IN THE UNITED STATES DISTRICT COURT
FOR THE DISTRICT OF MASSACHUSETTS

SINGULAR COMPUTING LLC, Plaintiff, v. GOOGLE LLC, Defendant.	C.A. No. 1:19-cv-12551-FDS Hon. F. Dennis Saylor IV

DEFENDANT GOOGLE LLC'S TRIAL BRIEF

TABLE OF CONTENTS

Page

I. INTRODUCTION 1

II. FACTUAL BACKGROUND 2

A. Accused Technology Terminology 2

B. Google's TPU Systems and Chips 4

C. Approximate Computing 8

D. The '273 and '156 Patents 9

III. PROCEDURAL BACKGROUND 11

IV. ARGUMENT 12

A. Google does not infringe the asserted claims. 12

1. Google's TPUs do not meet the "first operation" limitation. 12

2. Google's TPUs do not meet the "relative error" limitation. 13

3. Google's TPUs do not meet the "exceeds" limitation. 14

B. Google did not willfully infringe the asserted claims. 15

C. The asserted claims are invalid as anticipated and/or obvious. 16

D. Singular's damages request is massively inflated and untethered to the asserted claims. 17

V. CONCLUSION 20

TABLE OF AUTHORITIES

Page(s)

Federal Cases

Arctic Cat Inc. v. Bombardier Rec. Prods.,
876 F.3d 1350 (Fed. Cir. 2017).....19

Bayer HealthCare LLC v. Baxalta Inc.,
989 F.3d 964 (Fed. Cir. 2021).....15

Ericsson, Inc. v. D-Link Sys., Inc.,
773 F.3d 1201 (Fed. Cir. 2014).....17

Google LLC v. Singular Computing LLC,
No. 22-1866 (Fed. Cir.).....12

Silicon Graphics, Inc. v. ATI Techs., Inc.,
607 F.3d 784 (Fed. Cir. 2010).....3

VLSI Tech. LLC v. Intel Corp.,
87 F.4th 1332 (Fed. Cir. 2023) 17

Federal Statutes

35 U.S.C. § 101.....12

35 U.S.C. § 102.....16

35 U.S.C. § 271.....17

35 U.S.C. § 284.....17

35 U.S.C. § 287.....19

In advance of trial scheduled to commence on January 8, 2024, and pursuant to Local Rule 16.5(f), Defendant Google LLC ("Google") hereby submits its trial brief.

I. INTRODUCTION

Plaintiff Singular Computing LLC ("Singular") filed this patent infringement action in December 2019, targeting two versions of Google's Tensor Processing Units ("TPUs"). TPUs are Google-designed, application-specific integrated circuits that accelerate certain computing tasks used in machine learning, a subfield of artificial intelligence. After Google succeeded in invalidating all of the asserted claims of one of Singular's patents and many of Singular's asserted claims from the remaining patents via *inter partes* review, Singular has only two asserted claims remaining. Each of these remaining claims is related to computing using at least a certain number of "low precision high dynamic range (LPHDR) execution unit[s]."

The evidence adduced at trial will show that Google did not infringe the asserted claims for three reasons. *First*, both of Singular's asserted claims require computers that perform *incorrect* arithmetic on a certain percentage of inputs, but Google's TPUs generate *correct* results every time. The TPUs accordingly fail the asserted claims' "relative-error" limitation: that a first operation on a first input signal result in an output with 0.05% error for at least 5% of valid inputs. *Second*, while the claims require execution units that are adapted to execute a "*first operation*" on data, Singular alleges that Google's TPUs infringe the asserted claims based on *two distinct operations* that occur in separate parts of the TPU—rounding in one unit followed by multiplication in an entirely separate unit. Both because Google's TPUs do not meet the relative error limitation and, separately, because Singular's infringement allegations focus on two distinct operations rather than a single "first" operation, the TPUs do not meet requirement of having an "LPHDR execution unit." And *third*, the supposed LPHDR execution units Singular points to do not exceed the number of 32-bit execution units by at least 100, as required by the asserted claims.

Confronted with these arguments at trial, Singular will be unable to show infringement by Google, much less willful infringement. Singular's asserted claims are also invalid as anticipated and/or obvious based on the prior public use, public knowledge, and prior invention of a system for highly parallel low-precision, high-dynamic range computing, which was built using a code library called VFLOAT that the Patent Office did not previously consider.

Should a jury return a verdict of infringement and no invalidity, Singular is not entitled to its inflated request for reasonable royalty damages of up to $7.01 billion. This staggering amount, given Singular's expert Philip Green's opinions, necessarily wouldn't be tied to the smallest salable patent-practicing unit—i.e., functionality within the TPU *chip*, rather than the TPU *system*. Moreover, Singular's damages request has no basis since Google could have, among other things, modified its TPUs at minimal expense to avoid infringement, even under Singular's theory. As Google's damages expert Laura Stamm will explain, Singular's damages request is nonsensical.

Google looks forward to presenting its case to a jury and demonstrating that the evidence requires a verdict in its favor.

II. FACTUAL BACKGROUND

A. Accused Technology Terminology

This case concerns Google technology used for machine learning applications. Machine learning, a subfield of artificial intelligence, is the process through which a program or system trains a computer model, often called a neural network ("NN") model, from a given set of input data ("training"). A trained model can then make useful predictions about new, never-before-seen data that resembles the data used to train the model (a process known as "inference"). As a simple example, you could ***train*** a NN model by feeding it photos of cats and telling the model that the photos contain cats. The trained model could then be used for ***inference*** to identify a cat in new photos that it had never seen before.

Computers utilize different number formats. These include floating point formats, which operate similarly to scientific notation, a common method of representing very large or small numbers:

> In floating point format, data is represented by the product of a fraction, or mantissa, and a number raised to an exponent. For example, a number *n* can be represented in base 10 by $n = m \times 10^e$, where *m* is the mantissa and *e* is the exponent. If *m* equals 2 and *e* equals 1, *n* equals 20; if *m* equals 2 and *e* equals -1, then *n* equals 0.2.

Silicon Graphics, Inc. v. ATI Techs., Inc., 607 F.3d 784, 786 (Fed. Cir. 2010) (internal quotations and citations omitted).

Floating-point number formats, which have been used in computers for at least the last 60 years, assign a certain number of bits to represent the format's different numerical components: the fraction (or mantissa), exponent, and a sign (i.e., plus or minus). All floating-point number formats have a specified precision[1] level, which is the number of digits in the mantissa (*m*). The range of floating-point numbers is determined by the number of exponent (*e*) bits in the number format. The table below shows the relevant number formats in this case, floating-point 16 ("fp16"), floating-point 32 ("fp32"), brain-float floating point 16 ("bfloat16" or "bf16") and brain-float floating point 20 ("bfloat20" or "bf20"):

[1] "Precision" here has a different meaning than the way the Court has construed the term in its claim construction order.

	Sign Bits	**Exponent Bits (*impacts range*)**	**Mantissa Bits[2] (*impacts precision*)**	**Total Bits (Sign + Exponent + Mantissa)**
fp16	1	5	10 or 11	16 or 17
fp32	1	8	23 or 24	32 or 33
bfloat16	1	8	7 or 8	16 or 17
bfloat20	1	8	11 or 12	20 or 21

B. Google's TPU Systems and Chips

In 2013, Google began developing a single-purpose, high performance computing system for machine learning applications. In designing the chip that forms the basis of this system, Google incorporated a number of different features to maximize its computing speed. For example, Google chose to use a systolic array design (a well-known and highly efficient chip design where data is passed from one unit in the array to the next at a regular interval, rather than each unit having to do a more resource-intensive "fetch" of data from memory).

The design process began in November 2013, when Google formed a TPU hardware team to start development of the first version of the TPU. Google hired Norm Jouppi, a hardware engineer with over 30 years of industry experience at the Digital Equipment Corporation, Compaq Computer, and Hewlett-Packard, to provide technical leadership and pulled together a team of senior and experienced engineers to work on the project. This multi-year effort first led to TPUv1

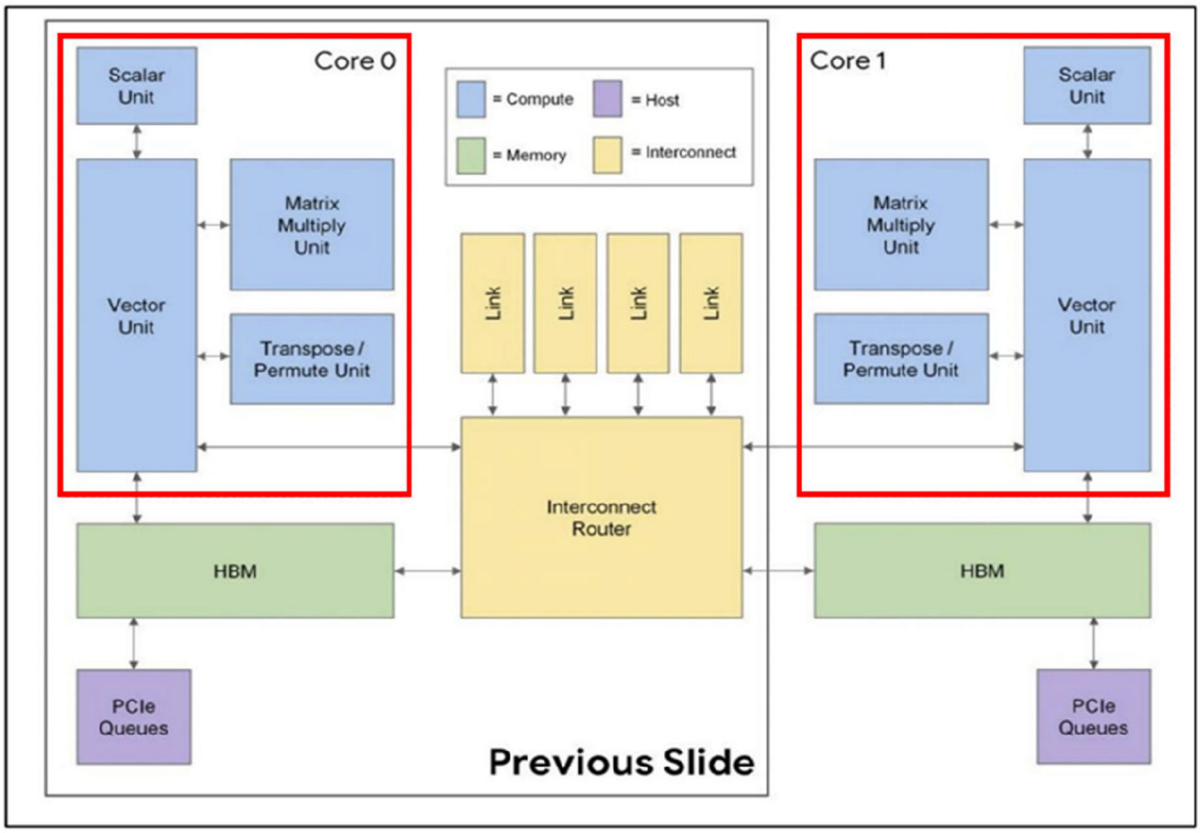
[2] The floating-point number formats that are commonly used in computers include a "hidden bit," and persons of skill in the art count the number of bits in the mantissa both with and without the hidden bit. The hidden bit is the value to the left of the "radix" point (the point referred to as the decimal point in commonly used base 10 numbers) in the mantissa portion of the number. That hidden bit is always "1" and thus is not stored, but is used when, for example, a multiplication is performed using the number.

("SeaStar"), which is not accused of infringement in this case. Because inference involves simpler mathematics than does training, the team designed TPUv1 to perform inference only.

Google began developing TPUv2 in December 2014. Unlike TPUv1, Google sought to develop TPUv2 and later versions of the TPU to perform both training and inference. On their own, TPUv2 and v3 chips are not particularly fast, performing at similar speeds as GPU chips. But Google also incorporated interconnects, which allows chips within a system to communicate directly with each other, thereby allowing Google to massively scale its computing power by linking multiple chips together. When TPUv2 and v3 chips are networked together, TPUv2 and v3 systems perform substantially better than a traditional computer at machine learning tasks.

Singular accuses TPUv2 and TPUv3 chips of infringement in this case. The two accused TPU chips consist of multiple non-accused components and software. Importantly, Singular's infringement allegations only target functionality specific to certain sub-components of the TPU *chips* contained within the larger TPU systems.

As shown in the following diagram, each TPUv2 chip contains two "Tensor Cores" (labeled Core 0 and Core 1 with sub-components shown in blue below):



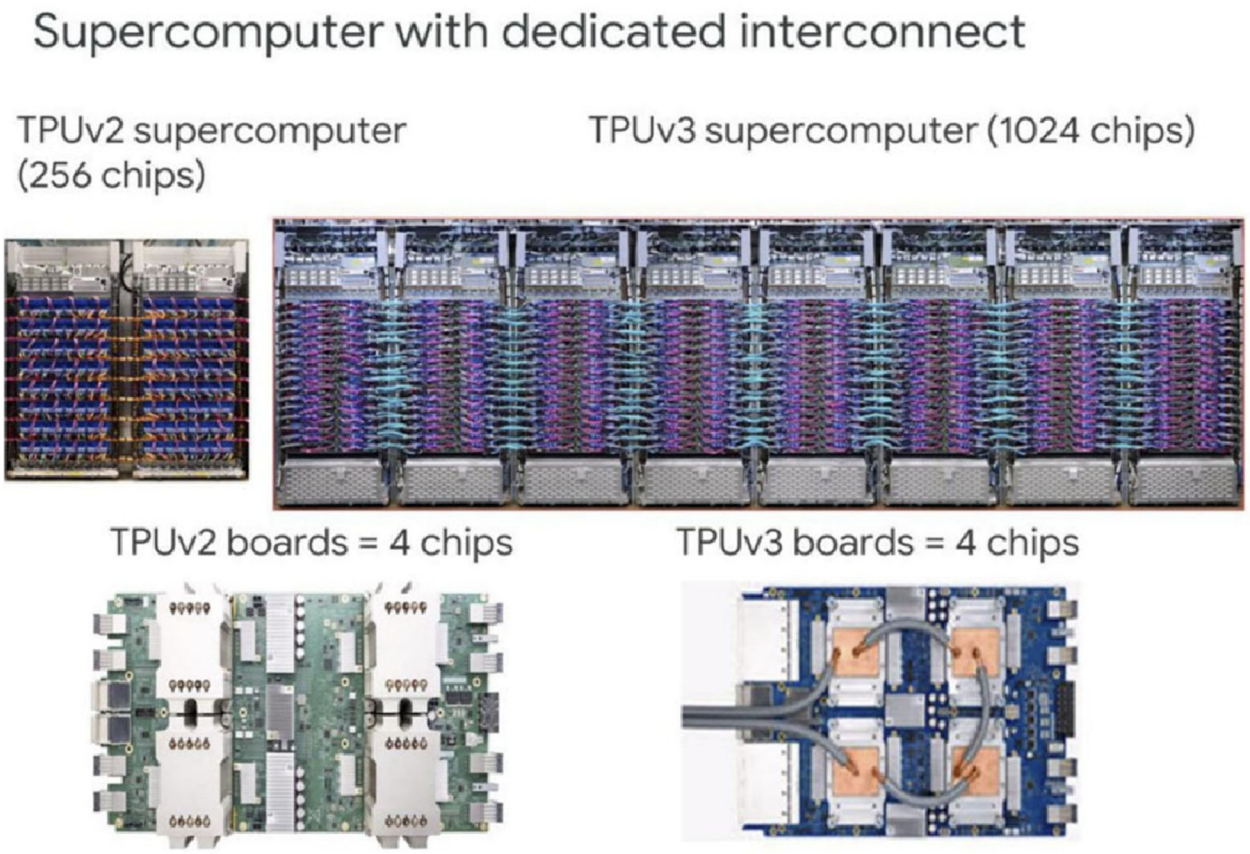
Dkt. 477 at 2 (red boxes added for emphasis).

Each Tensor Core itself contains several distinct units, including the Vector Unit (also called the "Vector Processing Unit" or "VPU"), a Matrix Multiply Unit ("MXU"), and a core sequencer. There is one VPU per Tensor Core. Each VPU contains 256 float conversion units (also referred to as "rounding circuits"), each of which can convert a fp32 value to a bfloat16 value using rounding. This rounding operation produces the same result, every time, given the same fp32 input. Each TPUv2 board contains four integrated circuits ("chips") known as "Jellyfish Chips" ("JFCs") that are attached to the TPU board. There are 2,048 rounding circuits on each TPUv2 board overall – 256 rounding circuits per VPU X 1 VPU per Tensor Core X 2 Tensor Cores per TPUv2 X 4 JFCs per TPUv2 board = 2,048 rounding circuits.

There is one MXU per Tensor Core in TPUv2. The MXU contains circuits organized into a systolic array and that perform arithmetic to carry out "matrix multiplication." The inputs are the result of the VPU's rounding operation, and then are sent through several other stages before reaching the MXU. The matrix multiplication includes both multiplication and addition. The multiplication occurs correctly and without error; in other words, Google's MXUs do correct multiplication, with the same result obtained from the same inputs every time. This is accomplished by having more bits in the multiplication's output than its input. Each TPUv2 board overall contains eight MXUs – 1 MXU per Tensor Core X 2 Tensor Cores per JFC X 4 JFCs per TPUv2 board = 8 MXUs.

The TPUv3 board is structurally similar to the TPUv2 board, with two important differences. The four integrated circuits on a TPUv3 board are known as "Dragonfish Chips" ("DFCs"). Like the TPUv2, each DFC in TPUv3 contains two Tensor Cores, but each has twice as many MXUs and twice as many rounding circuits as a TPUv2. There are thus sixteen MXUs and 4,096 rounding circuits on a TPUv3 board.

To maximize computational speed, Google designed its TPU systems to be networked so that multiple chips could work in conjunction to train a machine learning model. Four TPU chips are attached to the TPU board, and Google can combine those TPU boards to form scaled TPU pods with many TPU chips:



Dkt. 477 at 3.

To facilitate that level of scaling, each TPU chip and system contains many components beyond the MXU and VPU, including the Google-designed interconnects, interconnect routers, and high-bandwidth memory, that allow for high-speed inter-chip communication. The TPUs additionally use Google-designed software that further optimizes system-wide performance. There is no dispute that the interconnects, high-bandwidth memory, and software—along with many other aspects of both TPU chips and systems—are unrelated to the asserted claims and reflect Google's non-accused contributions to the TPU system and chips.

C. Approximate Computing

Dr. Bates's patents fall within the realm of approximate computing, a method of computing that generates incorrect and inconsistent results as a tradeoff for greater computational efficiency.

Dr. Bates claims to have invented a computer design that implements approximate computing and generates unusually wrong answers unusually often. For example, as Dr. Bates himself describes, using the system to repeatedly calculate 1 + 1 could generate different answers somewhere within the range of 1.98 to 2.02 each time the query was performed. Dr. Bates did not invent approximate computing, and his ideas were not new. In fact, as Dr. Bates says, the computer design that he proposed is based on a familiar and simple architecture in the history of computing that was studied in the 1980s and is still in use in certain of today's modern processors.

D. The '273 and '156 Patents

Singular accuses both the TPUv2 and TPUv3 chips of infringing two claims: (1) dependent claim 53 of U.S. Patent No. 8,407,273 (the "'273 patent"); and (2) dependent claim 7 of U.S. Patent No. 9,218,156 (the "'156 patent"). Claim 53 of the '273 patent depends from (i.e., incorporates the limitations of) claim 43, which in turn depends from independent claim 36. Claim 7 of the '156 patent depends from claim 3, which in turn depends from claim 2, which in turn depends from independent claim 1. With the limitations of the earlier claims from which they depend included, the asserted claims read:

Claim 53 of the '273 patent:

36. A device:

> comprising ***at least one first low precision high-dynamic range (LPHDR) execution unit*** adapted to execute a ***first operation on a first input signal*** representing a first numerical value to produce a first output signal representing a second numerical value,
>
> wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/65,000 through 65,000 and for ***at least X=5% of the possible valid inputs to the first operation***, the statistical mean, over repeated execution of the first operation on each specific input from the at least X % of the possible valid inputs to the first operation, of ***the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input***;

wherein the number of LPHDR execution units in the device exceeds the nonnegative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.

43. The device of claim 36, ***wherein the number of LPHDR execution units in the device exceeds by at least one hundred*** the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.

53. The device of claim 43, wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 to 1,000,000.

Claim 7 of the '156 patent:

1. A device comprising:

at least one first low precision high dynamic range (LPHDR) execution unit adapted to execute a ***first operation on a first input signal*** representing a first numerical value to produce a first output signal representing a second numerical value,

wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/65,000 through 65,000 and ***for at least X=5% of the possible valid inputs to the first operation***, the statistical mean, over repeated execution of the first operation on each specific input from the at least X% of the possible valid inputs to the first operation, of ***the numerical values represented by the first output signal of the LPHDR unit executing the first operation on that input differs by at least Y=0.05% from the result of an exact mathematical calculation of the first operation on the numerical values of that same input***; and

at least one first computing device adapted to control the operation of the at least one first LPHDR execution unit.

2. The method of claim 1, wherein the at least one first computing device comprises at least one of a central processing unit (CPU), a graphics processing unit (GPU), a field programmable gate array (FPGA), a microcode-based processor, a hardware sequencer, and a state machine.

3. The device of claim 2, ***wherein the number of LPHDR execution units in the device exceeds by at least one hundred*** the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide.

7. The device of claim 3, wherein the dynamic range of the possible valid inputs to the first operation is at least as wide as from 1/1,000,000 through 1,000,000.

(emphases added).

Both asserted claims cover a computing device that contains "execution units" that have two identifying features: (a) they perform arithmetic operations on numerical inputs that can be both very large and very small (i.e., high dynamic range); and (b) they perform the arithmetic operations in such a way as to introduce some minimum amount of error for some minimum percentage of valid inputs. The patents call these components low precision high dynamic range execution units ("LPHDR execution units").

The asserted claims have three limitations particularly relevant in this case. The LPHDR execution units must be (1) adapted to execute a "first operation" on their inputs (the "first operation" limitation), (2) with a certain amount of error compared to an "exact mathematical calculation" of that first operation on the same inputs (the "relative error" limitation), and there must be (3) at least 100 more LPHDR execution units than execution units "adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide" (the "exceeds" limitation).

Put simply, the asserted claims encompass a device containing a certain number of LPHDR execution units that are able to perform calculations using very big numbers or very small numbers, and when performing those calculations gets the math a certain amount wrong for a certain percentage of the valid inputs.

III. PROCEDURAL BACKGROUND

Singular filed its Complaint in this case in December 2019 and the operative First Amended Complaint in March 2020. *See* Dkts. 1, 37. Google filed IPR petitions against Singular's patents in the fall of 2020, and the case was stayed in summer 2021 pending completion of the IPR proceedings. The IPR trials resulted in invalidation of the majority of the asserted claims in Singular's patents. Singular now accuses Google of infringing only two claims, one in each remaining patent. Although those claims survived the IPR proceedings, they are the subject of a

pending appeal by Google that the Federal Circuit is scheduled to hear on January 9, 2024, the second day of trial. *See Google LLC v. Singular Computing LLC*, No. 22-1866 (Fed. Cir.), Dkt. 72. Singular is no longer appealing the IPR decisions invalidating the majority of the claims of the patents, making those decisions final. *See id.*, Dkt. 40 (dismissing cross-appeals).

After the PTAB issued its decisions in May 2022, proceedings in this Court resumed. The Court issued its *Markman* order in July 2022, *see* Dkt. 354, and the parties completed expert discovery in April 2023. In July 2023, the Court granted Plaintiff's motion for partial summary judgment of no invalidity under 35 U.S.C. § 101 and denied the parties' other motions for summary judgment. *See* Dkt. 551.

Each of the parties has submitted a proposed set of jury instructions, a verdict form, and voir dire questions. The parties have also jointly filed a pre-trial memorandum with exhibit and witness lists.

IV. ARGUMENT

Google has several defenses to Singular's claims. The evidence adduced at trial will show that Google has not infringed the patents, much less willfully, and that the claims are invalid as anticipated and/or obvious.

A. Google does not infringe the asserted claims.

Google's TPUv2 and TPUv3 do not infringe Singular's patents because the TPUs do not meet three of the limitations in the claims.

1. Google's TPUs do not meet the "first operation" limitation.

The first operation limitation requires that the LPHDR execution units be "adapted to execute a first operation on a first input signal representing a first numerical value to produce a first output signal representing a second numerical value." '156 patent, claims 1, 2, 3 & 7; '273 patent, claims 36, 43 & 53. In arguing that the TPUs infringe the asserted claims, Singular's

technical expert, Dr. Khatri, posits a "two-stage operation": (1) a round-to-nearest-even operation in the VPU, followed by (2) a multiplication operation in the MXU. But these are two separate, distinct operations that occur in physically distinct units on the TPU. Operations that occur in two distinct units of the TPU cannot be considered a single "first operation." Because what Singular accuses in Google's TPU is two distinct operations rather than a single "first" operation, it has not and cannot show that the TPUs meet any of the "first operation" limitations.

2. Google's TPUs do not meet the "relative error" limitation.

The relative error limitation requires that there be at least 0.05% error in the output from the LPHDR execution units executing the "first operation," for at least 5% of the valid inputs, relative to the results of an exact mathematical calculation. '156 patent, claim 7; '273 patent, claim 53. But whether you look at each operation separately or accept Dr. Khatri's "two-stage operation" theory, the Google TPUs do not meet this requirement because they perform correct and consistent math for each operation, with no error.

Looking at each operation separately, first the rounders in the VPU round a fp32 value to a bfloat16 value. But this rounding occurs through a mathematically correct "round-to-nearest-even" operation that produces the same result, every time, for any given input. In the second operation, a multiplication operation occurs in the MXU. Dr. Khatri has already conceded that the operation multiplies the bfloat16 values "without further loss of precision," such that the output is mathematically correct. Thus the two separate operations occur correctly and without error.

In Dr. Khatri's "two-stage operation" theory, he opines there is error by comparing (1) a one-stage exact multiplication calculation with (2) his two-stage first operation of the TPU. But this comparison goes against the plain language of the claim, which requires comparing the output of the LPHDR execution unit to an exact mathematical calculation of the *same* "first operation" that the LPHDR execution unit is adapted to perform. In other words, Dr. Khatri is improperly

making an apples-to-oranges comparison of the results of a *two*-stage first operation to a *one*-stage exact multiplication operation. Dr. Khatri's theory thus ignores the plain language of the claims.

Google's TPUs accordingly cannot satisfy the relative-error limitation.

3. Google's TPUs do not meet the "exceeds" limitation.

The exceeds limitation requires that "the number of LPHDR execution units in the device exceeds by at least one hundred the non-negative integer number of execution units in the device adapted to execute at least the operation of multiplication on floating point numbers that are at least 32 bits wide." '156 patent, claim 7; '273 patent, claim 53. Under Dr. Khatri's theory, this means that the TPU boards must include at least 8,300 LPHDR execution units.

The TPUs do not meet this requirement. As explained above, each TPUv2 board has 2,048 rounders in its VPUs, and each TPUv3 board has 4,096 rounders. Accepting Dr. Khatri's own theory that each LPHDR execution unit has two VPU rounders, that means that there are at most 1,024 LPHDR execution units per TPUv2 board, and 2,048 per TPUv3 board. Neither 1,024 nor 2,048 is greater than 8,300. Thus, the TPUs do not meet the exceeds limitation.

Dr. Khatri opines that the TPUs meet this limitation only through a "unique pairs" theory that relies on counting each physical rounding circuit in the VPU *128 times*—massively inflating the purported number of LPHDR execution units in each TPU board. No person of ordinary skill in the art, faced with the requirement that processing elements be counted, would accept that a single circuit can be counted 128 times to make up the required number of LPHDR execution units.

With Dr. Khatri's approach foreclosed, that leaves the truism that neither 1,024 nor 2,048 is greater than 8,300. The TPUs thus do not meet the "exceeds" limitation.

* * *

Because Google's TPUs do not meet three distinct limitations in the asserted claims (much less all of them), the TPUs do not infringe Singular's patents.

B. Google did not willfully infringe the asserted claims.

Singular contends that Google has and continues to willfully infringe the asserted claims. To prove that, Singular must "show [that Google] had a specific intent to infringe at the time of the challenged conduct," and Google must both have had knowledge of the asserted patents and have deliberately or intentionally infringed them. *Bayer HealthCare LLC v. Baxalta Inc.*, 989 F.3d 964, 987 (Fed. Cir. 2021). Singular seeks to offer three sets of communications between Dr. Bates and individuals at X and Google to purportedly show Google's knowledge of and intent to infringe Singular's patents. Those communications show no such thing.

The first set of communications between Dr. Bates and X—a division of Google distinct from the division developing TPUs—occurred from November 2010 through July 2011 and cannot support a willfulness finding. Neither of the asserted patents had issued at the time, so these communications obviously cannot show knowledge of the patents, and thus do not support willfulness. If anything, they show that Google engineers expressed concerns about the limited usefulness of computing that is based on math that is frequently and unpredictably wrong.

The second set of communications occurred in July 2013 and January 2014. While these communications occurred after Dr. Bates's first patent issued and some of the presentations generically mention "patents," he never specifically identified the issued patent in any of the correspondence. Messages, such as these, that fail to "specifically identify" the asserted patents cannot support a claim of willfulness. Again, if anything, these communications show that Google ***wasn't*** interested in Singular's technology—Google engineers had mixed feelings about the

technology and the company ultimately rejected it, explicitly telling Dr. Bates that his idea was not right for the type of applications Google was developing.

The final set of communications occurred in January 2017 and February 2017. These communications cannot support a finding of willfulness either. When Dr. Bates again reached out to X on January 5, 2017, the TPUv2 project had launched over two years prior and the chip and system were already in production, and TPUv3 had already been in development for months. At most, Singular could point to the fact that counsel at X became aware of Singular's patents in February 2017. But knowledge of the patents alone is insufficient to prove willful infringement. Singular will not be able to adduce any evidence at trial to make the additional required showing that Google believed its TPUs infringed the patents. These communications thus cannot support a finding of willful infringement either.

C. The asserted claims are invalid as anticipated and/or obvious.

Both claim 53 of the '273 patent and claim 7 of the '156 patent are also invalid as anticipated and/or as obvious in light of prior art. As Google's expert Dr. Miriam Leeser will explain at trial, in the early 2000s, she and one of her graduate students at Northeastern University developed a system capable of highly parallel low-precision, high-dynamic range computation (the "VFLOAT system"), which was in part based on a library of parameterized hardware modules they created for performing variable-precision arithmetic on floating-point numbers with custom formats named "VFLOAT." Dr. Leeser will explain how the VFLOAT system was reduced to practice, publicly known, and in public use by approximately 2002, and how the VFLOAT system anticipates and/or renders obvious the asserted claims under 35 U.S.C. §§ 102(a), 102(b), and 102(g)(2). For those reasons, Google will show at trial that the asserted patents are invalid.

D. Singular's damages request is massively inflated and untethered to the asserted claims.

Singular seeks damages of no less than a lump sum of between $1.63 and $5.19 billion,[3] based on the opinions of its expert, Philip Green, regarding the license and royalty rate that a hypothetical negotiation would have produced. As outlined above, Google does not infringe the asserted claims, so Singular is not entitled to any damages. To the extent either asserted claim is found to be infringed and not invalid, Google agrees that a lump-sum reasonable royalty is the appropriate measure of damages. 35 U.S.C. § 284. But not only does Mr. Green's multi-billion dollar range improperly invite the jury to speculate on what a reasonable royalty would be, his calculation of the reasonable royalty is off by several orders of magnitude.

First, the royalty base on which Mr. Green based his cost-savings damages opinions is overbroad. "Any reasonable royalty must seek to measure the value of the patented technology—it must be 'apportioned' to that value—by separating out and excluding other value in economic products or practices." *VLSI Tech. LLC v. Intel Corp.*, 87 F.4th 1332, 1345 (Fed. Cir. 2023). "[W]here multi-component products are involved . . . the ultimate combination of royalty base and royalty rate must reflect the value attributable to the infringing features of the product, and no more." *Ericsson, Inc. v. D-Link Sys., Inc.*, 773 F.3d 1201, 1226 (Fed. Cir. 2014). Mr. Green does not base his calculations on the "smallest salable patent-practicing unit"—which is at most the TPUv2 or v3 chip. Instead, Mr. Green derives his royalty base by analyzing the value of the entire TPU systems. These systems undisputedly contain myriad non-accused features and Google-

[3] Singular's damages expert also provides a lump-sum damages range of $3.3 to $7.01 billion. This larger range improperly uses TPU *global* deployment data instead of *domestic* deployment data, further inflating the calculations. Because an infringer is only liable to the extent that it uses, makes, offers to sell, or sells a patented invention within the United States, 35 U.S.C. § 271(a), or supplies components to assemble a patented invention abroad, *id.* § 271(f), non-domestic *use* of the TPUs cannot form part of the royalty calculation.

specific innovations that provide value unrelated to the asserted claims, including at least (1) the interconnects between TPU chips that allow the chips to be networked and work in conjunction for parallel scaling, (2) the interaction between hardware, software, and infrastructure in Google's TPU ecosystem, and (3) Google's innovations in its services and products. These non-accused features that make up TPU systems produced all of the speed and efficiency gains over alternative chips: While industry benchmarks show that individual TPUv2 and v3 *chips* perform no better than GPU chips, TPUv2 and v3 *systems* produced significant gains over a network of GPUs. Thus, Mr. Green's reliance on the value of TPU *systems*, rather than TPU *chips*, poisons Singular's entire damages analysis.

Second, Google would not have paid a massive royalty to a company lacking any commercial success—whether through licensing or from its own product. The evidence will show that Singular spent years trying to license its patents to dozens of companies including [redacted], all to no avail. The few that took evaluation licenses for Singular's S1 product rejected the technology, with one finding that implementation of an algorithm on the S1 was "much slower" than on existing chips. Singular's lack of success further undermines Mr. Green's damages calculations.

Third, rather than pay an enormous lump sum royalty for unproven technology, Google could have implemented relatively simple—and cheap—modifications to the TPUv2 and TPUv3 to avoid infringing the patents under Singular's infringement theory, the cost of which serves roughly as a ceiling on what Google would have agreed to pay in a hypothetical negotiation. As Dr. Walker will explain at trial, with slight modification and minimal cost, Google could have designed the TPUs to use the bfloat20 number format instead of bfloat16. Use of bfloat20, even under Dr. Khatri's theory, would not produce results that satisfy the relative error limitation. The

relatively easy conversion of the TPUs to use an available noninfringing alternative even further cuts against Singular's inflated request.

Finally, Singular's damages are limited by its failure to mark its own S1 product as required by 35 U.S.C. § 287. Under Section 287, a patentee who makes, offers for sale, or sells within the U.S. any patented article, or imports any patented article into the U.S., must mark its patented articles or provide pre-suit notice of infringement to recover pre-suit damages. *Arctic Cat Inc. v. Bombardier Rec. Prods.*, 876 F.3d 1350, 1366 (Fed. Cir. 2017). Singular and its expert Dr. Khatri both contend that the S1 practices the asserted claims, but it appears undisputed that the S1 was not marked with the patent numbers of the asserted patents. Section 287 is clear on the result here: "[N]o damages shall be recovered by [Singular] in any action for infringement, except on proof that the infringer was notified of the infringement and continued to infringe thereafter, in which event damages may be recovered only for infringement occurring after such notice." 35 U.S.C. § 287(a). The only "noti[ce] of infringement" that Singular can point to in this case is the filing of its complaint in December 2019, *see id.*, which means it may only recover damages for alleged infringement from that date forward.

To the extent that the jury returns a verdict of infringement, a proper reasonable royalty would consider the availability of the bfloat20 noninfringing alternative, Google's myriad non-accused contributions to improve the TPU system, and Singular's failure to monetize the technology and its willingness to license to Google, as Ms. Stamm will explain at trial. Even accepting for argument's sake that Google could not avail itself of bfloat20, comparison to fp16 (which other companies use to economically perform machine learning tasks) would be justified over Singular's flawed model.

V. CONCLUSION

For the foregoing reasons and more to be shown at trial, Singular will be unable to prove its claims.

Respectfully submitted,

Dated: December 29, 2023

By: */s/ Nathan R. Speed*

Gregory F. Corbett (BBO #646394)
gcorbett@wolfgreenfield.com
Nathan R. Speed (BBO #670249)
nspeed@wolfgreenfield.com
Elizabeth A. DiMarco (BBO #681921)
edimarco@wolfgreenfield.com
Anant K. Saraswat (BBO #676048)
asaraswat@wolfgreenfield.com
WOLF, GREENFIELD & SACKS, P.C.
600 Atlantic Avenue
Boston, MA 02210
Telephone: (617) 646-8000
Fax: (617) 646-8646

Robert Van Nest (admitted *pro hac vice*)
rvannest@keker.com
Michelle Ybarra (admitted *pro hac vice*)
mybarra@keker.com
Andrew Bruns (admitted *pro hac vice*)
abruns@keker.com
Vishesh Narayen (admitted *pro hac vice*)
vnarayen@keker.com
Christopher S. Sun (admitted *pro hac vice*)
csun@keker.com
Anna Porto (admitted *pro hac vice*)
aporto@keker.com
Deeva Shah (admitted *pro hac vice*)
dshah@keker.com
Stephanie J. Goldberg (admitted *pro hac vice*)
sgoldberg@keker.com
Eugene M. Paige (admitted *pro hac vice*)
epaige@keker.com
Rachael E. Meny (admitted *pro hac vice*)
rmeny@keker.com
Eric K. Phung (admitted *pro hac vice*)
ephung@keker.com
Kaiyi A. Xie (admitted *pro hac vice*)
kxie@keker.com
Spencer McManus (admitted *pro hac vice*)

smcmanus@keker.com
KEKER, VAN NEST & PETERS LLP
633 Battery Street
San Francisco, CA 94111-1809
Telephone: (415) 391-5400

Michael S. Kwun (admitted *pro hac vice*)
mkwun@kblfirm.com
Asim M. Bhansali (admitted *pro hac vice*)
abhansali@kblfirm.com
KWUN BHANSALI LAZARUS LLP
555 Montgomery Street, Suite 750
San Francisco, CA 94111
Telephone: (415) 630-2350

Matthias A. Kamber (admitted *pro hac vice*)
matthiaskamber@paulhastings.com
PAUL HASTINGS LLP
101 California Street, 48th Floor
San Francisco, CA 94111
Telephone: (415) 856-7000
Fax: (415) 856-7100

Ginger D. Anders (admitted *pro hac vice*)
Ginger.Anders@mto.com
J. Kain Day (admitted *pro hac vice*)
Kain.Day@mto.com
MUNGER, TOLLES & OLSON LLP
601 Massachusetts Avenue NW, Suite 500E
Washington, D.C. 20001
Tel: (202) 220-1100

Jordan D. Segall (admitted *pro hac vice*)
Jordan.Segall@mto.com
MUNGER, TOLLES & OLSON LLP
350 South Grand Avenue, 50th Floor
Los Angeles, CA 90071-3426
Tel: (213) 683-9100

Counsel for Defendant Google LLC

CERTIFICATE OF SERVICE

I certify that this document is being filed through the Court's electronic filing system, which serves counsel for other parties who are registered participants as identified on the Notice of Electronic Filing (NEF). Any counsel for other parties who are not registered participants are being served by first class mail on the date of electronic filing.

/s/ Nathan R. Speed
Nathan R. Speed