

**IN THE UNITED STATES DISTRICT COURT
FOR THE WESTERN DISTRICT OF TEXAS
MIDLAND/ODESSA DIVISION**

ADVANCED CLUSTER SYSTEMS, INC.,

Plaintiff,

v.

INTEL CORPORATION,

Defendant.

Civil Action No. 7:24-cv-00245-ADA

**PLAINTIFF ADVANCED CLUSTER SYSTEMS, INC.’S
PRELIMINARY INFRINGEMENT CONTENTIONS**

Pursuant to the Court’s Standing Order Governing Proceedings 4.4 – Patent Cases, dated January 23, 2024 (“OGP”), at p. 2 (Section II.2), Plaintiff Advanced Cluster Systems, Inc. (“ACS”) hereby serves on Defendant Intel Corporation (“Intel”) the following Preliminary Infringement Contentions.¹

The following disclosures are based on information currently available to ACS. Fact and expert discovery in this action has not begun, and ACS reserves the right to modify, amend, or otherwise supplement these disclosures as discovery reveals new documents and information. ACS further reserves the right to modify, amend, or otherwise supplement these disclosures should Defendant alter, supplement, or otherwise clarify documents and information currently available

¹ ACS anticipates that its contentions will require supplementation because Intel: (1) has not produced any versions, subroutines or APIs for the source code for the Accused Products; (2) has not provided any relevant and responsive technical documents for the Accused Products; and (3) has not produced documents or otherwise provided information or evidence responsive to ACS’s forthcoming Requests for Production or other discovery requests. *See* OGP at p. 11, n. 8.

to ACS. ACS notes that Defendant has not yet provided any versions, subroutines, or APIs for the source code for the Accused Products. Defendant has not yet produced any relevant and responsive technical documents that show the operation of the Accused Products and have not produced any documents responsive to ACS's forthcoming Requests for Production. Such responsive documents are relevant to several of the required disclosures below.

I. DISCLOSURES

A. ASSERTED CLAIMS

Pursuant to the Court's OGP, at p. 2 (Section II.2), ACS identifies the following asserted claims (the "Asserted Claims"):

U.S. Patent No.	Asserted Claims
10,333,768	1, 4, 20, 21, 26, 27, 29, 30, 31, 33, 34, 35, 36, 37, 39
11,563,621	1, 3, 4, 6, 7, 8, 9, 10, 11, 13, 15, 16, 17, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30
11,570,034	1, 2, 3, 8, 10, 24, 25, 27, 28, 30
11,811,582	1, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20
12,021,679	1, 2, 3, 8, 10, 17, 18

B. ACCUSED PRODUCTS AND METHODS

Based on presently available information, ACS contends that Intel infringes the Asserted Claims based on the following instrumentalities (the "Accused Products"). Specifically, ACS asserts that the Accused Products meet the limitations of the asserted claims identified below.

Asserted Patent	Asserted Claims	Accused Product(s)
10,333,768	1, 4, 20, 21, 26, 27, 29, 30, 31, 33, 35, 37, 39	All server and workstation products that utilize Intel's Habana AI accelerator architectures, including at least the Gaudi HLS-1 AI Training System, HLS-Gaudi2 server, and Gaudi 3 AI Accelerator HLB-325 Baseboard products, all variants thereof, and all products including to related to those products, as well as any products incorporating those items and any products that are substantially similar to those products (the "Accused Habana Server Products"). All AI accelerator products that utilize Intel's Habana AI accelerator architectures and are especially adapted for use in server and workstation products, including at least the Gaudi, Gaudi 2, Gaudi 3, and Goya AI accelerator products, all variants thereof, and any products that are substantially similar to those products (the "Accused Habana AI Accelerator Products"). <i>See Exhibit A1 hereto. Each of the Accused Habana AI Accelerator Products and each of the Accused Habana Server Products is configured and operates in a substantially similar manner in relevant part.</i>
10,333,768	1, 4, 20, 21, 26, 27, 29, 30, 31, 33, 34, 35, 36, 37, 39	All Xeon Scalable Processors (Skylake-SP architecture), 2nd Gen Intel Xeon Scalable Processors, 3rd Gen Intel Xeon Scalable Processors, 4th Gen Intel Xeon Scalable Processors, 5th Gen Intel Xeon Scalable Processors, and Intel Xeon 6 Processors, all variants thereof, all products incorporating those products, and all products that are substantially similar to those products (the "Accused Xeon Products").²

² Collectively, the Accused Habana Server Products, Accused Habana AI Accelerator Products, and the Accused Xeon Products are referred to herein as the Accused Products.

Asserted Patent	Asserted Claims	Accused Product(s)
		<i>See Exhibit B1 hereto. Each of the Accused Xeon Products is configured and operates in a substantially similar manner in relevant part.</i>
11,563,621	1, 3, 4, 6, 7, 8, 9, 11, 13, 15, 16, 17, 21, 23, 24, 26, 27, 28, 29, 30	All server and workstation products that utilize Intel's Habana AI accelerator architectures, including at least the Gaudi HLS-1 AI Training System, HLS-Gaudi2 server, and Gaudi 3 AI Accelerator HLB-325 Baseboard products, all variants thereof, and all products including to related to those products, as well as any products incorporating those items and any products that are substantially similar to those products (the "Accused Habana Server Products"). All AI accelerator products that utilize Intel's Habana AI accelerator architectures and are especially adapted for use in server and workstation products, including at least the Gaudi, Gaudi 2, Gaudi 3, and Goya AI accelerator products, all variants thereof, and any products that are substantially similar to those products (the "Accused Habana AI Accelerator Products"). <i>See Exhibit A2 hereto. Each of the Accused Habana AI Accelerator Products and each of the Accused Habana Server Products is configured and operates in a substantially similar manner in relevant part.</i>
11,563,621	1, 3, 4, 6, 7, 8, 9, 10 11, 13, 15, 16, 17, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30	All Xeon Scalable Processors (Skylake-SP architecture), 2nd Gen Intel Xeon Scalable Processors, 3rd Gen Intel Xeon Scalable Processors, 4th Gen Intel Xeon Scalable Processors, 5th Gen Intel Xeon Scalable Processors, and Intel Xeon 6 Processors, all

Asserted Patent	Asserted Claims	Accused Product(s)
		variants thereof, all products incorporating those products, and all products that are substantially similar to those products (the “Accused Xeon Products”). <i>See Exhibit B2 hereto. Each of the Accused Xeon Products is configured and operates in a substantially similar manner in relevant part.</i>
11,570,034	1, 2, 3, 8, 10, 24, 25, 27, 28, 30	All server and workstation products that utilize Intel’s Habana AI accelerator architectures, including at least the Gaudi HLS-1 AI Training System, HLS-Gaudi2 server, and Gaudi 3 AI Accelerator HLB-325 Baseboard products, all variants thereof, and all products including to related to those products, as well as any products incorporating those items and any products that are substantially similar to those products (the “Accused Habana Server Products”). All AI accelerator products that utilize Intel’s Habana AI accelerator architectures and are especially adapted for use in server and workstation products, including at least the Gaudi, Gaudi 2, Gaudi 3, and Goya AI accelerator products, all variants thereof, and any products that are substantially similar to those products (the “Accused Habana AI Accelerator Products”). <i>See Exhibit A3 hereto. Each of the Accused Habana AI Accelerator Products and each of the Accused Habana Server Products is configured and operates in a substantially similar manner in relevant part.</i> All Xeon Scalable Processors (Skylake-SP architecture), 2nd Gen Intel Xeon Scalable Processors, 3rd Gen Intel Xeon Scalable

Asserted Patent	Asserted Claims	Accused Product(s)
		Processors, 4th Gen Intel Xeon Scalable Processors, 5th Gen Intel Xeon Scalable Processors, and Intel Xeon 6 Processors, all variants thereof, all products incorporating those products, and all products that are substantially similar to those products (the “Accused Xeon Products”). <i>See Exhibit B3 hereto. Each of the Accused Xeon Products is configured and operates in a substantially similar manner in relevant part.</i>
11,811,582	1, 3, 4, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20	All server and workstation products that utilize Intel’s Habana AI accelerator architectures, including at least the Gaudi HLS-1 AI Training System, HLS-Gaudi2 server, and Gaudi 3 AI Accelerator HLB-325 Baseboard products, all variants thereof, and all products including to related to those products, as well as any products incorporating those items and any products that are substantially similar to those products (the “Accused Habana Server Products”). All AI accelerator products that utilize Intel’s Habana AI accelerator architectures and are especially adapted for use in server and workstation products, including at least the Gaudi, Gaudi 2, Gaudi 3, and Goya AI accelerator products, all variants thereof, and any products that are substantially similar to those products (the “Accused Habana AI Accelerator Products”). <i>See Exhibit A4 hereto. Each of the Accused Habana AI Accelerator Products and each of the Accused Habana Server Products is configured and operates in a substantially similar manner in relevant part.</i>

Asserted Patent	Asserted Claims	Accused Product(s)
11,811,582	1, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	All Xeon Scalable Processors (Skylake-SP architecture), 2nd Gen Intel Xeon Scalable Processors, 3rd Gen Intel Xeon Scalable Processors, 4th Gen Intel Xeon Scalable Processors, 5th Gen Intel Xeon Scalable Processors, and Intel Xeon 6 Processors, all variants thereof, all products incorporating those products, and all products that are substantially similar to those products (the “Accused Xeon Products”). <i>See Exhibit B4 hereto. Each of the Accused Xeon Products is configured and operates in a substantially similar manner in relevant part.</i>
12,021,679	1, 2, 3, 8, 10, 17, 18	All server and workstation products that utilize Intel’s Habana AI accelerator architectures, including at least the Gaudi HLS-1 AI Training System, HLS-Gaudi2 server, and Gaudi 3 AI Accelerator HLB-325 Baseboard products, all variants thereof, and all products including to related to those products, as well as any products incorporating those items and any products that are substantially similar to those products (the “Accused Habana Server Products”). All AI accelerator products that utilize Intel’s Habana AI accelerator architectures and are especially adapted for use in server and workstation products, including at least the Gaudi, Gaudi 2, Gaudi 3, and Goya AI accelerator products, all variants thereof, and any products that are substantially similar to those products (the “Accused Habana AI Accelerator Products”). <i>See Exhibit A5 hereto. Each of the Accused Habana AI Accelerator Products and each of the Accused Habana Server Products is</i>

Asserted Patent	Asserted Claims	Accused Product(s)
		configured and operates in a substantially similar manner in relevant part. All Xeon Scalable Processors (Skylake-SP architecture), 2nd Gen Intel Xeon Scalable Processors, 3rd Gen Intel Xeon Scalable Processors, 4th Gen Intel Xeon Scalable Processors, 5th Gen Intel Xeon Scalable Processors, and Intel Xeon 6 Processors, all variants thereof, all products incorporating those products, and all products that are substantially similar to those products (the “Accused Xeon Products”). <i>See Exhibit B5 hereto. Each of the Accused Xeon Products is configured and operates in a substantially similar manner in relevant part.</i>

ACS does not intend for the product names or model numbers in this chart of Accused Products, or in the appended infringement charts in Exhibits A1-A5 and B1-B5, to be a conclusive list. By identifying product names and model numbers in this chart of Accused Products, and in the appended infringement charts in Exhibits A1-A5 and B1-B5, ACS intends to include all model numbers and configurations for each identified Accused Product. ACS has identified additional model numbers where available based on currently available information. ACS reserves the right to supplement this table of Accused Products as additional information becomes available through discovery in this action.

C. INFRINGEMENT CHARTS

For the Asserted Patent, ACS attaches claim charts (*see* Exhibits A1-A5 and B1-B5) to these Preliminary Infringement Contentions identifying specifically where and how each

limitation of each Asserted Claim is found within each Accused Product. The attached claim charts are exemplary.

ACS has not completed its investigation of the facts relating to this case, has not completed discovery in this action, and has not completed preparation for trial. Therefore, ACS expressly reserves the right to amend, supplement, or otherwise modify the attached charts as additional information becomes available through discovery and/or as necessitated by any claim construction ruling from the Court. *See* OGP at p. 11, n. 8.

D. INDIRECT INFRINGEMENT

In addition to directly infringing the Asserted Patents by, for example, making, using, selling, offering for sale, and/or importing into the United States the Accused Products, Intel has indirectly infringed the Asserted Patents by actively inducing direct infringement of the Asserted Patents by others.

Defendant actively induces others to infringe the Asserted Patents by marketing and selling the Accused Products, knowing, no later than the date ACS served the original complaint in this case on September 26, 2024, and intending that the Accused Products would be used by customers, users, system builders, Intel Partners, and retailers/distributors in a manner that infringes the Asserted Patent.

ACS is informed and believes, and thereon alleges that Intel makes, uses, sells, offers for sale, and/or imports one or more of the Accused Products for or on behalf of third parties such as customers, users, system builders, partners, and retailers/distributors, knowing and intending that the Accused Products will be used by the third parties in a manner that practices the Asserted Claims as shown in Exhibits A1-A5 and B1-B5, respectively. For example, Intel published and provided marketing materials, technical specifications, whitepapers, datasheets, user manuals, and

development and testing information, and other resources on its website (<http://www.intel.com/>) that instructed and encouraged third parties to integrate the Accused Products into products using Intel's technology that were then made, used, sold, offered for sale and/or imported into the United States.³ Intel has also established the "Intel AI Partner ecosystem" to assist customers with training, professional services, and service and support.⁴ These activities were designed to bring infringing products that incorporate Intel's Accused Products to market in the United States. As a result of Intel's activities, the Accused Products have been used in a manner that directly infringes the Asserted Patents. Intel continues to engage in acts of inducement of infringement of the Asserted Patents.

Defendant induces others to directly infringe in numerous ways. As a non-exhaustive exemplary list, a direct infringer may make server products by combining Accused Habana AI Accelerator Products and/or Accused Xeon Products into an infringing system; use a server product owned by the direct infringer, use a server product owned by a third party, use a server product owned by Defendant; sell and/or offer to sell server products made by the direct infringer, sell and/or offer to sell server products acquired from another party, including possibly from Defendant; and/or import server products into the United States.

Based on currently available information, Defendant provides marketing materials, technical specifications, whitepapers, datasheets, user manuals, development and testing

³ *E.g.*, <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi-overview.html>; <https://www.intel.com/content/www/us/en/products/docs/processors/xeon/xeon-scalable-platform-where-to-buy.html>.

⁴ *E.g.*, <https://www.intel.com/content/www/us/en/partner-alliance/membership/benefits/partners-ai.html>.

information, and other resources that instruct and encourage customers, users, system builders, and/or retailers/distributors to make, use, sell, offer for sale, and import the server products in an infringing manner in the United States. Defendant instructs and encourages this infringement by, for example, establishing the Intel AI Partner ecosystem to assist Intel partners with marketing, training, professional services, sales and distribution, and service and support for the server products. The server products embody or practice the Asserted Claims as set forth in Exhibits A1-A5 and B1-B5.

Additionally, Defendant provides marketing materials, technical specifications, whitepapers, datasheets, user manuals, development and testing information, reference architectures, and other resources that instruct and encourage customers, users, system builders, Intel AI Partners, and retailers/distributors to use the Accused Products in an infringing manner in the United States. For example, Intel provides instructions for integrating the Accused Products using into other products that infringe the Asserted Patents. These third parties then use the materials and support provided by Defendant to make, use, sell, offer for sale, and/or import into the United States infringing products that use the Accused Products therein. Defendant instructs and encourages this by, for example, establishing the Intel AI Partner ecosystem to assist Intel Partners with marketing, training, professional services, sales and distributions, and service and support.

Intel also encourages others to provide, or provides itself, use of Accused Products in a compute as a service model. Their compute as a service model encourages use of Accused Products by a party different from the owner of the Accused Products. This model encourages many infringing users to share access to and use of the Accused Products. Intel also offers support for Accused Products allowing, and inducing, infringing use to continue.

In view of at least Defendant's knowledge of the Asserted Patents, Defendant knew or should have known that such activities would cause direct infringement of the Asserted Patents by Defendant's customers, end users, system builders, Intel AI Partners, and/or retailers/distributors of the Accused Products. Defendant's inducement constitutes infringement of the Asserted Patents in violation of 35 U.S.C. § 271(b).

ACS has not completed its investigation of the facts relating to this case, has not completed discovery in this action, and has not completed preparation for trial. Therefore, ACS expressly reserves the right to amend, supplement, or otherwise modify the attached charts as additional information becomes available through discovery. *See* OGP at p. 11, n. 8.

E. LITERAL INFRINGEMENT AND INFRINGEMENT UNDER THE DOCTRINE OF EQUIVALENTS

Based on currently available information, ACS alleges that each of the Accused Products literally infringes each of the Asserted Claims. *See Supra* §§ I.B, I.C & I.D, and Exhibits A1-A5 and B1-B5 hereto.

To the extent any claim limitations are not literally infringed by any Accused Product, ACS contends the limitations are infringed under the doctrine of equivalents because the identified features of the Accused Products are insubstantially different from the literal claim elements, for example because they perform substantially the same function in substantially the same way to achieve substantially the same result. ACS has not completed its investigation of the facts relating to this case and has not yet completed discovery in this action. Furthermore, the Asserted Claims have not yet been construed by the Court, nor has Defendant addressed claim construction. Defendant also has not yet set forth its bases, if any, for non-infringement of the Asserted Claims. Accordingly, ACS reserves the right to amend, supplement, or otherwise modify these

infringement contentions as additional information becomes available through the claim construction process and discovery, including the right to further supplement its contentions with regard to infringement pursuant to the doctrine of equivalents. *See* OGP at p. 11, n. 8.

F. PRIORITY DATE OF ASSERTED CLAIMS

ACS contends that the Asserted Claims of the Asserted Patents are entitled to a priority date no later than the filing of provisional application No. 60/813,738, which was filed on June 13, 2006, and of provisional application No. 60/850,908, which was filed on October 11, 2006. The Asserted Patents claim priority to both provisional applications on their faces as related U.S. Applications.

II. DOCUMENT PRODUCTION ACCOMPANYING DISCLOSURE

Pursuant to the Court's OGP, at p. 2 (Section II.2), concurrently with the service of its foregoing disclosures and contentions, ACS will be producing via an FTP link today a copy of the file history for each Asserted Patent and documents evidencing conception and reduction to practice for each claimed invention (ACS_0000167 – ACS_0076737).

These disclosures, contentions, and document production are based on information currently available to ACS. Fact and expert discovery in this action has not begun, and ACS reserves the right to modify, amend, or otherwise supplement its disclosures, contentions, and document production as new documents and information are revealed through discovery and/or to the extent Defendant alters, supplants, or otherwise clarifies documents and information currently available to ACS. *See* OGP at p. 11, n. 8.

Respectfully submitted,

By: /s/ Reynaldo C. Barceló

BARCELÓ, HARRISON & WALKER, LLP

Reynaldo C. Barceló (admitted *pro hac vice*)
rey@bhiplaw.com
2901 West Coast Hwy, Suite 200
Newport Beach, CA 92663
Telephone: (949) 340-9736

CROWELL & MORING LLP

Jon Gurka (admitted *pro hac vice*)
jgurka@crowell.com
Kainoa Asuega (admitted *pro hac vice*)
kasuega@crowell.com
3 Park Plaza, 20th Floor
Irvine, CA 92614
Telephone: (949) 263-8400

David Lindner (admitted *pro hac vice*)
dlindner@crowell.com
Andrew McElligott (admitted *pro hac vice*)
amcelligott@crowell.com
455 N Cityfront Plaza Drive, Ste 3600
Chicago, IL 60611
Telephone: (312) 321-4200

Michelle Wang
michellewang@crowell.com
Two Manhattan West
375 Ninth Avenue
New York, NY 10001
Telephone: (212) 223-4000

CHERRY JOHNSON SIEGMUND JAMES, PLLC

Mark D. Siegmund (TX Bar No. 24117055)

msiegmund@cjsjlaw.com

William D. Ellerman (TX Bar No. 24007151)

wellerman@cjsjlaw.com

7901 Fish Pond Rd., 2nd Floor

Waco, TX 76710

Telephone: (254) 732-2242

Attorneys for Plaintiff

Advanced Cluster Systems, Inc.

CERTIFICATE OF SERVICE

I hereby certify that on January 22, 2025, I served the foregoing document (including the accompanying claim chart exhibits A1-A5 and B1-B5 identified therein) via e-mail on Defendant's counsel of record at Intel-ACS@mto.com and BNash@mofo.com.

Dated: January 22, 2025

By: /s/ Reynaldo C. Barceló
Reynaldo C. Barceló

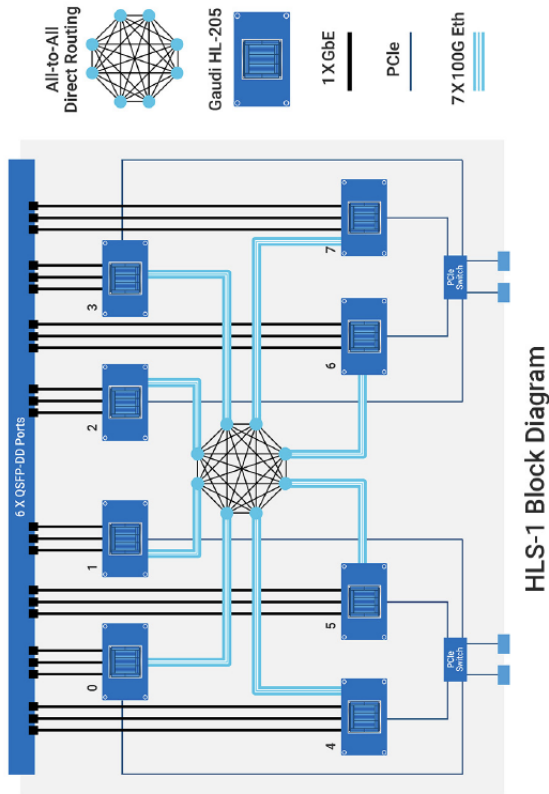
Exhibit A4
U.S. Pat. No. 11,811,582 in view of
Accused Habana Server Products and Accused Habana AI Accelerator Products

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
1. A computer cluster comprising:	Each of the Accused Habana Server Products¹ comprises a computer cluster. For example, as depicted below and as described further herein, Intel's Gaudi® HLS-1 AI Training System ("HLS-1") includes a computer cluster.  Each of the Accused Habana Server Products includes a plurality of nodes, each of which comprises an Accused Habana AI Accelerator Product.² For example, as depicted below, the Intel HLS-1 system

¹ The Accused Habana Server Products include, but are not limited to, all products including or related to Intel's Gaudi® HLS-1 AI Training System ("HLS-1") (<https://www.intel.com/content/www/us/en/content-details/784787/gaudi-hls-1-ai-training-system.html>), HLS-Gaudi2 server (https://habana.ai/wp-content/uploads/2023/10/HLS-Gaudi2_Datasheet_10_23.pdf), Gaudi 3 AI Accelerator HLB-325 Baseboard (<https://www.intel.com/content/www/us/en/content-details/817489/intel-gaudi-3-ai-accelerator-hlb-325-baseboard-product-brief.html>), as well as any products incorporating those items.

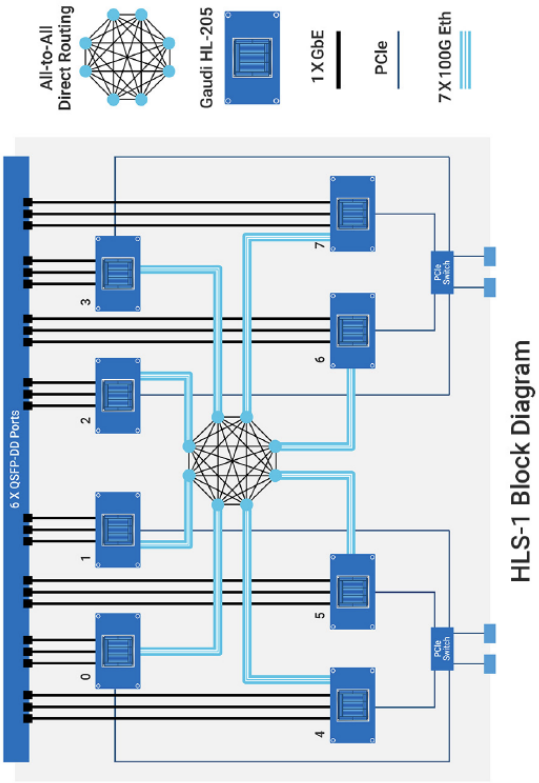
² Accused Habana AI Accelerator Products include, but are not limited to all products including or related to intel's Gaudi (https://docs.habana.ai/en/latest/Gaudi_Overview/Gaudi_Architecture.html; <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi.html>), Gaudi 2 (<https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi2.html>), Gaudi 3 (<https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi3.html>), and Goya (https://habana.ai/wp-content/uploads/pdf/habana_labs_goya_whitepaper.pdf) AI accelerator products, as well as any products incorporating those items.

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products comprises eight Gaudi HL-205 Mezzanine cards (designated as nodes “0”, “1” ... “7” in the figure below) connected in an all-to-all local bus routing by eight 7x 100G Ethernet connections.³
a first processor; a second processor; a third processor;	Each of the Accused Habana Server Products comprises a plurality of nodes, each of which comprises an Accused Habana AI Accelerator Product. For example, as depicted below, the Intel HLS-1 system comprises eight Gaudi HL-205 Mezzanine cards (designated as nodes “0”, “1” ... “7” in the figure below) connected in an all-to-all local bus routing by eight 7x 100G Ethernet connections.⁴

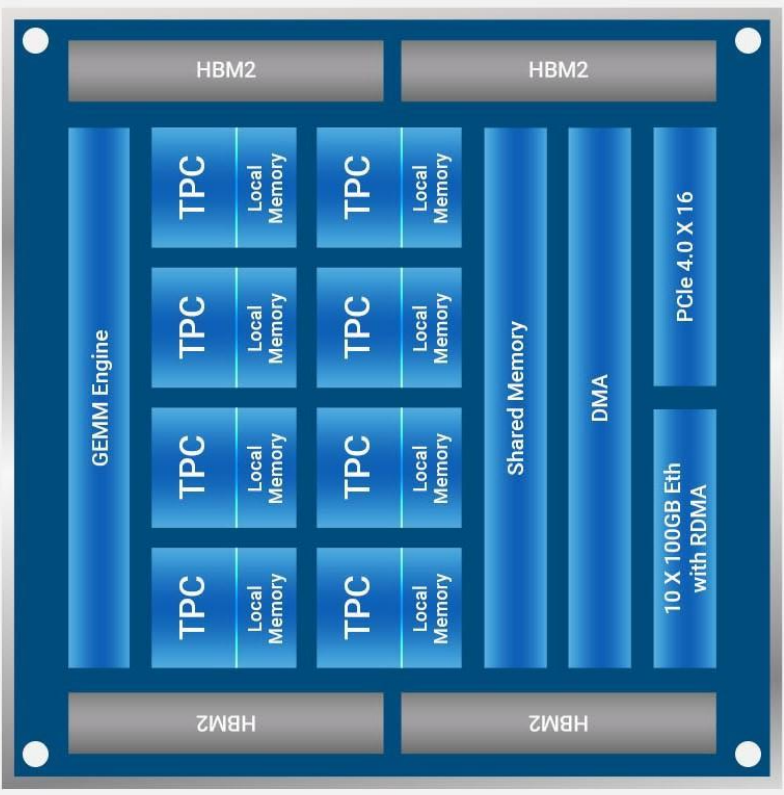


³ Gaudi® HLS-1 AI Training System (<https://www.intel.com/content/www/us/en/content-details/784787/gaudi-hls-1-ai-training-system.html>); see also HLS-1 introductory video (“Habana Labs: How to Scale AI Training with Gaudi Processors”), at <https://www.youtube.com/watch?v=u0siCfmCNfg>; see also <https://habana.ai/products/gaudi/>.

⁴ Gaudi® HLS-1 AI Training System (<https://www.intel.com/content/www/us/en/content-details/784787/gaudi-hls-1-ai-training-system.html>); see also HLS-1 introductory video (“Habana Labs: How to Scale AI Training with Gaudi Processors”), at <https://www.youtube.com/watch?v=u0siCfmCNfg>; see also <https://habana.ai/products/gaudi/>.

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	 HLS-1 Block Diagram Each of the plurality of Accused Habana AI Accelerator Product nodes in each Accused Habana Server Product comprises a hardware processor. Specifically, each Accused Habana AI Accelerator Product comprises a hardware processor that comprises multiple processor cores. For example, each Gaudi HL-205 Mezzanine card is a node that includes a Gaudi HL-2000 processor that contains a cluster of eight programmable Tensor Processing Cores (TPC) and a General Matrix Multiplication Engine (GEMM). The architecture of the Gaudi HL-2000 processor is depicted in the following figure:⁵

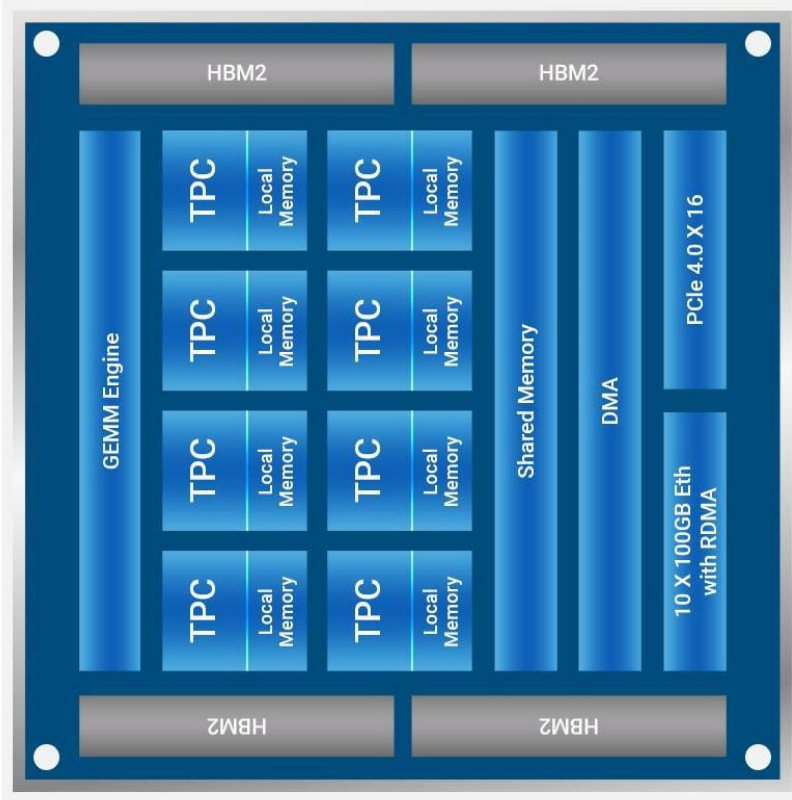
⁵ Gaudi™ Training Platform White Paper, Nov 2020, p. 6, (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>); see also https://docs.habana.ai/en/latest/Gaudi_Overview/Gaudi_Architecture.html; <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi2.html> (Gaudi 2), <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi3.html> (Gaudi 3), https://habana.ai/wp-content/uploads/pdf/habana_labs_goya_whitepaper.pdf (Goya).

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	 Thus, each of the Accused Habana Server Products comprises a first processor, a second processor, and a third processor. Each of the Accused Habana Server Products comprises at least one computer-readable medium in communication with at least one of the first processor, the second processor, or the third processor, and a first kernel residing in the at least one computer-readable medium, said first kernel configured to translate user commands into calls for execution of matrix operations on an array of data elements.

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products Each of the Accused Habana Server Products comprises a plurality of nodes, each of which comprises an Accused Habana AI Accelerator Product. For example, as depicted below, the Intel HLS-1 system comprises eight Gaudi HL-205 Mezzanine cards (designated as nodes “0”, “1” ... “7” in the figure below) connected in an all-to-all local bus routing by eight 7x 100G Ethernet connections.⁶ HLS-1 Block Diagram Each Accused Habana Server Product comprises at least one computer-readable medium in communication with at least one of the first processor, the second processor, or the third processor in the Accused Habana Server Product. For example, each Gaudi HL-205 Mezzanine card is a node that includes a Gaudi HL-2000 processor that contains a cluster of eight programmable Tensor Processing Cores (TPC) and a General Matrix Multiplication Engine (GEMM). The architecture of the Gaudi HL-2000 processor is depicted in the following figure:⁷
--	---

⁶ Gaudi® HLS-1 AI Training System (<https://www.intel.com/content/www/us/en/content-details/784787/gaudi-hls-1-ai-training-system.html>); see also HLS-1 introductory video (“Habana Labs: How to Scale AI Training with Gaudi Processors”), at <https://www.youtube.com/watch?v=u0siCfmCNfg>; see also <https://habana.ai/products/gaudi/>.

⁷ Gaudi™ Training Platform White Paper, Nov 2020, p. 6, (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>); see also

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	 The diagram illustrates the internal architecture of a Habana AI Accelerator. It features a central processing area with a GEMM Engine at the top. Below the engine are eight TPC (Tensor Processing Cluster) units, each paired with its own Local Memory. These are flanked by four HBM2 (High Bandwidth Memory 2) blocks. A Shared Memory block is positioned below the TPCs. At the bottom, there is a DMA (Direct Memory Access) block and a PCIe 4.0 X 16 interface. The entire system is connected to 10 X 100GB Eth with RDMA (Remote Direct Memory Access) capabilities. Each Accused Habana AI Accelerator Product is configured to access and be in communication with a non-transitory computer-readable medium (e.g., memory accessible by and in communication with by

https://docs.habana.ai/en/latest/Gaudi_Overview/Gaudi_Architecture.html; [https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/audi.html](https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/details/processors/ai-accelerators/audi.html) (Gaudi), <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/audi2.html> (Gaudi 2), <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/audi3.html> (Gaudi 3), https://habana.ai/wp-content/uploads/pdf/habana_labs_goya_whitepaper.pdf (Goya).

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products a Gaudi HL-2000 processor, including local memory, shared memory, and HBM2 memory, as well as memory external to the Gaudi HL-2000 processor).⁸ The computer-readable medium accessible to and in communication with each Accused Habana AI Accelerator Product in each Accused Habana Server Product comprises program code that is configured to translate user commands into calls for execution of matrix operations on an array of data elements (<i>e.g.</i>, executable SynapseAI® Software Suite program code stored in memory accessible by and in communication with the Gaudi HL-2000 processor in a Gaudi HL-205 mezzanine card). This is done, for example, by using the Habana Collective Communication Library (HCCL).⁹ HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation."¹⁰ For example, code stored in the memory accessible by and in communication with a Gaudi HL-2000 processor is configured to translate user commands into calls for execution of matrix operations on an array of data elements (<i>e.g.</i>, by performing matrix multiplications using the General Matrix Multiplication ("GEMM") engine, evaluating mathematical expressions using the Tensor Processing Cores in the Gaudi HL-2000 processor, and/or executing commands in the HCCL library). TPCs are configured to evaluate multiple mathematical expressions, such as square root, sine, cosine, tangent, and reciprocal, as well as matrix operations such as Vector Reduction Intrinsics.¹¹ Each of the plurality of Accused Habana AI Accelerator Product nodes in each Accused Habana Server Product comprises a first cluster node module residing in the at least one computer-readable medium, said first cluster node module configured to communicate with the first kernel and receive commands from a user interface.
--------------------------	--

⁸ GaudiTM Training Platform White Paper, Nov 2020, p. 6, (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

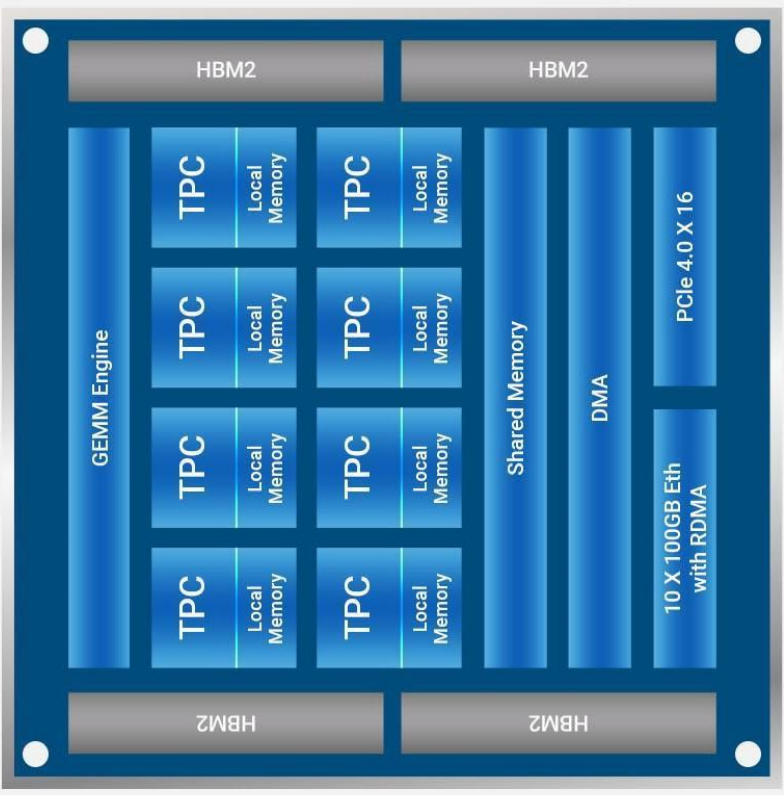
⁹ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

¹⁰ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

¹¹ See, *e.g.*, Built-in Functions ("(https://docs.habana.ai/en/latest/TPC_User_Guide/Built_in_Functions.html)").

U.S. Pat. No. 11,811,582 node module configured to communicate with the first kernel and receive commands from a user interface;	Accused Habana Server Products and Accused Habana AI Accelerator Products Each Accused Habana AI Accelerator Product comprises a hardware processor. For example, each Gaudi HL-205 Mezzanine card is a node that includes a Gaudi HL-2000 processor that contains a cluster of eight programmable Tensor Processing Cores (TPC) and a General Matrix Multiplication Engine (GEMM), as depicted in the following figure: ¹²
--	--

¹² Gaudi™ Training Platform White Paper, Nov 2020, p. 6, (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>); see also https://docs.habana.ai/en/latest/Gaudi_Overview/Gaudi_Architecture.html; <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi.html> (Gaudi), <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi2.html> (Gaudi 2), <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi3.html> (Gaudi 3), https://habana.ai/wp-content/uploads/pdf/habana_labs_goya_whitepaper.pdf (Goya).

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	 The diagram illustrates the internal architecture of a Habana AI Accelerator. It features a central processing area with a GEMM Engine at the top, followed by four TPC (Tensor Processing Core) units, each paired with its own Local Memory. This central block is flanked by four HBM2 (High Bandwidth Memory 2) modules. Below the TPCs is a Shared Memory block, followed by a DMA (Direct Memory Access) block. At the bottom, there are two PCIe 4.0 X16 interfaces, one labeled '10 X 100GB Eth with RDMA'. The entire system is housed within a blue frame representing the accelerator card. Each Accused Habana AI Accelerator Product is configured to access and be in communication with a non-transitory computer-readable medium (e.g., memory accessible by and in communication with by a Gaudi HL-2000 processor, including local memory, shared memory, and HBM2 memory, as well as memory external to the Gaudi HL-2000 processor).¹³

¹³ Gaudi™ Training Platform White Paper, Nov 2020, p. 6, (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

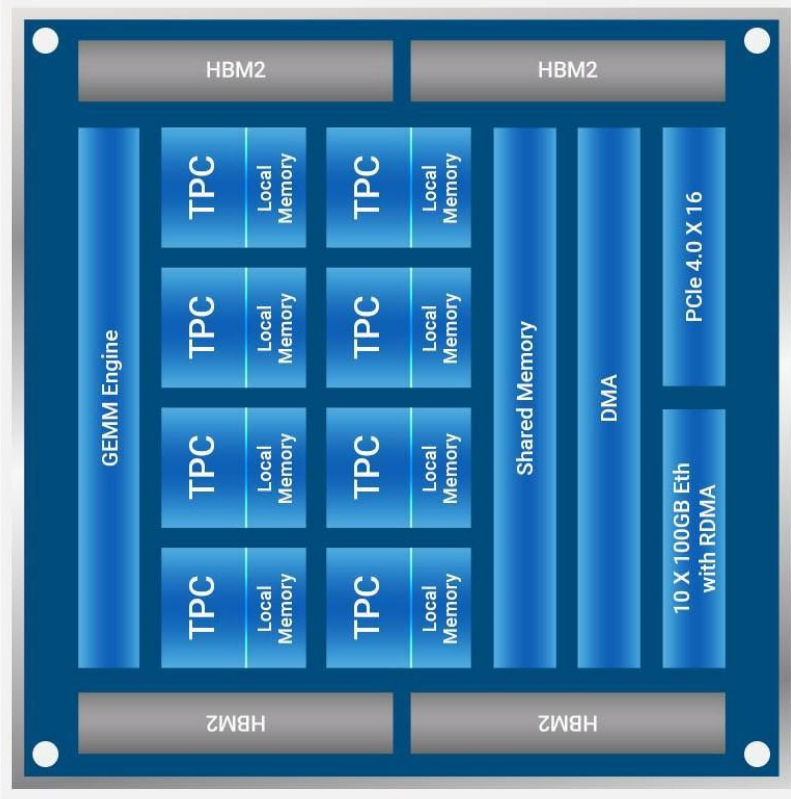
U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	For example, the Habana Collective Communication Library (HCCL) includes commands that cause a Gaudi HL-2000 processor of a Gaudi HL-205 mezzanine card in a HLS-1 system to communicate with the first kernel and receive commands from a user interface.¹⁴ HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation."¹⁵ The source code for Intel's HCCL Demo program¹⁶ includes the details for using HCCL commands to initialize a HCCL communicator on a HLS-1 system comprising multiple Gaudi HL-205 Mezzanine card nodes, communicate with the first kernel and receive commands from a user interface.
a second kernel residing in the at least one computer-readable medium, said second kernel configured to execute the matrix operations on the array of data elements and return a first result using the second processor;	Each of the plurality of Accused Habana AI Accelerator Product nodes in each Accused Habana Server Product comprises a second kernel residing in the at least one computer-readable medium, said second kernel configured to execute the matrix operations on the array of data elements and return a first result using the second processor. Each Accused Habana AI Accelerator Product comprises a hardware processor. For example, each Gaudi HL-205 Mezzanine card is a node that includes a Gaudi HL-2000 processor that contains a cluster of eight programmable Tensor Processing Cores (TPC) and a General Matrix Multiplication Engine (GEMM), as depicted in the following figure:¹⁷

¹⁴ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

¹⁵ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

¹⁶ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py, https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

¹⁷ Gaudi™ Training Platform White Paper, Nov 2020, p. 6, (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>); see also https://docs.habana.ai/en/latest/Gaudi_Overview/Gaudi_Architecture.html; <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/ai-accelerators/audi2.html> (Gaudi 2), <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/audi3.html> (Gaudi 3), https://habana.ai/wp-content/uploads/pdf/habana_labs_goya_whitepaper.pdf (Goya).

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	 The diagram illustrates the internal architecture of a Habana AI Accelerator. It features a central processing area with a GEMM Engine at the top and eight TPC (Tensor Processing Cluster) units arranged in a 2x4 grid below it. Each TPC unit is associated with its own Local Memory. This central processing area is flanked by four HBM2 (High Bandwidth Memory 2) blocks, two on each side. Below the TPC units is a Shared Memory block. At the bottom of the architecture are DMA (Direct Memory Access) and PCIe 4.0 X 16 interfaces, with a note indicating 10 X 100GB Eth with RDMA. The entire system is enclosed in a blue frame with four corner mounting holes. Each Accused Habana AI Accelerator Product is configured to access and be in communication with a non-transitory computer-readable medium (e.g., memory accessible by and in communication with by a Gaudi HL-2000 processor, including local memory, shared memory, and HBM2 memory, as well as memory external to the Gaudi HL-2000 processor).¹⁸

¹⁸ Gaudi™ Training Platform White Paper, Nov 2020, p. 6, (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

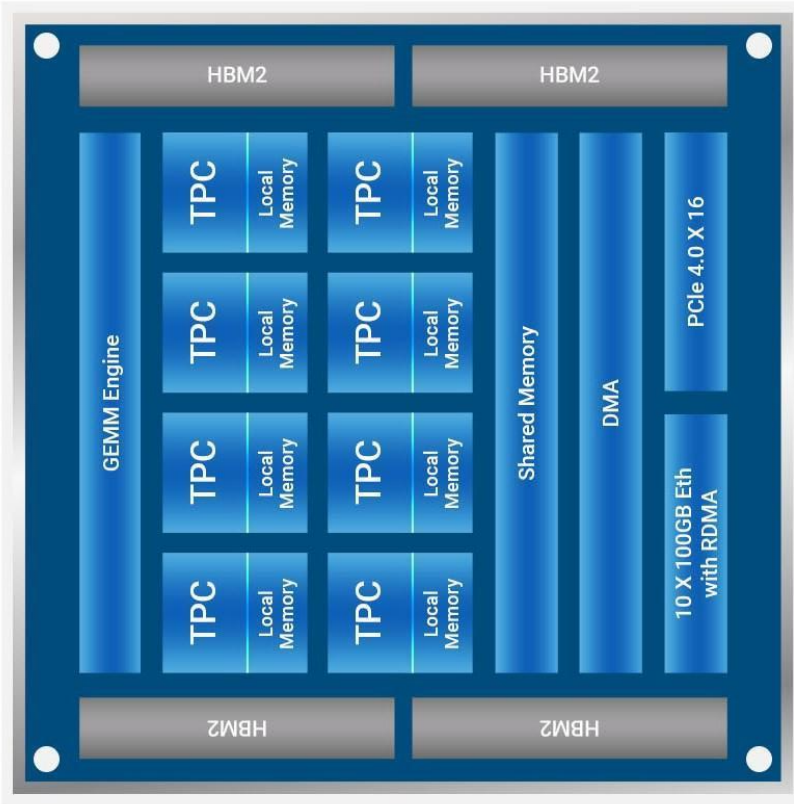
U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products For example, the Habana Collective Communication Library (HCCL) includes commands that cause a Gaudi HL-2000 processor of a second Gaudi HL-205 mezzanine card in a HLS-1 system to execute the matrix operations on the array of data elements and return a first result using the Gaudi HL-2000 processor of the second Gaudi HL-205 mezzanine card (<i>e.g.</i>, by performing matrix multiplications using the General Matrix Multiplication (“GEMM”) engine, evaluating mathematical expressions using the Tensor Processing Cores in the Gaudi HL-2000 processor, and/or executing commands in the HCCL library).¹⁹ TPCs are configured to evaluate multiple mathematical expressions, such as square root, sine, cosine, tangent, and reciprocal, as well as matrix operations such as Vector Reduction Intrinsic.²⁰ HCCL is included in Intel’s SynapseAI® Software Suite. SynapseAI® is “Habana’s complete software stack custom designed to support Habana’s Gaudi implementation.”²¹ The source code for Intel’s HCCL Demo program²² includes the details for using HCCL commands to initialize a HCCL communicator on a HLS-1 system comprising multiple Gaudi HL-205 Mezzanine card nodes and to execute the matrix operations on the array of data elements and return a first result using the Gaudi HL-2000 processor of the second Gaudi HL-205 mezzanine card in the HLS-1 system.
a second cluster node module residing in the at least one computer-readable medium, said second cluster node module configured to send commands to the second kernel and communicate	Each of the plurality of Accused Habana AI Accelerator Product nodes in each Accused Habana Server Product comprises a second cluster node module residing in the at least one computer-readable medium, said second cluster node module configured to send commands to the second kernel and communicate with the first cluster node module. Each Accused Habana AI Accelerator Product comprises a hardware processor. For example, each Gaudi HL-205 Mezzanine card is a node that includes a Gaudi HL-2000 processor that contains a

¹⁹ “Habana Collective Communications Library (HCCL) is Habana’s emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library.” (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

²⁰ See, *e.g.*, Built-in Functions (“https://docs.habana.ai/en/latest/TPC/User_Guide/Built_in_Functions.html”).

²¹ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

²² See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py, https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

U.S. Pat. No. 11,811,582 with the first cluster node module;	Accused Habana Server Products and Accused Habana AI Accelerator Products cluster of eight programmable Tensor Processing Cores (TPC) and a General Matrix Multiplication Engine (GEMM), as depicted in the following figure:²³ 
--	---

²³ Gaudi™ Training Platform White Paper, Nov 2020, p. 6, (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>); see also https://docs.habana.ai/en/latest/Gaudi_Overview/Gaudi_Architecture.html; <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/audi2.html> (Gaudi 2), <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/audi3.html> (Gaudi 3), https://habana.ai/wp-content/uploads/pdf/habana_labs_goya_whitepaper.pdf (Goya).

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	Each Accused Habana AI Accelerator Product is configured to access and be in communication with a non-transitory computer-readable medium (e.g., memory accessible by and in communication with a Gaudi HL-2000 processor, including local memory, shared memory, and HBM2 memory, as well as memory external to the Gaudi HL-2000 processor).²⁴ For example, the Habana Collective Communication Library (HCCL) includes commands that cause a Gaudi HL-2000 processor of a Gaudi HL-205 mezzanine card in a HLS-1 system to send commands to the second kernel and communicate with the first cluster node module.²⁵ HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation."²⁶ The source code for Intel's HCCL Demo program²⁷ includes the details for using HCCL commands to initialize a HCCL communicator on a HLS-1 system comprising multiple Gaudi HL-205 Mezzanine card nodes, send commands to the second kernel and communicate with the first cluster node module.
a third kernel residing in the at least one computer-readable medium, said third kernel configured to receive the first result and execute additional matrix operations on the first result and return a second result to one or more of the first, second, or third cluster node modules using the third processor. Each Accused Habana AI Accelerator Product comprises a hardware processor. For example, each Gaudi HL-205 Mezzanine card is a node that includes a Gaudi HL-2000 processor that contains a	Each of the plurality of Accused Habana AI Accelerator Product nodes in each Accused Habana Server Product comprises a third kernel residing in the at least one computer-readable medium, said third kernel configured to receive the first result and execute additional matrix operations on the first result and return a second result to one or more of the first, second, or third cluster node modules using the third processor. Each Accused Habana AI Accelerator Product comprises a hardware processor. For example, each Gaudi HL-205 Mezzanine card is a node that includes a Gaudi HL-2000 processor that contains a

²⁴ Gaudi™ Training Platform White Paper, Nov 2020, p. 6, (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

²⁵ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

²⁶ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

²⁷ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py, https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

U.S. Pat. No. 11,811,582	first, second, or third cluster node modules using the third processor; and
Accused Habana Server Products and Accused Habana AI Accelerator Products cluster of eight programmable Tensor Processing Cores (TPC) and a General Matrix Multiplication Engine (GEMM), as depicted in the following figure:²⁸	

²⁸ Gaudi™ Training Platform White Paper, Nov 2020, p. 6, (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>); see also https://docs.habana.ai/en/latest/Gaudi_Overview/Gaudi_Architecture.html; <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/audi2.html> (Gaudi 2), <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/audi3.html> (Gaudi 3), https://habana.ai/wp-content/uploads/pdf/habana_labs_goya_whitepaper.pdf (Goya).

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products Each Accused Habana AI Accelerator Product is configured to access and be in communication with a non-transitory computer-readable medium (<i>e.g.</i>, memory accessible by and in communication with a Gaudi HL-2000 processor, including local memory, shared memory, and HBM2 memory, as well as memory external to the Gaudi HL-2000 processor).²⁹ For example, the Habana Collective Communication Library (HCCL) includes commands that cause a Gaudi HL-2000 processor of a Gaudi HL-205 mezzanine card in a HLS-1 system to receive the first result and execute additional matrix operations on the first result and return a second result to one or more of the first, second, or third cluster node modules using the Gaudi HL-2000 processor of a third Gaudi HL-205 mezzanine card in a HLS-1 system (<i>e.g.</i>, by performing matrix multiplications using the General Matrix Multiplication (“GEMM”) engine, evaluating mathematical expressions using the Tensor Processing Cores in the Gaudi HL-2000 processor, and/or executing commands in the HCCL library).³⁰ TPCs are configured to evaluate multiple mathematical expressions, such as square root, sine, cosine, tangent, and reciprocal, as well as matrix operations such as Vector Reduction Intrinsic.³¹ HCCL is included in Intel’s SynapseAI® Software Suite. SynapseAI® is “Habana’s complete software stack custom designed to support Habana’s Gaudi implementation.”³² The source code for Intel’s HCCL Demo program³³ includes the details for using HCCL commands to initialize a HCCL communicator on a HLS-1 system comprising multiple Gaudi HL-205 Mezzanine card nodes, receive the first result, and execute additional matrix operations on the first result and return a second result to one or more of the first, second, or third cluster node modules using the Gaudi HL-2000 processor of a third Gaudi HL-205 mezzanine card in a HLS-1 system.
--------------------------	---

²⁹ GaudiTM Training Platform White Paper, Nov 2020, p. 6, (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

³⁰ “Habana Collective Communications Library (HCCL) is Habana’s emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library.” (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

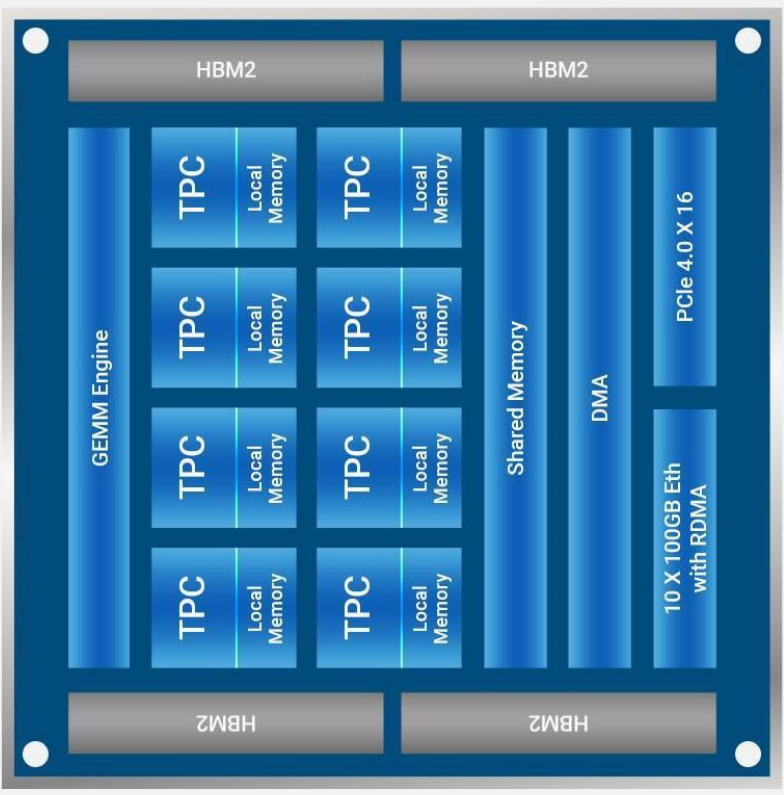
³¹ See, *e.g.*, Built-in Functions (“https://docs.habana.ai/en/latest/TPC_User_Guide/Built_in_Functions.html”).

³² Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

³³ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py, https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
a third cluster node module residing in the at least one computer-readable medium, said third cluster node module configured to send commands to the third cluster node module configured to send commands to the third cluster node module. Each Accused Habana AI Accelerator Product comprises a hardware processor. For example, each Gaudi HL-205 Mezzanine card is a node that includes a Gaudi HL-2000 processor that contains a cluster of eight programmable Tensor Processing Cores (TPC) and a General Matrix Multiplication Engine (GEMM), as depicted in the following figure:³⁴	Each of the plurality of Accused Habana AI Accelerator Product nodes in each Accused Habana Server Product comprises a third cluster node module residing in the at least one computer-readable medium, said third cluster node module configured to send commands to the third cluster node module configured to communicate with the first cluster node module and the second cluster node module. Each Accused Habana AI Accelerator Product comprises a hardware processor. For example, each Gaudi HL-205 Mezzanine card is a node that includes a Gaudi HL-2000 processor that contains a cluster of eight programmable Tensor Processing Cores (TPC) and a General Matrix Multiplication Engine (GEMM), as depicted in the following figure:³⁴

³⁴ Gaudi™ Training Platform White Paper, Nov 2020, p. 6, (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>); see also https://docs.habana.ai/en/latest/Gaudi_Overview/Gaudi_Architecture.html; <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi.html> (Gaudi), <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi2.html> (Gaudi 2), <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi3.html> (Gaudi 3), https://habana.ai/wp-content/uploads/pdf/habana_labs_goya_whitepaper.pdf (Goya).

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	 The diagram illustrates the internal architecture of the Habana AI Accelerator. It features a central processing unit (a dark blue rectangle) containing several components: a GEMM Engine at the top, four TPC (Tensor Processing Cluster) blocks arranged in a 2x2 grid, each with its own Local Memory, a Shared Memory block, a DMA (Direct Memory Access) block, and a PCIe 4.0 X 16 interface. The unit is flanked by four HBM2 (High Bandwidth Memory 2) blocks, two on each side. Below the unit, there is a block for 10 X 100GB Eth with RDMA. Each Accused Habana AI Accelerator Product is configured to access and be in communication with a non-transitory computer-readable medium (e.g., memory accessible by and in communication with a Gaudi HL-2000 processor, including local memory, shared memory, and HBM2 memory, as well as memory external to the Gaudi HL-2000 processor).³⁵

³⁵ Gaudi™ Training Platform White Paper, Nov 2020, p. 6, (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20Gaudi%20Training%20Whitepaper%20v1.2.pdf>).

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products For example, the Habana Collective Communication Library (HCCL) includes commands that cause a Gaudi HL-2000 processor of a Gaudi HL-205 mezzanine card in a HLS-1 system to send commands to the third kernel and configured to communicate with the first cluster node module and the second cluster node module.³⁶ HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation."³⁷ The source code for Intel's HCCL Demo program³⁸ includes the details for using HCCL commands to initialize a HCCL communicator on a HLS-1 system comprising multiple Gaudi HL-205 Mezzanine card nodes, send commands to the third kernel and communicate with the first cluster node module and the second cluster node module.
wherein the first cluster node module is configured to return, based on the second result, a final result.	Each of the plurality of Accused Habana AI Accelerator Product nodes in each Accused Habana Server Product comprises a first cluster node module configured to return, based on the second result, a final result. For example, the Habana Collective Communication Library (HCCL) includes commands that cause a Gaudi HL-2000 processor of a Gaudi HL-205 mezzanine card in a HLS-1 system to return, based on the second result, a final result.³⁹ HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation."⁴⁰

³⁶ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

³⁷ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

³⁸ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also

https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py, https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

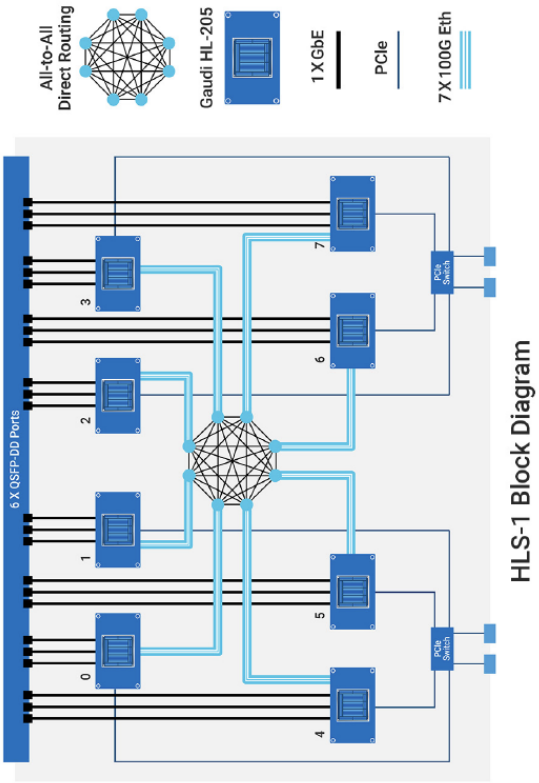
³⁹ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

⁴⁰ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	The source code for Intel's HCCL Demo program⁴¹ includes the details for using HCCL commands to initialize a HCCL communicator on a HLS-1 system comprising multiple Gaudi HL-205 Mezzanine card nodes and return, based on the second result, a final result.
3. The computer cluster of claim 1, wherein the at least one computer-readable medium comprises a first computer-readable medium coupled to at least the first processor and a second computer-readable medium coupled to at least the second processor, wherein the first kernel and the first cluster node module reside in the first computer-readable medium, and wherein the second kernel and the second cluster node module reside in the second computer-readable medium.	Each of the Accused Habana Server Products comprises a computer cluster with a first computer-readable medium coupled to a first processor and a second computer-readable medium coupled to a second processor, wherein a first kernel and a first cluster node module reside in the first computer-readable medium, and wherein a second kernel and a second cluster node module reside in the second computer-readable medium. Each of the Accused Habana Server Products comprises a plurality of nodes, each of which comprises an Accused Habana AI Accelerator Product. For example, as depicted below, the Intel HLS-1 system comprises eight Gaudi HL-205 Mezzanine cards (designated as nodes "0", "1" ... "7" in the figure below) connected in an all-to-all local bus routing by eight 7x 100G Ethernet connections.⁴²

⁴¹ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py, https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

⁴² Gaudi® HLS-1 AI Training System (<https://www.intel.com/content/www/us/en/content-details/784787/audi-hls-1-ai-training-system.html>); see also HLS-1 introductory video ("Habana Labs: How to Scale AI Training with Gaudi Processors"), at <https://www.youtube.com/watch?v=u0siCfmCNfg>; see also <https://habana.ai/products/gaudi/>.

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	 HLS-1 Block Diagram Each Accused Habana Server Product comprises a first computer-readable medium coupled to a first processor and a second computer-readable medium coupled to a second processor, wherein a first kernel⁴³ and a first cluster node module reside in the first computer-readable medium, and wherein a second kernel and a second cluster node module reside in the second computer-readable medium of the Accused Habana Server Product. For example, each Gaudi HL-205 Mezzanine card is a node that includes a Gaudi HL-2000 processor that contains a cluster of eight programmable Tensor Processing Cores (TPC) and a General Matrix Multiplication Engine (GEMM). The architecture of the Gaudi HL-2000 processor is depicted in the following figure:⁴⁴

⁴³ <https://developer.habana.ai/get-started/kernel-libraries/>.

⁴⁴ Gaudi™ Training Platform White Paper, Nov 2020, p. 6, (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>); see also https://docs.habana.ai/en/latest/Gaudi_Overview/Gaudi_Architecture.html; <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi2.html> (Gaudi 2), <https://www.intel.com/content/www/us/en/products/details/processors/ai-accelerators/gaudi3.html> (Gaudi 3), https://habana.ai/wp-content/uploads/pdf/habana_labs_goya_whitepaper.pdf (Goya).

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	Each Accused Habana AI Accelerator Product includes computer readable memory coupled with corresponding processors (<i>e.g.</i>, local memory, shared memory, external to the Gaudi HL-2000 processor, or HBM2 memory accessible by and in communication with, the Gaudi HL-2000 processor and its tensor processing cores.⁴⁵ Cluster node module and kernel software (<i>e.g.</i>, executable SynapseAI® Software Suite program code, which may include software of the Habana Collective Communication Library (HCCL)), necessarily

⁴⁵ Gaudi™ Training Platform White Paper, Nov 2020, p. 6, (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	resides in memory accessible by and in communication with the executing Gaudi HL-2000 processor of a Gaudi HL-205 mezzanine card of an Accused Habana AI Accelerator Product in each Accused Habana Server Product. ⁴⁶ HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation." ⁴⁷
4. The computer cluster of claim 1, further comprising a cluster configuration module that initializes at least one of the first cluster node module, the second cluster node module, or the third cluster node module.	Each Accused Habana AI Accelerator Product comprises a cluster configuration module that initializes at least one of the first cluster node module, the second cluster node module, or the third cluster node module. For example, each Gaudi HL-205 Mezzanine card in each HLS-1 system accesses and executes software for configuration of the computer cluster node modules residing in the memory of the eight Gaudi HL-205 Mezzanine cards in the HLS-1 system. This is done, for example, by using the Habana Collective Communication Library (HCCL). ^{48,49} HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation." ⁵⁰ The source code for Intel's HCCL Demo program ⁵¹ includes the details for using HCCL commands to initialize at least one of a first cluster node module, a second cluster node module, or a third cluster node module.
6. The computer cluster of claim 1, further comprising a data structure in which messages originating from	Each Accused Habana Server Product comprises a data structure in which messages originating from the second and third cluster node modules are stored. For example, the Habana Collective Communication Library (HCCL) includes commands that cause a Gaudi HL-2000 processor of a Gaudi HL-205 mezzanine card in a HLS-1 system to create a data structure in which messages

⁴⁶ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

⁴⁷ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

⁴⁸ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

⁴⁹ https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/Using_HCCL.html.

⁵⁰ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

⁵¹ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py, https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
the second and third cluster node modules are stored.	originating from the second and third cluster node modules are stored. ^{52,53} HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation." ⁵⁴ The source code for Intel's HCCL Demo program ⁵⁵ includes the details for using HCCL commands to initialize a HCCL communicator on a HLS-1 system comprising multiple Gaudi HL-205 Mezzanine card nodes and to create a data structure in which messages originating from the second and third cluster node modules are stored.
7. The computer cluster of claim 6, wherein the data structure of the first cluster node module comprises a received message queue.	Each of the plurality of Accused Habana AI Accelerator Product nodes in each Accused Habana Server Product comprises a first cluster node module comprising a received message queue in which messages originating from the second and third cluster node modules are stored. For example, the Habana Collective Communication Library (HCCL) includes commands that cause a Gaudi HL-2000 processor of a Gaudi HL-205 mezzanine card in a HLS-1 system to create a received message queue in which messages originating from the second and third cluster node modules are stored. ^{56,57} HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation." ⁵⁸ The source code for Intel's HCCL Demo program ⁵⁹ includes the details for using HCCL commands to initialize a HCCL communicator on a HLS-1 system comprising multiple Gaudi HL-205 Mezzanine card nodes and to create a received

⁵² "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

⁵³ https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/Using_HCCL.html.

⁵⁴ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

⁵⁵ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py.

⁵⁶ https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

⁵⁷ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

⁵⁸ https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/Using_HCCL.html.

⁵⁹ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

⁵⁹ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py.

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	message queue in which messages originating from the second and third cluster node modules are stored.
8. The computer cluster of claim 7, wherein the first cluster node module further comprises a message receiving queue, in which data specifying a location to which the first cluster node module can expect to receive an expression and an identifier of a processor from which the expression is expected to be sent are stored.^{60,61,62} HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation."⁶³ The source code for Intel's HCCL Demo program⁶⁴ includes the details for using HCCL commands to initialize a HCCL communicator on a HLS-1 system comprising multiple Gaudi HL-205 Mezzanine card nodes and to create a message receiving queue in which data specifying a location to which the first cluster node module can expect to receive an expression and an identifier of a processor from which the expression is expected to be sent are stored.	Each of the plurality of Accused Habana AI Accelerator Product nodes in each Accused Habana Server Product comprises a first cluster node module comprising a message receiving queue in which data specifying a location to which the first cluster node module can expect to receive an expression and an identifier of a processor from which the expression is expected to be sent are stored. For example, the Habana Collective Communication Library (HCCL) includes commands that cause a Gaudi HL-2000 processor of a Gaudi HL-205 mezzanine card in a HLS-1 system to create a message receiving queue in which data specifying a location to which the first cluster node module can expect to receive an expression and an identifier of a processor from which the expression is expected to be sent are stored.
9. The computer cluster of claim 1, wherein the second cluster node module further comprises a received message queue in	Each of the plurality of Accused Habana AI Accelerator Product nodes in each Accused Habana Server Product comprises a second cluster node module comprising a received message queue in which messages originating from the first and third cluster node modules are stored. For example, the Habana Collective Communication Library (HCCL) includes commands that cause a Gaudi HL-2000 processor of a Gaudi HL-205 mezzanine card in a HLS-1 system to create a received message queue

⁶⁰ https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/C_API.html

⁶¹ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

⁶² https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/Using_HCCL.html

⁶³ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

⁶⁴ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py, https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

U.S. Pat. No. 11,811,582 which messages originating from the first and third cluster node modules are stored.	Accused Habana Server Products and Accused Habana AI Accelerator Products in which messages originating from the first and third cluster node modules are stored.^{65,66} HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is 'Habana's complete software stack custom designed to support Habana's Gaudi implementation.'⁶⁷ The source code for Intel's HCCL Demo program⁶⁸ includes the details for using HCCL commands to initialize a HCCL communicator on a HLS-1 system comprising multiple Gaudi HL-205 Mezzanine card nodes and to create a received message queue in which messages originating from the first and third cluster node modules are stored.
10. The computer cluster of claim 1, wherein the first cluster node module comprises an advanced functions module, and wherein the advanced functions module comprises a call that evaluates an expression across the computer cluster in parallel.	Each Gaudi HL-2000 processor in a Gaudi HL-205 mezzanine card of the HLS-1 system comprises a first cluster node module that comprises an advanced functions module that comprises a call that evaluates an expression across the HLS-1 computer cluster in parallel. Each Accused Habana AI Accelerator Product in each Accused Habana Server Product comprises advanced functions module program code that, when executed, is capable of causing a hardware processor to evaluate mathematical expressions (<i>e.g.</i>, executable SynapseAI® Software Suite program code stored in memory accessible by the Gaudi HL-2000 processor in a Gaudi HL-205 mezzanine card). For example, when executed, the code stored in the memory accessible by a Gaudi HL-2000 processor causes the processor to evaluate mathematical expressions across the cluster in parallel (<i>e.g.</i>, by performing matrix multiplications using the General Matrix Multiplication ("GEMM") engine, evaluating mathematical expressions using the Tensor Processing Cores ("TPCs") in the Gaudi HL-2000 processor, and/or executing commands in the HCCL library). TPCs are configured to evaluate mathematical expressions, such as square root, sine, cosine, tangent, and reciprocal, as well as matrix operations such as Vector Reduction Intrinsics.⁶⁹

⁶⁵ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

⁶⁶ https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/Using_HCCL.html.

⁶⁷ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

⁶⁸ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py, https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

⁶⁹ See, *e.g.*, Built-in Functions ("https://docs.habana.ai/en/latest/TPC/TPC_User_Guide/Built_in_Functions.html").

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products A first Gaudi HL-205 accelerator node is responsible for interpreting user instructions and distributing calls to at least one of the other Gaudi HL-205 accelerators for execution. The other Gaudi HL-205 accelerator nodes in the cluster are responsible for executing the tasks that are distributed to them by the first node. When the user submits a job to the cluster, the first node parses the job and distributes the tasks to the other nodes for execution. The other nodes then execute the tasks and return the results to the first node or communicate the results of mathematical expression evaluation to other nodes. The first node then collects the results from the other nodes and returns them to the user. For example, the Habana Collective Communication Library (HCCL) includes commands that cause a Gaudi HL-2000 processor of a Gaudi HL-205 mezzanine card in a HLS-1 system to distribute calls to at least one other Gaudi HL-205 mezzanine card in the HLS-1 system for execution.⁷⁰ HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation."⁷¹ The Intel Gaudi Software Suite, which comprises Intel's SynapseAI® Software Suite and the Intel Gaudi graph compiler, "generates optimized binary code that implements the given model topology on Gaudi. It performs operator fusion, data layout management, parallelization, pipelining and memory management, as well as graph-level optimizations. The graph compiler uses the rich TPC kernel library⁷² which contains a wide variety of operations (for example, elementwise, non-linear, non-GEMM operators)."⁷³ The source code for Intel software such as Intel's HCCL Demo program⁷⁴ includes the details for using HCCL commands to initialize a HCCL communicator on a HLS-1 system comprising multiple Gaudi HL-205 Mezzanine card nodes, and implement an advanced functions module that comprises a call that evaluates an expression across the HLS-1 computer cluster in parallel.
--------------------------	--

⁷⁰ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

⁷¹ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

⁷² "Intel® Gaudi® processors TPC kernel library with supporting firmware, drivers and tools" (<https://developer.habana.ai/get-started/kernel-libraries/>).

⁷³ "Intel Gaudi Software Suite" (https://docs.habana.ai/en/latest/Gaudi_Overview/Intel_Gaudi_Software_Suite.html).

⁷⁴ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py, https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
12. The computer cluster of claim 10, wherein the advanced functions module comprises one or more calls that evaluate matrix operations across the computer cluster in parallel.	Each Gaudi HL-2000 processor in a Gaudi HL-205 mezzanine card of the HLS-1 system comprises a first cluster node module that comprises an advanced functions module that comprises one or more calls that evaluate matrix operations across the HLS-1 computer cluster in parallel. Each Accused Habana AI Accelerator Product in each Accused Habana Server Product comprises advanced functions module program code that, when executed, is capable of causing a hardware processor to evaluate mathematical expressions (<i>e.g.</i>, executable SynapseAI® Software Suite program code stored in memory accessible by the Gaudi HL-2000 processor in a Gaudi HL-205 mezzanine card). For example, when executed, the code stored in the memory accessible by a Gaudi HL-2000 processor causes the processor to evaluate mathematical expressions (<i>e.g.</i>, by performing matrix multiplications using the General Matrix Multiplication (“GEMM”) engine, evaluating mathematical expressions using the Tensor Processing Cores (“TPCs”) in the Gaudi HL-2000 processor, and/or executing commands in the HCCL library). TPCs are configured to evaluate mathematical expressions, such as square root, sine, cosine, tangent, and reciprocal, as well as matrix operations such as Vector Reduction Intrinsics.⁷⁵ A first Gaudi HL-205 accelerator node is responsible for interpreting user instructions and distributing calls to at least one of the other Gaudi HL-205 accelerators for execution. The other Gaudi HL-205 accelerator nodes in the cluster are responsible for executing the tasks that are distributed to them by the first node. When the user submits a job to the cluster, the first node parses the job and distributes the tasks to the other nodes for execution. The other nodes then execute the tasks and return the results to the first node or communicate the results of mathematical expression evaluation to other nodes. The first node then collects the results from the other nodes and returns them to the user. For example, the Habana Collective Communication Library (HCCL) includes commands that cause a Gaudi HL-2000 processor of a Gaudi HL-205 mezzanine card in a HLS-1 system to distribute calls to at least one other Gaudi HL-205 mezzanine card in the HLS-1 system for execution.⁷⁶ HCCL is

⁷⁵ See, *e.g.*, Built-in Functions (“https://docs.habana.ai/en/latest/TPC/TPC_User_Guide/Built_in_Functions.html”).

⁷⁶ “Habana Collective Communications Library (HCCL) is Habana’s emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library.” (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation."⁷⁷ The Intel Gaudi Software Suite, which comprises Intel's SynapseAI® Software Suite and the Intel Gaudi graph compiler, "generates optimized binary code that implements the given model topology on Gaudi. It performs operator fusion, data layout management, parallelization, pipelining and memory management, as well as graph-level optimizations. The graph compiler uses the rich TPC kernel library⁷⁸ which contains a wide variety of operations (for example, elementwise, non-linear, non-GEMM operators)."⁷⁹ The source code for Intel software such as Intel's HCCL Demo program⁸⁰ includes the details for using HCCL commands to initialize a HCCL communicator on a HLS-1 system comprising multiple Gaudi HL-205 Mezzanine card nodes, and implement an advanced functions module that comprises one or more calls that evaluate matrix operations across the HLS-1 computer cluster in parallel.
13. The computer cluster of claim 1, wherein the final result comprises at least a portion of the second result.	Each of the plurality of Accused Habana AI Accelerator Product nodes in each Accused Habana Server Product comprises a first cluster node module configured to return a final result based on, and including a portion of, the second result. For example, the Habana Collective Communication Library (HCCL) includes commands that cause a first cluster node module running on a Gaudi HL-2000 processor of a Gaudi HL-205 mezzanine card in a HLS-1 system to return a final result based on, and including a portion of, the second result.^{81,82} HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi

⁷⁷ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

⁷⁸ "Intel® Gaudi® processors TPC kernel library with supporting firmware, drivers and tools" (<https://developer.habana.ai/get-started/kernel-libraries/>).

⁷⁹ "Intel Gaudi Software Suite" (https://docs.habana.ai/en/latest/Gaudi_Overview/Intel_Gaudi_Software_Suite.html).

⁸⁰ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py, https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

⁸¹ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

⁸² https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/Using_HCCL.html.

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	implementation.” ⁸³ The source code for Intel’s HCCL Demo program ⁸⁴ includes the details for using HCCL commands to cause a first cluster node module to return a final result based on, and including a portion of, the second result.
14. The computer cluster of claim 1, wherein the first cluster node module is configured to return the final result to the user interface.	Each of the plurality of Accused Habana AI Accelerator Product nodes in each Accused Habana Server Product comprises a first cluster node module configured to return a final result to the user interface. For example, the Habana Collective Communication Library (HCCL) includes commands that cause a first cluster node module running on a Gaudi HL-2000 processor of a Gaudi HL-205 mezzanine card in a HLS-1 system to return a final result to the user interface. ^{85,86} HCCL is included in Intel’s SynapseAI® Software Suite. SynapseAI® is “Habana’s complete software stack custom designed to support Habana’s Gaudi implementation.” ⁸⁷ The source code for Intel’s HCCL Demo program ⁸⁸ includes the details for using HCCL commands to cause a first cluster node module to return a final result to the user interface.
15. The computer cluster of claim 1, further comprising a communications network interface for the cluster node modules to communicate, with each	One or more of the Accused Habana AI Accelerator Products in an Accused Habana Server Product comprise a communications network interface for cluster node modules to communicate results of the execution of matrix operations with each other using a peer-to-peer architecture. For example, the Intel HLS-1 system comprises eight Gaudi HL-205 Mezzanine cards (designated as nodes “0”, “1” ... “7” in the figure below) which communicate results of the execution of matrix operations with

⁸³ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

⁸⁴ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py, https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

⁸⁵ “Habana Collective Communications Library (HCCL) is Habana’s emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library.” (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

⁸⁶ https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/Using_HCCL.html.

⁸⁷ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

⁸⁸ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py, https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products each other via network connections that connect the nodes in an all-to-all local bus routing by eight 7x 100G Ethernet connections with a peer-to-peer architecture .⁸⁹ other using a peer-to-peer architecture, results of execution of the matrix operations. HLS-1 Block Diagram At least because of the architecture of the Gaudi HL-2000 processor along with related resources included in Intel’s SynapseAI® Software Suite, the cluster of eight Gaudi HL-205 Mezzanine cards in an HLS-1 system communicate results of the execution of matrix operations with each other using a communications network interface having a peer-to-peer architecture. For example, the Habana Collective Communication Library (HCCL) includes commands that cause a Gaudi HL-2000 processor of a second Gaudi HL-205 mezzanine card in a HLS-1 system to execute the matrix operations on the array of data elements and return a first result using the Gaudi HL-2000 processor of the second Gaudi HL-205 mezzanine card (e.g., by performing matrix multiplications using the General Matrix Multiplication (“GEMM”) engine, evaluating mathematical expressions
--	---

⁸⁹ Gaudi® HLS-1 AI Training System (<https://www.intel.com/content/www/us/en/content-details/784787/audi-hls-1-ai-training-system.html>); see also HLS-1 introductory video (“Habana Labs: How to Scale AI Training with Gaudi Processors”), at <https://www.youtube.com/watch?v=u0siCfmCNfg>; see also <https://habana.ai/products/audi/>.

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	using the Tensor Processing Cores in the Gaudi HL-2000 processor, and/or executing commands in the HCCL library).⁹⁰ TPCs are configured to evaluate multiple mathematical expressions, such as square root, sine, cosine, tangent, and reciprocal, as well as matrix operations such as Vector Reduction Intrinsic.⁹¹ HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation."⁹² The source code for Intel's HCCL Demo program⁹³ includes the details for using HCCL commands to cause a cluster node modules of Accused Habana AI Accelerator Products in an Accused Habana Server Product to communicate results of the execution of matrix operations with each other via a communications network interface having a peer-to-peer architecture.^{2,94}
16. The computer cluster of claim 15, wherein one or more of the cluster node modules are configured to: accept user instructions; after accepting the user instructions, communicate at least some of the user instructions using the communications network	One or more of the Accused Habana AI Accelerator Products in an Accused Habana Server Product comprise cluster node modules that are configured to accept user instructions and communicate at least some of the user instructions to each other using the communications network interface for the cluster node modules to communicate with each other, and after communicating at least some of the user instructions using the communications network interface, communicate at least some of the user instructions to one or more of the kernels. For example, the Intel HLS-1 system comprises eight Gaudi HL-205 Mezzanine cards (designated as nodes "0", "1" ... "7" in the figure below) in which the nodes are configured to communicate commands including user instructions with one another via network connections that connect the nodes in an all-to-all local bus routing by eight 7x 100G Ethernet connections.⁹⁵

⁹⁰ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

⁹¹ See, e.g., Built-in Functions ("https://docs.habana.ai/en/latest/TPC/TPC_User_Guide/Built_in_Functions.html").

⁹² Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

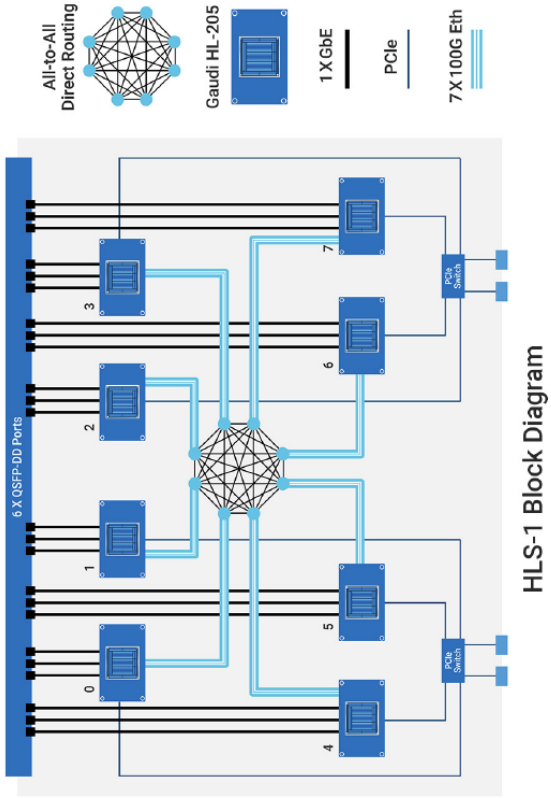
⁹³ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/nun_hccl_demo.py; https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

⁹⁴ https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/Using_HCCL.html.

⁹⁵ Gaudi® HLS-1 AI Training System (<https://www.intel.com/content/www/us/en/content-details/784787/gaudi-hls-1-ai-training-system.html>); see also HLS-1 introductory video ("Habana Labs: How to Scale AI Training with Gaudi Processors"), at <https://www.youtube.com/watch?v=u0siCfmCNfg>; see also <https://habana.ai/products/gaudi/>.

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
interface for the cluster node modules to communicate with each other; and after communicating at least some of the user instructions using the communications network interface, communicate at least some of the user instructions to one or more of the kernels.	The diagram illustrates the HLS-1 Block Architecture. It features eight Gaudi HL-205 Mezzanine cards, numbered 0 through 7, arranged in two rows. Each card is connected to a central network interface (represented by a mesh of nodes). The cards are also connected to a 6 X QSFP-DD Ports interface. A legend indicates the following components: All-to-All Direct Routing (represented by a mesh of nodes), Gaudi HL-205 (represented by a blue rectangle), 1 X GbE (represented by a single line), PCIe (represented by a double line), and 7 X 100G Eth (represented by a thick blue line). HLS-1 Block Diagram At least because of the architecture of the Gaudi HL-2000 processor along with related resources included in Intel’s SynapseAI® Software Suite, each of the cluster of eight Gaudi HL-205 Mezzanine cards in an HLS-1 system corresponds to a node that can access and execute kernel software residing in memory, and that is configured to enable cluster node modules to accept user instructions and communicate at least some of the user instructions to each other using the communications network interface shown above, and further communicate at least some of the user instructions to one or more of the kernels. The source code for Intel’s HCCL Demo program⁹⁶ includes the details for using HCCL commands to cause a cluster node modules of Accused Habana AI Accelerator Products in an Accused Habana Server Product to accept user instructions and communicate at least some of the user

⁹⁶ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py; https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
	instructions to each other using the communications network interface shown above, and further communicate at least some of the user instructions to one or more of the kernels.^{97,98}
17. The computer cluster of claim 15, wherein the communications network interface comprises at least one of: a local bus, a local area network (LAN), a wide area network (WAN), a storage area network (SAN), a TCP/IP interface over Ethernet, or a wireless network interface.	A communications network interface of the Accused Habana AI Accelerator Products comprises at least one of: a local bus, a local area network (LAN), a wide area network (WAN), a storage area network (SAN), a TCP/IP interface over Ethernet, or a wireless network interface (e.g., the all-to-all direct routing connectivity between the eight Gaudi HL-205 Mezzanine cards in the HLS-1 system shown in the following block diagram).⁹⁹  HLS-1 Block Diagram

⁹⁷ “Habana Collective Communications Library (HCCL) is Habana’s emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library.” (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

⁹⁸ https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/Using_HCCL.html.

⁹⁹ Gaudi® HLS-1 AI Training System (<https://www.intel.com/content/www/us/en/content-details/784787/audi-hls-1-ai-training-system.html>); see also HLS-1 introductory video (“Habana Labs: How to Scale AI Training with Gaudi Processors”), at <https://www.youtube.com/watch?v=u0siCfmCNfg>; see also <https://habana.ai/products/gaudi/>.

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
18. The computer cluster of claim 1, wherein at least one of the first, second, or third kernels comprises a single-node kernel.	Each of the plurality of Accused Habana AI Accelerator Product nodes in each Accused Habana Server Product is configured to access program code for a single-node kernel, and so at least one of first, second, or third kernels comprises a single-node kernel. For example, the Habana Collective Communication Library (HCCL) includes commands that cause a Gaudi HL-2000 processor of a Gaudi HL-205 mezzanine card in an HLS-1 system to access program code for a single-node kernel. ¹⁰⁰ The source code for Intel's HCCL Demo program ¹⁰¹ includes the details for using HCCL commands for that purpose. HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation." ¹⁰²
19. The computer cluster of claim 1, wherein the cluster node modules communicate with one another to act as a cluster.	Each of the plurality of Accused Habana AI Accelerator Product nodes in each Accused Habana Server Product comprises a plurality of cluster node modules that communicate with one another to act as a cluster. For example, the Habana Collective Communication Library (HCCL) includes commands that cause a Gaudi HL-2000 processor of any of a plurality of Gaudi HL-205 mezzanine cards in a HLS-1 system to communicate with one another to act as a cluster. ¹⁰³ HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation." ¹⁰⁴ The source code for Intel's HCCL Demo program ¹⁰⁵ includes the details for using HCCL commands to initialize a HCCL communicator on a HLS-1 system comprising multiple Gaudi HL-205 Mezzanine card nodes to communicate with one another to act as a cluster.

¹⁰⁰ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

¹⁰¹ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py.

¹⁰² Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

¹⁰³ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

¹⁰⁴ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

¹⁰⁵ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py.

U.S. Pat. No. 11,811,582	Accused Habana Server Products and Accused Habana AI Accelerator Products
20. The computer cluster of claim 1, wherein one or more of the cluster node modules are configured to receive a command to start a cluster initialization process for the computer cluster.	Each Accused Habana AI Accelerator Product comprises one or more cluster node modules that are configured to receive a command to start a cluster initialization process for the computer cluster. For example, each Gaudi HL-205 Mezzanine card in each HLS-1 system can access and execute software for configuration of the computer cluster node modules residing in the memory of the eight Gaudi HL-205 Mezzanine cards in the HLS-1 system. This is done, for example, by using the Habana Collective Communication Library (HCCL).^{106,107} HCCL is included in Intel's SynapseAI® Software Suite. SynapseAI® is "Habana's complete software stack custom designed to support Habana's Gaudi implementation."¹⁰⁸ The source code for Intel's HCCL Demo program¹⁰⁹ includes the details for using HCCL commands received by one or more of the cluster node modules to start a cluster initialization process for the computer cluster.

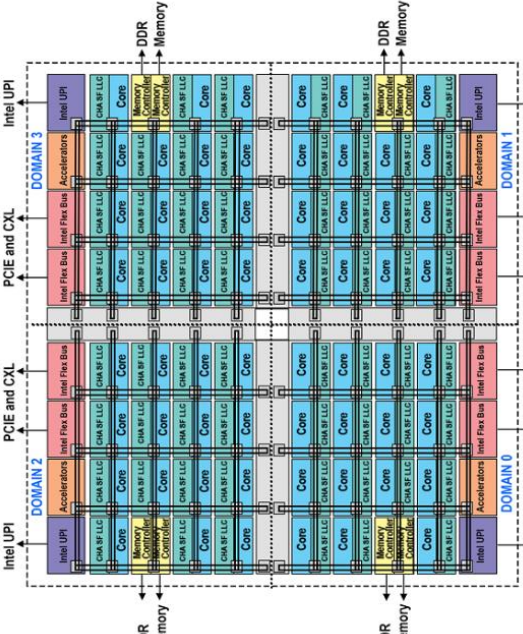
¹⁰⁶ "Habana Collective Communications Library (HCCL) is Habana's emulation layer of the NVIDIA Collective Communication Library (NCCL) and is included in the SynapseAI® Software library." (https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/index.html).

¹⁰⁷ https://docs.habana.ai/en/latest/API_Reference_Guides/HCCL_APIs/Using_HCCL.html.

¹⁰⁸ Gaudi® Training Platform White Paper, Nov 2020, p. 7 (<https://habana.ai/wp-content/uploads/pdf/2020/Habana%20GAUDI%20Training%20Whitepaper%20v1.2.pdf>).

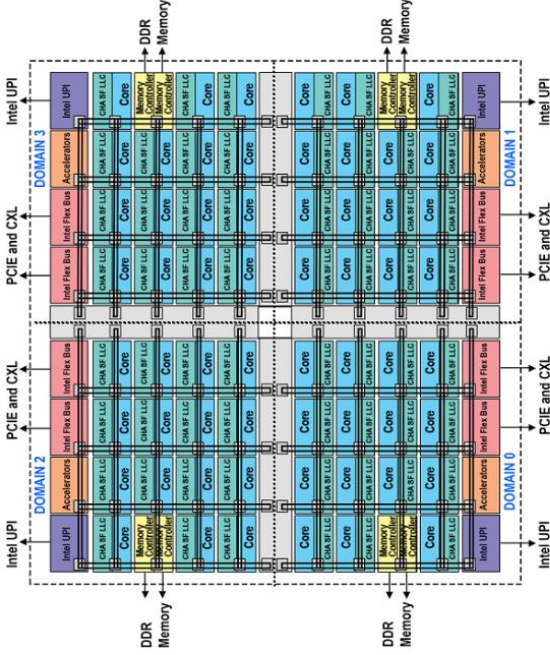
¹⁰⁹ See https://github.com/HabanaAI/hccl_demo/blob/main/hccl_demo.cpp; see also https://github.com/HabanaAI/hccl_demo/blob/main/run_hccl_demo.py, https://github.com/HabanaAI/hccl_demo/blob/main/README.md.

Exhibit B4
U.S. Pat. No. 11,811,582 in view of
Accused Xeon Products

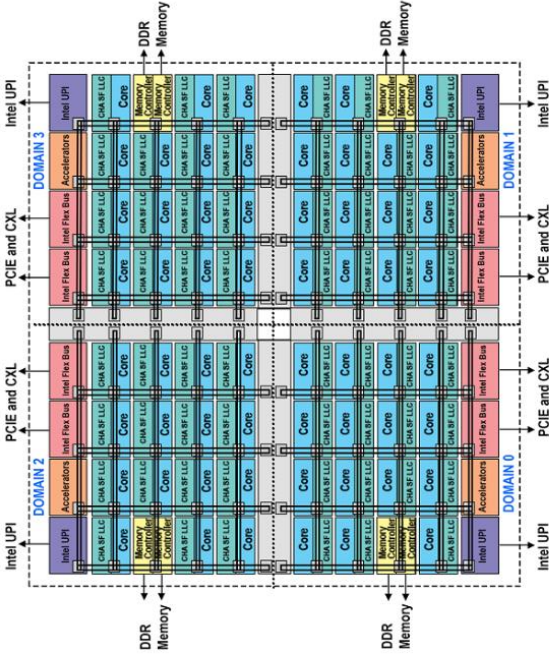
U.S. Pat. No. 11,811,582	Accused Xeon Products Each of the Accused Xeon Products¹ comprises a computer cluster. For example, as depicted below and as described further herein, Intel's 4th Gen Xeon Scalable Processor² includes a computer cluster.  Figure 3 – Block Diagram Representing Domains Of sub-NUMA With Four Clusters
--------------------------	--

¹ The Accused Xeon Products include, but are not limited to, all products including or related to Intel's Xeon Scalable Processors (Skylake-SP architecture) (<https://www.intel.com/content/www/us/en/developer/articles/technical/xeon-processor-scalable-family-technical-overview.html>), 2nd Gen Xeon Scalable Processors (<https://www.intel.com/content/www/us/en/developer/articles/technical/xeon-processor-scalable-family-technical-overview.html>), 3rd Gen Xeon Scalable Processors (<https://www.intel.com/content/www/us/en/developer/articles/technical/xeon-processor-scalable-family-technical-overview.html>), 4th Gen Xeon Scalable Processors (<https://www.intel.com/content/www/us/en/developer/articles/technical/xeon-processor-scalable-family-technical-overview.html>), 5th Gen Xeon Scalable Processors (<https://www.intel.com/content/www/us/en/developer/articles/technical/xeon-processor-scalable-family-technical-overview.html>), and Xeon 6 Processors (<https://www.intel.com/content/www/us/en/products/details/processors/xeon.html>), as well as any products incorporating those items.

² "Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family," <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-technical-overview.html>.

U.S. Pat. No. 11,811,582	a first processor; a second processor; a third processor;
Accused Xeon Products	Each of the Accused Xeon Products comprises a plurality of nodes. For example, as depicted below, the 4th Gen Xeon Scalable Processor comprises four nodes (described as domains in the figure below), which are configured to be connected in an all-to-all local bus interconnection configuration.³ Each of the 4th Gen Xeon Scalable Processors is configured to operate in a Sub-NUMA clustering 4 (SNC-4) affinity mode, comprising four domains of sub-NUMA clusters / nodes.  Figure 3 – Block Diagram Representing Domains Of sub-NUMA With Four Clusters Each Accused Xeon Product comprises a plurality of nodes, wherein each of the plurality of nodes comprises a hardware processor. For example, each 4th Gen Xeon Scalable Processor comprises a hardware processor that comprises multiple processor cores. The 4th Gen Xeon Scalable Processor is configured to operate in an SNC-4 mode where each of the four domains is a node that, for example, includes a plurality of processing cores (fifteen) as shown in the above figure.

³ “Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family,” <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

U.S. Pat. No. 11,811,582	at least one computer-readable medium in communication with at least one of the first processor, the second processor, or the third processor; a first kernel residing in the at least one computer-readable medium, said first kernel configured to translate user commands into calls for execution of matrix operations on an array of data elements;	Accused Xeon Products Thus, each of the Accused Xeon Products comprises a first processor, a second processor, and a third processor. Each of the Accused Xeon Products comprises at least one computer-readable medium in communication with at least one of the first processor, the second processor, or the third processor, and a first kernel residing in the at least one computer-readable medium, said first kernel configured to translate user commands into calls for execution of matrix operations on an array of data elements. Each of the Accused Xeon Products comprises a plurality of nodes. For example, as depicted below, the 4th Gen Xeon Scalable Processor⁴ comprises four nodes (described as domains in the figure below), which are configured to be connected in an all-to-all local bus interconnection configuration.  Figure 3 – Block Diagram Representing Domains Of sub-NUMA With Four Clusters
--------------------------	---	---

⁴ “Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family,” <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products Each of the 4th Gen Xeon Scalable Processors is configured to operate in a Sub-NUMA clustering 4 (SNC-4) affinity mode, comprising four domains of sub-NUMA clusters / nodes.⁵ Each Accused Xeon Product comprises a plurality of nodes, wherein each of the plurality of nodes comprises at least one computer-readable medium in communication with at least one of the first processor, the second processor, or the third processor. For example, the 4th Gen Xeon Scalable Processor is configured to operate in an SNC-4 mode where each of the four domains is a node that includes a plurality of processing cores (fifteen) as shown in the above figure. Each Accused Xeon Product is configured to access and be in communication with a non-transitory computer-readable medium (e.g., memory accessible by and in communication with the 4th Gen Xeon Scalable Processor, including local memory, shared memory, and HBM memory). The 4th Gen Xeon Scalable Processor is configured to access the computer-readable medium that comprises program code that is configured to translate user commands into calls for execution of matrix operations on an array of data elements. This is done, for example, by using Intel's oneCCL software, which is part of Intel's oneAPI Collective Communications Library.⁶ More particularly, Intel's Accused Xeon Products are configured to be supported by Intel's oneAPI software suite, which, for example, "maximize[s] application performance by activating the advanced capabilities of 4th and 5th gen Intel® Xeon® processors..."⁷ The oneCCL library is one of the 10 core elements of Intel's oneAPI software suite.⁸ Intel's oneCCL package comprises the oneCCL Software Development Kit (SDK) and the Intel® MPI Library Runtime components.⁹
--------------------------	--

⁵ Id.

⁶ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf>.

⁷ "Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors," <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>; "oneAPI Specification, Release 1.3-rev-1," pp. 388-389 <https://spec.oneapi.io/versions/latest/oneAPI-spec.pdf>; "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/overview.html>, <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/communicator.html>.

⁸ <https://www.intel.com/content/www/us/en/developer/articles/technical/oneapi-what-is-it.html>.

⁹ See "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/overview.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products
	For example, code stored in the memory accessible by and in communication with a 4th Gen Xeon Scalable Processor is configured to translate user commands into calls for execution of matrix operations. The processors are configured to evaluate multiple mathematical expressions and matrix operations.^{10 11} For example, Intel's 4th Gen Xeon Scalable Processors are configured such that each domain / node in a Sub-NUMA clustering 4 (SNC-4) affinity mode, includes at least two types of compute engines for evaluating mathematical expressions including matrix operations using multiple processors and cores: • Intel Advanced Matrix Extensions Accelerator (AMX): AMX is a specialized hardware accelerator for evaluating matrix multiplication expressions.¹² • Processing Cores (PCs): The PCs are general-purpose hardware accelerators that can be used to evaluate a variety of mathematical expressions.¹³
a first cluster node module residing in the at least one computer-readable medium, said first cluster node module configured to communicate with the first kernel and receive commands from a user interface;	Each of the Accused Xeon Products comprises a first cluster node module residing in the at least one computer-readable medium, said first cluster node module configured to communicate with the first kernel and receive commands from a user interface. For example, as depicted below, the 4th Gen Xeon Scalable Processor, configured in a Sub-NUMA clustering 4 (SNC-4) affinity mode, includes four domains where each domain includes fifteen processing cores, each of which is configured to access memory including, for example, LLC, DDR, and HBM memory.¹⁴

¹⁰ “Technical Overview Of The 4th Gen Intel Xeon Scalable processor family,” <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

¹¹ Processor cores are configured to evaluate multiple mathematical expressions, such as addition, subtraction, multiplication, division, and square root. See, e.g., “Intel® AVX-512 - FP16 Instruction Set for Intel® Xeon® Processor Based Products Technology Guide,” <https://www.intel.com/content/www/us/en/content-details/669773/intel-avx-512-fp16-instruction-set-for-intel-xeon-processor-based-products-technology-guide.html>.

¹² <https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/advanced-matrix-extensions/overview.html>.

¹³ <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

¹⁴ “Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family,” <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

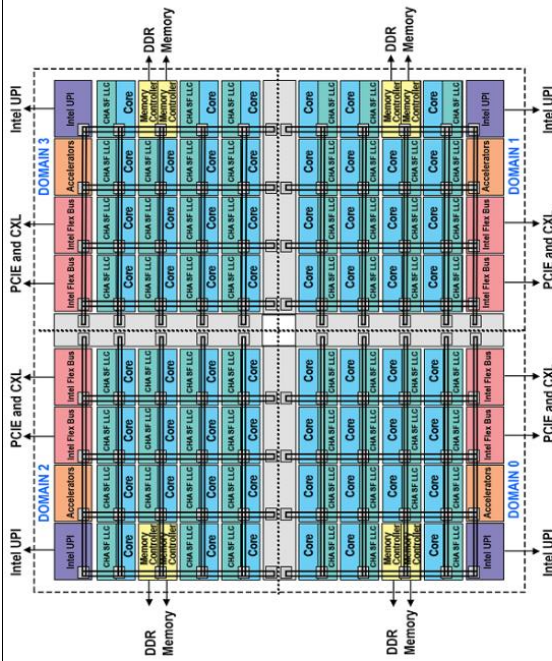


Figure 3 - Block Diagram Representing Domains Of sub-NUMA With Four Clusters

Each Accused Xeon Product is configured to access and be in communication with a non-transitory computer-readable medium (e.g., memory accessible by and in communication with the 4th Gen Xeon Scalable Processor, including local memory, shared memory, and HBM memory).

Intel's Accused Xeon Products are configured to be supported by Intel's oneAPI software suite, which, for example, "maximize[s] application performance by activating the advanced capabilities of 4th and 5th gen Intel® Xeon® processors...."¹⁵ The oneCCL library is one of the 10 core elements of

¹⁵ "Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors," <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>; "oneAPI Specification, Release 1.3-rev-1," pp. 388-389 <https://spec.oneapi.io/versions/latest/oneAPI-spec.pdf>; "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-12/overview.html>, <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-12/communicator.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products Intel's oneAPI software suite.¹⁶ Intel's oneCCL package comprises the oneCCL Software Development Kit (SDK) and the Intel® MPI Library Runtime components.¹⁷ “The Intel® oneAPI Math Kernel Library (oneMKL) is a computing math library of highly optimized and extensively parallelized routines for applications that require maximum performance. oneMKL contains the high-performance optimizations from the full Intel® Math Kernel Library for CPU architectures (with C/Fortran programming language interfaces) and adds to them a set of SYCL* interfaces for achieving performance on various CPU architectures and Intel Graphics Technology for certain key functionalities. oneMKL provides BLAS and LAPACK linear algebra routines, fast Fourier transforms, vectorized math functions, random number generation functions, and other functionality.”¹⁸ The mathematical routines addressed in oneMKL include matrix operations using, for example, arrays of data elements. The figure below depicts Intel's oneAPI software suite and its Collective Communications Library (oneCCL) and Math Kernel Library (oneMKL) in the context of the software suite for Intel's 4th Gen Xeon Scalable Processors:¹⁹
--------------------------	--

¹⁶ <https://www.intel.com/content/www/us/en/developer/articles/technical/oneapi-what-is-it.html>.

¹⁷ See “Intel® oneAPI Collective Communications Library Developer Guide and Reference,” <https://www.intel.com/content/www/us/en/docs/onecc/developer-guide-reference/2021-12/overview.html>.

¹⁸ “Intel® oneAPI Programming Guide,” <https://www.intel.com/content/www/us/en/docs/oneapi/programming-guide/2024-1/intel-oneapi-math-kernel-library-onemkl.html>.

¹⁹ “Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors,” <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqdl>.

U.S. Pat. No. 11,811,582 a second kernel residing in the at least one computer-readable medium, said second kernel configured to execute the matrix operations on the array of data elements and return a first result using the second processor;	Accused Xeon Products Each of the Accused Xeon Products comprises a second kernel residing in the at least one computer-readable medium, said second kernel configured to execute the matrix operations on the array of data elements and return a first result using the second processor. For example, as depicted below, the 4th Gen Xeon Scalable Processor, configured in a Sub-NUMA clustering 4 (SNC-4) affinity mode, includes four domains where each domain includes fifteen processing cores, each of which is configured to access memory including, for example, LLC, DDR, and HBM memory.²² Figure 3 - Block Diagram Representing Domains Of sub-NUMA With Four Clusters Each Accused Xeon Product is configured to access and be in communication with a non-transitory computer-readable medium (e.g., memory accessible by and in communication with the 4th Gen Xeon Scalable Processor, including local memory, shared memory, and HBM memory).
---	--

²² “Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family,” <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products
	Intel's Accused Xeon Products are configured to be supported by Intel's oneAPI software suite, which, for example, "maximize[s] application performance by activating the advanced capabilities of 4th and 5th gen Intel® Xeon® processors...."²³ The oneCCL library is one of the 10 core elements of Intel's oneAPI software suite.²⁴ Intel's oneCCL package comprises the oneCCL Software Development Kit (SDK) and the Intel® MPI Library Runtime components.²⁵ "The Intel® oneAPI Math Kernel Library (oneMKL) is a computing math library of highly optimized and extensively parallelized routines for applications that require maximum performance. oneMKL contains the high-performance optimizations from the full Intel® Math Kernel Library for CPU architectures (with C/Fortran programming language interfaces) and adds to them a set of SYCL* interfaces for achieving performance on various CPU architectures and Intel Graphics Technology for certain key functionalities. oneMKL provides BLAS and LAPACK linear algebra routines, fast Fourier transforms, vectorized math functions, random number generation functions, and other functionality."²⁶ The mathematical routines addressed in oneMKL include matrix operations using, for example, arrays of data elements. The figure below depicts Intel's oneAPI software suite and its Collective Communications Library (oneCCL) and Math Kernel Library (oneMKL) in the context of the software suite for Intel's 4th Gen Xeon Scalable Processors:²⁷

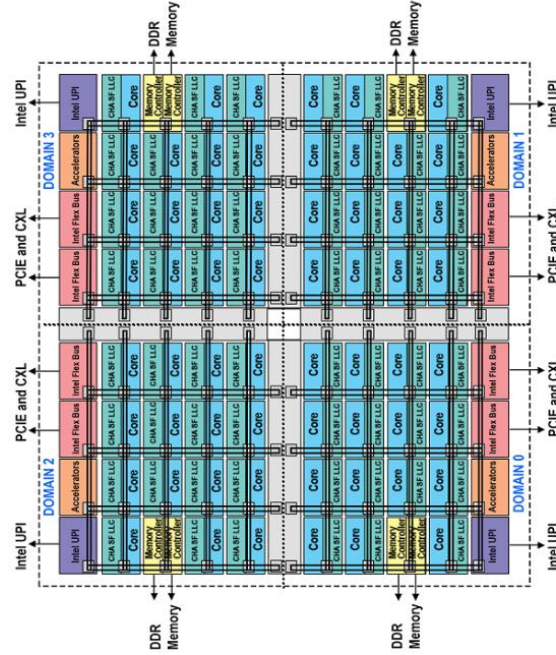
²³ "Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors," <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>; "oneAPI Specification, Release 1.3-rev-1," pp. 388-389 <https://spec.oneapi.io/versions/latest/oneAPI-spec.pdf>; "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-12/overview.html>, <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-12/overview.html>.

²⁴ <https://www.intel.com/content/www/us/en/developer/articles/technical/oneapi-what-is-it.html>.

²⁵ See "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-12/overview.html>.

²⁶ "Intel® oneAPI Programming Guide," <https://www.intel.com/content/www/us/en/docs/oneapi/programming-guide/2024-1/intel-oneapi-math-kernel-library-onemkl.html>.

²⁷ "Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors," <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>.

U.S. Pat. No. 11,811,582 a second cluster node module residing in the at least one computer-readable medium, said second cluster node module configured to send commands to the second kernel and communicate with the first cluster node module;	Accused Xeon Products Each of the Accused Xeon Products comprises a second cluster node module residing in the at least one computer-readable medium, said second cluster node module configured to send commands to the second kernel and communicate with the first cluster node module. For example, as depicted below, the 4th Gen Xeon Scalable Processor, configured in a Sub-NUMA clustering 4 (SNC-4) affinity mode, includes four domains where each domain includes fifteen processing cores, each of which is configured to access memory including, for example, LLC and DDR memory.³⁰  Figure 3 – Block Diagram Representing Domains of sub-NUMA With Four Clusters Each Accused Xeon Product is configured to access and be in communication with a non-transitory computer-readable medium. For example, the 4th Gen Xeon Scalable Processor, as shown above, is
--	---

³⁰ “Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family,” <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products configured to access and be in communication with a non-transitory computer-readable medium that includes, for example, local memory (e.g., LLC), and external memory (e.g., DDR memory). Intel's Accused Xeon Products are configured to be supported by Intel's oneAPI software suite, including Intel's oneCCL software.³¹ For example, Intel's 4th Gen Xeon Scalable Processor are configured to be supported by Intel's oneAPI software suite, which, for example, "maximize[s] application performance by activating the advanced capabilities of 4th and 5th gen Intel® Xeon® processors...."³² The oneCCL library is one of the 10 core elements of Intel's oneAPI software suite.³³ Intel's oneCCL package comprises the oneCCL Software Development Kit (SDK) and the Intel® MPI Library Runtime components.³⁴ "The Intel® oneAPI Math Kernel Library (oneMKL) is a computing math library of highly optimized and extensively parallelized routines for applications that require maximum performance. oneMKL contains the high-performance optimizations from the full Intel® Math Kernel Library for CPU architectures (with C/Fortran programming language interfaces) and adds to them a set of SYCL* interfaces for achieving performance on various CPU architectures and Intel Graphics Technology for certain key functionalities. oneMKL provides BLAS and LAPACK linear algebra routines, fast Fourier transforms, vectorized math functions, random number generation functions, and other functionality."³⁵ The mathematical routines addressed in oneMKL include matrix operations using, for example, arrays of data elements.
--------------------------	---

³¹ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> ("Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.").

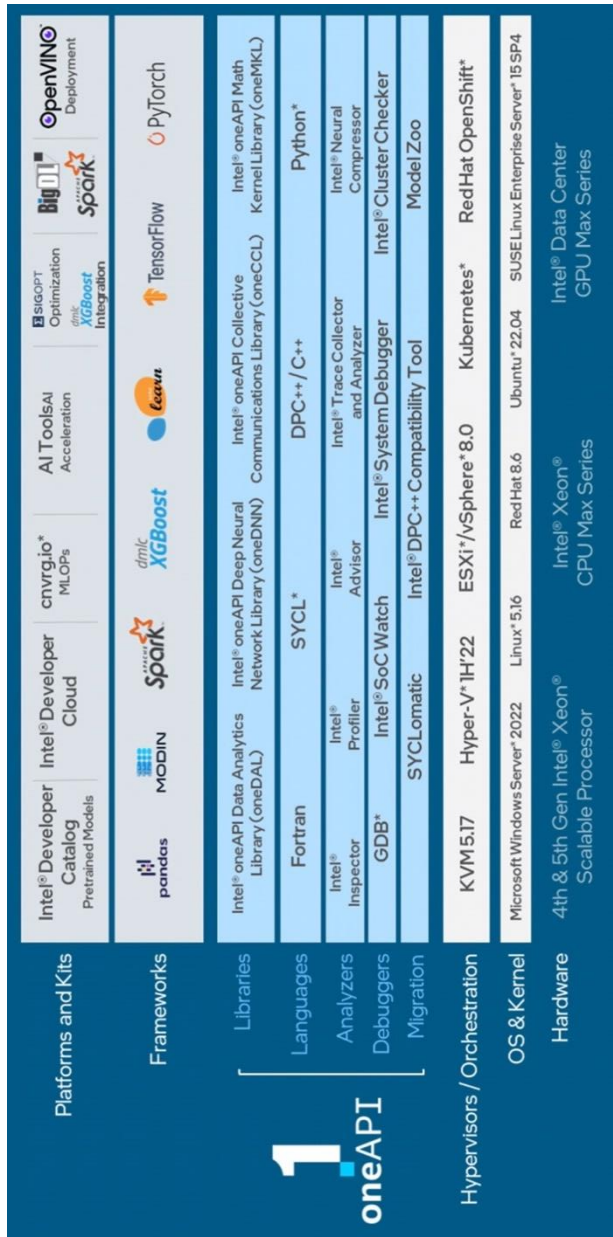
³² "Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors," <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>; "oneAPI Specification, Release 1.3-rev-1," pp. 388-389 <https://spec.oneapi.io/versions/latest/oneAPI-spec.pdf>; "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-12/overview.html>, <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/communicator.html>.

³³ <https://www.intel.com/content/www/us/en/developer/articles/technical/oneapi-what-is-it.html>.

³⁴ See "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/overview.html>.

³⁵ "Intel® oneAPI Programming Guide," https://www.intel.com/content/www/us/en/docs/oneapi/programming_guide/2024-1/intel-oneapi-math-kernel-library-onemkl.html.

The figure below depicts Intel's oneAPI software suite and its Collective Communications Library (oneCCL) and Math Kernel Library (oneMKL) in the context of the software suite for Intel's 4th Gen Xeon Scalable Processors:³⁶



The CCL library provides multiple commands for creating, managing, and using CCL communicators.³⁷ The CCL library includes software objects that are used to manage the communication between the domains / nodes accelerators in the 4th Gen Xeon Scalable Processor.³⁸

³⁶ “Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors,” <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>.

³⁷ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”).

³⁸ <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-9/device-communication.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products
	The source code for the CCL library includes the details for using oneCCL commands to initialize communications on a 4th Gen Xeon Scalable Processor, send commands to the second kernel and communicate with the first cluster node module.³⁹
a third kernel residing in the at least one computer-readable medium, said third kernel configured to receive the first result and execute additional matrix operations on the first result and return a second result to one or more of the first, second or third cluster node modules using the third processor. For example, as depicted below, the 4th Gen Xeon Scalable Processor, configured in a Sub-NUMA clustering 4 (SNC-4) affinity mode, includes four domains where each domain includes fifteen processing cores, each of which is configured to access memory including, for example, LLC, DDR, and HBM memory.⁴⁰	Each of the Accused Xeon Products comprises a third kernel residing in the at least one computer-readable medium, said third kernel configured to receive the first result and execute additional matrix operations on the first result and return a second result to one or more of the first, second or third cluster node modules using the third processor.

³⁹ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

⁴⁰ “Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family,” <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

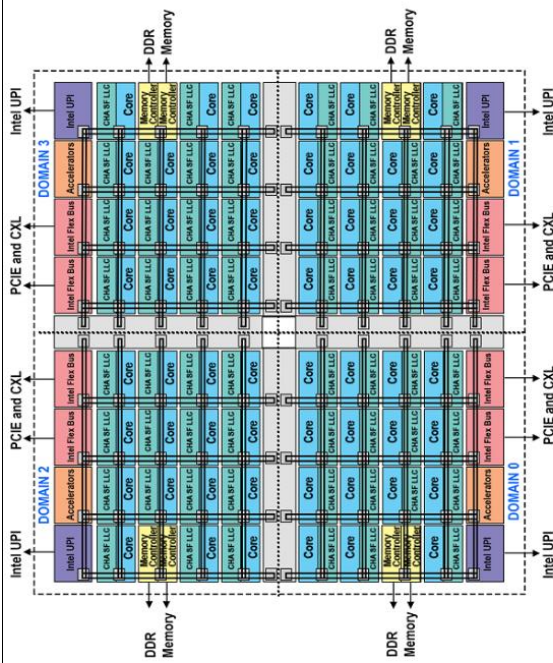


Figure 3 – Block Diagram Representing Domains Of sub-NUMA With Four Clusters

Each Accused Xeon Product is configured to access and be in communication with a non-transitory computer-readable medium (e.g., memory accessible by and in communication with the 4th Gen Xeon Scalable Processor, including local memory, shared memory, and HBM memory).

Intel's Accused Xeon Products are configured to be supported by Intel's oneAPI software suite, which, for example, "maximize[s] application performance by activating the advanced capabilities of 4th and 5th gen Intel® Xeon® processors...."⁴¹ The oneCCL library is one of the 10 core elements of

⁴¹ "Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors," <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>; "oneAPI Specification, Release 1.3-rev-1," pp. 388-389 <https://spec.oneapi.io/versions/latest/oneAPI-spec.pdf>; "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-12/overview.html>, <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-12/communicator.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products Intel's oneAPI software suite.⁴² Intel's oneCCL package comprises the oneCCL Software Development Kit (SDK) and the Intel® MPI Library Runtime components.⁴³ “The Intel® oneAPI Math Kernel Library (oneMKL) is a computing math library of highly optimized and extensively parallelized routines for applications that require maximum performance. oneMKL contains the high-performance optimizations from the full Intel® Math Kernel Library for CPU architectures (with C/Fortran programming language interfaces) and adds to them a set of SYCL* interfaces for achieving performance on various CPU architectures and Intel Graphics Technology for certain key functionalities. oneMKL provides BLAS and LAPACK linear algebra routines, fast Fourier transforms, vectorized math functions, random number generation functions, and other functionality.”⁴⁴ The mathematical routines addressed in oneMKL include matrix operations using, for example, arrays of data elements. The figure below depicts Intel's oneAPI software suite and its Collective Communications Library (oneCCL) and Math Kernel Library (oneMKL) in the context of the software suite for Intel's 4th Gen Xeon Scalable Processors:⁴⁵
--------------------------	--

⁴² <https://www.intel.com/content/www/us/en/developer/articles/technical/oneapi-what-is-it.html>.

⁴³ See “Intel® oneAPI Collective Communications Library Developer Guide and Reference,” <https://www.intel.com/content/www/us/en/docs/onecc/developer-guide-reference/2021-12/overview.html>.

⁴⁴ “Intel® oneAPI Programming Guide,” <https://www.intel.com/content/www/us/en/docs/oneapi/programming-guide/2024-1/intel-oneapi-math-kernel-library-onemkl.html>.

⁴⁵ “Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors,” <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqdl>.

U.S. Pat. No. 11,811,582	Accused Xeon Products The diagram illustrates the Intel oneAPI ecosystem. At the top, 'Platforms and Kits' includes Intel® Developer Catalog Pretrained Models, Intel® Developer Cloud, cnvrg.io* MLOps, AI Toolset Acceleration, SIGOPT Optimization, BigDL, Spark, and OpenVINO Deployment. Below this, 'Frameworks' lists pandas, MODIN, Spark, XGBoost, Leven, TensorFlow, and PyTorch. The 'Libraries' section includes Intel® oneAPI Data Analytics Library (oneDAL), Intel® oneAPI Deep Neural Network Library (oneDNN), Intel® oneAPI Collective Communications Library (oneCCL), and Intel® oneAPI Math Kernel Library (oneMKL). 'Languages' covers Fortran, SYCL*, and DPC++/C++. 'Analyzers' includes Intel® Inspector, Intel® Profiler, and Intel® Trace Collector and Analyzer. 'Debuggers' lists GDB* and Intel® System Debugger. 'Migration' includes SYCLomatic and Intel® DPC++ Compatibility Tool. 'Hypervisors / Orchestration' includes KVM 5.17, Hyper-V 1H'22, ESXi*/vSphere* 8.0, Kubernetes*, and RedHat OpenShift*. 'OS & Kernel' includes Microsoft Windows Server* 2022, Linux* 5.16, Red Hat 8.6, Ubuntu* 22.04, and SUSE Linux Enterprise Server* 15 SP4. 'Hardware' includes 4th & 5th Gen Intel® Xeon® Scalable Processor, Intel® Xeon® CPU Max Series, and Intel® Data Center GPU Max Series. The CCL library provides multiple commands for creating, managing, and using CCL communicators.⁴⁶ The CCL library includes software objects that are used to manage the communication between the domains / nodes accelerators in the 4th Gen Xeon Scalable Processor.⁴⁷ For example, the oneCCL library includes commands that cause the 4th Gen Xeon Scalable Processor to receive the first result and execute additional matrix operations on the first result and return a second result to one or more of the first, second, or third cluster node modules using the third domain of the 4th Gen Xeon Scalable Processor. The cores within the processor are configured to evaluate multiple mathematical expressions and matrix operations.⁴⁸
--------------------------	--

⁴⁶ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”).

⁴⁷ <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-9/device-communication.html>.

⁴⁸ <https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/advanced-matrix-extensions-overview.html>;
<https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products
	The source code for the CCL library includes the details for using oneCCL commands to initialize communications on a 4th Gen Xeon Scalable Processor, receive the first result, and execute additional matrix operations on the first result and return a second result to one or more of the first, second, or third cluster node modules using the third domain of the 4th Gen Xeon Scalable Processor.⁴⁹
a third cluster node module residing in the at least one computer-readable medium, said third cluster node module configured to send commands to the third cluster node module and configured to communicate with the first cluster node module and the second cluster node module. For example, as depicted below, the 4th Gen Xeon Scalable Processor, configured in a Sub-NUMA clustering 4 (SNC-4) affinity mode, includes four domains where each domain includes fifteen processing cores, each of which is configured to access memory including, for example, LLC and DDR memory.⁵⁰	

⁴⁹ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

⁵⁰ “Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family,” <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

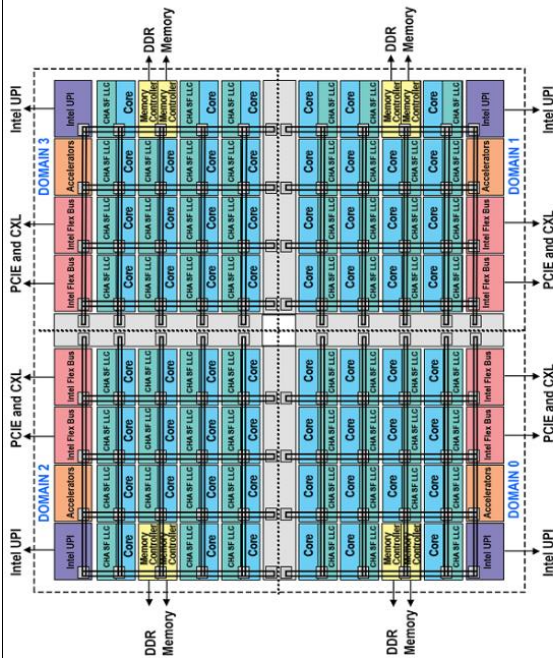


Figure 3 - Block Diagram Representing Domains Of sub-NUMA With Four Clusters

Each Accused Xeon Product is configured to access and be in communication with a non-transitory computer-readable medium. For example, the 4th Gen Xeon Scalable Processor, as shown above, is configured to access and be in communication with a non-transitory computer-readable medium that includes, for example, local memory (e.g., LLC), and external memory (e.g., DDR memory).

Intel's Accused Xeon Products are configured to be supported by Intel's oneAPI software suite, including Intel's oneCCL software.⁵¹ For example, Intel's 4th Gen Xeon Scalable Processor are configured to be supported by Intel's oneAPI software suite, which, for example, "maximize[s] application performance by activating the advanced capabilities of 4th and 5th gen Intel® Xeon®

⁵¹ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> ("Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.").

U.S. Pat. No. 11,811,582	Accused Xeon Products
	processors....”⁵² The oneCCL library is one of the 10 core elements of Intel’s oneAPI software suite.⁵³ Intel’s oneCCL package comprises the oneCCL Software Development Kit (SDK) and the Intel® MPI Library Runtime components.⁵⁴ “The Intel® oneAPI Math Kernel Library (oneMKL) is a computing math library of highly optimized and extensively parallelized routines for applications that require maximum performance. oneMKL contains the high-performance optimizations from the full Intel® Math Kernel Library for CPU architectures (with C/Fortran programming language interfaces) and adds to them a set of SYCL* interfaces for achieving performance on various CPU architectures and Intel Graphics Technology for certain key functionalities. oneMKL provides BLAS and LAPACK linear algebra routines, fast Fourier transforms, vectorized math functions, random number generation functions, and other functionality.”⁵⁵ The mathematical routines addressed in oneMKL include matrix operations using, for example, arrays of data elements. The figure below depicts Intel’s oneAPI software suite and its Collective Communications Library (oneCCL) and Math Kernel Library (oneMKL) in the context of the software suite for Intel’s 4th Gen Xeon Scalable Processors:⁵⁶

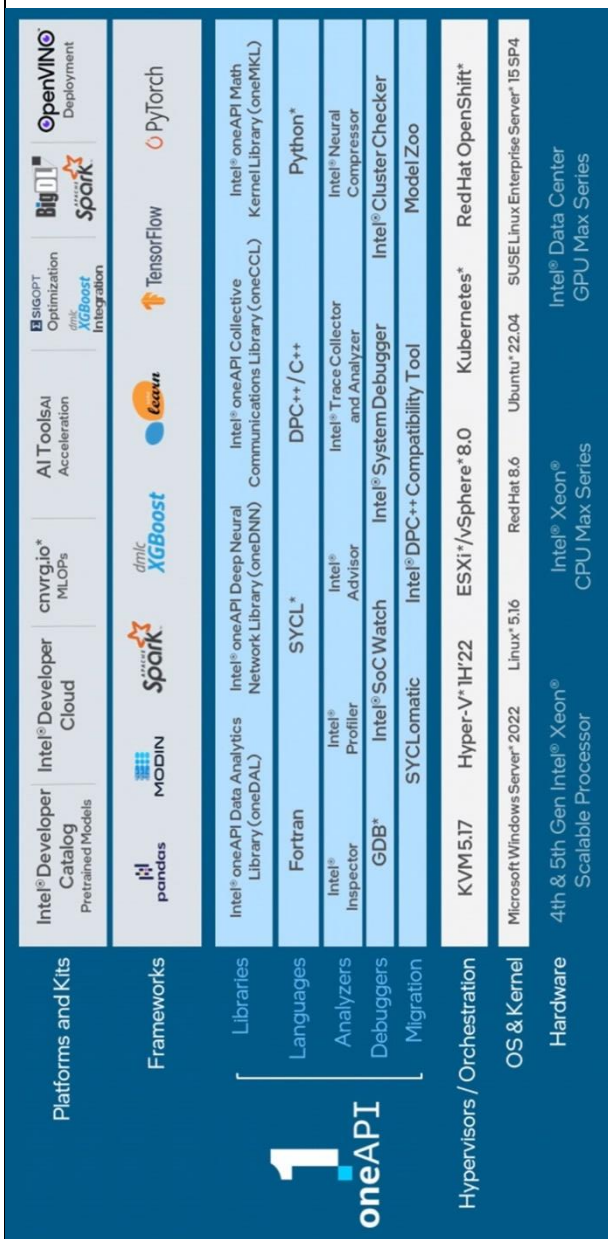
⁵² “Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors,” <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>; “oneAPI Specification, Release 1.3-rev-1,” pp. 388-389 <https://spec.oneapi.io/versions/latest/oneAPI-spec.pdf>; “Intel® oneAPI Collective Communications Library Developer Guide and Reference,” <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-12/overview.html>, <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-12/overview.html>.

⁵³ <https://www.intel.com/content/www/us/en/developer/articles/technical/oneapi-what-is-it.html>.

⁵⁴ See “Intel® oneAPI Collective Communications Library Developer Guide and Reference,” <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-12/overview.html>.

⁵⁵ “Intel® oneAPI Programming Guide,” <https://www.intel.com/content/www/us/en/docs/oneapi/programming-guide/2024-1/intel-oneapi-math-kernel-library-onemkl.html>.

⁵⁶ “Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors,” <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>.



The CCL library provides multiple commands for creating, managing, and using CCL communicators.⁵⁷ The CCL library includes software objects that are used to manage the communication between the domains / nodes accelerators in the 4th Gen Xeon Scalable Processor.⁵⁸

For example, the oneCCL library includes commands that cause the 4th Gen Xeon Scalable Processor to send commands to the third kernel and configured to communicate with the first cluster node module and the second cluster node module.⁵⁹

⁵⁷ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”)

⁵⁸ <https://www.intel.com/content/www/us/en/docs/oneapi/onecccl/developer-guide-reference/2021-9/device-communication.html>.

⁵⁹ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/onecccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”)

U.S. Pat. No. 11,811,582	Accused Xeon Products
	The source code for the CCL library includes the details for using oneCCL commands to initialize communications on a 4th Gen Xeon Scalable Processor and send commands to the third kernel and communicate with the first cluster node module and the second cluster node module. ⁶⁰
wherein the first cluster node module is configured to return, based on the second result, a final result.	Each of the Accused Xeon Products comprises a first cluster node module configured to return, based on the second result, a final result. For example, the oneCCL library includes commands that cause 4th Gen Xeon Scalable Processor to return based on the second result, a final result.⁶¹ The source code for the CCL library includes the details for using oneCCL commands to initialize communications on a 4th Gen Xeon Scalable Processor and return, based on the second result, a final result.⁶²
3. The computer cluster of claim 1, wherein the at least one computer-readable medium comprises a first computer-readable medium coupled to at least the first processor and a second computer-readable medium coupled to at least the	Each of the Accused Xeon Products comprises at least one computer-readable medium comprises a first computer-readable medium coupled to at least the first processor and a second computer-readable medium coupled to at least the second processor, wherein the first kernel and the first cluster node module reside in the first computer-readable medium, and wherein the second kernel and the second cluster node module reside in the second computer-readable medium. For example, as depicted below, the 4th Gen Xeon Scalable Processor, configured in a Sub-NUMA clustering 4 (SNC-4) affinity mode, includes four domains where each domain includes fifteen

⁶⁰ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

⁶¹ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

⁶² Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

U.S. Pat. No. 11,811,582	Accused Xeon Products Intel's oneAPI software suite.⁶⁵ Intel's oneCCL package comprises the oneCCL Software Development Kit (SDK) and the Intel® MPI Library Runtime components.⁶⁶ “The Intel® oneAPI Math Kernel Library (oneMKL) is a computing math library of highly optimized and extensively parallelized routines for applications that require maximum performance. oneMKL contains the high-performance optimizations from the full Intel® Math Kernel Library for CPU architectures (with C/Fortran programming language interfaces) and adds to them a set of SYCL* interfaces for achieving performance on various CPU architectures and Intel Graphics Technology for certain key functionalities. oneMKL provides BLAS and LAPACK linear algebra routines, fast Fourier transforms, vectorized math functions, random number generation functions, and other functionality.”⁶⁷ The mathematical routines addressed in oneMKL include matrix operations using, for example, arrays of data elements. The figure below depicts Intel's oneAPI software suite and its Collective Communications Library (oneCCL) and Math Kernel Library (oneMKL) in the context of the software suite for Intel's 4th Gen Xeon Scalable Processors:⁶⁸
--------------------------	--

Collective Communications Library Developer Guide and Reference,” <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/overview.html>, <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/communicator.html>.

⁶⁵ <https://www.intel.com/content/www/us/en/developer/articles/technical/oneapi-what-is-it.html>.

⁶⁶ See “Intel® oneAPI Collective Communications Library Developer Guide and Reference,” <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/overview.html>.

⁶⁷ “Intel® oneAPI Programming Guide,” <https://www.intel.com/content/www/us/en/docs/oneapi/programming-guide/2024-1/intel-oneapi-math-kernel-library-onemkl.html>.

⁶⁸ “Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors,” <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>.

U.S. Pat. No. 11,811,582	Accused Xeon Products  Platforms and Kits Intel® Developer Catalog Pretrained Models Intel® Developer Cloud cnvrg.io* MLOPs AI ToolsAI Acceleration BigDL™ Spark OpenVINO™ Deployment Frameworks MODIN Spark TensorFlow PyTorch Intel® oneAPI Data Analytics Library (oneDAL) Intel® oneAPI Deep Neural Network Library (oneDNN) Intel® oneAPI Collective Communications Library (oneCCL) Intel® oneAPI Math Kernel Library (oneMKL) Libraries Fortran SYCL* DPC++/C++ Python* Analyzers Intel® Inspector Intel® Profiler Intel® Trace Collector and Analyzer Intel® Neural Compressor Debuggers GDB* Intel® SoC Watch Intel® System Debugger Intel® Cluster Checker Migration SYCLomatic Intel® DPC++ Compatibility Tool Model Zoo Hypervisors / Orchestration KVM 5.17 Hyper-V™ 1H'22 ESXi* vSphere* 8.0 Kubernetes* RedHat OpenShift* OS & Kernel Microsoft Windows Server* 2022 Linux* 5.16 Red Hat 8.6 Ubuntu* 22.04 SUSE Linux Enterprise Server* 15 SP4 Hardware 4th & 5th Gen Intel® Xeon® Scalable Processor Intel® Xeon® CPU Max Series Intel® Data Center GPU Max Series
4. The computer cluster of claim 1, further comprising a cluster configuration module that initializes at least one of the first cluster	The CCL library provides multiple commands for creating, managing, and using CCL communicators.⁶⁹ The CCL library includes software objects that are used to manage the communication between the domains / nodes accelerators in the 4th Gen Xeon Scalable Processor.⁷⁰ The CCL library is responsible for distributing the workload across the domains / nodes and for collecting the results from the domains / nodes. Each of the Accused Xeon Products comprises a computer cluster that includes a cluster configuration module that initializes at least one of the first cluster node module, the second cluster node module, or the third cluster node module. For example, as depicted below, the 4th Gen Xeon Scalable Processor

⁶⁹ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”).

⁷⁰ <https://www.intel.com/content/www/us/en/docs/oneapi/oneccl/developer-guide-reference/2021-9/device-communication.html>.

U.S. Pat. No. 11,811,582	node module, the second cluster node module, or the third cluster node module.
Accused Xeon Products	comprises four nodes (described as domains in the figure below), which are configured to be connected in an all-to-all local bus interconnection configuration.⁷¹ The diagram illustrates a sub-NUMA architecture with four domains. Each domain (DOMAIN 0, 1, 2, 3) is a self-contained unit. Within each domain, there is an Intel UPI interface at the top, followed by an Intel Fixed Bus, Intel Accelerators, Intel Core, Intel Memory Controller, and Intel Memory. The domains are interconnected via PCIe and CXL interfaces. The diagram shows a 2x2 grid of domains, with each domain having its own set of components and being connected to the others through the interconnects. Figure 3 – Block Diagram Representing Domains Of sub-NUMA With Four Clusters Each of the 4th Gen Xeon Scalable Processors is configured to operate in a Sub-NUMA clustering 4 (SNC-4) affinity mode, comprising four domains of sub-NUMA clusters / nodes. Each Accused Xeon Product comprises a plurality of nodes, wherein each of the plurality of nodes comprises a hardware processor, wherein one or more of the nodes includes a cluster configuration module configured to receive a command to start a cluster initialization process for the computer cluster. For example, each 4th Gen Xeon Scalable Processor comprises a cluster configuration module configured to receive a command to start a cluster initialization process for the computer cluster that includes the four domains (node), thereby including at least four cluster node modules. This is done,

⁷¹ “Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family,” <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products
	for example, by using Intel's oneCCL software, which is part of Intel's oneAPI Collective Communications Library.⁷² More particularly, Intel's Accused Xeon Products are configured to be supported by Intel's oneAPI software suite, which, for example, "maximize[s] application performance by activating the advanced capabilities of 4th and 5th gen Intel® Xeon® processors...."⁷³ The oneCCL library is one of the 10 core elements of Intel's oneAPI software suite.⁷⁴ Intel's oneCCL package comprises the oneCCL Software Development Kit (SDK) and the Intel® MPI Library Runtime components.⁷⁵
6. The computer cluster of claim 1, further comprising a data structure in which messages originating from the second and third cluster node modules are stored.	Each of the Accused Xeon Products comprises a first cluster node module comprising a data structure in which messages originating from the second and third cluster node modules are stored. For example, the oneCCL library includes commands that cause 4th Gen Xeon Scalable Processor to create a data structure in which messages originating from the second and third cluster node modules are stored, where the data structure stores messages originating from the various cluster node modules are stored.⁷⁶ The source code for the CCL library includes the details for using oneCCL commands to initialize communications on a 4th Gen Xeon Scalable Processor and to create a data structure in which messages originating from the second and third cluster node modules are stored and where the data structure stores messages originating from the various cluster node modules are stored.⁷⁷

⁷² Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf>.

⁷³ "Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors," <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqdi>; "oneAPI Specification, Release 1.3-rev-1," pp. 388-389 <https://spec.oneapi.io/versions/latest/oneAPI-spec.pdf>; "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneapi/oneccl/developer-guide-reference/2021-12/overview.html>, <https://www.intel.com/content/www/us/en/developer/articles/technical/oneapi-what-is-it.html>.

⁷⁴ <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl/developer-guide-reference/2021-12/overview.html>.

⁷⁵ See "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneapi/oneccl/developer-guide-reference/2021-12/overview.html>.

⁷⁶ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> ("Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter."); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

⁷⁷ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> ("Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter."); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

U.S. Pat. No. 11,811,582	Accused Xeon Products
7. The computer cluster of claim 6, wherein the data structure of the first cluster node module comprises a received message queue.	Each of the Accused Xeon Products comprises a first cluster node module comprising a data structure wherein the data structure of the first cluster node module comprises a received message queue. For example, the oneCCL library includes commands that cause 4th Gen Xeon Scalable Processor to create a data structure in which messages originating from the second and third cluster node modules are stored, where the data structure comprises a received message queue.⁷⁸ The source code for the CCL library includes the details for using oneCCL commands to initialize communications on a 4th Gen Xeon Scalable Processor and to create a data structure in which messages originating from the second and third cluster node modules are stored and where the data structure comprises a received message queue.⁷⁹
8. The computer cluster of claim 7, wherein the first cluster node module further comprises a message receiving queue, in which data specifying a location to which the first cluster node module can expect to receive an expression and an identifier of a processor from which the expression is expected to be sent are stored. For example, the oneCCL library includes commands that cause 4th Gen Xeon Scalable Processor to create a data structure in which messages originating from the second and third cluster node modules are stored, where the data structure comprises a message receiving queue in which data specifying a location to which the first cluster node module can expect to receive an expression and an identifier of a processor from which the expression is expected to be sent are stored.⁸⁰	Each of the Accused Xeon Products comprises a first cluster node module wherein the first cluster node module further comprises a message receiving queue, in which data specifying a location to which the first cluster node module can expect to receive an expression and an identifier of a processor from which the expression is expected to be sent are stored. For example, the oneCCL library includes commands that cause 4th Gen Xeon Scalable Processor to create a data structure in which messages originating from the second and third cluster node modules are stored, where the data structure comprises a message receiving queue in which data specifying a location to which the first cluster node module can expect to receive an expression and an identifier of a processor from which the expression is expected to be sent are stored.⁸⁰

⁷⁸ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

⁷⁹ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

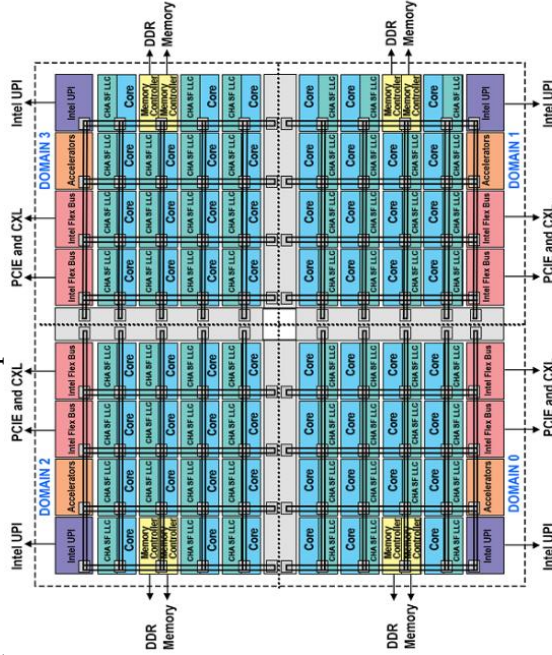
⁸⁰ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

U.S. Pat. No. 11,811,582	Accused Xeon Products
is expected to be sent are stored.	The source code for the CCL library includes the details for using oneCCL commands to initialize communications on a 4th Gen Xeon Scalable Processor and to create a data structure in which messages originating from the second and third cluster node modules are stored and where the data structure comprises a message receiving queue in which data specifying a location to which the first cluster node module can expect to receive an expression and an identifier of a processor from which the expression is expected to be sent are stored. ⁸¹
9. The computer cluster of claim 1, wherein the second cluster node module further comprises a received message queue in which messages originating from the first and third cluster node modules are stored.	Each of the Accused Xeon Products comprises a first cluster node module comprising a data structure wherein the second cluster node module further comprises a received message queue in which messages originating from the first and third cluster node modules are stored. For example, the oneCCL library includes commands that cause 4th Gen Xeon Scalable Processor to create a data structure in which messages originating from the second and third cluster node modules are stored, where the data structure comprises a received message queue in which messages originating from the various cluster node modules are stored.⁸² The source code for the CCL library includes the details for using oneCCL commands to initialize communications on a 4th Gen Xeon Scalable Processor and to create a data structure in which messages originating from the second and third cluster node modules are stored and where the data structure comprises a received message queue in which messages originating from the various cluster node modules are stored.⁸³

⁸¹ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

⁸² Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

⁸³ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

U.S. Pat. No. 11,811,582	Accused Xeon Products
10. The computer cluster of claim 1, wherein the first cluster node module comprises an advanced functions module, and wherein the advanced functions module comprises a call that evaluates an expression across the computer cluster in parallel.	Each of the Accused Xeon Products comprises a first cluster node module which comprises an advanced functions module, and wherein the advanced functions module comprises a call that evaluates an expression across the computer cluster in parallel. For example, as depicted below and as described further herein, Intel's 4th Gen Xeon Scalable Processor⁸⁴ includes a computer cluster comprising a computer cluster node (e.g., domain) for evaluating expressions in parallel with other computer cluster nodes.  Figure 3 – Block Diagram Representing Domains Of sub-NUMA With Four Clusters “The Intel® oneAPI Math Kernel Library (oneMKL) is a computing math library of highly optimized and extensively parallelized routines for applications that require maximum performance. oneMKL contains the high-performance optimizations from the full Intel® Math Kernel Library for CPU architectures (with C/Fortran programming language interfaces) and adds to them a set of SYCL.* interfaces for achieving performance on various CPU architectures and Intel Graphics Technology for

⁸⁴ “Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family,” <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products
	certain key functionalities. oneMKL provides BLAS and LAPACK linear algebra routines, fast Fourier transforms, vectorized math functions, random number generation functions, and other functionality.”⁸⁵ The mathematical routines addressed in oneMKL include matrix operations using, for example, arrays of data elements. The Intel® oneAPI Math Kernel Library (oneMKL), therefore, comprises calls that evaluate expressions across the computer cluster in parallel. The figure below depicts Intel’s oneAPI software suite and its Collective Communications Library (oneCCL) and Math Kernel Library (oneMKL) in the exemplary context of the software suite for Intel’s 4th Gen Xeon Scalable Processors.⁸⁶

⁸⁵ “Intel® oneAPI Programming Guide,” <https://www.intel.com/content/www/us/en/docs/oneapi/programming-guide/2024-1/intel-oneapi-math-kernel-library-onemkl.html>.

⁸⁶ “Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors,” <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>.

U.S. Pat. No. 11,811,582	Accused Xeon Products The diagram illustrates the Intel oneAPI ecosystem architecture, organized into several layers: Platforms and Kits: Includes Intel® Developer Catalog Pretrained Models, Intel® Developer Cloud, cnvrg.io* MLOps, AI Tools/AI Acceleration, SIGOPT Optimization (oneAPI XGBoost Integration), BigDL, Spark, and OpenVINO Deployment. Frameworks: Includes pandas, MODIN, Spark, oneAPI XGBoost, Leven, TensorFlow, and PyTorch. Libraries: Includes Intel® oneAPI Data Analytics Library (oneDAL), Intel® oneAPI Deep Neural Network Library (oneDNN), Intel® oneAPI Collective Communications Library (oneCCL), and Intel® oneAPI Math Kernel Library (oneMKL). Languages: Includes Fortran, SYCL*, and DPC++/C++. Analyzers: Includes Intel® Inspector, Intel® Profiler, and Intel® Trace Collector and Analyzer. Debuggers: Includes GDB*, Intel® SoC Watch, Intel® System Debugger, and Intel® Cluster Checker. Migration: Includes SYCLomatic and Intel® DPC++ Compatibility Tool. Hypervisors / Orchestration: Includes KVM 5.17, Hyper-V* 1H*22, ESXi*/vSphere* 8.0, Kubernetes*, and Red Hat OpenShift*. OS & Kernel: Includes Microsoft Windows Server* 2022, Linux* 5.16, Red Hat 8.6, Ubuntu* 22.04, and SUSE Linux Enterprise Server* 15 SP4. Hardware: Includes 4th & 5th Gen Intel® Xeon® Scalable Processor, Intel® Xeon® CPU Max Series, and Intel® Data Center GPU Max Series.
11. The computer cluster of claim 10, wherein the advanced functions module comprises a call that	The CCL library provides multiple commands for creating, managing, and using CCL communicators.⁸⁷ The CCL library includes software objects that are used to manage the communication between the domains / nodes accelerators in the 4th Gen Xeon Scalable Processor.⁸⁸ The CCL library is responsible for receiving first commands from a user interface without the first commands first passing through the first kernel, and after receiving the first commands from the user interface, send second commands to the first kernel. Each of the Accused Xeon Products comprises an advanced functions module which comprises a call that calculates a Fourier transform across the computer cluster in parallel.

⁸⁷ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”).

⁸⁸ <https://www.intel.com/content/www/us/en/docs/oneapi/oneccl/developer-guide-reference/2021-9/device-communication.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products Intel’s Accused Xeon Products are configured to be supported by Intel’s oneAPI software suite, including Intel’s oneCCL software.⁹⁰ For example, Intel’s 4th Gen Xeon Scalable Processor are configured to be supported by Intel’s oneAPI software suite, which, for example, “maximize[s] application performance by activating the advanced capabilities of 4th and 5th gen Intel® Xeon® processors....”⁹¹ The oneCCL library is one of the 10 core elements of Intel’s oneAPI software suite.⁹² Intel’s oneCCL package comprises the oneCCL Software Development Kit (SDK) and the Intel® MPI Library Runtime components.⁹³ “The Intel® oneAPI Math Kernel Library (oneMKL) is a computing math library of highly optimized and extensively parallelized routines for applications that require maximum performance. oneMKL contains the high-performance optimizations from the full Intel® Math Kernel Library for CPU architectures (with C/Fortran programming language interfaces) and adds to them a set of SYCL* interfaces for achieving performance on various CPU architectures and Intel Graphics Technology for certain key functionalities. oneMKL provides BLAS and LAPACK linear algebra routines, fast Fourier transforms, vectorized math functions, random number generation functions, and other functionality.”⁹⁴ The mathematical routines addressed in oneMKL include matrix operations using, for example, arrays of data elements. Utilizing the highly optimized FFT implementations within the Intel oneMKL library, oneMKL provides optimized routines for various FFT algorithms, leveraging advanced instruction sets (like AVX-512) and multi-core parallelism available on Xeon Scalable processors. The highly optimized Basic Linear Algebra Subprograms (BLAS) routines within oneMKL are used to perform the matrix multiplications involved in the butterfly operations of the
--------------------------	--

⁹⁰ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”).

⁹¹ “Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors,” <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>; “oneAPI Specification, Release 1.3-rev-1,” pp. 388-389 <https://spec.oneapi.io/versions/latest/oneapi-spec.pdf>; “Intel® oneAPI Collective Communications Library Developer Guide and Reference,” <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-12/overview.html>, <https://www.intel.com/content/www/us/en/docs/oneapi-what-is-it.html>.

⁹² <https://www.intel.com/content/www/us/en/developer/articles/technical/oneapi-what-is-it.html>.
⁹³ See “Intel® oneAPI Collective Communications Library Developer Guide and Reference,” <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-12/overview.html>.

⁹⁴ “Intel® oneAPI Programming Guide,” https://www.intel.com/content/www/us/en/docs/oneapi/programming_guide/2024-1/intel-oneapi-math-kernel-library-onemkl.html.

U.S. Pat. No. 11,811,582	Accused Xeon Products FFT algorithm. BLAS provides highly optimized implementations for matrix-vector and matrix-matrix multiplications. The full capabilities of the oneMKL library, including its optimized FFT routines, BLAS functions, and threading support are used to calculate a Fourier transform across the computer cluster in parallel. The figure below depicts Intel's oneAPI software suite and its Collective Communications Library (oneCCL) and Math Kernel Library (oneMKL) in the exemplary context of the software suite for Intel's 4th Gen Xeon Scalable Processors.⁹⁵ The diagram illustrates the Intel oneAPI software suite architecture, organized into several layers: Platforms and Kits: Includes Intel® Developer Catalog, Intel® Developer Cloud, and AI Tools/AI Acceleration. Frameworks: Includes pandas, MODIN, Spark, and XGBoost. Libraries: Includes Intel® oneAPI Data Analytics Library (oneDAL), Intel® oneAPI Deep Neural Network Library (oneDNN), Intel® oneAPI Collective Communications Library (oneCCL), and Intel® oneAPI Math Kernel Library (oneMKL). Languages: Includes Fortran, SYCL*, and DPC++/C++. Analyzers: Includes Intel® Inspector and Intel® Trace Collector and Analyzer. Debuggers: Includes GDB* and Intel® System Debugger. Migration: Includes SYCLomatic and Intel® DPC++ Compatibility Tool. Hypervisors / Orchestration: Includes KVM 5.17, Hyper-V* 1H'22, ESXi*/vSphere* 8.0, and Kubernetes*. OS & Kernel: Includes Microsoft Windows Server* 2022, Linux* 5.16, Red Hat 8.6, and Ubuntu* 22.04. Hardware: Includes 4th & 5th Gen Intel® Xeon® Scalable Processor, Intel® Xeon® CPU Max Series, and Intel® Data Center GPU Max Series.
--------------------------	--

⁹⁵ “Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors,” <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>.

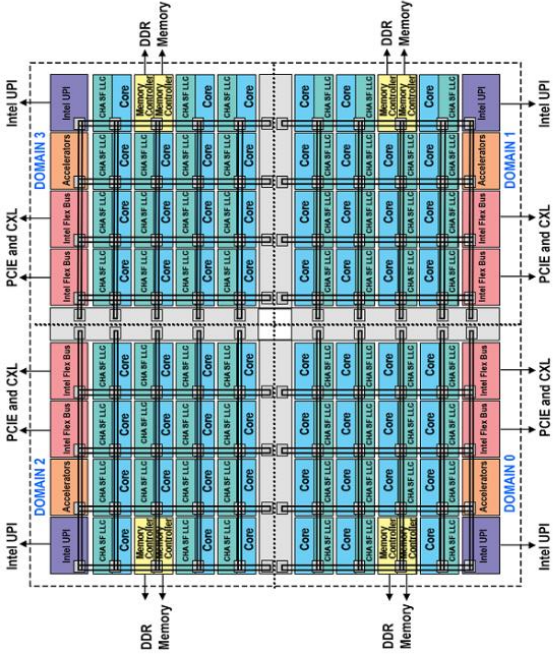
U.S. Pat. No. 11,811,582	Accused Xeon Products
	The CCL library provides multiple commands for creating, managing, and using CCL communicators.⁹⁶ The CCL library includes software objects that are used to manage the communication between the domains / nodes accelerators in the 4th Gen Xeon Scalable Processor.⁹⁷ The CCL library is responsible for receiving first commands from a user interface without the first commands first passing through the first kernel, and after receiving the first commands from the user interface, send second commands to the first kernel. The source code for the CCL library includes the details for using oneCCL commands to initialize communications on a 4th Gen Xeon Scalable Processor, receive first commands from a user interface without the first commands first passing through the first kernel, and after receiving the first commands from the user interface, send second commands to the first kernel. As described above, oneMKL is a computing math library of highly optimized and extensively parallelized routines for applications that require maximum performance that is an advanced functions module for calling Fast Fourier Transforms, for example. Among its functionality is Fast Fourier Transforms⁹⁸ that are used within the parallelized routines of oneMKL to calculate a Fourier transform across the computer cluster in parallel.⁹⁹
12. The computer cluster of claim 10, wherein the advanced functions module comprises one or more calls that evaluate matrix operations across the	Each of the Accused Xeon Products comprises an advanced functions module, and wherein the advanced functions module comprises one or more calls that evaluate matrix operations across the computer cluster in parallel.

⁹⁶ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”).

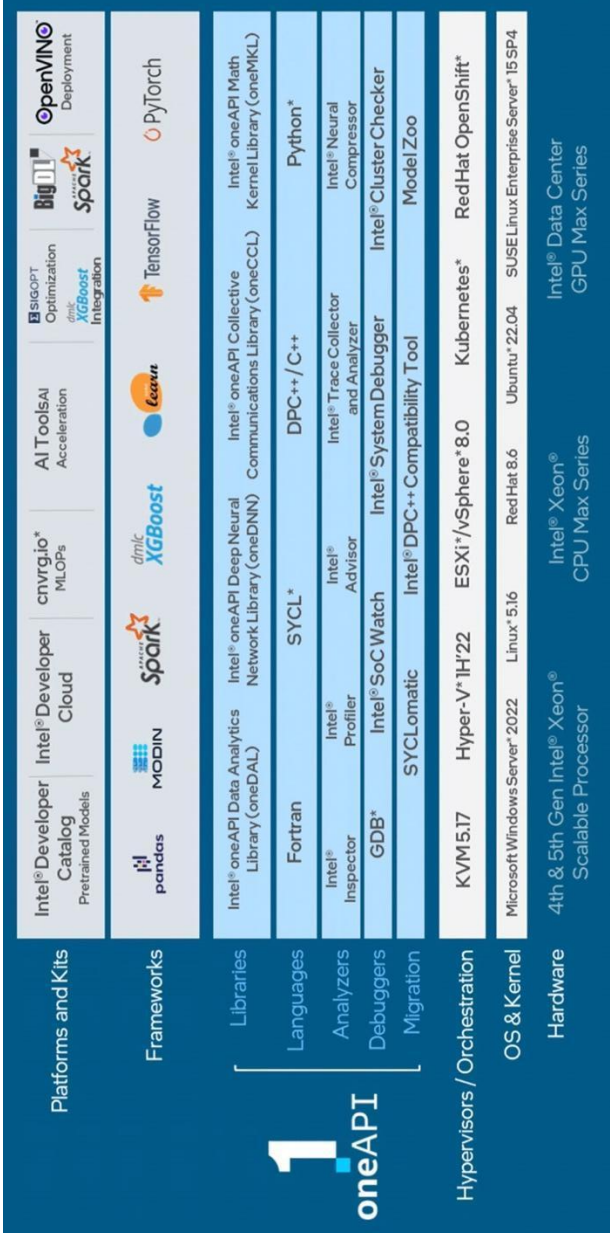
⁹⁷ <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-9/device-communication.html>.

⁹⁸ See, e.g., <https://www.intel.com/content/www/us/en/developer/articles/technical/performance-benchmarks-onekl-on-xeon-processors.html>; <https://www.intel.com/content/www/us/en/docs/onemkl/developer-reference-c/2023-1/cluster-fft-functions.html>.

⁹⁹ “Developer Reference for Intel® oneAPI Math Kernel Library for C,” <https://www.intel.com/content/www/us/en/docs/onemkl/developer-reference-c/2023-1/cluster-fft-functions.html> (“One or more [Cluster FFT] processes may be running in parallel on each cluster node.”).

U.S. Pat. No. 11,811,582	computer cluster in parallel.
Accused Xeon Products	For example, as depicted below and as described further herein, Intel's 4th Gen Xeon Scalable Processor¹⁰⁰ includes a computer cluster comprising a computer cluster node (e.g., domain) for evaluating expressions in parallel with other computer cluster nodes.  Figure 3 – Block Diagram Representing Domains Of sub-NUMA With Four Clusters “The Intel® oneAPI Math Kernel Library (oneMKL) is a computing math library of highly optimized and extensively parallelized routines for applications that require maximum performance. oneMKL contains the high-performance optimizations from the full Intel® Math Kernel Library for CPU architectures (with C/Fortran programming language interfaces) and adds to them a set of SYCL* interfaces for achieving performance on various CPU architectures and Intel Graphics Technology for certain key functionalities. oneMKL provides BLAS and LAPACK linear algebra routines, fast Fourier transforms, vectorized math functions, random number generation functions, and other

¹⁰⁰ “Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family,” <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products functionality.”¹⁰¹ The mathematical routines addressed in oneMKL include matrix operations using, for example, arrays of data elements. The Intel® oneAPI Math Kernel Library (oneMKL), therefore, comprises calls that evaluate matrix operations across the computer cluster in parallel. The figure below depicts Intel’s oneAPI software suite and its Collective Communications Library (oneCCL) and Math Kernel Library (oneMKL) in the exemplary context of the software suite for Intel’s 4th Gen Xeon Scalable Processors.¹⁰²  The diagram illustrates the Intel oneAPI software suite architecture. At the top, the 'Platforms and Kits' section includes Intel® Developer Catalog Pretrained Models, Intel® Developer Cloud, cnvrg.io* MLOPs, AI ToolsAI Acceleration, SIGOPT Optimization dnnl XGBoost Integration, BigDL, and OpenVINO Deployment. Below this, the 'Frameworks' section lists pandas, MODIN, Spark, dnnl XGBoost, Leven, TensorFlow, and PyTorch. The main body is divided into several categories: 'Libraries' (Intel® oneAPI Data Analytics Library (oneDAL), Intel® oneAPI Deep Neural Network Library (oneDNN), Intel® oneAPI Collective Communications Library (oneCCL), Intel® oneAPI Math Kernel Library (oneMKL)), 'Languages' (Fortran, SYCL*, DPC++/C++), 'Analyzers' (Intel® Inspector, Intel® Profiler, Intel® Trace Collector and Analyzer), 'Debuggers' (GDB*, Intel® SoC Watch, Intel® System Debugger, Intel® Cluster Checker), 'Migration' (SYCLomatic, Intel® DPC++ Compatibility Tool), 'Hypervisors / Orchestration' (KVM 5.17, Hyper-V* 1H'22, ESXi*/vSphere* 8.0, Kubernetes*, Red Hat OpenShift*), 'OS & Kernel' (Microsoft Windows Server* 2022, Linux* 5.16, Red Hat 8.6, Ubuntu* 22.04, SUSE Linux Enterprise Server* 15 SP4), and 'Hardware' (4th & 5th Gen Intel® Xeon® Scalable Processor, Intel® Xeon® CPU Max Series, Intel® Data Center GPU Max Series).
--------------------------	--

¹⁰¹ “Intel® oneAPI Programming Guide,” <https://www.intel.com/content/www/us/en/docs/oneapi/programming-guide/2024-1/intel-oneapi-math-kernel-library-onemkl.html>.

¹⁰² “Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors,” <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqdl>.

U.S. Pat. No. 11,811,582	Accused Xeon Products
	The CCL library provides multiple commands for creating, managing, and using CCL communicators.¹⁰³ The CCL library includes software objects that are used to manage the communication between the domains / nodes accelerators in the 4th Gen Xeon Scalable Processor.¹⁰⁴ The CCL library is responsible for receiving first commands from a user interface without the first commands first passing through the first kernel, and after receiving the first commands from the user interface, send second commands to the first kernel.
13. The computer cluster of claim 1, wherein the final result comprises at least a portion of the second result.	Each of the Accused Xeon Products comprises a computer cluster wherein the final result comprises at least a portion of the second result. For example, the oneCCL library includes commands that cause 4th Gen Xeon Scalable Processor to return a final result based on at least a portion of the second result.¹⁰⁵ The source code for the CCL library includes the details for using oneCCL commands to initialize communications on a 4th Gen Xeon Scalable Processor and return a final result based on at least a portion of the second result, a final result.¹⁰⁶
14. The computer cluster of claim 1, wherein the first cluster node module is configured to return the final result to the user interface.	Each of the Accused Xeon Products comprises a computer cluster wherein the first cluster node module is configured to return the final result to the user interface.

¹⁰³ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”).

¹⁰⁴ <https://www.intel.com/content/www/us/docs/oneccl/developer-guide-reference/2021-9/device-communication.html>.

¹⁰⁵ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

¹⁰⁶ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

U.S. Pat. No. 11,811,582	Accused Xeon Products
	For example, the oneCCL library includes commands that cause 4th Gen Xeon Scalable Processor to return a final result based on at least a portion of the second result.¹⁰⁷ This final result is then returned to the user interface by the first cluster node module. The source code for the CCL library includes the details for using oneCCL commands to initialize communications on a 4th Gen Xeon Scalable Processor and return a final result based on at least a portion of the second result, a final result.¹⁰⁸ This final result is then returned to the user interface by the first cluster node module.
15. The computer cluster of claim 1, further comprising a communications network interface for the cluster node modules to communicate, with each other using a peer-to-peer architecture, results of execution of the matrix operations.	Each of the Accused Xeon Products comprises a mechanism such as a communications network interface for the cluster node modules to communicate, with each other using a peer-to-peer architecture, results of execution of the matrix operations. For example, as depicted below, the 4th Gen Xeon Scalable Processor is configured to provide all-to-all communication using Sub-NUMA clustering 4 (SNC-4) affinity mode using four domains of sub-NUMA clusters / nodes.¹⁰⁹

¹⁰⁷ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

¹⁰⁸ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”); <https://github.com/oneapi-src/oneAPI-samples/tree/master/Libraries/oneCCL>.

¹⁰⁹ “Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family,” <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

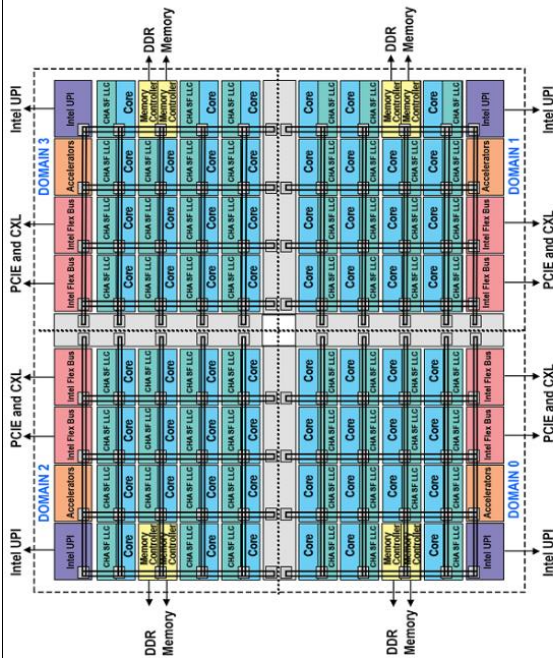


Figure 3 – Block Diagram Representing Domains Of sub-NUMA With Four Clusters

Each of the 4th Gen Xeon Scalable Processors is configured to operate in a Sub-NUMA clustering 4 (SNC-4) affinity mode, comprising four domains of sub-NUMA clusters / nodes.

Each Accused Xeon Product comprises a hardware processor that comprises multiple processor cores. For example, the 4th Gen Xeon Scalable Processor is configured to operate in an SNC-4 mode where each of the four domains is a node that includes a plurality of processing cores (fifteen) as shown in the above figure.

Each Accused Xeon Product comprises a user connection interface configured to receive a command to start a cluster initialization process for the computer cluster of each Accused Xeon Product. For example, each 4th Gen Xeon Scalable Processor comprises a user connection interface configured to receive a command to start a cluster initialization process for the computer cluster that includes the four domains (node). For example, Intel's 4th Gen Xeon Scalable Processors are configured such that each domain / node in a Sub-NUMA clustering 4 (SNC-4) affinity mode, includes at least two types

U.S. Pat. No. 11,811,582	Accused Xeon Products
	of compute engines for evaluating mathematical expressions including matrix operations using multiple processors and cores: • Intel Advanced Matrix Extensions Accelerator (AMX): AMX is a specialized hardware accelerator for evaluating matrix multiplication expressions.¹¹⁰ • Processing Cores (PCs): The PCs are general-purpose hardware accelerators that can be used to evaluate a variety of mathematical expressions.¹¹¹ This is done, for example, by using Intel's oneCCL software, which is part of Intel's oneAPI Collective Communications Library.¹¹² More particularly, Intel's Accused Xeon Products are configured to be supported by Intel's oneAPI software suite, which, for example, "maximize[s] application performance by activating the advanced capabilities of 4th and 5th gen Intel® Xeon® processors...."¹¹³ The oneCCL library is one of the 10 core elements of Intel's oneAPI software suite.¹¹⁴ Intel's oneCCL package comprises the oneCCL Software Development Kit (SDK) and the Intel® MPI Library Runtime components.¹¹⁵
16. The computer cluster of claim 15, wherein one or more of the cluster node modules are configured to: accept user instructions; after accepting the user instructions, communicate	Each of the Accused Xeon Products comprises one or more nodes configured to: accept user instructions; after accepting user instructions, communicate at least some of the user instructions using the communications network interface for the nodes to communicate with each other; and after communicating at least some of the user instructions using the communications network interface, communicate at least some of the user instructions to one or more single-node kernels.

¹¹⁰ <https://www.intel.com/content/www/us/en/products/docs/accelerator-engines/advanced-matrix-extensions/overview.html>.

¹¹¹ <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

¹¹² Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf>.

¹¹³ "Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors," <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>; "oneAPI Specification, Release 1.3-rev-1," pp. 388-389 <https://spec.oneapi.io/versions/latest/oneAPI-spec.pdf>; "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/overview.html>, <https://www.intel.com/content/www/us/en/developer/articles/technical/oneapi-what-is-it.html>.

¹¹⁴ <https://www.intel.com/content/www/us/en/developer/articles/technical/oneapi-what-is-it.html>.

¹¹⁵ See "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/overview.html>.

U.S. Pat. No. 11,811,582 at least some of the user instructions using the communications network interface for the cluster node modules to communicate with each other; and after communicating at least some of the user instructions using the communications network interface, communicate at least some of the user instructions to one or more of the kernels.	Accused Xeon Products For example, as depicted below, the 4th Gen Xeon Scalable Processor, configured in a Sub-NUMA clustering 4 (SNC-4) affinity mode, includes four domains where each domain includes fifteen processing cores, each of which is configured to access memory including, for example, LLC, DDR, and HBM memory.¹¹⁶ Figure 3 – Block Diagram Representing Domains Of sub-NUMA With Four Clusters Intel’s Accused Xeon Products are configured to be supported by Intel’s oneAPI software suite, which, for example, “maximize[s] application performance by activating the advanced capabilities of 4th and 5th gen Intel® Xeon® processors...”¹¹⁷ The oneCCL library is one of the 10 core elements of
---	---

¹¹⁶ “Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family,” <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

¹¹⁷ “Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors,” <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqjd>; “oneAPI Specification, Release 1.3-rev-1,” pp. 388-389 <https://spec.oneapi.io/versions/latest/oneAPI-spec.pdf>; “Intel® oneAPI Collective Communications Library Developer Guide and Reference,” <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-12/overview.html>, <https://www.intel.com/content/www/us/en/docs/oneapi/developer-guide-reference/2021-12/communicator.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products Intel's oneAPI software suite.¹¹⁸ Intel's oneCCL package comprises the oneCCL Software Development Kit (SDK) and the Intel® MPI Library Runtime components.¹¹⁹ “The Intel® oneAPI Math Kernel Library (oneMKL) is a computing math library of highly optimized and extensively parallelized routines for applications that require maximum performance. oneMKL contains the high-performance optimizations from the full Intel® Math Kernel Library for CPU architectures (with C/Fortran programming language interfaces) and adds to them a set of SYCL* interfaces for achieving performance on various CPU architectures and Intel Graphics Technology for certain key functionalities. oneMKL provides BLAS and LAPACK linear algebra routines, fast Fourier transforms, vectorized math functions, random number generation functions, and other functionality.”¹²⁰ The mathematical routines addressed in oneMKL include matrix operations using, for example, arrays of data elements. The figure below depicts Intel's oneAPI software suite and its Collective Communications Library (oneCCL) and Math Kernel Library (oneMKL) in the context of the software suite for Intel's 4th Gen Xeon Scalable Processors;¹²¹
--------------------------	--

¹¹⁸ <https://www.intel.com/content/www/us/en/developer/articles/technical/oneapi-what-is-it.html>.

¹¹⁹ See “Intel® oneAPI Collective Communications Library Developer Guide and Reference,” <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/overview.html>.

¹²⁰ “Intel® oneAPI Programming Guide,” <https://www.intel.com/content/www/us/en/docs/oneapi/programming-guide/2024-1/intel-oneapi-math-kernel-library-onemkl.html>.

¹²¹ “Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors,” <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>.

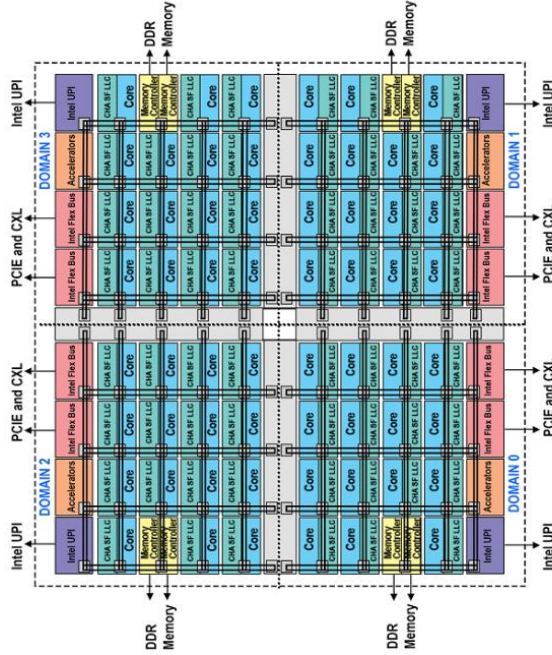
U.S. Pat. No. 11,811,582	Accused Xeon Products  Platforms and Kits Intel® Developer Catalog Pretrained Models Intel® Developer Cloud cnvrg.io* MLOPs AI Tools/AI Acceleration BigDL™ Spark OpenVINO™ Deployment Frameworks pandas MODIN Spark Intel® XGB Boost Leaven TensorFlow PyTorch Libraries Intel® oneAPI Data Analytics Library (oneDAL) Intel® oneAPI Deep Neural Network Library (oneDNN) Intel® oneAPI Math Kernel Library (oneMKL) Languages Fortran SYCL* DPC++/C++ Python* Analyzers Intel® Inspector Intel® Profiler Intel® Trace Collector and Analyzer Debuggers GDB* Intel® SoC Watch Intel® System Debugger Intel® Cluster Checker Migration SYCLomatic Intel® DPC++ Compatibility Tool Model Zoo Hypervisors / Orchestration KVM 5.17 Hyper-V™ 1H'22 ESXi* vSphere* 8.0 Kubernetes* Red Hat OpenShift* OS & Kernel Microsoft Windows Server* 2022 Linux* 5.16 Red Hat 8.6 Ubuntu* 22.04 SUSE Linux Enterprise Server* 15 SP4 Hardware 4th & 5th Gen Intel® Xeon® Scalable Processor Intel® Xeon® CPU Max Series Intel® Data Center GPU Max Series
--------------------------	--

The CCL library provides multiple commands for creating, managing, and using CCL communicators.¹²² The CCL library includes software objects that are used to manage the communication between the domains / nodes accelerators in the 4th Gen Xeon Scalable Processor.¹²³ The CCL library is responsible for distributing the workload across the domains / nodes and for collecting the results from the domains / nodes.

One of more of the cluster node modules is responsible for accepting and communicating at least some of the user instructions using the communications network interface for the cluster node modules to communicate with each other. The communications network is then used to communicate at least some of the user instructions to one or more of the kernels.

¹²² Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”).

¹²³ <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-9/device-communication.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products
17. The computer cluster of claim 15, wherein the communications network interface comprises at least one of: a local bus, a local area network (LAN), a wide area network (WAN), a storage area network (SAN), a TCP/IP interface over Ethernet, or a wireless network interface.	Each of the Accused Xeon Products comprises a computer cluster that includes a mechanism such as a communication network interface that comprises at least one of a local bus, a local area network (LAN), a wide area network (WAN), a storage area network (SAN), a TCP/IP interface over Ethernet, or a wireless network interface. For example, as depicted below, the 4th Gen Xeon Scalable Processor is configured to provide all-to-all communication using Sub-NUMA clustering 4 (SNC-4) affinity mode using four domains of sub-NUMA clusters / nodes.¹²⁴  As shown in the above figure, the 4th Gen Xeon Scalable Processor includes, among other things, a flex bus that is capable of providing a PCI express bus as well as CXL buses with up to 32 Gigabits per second.¹²⁵

¹²⁴ “Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family,” <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

¹²⁵ “Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family,” <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products
	oneCCL software, which is part of Intel's oneAPI Collective Communications Library.¹²⁷ More particularly, Intel's Accused Xeon Products are configured to be supported by Intel's oneAPI software suite, which, for example, "maximize[s] application performance by activating the advanced capabilities of 4th and 5th gen Intel® Xeon® processors...."¹²⁸ The oneCCL library is one of the 10 core elements of Intel's oneAPI software suite.¹²⁹ Intel's oneCCL package comprises the oneCCL Software Development Kit (SDK) and the Intel® MPI Library Runtime components.¹³⁰
19. The computer cluster of claim 1, wherein the cluster node modules communicate with one another to act as a cluster.	Each of the Accused Xeon Products comprises cluster node modules that communicate with one another to act as a cluster. For example, as depicted below, the 4th Gen Xeon Scalable Processor comprises four nodes (described as domains in the figure below), each of which is configured to access memory including, for example, LLC and DDR memory.¹³¹

¹²⁷ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf>.

¹²⁸ "Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors," <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87c1qd>; "oneAPI Specification, Release 1.3-rev-1," pp. 388-389 <https://spec.oneapi.io/versions/latest/oneAPI-spec.pdf>; "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/overview.html>, <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/communicator.html>.

¹²⁹ <https://www.intel.com/content/www/us/en/developer/articles/technical/oneapi-what-is-it.html>.

¹³⁰ See "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/overview.html>.

¹³¹ "Technical Overview Of The 4th Gen Intel® Xeon® Scalable processor family," <https://www.intel.com/content/www/us/en/developer/articles/technical/fourth-generation-xeon-scalable-family-overview.html>.

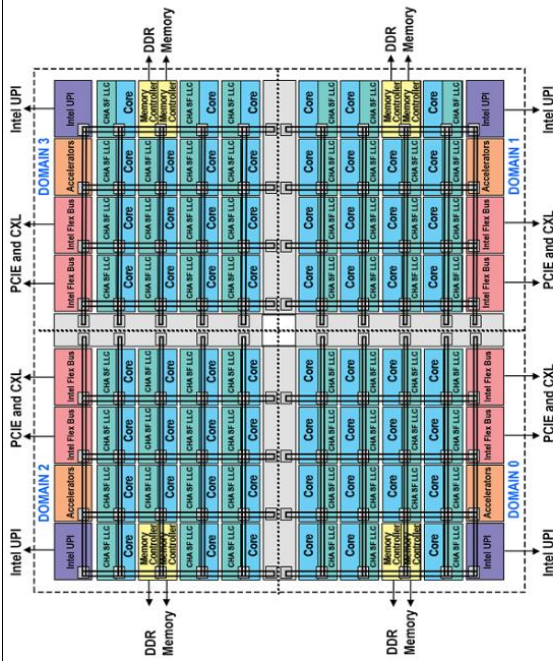


Figure 3 – Block Diagram Representing Domains Of sub-NUMA With Four Clusters

Each Accused Xeon Product comprises a plurality of nodes, wherein each of the plurality of nodes comprises at least one computer-readable medium in communication with at least one of the first processor, the second processor, or the third processor. For example, the 4th Gen Xeon Scalable Processor, as shown above, is configured to access and be in communication with a non-transitory computer-readable medium that includes, for example, local memory (e.g., LLC), and external memory (e.g., DDR memory). The 4th Gen Xeon Scalable Processor is, therefore, configured (see Memory Controller) to have memory accessible by and in communication with the processor.

Intel's Accused Xeon Products are configured to be supported by Intel's oneAPI software suite, including Intel's oneCCL software.¹³² For example, Intel's 4th Gen Xeon Scalable Processor are configured to be supported by Intel's oneAPI software suite, which, for example, "maximize[s] application performance by activating the advanced capabilities of 4th and 5th gen Intel® Xeon®

¹³² Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> ("Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.").

U.S. Pat. No. 11,811,582	
	Accused Xeon Products processors....”¹³³ The oneCCL library is one of the 10 core elements of Intel’s oneAPI software suite.¹³⁴ Intel’s oneCCL package comprises the oneCCL Software Development Kit (SDK) and the Intel® MPI Library Runtime components.¹³⁵ “The Intel® oneAPI Math Kernel Library (oneMKL) is a computing math library of highly optimized and extensively parallelized routines for applications that require maximum performance. oneMKL contains the high-performance optimizations from the full Intel® Math Kernel Library for CPU architectures (with C/Fortran programming language interfaces) and adds to them a set of SYCL* interfaces for achieving performance on various CPU architectures and Intel Graphics Technology for certain key functionalities. oneMKL provides BLAS and LAPACK linear algebra routines, fast Fourier transforms, vectorized math functions, random number generation functions, and other functionality.”¹³⁶ The mathematical routines addressed in oneMKL include matrix operations using, for example, arrays of data elements. The figure below depicts Intel’s oneAPI software suite and its Collective Communications Library (oneCCL) and Math Kernel Library (oneMKL) in the exemplary context of the software suite for Intel’s 4th Gen Xeon Scalable Processors.¹³⁷ For example, the oneCCL library includes commands that cause the 4th Gen Xeon Scalable Processor to send commands among the cluster node modules to act as a cluster.¹³⁸

¹³³ “Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors,” <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>; “oneAPI Specification, Release 1.3-rev-1,” pp. 388-389 <https://spec.oneapi.io/versions/latest/oneAPI-spec.pdf>; “Intel® oneAPI Collective Communications Library Developer Guide and Reference,” <https://www.intel.com/content/www/us/en/docs/oneapi/oneccl/developer-guide-reference/2021-12/overview.html>, <https://www.intel.com/content/www/us/en/developer/articles/technical/oneapi-what-is-it.html>.

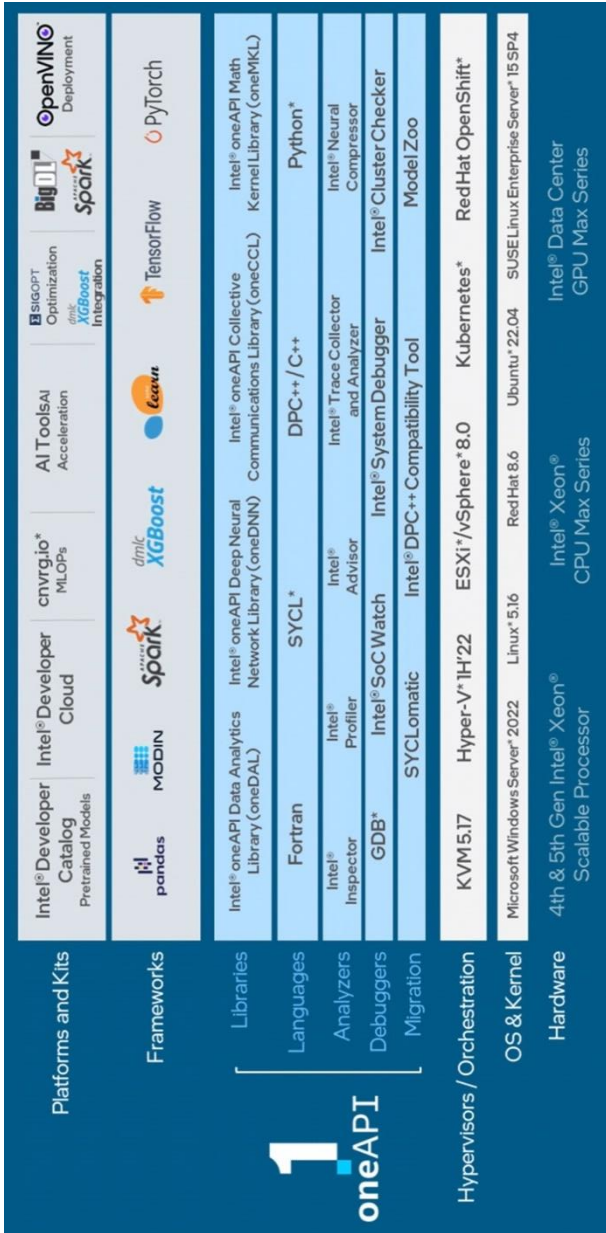
¹³⁴ <https://www.intel.com/content/www/us/en/docs/oneapi/oneccl/developer-guide-reference/2021-12/overview.html>.

¹³⁵ See “Intel® oneAPI Collective Communications Library Developer Guide and Reference,” <https://www.intel.com/content/www/us/en/docs/oneapi/oneccl/developer-guide-reference/2021-12/overview.html>.

¹³⁶ “Intel® oneAPI Programming Guide,” <https://www.intel.com/content/www/us/en/docs/oneapi/programming-guide/2024-1/intel-oneapi-math-kernel-library-onemkl.html>.

¹³⁷ “Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors,” <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>.

¹³⁸ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”).

U.S. Pat. No. 11,811,582	Accused Xeon Products  The diagram illustrates the Intel oneAPI architecture, showing a central '1 oneAPI' core. Above the core are 'Platforms and Kits' (Intel® Developer Catalog, Intel® Developer Cloud, AI Tools/ai Acceleration, cnvrg.io* MLOPs, SIGOPT Optimization, BigDL, Spark, OpenVINO Deployment) and 'Frameworks' (pandas, MODIN, Spark, XGBoost, TensorFlow, PyTorch). Below the core are 'Libraries' (Intel® oneAPI Data Analytics Library, Intel® oneAPI Deep Neural Network Library, Intel® oneAPI Math Kernel Library), 'Languages' (Fortran, SYCL*, DPC++/C++), 'Analyzers' (Intel® Inspector, Intel® Trace Collector and Analyzer), 'Debuggers' (GDB*, Intel® System Debugger), and 'Migration' (SYCLomatic, Intel® DPC++ Compatibility Tool, Model Zoo). At the bottom are 'Hypervisors / Orchestration' (KVM 5.17, Hyper-V* 1H*22, ESXi* / vSphere* 8.0, Kubernetes*), 'OS & Kernel' (Microsoft Windows Server* 2022, Linux* 5.16, Red Hat 8.6, Ubuntu* 22.04, SUSE Linux Enterprise Server* 15 SP4), and 'Hardware' (4th & 5th Gen Intel® Xeon® Scalable Processor, Intel® Xeon® CPU Max Series, Intel® Data Center GPU Max Series).
20. The computer cluster of claim 1, wherein one or more of the cluster node	The CCL library provides multiple commands for creating, managing, and using CCL communicators.¹³⁹ The CCL library includes software objects that are used to manage the communication between the domains / nodes accelerators in the 4th Gen Xeon Scalable Processor.¹⁴⁰ The CCL library is responsible for receiving commands from a user interface. The source code for the CCL library includes the details for using oneCCL commands to initialize communications on a 4th Gen Xeon Scalable Processor. Each of the Accused Xeon Products comprises computer cluster wherein one or more of the cluster node modules are configured to receive a command to start a cluster initialization process for the computer cluster. For example, as depicted below, the 4th Gen Xeon Scalable Processor comprises

¹³⁹ Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf> (“Enables efficient implementations of collectives that are heavily used for neural network training, including all-gather, all-reduce, and reduce-scatter.”).

¹⁴⁰ <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-9/device-communication.html>.

U.S. Pat. No. 11,811,582	Accused Xeon Products
	Collective Communications Library.¹⁴² More particularly, Intel's Accused Xeon Products are configured to be supported by Intel's oneAPI software suite, which, for example, "maximize[s] application performance by activating the advanced capabilities of 4th and 5th gen Intel® Xeon® processors...."¹⁴³ The oneCCL library is one of the 10 core elements of Intel's oneAPI software suite.¹⁴⁴ Intel's oneCCL package comprises the oneCCL Software Development Kit (SDK) and the Intel® MPI Library Runtime components.¹⁴⁵

¹⁴² Intel oneAPI Collective Communications Library, <https://www.intel.com/content/www/us/en/developer/tools/oneapi/oneccl.html#gs.ejo7gf>.
¹⁴³ "Software for 4th and 5th Gen Intel® Xeon® and Intel® Max Series Processors," <https://www.intel.com/content/www/us/en/developer/platform/4gen-xeon-max-series-cpu.html#gs.87ciqd>; "oneAPI Specification, Release 1.3-rev-1," pp. 388-389 <https://spec.oneapi.io/versions/latest/oneAPI-spec.pdf>; "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/overview.html>, <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/communicator.html>.
¹⁴⁴ <https://www.intel.com/content/www/us/en/developer/articles/technical/oneapi-what-is-it.html>.
¹⁴⁵ See "Intel® oneAPI Collective Communications Library Developer Guide and Reference," <https://www.intel.com/content/www/us/en/docs/oneccl/developer-guide-reference/2021-12/overview.html>.