

DATA CONVERSION TO/FROM SELECTED DATA TYPE WITH IMPLIED ROUNDING MODE

BACKGROUND

[0001] One or more aspects relate, in general, to facilitating processing within a computing environment, and in particular, to improving such processing.

[0002] In order to enhance processing in computing environments that are data and/or computational-intensive, co-processors are utilized, such as artificial intelligence accelerators (also referred to as neural network processors or neural network accelerators). Such accelerators provide a great deal of compute power used in performing, for instance, involved computations, such as computations on matrices or tensors.

[0003] Tensor computations, as an example, are used in complex processing, including deep learning, which is a subset of machine learning. Deep learning or machine learning, an aspect of artificial intelligence, is used in various technologies, including but not limited to, engineering, manufacturing, medical technologies, automotive technologies, computer processing, etc.

[0004] Performing tensor computations takes a great deal of time and compute power. Thus, improvements in facilitating such performance are sought including providing data in a format that facilitates such computations.

SUMMARY

[0005] Shortcomings of the prior art are overcome, and additional advantages are provided through the provision of a computer program product for facilitating processing within a computing environment. The computer program product includes one or more computer readable storage media and program instructions collectively stored on the one or more computer readable storage media to perform a method. The method includes obtaining an instruction to convert data from a source data type to a target data type. The source data type is selected from one or more source data types supported by the instruction, and the target data type is selected from one or more target data types supported by the instruction. Based on a selected data type of one of the source data type and the target data type, a determination is made of a rounding mode for use by the instruction. The rounding mode is implicitly set

based on the selected data type; it is assigned to the selected data type. A conversion of the data from the source data type to the target data type is performed. The conversion includes performing a rounding operation using the rounding mode implicitly set based on the selected data type. The performing the conversion provides a result in the target data type. The result is written to a select location.

[0006] By providing an implicitly set rounding mode that is assigned to a particular data type, the conversion process is simplified and may improve processing speed by limiting supported rounding modes. It also facilitates the writing of program code that converts data to be used on a select processor, such as a co-processor.

[0007] In one example, the selected data type is the target data type specified using the instruction. The target data type is, for instance, a machine-specific data type specified using the instruction. The machine-specific data type is, in one example, a neural network processing-data-type-1 data type. As an example, the rounding mode assigned to the neural network processing-data-type-1 data type is, for instance, a round to nearest ties away from zero. Further, in one example, the source data type is a standard data type specified using the instruction.

[0008] By providing an implicitly set rounding mode that is assigned to a particular data type, such as a machine-specific data type, the conversion process is simplified and may improve processing speed by limiting supported rounding modes. Efficiencies are provided, complexity is reduced, and system performance is increased. Performance of the conversion is improved, improving system performance and data availability.

[0009] In one example, the selected data type is the source data type specified using the instruction. The source data type is, for instance, a machine-specific data type specified using the instruction. The machine-specific data type is, in one example, a neural network processing-data-type-1 data type. As an example, the rounding mode assigned to the neural network processing-data-type-1 data type is a round to nearest ties away from zero. Further, in one example, the target data type is a standard data type specified using the instruction.

[0010] As an example, the data to be converted includes source data in the source data type. The source data includes a concatenation of data in one register specified using one field of the instruction and other data in another register specified using another field of the instruction. The performing the conversion includes, for instance, rounding the source data

based on the rounding mode to provide rounded data, and converting the rounded data to provide the result in the target data type.

[0011] As another example, the data to be converted includes source data in the source data type obtained using the instruction. The performing the conversion includes, for instance, rounding the source data based on the rounding mode to provide rounded data in a precision of a target data type number, and converting the rounded data to provide the result in the target data type.

[0012] Computer-implemented methods and systems relating to one or more aspects are also described and claimed herein. Further, services relating to one or more aspects are also described and may be claimed herein.

[0013] Additional features and advantages are realized through the techniques described herein. Other embodiments and aspects are described in detail herein and are considered a part of the claimed aspects.

BRIEF DESCRIPTION OF THE DRAWINGS

[0014] One or more aspects are particularly pointed out and distinctly claimed as examples in the claims at the conclusion of the specification. The foregoing and objects, features, and advantages of one or more aspects are apparent from the following detailed description taken in conjunction with the accompanying drawings in which:

FIGS. 1A-1D depict examples of binary floating-point formats, in accordance with one or more aspects of the present invention;

FIG. 2 depicts one example of a format of a Neural Network Processing (NNP)-data-type-1 data type, in accordance with one or more aspects of the present invention;

FIG. 3A depicts one example of a computing environment to incorporate and use one or more aspects of the present invention;

FIG. 3B depicts one example of further details of a memory of FIG. 3A, in accordance with one or more aspects of the present invention;

FIG. 3C depicts another example of further details of a memory of FIG. 3A, in accordance with one or more aspects of the present invention;

FIG. 3D depicts further details of a processor of FIG. 3A, in accordance with one or more aspects of the present invention;

FIG. 4A depicts one example of performing a conversion operation, in accordance with one or more aspects of the present invention;

FIG. 4B depicts further details relating to converting and rounding of the conversion operation of FIG. 4A, in accordance with one or more aspects of the present invention;

FIG. 5A depicts one example of a format of a Vector Floating Point Convert and Lengthen From Neural Network Processing High instruction, in accordance with one or more aspects of the present invention;

FIG. 5B depicts one example of a format of a Vector Floating Point Convert and Lengthen From Neural Network Processing Low instruction, in accordance with one or more aspects of the present invention;

FIG. 5C depicts one example of a format of a Vector Floating Point Convert and Round To Neural Network Processing instruction, in accordance with one or more aspects of the present invention;

FIG. 5D depicts one example of a format of a Vector Floating Point Convert From Neural Network Processing instruction, in accordance with one or more aspects of the present invention;

FIG. 5E depicts one example of a format of a Vector Floating Point Convert To Neural Network Processing instruction, in accordance with one or more aspects of the present invention;

FIG. 6A depicts one example of a format of a Neural Network Processing Assist instruction used in accordance with one or more aspects of the present invention;

FIG. 6B depicts one example of a general register used by the Neural Network Processing Assist instruction, in accordance with one or more aspects of the present invention;

FIG. 6C depicts examples of function codes supported by the Neural Network Processing Assist instruction, in accordance with one or more aspects of the present invention;

FIG. 6D depicts one example of another general register used by the Neural Network Processing Assist instruction, in accordance with one or more aspects of the present invention;

FIG. 6E depicts one example of a parameter block used by a query function of the Neural Network Processing Assist instruction, in accordance with one or more aspects of the present invention;

FIG. 6F depicts one example of a parameter block used by one or more non-query functions of the Neural Network Processing Assist instruction, in accordance with one or more aspects of the present invention;

FIG. 6G depicts one example of a tensor descriptor used by the Neural Network Processing Assist instruction, in accordance with one or more aspects of the present invention;

FIGS. 7A-7C depict examples of an input data layout used by the Neural Network Processing Assist instruction, in accordance with one or more aspects of the present invention;

FIGS. 8A-8C depict example output corresponding to the input data layout of FIGS. 7A-7C, in accordance with one or more aspects of the present invention;

FIGS. 9A-9B depict one example of facilitating processing within a computing environment, in accordance with one or more aspects of the present invention;

FIG. 10 depicts another example of a computing environment to incorporate and use one or more aspects of the present invention;

FIG. 11A depicts yet another example of a computing environment to incorporate and use one or more aspects of the present invention;

FIG. 11B depicts further details of the memory of FIG. 11A, in accordance with one or more aspects of the present invention;

FIG. 12 depicts one embodiment of a cloud computing environment, in accordance with one or more aspects of the present invention; and

FIG. 13 depicts one example of abstraction model layers, in accordance with one or more aspects of the present invention.

DETAILED DESCRIPTION

[0015] In accordance with one or more aspects of the present invention, a capability is provided to facilitate processing within a computing environment. In one aspect, the capability includes employing a co-processor, such as a neural network processor, of a particular machine model to perform functions, including, for instance, tensor computations. A specific model of a machine or processor has certain capabilities. Another model of the machine or processor may have additional, fewer and/or different capabilities and/or be of a different generation (e.g., a current or future generation) having additional, fewer and/or different capabilities. For greater efficiency and to preserve the accuracy of deep learning networks, a particular machine model may operate on a specific internal or machine-specific data format/type used to represent the input/output data (e.g., input/output tensor data). As an example, input data is converted from a standard format to a machine-specific format, and the output data is converted back to the standard format. The converting to/from the machine-specific format may include a rounding operation based on a pre-specified rounding mode. In accordance with one or more aspects of the present invention, the rounding mode is not selectable but instead, implied based on the machine-specific data format. This simplifies the conversion process and may improve processing speed by limiting the supported rounding modes. It also facilitates the writing of program code that converts data to be used on a co-processor, such as a neural network processor.

[0016] In one example, the converting includes using one or more conversion instructions. Each instruction is a single instruction (e.g., a single architected hardware machine instruction at the hardware/software interface) configured to perform, at the very least, a conversion function. Each instruction may be part of a general-purpose processor instruction set

architecture (ISA), which is dispatched by a program on a processor, such as a general-purpose processor or may be part of a special-purpose processor, such as a co-processor configured for certain functions, that is coupled to or part of the general-purpose processor. Other variations are also possible.

[0017] Each instruction indicates a source data type and a target data type and converts the source data type to the target data type. As one example, one of the data types is a standard data type, such as a binary floating-point data type. Binary floating-point data may be represented in a number of different formats including, for instance tiny, short, long or extended, examples of which are described with reference to FIGS. 1A-1D. Referring to FIG. 1A, a binary floating-point tiny format 100 includes, for example, a sign 102 (e.g., bit 0), an exponent + 15 (104) (e.g., bits 1-5), and a fraction 106 (e.g., bits 6-15). With reference to FIG. 1B, a binary floating-point short format 110 includes, for example, a sign 112 (e.g., bit 0), an exponent + 127 (114) (e.g., bits 1-8), and a fraction 116 (e.g., bits 9-31); and referring to FIG. 1C, a binary floating-point long format 120 includes, for example, a sign 122 (e.g., bit 0), an exponent + 1023 (124) (e.g., bits 1-11), and a fraction 126 (e.g., bits 12-63). Further, with reference to FIG. 1D, a binary floating-point extended format 130 includes, for example, a sign 132 (e.g., bit 0), an exponent + 16383 (134) (e.g., bits 1-15), and a fraction 136 (e.g., bits 16-127). Other formats are also possible.

[0018] As examples, when an operand in the binary floating-point short format is loaded into a floating-point register, it occupies the left half of the register and the right half remains unchanged. When an operand in the binary floating-point long format is loaded into a floating-point register, it occupies the entire register. An operand in the binary floating-point extended format occupies a register pair. The sign and biased exponent are in the leftmost 16 bits of the lower-numbered register of a register pair and are followed by the leftmost 48 bits of the fraction. The rightmost 64 bits of the fraction are in the higher-numbered register of the register pair.

[0019] Example properties of each of the formats are described below:

[0020]	<u>Property</u>	<u>Tiny</u>	<u>Short</u>	<u>Long</u>	<u>Extended</u>
[0021]	Format length (bits)	16	32	64	128
[0022]	Biased-exponent length (bits)	5	8	11	15

[0023]	Fraction length (bits)	10	23	52	112
[0024]	Precision (p)	11	24	53	113
[0025]	E _{max}	15	127	1023	16383
[0026]	E _{min}	-14	-126	-1022	-16382
[0027]	LUV bias	15	127	1023	16383
[0028]	N _{max}	$(1-2^{-11}) \times 2^{16}$	$(1-2^{-24}) \times 2^{128}$	$(1-2^{-53}) \times 2^{1024}$	$(1-2^{-113}) \times 2^{16384}$
[0029]		$\approx 3.4 \times 10^{-38}$	$\approx 1.8 \times 10^{-308}$	$\approx 1.2 \times 10^{-4932}$	
[0030]	N _{min}	1.0×2^{-14}	1.0×2^{-126}	1.0×2^{-1022}	1.0×2^{-16382}
[0031]		$\approx 1.2 \times 10^{-38}$	$\approx 2.2 \times 10^{-308}$	$\approx 3.4 \times 10^{-4932}$	
[0032]	D _{min}	1.0×2^{-24}	1.0×2^{-149}	1.0×2^{-1074}	1.0×2^{-16494}
[0033]		$\approx 1.4 \times 10^{-45}$	$\approx 4.9 \times 10^{-324}$	$\approx 6.5 \times 10^{-4966}$	

[0034] Where $\approx$ indicates that the value is approximate, E_{max} is maximum left-units-view (LUV) exponent, E_{min} is minimum left-units-view exponent, N_{max} is largest (in magnitude) representable finite number, and N_{min} is smallest (in magnitude) representable normal number.

[0035] For each format, the bias that is used to allow all exponents to be expressed as unsigned numbers is indicated above, as examples.

[0036] In each format, the binary point of a binary floating-point number is considered to be to the left of the leftmost fraction bit. To the left of the binary point there is an implied unit bit, which is considered to be one for normal numbers and zero for zeros and subnormal numbers. The fraction with the implied unit bit appended on the left is the significand of the number.

[0037] The value of a normal binary floating-point number is the significand multiplied by the radix 2 raised to the power of the unbiased exponent. The value of a subnormal binary floating-point number is the significand multiplied by the radix 2 raised to the power of the minimum exponent.

[0038] Example values of non-zero numbers for the various formats are described below:

<u>Number Class</u>	<u>Format</u>	<u>Value</u>
Normal	Tiny	$\pm 2^{e-15} \times (1.f)$
	Short	$\pm 2^{e-127} \times (1.f)$
	Long	$\pm 2^{e-1023} \times (1.f)$
	Extended	$\pm 2^{e-16383} \times (1.f)$
Subnormal	Tiny	$\pm 2^{-14} \times (0.f)$
	Short	$\pm 2^{-126} \times (0.f)$
	Long	$\pm 2^{-1022} \times (0.f)$
	Extended	$\pm 2^{-16382} \times (0.f)$

[0039] Where: e is biased exponent (shown in decimal), and f is fraction (in binary).

[0040] Further, example floating-point rounding and range actions are described below:

Nontrap Result (r) when Effective Rounding Method is

Range of v	Case	To Nearest with ties to even	To Nearest with ties away from 0	Toward 0	Toward $+\infty$	Toward $-\infty$	Prepare for shorter precision
$v < -N_{\max}$, $g < -N_{\max}$	Overflow	$-\infty^1$	$-\infty^1$	$-N_{\max}$	$-N_{\max}$	$-\infty^1$	$-N_{\max}$
$v < -N_{\max}$, $g = -N_{\max}$	Normal	$-N_{\max}$	$-N_{\max}$	$-N_{\max}$	$-N_{\max}$	-	$-N_{\max}$
$-N_{\max} \leq v \leq -N_{\min}$	Normal	g	g	g	g	g	g

$-N_{min} < v \leq$							
-Dmin	Tiny	d*	d*	d	d	d*	d
$-D_{min} < v <$							
-Dmin/2	Tiny	-Dmin	-Dmin	-0	-0	-Dmin	-Dmin
$v = -D_{min}/2$	Tiny	-0	-Dmin	-0	-0	-Dmin	-Dmin
$-D_{min}/2 <$							
$v < 0$	Tiny	-0	-0	-0	-0	-Dmin	-Dmin
$v = 0$	Exact zero difference ²	+0	+0	+0	+0	-0	+0
$0 < v <$							
+Dmin/2	Tiny	+0	+0	+0	+Dmin	+0	+Dmin
$v = +D_{min}/2$	Tiny	+0	+Dmin	+0	+Dmin	+0	+Dmin
$+D_{min}/2 < v$							
$< +D_{min}$	Tiny	+Dmin	+Dmin	+0	+Dmin	+0	+Dmin
$+D_{min} \leq v <$							
+Nmin	Tiny	d*	d*	d	d*	d	d
$+N_{min} \leq v$							
+Nmax	Normal	g	g	g	g	g	g
$+N_{max} < v,$							
$g = +N_{max}$	Normal	+Nmax	+Nmax	+Nmax	-	+Nmax	+Nmax
$+N_{max} < v,$							
$+N_{max} < g$	Overflow	$+\infty^1$	$+\infty^1$	+Nmax	$+\infty^1$	+Nmax	+Nmax

Where:

- This situation cannot occur.

* The rounded value, in the extreme case, may be Nmin. In this case, the exceptions are underflow, inexact and incremented.

¹ The non-trap result r is considered to have been incremented.

²	The exact-zero-difference case applies only to certain instructions, such as, e.g., Vector Floating Point Add, Vector Floating Point Subtract, Vector Floating Point Multiply and Add, Vector Floating Point Multiply and Subtract. For all other binary floating-point operations, a zero result is detected by inspection of the source operand elements without use of the R(v) function.
d	The denormalized value. The value derived when the precise intermediate value (v) is rounded to the format of the target including both precision and bounded exponent range. Except as explained in note *, this is a subnormal number.
g	The precision-rounded value. The value derived when the precise intermediate value (v) is rounded to the precision of the target, but assuming an unbounded exponent range.
v	Precise intermediate value. This is the value, before rounding, assuming unbounded precision and an unbounded exponent range. For Vector Load Rounded, v is the source value (a).
Dmin	Smallest (in magnitude) representable subnormal number in the target format.
Nmax	Largest (in magnitude) representable finite number in the target format.
Nmin	Smallest (in magnitude) representable normal number in the target format.

[0041] Another data type, in accordance with an aspect of the present invention, is a machine-specific data type, rather than a standard data type. A machine-specific data type is, for instance, a data type optimized for use by a particular processor running on a particular hardware machine. For instance, the machine-specific data type may preserve enhanced accuracy and/or be easier for the machine to implement. One example of a machine-specific data type is an NNP-data-type-1 data type (also referred to as a neural network processing-data-type-1 data type), as an example. This data type is a 16-bit floating-point format that provides a number of advantages for, e.g., deep learning training and inference computations, including, for instance: preserves the accuracy of deep learning networks; eliminates the subnormal format which simplifies rounding modes and handling of corner cases; automatic rounding to the nearest value for arithmetic operations; and special entities of infinity and not-a-number (NaN) are combined into one value (NINF), which is accepted and handled by arithmetic operations. NINF provides better defaults than other formats for exponent overflow and invalid operations (such as division by zero). This allows many programs to continue running without hiding such errors and without using specialized exception handlers. Other machine-specific data types are also possible.

[0042] One example of a format of the NNP-data-type-1 data type is depicted in FIG. 2. As depicted, in one example, NNP-data-type-1 data may be represented in a format 200, which includes, for instance, a sign 202 (e.g., bit 0), an exponent + 31 (204) (e.g., bits 1-6) and a fraction 206 (e.g., bits 7-15).

[0043] Example properties of the NNP-data-type-1 format are depicted below:

[0044]	<u>Property</u>	<u>NNP-data-type-1</u>
[0045]	Format length (bits)	16 bits
[0046]	Biased-exponent length (bits)	6 bits
[0047]	Fraction length (bits)	9 bits
[0048]	Precision (p)	10 bits
[0049]	Maximum left-units-view exponent (E _{max})	32
[0050]	Minimum left-units-view exponent (E _{min})	-31
[0051]	Left-units-view (LUV) bias	31
[0052]	N _{max}	$(1-2^{-9}) \times 2^{33} \approx 8.6 \times 10^9$
[0053]	N _{min}	$(1+2^{-9}) \times 2^{-31} \approx 4.6 \times 10^{-10}$
[0054]	D _{min}	---

[0055] Where $\approx$ indicates that the value is approximate, N_{max} is largest (in magnitude) representable finite number, and N_{min} is smallest (in magnitude) representable number.

[0056] In addition to (or instead of) the NNP-data-type-1 data type, another machine-specific data type that may be supported is a NNP-data-type-2 data type. Additional, fewer and/or other data types may also be supported.

[0057] Example properties of the NNP-data-type-2 format are depicted below:

[0058]	<u>Property</u>	<u>NNP-data-type-2</u>
[0059]	Format length (bits)	16 bits

[0060]	Biased-exponent length (bits)	5 bits
[0061]	Fraction length (bits)	10 bits
[0062]	Precision (p)	11 bits
[0063]	Maximum left-units-view exponent (E _{max})	16
[0064]	Minimum left-units-view exponent (E _{min})	-15
[0065]	Left-units-view (LUV) bias	15
[0066]	N _{max}	$(1-2^{-10}) \times 2^{16}$
[0067]	N _{min}	$(1+2^{-10}) \times 2^{-15}$
[0068]	D _{min}	---

[0069] Where $\approx$ indicates that the value is approximate, N_{max} is largest (in magnitude) representable finite number, and N_{min} is smallest (in magnitude) representable normal number.

[0070] Further details relating to the NNP-data-type-1 and NNP-data-type-2 data types are described below:

[0071] Biased Exponent: The bias that is used to allow exponents to be expressed as unsigned numbers is shown above. Biased exponents are similar to characteristics of the binary floating-point format, except that no special meanings are attached to biased exponents of all zeros and all ones, as described below with reference to the classes of the NNP-data-type-1 and NNP-data-type-2 data types.

[0072] Significand: The binary point of NNP-data-type-1 and NNP-data-type-2 numbers are considered to be to the left of the leftmost fraction bit. To the left of the binary point there is an implied unit bit, which is considered to be one for normal numbers and zero for zeros. The fraction with the implied unit bit appended on the left is the significand of the number.

[0073] The value of a normal NNP-data-type-1 or NNP-data-type-2 is the significand multiplied by the radix 2 raised to the power of the unbiased exponent.

[0074] Values of Non-Zero Numbers: The values of non-zero numbers are shown below:

[0075]	<u>Number Class</u>	<u>Value</u>
[0076]	Normal numbers	$\pm 2^{e-31} \times (1.f)$

Where e is biased exponent shown in decimal, and f is fraction in binary.

[0077] In one embodiment, there are three classes of NNP-data-type-1 data and NNP-data-type-2 data, including numeric and related non-numeric entities. Each data item includes a sign, an exponent and a significand. The exponent is biased such that all biased exponents are non-negative unsigned numbers and the minimum biased exponent is zero. The significand includes an explicit fraction and an implicit unit bit to the left of the binary point. The sign bit is zero for plus and one for minus.

[0078] All non-zero finite numbers permitted have a unique NNP-data-type-1 or NNP-data-type-2 representation. There are no subnormal numbers, which numbers might allow multiple representations for the same values, and there are no subnormal arithmetic operations. The three classes include, for instance:

[0079]	<u>Data Class</u>	<u>Sign</u>	<u>Biased Exponent</u>	<u>Unit Bit*</u>	<u>Fraction</u>
[0080]	Zero	$\pm$	0	0	0
[0081]	Normal numbers	$\pm$	0	1	Not 0
[0082]	Normal numbers	$\pm$	Not 0, not all ones	1	Any
[0083]	Normal numbers	$\pm$	All ones	-	Not all ones
[0084]	NINF	$\pm$	All ones	-	All ones

Where: – indicates does not apply, * indicates the unit bit is implied, NINF is not a number or infinity.

[0085] Further details regarding each of the classes are described below:

[0086] Zeros: Zeros have a biased exponent of zero and a zero fraction. The implied unit bit is zero.

[0087] Normal Numbers: Normal numbers may have a biased exponent of any value. When the biased exponent is 0, the fraction is to be non-zero. When the biased exponent is all

ones, the fraction is not to be all ones. Other biased exponent values may have any fraction value. The implied unit bit is one for all normal numbers.

[0088] NINF: A NINF is represented by a biased exponent of all ones and a fraction of all ones. A NINF represents a value not in the range of representable values in NNP-data-type-1 (i.e., 16-bit floating-point designed for deep learning (DL) that has 6 exponent bits and 9 fraction bits) or NNP-data-type-2. Normally, NINFs are just propagated during computations so that they will remain visible at the end.

[0089] Although binary floating-point data formats, the NNP-data-type-1 data type and the NNP-data-type-2 data type are provided as examples, additional, fewer and/or other data formats/types may also be used.

[0090] One embodiment of a computing environment to incorporate and use one or more aspects of the present invention is described with reference to FIG. 3A. As an example, the computing environment of FIG. 3A is based on the z/Architecture[®] instruction set architecture offered by International Business Machines Corporation, Armonk, New York. One embodiment of the z/Architecture instruction set architecture is described in a publication entitled, "z/Architecture Principles of Operation," IBM Publication No. SA22-7832-12, Thirteenth Edition, September 2019, which is hereby incorporated herein by reference in its entirety. The z/Architecture instruction set architecture, however, is only one example architecture; other architectures and/or other types of computing environments of International Business Machines Corporation and/or of other entities may include and/or use one or more aspects of the present invention. z/Architecture and IBM are trademarks or registered trademarks of International Business Machines Corporation in at least one jurisdiction.

[0091] In one example, a computing environment 10 includes a central electronics complex (CEC) 11. Central electronics complex 11 includes a plurality of components, such as, for instance, a memory 12 (a.k.a., system memory, main memory, main storage, central storage, storage) coupled to one or more processors, such as one or more general-purpose processors (a.k.a., central processing units (CPUs) 13) and one or more special-purpose processors (e.g., neural network processor 31), and to an input/output (I/O) subsystem 14.

[0092] As examples, the one or more special-purpose processors may be separate from the one or more general-purpose processors and/or at least one special-purpose processor may be embedded within at least one general-purpose processor. Other variations are also possible.

[0093] I/O subsystem 14 can be a part of the central electronics complex or separate therefrom. It directs the flow of information between main storage 12 and input/output control units 15 and input/output (I/O) devices 16 coupled to the central electronics complex.

[0094] Many types of I/O devices may be used. One particular type is a data storage device 17. Data storage device 17 can store one or more programs 18, one or more computer readable program instructions 19, and/or data, etc. The computer readable program instructions can be configured to carry out functions of embodiments of aspects of the invention.

[0095] Central electronics complex 11 can include and/or be coupled to removable/non-removable, volatile/non-volatile computer system storage media. For example, it can include and/or be coupled to a non-removable, non-volatile magnetic media (typically called a "hard drive"), a magnetic disk drive for reading from and writing to a removable, non-volatile magnetic disk (e.g., a "floppy disk"), and/or an optical disk drive for reading from or writing to a removable, non-volatile optical disk, such as a CD-ROM, DVD-ROM or other optical media. It should be understood that other hardware and/or software components could be used in conjunction with central electronics complex 11. Examples include, but are not limited to: microcode or millicode, device drivers, redundant processing units, external disk drive arrays, RAID systems, tape drives, and data archival storage systems, etc.

[0096] Further, central electronics complex 11 can be operational with numerous other general-purpose or special-purpose computing system environments or configurations. Examples of well-known computing systems, environments, and/or configurations that may be suitable for use with central electronics complex 11 include, but are not limited to, personal computer (PC) systems, server computer systems, thin clients, thick clients, handheld or laptop devices, multiprocessor systems, microprocessor-based systems, set top boxes, programmable consumer electronics, network PCs, minicomputer systems, mainframe computer systems, and distributed cloud computing environments that include any of the above systems or devices, and the like.

[0097] Central electronics complex 11 provides in one or more embodiments logical partitioning and/or virtualization support. As shown in FIG. 3B, memory 12 includes, for example, one or more logical partitions 20, a hypervisor 21 that manages the logical partitions, and processor firmware 22. One example of hypervisor 21 is the Processor Resource/System Manager (PR/SMTM), offered by International Business Machines Corporation, Armonk, New

York. As used herein, firmware includes, e.g., the microcode or millicode of the processor. It includes, for instance, the hardware-level instructions and/or data structures used in implementation of higher level machine code. In one embodiment, it includes, for instance, proprietary code that is typically delivered as microcode or millicode that includes trusted software, microcode or millicode specific to the underlying hardware and controls operating system access to the system hardware. PR/SM is a trademark or registered trademark of International Business Machines Corporation in at least one jurisdiction.

[0098] Each logical partition 20 is capable of functioning as a separate system. That is, each logical partition can be independently reset, run a guest operating system 23, such as the z/OS[®] operating system, offered by International Business Machines Corporation, Armonk, New York, or other control code 24, such as coupling facility control code (CFCC), and operate with different programs 25. An operating system or application program running in a logical partition appears to have access to a full and complete system, but in reality, only a portion of it is available. Although the z/OS operating system is offered as an example, other operating systems offered by International Business Machines Corporation and/or other companies may be used in accordance with one or more aspects of the present invention. z/OS is a trademark or registered trademark of International Business Machines Corporation in at least one jurisdiction.

[0099] Memory 12 is coupled to, e.g., CPUs 13 (FIG. 3A), which are physical processor resources that can be allocated to the logical partitions. For instance, a logical partition 20 may include one or more logical processors, each of which represents all or a share of a physical processor resource 13 that can be dynamically allocated to the logical partitions.

[00100] In another embodiment, central electronics complex 11 provides virtualization support (either with or without logical partitioning support). As shown in FIG. 3C, memory 12 of central electronics complex 11 includes, for example, one or more virtual machines 26, a virtual machine manager, such as a hypervisor 27, that manages the virtual machines, and processor firmware 28. One example of hypervisor 27 is the z/VM[®] hypervisor, offered by International Business Machines Corporation, Armonk, New York. The hypervisor is sometimes referred to as a host. z/VM is a trademark or registered trademark of International Business Machines Corporation in at least one jurisdiction.

[00101] The virtual machine support of the central electronics complex provides the ability to operate large numbers of virtual machines 26, each capable of operating with different

programs 29 and running a guest operating system 30, such as the Linux[®] operating system. Each virtual machine 26 is capable of functioning as a separate system. That is, each virtual machine can be independently reset, run a guest operating system, and operate with different programs. An operating system or application program running in a virtual machine appears to have access to a full and complete system, but in reality, only a portion of it is available. Although z/VM and Linux are offered as examples, other virtual machine managers and/or operating systems may be used in accordance with one or more aspects of the present invention. The registered trademark Linux[®] is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

[00102] In one example, a processor (e.g., processor 13 and/or processor 31) includes a plurality of functional components (or a subset thereof) used to execute instructions. As depicted in FIG. 3D, these functional components include, for instance, an instruction fetch component 350 to fetch instructions to be executed; an instruction decode unit 352 to decode the fetched instructions and to obtain operands of the decoded instructions; one or more instruction execute components 354 to execute the decoded instructions; a memory access component 356 to access memory for instruction execution, if necessary; and a write back component 358 to provide the results of the executed instructions. One or more of the components may access and/or use one or more registers 360 in instruction processing. Further, one or more of the components may, in accordance with one or more aspects of the present invention, include at least a portion of or have access to one or more other components used in performing conversion, rounding and/or processing of, e.g., one or more conversion instructions, and/or neural network processing assist processing of, e.g., a Neural Network Processing Assist instruction (or other processing that may use one or more aspects of the present invention), as described herein. The one or more other components include, for instance, a conversion and/or rounding component 370 and/or a neural network processing assist component 372 (and/or one or more other components).

[00103] In accordance with an aspect of the present invention, for greater efficiency, a select machine of a particular machine model may operate on a specific internal or machine-specific data format/type used to represent the input/output data. In such a case, as an example, input data is converted from a standard format to the machine-specific format, and the output data is converted back to the standard format. Other examples are also possible. In accordance with an aspect of the present invention, the converting to/from the machine-specific format may include rounding. The rounding mode to be used is implied (also referred

to as pre-selected) based on the machine-specific format to be used rather than being selectable. As examples, the rounding mode is assigned to the machine-specific format and not obtained from, e.g., an instruction operand or a control register, such as a floating-point control register, that allows for selectable rounding modes. Instead, in accordance with an aspect of the present invention, the rounding mode is pre-selected and assigned to the machine-specific format such that the machine-specific format uses only that rounding mode. This provides efficiencies, reduces complexity and increases system performance. Performance of the conversions is improved, improving system performance and data availability.

[00104] In one example, in accordance with an aspect of the present invention, a conversion operation is performed by, e.g., a general-purpose processor (e.g., processor 13) to enable a select processor of a particular machine model (e.g., neural network processor 31) to perform functions on data in a machine-specific format. This conversion operation includes performing a conversion, as indicated by the selected instruction, and may include rounding using a rounding mode determined based on a data type indicated by the instruction. Further details regarding this processing are described with reference to FIGS. 4A-4B.

[00105] Referring to FIG. 4A, in one example, an appropriate conversion instruction is selected and decoded 400. The conversion instruction selected depends, for instance, on the source and target data types, as well as other processing to be performed. Example conversion instructions include, for instance, a Vector Floating Point Convert and Lengthen From Neural Network Processing High instruction, a Vector Floating Point Convert and Lengthen From Neural Network Processing Low instruction, a Vector Floating Point Convert and Round to Neural Network Processing instruction, a Vector Floating Point Convert From Neural Network Processing instruction and a Vector Floating Point Convert To Neural Network Processing instruction. Additional, fewer and/or other conversion instructions may be used depending, for instance, on the data types to be converted, etc.

[00106] The decoding includes, for instance, obtaining one or more input operands of the instruction (e.g., text information provided in one or more masks) and determining whether and what type of machine-specific data type is specified by the one or more input operands. For instance, either a source data type input operand or a target data type input operand may specify a machine-specific data type. As one example, it is known by the instruction being decoded (e.g., indicated by the opcode) whether it is the source or target data type specifying

the machine-specific data type. In another example, an indicator in an operand of the instruction indicates which operand is specifying a machine-specific data type, if any. The indicator can be part of the operand specifying the data type or in a separate operand. In other embodiments, both the source and target data types specify machine-specific data types (e.g., one type and another type). Other examples are also possible.

[00107] A determination is made as to whether the instruction is specifying a first machine-specific data type, referred to herein as machine-specific data type-1 (e.g., NNP-data-type-1) 410. If the instruction is not specifying a first machine-specific data type, then a further determination is made as to whether the instruction is specifying a second machine-specific data type, referred to herein as machine-specific data type-2 (e.g., NNP-data-type-2) (420). If the first or second machine-specific data type is not being specified, then an unsupported data type exception is specified 430, in one example. Although a check is made of two machine-specific data types in this example, other embodiments may include additional or fewer checks for additional or fewer machine-specific data types. Many possibilities exist.

[00108] Returning to inquiry 410, if machine-specific data type-1 is specified, then the rounding mode is automatically and implicitly set to a first mode, such as round to nearest with ties away from zero. Further, returning to inquiry 420, if machine-specific data type-2 is specified, then the rounding mode is automatically and implicitly set to a second mode, such as round toward 0. Other examples are also possible.

[00109] Subsequent to setting the rounding mode, which is pre-specified based on the machine-specific data type that is selected, the instruction is executed and in particular, the conversion and rounding are performed 440, as described with reference to FIG. 4B.

[00110] Referring to FIG. 4B, in one embodiment, one or more input operands of the instruction are obtained, if not previously obtained, including, for instance, the data to be converted 450. A determination is made of the machine-specific data type specified using an operand of the instruction (assuming a machine-specific data type is specified) and thus, the rounding mode to be used 460. Rounding is performed based on the rounding mode pre-selected and assigned to the machine-specific data type 470. The data that has been rounded is converted from the source data type to the target data type 480, and results of the rounding and converting are provided 490. In one example, the results are an output tensor in the target data type. This completes the rounding/converting processing, in one example.

[00111] In other embodiments, rounding may be performed after the converting. Further, in other embodiments, the rounding mode may be pre-selected or assigned to a standard data type, rather than a machine-specific data type. In one or more embodiments, the rounding mode is implicitly set for one or more select data types, in which a select data type is a standard data type, a machine-specific data type and/or another data type. Other examples are also possible.

[00112] In one embodiment, each of the conversion instructions is, for instance, part of a vector facility, which provides, for instance, fixed sized vectors ranging from one to sixteen elements. Each vector includes data which is operated on by vector instructions defined in the facility. In one embodiment, if a vector is made up of multiple elements, then each element is processed in parallel with the other elements. Instruction completion does not occur, in one example, until processing of all the elements is complete. In other embodiments, the elements are processed partially in parallel and/or sequentially.

[00113] In one embodiment, there are 32 vector registers and other types of registers can map to a quadrant of the vector registers. For instance, a register file, which is an array of processor registers in a central processing unit (e.g., processor 13), may include 32 vector registers and each register is 128 bits in length. Sixteen floating-point registers, which are 64 bits in length, can overlay the vector registers. Thus, as an example, when floating-point register 2 is modified, then vector register 2 is also modified. Other mappings for other types of registers are also possible.

[00114] Vector data appears in storage, for instance, in the same left-to-right sequence as other data formats. Bits of a data format that are numbered 0-7 constitute the byte in the leftmost (lowest-numbered) byte location in storage, bits 8-15 form the byte in the next sequential location, and so on. In a further example, the vector data may appear in storage in another sequence, such as right-to-left.

[00115] Each of the vector instructions described herein has a plurality of fields, and a field may have a subscript number associated therewith. The subscript number associated with a field of the instruction denotes the operand to which the field applies. For instance, the subscript number 1 associated with vector register V_1 denotes that the register specified using V_1 includes the first operand, and so forth. A register operand is one register in length, which is, for instance, 128 bits.

[00116] Further, each of the vector instructions includes a register extension bit (RXB) field that includes the most significant bit for a vector register designated operand. Bits for register designations not specified by the instruction are to be reserved and set to zero. The most significant bit is concatenated, for instance, to the left of the four-bit register designation of the vector register field to create a five-bit vector register designation.

[00117] In one example, the RXB field includes four bits (e.g., bits 0-3), and the bits are defined, as follows:

[00118] 0 - Most significant bit for the first vector register designation (e.g., in bits 8-11) of the instruction.

[00119] 1 - Most significant bit for the second vector register designation (e.g., in bits 12-15) of the instruction, if any.

[00120] 2 - Most significant bit for the third vector register designation (e.g., in bits 16-19) of the instruction, if any.

[00121] 3 - Most significant bit for the fourth vector register designation (e.g., in bits 32-35) of the instruction, if any.

[00122] Each bit is set to zero or one by, for instance, the assembler depending on the register number. For instance, for registers 0-15, the bit is set to 0; for registers 16-31, the bit is set to 1, etc.

[00123] In one embodiment, each RXB bit is an extension bit for a particular location in an instruction that includes one or more vector registers. For instance, bit 0 of RXB is an extension bit for location 8-11, which is assigned to, e.g., V₁, and so forth. In particular, for vector registers, the register containing the operand is specified using, for instance, a four-bit field of the register field with the addition of its corresponding register extension bit (RXB) as the most significant bit. For instance, if the four-bit field is 0110 and the extension bit is 0, then the five bit field 00110 indicates register number 6. In a further embodiment, the RXB field includes additional bits, and more than one bit is used as an extension for each vector or location.

[00124] In the description herein of the conversion instructions and/or other instructions, specific locations, specific fields and/or specific sizes of the fields are indicated (e.g., specific bytes and/or bits). However, other locations, fields and/or sizes may be provided. Further,

although the setting of a bit to a particular value, e.g., one or zero, may be specified, this is only an example. The bit, if set, may be set to a different value, such as the opposite value or to another value, in other examples. Many variations are possible.

[00125] One example of a vector conversion instruction is the Vector Floating Point Convert and Lengthen From Neural Network Processing High instruction, an example of which is described with reference to FIG. 5A. In one embodiment, Vector Floating Point Convert and Lengthen From Neural Network Processing High instruction 500 has a VRR format indicating a vector register to register operand with an extended opcode. It includes, for instance, operation code (opcode) fields 502a, 502b (e.g., bits 0-7 and 40-47) indicating a conversion is to be performed; a first vector register (V_1) field 504 (e.g., bits 8-11) used to designate a first vector register; a second vector register (V_2) field 506 (e.g., bits 12-15) used to designate a second vector register; a first mask (M_4) field 508 (e.g., bits 28-31); a second mask (M_3) field 510 (e.g., bits 32-35); and a register extension bit (RXB) field 512 (e.g., bits 36-39), each of which is described herein. In one embodiment, the fields are separate and independent from one another; however, in other embodiments, more than one field may be combined. Further information regarding these fields is described below.

[00126] In one embodiment, vector register (V_1) field 504 is used to indicate a vector register that is to store the first operand. The first operand is a result obtained from the conversion. The second operand is contained in a vector register specified using vector register (V_2) field 506 and is, for instance, a value to be converted. In one example, each of vector register fields 504, 506 is used with RXB field 512 to designate the vector register. In another embodiment, a vector register field is used to designate one or more vector registers.

[00127] In one example, M_4 field 508 specifies the source data type. Example data types include, for instance:

[00128]	<u>M_4</u>	<u>Source Data Type</u>
[00129]	0	Neural Network Processing-data-type-1
[00130]	1	Neural Network Processing-data-type-2
[00131]	2-15	Reserved

[00132] If a reserved value is specified, an IEEE-inexact exception is recognized.

[00133] In one example, M_3 field 510 specifies the target data type. Example data types include, for instance:

[00134]	M_3	<u>Target Data Type</u>
[00135]	0-1	Reserved
[00136]	2	Binary Floating-Point short format
[00137]	3-15	Reserved

[00138] If a reserved value is specified, an IEEE-inexact exception is recognized.

[00139] In one embodiment, if the combination of non-reserved M_3 and M_4 values is not supported by the model (e.g., the select machine or processor of a particular model), an IEEE-inexact exception is recognized, in one example. For instance, the IEEE-inexact exception flag, also referred to herein as an indicator, is set, e.g., to one. If an IEEE-inexact exception is recognized, the first operand is, e.g., unpredictable. In one example, the first operand is set to a default value (e.g., zeros, the value to be converted, or another value). In other embodiments, the IEEE-inexact exception is recognized only for certain invalid data types, such as an invalid machine-specific data type (e.g., NNP-data-type-1, NNP-data-type-2), but not for other data types, such as an invalid standard data type (e.g., a binary floating-point format). Other examples are also possible. Further, in one embodiment, if the IEEE-inexact exception is recognized for only certain invalid data types (e.g., invalid machine-specific data types) and not other invalid data types (e.g., invalid standard types), then a specification exception may be recognized for the other invalid data types. Other variations are possible.

[00140] Although particular values and formats are provided as examples for the source and target data types, additional, fewer and/or other values and/or formats may be used. Further, although in this embodiment, rounding is not performed for this instruction, in other embodiments, rounding, in accordance with one or more aspects of the present invention, is performed. If rounding is to be performed, the rounding mode used is, for instance, the one assigned to the machine-specific data type specified by the instruction.

[00141] In execution of one embodiment of the Vector Floating Point Convert and Lengthen From Neural Network Processing High instruction, the values of the specified source data type elements in the leftmost half of the second operand are converted to the specified target format and placed in the corresponding double-wide elements in the first

operand. The rightmost half of the second operand is ignored. The format of the second operand elements is specified by the source data type in the M_4 field. The format of the first operand elements is specified by the target data type in the M_3 field.

[00142] As an example, M_4 specifies the Neural Network Processing-data-type-1 data type, and the values in the second operand that are in the NNP-data-type-1 data format are converted to the binary floating-point short format specified by M_3 and the result is placed in the first operand. The conversion includes, for instance:

[00143] Second Operand Value = Binary $1.1 \times 2^1 = 3.0$ decimal

[00144] In NNP-data-type-1:

[00145] Sign = 0 (for positive)

[00146] Exponent = $31 + 1$ (1 is the exponent in 2^1) = $32 = 100000$

[00147] Mantissa = 1.100000000 (fractional part of binary number (e.g., 1) with 9 bits of precision; note-the leading “1.” is implied and therefore, removed when placed in the actual NNP-data-type-1 format number)

[00148] This provides 0 100000 100000000 in binary

[00149] 0x4100 (100000100000000 converted to hex)

[00150] In short binary floating-point:

[00151] Sign = 0 (for positive)

[00152] Exponent = $127 + 1 = 128 = 10000000$

[00153] Mantissa = .10000000000000000000000 (fractional part, 23 bits, with implied “1.” also removed)

[00154] In hex: 0x40400000

[00155] In execution of one embodiment of the instruction, the condition code remains unchanged. Possible IEEE exceptions include invalid operation and inexact.

[00156] Example results of the Vector Floating-Point Convert and Lengthen From NNP High/Low instructions are indicated below:

Operand Element (a)	Invalid-Op. Mask (FPC 0.0)	Results
$-N_{\max} \leq v \leq -N_{\min}$	-	$T(v)$
-0	-	$T(-0)$
+0	-	$T(+0)$
$N_{\min} \leq v \leq N_{\max}$	-	$T(v)$
$\pm N_{\text{INF}}$	0	$T(\text{dNaN})$, $\text{SF} \leftarrow 1$
$\pm N_{\text{INF}}$	1	$\text{PIV}(1)$

Where:

-	The results do not depend on this condition or mask bit.
$N_{\max}$	Maximum positive number representable in the source format.
$N_{\min}$	Smallest positive number representable in the source format.
$\text{PIV}(h)$	Program interruption for vector-processing exception with VXC (Vector Exception Code) containing the element index and a VIC (Vector Index) of h in hex.
v	Precise intermediate floating-point value which is the source value (a).
SF_i	IEEE invalid-operation flag, FPC (Floating-Point Control Register) 1.0.
dNaN	Default NaN.
$T(x)$	The value x is converted to the target format and placed at the target operand element location.

[00157] Another example of a conversion instruction is a Vector Floating Point Convert and Lengthen From Neural Network Processing Low instruction. As depicted in FIG. 5B, a Vector Floating Point Convert and Lengthen From Neural Network Processing Low instruction 520 has a VRR format indicating a vector register to register operand with an extended opcode. It includes, for instance, operation code (opcode) fields 522a, 522b (e.g., bits 0-7 and 40-47) indicating a conversion to be performed; a first vector register (V_1) field 524 (e.g., bits 8-11) used to designate a first vector register; a second vector register (V_2) field 526 (e.g., bits 12-15) used to designate a second vector register; a first mask (M_4) field 528 (e.g., bits 28-31); a second mask (M_3) field 530 (e.g., bits 32-35); and a register extension bit (RXB) field 532 (e.g., bits 36-39), each of which is described herein. In one embodiment, the fields are separate and independent from one another; however, in other embodiments, more

than one field may be combined. Further information regarding these fields is described below.

[00158] In one embodiment, vector register (V_1) field 524 is used to indicate a vector register that is to store the first operand. The first operand is a result obtained from the conversion. The second operand is contained in a vector register specified using vector register (V_2) field 526 and includes, for instance, a value to be converted. In one example, each of vector register fields 524, 526 is used with RXB field 532 to designate the vector register. In another embodiment, a vector register field is used to designate one or more vector registers.

[00159] In one example, M_4 field 528 specifies the source data type. Example data types include, for instance:

[00160]	<u>M_4</u>	<u>Source Data Type</u>
[00161]	0	Neural Network Processing-data-type-1
[00162]	1	Neural Network Processing-data-type-2
[00163]	2-15	Reserved

[00164] If a reserved value is specified, an IEEE-inexact exception is recognized.

[00165] In one example, M_3 field 530 specifies the target data type. Example data types include, for instance:

[00166]	<u>M_3</u>	<u>Target Data Type</u>
[00167]	0-1	Reserved
[00168]	2	Binary Floating-Point short format
[00169]	3-15	Reserved

[00170] If a reserved value is specified, an IEEE-inexact exception is recognized.

[00171] In one embodiment, if the combination of non-reserved M_3 and M_4 values is not supported by the model (e.g., the select machine or processor of a particular model), an IEEE-inexact exception is recognized, in one example. For instance, the IEEE-inexact exception flag, also referred to herein as an indicator, is set, e.g., to one. If an IEEE-inexact exception is

recognized, the first operand is, e.g., unpredictable. In one example, the first operand is set to a default value (e.g., zeros, the value to be converted, or another value). In other embodiments, the IEEE-inexact exception is recognized only for certain invalid data types, such as an invalid machine-specific data type (e.g., NNP-data-type-1, NNP-data-type-2), but not for other data types, such as an invalid standard data type (e.g., a binary floating-point format). Other examples are also possible. Further, in one embodiment, if the IEEE-inexact exception is recognized for only certain invalid data types (e.g., invalid machine-specific data types) and not other invalid data types (e.g., invalid standard types), then a specification exception may be recognized for the other invalid data types. Other variations are possible.

[00172] Although particular values and formats are provided as examples for the source and target data types, additional, fewer and/or other values and/or formats may be used. Further, although in this embodiment, rounding is not performed for this instruction, in other embodiments, rounding, in accordance with one or more aspects of the present invention, is performed. If rounding is to be performed, the rounding mode used is, for instance, the one assigned to the machine-specific data type specified by the instruction.

[00173] In execution of one embodiment of the Vector Floating Point Convert and Lengthen From Neural Network Processing Low instruction, the values of the specified source data type elements in the rightmost half of the second operand and extended with zeros are converted to the specified target format and placed in the corresponding double-wide elements in the first operand. The leftmost half of the second operand is ignored. The format of the second operand elements is specified by the source data type in the M₄ field. The format of the first operand elements is specified by the target data type in the M₃ field.

[00174] In execution of one embodiment of the instruction, the condition code remains unchanged. Possible IEEE exceptions include invalid operation and inexact. Example results of the Vector Floating Point Convert and Lengthen From NNP Low instruction are indicated above.

[00175] Another example of a conversion instruction is a Vector Floating Point Convert and Round to Neural Network Processing instruction. As depicted in FIG. 5C, a Vector Floating Point Convert and Round to Neural Network Processing instruction 540 has a VRR format indicating a vector register to register operand with an extended opcode. It includes, for instance, operation code (opcode) fields 542a, 542b (e.g., bits 0-7 and 40-47) indicating a conversion is to be performed, as well as rounding; a first vector register (V₁) field 544 (e.g.,

bits 8-11) used to designate a first vector register; a second vector register (V_2) field 546 (e.g., bits 12-15) used to designate a second vector register; a third vector register (V_3) field 548 (e.g., bits 16-19) used to designate a third vector register; a first mask (M_5) field 550 (e.g., bits 28-31); a second mask (M_4) field 552 (e.g., bits 32-35); and a register extension bit (RXB) field 554 (e.g., bits 36-39), each of which is described herein. In one embodiment, the fields are separate and independent from one another; however, in other embodiments, more than one field may be combined. Further information regarding these fields is described below.

[00176] In one embodiment, vector register (V_1) field 544 is used to indicate a vector register that is to store the first operand. The first operand is a result obtained from the conversion and rounding. The second operand contained in a vector register specified using vector register (V_2) field 546 and the third operand contained in a vector register specified using vector register (V_3) field 548 are concatenated to create a source vector. In one example, each of vector register fields 544, 546, 548 is used with RXB field 554 to designate the vector register. In another embodiment, a vector register field is used to designate one or more vector registers.

[00177] In one example, M_5 field 550 specifies the source data type. Example data types include, for instance:

[00178]	<u>M_5</u>	<u>Source Data Type</u>
[00179]	0-1	Reserved
[00180]	2	Binary Floating-Point short format
[00181]	3-15	Reserved

[00182] If a reserved value is specified, an IEEE-inexact exception is recognized.

[00183] In one example, M_4 field 552 specifies the target data type. Example data types include, for instance:

[00184]	<u>M_4</u>	<u>Target Data Type</u>
[00185]	0	Neural Network Processing-data-type-1
[00186]	1	Neural Network Processing-data-type-2

[00187] 2-15 Reserved

[00188] If a reserved value is specified, an IEEE-inexact exception is recognized.

[00189] In one embodiment, if the combination of non-reserved M_4 and M_5 values is not supported by the model (e.g., the select machine or processor of a particular model), an IEEE-inexact exception is recognized, in one example. For instance, the IEEE-inexact exception flag, also referred to herein as an indicator, is set, e.g., to one. If an IEEE-inexact exception is recognized, the first operand is, e.g., unpredictable. In one example, the first operand is set to a default value (e.g., zeros, the value to be converted, or another value). In other embodiments, the IEEE-inexact exception is recognized only for certain invalid data types, such as an invalid machine-specific data type (e.g., NNP-data-type-1, NNP-data-type-2), but not for other data types, such as an invalid standard data type (e.g., a binary floating-point format). Other examples are also possible. Further, in one embodiment, if the IEEE-inexact exception is recognized for only certain invalid data types (e.g., invalid machine-specific data types) and not other invalid data types (e.g., invalid standard types), then a specification exception may be recognized for the other invalid data types. Other variations are possible.

[00190] Although particular values and formats are provided as examples for the source and target data types, additional, fewer and/or other values and/or formats may be used.

[00191] In execution of one embodiment of the Vector Floating Point Convert and Round to Neural Network Processing instruction, a source vector is created from the concatenation of the second operand followed by the third operand. In a left to right fashion, the source data type elements of the source vector are rounded to the precision and converted to the format of a target data type number. The half-sized target data type number is placed in the corresponding element of the vector register specified first operand. In accordance with an aspect of the present invention, the rounding mode used for the rounding is based on the machine-specific data type of the target data type. Thus, in one example, the rounding mode is the mode assigned to Neural Network Processing-data-type-1, since this is the machine-specific data type specified for use in the conversion by the instruction. The rounding mode assigned to Neural Network Processing-data-type-1 in this example is round to nearest with ties away from zero.

[00192] In another example, the Neural Network Processing-data-type-2 data type is specified, and the rounding mode assigned to the Neural Network Processing-data-type-2 is round toward 0 with ties to even.

[00193] An example conversion from a short format binary floating-point number to a NNP-data-type-2 value is illustrated below:

[00194] Second operand value: $+1.0000000001 \times 2^{-16}$ (binary) in BFP short format.

[00195] Sign = 0 (for positive)

[00196] Exponent = $-16+127 = 111$ (decimal)

[00197] Mantissa = 1.0000000001 (fractional part with 10 bits of precision; note the leading “1.” is implied and therefore removed in the BFP Short format number.

[00198] This provides 0 01101111 000000000100000000000000 or 0x37802000.

[00199] In NNP-data-type-2, this number cannot be directly represented since its exponent value is less than 2^{-15} . In fact this value is exactly half of NMin. When using the round to nearest with ties to even, this value becomes +0. So the result would be 0x0000.

[00200] In execution of one embodiment of the instruction, the condition code remains unchanged. Possible IEEE exceptions include invalid operation, overflow, underflow and inexact.

[00201] Example results of the Vector Floating Point Convert and Round to Neural Network Processing instruction are indicated below:

Operand Element (a)	Invalid	Overflow	Underflow	Results
	Op. Mask (FPC 0.0)	Mask (FPC 0.2)	Mask (FPC 0.3)	
$v < -N_{\max}, g < -N_{\max}$	-	0	-	T(-NINF), SFo←1
$v < -N_{\max}, g < -N_{\max}$	-	1	-	PIV(3)
$v < -N_{\max}, g = -N_{\max}$	-	-	-	T(-Nmax)
$-N_{\max} \leq v \leq -N_{\min}$	-	-	-	T(g)
$-N_{\min} < v < 0, g < -N_{\min}$	-	-	0	T(-0), SFu←1
$-N_{\min} < v < 0, g < -N_{\min}$	-	-	1	PIV(4)

$-N_{min} < v < 0, g = -N_{min}$	-	-	0	$-N_{min}, SF_u \leftarrow 1$
$-N_{min} < v < 0, g = -N_{min}$	-	-	1	PIV(4)
-0	-	-	-	T(-0)
+0	-	-	-	T(+0)
$0 < v < N_{min}, g < N_{min}$	-	-	0	$T(+0), SF_u \leftarrow 1$
$0 < v < N_{min}, g < N_{min}$	-	-	1	PIV(4)
$0 < v < N_{min}, g = N_{min}$	-	-	0	$T(+N_{min}), SF_u \leftarrow 1$
$0 < v < N_{min}, g = N_{min}$	-	-	1	PIV(4)
$N_{min} \leq v \leq N_{max}$	-	-	-	T(g)
$N_{max} < v, g = N_{max}$	-	-	-	T(+N _{max})
$N_{max} < v, g > N_{max}$	-	0	-	$T(\pm N_{INF}^1), SF_o \leftarrow 1$
$N_{max} < v, g > N_{max}$	-	1	-	PIV(3)
$\pm\infty$	-	0	-	$T(\pm N_{INF}^1), SF_o \leftarrow 1$
$\pm\infty$	-	1	-	PIV(3)
$\pm NaN$	0	-	-	$T(\pm N_{INF}^1), SF_i \leftarrow 1$
$\pm NaN$	1	-	-	PIV(1)

Where:

¹	The sign bit of the result is the same as the sign bit of the source.
-	The results do not depend on this condition or mask bit.
N _{max}	Maximum positive number representable in the target format.
N _{min}	Smallest positive number representable in the target format.
PIV(h)	Program interruption for vector-processing exception with VXC (Vector Exception Code) containing the element index and a VIC (Vector Index) of h in hex.
v	Precise intermediate floating-point value which is the source value (a).
g	The precision-rounded value. The value derived when the precise intermediate value (v) is rounded to the precision of the target, but assuming an unbounded exponent range. The rounding mode used is round to nearest with ties away from 0.

SFi	IEEE invalid-operation flag, FPC (Floating-Point Control Register) 1.0.
SFo	IEEE overflow flag, FPC 1.2.
SFu	IEEE underflow flag, FPC 1.3.
T(x)	The value x is converted to the target format and placed at the target operand element location.

[00202] Another example of a conversion instruction is a Vector Floating Point Convert From Neural Network Processing instruction. As depicted in FIG. 5D, a Vector Floating Point Convert From Neural Network Processing instruction 560 has a VRR format indicating a vector register to register operand with an extended opcode. It includes, for instance, operation code (opcode) fields 562a, 562b (e.g., bits 0-7 and 40-47) indicating a conversion is to be performed; a first vector register (V_1) field 564 (e.g., bits 8-11) used to designate a first vector register; a second vector register (V_2) field 566 (e.g., bits 12-15) used to designate a second vector register; a first mask (M_4) field 568 (e.g., bits 28-31); a second mask (M_3) field 570 (e.g., bits 32-35); and a register extension bit (RXB) field 572 (e.g., bits 36-39), each of which is described herein. In one embodiment, the fields are separate and independent from one another; however, in other embodiments, more than one field may be combined. Further information regarding these fields is described below.

[00203] In one embodiment, vector register (V_1) field 564 is used to indicate a vector register that is to store the first operand. The first operand is a result obtained from the conversion to be performed. The second operand is contained in a vector register specified using vector register (V_2) field 566. In one example, each of vector register fields 564, 566 is used with RXB field 572 to designate the vector register. In another embodiment, a vector register field is used to designate one or more vector registers.

[00204] In one example, M_4 field 568 specifies the source data type. Example data types include, for instance:

[00205]	<u>M_4</u>	<u>Source Data Type</u>
[00206]	0	Neural Network Processing-data-type-1
[00207]	1	Neural Network Processing-data-type-2
[00208]	2-15	Reserved

[00209] If a reserved value is specified, an IEEE-inexact exception is recognized.

[00210] In one example, M_3 field 570 specifies the target data type. Example data types include, for instance:

[00211]	M_3	<u>Target Data Type</u>
[00212]	0	Reserved
[00213]	1	Binary Floating-Point tiny format
[00214]	2-15	Reserved

[00215] If a reserved value is specified, an IEEE-inexact exception is recognized.

[00216] In one embodiment, if the combination of non-reserved M_3 and M_4 values is not supported by the model (e.g., the select machine or processor of a particular model), an IEEE-inexact exception is recognized, in one example. For instance, the IEEE-inexact exception flag, also referred to herein as an indicator, is set, e.g., to one. If an IEEE-inexact exception is recognized, the first operand is, e.g., unpredictable. In one example, the first operand is set to a default value (e.g., zeros, the value to be converted, or another value). In other embodiments, the IEEE-inexact exception is recognized only for certain invalid data types, such as an invalid machine-specific data type (e.g., NNP-data-type-1, NNP-data-type-2), but not for other data types, such as an invalid standard data type (e.g., a binary floating-point format). Other examples are also possible. Further, in one embodiment, if the IEEE-inexact exception is recognized for only certain invalid data types (e.g., invalid machine-specific data types) and not other invalid data types (e.g., invalid standard types), then a specification exception may be recognized for the other invalid data types. Other variations are possible.

[00217] Although particular values and formats are provided as examples for the source and target data types, additional, fewer and/or other values and/or formats may be used.

[00218] In execution of one embodiment of the Vector Floating Point Convert From Neural Network Processing instruction, the specified source data type elements of the second operand are rounded to the precision of a target data type number. The result is converted to the specified target data type and placed in the corresponding element of the first operand location. In accordance with an aspect of the present invention, the rounding mode used for rounding is based on the machine-specific data type of the source data type. Thus, in one

example, the rounding mode is the mode assigned to the Neural Network Processing-data-type-1, since this is the machine-specific data type specified for use in the conversion by the instruction. If another data type is selected (e.g., NNP-data-type-2 or another data type), the rounding mode is the rounding mode assigned to that other data type.

[00219] In execution of one embodiment of the instruction, the condition code remains unchanged. Possible IEEE exceptions include invalid operation, overflow, underflow and inexact.

[00220] Example Results of a Vector Floating Point Convert From Neural Network Processing is indicated below:

Operand Element (a)	Invalid Op. Mask (FPC 0.0)	Overflow Mask (FPC 0.2)	Underflow Mask (FPC 0.3)	Results
$v < -N_{\max}$	-	0	-	$T(-\infty)$, $SF_o \leftarrow 1$
$v < -N_{\max}$	-	1	-	PIV(3)
$-N_{\max} \leq v \leq -N_{\min}$	-	-	-	$(T(g))$
$-N_{\min} < v \leq -D_{\min}$, $g = -N_{\min}$	-	-	0	$T(-N_{\min})$, $SF_u \leftarrow 1$
$-N_{\min} < v \leq -D_{\min}$, $g = -N_{\min}$	-	-	1	PIV(4)
$-N_{\min} < v \leq -D_{\min}$, $g < -N_{\min}$	-	-	0	$T(d)$, $SF_u \leftarrow 1$
$-N_{\min} < v \leq -D_{\min}$, $g < -N_{\min}$	-	-	1	PIV(4)
$-D_{\min} < v \leq -D_{\min}/2$	-	-	0	$T(-D_{\min})$, $SF_u \leftarrow 1$
$-D_{\min} < v \leq -D_{\min}/2$	-	-	1	PIV(4)
$-D_{\min}/2 < v < 0$	-	-	0	$T(-0)$, $SF_u \leftarrow 1$
$-D_{\min}/2 < v < 0$	-	-	1	PIV(4)
-0	-	-	-	$T(-0)$
+0	-	-	-	$T(+0)$
$0 < v < D_{\min}/2$	-	-	0	$T(+0)$, $SF_u \leftarrow 1$
$0 < v < D_{\min}/2$	-	-	1	PIV(4)
$D_{\min}/2 \leq v < D_{\min}$	-	-	0	$T(+D_{\min})$, $SF_u \leftarrow 1$

$D_{min}/2 \leq v < D_{min}$	-	-	1	PIV(4)
$D_{min} \leq v < N_{min}, g < N_{min}$	-	-	0	T(d), SFu←-1
$D_{min} \leq v < N_{min}, g < N_{min}$	-	-	1	PIV(4)
$D_{min} \leq v < N_{min}, g = N_{min}$	-	-	0	T(+Nmin), SFu←-1
$D_{min} \leq v < N_{min}, g + N_{min}$	-	-	1	PIV(4)
$N_{min} \leq v \leq N_{max}$	-	-	-	T(g)
$N_{max} < v$	-	0	-	T(+∞), SFO←-1
$N_{max} < v$	-	1	-	PIV(3)
+NINF	0	-	-	T(dNaN), SFi←-1
-NINF	1	-	-	PIV(1)

Where:

-	The results do not depend on this condition or mask bit.
d	The denormalized value. The value derived when the precise intermediate value(v) is rounded to the format of the target, including both precision and bounded exponent range.
Dmin	Smallest (in magnitude) representable subnormal number in the target format.
Nmax	Largest (in magnitude) representable finite number in the target format.
Nmin	Smallest (in magnitude) representable normal number in the target format.
PIV(h)	Program interruption for vector-processing exception with VXC (Vector Exception Code) containing the element index and a VIC (Vector Index) of h in hex.
v	Precise intermediate floating-point value which is the source value (a).
g	The precision-rounded value. The value derived when the precise intermediate value (v) is rounded to the precision of the target, but assuming an unbounded exponent range. The rounding mode used is round to nearest with ties away from 0.
SFi	IEEE invalid-operation flag, FPC (Floating-Point Control Register) 1.0.

SFo	IEEE overflow flag, FPC 1.2.
SFu	IEEE underflow flag, FPC 1.3.
dNaN	Default NaN.
T(x)	The value x is converted to the target format and placed at the target operand element location.

[00221] Another example of a conversion instruction is a Vector Floating Point Convert To Neural Network Processing instruction. As depicted in FIG. 5E, a Vector Floating Point Convert to Neural Network Processing instruction 580 has a VRR format indicating a vector register to register operand with an extended opcode. It includes, for instance, operation code (opcode) fields 582a, 582b (e.g., bits 0-7 and 40-47) indicating a conversion to be performed; a first vector register (V_1) field 584 (e.g., bits 8-11) used to designate a first vector register; a second vector register (V_2) field 586 (e.g., bits 12-15) used to designate a second vector register; a first mask (M_4) field 588 (e.g., bits 28-31); a second mask (M_3) field 590 (e.g., bits 32-35); and a register extension bit (RXB) field 592 (e.g., bits 36-39), each of which is described herein. In one embodiment, the fields are separate and independent from one another; however, in other embodiments, more than one field may be combined. Further information regarding these fields is described below.

[00222] In one embodiment, vector register (V_1) field 584 is used to indicate a vector register that is to store the first operand. The first operand is a result obtained from the conversion. The second operand is contained in a vector register specified using vector register (V_2) field 586. In one example, each of vector register fields 584, 586 is used with RXB field 592 to designate the vector register. In another embodiment, a vector register field is used to designate one or more vector registers.

[00223] In one example, M_4 field 588 specifies the source data type. Example data types include, for instance:

[00224]	<u>M_4</u>	<u>Source Data Type</u>
[00225]	0	Reserved
[00226]	1	Binary Floating-Point tiny format
[00227]	2-15	Reserved

[00228] If a reserved value is specified, an IEEE-inexact exception is recognized.

[00229] In one example, M_3 field 590 specifies the target data type. Example data types include, for instance:

[00230]	M_3	<u>Target Data Type</u>
[00231]	0	Neural Network Processing-data-type-1
[00232]	1	Neural Network Processing-data-type-2
[00233]	2-15	Reserved

[00234] If a reserved value is specified, an IEEE-inexact exception is recognized.

[00235] In one embodiment, if the combination of non-reserved M_3 and M_4 values is not supported by the model (e.g., the select machine or processor of a particular model), an IEEE-inexact exception is recognized, in one example. For instance, the IEEE-inexact exception flag, also referred to herein as an indicator, is set, e.g., to one. If an IEEE-inexact exception is recognized, the first operand is, e.g., unpredictable. In one example, the first operand is set to a default value (e.g., zeros, the value to be converted, or another value). In other embodiments, the IEEE-inexact exception is recognized only for certain invalid data types, such as an invalid machine-specific data type (e.g., NNP-data-type-1, NNP-data-type-2), but not for other data types, such as an invalid standard data type (e.g., a binary floating-point format). Other examples are also possible. Further, in one embodiment, if the IEEE-inexact exception is recognized for only certain invalid data types (e.g., invalid machine-specific data types) and not other invalid data types (e.g., invalid standard types), then a specification exception may be recognized for the other invalid data types. Other variations are possible.

[00236] Although particular values and formats are provided as examples for the source and target data types, additional, fewer and/or other values and/or formats may be used.

[00237] In execution of one embodiment of the Vector Floating Point Convert to Neural Network Processing instruction, the binary floating-point elements of the second operand are rounded to the precision of the target data type number. The result is converted to the target data type and placed in the corresponding element of the first operand location. In accordance with an aspect of the present invention, the rounding mode used for rounding is based on the machine-specific data type of the target data type. Thus, in one example, the rounding mode is the mode assigned to the Neural Network Processing-data-type-1, since this is the machine-specific data type specified for use in the conversion by the instruction. If another data type is

selected (e.g., NNP-data-type-2 or another data type), the rounding mode is the rounding mode assigned to that other data type.

[00238] In one execution of one embodiment of the instruction, the condition code remains unchanged. Possible IEEE exceptions include invalid operation, overflow and inexact.

[00239] Example results of a Vector Floating-Point Convert To Neural Network Processing is indicated below:

Operand Element (a)	Invalid Op.	Overflow Mask	Results
	Mask (FPC 0.0)	(FPC 0.2)	
$-N_{\max} \leq v < -N_{\min}$	-	-	T(g)
$-N_{\min} \leq v \leq -D_{\min}$	-	-	T(v)
-0	-	-	T(-0)
+0	-	-	T(+0)
$D_{\min} \leq v \leq N_{\min}$	-	-	T(v)
$N_{\min} < v \leq N_{\max}$	-	-	T(g)
$+\infty$	-	0	T($\pm N_{\text{INF}}$), $SF_o \leftarrow 1$
$+\infty$	-	1	PIV(3)
+NaN	0	-	T($\pm N_{\text{INF}}$), $SF_i \leftarrow 1$
+NaN	1	-	PIV(1)

Where:

-	The results do not depend on this condition or mask bit.
$D_{\min}$	Smallest (in magnitude) representable subnormal number in the source format.
$N_{\max}$	Largest (in magnitude) representable subnormal number in the source format.
$N_{\min}$	Smallest (in magnitude) representable normal number in the source format.
PIV(h)	Program interruption for vector-processing exception with VXC (Vector Exception Code) containing the element index and a VIC (Vector Index) of h in hex.
v	Precise intermediate floating-point value which is the source value (a).

g	The precision-rounded value. The value derived when the precise intermediate value (v) is rounded to the precision of the target, but assuming an unbounded exponent range. The rounding mode used in round to nearest with ties away from 0.
SFi	IEEE invalid-operation flag, FPC (Floating-Point Control Register) 1.0.
SFo	IEEE overflow flag, FPC 1.2.
T(x)	The value x is converted to the target format and placed at the target operand element location.

[00240] One or more of the above conversion instructions is used, in accordance with one or more aspects of the present invention, to provide data to a processor of a select machine model in a format, e.g., optimized for that processor/machine model, such that the processor of that machine model can perform functions using that data. The processor of that machine model provides output in that data type, which may then be converted using one or more of the above conversion instructions to a standard data type or other data type for use by programs and/or other processors.

[00241] In one example, data is converted (and may be rounded) by, e.g., a general-purpose processor, from a standard data type (e.g., a binary floating-point format) to a machine-specific data type (e.g., NNP-data-type-1, NNP-data-type-2) of a select processor (e.g., a special-purpose processor) of a particular machine model. One or more functions are performed by the select processor using the converted data. When performance of the one or more functions is complete, the converted data, in e.g., the machine-specific format, is converted back to the standard format by, e.g., a general-purpose processor, for use, e.g., by the general-purpose processor, other processors and/or other processing.

[00242] As one example, the select processor is a co-processor, such as a neural network processor (e.g., neural network processor 31) used to perform, e.g., tensor computations and/or other operations, and data converted to the NNP-data-type-1 format or NNP-data-type-2, as examples, is provided to the select processor via, for instance, an instruction, such as a Neural Network Processing Assist instruction, an example of which is described below.

[00243] One example of the Neural Network Processing Assist instruction is described with reference to FIG. 6A. In one example, the instruction is initiated on a general-purpose processor (e.g., processor 13) and a function specified by the instruction is either executed on

the general-purpose processor and/or a special-purpose processor (e.g., neural network processor 31) depending on the function. The instruction then completes on the general-purpose processor. In other examples, the instruction is initiated, executed and completed on one or more general-purpose processors or one or more special-purpose processors. Other variations are possible. In the description herein, specific locations, specific fields and/or specific sizes of the fields are indicated (e.g., specific bytes and/or bits). However, other locations, fields and/or sizes may be provided. Further, although the setting of a bit to a particular value, e.g., one or zero, may be specified, this is only an example. The bit, if set, may be set to a different value, such as the opposite value or to another value, in other examples. Many variations are possible.

[00244] In one example, referring to FIG. 6A, a Neural Network Processing Assist instruction 600 has an RRE format that denotes a register and register operation with an extended operation code (opcode). As shown in FIG. 6A, in one example, Neural Network Processing Assist instruction 600 includes an operation code (opcode) field 602 (e.g., bits 0-15) indicating a neural network processing assist operation. In one example, bits 16-31 of the instruction are reserved and are to contain zeros.

[00245] In one example, the instruction uses a plurality of general registers implicitly specified by the instruction. For instance, Neural Network Processing Assist instruction 600 uses implied registers general register 0 and general register 1, examples of which are described with reference to FIGS. 6B and 6D, respectively.

[00246] Referring to FIG. 6B, in one example, general register 0 includes a function code field, and status fields which may be updated upon completion of the instruction. As an example, general register 0 includes a response code field 610 (e.g., bits 0-15), an exception flags field 612 (e.g., bits 24-31) and a function code field 614 (e.g., bits 56-63). Further, in one example, bits 16-23 and 32-55 of general register 0 are reserved and are to contain zeros. One or more fields are used by a particular function performed by the instruction. Not all fields are used by all of the functions, in one example. Each of the fields is described below:

[00247] Response Code (RC) 610: This field (e.g., bit positions 0-15) contains the response code. When execution of the Neural Network Processing Assist instruction completes with a condition code of, e.g., one, a response code is stored. When an invalid input condition is encountered, a non-zero value is stored to the response code field, which indicates the cause of

the invalid input condition recognized during execution and a selected condition code, e.g., 1, is set. The codes stored to the response code field are defined, as follows, in one example:

[00248]	<u>Response Code</u>	<u>Meaning</u>
[00249]	0001	The format of the parameter block, as specified by the parameter block version number, is not supported by the model.
[00250]	0002	The specified function is not defined or installed on the machine.
[00251]	0010	A specified tensor data layout format is not supported.
[00252]	0011	A specified tensor data type is not supported.
[00253]	0012	A specified single tensor dimension is greater than the maximum dimension index size.
[00254]	0013	The size of a specified tensor is greater than the maximum tensor size.
[00255]	0014	The specified tensor address is not aligned on a 4 K-byte boundary.
[00256]	0015	The function-specific-save-area-address is not aligned on a 4 K-byte boundary.
[00257]	F000-FFFF	Function specific response codes. These response codes are defined for certain functions.

[00258] Exception Flags (EF) 612: This field (e.g., bit positions 24-31) includes the exception flags. If an exception condition is detected during execution of the instruction, the corresponding exception flag control (e.g., bit) will be set to, e.g., one; otherwise, the control remains unchanged. The exception flags field is to be initialized to zero prior to the first invocation of the instruction. Reserved flags are unchanged during the execution of the instruction. The flags stored to the exception flags field are defined as follows, in one example:

[00259]	<u>EF (Bit)</u>	<u>Meaning</u>
---------	-----------------	----------------

[00260] 0 Range Violation. This flag is set when a non-numeric value was either detected in an input tensor or stored to the output tensor. This flag is, e.g., only valid when the instruction completes with condition code, e.g., 0.

[00261] 1-7 Reserved.

[00262] Function Code (FC) 614: This field (e.g., bit positions 56-63) includes the function code. Examples of assigned function codes for the Neural Network Processing Assist instruction are depicted in FIG. 6C. All other function codes are unassigned. If an unassigned or uninstalled function code is specified, a response code of, e.g., 0002 hex and a select condition code, e.g., 1, are set. This field is not modified during execution.

[00263] As indicated, in addition to general register 0, the Neural Network Processing Assist instruction also uses general register 1, an example of which is depicted in FIG. 6D. As examples, bits 40-63 in the 24-bit addressing mode, bits 33-63 in the 31-bit addressing mode or bits 0-63 in the 64-bit addressing mode include an address of a parameter block 620. The contents of general register 1 specify, for instance, a logical address of a leftmost byte of the parameter block in storage. The parameter block is to be designated on a doubleword boundary; otherwise, a specification exception is recognized. For all functions, the contents of general register 1 are not modified.

[00264] In the access register mode, access register 1 specifies an address space containing the parameter block, input tensors, output tensors and the function specific save area, as an example.

[00265] In one example, the parameter block may have different formats depending on the function specified by the instruction to be performed. For instance, the query function has a parameter block of one format and other functions of the instruction have a parameter block of another format. In another example, all functions use the same parameter block format. Other variations are also possible.

[00266] As examples, a parameter block and/or the information in the parameter block is stored in memory, in hardware registers and/or in a combination of memory and/or registers. Other examples are also possible.

[00267] One example of a parameter block used by a query function, such as the NNPA-Query Available Functions (QAF) operation, is described with reference to FIG. 6E. As shown, in one example, a NNPA-Query Available Functions parameter block 630 includes, for instance:

[00268] Installed Functions Vector 632: This field (e.g., bytes 0-31) of the parameter block includes the installed functions vector. In one example, bits 0-255 of the installed functions vector correspond to function codes 0-255, respectively, of the Neural Network Processing Assist instruction. When a bit is, e.g., one, the corresponding function is installed; otherwise, the function is not installed.

[00269] Installed Parameter Block Formats Vector 634: This field (e.g., bytes 32-47) of the parameter block includes the installed parameter block formats vector. In one example, bits 0-127 of the installed parameter block formats vector correspond to parameter block formats 0-127 for the non-query functions of the Neural Network Processing Assist instruction. When a bit is, e.g., one, the corresponding parameter block format is installed; otherwise, the parameter block format is not installed.

[00270] Installed Data Types 636: This field (e.g., bytes 48-49) of the parameter block includes the installed data types vector. In one example, bits 0-15 of the installed data types vector correspond to the data types being installed. When a bit is, e.g., one, the corresponding data type is installed; otherwise, the data type is not installed. Example data types include (additional, fewer and/or other data types are possible):

[00271]	<u>Bit</u>	<u>Data Type</u>
[00272]	0	NNP-data-type-1
[00273]	1	NNP-data-type-2
[00274]	2-15	Reserved

[00275] Installed Data Layout Formats 638: This field (e.g., bytes 52-55) of the parameter block includes the installed data layout formats vector. In one example, bits 0-31 of the installed data layout formats vector correspond to data layout formats being installed. When a bit is, e.g., one, the corresponding data layout format is installed; otherwise, the data layout format is not installed. Example data layout formats include (additional, fewer and/or other data layout formats are possible):

[00276] Bit Data Layout Format

[00277] 0 4D-feature tensor

[00278] 1 4D-kernel tensor

[00279] 2-31 Reserved

[00280] Maximum Dimension Index Size 640: This field (e.g., bytes 60-63) of the parameter block includes, e.g., a 32-bit unsigned binary integer that specifies a maximum number of elements in a specified dimension index size for any specified tensor. In another example, the maximum dimension index size specifies a maximum number of bytes in a specified dimension index size for any specified tensor. Other examples are also possible.

[00281] Maximum Tensor Size 642: This field (e.g., bytes 64-71) of the parameter block includes, e.g., a 32-bit unsigned binary integer that specifies a maximum number of bytes in any specified tensor including any pad bytes required by the tensor format. In another example, the maximum tensor size specifies a maximum number of total elements in any specified tensor including any padding required by the tensor format. Other examples are also possible.

[00282] Installed-NNP-Data-Type-1-Conversions Vector 644: This field (e.g., bytes 72-73) of the parameter block includes the installed-NNP-Data-Type-1-conversions vector. In one example, bits 0-15 of the installed-NNP-Data-Type-1-conversions vector correspond to installed data type conversion from/to NNP-data-type-1 format. When a bit is one, the corresponding conversion is installed; otherwise, the conversion is not installed. Additional, fewer and/or other conversions may be specified.

[00283] Bit Data Type

[00284] 0 Reserved

[00285] 1 BFP tiny format

[00286] 2 BFP short format

[00287] 3-15 Reserved

[00288] In one or more embodiments, an Installed-NNP-Data-Type-Conversions Vector is added to the query parameter block for each machine-specific data type supported. For instance, in one example, an Installed-NNP-Data-Type-2-Conversions Vector is provided for NNP-data-type-2. Other variations are possible.

[00289] Although one example of a parameter block for a query function is described with reference to FIG. 6E, other formats of a parameter block for a query function, including the NNPA-Query Available Functions operation, may be used. The format may depend, in one example, on the type of query function to be performed. Further, the parameter block and/or each field of the parameter block may include additional, fewer and/or other information.

[00290] In addition to the parameter block for a query function, in one example, there is a parameter block format for non-query functions, such as non-query functions of the Neural-Network Processing Assist instruction. One example of a parameter block used by a non-query function, such as a non-query function of the Neural Network Processing Assist instruction, is described with reference to FIG. 6F.

[00291] As shown, in one example, a parameter block 650 employed by, e.g., the non-query functions of the Neural Network Processing Assist instruction includes, for instance:

[00292] Parameter Block Version Number 652: This field (e.g., bytes 0-1) of the parameter block specifies the version and size of the parameter block. In one example, bits 0-8 of the parameter block version number are reserved and are to contain zeros, and bits 9-15 of the parameter block version number contain an unsigned binary integer specifying the format of the parameter block. The query function provides a mechanism of indicating the parameter block formats available. When the size or format of the parameter block specified is not supported by the model, a response code of, e.g., 0001 hex is stored in general register 0 and the instruction completes by setting a condition code, e.g., condition code 1. The parameter block version number is specified by the program and is not modified during the execution of the instruction.

[00293] Model Version Number 654: This field (e.g., byte 2) of the parameter block is an unsigned binary integer identifying the model which executed the instruction (e.g., the particular non-query function). When a continuation flag (described below) is one, the model version number may be an input to the operation for the purpose of interpreting the contents of

a continuation state buffer field (described below) of the parameter block to resume the operation.

[00294] Continuation Flag 656: This field (e.g., bit 63) of the parameter block, when, e.g., one, indicates the operation is partially complete and the contents of the continuation state buffer may be used to resume the operation. The program is to initialize the continuation flag to zero and not modify the continuation flag in the event the instruction is to be re-executed for the purpose of resuming the operation; otherwise, results are unpredictable.

[00295] If the continuation flag is set at the beginning of the operation and the contents of the parameter block have changed since the initial invocation, results are unpredictable.

[00296] Function-specific-save-area-address 658: This field (e.g., bytes 56-63) of the parameter block includes the logical address of the function specific save area. In one example, the function-specific-save-area-address is to be aligned on a 4 K-byte boundary; otherwise, a response code of, e.g., 0015 hex is set in general register 0 and the instruction completes with a condition code of, e.g., 1. The address is subject to the current addressing mode. The size of the function specific save area depends on the function code.

[00297] When the entire function specific save area overlaps the program event recording (PER) storage area designation, a PER storage alteration event is recognized, when applicable, for the function specific save area. When only a portion of the function specific save area overlaps the PER storage area designation, it is model-dependent which of the following occurs:

[00298] * A PER storage alteration event is recognized, when applicable, for the entire function specific save area.

[00299] * A PER storage alteration event is recognized, when applicable, for the portion of the function specific save area that is stored.

[00300] When the entire parameter block overlaps the PER storage area designation, a PER storage alteration event is recognized, when applicable, for the parameter block. When only a portion of the parameter block overlaps the PER storage area designation, it is model-dependent which of the following occurs:

[00301] * A PER storage alteration event is recognized, when applicable, for the entire parameter block.

[00302] * A PER storage alteration event is recognized, when applicable, for the portion of the parameter block that is stored.

[00303] A PER zero-address detection event is recognized, when applicable, for the parameter block. Zero address detection does not apply to the tensor addresses or the function-specific-save-area-address, in one example.

[00304] Output Tensor Descriptors (e.g., 1-2) 660/Input Tensor Descriptors (e.g., 1-3) 665: One example of a tensor descriptor is described with reference to FIG. 6G. In one example, a tensor descriptor 660, 665 includes:

[00305] Data Layout Format 682: This field (e.g., byte 0) of the tensor descriptor specifies the data layout format. Valid data layout formats include, for instance (additional, fewer and/or other data layout formats are possible):

[00306]	<u>Format</u>	<u>Description</u>	<u>Alignment (bytes)</u>
[00307]	0	4D-feature tensor	4096
[00308]	1	4D-kernel tensor	4096
[00309]	2-255	Reserved	--

[00310] If an unsupported or reserved data layout format is specified, the response code of, e.g., 0010 hex, is stored in general register 0 and the instruction completes by setting condition code, e.g., 1.

[00311] Data Type 684: This field (e.g., byte 1) specifies the data type of the tensor. Examples of supported data types are described below (additional, fewer and/or other data types are possible):

[00312]	<u>Value</u>	<u>Data Type</u>	<u>Data Size (bits)</u>
[00313]	0	NNP data-type-1	16
[00314]	1	NNP data-type-2	16
[00315]	2-255	Reserved	--

[00316] If an unsupported or reserved data type is specified, a response code of, e.g., 0011 hex is stored in general register 0 and the instruction completes by setting condition code, e.g., 1.

[00317] Dimension 1-4 Index Size 686: Collectively, dimension index sizes one through four specify the shape of a 4D tensor. Each dimension index size is to be greater than zero and less than or equal to the maximum dimension index size (640, FIG. 6E); otherwise, a response code of, e.g., 0012 hex is stored in general register 0 and the instruction completes by setting condition code, e.g., 1. The total tensor size is to be less than or equal to the maximum tensor size (642, FIG. 6E); otherwise, a response code, e.g., 0013 hex is stored in general register 0 and the instruction completes by setting condition code, e.g., 1.

[00318] In one example, to determine the number of a 4D-feature tensor with elements of NNPA-data-type-1 (i.e., total tensor size), the following is used: $\text{dimension-index-4} * \text{dimension-index-3} * \text{ceil}(\text{dimension-index-2}/32) * 32 * \text{ceil}(\text{dimension-index-1}/64) * 64 * 2$.

[00319] Tensor Address 688: This field (e.g., bytes 24-31) of the tensor descriptor includes a logical address of the leftmost byte of the tensor. The address is subject to the current addressing mode.

[00320] If the address is not aligned on the boundary of the associated data layout format, a response code of, e.g., 0014 hex, is stored in general register 0 and the instruction completes by setting condition code, e.g., 1.

[00321] In the access register mode, access register 1 specifies the address space containing all active input and output tensors in storage.

[00322] Returning to FIG. 6F, parameter block 650 further includes, in one example, function-specific-parameters 1-5 (670), which may be used by specific functions. Further, parameter block 650 includes, in one example, a continuation state buffer field 675, which includes data (or a location of data) to be used if operation of this instruction is to be resumed.

[00323] As an input to the operation, reserved fields of the parameter block should contain zeros. When the operation ends, reserved fields may be stored as zeros or remain unchanged.

[00324] Although one example of a parameter block for a non-query function is described with reference to FIG. 6F, other formats of a parameter block for a non-query function, including a non-query function of the Neural Network Processing Assist instruction, may be

used. The format may depend, in one example, on the type of function to be performed. Further, although one example of a tensor descriptor is described with reference to FIG. 6G, other formats may be used. Further, different formats for input and output tensors may be used. Other variations are possible.

[00325] Further details regarding one or more functions supported by one embodiment of the Neural Network Processing Assist instruction are described below:

[00326] Function Code 0: NNPA-QAF (Query Available Functions)

[00327] The Neural Network Processing Assist (NNPA) query function provides a mechanism to indicate selected information, such as, for instance, the availability of installed functions, installed parameter block formats, installed data types, installed data layout formats, maximum dimension index size and maximum tensor size. The information is obtained and placed in a selected location, such as a parameter block (e.g., parameter block 630). When the operation ends, reserved fields of the parameter block may be stored as zeros or may remain unchanged.

[00328] In execution of one embodiment of the query function, a processor, such as general-purpose processor 13, obtains information relating to a specific model of a selected processor, such as a specific model of a neural network processor, such as neural network processor 31. The obtained information is placed in a parameter block (e.g., parameter block 630) or other structure that is accessible to and/or for use with one or more applications that may use this information in further processing. In one example, the parameter block and/or information of the parameter block is maintained in memory. In other embodiments, the parameter block and/or information may be maintained in one or more hardware registers. As another example, the query function may be a privileged operation executed by the operating system, which makes available an application programming interface to make this information available to the application or non-privileged program. In yet a further example, the query function is performed by a special-purpose processor, such as neural network processor 31. Other variations are possible.

[00329] The information is obtained, e.g., by the firmware of the processor executing the query function. The firmware has knowledge of the attributes of the specific model of the specific processor (e.g., neural network processor). This information may be stored in, e.g., a

control block, register and/or memory and/or otherwise be accessible to the processor executing the query function.

[00330] The obtained information includes, for instance, model-dependent detailed information regarding at least one or more data attributes of the specific processor, including, for instance, one or more installed or supported data types, one or more installed or supported data layout formats and/or one or more installed or supported data sizes of the selected model of the specific processor. This information is model-dependent in that other models (e.g., previous models and/or future models) may not support the same data attributes, such as the same data types, data sizes and/or data layout formats. When execution of the query function (e.g., NNPA-QAF function) completes, condition code 0, as an example, is set. Condition codes 1, 2 and 3 are not applicable to the query function, in one example. Further information relating to the obtained information is described below.

[00331] In one example, the query function obtains an indication of the data types installed on the model-dependent processor and places the indication in the parameter block by, e.g., setting one or more bits in installed data types field 636 of parameter block 630. Example data types that may be installed are the NNP-data-type-1 data type and the NNP-data-type-2 data type, examples of which are described above. Additional, fewer and/or other data types may be installed. A data type to be supported may have one or more assigned bits in the query parameter block (e.g., in installed data types field 636). For instance, machine-specific or non-standard data types supported by a particular processor are indicated in the installed data types field but standard data types are not indicated. In other embodiments, one or more standard data types are also indicated. Other variations are possible.

[00332] In one particular example, bit 0 of installed data types field 636 is reserved for the NNP-data-type-1 data type, and when it is set to, e.g., 1, it indicates that the processor supports NNP-data-type-1; and bit 1 is reserved for NNP-data-type-2 and when it is set to, e.g., 1, it indicates that the processor supports NNP-data-type-2. As an example, the bit vector of installed data types is configured to represent up to 16 data types, in which a bit is assigned to each data type. However, a bit vector in other embodiments may support more or fewer data types. Further, a vector may be configured in which one or more bits are assigned to a data type. Many examples are possible and/or additional, fewer and/or other data types may be supported and/or indicated in the vector.

[00333] Further, in one example, the query function obtains an indication of installed data layout formats (another data attribute) and places the information in the parameter block by, e.g., setting one or more bits in installed data layout formats field 638. Example data layout formats include, for instance, a 4D-feature tensor layout and a 4D-kernel tensor layout. The 4D-feature tensor layout is used, in one example, by the functions indicated herein, and in one example, the convolution function uses the 4D-kernel tensor layout. These data layout formats arrange data in storage for a tensor in a way that increases processing efficiency in execution of the functions of the Neural Network Processing Assist instruction. For instance, to operate efficiently, the Neural Network Processing Assist instruction uses input tensors provided in particular data layout formats. Although example layouts are provided, additional, fewer and/or other layouts may be provided for the functions described herein and/or other functions.

[00334] The use or availability of layouts for a particular processor model is provided by the vector of installed data layout formats (e.g., field 638 of parameter block 630). The vector is, for instance, a bit vector of installed data layout formats that allows the CPU to convey to applications which layouts are supported. For instance, bit 0 is reserved for the 4D-feature tensor layout, and when it is set to, e.g., 1, it indicates that the processor supports a 4D-feature tensor layout; and bit 1 is reserved for the 4D-kernel tensor layout, and when it is set to, e.g., 1, it indicates that the processor supports a 4D-kernel tensor layout. In one example, the bit vector of installed data layout formats is configured to represent up to 16 data layouts, in which a bit is assigned to each data layout. However, a bit vector in other embodiments may support more or fewer data layouts. Further, a vector may be configured in which one or more bits are assigned to data layouts. Many examples are possible. Further details regarding the 4D-feature tensor layout and the 4D-kernel tensor layout are described below. Again, other layouts may be used now or in the future to optimize performance.

[00335] In one example, the Neural Network Processing Assist instruction operates with 4D-tensors, i.e., tensors with 4 dimensions. These 4D-tensors are obtained from generic input tensors described herein in, e.g., row-major, i.e., when enumerating the tensor elements in increasing memory address order, the inner dimension called E1 will be stepped up first through the E1-index-size values starting with 0 through the E1-index-size -1, before the index of the E2 dimension will be increased and the stepping through the E1 dimension is repeated. The index of the outer dimension called the E4 dimension is increased last.

[00336] Tensors that have a lower number of dimensions (e.g., 3D- or 1D-tensors) will be represented as 4D-tensors with one or more dimensions of the 4D-tensor exceeding the original tensor dimensions set to 1.

[00337] The transformation of a row-major generic 4D-tensor with dimensions E4, E3, E2, E1 into a 4D-feature tensor layout (also referred to herein as NNPA data layout format 0 4D-feature tensor) is described herein:

[00338] A resulting tensor can be represented, for instance, as a 4D-tensor of, e.g., 64-element vectors or a 5D-tensor with dimensions:

[00339] $E4, [E1/64], E3, [E2/32] * 32, 64$, where $\lceil \cdot \rceil$ refers to a ceil function. (Stated another way: $E4 * E3 * \text{ceil}(E2/32) * 32 * \text{ceil}(E1/64) * 64$ elements)

[00340] An element $[e4][e3][e2][e1]$ of the generic tensor may be mapped to the following element of the resulting 5D-tensor:

[00341] $[e4][\lfloor e1/64 \rfloor][e3][e2][e1 \text{ MOD } 64]$, where $\lfloor \cdot \rfloor$ is a floor function and mod is modulo. (Stated another way: element $(E3 * e2_limit * e1_limit * e4x) + (e2_limit * e3x * 64) + (e2x * 64) + (\lfloor e1x/64 \rfloor * e2_limit * E3 * 64) + (e1x \text{ mod } 64)$, where $e2_limit = [E2/32] * 32$ and $e1_limit = [E1/64] * 64$.)

[00342] The resulting tensor may be larger than the generic tensor. Elements of the resulting tensor with no corresponding elements in the generic tensor are called pad elements.

[00343] Consider the element $[fe4][fe1][fe3][fe2][fe0]$ of a NNPA data layout format 0 4D-feature tensor of a 64-element vector or its equivalent representation as a 5D-tensor of elements. This element is either a pad element or its corresponding element in the generic 4D tensor with dimensions E4, E3, E2, E1 can be determined with the following formula:

[00344] · if $fe2 \geq E2$ then this is an E2 (or page)-pad element

[00345] · else if $fe1 * 64 + fe0 \geq E1$ then this is an E1 (or row)-pad element

[00346] · else corresponding element in generic 4D tensor is:

[00347] $[fe4][fe3][fe2][fe1 * 64 + fe0]$

[00348] For convolutional neural network based artificial intelligence models, the meaning of the 4 dimensions of a feature tensor can generally be mapped to:

[00349] · E4: N – Size of mini-batch

[00350] · E3: H – Height of the 3D-tensor/image

[00351] · E2: W – Width of the 3D-tensor/image

[00352] · E1: C – Channels or classes of the 3D-tensor

[00353] For machine learning or recurrent neural network based artificial intelligence models, the meaning of the 4 dimensions of a 4D-feature tensor may generally be mapped to:

[00354] · E4: T – Number of time-steps or models

[00355] · E3: Reserved, generally set to 1

[00356] · E2: N_{mb} – Minibatch size

[00357] · E1: L – Features

[00358] The NNPA data layout format 0 provides, e.g., two dimensional data locality with 4k-Bytes blocks of data (pages) as well as 4k-Byte block data alignment for the outer dimensions of the generated tensor.

[00359] Pad element bytes are ignored for the input tensors and unpredictable for output tensors. PER storage-alteration on pad bytes is unpredictable.

[00360] One example of an input data layout for a 4D-feature tensor layout, which has dimensions E1, E2, E3 and E4, is shown in FIGS. 7A-7C, and an example output for the 4D-feature tensor layout is depicted in FIGS. 8A-8C. Referring to FIG. 7A, a 3D-tensor 700 is shown, which has dimensions E1, E2 and E3. In one example, each 3D-tensor includes a plurality of 2D-tensors 702. The numbers in each 2D-tensor 702 describe memory offsets of where each of its elements would be in memory. The inputs are used to lay-out the data of the original tensor (e.g., original 4D-tensor of FIGS. 7A-7C) in memory, as shown in FIGS. 8A-8C, which correspond to FIGS. 7A-7C.

[00361] In FIG. 8A, as an example, a unit of memory 800 (e.g., a memory page) includes a pre-selected number (e.g., 32) of rows 802, each of which is identified by, e.g., e2_page_idx; and each row has a pre-selected number (e.g., 64) of elements 804, each identified by, e.g., e1_page_idx. If a row does not include the pre-selected number of elements, it is padded 806, referred to as row or E1 padding; and if the memory unit does not have a pre-selected number of rows, it is padded 808, referred to as page or E2 padding. As examples, the row padding is, e.g., a pre-specified value, such as zeros, spaces, or other values, etc., and the page padding is, e.g., unpredictable data, existing values, zeros, any value, etc.

[00362] In one example, output elements of a row are provided in memory (e.g., in a page) based on element positions in the E1 direction of its corresponding input. For instance, referring to FIG. 7A, element positions 0, 1 and 2 of the three matrices shown (e.g., element positions at a same location in each matrix) are shown in row 0 of page 0 of FIG. 8A, etc. In this example, the 4D-tensor is small and all of the elements of each 2D-tensor representing the 4D-tensor fits in one page. However, this is only one example. A 2D-tensor may include one or more pages. If a 2D-tensor is created based on a reformatting of a 4D-tensor, then the number of pages of the 2D-tensor is based on the size of the 4D-tensor. In one example, one or more ceil functions are used to determine the number of rows in a 2D-tensor and the number of elements in each row, which will indicate how many pages are to be used. Other variations are possible.

[00363] In addition to the 4D-feature tensor layout, in one example, a neural network processor may support a 4D-kernel tensor layout, which re-arranges the elements of a 4D-tensor to reduce the number of memory accesses and data gathering steps when executing certain artificial intelligence (e.g., neural network processing assist) operations, such as a convolution. As an example, a row-major generic 4D-tensor with dimensions E4, E3, E2, E1 is transformed into a NNPA data layout format 1 4D-kernel tensor (4D-kernel tensor), as described herein:

[00364] A resulting tensor can be represented as a 4D-tensor of, e.g., 64-element vectors or a 5D-tensor with dimensions:

[00365] $[E1/64], E4, E3, [E2/32] * 32, 64$, where $[]$ refers to a ceil function. (Stated another way: $E4 * E3 * \text{ceil}(E2/32) * 32 * \text{ceil}(E1/64) * 64$ elements.)

[00366] An element $[e_4][e_3][e_2][e_1]$ of the generic tensor may be mapped to the following element of the resulting 5D-tensor:

[00367] $\lfloor \frac{e_1}{64} \rfloor [e_4][e_3][e_2][e_1 \bmod 64]$, where $\lfloor \cdot \rfloor$ refers to a floor function and mod is modulo. Stated another way: element $(\lfloor \frac{e_1}{64} \rfloor * E_4 * E_3 * e_2_limit * 64) + (e_4 * E_3 * e_2_limit * 64) + (e_3 * e_2_limit * 64) + (e_2 * 64) + (e_1 \bmod 64)$, where $e_2_limit = \lfloor \frac{E_2}{32} \rfloor * 32$ and $e_1_limit = \lfloor \frac{E_1}{64} \rfloor * 64$.

[00368] The resulting tensor may be larger than the generic tensor. Elements of the resulting tensor with no corresponding elements in the generic tensor are called pad elements.

[00369] Consider the element $[fe_1][fe_4][fe_3][fe_2][fe_0]$ of a NNPA data layout format 1 4D-feature tensor of 64-element vectors or its equivalent representation as a 5D-tensor of elements. This element is either a pad element or its corresponding element in the generic 4D tensor with dimensions E_4, E_3, E_2, E_1 can be determined with the following formula:

[00370] · if $fe_2 \geq E_2$ then this is an E_2 (or page)-pad element

[00371] · else if $fe_1 * 64 + fe_0 \geq E_1$ then this is an E_1 (or row)-pad element

[00372] · else corresponding element in generic 4D tensor is

[00373] $[fe_4][fe_3][fe_2][fe_1 * 64 + fe_0]$

[00374] For convolutional neural network based artificial intelligence models, the meaning of the 4 dimensions of a kernel tensor can generally be mapped to:

[00375] · E_4 : H – Height of the 3D-tensor/image

[00376] · E_3 : W – Width of the 3D-tensor/image

[00377] · E_2 : C – Number of Channels of the 3D-tensor

[00378] · E_1 : K – Number of Kernels

[00379] The NNPA data layout format 1 provides, e.g., two dimensional kernel parallelism within 4k-Byte blocks of data (pages) as well as 4k-Byte block data alignment for the outer dimensions of the generate tensor for efficient processing.

[00380] Pad bytes are ignored for the input tensors. PER storage-alteration on pad bytes is unpredictable.

[00381] Again, although example data layout formats include a 4D-feature tensor layout and a 4D-kernel tensor layout, other data layout formats may be supported by the processor (e.g., neural network processor 31). An indication of supported data layouts is obtained and placed in the query parameter block by setting one or more bits in, e.g., field 638.

[00382] The query parameter block also includes, in accordance with one or more aspects of the present invention, other data attribute information, which includes, e.g., supported size information for the data. A processor, such as a neural network processor, typically has limitations based on internal buffer sizes, processing units, data bus structures, firmware limitations, etc. that can limit the maximum size of tensor dimensions and/or the overall size of a tensor. Therefore, the query function provides fields to convey these limits to applications. For instance, the processor, based on executing the query function, obtains various data sizes, such as a maximum dimension index size (e.g., 65,536 elements) and a maximum tensor size (e.g., 8 GB), and includes this information in fields 640 and 642, respectively, of the parameter block (e.g., parameter block 630). Additional, fewer and/or other size information may also be supported by the processor (e.g., neural network processor 31), and thus, obtained and placed in the parameter block, e.g., fields 640, 642 and/or other fields. In other embodiments, the limitations could be smaller or larger, and/or the sizes may be in other units, such as bytes instead of elements, elements instead of bytes, etc. Further, other embodiments allow for different maximum sizes of each dimension, rather than the same maximum for all dimensions. Many variations are possible.

[00383] In accordance with one or more aspects of the present invention, a query function is provided that conveys detailed information relating to a specific model of a selected processor (e.g., neural network processor 31). The detailed information includes, for instance, model-dependent information relating to a specific processor. (A processor may also support standard data attributes, such as standard data types, standard data layouts, etc., which are implied and not necessarily presented by the query function; although, in other embodiments, the query function may indicate all or various selected subsets of data attributes, etc.) Although example information is provided, other information may be provided in other embodiments. The obtained information, which may be different for different models of a processor and/or different processors, is used to perform artificial intelligence and/or other

processing. The artificial intelligence and/or other processing may employ one or more non-query functions of, e.g., the Neural Network Processing Assist instruction. A specific non-query function employed in the processing is performed by executing the Neural Network Processing Assist instruction one or more times and specifying the non-query specific function.

[00384] As indicated by the query function, example non-query functions supported by the Neural Network Processing Assist instruction include NNPA-ADD (Addition), NNPA-SUB (Subtraction), NNPA-MUL (Multiplication), NNPA-DIV (Division), NNPA-MIN (Minimum), NNPA-MAX (Maximum), NNPA-LOG (Logarithm), NNPA-EXP (Exponent), NNPA-RELU (Rectified Linear Unit), NNPA-TANH, NNPA-SIGMOID, NNPA-SOFTMAX, NNPA-BATCH-NORM (Batch Normalization), NNPA-MAXPOOL2D, NNPA-AVGPOOL2D, NNPA-LSTMACT (Long Short-Term Memory Activation), NNPA-GRUACT (Gated Recurrent Unit Activation), NNPA-CONVOLUTION, NNPA-MATMUL-OP (Matrix Multiplication Operation), and NNPA-MATMUL-OP-BCAST23 (Matrix Multiplication Operation Broadcast 23). Additional, fewer and/or other functions may also be supported.

[00385] One or more of the functions is executed by a processor, such as neural network processor 31. For instance, the Neural Network Processing Assist instruction specifying a particular function to be performed is initiated by a general-purpose processor (e.g., processor 13) and if the function specified by the instruction is a non-query function, information is provided by the general-purpose processor to, e.g., a neural network processor (e.g., neural network processor 31) to allow the neural network processor to perform the function. The information includes, for instance, an indication of the function to be performed and/or one or more addresses of input data (e.g., input tensor data) for the function. In one example, the input data has been converted from, e.g., a standard format, such as a binary floating-point format, to a non-standard format, such as a machine-specific format (e.g., NNP-data-type-1 or NNP-data-type-2) used by, for instance, neural network processor 31. The function is performed using the input data and produces output data (e.g., output tensor data) in the machine-specific format. That data (or a location of the data) is then provided to another processor (e.g., general processor 13) for conversion back to the standard format. The output data in the standard format may then be used in other processing. In other examples, the instruction may be initiated, executed and completed on one or more general-purpose processors or one or more special-purpose processors. Other variations are possible.

[00386] Further details regarding an example non-query function supported by the Neural Network Processing Instruction and performed by, e.g., neural network processor 31, are described below (additional, fewer and/or other functions may be supported in other embodiments):

[00387] Function Code 16: NNPA-ADD (Addition)

[00388] When the NNPA-ADD function is specified, each element of the input tensor 1 described by tensor descriptor 1 is added to the corresponding element of the input tensor 2 described by tensor descriptor 2, and the resulting sum is placed in the corresponding element of the output tensor described by the output tensor descriptor.

[00389] In one example, if the specified data layout in any of the specified tensor descriptors does not specify a 4-D feature tensor (e.g., data layout = 0) or if the data type in any specified tensor descriptor does not specify NNP-data-type-1 (e.g., data type = 0), response code, e.g., 0010 hex or 0011 hex, respectively, is set in general register 0 and the instruction completes with condition code, e.g., 1.

[00390] The shape, the data layout and the data type of input tensor 1, input tensor 2 and the output tensor are to be the same, in one example; otherwise, a general operand data exception is recognized.

[00391] The output tensor descriptor 2, input tensor descriptor 3, function-specific-parameters 1-5 and function-specific-save-area-address fields are ignored, in one example.

[00392] Although the NNPA-ADD function is provided as one example, other and/or additional functions may be specified, as described herein.

[00393] For the Neural Network Processing Assist instruction, in one embodiment, if the output tensor overlaps with any input tensor or the parameter block, results are unpredictable.

[00394] A specification exception is recognized when execution of the Neural Network Processing Assist instruction is attempted and the parameter block is not designated on, e.g., a doubleword boundary, as an example.

[00395] A general operand data exception is recognized when execution of the Neural Network Processing Assist instruction is attempted and there are, for instance, tensor descriptor inconsistencies.

[00396] Resulting Condition Codes for the Neural Network Processing Assist instruction include, for instance: 0 – Normal completion; 1 – Response code is set; 2 --; 3 – CPU-determined amount of data processed.

[00397] In one embodiment, the priority of execution for the Neural Network Processing Assist instruction includes, for instance:

[00398] 1.-7. Exceptions with the same priority as the priority of program interruption conditions for the general case.

[00399] 8.A Condition code 1 due to an unassigned or uninstalled function code specified.

[00400] 8.B Specification exception due to parameter block not designated on doubleword boundary.

[00401] 9. Access exceptions for an access to the parameter block.

[00402] 10. Condition code 1 due to specified format of the parameter block not supported by the model.

[00403] 11.A Condition code 1 due to the specified tensor data layouts are not supported.

[00404] 11.B General operand data exception due to differing data layouts between tensor descriptors.

[00405] 12.A Condition code 1 due to conditions other than those included in items 8.A, 10 and 11.A above and 12.B.1 below.

[00406] 12.B.1 Condition code 1 due to invalid output tensor data type for NNPA-RELU and NNPA-CONVOLUTION.

[00407] 12.B.2 General operand data exception for invalid value for NNPA-RELU function-specific-parameter 1 and NNPA-CONVOLUTION function-specific-parameter 4.

[00408] 13.A Access exceptions for an access to the output tensor.

[00409] 13.B Access exceptions for an access to the input tensors.

[00410] 13.C Access exceptions for an access to the function-specific-save-area.

[00411] 14. Condition code 0.

[00412] As described herein, in one or more aspects, a capability is provided to perform conversions to/from machine-specific data types in which a particular machine-specific data type has an assigned rounding mode such that the rounding mode is not selectable, facilitating the conversions and processing.

[00413] In one example, a machine-specific data type is a data type optimized for use by a particular processor running on a particular hardware machine. For instance, the machine-specific data type may preserve enhanced accuracy and/or be easier for the machine to implement. Examples of machine-specific data types are the NNP-data-type-1 data type and the NNP-data-type-2 data type that may be used by a model of the neural network processor. As an example, the Neural Network Processing Assist instruction operates on a selected machine-specific data type (also referred to as an internal data type) to represent the input/output data for greater efficiency. Thus, the input/output data is converted from a standard format (e.g., IEEE 754 binary floating-point, etc.) to the selected machine-specific format available on the machine (e.g., NNP-data-type-1, NNP-data-type-2, etc.). The machine-specific data types may vary from machine model to machine model in order to optimize for new and/or other processor architectures, and therefore, there may be a plurality of NNP-data-types (e.g., NNP-data-type-1 to NNP-data-type-x) and/or other machine-specific data types.

[00414] One or more aspects of the present invention are inextricably tied to computer technology and facilitate processing within a computer, improving performance thereof. In one or more aspects, a conversion operation is provided that facilitates data conversion, including conversion operations that perform rounding, improving system performance and reducing use of system resources. The converted and/or rounded data, instruction(s) used to perform the conversion operations and/or instructions to use the data may be used in many technical fields, such as in computer processing, medical processing, engineering, automotive technologies, manufacturing, etc. By providing optimizations in converting and/or rounding of conversion operations, these technical fields are improved by, e.g., reducing execution time.

[00415] Further details of one or more embodiments of facilitating processing within a computing environment, as it relates to one or more aspects of the present invention, are described with reference to FIGS. 9A-9B.

[00416] Referring to FIG. 9A, an instruction to convert data from a source data type to a target data type is obtained 900. The source data type is selected from one or more source data types supported by the instruction 902, and the target data type is selected from one or more target data types supported by the instruction 904. Based on a selected data type of one of the source data type and the target data type, a determination is made of a rounding mode for use by the instruction 906. The rounding mode is implicitly set based on the selected data type 908; it is assigned to the selected data type 910. A conversion of the data from the source data type to the target data type is performed 912. The conversion includes performing a rounding operation using the rounding mode implicitly set based on the selected data type 914. The performing the conversion provides a result in the target data type 916. The result is written to a select location 918.

[00417] By providing an implicitly set rounding mode that is assigned to a particular data type, the conversion process is simplified and may improve processing speed by limiting supported rounding modes. It also facilitates the writing of program code that converts data to be used on a select processor, such as a co-processor. Further, by providing selectable source and/or target data types, a reduced number of conversion instructions may be used, saving an operation code space and facilitating the decoding of the instructions and the providing of software to run on an accelerator.

[00418] In one example, the selected data type is the target data type specified using the instruction 920. The target data type is, for instance, a machine-specific data type specified using the instruction 922. The machine-specific data type is, in one example, a neural network processing-data-type-1 data type 924. As an example, the rounding mode assigned to the neural network processing-data-type-1 data type is, for instance, a round to nearest ties away from zero 926. Further, in one example, the source data type is a standard data type specified using the instruction 928.

[00419] By providing an implicitly set rounding mode that is assigned to a particular data type, such as a machine-specified data type, the conversion process is simplified and may improve processing speed by limiting supported rounding modes. Efficiencies are provided,

complexity is reduced, and system performance is increased. Performance of the conversion is improved, improving system performance and data availability.

[00420] In another example, referring to FIG. 9B, the selected data type is the source data type specified using the instruction 930. The source data type is, for instance, a machine-specific data type specified using the instruction 932. The machine-specific data type is, in one example, a neural network processing-data-type-1 data type 934. As an example, the rounding mode assigned to the neural network processing-data-type-1 data type is, for instance, a round to nearest ties away from zero 936. Further, in one example, the target data type is a standard data type specified using the instruction 938.

[00421] As an example, the data to be converted includes source data in the source data type 940. The source data includes a concatenation of data in one register specified using one field of the instruction and other data in another register specified using another field of the instruction 942. The performing the conversion includes, for instance, rounding the source data based on the rounding mode to provide rounded data 944, and converting the rounded data to provide the result in the target data type 946.

[00422] As another example, the data to be converted includes source data in the source data type obtained using the instruction 950. The performing the conversion includes, for instance, rounding the source data based on the rounding mode to provide rounded data in a precision of a target data type number 952, and converting the rounded data to provide the result in the target data type 954.

[00423] Other variations and embodiments are possible.

[00424] Aspects of the present invention may be used by many types of computing environments. Another example of a computing environment to incorporate and use one or more aspects of the present invention is described with reference to FIG. 10. As an example, the computing environment of FIG. 10 is based on the z/Architecture[®] instruction set architecture offered by International Business Machines Corporation, Armonk, New York. The z/Architecture instruction set architecture, however, is only one example architecture. Again, the computing environment may be based on other architectures, including, but not limited to, the Intel[®] x86 architectures, other architectures of International Business Machines Corporation, and/or other architectures of other companies. Intel is a trademark or registered trademark of Intel Corporation or its subsidiaries in the United States and other countries.

[00425] Referring to FIG. 10, in one example, a computing environment 1000 includes, for instance, a computer system 1002 shown, e.g., in the form of a general-purpose computing device. Computer system 1002 may include, but is not limited to, one or more general-purpose processors or processing units 1004 (e.g., central processing units (CPUs)), at least one special-purpose processor, such as a neural network processor 1005, a memory 1006 (a.k.a., system memory, main memory, main storage, central storage or storage, as examples), and one or more input/output (I/O) interfaces 1008, coupled to one another via one or more buses and/or other connections. For instance, processors 1004, 1005 and memory 1006 are coupled to I/O interfaces 1008 via one or more buses 1010, and processors 1004, 1005 are coupled to one another via one or more buses 1011.

[00426] Bus 1011 is, for instance, a memory or cache coherence bus, and bus 1010 represents, e.g., one or more of any of several types of bus structures, including a memory bus or memory controller, a peripheral bus, an accelerated graphics port, and a processor or local bus using any of a variety of bus architectures. By way of example, and not limitation, such architectures include the Industry Standard Architecture (ISA), the Micro Channel Architecture (MCA), the Enhanced ISA (EISA), the Video Electronics Standards Association (VESA) local bus, and the Peripheral Component Interconnect (PCI).

[00427] As examples, one or more special-purpose processors (e.g., neural network processors) may be separate from but coupled to one or more general-purpose processors and/or may be embedded within one or more general-purpose processors. Many variations are possible.

[00428] Memory 1006 may include, for instance, a cache 1012, such as a shared cache, which may be coupled to local caches 1014 of processors 1004 and/or to neural network processor 1005, via, e.g., one or more buses 1011. Further, memory 1006 may include one or more programs or applications 1016 and at least one operating system 1018. An example operating system includes a z/OS[®] operating system, offered by International Business Machines Corporation, Armonk, New York. Other operating systems offered by International Business Machines Corporation and/or other entities may also be used. Memory 1006 may also include one or more computer readable program instructions 1020, which may be configured to carry out functions of embodiments of aspects of the invention.

[00429] Computer system 1002 may communicate via, e.g., I/O interfaces 1008 with one or more external devices 1030, such as a user terminal, a tape drive, a pointing device, a display,

and one or more data storage devices 1034, etc. A data storage device 1034 may store one or more programs 1036, one or more computer readable program instructions 1038, and/or data, etc. The computer readable program instructions may be configured to carry out functions of embodiments of aspects of the invention.

[00430] Computer system 1002 may also communicate via, e.g., I/O interfaces 1008 with network interface 1032, which enables computer system 1002 to communicate with one or more networks, such as a local area network (LAN), a general wide area network (WAN), and/or a public network (e.g., the Internet), providing communication with other computing devices or systems.

[00431] Computer system 1002 may include and/or be coupled to removable/non-removable, volatile/non-volatile computer system storage media. For example, it may include and/or be coupled to a non-removable, non-volatile magnetic media (typically called a "hard drive"), a magnetic disk drive for reading from and writing to a removable, non-volatile magnetic disk (e.g., a "floppy disk"), and/or an optical disk drive for reading from or writing to a removable, non-volatile optical disk, such as a CD-ROM, DVD-ROM or other optical media. It should be understood that other hardware and/or software components could be used in conjunction with computer system 1002. Examples, include, but are not limited to: microcode or millicode, device drivers, redundant processing units, external disk drive arrays, RAID systems, tape drives, and data archival storage systems, etc.

[00432] Computer system 1002 may be operational with numerous other general-purpose or special-purpose computing system environments or configurations. Examples of well-known computing systems, environments, and/or configurations that may be suitable for use with computer system 1002 include, but are not limited to, personal computer (PC) systems, server computer systems, thin clients, thick clients, handheld or laptop devices, multiprocessor systems, microprocessor-based systems, set top boxes, programmable consumer electronics, network PCs, minicomputer systems, mainframe computer systems, and distributed cloud computing environments that include any of the above systems or devices, and the like.

[00433] Another embodiment of a computing environment to incorporate and use one or more aspects of the present invention is described with reference to FIG. 11A. In this example, a computing environment 36 includes, for instance, a native central processing unit (CPU) 37, a memory 38, and one or more input/output devices and/or interfaces 39 coupled to one another via, for example, one or more buses 40 and/or other connections. As examples,

computing environment 36 may include a PowerPC[®] processor offered by International Business Machines Corporation, Armonk, New York; an HP Superdome with Intel[®] Itanium[®] II processors offered by Hewlett Packard Co., Palo Alto, California; and/or other machines based on architectures offered by International Business Machines Corporation, Hewlett Packard, Intel Corporation, Oracle, and/or others. PowerPC is a trademark or registered trademark of International Business Machines Corporation in at least one jurisdiction. Itanium is a trademark or registered trademark of Intel Corporation or its subsidiaries in the United States and other countries.

[00434] Native central processing unit 37 includes one or more native registers 41, such as one or more general purpose registers and/or one or more special purpose registers used during processing within the environment. These registers include information that represents the state of the environment at any particular point in time.

[00435] Moreover, native central processing unit 37 executes instructions and code that are stored in memory 38. In one particular example, the central processing unit executes emulator code 42 stored in memory 38. This code enables the computing environment configured in one architecture to emulate another architecture. For instance, emulator code 42 allows machines based on architectures other than the z/Architecture instruction set architecture, such as PowerPC processors, HP Superdome servers or others, to emulate the z/Architecture instruction set architecture and to execute software and instructions developed based on the z/Architecture instruction set architecture.

[00436] Further details relating to emulator code 42 are described with reference to FIG. 11B. Guest instructions 43 stored in memory 38 comprise software instructions (e.g., correlating to machine instructions) that were developed to be executed in an architecture other than that of native CPU 37. For example, guest instructions 43 may have been designed to execute on a processor based on the z/Architecture instruction set architecture, but instead, are being emulated on native CPU 37, which may be, for example, an Intel Itanium II processor. In one example, emulator code 42 includes an instruction fetching routine 44 to obtain one or more guest instructions 43 from memory 38, and to optionally provide local buffering for the instructions obtained. It also includes an instruction translation routine 45 to determine the type of guest instruction that has been obtained and to translate the guest instruction into one or more corresponding native instructions 46. This translation includes,

for instance, identifying the function to be performed by the guest instruction and choosing the native instruction(s) to perform that function.

[00437] Further, emulator code 42 includes an emulation control routine 47 to cause the native instructions to be executed. Emulation control routine 47 may cause native CPU 37 to execute a routine of native instructions that emulate one or more previously obtained guest instructions and, at the conclusion of such execution, return control to the instruction fetch routine to emulate the obtaining of the next guest instruction or a group of guest instructions. Execution of the native instructions 46 may include loading data into a register from memory 38; storing data back to memory from a register; or performing some type of arithmetic or logic operation, as determined by the translation routine.

[00438] Each routine is, for instance, implemented in software, which is stored in memory and executed by native central processing unit 37. In other examples, one or more of the routines or operations are implemented in firmware, hardware, software or some combination thereof. The registers of the emulated processor may be emulated using registers 41 of the native CPU or by using locations in memory 38. In embodiments, guest instructions 43, native instructions 46 and emulator code 42 may reside in the same memory or may be disbursed among different memory devices.

[00439] Example instructions that may be emulated are the Vector Floating-Point Convert and Lengthen From Neural Network Processing High instruction, Vector Floating-Point Convert and Lengthen From Neural Network Processing Low instruction, Vector Floating-Point Convert and Round to Neural Network Processing instruction, Vector Floating-Point Convert From Neural Network Processing instruction, Vector Floating-Point Convert To Neural Network Processing instruction and/or the Neural Network Processing Assist instruction described herein, in accordance with one or more aspects of the present invention.

[00440] The computing environments described above are only examples of computing environments that can be used. Other environments, including but not limited to, non-partitioned environments, partitioned environments, cloud environments and/or emulated environments, may be used; embodiments are not limited to any one environment. Although various examples of computing environments are described herein, one or more aspects of the present invention may be used with many types of environments. The computing environments provided herein are only examples.

[00441] Each computing environment is capable of being configured to include one or more aspects of the present invention.

[00442] One or more aspects may relate to cloud computing.

[00443] It is to be understood that although this disclosure includes a detailed description on cloud computing, implementation of the teachings recited herein are not limited to a cloud computing environment. Rather, embodiments of the present invention are capable of being implemented in conjunction with any other type of computing environment now known or later developed.

[00444] Cloud computing is a model of service delivery for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, network bandwidth, servers, processing, memory, storage, applications, virtual machines, and services) that can be rapidly provisioned and released with minimal management effort or interaction with a provider of the service. This cloud model may include at least five characteristics, at least three service models, and at least four deployment models.

[00445] Characteristics are as follows:

[00446] On-demand self-service: a cloud consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with the service's provider.

[00447] Broad network access: capabilities are available over a network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).

[00448] Resource pooling: the provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to demand. There is a sense of location independence in that the consumer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or datacenter).

[00449] Rapid elasticity: capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out and rapidly released to quickly scale in. To the

consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.

[00450] Measured service: cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer of the utilized service.

[00451] Service Models are as follows:

[00452] Software as a Service (SaaS): the capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through a thin client interface such as a web browser (e.g., web-based e-mail). The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings.

[00453] Platform as a Service (PaaS): the capability provided to the consumer is to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including networks, servers, operating systems, or storage, but has control over the deployed applications and possibly application hosting environment configurations.

[00454] Infrastructure as a Service (IaaS): the capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls).

[00455] Deployment Models are as follows:

[00456] Private cloud: the cloud infrastructure is operated solely for an organization. It may be managed by the organization or a third party and may exist on-premises or off-premises.

[00457] Community cloud: the cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns (e.g., mission, security requirements, policy, and compliance considerations). It may be managed by the organizations or a third party and may exist on-premises or off-premises.

[00458] Public cloud: the cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services.

[00459] Hybrid cloud: the cloud infrastructure is a composition of two or more clouds (private, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability (e.g., cloud bursting for load-balancing between clouds).

[00460] A cloud computing environment is service oriented with a focus on statelessness, low coupling, modularity, and semantic interoperability. At the heart of cloud computing is an infrastructure that includes a network of interconnected nodes.

[00461] Referring now to FIG. 12, illustrative cloud computing environment 50 is depicted. As shown, cloud computing environment 50 includes one or more cloud computing nodes 52 with which local computing devices used by cloud consumers, such as, for example, personal digital assistant (PDA) or cellular telephone 54A, desktop computer 54B, laptop computer 54C, and/or automobile computer system 54N may communicate. Nodes 52 may communicate with one another. They may be grouped (not shown) physically or virtually, in one or more networks, such as Private, Community, Public, or Hybrid clouds as described hereinabove, or a combination thereof. This allows cloud computing environment 50 to offer infrastructure, platforms and/or software as services for which a cloud consumer does not need to maintain resources on a local computing device. It is understood that the types of computing devices 54A-N shown in FIG. 12 are intended to be illustrative only and that computing nodes 52 and cloud computing environment 50 can communicate with any type of computerized device over any type of network and/or network addressable connection (e.g., using a web browser).

[00462] Referring now to FIG. 13, a set of functional abstraction layers provided by cloud computing environment 50 (FIG. 12) is shown. It should be understood in advance that the components, layers, and functions shown in FIG. 13 are intended to be illustrative only and

embodiments of the invention are not limited thereto. As depicted, the following layers and corresponding functions are provided:

[00463] Hardware and software layer 60 includes hardware and software components. Examples of hardware components include: mainframes 61; RISC (Reduced Instruction Set Computer) architecture based servers 62; servers 63; blade servers 64; storage devices 65; and networks and networking components 66. In some embodiments, software components include network application server software 67 and database software 68.

[00464] Virtualization layer 70 provides an abstraction layer from which the following examples of virtual entities may be provided: virtual servers 71; virtual storage 72; virtual networks 73, including virtual private networks; virtual applications and operating systems 74; and virtual clients 75.

[00465] In one example, management layer 80 may provide the functions described below. Resource provisioning 81 provides dynamic procurement of computing resources and other resources that are utilized to perform tasks within the cloud computing environment. Metering and Pricing 82 provide cost tracking as resources are utilized within the cloud computing environment, and billing or invoicing for consumption of these resources. In one example, these resources may include application software licenses. Security provides identity verification for cloud consumers and tasks, as well as protection for data and other resources. User portal 83 provides access to the cloud computing environment for consumers and system administrators. Service level management 84 provides cloud computing resource allocation and management such that required service levels are met. Service Level Agreement (SLA) planning and fulfillment 85 provide pre-arrangement for, and procurement of, cloud computing resources for which a future requirement is anticipated in accordance with an SLA.

[00466] Workloads layer 90 provides examples of functionality for which the cloud computing environment may be utilized. Examples of workloads and functions which may be provided from this layer include: mapping and navigation 91; software development and lifecycle management 92; virtual classroom education delivery 93; data analytics processing 94; transaction processing 95; and conversion, rounding and/or neural network processing assist processing 96.

[00467] Aspects of the present invention may be a system, a method, and/or a computer program product at any possible technical detail level of integration. The computer program

product may include a computer readable storage medium (or media) having computer readable program instructions thereon for causing a processor to carry out aspects of the present invention.

[00468] The computer readable storage medium can be a tangible device that can retain and store instructions for use by an instruction execution device. The computer readable storage medium may be, for example, but is not limited to, an electronic storage device, a magnetic storage device, an optical storage device, an electromagnetic storage device, a semiconductor storage device, or any suitable combination of the foregoing. A non-exhaustive list of more specific examples of the computer readable storage medium includes the following: a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), a static random access memory (SRAM), a portable compact disc read-only memory (CD-ROM), a digital versatile disk (DVD), a memory stick, a floppy disk, a mechanically encoded device such as punch-cards or raised structures in a groove having instructions recorded thereon, and any suitable combination of the foregoing. A computer readable storage medium, as used herein, is not to be construed as being transitory signals per se, such as radio waves or other freely propagating electromagnetic waves, electromagnetic waves propagating through a waveguide or other transmission media (e.g., light pulses passing through a fiber-optic cable), or electrical signals transmitted through a wire.

[00469] Computer readable program instructions described herein can be downloaded to respective computing/processing devices from a computer readable storage medium or to an external computer or external storage device via a network, for example, the Internet, a local area network, a wide area network and/or a wireless network. The network may comprise copper transmission cables, optical transmission fibers, wireless transmission, routers, firewalls, switches, gateway computers and/or edge servers. A network adapter card or network interface in each computing/processing device receives computer readable program instructions from the network and forwards the computer readable program instructions for storage in a computer readable storage medium within the respective computing/processing device.

[00470] Computer readable program instructions for carrying out operations of the present invention may be assembler instructions, instruction-set-architecture (ISA) instructions, machine instructions, machine dependent instructions, microcode, firmware instructions, state-

setting data, configuration data for integrated circuitry, or either source code or object code written in any combination of one or more programming languages, including an object oriented programming language such as Smalltalk, C++, or the like, and procedural programming languages, such as the "C" programming language or similar programming languages. The computer readable program instructions may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider). In some embodiments, electronic circuitry including, for example, programmable logic circuitry, field-programmable gate arrays (FPGA), or programmable logic arrays (PLA) may execute the computer readable program instructions by utilizing state information of the computer readable program instructions to personalize the electronic circuitry, in order to perform aspects of the present invention.

[00471] Aspects of the present invention are described herein with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems), and computer program products according to embodiments of the invention. It will be understood that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the flowchart illustrations and/or block diagrams, can be implemented by computer readable program instructions.

[00472] These computer readable program instructions may be provided to a processor of a computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks. These computer readable program instructions may also be stored in a computer readable storage medium that can direct a computer, a programmable data processing apparatus, and/or other devices to function in a particular manner, such that the computer readable storage medium having instructions stored therein comprises an article of manufacture including instructions which implement aspects of the function/act specified in the flowchart and/or block diagram block or blocks.

[00473] The computer readable program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other device to cause a series of operational steps to be performed on the computer, other programmable apparatus or other device to produce a computer implemented process, such that the instructions which execute on the computer, other programmable apparatus, or other device implement the functions/acts specified in the flowchart and/or block diagram block or blocks.

[00474] The flowchart and block diagrams in the Figures illustrate the architecture, functionality, and operation of possible implementations of systems, methods, and computer program products according to various embodiments of the present invention. In this regard, each block in the flowchart or block diagrams may represent a module, segment, or portion of instructions, which comprises one or more executable instructions for implementing the specified logical function(s). In some alternative implementations, the functions noted in the blocks may occur out of the order noted in the Figures. For example, two blocks shown in succession may, in fact, be accomplished as one step, executed concurrently, substantially concurrently, in a partially or wholly temporally overlapping manner, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts or carry out combinations of special purpose hardware and computer instructions.

[00475] In addition to the above, one or more aspects may be provided, offered, deployed, managed, serviced, etc. by a service provider who offers management of customer environments. For instance, the service provider can create, maintain, support, etc. computer code and/or a computer infrastructure that performs one or more aspects for one or more customers. In return, the service provider may receive payment from the customer under a subscription and/or fee agreement, as examples. Additionally or alternatively, the service provider may receive payment from the sale of advertising content to one or more third parties.

[00476] In one aspect, an application may be deployed for performing one or more embodiments. As one example, the deploying of an application comprises providing computer infrastructure operable to perform one or more embodiments.

[00477] As a further aspect, a computing infrastructure may be deployed comprising integrating computer readable code into a computing system, in which the code in combination with the computing system is capable of performing one or more embodiments.

[00478] As yet a further aspect, a process for integrating computing infrastructure comprising integrating computer readable code into a computer system may be provided. The computer system comprises a computer readable medium, in which the computer medium comprises one or more embodiments. The code in combination with the computer system is capable of performing one or more embodiments.

[00479] Although various embodiments are described above, these are only examples. For instance, computing environments of other architectures can be used to incorporate and/or use one or more aspects. Further, different instructions or operations may be used. Additionally, different types of registers and/or different register may be used. Further, other data formats, data layouts and/or data sizes may be supported. In one or more embodiments, one or more general-purpose processors, one or more special-purpose processors or a combination of general-purpose and special-purpose processors may be used. Many variations are possible.

[00480] Various aspects are described herein. Further, many variations are possible without departing from a spirit of aspects of the present invention. It should be noted that, unless otherwise inconsistent, each aspect or feature described herein, and variants thereof, may be combinable with any other aspect or feature.

[00481] Further, other types of computing environments can benefit and be used. As an example, a data processing system suitable for storing and/or executing program code is usable that includes at least two processors coupled directly or indirectly to memory elements through a system bus. The memory elements include, for instance, local memory employed during actual execution of the program code, bulk storage, and cache memory which provide temporary storage of at least some program code in order to reduce the number of times code must be retrieved from bulk storage during execution.

[00482] Input/Output or I/O devices (including, but not limited to, keyboards, displays, pointing devices, DASD, tape, CDs, DVDs, thumb drives and other memory media, etc.) can be coupled to the system either directly or through intervening I/O controllers. Network adapters may also be coupled to the system to enable the data processing system to become coupled to other data processing systems or remote printers or storage devices through

intervening private or public networks. Modems, cable modems, and Ethernet cards are just a few of the available types of network adapters.

[00483] The terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting. As used herein, the singular forms "a", "an" and "the" are intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms "comprises" and/or "comprising", when used in this specification, specify the presence of stated features, integers, steps, operations, elements, and/or components, but do not preclude the presence or addition of one or more other features, integers, steps, operations, elements, components and/or groups thereof.

[00484] The corresponding structures, materials, acts, and equivalents of all means or step plus function elements in the claims below, if any, are intended to include any structure, material, or act for performing the function in combination with other claimed elements as specifically claimed. The description of one or more embodiments has been presented for purposes of illustration and description, but is not intended to be exhaustive or limited to in the form disclosed. Many modifications and variations will be apparent to those of ordinary skill in the art. The embodiment was chosen and described in order to best explain various aspects and the practical application, and to enable others of ordinary skill in the art to understand various embodiments with various modifications as are suited to the particular use contemplated.

CLAIMS

1. A computer program product for facilitating processing within a computing environment, the computer program product comprising:

one or more computer readable storage media and program instructions collectively stored on the one or more computer readable storage media to perform a method comprising:

obtaining an instruction to convert data from a source data type to a target data type, the source data type being selected from one or more source data types supported by the instruction and the target data type being selected from one or more target data types supported by the instruction;

determining, based on a selected data type of one of the source data type and the target data type, a rounding mode for use by the instruction, the rounding mode implicitly set based on the selected data type, wherein the rounding mode is assigned to the selected data type;

performing a conversion of the data from the source data type to the target data type, the conversion including performing a rounding operation using the rounding mode implicitly set based on the selected data type, wherein the performing the conversion provides a result in the target data type; and

writing the result to a select location.

2. The computer program product according to the preceding claim, wherein the selected data type is the target data type specified using the instruction, the target data type comprising a machine-specific data type specified using the instruction.

3. The computer program product according to the preceding claim, wherein the machine-specific data type comprises a neural network processing-data-type-1 data type, and wherein the rounding mode assigned to the neural network processing-data-type-1 data type is a round to nearest ties away from zero.

4. The computer program product according to any of the two preceding claims, wherein the source data type is a standard data type specified using the instruction.

5. The computer program product according to claim 1, wherein the selected data type is the source data type specified using the instruction, the source data type comprising a machine-specific data type specified using the instruction.

6. The computer program product according to the preceding claim, wherein the machine-specific data type comprises a neural network processing-data-type-1 data type, and wherein the rounding mode assigned to the neural network processing-data-type-1 data type is a round to nearest ties away from zero.

7. The computer program product according to any of the two preceding claims, wherein the target data type is a standard data type specified using the instruction.

8. The computer program product according to any of the preceding claims, wherein the data to be converted includes source data in the source data type, the source data including a concatenation of data in one register specified using one field of the instruction and other data in another register specified using another field of the instruction.

9. The computer program product according to the preceding claim, wherein the performing the conversion includes rounding the source data based on the rounding mode to provide rounded data and converting the rounded data to provide the result in the target data type.

10. The computer program product according to any of the preceding claims, wherein the data to be converted includes source data in the source data type obtained using the instruction, and wherein the performing the conversion includes rounding the source data based on the rounding mode to provide rounded data in a precision of a target data type number and converting the rounded data to provide the result in the target data type.

11. A computer system for facilitating processing within a computing environment, the computer system comprising:

a memory; and

a processor in communication with the memory, wherein the computer system is configured to perform a method, said method comprising:

obtaining an instruction to convert data from a source data type to a target data type, the source data type being selected from one or more source

data types supported by the instruction and the target data type being selected from one or more target data types supported by the instruction;

determining, based on a selected data type of one of the source data type and the target data type, a rounding mode for use by the instruction, the rounding mode implicitly set based on the selected data type, wherein the rounding mode is assigned to the selected data type;

performing a conversion of the data from the source data type to the target data type, the conversion including performing a rounding operation using the rounding mode implicitly set based on the selected data type, wherein the performing the conversion provides a result in the target data type; and

writing the result to a select location.

12. The computer system according to the preceding claim, wherein the selected data type is the target data type specified using the instruction, the target data type comprising a machine-specific data type specified using the instruction.

13. The computer system according to the preceding claim, wherein the machine-specific data type comprises a neural network processing-data-type-1 data type, and wherein the rounding mode assigned to the neural network processing-data-type-1 data type is a round to nearest ties away from zero.

14. The computer system according to claim 11, wherein the selected data type is the source data type specified using the instruction, the source data type comprising a machine-specific data type specified using the instruction.

15. The computer system according to the preceding claim, wherein the machine-specific data type comprises a neural network processing-data-type-1 data type, and wherein the rounding mode assigned to the neural network processing-data-type-1 data type is a round to nearest ties away from zero.

16. A computer-implemented method of facilitating processing within a computing environment, the computer-implemented method comprising:

obtaining an instruction to convert data from a source data type to a target data type, the source data type being selected from one or more source data types supported

by the instruction and the target data type being selected from one or more target data types supported by the instruction;

determining, based on a selected data type of one of the source data type and the target data type, a rounding mode for use by the instruction, the rounding mode implicitly set based on the selected data type, wherein the rounding mode is assigned to the selected data type;

performing a conversion of the data from the source data type to the target data type, the conversion including performing a rounding operation using the rounding mode implicitly set based on the selected data type, wherein the performing the conversion provides a result in the target data type; and

writing the result to a select location.

17. The computer-implemented method according to the preceding claim, wherein the selected data type is the target data type specified using the instruction, the target data type comprising a machine-specific data type specified using the instruction.

18. The computer-implemented method according to the preceding claim, wherein the machine-specific data type comprises a neural network processing-data-type-1 data type, and wherein the rounding mode assigned to the neural network processing-data-type-1 data type is a round to nearest ties away from zero.

19. The computer-implemented method according to claim 16, wherein the selected data type is the source data type specified using the instruction, the source data type comprising a machine-specific data type specified using the instruction.

20. The computer-implemented method according to the preceding claim, wherein the machine-specific data type comprises a neural network processing-data-type-1 data type, and wherein the rounding mode assigned to the neural network processing-data-type-1 data type is a round to nearest ties away from zero.