UNITED STATES DISTRICT COURT
WESTERN DISTRICT OF TEXAS
MIDLAND-ODESSA DIVISION

NEURAL AI, LLC, Plaintiff, v. NVIDIA CORPORATION, Defendant.	Case No. 7:24-cv-00221-ADA-DTG JURY TRIAL DEMANDED

PLAINTIFF NEURAL AI LLC'S RESPONSIVE CLAIM CONSTRUCTION BRIEF

TABLE OF CONTENTS

I. INTRODUCTION 1

II. PATENT BACKGROUND 1

III. LEVEL OF ORDINARY SKILL 3

IV. LEGAL PRINCIPLES OF CLAIM CONSTRUCTION 3

V. DISPUTED CLAIM TERMS 5

A. "accelerator" ('867 Patent cl. 16, '438 Patent cls. 1, 12, 44) 5

1. The Accelerator Need Not be Separate from the CPU 5
2. The Accelerator Is Not Limited to Hardware 8

B. "accelerator controller" ('867 patent, cls. 16, 17, 19) / "controller" ('438 patent, cls. 1, 7–10, 21, 40, 42, 44, 50–53) 10

C. "partition" ('867 patent, cls. 16, 18; '438 patent, cls. 26, 27, 40, 43, 49; '461 patent, cls. 21, 30) 13

D. "bank" ('438 patent, cls. 6, 20) 16

E. "swapping the first pointer with the second pointer" ('867 patent, cl. 16) 19

F. "input data at first rate; and . . . sequence of computations at a second rate" ('438 patent, cls. 3, 14, 46) 22

1. The First Rate Need Not Be "An Amount of Data Received Per Unit Time" 23
2. The Second Rate Need Not Be "Number of Computations Per Unit Time" 24

G. Order of Steps ('867 patent, cl. 16; '438 patent, cls. 12, 21) 25

VI. CONCLUSION 29

TABLE OF AUTHORITIES

Page(s)

Cases

3M Innovative Props. Co. v. EnvisionWare, Inc.,
2010 WL 5067449 (D. Minn. Dec. 6, 2010).....13

3M Innovative Props. Co. v. Tredegar Corp.,
725 F.3d 1315 (Fed. Cir. 2013).....4, 19

Altiris, Inc. v. Symantec Corp.,
318 F.3d 1363 (Fed. Cir. 2003).....25

Amgen Inc. v. Hoechst Marion Roussel, Inc.,
314 F.3d 1313 (Fed. Cir. 2003).....14

Ancora Technologies, Inc. v. LG Electronics Inc.,
No. 1-20-CV-00034-ADA, 2020 WL 4825716 (W.D. Tex. Aug. 19, 2020) 26, 27

*Apple Inc. v. MPH Techs. O*y,
28 F.4th 254 (Fed. Cir. 2022) 4

Apple Inc. v. Omni MedSci, Inc.,
No. 2023-1034, 2024 WL 3084509 (Fed. Cir. Jun. 21, 2024).....17

Apple Inc. v. Wi-LAN Inc.,
25 F.4th 960 (Fed. Cir. 2022) 15

Becton, Dickinson & Co. v. Tyco Healthcare Grp., LP,
616 F.3d 1249 (Fed. Cir. 2010).....6

CA, Inc. v. Netflix, Inc.,
No. 2023-1768, 2025 WL 303436 (Fed. Cir. Jan. 27, 2025).....9

E-Pass Techs., Inc. v. 3Com Corp.,
343 F.3d 1364 (Fed. Cir. 2003).....7

Hill-Rom Servs., Inc. v. Stryker Corp.,
755 F.3d 1367 (Fed. Cir. 2014).....4

Interactive Gift Express, Inc. v. Compuserve Inc.,
256 F.3d 1323 (Fed. Cir. 2001).....25

Intervet Am., Inc. v. Kee-Vet Labs., Inc.,
887 F.2d 1050 (Fed. Cir. 1989).....5

Iris Connex, LLC v. Acer Am. Corp.,
2016 WL 4596043 (E.D. Tex. Sept. 2, 2016) 15

Kaneka Corp. v. Xiamen Kingdomway Grp. Co.,
790 F.3d 1298 (Fed. Cir. 2015) 26

Liebel-Flarsheim Co. v. Medrad, Inc.,
358 F.3d 898 (Fed. Cir. 2004) 8

Merck & Co., Inc. v. Teva Pharms. USA, Inc.,
347 F.3d 1367 (Fed. Cir. 2005) 4

Mformation Techs., Inc. v. Research in Motion Ltd.,
764 F.3d 1392 (Fed. Cir. 2014) 25

Nat'l Cheng Kung Univ. v. Samsung Elecs. Co.,
2014 WL 2885380 (E.D. Tex. June 25, 2014) 13

O2 Micro Int'l Ltd. v. Beyond Innovation Tech. Co.,
521 F.3d 1351 (Fed. Cir. 2008) 16

Oatey Co. v. IPS Corp.,
514 F.3d 1271 (Fed. Cir. 2008) 9, 15

Omega Eng'g, Inc., v. Raytek Corp.,
334 F.3d 1314 (Fed. Cir. 2003) 4

Orenshteyn v. Citrix Sys., Inc.,
341 F. App'x 621 (Fed. Cir. 2009) 12

Phillips v. AWH Corp.,
415 F.3d 1303 (Fed. Cir. 2005) (en banc) 3, 4, 5, 15

Promptu Sys. Corp. v. Comcast Corp.,
92 F.4th 1372 (Fed. Cir. 2024) 8

Regents of Univ. of Minnesota v. AGA Med. Corp.,
717 F.3d 929 (Fed. Cir. 2013) 11

Renishaw PLC v. Marposs Societa' per Azioni,
158 F.3d 1243, 1248 (Fed. Cir. 1998). 4, 28

Retractable Techs., Inc. v. Becton, Dickinson & Co.,
653 F.3d 1296 (Fed. Cir. 2011) 4, 19, 20

Signify N. Am. Corp. v. Lepro Innovation Inc.,
No. 222CV02095JADDJA, 2023 WL 8435567 (D. Nev. Dec. 4, 2023) 13

Sisvel Int'l S.A. v. Sierra Wireless, Inc.,
81 F.4th 1231, 1236 (Fed. Cir. 2023) 4, 6, 20, 23

Tandon Corp. v. U.S. Int'l Trade Comm'n,
831 F.2d 1017 (Fed. Cir. 1987).....17

Thorner v. Sony Computer Ent. Am. LLC,
669 F.3d 1362 (Fed. Cir. 2012)..... *passim*

Traxcell Techs., LLC v. Cellco P'ship,
No. 6:20-CV-01175-ADA, 2023 WL 2415583 (W.D. Tex. Mar. 8, 2023).....5

Vitronics Corp. v. Conceptronic,
90 F.3d 1576 (Fed. Cir. 1996).....5

Voice Tech Corp. v. Unified Pats., LLC,
110 F.4th 1331 (Fed. Cir. 2024) 14

I. INTRODUCTION

Claim construction is not necessary for the seven terms Defendant NVIDIA Corporation ("NVIDIA") disputes before the Court. Each term has a plain and ordinary meaning that would be understood by a person of ordinary skill in the art at the time of the effective dates of the Asserted Patents. Nothing in the claims themselves, the specifications, or the file histories demonstrate that the patentee either deviated from the plain and ordinary meaning of the disputed terms or made a clear and unmistakable disavowal of claim scope. In these circumstances, the Federal Circuit has repeatedly held that the plain and ordinary meaning must control, that adding unclaimed limitations under the guise of construction is improper, and that extrinsic evidence cannot be used to create an ambiguity or contradict the intrinsic evidence.

NVIDIA nonetheless asks the Court to adopt results-oriented constructions that narrow and distort clear claim language. NVIDIA's departure from the plain and ordinary meaning violates numerous canons of construction. As the patentee never disavowed claim scope or acted as its own lexicographer with respect to the claim terms at issue, NVIDIA's constructions are unwarranted. NVIDIA's proposed constructions also do nothing to assist the jury in understanding Plaintiff Neural AI, LLC's ("Neural AI's") patents. Instead, NVIDIA attempts to sow confusion and manufacture non-infringement arguments by rewriting unambiguous claim language and using vague phrases divorced from the claims.

NVIDIA cannot overcome the heavy presumption that the claim terms carry their ordinary meaning. As such, NVIDIA's constructions must be rejected. The Court should adopt each term's plain and ordinary meaning.

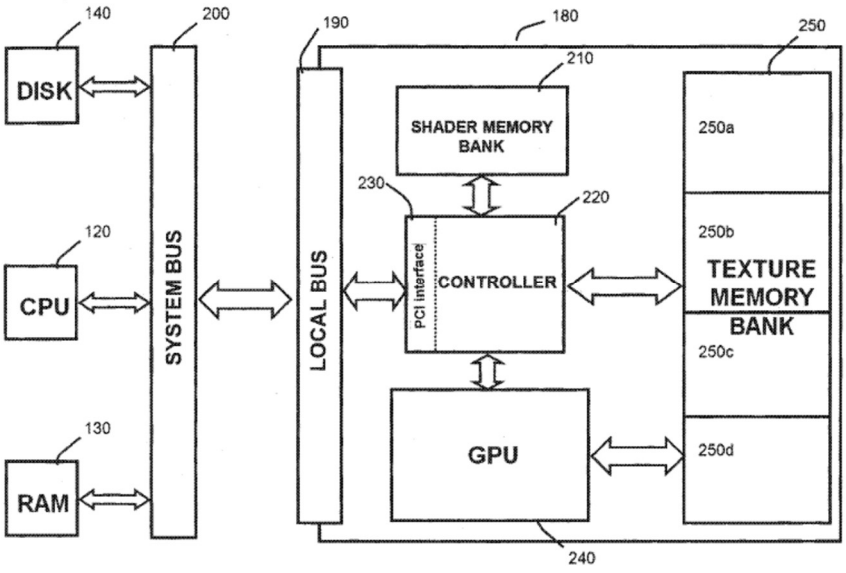
II. PATENT BACKGROUND

U.S. Patent Nos. 8,648,867 ("'867 Patent"), RE49,461 ("'461 Patent"), and RE48,438 ("'438 Patent") (collectively, the "Asserted Patents") are each titled "Graphic Processor Based

NVIDIA Ex1041
NVIDIA Corp. v. Neural AI, LLC
IPR2025-00608
Page 6 of 37

Accelerator System and Method." The Asserted Patents regard a computer "accelerator system" having "one or more graphics processing units (GPU)." Ex. 1 ('867 Patent), Abstract.[1] The accelerator system includes a "controller [that] handles most of the primitive operations needed to set up and control GPU computation." *Id.* As a result, the computer's central processing unit (CPU) is freed from this function and is dedicated to other tasks. *Id*. The "input/output data" processed by the accelerator is "exchanged between CPU" and GPU(s) by the controller. *Id.* at Abstract, 16:26–33. In the accelerator system, "results from the previous time step are used but not changed," and thus beneficially "the results are preferably transferred back to CPU in parallel with the computation." *Id.* at Abstract, 2:24–26.

As illustrated in Fig. 2 of the '867 Patent below, an example embodiment can include a CPU 120, controller 220, GPU 240, and memories 210 and 250 associated with the GPU. '867 Patent, Fig. 2, 4:51–60.



'867 Patent, Fig. 2

[1] The Asserted Patents share a common specification.

The ’867 Patent invention can include the CPU “transfer[ring] the input data for a simulation to the accelerator, after which the accelerator executes simulation computations to generate the output data, which is transferred to the [CPU].” *Id.* at 2:41–45.

The ’438 Patent further describes a novel accelerator system where “the sequence of computations represent[s] [an] artificial neural network.” Ex. 2 (’438 Patent), 14:62–64. For instance, “[c]omputational elements are same or similar within a large population and are computed in parallel. An example of such a population is a layer of neurons in an artificial neural network (ANN), where all neurons are described by the same equation.” *Id.* at 5:39–44.

The ’461 Patent describes an improved computer system that can include “queueing a user command received by [a] user interaction stream during execution of the computations representing the artificial neural network.” Ex. 3 (’461 Patent), 16:65–67. In one example embodiment of the computer system, a user “interacts with the software through the user interaction section” of a “graphic user interface [] executed on the CPU.” *Id*. at 8:17–19.

III. LEVEL OF ORDINARY SKILL

A person of ordinary skill in the art at the time of the invention would have the equivalent of a Bachelor of Science degree in electrical engineering, computer science, or a related field and at least two years of experience in the field of computer science and/or electrical and computer engineering. Alternatively, the person of ordinary skill in the art would have a master’s degree in electrical and computer engineering, computer science, or similar discipline.

IV. LEGAL PRINCIPLES OF CLAIM CONSTRUCTION

Generally, the words in a claim should be given their “ordinary and customary meaning,” and “[t]he construction that stays true to the claim language and most naturally aligns with the patent’s description of the invention will be, in the end, the correct construction.” *Phillips v. AWH Corp.*, 415 F.3d 1303, 1313, 1316 (Fed. Cir. 2005) (en banc) (citation omitted). The plain and

ordinary meaning is that which someone of ordinary skill in the art at the time of the effective date of the patent application would ascribe to a term when read in the context of the claim, specification, and prosecution history. *Apple Inc. v. MPH Techs. O*y, 28 F.4th 254, 259 (Fed. Cir. 2022). The two exceptions for deviating from a term's plain and ordinary meaning and adopting a construction are "1) when a patentee sets out a definition and acts as his own lexicographer, or 2) when the patentee disavows the full scope of a claim term either in the specification or during prosecution." *Sisvel Int'l S.A. v. Sierra Wireless, Inc.*, 81 F.4th 1231, 1236 (Fed. Cir. 2023) (citation omitted).

The "standards for finding lexicography and disavowal are exacting." *Hill-Rom Servs., Inc. v. Stryker Corp.*, 755 F.3d 1367, 1371 (Fed. Cir. 2014). Lexicography requires the patentee to "clearly express an intent to redefine a term." *Id.* (citation omitted). "[T]o disavow claim scope, the specification must contain expressions of manifest exclusion or restriction, representing a clear disavowal of claim scope."). *Retractable Techs., Inc. v. Becton, Dickinson & Co*., 653 F.3d 1296, 1306 (Fed. Cir. 2011) (quotation omitted). And any disavowal of claim scope made during prosecution "must be both clear and unmistakable." *3M Innovative Props. Co. v. Tredegar Corp*., 725 F.3d 1315, 1325 (Fed. Cir. 2013). Courts accordingly "indulge a 'heavy presumption' that claim terms carry their full ordinary and customary meaning." *Omega Eng'g, Inc., v. Raytek Corp*., 334 F.3d 1314, 1323 (Fed. Cir. 2003) (citation omitted).

Claim construction "begins and ends in all cases with the actual words of the claim." *Renishaw PLC v. Marposs Societa' per Azioni*, 158 F.3d 1243, 1248 (Fed. Cir. 1998). Courts look to "the context in which a term is used in the asserted claim," as well as "[o]ther claims of the patent in question." *Phillips*, 415 F.3d at 1314. In addition, "claims must be construed so as to be consistent with the specification." *Merck & Co. v. Teva Pharms. USA, Inc.*, 347 F.3d 1367, 1371

(Fed. Cir. 2003). But "limitations appearing in the specification will not be read into claims." *Intervet Am., Inc. v. Kee-Vet Labs., Inc.*, 887 F.2d 1050, 1053 (Fed. Cir. 1989).

The patent prosecution history may shed light on the meaning of terms; however, "because the prosecution history represents an ongoing negotiation between the PTO and the applicant, rather than the final product of that negotiation, it often lacks the clarity of the specification and thus is less useful for claim construction purposes." *Phillips*, 415 F.3d at 1317. Courts may also rely on extrinsic evidence such as "expert and inventor testimony, dictionaries, and learned treatises," but such evidence is "less significant than the intrinsic record." *See id*. (citation omitted). "[E]xtrinsic evidence in general, and expert testimony in particular, may be used only to help the court come to the proper understanding of the claims; it may not be used to vary or contradict the claim language." *Vitronics Corp. v. Conceptronic*, 90 F.3d 1576, 1584 (Fed. Cir. 1996).

V. DISPUTED CLAIM TERMS

A. "accelerator" ('867 Patent cl. 16, '438 Patent cls. 1, 12, 44)

Neural AI's Construction	NVIDIA's Construction
Plain and ordinary meaning	Hardware separate from the CPU

The term "accelerator" does not require construction because its meaning is understood by POSITAs. NVIDIA's construction attempts to add unclaimed limitations: that the "accelerator" is (1) separate from the CPU and (2) hardware. Both limitations are unfounded.

1. The Accelerator Need Not be Separate from the CPU

The claims are silent as to the location of the accelerator relative to the CPU and are agnostic on whether they are "separate from" one another. Because the claims do not recite the accelerator's location, "the Court [should] decline[] to impose a requirement on where the [accelerator] is located." *Traxcell Techs., LLC v. Cellco P'ship*, No. 6:20-CV-01175-ADA, 2023

WL 2415583, at *4–5 (W.D. Tex. Mar. 8, 2023) (rejecting the construction that the second processor was "separate from" the wireless device and adopting plain and ordinary meaning).

NVIDIA attempts to support its construction with the claim language but conflates logical distinction with physical separation. Op. Br. at 5. For example, Claim 16 of the '836 Patent teaches a method performed on "a computer system including a central processing unit and an accelerator," saying nothing whatsoever about physical separation.[2] Ex. 1, 16:21–22. NVIDIA cites *Becton, Dickinson & Co. v. Tyco Healthcare Grp., LP*, 616 F.3d 1249, 1254 (Fed. Cir. 2010)), but *Becton* held that elements listed separately in a claim are ***logically distinct*** and "logically cannot be one and the same." 616 F.3d at 1254 (quotation omitted). NVIDIA's citation does not support the incorrect proposition that the ordering, listing, or indentation of claim elements limits ***physical structure***.

NVIDIA's attempt to limit the term "accelerator" with an embodiment disclosed in the specification, Op. Br. at 5–6, fails under well-established law. *See Sisvel*, 81 F.4th at 1236 (warning that, even when the specification "describes very specific embodiments," courts should not "confin[e] the claims to those embodiments." (citation omitted)). The canon against reading limitations from embodiments applies even when "the only embodiments, or all the embodiments contain a particular limitation. [Courts] do not read limitations from the specification into claims." *Thorner v. Sony Computer Ent. Am. LLC*, 669 F.3d 1362, 1366 (Fed. Cir. 2012).

Even if the Court were to entertain NVIDIA's attempt to limit claims to a particular embodiment, the embodiment NVIDIA cites actually undermines its argument. NVIDIA's cited embodiment teaches that the CPU is "***preferably*** independent" (and not "***necessarily***

[2] NVIDIA incorrectly states that Claim 16 discloses that the accelerator and CPU are "operably coupled." *See* Op. Br. at 5. This language appears nowhere in Claim 16, although it is in unasserted Claims 1 and 9. Regardless, a POSITA would not understand "operably coupled" elements to require physical separation. *See* Ex. 5 ¶ 34.

independent") of the accelerator. '867 Patent at 4:8–12 (emphasis added). Therefore, the cited embodiment does not support NVIDIA's construction that requires the CPU to be independent.

NVIDIA's construction fares no better with extrinsic evidence. Neural AI's expert Dr. Chandra Bajaj explains that requiring the accelerator to be "separate from" the CPU would preclude various implementations of the claim scope that would fall within the plain and ordinary meaning of the term as understood by a POSITA. *See* Ex. 5 (Bajaj Declaration), ¶ 33. For example, the claim language, as understood by a POSITA, would include "well-understood implementations," such as "many core CPU's and multiple CPU configurations" among others. *Id.* And NVIDIA's proffered dictionary definition is silent on the physical location of the accelerator relative to the CPU. ECF No. 71-1 at 7 (defining "accelerator" as "[a] special circuit board that is placed within a computer to speed up some aspect of its operation"). Other contemporaneous definitions of "accelerator" are also agnostic on its location. *See* Ex. 4 (Microsoft Computer Dictionary 2002), at 13 ("accelerator *n.* . . . 2. In hardware, a device that speeds or enhances the operation of one or more subsystems, leading to improved program performance.").

NVIDIA asserts that rejecting its physical separation limitation would "eviscerate" the invention's purpose, which NVIDIA claims is limiting computations on the CPU. Op. Br. at 6. NVIDIA fails to connect the dots on how physical separation between the accelerator and the CPU affects the invention's purpose. The disclosed patent inventions could still limit computations on the CPU, as the patent's controller handles most of the primitive operations needed to set up and control GPU computation, freeing up the CPU, regardless of where the accelerator and CPU are located and independent of whether they are "separate from" each other. In any event, the Court cannot read unclaimed limitations into the term based on NVIDIA's perceived purpose of the invention. *See E-Pass Techs., Inc. v. 3Com Corp.*, 343 F.3d 1364, 1370 (Fed. Cir. 2003) ("The

court's task is not to limit claim language to exclude particular devices because they do not serve a perceived 'purpose' of the invention. . . . An invention may possess a number of advantages or purposes, and there is no requirement that every claim directed to that invention be limited to encompass all of them."). If anything, NVIDIA's "separation" limitation—which is at odds with the claim language and appears designed to bolster non-infringement arguments NVIDIA hopes to make—merely begs the question of ***how much separation*** is required. NVIDIA's proposed addition of "separate from" will create uncertainty and confuse the jury. *See Promptu Sys. Corp. v. Comcast Corp.*, 92 F.4th 1372, 1381 (Fed. Cir. 2024) ("[A] claim construction, if needed at all, should help resolve, *not add to*, uncertainty in the understanding the finder of fact is to use in applying a claim term." (emphasis added)).

2. The Accelerator Is Not Limited to Hardware

The claims do not limit the accelerator to hardware, and merely require the accelerator—which, for example, "receiv[es]" "input data"—to be functionally programmable. Ex. 1, 16:23–24. The plain language permits the claimed functions to be performed by hardware, software, or a combination of hardware and software.

NVIDIA contends that because Claims 1 and 44 of the '438 Patent disclose an accelerator comprising what NVIDIA asserts are "specific hardware sub-components," such as a "graphics processing unit" (Claim 1) and "accelerator memory" (Claim 44), the accelerator must be "hardware based." Op. Br. at 7. As a threshold matter, NVIDIA fails to explain why "hardware based" components must be hardware *alone*, as opposed to a combination of hardware and software. NVIDIA does not justify its "hardware based" requirement either, and merely attempts to rely on a strict "hardware implementation" as a disclosure in the specification. *Id.* First, it is well established that specification examples are not limiting. *See Liebel-Flarsheim Co. v. Medrad, Inc.*, 358 F.3d 898, 913 (Fed. Cir. 2004) ("[I]t is improper to read limitations from a preferred

NVIDIA Ex1041
NVIDIA Corp. v. Neural AI, LLC
IPR2025-00608
Page 13 of 37

embodiment described in the specification."). Second and in fact, the specification discloses that the patent's disclosed "architectur[al]" inventions such as the accelerator may be software implementations executing on backend processors. Ex. 2, 9:51–62 & FIG. 3. "SANNDRA (Synchronous Artificial Neuronal Network Distributed Runtime Algorithm; http://www.kinness.net/Docs/SANDRA/html) was developed to accelerate and optimize processing of numerical integration of large non-homogenous systems of differential equations. . . . [A] GPU based backend for SANNDRA-2.x.x can serve as an example [of] practical software implementation of the method and architecture described [in the specification] and pictorially represented in FIG. 3." *Id.* NVIDIA's construction of the accelerator as hardware reads out preferred embodiments in the specification. *See Oatey Co. v. IPS Corp.*, 514 F.3d 1271, 1276 (Fed. Cir. 2008) ("[I]t is incorrect to construe the claims to exclude [] embodiment[s]."); *see also CA, Inc. v. Netflix, Inc.*, No. 2023-1768, 2025 WL 303436, at *4 (Fed. Cir. Jan. 27, 2025) (explaining as "disfavor[ed]" constructions that "exclude embodiments expressly disclosed in a patent's specification"). Further, Claim 16 of the '867 Patent does not disclose any requirements for an accelerator, such as a "graphics processing unit," "accelerator memory," or any other term NVIDIA can plausibly claim is "hardware," instead teaching an "accelerator" in functional language, as the disclosed method of receiving, by an accelerator, a first input data from the central processing unit. Ex. 1, 16:23–24.

Extrinsic evidence reinforces that an "accelerator" is not limited to hardware. Even NVIDIA's selected definition contemplates a ***combination*** of hardware and software, ECF No. 71-1 at 7, and other contemporaneous definitions are functional as opposed to hardware-based, *see* Ex. 6 (Authoritative Dictionary of IEEE Standards Terms 2000), at 4 ("accelerator . . . (2) A circuit or device that accelerates some unit in a computer"). NVIDIA's improper limitation would

preclude implementations that would fall within the claim scope as understood by a POSITA. Ex. 5 ¶ 32 (explaining that NVIDIA's proposed construction "omits [an in-scope] class of computer architecture referred to as programmable firmware").

The Court should reject NVIDIA's myopic reading of the term "accelerator" and apply the term's plain and ordinary meaning.

B. "accelerator controller" ('867 patent, cls. 16, 17, 19) / "controller" ('438 patent, cls. 1, 7–10, 21, 40, 42, 44, 50–53)

Neural AI's Construction	NVIDIA's Construction
Plain and ordinary meaning	specialized hardware and/or software located on the accelerator

The terms "accelerator controller" or "controller" do not require any construction. NVIDIA again seeks to add unclaimed limitations, and its proposed construction should be rejected.

Tellingly, NVIDIA does not even cite the claims to support its arguments regarding the controller. *See* Op. Br. at 8–10. That is because nothing in the claims requires the controller to be "specialized" or to have a particular location, including on the accelerator. Nor does the specification contain any express intent to redefine the term "controller." *See Thorner*, 669 F.3d at 1365 (holding that, to re-define claim language, a patentee must do more than "simply disclose a single embodiment or use a word in the same manner in all embodiments, the patentee must clearly express an intent to redefine the term"). NVIDIA's attempts to pull limitations from (a) the patent's abstract and summary sections and (b) from a single embodiment both fail.

First, NVIDIA's cherry-picked citations to the abstract and summary sections, the *only* references to the controller as "specialized," are insufficient to read that limitation into the claim. No "specialization" requirement is disclosed in the specification or claims. In fact, elsewhere in the specification, the controller is described as "general purpose," Ex. 1, 13:53–54, as opposed to

"specialized." And within the claims themselves, the controller is not limited. *See* Ex. 1, cls. 16, 17, 19; Ex. 2, cls. 1, 7–10, 21, 40, 42, 44, 50–53.

Second, NVIDIA's location limitation fails. Hornbook law prevents NVIDIA from reading specific features of a preferred embodiment into the claims. NVIDIA acknowledges that the "Federal Circuit has warned against reading in limitations from embodiments into the claims," Op. Br. at 9, but ignores the Circuit's guidance and overstates *Regents of Univ. of Minnesota v. AGA Med. Corp*., 717 F.3d 929, 936 (Fed. Cir. 2013). *See id.* In *Regents*, the parties disputed a claim construction regarding a "conjoint disk," which had been construed by the trial court to cover only "two physically separate disks that are attached to one another." 717 F.3d at 935. Contrary to NVIDIA's brief, *Regents* did not agree the invention should be limited to "attaching two disks together" solely because "it is the *only* [embodiment] described." Op. Br. at 9 (quoting *Regents*). Instead, the "attaching two disks together" limitation was deemed appropriate because the plain claim language "fully supports a requirement of separateness," reciting (for example) "affixing the membranes of the first and second occluding disks to one another." *Regents*, 717 F.3d at 935 (quotations and alterations removed). Moreover, in order to overcome prior art in prosecution, the patentee had disclaimed a more expansive scope and limited the relevant invention to "attaching . . . two disks . . . to form a conjoint disk." *Id.* at 937. But here, unlike in *Regents*, the claim language is silent on the controller's location relative to the accelerator, and nothing in the prosecution history supports NVIDIA's unfounded limitation.

Moreover, NVIDIA is simply incorrect that "[t]here is no embodiment in which the controller is *not* located on the accelerator." Op. Br. at 9. On the contrary, the specification discloses Figure 3, which "is a block/flow diagram," Ex. 1, 3:28, illustrating a "method that is used to control the computation," *id.* at 7:24–30, and which can be executed entirely via a "practical

software implementation," *id.* at 9:25–35, that is agnostic as to the physical location of any components. In any event, it is well established that it is "not enough that the only embodiments, or all of the embodiments, contain a particular limitation." *Thorner*, 669 F.3d at 1366 (internal marks omitted).

The extrinsic evidence further indicates that NVIDIA's construction is inappropriate. To start, dictionary definitions do not require that the "controller" has any "specialized" hardware and instead indicate that the controller derives its meaning from the role it plays in a system, i.e., providing access or coordinating operations, rather than its makeup as hardware, software, or some combination. *See* Ex. 4, at 128 ("controller *n.* A device that other devices rely on for access to a computer subsystem."); Ex. 7 (IBM Dictionary of Computing 1994), at 145 ("controller A device that coordinates and controls the operation of one or more input/output devices, such as workstations, and synchronizes the operation of such devices with the operation of the system as a whole.").[3] Neural AI's expert Dr. Bajaj corroborates that a POSITA would understand the controller could be a component located "in any software or hardware and software combination," as well as firmware. *See* Ex. 5 ¶¶ 39, 41.

Neither the intrinsic nor extrinsic evidence dictates limitations to the term "controller," and the Court should reject NVIDIA's construction. Indeed, Courts routinely construe the term "controller" to have its plain and ordinary meaning. *See, e.g.*, *Orenshteyn v. Citrix Sys., Inc.*, 341 F. App'x 621, 624–25 (Fed. Cir. 2009) (construing "controller" to have no limitations and

[3] NVIDIA protests the IBM definition with circular logic. NVIDIA claims this definition is inapplicable because the definition does not place the accelerator controller on the accelerator itself. Op. Br. at 10. This argument illogically assumes that NVIDIA's construction is correct to then explain away contrary evidence. Instead, the IBM definition illustrates that a POSITA at the time of the effective date of the patent would understand that a "controller" is *not* limited to a location on the accelerator.

overruling district court's construction that a controller cannot include a general purpose CPU); *Signify N. Am. Corp. v. Lepro Innovation Inc.*, No. 222CV02095JADDJA, 2023 WL 8435567, at *7–8 (D. Nev. Dec. 4, 2023) (construing "controller" to have its plan and ordinary meaning where, as here, the specification did not otherwise define the term); *3M Innovative Props. Co. v. EnvisionWare, Inc.*, 2010 WL 5067449, at *2 (D. Minn. Dec. 6, 2010) (construing "controller" to have its plain and ordinary meaning not requiring express construction because "[j]urors will readily understand the meaning of the word 'controller' as it is used in the patent").

C. "partition" ('867 patent, cls. 16, 18; '438 patent, cls. 26, 27, 40, 43, 49; '461 patent, cls. 21, 30)

Neural AI's Construction	NVIDIA's Construction
Plain and ordinary meaning	A separated subdivision of memory designed to hold specified data

"Partition" is an easy word to understand, having the same meaning in the art as to ordinary laypeople. NVIDIA nonetheless urges a results-oriented construction that limits this term to a hardware implementation (a "separated subdivision of memory") for data that must be "specified." This is unsupported by the intrinsic or extrinsic record. Moreover, NVIDIA's construction invites juror confusion, begging the question of what qualifies as "separated," what it means to be "separated" memory yet also a "subdivision" of memory, and what qualifies as "specified data." For a simple term like "partition," inserting unnecessary words is not "helpful or warranted"; such changes "tend to confuse rather than clarify." *Nat'l Cheng Kung Univ. v. Samsung Elecs. Co*., 2014 WL 2885380, at *6 (E.D. Tex. June 25, 2014).

First, NVIDIA argues that where the claims recite a "first partition" to which "first input data" is "transferr[ed]" and a "second partition" in which "first output data" is "stor[ed]," Ex. 1, cl. 16, the "claim language explicitly states the partitions are subdivisions of memory designed to hold specific data, Op. Br. at 10. But the claim language says nothing of the sort. It is silent on

whether the partitions are "separated subdivisions," or whether the data in "first input data" and "first output data" is "specified." To the extent "hold[ing]" data (NVIDIA's construction) is the same as "storing" data (the claim language), only the "first output data" is "stor[ed]"'; "the first input data" is "transferr[ed]" according to the plain language. As support for its "specialized" limitation, NVIDIA notes that *some* claims "further recite the existence of third and fourth partitions for storing 'internal variables' and 'data used as input at a particular computation cycle of the numerical simulation." *See* Op. Br. at 11 (citing '867 Patent, Claims 5, 13, and 18). But the existence of more specific claims with additional limitations actually *undermines* NVIDIA's proposed construction. *See Voice Tech Corp. v. Unified Pats., LLC*, 110 F.4th 1331, 1342–43 (Fed. Cir. 2024) ("Such a construction would improperly impute other limitations [stated in other claims] into the claim term, rendering those other limitations superfluous."); *Amgen Inc. v. Hoechst Marion Roussel, Inc.*, 314 F.3d 1313, 1326 (Fed. Cir. 2003) ("When a patent claim does not contain a certain limitation and another claim does, that limitation cannot be read into the former claim." (cleaned up)). When the patentee wanted additional limitations on the type of data, those limitations were claimed explicitly.

Second, the specification does not support NVIDIA. Instead, the specification states the inventions' "partitioning scheme" should be implemented flexibly and "is also altered based on new designs or needs of the algorithms being employed." Ex. 1, 5:49–50. This can explicitly be logical, as opposed to physical separation. For example, the specification describes "a single, logically partitioned memory component," *id.* at 5:16, thereby teaching that memory can be one single component that is ***logically partitioned*** as opposed to physically separated in a "separated subdivision." *See* Ex. 5 ¶ 44 ("[T]he first partition could be all in the first memory module, chip, or memory bank, and the second partition could also be collocated or could be stored elsewhere,

and NVIDIA's construction improperly reduces this flexibility."). NVIDIA tries to limit the invention to Figure 2, Op. Br. at 12, but an embodiment cannot limit the claim language, especially not when doing so would exclude other preferred embodiments that contain a "single, logically partitioned memory component." *See Apple Inc. v. Wi-LAN Inc.*, 25 F.4th 960, 967 (Fed. Cir. 2022) (explaining embodiments do not limit a claim); *Oatey*, 514 F.3d at 1276–77 (explaining claims should not be interpreted in a way that excludes embodiments).

Third, the extrinsic evidence, including testimony of Neural AI's expert Dr. Bajaj and multiple technical dictionaries, confirm that the plain and ordinary meaning of "partition" does not include the arbitrary limitations NVIDIA proposes. *See* Ex. 5 ¶ 42, 45; *see also* Ex. 4, at 392 ("partition *n*. 1. ***A logically distinct portion of memory*** or a storage device that functions as though it were a physically separate unit." (emphasis added)); Ex. 6, at 797 ("partition … (B) A portion of a storage medium that is set aside for some special purpose . . . (C) ***A portion of a storage medium that is treated as if it were an individual medium***, as in a partition of hard disk." (emphasis added)).

With no intrinsic or extrinsic support, NVIDIA relies on a mischaracterization of Neural AI's *preliminary infringement contentions* to support its improper claim construction. Op. Br. at 10, 13. Even if NVIDIA were accurately characterizing the contentions, infringement contentions are not proper claim construction evidence and cannot justify a particular construction. *See Iris Connex, LLC v. Acer Am. Corp.*, 2016 WL 4596043, at *9–11, 13 (E.D. Tex. Sept. 2, 2016) (rejecting the argument "that the Court may consider the accused products in order to understand the claim construction disputes"); *Phillips*, 415 F.3d at 1317 (only authorizing claim construction evidence of "the patent and prosecution history," as well as "expert and inventor testimony, dictionaries, and learned treatises.")). NVIDIA fails to explain why a description of accused

products in the infringement contentions—mere embodiments—should limit the scope of the claim language. NVIDIA's invocation of *O2 Micro* is inapposite. *See* Op. Br. at 13 (quoting *O2 Micro Int'l Ltd. v. Beyond Innovation Tech. Co*., 521 F.3d 1351, 1361 (Fed. Cir. 2008)). *O2 Micro* explained that "'plain and ordinary meaning' may be inadequate when . . . [the] term's 'ordinary' meaning does not resolve the parties' dispute," meaning "a fundamental dispute regarding the scope of a *claim term*." *id.* at 1361–62 (emphasis added). By using the infringement contentions, NVIDIA raises a dispute about *infringement*, not a dispute about the claim scope.

Even if it were proper to rely on infringement contentions to inform the scope of a claim term, the portion of the preliminary infringement contentions that NVIDIA cites does not state that all data copied to any "destination memory address" is stored within a partition, as NVIDIA suggests. *See, e.g.*, ECF No. 71-6 at 46. Instead, it alleges that the particular CUDA platform embodiment has various "memory management functions" and parameters like "Destination memory address" known as "dst" that enable the Accused Products to store data in a partition within the device. *Id.* at 46, 48.

The Court should reject NVIDIA's proposed construction for "partition" and instead adopt the term's plain and ordinary meaning.

D. "bank" ('438 patent, cls. 6, 20)

Neural AI's Construction	NVIDIA's Construction
Plain and ordinary meaning	Partition

As the Asserted Patents show, when the patentee intended to use the term "partition," it did so. Nothing in the intrinsic record suggests the patentee redefined the word "bank" to mean "partition." There is no basis to deviate from the plain meaning. Even if there were, the intrinsic record certainly does not support replacing the term with an ***entirely different claim term***, as NVIDIA requests. Indeed, NVIDIA's proposed construction rewrites the claim twice by (1)

changing the plain claim language "bank" to a different word, "partition," (2) then defining "partition" with results-oriented limitations, as described above. *See supra* pp. 13–16. The Court should reject NVIDIA's effort to construe "bank" as "partition."

It is a fundamental and longstanding canon of claim construction that different claim terms are presumed to have different meanings. *See Tandon Corp. v. U.S. Int'l Trade Comm'n*, 831 F.2d 1017, 1023 (Fed. Cir. 1987) ("There is presumed to be a difference in meaning and scope when different words or phrases are used in separate claims."); *Apple Inc. v. Omni MedSci, Inc.*, No. 2023-1034, 2024 WL 3084509, at *3 (Fed. Cir. Jun. 21, 2024) (noting it is "[w]ell-settled" that "different claim terms are presumed to have different meanings" (citation omitted)). Here, certain Asserted Patent claims use the term "bank," while other claims disparately recite "partition." *Compare* Ex. 2, cl. 6 (disclosing, in that invention, that "accelerator memory comprises . . . a first memory ***bank*** to store parameters"), *with id.* at cl. 26 ("storing, in a first memory ***partition*** of the memory, parameters"). Equating the terms would fundamentally change the meaning of the asserted claims and violate well-established law.

The intrinsic record confirms that "bank" is used in the claims in accordance with its plain meaning and should not be construed as "partition." On the contrary, the intrinsic evidence repeatedly contrasts "bank" with "partition." For example, the specification describes: "In some example[s], these memory banks separated in the hardware, as shown, or alternatively implemented as a single, logically partitioned memory component." Ex. 2, 5:30–34. This passage reflects that multiple memory banks can be either contiguous or noncontiguous, as well as physically separated or "a single, logically partitioned memory component." *Id.* In another example embodiment, the specification describes that a bank can include multiple partitions: "texture memory bank 250 is preferably further partitioned into four sections." *Id.* at 5:58–59; *see*

also id. at 2:29–31 (describing "two or more associated memory banks that are logically or physically partitioned").

Neural AI's expert Dr. Bajaj explained in detail why a POSITA would have understood "bank" differently than "partition":

> [T]he plain and ordinary meaning of "bank" can include computer devices grouped together for use or inclusion within a single device. The plain meaning of bank in the computer engineering field of the Asserted Patents is also well known to refer to, for example, a memory bank that has virtual or logical separation, such as noncontiguous memory cells spread over a memory chip or multiple computer memories; there are varieties of implementation options for a memory bank that would also have been well known to a POSITA.

Ex. 5 ¶ 48; *see also id.* at ¶ 42 (describing plain and ordinary meaning of "partition"). As Dr. Bajaj further opined, the terms partition and bank *may* refer to similar things, but not always, because they are not synonymous and the correct usage depends on the context:

> [T]hese terms are used for conveniently referring to hierarchies of memory, where a memory card or chip may be described as having one, two, three, or more memory "banks" as a relative reference for the size of the memory chip, such as in comparison to other memories on the same chip or different memory chips. And any such memory bank could be described has having one, multiple, or a part of a memory partition, which, as described above, is ordinarily referred to in reference to a logical division of data.

Id. at ¶ 49.

Technical dictionaries not only define the terms "bank" and "partition" independently, but definitions for "bank" also omit the other limitations as to a "separate subdivision" and "specified data" that NVIDIA proposes should apply for the term "partition," and therefore also to "bank." *See, e.g.*, Ex. 4, at 50–51 ("bank *n.* 1. Any group of similar electrical devices connected together for use as a single device. . . . 2. A section of memory, usually of a size convenient for a CPU to address."); Ex. 6, at 85 ("bank … (2) … (B) Any group of similar devices that are connected together for use as a single device."). These technical dictionaries are consistent with Dr. Bajaj's explanation as to the plain and ordinary meaning of "bank." *See* Ex. 5 ¶ 48.

The Court should reject NVIDIA's proposal asking the jury to defy its own common sense to conclude that different words are in fact the same.

E. "swapping the first pointer with the second pointer" ('867 patent, cl. 16)

Neural AI's Construction	NVIDIA's Construction
Plain and ordinary meaning	swapping, by the accelerator controller, the first pointer with the second pointer

The term "swapping the first pointer with the second pointer" does not require any construction by the Court, because the term has a plain and ordinary meaning. Since the patentee neither created its own definition of "swapping the first pointer with the second pointer" that deviates from the plain and ordinary meaning nor disavowed the full scope of the term, no construction is necessary.

NVIDIA attempts to improperly limit the scope of the claim by asking this Court to insert "by the accelerator controller" into the claim language. NVIDIA argues that the specification and the file history limited the invention's "swapping the first pointer with the second pointer" to instances in which the accelerator controller performs the swap. But "the standard for disavowal of claim scope is . . . exacting." *Thorner*, 669 F.3d at 1366. "To disavow claim scope, the specification must contain expressions of manifest exclusion or restriction, representing a clear disavowal of claim scope." *Retractable*, 653 F.3d at 1306 (internal marks omitted). And any disavowal of claim scope made during patent prosecution "must be both clear and unmistakable." *3M Innovative Props.*, 725 F.3d at 1325. NVIDIA's cited examples in the specification and file history do not meet that standard.

For the specification, NVIDIA broadly asserts the specification contains several embodiments requiring the accelerator controller to perform the pointer swap. Op. Br. at 16. But NVIDIA only specifically cites to Figure 4, which NVIDIA describes as depicting the swapping "as part of the Computational Substream 403, which is executed by the accelerator controller." *Id.*

NVIDIA's embodiment argument fails as a matter of law. For a disavowal of claim scope, it is "not enough that the only embodiments, or all of the embodiments, contain a particular limitation." *Thorner*, 669 F.3d at 1366. Indeed, the Federal Circuit has repeatedly warned against construction that limits a term to just an embodiment in the specification. *Sisvel*, 81 F.4th at 1236 ("[A]lthough the specification often describes very specific embodiments of the invention, we have repeatedly warned against confining the claims to those embodiments." (citation omitted)). Figure 4 cannot amount to a clear disavowal of claim scope because an embodiment is not an "expression[] of manifest exclusion or restriction, representing a clear disavowal of claim scope." *Retractable*, 653 F.3d at 1306 (citation omitted). In fact, the specification expressly states that Figure 4 is *not* intended to limit the scope of the claimed invention: "FIG. 4 is a representation of ***one of several ways*** in which a system and method for processing numerical techniques can be implemented." Ex. 1, 11:27–29 (emphasis added).

NVIDIA's file history argument is complete misdirection. NVIDIA asks the Court to construe "swapping the first pointer with the second pointer" in ***Claim 16***. But NVIDIA points to two statements from the file history about ***Claim 1***:

- "[Prior art] did not 'disclose or suggest an accelerator controller . . . to swap the first pointer and the second pointer at the conclusion of the second computational cycle such that the second output data becomes an input for a third computational cycle of the plurality of computational cycles.'" Op. Br. at 18.
- "According to the applicant, 'at best the disclosure of [the applied prior art reference] suggests an element that simply moves data into and out of a video card,' but '[t]here is no disclosure whatsoever in [the reference] of an accelerator controller that swaps respective pointers referencing a portion of the input data and the output data . . . as recited in claim 1.'" *Id.*

NVIDIA again fails to alert the Court to the fact that these excerpts come from the patentee's explanation for why ***Claim 1*** overcame the cited references. Unlike method Claim 16—which contains the phrase "swapping the first pointer with the second pointer" at issue—Claim 1

expressly requires the "accelerator controller . . . to swap the first pointer and the second pointer." *Compare* Ex. Ex. 1, 16:36–39 (Claim 16) *with id.* at 14:40–52 (Claim 1). ***In fact, the language NVIDIA quotes from the file history is actually just an excerpt of Claim 1 itself***. The patentee never even used the fact that the accelerator is performing the pointer swapping as the basis for distinguishing the prior art. The patentee distinguished the prior art based on the accelerator controller itself and then simply quoted Claim 1, which already linked the accelerator controller to the swapping.

In an entirely different paragraph, the patentee distinguished method Claim 16—then Claim 17—from the prior art, and did so without in any way linking the accelerator controller to the pointer swapping functionality. ECF No. 71-11 at 13. Given the claim language, it makes sense that the patentee did not require the pointer swapping to be performed by the accelerator controller. Claim 16 recites a method for performing computations in a computer system. While other limitations are limited to specific hardware, "swapping the first pointer with the second pointer" is not. As Neural AI's expert Dr. Bajaj explains, there is no technical reason why a POSITA would understand that "swapping the first pointer with the second pointer" in Claim 16 must be performed by the accelerator controller. Ex. 5 ¶ 58. The swapping need not even be limited to hardware and could instead be performed by software. *Id.*

NVIDIA also points to the Brief Summary of the Claimed Subject Matter in the patentees' remarks, Op. Br. at 18, but this too fails to constitute a clear and unmistakable disavowal of claim scope. NVIDIA is correct that the patentee did provide a summary exemplar of the invention that stated the "accelerator controller swaps the first pointer and the second pointer." ECF No. 71-11 at 10. But the patentee did not narrow the claims to swaps performed by the accelerator controller to overcome the prior art. Rather, the patentee stated the examiner should *not* rely on the exemplar

summary and should instead "rely upon the language specifically used in the claims to determine whether each of the claims distinguishes over the prior art of record." *Id.* An embodiment that the patentee expressly indicates should not limit claim scope cannot possibly be a clear and unmistakable disavowal of claim scope.

Ultimately, it is completely inaccurate for NVIDIA to suggest to the Court that the patentee overcame prior art by narrowing Claim 16 to require pointer swaps to be performed by the accelerator controller. That never happened. The Court should reject NVIDIA's construction based on its re-written version of the file history.

F. "input data at first rate; and . . . sequence of computations at a second rate" ('438 patent, cls. 3, 14, 46)

Neural AI's Construction	NVIDIA's Construction
Plain and ordinary meaning	input data at an amount of data received per unit time; and . . . sequence of computations at a number of computations per unit time

Asserted Claims 3, 14, and 46 of the '438 Patent each recite input data "at a first rate" as well as a "sequence of computations at a second rate different than the first rate." *See, e.g.*, Ex. 2, 15:24–28.

"Rate" is a common word that is easily understood. Yet NVIDIA tries to strip the simplicity and clarity from the term by offering a nonsensical construction that defines "rate" in *two* different ways. NVIDIA says "rate" means an "amount of data," but it also means a "number of computations." According to NVIDIA, "rate" cannot apply to anything else besides an amount of data or number of computations. The patentee neither defined "rate" in a way that deviated from the plain meaning nor disavowed the full scope of the term. Nothing in the claim language, specification, or any extrinsic evidence cited by NVIDIA supports these arbitrary limitations. There is no basis for NVIDIA's tortured construction.

1. The First Rate Need Not Be "An Amount of Data Received Per Unit Time"

NVIDIA attempts to support its "first rate" limitation by pointing to certain "specification[] examples" where "first rate" is purportedly used "to describe how much data is transferred to the CPU or GPU over a specific amount of time." Op. Br. at 19. But it is Hornbook law that the presence of preferred embodiments in the specification is not limiting. *See e.g.*, *Sisvel*, 81 F.4th at 1236 (highlighting that the Federal Circuit has "repeatedly warned against confining the claims to th[e] embodiments" (citation omitted)). The embodiment cannot limit "first rate" to an amount of data.

NVIDIA also asserts that "in the art of computer architecture," the word "rate" "capture[s] the clear notion of data transfer throughput." Op. Br. at 20. NVIDIA does not cite any authority for this proposition. NVIDIA's unsupported assertion is contradicted by its defining "second rate" as a "number of computations" and further contradicted by the extrinsic evidence. *See* Ex. 6, at 920 ("rate The change in a value over a specified period of time."). In his declaration, Neural AI's expert states the obvious: "rate" has no specialized meaning "in the context of computer technology." Ex. 5 ¶ 55. A POSITA would not understand the term "rate" to be limited to the particular units NVIDIA proposes.

NVIDIA's similarly unsupported argument about the purported term of art "input rate" is sleight of hand. Op. Br. at 20. The term "input rate" does not appear in the claims. Rather, the claims refer to "input data," and separately, a "first rate" and "second rate." *See* Ex. 7, at 341 ("input data (1) Data that are entered into a data processing system or any of its parts for storage or processing. (T) (2) Data received or to be received by a functional unity or by any part of a functional unit. (3) Data to be processed."). NVIDIA's mishmash of claim language does not support its construction.

The plain and ordinary meaning of "rate"—to both laypeople and POSITAs—is simply the change in a value over a period of time. Nothing in the claim language requires the "rate" of "first rate" to be any particular value or type of unit—much less than "an amount of data" (as NVIDIA contends). *See also* Ex. 5 ¶ 53 (a POSITA would not understand the general term "rate" to refer to "*particular* units of measure per unit time" in the '438 Patent claims (emphasis added)).

2. The Second Rate Need Not Be "Number of Computations Per Unit Time"

Just as "rate" does not mean "an amount of data over time," the same word does not mean "a number of computations per time" later in the same claim. With respect to the second "rate," NVIDIA shows its hand by ***citing to no evidence whatsoever—intrinsic or extrinsic—in support.*** Instead, NVIDIA's brief asserts as a matter of its own say-so that "the ordinary meaning of this phrase" in "technical and scientific fields" is that "a number of computations are performed in every unit of time." Op. Br. at 20. Other than NVIDIA's belief that, for the value over time in "second rate," "the quantity here [must] be[] the number of computations," *id.*, it offers no support.

Indeed, NVIDIA does not even attempt to address the glaring contradiction in its argument, where NVIDIA asserts the word "rate" has a specialized meaning in the art when used in "first rate" (i.e., a value that must be an amount of data), but somehow a completely different specialized meaning in the art when used in "second rate" (i.e., a value consisting, instead, of a number of computations). Nothing in the claim language requires the "first" and "second" rates to measure different units of value over time, as long as the values themselves (i.e., the "rate[s]") are "different." Ex. 2, cls. 3, 14, 46; *see* Ex. 5 ¶ 56 (explaining it was well known in the art for first and second rates to measure the same or different types of units). The Court should reject this inconsistent construction in favor of the term's well-understood plain and ordinary meaning.

G. Order of Steps (’867 patent, cl. 16; ’438 patent, cls. 12, 21)

Neural AI’s Construction	NVIDIA’s Construction
plain and ordinary meaning: no order required except “swapping” (’867, cl 16) performed last	steps must be performed in the order listed

“As a general rule, ‘unless the steps of a method claim actually recite an order, the steps are not ordinarily construed to require one.’” *Mformation Techs., Inc. v. Research in Motion Ltd.*, 764 F.3d 1392, 1398 (Fed. Cir. 2014) (quoting *Interactive Gift Express, Inc. v. Compuserve Inc.*, 256 F.3d 1323, 1342 (Fed. Cir. 2001)). The Court should deviate from this rule *only* if (1) logic or grammar require a certain order, or (2) if the specification requires such a narrow construction. *See Altiris, Inc. v. Symantec Corp*., 318 F.3d 1363, 1369–70 (Fed. Cir. 2003). Otherwise, “the sequence in which such steps are written is not a requirement.” *Id.* at 1370. The method steps of Claim 16 of the ’867 patent and Claims 12 and 21 of the ’438 patent do not recite the steps in a required order. Defendant asks the Court to impose one based on the claim language, Op. Br. at 20–24, but the claim language does not compel NVIDIA’s construction.

Claim 16 of the ’867 Patent consists of the following steps:

- receiving, by an accelerator, first input data from the central processing unit;
- transferring, by an accelerator controller, the first input data into a first partition, referenced by first pointer, of an accelerator memory before a first computational cycle of the numerical simulation;
- performing, by at least one graphics processing unit during the first computational cycle, at least one calculation on the first portion of the input data as to generate first output data;
- storing, by the accelerator controller, the first output data into a second partition, referenced by a second pointer, of the accelerator memory; and
- swapping the first pointer with the second pointer at the end of the first computational cycle, such that the first output data becomes an input for a second computational cycle of the numerical simulation.

Ex. 1, 16:20–40. Neural AI acknowledges that the swapping step of this claim involves outcomes from the previous steps and must be performed last. The claim explicitly states that this swapping step comes "at the end of the first computational cycle," *id.* at 16:37–38, but provides no such explicit limitation on the order of steps for the first three. When the patentee wanted to provide a required order, the claims say so.

However, NVIDIA over-extends the logical order of the swapping step to the remaining steps and argues that *all* steps must be performed in the recited order. Op. Br. at 21–22. Not so. The intermediary steps do not require an order and, according to the plain language, can be performed in parallel. NVIDIA's construction would exclude the simultaneous nature of the claims. Parallel execution is important here—as the specification repeatedly emphasizes—where the method described in the claim is performed in computational cycles executed in parallel loops, such that some steps of the first computational cycle are in parallel to the calculations performed in the second computational cycle. *See, e.g.*, Ex. 1, 13:44–46 ("[W]hatever data transfer remains necessary will take place in parallel with the computation, thus reducing the impact of this transfer on the performance"); *id.* at 13:55–59 ("The GPU 240 is inherently parallel and is well suited to perform parallel computations. In parallel with the GPU 240 performing the next calculation, the controller 220 is uploading the data from the previous calculation into main memory.").

Nor does the claim specify when the steps *begin*. In *Ancora Technologies, Inc. v. LG Electronics Inc.*, No. 1-20-CV-00034-ADA, 2020 WL 4825716 (W.D. Tex. Aug. 19, 2020), this Court recognized that there is a distinction between an order of when steps are completed versus when they begin. 2020 WL 4825716, at *5 ("While the language of the claim requires an order to the completion of certain limitations in the '941 Patent, neither the claim language nor the specification requires an order to when the steps begin."); *see also Kaneka Corp. v. Xiamen*

Kingdomway Grp. Co., 790 F.3d 1298, 1306–07 (Fed. Cir. 2015) ("We also disagree with the district court's conclusion and Defendants' arguments on appeal suggesting that the claimed order requires that each step occur independently or separately."). There, the Court found "no reason to impose all of the limitations on the claim steps that Defendants urge" because they "provided no evidence to show why [certain steps] could not be performed concurrently." 2020 WL 4825716, at *5. Like in *Ancora*, there is no reason for the Court to impose the limitations NVIDIA urges here, especially where three out of the four steps are silent on the necessary order.

The asserted claims in the '438 patent are even more explicit in describing that they are free from a specific order. The clear language of Claim 12 of the '438 states that the steps can be performed "in parallel;" the claim recites the following:

- A method of performing a sequence of computations representing an artificial neural network on a computer system comprising a central processing unit (CPU), a main memory operably coupled to the central processing unit via a bus, an accelerator operably coupled to the CPU and the main memory via the bus, the accelerator comprising a graphics processing unit (GPU) and an accelerator memory, the method comprising:

- (A) performing, by the GPU, the sequence of computations on a first portion of the input data so as to generate a first portion of the output data, the first portion of the output data representing an output of a neuron in a first layer of the artificial neural network, intermediate computations in the sequence of computations yielding intermediate results, wherein performing the sequence of computations on the first portion of the input data comprises (i) assigning an output variable to a first texture and a second texture, the output variable being included in a first computational element of a plurality of computational elements, the plurality of computational elements representing the sequence of computations and (ii) accumulating a first value for the output variable in the first texture during a first time step;

- (B) ***in parallel*** with performing the sequence of computations by the GPU in (A), transferring a second portion of the input data from the main memory to the accelerator via the bus; and

- (C) ***in parallel*** with performing the sequence of computations by the GPU in (A), transferring a second portion of the output data from the accelerator memory to the main memory via the bus, the second portion of the output

> data representing an output of a neuron in a second layer in the artificial neural network; and
>
> - (D) performing, by the GPU, the sequence of computations on the second portion of the input data, wherein performing the sequence of computations on the second portion of the input data comprises (i) accumulating a second value for the output variable in the second texture during a second time step and (ii) making the first value of the output variable in the first texture accessible to other computational elements in the plurality of computational elements during the second time step.

Ex. 2, 15:66–16:43 (emphasis added). Nothing in this claim requires that the steps be performed in the order listed, and to the contrary, multiple steps can and should be performed "in parallel." NVIDIA acknowledges, as it must based on the clear claim language, that part of steps B and C can occur in parallel. *See* Op. Br. at 22–23. This unequivocal claim language controls. *See Renishaw*, 158 F.3d at 1248 (claim construction "begins and ends" with the claim language).

Further, Claim 21 of the '438 patent recites its steps without dictating a particular order:

> - A method of performing a sequence of computations representing an artificial neural network, the method comprising:
>
> - receiving, at a central processing unit (CPU), first input data acquired from an external system in real time;
>
> - initializing, by a controller operably coupled to a graphics processing unit (GPU), textures and shaders in a memory operably coupled to the GPU;
>
> - transferring the first input data received by the CPU to the memory operably coupled to the GPU;
>
> - performing, by the graphics processing unit (GPU), a first computation in the sequence of computations on the first input data based on the textures and shaders to generate first output data, computations in the sequence of computations representing respective layers of neurons in the artificial neural network, an output of the first computation in the sequence of computations representing an output of a first neuron in a first layer in the artificial neural network;
>
> - storing, in the memory operably coupled to the GPU, the first input data and the first output data; and
>
> - transferring second input data acquired from the external system in real time into the memory operably coupled to the GPU after the GPU starts the first computation

> and before the GPU starts a second computation of the sequence of computations, an output of the second computation in the sequence of computations representing an output of a second neuron in a second layer in the artificial neural network.

Ex. 2, 17:16–45. The claim language does not require each step to be performed in the order listed, and even NVIDIA acknowledges that the second step "does not necessarily have to occur after the first input data is received in the first step." Op. Br. at 23.

The text, grammar, and logic of the claim language make clear that parallel execution is a core feature of the methods described in the above claims, such that the steps need not be performed in the order listed. The specification confirms this. Nonetheless, and even in the face of NVIDIA's own admissions that certain steps may be taken in parallel or out of order, NVIDIA still urges its proposed construction that *all* steps must occur in the order listed in the claims. *Id.* at 24. NVIDIA's position is contradictory and illogical, refusing to engage with the nuance of the claimed methods. NVIDIA's proposed construction precludes, without justification, both (a) fragmented mathematical computations where the sub-steps run out of sequence and (b) parallel execution of portions of the mathematical computations. Neural AI's expert Dr. Bajaj validates that a POSITA would not understand the claims as precluding such out-of-sequence or parallel execution, but rather would embrace them. Ex. 5 ¶¶ 61–62.

NVIDIA gives the Court no reason to deviate from the general rule that method steps do not recite a particular order, and the claimed steps should not be construed to require one. The Court should adopt the plain and ordinary meaning of the claims, which acknowledge parallel and out-of-sequence execution, except that the swapping step of patent '867, Claim 16 occurs after the other steps of that claim.

VI. CONCLUSION

The Court should adopt Neural AI's proposed constructions of the disputed terms.

Dated: June 10, 2025

Respectfully submitted,

/s/ *Max Tribble*
Max L. Tribble
Texas State Bar 20213950
Brian D. Melton
Texas State Bar 24010620
Rocco Magni
Texas State Bar 24092745
Samuel Drezdzon
Texas State Bar 24117374
SUSMAN GODFREY L.L.P.
1000 Louisiana
Suite 5100
Houston, TX 77002
Telephone: (713) 651-9366
Facsimile: (713) 654-6666
mtribble@susmangodfrey.com
bmelton@susmangodfrey.com
rmagni@susmangodfrey.com
sdrezdzon@susmangodfrey.com

Tamar Lusztig
NY State Bar 5125174
Emily Portuguese
NY State Bar 5920327
One Manhattan West, 50th Floor
New York, NY 10001
tlusztig@susmangodfrey.com
eportuguese@susmangodfrey.com

Mark D. Siegmund
Texas State Bar No. 24117055
CHERRY JOHSON SIEGMUND JAMES PC
Bridgeview Center
7901 Fish Pond Road, 2nd Floor
Waco, Texas 76710

Max Ciccarelli
Texas State Bar No. 00787242
CICCARELLI LAW FIRM LLC
100 N. 6th Street, Suite 502
Waco, Texas 76701

Max@CiccarelliLawFirm.com

Attorneys for Plaintiff Neural AI, LLC

NVIDIA Ex1041
NVIDIA Corp. v. Neural AI, LLC
IPR2025-00608
Page 36 of 37

CERTIFICATE OF SERVICE

I hereby certify that on June 10, 2025, I caused the foregoing to be served via ECF on all counsel of record.

/s/ Emily Portuguese
Emily Portuguese

NVIDIA Ex1041
NVIDIA Corp. v. Neural AI, LLC
IPR2025-00608
Page 37 of 37