

Google LLC
v.
Sandpiper CDN, LLC

IPR2025-00860
U.S. Patent No. 10,924,573



Patent Owner's Demonstratives
August 26, 2026

Sandpiper EX2028
Google v. Sandpiper
IPR2025-00860

1

Instituted Grounds

Claims	Ground #	References
1-20	1	Newton-471
1-20	2	Dilley, Pai, Wang

Agenda

- ▶ **The '573 Patent**
- ▶ Independent Claims 1, 11, 19 (Ground 2)
 - ▶ “a popularity designation”
- ▶ Independent Claim 19 (Ground 2)
 - ▶ “requesting the resource ... based at least on the popularity designation”
- ▶ Independent Claims 1, 11 (Ground 2)
 - ▶ “a redirect instruction”
- ▶ Dependent claims 5, 15 (Ground 2)
 - ▶ peer-to-peer redirect
- ▶ Claim 1-20 (Ground 1) (LEAP)
 - ▶ “instruction to not cache”

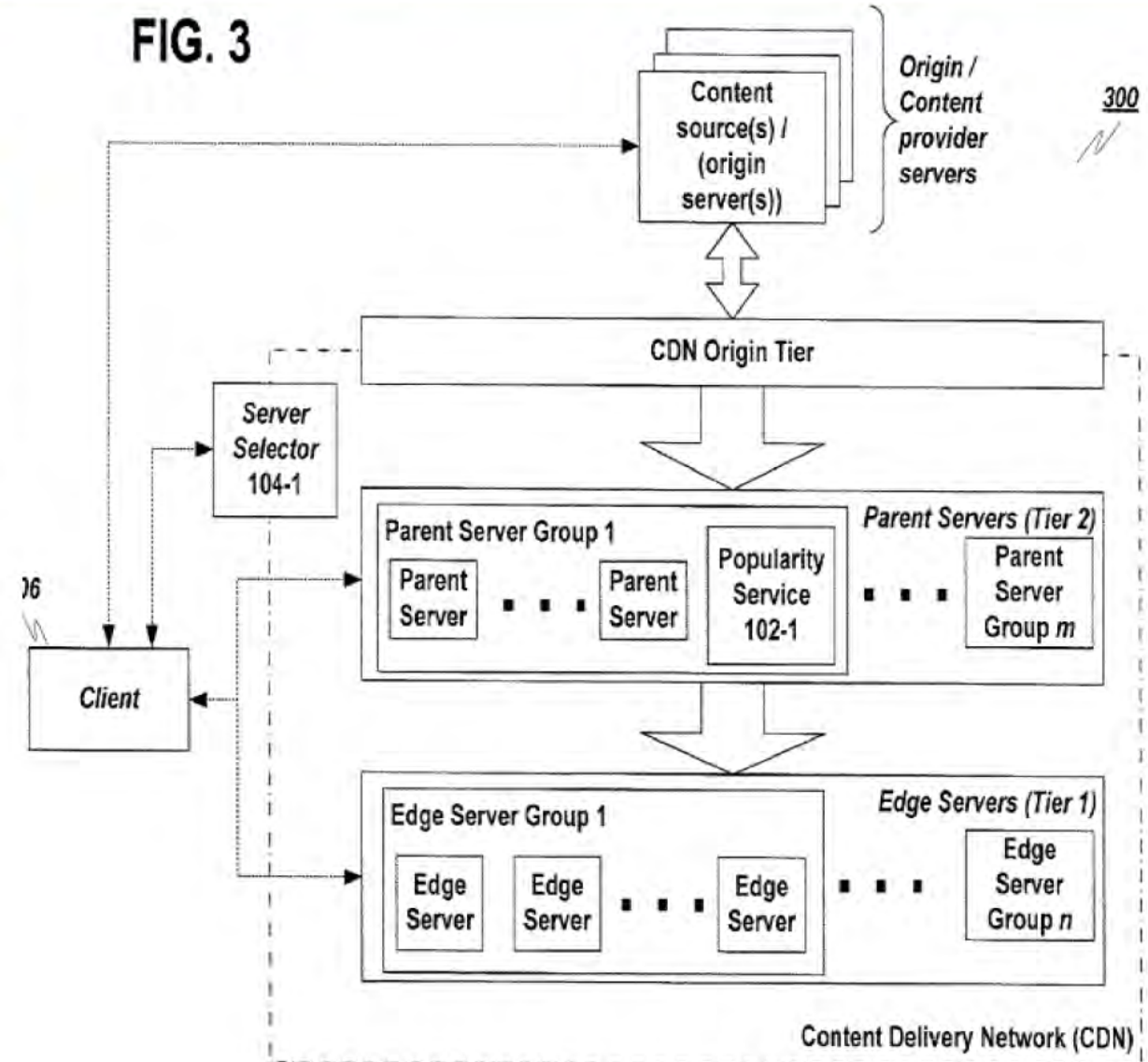
The '573 patent discloses a two-stage content delivery process based on a popularity designation.

'573 Patent

FIG. 3 shows a content delivery framework 300 with a two-level hierarchical CDN consisting of a tier of edge servers (Tier 1) and a tier of parent servers (Tier 2). Some or all of the edge servers can communicate with some or all of the parent servers. The edge servers are divided into n edge server groups, and the parent servers are divided into m parent server groups. In a presently preferred embodiment,

EX1001, '573 patent, 5:27-33.

FIG. 3



The '573 Patent discloses a two-stage content delivery process based on a popularity designation.

'573 Patent

Certain publishers can have large content libraries in which only a small proportion of the content (the so-called “short head”) is popular enough to benefit from serving through a caching CDN, while the majority of the content (the so-called “long tail”) is accessed only occasionally and not generally worth caching (or even serving from an edge server). This situation would be typical for a content pub-

EX1001, 2:49-55.

If the object is not available on the edge server 108 or on a peer, the edge server 108 ascertains whether this object is served based on its popularity (i.e., whether this object has been designated so that the object's popularity will be used to determine where it will be served from) (at 506). If so, then the request is sent to the popularity service 102 associated with the edge server 108, in this case, to the popularity server for the this group (S3a).

EX1001, 6:27-34.

The eligibility-based step (“popularity designation” determination of box 506) is not disclosed in any of the prior-art references.

POR, 2.

Patent Owner's Demonstrative Exhibit – Not Evidence

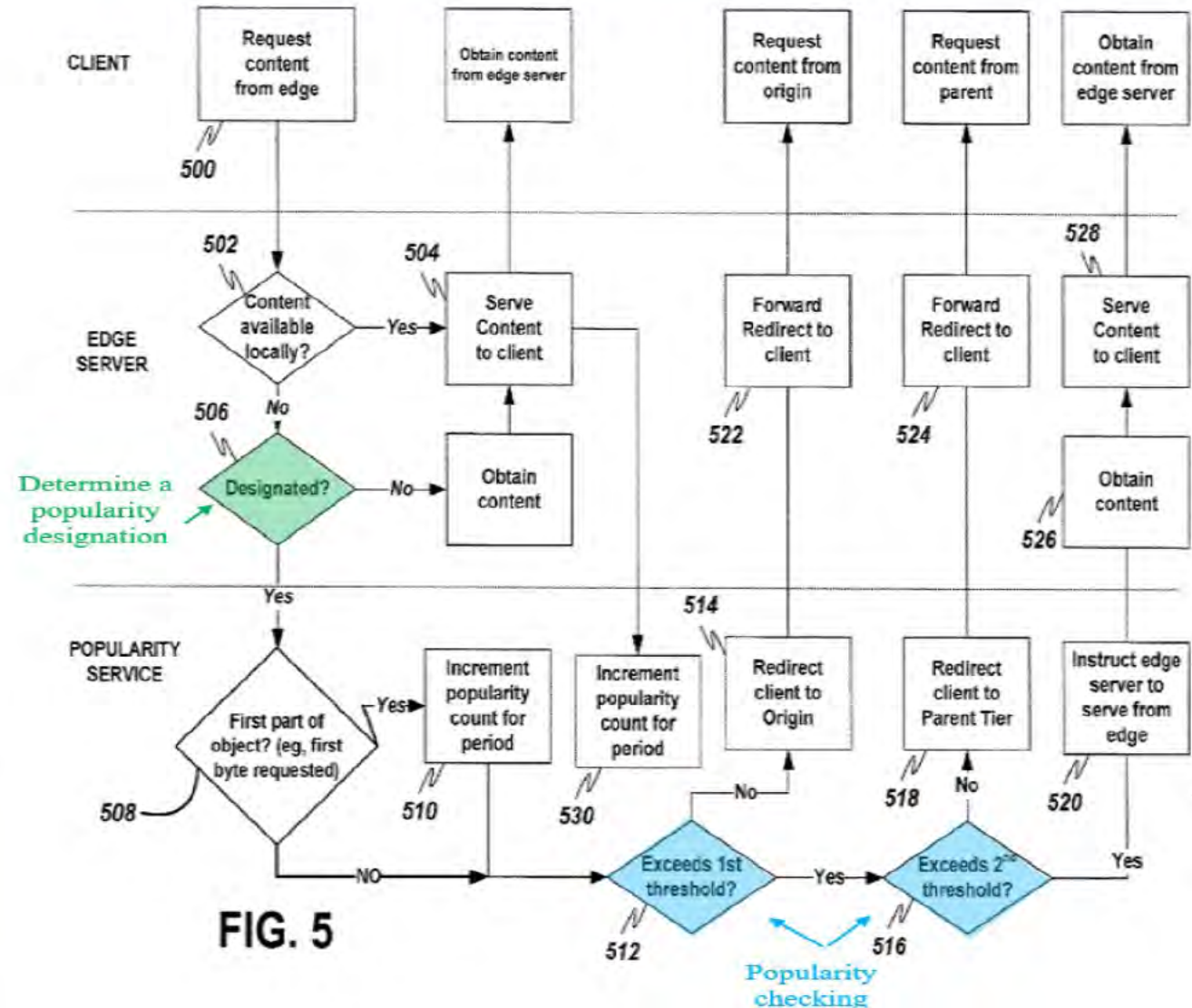


FIG. 5

Popularity checking

EX1001, FIG. 5 (annotated).

Agenda

- ▶ The '573 Patent
- ▶ **Independent Claims 1, 11, 19 (Ground 2)**
 - ▶ **"a popularity designation"**
- ▶ Independent Claim 19 (Ground 2)
 - ▶ "requesting the resource ... based at least on the popularity designation"
- ▶ Independent Claims 1, 11 (Ground 2)
 - ▶ "a redirect instruction"
- ▶ Dependent claims 5, 15 (Ground 2)
 - ▶ peer-to-peer redirect
- ▶ Claim 1-20 (Ground 1)
 - ▶ "instruction to not cache"

The '573 Patent, Claim 1

Claim 1	
1(preamble)	A method of content delivery in a content delivery network comprising:
1(a)	receiving, at a first server of a first tier of servers of the content delivery network, a request from a requesting device for a resource available from the content delivery network;
1(b)	accessing a popularity service associated with the content delivery network to determine a popularity designation associated with the requested resource;
1(c)	requesting the resource from a second server of the content delivery network;
1(d)	processing, at the first server of the first tier of servers, a redirect instruction from the second server of the content delivery network to obtain the resource from a content server of the content delivery network;
1(e)	receiving an instruction to not cache the portion of the resource at the first server of the first tier of servers when the portion of the resource is obtained from the content server of the content delivery network; and
1(f)	providing the obtained resource to the requesting device.

EX1001, Claim 1.

The '573 Patent, Claim 11

Claim 11	
11(preamble)	A content delivery network comprising:
11(a)	a first tier of servers comprising a first plurality of servers, an edge server of the first tier of servers receiving a request from a requesting device in communication with the first server for a resource available from the content delivery network;
11(b(i))	a second tier of servers comprising a second plurality of servers,
11(b(ii))	a first server of the second plurality of servers receiving a request from the edge server for the content and, in response, transmitting a redirect instruction intended for the requesting device to the edge server to obtain the resource from a content server and transmitting an instruction to not cache the portion of the resource at the edge server when the portion of the resource is obtained from the content server; and
11(c)	a popularity service tracking a popularity designation associated with the requested resource;
11(d)	wherein the edge server processes the redirect instruction to obtain the content from the content server and provides the obtained content to the requesting device.

EX1001, Claim 11.

The '573 Patent, Claim 19

Claim 19	
19(preamble)	A method of content delivery in a content delivery network comprising:
19(a)	receiving, at a first server of a first tier of servers of the content delivery network, a request from a requesting device for a resource available from the content delivery network;
19(b)	accessing a popularity service associated with the first server of the first tier of the content delivery network to determine a popularity designation associated with the requested resource;
19(c)	receiving an instruction to not cache the resource at the first server of the first tier of servers when the portion of the resource is obtained from the content server of the content delivery network;
19(d)	requesting the resource from a second server of the content delivery network based at least on the popularity designation associated with the requested resource; and
19(e)	providing the obtained resource to the requesting device.

EX1001, Claim 19.

Petitioner relies on Pai alone for the “popularity designation.”

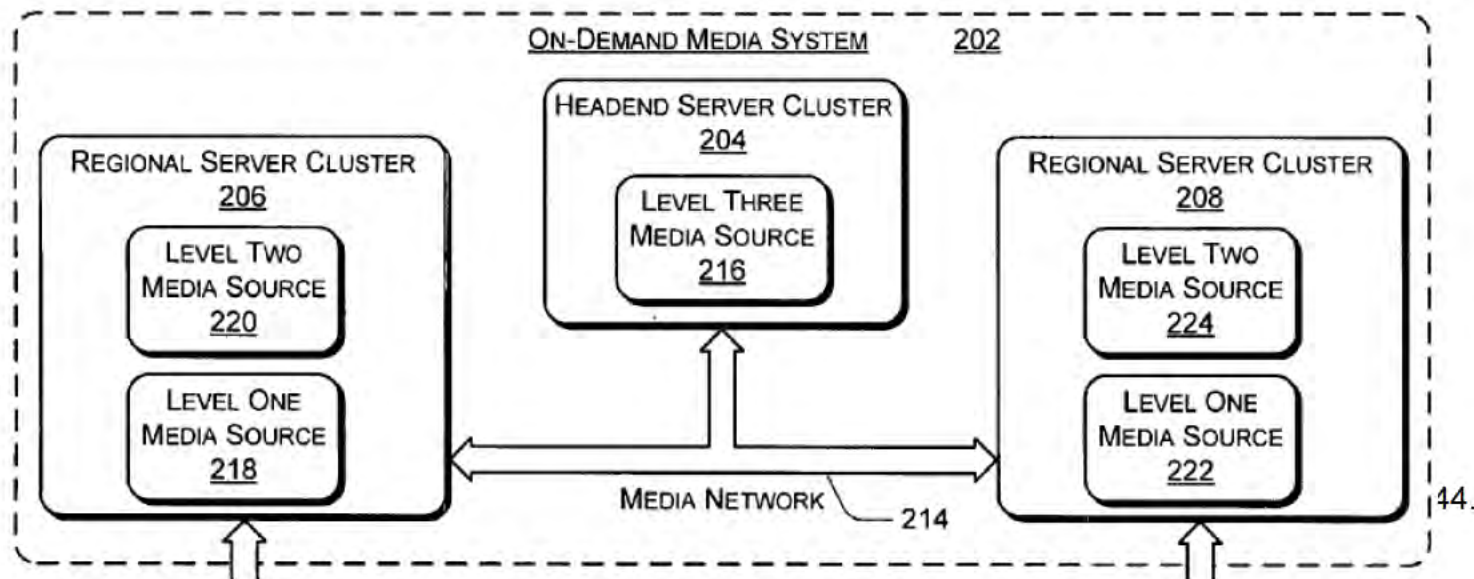
Petitioner's Expert

236. To the extent Dilley does not alone teach accessing a popularity service to determine a popularity designation associated with the requested resource, Pai does. As explained in Section VII.B.5 above, while Dilley teaches the advantages of using a hierarchical infrastructure to distribute popular content closer to the edge, it leaves to a POSITA how the popularity service would work in application, and a POSITA would have been motivated to implement Pai's teachings regarding its popularity service in Dilley's system.

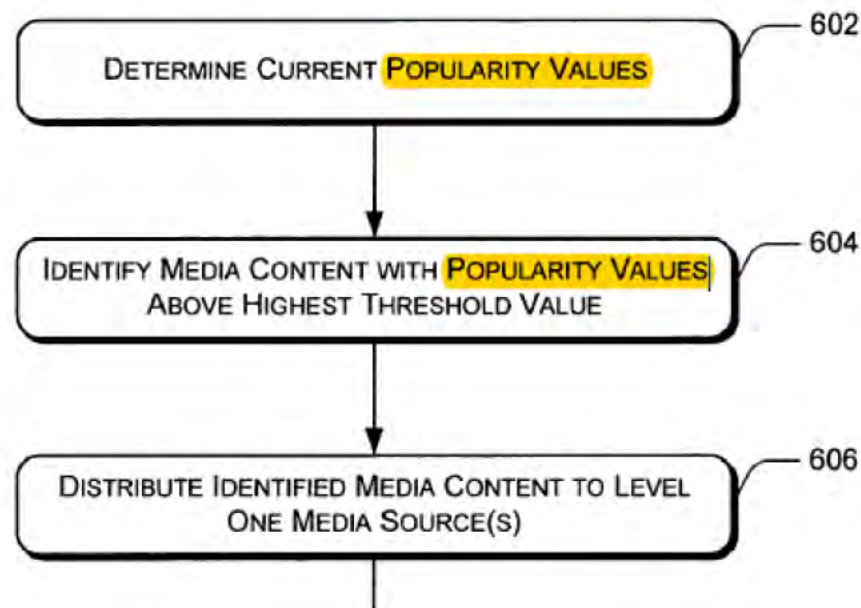
EX1006, Mowry Decl., ¶236.

Pai teaches media storage based on “popularity values.”

Pai’s on-demand media system stores media content based on popularity values.



EX1004, Pai, FIG. 2 (excerpt).



EX1004, FIG. 6 (excerpt).

[0039] At block 602 current popularity values associated with currently maintained media content are determined. For example, a server may re-calculate a popularity value for a particular media item based on a number of times that the media item has been requested within a particular time period. For example, a media item that has been requested more often (or by more users concurrently) may be assigned a higher popularity value than a media item that has been requested less often (or by fewer concurrent users).

EX1004, ¶39.

Pai's popularity value is a numerical value.

Petitioner conflates Pai's "popularity value" with the claimed "popularity designation."

Petition

Pai's popularity service monitors and calculates a content's popularity and, based on this calculated popularity value, dynamically adjusts the content's storage location to improve content delivery to users while efficiently managing the system's limited storage capacity. EX-1004, ¶¶[0013], [0020]. For instance, Pai's

Pet., 44.

Through monitoring and dynamically adjusting a content's popularity value, Pai discloses accessing a popularity service to determine the popularity associated with the requested resource. EX-1004, ¶¶[0039], [0047]. Indeed, content in Pai is

Pet., 45.

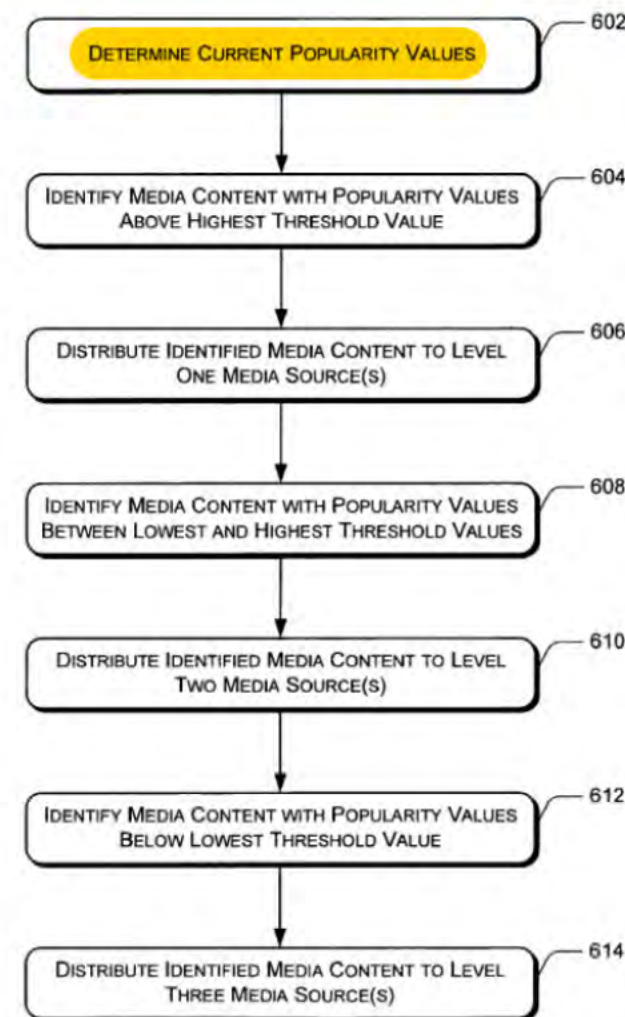


Figure 6

EX1004, Pai, FIG. 6.

The “popularity designation” determination is an eligibility check

A two-step process includes a prerequisite eligibility test.

'573 patent

The popularity service may be used by certain, though not necessarily all, content. When only certain content uses the popularity service, content should be designated in order to use the popularity service.

EX1001, 5:47-50.

When a client 106 requests content (e.g., using an HTTP GET request), that request is directed (e.g., by server selector 104-1) to an edge server in order for the content to be served to the client. For certain designated content a popularity check is interposed into the fill side of the caching operation. FIG. 4 illustrates the flow of messages and data in

EX1001, 5:59-64.

If the object is not available on the edge server 108 or on a peer, the edge server 108 ascertains whether this object is served based on its popularity (i.e., whether this object has been designated so that the object's popularity will be used to determine where it will be served from) (at 506). If so, then the request is sent to the popularity service 102 associated with the edge server 108, in this case, to the popularity server for the this group (S3a).

EX1001, 6:27-34.

Designation check happens *before* numerical popularity checking.

POR, 27; EX2016, Shenoy Decl., ¶65.

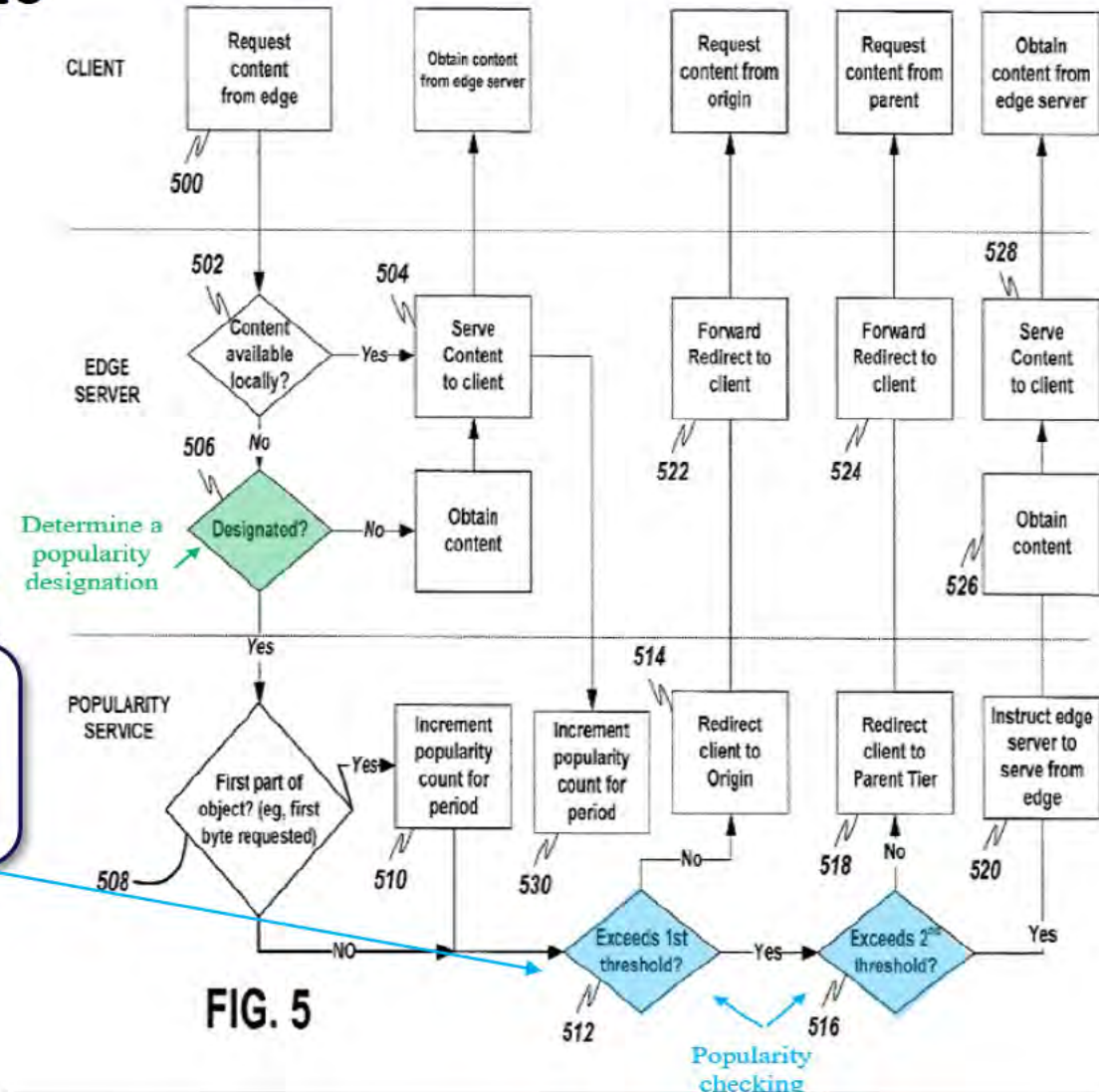


FIG. 5

Popularity checking

Pai's "popularity value" is not a "popularity designation."

Pai fails to teach the initial eligibility test step required by the claim.

Pai

[0039] At block 602 current popularity values associated with currently maintained media content are determined. For example, a server may re-calculate a popularity value for a particular media item based on a number of times that the media item has been requested within a particular time period. For example, a media item that has been requested more often (or by more users concurrently) may be assigned a higher popularity value than a media item that has been requested less often (or by fewer concurrent users).

EX1004, ¶39.

Petition

Pai assigns popularity values to each content item and regularly updates those values. EX-1004, ¶[0026]. Pai dynamically redistributes content between different storage levels (i.e., level 1, 2, and 3) on the network based on popularity thresholds. EX-1004, ¶¶[0018], [0038]-[0045].

Pet., 35.

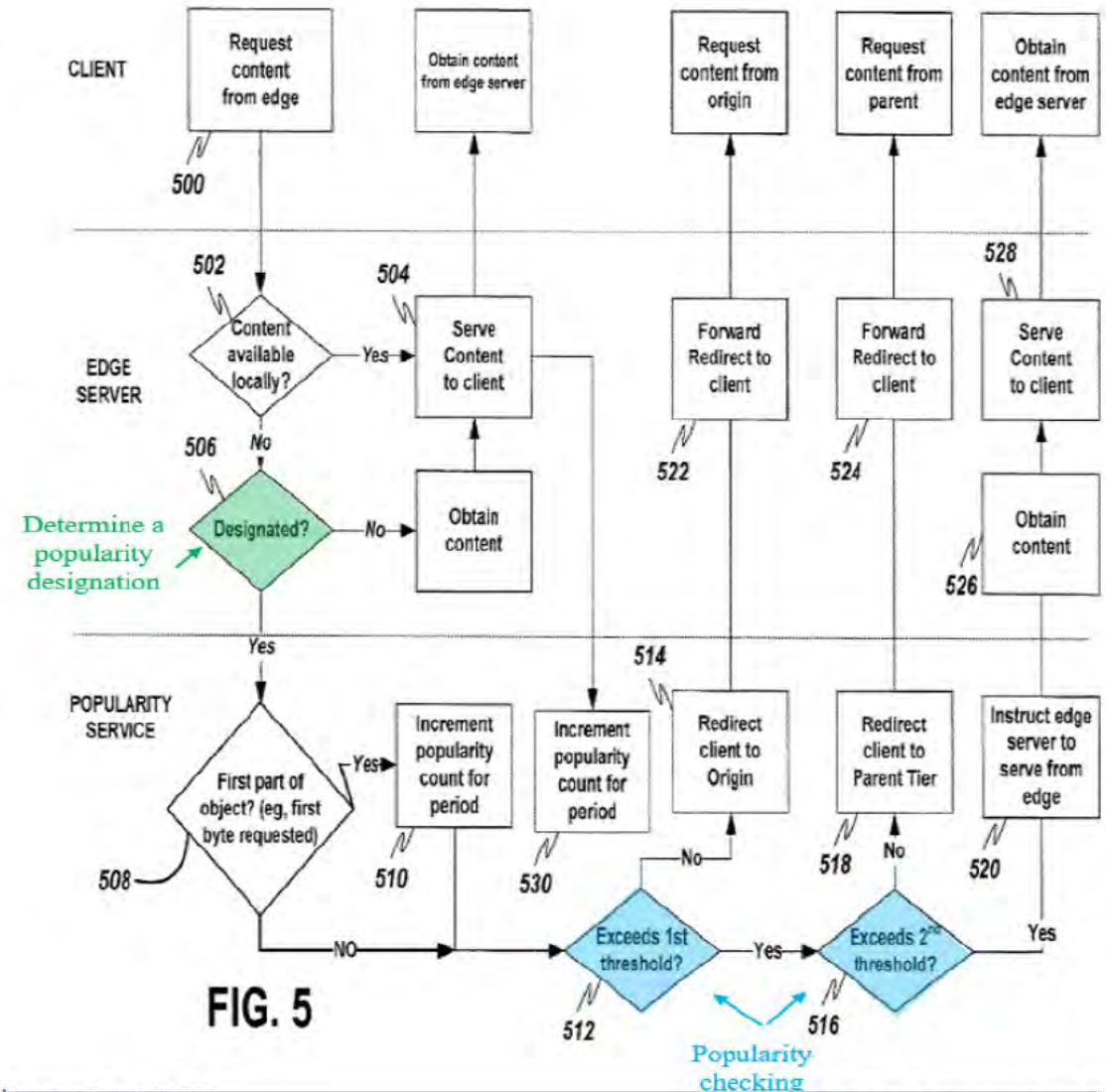


FIG. 5

Petitioner's Expert also conflates “popularity value” with the claimed “popularity designation.”

Petitioner's expert concedes that Pai's popularity value is “the same concept” as using the threshold levels described in the '573 patent.

Petitioner's Expert

Q In your proposed combination of Dilley, Pai, and Wang, what is the thing that determines a popularity designation?

A So there's a lengthy discussion between paragraphs 237 and 241 of my declaration, where I go into detail about how Pai -- that Pai teaches how to dynamically determine popularity values to determine popularity. And there's a lot of discussion in my report about Pai. It, in fact, uses the same concept of two threshold levels that were described in the '573 patent. There's a lower threshold and a higher threshold, so you end

EX2017, 92:11-22.

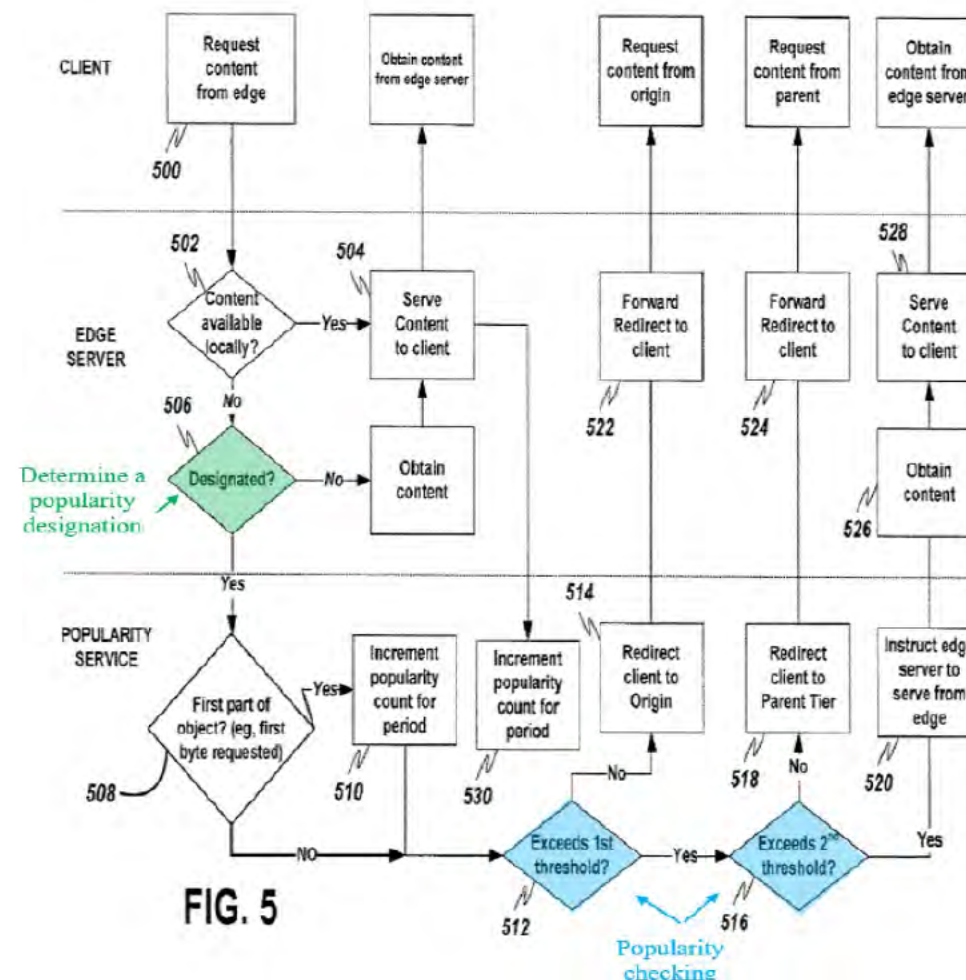


FIG. 5

EX1001, FIG. 5 (annotated).

Petitioner's Reply does not cure the Petition's deficiencies.

Petitioner's Reply

redistributes content using popularity thresholds. ID at 22-23. To rebut this, Patent Owner reads into the claims an “eligibility” determination—determining if content is “eligible” for popularity-based handling before popularity analysis occurs. POR at 26 (arguing Pai does not assess eligibility *before* assigning popularity value). This claim rewriting should be rejected.

Pet. Reply, 11.

Reading the claim term in view of specification is not “importing” a limitation.

PO Sur-reply, 7.



Claim terms cannot be determined in a vacuum; they must be read “in view of the specification,” which serves as “the single best guide” to their meanings.

Phillips v. AWH Corp., 415 F.3d 1303, 1315 (Fed. Cir. 2005) (en banc)

Petitioner's Reply does not cure the Petition's deficiencies.

Petitioner's Reply

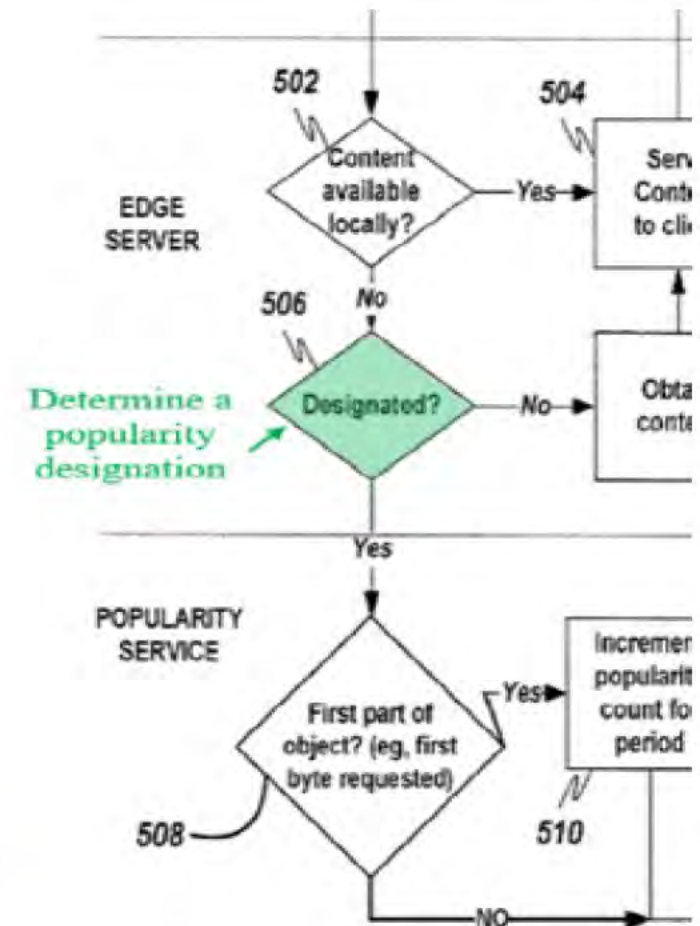
Reading the claim term in view of specification is not
“importing” a limitation.

First, the plain claim language does not recite a determination of “eligibility” for popularity-based handling. EX-1001, 13:35-54, 14:30-49, 15:13-16:14. Claim 1 recites “accessing a popularity service ... to determine a popularity designation.” *Id.* at 13:41-43. Claim 11 recites “a popularity service tracking a popularity designation” *Id.* at 14:45-46. And claim 19 has similar concepts. *Id.* at 16:1-4. Each requires determining the resource’s popularity and using the determination to route and cache resources. Nothing in this plain claim language requires assessing eligibility for the popularity analysis *before* assessing what the popularity designation may be.

Pet. Reply, 11.

Ample intrinsic evidence establishes that “determining a popularity designation” is a prerequisite eligibility check.

POR, 26-27; EX2016, ¶¶64-65.



EX1001, FIG. 5 (excerpt).

Petitioner's Reply misinterprets dependent claim differentiation.

Dependent claims add further limitation, not to redefine the independent claim term of “popularity designation.”

Petitioner's Reply

The dependent claims confirm this understanding, explaining that “popularity designation” simply requires making the popularity threshold determination. Claims 2 and 3, for example, provide the popularity designation “indicates that the requested resource is [or is not] popular.” *Id.* at 13:55-64. And claim 6 states that the popularity designation “comprises an indication whether a current popularity value ... exceeds popularity threshold.” *Id.* at 14:4-8. Claims 12, 13, and 16 recite similar limitations.

Pet. Reply, 12.

The “popularity designation” can have multiple functions besides indicating baseline eligibility, such as those claimed by dependent claims 2, 3, 6, 12, 13, and 16.

PO Sur-reply, 8.



“[A] claim in dependent form shall contain a reference to a claim previously set forth and then specify a further limitation of the subject matter claimed.”

35 U.S.C. § 112(d).

Petitioner's Reply mischaracterizes Figure 5 of the '573 patent.

The edge server accessing the popularity service to perform the eligibility determination meets the claim limitation.

1(b)

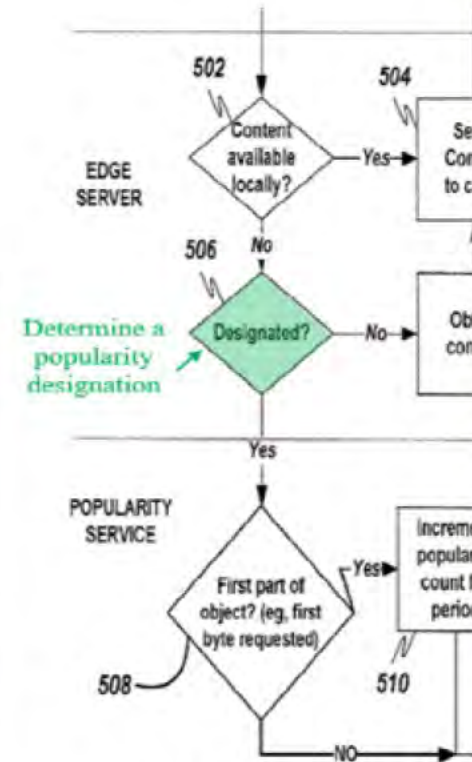
accessing a popularity service associated with the content delivery network to determine a popularity designation associated with the requested resource;

EX1001, Claim 1.

Petitioner's Reply

Second, the '573 specification does not establish that "popularity designation" requires any eligibility assessment. Patent Owner cites the yes/no "Designated?" box in Figure 5, but that does not support its position. POR at 27. Again, the claims say: "accessing a popularity service ... to determine a popularity designation." EX-1001, 13:41-43. The "Designated?" box (506) appears in the edge server *prior to* accessing a popularity service and, thus, cannot be the claimed "popularity designation." That

Pet. Reply, 12.



EX1001, FIG. 5 (excerpt).

1

Edge server performs the determination (Box 506) by accessing a popularity service.

PO Sur-reply, 9-10.

2

The '573 patent allows architectural flexibility: "the popularity service may [be] located anywhere in the system, including in the edge tier."

PO Sur-reply, 11 (citing EX1001, 5:44-46).

Agenda

- ▶ The '573 Patent
- ▶ Independent Claims 1, 11, 19 (Ground 2)
 - ▶ “a popularity designation”
- ▶ **Independent Claim 19 (Ground 2)**
 - ▶ “requesting the resource ... based at least on the popularity designation”
- ▶ Independent Claims 1, 11 (Ground 2)
 - ▶ “a redirect instruction”
- ▶ Dependent claims 5, 15 (Ground 2)
 - ▶ peer-to-peer redirect
- ▶ Claim 1-20 (Ground 1)
 - ▶ “instruction to not cache”

The '573 Patent, Claim 19

Claim 19	
19(preamble)	A method of content delivery in a content delivery network comprising:
19(a)	receiving, at a first server of a first tier of servers of the content delivery network, a request from a requesting device for a resource available from the content delivery network;
19(b)	accessing a popularity service associated with the first server of the first tier of the content delivery network to determine a popularity designation associated with the requested resource;
19(c)	receiving an instruction to not cache the resource at the first server of the first tier of servers when the portion of the resource is obtained from the content server of the content delivery network;
19(d)	requesting the resource from a second server of the content delivery network based at least on the popularity designation associated with the requested resource; and
19(e)	providing the obtained resource to the requesting device.

EX1001, Claim 19.

Dilley selects a second server based on metadata, not popularity.

Petitioner's Expert

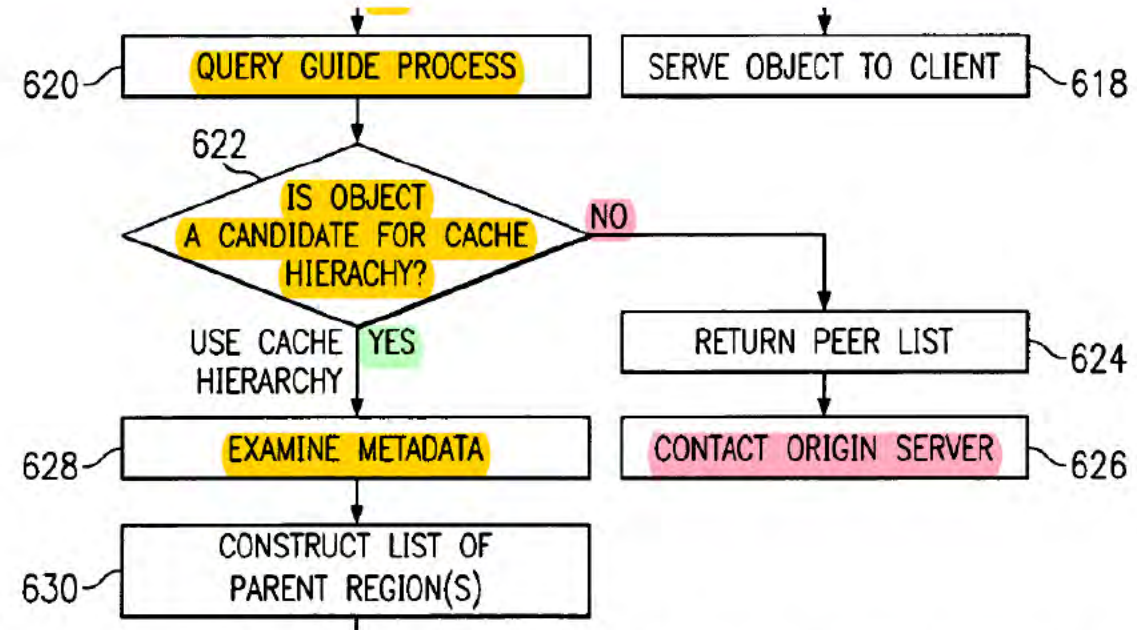
origin servers from becoming "overwhelmed with requests for popular content." a POSITA would have understood that content with a popularity designation, identifying that content as unpopular, would be served from the parent or origin servers, while popular content would be served from the edge. EX-1003, 5:10-21. Thus, based on an unpopular designation, Dilley would retrieve content from the second server.

EX1006, ¶353.

Dilley

At step 620, the edge server queries its associated guide process to identify where to go to get the object. At step 622, the guide process makes a determination whether the object is a candidate for using the hierarchy. As noted above, preferably this determination is made by having the guide process examine object metadata for a given tag, e.g., "use-hierarchy," in a data string. An illustrative data string is

EX1003, Dilley, 8:22-28.



EX1003, FIG. 6 (excerpt and annotated).

Dilley's guide process examines the object metadata tag for selecting the second server.

POR, 43-45.

Petitioner improperly relied on its expert to bridge the gaps.

Petitioner's Expert

Petitioner's references do not meet the claimed "requesting the resource ... based ... on the popularity designation."

354. Pai further dynamically manages and distributes content based on popularity values. EX-1004, ¶[0013]. As Pai explains, as content is consumed, its popularity value can change over time, moving from media source to media source. EX-1004, ¶¶[0020], [0026], [0029]. Pai explains that media with popularity values exceeding the highest threshold will be distributed and managed in one tier—media source one—while others would be distributed and managed in other tiers of media sources. EX-1004, ¶¶[0038]-[0045]. Thus, a POSITA would have understood that when Pai's system receives a request for content that does not exceed the highest popularity threshold, the request would need to be forwarded to a second server (i.e., a server of media level two for mid-tier popular content or a server of media level three for unpopular content).

EX1006, ¶354.

Gaps:

Pai's *content placement* based on *popularity values* and the claimed *content delivery* based on the *popularity designation*.

PO Sur-reply, 12.

Pai fails to disclose “a popularity designation,” thus Dilley and Pai fail to teach “request ... based ... on the popularity designation.”

As already established, Pai fails to disclose “a popularity designation.”

POR

In contrast, Pai discloses a system that distributes or redistributes *all* media content based on “current popularity values.” EX1004, ¶39. While Pai evaluates these numeric popularity values to manage storage of *all* media content across media sources, Pai lacks any teaching of *determining* a “popularity designation” for its media content. EX1004, ¶¶39-51. Indeed, Pai has no popularity designation. Pai has no reason to check for one, and there is no step or action to perform based on the popularity designation. Petitioner effectively admits this deficiency by pointing only to Pai’s determination of a “popularity value” associated with media content. Pet., 45; EX1004, ¶39. A POSA would instantly distinguish Pai’s “popularity value” from the claimed “popularity designation.” EX2016, ¶¶66-67.
POR, 28.

Pet., 54-55.

Patent Owner's Demonstrative Exhibit – Not Evidence

1 Pai does not disclose “a popularity designation.”

POR, 26-28; EX2016, ¶¶63-68.

2 Petitioner’s reply repeats its flawed conflation of Pai’s numerical value with the claimed “popularity designation.”

Pet. Reply, 14; PO Sur-reply, 13.

3 In its reply, Petitioner does not provide any evidence to rebut.

PO Sur-reply, 13.

Thus, Dilley and Pai fail to teach he claimed “request ... based ... on the popularity designation.”

PO Sur-reply, 12.

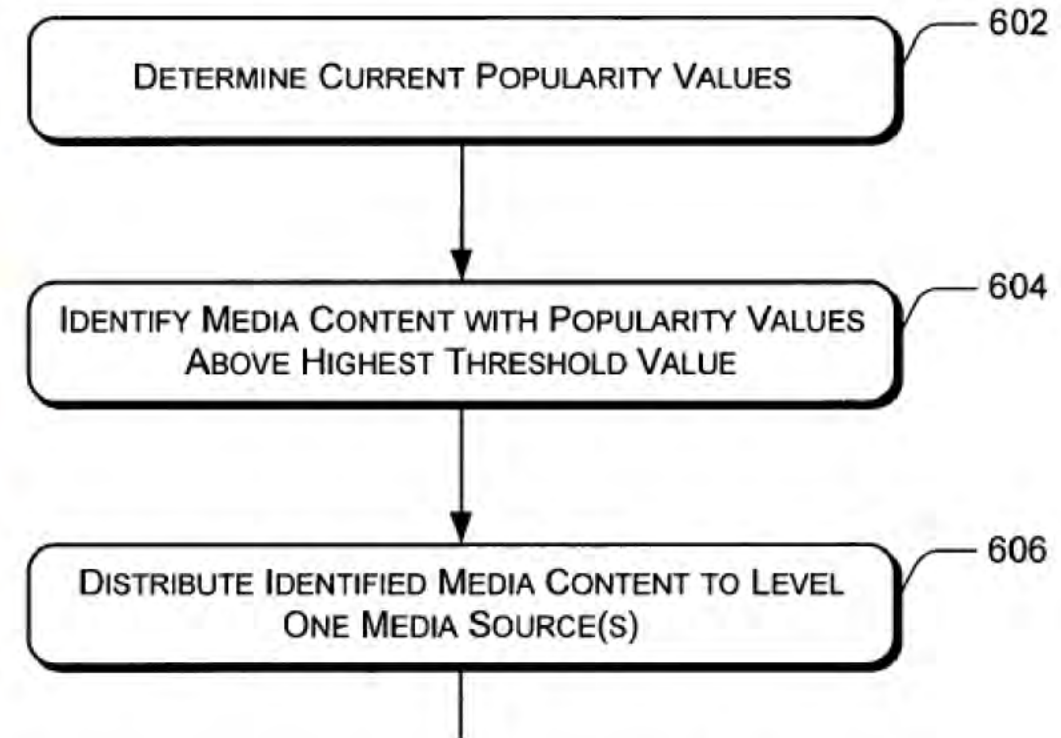
Pai discloses media storage based on popularity values

Pai

[0013] The embodiments of popularity-based on-demand media distribution described below provide techniques for distributing on-demand media content within an on-demand media system based on popularity of the media content. An

maintained with less replication. Available media content is dynamically distributed among the multiple media sources based on popularity of the media content, determined, for example, by monitoring a number of received requests for the media content over time.

EX1004, ¶13.



EX1004, FIG. 6 (truncated).

Petitioner relies on impermissible hindsight to combine Dilley and Pai

POR

popularity designation. See Section VIII.C. There is simply no suggestion in either reference that a background storage metric (i.e., Pai's popularity value) could, or should be, used as a real-time routing trigger. Petitioner's leap from Pai's storing based on "popularity values" to the claimed "*requesting ... based ... on [] popularity designation*" is a narrative driven by what is claimed, but is not supported by the prior art without distorting its teachings to meet the claims. EX2016, ¶78.

POR, 34, citing EX2016, ¶78.

1 Both Dilley and Pai fail to disclose "a popularity designation"

POR, 31-32, 34; EX2016, ¶¶73-74, 76.

2 Petitioner conflates Pai's internal distribution based on popularity values with external delivery based on popularity designations

Pet., 84; POR, 34; EX2016, ¶78.

Petitioner modifies the prior art to perform a step in order to achieve a hindsight-driven result that neither reference suggests.

POR, 35.

Petitioner's Reply does not cure the hindsight deficiency

Petitioner Reply

Patent Owner's attempt to undermine this obviousness showing by arguing it conflates "internal distribution" and "external delivery" ignores that such systems in fact are "closely related" (EX-2017, 60:16-61:3) and focused on improving content management of network resources. Pet. at 36-39; POR at 34-36. Neither Dilley nor Pai draws the rigid distinction Patent Owner asserts. Moreover, as explained above and in the Petition, it is not a "leap" to use Pai's popularity distribution together with Dilley's retrieval of content based on popularity. *See supra* Sections III.A.1-III.A.2, III.B.1; Pet. at 37-39.

Pet. Reply, 16.



Disclosing closely related subject matter is legally insufficient to establish an obviousness ground.

Unified Patents Inc. v. DivX, LLC, IPR2019-01379, Paper 52 (P.T.A.B. Feb. 98, 2021) (persuasive).

Agenda

- ▶ The '573 Patent
- ▶ Independent Claims 1, 11, 19 (Ground 2)
 - ▶ “a popularity designation”
- ▶ Independent Claim 19 (Ground 2)
 - ▶ “requesting the resource ... based at least on the popularity designation”
- ▶ **Independent Claims 1, 11 (Ground 2)**
 - ▶ **“a redirect instruction”**
- ▶ Dependent claims 5, 15 (Ground 2)
 - ▶ peer-to-peer redirect
- ▶ Claim 1-20 (Ground 1)
 - ▶ “instruction to not cache”

The '573 Patent, Claim 1

Claim 1	
1(preamble)	A method of content delivery in a content delivery network comprising:
1(a)	receiving, at a first server of a first tier of servers of the content delivery network, a request from a requesting device for a resource available from the content delivery network;
1(b)	accessing a popularity service associated with the content delivery network to determine a popularity designation associated with the requested resource;
1(c)	requesting the resource from a second server of the content delivery network;
1(d)	processing, at the first server of the first tier of servers, a redirect instruction from the second server of the content delivery network to obtain the resource from a content server of the content delivery network;
1(e)	receiving an instruction to not cache the portion of the resource at the first server of the first tier of servers when the portion of the resource is obtained from the content server of the content delivery network; and
1(f)	providing the obtained resource to the requesting device.

EX1001, Claim 1.

The '573 Patent, Claim 11

Claim 11	
11(preamble)	A content delivery network comprising:
11(a)	a first tier of servers comprising a first plurality of servers, an edge server of the first tier of servers receiving a request from a requesting device in communication with the first server for a resource available from the content delivery network;
11(b(i))	a second tier of servers comprising a second plurality of servers,
11(b(ii))	a first server of the second plurality of servers receiving a request from the edge server for the content and, in response, transmitting a redirect instruction intended for the requesting device to the edge server to obtain the resource from a content server and transmitting an instruction to not cache the portion of the resource at the edge server when the portion of the resource is obtained from the content server; and
11(c)	a popularity service tracking a popularity designation associated with the requested resource;
11(d)	wherein the edge server processes the redirect instruction to obtain the content from the content server and provides the obtained content to the requesting device.

EX1001, Claim 11.

Wang's CDN architecture (LDAP exchange)

PO Expert

62. Wang describes a CDN with a hierarchical service network that utilizes the light weight directory access protocol (LDAP) to exchange directory information between service managers of the network. EX1005, Abstract, ¶¶4, 8, etc.). *Id.*, ¶123. If Wang's level 3 service manager returns the origin server IP address along with an indication of cacheable or not, the caching service engine will – based on the indication - either “http-redirect” the client to the origin server, or “initiate” a new http session with the origin server on behalf of the client. *Id.*, ¶¶123-124.

EX2016, ¶62

Petitioner's original obviousness rationale fails by manufacturing a non-existent gap in Dilley

Petitioner's Expert

252. While Dilley discloses that the content can be retrieved from the origin server when it is unavailable at the edge and parent servers, it does not provide specifics on how that retrieval happens. Instead, it leaves it open to a POSITA how the communications and instructions among the tiered servers would work.

EX1006, ¶252.

253. A POSITA would have understood that, in the example when the second server is a parent server, Dilley's parent server could have sent a redirect instruction to the edge server, directing the edge server to obtain the resource from the origin content server instead. EX-1003, 7:15-30. A POSITA would have understood this and have been motivated to implement the redirect instruction this way from the description in Wang. See e.g., EX-1005, ¶¶[0123]-[0130].

EX1006, ¶253.

1 Petitioner alleges that Dilley's parent server "does not provide specifics" on how to retrieve content from the origin server."

2 Petitioner's alleged gap is used to justify introducing Wang's redirect instruction.

However, Dilley explicitly describes the specifics on how to retrieve content from the origin server.

POR, 43-45; PO Sur-reply, 16-17.

Dilley provides specifics on retrieving content from the origin server

Dilley uses metadata tag to turn on/off cache hierarchy

The metadata tag “use-hierarchy” turns on the cache hierarchy feature, specifies the policy or type of hierarchy parent, and lists the base domain to be used.

The default setting for cache hierarchy is OA:

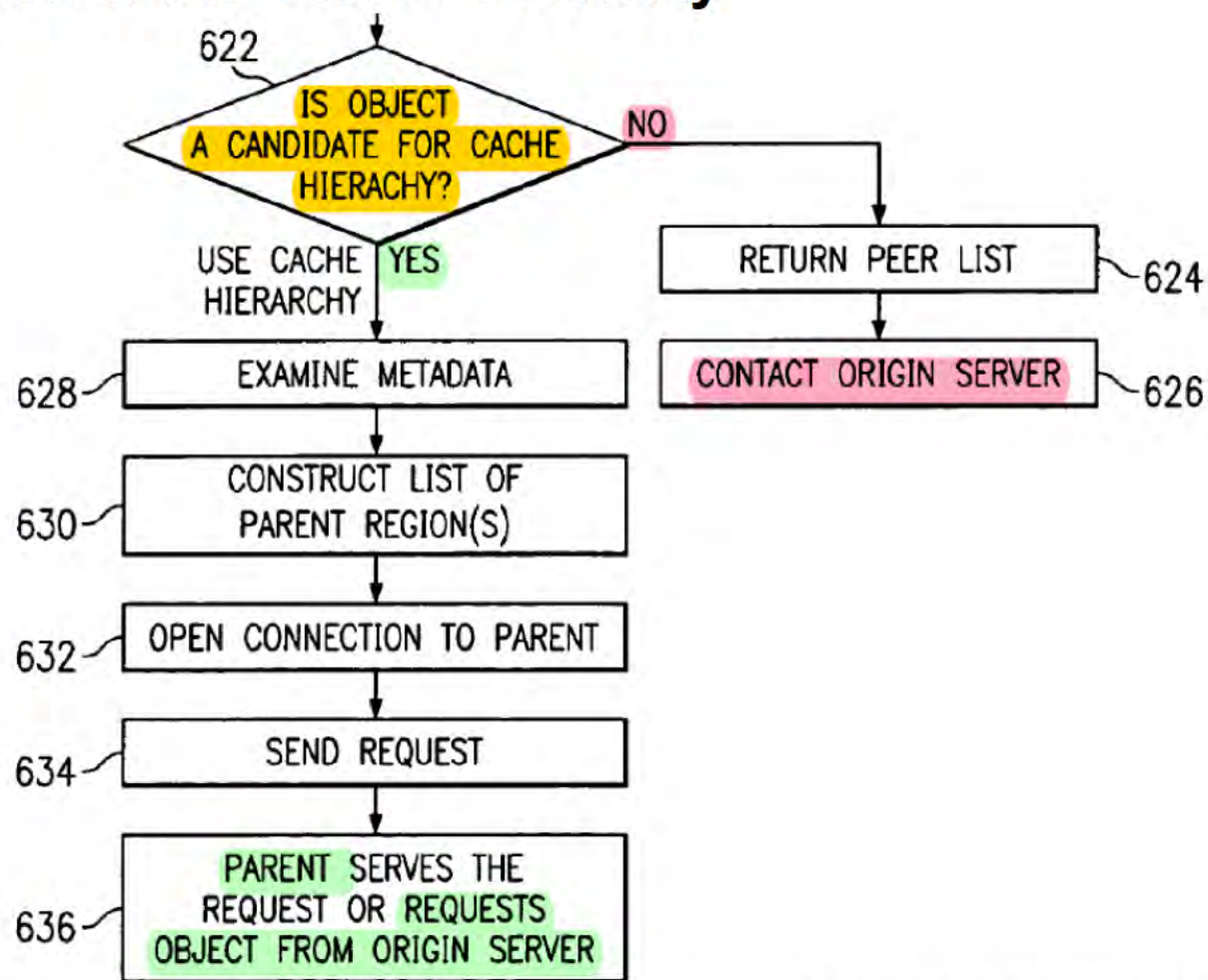
```
<md name=“use-hierarchy”>OA</md>
```

OA indicates that the edge server is to go direct to the (O)rigin-server specified in the (A)RL, where the ARL is a modified URL (or portion thereof) that points an end user browser to attempt to retrieve the object from the CDN. This setting is essentially the equivalent of turning the cache hierarchy off.

To turn on cache hierarchy for a typical CDN customer, the metadata tag may be of the following format:

```
<md name=“use-hierarchy”>coremap</md>
```

EX1003, 6:25-40.



EX1003, FIG. 6 (truncated).

Dilley explicitly provides when to retrieve content from origin server

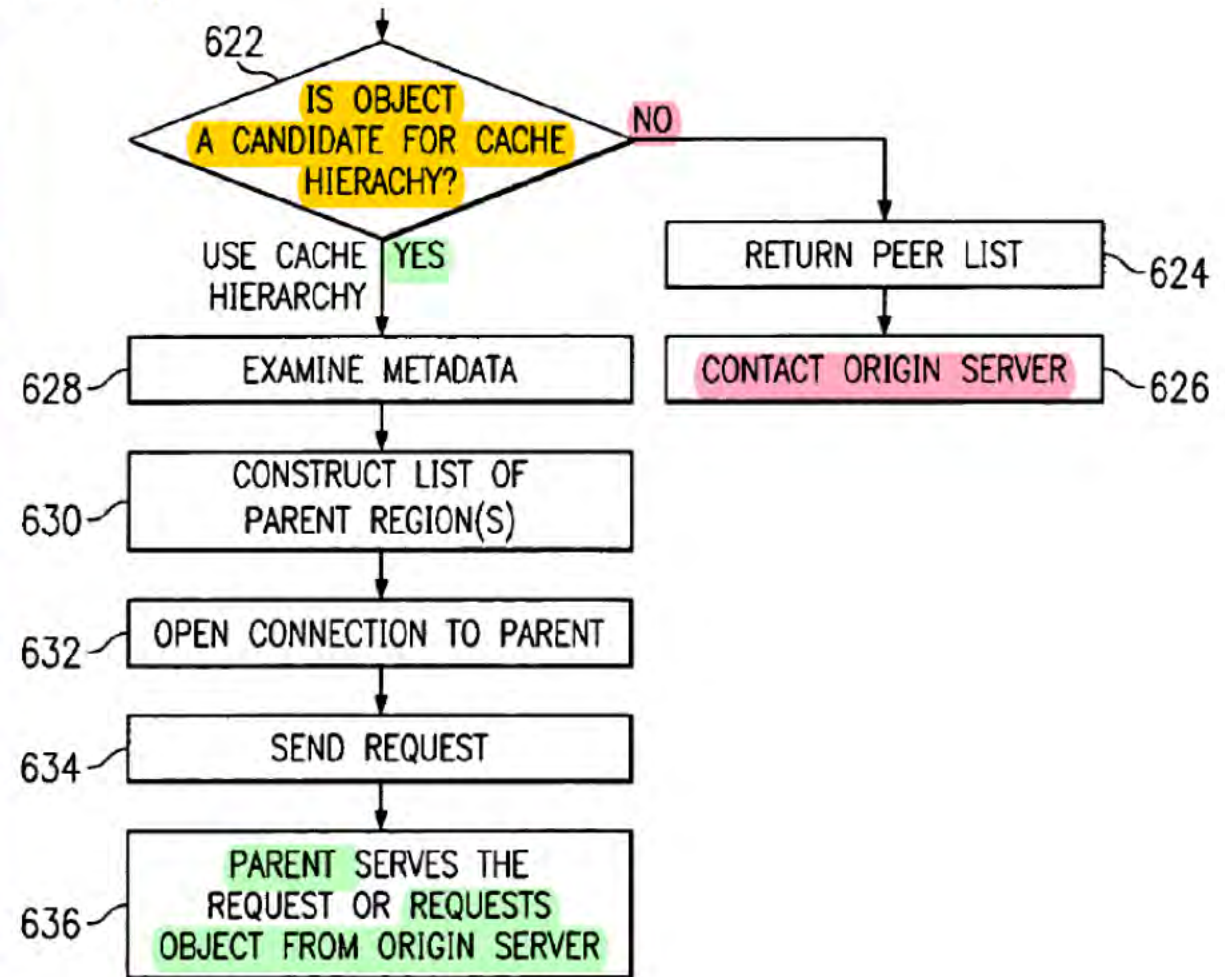
Dilley's default mechanism: edge server contacts origin server directly (cache hierarchy turned off)

edge regions. Preferably, only a relatively small, specific subset of objects are serviced using the hierarchy, while most objects are fetched using a default mechanism where the edge server contacts the origin server to handle a request that cannot be serviced in the edge region. This selectivity is accomplished preferably using object "metadata."

EX1003, 6:10-15.

cache hierarchy. If the object is not a candidate for distribution through the hierarchy, the guide process at step 624 returns a peer list that includes just a given directive (e.g., DIRECT), which causes the server, at step 626 to talk to the origin server to obtain the object as usual. If, however, the object is a candidate for tiered distribution, the guide process examines the "use-hierarchy" metadata information at step 628 and, at step 630, the process constructs a list of one or more parent region(s) or a dedicated parent region. At step 632, the server opens a connection (or uses a persistent one) to a parent region on the list and, at step 634, it sends the request to a server in the parent region. The requesting server

EX1003, 8:34-45.



EX1003, FIG. 6 (truncated).

Modifying Dilley with Wang is a superfluous and redundant

Dilley's default mechanism has already bypassed the parent server:

111. Dilley already discloses a “guide process” that evaluates metadata to determine if an object request should bypass the parent server and the object be retrieved directly from the origin server. EX1003, 7:46-8:2, 9:7-13. Petitioner and Dr. Mowry's proposal to implement Wang's redirection procedure in Dilley's parent server creates a superfluous mechanism to achieve a bypass result that Dilley already performs at the edge server level.

EX2016, ¶111.

A POSITA would not introduce unnecessary logic to a parent server to handle requests that, by definition, already bypass that server under Dilley's default routing mechanism.

PO Sur-reply, 18-19.

Dilley explicitly provides when to retrieve content from origin server

Petitioner Reply:

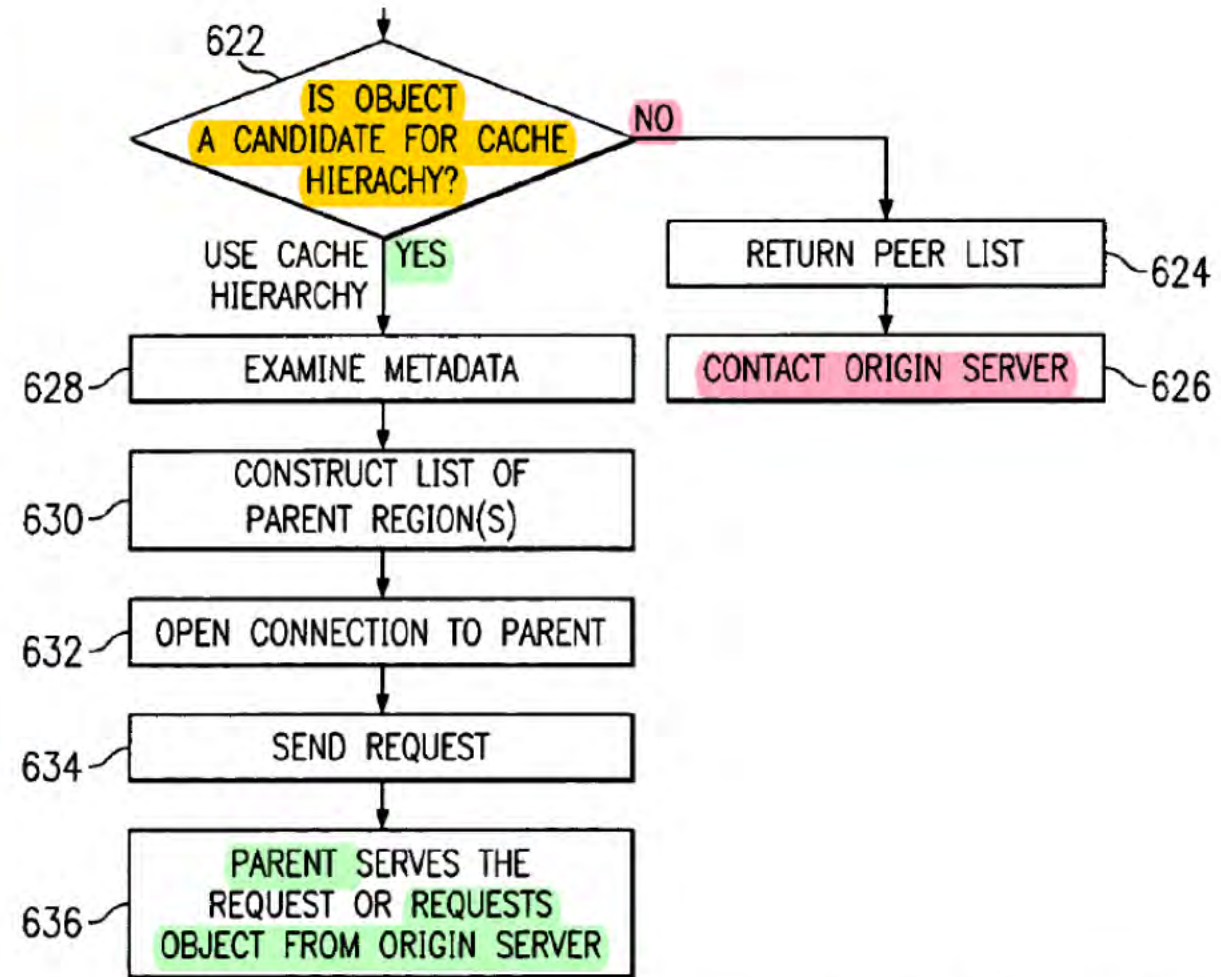
Tellingly, Patent Owner relies upon instances where Dilley has essentially “turn[ed] the cache hierarchy off.” EX-1003, 6:30-40. But these passages do not address the relevant situation: when content *is* a candidate for cache hierarchy and is not located at the edge. In that circumstance, Wang supplies the necessary details regarding how

Pet. Reply, 18.

Dilley explicitly describes the exact scenario when the content is a candidate for cache hierarchy

At step 636, the parent serves the request either from its cache if it has it, or by making a request to the origin server. Preferably, a parent will not make a request to another server in the hierarchy; rather, its guide process by default will return only to the origin server as a valid destination. This completes the process.

EX1003, 8:53-58.



EX1003, FIG. 6 (truncated).

Petitioner's Reply shifts from gap-filling to known-alternative theory.

Petitioner's new “known alternative” theory:

841 F.3d 995, 1002-03 (Fed. Cir. 2016). In other words, whether Dilley provides one origin-retrieval path is of no moment, as Wang teaches another well-known path that would be predictably used with Dilley's hierarchy. EX-1006, ¶¶253-257; EX-1005,

Pet. Reply, 19.



A petitioner may not raise in its Reply an entirely new rationale.

37 C.F.R. 42.23(b); Yita LLC v. MacNeil IP LLC, 69 F.4th 1356, 1365-66 (Fed. Cir. 2023).



A petitioner may not introduce an “entirely new rationale” and treat motivation to combine as a moving target to escape its initial failed burden of proof.

Intelligent Bio-Systems, Inc. v. Illumina Cambridge Ltd., 821 F.3d 1359, 1369-70 (Fed. Cir. 2016)

Petitioner relies on hindsight to manufacture unconvincing benefits

Petition:

be expecting the response directly from the origin server. Such a design would decrease latency as it would remove the parent server as a “middleman,” and would allow the edge server to communicate with and receive content directly from the origin server. EX-1006, ¶¶255-256. In addition, this process would allow the parent server to remain available to service other requests received by the CDN. EX-1006, ¶257.

Pet., 51, citing EX1006, ¶¶255-257.

Dilley:

edge servers with the requested content. Preferably, a tiered distribution hub 304 maintains a persistent connection to the origin server 308, although this is not a requirement. The

EX1003, 5:10-12.

sensitive objects. As a result, the cache hierarchy protects the content providers from requests for large, popular objects causing the CDN to make more requests than the content provider can handle. In addition, the cache hierarchy reduces average latency and improves connectivity and content access in what might otherwise be poorly-connected edge regions. Preferably, only a relatively small, specific

EX1003, 6:4-10.

- Petitioner’s “fewer nodes equals less latency” is a simplistic generalization that fails to account for specific technical context of Dilley.
- Dilley’s “middleman” “maintains a persistent connection” to the origin server, reduces latency and improves connectivity for “poorly-connected” edge server.

POR, 46, citing EX1003, 5:10-12, 6:7-11.

Combining Wang with Dilley offers no latency benefit

Removing Dilley's middleman would increase latency

PO's Expert

106. Dilley expressly teaches that its parent server is not a bottleneck, but a functional necessity that “reduces average latency” by providing a persistent connection for “poorly-connected” edge servers. EX1003, 6:7-11. A POSA would have understood that removing the parent server - and thus the “middleman” providing that persistent connection - would inevitably increase latency by forcing a direct, unstable connection between the poorly-connected edge server and the origin server.

EX2016, ¶106.

Dilley:

sensitive objects. As a result, the cache hierarchy protects the content providers from requests for large, popular objects causing the CDN to make more requests than the content provider can handle. In addition, the cache hierarchy reduces average latency and improves connectivity and content access in what might otherwise be poorly-connected edge regions. Preferably, only a relatively small, specific subset of objects are serviced using the hierarchy, while most objects are fetched using a default mechanism where the edge server contacts the origin server to handle a request that cannot be serviced in the edge region. This selectivity is accomplished preferably using object “metadata.”

EX1003, 6:4-15.

Petitioner's Reply does not cure the deficiencies

Petitioner Reply

1003, 5:10-12. It is optional. And Patent Owner's argument conflicts with Dilley's explanation that most content is fetched using a policy that bypasses the hierarchy, having the allegedly "poorly connected" edge servers request content directly from the origin without issue. EX-1003, 6:10-15, 31-36. This direct retrieval disclosure also defeats Patent Owner's further argument that the combination would allegedly eliminate the connection to Dilley's allegedly "poorly connected regions." POR at 49; EX-1003, 6:10-15. Patent Owner also contends the combination does not provide

Pet. Reply, 20-21.

Dilley:

sensitive objects. As a result, the cache hierarchy protects the content providers from requests for large, popular objects causing the CDN to make more requests than the content provider can handle. In addition, the cache hierarchy reduces average latency and improves connectivity and content access in what might otherwise be poorly-connected edge regions. Preferably, only a relatively small, specific subset of objects are serviced using the hierarchy, while most objects are fetched using a default mechanism where the edge server contacts the origin server to handle a request that cannot be serviced in the edge region. This selectivity is accomplished preferably using object "metadata."

EX1003, 6:4-15.

- Dilley contains no such disclosure of "poorly connected" edge servers request content directly from the origin without issue".
- Dilley *selectively* activate the cache hierarchy for a specific subset of objects to protect content providers, reduce latency, and improve connectivity.

Modifying with Wang disrupts Dilley's core principle of operation

Petitioner Reply

parent, to edge. POR at 48. Under the Wang-inspired modification, the second server need only transmit a small instructional message; the resource travels directly from origin to edge, preventing the second server from obtaining and processing large or popular content unnecessarily. EX-1005, ¶¶123-124. As Dr. Mowry explained, the

Pet. Reply, 21.

Dilley:

In effect, the tiered distribution hubs act as accelerators for the end users, ensuring that the users obtain their desired content quickly and reliably, and the hubs acts as buffers for the origin site's internal infrastructure, guaranteeing that the site is not overwhelmed with requests for popular content or large files.

EX1003, 5:16-21.

- Dilley explicitly teaches using second servers as “accelerators” for end users, and act as “buffers” to protect the origin server.
- Petitioner's argument directly contradicts Dilley's core design principle and destroys the protective buffer functionality.

PO Sur-reply, 21-22.

Petitioner's "service availability" argument subverts the purpose of Dilley's buffering tier

Petitioner Reply

Patent Owner's parent-server-availability argument is also misplaced. POR at 49. Just because servers can handle multiple requests concurrently does not mean no benefits exist when preventing a parent server from processing and relaying large, popular objects unnecessarily. Freeing parent resources from that role leaves them available for other requests and reduces the amount of data that they must manage. POR at 51. That is important in a CDN designed to protect upstream infrastructure from requests for popular content. EX-1003, 2:26-31, 2:67-3:2.

Pet. Reply, 21-22.

Dilley:

In effect, the tiered distribution hubs act as accelerators for the end users, ensuring that the users obtain their desired content quickly and reliably, and the hubs acts as buffers for the origin site's internal infrastructure, guaranteeing that the site is not overwhelmed with requests for popular content or large files.

EX1003, 5:16-21.

- Dilley explicitly teaches that parent servers intentionally absorb the heavy computing, bandwidth, and processing burdens associated with large, popular objects to protect the origin server.
- Petitioner's argument forces the downstream edge tier to hit the origin server repeatedly for subsequent cache files.

Agenda

- ▶ The '573 Patent
- ▶ Independent Claims 1, 11, 19 (Ground 2)
 - ▶ “a popularity designation”
- ▶ Independent Claim 19 (Ground 2)
 - ▶ “requesting the resource ... based at least on the popularity designation”
- ▶ Independent Claims 1, 11 (Ground 2)
 - ▶ “a redirect instruction”
- ▶ **Dependent claims 5, 15 (Ground 2)**
 - ▶ **peer-to-peer redirect**
- ▶ Claim 1-20 (Ground 1)
 - ▶ “instruction to not cache”

The '573 Patent, Claims 5 and 15

Claim 1 (d)

processing, at the first server of the first tier of servers, a redirect instruction from the second server of the content delivery network to obtain the resource from a content server of the content delivery network;

Claim 5

The method as recited in claim 1 wherein the second server is a peer server of the first tier of servers of the content delivery network.

Claim 11 (b(ii))

a first server of the second plurality of servers receiving a request from the edge server for the content and, in response, transmitting a redirect instruction intended for the requesting device to the edge server to obtain the resource from a content server and transmitting an instruction to not cache the portion of the resource at the edge server when the portion of the resource is obtained from the content server; and

Claim 15

The content delivery network as recited in claim 11 wherein the first server of the second plurality of servers is a peer server of the first tier of servers of the content delivery network.

Dilley and Wang fails to teach a peer-to-peer redirect

Petitioner Reply

claim 1d (POR at 51) fails for the same reasons above. Section III.D.2. Next, Dilley expressly teaches peer-server queries when the first edge server does not have the requested object. EX-1003, Fig. 6, 8:11-21. Wang teaches the well-known technique of returning redirect and cache/no-cache instructions in that circumstance. EX-1005,

The skilled artisan would have readily applied the same redirect technique to any “second server” (peer server) in Dilley’s hierarchy when that peer cannot satisfy the request locally. Pet. at 69; EX-1006, ¶258. It is not that a peer is a “middleman” in the same sense as a parent; rather, the peer is an intermediate node in the request path, and redirecting a first server toward a content source avoids futile peer-to-peer handoffs and resource expenditure. Pet. at 69; EX-1006, ¶¶258-260.

1 Dilley teaches peer-server queries.

2 Wang teaches redirect instruction.

Thus, Dilley and Wang teaches peer-to-peer redirect to origin server.

Pet. Reply, 23.

Pet. Reply, 23.

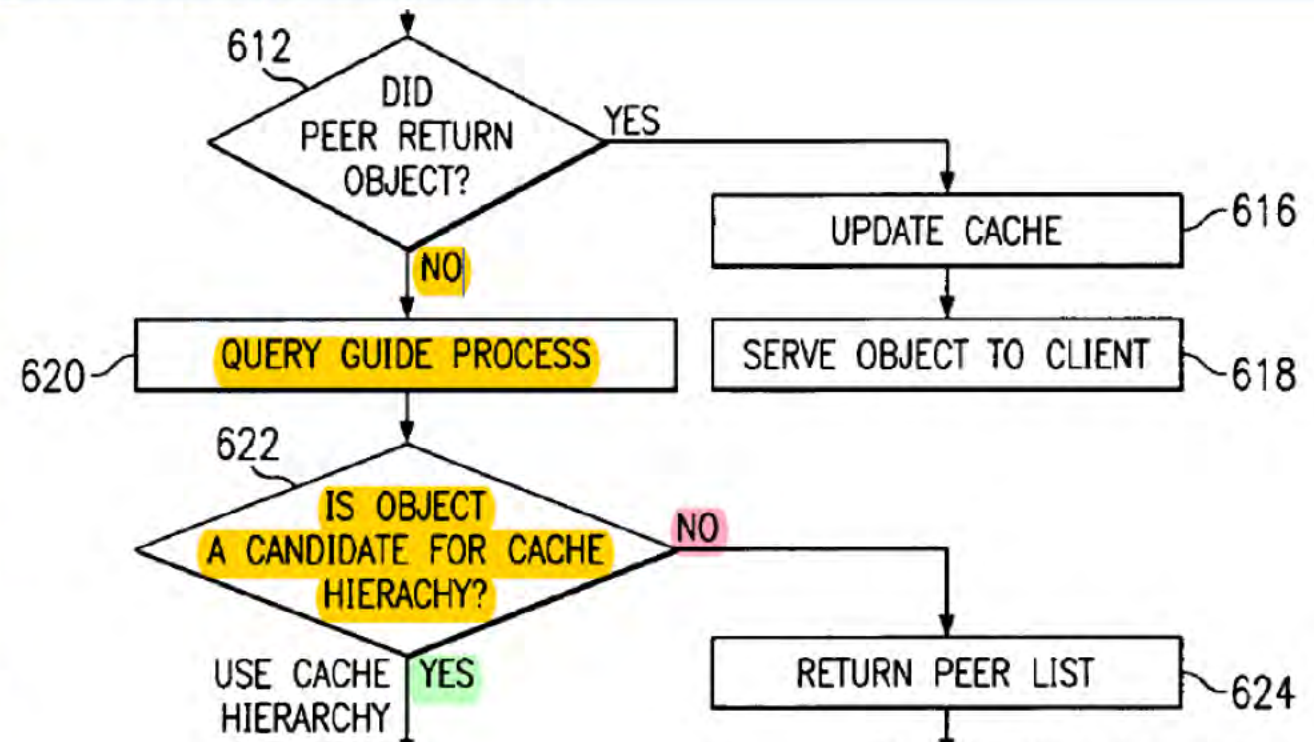
Petitioner's alleged peer-to-peer redirect defeats Dilley's core principle of operation

Dilley:

the requesting client. This completes the processing. If, however, the outcome of the test at step 612 indicates that a peer has not returned a hit, i.e., no peer copy of the object is available, the process continues.

At step 620, the edge server queries its associated guide process to identify where to go to get the object. At step 622, the guide process makes a determination whether the object is a candidate for using the hierarchy. As noted above, preferably this determination is made by having the guide process examine object metadata for a given tag, e.g., "use-hierarchy," in a data string. An illustrative data string is

EX1003, 8:18-27.



EX1003, FIG. 6 (truncated).

- Dilley explicitly teaches checking metadata tag for routing when a peer query fails to return a valid object.
- The routing mechanism serves Dilley's goal of using parent servers as buffer to protect the origin server.
- Imputing a peer-to-peer redirect (bypasses the parent tier) actively destroys the protective buffering advantages taught by Dilley

Agenda

- ▶ **The '573 Patent**
- ▶ Independent Claims 1, 11, 19 (Ground 2)
 - ▶ “a popularity designation”
- ▶ Independent Claim 19 (Ground 2)
 - ▶ “requesting the resource ... based at least on the popularity designation”
- ▶ Independent Claims 1, 11 (Ground 2)
 - ▶ “a redirect instruction”
- ▶ Dependent claims 5, 15 (Ground 2)
 - ▶ peer-to-peer redirect
- ▶ **Claim 1-20 (Ground 1)**
 - ▶ **“instruction to not cache”**

The '573 Patent, Claim 1

Claim 1	
1(preamble)	A method of content delivery in a content delivery network comprising:
1(a)	receiving, at a first server of a first tier of servers of the content delivery network, a request from a requesting device for a resource available from the content delivery network;
1(b)	accessing a popularity service associated with the content delivery network to determine a popularity designation associated with the requested resource;
1(c)	requesting the resource from a second server of the content delivery network;
1(d)	processing, at the first server of the first tier of servers, a redirect instruction from the second server of the content delivery network to obtain the resource from a content server of the content delivery network;
1(e)	receiving an instruction to not cache the portion of the resource at the first server of the first tier of servers when the portion of the resource is obtained from the content server of the content delivery network; and
1(f)	providing the obtained resource to the requesting device.

EX1001, Claim 1.

The '573 Patent's Priority Date

Petition

The priority date for the challenged claims is February 25, 2019, the filing date of the '791 application, because new matter—specifically, the “instruction to not cache” limitation—was introduced on that date and has no support in any application to which the '573 patent claims priority. EX-1006, ¶¶54-56.

Pet., 8.

The '538 patent's disclosure is substantively the same as the '412 provisional application that is the earliest claimed priority application. But neither the '412 provisional nor, by extension, the intervening three applications to which the '573 patent claims priority support the “instruction not to cache” limitation found in each challenged claim, either directly or by dependency. EX-1006, ¶59.

Pet., 10.

Appl. No.: 16/284,791

Filed: Feb. 25, 2019

EX1001, codes (21, 22).

Related U.S. Application Data

Continuation of application No. 15/645,584, filed on Jul. 10, 2017, now Pat. No. 10,218,806, which is a continuation of application No. 12/880,324, filed on Sep. 13, 2010, now Pat. No. 9,762,692, which is a continuation-in-part of application No. 12/408,681, filed on Mar. 21, 2009, now Pat. No. 8,930,538.

Provisional application No. 61/042,412, filed on Apr. 4, 2008.

DI, 10, citing EX1001, codes (60, 63).

Written description support for “instruction to not cache”

Institution Decision

See Ex. 1011 ¶¶ 68–69. In light of these disclosures, we preliminarily find support for Dr. Shenoy’s testimony that, even with the words “instruction to not cache” are not literally recited in the Provisional Application, caching and not caching are alternative options that indicate the instruction to the edge server that will be implemented for each option. See Ex. 2014 ¶ 56. Further, these disclosures provide support for Dr. Shenoy’s testimony that the disclosures “reasonably convey to those skilled in the art that Patent Owner had possession of the edge server receiving an instruction to not cache the resource (and instead simply forward the redirect).” *Id.*

DI, 14.

alternative options. Further, there is sufficient evidence, in the context of the disclosures of written description support in the Provisional Application, providing support for Dr. Shenoy’s testimony that Patent Owner had possession of an edge server receiving an instruction to not cache the resource. Nonetheless, we invite additional briefing on these issues, including issues relating to alternative options of caching and not caching.

On this preliminary record, we find that the ’573 patent has a priority date of April 2, 2008, which predates the earliest possible date on which Newton-471 could be effective as prior art, September 30, 2014 (the date on which the earliest application in the priority chain of Newton-471 was filed). Accordingly, we preliminarily find that Newton-471 is not prior art to the challenged claims. Thus, on the present record, Petitioner has not

DI, 15.

Petitioner's expert agrees with the Board's decision

Petitioner's Expert

Q Did you read the portions of the institution decision where the board found, at least preliminarily, that the claims were entitled to the April 4th, 2008 priority date?

A Yes, I did.

EX2017, 14:13-17.

Q Do you think the board was wrong?

MR. ANDERSON: Objection. Form.

A Well, I describe my opinions about -- I recognize that the board makes decisions and that given the board's decision about the priority date, my combination of Dilley, Pai, and Wang, that's all prior to the priority date. Those references are before the April 4th, 2008 priority date. So I accept the decision of the board in that they get to set the rules, so that's their right to do that, certainly.

EX2017, 14:18-15:6.

The '573 patent and '412 application's written description describe two alternative instructions

Patent Owner's Expert

52. For example, the '573 patent and the '412 application both disclose:

If the edge server 108 receives a redirect from the popularity service 102 without the “follow me” flag set (cases 1 and 2 above), it simply forwards the redirect to the client 104 (S4a, 522, 524). If the edge server 108 receives a “follow me” redirect, it obtains and caches the resource (at 526) and serves it to the client (at 528).

Instruction not to cache

POR, 15.

Instruction to cache

POR, 15.

EX1001, 7:25-30; EX1011, ¶44. Accordingly, the written description of the '573 patent and the '412 application teach two alternate instructions that lead to two corresponding actions. EX1001, 7:25-30; EX1011, ¶44.

EX2014, ¶52, citing EX1001, 7:25-30; EX1011, ¶44.



A specification that “reasonably conveys to those skilled in the art that the inventor had possession of the claimed subject matter as of the filing date” has adequate written description of the claimed invention.

Novartis Pharm. Corp. v. Accord Healthcare, Inc., 38 F.4th 1031, 1021 (Fed. Cir. 2022) (quoting *Ariad Pharm. Inc. v. Eli Lilly & Co.*, 598 F.3d 1336, 1351 (Fed. Cir. 2010)).

Petitioner's Reply improperly reads extraneous limitations ("retrieve-and-do-not-cache") into a plainly recited term ("instruction to not cache")

Petitioner Reply

resource" and then serves it to the client. EX-1011, ¶44. If, however, the edge server receives a redirect *without* a follow me tag, it "forwards the redirect to the client."

Id. Importantly, when the follow-me tag is missing, the edge server *does not obtain the resource*, so whether it should cache the content is irrelevant; instead, the client is redirected elsewhere to obtain the content. And when the follow-me tag is present

cached. *Id.*, ¶44, Fig. 5. Thus, the '573 specification provides: (1) forward-only, with no edge retrieval, and (2) retrieve-and-cache. There is no disclosure of any *retrieve-and-do-not-cache action*, much less one responding to an "instruction to not cache."

Id., ¶44, Fig. 5.

Pet. Reply, 5.

1 It is not required to retrieve the content to comply with the "instruction to not cache."

PO Sur-reply, 4.

2 Redirect without edge retrieval is a direct implementation of not caching and achieves the intended goal of eliminating wasteful bandwidth consumption.

PO Sur-reply, 4, citing EX1001, 10:29-37; EX1011, ¶69.

3 Distortion of the claim language and improperly imports the "retrieve-and-do-not-cache" limitation into the "not cache" phase.

PO Sur-reply, 3.