

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

APPLE INC., AMAZON WEB SERVICES, INC., AND
AMAZON.COM SERVICES LLC,
Petitioner,

v.

ZENTIAN LIMITED,
Patent Owner.

IPR2023-00037
Patent 10,971,140 B2

APPLE INC.'S NOTICE OF APPEAL

via P-TACTS
Patent Trial and Appeal Board

via Email
Director of the United States Patent and Trademark Office
c/o Office of the General Counsel
efileSO@uspto.gov

via CM/ECF
United States Court of Appeals for the Federal Circuit

INTRODUCTION

Apple Inc.’s appeal stems from the Patent Trial and Appeal Board’s Final Written Decision (Paper 34) (the “Final Written Decision”) entered on June 11, 2024, in the above-captioned *inter partes* review of United States Patent No. 10,971,140 (“the ’140 patent”). This notice is timely filed within 63 days of the Final Written Decision. 37 C.F.R. § 90.3(a)(1).

APPLE INC.’S APPEAL

Please take notice that under 35 U.S.C. §§ 141(c), 142, 319; 37 C.F.R. §§ 90.2(a), 90.3(a); and Federal Rule of Appellate Procedure/Federal Circuit Rule 15, Apple Inc. hereby appeals to the United States Court of Appeals for the Federal Circuit from the Final Written Decision and all underlying decisions and orders in this action on which that decision is based.

APPLE INC.’S ISSUES ON APPEAL

Apple Inc.’s issues on appeal include at least (37 C.F.R. § 90.2(a)(3)(ii)):

(i) the Board’s conclusion that claims 1-8 of the ’140 patent had not been proven unpatentable; and

(ii) any findings, analysis, or determinations supporting or related to that issue, including on motivation to combine, as well as all other issues decided adversely to Apple Inc. in any orders, decisions, rulings, or opinions in this proceeding.

Simultaneously with this submission, Apple Inc. is filing a true and correct

copy of this Notice of Appeal with the Director of the United States Patent and Trademark Office and a true and correct copy (or copies) of the same, along with the required filing fee, with the Clerk of the United States Court of Appeals for the Federal Circuit as set forth in the accompanying Certificate of Filing.

Dated: August 12, 2024

Respectfully submitted,

By: /Jennifer C. Bailey/
Jennifer C. Bailey, Reg. No. 52,583
ERISE IP, P.A.
7015 College Blvd., Suite 700
Overland Park, KS 66211
P: (913) 777-5600
F: (913) 777-5601
Jennifer.Bailey@eriseip.com

Attorney for Apple Inc.

CERTIFICATE OF FILING

The undersigned hereby certifies that, in addition to being electronically filed through P-TACTS, a true and correct copy of the above-captioned APPLE INC.'S NOTICE OF APPEAL is being filed via email with the Director on August 12, 2024, at efileSO@uspto.gov, the email address indicated on the United States Patent and Trademark Office's web page for the Office of the General Counsel.

The undersigned also hereby certifies that a true and correct copy of the above-captioned APPLE INC.'S NOTICE OF APPEAL and the filing fee is being filed via CM/ECF with the Clerk's Office of the United States Court of Appeals for the Federal Circuit on August 12, 2024.

Dated: August 12, 2024

Respectfully submitted,

By: Jennifer C. Bailey
Jennifer C. Bailey, Reg. No. 52,583
ERISE IP, P.A.
7015 College Blvd., Suite 700
Overland Park, KS 66211
P: (913) 777-5600
F: (913) 777-5601
Jennifer.Bailey@eriseip.com

Attorney for Apple Inc.

CERTIFICATION OF SERVICE

The undersigned hereby certifies that on August 12, 2024, the foregoing
APPLE INC.'S NOTICE OF APPEAL was served in its entirety via electronic mail
on the following:

Peter Knops
Kayvan Noroozi
NOROOZI PC
peter@noroozipc.com
kayvan@noroozipc.com

Katherine Rhoades
BARLIT BECK LLP
katherine.rhoades@bartlitbeck.com

Dated: August 12, 2024

Respectfully submitted,

By: /Jennifer C. Bailey/
Jennifer C. Bailey, Reg. No. 52,583
ERISE IP, P.A.
7015 College Blvd., Suite 700
Overland Park, KS 66211
P: (913) 777-5600
F: (913) 777-5601
Jennifer.Bailey@eriseip.com

Attorney for Apple Inc.

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

APPLE INC., AMAZON WEB SERVICES, INC., and
AMAZON.COM SERVICES LLC

Petitioner,

v.

ZENTIAN LIMITED,
Patent Owner.

IPR2023-00037¹
Patent 10,971,140 B2

Before KEVIN F. TURNER, JEFFREY S. SMITH, and
CHRISTOPHER L. OGDEN, *Administrative Patent Judges*.

Opinion for the Board filed by *Administrative Patent Judge* SMITH.

Opinion Concurring filed by *Administrative Patent Judge* OGDEN

SMITH, *Administrative Patent Judge*.

JUDGMENT
Final Written Decision
Determining No Challenged Claims Unpatentable
35 U.S.C. § 318(a)

¹ IPR2023-01197 has been joined with this proceeding.

I. INTRODUCTION

A. *Background and Summary*

Petitioner filed a Petition (Paper 1, “Pet.”) requesting *inter partes* review of claims 1–8 of U.S. Patent No. 10,971,140 B2 (Ex. 1001, “the ’140 patent”). We issued an Institution Decision (Paper 10, “Dec.”) instituting the petitioned review. Patent Owner then filed a Patent Owner Response (Paper 19, “PO Resp.”) to the Petition. Petitioner filed a Reply (Paper 22, “Reply”) to the Patent Owner Response. Patent Owner filed a Sur-reply (Paper 27, “PO Sur-Reply”) to the Reply. An oral hearing was held on March 11, 2024, for which the transcript was entered into the record (Paper 33).

We have jurisdiction under 35 U.S.C. § 6(b)(4) and § 318(a). This Decision is a final written decision under 35 U.S.C. § 318(a) and 37 C.F.R. § 42.73 as to the patentability of claims 1–8 of the ’140 patent. We determine Petitioner has not shown by a preponderance of evidence that claims 1–8 are unpatentable.

B. *Related Matters*

The parties indicate that the following matters relate to the ’140 patent: *Zentian Ltd v. Apple Inc.*, 6:22-cv-00122 (W.D. Tex. Feb. 2, 2022); *Zentian Ltd v. Amazon.com, Inc.*, 6:22-cv-00123 (W.D. Tex. Feb. 2, 2022); *Apple Inc. v. Zentian Ltd.*, *Inter Partes* Review No. IPR2023-00033; *Apple Inc. v. Zentian Ltd.*, *Inter Partes* Review No. IPR2023-00034; *Apple Inc. v. Zentian Ltd.*, *Inter Partes* Review No. IPR2023-00035; and *Apple Inc. v. Zentian Ltd.*, *Inter Partes* Review No. IPR2023-00036. Paper 4, 1; Pet. 64.

C. The '140 Patent

The '140 patent is related to a speech recognition circuit which uses parallel processors for processing the input speech data in parallel.

Ex. 1001, 1:18–20.

The patent describes that in speech recognition, there are generally two processes: “front end processing to generate processed speech parameters such as feature vectors, followed by a search process which attempts to find the most likely set of words spoken from a given vocabulary (lexicon).” *Id.* at 1:21–26. According to the '140 patent, “for large vocabulary, speaker independent speech recognition, it is the search process that presents the biggest challenge.” *Id.* at 1:28–30.

The '140 patent describes that in order to speed up the search function, parallel processing techniques have been suggested. *Id.* at 1:45–47. The patent further describes that “one algorithm for performing the search is the Viterbi algorithm,” which “is a parallel or breadth first search through a transition network of states of Hidden Markov Models.” *Id.* at 1:36–39. This search algorithm is computationally intensive. *Id.* at 1:44. In one paper cited by the '140 patent, “a multi-threaded implementation of a fast beam search algorithm is disclosed.” *Id.* at 1:47–52. This “multi-threading implementation requires a significant amount of communication and synchronization among threads.” *Id.* at 1:52–54. In another cited paper, “the parallel processing of input speech parameters is disclosed in which a lexical network is split statically among processors.” *Id.* at 1:54–58.

To implement parallel processing of the search function, the '140 patent describes a special circuit, in which a “plurality of lexical tree processors are connected in parallel to the input port and perform parallel lexical tree processing for word recognition by accessing the lexical data in

the lexical memory arrangement.” *Id.* at 2:4–8. In addition, a “controller controls the lexical tree processors to process lexical trees identified in the results memory arrangement by performing parallel processing of a plurality of lexical tree data structures.” *Id.* at 2:12–15.

Figure 2 is a diagram of the circuit of the ’140 patent, and is reproduced below.

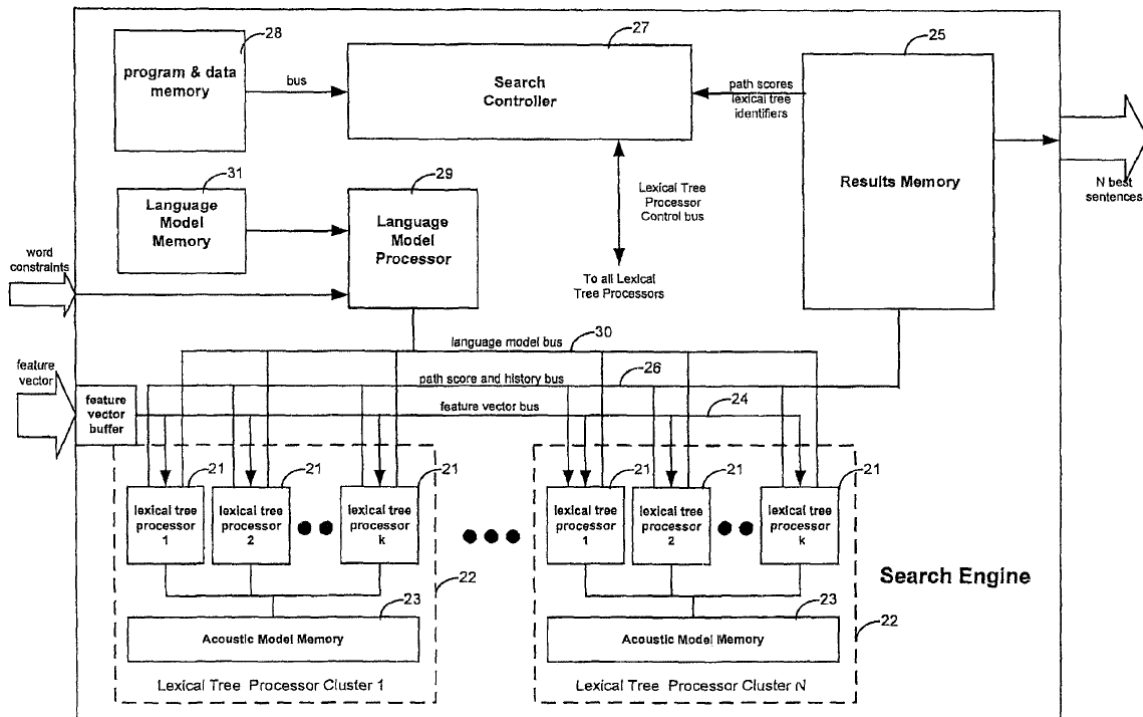


Figure 2, showing a plurality k of lexical tree processors 21, arranged in a lexical tree processor cluster 22, with acoustic model memory 23.

D. Illustrative Claim

Challenged claim 1 of the ’140 patent recites:

1. [Pre] A speech recognition circuit comprising:
 - [a] one or more clusters of processors, each of the one or more clusters of processors comprising:
 - a plurality of processors; and
 - [b] an acoustic model memory storing acoustic model data, [c] wherein each of the plurality of processors is configured to compute a probability using the

acoustic model data in the acoustic model memory,
[d] wherein:

the speech recognition circuit is configured to
generate an initial score for an audio sample;
and

[e] the initial score is used to determine whether to
continue processing to determine a final
score via processing a larger amount of
model data than that was processed to
generate the initial score.

Ex. 1001, 12:13–26; Pet. 66–67 (showing Petitioner’s bracketed claim
annotations).

E. Evidence

Petitioner relies on the following prior art:

U.S. Patent No. 6,374,219 B1, issued April 16, 2002 (Ex. 1004,
“Jiang”);

U.S. Patent No. 5,428,803, issued June 27, 1995 (Ex. 1005, “Chen”);

U.S. Patent Appl. Publ. No. 2001/0053974 A1, published December
20, 2001 (Ex. 1008, “Lucke”);

U.S. Patent No. 5,983,180, issued November 9, 1999 (Ex. 1009,
“Robinson”);

U.S. Patent No. 5,036,539, issued July 30, 1991 (Ex. 1010,
“Wrench”).

F. Prior Art and Asserted Grounds

Petitioner asserts that claims 1–8 of the ’140 patent are unpatentable
on the following grounds:

Claim(s) Challenged	35 U.S.C. §	Reference(s)/Basis
1–3, 5, 7, 8	103(a)	Jiang, Chen
1–3, 5, 7, 8	103(a)	Jiang, Chen, Lucke
4	103(a)	Jiang, Chen, Robinson

Claim(s) Challenged	35 U.S.C. §	Reference(s)/Basis
4	103(a)	Jiang, Chen, Lucke, Robinson
6	103(a)	Jiang, Chen, Wrench
6	103(a)	Jiang, Chen, Lucke, Wrench

Pet. 6. Petitioner relies on the Declaration testimony of Christopher Schmandt. Ex. 1003.

II. ANALYSIS

A. Legal Standards

“In an [*inter partes* review], the petitioner has the burden from the onset to show with particularity why the patent it challenges is unpatentable.” *Harmonic Inc. v. Avid Tech., Inc.*, 815 F.3d 1356, 1363 (Fed. Cir. 2016) (citing 35 U.S.C. § 312(a)(3) (requiring *inter partes* review petitions to identify “with particularity . . . the evidence that supports the grounds for the challenge to each claim”)); *see also* 37 C.F.R. § 42.104(b) (requiring a petition for *inter partes* review to identify how the challenged claim is to be construed and where each element of the claim is found in the prior art patents or printed publications relied upon).

A claim is unpatentable under 35 U.S.C. § 103(a) if “the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains.” *KSR Int’l Co. v. Teleflex Inc.*, 550 U.S. 398, 406 (2007). The question of obviousness is resolved on the basis of underlying factual determinations, including: (1) the scope and content of the prior art; (2) any differences between the claimed subject matter and the prior art; (3) the level of skill in the art; and (4) when in evidence, objective evidence of obviousness or nonobviousness, i.e., secondary considerations. *See*

Graham v. John Deere Co., 383 U.S. 1, 17–18 (1966). An obviousness analysis “need not seek out precise teachings directed to the specific subject matter of the challenged claim, for a court can take account of the inferences and creative steps that a person of ordinary skill in the art would employ.” *KSR*, 550 U.S. at 418.

Additionally, the obviousness inquiry typically requires an analysis of “whether there was an apparent reason to combine the known elements in the fashion claimed by the patent at issue.” *KSR*, 550 U.S. at 418 (citing *In re Kahn*, 441 F.3d 977, 988 (Fed. Cir. 2016) (requiring the fact-finder to provide “some articulated reasoning with some rational underpinning to support the legal conclusion of obviousness”)). Furthermore, Petitioner does not satisfy its burden of proving obviousness by employing “mere conclusory statements,” but “must instead articulate specific reasoning, based on evidence of record, to support the legal conclusion of obviousness.” *In re Magnum Oil Tools Int’l, Ltd.*, 829 F.3d 1364, 1380 (Fed. Cir. 2016).

B. Claim Construction

We construe the challenged claims

using the same claim construction standard that would be used to construe the claim in a civil action under 35 U.S.C. 282(b), including construing the claim in accordance with the ordinary and customary meaning of such claim as understood by one of ordinary skill in the art and the prosecution history pertaining to the patent.

37 C.F.R. § 42.100(b) (2022). “In determining the meaning of [a] disputed claim limitation, we look principally to the intrinsic evidence of record, examining the claim language itself, the written description, and the prosecution history, if in evidence.” *DePuy Spine, Inc. v. Medtronic*

Sofamor Danek, Inc., 469 F.3d 1005, 1014 (Fed. Cir. 2006). Claim terms are given their ordinary and customary meaning as would be understood by a person of ordinary skill in the art at the time of the invention and in the context of the entire patent disclosure. *Phillips v. AWH Corp.*, 415 F.3d 1303, 1313 (Fed. Cir. 2005) (en banc). “There are only two exceptions to this general rule: 1) when a patentee sets out a definition and acts as his own lexicographer, or 2) when the patentee disavows the full scope of a claim term either in the specification or during prosecution.” *Thorner v. Sony Comput. Entm’t Am. LLC*, 669 F.3d 1362, 1365 (Fed. Cir. 2012).

Petitioner “applies the plain and ordinary meaning of all claim terms [as understood by a person of ordinary skill in the art for all terms].” Pet. 7. Patent Owner does not offer an explicit claim construction. *See generally* PO Resp. Although neither party explicitly construes the claims, the parties implicitly construe the claims in their respective arguments over level of ordinary skill and patentability. We agree with Petitioner that the proper standard for construing the claims is the plain and ordinary meaning standard. To the extent necessary, we resolve the implicit claim construction disputes by applying the plain and ordinary meaning to the claims, as discussed below.

C. Level of Ordinary Skill in the Art

Petitioner contends that a person having ordinary skill in the art would have had “a master’s degree in computer engineering, computer science, electrical engineering, or a related field, with at least two years of experience in the field of speech recognition, or a bachelor’s degree in the same fields with at least four years of experience in the field of speech recognition,” and that “[a]dditional education or experience might substitute for the above requirements.” Pet. 5 (citing Ex. 1003 ¶¶ 24–26).

Patent Owner contends that the field of electrical engineering has many specialties, such as digital signal processing and computer architecture. PO Resp. 6 (citing Ex. 2020 ¶ 17). Patent Owner contends that a person of ordinary skill in the field of speech recognition would pursue studies specific to digital signal processing, and a person of ordinary skill in the field of computer architecture would take a different set of courses. *Id.* Patent Owner contends that a person of ordinary skill in speech recognition would not be expected to have also specialized in parallel processing architectures and methods, or in high performance computing. *Id.* at 7 (citing Ex. 2020 ¶ 18). Patent Owner contends that Petitioner's declarant, Mr. Schmandt, has not taught a course that would have equipped a masters-level student in speech recognition to apply a parallel processing architecture to known speech recognition techniques. *Id.* Patent Owner contends that although a person with a master's degree in one of the fields identified by Mr. Schmandt could be a person of ordinary skill in speech recognition, or a person of ordinary skill in high performance computing and parallel processing, that person would not be both. *Id.* (citing Ex. 2020 ¶ 19).

We do not need to resolve this dispute, because even under Petitioner's definition of the level of ordinary skill, Petitioner has not shown that the challenged claims are unpatentable. We apply Petitioner's definition that a person having ordinary skill in the art would have had a master's degree in computer engineering, computer science, electrical engineering, or a related field, with two years of experience in the field of speech recognition, or a bachelor's degree in the same fields with four years of experience in the field of speech recognition, and that additional education or experience might substitute for the above requirements.

D. Obviousness over Jiang and Chen

Petitioner contends that claims 1–3, 5, 7, and 8 would have been obvious over the combined teachings of Jiang and Chen. Pet. 9–53.

1. Overview of Jiang

Jiang, entitled “System for Using Silence in Speech Recognition,” is related to “computer speech recognition performed by conducting a prefix tree search of a silence bracketed lexicon.” Ex. 1004, code (54), 1:16–18.

Figure 2 shows a block diagram of speech recognition system 60, and is reproduced below.

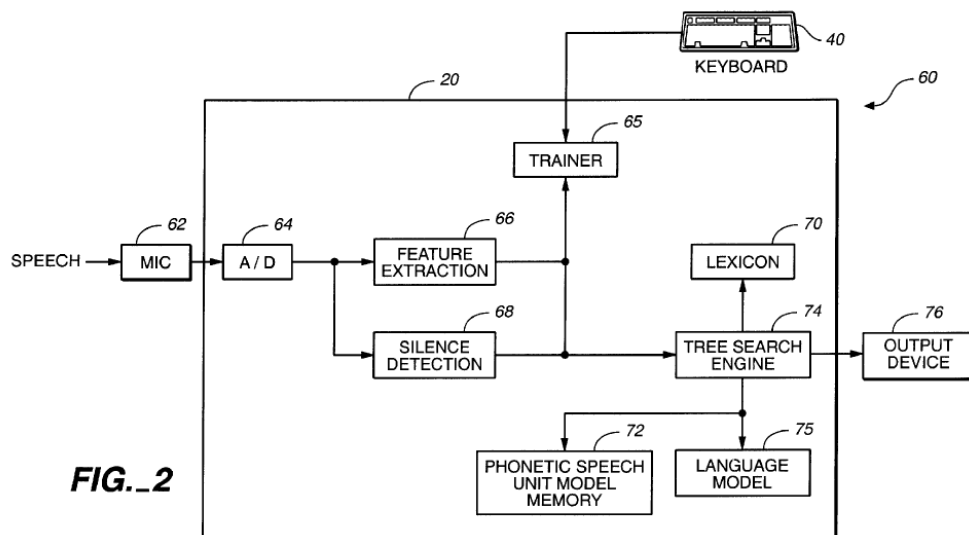


Figure 2 above shows microphone 62, analog-to-digital (A/D) converter 64, training module 65, feature extraction module 66, silence detection module 68, lexicon storage module 70, phonetic speech unit storage module 72, tree search engine 74, and output device 76. *Id.* at 6:20–25.

Jiang discloses that the speech recognition system:

recognizes speech based on an input data stream indicative of the speech. Possible words represented by the input data stream are provided as a prefix tree including a plurality of phoneme branches connected at nodes. The plurality of phoneme branches

are bracketed by at least one input silence branch corresponding to a silence phone on an input side of the prefix tree and at least one output silence branch corresponding to a silence phone on an output side of the prefix tree.

Id. at 4:7–16. Lexicon storage module 70 contains information, which is representative of all of the words in the vocabulary of speech recognition system 60. *Id.* at 7:65–67. The words are presented to tree search engine 74 in the form of a prefix tree, which can be traversed from a root to a leaf to arrive at the word most likely indicative of the utterance of the user. *Id.* at 7:67–8:4. As the tree is traversed from an input node to an output node, a score is assigned to each node connected to a phoneme branch then under consideration. *Id.* at 8:16–20.

A pruning technique can be used by comparing the score at a given node with the largest score from the other nodes of a frame being considered. *Id.* at 8:52–55. If the score at that node is sufficiently lower than the largest score, that branch is pruned from the tree, thereby drastically reducing the search space. *Id.* at 8:55–64.

2. Overview of Chen

Chen “relates generally to parallel processing computer systems for performing multiple-instruction-multiple-data (MIMD) parallel processing.” Ex. 1005, 1:29–31. Chen describes an architecture for high performance MIMD multiprocessors, which organizes the multiprocessors into four or more physically separable clusters, and provides for a shared memory model to be used with programs executed in the floating shared memory space, and a distributed memory model to be used with any programs executed across non-adjacently interconnected clusters. *Id.* at 1:32–43.

Figure 4 is a block diagram of the architecture of the Chen system, annotated by Petitioner, and is reproduced below.

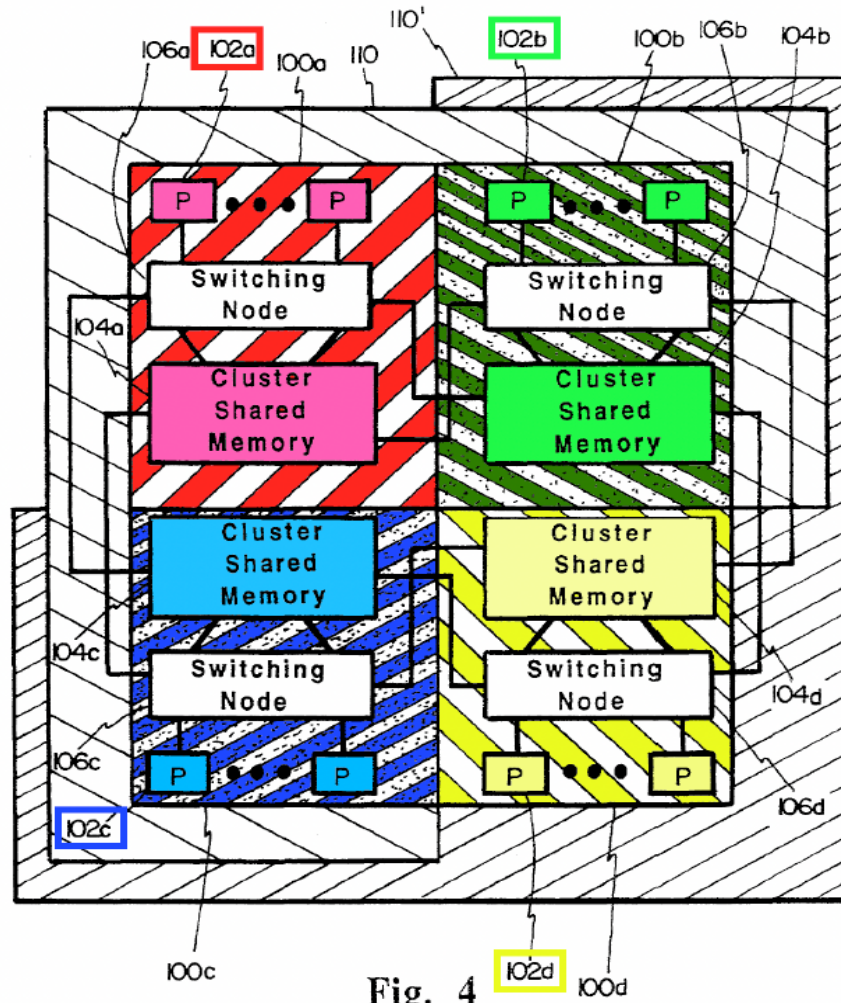


Fig. 4

Figure 4 above shows computer clusters 100a (in red), 100b (in green), 100c (in blue), and 100d (in green), processors (P) 102a–d, and memories 104a–d. Chen’s architecture provides all three of the major memory types for parallel processing:

A true shared memory model with symmetrical access to a common shared memory 104a is provided for all processors 102a in cluster 100a. An extended shared memory model is provided for all of the processors 102a, 102b and 102c that adjacently access the cluster shared memories 104a, 104b and 104c in the floating shared memory 110, for example. Finally, a distributed shared memory model is provided for all processors 102a that need to access the cluster shared memory 104d of a non-adjacently connected cluster 100d, for example.

Id. at 9:44–56.

3. *Claim 1*

Claim 1(a): “one or more clusters of processors”

Claim 1 recites “one or more clusters of processors, each of the one or more clusters of processors comprising: a plurality of processors.”

Petitioner contends that the combination of Jiang and Chen teaches this limitation. Petitioner contends that the “’140 Patent describes a ‘cluster’ as at least including processing architectures comprising a group of processors and a memory.” Pet. 14 (citing Ex. 1001, 3:14–19, 3:53–57, 5:14–21, Fig. 2).

Petitioner contends that Jiang teaches implementing its speech recognition system on multiprocessor systems, but does not specifically discuss clusters of processors. *Id.* (citing Ex. 1001, 4:51–5:3, 6:39–42). Petitioner contends that Chen teaches clusters of processors, with each cluster comprising a plurality of processors and memory. *Id.* (citing Ex. 1005, 9:5–43). In particular, Petitioner contends that Figure 4 of Chen shows four clusters, 100a, 100b, 100c, and 100d, that together comprise a parallel processing computer system. Petitioner contends that each cluster such as 100a includes two or more processors 102a that are symmetrically connected to cluster shared memory 104a via connection node 106a. Pet. 15 (citing Ex. 1004, 9:5–19, Fig. 4; Ex. 1003 ¶ 73); *see id.* at 16–17 (citing Ex. 1004, Abstract, 5:9–17, 6:17–21, 9:5–43, 10:14–35, Figs. 4, 5A, 6A, 6B). Petitioner contends that Chen teaches “each of the one or more clusters of processors comprising: a plurality of processors” as claimed in disclosing clusters 100a, 100b, 100c, and 100d, each comprising respective processors 102a, 102b, 102c, and 102d. Pet. 14.

Petitioner contends that a person of ordinary skill in the art would have replaced Jiang's tree search engine 74 with Chen's clusters of processors in order to replicate the tree search engine's functionality amongst each of Chen's processors for the benefit of processing hidden Markov models in parallel. Pet. 17–19 (citing Ex. 1003 ¶¶ 71–72, 74).

Patent Owner disputes Petitioner's reason to combine Jiang and Chen. PO Resp. 17–28. We agree with Patent Owner for the reasons given below in our analysis of reasons to combine Jiang and Chen.

Claim 1(b): “an acoustic model memory storing acoustic model data”

Claim 1 recites “each of the one or more clusters of processors comprising . . . an acoustic model memory storing acoustic model data.”

Petitioner contends that Jiang's phonetic speech unit model memory 72 stores phonetic speech unit models such as hidden Markov models that represent phonemes. Pet. 19–22 (citing Ex. 1004, Fig. 2, 6:18–28, 7:29–54). Petitioner, relying on the testimony of Mr. Schmandt, contends that a person of ordinary skill in the art would have recognized that Jiang's hidden Markov model (HMM) based phonetic model teaches the claimed acoustic model. Pet. 21 (citing Ex. 1003 ¶ 75). Petitioner contends that Jiang's phonetic speech unit model memory 72 therefore teaches “an acoustic model memory storing acoustic model data” as claimed. Pet. 22. Petitioner contends that Chen teaches “one or more clusters of processors,” where each cluster includes a cluster shared memory. Pet. 22–24 (citing Ex. 1005, Fig. 4, Abstract, 5:9–17, 6:17–21, 9:5–43, 10:14–35; Ex. 1003 ¶ 76).

Petitioner, relying on the testimony of Mr. Schmandt, contends that a person of ordinary skill in the art would have stored at least a portion of Jiang's acoustic model data in the cluster shared memory of each cluster so that each processor would have access to the necessary acoustic model data

for performing speech recognition for the benefits of reducing the time required to recognize a user's speech, enabling a more relaxed pruning threshold, and improving accuracy. Pet. 24–25 (citing Ex. 1003 ¶¶ 72, 74, 76).

Patent Owner disputes Petitioner's reason to combine Jiang and Chen. PO Resp. 17–28. We agree with Patent Owner for the reasons given below in our analysis of the reasons to combine Jiang and Chen.

Claim 1(c): “wherein each of the plurality of processors”

Claim 1 recites “wherein each of the plurality of processors is configured to compute a probability using the acoustic model data in the acoustic model memory.”

Petitioner contends that Jiang teaches a processor “configured to compute a probability using the acoustic model data” in disclosing that tree search engine 74 determines a most likely phoneme represented by a codeword based upon hidden Markov models stored in memory 72. Pet. 25–30 (citing Ex. 1004, 1:28–32, 1:33–39, 2:1–15, 7:30–45, 8:16–52; Ex. 1003 ¶¶ 77–82). Petitioner contends that Chen teaches that for each cluster of processors, the processors in each cluster may perform processing in parallel, such that each processor of each of the clusters performs the intended operation. Pet. 30 (citing Ex. 1005, 5:10–17, 18:64–19:9; Ex. 1003 ¶ 83). Petitioner contends that in the system of Jiang as modified by Chen, each of Chen's processors 102a through 102d would have computed probability scores as taught by Jiang using Jiang's acoustic model data stored in shared cluster memories 104a through 104d. *Id.*

Petitioner contends that a person of ordinary skill would have modified Jiang to include Chen's clusters of processors and would have caused each processor to perform a lexical tree search using acoustic model

data for the reasons given in Petitioner’s analysis of limitations 1(a) and 1(b). Pet. 31.

Patent Owner disputes Petitioner’s reason to combine Jiang and Chen. PO Resp. 17–28. We agree with Patent Owner for the reasons given below in our analysis of the reasons to combine Jiang and Chen.

Claim 1(d): “wherein: the speech recognition circuit is configured . . .” and claim 1(e): “the initial score is used to determine”

Claim 1 recites “wherein: the speech recognition circuit is configured to generate an initial score for an audio sample.” Claim 1 further recites “the initial score is used to determine whether to continue processing to determine a final score via processing a larger amount of model data than that was processed to generate the initial score.”

Summary of contentions regarding limitations 1(d) and 1(e)

Petitioner contends that Jiang teaches an audio sample in disclosing that a microphone outputs an acoustic signal that is processed to generate a codeword, where the codeword corresponds to the “audio sample” as claimed. Pet. 31–32 (citing Ex. 1004, 1:58–67, 2:1–15, 6:46–7:1, 7:5–15, 7:29–47, 8:16–24, 10:1–11; Ex. 1003 ¶¶ 78–82, 85, 87). Petitioner contends that Jiang teaches the claimed “speech recognition circuit is configured to generate an initial score” in disclosing a tree search engine that, in traversing prefix tree 77 shown in Figure 3, computes scores for nodes 82, 84, and 86, which are the nodes associated with the initial branches of the prefix tree leaving root node 78, such that the score for each of nodes 82, 84, and 86 is an “initial score” representing the likelihood that the phoneme represented by the first codeword under consideration corresponds to the phoneme represented by the branch associated with the node. Pet. 32–36 (citing Ex.

1004, 1:19–33, 2:1–15, 7:29–54, Ex. 1004, 8:5–51, Fig. 3; Ex. 1003 ¶¶ 84, 87).

Petitioner contends that Jiang teaches “the initial score is used to determine whether to continue processing to determine a final score via processing a larger amount of model data than was processed to generate the initial score” as claimed in disclosing a pruning technique that compares, at each node, the scores assigned to that node with the largest score on any of the other nodes corresponding to the frame being considered, and, if the score at a particular node is sufficiently low compared to the largest score, then pruning the corresponding branch from the prefix tree and no longer considering the pruned branch in further processing. Pet. 38–39 (citing Ex. 1004, 8:52–64, 10:51–11:24); *see id.* at 39–44. Petitioner, relying on testimony of Mr. Schmandt, contends that because the initial score is used to determine whether to continue processing to determine a final score or to prune the branch, the final score is determined by processing a larger amount of model data than was processed to generate the initial score. Pet. 42–44 (citing Ex. 1003 ¶¶ 90–91).

In its Response, Patent Owner contends that neither the Petition nor Mr. Schmandt’s Declaration explains how Jiang’s tree search engine modified in view of Chen would operate to meet limitations 1(d) and 1(e). PO Resp. 29–32 (citing Ex. 2017, 79:11–23, 81:8–22, 83:15–18). Patent Owner contends that Mr. Schmandt testified during cross-examination that one processor in one cluster in Chen would have been used to arrive at limitations 1(d) and 1(e). *Id.* at 32 (citing Ex. 2017, 102:5–103:3 (Mr. Schmandt testifying that after the processors finish computing their respective scores by operating on the acoustic data in parallel, one processor can look at all of the scores and perform pruning)); *see* Ex. 2017, 45:10–

46:10, 97:8–17, 98:11–99:7, 101:10–23. Patent Owner contends Mr. Schmandt is incorrect, because Chen teaches that not all the clusters are interconnected, and that processors 102a in cluster 100a cannot directly or adjacently access memory 104d of cluster 100d, while processors 102d cannot directly or adjacently access memory 104a. PO Resp. 33 (citing Ex. 1003 ¶ 95). According to Patent Owner, because the scores for the processors in each cluster are stored in that cluster’s memory, and because no one processor could access all of the scores stored in all of the memories, no one processor could perform “the initial score is used to determine whether to continue processing” the model data as claimed. *Id.* at 35 (citing Ex. 1005, 9:10–39; Ex. 2020 ¶ 51).

In Reply, Petitioner contends that the Petition states that “*Jiang’s* algorithm of generating an initial score and using the initial score to determine whether to continue processing (the functionality of Claims 1(d)–1(e)) would be *replicated* across *Chen’s* processors.” Reply 2 (citing Pet. 17–18, 24, 30; Ex. 1003, 74, 76–77, 83). Petitioner contends that the claim does not require any particular processor from the plurality of processors to perform any part of the functionality of limitations 1(d) and 1(e). *Id.* at 2–3. Petitioner contends that Mr. Schmandt confirmed during cross-examination that although paragraphs 83 through 87 of his Declaration did not explain how the combination of Jiang and Chen operated, other portions of his Declaration did discuss the combination. *Id.* at 2 (citing Ex. 2017, 44:2–11, 45:10–46:10, 71:6–23, 73:12–17, 79:11–23, 81:8–22, 83:15–18, 86:25–87:23, 93:7–94:10, 94:16–23, 95:19–96:10, 97:8–17, 99:17–102:12). Petitioner contends that in the combination of Jiang and Chen, each processor operates on a portion of the lexical trees, which does not require inter-cluster communication. *Id.* at 16. Petitioner contends that once the

scores are calculated, inter-cluster processing of the results would be communicated. *Id.* Petitioner contends that the combination of Jiang and Chen as proposed in the Petition provides the framework for inter-cluster communication of the results. *Id.* at 17–18.

In its Sur-Reply, Patent Owner contends that Petitioner’s argument in its Reply, that “Jiang’s algorithm of generating an initial score and using the initial score to determine whether to continue processing (the functionality of claims 1(d)–1(e)) would be *replicated* across Chen’s processors,” fails for several reasons. PO Sur-Reply 3 (quoting Reply 2). Patent Owner contends that Petitioner’s theory in its Reply requires every processor in Chen’s four clusters to perform limitations 1(d) and 1(e), in contrast to Mr. Schmandt’s deposition theory, which required only one processor to do so, thus abandoning Mr. Schmandt’s deposition theory and discrediting Mr. Schmandt. *Id.* Patent Owner contends that Mr. Schmandt testified at his deposition that there would be no advantage from using all of Chen’s processors to calculate the initial score and determine whether to continue processing because “[a]t that point, we have *finished* our parallel processing operation and we’re back to serial operation.” *Id.* at 3–4 (quoting Ex. 2017, 102:5–103:3).

Patent Owner contends that Petitioner’s theory would require every processor in Chen’s four clusters to calculate the same initial score and make the same determination whether to continue processing towards a final score, which makes no sense. PO Sur-Reply 3. Patent Owner contends that Petitioner’s theory in its Reply requires each processor to obtain necessary information stored across all of Chen’s four memories 104a through 104d, even though none of Chen’s processors can directly or adjacently access all cluster memories. *Id.* at 4. Patent Owner contends that Petitioner’s theory

in Reply was not set forth in the Petition, and that Petitioner's citations are to aspects of the Petition discussing other limitations, not limitations 1(d) and 1(e). *Id.*

Analysis of the contentions regarding limitations 1(d) and 1(e)

We agree with Patent Owner that the Petition's combination of Jiang and Chen requires implementing Jiang's algorithm on a plurality of clusters of processors as shown in Figure 4 of Chen. PO Sur-Reply 27; Ex. 2020 ¶ 44; *see* Pet. 15–17, 19–24; Ex. 1003 ¶¶ 73–76, 83. We highlight that Chen discloses that its system includes at least four clusters of processors. Ex. 1005, 9:13–14 (“Four clusters 100a, 100b, 100c and 100d together comprise a parallel processing computer system.”); *see id.* at 11:51–53 (“In order to provide a common shared memory, there should be four or more groups of processors.”).

Even though Petitioner's combination requires implementing Jiang's algorithm on Chen's plurality of clusters of processors, the Petition contends that Jiang alone teaches limitations 1(d) and 1(e) in disclosing assigning an initial score to each of the initial nodes of a prefix tree, then using these initial scores to determine, via pruning, whether to continue processing branches connected to the nodes. Pet. 38 (citing Ex. 1004, 8–64); *see id.* at 31–44. We agree with Patent Owner that the Petition does not explain how Jiang's pruning would be implemented in Chen's multiple clusters of processors to generate an initial score for an audio sample, and to use the initial score “to determine whether to continue processing” as required by limitations 1(d) and 1(e). In contending that Jiang alone teaches limitations 1(d) and 1(e), the Petition quotes from Jiang's description of its pruning technique:

Pruning is accomplished by **comparing, at each node, the scores assigned to that node with the largest score on any of the other nodes** corresponding to the frame being considered. **If the score at a particular node is sufficiently low** compared to the maximum score for other corresponding nodes in tree 77, it is assumed that the branches leading to the node under consideration (the node with the lower score) will not likely be part of the completed most likely phoneme sequence (i.e., the phonemes in that sequence will not likely be part of the final word recognized by the system). Thus, **that branch is dropped (or pruned) from tree 77 and is no longer considered in further processing.**

Pet. 38–39 (quoting Ex. 1004, 8:52–64) (emphasis in original). Petitioner concludes that “[p]er Mr. Schmandt, *Jiang* thus teaches an initial score . . . used to determine whether to continue processing to determine a final score *Dec.*, 90–91. As such, *Jiang* teaches claim 1(e).” Pet. 43–44.

The Petition explains how Jiang’s pruning algorithm generates an initial score and uses the initial score to prune, or “determine whether to continue processing” as recited in claim 1, but does not explain how Jiang’s pruning algorithm would be implemented on Chen’s four clusters of processors shown in Figure 4, as required by the Petition. That is, the Petition does not explain how each of the processors in the four clusters of Chen, after generating a score for one of the initial nodes, would compare the score with the maximum score of all corresponding initial nodes in tree 77. In fact, the Petition, in relying on Jiang alone to teach this limitation, does not even contend that Jiang’s pruning technique would be performed by any of Chen’s clusters of processors, let alone explain how pruning would have been performed by Chen’s parallel processing computer system.

In Reply, Petitioner contends that the Petition specifies that Jiang’s pruning algorithm of generating an initial score and using the initial score to

determine whether to continue processing, which is the functionality of limitations 1(d) and 1(e), would be replicated amongst each of Chen's processors. Reply 1–2 (citing Pet. 17–18, 24, 30; Ex. 1003 ¶¶ 74, 76–77, 83). Pages 17, 18, 24, 25 and 30 of the Petition contend that a person of ordinary skill in the art would have replicated Jiang's tree search functionality amongst each of the plurality of Chen's processors 102a–d in respective clusters 100a–d in order to compute probability scores as taught by Jiang using Jiang's acoustic model data stored in Chen's shared cluster memories 104a–d of each of the clusters, as required by limitations 1(a), 1(b), and 1(c). Paragraphs 74, 76–77, and 83 of Mr. Schmandt's Declaration contain similar testimony.

Thus, even in Reply, Petitioner's contention and the supporting testimony of Mr. Schmandt's Declaration, at best, explain how Chen's processors could have been configured to calculate probability scores for nodes of Jiang's prefix tree to meet limitations 1(a), 1(b), and 1(c). Petitioner's contention in Reply does not explain how Chen's processors 102a–d in clusters 100a–d would have performed Jiang's pruning technique, which determines the maximum score of the initial scores generated for each of the initial nodes in tree 77, and compares the initial score for a particular node to the maximum score of all the initial nodes in order “to determine whether to continue processing” as recited in limitation 1(e).

Patent Owner in its Sur-Reply identifies several unanswered questions raised by Petitioner's contention in Reply that Jiang's algorithm would have been replicated across Chen's processors to provide the functionality of limitations 1(d) and 1(e), such as why would every processor in Chen's four clusters have performed limitations 1(d) and 1(e); why would every processor have calculated the same initial score and make the same

determination whether to continue processing; would there have been an advantage in using all of Chen’s processors to perform limitations 1(d) and 1(e) when, according to Mr. Schmandt, one processor may have been enough; and how would each processor have obtained necessary information stored across all memories 104a–d in Chen, given that none of Chen’s processors can directly or adjacently access all cluster memories. PO Sur-Reply 3–4 (citing Reply 2). We decline to speculate about the answers to these questions. We find that Patent Owner’s questions raised by Petitioner’s contention in Reply emphasize the importance of the requirement that the initial Petition identify “with particularity” the “evidence that supports the grounds for the challenge to each claim.” 35 U.S.C. § 312(a)(3). Here, the Petition does not identify “with particularity” how the combination of Jiang and Chen teaches limitations 1(d) and 1(e).

Patent Owner contends that Petitioner’s Reply theory was not set forth in the Petition, and is therefore untimely. PO Sur-Reply 4. We agree with Patent Owner. The Petition relies on Jiang alone to teach limitations 1(d) and 1(e), as discussed above. Petitioner’s contention that Jiang’s algorithm would have been replicated across Chen’s processors for performing limitations 1(d) and 1(e) is not found in the Petition. Rather, as discussed above, the Petition describes how Jiang’s algorithm would have been replicated for performing limitations 1(a), 1(b), and 1(c). Petitioner’s new theory in Reply is untimely. A “[p]etitioner may not submit new evidence or argument in reply that it could have presented earlier, e.g., to make out a *prima facie* case of unpatentability.” Patent Trial and Appeal Board Consolidated Trial Practice Guide (Nov. 2019) 73;² *see also Intelligent Bio-*

² Available at www.uspto.gov/TrialPracticeGuideConsolidated.

Sys., Inc. v. Illumina Cambridge Ltd., 821 F.3d 1359, 1369–70 (Fed. Cir. 2016) (holding that the Board did not err in refusing the reply brief as improper under 37 C.F.R. § 42.23(b) because petitioner relied on an entirely new rationale to explain why one of skill in the art would have combined the references at issue); 37 C.F.R. § 42.23(b); *Apple Inc. v. e-Watch, Inc.*, IPR2015-00412, Paper 50, 44 (PTAB May 6, 2016) (“‘Respond,’ in the context of 37 C.F.R. § 42.23(b), does not mean embark in a new direction with a new approach as compared to the position originally taken in the Petition. Accepting such belatedly presented new arguments would be unjust to the Patent Owner and we decline to do so.”).

Petitioner contends that Mr. Schmandt explained during his deposition that, even though paragraphs 83–87 of his Declaration do not mention Chen, other parts of his Declaration explain how the combination of Jiang and Chen would have operated. Reply 2 (citing Ex. 2017, 44:2–11, 45:10–46:10, 71:6–23, 73:12–17, 79:11–23, 81:8–22, 83:15–18, 86:25–87:23, 93:7–94:10, 94:16–23, 95:19–96:10, 97:8–17, 99:17–102:12). We disagree. During his deposition, Mr. Schmandt did not identify where his Declaration explains how the combination of Jiang and Chen teaches the limitations of claim 1(d) and 1(e). Neither Petitioner nor Mr. Schmandt identifies any paragraph of Mr. Schmandt’s Declaration that explains how the processors in Chen’s cluster of processors would have performed Jiang’s pruning technique in order to meet limitations 1(d) and 1(e). Rather, both the Petition and Mr. Schmandt’s Declaration rely on Jiang alone to teach limitations 1(d) and 1(e). *See* Pet. 38–44; Ex. 1003 ¶¶ 84–91; Ex. 2017, 68:17–25, 79:11–23, 81:8–22, 83:10–18.

Petitioner contends that Patent Owner’s contention “turns on bodily incorporation of *Chen*’s architecture.” Reply 3. We disagree. Patent

Owner's contention turns on the Petition relying on Jiang alone to teach limitations 1(d) and 1(e). *See* Pet. 38–44. As discussed above, the Petition and Mr. Schmandt explain how implementing Jiang's algorithm using each of Chen's processors 102a–102d teaches limitations 1(a), 1(b), and 1(c). Yet, inexplicably, the Petition relies on Jiang alone to teach limitations 1(d) and 1(e). We agree with Patent Owner, that neither the Petition nor Mr. Schmandt's Declaration explains how Jiang's pruning technique would have been combined with each of Chen's processors 102a–102d, which specific elements of Chen would perform Jiang's pruning technique, or how the combination of Jiang and Chen would have operated or read on limitations 1(d) and 1(e). *See* PO Resp. 29 (citing *ActiveVideo Networks, Inc. v. Verizon*, 694 F.3d 1312, 1327 (Fed. Cir. 2012)). We also agree with Patent Owner that Mr. Schmandt's declaration admits that Jiang's algorithm would have to be modified to perform pruning using Chen's clusters of processors. PO Sur-Reply 1 (citing Ex. 2017, 46:12–22, 68:17–25, 83:10–14).

Petitioner contends that “[n]either the Challenged Claims nor caselaw requires [Petitioner] to explain how the combination utilizes Chen's clusters of processors, memories, etc.” Reply 3. We disagree. “[T]he burden of persuasion is on [P]etitioner to prove unpatentability by a preponderance of the evidence, and that burden never shifts to [Patent Owner].” *Dynamic Drinkware, LLC v. Nat'l Graphics, Inc.*, 800 F.3d 1375, 1378 (Fed. Cir. 2015)). “Failure to prove the matter as required by the applicable standard means that the party with the burden of persuasion loses on that point – thus, if the fact trier of the issue is left uncertain, the party with the burden loses.” *Id.* at 1378–79. Here, Petitioner has the burden to show how the combination of Jiang and Chen would have operated to teach limitations 1(d) and 1(e). However, the Petition, in relying on Jiang alone to teach

limitations 1(d) and 1(e), does not show, by a preponderance of the evidence, how Jiang's pruning technique would have been performed by Chen's processors 102a–d. Therefore, Petitioner, who is the party with the burden of persuasion, loses on that point.

With respect to Mr. Schmandt's deposition testimony that one processor in one cluster in Chen would have compiled the scores from memories 104a–d in order to perform Jiang's pruning, because “[a]t that point we have finished our parallel processing operation and we're back to serial operations” (Ex. 2017, 102:25–103:3), Mr. Schmandt does not testify in his Declaration that Chen's parallel processing computer system would have performed any portion of Jiang's algorithm using serial operations. Similarly, the Petition does not contend that Chen's parallel processing computer system would have performed any portion of Jiang's algorithm using serial operations. Further, Petitioner in Reply does not rely on this new theory from Mr. Schmandt's deposition. Rather, Petitioner contends that a person of ordinary skill in the art would have included the plurality of processors of Chen to perform Jiang's tree search algorithm by “replicating the tree search engine's functionality [taught by Jiang] amongst each of the plurality of *Chen's* processors.” Reply 1 (citing Pet. 17–18).

We agree with Patent Owner that Petitioner, in Reply, requires every processor in Chen's clusters to perform Jiang's pruning algorithm to meet limitations 1(d) and 1(e), “thus abandon[ing] its own expert's testimony” that one processor would have performed limitations 1(d) and 1(e). PO Sur-Reply 3; *see* Reply 1–3. Given that Mr. Schmandt's deposition theory is inconsistent with Mr. Schmandt's Declaration theory of using Chen's system to perform Jiang's tree search algorithm in parallel, that Mr. Schmandt's deposition theory is inconsistent with the Petition's theory of using Chen's

system to perform Jiang's tree search algorithm in parallel, and that Petitioner in Reply does not rely on Mr. Schmandt's deposition theory of using Chen's system to perform Jiang's pruning algorithm using serial operations, we also do not rely on Mr. Schmandt's deposition theory of using Chen's system to perform Jiang's pruning algorithm using serial operations.

In sum, we determine that Petitioner has not proven that the combination of Jiang and Chen teaches limitations 1(d) and 1(e). Thus, Petitioner has not proven that challenged claims 1–3, 5, 7, and 8 would have been obvious over the combination of Jiang and Chen.

Contentions over reasons to combine Jiang and Chen

The Petition contends that a person of ordinary skill in the art “would have recognized that the increased computing resources provided by *Chen*'s cluster of processors would have increased the speed of the system, enabling a reduced cost system that recognized speech at a faster rate and/or higher level of accuracy.” Pet. 10 (citing Ex. 1003 ¶ 67). The Petition contends that “the more efficient processing architecture of *Chen* would have been simply substituted for *Jiang*'s architecture to achieve a compatible improvement in processing speed and power.” *Id.* The Petition contends that utilizing Chen's clusters of processors would have enabled Jiang's speech recognition circuit to efficiently process state sequences of hidden Markov models (HMMs) in parallel, resulting in improved memory management and allowing the pruning threshold to be relaxed. *Id.* at 18 (citing Ex. 1003 ¶ 74). The Petition contends that “the more efficient processing architecture of *Chen* would have simply been substituted for *Jiang*'s to achieve a compatible improvement in

processing speed and power.” *Id.* at 19; *see id.* at 25 (citing Ex. 1003 ¶ 74).

The Petition contends that a person of ordinary skill “would have been further motivated by *Hon*’s [Ex. 1006] express praise for the scalability and flexibility provided by clustered-processing computer platforms to speech recognition tasks.” *Id.* (citing Ex. 1003 ¶¶ 71–72; Ex. 1006, 24–25). The Petition contends that the “modification to *Jiang* to include multiple clusters of processors each with a memory (as taught by *Chen*) would have reduced the time required to recognize a user’s speech, speeding the system’s response and enabling a more relaxed pruning threshold.” *Id.* at 24 (citing Ex. 1003 ¶ 74).

Patent Owner contends that Mr. Schmandt’s Declaration assumes that the clustered processor and memory arrangement of *Chen* would have necessarily improved the speed and power of *Jiang*’s speech recognition, and thus allowed for a relaxed pruning threshold. PO Resp. 17 (citing Ex. 1003 ¶¶ 67, 74). Patent Owner contends that *Chen*’s distributed memory consisting of multiple private memories with one memory per cluster would require significant reprogramming effort to find the optimal approach to partitioning the processes and data of *Jiang*’s speech recognition techniques. PO Resp. 24–25 (citing Ex. 2020 ¶ 43). Patent Owner, relying on the testimony of Dr. Anderson, contends that “speech recognition search requires information to be shared among multiple recognition paths (candidate HMM state sequences) during the Viterbi decoding, [and that *Chen*’s architecture, which does not have direct communication between the clusters, presents a] poor choice for the

speech recognition search stage.” PO Resp. 25–26 (citing Ex. 2020 ¶ 43). Patent Owner also relies on additional evidence, such as Mathew I (Ex. 2021), to show that an ordinary artisan could not have assumed that Chen’s complex, clustered parallel processing architecture with multiple distributed memories would have improved the speed of Jiang’s speech recognition system or relaxed its pruning threshold. *Id.* at 18–19 (citing Ex. 2020 ¶¶ 36–37; Ex. 2021, 11).

Petitioner in Reply contends that “Mr. Schmandt opined ‘solitary, single-core processors sophisticated enough (for the time) to provide the necessary computing power would incur a high financial cost.’” Reply 19 (citing Ex. 1003 ¶ 67). Petitioner contends that Patent Owner does not explain why Chen’s clusters of processors would have been slower than Jiang’s system, but the clusters of processors described by the ’140 patent would not have been slower. Reply 20. Petitioner contends that Patent Owner’s reliance on Mathew I to show that the combination of Jiang and Chen would have required a high synchronization overhead, which would not have resulted in improved speed, is unsupported by evidence. *Id.* at 20–21.

Petitioner contends that clustered architectures were known for increased speed, therefore, Chen’s known clusters of processors would have improved Jiang’s speech recognition circuit in the same way, namely, by increasing speed. *Id.* at 21–23. Petitioner contends that Hon expressly teaches the benefit of scalability by adding more clusters. *Id.* at 23. Petitioner contends that the significant reprogramming effort required to port an existing program from a shared memory model to a distributed memory model as taught by Chen (Ex. 1005, 2:2–9) is solved by Chen’s architecture, not caused

by Chen’s architecture. *Id.* at 25. Petitioner contends that Chen’s architecture is a shared memory model within a cluster, and that each cluster does not need to continually access memories of other clusters to perform speech recognition. *Id.* at 26–27.

Patent Owner in Sur-Reply contends that Mr. Schmandt did not provide sufficient evidence to support his testimony about a cost benefit from replacing Jiang’s solitary processor with Chen’s parallel processing computer system, therefore, “[t]here is thus no basis for Petitioner’s ‘cost’ motivation.” PO Sur-Reply 22 (citing Ex. 2020 ¶¶ 38–39); *see id.* at 23–24. Patent Owner contends that Petitioner has not provided evidence that using Chen’s clustered architecture would have been faster than Jiang’s single processor. *Id.* at 24.

Patent Owner contends that Mathew I and Dr. Anderson provide evidence that in the context of parallel processing speech recognition systems, using more processors does not necessarily result in a speed benefit, and contends that Petitioner has not provided a basis for asserting that the combination of Jiang and Chen would have been “higher speed than Jiang alone.” *Id.* at 24–25 (quoting Reply 24). Patent Owner contends that Petitioner’s evidence neither shows a cost benefit nor a processing improvement. *Id.* at 25. Patent Owner contends that Petitioner ignores Dr. Anderson’s testimony that a person of ordinary skill in the art would not have desired the combination, given the significant reprogramming effort, onerous messaging overhead, and lack of direct communication across all clusters. *Id.* (citing Ex. 2020 ¶¶ 42–45).

Patent Owner contends that using Chen’s clustered architecture was not a known technique for improving a speech recognition

system. *Id.* at 26. Patent Owner contends that Hon's teaching of flexibility relates to the advantage of a programmable pipeline, not clustered architecture, and that Jiang already has the advantage of scalability. *Id.* Patent Owner contends that Petitioner is incorrect in contending that the combination does not use a distributed memory model, and that all of Chen's processors can access the cluster memory of every cluster using a shared memory model. *Id.* at 27–28 (citing Reply 26–27).

Analysis of reasons to combine Jiang and Chen

We disagree with Petitioner's contention that to perform speech recognition, the processors in each cluster would not need to access the memories of the other clusters. *See* Reply 26. Petitioner's contention is inconsistent with Petitioner's contention that "[o]nce the cumulative probability score is calculated, inter-cluster processing of the results for each tree search would be communicated." Reply 4; *see id.* at 16. Petitioner's contention is inconsistent with the intrinsic evidence, which discloses that parallel implementation of a search algorithm for speech recognition requires a significant amount of communication and synchronization. Ex. 1001, 1:47–52. Petitioner's contention is inconsistent with Mr. Schmandt's testimony that

if we're going to be doing pruning based on thresholds, then **we have to know the scores that we got on each of the trees**, so I think that's a safe assumption, that that's in there. So when each of the processors has performed its task on its set of trees, it can produce a set of numbers, and **looking at the numbers across all of the processors, across all of the results**, we can choose a pruning threshold just like Jiang chooses a pruning threshold after it performs all these operations in serial as opposed to parallel.

Ex 2017, 101:10–23 (emphasis added). Petitioner’s contention is also inconsistent with Jiang’s teaching that pruning is performed by comparing a score at a particular node to the maximum score for other corresponding nodes in the prefix tree. Ex. 1004, 8:55–64. We rely on Dr. Anderson’s testimony, that a parallel implementation of Jiang’s algorithm “requires extensive communication between computational components” and that “all four memories 104(a)-(d) would contain necessary information” to perform limitations 1(d) and 1(e), in finding that performing Jiang’s algorithm using Chen’s clusters of processors would have required the processors to access data in all of the cluster memories 104a-d in order to determine the scores of the corresponding nodes of the tree. Ex. 2020 ¶¶ 29, 51.

We disagree with Petitioner’s contention that performing Jiang’s algorithm using the processors of Chen would have resulted in the benefit of an increase in speed. Pet. 10. We credit Dr. Anderson’s testimony that Mr. Schmandt’s assumptions that the clustered processor and memory arrangement of Chen would have necessarily improved the speed and power of Jiang’s speech recognition are incorrect. Ex. 2020 ¶¶ 35 (citing Ex. 1003 ¶¶ 67, 74), 36 (citing Ex. 2021, 11).

Petitioner contends that Dr. Anderson’s testimony is conclusory and that Patent Owner does not provide evidence other than Mathew I to support its contention that the combination would have a high synchronization overhead. Reply 20–21. We disagree, and find that Dr. Anderson’s testimony is sufficiently supported by evidence, including Jiang, Chen, and Mr. Schmandt’s Declaration. For example, Dr. Anderson testifies that the Viterbi algorithm compares

the likelihoods of all possible sequences for a given frame, given the most likely outcomes from the previous frame. Ex. 2020 ¶ 29 (citing Ex. 1004, 1:19–39, 2:1–26). Dr. Anderson testifies that to perform these comparisons requires extensive communication between computational components that can have a drastic impact on system performance. *Id.* ¶¶ 29, 43. Dr. Anderson testifies that as a result, performing Viterbi decoding when the information needed may be in a cluster that is not directly accessible, as in Chen, is not likely to be successful in a practical speech recognition system. *Id.* (citing Ex. 1003 ¶ 95; Ex. 1005, 9:10–39); *see id.* ¶ 43. Thus, Dr. Anderson further supports his testimony with citations to Jiang, Chen, and Mr. Schmandt’s Declaration.

In addition, Mr. Schmandt does not support with objective evidence his testimony that using Chen’s clusters of processors would have increased the computing power for implementing Jiang’s speech recognition techniques to desirably result in improved processing speed and power. *See* Ex. 1003 ¶¶ 67, 74. Although Mr. Schmandt mentions Hon (Ex. 1006), Tremblay (Ex. 1027), Erman (Ex. 1025), Wulf (Ex. 1024), Garde (Ex. 1028), and Deo (Ex. 1026) in his testimony, Mr. Schmandt does not persuasively explain how these background references would have taught a person of ordinary skill in the art that Chen’s parallel processing computer system, which includes an extended shared memory model and a distributed memory model, would have increased the speed of Jiang’s speech recognition algorithm over Jiang’s shared memory model.

We agree with Dr. Anderson that Jiang’s existing architecture relies on a shared memory structure in which all components have

access to the same memory. Ex. 2020 ¶ 42 (citing Ex. 1004, Fig. 1); *see* Ex. 1004, 5:4–9. Chen describes this type of memory as a true shared memory, and that a system that uses a shared memory model is the most attractive from a programmer’s perspective. Ex. 1005, 1:66–2:6, 2:24–29 (“Thus, it would be desirable to have a parallel processing system” in which “processors all have equal and symmetric access to any and all memory locations in a common shared memory.”), 9:46–48. Jiang discloses that its CPU 21 may include one or more processors. Ex. 1004, 6:39–41, Fig. 1; Pet. 14. Dr. Anderson testifies that performing the Viterbi algorithm requires extensive communication. Ex. 2020 ¶¶ 29, 43; *see* Ex. 1001, 1:47–52. Petitioner and Mr. Schmandt have not persuasively explained why a person of ordinary skill in the art would have replaced Jiang’s system of multiple processors and a true shared memory, which is the most attractive from a programmer’s perspective (Ex. 1005, 1:66–2:5), with Chen’s clusters of processors that are effective for applications that require little to no communication (Ex. 2019, 533).

We agree with Dr. Anderson that Petitioner and Mr. Schmandt have not explained why Chen’s clusters of processors, coupled to both extended shared memories and distributed memories, would have performed the Viterbi algorithm faster than Jiang’s embodiment of multiple processors coupled to a true shared memory, or even Jiang’s embodiment of a single processor coupled to the shared memory. Ex. 2020 ¶¶ 41–45. For example, Mr. Schmandt has not provided persuasive evidence that the number of inter-cluster memory accesses made by Chen’s clusters of processors when performing Jiang’s algorithm would have been small enough to avoid the latency issues

described by Chen. Ex. 1005, 7:42–8:4. This is especially pertinent here because a parallel implementation of the Viterbi algorithm “requires a significant amount of communication.” Ex. 1001, 1:47–53; Ex. 2020 ¶¶ 29, 43. Mr. Schmandt has not persuasively explained how using Chen’s clusters of processors to perform Jiang’s Viterbi algorithm would have resulted in increased speed and efficiency, given the latencies caused by processors in one cluster accessing memories in other clusters, and given that Jiang does not have this latency because Jiang uses a shared memory in which all components have access to the same memory. Ex. 1005, 7:42–8:4; Ex. 2020 ¶ 42.

The evidence regarding whether a person of ordinary skill in the art would have used clusters of processors to perform Jiang’s speech recognition algorithm because the clusters would have increased performance speed due to the increased number of processors, or would not have used clusters because the clusters would have decreased performance speed due to the increased latency from inter-cluster memory accesses and other increases in complexity, is inconclusive at best. Given that Petitioner has the burden of establishing by a preponderance of the evidence that using clusters of processors would have increased the speed of performing Jiang’s algorithm, we rely on the testimony of Dr. Anderson and the disclosure of Chen in finding that Petitioner has not met this burden.

We disagree with Petitioner’s contention that using commercially available processors in Chen’s clusters of processors would have been less expensive than Jiang’s embodiment of using a single processor. Reply 19; *see* Pet. 9–10; Ex. 1003 ¶¶ 67, 74. Chen’s parallel processing computer system has four clusters of

processors, and each cluster has at least two processors. Ex. 1005, 9:5–19; Fig. 4. Thus, Chen’s system requires at least eight processors. Although Mr. Schmandt testifies that a solitary processor with necessary computing power would incur a high financial cost (Ex. 1003 ¶ 67), Mr. Schmandt does not testify that the cost of Jiang’s solitary processor with the necessary computing power would have been greater than Chen’s system that includes at least eight processors. *See* PO Sur-Reply 23–24 (citing Ex. 2017, 62:2–9). Rather, Mr. Schmandt testifies that using off-the-shelf components in Chen’s clusters of processors would have been attractive in terms of cost compared to competing clustered architectures. Ex. 2017, 62:10–63:7.

Further, Jiang discloses that its speech recognition system “includes a general purpose computing device in the form of a conventional personal computer.” Ex. 1004, 5:4–6. Even considering Jiang’s embodiment that uses multiple processors, such processors would have been conventional processors found in conventional personal computers, and therefore would not have been any more expensive than those of Chen. *See id.* We find that the Petition does not show by a preponderance of the evidence that using Chen’s system that includes eight processors in place of Jiang’s single processor would have yielded the benefit of reducing cost over Jiang’s single processor included in a conventional personal computer.

Petitioner contends that Chen’s clustered architecture provides the known advantages of increasing speed and lowering cost compared to a special-purpose processor. Reply 22. Petitioner contends that using the known technique of Chen’s clustering of

processors would have improved Jiang's speech recognition device in the same way, namely, to increase speed and reduce cost. *Id.*

However, as discussed above, Petitioner has not shown that Chen's clusters of processors would have increased the speed of performing Jiang's speech recognition algorithm, and has not shown that Chen's clusters of processors would have been less expensive than the processor of a conventional personal computer as taught by Jiang. We find that Chen's technique of clustering processors, which requires at least eight processors and also increases latency compared to a true shared memory model due to inter-cluster memory accesses, was not known to increase speed or reduce cost of performing a speech recognition algorithm and therefore would not have improved Jiang's speech recognition system in the same way.

Petitioner contends that Patent Owner has not responded to the motivation to use clustered architecture in speech recognition provided by the multiple background references of Tremblay (Ex. 1027), Wulf (Ex. 1024), Garde (Ex. 1028), and Deo (Ex. 1026). Reply 23 (citing Ex. 1003 ¶¶ 71–72). However, Petitioner does not persuasively explain how the background references would have motivated a person of ordinary skill in the art to replace Jiang's system that uses a shared memory with Chen's parallel processing computer system, including an extended shared memory model and a distributed memory model, to perform Jiang's speech recognition algorithm, given that the extended shared memory model and the distributed memory model increase latency and complexity. As discussed above, Petitioner has not persuasively explained how using memory models that increase latency and complexity would have

increased performance speed over a system that uses a true shared memory model.

We disagree with Petitioner's contention that using clusters of processors would have provided the benefits of flexibility and scalability to Jiang's speech recognition system as taught by Hon. Reply 23 (citing Ex. 1006, 24–25); *see* Pet. 23. Given that Petitioner has not shown that increasing the number of Chen's clusters would have increased speed or reduced cost, we are not persuaded that a person of ordinary skill would have wanted the scalability from adding additional clusters to Jiang's speech recognition system. Further, we agree with Patent Owner that Hon's teaching of flexibility is tied to the benefits of a programmable pipeline rather than clusters of processors, and that Jiang already has Hon's benefit of scalability. PO Sur-Reply 26 (citing Ex. 2020 ¶¶ 40–41).

In sum, we determine that Petitioner has not proven that an ordinarily skilled artisan would have been motivated to modify Jiang based on Chen's teachings. Thus, in addition determining that Petitioner has not proven that the combination of Jiang and Chen teaches limitations 1(d) and 1(e) as discussed above, we determine that Petitioner has not proven that challenged claims 1–3, 5, 7, and 8 would have been obvious over the combination of Jiang and Chen.

E. Additional Grounds and Claims

Petitioner challenges claims 1–3, 5, 7, and 8 as unpatentable over Jiang, Chen, and Lucke. Pet. 53–58. Petitioner relies on Lucke to teach that a person of ordinary skill would have caused Jiang's pruning technique to continue processing when scores are above a threshold. Pet. 53. However, Petitioner does not rely on Lucke to show how Jiang's pruning technique would have been performed by Jiang's clusters of processors to meet

limitations 1(d) and 1(e). Petitioner also does not rely on Lucke to show why a person of ordinary skill in the art would have replaced Jiang's single processor that directly accesses a memory with Chen's clusters of processors that require at least eight processors and include an extended shared memory model and a distributed memory model, both of which increase latency as compared to Jiang's memory access. Thus, Petitioner has not proven that challenged claims 1–3, 5, 7, and 8 would have been obvious over the combination of Jiang, Chen, and Lucke.

Petitioner challenges claim 4 as unpatentable over Jiang, Chen, and Robinson. Pet. 59–60. Petitioner also challenges claim 4 as unpatentable over Jiang, Chen, Lucke, and Robinson. Pet. 60. Petitioner relies on Robinson to teach a Gaussian mixture model. Pet. 59. However, Petitioner does not rely on Robinson to show how Jiang's pruning technique would have been performed by Jiang's clusters of processors to meet limitations 1(d) and 1(e). Petitioner also does not rely on Robinson to show why a person of ordinary skill in the art would have replaced Jiang's single processor of a conventional personal computer that directly accesses a memory with Chen's clusters of processors that require at least eight processors and include an extended shared memory model and a distributed memory model, both of which increase latency as compared to Jiang's memory access. Thus, Petitioner has not proven that challenged claim 4 would have been obvious over the combination of Jiang, Chen, and Robinson. Petitioner also has not proven that challenged claim 4 would have been obvious over the combination of Jiang, Chen, Lucke, and Robinson.

Petitioner challenges claim 6 as unpatentable over Jiang, Chen, and Wrench. Pet. 60–62. Petitioner also challenges claim 6 as unpatentable over

Jiang, Chen, Lucke, and Wrench. Pet. 62. Petitioner relies on Wrench to teach a controller that initiates speech recognition processing in speech recognition processors. Pet. 61. However, Petitioner does not rely on Wrench to show how Jiang's pruning technique would have been performed by Jiang's clusters of processors to meet limitations 1(d) and 1(e). Petitioner also does not rely on Wrench to show why a person of ordinary skill in the art would have replaced Jiang's single processor of a conventional personal computer that directly accesses a memory with Chen's clusters of processors that require at least eight processors and include an extended shared memory model and a distributed memory model, both of which increase latency as compared to Jiang's memory access. Thus, Petitioner has not proven that challenged claim 6 would have been obvious over the combination of Jiang, Chen, and Wrench. Petitioner also has not proven that claim 6 would have been obvious over the combination of Jiang, Chen, Lucke, and Wrench.

III. CONCLUSION

In summary, we determine a preponderance of the evidence does not establish that claims 1–8 are unpatentable, as shown in the following table:³

³ Should Patent Owner wish to pursue amendment of the challenged claims in a reissue or reexamination proceeding subsequent to the issuance of this decision, we draw Patent Owner's attention to the April 2019 *Notice Regarding Options for Amendments by Patent Owner Through Reissue or Reexamination During a Pending AIA Trial Proceeding*. See 84 Fed. Reg. 16,654 (Apr. 22, 2019). If Patent Owner chooses to file a reissue application or a request for reexamination of the challenged patent, we remind Patent Owner of its continuing obligation to notify the Board of any such related matters in updated mandatory notices. See 37 C.F.R. § 42.8(a)(3), (b)(2).

Claim(s)	35 U.S.C. §	Reference(s)/Basis	Claims Shown Unpatentable	Claims Not Shown Unpatentable
1–3, 5, 7, 8	103(a)	Jiang, Chen		1–3, 5, 7, 8
1–3, 5, 7, 8	103(a)	Jiang, Chen, Lucke		1–3, 5, 7, 8
4	103(a)	Jiang, Chen, Robinson		4
4	103(a)	Jiang, Chen, Lucke, Robinson		4
6	103(a)	Jiang, Chen, Wrench		6
6	103(a)	Jiang, Chen, Lucke, Wrench		6
Overall Outcome				1–8

IV. ORDER

In consideration of the foregoing, it is hereby:

ORDERED that claims 1–8 of the '140 patent have not been proven by a preponderance of evidence to be unpatentable; and

FURTHER ORDERED that, because this is a final written decision, parties to this proceeding seeking judicial review of the Decision must comply with the notice and service requirements of 37 C.F.R. § 90.2.

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

APPLE INC.,
Petitioner,

v.

ZENTIAN LIMITED,
Patent Owner.

IPR2023-00037
Patent 10,971,140 B2

Before KEVIN F. TURNER, JEFFREY S. SMITH, and
CHRISTOPHER L. OGDEN, *Administrative Patent Judges*.

OGDEN, *Administrative Patent Judge*, concurring.

I concur in the result and I concur in the majority decision that Petitioner has not proven that the combination of Jiang and Chen teaches limitations 1(d) and 1(e) because it did not explain in its Petition how a person of ordinary skill in the art would have implemented Jiang's pruning algorithm on Chen's parallel-processing system, or the specific motivation to do so.

For the reasons given in the majority decision, I agree that Petitioner's efforts to provide a rationale combining Jiang and Chen for these limitations in its Reply are untimely. I agree that, without any such timely explanation

for limitations 1(d) and 1(e), Petitioner has not met its burden of persuasion to show that claims 1–8 are unpatentable as obvious under any of the grounds of the Petition.

Having determined that Petitioner’s rationale for combining Jiang and Chen in the context of limitations 1(d) and 1(d) is untimely, I would not reach the question of whether Petitioner has shown that there was a reason to combine Jiang and Chen at the time of the claimed invention. Thus, I respectfully do not join the majority’s analysis or determinations under the heading “Contentions over reasons to combine Jiang and Chen.”

IPR2023-00037
Patent 10,971,140 B2

FOR PETITIONER:

Jennifer Bailey
Adam Seitz
Adam Sandwell
ERISE IP, P.A.
jennifer.bailey@eriseip.com
adam.seitz@eriseip.com
adam.sandwell@eriseip.com

J. David Hadden
Dargaye Churnet
Saina Shamilov
FENWICK & WEST LLP
dhadden@fenwick.com
dchurnet@fenwick.com
sshamilov@fenwick.com

FOR PATENT OWNER:

Peter Knops
Kayvan Noroozi
NOROOZI PC
peter@noroozipc.com
kayvan@noroozipc.com

Katherine Rhoades
BARLIT BECK LLP
katherine.rhoades@bartlitbeck.com