UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

CROWDSTRIKE, INC.,
Petitioner

v.

SKYSONG INNOVATIONS, LLC,
Patent Owner

Inter Partes Review Case No. IPR2025-01399
U.S. Patent No. 11,775,831 B2

PATENT OWNER'S PRELIMINARY RESPONSE

TABLE OF CONTENTS

I. INTRODUCTION 1

II. RELATED PROCEEDINGS 1

III. SKYSONG BACKGROUND 1

IV. THE ASSERTED PRIOR ART 2

A. *Ujiie* (EX1004) 3

B. *Moons* (EX1005) 5

C. *Kaul* (EX1006) 6

D. *Ishii* (EX1007) 7

E. *Ravindran* (EX1008) 8

V. THE '831 PATENT 10

A. Technical Background 10

B. Prosecution History 12

C. Priority Date 13

VI. LEVEL OF ORDINARY SKILL IN THE ART 14

VII. PATENT OWNER'S CLAIM CONSTRUCTIONS 14

VIII. PETITIONER HAS FAILED TO DEMONSTRATE A REASONABLE LIKELIHOOD OF PREVAILING AS TO ANY CHALLENGED CLAIM. 15

A. GROUND 1: Obviousness in Light of *Ujiie* and *Moons* 15

1. Claim 1 15

2. Claim 10 26

3. Claim 11 26

4. Motivation To Combine 27

B. GROUND 2: Obviousness in Light of *Ujiie*, *Moons*, and *Kaul* 32

1. Claim 2 32

2. Claim 3 35

3. Claim 4 36

4. Claim 5 36

5. Claim 6 37

6. Claim 8 37

7. Claim 9 38

8. Motivation To Combine 38

C. GROUND 3: Obviousness in Light of *Ujiie*, *Moons*, *Kaul*, and *Ishii* 40

1. Claim 2 41

2. Claim 3 42

3. Motivation To Combine 43

D. GROUND 4: Obviousness in Light of *Ujiie*, *Moons*, *Kaul*, and *Ravindran* 44

1. Claim 7 44

2. Motivation To Combine 44

A. GROUND 5: Obviousness in Light of *Ujiie*, *Moons*, *Kaul*, *Ishii*, and *Ravindran* 45

1. Claim 7 46

2. Motivation To Combine....46
IX. SECONDARY CONSIDERATIONS OF NON-OBVIOUSNESS....46
A. Commercial Success....47
B. Long-Felt Need and Industry Support....48
C. Failure Of Others....48
D. Unexpected Results....49
E. Skepticism of Experts....49
F. Praise By Others....49
G. Copying By Competitors....49
X. CONCLUSION....49

TABLE OF AUTHORITIES

CASES

Graham v. John Deere Co. of Kan. City,
383 U.S. 1 (1966).....64

Phillips v. AWH Corp.,
415 F.3d 1303 (Fed. Cir. 2005) (*en banc*).....13

TABLE OF ABBREVIATIONS

Patent Owner	Skysong Innovations, LLC ("Skysong")
Petitioner	CrowdStrike, Inc. ("CrowdStrike")
PTAB	Patent Trial and Appeals Board
'831 patent	U.S. Patent No. 11,775,831 (EX1001)
Ujiie	Takayuki Ujiie et al., *Approximated Prediction Strategy for Reducing Power Consumption of Convolutional Neural Network Processor*, 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (2016) (EX1004)
Moons	Bert Moons et al., *Energy-Efficient ConvNets Through Approximate Computing*, IEEE Winter Conference on Applications of Computer Vision (2016) (EX1005)
Kaul	Himanshu Kaul et al., *A 21.5M-Query-Vectors/s 3.37nJ/Vector Reconfigurable k-Nearest-Neighbor Accelerator with Adaptive Precision in 14nm Tri-Gate CMOS*, 2016 IEEE International Solid-State Circuits Conference (2016) (EX1006)
Ishii	U.S. Patent Application Publication No. 2016/0259995 A1 to Ishii et al. (EX1007)
Ravindran	U.S. Patent Application Publication No. 2016/0259994 A1 to Ravindran et al. (EX1008)

LIST OF EXHIBITS

Exhibit No.	Descriptions
2001	Skysong's First Amended Complaint against CrowdStrike
2002	Civil Docket For Case # 7:25-cv-00040-DC-DTG (Crowdstrike)
2003	*Skysong v. CrowdStrike*, Dkt. No. 55, Scheduling Order
2004	Civil Docket For Case # 2:25-cv-00098-JRG (Fortinet)
2005	Table of CrowdStrike IPRs and Parallel District Litigation Schedules
2006	Revision to Rules of Practice Before the Patent Trial and Appeal Board
2007	Declaration of Kyle N. Siegal regarding discretionary denial
2008	Executive Order No. 14,028
2009	Default Protective Order
2010	Patent Owner's discretionary denial brief, redacted
2011	Declaration of Kyle N. Siegal regarding discretionary denial, redacted
2012	Declaration of Dr. Nenad Medvidovic
2013	M. Grootendorst, A Visual Guide to Quantization
2014	Cameron Nicklaus Christou, *Optimal Dither and Noise Shaping in Image Processing*, Thesis presented to the University of Waterloo, Waterloo, Ontario, Canada, (2008)

LIST OF CHALLENGED CLAIMS

Limitation	Claim Language
Claim 1	
1	One or more non-transitory computer-readable storage media storing instructions which, when executed by at least one processor, cause the at least one processor to perform operations comprising:
1.a	in one or more layers of a convolutional neural network (CNN), performing a first iteration that includes computing a value based on a first set of most significant bits (MSBs) for each of a plurality of data sets;
1.b	examining a first set of values computed for the plurality of data sets in the first iteration to determine whether a maximum value is present among the first set of values;
1.c	responsive to identifying the maximum value, performing a full precision computation of the value for a data set, of the plurality of data sets, that exhibited the maximum value; and
1.d	propagating the full precision computation of the value to a subsequent layer of the CNN.
Claim 2	
2	The one or more non-transitory computer-readable storage media of claim 1, further comprising: responsive to determining that the first set of values are the same, performing, by the at least one processor, a second iteration that includes computing the value based on a second set of MSBs for each of the plurality of data sets, the second set of MSBs being larger than the first set of MSBs.
Claim 3	
3	The one or more non-transitory computer-readable storage media of claim 2, further storing instructions, which, when executed by the at least one processor, cause the at least one processor to perform operations comprising:
3.a	examining a second set of values computed for the plurality of data sets in the second iteration to determine whether the maximum value is present among the second set of values; and
3.b	responsive to identifying the maximum value among the second set of values, performing, by the at least one processor, the full precision computation of the value for a data set, of the plurality of data sets, that exhibited the maximum value in the second iteration.

Claim 4	
4	The one or more non-transitory computer-readable storage media of claim 2, wherein the computing in each of the first iteration and the second iteration employs a convolution and a pooling.
Claim 5	
5	The one or more non-transitory computer-readable storage media of claim 4, wherein the convolution is a N×N convolution, where N is any integer.
Claim 6	
6	The one or more non-transitory computer-readable storage media of claim 4, wherein the pooling is a N×N pooling, where N is any integer.
Claim 7	
7	The one or more non-transitory computer-readable storage media of claim 4, wherein the convolution is a 3×3 convolution, and the pooling is a 2×2 pooling.
Claim 8	
8	The one or more non-transitory computer-readable storage media of claim 2, wherein at least one of the first iteration and the second iteration is performed with a precision less than that of the full precision computation.
Claim 9	
9	The one or more non-transitory computer-readable storage media of claim 8, wherein the precision is 8-bit precision.
Claim 10	
10	The one or more non-transitory computer-readable storage media of claim 1, wherein the CNN is employed to analyze an image.
Claim 11	
11	The one or more non-transitory computer-readable storage media of claim 1, wherein:
11.a	the first iteration computes a value that approximates the full precision computation of the value; and
11.b	the full precision computation is performed on the data set that includes less data than the plurality of data sets.

I. INTRODUCTION

Patent Owner Skysong respectfully submits this Preliminary Response to the Petition for *Inter Partes* Review dated September 26, 2025 ("Petition") of U.S. Patent No. 11,775,831 (EX1001, "' 831 patent"). This response is timely filed. *See* Paper 3.

II. RELATED PROCEEDINGS

The '831 patent is the subject of the following related litigation:

- *Skysong Innovations, LLC v. CrowdStrike, Inc. and CrowdStrike Holdings, Inc.*, Case No. 7:25-cv-00040, filed January 31, 2025, in the United States District Court for the Western District of Texas; and
- *Skysong Innovations, LLC v. Fortinet, Inc.*, Case No. 2:25-cv-00098, filed January 31, 2025, in the United States District Court for the Eastern District of Texas.

III. SKYSONG BACKGROUND

Skysong is a nonprofit organization, and the exclusive intellectual property management and technology transfer organization for Arizona State University ("ASU"). Skysong works closely with ASU and its technology industry partners to translate ASU's innovations into commercial applications, fostering economic development and public benefit. *See* EX2001 (Skysong's First Amended Complaint), ¶¶3–4. Skysong significantly contributes to the mission statement of

the University and has become a core component of Arizona's innovation ecosystem and a driver of high-tech industry in the state.

The '831 patent stems from research conducted by Dr. Jae-sun Seo and his team. *See* EX2007, ¶8. It discloses methods for "efficiently reducing the amount of total computation in convolutional neural networks (CNNs) without affecting the output result or classification accuracy." *See* EX1001, abstract.

IV. THE ASSERTED PRIOR ART

Dr. Nenad Medvidovic, Department Chair of the Computer Science Department of the University of Southern California and leading artificial intelligence and cybersecurity authority, has been retained by Skysong as an independent expert to assist in the technical analysis of the '831 patent and the asserted prior art. He has provided a declaration (EX2012, "*Declaration Of Nenad Medvidovic, Ph. D., In Support Of Patent Owner's Preliminary Response*") setting forth his opinions, including a synopsis of the relevant aspects of the asserted prior art. His synopsis of the asserted prior art is relevant and reproduced here in most relevant part for reference:

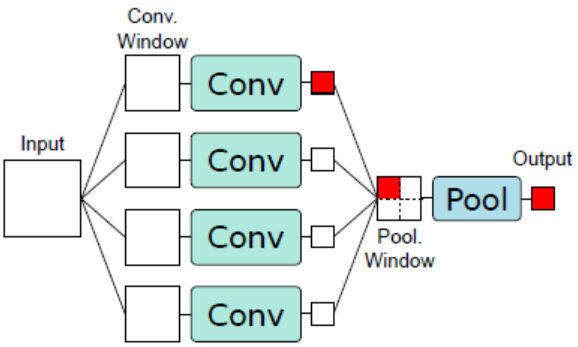
A. *Ujiie* (EX1004)[1]

Ujiie refers to Takayuki Ujiie *et al*., *Approximated Prediction Strategy for Reducing Power Consumption of Convolutional Neural Network Processor,* 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (2016).

Ujiie replaces conventional CNN architecture requiring exact convolutions against all the feature windows with an architecture that uses an approximate convolution to predict a relevant feature window on which the only relevant full precision convolution will be performed. *Ujiie* describes a method, LazyConvPool (LCP), designed to make CNNs more energy-efficient by avoiding unnecessary computations. Instead of performing full convolutions on every pixel in a pooling window, LCP first uses a lightweight, approximate convolution called AppConv to predict which feature will be selected by the max pooling operation. Only the predicted maximum feature is then processed with a full-precision convolution. To make AppConv efficient, a Sign Connect unit is used, which replaces multiplications with simple sign-based operations, reducing hardware complexity and power consumption.

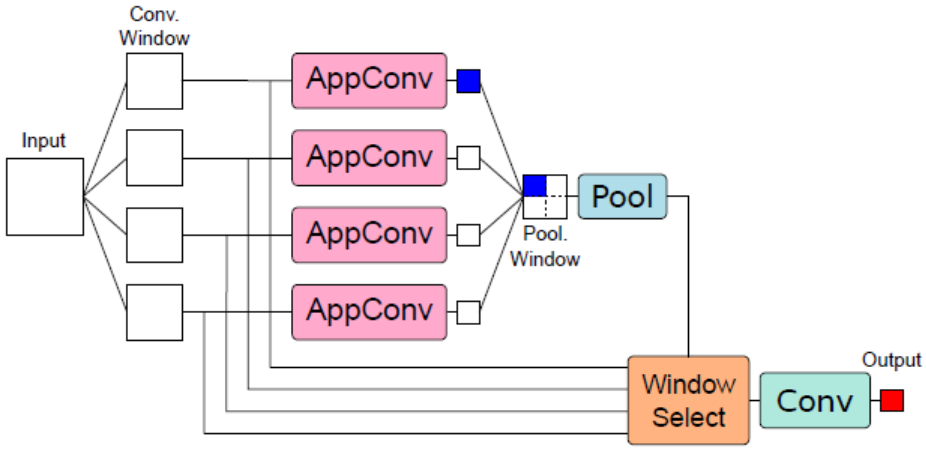
This is demonstrated in Figs. 2(a) and (b). Fig. 2(a) shows a conventional CNN architecture in which a full convolution is executed on all feature windows and the relevant feature window is then identified and used for the pooling operation:

[1] EX2012 at ¶¶90–95 (*Ujiie*).



(a) Conventional

Ujiie Fig. 2(a). The red indicated square if the feature window that is identified as the relevant one to forward to the pooling operation. Fig. 2(b) shows the operation of *Ujiie's* modified system, in which approximate convolutions (without multiplication overhead) are used to identify the relevant feature window to forward to the pooling layer:



(b) LCP

Ujiie Fig. 2(a). The blue indicated square is identified as the relevant feature map for pooling. This predicted feature map is used for the full precision convolution to determine the final full-precision output, indicated as the red square.

LCP, as described in *Ujiie*, predicts which feature in the pooling window matters using a very lightweight approximation and computes full convolution only for that feature. It focuses on skipping entire unnecessary convolutions rather than gradually refining precision.

Ujiie does not teach, and indeed teaches away from, the '831 patent, in which its MSB-based method performs iterative convolutions starting from the most significant bits, gradually increasing precision until the maximum is found. Unlike *Ujiie*, the '831 patent focuses on bit-level approximation and progressive refinement.

B. *Moons* (EX1005)[2]

Moons refers to Bert Moons *et al*., *Energy-Efficient ConvNets Through Approximate Computing,* IEEE Winter Conference on Applications of Computer Vision (2016).

Moons discloses quantization, i.e., precision scaling and computational skipping, as techniques to reduce energy consumption in CNN computations for small wearable devices. High-precision representations, like 32-bit floating point or 16-bit fixed point, are often unnecessary for accurate classification, so weights and inputs can be quantized to lower bit widths. Per-layer quantization of each layer's weights and inputs to fewer bits, sometimes as low as 4–10 bits, allows each layer to use a minimal number of bits, optimizing the energy-accuracy trade-off. The

[2] EX2012 at ¶¶96–99 (*Moons*).

process typically involves a greedy search to determine the optimal bit width for each layer, with out-of-range values clipped as necessary. Lower bit-width computations reduce switching activity in digital circuits, leading to significant energy savings. To maintain precision, sign bits must remain correctly positioned, and intermediate results are truncated or rescaled between layers using arithmetic shifts.

Unlike the teachings of the '831 patent, *Moons* discloses precision scaling and quantization that predefine the number of bits for each layer and rescale inputs and weights accordingly, aiming for fixed, low-precision computation from the start rather than iterative approximation.

Moons differs significantly from the '831 patent. It aims to maximize power savings at the expense of accuracy because it seeks to improve small wearable devices, it requires specific hardware adaptations, and it uses scaling and quantization, both of which independently modify the MSBs of the input data set.

C. *Kaul* (EX1006)[3]

Kaul refers to Himanshu Kaul *et al*., *A 21.5M-Query-Vectors/s 3.37nJ/Vector Reconfigurable k-Nearest-Neighbor Accelerator with Adaptive Precision in 14nm Tri-Gate CMOS*, 2016 IEEE International Solid-State Circuits Conference (2016).

[3] EX2012 at ¶¶100–102 (*Kaul*).

Kaul discloses an energy-efficient kNN hardware accelerator that saves power by starting distance calculations with coarse MSB-level approximations and refining them only when needed. Each iteration allows some vectors to be discarded quickly, so full-precision computation is done only for a smaller set of likely nearest neighbors. The hardware design uses compact arithmetic units, narrow data paths, partial-squaring, and low-voltage-friendly circuits to further reduce area and energy. A search-control system keeps track of which vectors remain in contention, enabling fast nearest and next-nearest neighbor detection. Overall, the accelerator achieves very high energy efficiency, high throughput, and reliable operation even at low supply voltages.

Kaul differs from the ’831 patent by using a sliding two bit wide window to incrementally compare vectors. Notably, *Kaul* is a nearest-neighbor algorithm, meaning it looks for the minimum distance, instead of the maximum value as the ’831 patent. Unlike the ’831 patent, *Kaul* does not use one set of MSBs in an initial iteration, and a larger set of MSBs in a second iteration. Instead, *Kaul* always uses only two bits in each iteration.

D. *Ishii* (EX1007)[4]

Ishii refers to U.S. Patent Application Publication No. 2016/0259995 A1 to Ishii *et al*.

[4] EX2012 at ¶¶103–106 (*Ishii*).

Ishii discloses an image-recognition a convolutional neural network that uses standard, parallel convolutions using multiple filters of differing scale/resolution (or color parameters) on the same image. For each pixel, the system selects one feature (e.g., the maximum response) from among all filter outputs. After performing all convolutions at their full resolution, the system simply applies max-pooling to pick the largest activation value, and in case of a tie, it arbitrarily selects one of the equal values. This differs from the '831 patent, which teaches that, in case of a tie, additional MSBs are used to determine the true maximum.

Ishii does not teach iterative approximation. Instead, it discloses computing all convolutions at full precision and performing max-pooling directly on those results.

Ishii does not use any feedback from the pooling stage to influence which computations to perform next, which is unlike the '831 patent that loops back the pooling outcome to selectively trigger full-precision convolution only for the candidate that appears to hold the maximum.

E. *Ravindran* (EX1008)[5]

Ravindran refers to U.S. Patent Application Publication No. 2016/0259994 A1 to Ravindran *et al*.

[5] EX2012 at ¶¶107–110 (*Ravindran*).

Ravindran describes a conventional deep-learning image-processing system that employs standard convolutional neural network layers built from full-precision convolutions and pooling operations. It builds multiple candidate CNNs (varying architecture, hyperparameters, preprocessing), trains and validates them on task-specific image data, and selects those meeting performance thresholds. The top-performing CNNs are then combined into an ensemble model whose aggregated predictions classify new images. This framework adapts automatically to different input types and use-cases, reducing development time and improving generalization by leveraging ensemble learning rather than a single fixed CNN. In one example, the system uses multiple convolutional layers with different kernel sizes including 3×3 filters in some layers and 2×2 max-pooling to explore and select optimal CNN architectures.

Ravindran does not teach MSB-Only approximation or iterative refinement. Instead, *Ravindran* performs full-precision 3×3 convolution and 2×2 pooling in the normal CNN pipeline. It does not teach using only the most significant bits to generate approximate convolution results, nor does it iteratively add more bits until a maximum is found—core requirements of the ’831 patent’s cascading convolution architecture.

Ravindran has no feedback loop and no mechanism where pooling decisions determine which convolution is refined. Instead, it computes all convolutions fully and independently.

V. THE '831 PATENT

A. Technical Background

Dr. Medvidovic provides a helpful explanation of the technical background of the '831 patent in paragraphs 73–79 of his declaration, reproduced in full here:

> 73. The '831 patent, titled "*Cascaded Computing For Convolutional Neural Networks*," is a continuation of U.S. Patent Application No. 16/335,775, which in turn claims priority through PCT/US2017/052736 to U.S. Provisional Patent Application No. 62/399,753, filed September 26, 2016. EX1001 at 1:7–18. The patent addresses the high computational cost of conventional convolutional neural networks ("CNNs") by reducing total computation without sacrificing classification accuracy. *Id.* at 1:21–31. In general terms, the claims recite performing iterative, MSB-based approximate computations across multiple data sets, identifying a maximum value, then performing a full-precision computation only for the data set exhibiting that maximum and propagating that full-precision result forward in the network. *Id.* at 1:32–2:40.
>
> 74. The '831 patent teaches techniques "for efficiently reducing the amount of total computation in convolutional neural networks (CNNs) without affecting the output result or classification accuracy. Computation redundancy in CNNs is reduced by exploiting the computing nature of the convolution and subsequent pooling …." *Id.* at abstract.
>
> 75. The prior art suffered from complex problems. For example, CNNs "have gained popularity in many computer vision applications (image, video, speech, etc.), because of their ability to train and classify with high accuracy. Due to multiple layers of convolution and pooling operations that are compute-/memory-intensive, it is difficult to perform real-time classification with low power consumption on today's computing systems." *Id.* at 1:22–29.

76. To address these shortcomings of the prior art, the ’831 patent discloses a “technique for efficiently reducing the amount of total computation in CNNs without affecting the output result or classification accuracy.” *Id*. at 1:34–36.

77. One embodiment of the ’831 patent teaches “in one or more layers of a convolutional neural network (CNN), performing a first iteration that includes computing a value based on a first set of most significant bits (MSBs) for each of a plurality of data sets; examining a first set of values computed for the plurality of data sets in the first iteration to determine whether a maximum value is present among the first set of values; responsive to identifying the maximum value, performing a full precision computation of the value for a data set, of the plurality of data sets, that exhibited the maximum value; and propagating the full precision computation of the value to a subsequent layer of the CNN.” *Id*. at 1:39–50.

78. Independent claim 1, for example, utilizes “one or more non-transitory computer-readable storage media storing instructions.” *Id*., claim 1. When executed, at least one processor performs operations in one or more layers of a CNN comprising:

- performing a first iteration that includes computing a value based on a first set of most significant bits (MSBs) for each of a plurality of data sets;
- examining a first set of values computed for the plurality of data sets in the first iteration to determine whether a maximum value is present among the first set of values;
- responsive to identifying the maximum value, performing a full precision computation of the value for a data set, of the plurality of data sets, that exhibited the maximum value; and
- propagating the full precision computation of the value to a subsequent layer of the CNN.

Id., claim 1.

79. In summary, the ’831 patent teaches that a highly innovative “MSB-driven convolution may produce similar results given that the lower precision bits in the data set would not have altered the maximum

> determination computation even if they had been included in the original computation. Accordingly, implementations initially perform a minimal amount of computation to determine and compare maximums among the data sets, and avoid the initial full convolution performed by conventional CNNs." *Id*. at 5:24–32.

EX2012 at ¶¶73–79.

B. Prosecution History

The provisional application that resulted in the '831 patent was filed on September 26, 2016, and the corresponding non-provisional application on January 13, 2023. EX1001.

On March 30, 2023, a preliminary amendment was submitted. EX1002 at 71.

On April 5, 2023, a second preliminary amendment was submitted. *Id*. at 86.

On April 17, 2023, a third preliminary amendment was submitted. *Id*. at 92.

On July 17, 2023, an interview with the examiner was held. *Id*. at 105.

On July 24, 2023, a notice of allowance issued. *Id*. at 136. The examiner specifically noted that the prior art does not disclose "in one or more layers of a convolutional neural network (CNN), performing a first iteration that includes computing a value based on a first set of most significant bits (MSBs) for each of a plurality of data sets; examining a first set of values computed for the plurality of data sets in the first iteration to determine whether a maximum value is present among the first set of values; responsive to identifying the maximum value, performing a full precision computation of the value for a data set, of the plurality

of data sets, that exhibited the maximum value; and propagating the full precision computation of the value to a subsequent layer of the CNN." *Id*. at 146.

C. Priority Date

The '831 patent derives priority from provisional application No. No. 62/399,753, filed on September 26, 2016. The provisional application fully discloses the subject matter of the claims in the '831 patent. EX2012 at ¶81. Petitioner alleges the following priority dates for the asserted prior art:

a. *Ujiie* (EX1004): August 5, 2016 (Pet. at 7)

b. *Moons* (EX1005): February 17, 2016 (Pet. at 9)

c. *Kaul* (EX1006): February 17, 2016 (Pet. at 10)

d. *Ishii* (EX1007): February 22, 2016 (Pet. at 11)

e. *Ravindran* (EX1008): March 4, 2015 (Pet. at 12)

Since these dates predate the provisional patent application as well, Patent Owner's analysis relies on the specification of the '831 patent itself. However, Patent Owner is still investigating the veracity of Petitioner's alleged dates. To the extent these dates are inaccurate and in fact post-date the provisional application, Patent Owner reserves the right to identify the respective portions of the provisional application supporting the claims of the '831 patent.

VI. LEVEL OF ORDINARY SKILL IN THE ART

In consideration of the type of problems encountered in the art, the prior solutions to those problems found in the prior art references, the rapidity with which innovations are made, the sophistication of the technology, the level of education of active workers in the field, and in reliance on Dr. Medvidovic's expert opinion, Patent Owner defines the relevant skill in the art as follows: A person of ordinary skill in the art ("POSITA") at the time of the filing of the '831 patent would have at least a Bachelor's degree in computer science (CS) or computer engineering (CE), and 2–4 years of experience working in the field of software development, distributed systems, or artificial intelligence (AI). A graduate degree in CS, CE, data science, AI, or software engineering may be substituted wholly or in part for work experience. EX2012 at ¶40. Dr. Medvidovic qualifies as a POSITA. *Id*. at ¶41.

VII. PATENT OWNER'S CLAIM CONSTRUCTIONS

Patent claims are construed using the framework set forth in *Phillips v. AWH Corp.*, 415 F.3d 1303, 1312–15 (Fed. Cir. 2005) (*en banc*). CrowdStrike relies on the plain and ordinary meaning of all claim terms. Pet. at 6. Patent Owner applies the ordinary meaning of the terms as a POSITA would have understood them at the time of the invention.

VIII. PETITIONER HAS FAILED TO DEMONSTRATE A REASONABLE LIKELIHOOD OF PREVAILING AS TO ANY CHALLENGED CLAIM.

A. GROUND 1: Obviousness in Light of *Ujiie* and *Moons*

Petitioner asserts that claims 1, 10, and 11 are obvious in light of *Ujiie* and *Moons*. These references differ significantly from and do not teach or suggest the invention claimed in the '831 patent.

1. Claim 1

> **Claim 1.** One or more non-transitory computer-readable storage media storing instructions which, when executed by at least one processor, cause the at least one processor to perform operations comprising:
>
> > ***a) Claim 1.a: in one or more layers of a convolutional neural network (CNN), performing a first iteration that includes computing a value based on a first set of most significant bits (MSBs) for each of a plurality of data sets;***

Petitioner admits that "*Ujiie* does not expressly discuss approximating convolutions by lowering precision 'based on a first set of most significant bits.'" Pet. at 16. Instead, Petitioner identifies *Moons'* "precision scaling" as "approximating convolutions" "which lowers the bit count of both the kernel weights and the original data inputs." *Id.* However, as explained by Dr. Medvidovic, *Moons'* quantization technique, including precision scaling, "differs significantly from the invention of the '831 patent." EX2012 at ¶115. Dr. Medvidovic identifies three independent reasons: (1) *Ujiie* and *Moons* disclose approximations while the '831 patent retains perfect classification accuracy; (2) the combination of *Ujiie* and *Moons* requires hardware updates while the '831 patent teaches software specifically

for use in devices with limited hardware capabilities; and (3) *Moons'* techniques do not rely or even retain the MSB's as claimed.

(1) Ujiie And Moons Only Disclose Approximations

Dr. Medvidovic first explains that the '831 patent does not affect the accuracy or precision of the CNN. EX2012 at ¶116. This contrasts with *Ujiie* and *Moons*, which disclose only approximations.

The '831 patent improves CNNs by "efficiently reducing the amount of total computation in convolutional neural networks (CNNs) *without affecting the output result or classification accuracy*.") '831 patent at abstract; *see also id.* ("the total number of bits used to perform the convolution is reduced *without affecting the output features or the final classification accuracy*."); *id.* at 1:24–36; *id.* at 3:8–10. The '831 identifies and avoids "redundant operations for which the results may not be employed subsequently in the analysis, and which therefore may not affect the overall outcome of the analysis," but it still performs "all key operations at full precision." EX2012 at ¶116 (citing '831 patent at 3:14–18 (redundant operations are skipped); 1:45–50 (relevant operations are performed at full precision); *see also* 2:8–10 (same); 4:1–9 (same)). "This is a key insight of the '831 patent." *Id*.

Moons has a different aim and does not disclose full precision. Unlike the '831 patent, *Moons* "maximize[s] power savings at the expense of accuracy, because it aims at improving small wearable electronic devices." *Id.* at ¶117. This is stated

in *Moons*: "While high energy consumption is no big concern during the network's training phase—which typically takes place on a computer cluster—it poses a problem when the network needs to be evaluated on mobile hardware like smartphones, smart glasses and other wearable devices." *Moons* at 1. *Moons* thus investigates "how energy consumption can be reduced through principles of *approximate computing*." *Moons* at 1. *Moons*' approach compromises accuracy to gain "drastic" power savings. "In order to be able to embed ConvNet-based classification into wearable platforms and embedded systems such as smartphones or ubiquitous electronics for the internet-of-things, their *energy consumption should be reduced drastically*." *Id*. (citing *Moons* at 1 (emphasis added)).

As Dr. Medvidovic explains, "*Moons* uses quantization, which is a lossy compression technique." *Id.* at ¶118. "*Moons* cannot guarantee that its techniques will not affect the overall outcome of the analysis." To the contrary, *Moons* explicitly states that its "main strategy is to quantize its weights and the inputs to its layers" in order to "reduce the energy consumption of the ConvNet's computations." *Id.* (citing *Moons* at 1). "Such quantization leads to a network that is only an approximation of the original network." *Moons* at 1.

Moons explicitly admits that its ("[p]recision scaling is an *approximate* computing technique allowing major energy savings in digital circuits.") *Moons* at 5 (emphasis added). Further, Petitioner admits that "*Moons'* goal is to save energy

while *minimizing accuracy loss*." Pet. at 17 (emphasis added). "Although *Moons* teaches that some energy gains could be made without losing accuracy, significantly better gains are achieved by sacrificing some classification accuracy: 'We can gain energy…up to 30x without losing classification accuracy" but "more than 100x at 99% classification accuracy.'" EX2012 at ¶119 (citing *Moons* at 1). *Moons'* "analysis shows that it is possible to quantize existing networks and drastically reduce the number of bits that encode the weights and inputs of each layer, *with only a minimal loss in accuracy*.") (emphasis added). "Especially given that *Moons* explicitly teaches 'approximate computing to reduce energy consumption in state-of-the-art ConvNet accelerators' and that 'energy consumption [for small wearable devices] should be reduced drastically,' a POSITA would understand that *Moons* was sacrificing some accuracy for drastic power consumption reduction." *Id.* (citing *Moons* at 1).

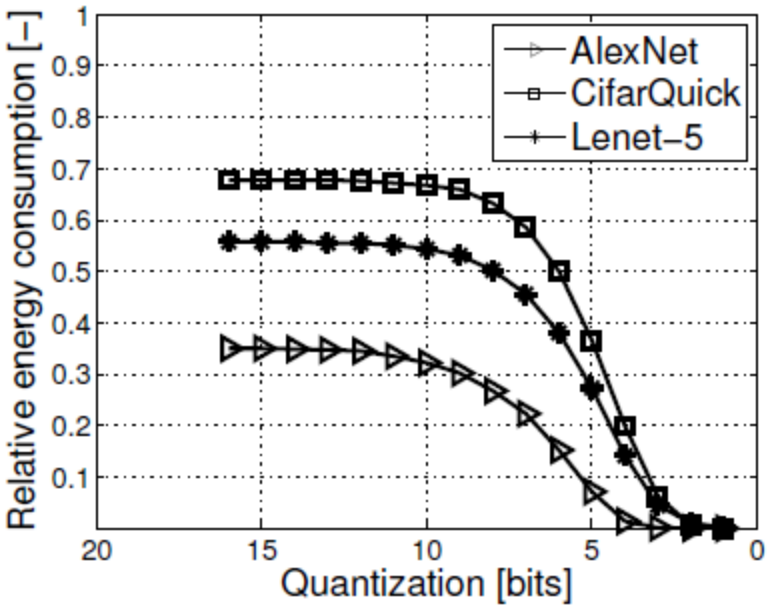
(2) Ujiie And Moons Require Hardware

Dr. Medvidovic further explains that the '831 patent discloses a software solution that teaches away from requiring hardware changes. EX2012 at ¶120. This contrasts with *Ujiie* and *Moons*, which disclose only approximations.

Moons explicitly requires an updated hardware architecture. *Id*. (citing *Moons* at 1 ("We show how this approximate computing / precision scaling can lead to reduced energy consumption in ConvNet accelerators. This is achieved through a

combination of algorithmical and circuit-level techniques.")). "With an appropriate hardware architecture this knowledge can be used to reduce energy consumption in many neural network applications and even dynamically scale energy consumption depending on the current application." *Moons* at 1. "By combining techniques both at the system- and circuit level, we can gain energy in the systems arithmetic." *Moons* at abstract. "*Moons* thus teaches using an updated hardware architecture." EX2012 at ¶120.

As Dr. Medvidovic explains, "*Moons* teaches two techniques to achieve its energy savings objective, precision scaling and computational skipping." *Id.* at ¶121 (citing *Moons* at 5–6). "Precision scaling increases the number of zeros in the dataset." *Id.* (citing *Moons* at 6 ("the number of zeroed values is significantly larger under precision scaling")). To account for this effect, *"Moons* teaches the additional method of computational skipping, which skips '0' inputs because they do not affect the calculations." *Id. Moons* figure 5 supports this and shows the "relative energy reduction compared to the case without computation skipping:"



(b)

Moons, Fig. 5(b). For these reasons, "*Moons* thus teaches that both precision scaling and computation skipping should be used to achieve maximal energy savings." *Id.*

For these Reasons, *Moons* explicitly requires hardware modifications:

> It is very difficult to achieve such computation skipping in a software solution, since checking for zero-valued inputs in a scalar core consumes time. Using techniques for sparse matrix multiplication is not possible either, because the used convolutional kernels are typically very small on the order of 11 × 11, 5 × 5 or 3 × 3. However, *computation skipping can be achieved in a hardware accelerator with dedicated hardware support*. Flags can indicate if upcoming data is zero and prevent circuitry from switching if this is the case. As ConvNet-weights are fixed, these flags can be computed beforehand.

Moons at 6. "*Moons* thus specifically discourages software based approaches in favor of a hardware accelerator." EX2012 at ¶122.

Ujiie also requires a specific hardware architecture. EX2012 at ¶123 (citing *Ujiie* at abstract (*Ujiie* proposes "an efficient computing method, LazyConvPool

(LCP), *and its hardware architecture* to reduce power consumption of CNN-based image recognition.") (emphasis added); *see also Ujiie* at 4–5 ("Hardware architecture").

The '831 patent, on the other hand, teaches away from imposing any hardware requirements. EX2012 at ¶124 (citing '831 patent at 3:56–59 (the same hardware engine may be reused . . . , thus reducing the hardware footprint) and 8:40–44 (same); 4:16–17 (the invention is useful in "devices that have limited hardware capabilities"); and 4:46–49 ("Implementations may be especially useful in environments with limited hardware resources, given that implementations enable the hardware footprint and/or processing power to be substantially reduced.")) Per Dr. Medvidovic, "this is a critical difference." *Id.*

(3) Ujiie and Moons Do Not Use The MSBs

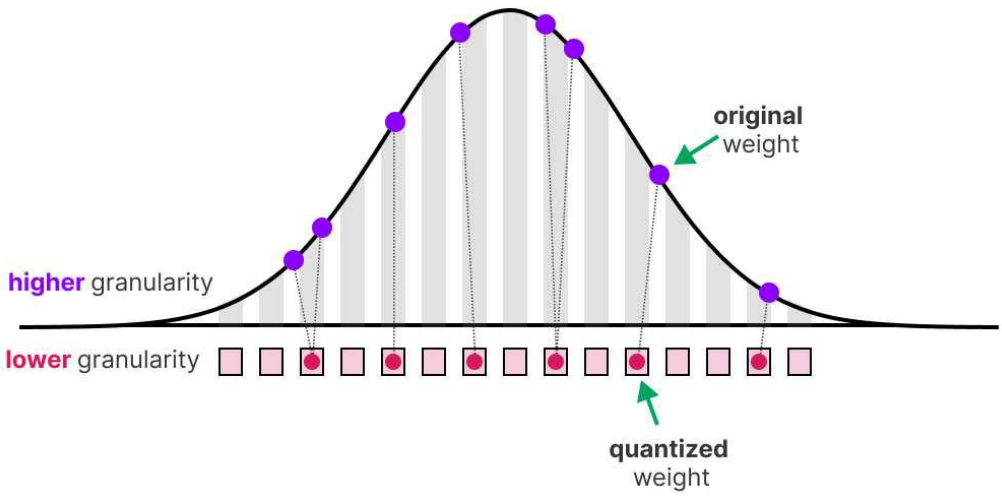
Dr. Medvidovic further explains that the combination of *Ujiie* and *Moons* does not utilize a set of most significant bits as claimed. EX2012 at ¶125.

Petitioner relies only on *Moons* for this claim limitation, but *Moons* "precision scaling" does not "comput[e] a value based on a first set of most significant bits (MSBs) for each of a plurality of data sets." '831 patent, claim 1. Dr. Medvidovic explains that "*Moons* describes its precise implementation or precision scaling as *quantization*." EX2012 at ¶125 (citing *Moons* at 2–6 (chapter 3 "Analysis and experiments")). Specifically, he points out that "*Moons* 'examines the performance-

and energy-related effects of quantization on three different [CNN] architectures for image classification.'" *Id.* (citing *Moons* at 2). This is confirmed by *Moons*, which states that "to reduce the energy consumption of the ConvNet's computations, our main strategy is to quantize its weights and the inputs to its layers." *Moons* at 3.

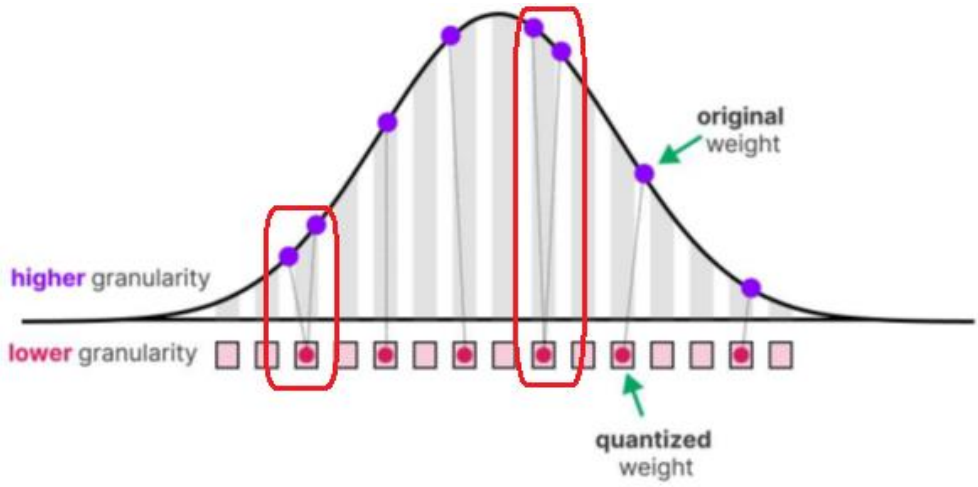
Dr. Medvidovic confirms that "quantization does not preserve the MSBs of the plurality of data sets that are quantized," and explains generally how quantization works and why it does not preserve the MSBs:

> Quantization is used to minimize the number of bits used to represent a data set. All values in the data set are mapped from a higher precision / higher bit representation into a lower precision / lower bit representation. For example:



EX2012 at ¶126 (citing M. Grootendorst, *A Visual Guide to Quantization*, available at https://www.maartengrootendorst.com/blog/quantization (EX2013) at 7. Dr. Medvidovic further confirms that this visual guide "accurately reflects the concepts of quantization as a POSITA would have understood it at the relevant time." *Id.* He continues:

In some instances, adjacent (but different) values in the original data set are mapped to the same value in the lower granularity data set. Examples are indicated in the annotated version of the same figure below:



Id. Dr. Medvidovic's annotated example confirms that "quantization used to downscale the values may change the MSBs significantly." *Id.*

Dr. Medvidovic further illustrates this fact using a simple example, in which the values 63 and 65 are both downscaled to 64:

Decimal	Binary						
	2^6 (64)	2^5 (32)	2^4 (16)	2^3 (8)	2^2 (4)	2^1 (2)	2^0 (1)
63	**0**	**1**	**1**	1	1	1	1
64	**1**	**0**	**0**	0	0	0	0
65	1	0	0	0	0	0	1

EX2012 at ¶127. "As is readily apparent, in the example of quantizing the number 63 to 64, the MSBs change from **011** to **100**." *Id.* A POSITA would not have understood *Moons* quantization scheme to teach or suggest using only—or even preserving—the MSBs for abbreviated calculations. To the contrary, given the number of parameters of the CNNs discussed in *Moons* (LeNet-5 with 60,000

parameters, CifarQuick with 36.7 million parameters, and AlexNet with ~61 million parameters (*Moons* at 2–3), a POSITA would know that *Moons'* quantization does not rely on the MSBs of the input data as claimed. *Id*.

Dr. Medvidovic provides another useful example to demonstrate the effect of quantization on the data:

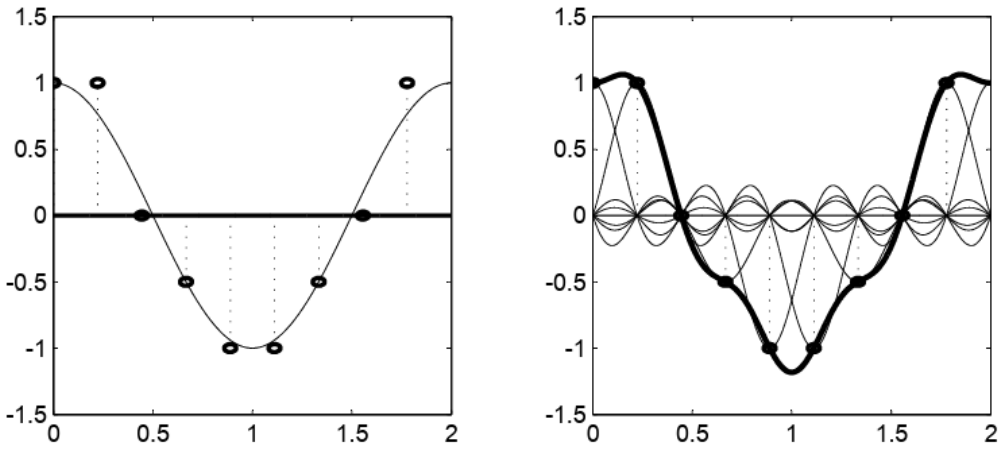


Figure 1.3: a) Quantized cosine signal. b) Cosine signal reconstructed from quantized points.

EX2012 at ¶128 (citing Cameron Nicklaus Christou, *Optimal Dither and Noise Shaping in Image Processing*, Thesis presented to the University of Waterloo, Waterloo, Ontario, Canada, (2008) (EX2014)). Figure 1.3.a shows an original input signal and its quantized data points. Figure 1.3.b, in contrast, shows the reconstructed signal, using the quantized data points. "The shape of the reconstructed signal is noticeably different from the original." *Id*. Notably, *Moons* seeks to minimize the quantization bits used, because "increasing quantization bits

comes at a heavy cost of increasing the computational complexity of *Moons'* CNN."

The changes to the MSB's caused by *Moons'* quantization scheme are multiplied because *Moons* scales the input data even before quantization. EX2012 at ¶129 (citing *Moons* at 3 ("Before quantizing the weights and inputs, it is important to rescale them properly according to the distribution of their values. If there is a mismatch between the interval in which these values lie and the interval over which we quantize, the accuracy will drop even at very low quantization settings.")). *Moons* therefore rescales "all layer inputs and weights in the network with a single value that corresponds to the maximal input or weight value (rounded to the next factor of two) observed during a complete run over the data." *Id.* (quoting *Moons* at 3). "This rescaling is accomplished 'multiplying the results with a factor of 2, or shifting data in a fixed point number representation.'" *Id.* Dr. Medvidovic confirms that "this operation changes the MSBs in nearly every instance."

> Only numbers for which all MSBs and the next most significant bit are "0" would not experience a change in MSB, which would be a very small fraction of the total value range. For example, the 8 bit unsigned integer range shown in Fig. 1B of the '831 patent allows representation of the numbers 0–255. With 2 MSBs (as indicated in Fig. 2B), *Moons* scaling would change the 2 MSB's for numbers 32–255. Only the range 0–31 would be an exception because all three MSBs are "0," which would [incidentally] keep the two MSBs static at "0" even after scaling.

EX2012 at ¶129. A POSITA would understand that *Moons'* precision scaling does not preserve the MSBs for the data sets as claimed in the '831 patent. *Id.*

Finally, the only reference to "MSB" in *Moons* relates only to considerations regarding signed binary numbers, because the MSB is customarily used to indicate the sign, i.e., positive or negative. *Moons* at 5 ("the positioning of the sign-bit in the fixed number representation is crucial. The MSB-bit should always be placed at the MSB position, otherwise toggling sign bits will lead to high switching activity."). This disclosure does not teach or suggest this claim element.

b) Claim 1.c: responsive to identifying the maximum value, performing a full precision computation of the value for a data set, of the plurality of data sets, that exhibited the maximum value; and

Ujiie explicitly notes that it experiences "performance degradation due to the prediction inaccuracy." *Ujiie* at 6. Thus, "*Ujiie* does not always identify the correct maximum value performing a full precision computation." EX2012 at ¶130.

2. Claim 10

Claim 10. The one or more non-transitory computer-readable storage media of claim 1, wherein the CNN is employed to analyze an image.

Claim 10 depends on claim 1 and is not disclosed or suggested for the reasons set forth regarding claim 1.

3. Claim 11

Claim 11. The one or more non-transitory computer-readable storage media of claim 1, wherein:

Claim 11 depends on claim 1 and is not disclosed or suggested for the reasons set forth regarding claim 1.

4. Motivation To Combine

A POSITA would not be motivated to combine *Ujiie* and *Moons* for multiple reasons.

First, as Dr. Medvidovic explains, *Ujiie* operates at full precision with all bits, while *Moons* only offers an approximation, due to its use of scaling and quantization. EX2012 at ¶133 (citing *Moons* at 3). "Indeed, in the MNIST architecture, *Ujiie* reports a higher accuracy than the conventional CNN." *Id.* (citing *Ujiie* at 5 and Table 3). *Ujiie* thus teaches full precision (or even better), while *Moons* is willing to trade accuracy for power savings. *Id.* "This is acceptable for *Moons* because it aims at small wearable devices that require drastic power reduction to function, and are therefore willing to accept the corresponding accuracy tradeoff." *Id.* (citing *Moons* at 1).

Second, *Ujiie* and *Moons* teaches away from the '831 patent because "they both require changes to the hardware architecture." EX2012 at ¶134 (citing *Moons* at 1 and *Ujiie* at abstract (referencing hardware requirements)). The '831 patent, on the other hand, teaches away from using specialized hardware. *Id.* at ¶124 (citing '831 patent at 3:56–59 (it is desirable to reduce the hardware footprint) and 8:40–44 (same); 4:16–17 ("the invention is useful in 'devices that have limited hardware

capabilities'"); and 4:46–49 ("Implementations may be especially useful in environments with limited hardware resources, given that implementations enable the hardware footprint and/or processing power to be substantially reduced.")) Per Dr. Medvidovic, "this is a critical difference." *Id.*

Third, Petitioner's argument that "a POSITA would have been motivated to replace *Ujiie's* Sign Connect approximations (i.e., using a single signed bit for kernel data and full bit-resolution for image data) with *Moons'* precision scaling quantization techniques (i.e., reduced bit-resolution for both kernel data and image data) to increase system accuracy" (Pet. at 19) is not correct. "To the contrary, *Ujiie* operates at full precision" and "avoids the degradation of recognition performance by using 'the approximation only to predict the region to be selected in the max pooling.'" EX2012 at ¶135 (citing *Ujiie* at 3). "Once the predicted region has been identified, a full-precision convolution followed by full-precision max pooling is performed on that region." *Id*. *Ujiie* therefore teaches away from using an approximation that degrades accuracy as taught in *Moons*. *Id.* Dr. Medvidovic provides support directly from *Moons,* which investigates "how energy consumption can be reduced through principles of *approximate computing*." *Id.* (citing *Moons* at 1). "*Moons* is willing to compromise accuracy because 'in order to be able to embed ConvNet-based classification into wearable platforms and embedded systems such as smartphones or ubiquitous electronics for the internet-of-things, their *energy*

consumption should be reduced drastically.'" *Id.* (citing *Moons* at 1 (emphasis added)).

Petitioner's citation to the declaration of Dr. Saber does not support its point that a POSITA implementing *Ujiie* would have been motivated to accept *Moons'* approximations. Pet. at 19 (citing EX1003 at ¶¶ 80 and 81). The cited part of *Ujiie's* is significantly mischaracterized and omits a key point emphasized by Dr. Medvidovic in the full citation shown below:

> The difference shows the performance degradation due to the prediction inaccuracy *using Ternary Connect* for the complex data such as color images.

EX2012 at ¶136 (citing *Ujiie* at 6). Petitioner's cited prediction inaccuracy refers to the old Ternary Connect method, not to *Ujiie's* new Sign Connect method. *Id.* "To the contrary, *Ujiie* explicitly states that 'for LCP with Sign Connect, the accuracy is almost equal to the conventional result in MNIST data set.'" *Id.* *Ujiie's* table 3 confirms this point:

Method	MNIST	CIFAR-10
(a) Conventional	0.986	0.860
(b) LCP	0.986	0.860
(c) LCP + TC	0.984	0.825
(d) LCP + SC	0.988	0.849

Table 3: Classification accuracy on datasets.

EX2012 at ¶137 (citing *Ujiie*, Table 3 (showing four types of CNN: (a) the conventional CNN, (b) CNN with LCP that computes AppConv by the exact

convolution, (c) AppConv by Ternary Connect, and (d) AppConv by Sign Connect) (emphasis added)).

> For the CIFAR-10 CNN, *Ujiie's* LCP + SC algorithm offers much better accuracy (0.849) then the alternative LCP + TC (0.825). Further, for the MNIST CNN, *Ujiie* even improves performance (0.988) over the conventional CNN at (0.986). Thus, a POSITA would have had absolutely no motivation to modify *Ujiie's* LCP + SC algorithm with *Moons* approximation with admittedly reduced accuracy. There is no support that *Moons* would have done any better than *Ujiie* in the CIFAR-10 architecture, and it is impossible for *Moons* to have exceeded *Ujiie's* performance in the MNIST architecture, because *Moons* scaling and quantization necessarily decreases accuracy due to the downscaling of the data as explained above regarding claim element 1(a).

EX2012 at ¶137. Thus, a POSITA would not have been motivated to combine *Ujiie* with *Moons*. Notably, *Moons'* primary objective of power savings for "small wearable devices" (which is the reason for *Moons'* willingness to sacrifice accuracy) is not a concern raised by *Ujiie*.

Fourth, Petitioner admits that "modifying *Ujiie* to perform multiplications of low bit-width may lead to an increase in energy consumption," and Dr. Medvidovic confirms that this is correct. EX2012 at ¶138. "*Ujiie* intentionally seeks to avoid multiplication, and thus teaches away from *Moons*." *Id.* (citing *Ujiie* at abstract (proposing "Sign Connect, which is a low computational-cost approximated prediction *without any multiplications*.") (emphasis added); *Id.* at 1 (*Ujiie's* Sign Connect "gives a significantly lightweight approximation of the convolution with no multiplication operation, while keeping prediction accuracy."); *Ujiie* at 4 ("With

Sign Connect, multiplication could be drastically simplified, and the total performance becomes deterministic."); *Id.* at 6 ("a new approach to reduce the number of multiplications in CNN based on an approximated prediction is proposed.")). "Thus, *Ujiie* teaches reducing the number of multiplications, and a POSITA would be motivated to avoid increasing the number of multiplications by the implementation of *Moons*." *Id.*

Fifth, Petitioner simply states that "the tradeoff for computational energy would be a desirable tradeoff for the enhanced accuracy *Moons's* [sic] process realizes." Pet. at 21. This is completely unsupported." EX2012 at ¶139. "To the contrary, *Moons* stresses the importance of drastically reducing energy consumption of its small wearable devices, even at the expense of slight accuracy loss. *Id.* (citing *Moons* at 1 ("In order to be able to embed ConvNet-based classification into wearable platforms and embedded systems such as smartphones or ubiquitous electronics for the internet-of-things, *their energy consumption should be reduced drastically*.") (emphasis added); *see also id.* (*Moons* explicitly teaches "*approximate computing to reduce energy consumption*")(emphasis added))). Thus, a POSITA looking at *Moons* would have been motivated to reduce power consumption at the expense of accuracy, unlike *Ujiie*.

B. GROUND 2: Obviousness in Light of *Ujiie*, *Moons*, and *Kaul*

Petitioner asserts that claims 2–6, 8, and 9 are obvious in light of *Ujiie*, *Moons*, and *Kaul*.

1. Claim 2

> **Claim 2**. The one or more non-transitory computer-readable storage media of claim 1, further comprising: responsive to determining that the first set of values are the same, performing, by the at least one processor, a second iteration that includes computing the value based on a second set of MSBs for each of the plurality of data sets, the second set of MSBs being larger than the first set of MSBs.

Claim 2 depends on claim 1 and is not disclosed or suggested for the reasons set forth regarding claim 1.

Petitioner admits that "*Ujiie-Moons* does not discuss what happens when max pooling does not identify a maximum." Pet. at 34. "The method disclosed in *Kaul* does not supplement this missing disclosure to teach or suggest this claim." EX2012 at ¶142.

(1) Kaul Does Not Identify A Maximum At All

Petitioner ignores that *Kaul* does not identify a maximum at all. Instead, "*Kaul* performs a k-nearest-neighbor search which does not rely on finding a maximum value." EX2012 at ¶143 (citing *Kaul* at 1 ("the sort network finds the minimum within current accuracy bounds, eliminating vectors outside these bounds for up to 6×compute reduction;" "the precise signals indicate whether all other psums exceed the minimum by more than 1, signifying that the same minimum has

been found as with full-precision calculations;" and "the compare psum circuit uses the global minimum to determine vector elimination and nearest neighbor selection.")). Thus, *Kaul* seeks to identify a minimum, not a maximum. Notably, Dr. Medvidovic confirms Petitioner's point that "*Ujiie's* native prediction strategy requires identifying a maximum." *Id*. (citing Pet. at 39). *Kaul* makes no reference to finding a maximum, and neither the Petitioner nor Dr. Saber substantiate the argument that *Kaul's* minimum detection solution satisfies *Ujiie's* maximum detection requirement. Contrary to Petitioner's unspoken assumption, "these two are not equivalent." *Id.*

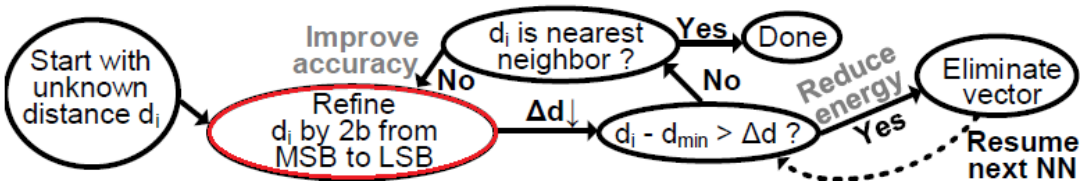
(2) The Concept Of Iteratively Increasing Bit-Widths In Neural Network Computations Was New

Petitioner argues that "the concept of iteratively increasing bit-widths in neural network computations is not new," but cites in support only *Kaul* and *Kim*. Pet. at 32 (citing *Kaul* (EX1006)and *Kim* (EX1012)). However, *Kaul* relates only to k nearest-neighbor (kNN) minimum distance computations (*Kaul* at 1) and *Kim* was published in 2021 (*Kim* at 1), several years after the priority date of the '831 patent. Neither disclose using iteratively larger MSB portions to determine a maximum.

(3) Kaul Does Not Teach Iteratively Increasing Bit Width In Neural Network Computations

Petitioner further argues that "*Kaul* teaches iteratively increasing bit width in neural network computations when a first iteration does not provide an answer." *Kaul* does not support this assertion. "Instead of increasing the bit width of the snippet of the original data to be analyzed, *Kaul* only ever analyzes a 2-bit window." EX2012 (citing *Kaul* at 1 ("Iterative refinement of accumulated distance is interleaved as needed with successive sort operations on 2b windows from MSB to LSB.")). "Every iteration only uses the next set of two bits." *Id.* (citing *Kaul* at 1 ("The partial distance compute circuits refine the remaining distances, *with a successive 2b of the accumulated distance available for further sorting in each iteration*.") (emphasis added)). Therefore, *Kaul* does not disclose this claim element, which requires "a second iteration that includes computing the value based on a second set of MSBs for each of the plurality of data sets, the second set of MSBs being larger than the first set of MSBs." "To the contrary, every calculation in *Kaul* is only on a two bit (2b) set." *Id.*

Dr. Medvidovic explains this in reference to *Kaul's* Fig. 14.4.1:



EX2012 at ¶145 (citing *Kaul*, Fig. 14.4.1 (emphasis added)). "As shown, each iteration only reviews a 2 bit window." *Id.* "During each iteration, the smallest vectors are discarded, and the next two bits are reviewed." *Id.* This is significantly different from performing convolutions of increasing accuracy based on larger and larger sets of MSBs. *Id*. First, *Kaul* requires a hardware solution to efficiently implement the sliding two bit window. *Id.* (citing *Kaul* at 1 (disclosing a "special-purpose on-die kNN accelerator")). "A POSITA would understand that *Kaul's* solution could not efficiently be implemented in software without undue experimentation." *Id.* *Kaul* provides no detail how to implement a software-only solution, and per Dr. Medvidovic, it would require "significant experimentation to implement *Kaul* as proposed by Petitioner." *Id.* Further, "the speed and power gains of *Kaul's* specialized hardware implementation may actually be lost" in a software solution. *Id*.

Finally, "*Kaul* does not teach computing a convolution," but "only teaches sorting vectors using its two bit sliding window." EX2012 at ¶146 (citing *Kaul* at 1).

2. Claim 3

> **Claim 3**. The one or more non-transitory computer-readable storage media of claim 2, further storing instructions, which, when executed by the at least one processor, cause the at least one processor to perform operations comprising:

Claim 3 depends on claim 2 and is not disclosed or suggested for the reasons set forth regarding claim 2.

3. Claim 4

> **Claim 4**. The one or more non-transitory computer-readable storage media of claim 2, wherein the computing in each of the first iteration and the second iteration employs a convolution and a pooling.

Claim 4 depends on claim 2 and is not disclosed or suggested for the reasons set forth regarding claim 2.

Petitioner argues only that *Ujiie* teaches performing convolution and pooling. Pet. at 46. "However, that is only true for a single iteration shown in *Ujiie*." EX2012 at ¶149. But the combination of *Ujiie*, *Moons*, and *Kaul* does not disclose using a convolution to determine the initial maximum value. *Id*. "Instead, the combination relies on *Kaul's* two-bit sliding window hardware accelerator to identify the relevant value for pooling." Thus, this claim limitation requiring that "each of the first iteration and the second iteration employs a convolution and a pooling" is not met. *Id.* Instead, *Kaul* would locate the relevant value using simple hardware comparisons, and then "that value would be used for a single full-precision convolution and a single pooling operation." *Id.*

4. Claim 5

> **Claim 5**. The one or more non-transitory computer-readable storage media of claim 4, wherein the convolution is a N×N convolution, where N is any integer.

Claim 5 depends on claim 4 and is not disclosed or suggested for the reasons set forth regarding claim 4.

5. Claim 6

> **Claim 6**. The one or more non-transitory computer-readable storage media of claim 4, wherein the pooling is a N×N pooling, where N is any integer.

Claim 6 depends on claim 4 and is not disclosed or suggested for the reasons set forth regarding claim 4.

6. Claim 8

> **Claim 8**. The one or more non-transitory computer-readable storage media of claim 2, wherein at least one of the first iteration and the second iteration is performed with a precision less than that of the full precision computation.

Claim 8 depends on claim 2 and is not disclosed or suggested for the reasons set forth regarding claim 2.

Petitioner argues only that *Ujiie* performs "an approximate convolution 'AppConv' followed by an exact full precision convolution." Pet. at 48. "However, that is only true for a single iteration shown in *Ujiie*." EX2012 at ¶153. *Ujiie*, *Moons*, and *Kaul* do not rely on a convolution to determine the maximum value for a full-precision convolution. *Id.* "Instead, the combination relies on *Kaul's* two-bit sliding window hardware accelerator to identify the relevant value for pooling." *Id.* "That does not meet this claim's requirement of an iteration of convolution / pooling using less than all bits." *Id.* Instead, *Kaul* would locate the relevant value using

simple hardware comparisons, and then "that value would be used for a *full-precision* convolution and pooling operation." *Id.*

7. Claim 9

Claim 9. The one or more non-transitory computer-readable storage media of claim 8, wherein the precision is 8-bit precision.

Claim 9 depends on claim 8 and is not disclosed or suggested for the reasons set forth regarding claim 8.

8. Motivation To Combine

A POSITA would not have been motivated to further combine *Ujiie* and *Moons* with *Kaul*. EX2012 at 161.

As Dr. Medvidovic explains, "the combination of *Ujiie* and *Moons* with *Kaul* teaches away from the '831 patent, because they all require changes to the hardware architecture, which is not a requirement of the '831 patent." EX2012 at ¶156 (citing *Moons* at 1 and *Ujiie* at abstract (both referencing hardware requirements), and *Kaul* (disclosing a reconfigurable k-nearest-neighbor hardware accelerator)).

Further, a POSITA would not have been motivated to make the asserted combination because *Kaul* does not address finding the maximum value needed for the convolution. *Id.* at ¶157. "Instead, *Kaul* performs a k-nearest-neighbor search which does not rely on finding a maximum value." *Id.* (citing *Kaul* at 1 ("the sort network finds the minimum within current accuracy bounds, eliminating vectors outside these bounds for up to 6×compute reduction;" "the precise signals indicate

whether all other psums exceed the minimum by more than 1, signifying that the same minimum has been found as with full-precision calculations;" and "the compare psum circuit uses the global minimum to determine vector elimination and nearest neighbor selection.")). *Kaul* pertains to finding a minimum value, and neither the Petitioner nor Dr. Saber explain how the *maximum* convolution value problem faced by *Ujiie* and *Moons* could be solved by *Kaul's* minimum distance hardware accelerator. *Id.*

Further, Petitioner argues that *Ujiie* and *Moons* could be modified by inserting *Kaul's* hardware accelerator as shown in an annotated version of *Ujiie's* figure 4:

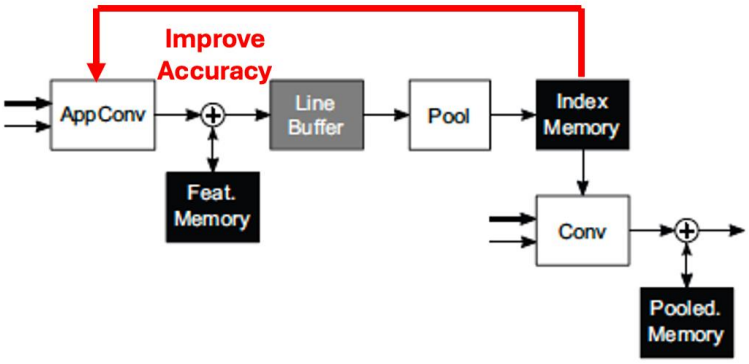


Figure 4: The proposed architecture for LCP-based CNN processor.

Pet. at 38 (citing *Ujiie*, Fig. 4, and indicating addition of *Kaul's* hardware accelerator in red). However, Dr. Medvidovic explains that this "hypothetical addition of *Kaul's* hardware accelerator is unrealistic and anything but straightforward." EX2012 at ¶159. "Significant hardware changes would be required," because the specialized hardware taught by *Ujiie* and *Moons* could not simply use *Kaul's* accelerator. *Id.*

"Instead, the three hardware systems would have to be merged, because Petitioner argues that a number of steps occur *between iterations* in *Kaul's* 2-bit sliding window." *Id.* "There is a stark difference between simply plugging in a hardware accelerator to accomplish the task it was designed for (determine the minimum distance neighbor) and modifying the hardware architecture to accomplish several additional steps within each iteration." *Id.* This would have been a non-trivial process which a POSITA would not have been motivated to undertake without clear motivation or understanding of potential benefits. *Id*. No such motivation is provided by any of the asserted references. Petitioner's argument relies on the blueprint of the '831 patent itself for motivation.

Finally, *Moons* and *Kaul* both aim at the same purpose. EX2012 at ¶160. "*Moons* seeks to identify the most relevant value (the maximum) for a full-precision convolution by using approximate convolutions on a scaled and quantized dataset." *Id.* (citing *Moons* at 3). "Similarly, *Kaul* aims at identifying the most relevant value (the minimum) for further processing." *Id.* A POSITA would not be motivated to implement both methods in *Ujiie* to accomplish the same goal.

C. GROUND 3: Obviousness in Light of *Ujiie*, *Moons*, *Kaul*, and *Ishii*

Petitioner asserts that claims 2–6, 8, and 9 are obvious in light of *Ujiie*, *Moons*, *Kaul*, and *Ishii*. However, Petitioner only references *Ishii* regarding claims 2 and 3.

No argument relying on *Ishii* is provided regarding claims 4–6, 8, and 9. Patent Owner therefore relies on its analysis provided in ground 2 regarding these claims.

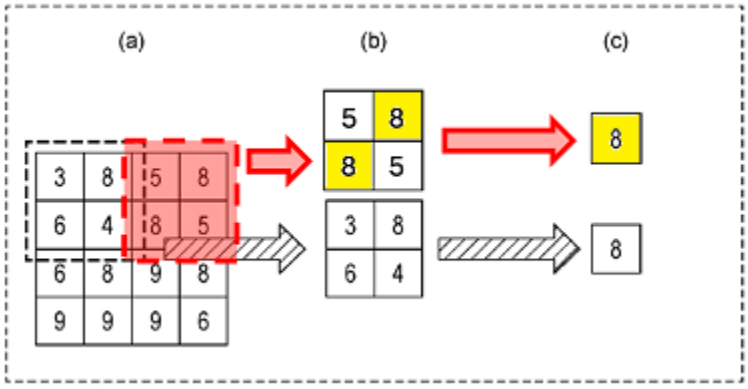
1. Claim 2

> **Claim 2**. The one or more non-transitory computer-readable storage media of claim 1, further comprising: responsive to determining that the first set of values are the same, performing, by the at least one processor, a second iteration that includes computing the value based on a second set of MSBs for each of the plurality of data sets, the second set of MSBs being larger than the first set of MSBs.

Claim 2 depends on claim 1 and is not disclosed or suggested for the reasons set forth regarding claim 1 in ground 1 above, and regarding claim 2 in ground 2 above.

Petitioner admits that "*Ujiie* does not expressly describe checking if the output of its max pooling process identifies a tie." Pet. at 40. Petitioner then notes that "*Ishii* demonstrates a scenario in which the results of max-pooling include a tie between two values." Pet. at 40. However, Petitioner further admits that, in the event of a tie, "*Ishii* simply *arbitrarily* selects one of the identified tied values." EX2012 at ¶163 (citing Pet at 42 ("in a case where there are two or more maximum values, the output selecting unit 124 *need just select any one of the two or more maximum values* of the processing results.")) Therefore, "*Ishii* does not teach using more MSBs of each of the tied values—or in *Ishii's* case, a higher resolution version of the image—to identify the true maximum of the two or more tied values." *Id*.

Instead, "in the event of max-pooling, if two or more values are the same, the system will proceed with one of those values." Pet. at 11.



Id. (citing *Ishii*, Fig. 5, annotated in red) (further annotated in yellow)). "Petitioner indicates the subject 2x2 feature map in red, which shows that the maximum value '8' appears twice." *Id*. "*Ishii* then simply uses the value '8' (indicated in yellow) to proceed, without another iteration using more MSBs or a higher resolution to determine which of the two values is the true maximum." *Id.* This algorithm differs significantly from and does not teach or suggest the claimed invention, which requires further evaluation to determine which of the tied values is the true maximum. "A POSITA would have known that *Ishii's* arbitrary selection risks losing accuracy, in the event that the value chosen is not the true maximum. In the above two-value-tie, there is a 50% chance that the wrong value is chosen and accuracy is lost." *Id.*

2. Claim 3

Claim 3. The one or more non-transitory computer-readable storage media of claim 2, further storing instructions, which, when executed by the

at least one processor, cause the at least one processor to perform operations comprising:

Claim 3 depends on claim 2 and is not disclosed or suggested for the reasons set forth regarding claim 2.

b. responsive to identifying the maximum value among the second set of values, performing, by the at least one processor, the full precision computation of the value for a data set, of the plurality of data sets, that exhibited the maximum value in the second iteration.

Petitioner argues that "*Ujiie-Moons-Kaul-Ishii* teaches performing a full precision computation on the dataset that exhibited the maximum value." Pet. at 45. However, as explained regarding claim 2, "*Ishii* does not disclose determining the true maximum." EX2012 at ¶165. "Instead, in case of a tie, it simply chooses an arbitrary value." *Id.*; *see also* subsection 3 below (Motivation To Combine).

3. Motivation To Combine

A POSITA would not have been motivated to further combine *Ujiie*, *Moons*, and *Kaul* with *Ishii*, because "*Ishii* does not determine the true maximum value as explained regarding claim 2." EX2012 at ¶166. Instead, in a tie, *Ishii* simply chooses one of the identified values. *Id.* Thus, *Ishii* may arbitrarily choose a value other than the true maximum. This would not work for *Ujiie*. "As Petitioner notes, "*Ujiie's* native prediction strategy requires identifying a maximum." *Id.* (citing Pet. at 39). *Ishii's* arbitrary selection of one of the tied values falls short, and Petitioner admits that it actually teaches that full accuracy is not needed. *See* Pet at 42 ("in a

case where there are two or more maximum values, the output selecting unit 124 *need just select any one of the two or more maximum values* of the processing results."). Thus, "a POSITA would not have been motivated to use *Ishii* to meet *Ujiie's* true maximum requirement." *Id.*

D. GROUND 4: Obviousness in Light of *Ujiie*, *Moons*, *Kaul*, and *Ravindran*

Petitioner asserts that claim 7 is obvious in light of *Ujiie*, *Moons*, *Kaul*, and *Ravindran*.

1. Claim 7

> **Claim 7**. The one or more non-transitory computer-readable storage media of claim 4, wherein the convolution is a 3×3 convolution, and the pooling is a 2×2 pooling.

Claim 7 depends on claim 4 and is not disclosed or suggested for the reasons set forth regarding claim 4.

2. Motivation To Combine

A POSITA would not have been motivated to further combine *Ujiie*, *Moons*, and *Kaul* with *Ravindran* because "*Ravindran* teaches using an ensemble of CNNs to generate predictions, which are then 'aggregated to accurately classify the objects in the digital image.'" EX2012 at ¶169. While *Ujiie*, *Moons*, and *Kaul* are all concerned with power reduction, *Ravindran* does not value that concern, and makes no reference to power consumption. *Id.* To the contrary, "*Ravindran* aims to use 'ensemble of deep learning models, such as convolutional neural networks (CNNs),

to accurately and automatically perform image processing to detect particular attributes of objects in a digital image, and to classify the objects according to the detected attributes.'" *Id.* (citing *Ravindran* at ¶[0010] (disclosing "a method for building and fine tuning CNNs proven to output an accurate classification of an image.")). Unlike *Ujiie*, *Moons*, and *Kaul*, which seek to reduce power requirements, a POSITA would have known that "*Ravindran's* teaching of using multiple CNNs will increase power consumption in order to gain accuracy." *Id.* "Indeed, *Ravindran* explicitly aims to generate 'an ensemble of the most accurate intermediate CNNs' which may be 'the top 5 most accurate intermediary CNNs.'" *Id.* (citing *Ravindran* at ¶[0010]; *see also* [0012] (*Ravindran* aims to "exhibit[] superior classification accuracy.")). "This is the opposite goal of *Ujiie*, *Moons*, and *Kaul*." *Id.* Since each of *Ujiie*, *Moons*, and *Kaul* stress the importance of energy savings, a POSITA would not be motivated to use *Ravindran's* "ensemble of CNNs" which would drastically increase power requirements. *Id*. (citing *Ravindran* at abstract).

A. GROUND 5: Obviousness in Light of *Ujiie*, *Moons*, *Kaul*, *Ishii*, and *Ravindran*

Petitioner asserts that claim 7 is obvious in light of *Ujiie*, *Moons*, *Kaul*, *Ishii*, and *Ravindran*. However, Petitioner does not reference *Ishii* regarding claim 7. Patent Owner therefore relies on its analysis provided in ground 4 regarding claim 7.

1. Claim 7

Claim 7. The one or more non-transitory computer-readable storage media of claim 4, wherein the convolution is a 3×3 convolution, and the pooling is a 2×2 pooling.

Claim 7 depends on claim 4 and is not disclosed or suggested for the reasons set forth regarding claim 4.

Claim 7 is not taught or suggested by the combination of *Ujiie*, *Moons*, *Kaul*, and *Ravindran*. Petitioner makes no argument that *Ishii* overcomes this shortcoming.

2. Motivation To Combine

A POSITA would not have been motivated to further combine *Ujiie*, *Moons*, *Kaul*, and *Ishii* with *Ravindran* for the reasons set forth in ground 4 above. "*Ravindran* teaches away from the combination because it identifies accuracy as a paramount concern over power consumption." EX2012 at ¶173. Further, Dr. Medvidovic explains that "careful analysis of the five asserted references confirms that Petitioner is simply backfilling concepts to plug holes in its arguments *purely* due to hindsight based on the '831 patent." *Id.*

IX. SECONDARY CONSIDERATIONS OF NON-OBVIOUSNESS

Strong secondary considerations of non-obviousness support the '831 patent. These must be evaluated in assessing obviousness, to provide real-world insight into the novelty of the invention. *Graham v. John Deere Co. of Kan. City*, 383 U.S. 1,

17–18 (1966). These important factors counter hindsight bias and are particularly important in close cases. *Id*.

A. Commercial Success

Beginning in May 2015, [redacted] entered a sponsored research agreement specifically related to research by Prof. Jae-sun Seo that led to the '831 patent. EX2007 at ¶¶ 8–9. In exchange for [redacted] funding, ASU granted [redacted] a non-exclusive, paid-up, royalty-free license to University Foreground Intellectual Property for [redacted] internal research and development uses, including a limited right to sublicense to [redacted] affiliates, conditioned on [redacted] paying its pro rata share of patent expenses in each jurisdiction it elected to maintain. ASU also granted [redacted] a two-year option to negotiate an exclusive worldwide royalty-bearing license, during which [redacted] was required to bear all patent-prosecution costs. EX2007 at ¶10. This "pay upfront + internal-use NERF + option" structure reflects [redacted] substantial, at-risk investment made based on the commercial viability and success of the '831 patent. EX2007 at ¶11.

This evidence of commercial success is directly tied to the patented technology: [redacted] deemed the '831 patent's computationally efficient high accuracy CNNs commercially promising and relevant to its research.

B. Long-Felt Need and Industry Support

The ’831 patent addressed a long-felt need for effective computationally efficient high accuracy CNNs in a rapidly evolving technology landscape. EX2007 at ¶8–11. For that reason, Dr. Seo’s research was funded by a technology leader in the AI space heavily focused on mission-critical AI technology. *Id*. This support highlights the invention’s role in solving persistent challenges in CNN AI technology, where prior approaches lacked the high efficiency / high accuracy solution offered by the ’831 patent.

C. Failure Of Others

The prior art at the time failed to achieve the inventions of the ’831 patent, and others explicitly discouraged various concepts claimed in the ’831 patent. For example, *Ujiie* teaches away from the ’831 patent by discouraging using convolutions at all to determine a likely maximum value. *Ujiie* at 3. *Ujiie* and *Moons* also teach away from the ’831 patent because “they both require changes to the hardware architecture.” EX2012 at ¶134 (citing *Moons* at 1 and *Ujiie* at abstract (referencing hardware requirements)). The ’831 patent, on the other hand, teaches away from using specialized hardware, which is a critical difference. EX2012 at ¶124. Thus, others attempted to solve the problem but failed.

D. Unexpected Results

Ujiie and *Moons* teach away from various aspects of the invention, confirming that the results were surprising and not predicted by others in the field.

E. Skepticism of Experts

The same teaching away confirms that leading experts in the field doubted the invention.

F. Praise By Others

The fact that Dr. Seo's research that led to the '831 patent was funded by [redacted]—one of the primary industry players in the area—supports that industry experts acknowledged the significance of the invention.

G. Copying By Competitors

Skysong's suits against CrowdStrike and Fortinet confirm that competitors have copied the invention, which suggests they recognized its value and that any non-infringing alternatives were not obvious to them.

X. CONCLUSION

The Board should deny institution because CrowdStrike has failed to demonstrate a reasonable likelihood of prevailing as to any challenged claim.

Dated: January 5, 2026

Respectfully submitted,

/s/ Jay P. Kesan

Jay P. Kesan (Reg. No. 37,488)

KEY KESAN DALLMANN PLLC
1050 Connecticut Avenue, N.W. Suite 500 Washington, D.C. 20036
jay.kesan@kkd-law.com
Telephone: (202) 772-1100

Alfonso G. Chan
Andrew Cochran
Cahill Gordon & Reindel LLP
900 16th Street, N.W., Suite 500
Washington, D.C. 20006
Telephone: (202) 862-8942

Attorneys for Patent Owner
SKYSONG INNOVATIONS, LLC

CERTIFICATE OF SERVICE

Pursuant to 37 C.F.R. §§ 42.6(e)(4) and 42.25(b), the undersigned certifies that on January 5, 2026, a complete copy of Patent Owner's Preliminary Response was filed electronically through the Patent Trial and Appeal Board's PTABE2E System and provided, via electronic service, to the Petitioner by serving the correspondence address of record.

Dated: January 5, 2026

/s/ Jay P. Kesan
Jay P. Kesan

CERTIFICATE OF WORD COUNT

Pursuant to 37 C.F.R. § 42.24, the undersigned certifies that the foregoing Patent Owner's Preliminary Response contains less than 14,000 words excluding the table of contents, table of authorities, certificate of service, certificate of word count, and appendix of exhibits. Patent Owner has relied on the word count feature of the word processing software used to create this paper in making this certification.

Dated: January 5, 2026

/s/ Jay P. Kesan
Jay P. Kesan