

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

DECLARATION OF DR. PRASHANT SHENOY

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

Sandpiper EX2016
Google v. Sandpiper
IPR2025-00860

TABLE OF CONTENTS

I.	INTRODUCTION	1
II.	LIST OF MATERIALS CONSIDERED	7
III.	BACKGROUND OF THE TECHNOLOGY.....	9
IV.	OVERVIEW OF THE '573 PATENT	10
V.	PROSECUTION HISTORY	12
VI.	UNDERSTANDING OF RELEVANT LEGAL STANDARDS	14
A.	Person of Ordinary Skill in the Art (“POSA”).....	14
B.	Claim Construction.....	15
C.	Obviousness.....	15
D.	Written Description	17
VII.	ARGUMENT.....	18
A.	Ground 1	18
B.	Ground 2.....	21
1.	Overview of References.....	21
a.	Overview of Dilley	22
b.	Overview of Pai	23
c.	Overview of Wang.....	23
2.	The combination of Dilley, Pai, and Wang does not teach or suggest the feature of “determin[ing] a popularity designation.”	24
3.	The references do not teach or suggest “requesting the resource from a second server... based at least on the popularity designation” as required by claim 19.	27
a.	Petitioner relies on conclusory testimony to fill gaps and conflates internal distribution with external delivery.	28
b.	The Petition arguments are based on impermissible hindsight.	30

c.	Petitioner and Dr. Mowry’s motivation to combine Dilley and Pai is based on mischaracterization of Dilley.	31
d.	A POSA would not have been motivated to combine Dilley and Pai because their teachings are incompatible, rendering one or both inoperable for their intended purpose or changing their principle of operation.	32
4.	The references fail to render obvious the claimed “redirect instruction” that is sent from the second server of the content delivery network to the first server (Claims 1(d)/11(d)).	33
a.	Petitioner and Dr. Mowry erroneously allege that Dilley’s second server instructs the edge server to retrieve content from the origin server, or that Dilley does not provide specifics on how that retrieval happens.	35
b.	Dilley’s edge server, not the second server, determines where the edge server should fetch content based on the metadata.	36
c.	A POSA would not have been motivated to implement Wang's redirect procedure in Dilley's system because there are no benefits and no expectation of success.	39
i.	Petitioner and Dr. Mowry’s theory of motivation relies on improper hindsight reconstruction.	39
ii.	Petitioner’s alleged benefit of reducing latency or improving server availability are technically illusory.	40
iii.	The proposed modification results in a superfluous and redundant architecture.	43
5.	The references fail to teach or suggest that the second server is a peer server that sends a redirect instruction to the first server, as recited in dependent claims 5 and 15.	44
VIII.	CONCLUSION.....	46

I, Prashant Shenoy, Ph.D., do hereby declare as follows.

I. INTRODUCTION

1. I am over the age of twenty-one and otherwise competent to make this Declaration. This Declaration is based on my personal knowledge as an expert in the relevant field. I could testify competently if asked to do so.

2. I have been retained by Sterne, Kessler, Goldstein & Fox, P.L.L.C. as a technical expert witness on behalf of Sandpiper CDN, LLC to provide this Declaration concerning technical subject matter relevant to the *inter partes* review of U.S. Patent No. 10,924,573 (“’573 patent”) in IPR2025-00860.

3. I am being compensated for the time spent on this matter at my standard consulting rate of \$650.00/hour. My compensation is not dependent on, and in no way affects, the substance of my statements in this Declaration. My findings are based on my education, experience, and background in the fields discussed below. I reside in Amherst, Massachusetts, USA.

4. In formulating my opinions, I have relied upon my experience in the relevant art, and I have considered each document cited herein in, in light of the general knowledge in the art as of April 4, 2008, the effective filing date of the ’573 patent. For example, I have considered the documents listed in Section III in addition to any other documents cited in this declaration. To the best of my knowledge, these documents are true and accurate copies of what they purport to

be. An expert in the field would reasonably rely on them to formulate opinions such as those set forth in this declaration.

5. I will cite to the specification of a U.S. patent using the following format: (Patent, Col: Line Number(s)). For example, the citation to the '573 patent, 1:1-10 points to the '573 patent specification at column 1, lines 1-10.

6. I am familiar with the technology at issue and the state of the art at the time the application leading to the '573 patent was filed, including content delivery networks, web protocols, network protocols, and proxy servers.

7. I have been asked to provide my technical review, analysis, insights, and opinions regarding the references. I reserve the right to supplement this Declaration in response to additional evidence that may come to light.

8. I earned a Bachelor of Technology degree from the Department of Computer Science and Engineering at the Indian Institute of Technology in Bombay, India. I later received my Master of Science (M.S.) and Doctor of Philosophy (Ph.D.) degrees in 1994 and 1998, respectively, in Computer Science from the University of Texas at Austin.

9. Over the years, I have led research and industry collaboration projects in areas such as storage and file systems, Web and Internet systems, content distribution, streaming media, database systems, cloud computing, virtualization, RFID and sensor networks, and energy systems. During that time, I have consulted

with both large companies and startups in a technical advisory role related to product architecture, IP issues, and applied research.

10. I am a tenured professor in the College of Information and Computer Sciences at the University of Massachusetts Amherst (“UMass”), where I currently hold the position of Distinguished Professor of Computer Science. I previously served as an Associate Dean in the College of Information and Computer Sciences from 2016 to 2025. I have been a faculty member at UMass since 1998.

11. I received my Bachelor of Technology (B.Tech.) degree in 1993 from the Department of Computer Science and Engineering at The Indian Institute of Technology (“IIT”), Mumbai, India. IIT awarded me the Institute Medal for being the top-ranking undergraduate student. In 1994, I received my Master of Science (M.S.) degree in Computer Sciences from The University of Texas at Austin. In 1998, I received my Doctor of Philosophy (Ph.D.), also from the Department of Computer Sciences at The University of Texas at Austin. My dissertation, which was in the area of file systems, was recognized by the department as the Best Doctoral Dissertation of 1998-1999.

12. I have been involved in the research and development of distributed computing systems for more than 30 years. I have been a faculty member at UMass since 1998 where I perform research and teach in the areas of distributed systems,

operating systems, computer networking, data management, and multimedia systems, among others.

13. A major focus of my research has been the design of large-scale distributed computing systems. Over the years, I have led research and industry collaboration projects in areas such as computer networking, distributed systems, storage and file systems, Web and Internet systems, content distribution, streaming media, database systems, cloud computing, virtualization, RFID and sensor networks, and energy systems. During that time, I have consulted with both large companies and startups in a technical advisory role related to product architecture, IP issues, and applied research.

14. For example, starting in the early 1990s, my research focused on the development of storage and file systems. During that time, I developed multimedia file system and data management technologies that were designed for efficient storage and retrieval of large amounts of video and audio data. I collaborated with AT&T Bell Laboratories to design a new media server for video streaming. These efforts resulted in U.S. Patent No. 6,079,028 on fault-tolerant architectures for continuous media servers.

15. In the late 1990s, I led the design of a new operating system called QLinux that was based on the open-source Linux operating system. As part of this research, I developed resource management and scheduling techniques that provide

isolation and service guarantees to applications. This technology provided the underpinnings for fair resource management and isolation techniques used in OS virtualization. It was adopted by a startup company for providing resource isolation in its web hosting products.

16. In the early 2000s, I led several projects on web systems and cloud computing. For example, I led research projects on web technologies such as web caching and scale web applications. I also led research projects on cloud computing platforms (which were known as web hosting platforms at the time) involving virtualization, dynamic resource management, replication, fault tolerance, and storage. My group at UMass collaborated with technology companies such as IBM, AT&T, Sprint, Intel, and many others as part of this research.

17. My research on virtualization has involved design of technologies for virtual machines and container resource management and their deployment in cloud computing systems and data centers. I am co-inventor on U.S. Patent No. 8,473,557, jointly assigned to AT&T and UMass, on virtual machine migration.

18. Over the years, I have authored or co-authored more than 300 articles in peer-reviewed journals and conference proceedings that have been cited over 27,000 times. Seventeen papers authored by myself and members of my research group on this project won awards for their technical contributions. Two papers

have won Test-of-Time awards, an award given to a publication that has had the most impact over a ten-year period since its publication.

19. I have served as an Editor for six leading journals and have served as the Technical Program Chair for twenty conferences and workshops in my field. I founded the ACM Special Interest Group on Energy Systems and also served on steering committee of conferences and journals.

20. I am also a Fellow of the ACM, IEEE, AAAS, and AAIA, four professional societies in computing and science.

21. I am a named inventor on three U.S. Patents and one published application (Nos. 2014/0040206, 8,510,596, 8,473,557, and 6,079,028). These documents are generally related to distributed and operating systems, virtualization, and cloud data centers.

22. As a faculty member at UMass, I have also taught undergraduate and graduate courses in operating systems, distributed systems, networking, and related topics for more than twenty years.

23. These courses teach concepts related to operating systems, virtual machines, containers, cloud computing, file systems, and distributed systems. They also involve hands-on course projects on these topics.

24. My *Curriculum Vitae* is attached as EX2015, which contains further details on my education, experience, publications, and other qualifications to render an expert opinion.

II. LIST OF MATERIALS CONSIDERED

Exhibit No.	Description
1001	U.S. Patent No. 10,924,573 (“the ’573 patent”)
1002	U.S. Patent Publication No. 2016/0094471 to Newton et al. (“Newton-471”)
1003	U.S. Patent No. 7,133,905 to Dilley (“Dilley”)
1004	U.S. Patent Publication No. 2006/0271972 to Pai (“Pai”)
1005	U.S. Patent Publication No. 2005/0198250 (“Wang”)
1006	Declaration of Dr. Todd Mowry
1007	Prosecution History of the ’573 patent
1008	U.S. Patent No. 10,218,806 to Newton et al.
1009	U.S. Patent No. 9,762,692 to Fullagar et al.
1010	U.S. Patent No. 8,930,538 to Fullagar et al.
1011	U.S. Provisional Patent Application No. 61/042,412 (“the ’412 provisional”)
1012	U.S. Patent No. 8,667,175 to Thomas et al.
1013	U.S. Patent Publication No. 2006/0015574 to Seed et al.
1014	<i>Curriculum Vitae</i> of Dr. Todd Mowry
1015	Order Granting Stipulation Regarding Amended Complaint and Case Schedule, <i>Sandpiper CDN, LLC v. Google LLC</i> , 2:24-cv-03951, Dkt. 54 (C.D. Cal. Jan. 21, 2025)

Exhibit No.	Description
1016	Sandpiper's Disclosures of Asserted Claims & Infringement Contentions dated Jan. 14, 2025 in <i>Sandpiper CDN, LLC v. Google LLC</i> , 2:24-cv-03951.
1017	Exhibit D to Sandpiper's Disclosures of Asserted Claims & Infringement Contentions dated Jan. 14, 2025 in <i>Sandpiper CDN, LLC v. Google LLC</i> , 2:24-cv-03951.
1018	Letter to Robert H. Reckers <i>re Sandpiper CDN LLC v. Google LLC</i> , No. 2:24-3951-AB-RAO (C.D. Cal.) (Apr. 29, 2025)
1019	Duane Wessels, Web Caching, O'REILLY (Torkington et al. 1st ed. 2001)
1020	Lv, Z., Yang, et al., Multimedia Delivery Grid: A Novel Multimedia Delivery Scheme, In: Li, M., Sun, XH., Deng, Qn., Ni, J. (eds) GRID AND COOPERATIVE COMPUTING. GCC 2003. LECTURE NOTES IN COMPUTER SCIENCE, vol 3032. Springer, Berlin, Heidelberg (2004), https://doi.org/10.1007/978-3-54024679-4_46
1021	Prosecution History of U.S. Patent Application No. 12/408,681
1022	U.S. Patent Publication No. 2007/0055764 A1
2001	"Cloud CDN: content delivery network," <i>Google Cloud</i> , accessed at https://cloud.google.com/cdn (last accessed July 2, 2025)
2004	U.S. Publication No. 2004/0221019 A1 to Swildens et al., published November 4, 2004
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)

Exhibit No.	Description
2019	Welsh, M. et al., A Design Framework for Highly Concurrent Systems, Technical Report No. UCB/CSD-00-1108, EECS Department, University of California, Berkeley, accessed at http://ww2.eecs.berkeley.edu/Pubs/TechRpts/2000/CSD-00-1108.pdf (last accessed March 1, 2026)

III. BACKGROUND OF THE TECHNOLOGY

25. Content delivery networks (CDNs) provide critical services that enable content providers to quickly deliver online content to millions of consumers simultaneously over the Internet. In the early 1990s, the World Wide Web saw increasing adoption, becoming a household staple. This mass adoption led to data congestion issues from the ever-growing number of users seeking to simultaneously access Internet content.

26. A typical computer server in the 1990s could only handle a limited number of simultaneous connections before becoming overloaded. Moreover, signals take time to traverse routers and move through content-congested internet cables. Consumers living far from the physical server(s) hosting content experienced sluggish load times and high latency due to problems such as overloaded servers, congested network segments, and geographic separation.

27. The availability of CDN servers around the world allowed service providers to replicate appropriate content from customers' origin servers to appropriate CDN servers, and transparently rendezvous end-users requesting that

content to the “best” CDN server to deliver that content, all while providing their customers with control over their content and user experience. This service and its architecture was quickly adopted by many others in the industry.

28. End-users would connect to an optimal edge server—one that that was closer to them or that had available capacity. Distributing content across a network of servers alleviated data congestion issues and allowing consumers to connect to edge servers located near them reduced latency. Systems and methods for propagating data from origin servers to edge servers, storing this data on edge servers, and using it to serve client requests are known as “caching.”

29. In the early-to-mid 2000s, demand for CDNs exploded. Content providers like Google and YouTube use CDN technology to meet their enormous data streaming needs, as well as for video search results. EX2001.

IV. OVERVIEW OF THE '573 PATENT

30. U.S. Patent No. 10,924,573 (“the ’573 patent”) is entitled “Handling Long-Tail Content in a Content Delivery Network (CDN),” and it claims priority to U.S. Provisional Patent Application No. 61/042,412, filed on April 4, 2008.

31. CDN technology generally includes multiple servers used to serve content, which can create traffic and bandwidth problems in responding to requests for content. For example, only popular content (the so-called “short head”) benefits from serving through a caching CDN while other content is not generally worth

caching. EX1001, 2:49-55. Technical challenges that arise in the context of content delivery using a network and multiple servers include managing content in a caching CDN, handling requests for resources, and timely providing resources to requesting devices.

32. The inventors of the '573 patent described solutions to these problems in the '573 patent. The '573 patent describes a CDN including a tiered server system. EX1001, Abstract. In some embodiments, a first-tier server attempts to respond to a user request for content, and the first-tier server determines whether the requested content is subject to popularity-based handling by accessing a popularity service to determine a popularity designation for the resource. EX1001, 6:27-38. Based on that designation, the system determines how the content should be obtained. *Id.* Specifically, if the first-tier server does not have the content, and the content is designated, that server will request the content from a higher-tier server and subsequently cache the content for future delivery, for example. The disclosed invention addresses the reality of storage limitations by setting forth a framework that automatically caches only popular content, thus both speeding up content delivery and preserving memory space for popular content, in some cases. EX1001, 6:65-7:30.

33. The inventors devised solutions as shown in the embodiments claimed by the '573 patent, for example methods for using tiers of servers and specific

processes for obtaining content for a requesting device. EX1001, 13:35-54. The solutions recited in the claims of the '573 patent include particular combinations that were not conventional at the time of the invention of the '573 patent, such as specific technical improvements to content delivery networks, and the '573 patent describes how the solutions are implemented.

V. PROSECUTION HISTORY

34. The '573 patent was examined as U.S. Application No. 16/284,791 ("791 application"), which was filed on February 25, 2019. EX1001, (21), (22).

35. In a Non-Final Office Action mailed on July 24, 2020, Examiner Winder rejected prosecution claim 11 in view of U.S. Patent No. 8,930,538 ("Fullagar"), and claims 1-2, 4-7, 11-12, and 14-18 in view of U.S. Patent No. 10,348,848 ("Newton"). EX1007, 60-71. Additionally, claims 1, 3-8, 10-11, and 14-20 were rejected in view of U.S. Patent No. 7,552,233 ("Raju") and U.S. Publication No. 2004/0221019 ("Swildens"). Claims 9 and 12-13 were rejected in view of Raju, Swildens, and U.S. Patent No. 7,024,466 ("Outten"). EX1007, 71-77.

36. On September 17, 2020, Applicant filed a terminal disclaimer, as well as various amendments and remarks in response to the Examiner's rejections. EX1007, 40-50. Specifically, Applicant amended claim 1 to include subject matter disclosed in the originally filed claim 2. Independent claims 11 and 19 were

amended similarly. For example, claims 1 and 2 are amended as follows:

1. (Currently Amended) A method of content delivery in a content delivery network comprising:

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;

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

requesting the resource from a second server of the content delivery network;

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;

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

providing the obtained resource to the requesting device.

2. (Currently Amended) The method as recited in claim 1 wherein the popularity designation associated with the requested resource indicates that the requested resource is not popular, ~~the method further comprising: receiving an instruction to not cache the resource at the first server of the first tier of servers when the resource is obtained from the content server of the content delivery network.~~

EX1007, 40-41.

37. This response successfully overcame the rejections. On December 31,

2020, the Examiner issued a Notice of Allowance. Examiner Winder noted that the closest art (Raju and Swildens) does not teach “sending an instruction to not cache the content item” as recited in the claims. EX1007, 17.

VI. UNDERSTANDING OF RELEVANT LEGAL STANDARDS

38. In formulating my opinions set forth in this Declaration, I applied the following legal principles.

A. Person of Ordinary Skill in the Art (“POSA”)

39. I understand that patent law analyses are performed from the viewpoint of a POSA. I understand that a POSA is a hypothetical person who is presumed to be aware of all pertinent art, thinks along conventional wisdom in the art, and is a person of ordinary creativity. For purposes of this declaration, I have applied Petitioner’s articulation of a POSA. *See* Pet., 12; EX1006, ¶¶70-72 (“A POSITA at the time of the alleged invention (including either April 4, 2008, or February 25, 2019), would have had at least a bachelor’s degree in computer science, electrical engineering, or a related field, and at least two years of work or research experience in the field of content delivery management or networks. Work experience can substitute for formal education and additional formal education can substitute for work experience.”). It is my opinion that the definition of a POSA would not materially affect the limited analysis required for deciding institution and presented in this declaration. If called to give further opinions in

this proceeding, I reserve the right to reconsider the definition of a POSA and arrive at my own conclusion.

B. Claim Construction

40. I understand that in an IPR, patent claim terms are given their ordinary and customary meaning as understood by a POSA at the time of the invention, in view of the patent's specification and its prosecution history—unless the patent explicitly defines the claim term. When a patent explicitly defines a claim term in the specification, that is the definition that informs the analysis, even if it is different than the ordinary or customary meaning in the art.

41. I understand that a dependent claim contains all limitations of the claim from which it depends.

C. Obviousness

42. I understand that an obviousness analysis involves comparing a claim to the prior art to determine whether the claimed subject matter would have been obvious to a POSA in view of the prior art and general knowledge in the prior art at the time of the patent's priority date (here, June 28, 2005).

43. I understand that obviousness must be assessed from the viewpoint of a POSA at the time of the invention. I further understand that to establish obviousness, a party must perform the following factual inquiries: (a) determining the scope and content of the prior art; (b) ascertaining the differences between the

claimed invention and the prior art; and (c) resolving the level of skill in the art. I understand that, in determining the scope and content of the prior art and ascertaining the differences between the claimed invention, a patent challenger must specify where each element of the claim is found in the prior art.

44. I understand that one way of establishing obviousness is by establishing that a POSA would have had both (i) a reason to modify or combine the teachings of the prior art to achieve the claimed invention and (ii) a reasonable expectation of success in doing so. I understand that the reason to combine prior-art references can come from a variety of sources, not just the prior art itself or the specific problem the patentee was trying to solve. And I understand that the references themselves need not provide a specific hint or suggestion of the alteration needed to arrive at the claimed invention; the analysis may include recourse to logic, judgment, and common sense available to a person of ordinary skill that does not need to be explicit in any reference.

45. I understand that a “reasonable expectation of success” is assessed in view of the prior art and general knowledge in the art from the viewpoint of a POSA at the time of the patent’s priority date, and it does not require an absolute certainty of success. Furthermore, I understand that determining whether the prior art would have suggested to those of ordinary skill in the art to carry out the

claimed invention with a reasonable expectation of success must be done after determining that the claimed elements are present in the prior art.

46. I also understand that, before reaching a conclusion that the claimed invention would have been obvious, one must consider any objective evidence of non-obviousness if it is available. The objective evidence of non-obviousness can include evidence of commercial success attributable to the claimed invention, evidence of industry praise for the claimed invention, evidence of a long-felt need that was solved by the claimed subject matter, evidence that others copied the claimed subject matter, or evidence that the claimed subject matter achieved an unexpected, superior result relative to the closest prior art. I understand that such evidence must have a nexus, or causal relationship, to the claimed subject matter beyond what was available in the prior art and must be commensurate in scope with the patent claim(s) at issue.

D. Written Description

47. I understand that written description of a subject matter is assessed from the viewpoint of a POSA. I understand that to satisfy written description, the disclosures in the application must convey to a POSA that the inventors had “possession” of the claimed subject matter. I understand that the inventor need not physically possess the described subject matter to have “possession” in the context of written description. I understand that possession for written description purposes

is determined by the sufficiency of the disclosure in the application taken together, and in the context of, the state of the art at the time of the filing. I understand that the state of the art at the time of filing is determined by the general knowledge of a POSA, as evidenced by the published literature at the time.

48. I understand that the written description inquiry proceeds the same way whether one is analyzing the patent specification for determining the validity of the claims or analyzing a priority document for determining priority benefit.

VII. ARGUMENT

49. Upon considering the materials listed above and discussed herein, based on my extensive experience and education in the relevant art, in view of my understanding of the governing legal principles and the arguments presented in the Petition, I have formed the following opinions.

A. Ground 1

50. I understand that the Petition presents a priority-date challenge, and that Ground 1 is contingent on the success of that priority challenge. I understand that the Board needs to resolve priority challenges as part of the patentability determination. Specifically, Petitioner alleges that the '573 patent is not entitled to its claimed priority date because new matter, e.g., the “*instruction to not cache*” limitation was introduced in February 2019 and “has no support in any application to which the '573 Patent claims priority.” Pet., 8.

51. I do not agree. It is my opinion that the record clearly establishes that the instruction not to cache limitation has written description support throughout the priority chain. It is supported by the written description of the Provisional Application No. 61/042,412 (“’412 application”) to which the ’573 patent claims priority, as well as all the written descriptions of the patents in the priority chain of the ’573 patent.

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).

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.

53. A redirect from the popularity service without the follow-me tag is an instruction to not cache and to forward the redirect. *Id.* Alternatively, a redirect with the follow-me tag is an instruction to cache the resource and serve it to the client. *Id.* The specification thus discloses both the function of the instruction to not cache (forwarding without caching) and the form the instruction takes (a redirect without the follow-me tag).

54. Additionally, the written description of the '573 patent, the Provisional Application No. 61/042,412 ("412 application") to which the '573 patent claims priority, as well as all the written descriptions of the patents in the priority chain of the '573 patent, describe several technical reasons for not to cache unpopular resources at an edge server. For example, the '573 patent and the '412 application both disclose:

There are several reasons why a publisher/subscriber may not want the so-called "long-tail" content served from a caching CDN service, for example:

If cache fills are done from the subscriber's origin server and the subscriber pays for cache fill bandwidth, unnecessary fills for unpopular resources increase bandwidth costs without providing any value.

Serving an unpopular resource through a CDN cache adds latency (due to the cache fill), and risks forcing actually popular resources out of cache, potentially causing thrashing. The result is lower efficiency and the risk of degraded service.

For related or similar reasons, a CDN provider generally also does not want to serve long-tail content from an edge cache:

If cache fills are done from origin storage, unnecessary cache fills consume bandwidth both from the storage and cache systems, increasing (doubling) the internal cost to serve. This lowers efficiency, requiring relatively more bandwidth internally to serve the same content to the outside.

The second argument above also applies from the CDN's perspective: a CDN operator wants to minimize latency as well as the risk of thrashing in order to satisfy all of its subscribers.

EX1001, 10:17-41; EX1011, ¶¶68-69.

55. Accordingly, the written description of the '573 patent and the '412 application disclose that caching unpopular resources at an edge server "increase[s]

internal cost to serve,” “lowers efficiency,” “adds latency,” and “risks forcing actually popular resources out of cache.” *Id.* Therefore, the written description of the ’573 patent and the ’412 application not only teach two alternate instructions that lead to two corresponding actions, e.g., cache or not to cache at an edge server, but also teach not to cache unpopular resources at an edge server for various reasons. EX1001, 7:25-30, 10:17-41; EX1011, ¶¶44, 68-69.

56. The fact that the words “instruction to not cache” are not literally present in the ’573 patent and the ’412 application specification would not matter to a POSA’s understanding. The ’573 patent and the ’412 application specification present caching and not caching as alternative options and describes an embodiment that illustrates what instruction to the edge server will implement each option. EX1001, 7:25-30; EX1011, ¶44. The ’573 patent and the ’412 application specification also describe technical reasons for not to cache unpopular resources at an edge server. EX1001, 10:17-41; EX1011, ¶¶68-69. Those 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) as of the earliest priority date.

B. Ground 2

1. Overview of References

57. The three cited prior art references present three different approaches

to the issue of storing and retrieving content online. They differ from each other and from the system described and claimed in the '573 patent in both structure and function.

a. Overview of Dilley

58. Dilley describes a tiered CDN using a cache hierarchy feature.

EX1003, Abstract, 5:31-36. Dilley's tiered CDN includes edge servers, distribution hubs (or parent servers), and origin servers. *Id.*, 4:66-5:6. In this system, a client request for an object is first routed to an edge server. *Id.*, 8:3-10. If the edge server does not have the requested object locally, Dilley teaches two "request routing mechanism[s]" for the edge server to fulfill the request: (1) a default mechanism where the edge server contacts the origin server directly, or (2) serve the requested object using a parent server in the cache hierarchy. *Id.*, 6:10-15, 8:34-42.

Importantly, Dilley selects the routing mechanism based on "object-specific metadata." *Id.*, 4:23-26, 6:14-15, 8:23-48.

59. More specifically, Dilley's system uses "object-specific metadata" to handle a request for an object. *Id.*, 4:23-26. Dilley's metadata includes a "use-hierarchy" tag. *Id.*, 6:30-61. Dilley explains that the default setting of "OA" for the tag indicates that the edge server should "go direct to the (O)rigin-server specified in the (A)RL" without using the cache hierarchy feature. *Id.* Setting the tag to be "coremap" turns on the cache hierarchy feature and instructs the edge server to do

a DNS lookup for a parent server. *Id.*

60. As taught in Dilley, popular content or large files are serviced using the cache hierarchy so that the origin servers are not “overwhelmed.” *Id.*, 5:15-21, 6:4-7. Additionally, the use of this cache hierarchy “reduces average latency” and “improves connectivity” for “poorly-connected edge regions.” *Id.*, 6:7-10.

b. Overview of Pai

61. Pai discloses a storage optimization system that stores (i.e., distribute or redistribute) media content within an on-demand media system based on popular value of the media content. EX1004, ¶¶13, 20, FIGs. 6, 7. Recognizing that media content is “very expensive to store,” Pai proposes storing more popular content closer to the client. *Id.*, ¶¶2, 13. Pai also teaches dynamically redistributing media content to media source of different storage levels based on updated popularity values. *Id.*, ¶¶39- 51, FIGs. 6, 7.

c. Overview of Wang

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, 22. For instance, when a user requests content, an integrated service switch directs the request to a caching service engine, if the content is not found, the network propagates LDAP search requests up the hierarchy (e.g., from Level 1 to Level 2,

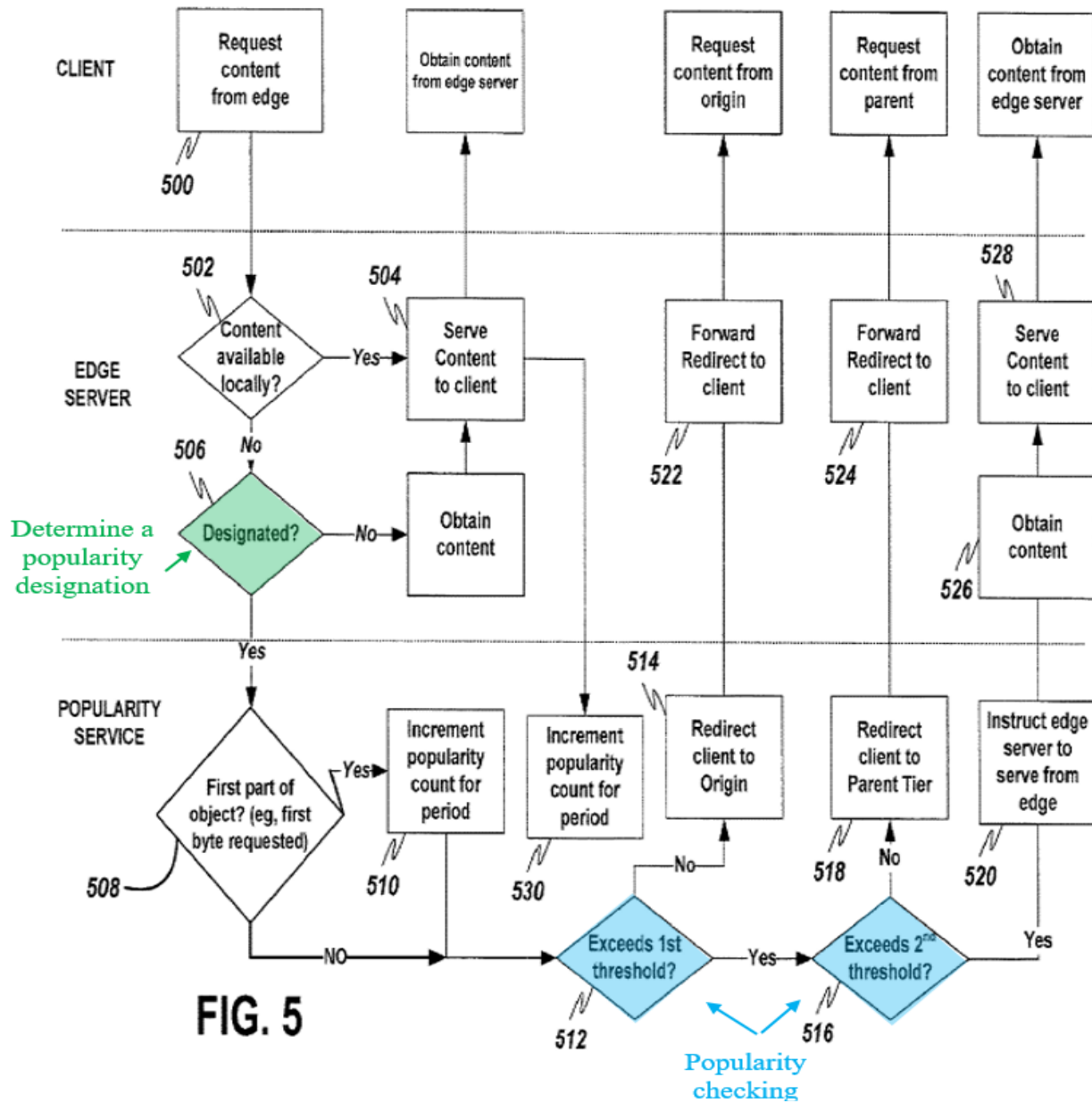
etc.). *Id.*, ¶123. If Wang’s level 3 service manager returns the origin server IP address along with an indication of cacheable or not, the caching service engine will – based on the indication - either “http-redirect” the client to the origin server, or “initiate” a new http session with the origin server on behalf of the client. *Id.*, ¶¶123-124.

2. The combination of Dilley, Pai, and Wang does not teach or suggest the feature of “determin[ing] a popularity designation.”

63. It is my opinion that the proposed combination do not teach or suggest “accessing a popularity service associated with the content delivery network to determine a popularity designation associated with the requested resource” as recited in independent claims 1, 11, and 19. Petitioner relies on *Pai alone* for this teaching but fails to point to any disclosure in *Pai* that satisfies the limitation.

64. The ’573 patent describes a specific operation. The system determines a “popularity designation” for a requested resource – i.e., whether the specific resource is even eligible for popularity-based routing. EX1001, 6:27-31 (“whether this object has been designated so that the object’s popularity will be used to determine where it will be served from”). After determining a resource is so “designated,” the system performs popularity-based routing: comparing the resource’s popularity (e.g., a “popularity count”) against specific thresholds to determine which server to fulfill the request. EX1001, 6:53-7:24, FIG. 5.

65. As the '573 patent explains, “not necessarily all” content uses the popularity service, e.g., a popularity check is performed only for “certain designated content.” EX1001, 5:47-50, 62-64. The '573 patent describes that this “determining a popularity designation” (Step 506 of Figure. 5) occurs before the “popularity checking” (Steps 512 and 516 of Figure. 5) is ever initiated. EX1001, 6:27-41; FIG. 5.



EX1001, FIG. 5 (annotated).

66. In contrast, Pai discloses a system that uses a currently known popular value of media content to distribute or redistribute the media content within an on-demand media system. EX1004, ¶13. While Pai evaluates “current popularity values” of the media content to manage storage (i.e., distributing or redistributing)

of *all* media content across media sources, Pai lacks the claimed **determining** a “popularity designation” for its media content. EX1004, ¶¶39- 51.

67. Specifically, Pai discloses a mechanism of dynamically distributing/redistributing *all* media content among multiple media sources based on known popularity. EX1004, ¶¶13, 20, FIGs. 6, 7. Pai moves or redistributes media content to a different media source based on “current popularity values.” *Id.*, ¶¶39- 51, FIGs. 6, 7. However, Pai does not perform the “designated” Step 506 to see if media content should be subject to popularity value-based logic.

68. Because Pai treats *all* media content as subject to its distribution/redistribution logic, it never “determines a popularity designation” for its media content as required by the claim. Accordingly, Pai does not teach or suggest to “determine a popularity designation,” and Dilley or Wang do not fix the deficiencies of Pai. Ground 2 therefore fails.

3. The references do not teach or suggest “requesting the resource from a second server... based at least on the popularity designation” as required by claim 19.

69. It is my opinion that the Petition fails to demonstrate that the combination of Dilley, Pai and Wang renders obvious the limitation: “*requesting the resource from a second server ... based at least on the popularity designation associated with the requested resource,*” as recited in claim 19.

70. First, Petitioner relies exclusively on the testimony of Dr. Mowry to

fill critical gaps between the prior art's teachings and the "*popularity designation*" limitation. Pet., 83-84 (citing EX1006, ¶¶352-54). However, as I explain in detail below, none of the references discloses this feature.

71. Second, in my opinion, a POSA would not have been motivated to combine these references because their respective teachings are technically incompatible. As discussed below, the proposed combination would fundamentally alter the principle of operation of the references, further demonstrating that there was no reasonable rationale to combine them.

a. Petitioner relies on conclusory testimony to fill gaps and conflates internal distribution with external delivery.

72. It is my opinion that Petitioner relies exclusively on the conclusory testimony of Dr. Mowry to allege that the combination of Dilley and Pai teaches requesting content from the second server based on a "*popularity designation*." However, neither reference, alone nor in combination, discloses this claim feature.

73. First, as established in Section VII.B.1.a, Dilley determines *a second server* (a parent server) based on object-specific metadata, not based on a popularity designation. Dilley's object-specific metadata includes a "use-hierarchy" tag, where specific values identify different destination, e.g., "OA" indicates the second server being an "(O)rigin-server specified in the (A)RL" and "coremap" indicate a hostname or IP address of a parent server (i.e., a second

server). EX1003, 6:30-61, 7:14-17. Thus, by evaluating the metadata associated with a requested object, Dilley's edge server "determines the name of the server region through which the request is funneled to fetch or validate the object being requested." *Id.*, 7:14-17.

74. Petitioner relies only on Dr. Mowry's unsubstantiated testimony to assert that "[t]hus, based on an unpopular designation, Dilley would retrieve content from the second server." Pet., 84 (citing EX1006, ¶¶352-53). However, Dilley's object-specific metadata describes control information (e.g., user-hierarchy tag) and server identities, which are fundamentally distinct from the recited "*popularity designation*." Thus, Dilley cannot teach "*requesting the resource from a second server... based at least on the popularity designation*."

75. Second, Pai is also deficient. Pai discloses a mechanism for **distributing or storing** media content among servers within a CDN based on popularity values. EX1004, ¶¶13, 20, 39, FIGs. 6, 7. In other words, Pai's distribution of media content is merely storing content **within the CDN**. However, the claimed feature requires *requesting the resource* from a specific *second server* for delivery to end-users **outside the CDN**.

76. Additionally, as established in Section VII.B.2 above, Pai does not teach the claimed popularity designation. Petitioner again exclusively relies on the conclusory testimony of Dr. Mowry to allege 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[.]” Pet., 84 (citing EX1006, ¶354). However, Dr. Mowry conflates Pai’s popularity value with the claimed popularity designation. More importantly, nowhere in Pai teaches forwarding the request to a second server based on popularity threshold. Indeed, Pai’s internal storage optimization does not teach selecting a delivery server *based at least on the popularity designation*.

77. Therefore, neither Dilley nor Pai teaches “*requesting the resource from a second server ... based at least on the popularity designation associated with the requested resource*” as recited in claim 19. Wang does not cure the deficiency of Dilley and Pai.

b. The Petition arguments are based on impermissible hindsight.

78. First, it is my opinion that Petitioner’s proposed combination of Dilley and Pai is a product of impermissible hindsight. Petitioner improperly conflates **internal distribution** (a background storage logic that stores content closer to end-users) with **external delivery** (real-time requesting and serving content to an end-user). This fundamental mismatch is exacerbated by Petitioner and Dr. Mowry’s failure to distinguish between Pai’s numerical “popularity value” with the claimed *popularity designation*. See Section VII.B.2, ¶¶63-65. There is no suggestion in either reference that a background storage metric could or should be used as a real-

time routing trigger. Petitioner’s jump from “storing based on popularity values” to “requesting based on popularity designation” is a leap found only in the challenged claims, not the prior art.

79. Second, Petitioner effectively admits that the combination only teaches storing popular content at Dilley’s edge “closer to the requesting end-users.” Pet., 84. The storage-centric logic remains structurally and functionally distinct from the claim requirement of *requesting the resource from a second server* where that specific second server is selected *based at least on the popularity designation*.

c. Petitioner and Dr. Mowry’s motivation to combine Dilley and Pai is based on mischaracterization of Dilley.

80. In my opinion, Petitioner and Dr. Mowry’s motivation to combine Dilley and Pai rests on a fundamental mischaracterization of Dilley. Petitioner erroneously alleges that Dilley determines where to “cache content” as a function of metadata associated with the requested object. Pet., 37. Next, Petitioner conflates Pai’s popularity value with the claimed *popularity designation*. Section VII.B.2., ¶¶63-65. Building on these errors, Petitioner contends that Dilley’s “object specific metadata” can be used to facilitate control over “a tiered distribution of content” using Pai’s popularity-based method. *Id.*

81. As I have already established in VII.B.1.a, Dilley’s “object specific

metadata” is used to select which server to fetch content, not to determine which server to “cache content.” EX1003, 4:23-26, 6:14-15, 8:23-48. Accordingly, Pai’s popularity-based content storage method is incompatible with Dilley’s teaching of fetching content using object specific metadata. In my view, Petitioner improperly conflates Pai’s popularity value with the claimed popularity designation, and Pai’s internal storage/distribution of content with Dilley’s external delivery/routing of content.

82. Therefore, it is my opinion that the combination of Dilley and Pai does not teach “*requesting the resource from a second server ... based at least on the popularity designation associated with the requested resource.*”

d. A POSA would not have been motivated to combine Dilley and Pai because their teachings are incompatible, rendering one or both inoperable for their intended purpose or changing their principle of operation.

83. In my opinion, Petitioner intentionally conflates Dilley’s edge servers with its distribution hubs or parent servers, and a POSA would not have been motivated to combine Dilley and Pai..

84. Specifically, Petitioner incorrectly alleges that “[b]ecause the edge servers are used to prevent the origin servers from becoming ‘overwhelmed with requests for popular content,’” unpopular content would be served from parent servers while “popular content would be served from the edge.” Pet., 83-84 (citing

EX1003, 5:10-21). This is a mischaracterization of Dilley

85. In reality, the very passage Petitioner cites clarifies that Dilley’s “**distribution hubs**” or parent servers – which are distinct from edge servers – act as buffers to guarantee that the origin site is not “overwhelmed with requests for popular content or large files.” EX1003, 5:10-21. Dilley explicitly distinguishes its “edge servers 302a-n” from “distribution hubs 304a-n,” teaching that its tiered infrastructure uses “a set of edge servers and a hierarchical set of ‘parent’ or hub regions that service the edge servers.” *Id.*, 4:63-5:7, Figure 3. On the other hand, Pai teaches that level one media sources store the most popular content while level two media sources (the second server) store less popular content. EX1004, ¶¶38-45.

86. Therefore, Pai’s system stores less popular content in a second server, while Dilley stores popular content in a second server (e.g., hubs or parent servers). It is my conclusion that this technical incompatibility would render one or both references inoperable for their intended purpose and/or change their principle of operation, which demonstrates the lack of any reasonable motivation for a POSA to combine these references.

4. The references fail to render obvious the claimed “redirect instruction” that is sent from the second server of the content delivery network to the first server (Claims 1(d)/11(d)).

87. In my opinion, the proposed combination of Dilley and Wang fails to

render obvious the challenged independent claims 1 and 11 at least because the combination fails to teach or suggest “*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,*” as recited in independent claim 1, or “*transmitting a redirect instruction intended for the requesting device to the edge server to obtain the resource from a content server,*” as recited in independent claim 11.

88. Petitioner and Dr. Mowry improperly assume that Dilley’s second server (“a peer server or a parent server”) would “instruct the first server to request the content from a server higher on the hierarchy, such as an origin server.” Pet., 50; EX1006, ¶251. Dr. Mowry further alleged that “[w]hile Dilley discloses that the content can be retrieved from the origin server ... it does not provide specifics on how that retrieval happens.” EX1006, ¶252. Based on those false premises, Petitioner concludes that a POSA would have understood that a parent server would send a “redirect instruction” to an edge server to obtain the resource from the origin server and would thus have been motivated to implement such a redirect instruction from Wang. Pet., 50 (citing EX1006, ¶¶253-54).

89. Furthermore, Petitioner and Dr. Mowry argue that combining Dilley Pai and Wang would “decrease latency” by removing the parent server as a “middleman,” supposedly allowing “the edge server to communicate with and

receive content directly from the origin server.” Pet., 51; EX1006, ¶¶255-57.

90. However, Petitioner and Dr. Mowry’s understanding of Dilley is technically misplaced. Dilley’s established retrieval path for requested objects already functions as an efficient, automated fetch process. The proposed combination would replace an automated hierarchical fetch with a multi-step redirect process, which offers no actual latency benefit and would likely increase overhead. In my view, a POSA would not have been motivated to combine Dilley and Wang to implement the claimed “*redirect instruction*” because doing so provides no technical advantage over Dilley’s existing process.

- a. Petitioner and Dr. Mowry erroneously allege that Dilley’s second server instructs the edge server to retrieve content from the origin server, or that Dilley does not provide specifics on how that retrieval happens.**

91. In my opinion, Petitioner and Dr. Mowry erroneously allege that Dilley’s second server “instruct[s] the first server” to request content from an origin server if the content is not available at the first and second servers. EX1006, ¶251 (citing EX1003, 7:31-40, 8:53-58). To the contrary, Dilley explicitly teaches that it is the second server that requests the content from the origin server. Dilley states that the origin server is contacted “when a server in a parent region needs to fetch content” and that the “parent serves the request... by making a request to the origin server.” EX1003, 7:31-40, 8:53-58. There is no disclosure or suggestion in

Dilley that a second server provides a redirect instruction to the first server to facilitate a direct request to the origin server. *Id.*

92. Additionally, Petitioner and Dr. Mowry's argument rests on the incorrect premise that Dilley "does not provide specifics" on how content is retrieved from the origin server. EX1006, ¶252. As I will show in more detail in Section VII.B.4.b below, Dilley provides a clear, specific disclosure of how to retrieve content from an origin server: Dilley's edge server evaluates **object-specific metadata** to determine whether to retrieve content from an origin server directly, or through a parent server, who fetches the object from the origin server. By ignoring this established mechanism, Petitioner incorrectly argues that there exists a gap in Dilley and introduces Wang's redirect mechanism to address it. Because Dilley already teaches a complete retrieval path, a POSA would have seen no reason, and had no motivation, to implement a redirect instruction to facilitate a process that Dilley's existing hierarchy already performs.

93. Thus, it is my opinion that Petitioner and Dr. Mowry's conclusion that "a POSITA would have been motivated to implement this redirect instruction from Wang" is incorrect. Pet., 50.

b. Dilley's edge server, not the second server, determines where the edge server should fetch content based on the metadata.

94. In my opinion, contrary to Petitioner and Dr. Mowry's allegations,

Dilley provides clear, specific disclosure of how to retrieve content from the origin server by using metadata associated with the content, and Dilley does not disclose or even suggest that the second server would instruct the first server to request content from the origin server because Dilley's system uses metadata associated with the content for routing decisions. EX1003, 7:14-18.

95. Dilley discloses a tiered content distribution system using a cache hierarchy feature. EX1003, 5:31-36. Dilley explains, only "a relatively small, specific subset of objects" are serviced using the hierarchy, while most objects are fetched by the edge server contacting the origin server. *Id.*, 6:10-14. More specifically, Dilley's system uses "object-specific metadata" to handle a request for an object. *Id.*, 4:23-26. The object-specific metadata includes a set of "all control options and parameters for the object." *Id.*, 4:27-33. For example, Dilley's metadata includes a "use-hierarchy" tag. *Id.*, 6:30-61. Dilley explains that the default setting of "OA" for the tag indicates that the edge server should "go direct to the (O)origin-server specified in the (A)RL" without using the cache hierarchy feature. *Id.* Setting the tag to be "coremap" turns on the cache hierarchy feature and instructs the edge server to do a DNS lookup for a parent server. *Id.*

96. Additionally, Dilley discloses a "request routing mechanism" that is implemented as a software "running in each edge server" and includes "a guide process," and the guide process instructs the edge server where to get the content.

EX1003, 7:46-8:2, 8:22-42. The edge server (“*the first server*”) determines whether it should fetch an object directly from the origin server or from a parent server by “quer[ying] its associated guide process.” *Id.*, 8:22-38, FIG. 6. The “guide process” of the edge server “examine[s] object metadata for a given tag” and makes the determination to request the object from a parent server (if the object is a candidate to use the cache hierarchy feature), or to request the object from the origin server (if the object is not a candidate for distribution through the hierarchy). *Id.*, *see also id.*, 7:14-18 (“application of metadata determines the name of the server region through which the request is funneled to fetch or validate the object being requested. This is the origin server by default, but it can also be a cache hierarchy parent[.]”).

97. Dilley further teaches that if the guide process instructs the edge server to request the object from a parent server, and that parent server does not have the object, a guide process of the parent server instructs the parent server to “make[] a request to the origin server.” *Id.*, 8:22-38, 8:53-58, FIG. 6. In other words, the parent server fetches the object from the origin server, instead of instructing the edge server to do so.

98. Therefore, it is my opinion that Dilley provides a clear, specific disclosure of how to retrieve content from the origin server based on metadata associated with the content, e.g., (1) contacting the origin server directly if the

requested object is not a candidate for using the cache hierarchy feature, or (2) contacting a parent server in the cache hierarchy and the parent server fetches the content from the origin server if it does not have it. *Id.*, 8:22-38, 8:53-58. Dilley does not even suggest or imply that the parent server would instruct the edge server to retrieve the content directly from the origin server. Accordingly, a POSA would not have been looking to other references, such as Wang, for a teaching of the claimed “*redirect instruction*” to facilitate a process already fully addressed by Dilley’s internal architecture.

c. A POSA would not have been motivated to implement Wang's redirect procedure in Dilley's system because there are no benefits and no expectation of success.

99. The Petitioner’s proposed combination of Dilley and Wang is a result of hindsight-driven reconstruction that ignores the technical realities and express objectives of the prior art. Because the combination results in a redundant architecture with no reasonable expectation of success, it would not have been obvious to a POSA at the time of the invention.

i. Petitioner and Dr. Mowry’s theory of motivation relies on improper hindsight reconstruction.

100. Petitioner and Dr. Mowry contend that removing Dilley’s parent server as a “middleman” would “decrease latency.” Pet., 50. In my opinion, Petitioner and Dr. Mowry’s theory of removing a middleman reduces latency is a simplistic generalization that fails to account for the specific technical context of

the prior art. Pet., 51; EX1006, ¶¶255-56. Dilley teaches that its “middleman” parent server in the cache hierarchy “maintains a persistent connection to the origin server,” which “reduces average latency and improves connectivity” for a “poorly-connected” edge server. EX1003, 5:10-12, 6:7-11.

101. Because Petitioner and Dr. Mowry offer no contemporaneous evidence that a POSA would have disregarded Dilley’s express stability benefits in favor of Wang’s redirect, the proposed combination is an improper hindsight reconstruction. In my view, a POSA would have seen that the removal of the parent server (the “middleman”) is not an optimization, but a degradation of Dilley’s fundamental “cache hierarchy” principle of operation.

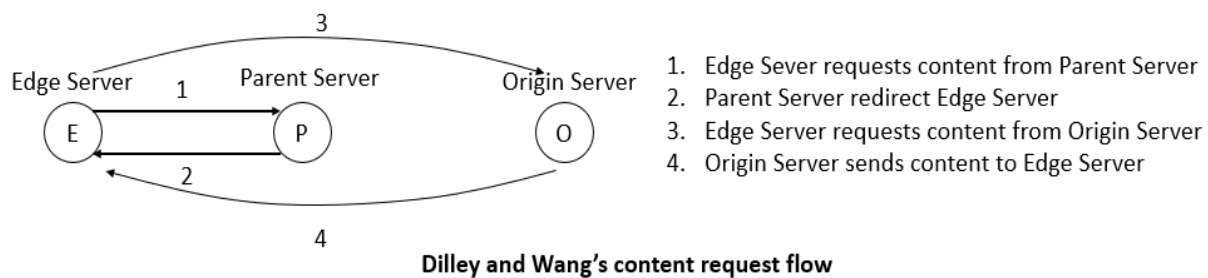
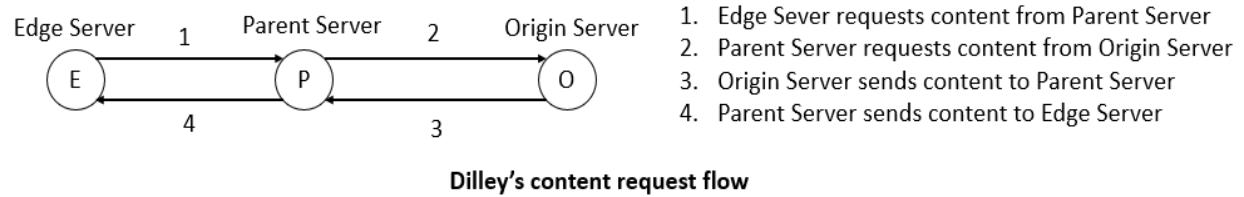
ii. Petitioner’s alleged benefit of reducing latency or improving server availability are technically illusory.

102. In my opinion, Petitioner’s alleged benefits of decreased latency or improved server availability are technically illusory and contrary to the express teachings of the prior art.

103. First, Petitioner and Dr. Mowry contend that bypassing Dilley’s parent server to request objects directly from the origin would necessarily “decrease latency,” Pet., 50; EX1006, ¶¶255-57. In my opinion, this assertion is technically unsound.

104. As illustrated in the diagrams below, the proposed combination of

Dilley and Wang requires **four discrete steps** to retrieve content from the origin server, which is **identical** to Dilley's standalone system.



105. A POSA would have understood that removing Dilley's parent server as a "middleman" does not inherently decrease latency. In fact, as I further demonstrate below, integrating Wang's redirect procedure into Dilley's cache hierarchy can actually **increase** latency.

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.

107. Second, Petitioner and Dr. Mowry contend that another benefit of combining Wang and Dilley would be improving parent server availability because this allows “the parent server to remain available to service other requests received by the CDN.” Pet., 51; EX1006, ¶257. However, a POSA would have known that a parent server at the time of the invention was inherently capable of servicing multiple concurrent requests. Because concurrent processing was already a baseline technical reality, the “benefit” alleged by Petitioner provides no actual improvement or motivation to modify the prior art. In fact, as early as 2000, a seminal paper from UC Berkeley researchers describes “highly concurrent” server models for large-scale internet services. EX2019.

108. Furthermore, as I have previously established, Dilley’s parent server provides essential service for an edge server in a poorly connected region. A POSA would have understood that terminating such service, merely to accommodate other requests, would run counter to the express teachings and intent of Dilley. Thus, a POSA would have had no motivation to implement Wang’s redirect instruction for the sake of server availability.

109. Because the proposed combination would likely result in a **degradation of system performance** rather than an improvement, a POSA would not have expected success in achieving the purported benefits.

iii. The proposed modification results in a superfluous and redundant architecture.

110. As I discussed in Section VII.B.1.a, Dilley discloses two “request routing mechanism[s],” e.g., (1) retrieving the requested object from the origin server directly, and (2) retrieving the requested object using the cache hierarchy. EX1003, 6:10-15, 8:34-42. For the second mechanism, as established in Section VII.B.4.c.ii, there is no reasonable expectation of success for the combination to improve performance. Regarding the first mechanism, it is my opinion that a POSA would have had no motivation to implement a modification that is entirely redundant of existing functionality in the prior art.

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.

112. In my view, a POSA would not have been motivated to add unnecessary logic to the parent server to handle requests that, by definition, would have already bypassed that parent server under Dilley’s existing “guide process.” Petitioner has failed to prove a rational reason to combine these specific references.

5. The references fail to teach or suggest that the second server is a peer server that sends a redirect instruction to the first server, as recited in dependent claims 5 and 15.

113. Claim 5 depends from claim 1 and further recites that the “*second server*” (which provides the redirect instruction) is a “*peer server*” of the first tier. Petitioner only relies on Dilley in its arguments. Thus, the reference must teach or suggest a peer-to-peer redirect that instructs the first server *to obtain the resource from a content server of the content delivery network*. Claim 15 depends from claim 11 and recites similar limitations.

114. First, Petitioner relies on the same incorrect arguments presented for claim limitation 1(d) to allege that a peer server sends the redirect instruction to the first server when the request resource is unavailable. Pet., 69. The combination of Dilley and Wang fails to teach or suggest any “second server” -- a peer server or a parent server – that provides a redirect instruction to the first server. *Supra* VII.B.4.

115. Second, Petitioner’s reliance on Dilley is incorrect. Petitioner cites portions of Dilley to suggest that the second server can be a peer server or a parent server, and “[t]his second server will either serve the content from its cache or instruct the first server to request the content from a server higher on the hierarchy, such as an origin server.” Pet., 50 (citing EX1003, 7:31-40, 8:53-58). However, these citations exclusively describe actions performed by a **parent server**, not a peer server. Indeed, Petitioner’s own argument admits that “[a] POSITA would

have understood that to implement the described process, **when the second server is a parent server**, Dilley would send a redirect instruction to the edge server, directing the edge server to obtain the resource from the origin server instead.” Pet., 50 (emphasis added). Dilley actually teaches that if “no peer copy of the object is available,” the edge server queries its **own** guide process to identify where to go get the object – it does not receive a redirect from a peer server. EX1003, 8:18-21, FIG. 6.

116. Third, Petitioner’s purported rationale for combining Wang and Dilley is logically inapplicable to peer servers. Petitioner argues that the combination removes **the parent server** as a “middleman” to decrease latency and increase availability of **the parent server**. Pet., 51 (citing EX1006, ¶¶255-257). These purported benefits are irrelevant to a peer server, which, by definition, is not a “middleman” in a parent-child hierarchy.

117. Thus, it is my opinion that the prior art fails to teach that the “*second server is a peer server*” providing “*a redirect instruction*” to the first sever.

VIII. CONCLUSION

118. In signing this declaration, I recognize that the declaration will be filed as evidence in an *inter partes* review before the Patent Trial and Appeal Board of the United States Patent and Trademark Office. I also recognize that I may be subject to cross-examination in the case and that cross-examination will take place within the United States. If cross-examination is required, I will appear for cross-examination within the United States during the time allotted.

119. I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Executed on this 3rd day of March 2026, in Amherst, MA.



Dr. Prashant Shenoy