

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

GOOGLE LLC,
Petitioner

v.

SANDPIPER CDN, LLC,
Patent Owner

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

PATENT OWNER'S SUR-REPLY TO PETITIONER'S REPLY

Mail Stop "PATENT BOARD"
Patent Trial and Appeal Board
U.S. Patent and Trademark Office
P.O. Box 1450
Alexandria, VA 22313-1450

TABLE OF CONTENTS

I.	PETITIONER’S INCONSISTENT POSITIONS BETWEEN THE DISTRICT COURT AND PTAB WARRANT TERMINATION OF THE PROCEEDING.	1
II.	GROUND 1	3
A.	The ’573 patent provides ample written description support for the claimed “instruction to not cache.”	3
1.	Petitioner improperly reads extraneous limitations into a plainly recited term.	3
2.	The specification explicitly contemplates avoiding retrieval to maximize efficiency.	4
III.	GROUND 2	5
A.	None of the asserted references teach “a popularity designation” (independent claims 1, 11, 19).	5
1.	Intrinsic evidence establishes that determining a popularity designation is an eligibility check.	5
2.	Petitioner’s Reply mischaracterizes the claims and improperly equates relative scope with definiteness.	7
3.	Petitioner’s Reply mischaracterizes the Figure 5 architecture of the ’573 patent.	9
B.	The Unbridged Gap Between Pai’s Internal “Content Placement” Based on Popularity Value and the Claimed “Content Delivery” Based on a Popularity Designation (Claims 1 and 19).....	12
1.	“Closely related” does not satisfy the particularity requirements of obviousness.....	12
2.	Dilley and Pai fail to teach content delivery based on a popularity designation.....	13
3.	Petitioner fails to rebut Patent Owner’s showing that its theories use impermissible hindsight.....	14
C.	None of the asserted references teaches “a redirect instruction” (independent claims 1, 11).	15

1.	Petitioner’s original obviousness rationale fails because the alleged gap in Dilley that Petitioner uses to justify the introduction of Wang’s redirect instruction does not exist.	16
2.	Petitioner’s new “known alternative” theory is legally improper and substantively flawed.	18
3.	Petitioner’s proposed combination fails because Wang’s redirect mechanism actively degrades system performance and offers no latency benefit.	19
a.	Petitioner’s argument mischaracterizes the stated purpose and selective activation of Dilley’s hierarchy.	20
b.	Eliminating the parent tier induces latency and threatens upstream infrastructure thrashing.....	21
c.	Petitioner’s “server availability” argument is technically illusory and subverts the purpose of a buffering tier.	23
D.	The asserted references fail to teach a peer-to-peer redirect to the origin server (claims 5 and 15).....	24
IV.	CONCLUSION.....	26

TABLE OF AUTHORITIES

Cases

<i>Intelligent Bio-Sys., Inc. v. Illumina Cambridge Ltd.</i> , 821 F.3d 1359 (Fed. Cir. 2016)	18
<i>In re Kahn</i> , 441 F.3d 977 (Fed. Cir. 2006)	19
<i>KSR Int'l Co. v. Teleflex Inc.</i> , 550 U.S. 398 (2007).....	18
<i>Phillps v. AWH Corp.</i> , 415 F.3d 1303 (Fed. Cir. 2005) (<i>en banc</i>)	7, 8
<i>In re Ratti</i> , 270 F.2d 810 (C.C.P.A. 1959)	23, 25
<i>Terumo BCT Inc. v. Haemonetics Corp.</i> , IPR2025-01374, Paper 20 (PTAB May 12, 2026)	2
<i>Tesla, Inc. v. IV II Ltd.</i> , IPR2025-00340, Paper 18 (P.T.A.B. Nov. 5, 2025).....	2
<i>Unified Patents Inc. v. DivX, LLC</i> , IPR2019-01379, Paper 52 (P.T.A.B. Feb. 8, 2021).....	13
<i>Yita LLC v. MacNeil IP LLC</i> , 69 F.4th 1356 (Fed. Cir. 2023)	18

Statutes

35 U.S.C. § 112(d)	8
--------------------------	---

Other Authorities

37 C.F.R. § 42.23(b)	18
----------------------------	----

PATENT OWNER'S EXHIBIT LIST

Exhibit No.	Description
2001	“Cloud CDN: content delivery network,” <i>Google Cloud</i> , accessed at https://cloud.google.com/cdn (last accessed July 2, 2025)
2002	<i>Intentionally Left Blank</i>
2003	<i>Intentionally Left Blank</i>
2004	U.S. Publication No. 2004/0221019 A1 to Swildens et al., published November 4, 2004
2005	<i>Intentionally Left Blank</i>
2006	<i>Sandpiper CDN, LLC v. Google LLC</i> , No. 2:24-cv-03951, ECF No. 82 (C.D. Cal. June 24, 2025)
2007	<i>Sandpiper CDN, LLC v. Google LLC</i> , No. 2:24-cv-03951, ECF No. 28 (C.D. Cal. Sept. 16, 2024)
2008	FAQs for Interim Processes for PTAB Workload Management, accessed at https://www.uspto.gov/patents/ptab/faqs/interim-processes-workload-management
2009	<i>Intentionally Left Blank</i>
2010	<i>Intentionally Left Blank</i>
2011	<i>Intentionally Left Blank</i>
2012	<i>Intentionally Left Blank</i>
2013	Interim Processes for PTAB Workload Management Memorandum (March 26, 2025)
2014	Declaration of Dr. Prashant Shenoy
2015	Curriculum Vitae of Dr. Prashant Shenoy
2016	Declaration of Prashant Shenoy in Support of Patent Owner's Response

Exhibit No.	Description
2017	Deposition Transcript of Dr. Todd Mowry (Feb. 24, 2026)
2018	“Level 3 Lands Netflix Streaming Business, Will Double Its Storage Capability,” TechCrunch, available at https://techcrunch.com/2010/11/11/level-3-lands-netflix-streaming-business-will-double-its-storage-capability/ (last accessed March 1, 2026)
2019	<i>Intentionally Left Blank</i>
2020	“Connect – Most Innovative New Products Award Winners Announced,” Connect NewsLetter, October 26, 1999, available at https://library.ucsd.edu/dc/object/bb60436752/_1.pdf (last accessed March 1, 2026)
2021	“Sandpiper Named ‘Best of the Best’ at Networld + Interop 99 Las Vegas,” PRNewswire, May 12, 1999, available at https://ubmchannelnews.mediaroom.com/index.php?s=29573&item=85841 (last accessed March 1, 2026)
2022	Entrepreneur of the Year, June 19, 2000
2023	“Sandpiper Networks’ Footprint Takes Aim at Internet Traffic,” Wall Street Journal Interactive Edition, June 17, 1999, available at https://url.us.m.mimecastprotect.com/s/wTVWCxklA1iDQQWjSYi2Syt1OK?domain=wsj.com (last accessed March 3, 2026)
2024	“Web-Page Distribution System Could Unclog Internet Traffic Jams,” The New York Time, April 26, 1999, available at https://archive.nytimes.com/www.nytimes.com/library/tech/99/04/biztech/articles/26sand.html (last accessed March 3, 2026)
2025	“Sandpiper Puts Its Footprint in Marketplace,” Los Angeles Times, September 28, 1998
2026	“CenturyLink named a Leader in IDC MarketScape: Worldwide Commercial CDN Vendors Report,” PR Newswire, available at https://www.prnewswire.com/news-releases/centurylink-named-a-leader-in-idc-marketscape-worldwide-commercial-cdn-vendors-report-300933112.html (last accessed March 1, 2026)

Exhibit No.	Description
2027	Google's Preliminary Claim Constructions, (<i>Sandpiper CDN, LLC v. Google LLC</i> , No. 2:24-cv-03951)

Petitioner's Reply fails to cure the Petition's flawed arguments. The Reply relies heavily on attorney argument, much of which is either conclusory or incorrect, both legally and factually. Petitioner fails to rebut the well-supported opinions of Patent Owner's expert, Dr. Shenoy, who explained in detail why none of Petitioner's grounds renders the challenged claims obvious. Indeed, Petitioner did not even take Dr. Shenoy's deposition or submit any Reply declaration countering Dr. Shenoy's opinions with evidence or expert opinion. Petitioner thus fails to engage with Patent Owner's evidence at all, much less does Petitioner's Reply overcome it. The Board should uphold the patentability of the claims.

I. PETITIONER'S INCONSISTENT POSITIONS BETWEEN THE DISTRICT COURT AND PTAB WARRANT TERMINATION OF THE PROCEEDING.

As a preliminary matter, Petitioner has taken the position in the parallel district court litigation that the "receiving an instruction to cache" limitation of claim 3 is indefinite, yet it conveniently assumes the limitation is definite when arguing for obviousness in the Petition. POR, 53-54; EX2027, 3-4. Petitioner attempts to excuse its double-dealing by arguing that its district court indefiniteness position is merely an identification of a structural "internal inconsistency" because dependent Claim 3 adds a cache instruction to a parent claim that already recites a no-cache instruction. Reply, 24-25. That defense

merely downplays, rather than justifies, Petitioner's choice to pursue different positions in different proceedings.

As a result, Petitioner cannot escape the fact that it is advancing entirely irreconcilable legal theories to gain a tactical upper hand across separate forums. In the parallel district court litigation, Petitioner has affirmatively flagged this limitation as legally indefinite and incapable of being understood by a POSITA. EX2027, 3-4. Yet, in this proceeding, Petitioner's expert bypasses that assertion entirely, "assum[ing]" the claim is definite so that Petitioner can attempt to invalidate it via an obviousness challenge. EX1006, ¶¶288-298.

Petitioner argues that the Board has noted that "identifying an incompatibility" may be proper. Reply, 24 (citing *Tesla, Inc. v. IV II Ltd.*, IPR2025-00340, Paper 18 at 3-4 (P.T.A.B. Nov. 5, 2025)). However, Petitioner misinterprets the caselaw. In *Tesla*, the Board emphasized that a petitioner "advanc[ing] different [claim construction] positions ... is *required* to explain why those different positions are warranted," such as pointing to true claim boundary uncertainty. *Tesla*, IPR2025-00340, Paper 18 at 3. Alleged claim incompatibility does not fit within the framework articulated in *Tesla*.

Because Petitioner has failed to provide a valid, uniform construction for Claim 3 across different fora, its Petition should be terminated. *See, e.g., Terumo BCT Inc. v. Haemonetics Corp.*, IPR2025-01374, Paper 20 (PTAB May 12, 2026).

II. GROUND 1

A. The '573 patent provides ample written description support for the claimed “instruction to not cache.”

The Board has correctly agreed that there is support for “instruction to not cache.” DI, 13-14. However, in its Reply, Petitioner attempts to manufacture a written-description deficiency by adding restrictive operational steps that are entirely absent from the plain language of the claims and are directly counter to the explicit goals of the '573 patent specification.

1. Petitioner improperly reads extraneous limitations into a plainly recited term.

Petitioner mistakenly construes the plain phrase “instruction to not cache” as requiring an instruction to *retrieve content first, and subsequently refrain from caching it*. Reply, 5. This is a distortion of the claim language. “Instruction to not cache” means exactly what it says: an instruction dictating that caching not take place. It does not dictate a temporal sequence requiring a server to pull a resource into memory or fetch it from an origin server prior to enforcing the non-caching directive. By arguing that the specification fails to describe a specific “retrieve-then-do-not-cache” routine, Petitioner improperly attempts to import limitations into a phrase that clearly does not include them.

2. The specification explicitly contemplates avoiding retrieval to maximize efficiency.

A POSITA would have understood that when a server receives an instruction not to cache an item, it is not required to retrieve the content in order to comply with that instruction. In fact, avoiding retrieval entirely is the most textually and technically accurate embodiment of a “not to cache” instruction. The ’573 patent, and Provisional Application No. 61/042,412 to which the ’573 patent claims priority, explicitly detail the technical problem of “serv[ing] long-tail content from an edge cache,” explaining that doing so creates significant inefficiencies because “unnecessary cache fills consume bandwidth both from the storage and cache systems.” EX1001, 10:29-37; EX1011, ¶69.

The patent explicitly recognizes that unnecessary caching exhausts vital bandwidth between the origin server and the edge server. The ’573 patent teaches that an edge server can minimize network strain by not retrieving the content upon receiving an instruction not to cache. EX1001, 10:29-37; EX1011, ¶69. This forward-only or redirect behavior without edge retrieval is a direct implementation of the patent’s goal to eliminate wasteful bandwidth consumption.

Accordingly, the Board should maintain its findings that the ’573 patent is entitled to its April 4, 2008, priority date and that Newton-471 is not prior art. Ground 1 thus fails as a matter of law.

III. GROUND 2

A. None of the asserted references teach “a popularity designation” (independent claims 1, 11, 19).

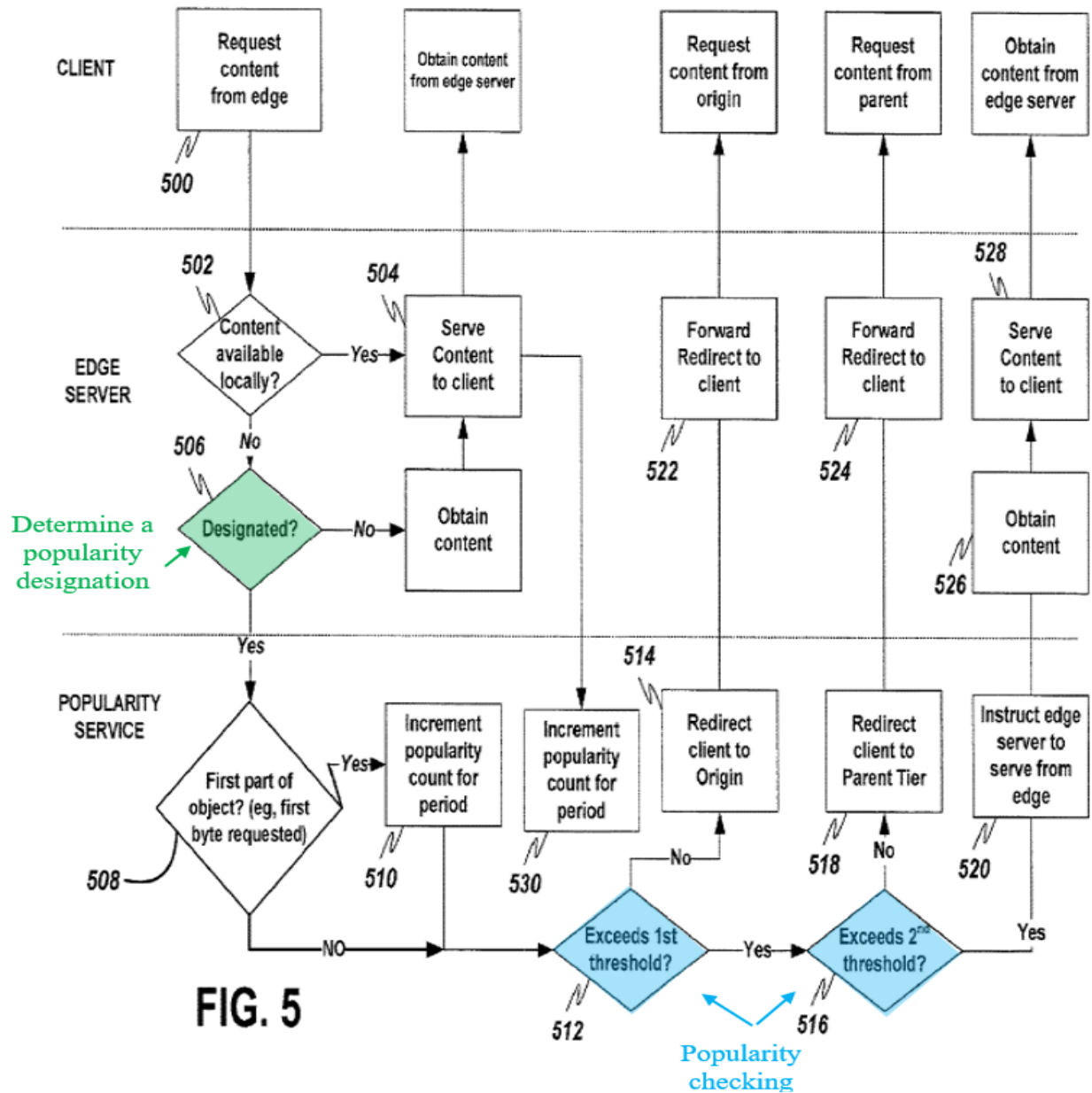
As the POR demonstrates, the Petition fails to establish that the asserted art teaches “*accessing a popularity service associated with the content delivery network to determine a popularity designation associated with the requested resource,*” as required by independent claims 1, 11, and 19. POR, 26-29; EX2016, ¶¶63-68. Petitioner’s Reply fails to cure this core deficiency. Instead, Petitioner continues to conflate Pai’s “popularity value” with the claimed “popularity designation” that governs initial handling. Reply, 10-13. As Dr. Shenoy explains, Pai’s popularity value is not the claimed popularity designation. EX2016, ¶¶63-68. Petitioner does not introduce any evidence to rebut Dr. Shenoy’s testimony on this point, as Petitioner did not depose Dr. Shenoy or submit a rebuttal declaration from its own expert. Petitioner’s attorney argument is not rebuttal evidence. The Board should credit Dr. Shenoy’s unrebutted opinion.

1. Intrinsic evidence establishes that determining a popularity designation is an eligibility check.

Petitioner contends in the Reply that the specification does not establish that a “popularity designation” requires an eligibility assessment. Reply, 12. Not so.

Ample intrinsic evidence demonstrates that “determining a popularity designation” is a prerequisite check to establish whether a requested resource is even eligible for popularity-based routing. POR, 26-27; EX1001, 5:47-50, 5:62-64,

6:53-7:24, FIG. 5. The '573 patent explicitly discloses that the edge server first ascertains “whether this object has been **designated** so that the object’s popularity will be used to determine where it will be served from.” EX1001, 6:27-31 (emphasis added).



EX1001, FIG. 5 (annotated).

The specification of the '573 patent confirms that “content should be **designated** in order to use the popularity service,” e.g., to perform a “popularity check.” EX1001, 5:47-50 (emphasis added), 5:62-64. As illustrated in the annotated version of Figure 5 above, the “Designated?” qualification (Step 506) is a structural gating step completed *before* the separate numerical popularity checking logic (Steps 512, 516) is ever initiated. POR, 27; EX1001, 6:27-38, FIG. 5. Pai lacks any mechanism recognizing or establishing an initial eligibility “designation.” POR, 28; EX1004, ¶¶39-51.

2. Petitioner’s Reply mischaracterizes the claims and improperly equates relative scope with definiteness.

Petitioner’s Reply contends that Patent Owner improperly imports an unrecited limitation because “the plain claim language does not recite a determination of ‘eligibility’ for popularity-based handling.” Reply, 11. However, claim terms cannot be determined in a vacuum; they must be assessed “in view of the specification,” which serves as the single best guide to their meaning. *Phillips v. AWH Corp.*, 415 F.3d 1303, 1315 (Fed. Cir. 2005) (*en banc*). As clearly established by the POR, and discussed earlier, the specification of the '573 patent makes clear that the determination of a popularity designation is a gating step that is, by definition, executed before performing popularity value checking. POR, 26-27; *supra* Section II.A.1. Petitioner again conflates the popularity designation with

a popularity value by alleging that “[e]ach [independent claim] requires determining the resource’s popularity.” Reply, 11.

Petitioner’s Reply further alleges that dependent claims confirm the understanding that “‘popularity designation’ simply requires making the popularity threshold determination.” Reply, 12. However, under 35 U.S.C. § 112(d), a dependent claim functions strictly to add a “further limitation,” not to redefine or narrow the meaning of individual terms within its parent independent claim. Under the doctrine of claim differentiation, the presence of an added limitation in a dependent claim creates a strong presumption that the independent claim is broader in scope and lacks that specific restriction. *Phillips*, 415 F.3d at 1315.

In fact, Petitioner improperly narrows the meaning of “popularity designation” as merely a restrictive “binary” eligibility determination. Reply, 12. On the contrary, the “popularity designation” can have multiple functions besides indicating baseline eligibility. As Petitioner itself pointed out, the popularity designation indicates whether the requested resource is popular or not, and it “comprises” an indication of “whether a current popularity value ... exceeds popularity threshold.” Reply, 12-13; EX1001, Claims 2, 3, 12, 13, 16. Importantly, the use of the open-ended term “comprises” in dependent claim 16 confirms that it does not restrict or exhaustively define the scope of the parent claim’s term.

Additionally, the '573 patent explains that some content (the “short head”) is “popular enough” to benefit from a caching CDN, and some content (the “long tail”) is “not generally worth caching.” EX1001, 2:49-59. This indicates that popularity is a spectrum which might support various different responses. That is why each content “should be designated” before using the popularity service, and only popular content are designated. EX1001, 2:49-59, 5:47-50. Therefore, in addition to indicating initial eligibility, the broad “popularity designation” indicates whether the content is popular or not, and whether a current popularity value exceeds popularity threshold (e.g., whether the content is “popular enough” to warrant an edge tier presence).

3. Petitioner’s Reply mischaracterizes the Figure 5 architecture of the ’573 patent.

Petitioner attempts to dismiss Figure 5 by arguing that because the yes/no “Designated?” box (Step 506) resides at the edge server prior to reaching the standalone popularity service flowchart block, it cannot be the claimed “popularity designation.” Reply, 12. This arguments rests on a fundamental misreading of both the claim language and the explicit architecture of the ’573 patent.

The plain language of the independent claims requires “*accessing a popularity service associated with the content delivery network to determine a popularity designation associated with the requested resource.*” Crucially, the claims recite a functional relationship—doing one action to achieve another—but

they do not mandate that the popularity service itself must independently generate or output the designation in structural isolation. Rather, as Figure 5 illustrates, the edge server performs the determination of a popularity designation (Box 506) **by accessing** a popularity service. EX1001, FIG. 5.

In other words, the edge server executes a dual operational requirement: 1) it accesses the popularity service and 2) it determines the “popularity designation.” The patent does not require the standalone popularity service to independently execute the designation step from a distance; rather, the edge server utilizes its access to the popularity service as the underlying mechanism to arrive at the eligibility determination.

Thus, Petitioner’s structural timeline is flawed. Even though the “Designated?” box (Step 506) is depicted visually at the edge server level in Figure 5, the edge server is actively “accessing” or querying the network’s popularity service functionality to resolve that very gating step.¹ EX1001, 6:27-38, FIG. 5. An edge server can always access, query, or communicate with a popularity service to

¹ Figure 5 shows Step 506 at the edge server prior to a popularity service block, not prior to accessing the popularity service functionality. Edge servers can always access a popularity service even if they are not co-located.

render a local decision, regardless of whether the components are physically distinct or where they sit on a flowchart.

Additionally, Petitioner's attempt to manufacture a rigid spatial or sequential conflict ignores the clear architectural flexibility disclosed in the specification. The '573 patent explicitly states that "[a]lthough shown in a parent tier, the popularity service may [be] located anywhere in the system, including in the edge tier." EX1001, 5:44-46. Because the popularity service can be entirely co-located with, or seamlessly embedded inside, the edge tier, the structural sequence and layout illustrated in Figure 5 perfectly align with the claimed steps. The edge server's localized execution of Step 506 via direct, co-located access to the popularity service completely undercuts Petitioner's temporal criticisms. Petitioner's artificial distinction between the edge server's actions and the service's location must be rejected.

Because the asserted references fail to reveal or suggest the critical "popularity designation" gating step required by independent claims 1, 11, and 19, Patent Owner respectfully requests that the Board reject Petitioner's obviousness theories and uphold the patentability of all challenged claims.

B. The Unbridged Gap Between Pai’s Internal “Content Placement” Based on Popularity Value and the Claimed “Content Delivery” Based on a Popularity Designation (Claims 1 and 19)

The POR establishes that Petitioner fails to bridge the gap between **content placement** based on popularity **value**, as taught by Pai, and **content delivery** dictated by popularity **designation**, as claimed. POR, 22, 29. As Dr. Shenoy explains, and Petitioner does not provide any evidence to rebut, Dilley and Pai fail to teach content delivery based on a popularity designation. EX2016, ¶¶69-86.

Importantly, Petitioner has already conceded that Pai’s popularity logic is confined strictly to internal “content placement” in the tiered cache hierarchy of Dilly. DI, 24; EX1006, ¶209. To bridge this missing link and meet independent Claim 19’s recited “*requesting the resource from a second server ... based at least on the popularity designation,*” Petitioner resorts to a wave of the hand, asserting that because storage placement and real-time retrieval are “closely related,” Pai’s placement data can logically inform an external retrieval request. Reply, 7-8, 15-16. This argument fails as a matter of law and technology.

1. “Closely related” does not satisfy the particularity requirements of obviousness.

Petitioner’s reliance on a “close relationship” between two distinct networking actions is an explicit admission that a gap (albeit an allegedly narrow one) exists between the prior-art disclosures and the claim limitations. Reply, 7-8. Showing that a prior art feature is merely “closely related” to a claim limitation,

therefore, does not meet a Petitioner's burden to prove that the prior art actually teaches or suggests the requirement. *See Unified Patents Inc. v. DivX, LLC*, IPR2019-01379, Paper 52 (P.T.A.B. Feb. 8, 2021) (persuasive) (holding that disclosing closely related subject matter is legally insufficient to establish an obviousness ground). Internal storage allocation (where an item sits in a cache hierarchy) is architecturally distinct from active, external request routing (generating a real-time server request based on an operational trigger). POR, 32. Petitioner cannot leverage vague commonalities to gloss over missing structural components.

2. Dilley and Pai fail to teach content delivery based on a popularity designation.

The POR has demonstrated that Pai does not disclose “a popularity designation.” POR, 26-28, 32; EX2016, ¶¶69-86. As discussed above, the '573 patent utilizes a popularity designation as an initial eligibility gate; if a resource lacks this designation, it is handled by the edge server without popularity processing. EX1001, 6:39-52, FIG. 5; *supra* Section II.A.

Pai, by contrast, teaches a system where *all* media content is inherently subject to distribution and redistribution based on numerical “popularity values,” and using the numeric popularity value to manage storage of media content. POR, 28-29; EX1004, ¶¶39-51. Pai lacks any concept of an eligibility “designation,” has no structural reason to check for one, and discloses no step or routing action to

perform based on such a designation. POR, 28-29; EX1004, ¶¶39-51. Because Pai's system does not possess "a popularity designation," it cannot take an operational routing action based on a parameter it completely lacks. POR, 28-29. Accordingly, the combination of Dilley and Pai fails to teach content delivery based on a popularity designation.

Petitioner ignores this fundamental distinction, choosing instead to repeat its flawed conflation of Pai's numerical values with the claimed "popularity designation," while still relying on conclusory expert testimony to fill the gap. Reply, 14-15; EX1006, ¶¶354-355. In fact, Petitioner effectively concedes the existence of this architectural gap by admitting that the Petition merely "uses Pai's popularity-based distribution system to yield a combination where server selection for retrieval is **based on popularity driven placement.**" Reply, 14 (emphasis added). This is the definition of an impermissible hindsight reconstruction.

3. Petitioner fails to rebut Patent Owner's showing that its theories use impermissible hindsight.

As established by the record, Pai's popularity metrics dictate where files are stored across different storage tiers. POR, 20-21; EX1004, ¶¶13, 20, 39-51. It does not, however, utilize that metric as a real-time routing trigger to generate a retrieval request to a second server. POR, 34. Patent Owner demonstrated in its POR with supporting expert testimony that Petitioner's proposed combination of Dilley and

Pai relies entirely on impermissible hindsight to bridge this technical gap. POR, 34-35, EX2016, ¶¶78-79.

Rather than engaging with this substantive defense or pointing to any objective evidence in the prior art that would suggest modifying an internal placement scheme into an external request routing mechanism, Petitioner's Reply completely ignores the argument. Instead, Petitioner simply doubles down on the unsupported assertion that the mechanisms are "closely related." Reply, 16.

Petitioner cannot meet its burden of proving obviousness by rewriting the references through the convenient lens of the '573 patent's own disclosure while leaving Patent Owner's detailed hindsight defense unaddressed.

C. None of the asserted references teaches "a redirect instruction" (independent claims 1, 11).

As established in the POR, Petitioner advanced a flawed obviousness theory in the Petition. POR, 36-38. Specifically, Petitioner first erroneously alleged that Dilley's second server would "instruct the first server to request the content from a server higher on the hierarchy, such as [a parent or] an origin server." Pet., 50. Petitioner then relied on conclusory expert testimony to assert that Dilley "does not provide specifics on how that retrieval happens." EX1006, ¶252. Based on these two unsupported premises, Petitioner alleged that "a POSITA would have been motivated to implement this redirect instruction from Wang." Pet., 50. Petitioner's original theory attempted to manufacture a "gap" regarding how Dilley's parent

server instructs an edge server to retrieve content in order to justify introducing Wang to fill the “gap.” POR, 42. As Dr. Shenoy explains—and Petitioner does not introduce any evidence to rebut—cited references fail to teach a redirect instruction. EX2016, ¶¶87-112.

In an attempt to salvage its flawed obviousness theory, Petitioner’s Reply pivots to a completely new theory not found in the Petition: arguing that even if Dilley is not silent and uses an entirely different mechanism, Wang simply supplies a “[k]nown [a]lternative” path that a POSITA would predictably substitute. Reply, 18-19. This untimely new argument fails as a matter of law and technology.

1. Petitioner’s original obviousness rationale fails because the alleged gap in Dilley that Petitioner uses to justify the introduction of Wang’s redirect instruction does not exist.

In the POR, Patent Owner demonstrated that Petitioner’s alleged “gap” in Dilly does not exist, and therefore does not need to be filled by a secondary reference. Dilley relies on a complete, autonomous edge-based metadata routing mechanism. POR, 43-45. Petitioner misapprehends Patent Owner’s argument, alleging in its Reply that the passages of Dilley cited by Patent Owner do not address the scenario “when content is a candidate for cache hierarchy and is not located at the edge.”² In that circumstance, Wang supplies the necessary details

² Patent Owner believes Petitioner meant “parent” here, instead of “edge.” In

regarding how a second server resolves a cache-miss scenario by causing the first server to retrieve content from the origin.” Reply, 18-19. Not so.

On the contrary, Dilley explicitly describes this exact scenario, but takes a different approach from Pai or the challenged claims. Specifically, Dilly discloses that if the cache hierarchy feature is active for an object (e.g., the object’s metadata includes a “use-hierarchy” tag), then the edge server requests the object from a parent server. EX1003, 8:22-28, 8:38-45. If the parent server’s cache does not have the requested object, the parent server itself “mak[es] a request to the origin server.” EX1003, 8:53-58.

Therefore, in this exact scenario, Dilley’s parent server fetches the object from the origin server instead of instructing the edge server to do so. POR, 45. Because Dilley teaches a complete and functional retrieval path, there is no missing operational detail for Wang to supply, and no reason to arrive at the claim limitation.

Dilley’s terminology, the edge server sends a request to a parent server in the cache hierarchy.

2. Petitioner’s new “known alternative” theory is legally improper and substantively flawed.

Petitioner’s late-stage pivot to a “known alternative” theory in its Reply fundamentally shifts its motivation to combine Dilley and Wang. Reply, 19. This shifting argument fails as a matter of law.

First, Petitioner’s late-stage pivot constitutes an impermissible new rationale that cannot be raised for the first time in a Reply. Such shifting arguments are improper. *See* 37 C.F.R. § 42.23(b); *Yita LLC v. MacNeil IP LLC*, 69 F.4th 1356, 1365–66 (Fed. Cir. 2023) (explaining 37 C.F.R. § 42.23(b) means that an IPR petitioner may not raise in its reply an entirely new rationale). A petitioner may not treat its motivation to combine as a moving target to escape its initial failed burden of proof. *See Intelligent Bio-Sys., Inc. v. Illumina Cambridge Ltd.*, 821 F.3d 1359, 1369–70 (Fed. Cir. 2016).

Second, merely labeling Wang as a “known alternative” does not satisfy the legal requirement to show a specific, rational reason why a POSITA would modify Dilley. Under *KSR*, there must be some articulation of a predictable benefit or reason to choose the alternative. *See KSR Int’l Co. v. Teleflex Inc.*, 550 U.S. 398, 416–18 (2007). Petitioner fails to provide one.

In fact, implementing Wang’s upstream HTTP redirection into Dilley’s metadata-driven architecture creates a redundant and superfluous architecture. POR, 49-50. As established in POR, Dilley’s edge server already executes a

localized “guide process” to determine direct-to-origin routing, adding a secondary redirect logic at the parent tier forces the system to execute the exact same bypass result through a more complex, multi-step transaction. POR, 49-50. A POSITA would not introduce unnecessary logic to an upstream server to handle requests that, by definition, already bypass that server under Dilley’s baseline rules.

Additionally, as demonstrated below, Wang’s redirect mechanism actively disrupts the network’s structural design of Dilley and provides absolutely no predictable performance benefit. A POSITA would therefore have no motivation to modify Dilley by Wang’s teaching. *See In re Kahn*, 441 F.3d 977, 986–88 (Fed. Cir. 2006) (holding that there must be an articulated reasoning with a clear “rational underpinning” demonstrating why a skilled artisan would choose to implement such a configuration). Because Petitioner’s proposed modification introduces architectural redundancy without a corresponding technical benefit, its stated motivation lacks the required rational underpinning as a matter of law.

3. Petitioner’s proposed combination fails because Wang’s redirect mechanism actively degrades system performance and offers no latency benefit.

Petitioner’s assertion that incorporating Wang’s redirect mechanism into Dilley’s tiered CDN would “decrease latency” and “improve server availability” relies on a fundamental misunderstanding of hierarchical caching dynamics and directly conflicts with the explicit disclosures of the prior art. POR, 47-49.

a. Petitioner’s argument mischaracterizes the stated purpose and selective activation of Dilley’s hierarchy.

As established by the POR, Dilley’s cache hierarchy is intentionally designed to achieve two specialized goals: (1) shield the origin infrastructure from being overwhelmed by requests for large or popular objects and (2) “reduce[] average latency and improve[] connectivity and content access in what might otherwise be poorly-connected edge regions.” POR, 48; EX1003, 5:16-21, 6:4-15. Dilley explains that while the vast majority of standard content bypasses the hierarchy by default, the hierarchy is selectively activated for a specific subset of objects to protect the origin server and maximize connectivity for these under-connected edge regions. POR, 48; EX1003, 5:16-21, 6:4-15.

In the Reply, Petitioner attempts to dismiss this by claiming that Patent Owner’s argument conflicts with Dilley’s disclosure that most content is fetched by bypassing the hierarchy. Reply, 20. This is a non sequitur. The fact that standard content routinely bypasses the hierarchy under a general policy does not alter the fact that when Dilley does choose to engage the hierarchy, it does so precisely to at least aid poorly connected regions. Petitioner further mischaracterizes Dilley by asserting that these “poorly connected edge servers request content directly from the origin without issue.” Reply, 20-21. Dilley contains no such disclosure. To the contrary, Dilley explicitly routes content

through parent regional hubs partly because direct-to-origin communication could introduce severe latency for those specific regions. EX1003, 5:16-21, 6:4-15.

b. Eliminating the parent tier induces latency and threatens upstream infrastructure thrashing.

Petitioner contends that forcing the resource to travel directly from the origin to the edge—rather than traversing the parent server—saves bandwidth and decreases latency by “preventing the second server from obtaining and processing large or popular content unnecessarily.” Reply, 21. Petitioner further argues that Wang provides a technique by which an upstream server can instruct a downstream server (e.g., the first server) to retrieve content directly from an origin server with or without caching. Reply, 22. This argument completely subverts Dilley’s core design principles and betrays a fundamental misunderstanding of hierarchical network architecture.

First, bypassing the parent tier in a poorly connected edge region does not reduce latency; it increases it. POR, 48. Dilley explicitly teaches that parent servers “act as accelerators for the end users.” EX1003, 5:16-21. Forcing a poorly connected edge node to establish a brand-new, unoptimized direct connection to a distant origin server—as Wang’s redirect mechanism would demand—creates an immediate technical bottleneck. POR, 48. Moreover, Dilley’s hierarchy of servers is structurally deployed to ensure that regional hubs are positioned “closer to the requesting end users.” EX1003, 9:57-59. A POSITA would understand that for

data to traverse a network directly from the origin to the edge, it must still route through multiple intermediate network nodes. By systematically bypassing an optimized parent server that is physically and logically closer to the edge node, Wang's modification risks routing traffic through arbitrary intermediate servers that are further away, thereby increasing latency and degrading performance.

Second, Petitioner's argument that parent tiers should not "process large or popular content unnecessarily" directly contradicts Dilley's structural objective. Dilley dictates that parent servers must act "as buffers for the origin site's internal infrastructure" specifically to ensure the origin site is "not overwhelmed with requests for popular content or large files." EX1003, 5:16-21. By utilizing Wang's redirect to strip the parent tier of its caching and processing functionality for these popular and large assets, Petitioner effectively destroys the centralized protective buffer.

Even if Wang's method allows an individual edge server the optional capacity to cache the popular content locally upon direct retrieval, it does absolutely nothing to protect the origin server from the broader network traffic. Because the parent buffer has been completely bypassed and excluded from the distribution path, clients connected to *other* edge regions who subsequently request that same popular content cannot benefit from a localized edge cache. Instead, each of those separate edge servers will be structurally forced to generate independent,

redundant cache-fill requests straight back to the origin site. Additionally, Dilley already discloses caching content at edge server. EX1003, 8:15-18, FIG. 6.

Rather than achieving an elegant optimization, Petitioner's modification fragments the CDN's architecture—transforming Dilley's coordinated, multi-tiered buffering system into an isolated cluster of independent nodes that continuously expose the core origin infrastructure to severe data thrashing. Because Petitioner's proposed modification disrupts the basic operation and structural benefits of the primary reference, a POSITA would have had no rational motivation or reasonable expectation of success to combine these teachings. *See In re Ratti*, 270 F.2d 810, 813 (C.C.P.A. 1959) (holding that a proposed combination is improper and fails to establish obviousness where the modification changes the underlying principle of operation of the primary reference).

c. Petitioner's "server availability" argument is technically illusory and subverts the purpose of a buffering tier.

Petitioner also argues that bypassing the parent hub via a redirect frees upstream resources, asserting that even if parent servers possess concurrent processing capabilities, preventing them from handling large or popular objects maximizes their overall availability. Reply, 21-22. This reasoning is flawed.

Petitioner's premise that processing popular items at the parent hub is "unnecessary" completely misses the structural engineering of a tiered CDN. In

Dilley's system, parent servers intentionally absorb the heavy computing, bandwidth, and processing burdens associated with large, popular files specifically to protect the origin server from being overwhelmed with requests. EX1003, 5:16-21. Shifting that asset-handling burden away from the parent hub via an HTTP redirect does not elegantly preserve parent resources; rather, it forces the downstream edge tier to hit the origin infrastructure repeatedly for subsequent cache fills.

Accordingly, Petitioner's Reply introduces an impermissible shifting of legal theories compared to the Petition and fails to account for the complete, autonomous routing mechanisms disclosed in the prior art.

D. The asserted references fail to teach a peer-to-peer redirect to the origin server (claims 5 and 15).

Claims 5 and 15 require that the "second server" providing the redirect instruction be a "peer server" of the first tier. POR, 51; EX2016, ¶¶113-117. In the Reply, Petitioner simply argues that because Dilley supplies a peer-server query scenario and Wang supplies a redirect technique, a POSITA would have found it an obvious design choice to apply Wang's technique to Dilley's peer configuration to achieve a peer-to-peer redirection. Reply, 23. However, Petitioner's attempt to map this limitation to the combination of Dilley and Wang relies on a highly selective, distorted extraction of the prior art. Reply, 23.

While Petitioner correctly notes that Dilley teaches peer-server queries when an edge server lacks an object, Petitioner intentionally omits Dilley's mandatory next step when a peer query fails to return a valid object. Reply, 23. In Dilley's actual architecture, if "no peer copy of the object is available," the edge server continues to query its own localized "guide process" to establish a connection upward to a parent server. EX1003, 8:18-25, FIG. 6.

Petitioner's hypothesis that a peer server would instead issue a redirect instruction forcing the first edge server directly back to the origin site completely defeats Dilley's core principle of operation. Dilley explicitly mandates that parent servers must be utilized within the cache hierarchy to protect the origin server from unnecessary requests and avoid overwhelming its internal infrastructure. EX1003, 5:16-21. Imputing a peer-to-peer redirect that bypasses the parent tier entirely strips away the network's buffering capabilities, forcing low-level edge nodes to hit the origin server prematurely. Because doing so would actively destroy the protective buffering advantages taught by the primary reference, a POSITA reading Dilley and Wang would not have been motivated to apply Wang's redirect technique to Dilley's peer-server. *See In re Ratti*, 270 F.2d at 813.

In Dilley, the "second server" in the chain cannot be fluidly interchanged as either a peer server or a parent server to suit Petitioner's needs. Both tiers serve distinct structural roles and must be queried sequentially. EX1003, 8:18-25, FIG. 6.

Because neither reference teaches or suggests a peer server capable of issuing a redirect instruction to the origin, Petitioner has failed to meet its burden for dependent claims 5 and 15.

Accordingly, Petitioner has failed to meet its burden of proving unpatentability for any of the challenged claims. Patent Owner respectfully requests that the Board reject Petitioner's obviousness theories and uphold the patentability of all challenged claims of the '573 patent.

IV. CONCLUSION

The Board should uphold the patentability of all challenged claims of the '573 patent.

Respectfully submitted,

STERNE, KESSLER, GOLDSTEIN & FOX PLLC

/Jason D. Eisenberg/

Jason D. Eisenberg, Reg. No. 43,447

Daniel Block, Reg. No. 68,395

Kristina Caggiano Kelly, Reg. No. 58,164

Xiaowei Jin, Reg. No. 81,041

Counsel for Patent Owner

Date: June 24, 2026

1101 K Street, NW, 10th Floor
Washington, DC 20005
(202) 371-2600

CERTIFICATE OF WORD COUNT (37 C.F.R. § 42.24(d))

1. This Patent Owner Sur-Reply complies with the type-volume limitation of 5,600 words, comprising 5,283 words, excluding the parts exempted by 37 C.F.R. § 42.24(a)(1).
2. This Patent Owner's Response complies with the general format requirements of 37 C.F.R. § 42.6(a) and has been prepared using Microsoft® Word 2016 in 14-point Times New Roman font.

Respectfully submitted,

STERNE, KESSLER, GOLDSTEIN & FOX PLLC

/Jason D. Eisenberg/

Jason D. Eisenberg
Registration No. 43,447
Counsel for Patent Owner

Date: June 24, 2026

1101 K Street, NW, 10th Floor
Washington, DC 20005
(202) 371-2600

CERTIFICATE OF SERVICE (37 C.F.R. § 42.6(e))

I certify that the above-captioned **PATENT OWNER'S SUR-REPLY TO PETITIONER'S REPLY** was served in its entirety on June 24, 2026, upon the following parties via electronic mail:

Erika H. Arner (Lead Counsel)	erika.arner@finnegan.com
Kara A. Specht (Back-Up Counsel)	kara.specht@finnegan.com
Cara E. Regan (Back-Up Counsel)	cara.regan@finnegan.com
Sydney R. Kestle (Back-Up Counsel)	sydney.kestle@finnegan.com
William C. Neer (Back-Up Counsel)	william.neer@finnegan.com
Daniel C. Tucker (Back-Up Counsel)	daniel.tucker@finnegan.com
Christopher B. Anderson (Back-Up Counsel)	christopher.anderson@finnegan.com
	Google-Sandpiper-IPRs@finnegan.com

FINNEGAN, HENDERSON, FARABOW, GARRETT & DUNNER, LLP

Respectfully submitted,

STERNE, KESSLER, GOLDSTEIN & FOX PLLC

/Jason D. Eisenberg/

Jason D. Eisenberg
Registration No. 43,447
Counsel for Patent Owner

Date: June 24, 2026

1101 K Street, NW, 10th Floor
Washington, DC 20005
(202) 371-2600