UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

GOOGLE LLC,

Petitioner,

v.

SINGULAR COMPUTING LLC,

Patent Owner.

Patent No. 11,169,775
Filing Date: May 25, 2020
Issue Date: November 9, 2021

Inventor: Joseph Bates
Title: PROCESSING WITH COMPACT ARITHMETIC PROCESSING ELEMENT

PATENT OWNER'S PRELIMINARY RESPONSE

Case No. IPR2023-00397

TABLE OF CONTENTS

Page(s)

TABLE OF AUTHORITIES

Page(s)

Cases

Statutes

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. 11,169,775 (Ex. 1001, "the '775 Patent"), challenging Claims 1-5, 7-14, and 16-23 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 because they fail to disclose all of the elements required by the Challenged Claims for several reasons.

First, Google offers no evidence that it would have been obvious to use large numbers of ***low precision processing high dynamic range (LPHDR) elements*** in a computer architecture. The '775 Patent covers a computer architecture comprising a very large number of processing elements that each perform low precision high dynamic range (LPHDR) arithmetic operations that represent values from one millionth up to one million with a precision of about 0.1%, which, if represented and manipulated using floating point arithmetic, would have binary mantissas of no more than 10 bits and binary exponents of at least 5 bits ('775 Patent, 5:59 to 6:4). The '775 Patent teaches that POSAs, at the time of the invention, believed such a

massively parallel LPHDR computer architecture to be "not useful . . . and even worse." And yet, though Google does not dispute this understanding of a POSA's opinion at the time of the invention, its expert does not address why a POSA would have disregarded this understanding to incorporate Shirazi's LPHDR floating point values into each of the processing elements in Stuttard. Indeed, despite multiple IPR petitions, Google has neither disputed Singular's understanding that a POSA would have found a massively parallel LPHDR computer architecture useless or worse, nor has it provided any reference suggesting such an architecture.

Second, in an attempt to meet the Challenged Claims' unique combination of a large number of LPHDR processing elements arranged in a two-dimensional array, Google draws from two incompatible references: i) Stuttard, which describes modifications that can be made to a three-part hybrid computer architecture to enable all three parts to be operated in parallel under the same programming model and instruction set; and ii) Shirazi, which describes two custom reduced bit-width numerical formats that would enable a single arithmetic circuit to be implemented on a single field-programmable gate array (FPGA) chip. Google does not address the fundamental differences between the aims of these references. A POSA would not find them compatible.

Third, Google proposes to fundamentally redesign Stuttard's processing element topology to meet the two-dimensional processing element array limitations

of the claims, in a manner not suggested by Shirazi. Google makes no showing that a POSA would be motivated to do so.

The foregoing deficiencies in the Petition relate to the key features of the '775 Patent's claimed architecture. Accordingly, the Board should deny institution.

II. THE '775 PATENT

The '775 Patent is entitled "Processing with compact arithmetic processing element" and issued on November 9, 2021. The '775 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. Claims from two of the patents that have issued from this chain have been found valid by this Board in previous *inter-partes* reviews brought by the same Petitioner. *E.g.*, *Google LLC v. Singular Computing LLC*, IPR2021-00155, Paper No. 62 (P.T.A.B. May 11, 2022) (finding claims not unpatentable). As indicated above, the '775 Patent is directed to a computer processor that economically achieves a massive degree of parallelism through an array of multiplier circuits that each perform arithmetic on values having relatively high dynamic range exponents and low precision mantissas . *See* '775 Patent at 2:1-15; 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 '775 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 ’775 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 arithmetic:

> 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.

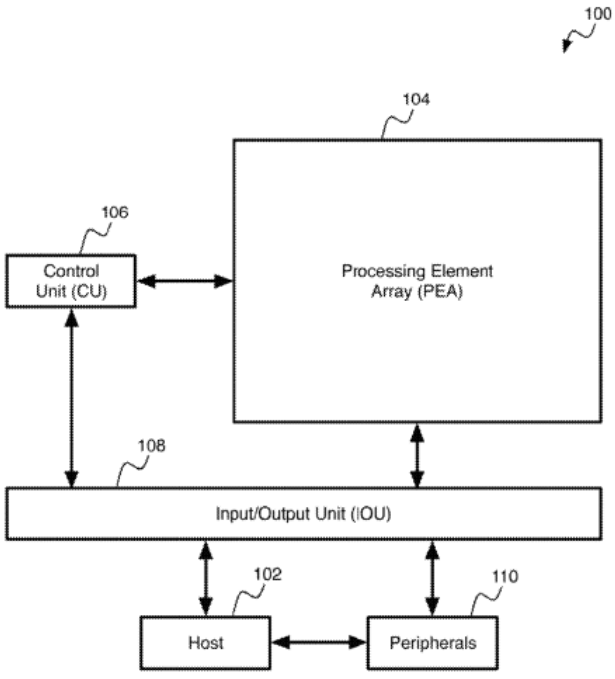
By using an LPHDR architecture as claimed,

> 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 '775 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 low precision high dynamic range (*i.e.*, shorter mantissa) formats.

In particular, the '775 Patent teaches a new computing architecture based on a massively parallel array of LPHDR processing elements (PEA). '775 Patent at 8:12-28. An embodiment is shown in Figure 1:



Id., Fig. 1.

In this embodiment, the "Host 102 is responsible for the 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 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 this claimed embodiment of the invention, "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 under the plain meaning 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. Telefex 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 PETITION SHOULD BE DENIED INSTITUTION BECAUSE PETITIONER HAS FAILED TO SHOW A REASONABLE LIKELIHOOD OF SUCCESS FOR ANY OF GROUNDS 1A-1C

A. Ground 1A: Claims 1-5 Are Not Obvious Over Stuttard in Combination with Shirazi

1. The Petition Fails to Demonstrate that A POSA Would Combine Stuttard and Shirazi as Google Proposes

Google's obviousness arguments rely on modifying Stuttard (Ex. 1006) to include portions of Shirazi's (Ex. 1036) functionality into Stuttard's row of PEs, and then further modifying Stuttard to turn a one-dimensional (1D) row of PEs into a two-dimensional (2D) array, even though both Shirazi and Stuttard lack such a 2D array teaching. However, the Petition fails to set out a sufficient motivation to make either modification, let alone such a combination.

(a) A POSA Would Not Have Stacked Large Numbers (*e.g.*, 5000) of Low Precision High Dynamic Range PEs in any Computer Architecture, Including the One Described by Stuttard

Google elides over a central advance of the '775 Patent: the discovery that large amounts of LPHDR arithmetic performed by multiple low precision processing elements in a massively parallel architecture are, in fact, useful and provide significant practical benefits. '775 Patent, 7:3-38. Google has presented no evidence that ***anyone*** had reached this discovery prior to Dr. Bates. Indeed, Shirazi

and Stuttard, though published 15 and 7 years before the '775 Patent, respectively, contain no hint that large numbers of low precision and high dynamic range calculations could provide useful results for ***any application***. This discovery, and the related architecture covered by the Challenged Claims, were non-obvious for so long in large part because it was surprising and counter-intuitive that performing so many individual arithmetic operations sequentially at low precision can lead to a final result with less error than each such individual arithmetic operation. '775 Patent, 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.,* 20:46-48 ("[t]o perform many calculations sequentially with 1% error and yet produce a final result with less than 1% error may seem counter-intuitive."); *id.* 24:33-42. In this IPR, and indeed in all of its other IPRs, Google simply assumes that a POSA would be aware of this discovery, despite providing ***no evidence*** whatsoever to this end.

Google points to the teachings of Shirazi and other references it has cited in its past IPRs (*i.e.*, Tong, Ex 1034) in an attempt to show that the discovery of the '775 Patent was, in fact, known by a POSA. But those references say nothing about scale in the number of processing elements performing imprecise LPHDR arithmetic in parallel, with all processing elements repeatedly performing imprecise LPHDR arithmetic. Indeed, by failing to proffer any contrary evidence, Google tacitly admits

that ***no one*** before Dr. Bates conceived of a computer architecture comprising large numbers of LPHDR processing elements in a massively parallel configuration. At a minimum, it was incumbent on Google to provide evidence *why* a POSA would have found it obvious to make this combination despite the '775 Patent's observation that such a combination would have been rejected by a POSA as useless. Google's failure to offer such evidence renders the Petition deficient because the only evidence of record is the '775 Patent's teaching that such an architecture was not obvious to a POSA.

(b) A POSA Would Not Discard Stuttard's Flexibility for Shirazi's Custom Formats

Separate and apart from the foregoing fundamental deficiency of the subject Petition, it also would not have been obvious for a POSA to adapt Stuttard to use Shirazi's reduced bit width formats because doing so would not have furthered any expressed goal of the Stuttard system. Petitioner quotes a part of Stuttard that allegedly states its disclosed architecture is for "high performance, high data rate processing," in an attempt to make the Shirazi number format seem like an obvious improvement to the Stuttard system. Pet. at 12 (quoting Stuttard, 8:24-28). But such a vague aspiration by itself does not amount to a motivation to adopt Shirazi's reduced bit widths (*i.e.*, to shrink PE size) in a Stuttard system, as Petitioner states.

Such a leap requires a very particular definition of "performance" and "data rate processing" for which there is no support in Stuttard.

More importantly, the part of Stuttard that Petitioner quotes was crudely cropped by Petitioner to obscure how a POSA would improve the Stuttard architecture. In full, the quote reads as follows: "The Multi-Threaded Array Processing (MTAP) architecture has been developed *to address a number of problems* in high performance, high data rate processing." Stuttard, 8:24-25 (emphasis added). That is, the Stuttard architecture is intended and designed to address programming problems, not the sort of performance problems that could be addressed by Shirazi. *Id.*, 1:25 to 2:8. For example, a POSA would not turn to Shirazi to solve the programming problem of a SIMD array controller not working with a standard compiler. *Id.*, 1:25-28. Nor would a POSA turn to Shirazi to solve Stuttard's other identified programming problem of having to generate code for programming the SIMD array controller completely separately from the main application code for programming a host processor. *Id.*, 1:27-28. Why a POSA trying to solve these programming problems would turn to the reduced bit width formats of an FPGA hardware design paper—Shirazi—is a mystery that is left unexplained in the Petition.

In keeping with the stated purpose of the Stuttard system, a POSA looking to improve it would have sought out improvements that make its SIMD array controller

more programmable. Examples of such improvements in Stuttard include making the processing elements (PEs) comprising the array as similar as possible to other more programmable parts of the architecture. *Id.*, 7: 14-16 ("[t]he functionality supported by each PE is then made as similar as possible to the mono execution unit: each PE has an ALU, a register file and memory, and supports a range of addressing modes for transferring data between memory and registers."). Doing so provides that, "[f]rom a programmer's perspective the MTAP processor appears as a single processor running a single C program." *Id.*, 10:28-32. Google does not even attempt to show how Shirazi's teaching of reduced bit formats would help a POSA in such programming-focused efforts.

Shirazi seeks to use reduced bit width floating point formats to fit an arithmetic circuit onto an FPGA. *Id.* To that end, Shirazi discusses two custom data formats, an 18-bit floating point format used specifically for a 2-D FFT application, and a 16-bit floating point format to be used for an FIR filter application. Shirazi, 156. The 18-bit format uses a 7-bit exponent and 10-bit mantissa; the 16-bit format uses a 6-bit exponent and 9-bit mantissa. *Id.* These formats in no way make the Stuttard system more programmable, meaning a POSA would not have been motivated to look to Shirazi to improve the Stuttard system.

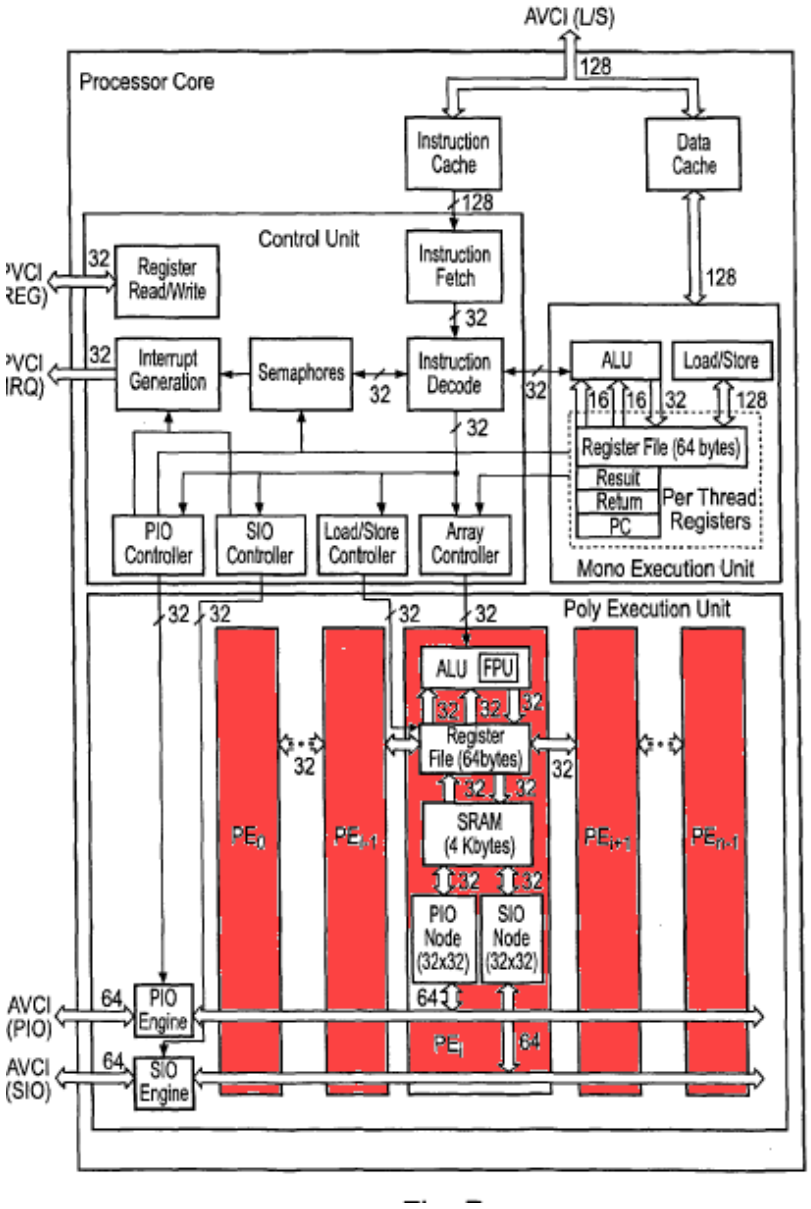
The Petition does not address the incompatibility between the goals of Stuttard and Shirazi. *See* Pet. at 6-20. Accordingly, Google has failed to demonstrate that a

POSA would be motivated to modify Stuttard's more programmable architecture to include Shirazi's custom number formats.

(c) A POSA Would Not Modify Stuttard From a One-Dimensional Array to a Two-Dimensional Array

Google further argues that a POSA would deploy Stuttard's array of PEs as a two-dimensional array to meet the '775 Patent's multiple "edge PE" limitations.[1] Pet. at 14-19. Stuttard however, describes its array of PEs as one-dimensional and shows it as such in the Figures (array of PEs denoted in red below).

[1] The "edge PE" limitations are "a first edge processing element positioned at a first edge of the processing element array, a second edge processing element positioned at the first edge of the processing element array,"; "a first processing element connection connecting the first edge processing element with the first interior processing element;" and "a second edge processing element connection connecting the second edge processing element with the second interior processing element" ('775 Patent, cl. 1) and similar limitations in Claims 7 and 16. Google denotes these limitations as [1C], [1D], [7C], [7E], [16B], and [16C].



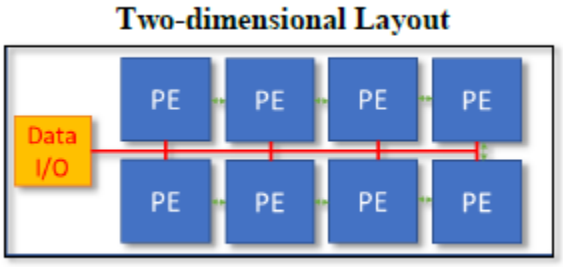
Google essentially argues that because it was known that arrays could be arranged in two dimensions, a POSA would choose to do so in Stuttard in particular. Pet. at 14-19. For example, Google cites passages from two other references to speculate that "other than linear interconnect[ed] topologies could be introduced" or that PEs could be arranged in "a 2D grid array." Pet. at 17-18 (annotations omitted).

But these statements show nothing about whether a POSA would modify Stuttard *in particular* into such an arrangement. For example, Ex. 1108, authored by Stuttard, does not mention a 2D arrangement at all; just that topologies other than linear ones could be introduced. Google does not explain why "other than linear" must be a 2D array; it could as easily be a ring, star, or other topology. *See* Pet. at 17 (citing Ex. 1108). Similarly, Ex. 1081 states that "[p]referably, a 1D line array is provided." Ex. 1081, 5:65. And even if one were to make the logical leap required to equate "other than linear" with "2D", Google does not address why the Stuttard authors, in full possession of the concept of a 2D arrangement (again, *see* Ex. 1108), described only a 1D arrangement with full connectivity between each PE and the shared bus, while not even mentioning a 2D arrangement in the publication at the heart of the subject Petition.

Google also does not address that even its own evidence at times teaches away from using a 2D arrangement. Pet. at 18 (citing Ex. 1108, 1081). Indeed, generalizing the array to 2D requires much engineering, rewiring, area, and power increases, and all this would require significant experimentation which Google does not flag, let alone address. *See* Pet. at 14-19. A POSA would not be motivated to conduct this exercise.

Google also states that a POSA would have made the modification of adopting a 2D array so as to have "facilitat[e] applications that benefited from full-2D

connectivity between PEs." Pet. at 18. In Google's words, "full-2D connectivity" means that every PE is directly connected to the four PEs surrounding it, *i.e.*, there exist "up, down, left, right" connections between PEs. *Id.* Indeed, Google characterizes that as one of the benefits of 2D connectivity that might motivate a POSA to alter the topology from 1D to 2D. *Id.* However, Google's proposal would have a central data bus shared between all PEs, in addition to a direct "swazzle" connection for each PE, connecting each PE to the PE to its immediate left and the PE to its immediate right.[2] Google shows its proposed alteration in the Figure below, with the shared bus in red and the swazzle connections in gray.



[2] As Stuttard explains, a "swazzle" connection is a direct connection between a PE and a PE to its left or right. Stuttard, 20:31-34 ("the PEs are able to communicate with one another via what is known as the *swazzle* path that connects the register file of each PE with the register files of its left and right neighbours.").

Pet. at 17.[3] Thus, Google's combination, having only a shared bus to effect "up" and "down" connections, would lack the *direct* "up" and "down" connections that form its primary reason for making the alteration.

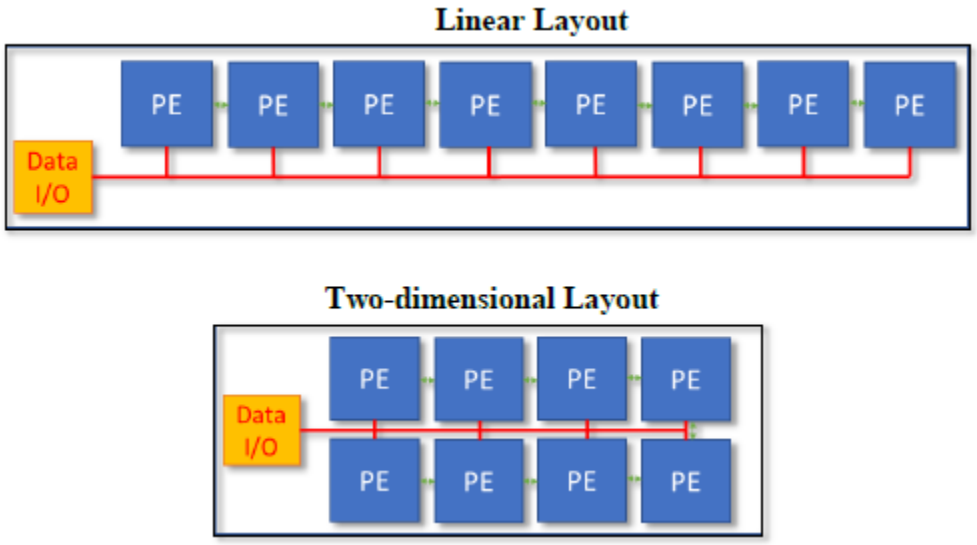
Moreover, even though Google states that a POSA would have made the modification of adopting a 2D topology to facilitate applications that "benefited from full-2D connectivity," Google fails to identify any such applications. Pet. at 18; Ex. 1003, ¶ 102.

Additionally, Google omits any explanation of how to prevent its proposed Stuttard modification of a single shared bus (red) from creating a bottleneck for data transfers to and from the PEs. Google might have made such an omission because dealing with the bottleneck issue necessarily involves dealing with the related issue of moving data into and out of the faster PEA without absorbing large amounts of hardware resources, which is a critical design goal that motivated the adoption of the 2D array in the '775 Patent, and which *is not mentioned at all in Stuttard or Shirazi*. That is, in contrast to Stuttard or Shirazi, the '775 Patent recognized the bottleneck

[3] Google's expert states that a 2D swizzle path ("an interconnection between North-East-West-South . . . would have also been a conventional, and advantageous interconnection. . . .") (Ex. 1003, ¶ 101), but provides no evidence or opinion beyond his bare conclusions. The Petition is silent as to such a connection.

issue and the related hardware resource absorption issue, by expressly avoiding both (i) a shared bus and (ii) any direct connections between the vast majority of PEs in the PE array (PEA), and the rest of the computer (*i.e.*, the IO unit). '775 Patent, 9:60-10:11. This teaching in the '775 Patent is manifested in both the edge PE and interior PE limitations of the Challenged Claims (*i.e.*, connecting the IOU to the PEs only at the edges of the PEA, and not connecting the IOU to the interior PEs without going through at least one edge PE). Indeed, as the '775 Patent notes, a shared connection between all PEs and the IO unit would require additional resources on each PE and create lengthy connections. *Id.* Thus, the '775 Patent connects its IO unit to only the edge PEs. *Id.*

Google's other purported reasons for modifying Stuttard's one-dimensional topology similarly lack merit. Google first suggests that "physically arranging PEs in a two-dimensional array reduces the overall length of each channel." Pet. at 16. As support, Google asks the Board to compare two Figures, created by Google, and measure the pictured length of a channel. *Id.* at 16-17.



Pet. at 17. According to Google, the Figures above demonstrate that "the channel's length in the 2D array can beneficially be approximately half as long as in the 1D line." Pet. at 16. But Google provides no evidence, not even from its expert, that its Figures accurately represent the relative scale and relationship between the PEs. *See id.; see also* Ex. 1003, ¶ 99. Instead, Google apparently asks the Board to take its pictorial representation of a Stuttard chip as somehow representative of the actual dimensions of the PEs in the chip itself, measure features in its not-to-scale Figure (*i.e.*, as opposed to in the chip) to ascertain dimensions on the chip, and finally compare the unspecified lengths of the channels illustrated within those Figures to one another. The Board should decline to do so.

Google also argues that "POSAs knew it was desirable and conventional for computing chips to have aspect ratios (width-to-length ratios) closer to 1, because

very high or very low aspect ratios cause 'mechanical stress[es]' and 'adversely change the electrical properties of the IC circuitry.'" Pet. at 17. But Google's evidence says nothing about the topology of an array of PEs; instead, it discusses the aspect ratio of a computer chip, which is a ***semiconductor die*** that can hold multiple functional blocks besides an array of PEs. Ex. 1067, [0007] ("An IC fabricated on a semiconductor die having a very large or very small aspect ratio often suffers from mechanical stress caused by the molding compound used to package the die."); *see* Ex. 1003, ¶ 100 (quoting Ex. 1067). For example, Figure 7 of Stuttard shows a computer chip that includes an array of PEs, ***as well as*** a control unit, a mono execution unit, an instruction cache, and data cache. Indeed, Ex. 1067—the only evidence Google's expert cites for its proposition regarding aspect ratios—is directed specifically to monolithic integrated circuits and not arrays of PEs. Ex. 1067, Abstract, [0002]. Google and its expert do not attempt to show that a computer chip using Stuttard's design would have an aspect ratio other than 1 without moving to a two-dimensional PE topology, or that moving to a two-dimensional topology would make the aspect ratio of the computer chip closer to 1. *See* Pet. at 17; Ex. 1003, ¶ 100.

The Petition thus fails to set out a sufficient motivation for a POSA to modify Stuttard from utilizing a one-dimensional array to utilizing a two-dimensional array. Google's arguments as to "edge processing element[s]" are based solely on that

modification. *See, e.g.*, Pet. at 21-24. Accordingly, the Petition fails to show that the "edge processing element[s]" limitations are rendered obvious.

2. Google's Combination Does Not Disclose or Render Obvious "a first interior processing element" and "a second interior processing element" as Recited in Claim 1

Google asserts that some PEs in its hypothetical and unsupported two-dimensional modification of Stuttard-Shirazi are "interior." *Id.*

As discussed above, however, Google has not shown that a POSA would be motivated to make its modified, two-dimensional array. Accordingly, Google's combination cannot render this limitation obvious.

The plain language of the claims makes clear the distinction, which Google ignores, between "edge" and "interior" processing elements. An "edge processing element" is at the "edge" of the processing element array and thus receives data from the host or the input-output unit. An "interior processing element" is not at the "edge" and thus is inside the array. The claims further recite that the input-output unit is connected to the first and second edge processing units, and that the edge processing elements themselves have ***separate*** connections to the interior processing elements. *E.g.*, '775 Patent, cl. 1. A plain reading of the claims shows that the edge processing elements are connected to the input-output unit (and thus, at the logical edge of the array), and that the interior elements are not.

The specification further confirms this understanding. The '775 Patent explains that only edge processing elements are connected to the input-output unit because outputting signals from inside the PEA is wasteful and difficult. *E.g.*, '775 Patent at 9:57-65 ("driving signals from inside the PEA 104 out to the IOU 108 ***usually requires a physically relatively large driving circuit or analogous mechanism***. Having those at every PE may absorb much of the available resources of the hardware implementation technology (such as VLSI area)."). Moreover, "having independent connections from every PE to the IOU **108** means many such connections, and long connections, which also may absorb much of the available hardware resources." *Id.*, 9:65-10:1. Therefore, "the connections between the PEs and the IOU **108** may be limited to those PEs at the edges of the PE array **104**." *Id.* 10:3-5. Accordingly, in the claimed embodiment, "the 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.*, 10:4-7.

Even if a POSA would make Google's combination, it still does not render this limitation obvious because its asserted interior PEs are still connected to what Google identifies as the input-output unit, and thus are not at the edge of the array, which means there would be no interior PEs. Indeed, Google argues that the data channels, "which are connected to every PE in the array," are part of the input-output unit. Pet. at 29-30. As explained above, however, an interior PE cannot be

connected to the input-output unit. Accordingly, all of the PEs in Google's combination, even if reconfigured as a 2D array, are "edge processing elements;" indeed, none of them are "interior" in any respect.

Similarly, Google argues that the "bus" that it alternatively identifies as the input-output unit is used to connect "PEs to off-chip peripherals and a host system." Pet. at 26. For the same reasons, therefore, the bus would connect to each PE, and thus no PEs are "interior processing elements" as required.

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

3. Google's Combination Does Not Disclose or Render Obvious "an input-output unit" as Recited in Claim 1

Google fails to identify the claimed input-output unit connected to first and second edge processing elements.

Google states that either "the bus" or, "alternatively, the bus plus the I/O engines and channels" make up the claimed input-output unit. Pet. at 30.

At the outset, Google presents no evidence that a POSA would recognize "the bus" to be an input-output unit. The '775 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 102. . . ." '775 Patent, 9:6-18. By using an input-output unit to account for the data rate difference, the '775 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.

Accordingly, the Petition fails to show that the input-output unit is rendered obvious.

B. Ground 1B: Claims 7-14 and 16-23 Are Not Obvious Over Stuttard in View of Shirazi

As discussed in detail below, Google has not shown that independent claims 7 and 16 would have been obvious over Stuttard in view of Shirazi. Accordingly, neither those claims nor their dependent claims are obvious.

1. A POSA would not Modify Stuttard as Google Proposes

Google's obviousness arguments under this Ground depend entirely on its Stuttard-Shirazi combination discussed above. Pet. at 51-55. For the same reasons set forth with respect to Ground 1A, Google has not shown that a POSA would be motivated to modify Stuttard to adopt Shirazi's bit-width floating point formats and to convert Stuttard's one-dimensional PE array to a two-dimensional array.

2. Google's Combination Does Not Render Obvious "a processing element array comprising a plurality of first processing elements, wherein the plurality of first processing elements is no less than 5000 in number, wherein each of a first subset of the plurality of first processing elements is positioned at a first edge of the processing element array, and wherein each of a second subset of the plurality of first

processing elements is positioned in the interior of the processing element array;" [7C]

(a) The 5000-PE Limitation Carries Patentable Weight and Is Not Obvious

Google asserts that Claim 7's requirement of no fewer than 5000 processing elements carries no patentable weight. Pet. at 51-55. However, as Petitioner, Google has the burden to show that the claim is obvious, and it has introduced 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 '775 Patent's specification in support. Pet. at 53-54 (citing Ex. 1003, ¶ 170). As stated above with regard to Ground 1 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. '775 Patent, 7:3-38 (noting the prevailing view "that massive amounts of LPHDR arithmetic are" not useful), *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.,* 20:46-48 ("[t]o perform many calculations sequentially with 1% error and yet produce a final result with less than 1% error may seem counter-intuitive), *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. '775 Patent, 28:13-29:4. In the absence of any evidence to the contrary, the Board should consider this limitation, like all limitations, as part of the patented invention.

Google secondarily argues that the claimed number of the claimed PEs is a "result-effective variable." Pet. at 54-55. But Google makes this argument using only art regarding the number of *full-precision* PEs, not the number of LPHDR PEs as claimed. *Id.* As the '775 Patent itself notes, 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 ('775 Patent, 7:3-38, 24:33-42), 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. Such an unexpected result arising from increasing the number of LPHDR PEs belies the notion that the number of claimed PEs is a result-effective variable arising from routine experimentation.

Only one reference cited by Google in support of its argument that the claimed number of LPHDR PEs is somehow result-effective, Ex. 1014, even mentions lower-precision. That reference makes specific note that "precision reduction tolerance may vary across different physics engines" and does not suggest including anywhere near 5000 PEs. Ex. 1014, 9.

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

(b) Google Has Not Shown that "each of a first subset of the plurality of first processing elements is positioned at a first edge of the processing element array, and wherein each of a second subset of the plurality of first processing elements is positioned in the interior of the processing element array" [7C] Is Obvious

Google relies on the same arguments as with respect to Claim 1 for this limitation. Pet. at 66-67. For the same reasons as discussed above, a POSA would not modify Stuttard to meet this limitation.

3. 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;"[7D]

Google relies on the same arguments as with Claim 1's input-output unit limitation. Pet. at 67. For the same reasons as discussed above, therefore, Google's Ground 1B combination does not render this limitation obvious.

4. Google's Combination Does Not Render Obvious the "wherein the plurality of first arithmetic units each comprises a first corresponding multiplier circuit adapted to receive" Limitation [7J]

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

5. Google's Combination Does Not Render Obvious Limitations [16B], [16C], [16F], and [16G]

Google expressly relies on its claim 7 arguments as to these limitations. Pet. at 80. For the same reasons as discussed above, therefore, Google's Ground 1B combination does not render these limitations, and claim 16 as a whole, 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
Enrique W. Iturralde (Reg. No. 72,883)
Back-up 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
Email: vrubino@fabricantllp.com
Email: eiturralde@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. 11,169,775 specified in 37 C.F.R. § 42.24 has 6,510 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. 11,169,775 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: plambrianankos@fabricantllp.com