



- (51) **International Patent Classification:**
G06N 3/063 (2006.01) *G06N 3/04* (2006.01)
- (21) **International Application Number:**
PCT/US2015/058397
- (22) **International Filing Date:**
30 October 2015 (30.10.2015)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (71) **Applicant:** HEWLETT PACKARD ENTERPRISE DEVELOPMENT LP [US/US]; 11445 Compaq Center Drive W., Houston, Texas 77070 (US).
- (72) **Inventors:** MURALIMANO HAR, Naveen; 1501 Page Mill Rd., Palo Alto, California 94304-1100 (US). STRACHAN, John Paul; 1501 Page Mill Rd., Palo Alto, California 94304-1100 (US). BALASUBRAMONIAN, Rajeev; 1501 Page Mill Rd., Palo Alto, California 94304-1100 (US). WILLIAMS, R. Stanley; 1501 Page Mill Rd., Palo Alto, California 94304-1100 (US).
- (74) **Agents:** PAGAR, Preetam B. et al.; Hewlett Packard Enterprise, 3404 E. Harmony Road, Mail Stop 79, Fort Collins, Colorado 80528 (US).
- (81) **Designated States** (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM,

AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) **Designated States** (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- as to the identity of the inventor (Rule 4.17(i))
- as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))

Published:

- with international search report (Art. 21(3))

(54) **Title:** HYBRID SYNAPTIC ARCHITECTURE BASED NEURAL NETWORK

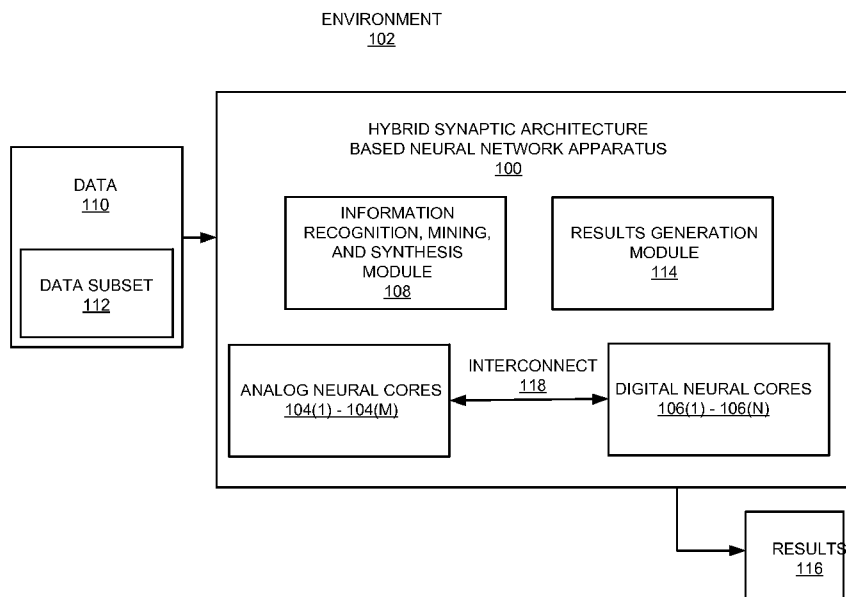


FIG. 2

(57) **Abstract:** According to an example, a hybrid synaptic architecture based neural network may be implemented by determining, from input data, information that is to be recognized, mined, and/or synthesized by a plurality of analog neural cores. Further, the hybrid synaptic architecture based neural network may be implemented by determining, based on the information, selected ones of the plurality of analog neural cores that are to be actuated to identify a data subset of the input data to generate, based on the analysis of the data subset, results of the recognition, mining, and/or synthesizing of the information.

HYBRID SYNAPTIC ARCHITECTURE BASED NEURAL NETWORK

BACKGROUND

[0001] With respect to machine learning and cognitive science, a neural network is a statistical learning model that is used to estimate or approximate functions that may depend on a large number of inputs. In this regard, artificial neural networks may include systems of interconnected neurons which exchange messages between each other. The interconnections may include numeric weights that may be tuned based on experience, which makes neural networks adaptive to inputs and capable of learning. For example, a neural network for character recognition may be defined by a set of input neurons which may be activated by pixels of an input image. The activations of the input neurons are then passed on to other neurons after the input neurons are weighted and transformed by a function. This process may be repeated until an output neuron is activated, whereby the character that is read may be determined.

BRIEF DESCRIPTION OF DRAWINGS

[0002] Features of the present disclosure are illustrated by way of example and not limited in the following figure(s), in which like numerals indicate like elements, in which:

[0003] Figure 1 illustrates a layout of a hybrid synaptic architecture based neural network apparatus, according to an example of the present disclosure;

[0004] Figure 2 illustrates an environment for the hybrid synaptic architecture based neural network apparatus of Figure 1, according to an example of the present disclosure;

[0005] Figure 3 illustrates details of an analog neural core for the hybrid synaptic architecture based neural network apparatus of Figure 1, according to an example of the present disclosure;

[0006] Figure 4 illustrates details of a digital neural core for the hybrid synaptic architecture based neural network apparatus of Figure 1, according to an example of the present disclosure;

[0007] Figure 5 illustrates a flowchart of a method for implementing the hybrid synaptic architecture based neural network apparatus of Figure 1, according to an example of the present disclosure;

[0008] Figure 6 illustrates another flowchart of a method for implementing the hybrid synaptic architecture based neural network apparatus of Figure 1, according to an example of the present disclosure;

[0009] Figure 7 illustrates another flowchart of a method for implementing the hybrid synaptic architecture based neural network apparatus of Figure 1, according to an example of the present disclosure;

[0010] Figure 8 illustrates a computer system, according to an example of the present disclosure; and

[0011] Figure 9 illustrates another computer system, according to an example

of the present disclosure.

DETAILED DESCRIPTION

[0012] For simplicity and illustrative purposes, the present disclosure is described by referring mainly to examples. In the following description, numerous specific details are set forth in order to provide a thorough understanding of the present disclosure. It will be readily apparent however, that the present disclosure may be practiced without limitation to these specific details. In other instances, some methods and structures have not been described in detail so as not to unnecessarily obscure the present disclosure.

[0013] Throughout the present disclosure, the terms "a" and "an" are intended to denote at least one of a particular element. As used herein, the term "includes" means includes but not limited to, the term "including" means including but not limited to. The term "based on" means based at least in part on.

[0014] With respect to neural networks, neuromorphic computing is described as the use of very-large-scale integration (VLSI) systems including electronic analog circuits to mimic neuro-biological architectures present in the nervous system. Neuromorphic computing may be used with recognition, mining, and synthesis (RMS) applications. Recognition may be described as the examination of data to determine what the data represents. Mining may be described as the search for particular types of models determined from the recognized data. Further, synthesis may be described as the generation of a potential model where a model does not previously exist. With respect to RMS applications and other types of applications, specialized neural chips, which may be several orders of magnitude more efficient than central processing unit (CPU) or graphics processor unit (GPU) computations, may provide for the scaling of neural networks to simulate billions of neurons and mine vast amounts of data.

[0015] With respect to machine readable instructions to control neural networks, neuromorphic memory arrays may be used for RMS applications and other types of applications by performing computations directly in such memory arrays. The type of memory employed in neuromorphic memory arrays may either

be analog or digital. In this regard, the choice of the type of memory may impact characteristics such as accuracy, energy, performance, etc., of the associated neuromorphic system.

[0016] In this regard, a hybrid synaptic architecture based neural network apparatus, and a method for implementing the hybrid synaptic architecture based neural network are disclosed herein. The apparatus and method disclosed herein may use a combination of analog and digital memory arrays to reduce energy consumption compared, for example, to state-of-the-art neuromorphic systems. According to examples, the apparatus and method disclosed herein may be used with memristor based neural systems, and/or use a memristor's high on/off ratio and tradeoffs between write latency and accuracy to implement neural cores with varying levels of accuracy and energy consumption. The apparatus and method disclosed herein may achieve a high degree of power efficiency, and may simulate an order of magnitude more neurons per chip compared to a fully digital design. For example, since more neurons per unit area may be simulated for an analog implementation, for the apparatus and method disclosed herein, a higher number of neurons per chip (e.g., a higher number of overall neural cores including analog neural cores and digital neural cores) may be simulated per chip compared to a fully digital design.

[0017] Figure 1 illustrates a layout of a hybrid synaptic architecture based neural network apparatus (hereinafter also referred to as "apparatus 100"), according to an example of the present disclosure. Figure 2 illustrates an environment 102 of the apparatus 100, according to an example of the present disclosure.

[0018] Referring to Figures 1 and 2, the apparatus 100 may include a plurality of analog neural cores 104, and a plurality of digital neural cores 106. The analog neural cores 104 may be designated as analog neural cores 104(1) – 104(M). Further, the digital neural cores 106 may be designated as digital neural cores 106(1) – 106(N).

[0019] An information recognition, mining, and synthesis module 108 may determine information that is to be recognized, mined, and/or synthesized from input data 110 (e.g., see Figure 2). The information recognition, mining, and synthesis module 108 may determine, based on the information, selected ones of the plurality of analog neural cores 104 that are to be actuated to identify a data subset 112 (e.g., see Figure 2) of the input data 110. The information recognition, mining, and synthesis module 108 may determine, based on the data subset 112, selected ones of the plurality of digital neural cores 106 that are to be actuated to analyze the data subset 112.

[0020] A results generation module 114 may generate, based on the analysis of the data subset 112, results 116 (e.g., see Figure 2) of the recognition, mining, and/or synthesizing of the information.

[0021] An interconnect 118 between the analog neural cores 104 and the digital neural cores 106 may be implemented by a CPU, a GPU, by a state machine, or other such techniques. For example, the state machine may detect an output of the analog neural cores 104 and direct the output to the digital neural cores 106. In this regard, the CPU, the GPU, the state machine, or other such techniques may be controlled and/or implemented as a part of the information recognition, mining, and synthesis module 108.

[0022] The modules and other elements of the apparatus 100 may be machine readable instructions stored on a non-transitory computer readable medium. In this regard, the apparatus 100 may include or be a non-transitory computer readable medium. In addition, or alternatively, the modules and other elements of the apparatus 100 may be hardware or a combination of machine readable instructions and hardware.

[0023] Figure 3 illustrates details of an analog neural core 104 for the apparatus 100, according to an example of the present disclosure.

[0024] Referring to Figure 3, the analog neural core 104 may include a plurality

of memristors to receive the input data 110, multiply the input data 110 by associated weights, and generate output data. The output data may represent the data subset 112 of the input data 110 or data that forms the data subset 112 of the input data 110.

[0025] For example, as shown in Figure 3, the analog neural core 104 may include a plurality of inputs x_i (e.g., x_1 , x_2 , x_3 , etc.) that are fed into an analog memory array 300 (e.g., a memristor array). The inputs x_i may represent, for example, pixels of a video stream, and generally any type of data that is to be analyzed (e.g., for recognition, mining, and/or synthesis) by the apparatus 100. The analog memory array 300 may include a plurality of weighted memristors including weights w_{ij} . For the example of x_i that represents pixels of a video stream, w_{ij} may represent a kernel that is used to convert an image to black/white, sharpen the image, etc. Each of the inputs x_i may be multiplied (e.g., to perform convolution by matrix multiplication) by a respective weight w_{ij} , and the resulting values may be added (i.e., summed) at 302 to generate output values y_j (e.g., y_1 , y_2 , etc.). Thus, *the output values y_j may be determined as $y_j = \sum_i w_{ij} * x_i$.* The accuracy of the values of the weights w_{ij} may directly correlate to the accuracy of the analog neural core 104. For example, an actual value of w_{ij} for the analog memory array 300 may be measured as $w_{ij} + \Delta$, compared to an ideal value. For the example of x_i that represents pixels of a video stream, the output values y_j may represent, for example, maximum values, a subset of values, etc., related to an image.

[0026] With respect to extraction of features from the data 110, the output values y_j may be compared to known values from a database to determine a feature that is represented by the output values y_j . For example, the information recognition, mining, and synthesis module 108 may compare the output values y_j to known values from a database to determine information (e.g., a feature) that is represented by the output values y_j . In this regard, the information recognition, mining, and synthesis module 108 may perform recognition, for example, by

examining the data 110 to determine what the data represents, mining to search for particular types of models determined from the recognized data, and synthesis to generate a potential model where a model does not previously exist.

[0027] For the analog neural core 104, instead of the use of the memristor array based analog memory array 300, the analog memory array 300 may be implemented by flash memory (used in an analog mode), and other types of memory.

[0028] Figure 4 illustrates details of a digital neural core 106 for the apparatus 100, according to an example of the present disclosure.

[0029] Referring to Figure 4, the digital neural core 106 may include a memory array 400 to receive input data, and a plurality of multiply-add-accumulate units 402 to process the input data received by the memory array 400 and associated weights from the memory array 400 to generate output data. For the interconnected example of Figure 1, the digital neural core 106 may include the memory array 400 to receive the output data of an associated analog neural core of the plurality of analog neural cores 104, and a plurality of multiply-add-accumulate units 402 to process the output data and associated weights from the memory array 400 to generate further output data.

[0030] For example, as shown in Figure 4, the digital neural core 106 may include the memory array 400 (i.e., a grid of memory cells) that models neurons and axons (e.g., N neurons, M axons). The memory array 400 may be connected to the set of multiply-add-accumulate units 402 to determine neural outputs. Each digital neural core 106 may include an input buffer to receive inputs x_i (e.g., x_1, x_2, x_3 , etc.). The positions of the inputs x_i (e.g., i) may be forwarded to a row decoder 404, where the positions i are used to determine an appropriate weight w_{ij} . The determined weight w_{ij} may be multiplied with the inputs x_i at each associated multiply-add-accumulate unit, and output to an output buffer as y_j (e.g., y_1, y_2 , etc.). With respect to the digital neural core 106, the overall latency of a calculation may be a function of the number of rows of the data that is loaded into the memory

array 400. A control unit 406 may control operation of the memory array 400 with respect to programming of the appropriate w_{ij} (e.g., in a memory mode of the digital neural core 106), control operation of the row decoder 404 with respect to selection of the appropriate w_{ij} , and control operation of the multiply-add-accumulate units 402 (e.g., in a compute mode of the digital neural core 106).

[0031] The output y_j (e.g., y_1 , y_2 , etc.) of the multiply-add-accumulate units 402 may be routed to other neural cores (e.g., other analog and/or neural cores), where, for a digital neural core, the output is fed as input to the row decoder 404 and the multiply-add-accumulate units 402 of the other neural cores.

[0032] For the digital neural core 106, the digital memory array 400 may be implemented by use of a variety of technologies. For example, the digital memory array 400 may be implemented by using memristor based memory, CPU based memory, GPU based memory, a process in memory based solution, etc. For example, with respect to the digital memory array 400, at first $w_{1,1}$ and a corresponding value for x_1 may be read, these values may be multiplied at the multiply-add-accumulate units 402, and so forth for further values of w_{ij} and x_i . In this regard, these operations may be performed by the digital memory array 400 implemented by using memristor based memory, CPU based memory, GPU based memory, a process in memory based solution, etc.

[0033] As disclosed herein, since the apparatus 100 may use a combination of analog neural cores 104 that include analog memory arrays and digital neural cores 106 that include digital memory arrays, the corresponding peripheral circuits may also use analog or digital functional units, respectively.

[0034] With respect to the use of the analog neural cores 104 and the digital neural cores 106 as disclosed herein, the choice of the neural core may impact the operating power and accuracy of the neural network. For example, a neural core using an analog memory array may consume an order of magnitude less energy compared to a neural core using a digital memory array. However, in certain instances, the use of the analog memory array 300 may degrade the accuracy of

the analog neural core 104. For example, if the value of the weights w_{ij} are inaccurate, these inaccuracies may further degrade the accuracy of the analog neural core 104.

[0035] The apparatus 100 may therefore selectively actuate a plurality of analog neural cores 104 to increase energy efficiency of the apparatus 100 or a component that utilizes the apparatus 100 and/or the plurality of analog neural cores 104, and selectively actuate a plurality of digital neural cores 106 to increase accuracy of the apparatus 100 or a component that utilizes the apparatus 100 and/or the plurality of digital neural cores 106. In this regard, according to examples, the apparatus 100 may include or be implemented in a component that includes a hybrid analog-digital neural chip. The hybrid analog-digital neural chip may be used to perform coarse level analysis on the data 110 (e.g., all or a relatively high amount of the data 110) using the analog neural cores 104. Based on the results of the coarse level analysis, the data subset 112 (i.e., a subset of the data 110) may be identified for fine grained analysis. For example, the digital neural cores 106 may be used to perform fine grained analysis on the data subset 112. In this regard, the digital neural cores 106 may be used to perform fine grained mining of the data subset 112. The data subset 112 may represent a region of interest related to an object of interest in the data 110.

[0036] According to examples, with respect to determining, based on the information, selected ones of the plurality of analog neural cores 104 that are to be actuated to identify the data subset 112 of the input data 110, the information recognition, mining, and synthesis module 108 may determine, based on the information, selected ones of the plurality of analog neural cores 104 that are to be actuated to identify the data subset 112 of the input data 110 to reduce an energy consumption of the apparatus 100.

[0037] According to examples, with respect to determining, based on the information, selected ones of the plurality of analog neural cores 104 that are to be actuated to identify the data subset 112 of the input data 110, the information

recognition, mining, and synthesis module 108 may determine, based on the information, selected ones of the plurality of analog neural cores 104 that are to be actuated to identify the data subset 112 of the input data 110 to meet an accuracy specification of the apparatus 100.

[0038] According to examples, with respect to accuracy of the apparatus 100, the information recognition, mining, and synthesis module 108 may increase a number of the selected ones of the plurality of digital neural cores 106 that are to be actuated to analyze the data subset 112 to increase an accuracy of the recognition, mining, and/or synthesizing of the information.

[0039] According to examples, with respect to energy consumption of the apparatus 100, the information recognition, mining, and synthesis module 108 may reduce an energy consumption of the apparatus 100 by decreasing a number of the selected ones of the plurality of digital neural cores 106 that are to be actuated to analyze the data subset 112.

[0040] The apparatus 100 may also selectively actuate a plurality of analog neural cores 104 to reduce the amount of data that is to be buffered for the digital neural cores 106. For example, instead of buffering all of the data for analysis by digital neural cores 106, the buffered data may be limited to the data subset 112 to thus increase energy efficiency of the apparatus 100 or a component that utilizes the apparatus 100. For example, with respect to reducing an amount of data received by the digital neural core input buffers, for an analog neural core input buffer associated with each of the analog neural cores 104 to receive the input data 110 for forwarding to the plurality of memristors, and a digital neural core input buffer associated with each of the digital neural cores 106 to receive the output data from the analog neural cores 104, the information recognition, mining, and synthesis module 108 may reduce an amount of data received by the digital neural core input buffers based on elimination of all but the data subset 112 that is to be analyzed by the selected ones of the plurality of digital neural cores 106.

[0041] The apparatus 100 may also selectively actuate the plurality of analog

neural cores 104 to increase performance aspects such as an amount of time needed to generate results. For example, based on the faster performance of the analog neural cores 104, the amount of time needed to generate results may be reduced compared to analysis of all of the data 110 by the digital neural cores 106.

[0042] According to examples, for the data 110 that includes a streaming video, for the apparatus 100 that operates as or in conjunction with an image recognition system, in order to identify certain aspects of the streaming video (e.g., a moving car, a number plate, or static objects such as buildings, building numbers, etc.), a hybrid analog-digital neural chip (that includes the analog neural cores 104 and the digital neural cores 106) may be used to perform coarse level analysis on the data 110 using the analog neural cores 104 to identify moving features that likely resemble a car. Based on the results of the coarse level analysis, the data subset 112 (i.e., a subset of the data 110 of moving features that likely resemble a car) may be identified for fine grained analysis. For example, the digital neural cores 106 may be used to perform fine grained analysis on the data subset 112 of moving features that likely resemble a car (e.g., a segment of a frame including the moving features that likely resemble a car). In this regard, the digital neural cores 106 may be used to perform fine grained mining of the data subset 112 of moving features that likely resemble a car. The fine grained analysis performed the digital neural cores 106 may be used to identify components such as number plates, face recognition of a person inside the car, etc. In this regard, as the input set to the digital neural cores 106 is smaller than the original streaming video, a number of the digital neural cores 106 that are utilized may be reduced, compared to use of the digital neural cores 106 for the entire analysis of the original streaming video.

[0043] The apparatus 100 may also include the selective feeding of results from the analog neural cores 104 to the digital neural cores 106 for processing. For example, if the output y_1 for the example of Figure 3 is determined to be an output corresponding to the data subset 112, that particular output may be fed to the digital neural cores 106 for processing, with the other output y_2 being discarded.

[0044] Figures 5-7 respectively illustrate flowcharts of methods 500, 600, and 700 for implementation of a hybrid synaptic architecture based neural network, corresponding to the example of the hybrid synaptic architecture based neural network apparatus 100 whose construction is described in detail above. The methods 500, 600, and 700 may be implemented on the hybrid synaptic architecture based neural network apparatus 100 with reference to Figures 1-4 by way of example and not limitation. The methods 500, 600, and 700 may be practiced in other apparatus. The example of Figure 6 may represent a method that is implemented on the apparatus 100 that includes a plurality of analog neural cores, a plurality of digital neural cores, a processor 902 (see Figure 9), and a memory 906 (see Figure 9) storing machine readable instructions that when executed by the processor cause the processor to perform the method 600. The example of Figure 7 may represent a non-transitory computer readable medium having stored thereon machine readable instructions to implement a hybrid synaptic architecture based neural network, the machine readable instructions, when executed, cause a processor (e.g., the processor 902 of Figure 9) to perform the method 700.

[0045] Referring to Figure 5, for the method 500, at block 502, the method may include determining, from input data 110, information that is to be recognized, mined, and/or synthesized by a plurality of analog neural cores 104 and a central processing unit (CPU) and/or a graphics processor unit (GPU).

[0046] At block 504, the method may include determining, based on the information, selected ones of the plurality of analog neural cores 104 that are to be actuated to identify a data subset 112 of the input data 110.

[0047] At block 506, the method may include discarding, based on the identification of the data subset 112, remaining data, other than the data subset 112, from further analysis.

[0048] At block 508, the method may include using, by a processor (e.g., the processor 902), the CPU and/or the GPU to analyze the data subset 112 (i.e., to

perform the digital neural processing) to generate, based on the analysis of the data subset 112, results 116 of the recognition, mining, and/or synthesizing of the information.

[0049] Referring to Figure 6, for the method 600, at block 602, the method may include determining information that is to be recognized, mined, and/or synthesized from input data 110.

[0050] At block 604, the method may include determining, based on the information, selected ones of the plurality of analog neural cores 104 that are to be actuated to identify a data subset 112 of the input data 110.

[0051] At block 606, the method may include determining, based on the data subset 112, selected ones of the plurality of digital neural cores 106 that are to be actuated to analyze the data subset 112.

[0052] At block 608, the method may include generating, based on the analysis of the data subset 112, results 116 of the recognition, mining, and/or synthesizing of the information.

[0053] Referring to Figure 7, for the method 700, at block 702, the method may include determining, from input data 110, information that is to be recognized, mined, and/or synthesized by a plurality of analog neural cores 104 and a plurality of digital neural cores 106.

[0054] At block 704, the method may include determining an energy efficiency parameter and/or an accuracy parameter related to the plurality of analog neural cores 104 and the plurality of digital neural cores 106. The energy efficiency parameter may represent, for example, an amount (or percentage) of energy efficiency that is to be implement for the apparatus 100. For example, a higher energy efficiency parameter may be determined to utilize a higher number of analog neural cores 104 compared to a lower energy efficiency parameter. The accuracy parameter may represent, for example, an amount (or percentage) of accuracy that is to be implement for the apparatus 100. For example, a higher

accuracy parameter may be selected to utilize a higher number of digital neural cores 106 compared to a lower energy efficiency parameter.

[0055] At block 706, the method may include determining, based on the information and the energy efficiency parameter and/or the accuracy parameter, selected ones of the plurality of analog neural cores 104 that are to be actuated to identify a data subset 112 of the input data 110.

[0056] At block 708, the method may include determining, based on the data subset 112, selected ones of the plurality of digital neural cores 106 that are to be actuated to analyze the data subset 112 to generate, based on the analysis of the data subset 112, results 116 of the recognition, mining, and/or synthesizing of the information.

[0057] Figure 8 shows a computer system 800 that may be used with the examples described herein. The computer system 800 may include components that may be in a server or another computer system. The computer system 800 may be used as a platform for the apparatus 100. The computer system 800 may execute, by a processor (e.g., a single or multiple processors) or other hardware processing circuit, the methods, functions and other processes described herein. These methods, functions and other processes may be embodied as machine readable instructions stored on a computer readable medium, which may be non-transitory, such as hardware storage devices (e.g., RAM (random access memory), ROM (read only memory), EPROM (erasable, programmable ROM), EEPROM (electrically erasable, programmable ROM), hard drives, and flash memory).

[0058] The computer system 800 may include a processor 802 that may implement or execute machine readable instructions performing some or all of the methods, functions and other processes described herein. Commands and data from the processor 802 may be communicated over a communication bus 804. The computer system may also include a main memory 806, such as a random access memory (RAM), where the machine readable instructions and data for the processor 802 may reside during runtime, and a secondary data storage 808,

which may be non-volatile and stores machine readable instructions and data. The memory and data storage are examples of computer readable mediums. The memory 806 may include a hybrid synaptic architecture based neural network implementation module 820 including machine readable instructions residing in the memory 806 during runtime and executed by the processor 802. The hybrid synaptic architecture based neural network implementation module 820 may include the modules of the apparatus 100 shown in Figures 1 and 2.

[0059] The computer system 800 may include an I/O device 810, such as a keyboard, a mouse, a display, etc. The computer system may include a network interface 812 for connecting to a network which may be further connected to analog neural cores and digital neural cores as disclosed herein with reference to Figures 1 and 2. Other known electronic components may be added or substituted in the computer system.

[0060] Figure 9 shows another computer system 900 that may be used with the examples described herein. The computer system 900 may represent a generic platform that includes components that may be in a server or another computer system. The computer system 900 may be used as a platform for the apparatus 100. The computer system 900 may execute, by a processor (e.g., a single or multiple processors) or other hardware processing circuit, the methods, functions and other processes described herein. These methods, functions and other processes may be embodied as machine readable instructions stored on a computer readable medium, which may be non-transitory, such as hardware storage devices (e.g., RAM, ROM, EPROM, EEPROM, hard drives, and flash memory).

[0061] The computer system 900 may include a processor 902 that may implement or execute machine readable instructions performing some or all of the methods, functions and other processes described herein. Commands and data from the processor 902 may be communicated over a communication bus 904. The computer system may also include a main memory 906, such as a RAM,

where the machine readable instructions and data for the processor 902 may reside during runtime, and a secondary data storage 908, which may be non-volatile and stores machine readable instructions and data. The memory and data storage are examples of computer readable mediums. The memory 906 may include a hybrid synaptic architecture based neural network implementation module 920 including machine readable instructions residing in the memory 906 during runtime and executed by the processor 902. The hybrid synaptic architecture based neural network implementation module 920 may include the modules of the apparatus 100 shown in Figures 1 and 2.

[0062] The computer system 900 may include an I/O device 910, such as a keyboard, a mouse, a display, etc. The computer system may include a network interface 912 for connecting to a network. Other known electronic components may be added or substituted in the computer system.

[0063] What has been described and illustrated herein is an example along with some of its variations. The terms, descriptions and figures used herein are set forth by way of illustration only and are not meant as limitations. Many variations are possible within the spirit and scope of the subject matter, which is intended to be defined by the following claims -- and their equivalents -- in which all terms are meant in their broadest reasonable sense unless otherwise indicated.

What is claimed is:

1. A hybrid synaptic architecture based neural network apparatus comprising:

a plurality of analog neural cores;

a plurality of digital neural cores;

a processor; and

a memory storing machine readable instructions that when executed by the processor cause the processor to:

determine information that is to be at least one of recognized, mined, and synthesized from input data;

determine, based on the information, selected ones of the plurality of analog neural cores that are to be actuated to identify a data subset of the input data;

determine, based on the data subset, selected ones of the plurality of digital neural cores that are to be actuated to analyze the data subset; and

generate, based on the analysis of the data subset, results of the at least one of the recognition, mining, and synthesizing of the information.

2. The hybrid synaptic architecture based neural network apparatus according to claim 1, wherein each of the analog neural cores comprises:

a plurality of memristors to receive the input data, multiply the input data by associated weights, and generate output data, wherein the output data represents the data subset of the input data or data that forms the data subset of the input data.

3. The hybrid synaptic architecture based neural network apparatus according to claim 2, wherein each of the digital neural cores comprises:

a memory array to receive the output data of an associated analog neural core of the plurality of analog neural cores; and

a plurality of multiply-add-accumulate units to process the output data and associated weights from the memory array to generate further output data.

4. The hybrid synaptic architecture based neural network apparatus according to claim 1, wherein each of the digital neural cores comprises:

a memory array to receive input data; and

a plurality of multiply-add-accumulate units to process the input data received by the memory array and associated weights from the memory array to generate output data.

5. The hybrid synaptic architecture based neural network apparatus according to claim 3, further comprising:

an analog neural core input buffer associated with each of the analog neural cores to receive the input data for forwarding to the plurality of memristors; and

a digital neural core input buffer associated with each of the digital neural cores to receive the output data from the analog neural cores,

wherein the memory further comprises machine readable instructions that when executed by the processor further cause the processor to:

reduce an amount of data received by the digital neural core input buffers based on elimination of all but the data subset that is to be analyzed by the selected ones of the plurality of digital neural cores.

6. The hybrid synaptic architecture based neural network apparatus according to claim 1, wherein the machine readable instructions to determine, based on the information, selected ones of the plurality of analog neural cores that are to be actuated to identify the data subset of the input data, further comprise machine readable instructions that when executed by the processor further cause the processor to:

determine, based on the information, selected ones of the plurality of analog neural cores that are to be actuated to identify the data subset of the input data to reduce an energy consumption of the apparatus.

7. The hybrid synaptic architecture based neural network apparatus according to claim 1, wherein the machine readable instructions to determine, based on the information, selected ones of the plurality of analog neural cores that are to be actuated to identify the data subset of the input data, further comprise machine readable instructions that when executed by the processor further cause the processor to:

determine, based on the information, selected ones of the plurality of analog neural cores that are to be actuated to identify the data subset of the input data to meet an accuracy specification of the apparatus.

8. The hybrid synaptic architecture based neural network apparatus according to claim 1, wherein the memory further comprises machine readable instructions that when executed by the processor further cause the processor to:

increase a number of the selected ones of the plurality of digital neural cores that are to be actuated to analyze the data subset to increase an accuracy of the at least one of the recognition, mining, and synthesizing of the information.

9. The hybrid synaptic architecture based neural network apparatus according to claim 1, wherein the memory further comprises machine readable instructions that when executed by the processor further cause the processor to:

reduce an energy consumption of the apparatus by decreasing a number of the selected ones of the plurality of digital neural cores that are to be actuated to analyze the data subset.

10. A method for implementing a hybrid synaptic architecture based neural network, the method comprising:

determining, from input data, information that is to be at least one of recognized, mined, and synthesized by a plurality of analog neural cores and at least one of a central processing unit (CPU) and a graphics processor unit (GPU);

determining, based on the information, selected ones of the plurality of analog neural cores that are to be actuated to identify a data subset of the input data;

discarding, based on the identification of the data subset, remaining data, other than the data subset, from further analysis; and

using, by a processor, the at least one of the CPU and the GPU to analyze the data subset to generate, based on the analysis of the data subset, results of the at least one of the recognition, mining, and synthesizing of the information.

11. The method of claim 10, wherein determining, based on the information, selected ones of the plurality of analog neural cores that are to be actuated to identify the data subset of the input data, further comprises:

determining, based on the information, selected ones of the plurality of analog neural cores that are to be actuated to identify the data subset of the input data to reduce an energy consumption related to the recognition, mining, and synthesizing of the information.

12. The method of claim 10, wherein determining, based on the information, selected ones of the plurality of analog neural cores that are to be actuated to identify the data subset of the input data, further comprises:

determining, based on the information, selected ones of the plurality of analog neural cores that are to be actuated to identify the data subset of the input data to meet an accuracy specification related to the recognition, mining, and synthesizing of the information.

13. A non-transitory computer readable medium having stored thereon machine readable instructions to implement a hybrid synaptic architecture based neural network, the machine readable instructions, when executed, cause a processor to:

determine, from input data, information that is to be at least one of recognized, mined, and synthesized by a plurality of analog neural cores and a plurality of digital neural cores;

determine at least one of an energy efficiency parameter and an accuracy parameter related to the plurality of analog neural cores and the plurality of digital neural cores;

determine, based on the information and the at least one of the energy efficiency parameter and the accuracy parameter, selected ones of the plurality of analog neural cores that are to be actuated to identify a data subset of the input data; and

determine, based on the data subset, selected ones of the plurality of digital neural cores that are to be actuated to analyze the data subset to generate, based on the analysis of the data subset, results of the at least one of the recognition, mining, and synthesizing of the information.

14. The non-transitory computer readable medium according to claim 13, further

comprising machine readable instructions to:

increase a number of the selected ones of the plurality of digital neural cores that are to be actuated to analyze the data subset to increase an accuracy of the at least one of the recognition, mining, and synthesizing of the information.

15. The non-transitory computer readable medium according to claim 13, further comprising machine readable instructions to:

reduce an energy consumption related to the recognition, mining, and synthesizing of the information by decreasing a number of the selected ones of the plurality of digital neural cores that are to be actuated to analyze the data subset.

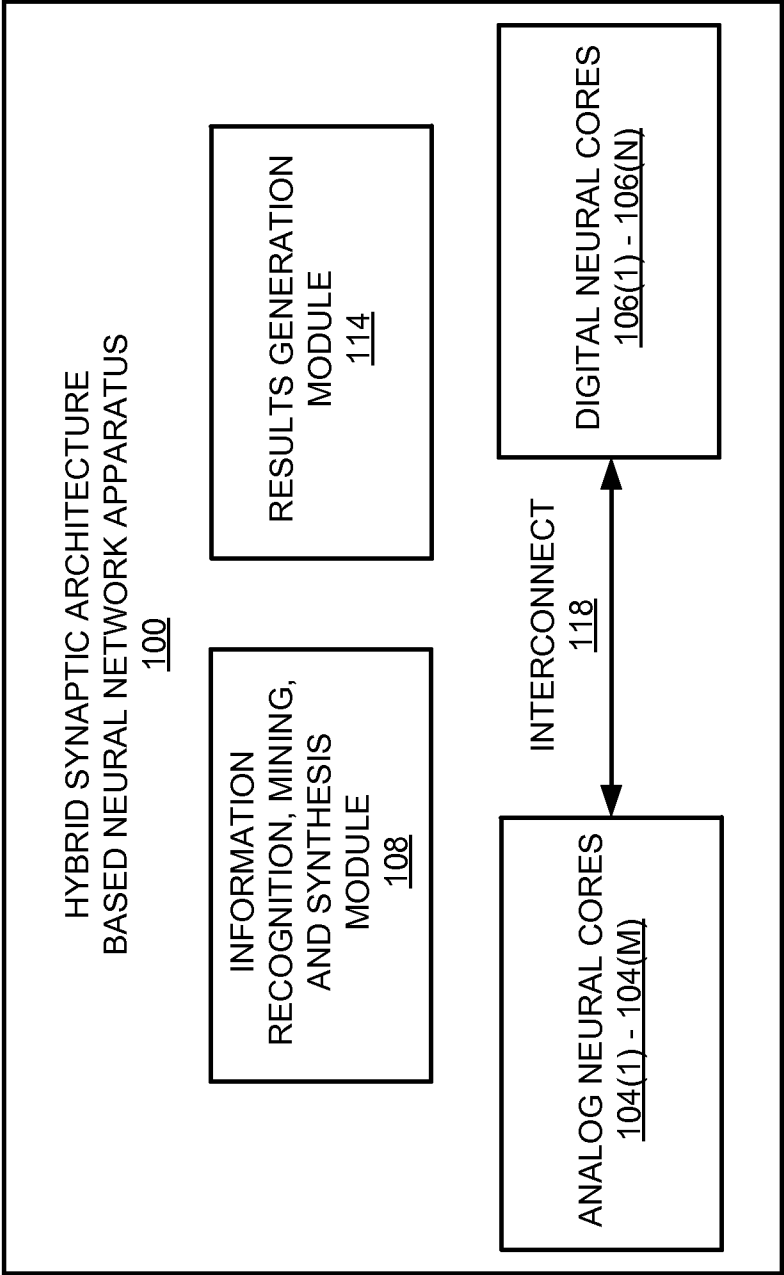


FIG. 1

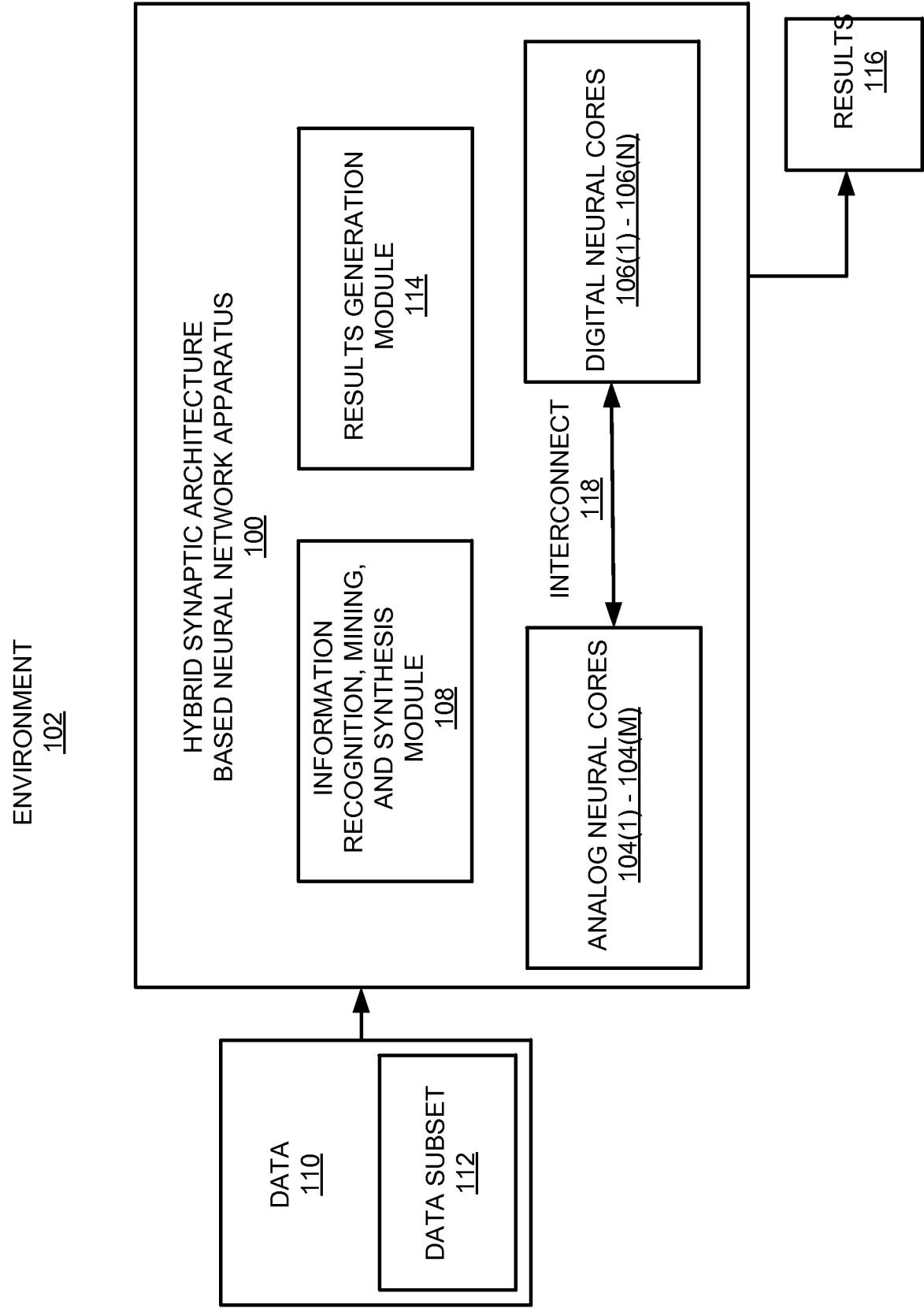
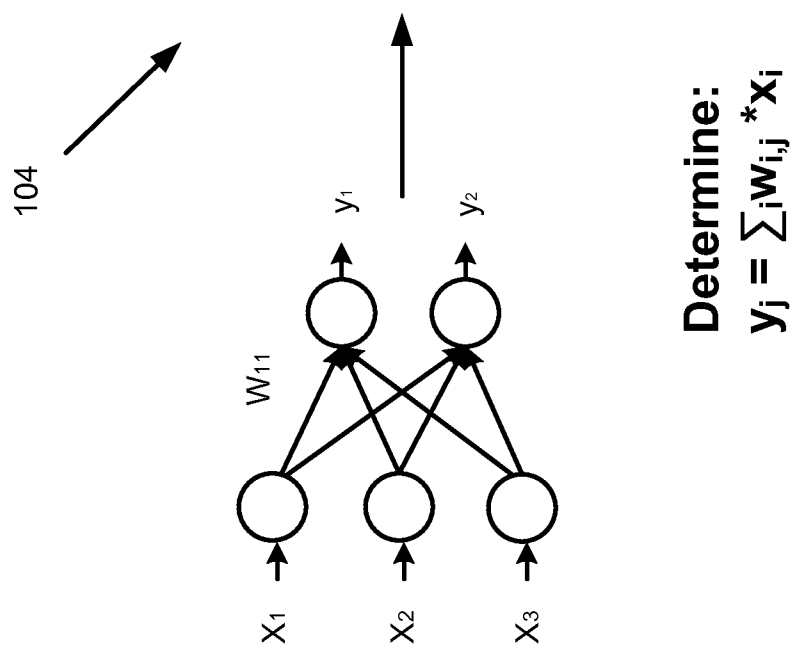
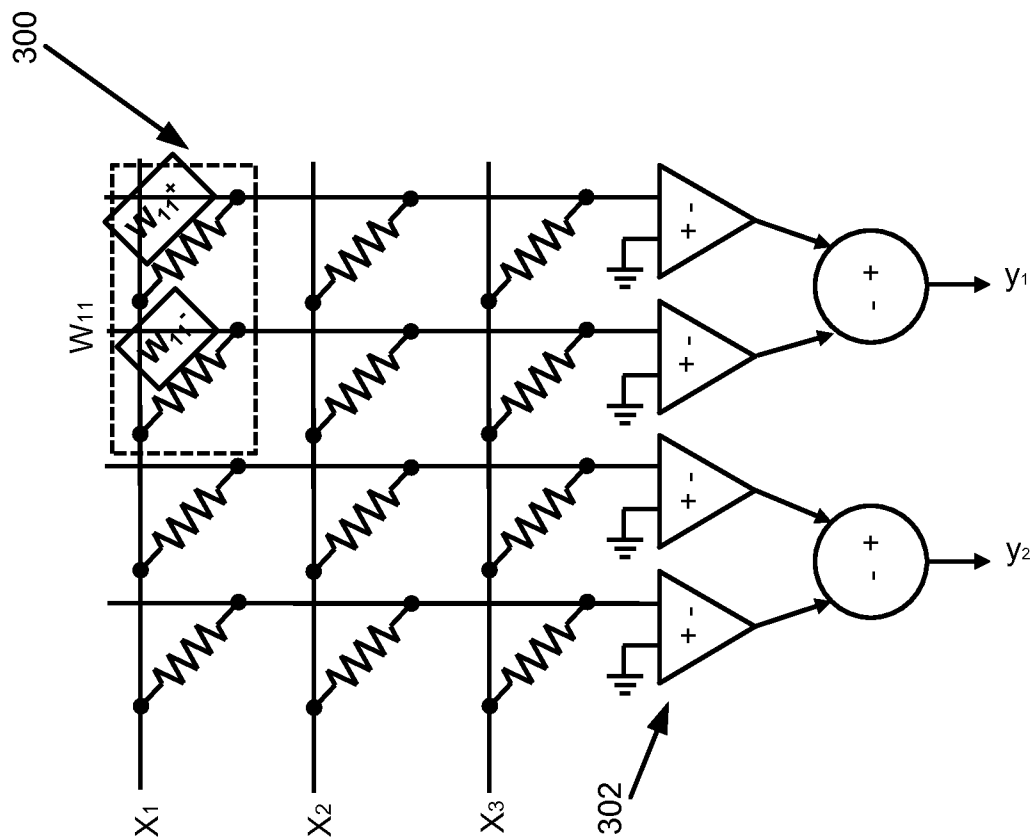


FIG. 2



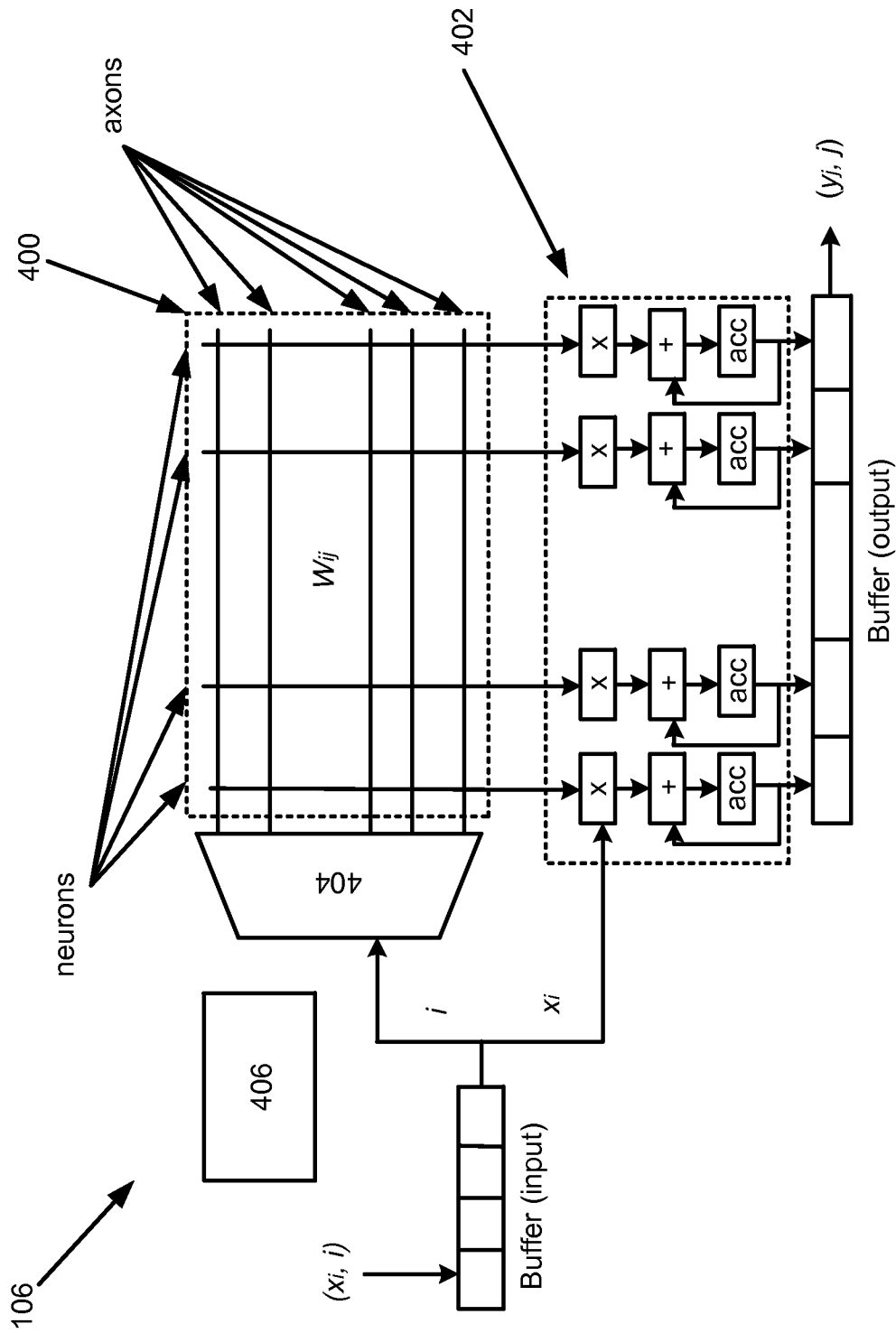
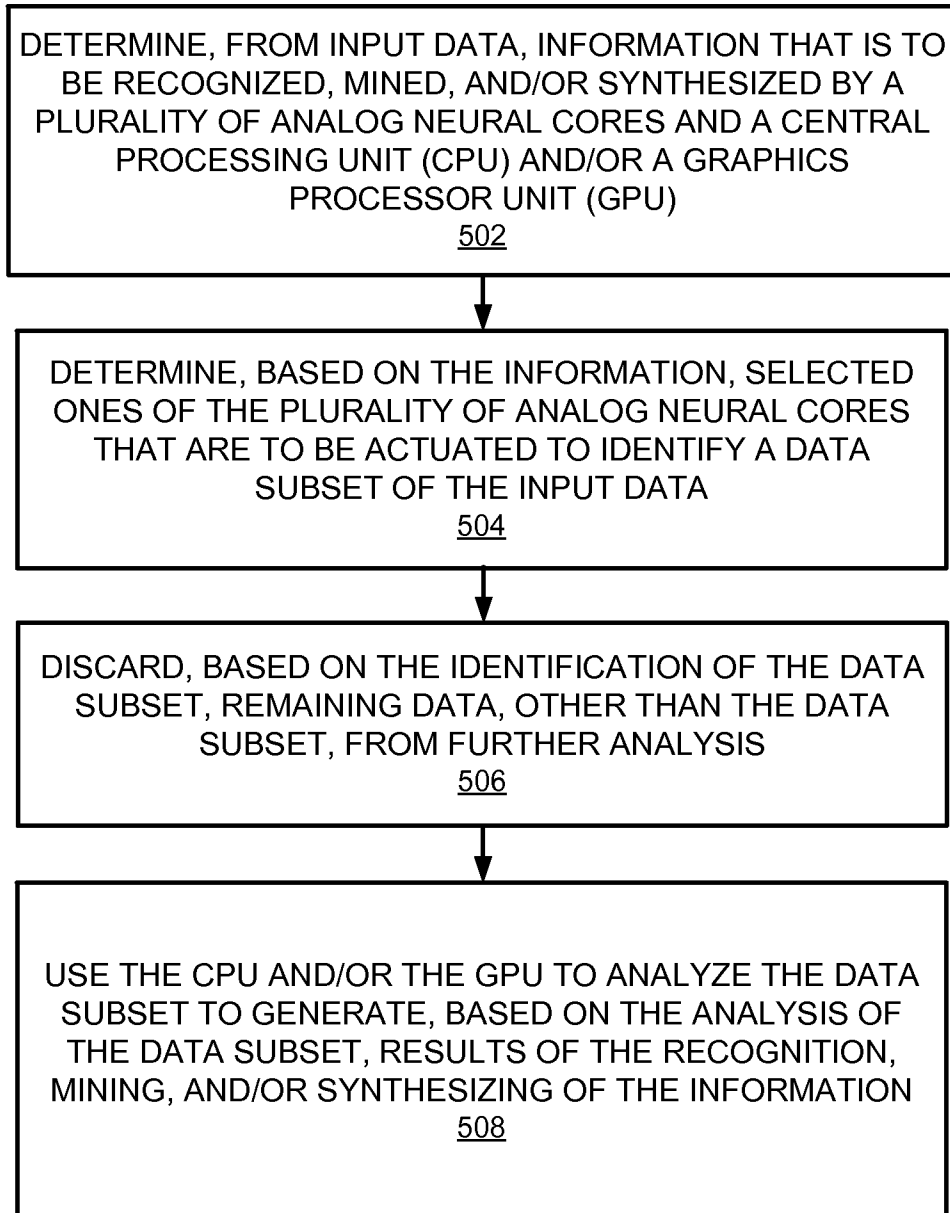
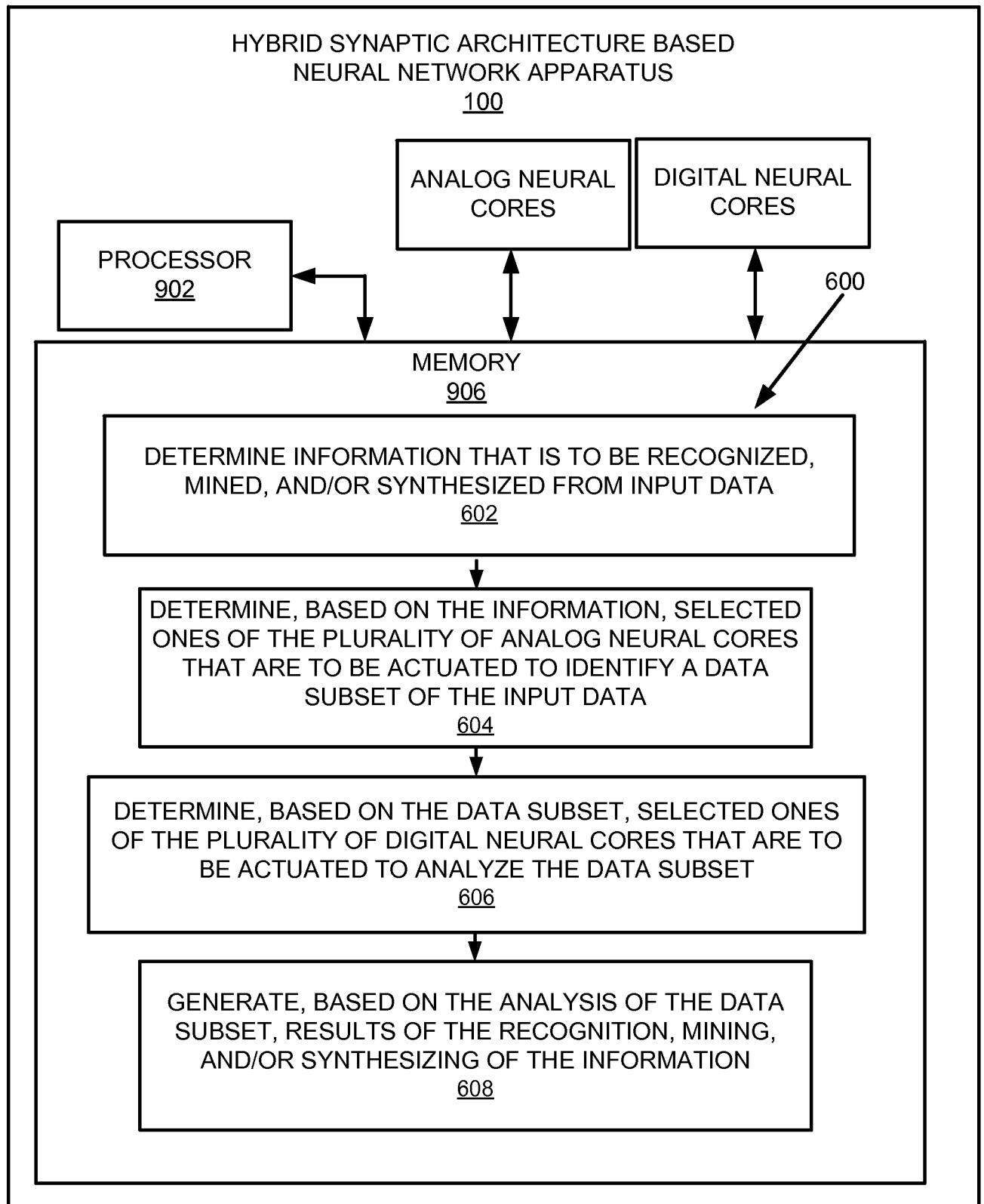


FIG. 4

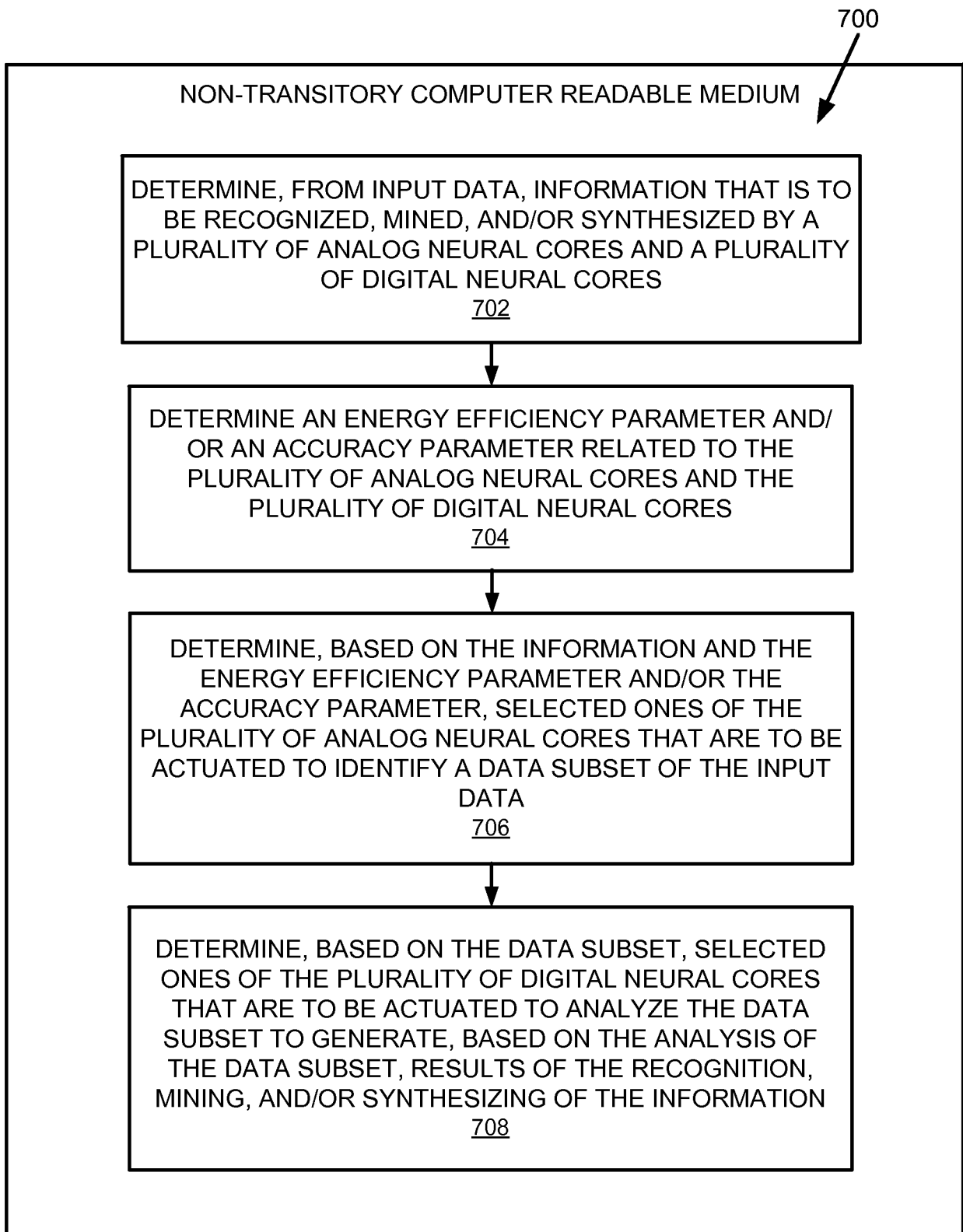
5/9

500**FIG. 5**

6/9



7/9

**FIG. 7**

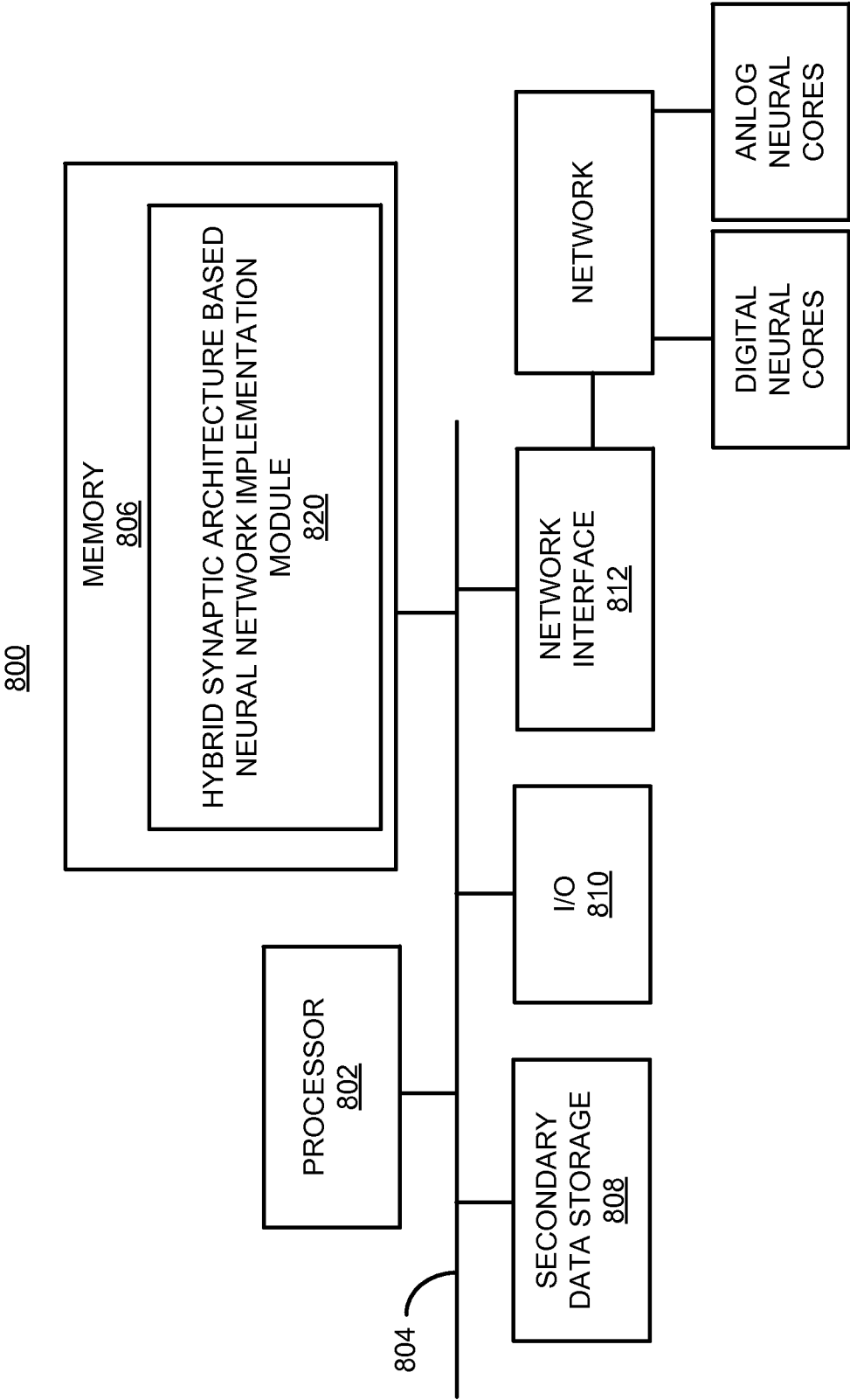


FIG. 8

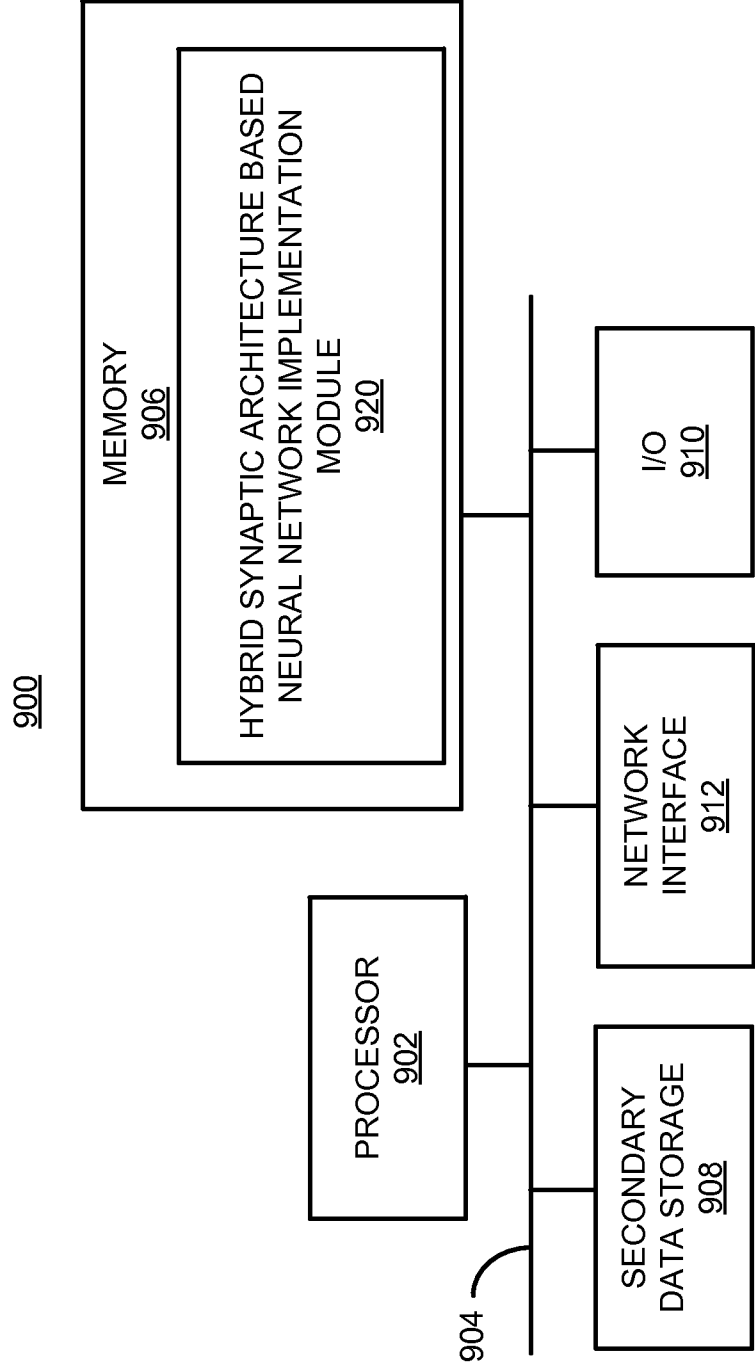


FIG. 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2015/058397**A. CLASSIFICATION OF SUBJECT MATTER****G06N 3/063(2006.01)i, G06N 3/04(2006.01)i**

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N 3/063; G06N 3/06; G06N 3/08; G06K 9/00; G06K 9/62; C40B 60/12; G06F 15/18; G06N 3/04

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & keywords: hybrid, neural network, analog, digital, cores, and similar terms.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2015-0046382 A1 (QUALCOMM INCORPORATED) 12 February 2015 See paragraphs [0037]-[0040], [0102]-[0103]; and figures 2, 10-11.	1-15
A	US 2010-0241601 A1 (JOHN C. CARSON et al.) 23 September 2010 See paragraphs [0050]-[0059], [0145]-[0157]; and figures 20-21.	1-15
A	US 2009-0048125 A1 (JAW-CHYNG LUE et al.) 19 February 2009 See paragraphs [0028], [0045]-[0050]; and figures 2, 4.	1-15
A	US 8401297 B1 (JOHN T. APOSTOLOS et al.) 19 March 2013 See column 2, line 46 - column 3, line 25; and figure 8.	1-15
A	US 2002-0059152 A1 (JOHN C. CARSON et al.) 16 May 2002 See paragraphs [0040]-[0052]; claims 1-4; and figures 2, 4.	1-15



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

28 July 2016 (28.07.2016)

Date of mailing of the international search report

28 July 2016 (28.07.2016)

Name and mailing address of the ISA/KR

International Application Division

Korean Intellectual Property Office

189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

BYUN, Sung Cheal

Telephone No. +82-42-481-8262



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2015/058397

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2015-0046382 A1	12/02/2015	US 2014-0351186 A1 US 2014-0351190 A1 US 2015-0046381 A1 WO 2014-189717 A1 WO 2014-189970 A2 WO 2014-189970 A3 WO 2015-020802 A2 WO 2015-020802 A3 WO 2015-020815 A2 WO 2015-020815 A3	27/11/2014 27/11/2014 12/02/2015 27/11/2014 27/11/2014 09/04/2015 12/02/2015 14/05/2015 12/02/2015 11/06/2015
US 2010-0241601 A1	23/09/2010	US 8510244 B2	13/08/2013
US 2009-0048125 A1	19/02/2009	US 7910873 B2	22/03/2011
US 8401297 B1	19/03/2013	GB 2518314 A WO 2014-005052 A2	18/03/2015 03/01/2014
US 2002-0059152 A1	16/05/2002	EP 1247194 A2 EP 1247194 A4 JP 2002-533823 A US 6389404 B1 WO 00-39658 A2 WO 00-39658 A3	09/10/2002 31/10/2007 08/10/2002 14/05/2002 06/07/2000 26/10/2000