

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

GOOGLE LLC,

Petitioner,

v.

SINGULAR COMPUTING LLC,

Patent Owner.

Patent No. 10,754,616

Filing Date: May 25, 2020

Issue Date: August 25, 2020

Inventor: Joseph Bates

Title: PROCESSING WITH COMPACT ARITHMETIC
PROCESSING ELEMENT

PATENT OWNER'S PRELIMINARY RESPONSE

Case No. IPR2023-00394

TABLE OF CONTENTS

	<u>Page(s)</u>
I. INTRODUCTION	1
II. THE '616 PATENT	3
III. LEVEL OF ORDINARY SKILL IN THE ART	9
IV. CLAIM CONSTRUCTION	9
V. LEGAL STANDARD	9
VI. THE BOARD SHOULD DENY INSTITUTION BECAUSE PETITIONER HAS FAILED TO SHOW A REASONABLE LIKELIHOOD OF SUCCESS FOR ANY OF GROUNDS 1-4.....	10
A. Ground 1A: Claims 1 and 4-5 Are Not Obvious Over Schmidt- 2000	10
1. Schmidt-2000's Floating Point Arithmetic System.....	11
2. Schmidt-2000 "Multiplier Circuit" as Presented by Google.....	15
3. Schmidt-2000, Whether or Not Modified by Google, Does Not Disclose or Render Obvious the "corresponding multiplier circuit" Limitation	16
4. Schmidt-2000 Does Not Disclose or Render Obvious "an input-output unit connected to the first edge processing element and the second edge processing element"	22
B. Ground 1B: Claims 7, 9, 14 Are Not Obvious Over Schmidt- 2000 in View of POSA's Background Knowledge of Semiconductor Process Advances.....	28
1. Google's Combination Does Not Render Obvious "an input-output unit connected to each of the first subset of the plurality of first processing elements;"	28

2.	Google’s Combination Does Not Render Obvious Claim 7’s “corresponding multiplier circuit” limitations	28
3.	Google’s Combination Does Not Render Obvious Claim 7’s “wherein the plurality of first processing elements is no less than 5000 in number” Limitation.....	29
(a)	The 5000-PE Limitation Carries Patentable Weight.....	29
(b)	No Evidence Presented to Motivate a 5000-PE Device	30
(c)	Application of Moore’s Law Is an Exercise in Hindsight.....	31
C.	Ground 2: Claims 2-3 Are Not Obvious Over Schmidt-2000 in View of Schmidt-1999	32
1.	Schmidt-2000 in View of Schmidt-1999 Does Not Disclose or Render Obvious a Control Unit	32
D.	Ground 3: Claims 8 and 10-13 Are Not Obvious Over Schmidt-2000 in View of Schmidt-1999 and Kan.....	33
1.	Schmidt-2000 in View of Schmidt-1999 and Kan Does Not Disclose or Render Obvious a Control Unit	33
E.	Ground 4: Claims 16-17 Are Not Obvious Over Schmidt-2000 in View of Schimmler-1996	35
1.	Google’s Combination Does Not Render Obvious “an input-output unit connected to the first edge processing element and the second edge processing element”	35
2.	Google’s Combination Does Not Render Obvious “an external input-output unit external to the computing chip”	35
3.	Google’s Combination Does Not Render Obvious Claim 16’s “multiplier circuit” Limitations.....	36

F.	Ground 5: Claims 18-22 and 24 Are Not Obvious Over Schmidt-2000 in View of Schimmler-1996, Schmidt-1999, and Kan	36
VII.	CONCLUSION.....	36

TABLE OF AUTHORITIES

	Page(s)
Cases	
<i>Apple Inc. v. Samsung Elecs. Co.</i> , 839 F.3d 1034 (Fed. Cir. 2016)	10
<i>Belden Inc. v. Berk-Tek LLC</i> , 805 F.3d 1064 (Fed. Cir. 2015)	31, 32
<i>Chemours Co. FC, LLC v. Daikin Indus., Ltd.</i> , 4 F.4th 1370 (Fed. Cir. 2021)	27, 28
<i>Forest Lab ’ys, LLC v. Sigmapharm Lab ’ys, LLC</i> , 918 F.3d 928 (Fed. Cir. 2019)	10
<i>Google LLC v. Singular Computing LLC</i> , IPR2021-00155, Paper No. 57 (P.T.A.B. May 11, 2022)	27
<i>Google LLC v. Singular Computing LLC</i> , IPR2021-00155, Paper No. 62 (P.T.A.B. May 11, 2022)	<i>passim</i>
<i>Graham v. John Deere Co. of Kansas City</i> , 383 U.S. 1 (1966).....	9, 10
<i>KSR Int’l Co. v. Teleflex Inc.</i> , 550 U.S. 398 (2007).....	10
<i>Pers. Web Techs., LLC v. Apple, Inc.</i> , 848 F.3d 987 (Fed. Cir. 2017)	31, 34
<i>Sanofi-Synthelabo v. Apotex, Inc.</i> , 550 F.3d 1075 (Fed. Cir. 2008)	30
<i>Trivascular, Inc. v. Samuels</i> , 812 F.3d 1056 (Fed. Cir. 2016)	27

Statutes

35 U.S.C. § 103	9
35 U.S.C. § 103(a)	1

EXHIBIT LIST

Exhibit No.	Description of Document
2001	Definition of “bus” from Microsoft Computer Dictionary, Fifth Edition

I. INTRODUCTION

On December 22, 2022, Google LLC (“Google” or “Petitioner”) submitted a Petition for *Inter Partes* Review (Paper No. 2, “Petition”) of U.S. Patent No. 10,754,616 (Ex. 1001, “the ’616 Patent”), challenging Claims 1-5, 7-14, 16-22, and 24 pursuant to §§ 311-319 and § 42.100, *et seq.* (“the Challenged Claims”).

The Petition should be denied because the Petitioner has failed to demonstrate a reasonable likelihood that the Challenged Claims are invalid as obvious under 35 U.S.C. § 103(a) based on the prior art references and obviousness grounds set forth in the Petition. The prior art references and obviousness grounds cannot render the Challenged Claims obvious for at least two reasons.

First, none of the cited references disclose or render obvious a “multiplier circuit adapted to receive as a first input *a floating point value having a mantissa of width no more than 11 bits*”, as required by all the Challenged Claims. Schmidt-2000’s actual floating point multiplier circuit receives as an input an ultra high-precision¹ floating point value having an overall width of 80 bits and a mantissa

¹ The Schmidt-2000 floating point multiplication operation is labeled herein as “ultra-high-precision” because the Schmidt-2000 system *increases* the precision and exponent from standard IEEE format to a 15-bit exponent and 64-bit mantissa, resulting in mantissa bit widths that are 41 bits more than a standard IEEE single-

width of 64-bits.² The multiplier circuit that Google presented and analyzed, which is a circumscribed version of the actual Schmidt-2000 multiplier circuit analyzed over only a subset of the steps it requires to perform a single floating point multiplication operation, only receives a part of a floating-point value that is itself an 8-bit portion of a 64 bit mantissa. Neither Schmidt-2000's actual multiplier circuit, nor Google's proposed multiplier circuit, is "a multiplier circuit adapted to receive *a floating-point value having a mantissa of width no more than 11 bits.*"

Second, the cited references do not disclose or render obvious an input-output unit as required by the Challenged Claims. Indeed, Schmidt-2000 boasts about having "no interface processors at all." Schmidt-2000 at 168. Instead, Google argues that a collection of memory registers and a data bus comprise an input-output unit. But memory registers and a bus, even if connected as Google proposes, do not constitute an input-output unit.

precision floating point value mantissa width, and 12 bits more than a standard IEEE double-precision floating point value mantissa width. Ex. 1007, "Schmidt-2000," 167.

² Schmidt-2000 also mentions that in an alternative embodiment, its circuit could be adapted to operate on full-precision floating point values having 23-bit mantissas. The arguments in this POPR apply equally to this alternative embodiment.

These two missing elements in combination are integral to the '616 Patent's claimed architecture, enabling the economical delivery of massive amounts of computing power (1:16-27). Accordingly, the Board should deny institution.

II. THE '616 PATENT

The '616 Patent is entitled "Processing with compact arithmetic processing element" and issued on August 25, 2020. The '616 Patent claims priority, through a chain of parent and grandparent applications, to U.S. Provisional Patent Application No. 61/218,691, filed on Jun. 19, 2009. In previous *inter-partes* reviews brought by the same Petitioner, the Board found claims from three of the patents that have issued from this chain to not be invalid. *See, e.g., Google LLC v. Singular Computing LLC*, IPR2021-00155, Paper No. 62 (P.T.A.B. May 11, 2022) (finding claims not unpatentable). At a high level, the '616 Patent is directed to a computer processor that economically achieves a massive degree of parallelism by incorporating an array of multiplier circuits that each perform arithmetic on values having relatively high dynamic range exponents and low precision mantissas. *See* '616 Patent at 1:63-2:15; 5:66-6:19; 6:49-61; 7:3-9; 7:19-38; 8:38-42; 13:52-60; 16:41-55; 25:49-55.

The '616 Patent's inventor, Dr. Joseph Bates, recognized that even though then-modern processors contained hundreds of millions of transistors, they could perform only a handful of operations per clock cycle:

Consider a modern silicon microprocessor chip containing about one billion transistors, clocked at roughly 1 GHz. On each cycle the chip delivers approximately one useful arithmetic operation to the software it is running. For instance, a value might be transferred between registers, another value might be incremented, perhaps a multiply is accomplished. This is not terribly different from what chips did 30 years ago, though the clock rates are perhaps a thousand times faster today.

Id. at 1:40-48.

As Dr. Bates explained, a large portion of this inefficiency comes from using transistor-intensive arithmetic units that perform arithmetic on values having relatively high dynamic range exponents and high-precision mantissas:

As described above, today's CPU chips make inefficient use of their transistors. For example, a conventional CPU chip containing a billion transistors might enable software to perform merely a few operations per clock cycle. Although this is highly inefficient, those having ordinary skill in the art design CPUs in this way for what are widely accepted to be valid reasons. For example, such designs satisfy the (often essential) requirement for software compatibility with earlier designs. Furthermore, they deliver great precision, performing exact arithmetic with integers typically 32 or 64 bits long and performing rather accurate and widely standardized arithmetic with 32 and 64 bit floating point numbers. Many applications need this kind of precision. As a result, conventional CPUs typically are designed

to provide such precision, using on the order of a million transistors to implement the arithmetic operations.

Id. at 2:62-3:10.

However, Dr. Bates realized that such high-precision computing was not necessary for all applications, including many valuable ones:

There are many economically important applications, however, which are not especially sensitive to precision and that would greatly benefit, in the form of application performance per transistor, from the ability to draw upon a far greater fraction of the computing power inherent in those million transistors. Current architectures for general purpose computing fail to deliver this power.

Id. at 3:11-17.

The '616 Patent thus is directed away from prior art computers which are based on high-precision execution units that take up space and are wasteful of transistors, and towards computers based on low-precision high-dynamic range (LPHDR) execution units, such as those based on LPHDR floating-point numerics:

Embodiments of the present invention efficiently provide computing power using a fundamentally different approach than those described above. In particular, embodiments of the present invention are directed to computer processors or other devices which use low precision high dynamic range (LPHDR)

processing elements to perform computations (such as arithmetic operations).

One variety of LPHDR arithmetic represents values from one millionth up to one million with a precision of about 0.1%. If these values were represented and manipulated using the methods of floating point arithmetic, they would have binary mantissas of no more than 10 bits plus a sign bit and binary exponents of at least 5 bits plus a sign bit.

Id. at 5:56-69.

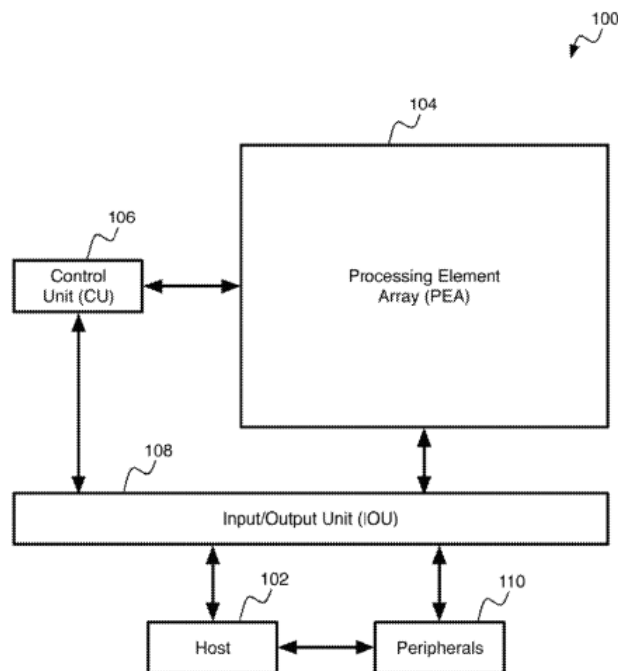
By using an LPHDR architecture as claimed,

the area of the arithmetic circuits remains relatively small and a greater number of computing elements can be fit into a given area of silicon. This means the machine can perform a greater number of operations per unit of time or per unit power, which gives it an advantage for those computations able to be expressed in the LPHDR framework.

Id. at 6:11-17. Indeed, “[b]ecause LPHDR processing elements are relatively small, a single processor or other device may include a very large number of LPHDR processing elements, adapted to operate in parallel with each other.” *Id.* at 6:49-52. As a result, “embodiments of the present invention may be implemented as any kind of machine which uses LPHDR arithmetic processing elements to provide

computing using a small amount of resources (*e.g.*, transistors or volume) compared with traditional architectures.” *Id.* at 8:1-5. In short, the ’616 Patent achieves the economical delivery of more computing power by teaching a computer architecture having large numbers of processing elements that each process numerical values having relatively high dynamic range low precision (*i.e.*, shorter mantissa) formats.

The ’616 Patent leverages the relatively small size enabled by the use of LPHDR arithmetic to teach a new computing system based on a massively parallel array of LPHDR processing elements (PEA). ’616 Patent at 8:12-28. An embodiment is shown in Figure 1:



Id., Fig. 1.

In an embodiment, the “Host 102 is responsible for overall control of the computing system 100.” *Id.*, 8:29-30. The Host seeks “to have the PEA 104 perform massive amounts of computations in a useful way . . . by causing the PE’s to perform computations, typically on data stored in each PE, in parallel with one another.” *Id.*, 8:37-40. To most efficiently control the PEA, the system may include a specialized control unit (CU) having “the primary task of retrieving and decoding instructions from an instruction memory . . . and issuing the partially decoded instructions to all the PEs in the PEA 104.” *Id.*, 8:49-54. The Host loads a program for execution by the PEA into the CU instruction memory and then instructs the CU to interpret the program’s instructions and cause the PEA to execute the instruction. *Id.*, 8:61-65.

The massively parallel PEA’s ability to process data far faster than the Host 102 requires an I/O unit (IOU) 108 to interface the Host and various peripherals in the Host system, with the PEA. *Id.*, 9:7-15. Getting data into and out of the faster PEA without absorbing large amounts of hardware resources is a critical design goal of the architecture covered by the Challenged Claims, that is achieved using the IOU, and specifically by connecting the IOU to the PEs only at the edges of the PEA. *Id.*, 10:1-3. Data may be moved between PEs only to their “nearest neighbors” so that “there are no long distance transfers,” as shown in Figures 2 and 3. *Id.*, 9:39-44, Figs. 2, 3. In that case, “data is read and written at the edges of the array and CU

instructions are performed to shift data between the edges and interior of the PEA 104.” *Id.*, 10:3-8.

III. LEVEL OF ORDINARY SKILL IN THE ART

For the purposes of this Preliminary Response only, Patent Owner utilizes Petitioner’s proposed level of ordinary skill in the art: “a bachelor’s degree in Electrical Engineering, Computer Engineering, Applied Mathematics, or the equivalent, and two years of academic or industry experience in computer architecture. More education could substitute for experience, and vice versa.” Pet. at 4.

IV. CLAIM CONSTRUCTION

For the purposes of this Preliminary Response only, Patent Owner agrees that claim construction is unnecessary. As discussed below, Petitioner’s references fail to disclose or render obvious numerous limitations of the claims.

V. LEGAL STANDARD

The question of obviousness is resolved on the basis of underlying factual determinations, including: (1) the scope and content of the prior art, (2) any differences between the claimed subject matter and the prior art, (3) the level of skill in the art, and (4) where in evidence, so called secondary considerations. *Graham v. John Deere Co. of Kansas City*, 383 U.S. 1, 17–18 (1966). A claim is only unpatentable under 35 U.S.C. § 103 if “the differences between the subject matter

sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains.” *KSR Int’l Co. v. Teleflex Inc.*, 550 U.S. 398, 406 (2007) (quoting 35 U.S.C. § 103).

“An invention is not obvious simply because all of the claimed limitations were known in the prior art at the time of the invention. Instead, we ask ‘whether there is a reason, suggestion, or motivation in the prior art that would lead one of ordinary skill in the art to combine the references, and that would also suggest a reasonable likelihood of success.’” *Forest Lab’ys, LLC v. Sigmapharm Lab’ys, LLC*, 918 F.3d 928, 934 (Fed. Cir. 2019). “Of course, concluding that the references are within the scope and content of the prior art to be considered for obviousness does not end the inquiry. *Graham* makes clear that the obviousness inquiry requires a determination whether the claimed invention would have been obvious to a skilled artisan.” *Apple Inc. v. Samsung Elecs. Co.*, 839 F.3d 1034, 1050 n.14 (Fed. Cir. 2016).

VI. THE BOARD SHOULD DENY INSTITUTION BECAUSE PETITIONER HAS FAILED TO SHOW A REASONABLE LIKELIHOOD OF SUCCESS FOR ANY OF GROUNDS 1-4

A. Ground 1A: Claims 1 and 4-5 Are Not Obvious Over Schmidt-2000

Schmidt-2000 does not render obvious the claims addressed in Ground 1 because Schmidt-2000 does not teach, disclose, or fairly suggest, on its own or in

combination with the standard knowledge and background of a person of ordinary skill, any of the “multiplier circuits” of the Challenged Claims or the input-output unit.

1. Schmidt-2000’s Floating Point Arithmetic System

KPROC – An Instruction Systolic Architecture for Parallel Prefix Applications (Ex. 1007, “Schmidt-2000”) discloses “a parallel computer with 1024 floating-point processors on a single chip,” which it calls a KPROC (“KiloPROCessor”). Schmidt-2000, 163.

Schmidt-2000’s KPROC is “tailored towards an efficient implementation of complex operations like double precision floating point addition or 1024-bit integer multiplication.” *Id.* The KPROC “is a quadratic array of identical processors, each connected to its four direct neighbors by single data wires.” *Id.*

Schmidt-2000 explains that:

The goal of integration of 1024 processors on a single chip has created a number [of] requirements for the implementation:

- The whole system should dissipate less than 5W. Since it consists of 1024 processors the individual processor was limited to slightly above 4mW.
- The chip area should not exceed 1.2 cm. Having 1024 processors plus padframe and clock distribution on one chip the area for one processor was limited to about 0.1 mm.

- 64-Bit and 32-Bit floating-point formats of IEEE should be supplied as well as all kinds of integer and word formats.
- The internal global clock rate should be 400 MHz.

Schmidt-2000, 164 (footnotes omitted).

The KPROC includes 1024 processors (arranged in a 32 X 32 array) that perform bit-serial operations. Schmidt-2000, 163. In Schmidt-2000's "bit-serial" processor, each operation (including each multiplication operation) is performed "one bit per clock cycle." *Id.* Schmidt-2000 uses the bit-serial data organization to save space while preserving the ability to perform operations on, *e.g.*, IEEE-format 32- and 64-bit numbers. *Id.*, 164.

Schmidt-2000 performs floating point arithmetic on double-precision (64-bits wide, including sign, exponent and mantissa bits) and single-precision (32-bits wide, including sign, exponent and mantissa bits) IEEE floating point formats. *Id.*

Because of the density of processors on the KPROC chip, the designers were concerned about needing to inspect too many bits in certain special cases encoded into the IEEE double-precision floating point format as "this implies large circuitry and a high power consumption in floating point computations." *Id.*, 167. The KPROC overcomes this by taking as inputs double-precision IEEE standard floating point format values and converting them to its own internal ultra high-precision format prior to multiplication. This format (i) takes all special case information and encodes it into three bits that are added to the standard 11-bit exponent portion of

the numeric and further (ii) *expands* the mantissa (“significant”) portion to 64 bits (an increase from the double-precision standard 52 bits). *Id.* Then, “floating point multiplication is performed in the usual way. Exponents are added and mantissas are multiplied *using the integer units.*” *Id.* (emphasis added.) In other words, the bit-serial integer units multiply the mantissas one bit at a time. Afterwards, the resulting value is converted back into IEEE double-precision standard format and output. *Id.* Importantly, these custom Schmidt-2000 ultra high-precision formatted floating point operands “even gives additional precision” over and above the already-high precision provided by the IEEE double-precision standard floating point format. Schmidt-2000, 167.

Just as importantly, as Google admits in its Petition, the Schmidt-2000 multiplier has a parallel-serial design. That is, rather than achieving compactness using processing elements that operate on high dynamic range values with low precision (*i.e.*, relatively short) mantissas as claimed by the ’616 Patent, Schmidt-2000 achieves compactness using processing elements that operate on values with very high precision (*i.e.*, relatively long, greater-than-double-precision) mantissas, but that do so only using bitwise multiplication (*i.e.*, one bit multiplication per clock cycle, on all 64 mantissa bits). Schmidt-2000, 163 (“Due to the limited area the processor had to be very compact. In fact, this was the reason for the choice of a bit serial data organisation.”); *see also id.*, 164-66. This allows for smaller processing

elements but results in arithmetic operations that take far more time (*i.e.*, far more clock cycles) to complete.³ In other words, while compactness is achieved at the price of lower precision in the '616 Patent, compactness is achieved at the price of longer compute times (but with ultra high precision) in Schmidt-2000.

More specifically, the Schmidt-2000 multiplier circuit first loads a 16-bit portion of one 64-bit mantissa (*i.e.*, corresponding to the first of two custom Schmidt-2000 ultra high-precision formatted floating point operands) over 16 clock cycles. A 16-bit portion of a different 64-bit mantissa, (*i.e.*, corresponding to a second of the two custom Schmidt-2000 ultra high-precision formatted floating point operands), is then serially broadcast to all 16 bits of the first already-loaded part of the first operand over the next 16 clock cycles, with a bitwise multiplication occurring at each clock cycle. The product is accumulated. Then, the foregoing 32 clock-cycle procedure is repeated for the remaining three un-multiplied 16-bit portions of the 64 mantissa bits of both the first and second operands. As a result, the Schmidt-2000 multiplier circuit requires a total of at least 64 clock cycles to perform a single double-precision floating point multiplication operation. That is,

³ The '616 Patent contrasts its invention to this type of small bit-width processor, as it inefficiently requires many operations to complete conventional calculations. '616 Patent, 3:46-65.

the multiplier circuit performs “the same procedure . . . more than one time[], . . . add[ing] the intermediate results” to arrive at the full product. Schmidt-2000, 166 (description of “Multiplier”); Ex. 1003, Franzon Declaration, ¶ 131; Pet. at 35.

2. Schmidt-2000 “Multiplier Circuit” as Presented by Google

Google does not compare the Challenged Claims to the actual Schmidt-2000 multiplier circuit, nor even a subset of that circuit. Doing so would be illogical, given that circuit’s ultra high-precision formatted floating point operands which “even gives additional precision” over and above the already-high precision provided by the IEEE double-precision standard floating point format Schmidt-2000, 167, and given the ’616 Patent’s teaching away from prior art computers which are based on high precision execution units. Instead, it asks the Board to find obviousness based on a particular subset of the steps of Schmidt-2000’s floating point multiplication process.

As discussed above, to multiply two floating point numbers that each have an ultra high-precision 64 bit mantissa, the actual Schmidt-2000 KPROC divides each mantissa into four 16-bit portions, multiplies a first pair of such portions together one bit at a time, and stores the intermediate product. It then repeats this intermediate processing step three times, each time on a pair of the three other portions of the input’s ultra high-precision 64 bit mantissas. The stored intermediate products are

then accumulated to arrive at the full product over a total of at least 64 clock cycles. Google takes this ultra high-precision multiplier circuit and imagines it multiplying an 8-bit portion (*i.e.*, instead of a 16-bit portion) of the input's ultra high-precision 64 bit mantissa in each of these intermediate steps. Pet. at 7-10. Google then, limiting its analysis to one of these intermediate steps, declares the Schmidt-2000's use of integer units to perform the bit-serial multiplication on those 8-bit portions of the input's ultra high-precision 64 bit mantissa, as rendering the claims obvious. *Id.*; *see also id.* at 11-27.

**3. Schmidt-2000, Whether or Not Modified by Google,
Does Not Disclose or Render Obvious the
“corresponding multiplier circuit” Limitation**

Claim 1 requires a “corresponding multiplier circuit” that receives as input “floating point value[s] having a [] binary mantissa of width no more than 11 bits and a [] binary exponent of width at least 6 bits.” ’616 Patent, cl. 1. The actual Schmidt-2000, which performs standard floating-point arithmetic on floating point values having mantissas of at least 23 bits using bit serial multiplication does not disclose or render obvious such a multiplier circuit.

Schmidt-2000 is specifically designed for IEEE standard-precision or double-precision floating point numbers. Schmidt-2000, 164 (noting as a “requirement” that “64-Bit and 32-Bit floating-point formats of IEEE should be supplied”), *Id.* at 164. As Schmidt-2000 explains, standard IEEE double-precision floating point

values have 11 exponent bits and 52 bits of mantissa. *Id.*, 167. Standard IEEE full-precision floating point values have 8 bits of exponent and 23 bits of mantissa. *E.g.*, Ex. 1013, 395-96 (“The IEEE standard single precision FP representation has a single sign bit (S), eight exponent bits (E), and 23 mantissa bits (M).”). Moreover, even after taking IEEE standard or double precision floating point inputs, Schmidt-2000’s floating point multiplication process never reduces the number of bits in the mantissa. Instead, as mentioned above, it *expands* the number of mantissa bits “which even gives additional precision” (Schmidt-2000 at 167), again in contrast to the low precision operation taught and claimed by the ’616 Patent. Such a full or double-precision multiplier circuit, operating on ultra-high precision mantissas, as described above, not only fails to meet the Challenged Claim’s mantissa width limitation, but also flies in the face of the central teaching of the ’616 Patent, which is to eschew performing arithmetic on full or high precision numbers as wasteful of computer resources. ’616 Patent, 4:65-5:2, 5:30-33.

Knowing this, Google imagines Schmidt-2000 as multiplying an 8-bit portion (*i.e.*, instead of a 16-bit portion) of its input’s ultra high-precision mantissa during one of the above-described intermediate processing steps. Google then argues that “[t]he integer adder, integer multiplier, ‘floating point control’ and ‘shift/add’ units that are used to perform floating-point multiplication . . . meet the claimed ‘multiplier circuit’.” Pet. at 25. Below is Google’s annotated Fig. 3.1.

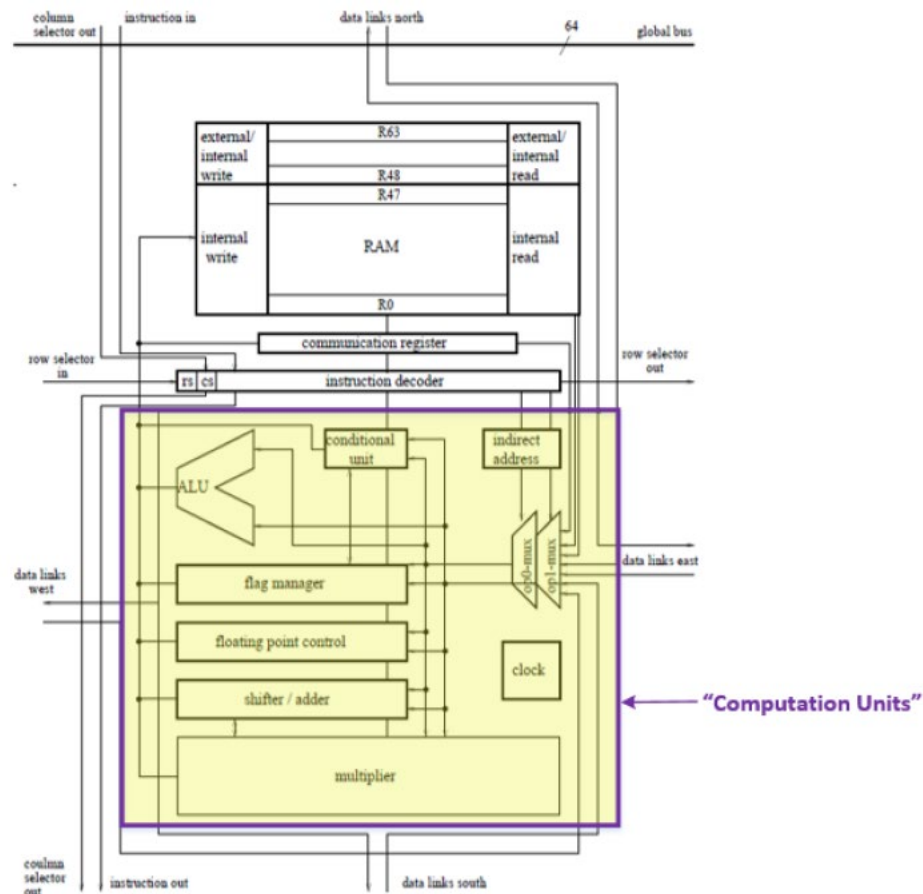


FIG. 3.1. Architecture of one processor.

Id.

At the outset of this argument however, Google fails to identify the input to its modified multiplier circuit. *Id.* at 24-27. This failure is critical, because the limitation requires that the first and second inputs to the multiplier circuit are each a ***floating point value*** having a mantissa of 11 or fewer bits, and an exponent of at least 6 bits. Google scrupulously avoids such an identification because its theory is that the multiplier circuit meets the limitation when it performs a bit-serial multiplication of an 8-bit portion of the 64-bit mantissa part of the floating point

value. But simply identifying that the system is built to an 8-bit or 16-bit architecture does not demonstrate that any input to the multiplier circuit is formatted as a floating point value with a “binary mantissa of width no more than 11 bits and a [] binary exponent of width at least 6 bits.”

Google leverages its failure to identify the input to make the unsupported statement that a supposed 8-bit modification of the KPROC system would multiply “two floating point values, each having an 8-bit exponent and a mantissa of up to 8 bits.” Pet. at 27. Google appears to argue that because the KPROC’s floating-point multiplier process operates on entire custom Schmidt-2000 ultra high-precision floating point numerical values by (i) chopping up the mantissa parts of those values into 8-bit or 16-bit portions and (ii) then performing arithmetic operations on *one of those portions* in bit-serial fashion, Schmidt-2000 therefore discloses multiplying floating point values having the mantissa and exponent limitations of the Challenged Claims, including a mantissa with a bit width of less than 11 bits (*i.e.*, when 8-bit portions of the mantissa are assumed to be multiplied) *Id.* at 25-27.

In making this argument, Google elevates form over substance. According to Google, since the Schmidt-2000 multiplier circuit, during a subset of the steps required to complete a single floating point operation, acts on an 8-bit integer portion of a larger floating point mantissa (*i.e.*, the ultra-high-precision mantissa of width 64 bits), that circuit is, in fact, operating on an 8-bit mantissa during that subset of steps

(“a mantissa of up to 8 bits (the first mantissa partition multiplied by the 8-bit multiplier)”). Pet. at 26-27. Google provides no evidence, however, that a POSA would understand an integer representation of a portion of a mantissa, to itself be a mantissa of a floating-point number.

Even if such an understanding was established, the imagined 8-bit multiplier circuit does not receive a floating-point value during the time required to complete a single floating-point operation. Instead, during that time, it receives a sequence of bit-serial *portions* of a Schmidt-2000 ultra high-precision formatted floating point mantissa, which portions are each an 8-bit integer that is multiplied by a part of Schmidt-2000 that is an integer multipliers. Schmidt-2000, 166-67 (“mantissas are multiplied using the integer units discussed above”).

Whether considering the actual Schmidt-2000 multiplier circuit, or the imagined 8-bit version of that circuit described by Google, no apparatus in Schmidt-2000 *can* be considered “*a multiplier circuit adapted to receive as an input a floating point value having a mantissa of width no more than 11 bits,*” as required by the Challenged Claims:

- If one identifies the multiplier circuit, operating over the full period required to perform a floating point multiplication operation (*i.e.*, at least 64 clock cycles as discussed above) on an entire ultra high precision input, as the “computation unit” identified in Google’s

annotation of Schmidt-2000's Fig. 3.1, then the multiplier circuit receives as an input a floating point value having *a mantissa of width 64 bits*. This is far greater than the stated 11-bit mantissa width limitation in the Challenged Claims, and flies in the face of the teachings of the '616 Patent.

- If one identifies the multiplier circuit, operating over the period required to multiply a single 8-bit portion of the mantissa part of the ultra-high precision input, as the “computation unit” identified in Google's annotation of Schmidt-2000's Fig. 3.1, then the multiplier circuit receives an 8-bit integer portion of a floating point value's 64-bit mantissa. Such a multiplier circuit is not even adapted to receive as an input a floating point value having a binary mantissa— it only is adapted to receive a portion of such a value— let alone a floating point value having the prescribed mantissa bit widths of the Challenged Claims.

Put another way, Google cannot invalidate the floating point multiplier circuit of the Challenged Claims by considering the Schmidt-2000 multiplier circuit over only a subset of the steps that it actually requires to complete a single floating-point multiplication operation. If Google chooses to consider the Schmidt-2000 multiplier circuit over only the subset of steps in which it receives an 8-bit portion of the

mantissa part of ultra high-precision floating point input, then Google must also acknowledge that during that subset of steps, that circuit is performing only integer multiplication and is, therefore, not “adapted to receive as a first input to the corresponding multiplier circuit a first floating point value.” For this reason, Google’s argument relying only on a subset of steps is fatally flawed irrespective of the number of bits in the subset that it chooses.

Accordingly, Google has not shown that this limitation is disclosed or rendered obvious.

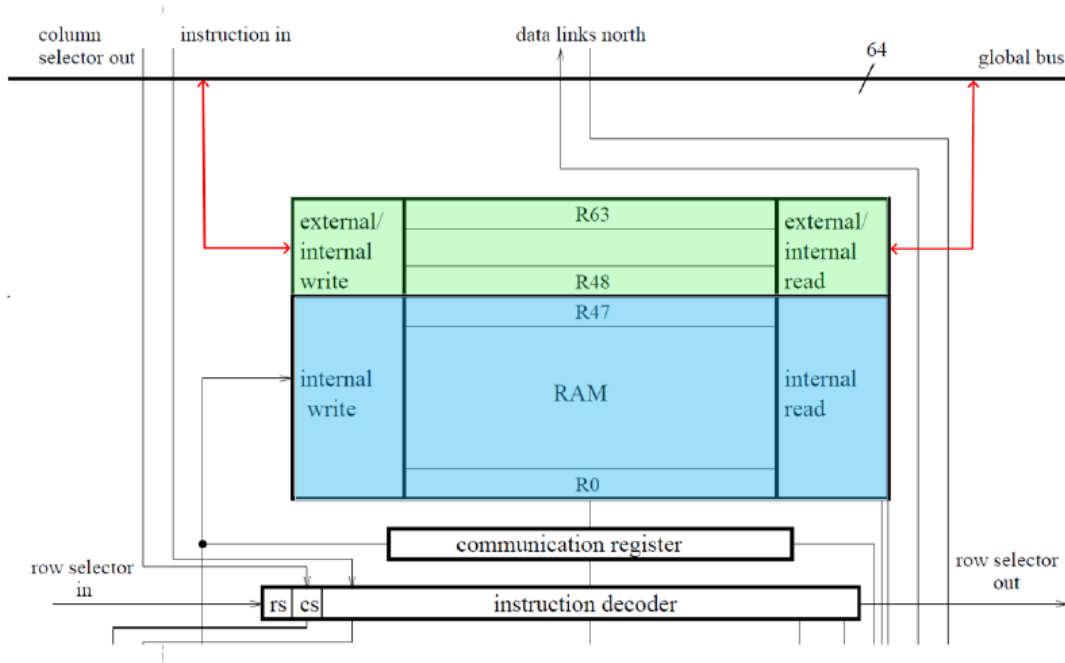
**4. Schmidt-2000 Does Not Disclose or Render Obvious
“an input-output unit connected to the first edge
processing element and the second edge processing
element”**

Schmidt-2000 neither discloses nor renders obvious “an input-output unit connected to the first edge processing element and the second edge processing element” as recited in Claim 1. Indeed, Schmidt explains that part of its advance over earlier systems is that the KPROC architecture does not require any sort of dedicated input-output unit. Schmidt-2000, 168 (“Random access data exchange with the external memory is possible, because *KPROC has no interface processors at all.*” (emphasis added)). As Schmidt-2000 explains, “[i]nstead, every processor has two memory banks, from which one is connected to the outside, while the other serves as local storage.” *Id.* Thus, because Schmidt-2000 provides direct access to

each processor's internal memory, there is no input-output unit that is connected to both a first and second edge processing unit as claimed.

This lack of a dedicated I/O unit in Schmidt-2000 is not surprising given its focus on achieving high density computation using parallel-serial multiplication, which itself takes far longer to perform than non-serial multiplication (as taught by the '616 Patent). By relying on far slower multiplication operations, Schmidt-2000 avoids the sort of speed mismatch between the host and the arithmetic circuits that arises under the architecture taught and claimed by the '616 Patent, which obviates the need in Schmidt-2000 for any dedicated I/O unit.

The details of Google's argument regarding the I/O unit reveals the lack of such a unit in Schimdt-2000. Google argues that 16 of the general-purpose registers in each of the processors on the KPROC chip (each such processor, allegedly corresponding to a "processing element" of the Challenged Claims) are "the memory bank connected to the outside for off-chip communication," and thus comprise the I/O unit (Pet. at 17). Google further inserts its own connection line from the registers to the KPROC's global bus. *Id.* Google's interpretation is shown in its modified Fig. 3.1:



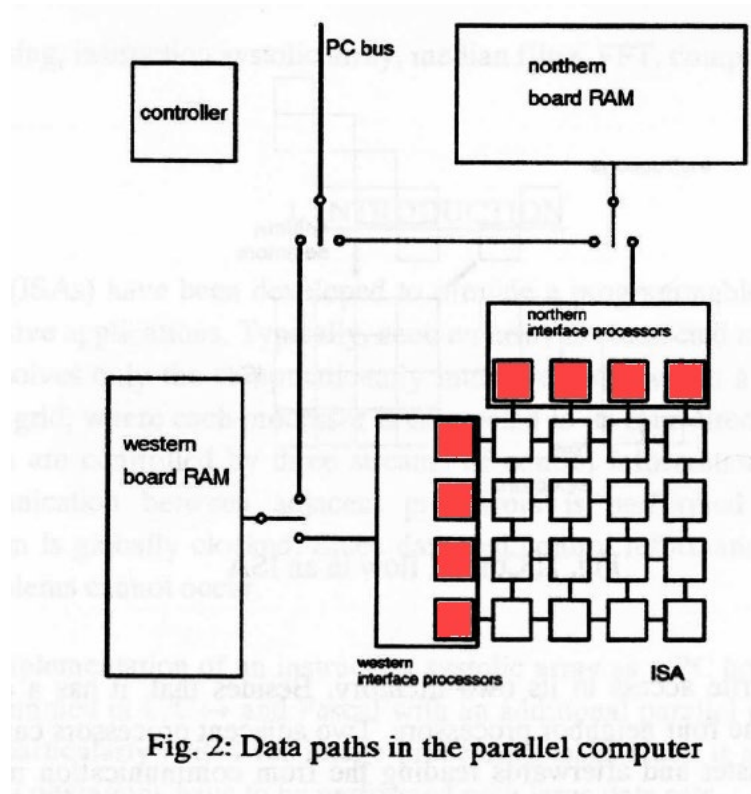
Pet. at 17. Based on its modified Fig. 3.1, Google argues that “registers R48-R63 of each processor and the global data bus” are “each individually or both together” the claimed input-output unit. *Id.* at 18-19.

Google however, previously identified these same registers, R48-R63, as part of the claimed first and second edge processing elements. *Id.* at 12-14. These registers cannot also be part of the input-output unit because the claim requires that the input-output unit be distinct from the edge processing elements, as the input-output unit is “connected to the first edge processing element and the second edge processing element.” The registers cannot be both internal to a processing element and the input-output unit connected to that processing element. Schmidt-2000, 168, Fig. 3.1. Otherwise, the input-output unit would be connected to itself.

Similarly, the global data bus cannot be the input-output unit. First, Google does not make any substantive argument to support its conclusory statement that the global bus is the input-output unit. Pet. at 17-20. The '616 Patent explains that part of the necessity for the claimed IO unit is that the PE array has the ability to “process data far faster than the Host.” '616 Patent, 9:6-18. By using an input-output unit to account for the data rate difference, the '616 Patent overcomes this issue. By contrast, there is no disclosure that Stuttard’s bus can handle a host that processes data slower than a PE array. Indeed, a “bus” is simply a “set of hardware lines (conductors) used for data transfer among the components of a computer system.” Ex. 2001, 77.

Second, Google’s support, *i.e.*, the Schimmler reference, (not mentioned in its Petition but provided by Google’s expert in his Declaration) actually counsels against such an understanding. *See* Pet. at 18 (citing Ex. 1003, ¶¶ 89-92). Schimmler (Ex. 1011) is a prior article from one of the authors of Schmidt-2000.

Unlike Schmidt-2000, Schimmler uses dedicated interface processors to allow for communications between the external host and a systolic array. Ex. 1011, 138. Each interface processor is connected to exactly one edge array processor. *Id.* (“Each interface processor is connected to its adjacent array processor by a single wire for data transfer in each direction.”).



Id., Fig. 2 (highlighting added to show interface processors).

Schmidt-2000 removes the interface processors from Schimmler and characterizes that removal as a benefit of the KPROC system. Schmidt-2000, 168 (“the KPROC architecture has a number of new features, that allows for more efficient parallel prefix operation: Random access data exchange with the external

memory is possible, because KPROC has no interface processors at all.”). Schmidt-2000 thus teaches away from including any sort of separate input-output unit.⁴

In view of this teaching away from any sort of input-output unit, a POSA would not be motivated to undo one of Schmidt-2000’s explicitly touted advances (removing interface processors) by introducing a new component into the Schmidt-2000 KPROC system that was purposely removed in the first place by its designers. *Trivascular, Inc. v. Samuels*, 812 F.3d 1056, 1068 (Fed. Cir. 2016) (finding insufficient motivation to combine two references by “substitut[ing] the recessed barbs of [one reference] with the protuberances of [another reference], since [the petitioner’s] proposed substitution would destroy the basic objective of the barbs”); *Chemours Co. FC, LLC v. Daikin Indus., Ltd.*, 4 F.4th 1370, 1376-77 (Fed. Cir. 2021) (finding insufficient motivation to “increase [a prior art reference’s] melt flow rate to the claimed range,” in part because the reference “includes numerous examples of processing techniques that are typically used to increase melt flow rate, which [the reference] cautions should not be used due to the risk of obtaining a broader molecular weight distribution”); *Google LLC v. Singular Computing LLC*,

⁴ Because Schimmler’s interface processors only connect to one array processor, as opposed to more than one edge array processor, they would also not render obvious the input-output unit limitation.

IPR2021-00155, Paper No. 57 at 75 (P.T.A.B. May 11, 2022) (citing *Chemours* and *Trivascular* and finding no motivation to combine where the combination would create “concerns that [the prior art] explicitly counsels against”).

Accordingly, Google has failed to show that Schmidt-2000 renders this limitation obvious.

B. Ground 1B: Claims 7, 9, 14 Are Not Obvious Over Schmidt-2000 in View of POSA’s Background Knowledge of Semiconductor Process Advances

Claims 9 and 14 are dependent on Claim 7. As shown below, Google has failed to show that its combination renders multiple limitations of Claim 7 obvious.

1. Google’s Combination Does Not Render Obvious “an input-output unit connected to each of the first subset of the plurality of first processing elements;”

Google relies on the same arguments as with Claim 1’s input-output unit limitation. Pet. at 42. For the same reasons as discussed above, therefore, Google’s Ground 1B combination does not render this limitation obvious.

2. Google’s Combination Does Not Render Obvious Claim 7’s “corresponding multiplier circuit” limitations

Google relies on the same arguments as with Claim 1’s multiplier circuit limitations. Pet. at 42. For the same reasons as discussed above, therefore, Google’s Ground 1B combination does not render this limitation obvious.

**3. Google's Combination Does Not Render Obvious
Claim 7's "wherein the plurality of first processing
elements is no less than 5000 in number" Limitation**

**(a) The 5000-PE Limitation Carries Patentable
Weight**

Google asserts that Claim 7's requirement of no fewer than 5000 processing elements carries no patentable weight. Pet. at 30-31. However, as Petitioner, Google has the burden to show that the claim is obvious and introduces no affirmative evidence that the limitation lacks such weight. Indeed, Google and its expert merely assert that the limitation does not produce a new and unexpected result but cite only the '616 Patent's specification in support. Pet. at 30-31 (citing Ex. 1003, ¶ 53). However, the patent specification actually spells out the new and unexpected result arising from the patent's underlying main discovery, namely that large amounts of LPHDR arithmetic performed by multiple LPHDR processing elements in a massively parallel architecture are, in fact, useful and provide significant practical benefits. '616 Patent, 7:3-38, 24:33-42. Indeed, at the time of the invention, a POSA would not expect large amounts of LPHDR arithmetic performed by multiple LPHDR processing elements in a massively parallel architecture to provide useful results (*id.*), and in fact would cause results that were simply too imprecise. *Id.*, 18:65-21:5 (noting the "surprising" result of performing many low-precision calculations is a result with less error than each individual calculation), *id.*, 24:33-42. The specification then goes on to provide various examples of the number of

such LPHDR PEs that could be deployed on a chip including 5000 (*id.*, 28:13 to 29:4). In the absence of any evidence to the contrary, the Board should consider this limitation, like all limitations, as part of the patented invention. *Sanofi-Synthelabo v. Apotex, Inc.*, 550 F.3d 1075, 1085 (Fed. Cir. 2008) (“The determination of obviousness is made with respect to the subject matter as a whole, not separate pieces of the claim.”).

(b) No Evidence Presented to Motivate a 5000-PE Device

Google points to the statement that one would build a large array of many KPROC chips as evidence that a single KRPOC chip should be built with a larger array. Pet. at 31-32. However, neither Google nor its expert provide evidence that i) using multiple KPROC chips is equivalent to a 5000+ processing element array or would motivate a POSA to ii) expand the number of processing elements in a single KPROC chip to include over 5000 PEs. Indeed, the two disparate approaches present distinct questions with distinct advantages and disadvantages to each, and a complex series of engineering cost/benefit tradeoffs which would result in significant analysis to optimize. Neither Google nor its expert analyze or even discuss this issue; instead, they simply assume that incorporating thousands of processing elements in one chip was possible and desirable and limit their analysis purely to space considerations. Pet. at 31-38.

Similarly, Schimmler-1996's embodiment that included 16 separate chips, each with a 64-processor array, does not suggest attempting to create a single chip with 5000 processors in a single array. Again, Schimmler-1996 merely presents a different possible solution and, at most, reflects that at the time, a POSA would have been motivated to assemble more chips on a board, not to create a single chip with a much larger array. Google's failure to demonstrate why a POSA would make its proposed combination is fatal to its obviousness arguments. *Pers. Web Techs., LLC v. Apple, Inc.*, 848 F.3d 987, 993-94 (Fed. Cir. 2017) (holding that testimony that references could be combined was "not enough: it does not imply a motivation to pick out those two references and combine them to arrive at the claimed invention"); *Belden Inc. v. Berk-Tek LLC*, 805 F.3d 1064, 1073 (Fed. Cir. 2015) ("[O]bviousness concerns whether a skilled artisan not only *could have made* but *would have been motivated to make* the combinations or modifications of prior art to arrive at the claimed invention.").

(c) Application of Moore's Law Is an Exercise in Hindsight

Google admits that Moore's Law was merely a prediction. Pet. at 33. Google's entire argument assumes that Moore's law was accurate, but that could be known only with the exercise of hindsight. Even if the technology as existed in 2009 was sufficient to yield a 5000-PE array, this possibility does not, *ipso facto*, supply

the motivation to actually create that processor in the first instance. *Belden*, 805 F.3d at 1073.

C. Ground 2: Claims 2-3 Are Not Obvious Over Schmidt-2000 in View of Schmidt-1999

Claims 2 and 3 depend from Claim 1. Google’s Ground 2 arguments do not remedy or address the deficiencies in its Ground 1A arguments. For the same reasons as discussed above, therefore, Google’s Ground 2 combination does not render Claims 2-3 obvious.

1. Schmidt-2000 in View of Schmidt-1999 Does Not Disclose or Render Obvious a Control Unit

In Grounds 1A and 1B, Google describes Schmidt-2000’s implementation of the KPROC chip as a coprocessor to a host computer. Pet. at 11-12, 40. According to Schmidt-2000, the KPROC chip is “controlled by three streams of control information: instructions, row selectors and column-selectors.” Schmidt-2000 at 163. Given that the array KPROC is disclosed as “connected to a sequential host,” and the array KPROC is implemented as a coprocessor to the host, the control information must emanate from the host.

Ignoring its own theory in Ground 1 which renders the KPROC chip a functioning coprocessor solely by its connection to the host, Google posits that a separate controller is the obvious and conventional way to issue instructions to the KPROC chip. Pet. at 48. Either Google is admitting that its Ground 1 combination

does not have a reasonable likelihood of success because the host cannot issue instructions to the KPROC coprocessor, or the control unit of Ground 2 is superfluous and, therefore, non-obvious because the host computer is already issuing instructions to the KPROC coprocessor. Regardless, Google ignores the role of the host in the system of Ground 1 and provides no motivation to add a control unit in view of host computer's ability to issue control information.

D. Ground 3: Claims 8 and 10-13 Are Not Obvious Over Schmidt-2000 in View of Schmidt-1999 and Kan

Claims 8 and 10-13 each ultimately depend from Claim 7. Google's Ground 3 arguments do not remedy or address the deficiencies in its Ground 1B arguments. For the same reasons as discussed above, therefore, Google's Ground 3 combination does not render Claims 8 and 10-13 obvious.

1. Schmidt-2000 in View of Schmidt-1999 and Kan Does Not Disclose or Render Obvious a Control Unit

In Ground 1, Google relied on Schmidt-2000's implementation of the KPROC chip as a co-processor to a host computer to meet limitation [1A]'s requirement for a host computer. According to Schmidt-2000, the KPROC chip "solves *only* computationally intensive tasks within a global application," Schmidt-2000 at 163, leaving the remainder of the tasks to be executed on the host computer. According to Ground 1, the KPROC chip, which is disclosed as operating on 32-bit (and 64-bit) IEEE floating-point formats should be modified to replace its 32-bit processing

elements with low precisions, high dynamic range processing elements. Presumably, the 32-bit computations are executed by the host computer in the Ground 1 combination. Claim 8 and its dependents, however, require a plurality of 32-bit processors.

Google turns to Kan for its disclosure of a MIMD unit that is implemented in a system alongside a SIMD unit. Pet. at 55. Kan teaches that its system controller selects the SIMD unit or MIMD unit depending on the type of processing that is more suitable for specific tasks. Google then posits that a POSA would be motivated to combine the MIMD unit from Kan into the KPROC chip and spends several pages in its Petition explaining why that modification would be successful.

The Petition never addresses *why* a POSA would go to the trouble and expense of integrating Kan's MIMD unit into the KPROC chip when Kan already discloses a functioning system with both MIMD and SIMD units. The Petition does not articulate a single advantage or improvement over Kan that results from its proposed modification of KPROC with Kan. Specifically, the Petition does not identify any advantages that the KPROC chip has over Kan's SIMD unit, such that a POSA would replace Kan's SIMD unit with the KPROC chip.

Absent such evidence, Google has failed to show that a POSA would make this combination. *Pers. Web Techs., LLC*, 848 F.3d at 993-94 (holding that testimony that references could be combined was "not enough: it does not imply a

motivation to pick out those two references and combine them to arrive at the claimed invention”). Accordingly, Google has not shown a reasonable likelihood of success as to this Ground.

E. Ground 4: Claims 16-17 Are Not Obvious Over Schmidt-2000 in View of Schimmler-1996

1. Google’s Combination Does Not Render Obvious “an input-output unit connected to the first edge processing element and the second edge processing element”

Google expressly relies on its Ground 1 arguments as to this limitation. Pet. at 71. For the same reasons as discussed above, its combination does not render this limitation obvious.

2. Google’s Combination Does Not Render Obvious “an external input-output unit external to the computing chip”

Google relies on the R48-R63 registers on each KPROC in an array, plus the global bus to meet this limitation. *Id.* at 72; *see also id.* at 68-69. But the global bus is not external to any computing chip. Indeed, Google argues with respect to the input-output unit that the global bus is part of the computing chip. Accordingly, its combination does not render this limitation obvious.

**3. Google's Combination Does Not Render Obvious
Claim 16's "multiplier circuit" Limitations**

Google relies on the same arguments as with Claim 1's multiplier circuit limitations. Pet. at 71-72. For the same reasons as discussed above, therefore, Google's Ground 4 combination does not render this limitation obvious.

**F. Ground 5: Claims 18-22 and 24 Are Not Obvious Over
Schmidt-2000 in View of Schimmler-1996, Schmidt-1999,
and Kan**

Claims 18-22 and 24 each ultimately depend from Claim 16. Google's Ground 5 arguments do not remedy or address the deficiencies in its Ground 4 arguments. For the same reasons as discussed above, therefore, Google's Ground 5 combination does not render Claims 18-22 and 24 obvious.

VII. CONCLUSION

For the foregoing reasons, Patent Owner respectfully requests that the Board deny institution of the Petition in its entirety.

Respectfully submitted,

Dated: May 8, 2023

By: /Peter Lambrianakos /
Peter Lambrianakos (Reg. No. 58,279)
Lead Counsel for Patent Owner
Vincent J. Rubino, III (Reg. No. 68,594)
Back-up Counsel for Patent Owner
Richard M. Cowell (Reg. No. 79,143)
Back-up Counsel for Patent Owner
FABRICANT LLP
411 Theodore Fremd Avenue
Suite 206 South
Rye, New York 10580

IPR2023-00394
PATENT NO. 10,754,616

Tel. 212-257-5797
Fax. 212-257-5796
Email: plambrianakos@fabricantllp.com
Email: vrubino@fabricantllp.com
Email: rcowell@fabricantllpp.com

CERTIFICATE OF WORD COUNT

The undersigned hereby certifies that the portions of the above-captioned PATENT OWNER'S PRELIMINARY RESPONSE TO PETITION FOR *INTER PARTES* REVIEW OF U.S. PATENT NO. 10,754,616 specified in 37 C.F.R. § 42.24 has 7,454 words in compliance with the 14,000 word limit set forth in 37 C.F.R. § 42.24. This word count was prepared using Microsoft Word for Office 365.

Respectfully Submitted,

Dated: May 8, 2023

By: /Peter Lambrianakos /
Peter Lambrianakos (Reg. No. 58,279)
Lead Counsel for Patent Owner
FABRICANT LLP
411 Theodore Fremd Avenue,
Suite 206 South
Rye, New York 10580
Tel. 212-257-5797
Fax. 212-257-5796
Email: plambrianakos@fabricantllp.com

CERTIFICATE OF SERVICE

A copy of PATENT OWNER'S PRELIMINARY RESPONSE TO PETITION FOR *INTER PARTES* REVIEW OF U.S. PATENT NO. 10,754,616 and EXHIBIT 2001 have been served on Petitioner's counsel of record as follows:

Elisabeth H. Hunt
Email: EHunt-PTAB@WolfGreenfield.com
Marc S. Johannes
Email: MJohannes-PTAB@WolfGreenfield.com
Gregory S. Nieberg
Email: GNieberg-PTAB@WolfGreenfield.com
WOLF, GREENFIELD & SACKS, P.C.
600 Atlantic Avenue
Boston, Massachusetts 02210-2206

Attorneys for Google LLC

May 8, 2023

By: /Peter Lambrianakos /
Peter Lambrianakos (Reg. No. 58,279)
FABRICANT LLP
411 Theodore Fremd Avenue
Suite 206 South
Rye, New York 10580
Tel. 212-257-5797
Fax. 212-257-5796
Email: plambrianakos@fabricantllp.com