologies

This section provides an overview of how the specific technologies that implement server-push mechanisms employ HTTP to asynchronously deliver messages from the server to the client.

4.1. Bayeux

The Bayeux protocol [BAYEUX] was developed in 2006-2007 by the Dojo Foundation. Bayeux can use both the long polling and HTTP streaming mechanisms.

In order to achieve bidirectional communications, a Bayeux client will use two HTTP connections to a Bayeux server so that both server-to-client and client-to-server messaging can occur asynchronously.

The Bayeux specification requires that implementations control pipeling of HTTP requests, so that requests are not pipelined inappropriately (e.g., a client-to-server message pipelined behind a long poll request).

In practice, for JavaScript clients, such control over pipelining is not possible in current browsers. Therefore JavaScript implementations of Bayeux attempt to meet this requirement by limiting themselves to a maximum of two outstanding HTTP requests at any one time, so that browser connection limits will not be applied and the requests will not be queued or pipelined. While broadly effective, this mechanism can be disrupted by non-Bayeux JavaScript simultaneously issuing requests to the same host.

Bayeux connections are negotiated between client and server with handshake messages that allow the connection type, authentication method, and other parameters to be agreed upon between the client and the server. Furthermore, during the handshake phase, the client and the server reveal to each other their acceptable bidirectional techniques and the client selects one from the intersection of those sets.

For non-browser or same-domain Bayeux, clients use HTTP POST requests to the server for both the long poll request and the request to send messages to the server. The Bayeux protocol packets are sent as the body of the HTTP messages using the "text/json; charset=utf-8" MIME content type.

For browsers that are operating in cross-domain mode, Bayeux clients use the "script src Ajax" ("AJAST") mechanism as described at <http://en.wikipedia.org/wiki/AJAST_(programming)>.

Client-to-server messages are sent as encoded JSON on the URL query parameters.

Server-to-client messages are sent as a JavaScript program that wraps the message JSON with a JavaScript function call to the already loaded Bayeux implementation.

4.2. BOSH

BOSH, which stands for Bidirectional-streams Over Synchronous HTTP [BOSH], was developed by the XMPP Standards Foundation in 2003-2004. The purpose of BOSH is to emulate normal TCP connections over HTTP (TCP is the standard connection mechanism used in the Extensible Messaging and Presence Protocol as described in [XMPP]). BOSH employs the long polling mechanism by allowing the server (called a "BOSH connection manager") to defer its response to a request until it actually has data to send to the client from the application server itself (typically an XMPP server). As soon as the client receives a response from the connection manager, it sends another request to the connection manager, thereby ensuring that the connection manager is (almost) always holding a request that it can use to "push" data to the client.

In some situations, the client needs to send data to the server while it is waiting for data to be pushed from the connection manager. To prevent data from being pipelined behind the long poll request that is on hold, the client can send its outbound data in a second HTTP request. BOSH forces the server to respond to the request it has been holding on the first connection as soon as it receives a new request from the client, even if it has no data to send to the client. It does so to make sure that the client can send more data immediately if necessary even in the case where the client is not able to pipeline the requests, respecting at the same time the two-connection limit discussed here under Section 5.1.

The number of long polling request-response pairs is negotiated during the first request sent from the client to the connection manager. Typically BOSH clients and connection managers will negotiate the use of two pairs, although it is possible to use only one pair or to use more than two pairs.

The roles of the two response-response pairs typically switch whenever the client sends data to the connection manager. This means that when the client issues a new request, the connection manager immediately answers to the blocked request on the other TCP connection, thus freeing it; in this way, in a scenario where only the client sends data, all the even requests are sent over one connection and the odd ones are sent over the other connection.

BOSH is able to work reliably both when network conditions force every HTTP request to be made over a different TCP connection and when it is possible to use HTTP/1.1 and then relay on two persistent TCP connections.

If the connection manager has no data to send to the client for an agreed amount of time (also negotiated during the first request), then the connection manager will respond to the request it has been holding with no data, and that response immediately triggers a fresh client request. The connection manager does so to ensure that if a network connection is broken then both parties will realise that fact within a reasonable amount of time.

Moreover BOSH defines the negotiation of an "inactivity period" value that specifies the longest allowable inactivity period (in seconds). This enables the client to ensure that the periods with no requests pending are never too long.

BOSH allows data to be pushed immediately when HTTP Pipelining is available. However if HTTP Pipelining is not available and one of the endpoints has just pushed some data, BOSH will usually need to wait for a network round trip time until it is able to push again.

BOSH uses standard HTTP POST request and response bodies to encode all information.

BOSH normally uses HTTP Pipelining over a persistent HTTP/1.1 connection. However, a client can deliver its POST requests in any way permitted by HTTP 1.0 or HTTP 1.1.

BOSH clients and connection managers are not allowed to use Chunked Transfer Coding, since intermediaries might buffer each partial HTTP request or response and only forward the full request or response once it is available.

BOSH allows the usage of the Accept-Encoding and Content-Encoding headers in the request and in the response respectively, and then compresses the response body accordingly.

Each BOSH session can share the HTTP connection(s) it uses with other HTTP traffic, including other BOSH sessions and HTTP requests and responses completely unrelated to the BOSH protocol (e.g., web page downloads).

4.3. Server-Sent Events

W3C Server-Sent Events specification [W3C.WD-eventsource-20090423] defines an API that enables servers to push data to Web pages over HTTP in the form of DOM events.

The data is encoded as text/event-stream content and pushed using a HTTP streaming mechanism, but the specification suggests to disable HTTP chunking for serving event streams unless the rate of messages is high enough to avoid the possible negative effects of this technique as described here under Section 3.2.

However it is not clear the benefit of using EOF rather than chunking with regards to intermediaries, unless they are HTTP/1.0.

5. HTTP Best Practices

5.1. Two Connection Limit

HTTP [HTTP-1.1] section 8.1.4 recommends that a single user client should not maintain more than two connections to any server or proxy, to prevent the server from being overloaded.

Web applications need to limit the number of long poll requests initiated, ideally to a single long poll that is shared between frames, tabs, or windows of the same browser. However the security constraints of the browsers make such sharing difficult.

A possible best practice is for a server to use cookies to detect multiple long poll requests from the same browser and to avoid deferring both requests since this might cause connection starvation and/or pipeline issues.

5.2. Pipelined Connections

HTTP [HTTP-1.1] permits optional request pipelining over persistent connections. Multiple requests can be enqueued before the responses arrive.

There is a possible open issue regarding the inability to control "pipelining". Normal requests can be pipelined behind a long poll, and are thus delayed until the long poll completes.

5.3. Proxies

Most proxies work well with long polling, because a complete HTTP response must be sent either on an event or a timeout. Proxies should return that response immediately to the user-agent, which immediately acts on it.

The HTTP streaming mechanism uses partial responses and sends some JavaScript in an HTTP/1.1 chunk as described under Section 3. This mechanism can face problems caused by two factors: (1) it relies on proxies to forward each chunk (even though there is no requirement for them to do so, and some caching proxies do not), and (2) it relies on user-agents to execute the chunk of JavaScript as it arrives (even though there is also no requirement for them to do so).

A "reverse proxy" basically is a proxy that pretends to be the actual server (as far as any client or client proxy is concerned), but it passes on the request to the actual server that is usually sitting behind another layer of firewalls. Any short polling or long polling Comet solution should work fine with this, as will most streaming Comet connections. The main downside is performance, since most proxies are not designed to hold many open connections (as a dedicated Comet server is).

Reverse proxies can come to grief when they try to share connections to the servers between multiple clients. As an example, Apache with mod_jk shares a small set of connections (often 8 or 16) between all clients. If long polls are sent on those shared connections, then the proxy can be starved of connections, which means that other requests (either long poll or normal) can be held up. Thus Comet mechanisms currently need to avoid any connection sharing -- either in the browser or in any intermediary -- because the HTTP assumption is that each request will complete as fast as possible.

Much of the "badness" of both long polling and HTTP streaming for servers and proxies results from using a synchronous programming model for handling requests, since the resources allocated to each request are held for the duration of the request. Asynchronous proxies and servers can handle Comet long polls with few resources above that of normal HTTP traffic. Unfortunately some synchronous proxies do exist (e.g., apache mod_jk) and many HTTP application servers also have a blocking model for their request handling (e.g., the Java servlet 2.5 specification).

5.4. HTTP Responses

The server responds to a request successfully received by sending a 200 OK answer, but only when a particular event, status, or timeout has occurred. The 200 OK body section contains the actual event, status, or timeout that occurred.

5.5. Timeouts

The long polling mechanism allows the server to respond to a request only when a particular event, status, or timeout has occurred. In order to minimize as much as possible both latency in server-client message delivery and the processing/network resources needed, the long polling request timeout should be set to a high value.

However, the value timeout value has to be chosen carefully; indeed, there can be problem if this value is set too high (e.g., the client might receive a 408 Request Timeout answer from the server or a 504 Gateway Timeout answer from a proxy). The default timeout value in a browser is 300 seconds, but most network infrastructures have proxies and server that do not have such a long timeout.

Several experiments have shown success with timeouts as high as 120 seconds, but generally 30 seconds is a safer value. Therefore it is recommended that all network equipment wishing to be compatible with the long polling mechanism should implement a timeout substantially greater than 30 seconds (where "substantially" means several times more than the medium network transit time).

5.6. Network Impact

To follow.

6. Future Work

This document focuses on best practices for bidirectional HTTP in the context of HTTP as it exists today. Future documents might define additions to HTTP that could enable improved mechanisms for bidirectional HTTP. Examples include:

- An HTTP extension for long polling, including request tracking, duplication, and retry methods.

- A method for monitoring the state of multiple resources.

- A request header to determine timeouts.

- A request header to determine the longest acceptable polling interval.

- Improved rendezvous logic between the user agent, a proxy / connection manager, and the backend application server.

- Improved addressing for the entities involved in bidirectional HTTP, possibly including the use of URI templates.

- Possible improvements/extensions to XMLHttpRequest (XHR) API [W3C.WD-XMLHttpRequest2-20090820] to expose connection-handling details (e.g., use of pconns, pipelining, etc.)

7. Acknowledgments

Thanks to Joe Hildebrand, Mark Nottingham, and Martin Tyler for their feedback.

8. IANA Considerations

This document does not require any actions by the IANA.

9. Security Considerations

To follow.

10. Informative References

[BAYEUX] Russell, A., Wilkins, G., Davis, D., and M. Nesbitt, "Bidirectional-streams Over Synchronous HTTP (BOSH)", 2007.

[BOSH] Paterson, I., Smith, D., and P. Saint-Andre, "Bidirectional-streams Over Synchronous HTTP (BOSH)", XSF XEP 0124, February 2007.

[COMET] Russell, A., "Comet: Low Latency Data for the Browser", March 2006.

[HTTP-1.0] Berners-Lee, T., Fielding, R., and H. Nielsen, "Hypertext Transfer Protocol -- HTTP/1.0", RFC 1945, May 1996.

[HTTP-1.1] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., and T. Berners-Lee, "Hypertext Transfer Protocol -- HTTP/1.1", RFC 2616, June 1999.

[W3C.WD-XMLHttpRequest2-20090820] Kesteren, A., "XMLHttpRequest Level 2", World Wide Web Consortium WD WD-XMLHttpRequest2-20090820, August 2009, <http://www.w3.org/TR/2009/WD-XMLHttpRequest2-20090820>.

[W3C.WD-eventsource-20090423] Hickson, I., "Server-Sent Events", World Wide Web Consortium WD WD-eventsource-20090423, April 2009, <http://www.w3.org/TR/2009/WD-eventsource-20090423>.

[XMPP] Saint-Andre, P., Ed., "Extensible Messaging and Presence Protocol (XMPP): Core", RFC 3920, October 2004.

Authors’ Addresses

Salvatore Loreto
Ericsson
Hirsalantie 11
Jorvas 02420
Finland

Email: salvatore.loreto@ericsson.com

Peter Saint-Andre
Cisco

Email: psaintan@cisco.com

Greg Wilkins
Webtide

Email: gregw@webtide.com

Stefano Salsano
Univ. of Rome "Tor Vergata"
Via del Politecnico, 1
Rome 00133
Italy

Email: stefano.salsano@uniroma2.it