CLAIMS:

1. A method for managing caching for a content delivery system, the method comprising:

receiving, by a business logic agent from a caching agent, a content request indicating that the caching agent is requesting content data for a client, wherein the business logic agent and the caching agent are operatively coupled via a network;

filling, by the business logic agent, the content data in a first cache storage of the business logic agent, wherein the content data is received from an upstream node of the content delivery system;

providing, by the business logic agent, the cached content data to the caching agent; and

while a second cache storage of the caching agent is being filled with the content data, maintaining, by the business logic agent, the cached content data in response to receiving additional content requests from the caching agent, wherein the additional content requests indicate that the caching agent is requesting the same content data for additional clients.

2. The method of claim 1, wherein

the content delivery system is a content delivery network (CDN);

the business logic agent and the caching agent are on a same edge node of the CDN;

the upstream node is an origin server that stores the content data or a node between the edge node and the origin node in the CDN that stores the content data; and

the network is a local host connection,

and optionally wherein the local host connection is over at least one of a hypertext transfer protocol (HTTP) connection and a HTTP/2 connection.

3. The method of claim 1 or 2, wherein

the content delivery system is a content delivery network (CDN);

the caching agent is on an edge node of the CDN;

the business logic agent is on an intermediate node of the CDN, wherein the intermediate node is different from the edge node; and

the upstream node is an origin server that stores the content data or a node between the edge node and the origin node in the CDN that stores the content data.

4. The method of any one of claims 1 to 3, wherein

the caching agent comprises a hypertext transfer protocol (HTTP) service engine configured to service HTTP requests received from clients; and

the caching agent comprises a caching engine, wherein the caching engine comprises the second cache storage.

5. The method of any one of claims 1 to 4, further comprising:

receiving, by the business logic agent from the caching agent, an authentication request before the content request is received;

authenticating that the content data is provided by a customer of the content delivery system; and

sending, by the business logic agent to the caching agent, an authentication response in response to authenticating that the content data is provided by the customer, wherein the authentication response comprises an authorization.

6. The method of any one of claims 1 to 5, wherein either:

(i) the content request identifies the upstream node; and

the business logic agent receives the content data from the upstream node based on the content request; or

(ii) the content request is addressed to the upstream node.

7. The method of any one of claims 1 to 6, wherein

the content request comprises an address of the content data and a range of the content data;

the address of the content data comprises a Uniform Resource Locator (URL); and

the business logic agent determines whether any of the additional content requests is requesting the same content data as requested by the content request based on the address and the range.

8. The method of any one of claims 1 to 7, wherein the content request comprises at least one of:

(i) a fill identification (ID), and the method further comprises:

determining, by the business logic agent, that a new copy of the content data is needed based on the fill ID; and

filling, by the business logic agent, the new copy of the content data in the first cache storage in response to determining that the new copy is needed; and

(ii) a conditional request header, and the method further comprises:

determining, by the business logic agent, a key corresponding to the content request based on the conditional request header.

9. The method of any one of claims 1 to 8, wherein while the second cache storage of the caching agent is being filled with the content data, maintaining the cached content data comprises maintaining the cached content data in response to determining that two consecutive ones of the additional content requests are received within a predetermined time interval.

10. The method of any one of claims 1 to 9, further comprising determining that no new additional content requests have been received for a predetermined time interval since a most recent one of the additional content requests has been received.

11. A method for managing caching for a content delivery system, the method comprising:

receiving, by a caching agent from a client, a request for content data;

determining, by the caching agent, that the caching agent needs to fill the content data;

sending, by the caching agent to a business logic agent, a content request indicating that the caching agent is requesting the content data for the client, wherein the business logic agent and the caching agent are operatively coupled via a network, and the business logic agent fills the content data in a first cache storage of the business logic agent responsive to the content request;

filling, by the caching agent, the content data from an upstream node in a second cache storage of the caching agent;

receiving, by the caching agent, additional content requests from additional clients, wherein the additional content requests indicate that the same content data is requested by additional clients; and

while the second cache storage is being filled with the content data, the cached content data is provided by the business logic agent to the additional clients.

12. The method of claim 11, wherein

the content delivery system is a content delivery network (CDN);

the business logic agent and the caching agent are on a same edge node of the CDN;

the upstream node is an origin server that stores the content data or a node between the edge node and the origin node in the CDN that stores the content data; and

the network is a local host connection,

and optionally wherein the local host connection is over at least one of a hypertext transfer protocol (HTTP) connection and a HTTP/2 connection.

13. The method of claim 11 or 12, wherein

the content delivery system is a content delivery network (CDN);

the caching agent is on an edge node of the CDN;

the business logic agent is on an intermediate node of the CDN, wherein the intermediate node is different from the edge node; and

the upstream node is an origin server that stores the content data or a node between the edge node and the origin node in the CDN that stores the content data.

14. The method of any one of claims 11 to 13, wherein

the caching agent comprises a hypertext transfer protocol (HTTP) service engine configured to service HTTP requests received from clients; and

the caching agent comprises a caching engine, wherein the caching engine comprises the second cache storage.

15. The method of any one of claims 11 to 14, further comprising:

sending, by the caching agent to the business logic agent, an authentication request before the content request is sent;

receiving, by the caching agent from the business logic agent to an authentication response, wherein the authentication response comprises an authorization.

16. The method of any one of claims 11 to 15, further comprising either:

(i) selecting, by the caching agent, the upstream node, wherein the content request identifies the upstream node; or

(ii) wherein the content request is addressed to the upstream node.

17. The method of any one of claims 11 to 16, wherein

the content request comprises an address of the content data and a range of the content data;

the address of the content data comprises a Uniform Resource Locator (URL); and

the business logic agent determines whether any of the additional content requests is requesting the same content data as requested by the content request based on the address and the range.

18. The method of any one of claims 11 to 17, wherein the content request comprises at least one of:

a fill identification (ID), and the fill ID indicates whether a new copy of the content data is needed to be filled in the first cache storage of the business logic agent; and

a conditional request header, wherein a key corresponding to the content request is determined by the business logic agent based on the conditional request header.

19. A node of a content delivery system, comprising:

a network device providing a network;

at least one processor implementing a business logic agent and a caching agent, the business logic agent and the caching agent are operatively coupled via the network; and

at least one memory implementing a first cache storage of the business logic agent and a second cache storage of the caching agent, wherein the at least one processor:

sends, using the caching agent to the business logic agent, a content request indicating that the caching agent is requesting content data for a client;

receives, using the business logic agent, the content data from an upstream node of the content delivery system;

fills, using the business logic agent, the content data in the first cache storage;

provides, using the business logic agent, the cached content data to the client; and

while the second cache storage is being filled with the content data, maintains, using the business logic agent, the cached content data in response to the business logic agent receiving additional content requests from the caching agent, wherein the additional content requests indicate that the caching agent is requesting the same content data for additional clients.

20. A non-transitory computer-readable medium comprising computer-readable instructions such that, when executed, cause a processor of a node of a content delivery system to:

send, using a caching agent to a business logic agent, a content request indicating that the caching agent is requesting content data for a client;

receive, using the business logic agent, the content data from an upstream node of the content delivery system;

fill, using the business logic agent, the content data in the first cache storage;

provide, using the business logic agent, the cached content data to the client; and

while the second cache storage is being filled with the content data, maintain, using the business logic agent, the cached content data in response to the business logic agent receiving additional content requests from the caching agent, wherein the additional content requests indicate that the caching agent is requesting the same content data for additional clients.