Product Guide of MLU370-X4 Intelligent Accelerating Card

Release *0.9.4*

Preliminary

Cambricon

April 25, 2021

Directory

1. Foreword

1.1. Copyright

DISCLAIMER

Cambricon Technologies Corporation Limited (hereinafter referred to as eCambricon"ambricongies Corporatioon, warranty (express, implied, or statutory) or guarantee regarding the information contained herein, and expressly disclaims any and all implied warranties of merchantability, title, noninfringement of intellectual property or fitness for a particular purpose, and Cambricon DOES NOT assume any liability arising out of the application or use of any product or services. Cambricon shall have no liability related to any defaults, damages, costs or problems which may be based on or attributable to: (i) the use of the Cambricon product in any manner that is contrary to this guide, or (ii) customer product designs.

LIMITATION OF LIABILITY

In no event shall Cambricon be liable for any damages whatsoever (including, without limitation, damages for loss of profits, business interruption and loss of information) arising out of the use of or inability to use this guide, even if Cambricon has been advised of the possibility of such damages. Notwithstanding any damages that customer might incur for any reason whatsoever, Cambricon a aggregate and cumulative liability towards customer for the product described in this guide shall be limited in accordance with the Cambricon terms and conditions of sale for the product.

ACCURACY OF INFORMATION

Information provided in this document is proprietary to Cambricon, and Cambricon reserves the right to make any changes to the information in this document or to any products and services at any time without notice. The information contained in this guide and all other information contained in Cambricon documentation referenced in this guide is provided provided Cambricon does not warrant the accuracy or completeness of the information, text, graphics, links or other items contained within this guide. Cambricon may make changes to this guide, or to the products described therein, at any time without notice, but makes no commitment to update this guide.

Performance tests and ratings set forth in this guide are measured using specific chips or computer systems or components. The results shown in this guide reflect approximate performance of Cambricon products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. As set forth above, Cambricon makes no representation, warranty or guarantee that the product described in this guide will be suitable for any specified use. Cambricon does not represent or warrant that it tests all parameters of each product. It is customer customerable for any specified use. formance. A is suitable and fit for the application planned by the customer and to do the necessary testing for the application in order to avoid a default of the application or the product.

Weaknesses in customer warrant that it tests all parameters of each productlity of Cambricon product and may result in additional or different conditions and/or requirements beyond those contained in this guide.

IP NOTICES

1.2. Version Record

Table 1.1 Version Record

Name of the Document	Product Guide of MLU370-X4_ Intelligent Accelerator Card
Version Number	V0.9.4
Author	Cambricon
Date	2021.04.25

1.3. Update History

V0.9.4

Update time：

Updated Content:

-Preliminary Version

2. Outline

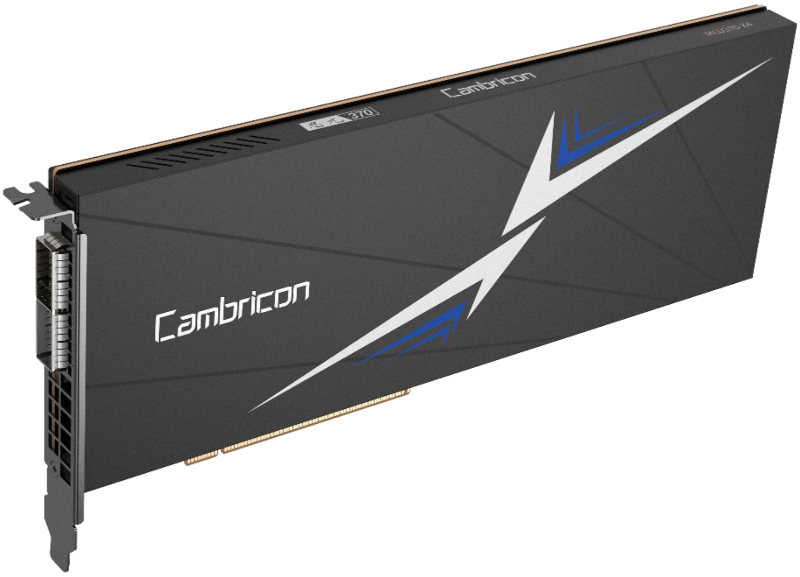

Fig. 1.1 MLU370-X4 Intelligent Accelerating Card

Fully Upgraded AI Accelerator Card with Data Center Integrating Training and Inference

MLU370-X4 intelligent accelerator card is based on the new generation of Cambricon SIYUAN 370 chip with PCIe 4.0 X16 interface. It is a full-height, full-length, single-width (FHFL-SS) standard PCIe size accelerator card, suitable for the latest CPU platforms in the industry. In addition, it can be easily mounted on the most advanced artificial intelligence server to quickly realize the deployment of AI computing power. The power consumption of the MLU370-X4 accelerator card is only 150W, which can provide powerful computing power support for highly diversified artificial intelligence applications such as computer vision, natural language processing, speech and traditional machine learning, and achieve AI computing with high energy efficiency.

Cambricon SIYUAN 370 Chip

The Cambricon SIYUAN 370 chip is manufactured using advanced TSMC's 7nm technology, and its performance indicators are comprehensively improved compared to the previous generation. The SIYUAN 370 chip contains up to 24 MLU-Cores, and adopts the MLUv03 architecture to ensure multi-core parallel efficiency; 24G memory can provide 3 times the memory bandwidth of the previous generation, effectively solving the bandwidth bottleneck in the AI computing process; the new platform vMLU can support 8 instances on one chip, helping customers achieve cloud virtualization and container-level resource isolation; the SIYUAN 370 chip provides comprehensive AI precision support for INT16, INT8, INT4, FP32, FP16, BF16, *etc.,* to meet the computing power requirements of diverse neural networks, and has both versatility and performance.

Significantly Improved AI Computing Power

Cambricon MLU370-X4 not only greatly improves fixed-point computing power, but also fully upgrades floating-point computing power, and the built-in hardware video and image codec capabilities are further enhanced. When INT8 precision is adopted for AI inference computations, the performance of non-sparse network is 2 times higher than that of the previous generation of the accelerator card. Besides, the computing power of floating-point precision such as FP32, FP16 and BF16 is also significantly enhanced, where FP16 precision can provide up to 96 TFLOPS peak computing power, which enables MLU370-X4 to be more widely used in AI scenarios that require floating-point operations. Its built-in brand-new hardware video and picture codec can provide 1.4 times the video performance of the previous generation of the accelerator card, and can process up to 16 channels of 8k 30fps high-definition video at the same time. When the system processes this type of application, it effectively reduces the CPU pre-processing load and PCIe bandwidth occupation, helping the application performance to be improved.

Cambricon Neuware End-Cloud Integrated Software Stack

The Cambricon Neuware Software Stack adopts an end-cloud integrated architecture, which supports the full range of Cambricon’s products to share the same software interface and complete ecology, and can facilitate the development, migration and optimization of AI applications. The Cambricon inference & acceleration engine (MagicMind) dedicated to MLU370 provides end-to-end model representation, model optimization and deployment capabilities, supports multiple frameworks, algorithm models in multiple business scenarios, and supports multiple AI computing hardware platforms (MLU&CPU).

New Platform vMLU Brings More Virtualized Instance Support

Cambricon virtualization technology vMLU supports the realization of 8 isolated AI computing instances on MLU370-X4. Each instance has exclusive computing, memory, and codec resources, and can still maintain a high efficiency of no less than 90% in a virtualized environment, realize cloud virtualization and container-level resource isolation, and help customers make full use of hardware resources.

MLU-Link^{TM} and ROCE v2, Set up Training Clusters Flexibly

The Cambricon MLU-Link group multi-core interconnection technology supports the interconnection between SIYUAN chips and cross-system interconnection, and can realize the vertical expansion of the computing center and meet the needs of super-large AI model training. MLU370-X4 supports a maximum of 2*200Gbps MLU-Links data communication bandwidth between chips, and can build a training cluster without relying on switches; it can also support a separate ROCEv2 network with 2*100Gbps bandwidth, as well as a hybrid networking of MLU-LinkTM and ROCE v2, so that the large-scale expansion of the training cluster can be realized.

3. Specifications of MLU370-X4

3.1. Performance Specifications

Table 3.1 MLU370-X4 Intelligent Accelerator Card Hardware Specifications

Type of Board Card	MLU370-X4
Core Architecture	Cambricon MLUv03
Core Frequency	1 GHz
Computation Accuracy Supporting	INT16, INT8, INT4, FP32, FP16/BF16
Video Decoding	support
Memory Capacity	24GB
Memory Bit Wide	384-bit
Memory Bandwidth	300GB/s
System Interface	PCI Express 4.0 x16 support lane reversal
PCI Identifier	PCIe Vendor ID 0xCABC PCIe Device ID 0x0370 PCIe Sub-Vendor ID 0xCABC PCIe Sub-System ID 0x0057
Shape	FHFL Single Slot
TDP Power Consumption	150W
ECC Protection	yes
Heat Dissipation Scheme	passive

3.2. Software Specifications

Table 3.2 describes the software specifications of MLU370-X4 intelligent accelerator card.

Table 3.2 MLU370-X4 Intelligent Accelerator Card Software Specifications

Interface	Description

Interface	Description
PCIE Base address	PF （one, 64bit）： BAR0: 256 MB prefetchable BAR2: 256 MB prefetchable BAR4: 256 MB prefetchable VF（four, 64bit）： BAR0: 256 MB prefetchable BAR2: 256 MB prefetchable BAR4: 256 MB prefetchable
SMBus（8bit address）	0x8E(Write） 0x8F（Read）

The bit width of the SMBUS register is 32 bits. Table 3.3 describes the reading process of the register (S: Slave, M: Master).

Table 3.3 Reading and Writing Process of SMBus Register

Direction	Bits	Content
M->S	1	START
M->S	8	SLAVE ADDRESS(Write)
S->M	1	ACK
M->S	8	REGISTER ADDRESS
S->M	1	ACK
M->S	1	RE START
M->S	8	SLAVE ADDRESS(Read)
S->M	1	ACK
S->M	8	DATA[7:0]
S->M	1	ACK
S->M	8	DATA[15:8]
S->M	1	ACK
S->M	8	DATA[23:16]
S->M	1	ACK
S->M	8	DATA[31:24]

Direction	Bits	Content
M->S	1	NACK
M->S	1	STOP

The definition, address, and description of the SMBUS register are shown in the Table 3.4.

Table 3.4 Description of SMBus Register

Definition of the Register	Address	Access	Description
Power Consumption of the Board Card	0x01	RO	[31:0] Power Consumption of the Board Card; Data Type: float; Unit: W
Temperature of the Board Card	0x02	RO	[31:0] Temperature of the Board Card; Data Type: float; Unit: °C
Temperature of the Chip	0x03	RO	[31:0] Temperature of the Chip; Data Type: float; Unit: °C
Temperature of DDR Particles	0x04	RO	[31:0] Temperature of DDR Particle; Data Type: float; Unit: °C
power brake	0x05	RW	Writing 0x04, the main frequency is reduced to 25% of the current; Writing 0x01, restore the level before frequency reduction
Setting State of the Power Consumption of the Board Card	0x19	RO	[31:16] power capping setting power consumption value [15:0] TDP Power Consumption Data Type: uint16_t Unit: W
State Information	0x20	RO	Bit0：whether the power brake may enable Bit1：over-temperature and frequency reduction state Bit[5:2]： reserved Bit6：whether the power capping may enable Bit7：whether the frequency capping IPU may enable Bit[17:7]： reserved Bit18：power capping do not preserve while power off in-band .0：disable；1：enable

Definition of the Register	Address	Access	Description
			Bit19： power capping preserve while power off in-band .0：disable； 1： enable Bit20： power capping do not preserve while power off out of band. 0： disable； 1： enable Bit21： power capping preserve while poweroff out of band 0：disable； 1： enable Bit[31:22]： reserved
Temperature Threshold Information	0x23	RO	[31:16] reserved [15:8] over-temperature power-off temperature [7:0] over-temperature frequency reduction Data Type:uint8_t Unit: ℃
Power capping	0x29	RW	[31:16] reserved [15] feature flag of power capping, 0： temporary effect; 1:power down and save [14:0] Power Capping Value of the Board Card Data Type: uint15_t Unit: W (If the value is 0, the power capping is released.)
PCIE Vendor ID and Device ID	0xA0	RO	[31:16] Device ID:0x0370 [15:0] Vendor ID :0xCABC
PCIE Sub-Vendor ID and Sub-System ID	0xA1	RO	[31:16] Sub-System ID : 0x0057 [15:0] Sub-Vendor ID:0xCABC
PCIE_negotiated_speed	0xA2	RO	[7:0] display PCIE negotiated speed, for example, 0x04 means gen4 16GT/s, 0x03 means gen3 8GT/s, 0x02 means gen2 5GT/s, 0x01 means gen1 2.5GT/s
PCIE_negotiated_link_width	0xA3	RO	[7:0] display PCIE negotiated width, for example, 0x16 means X16； 0x08 means X8, 0x04 means X4, 0x02 means X2, 0x01 means X1
Type of the Board Card	0xF0	RO	[7:0] display the type of the board card, for example, 0x57 means X4 model.

Definition of the Register	Address	Access	Description
Equipment Manufacturer	0xF1	RO	[3:0] display the serial number of the equipment manufacturer
Hardware Version Number	0xF2	RO	[7:0] display the hardware version number, for example, 0x11means the hardware version V1.1.
Firmware Version Number	0xF3	RO	[11:0] display the firmware version number, for example, 0x113 means that the main version number is 1, the sub-version number is 1, and the patch number is 3.
Manufacturing Time	0xF4	RO	[15:0] display the manufacturing time, for example, 0x2101means that the manufacturing time is January, 2021.
Serial Number	0xF5	RO	[19:0] display the serial number of the equipment, for example, 0x00030 means that the serial number is 00030.
Lower SN Number	0xF6	RO	[31:0] low 8-bit data of SN number, for example, the low 8-bit data of SN: 572101300030 is saved as 0x01300030.
Higher SN Number	0xF7	RO	[31:16] reserved [15:0] high 4-bit data of SN number, for example, the high 4-bit data of SN： 572101300030 is saved as 0x5721.
Part_number_1	0xF8	RO	[31:0] “MLU3” high 8-bit data of Part_number (the ASCII code corresponding to the character)
Part_number_2	0xF9	RO	[31:0] “70-X” middle 8-bit data of Part_number (the ASCII code corresponding to the character)
Part_number_3	0xFA	RO	[7:0] “4” low 8-bit data of Part_number (the ASCII code corresponding to the character)

Table 3.5 shows how to obtain SN information.

Table 3.5 SN Number Decomposition

SN Number	[47:40]	[39:24]	[23:20]	[19:0]
0x572101300030	Type of the Board Card *e.g.,* 0x57	Manufacturing Time *e.g.,* 0x2101	Equipment Manufacturer *e.g.,* 0x3	Serial Number *e.g.,* 0x00030

3.3. Specifications of the Use Environment

Table 3.6 describes the specifications of the use environment of MLU370-X4 intelligent accelerator card.

Table 3.6 Specifications of the Use Environment of MLU370-X4 Intelligent Accelerating Card

Item	Value
Operating Temperature	0°C～45°C
Storage Temperature	-40°C～75°C
Operating Humidity	5%—95% Relative Humidity
Storage Humidity	5%—95% Relative Humidity

3.4. Specifications of Structure and Dimension

The size structure and size of MLU370-X4 intelligent accelerator card are shown in Fig. 3.1:

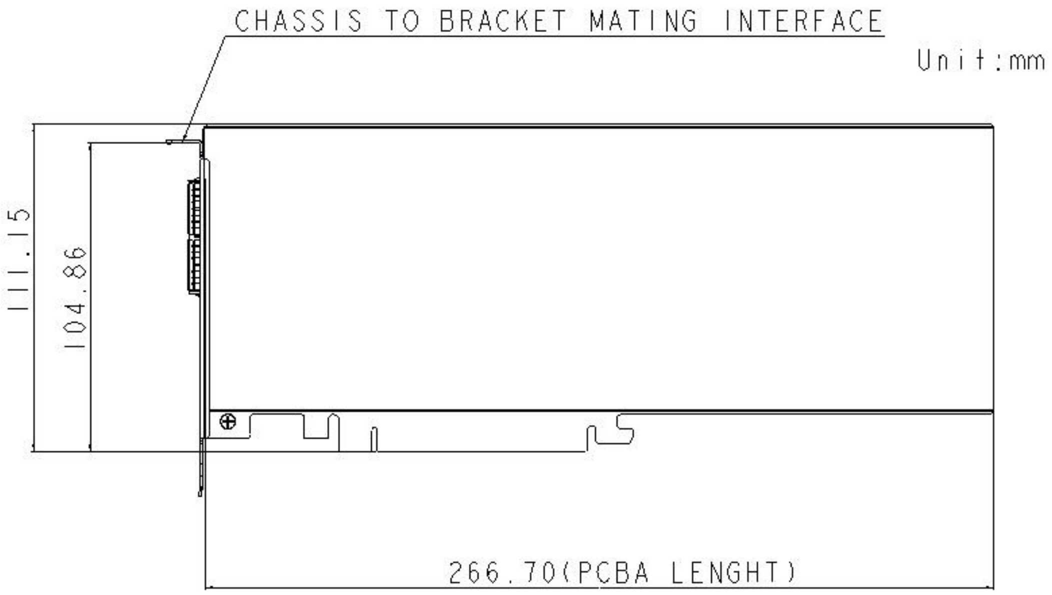

Fig. 3.1 Size of MLU370-X4 Intelligent Accelerating Card

Toolless design is applied to the top cover of MLU370-X4. After the bracket is disassembled, the top cover can be taken off directly for convenient disassembly and assembly.

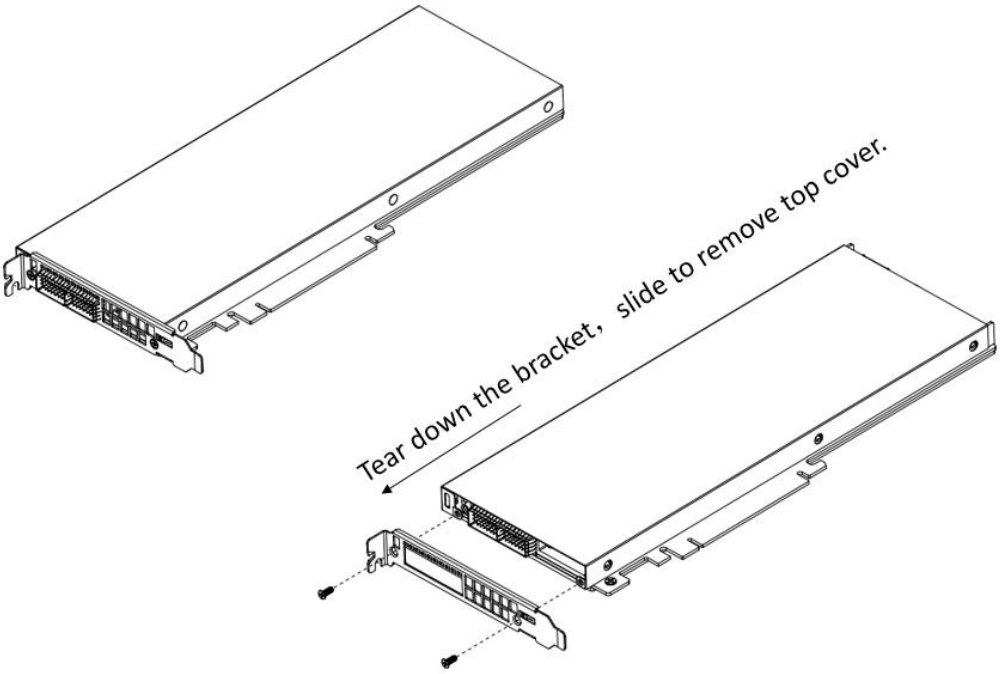

Fig. 3.2 Toolless design apply to MLU370-X4 top cover

3.5. Size and Weight of the Package

The size and weight information of the package of MLU370-X4 intelligent accelerator card is shown in Table 3.7:

Table 3.7 Size and Weight of the Package of MLU370-X4

Type	Weight	Size	Remark
Single Card	727g	266.7 mm*111.15mm*18.3mm	NA
Whole Case of Industry	14.1kg	600mm*400mm*253mm	16 Cards Per Box

Remarks: the weight is an actual measured value, tolerance +-10%

3.6. Heat Dissipation Specifications

3.6.1. MLU370-X4's Board Card Power Consumption and Temperature Definitions

Table 3.8 Specification of the Use Environment of MLU370-X4 Intelligent Accelerating Card

Items	Parameters
Thermal Design Power (TDP) of Whole Board Card	150W
Recommended Operating Tj(Junction temperature) of MLU	0-90℃
Frequency Drop Tj of MLU	92℃
Frequency Drop Range of MLU	50%
Shutdown Tj of MLU	95℃

3.6.2. Resistance Curve of the Radiator of MLU370-X4

The resistance curve measured by the radiator of MLU370-X4 is shown in Fig. 3.3:

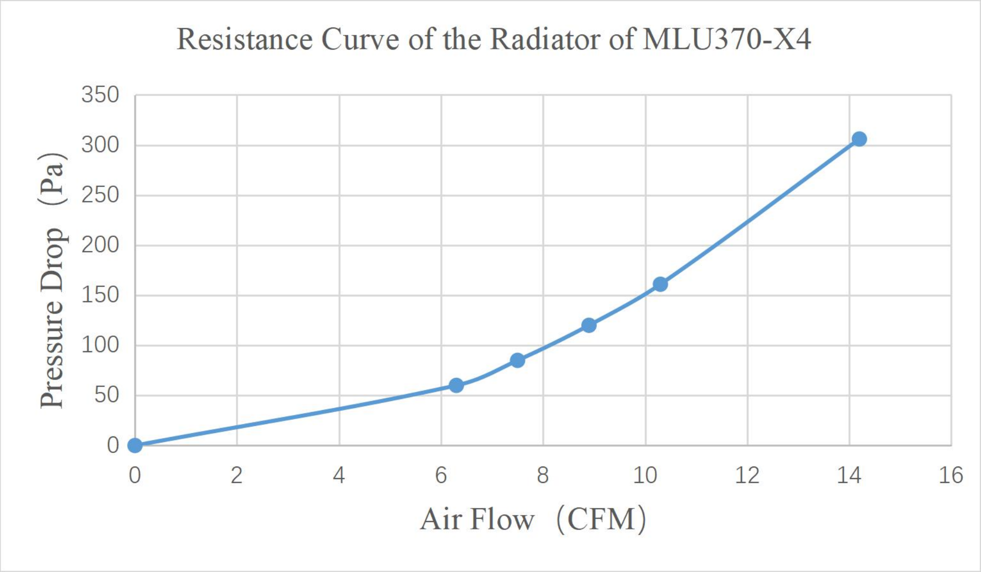

Fig. 3.3 Resistance Curve of the Radiator of MLU370-X4

The comparison table of the air flow of heat dissipation and pressure drop of the board card is shown in Table 3.9:

Table 3.9 MLU370-X4 Board Card Air Flow of the Radiator- Pressure Drop of

Air Flow（CFM）	Wind Pressure (Pa)
6.3	60
7.5	85

8.9	120
10.3	161
14.2	306

3.6.3. MLU370-X4 Supported Card Direction

The air inlet direction of MLU370-X4 is shown in Fig 3.4:

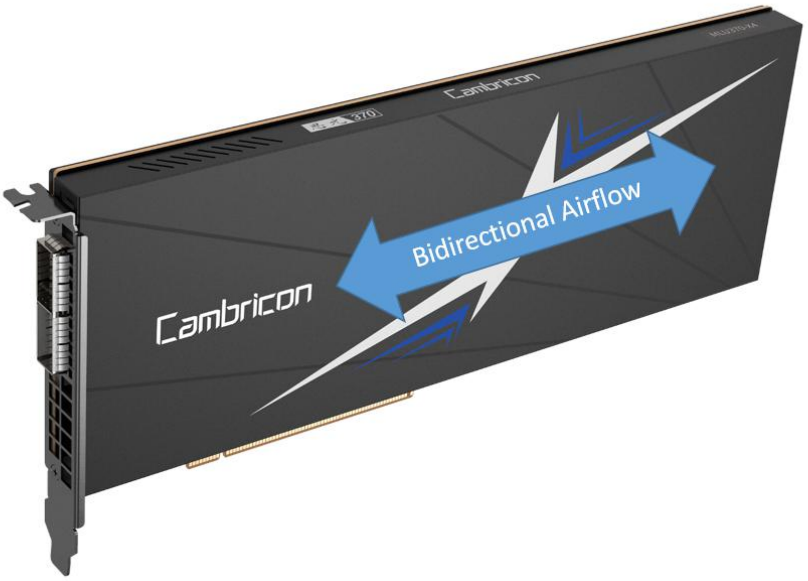

Fig. 3.4 Airflow Direction for PCIE Card

3.6.4. MLU370-X4 Supported Ambient Temperature for Working and Minimum Airflow Volume Requirements of the Radiator at Different Temperature

MLU370-X4 can work (TDP mode) at the ambient temperature of 0-45℃（air intake temperature of the radiator of board card). The minimum airflow requirements under main temperature conditions are shown in the Table 3.10:

Table 3.10 MLU370-X4 Minimum Air Flow Requirement of the Radiator vs Ambient Thermometer

Temperature of the Inlet（°C）	Minimum Air Flow Requirement of the Radiator（CFM)
25	6.3

30	7.5
35	8.9
40	10.3
45	14.2

3.6.5. A Curve for the Average Temperature of the Inlet and the Minimum Air Flow Requirement through the Radiator of MLU370-X4

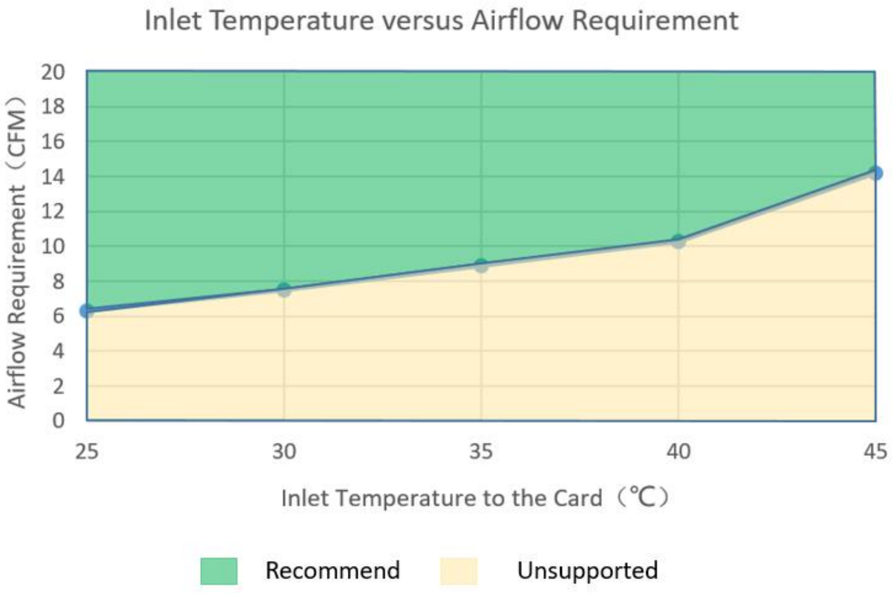

Fig. 3.5 Inlet Temperature versus Airflow Requirement

3.7. Power specifications and Electrical Specifications

The input voltage of the power interface and current specifications are shown in Table 3.11 and Table 3.12.

Table 3.11 Power Interface and Input Voltage

Power Interface	Minimum Voltage	Normal Voltage	Maximum Voltage
PCIe Gold Finger (12V）	11.04V	12V	12.96V
CPU 8-pin connector（12V）	11.04V	12V	12.96V

PCIe Gold Finger (3V3）	3.0V	3.3V	3.63V

Table 3.12 Current Specification

Power Interface	Peak Current	Moving Average
PCIe Gold Finger (12V）	20A	200us
	17A	1ms
	13A	5ms
CPU 8-pin（12V）	33A	200us
	30A	1ms
	25A	5ms

The specification of Power Capping is shown in Table 3.13：

Table 3.13 Power Capping

Item	Value
Power Capping Threshold	150W
Power Capping Response time （typical）	50ms
Power Capping Response time （max）	100ms

The specification of Power Brake is shown in Table 3.14：

Table 3.14 Power Brake

Item	Value
PB# PCIe pin assignment	B30
Power Brake response time（typical）	150us
PB# input insertion low time （min）	250ms
Power brake hardware slowdown factor	4x

4. Development Environment of Cambricon NeuWare

NeuWare can fully support all kinds of mainstream programming frameworks, such as TensorFlow, Caffe, PyTorch, and MXNet. With the above mentioned programming frameworks, users can easily and conveniently develop and deploy their deep learning applications on Cambricon MLU370-X4. At the same time, NeuWare provides complete runtime system and driver software to speed up the system integration procedure.

NeuWare further provides a series of tools including application development, function debugging and performance optimization. The application development tools include machine learning library, runtime library, compiler, model retraining tools and domain-specific (e.g., video analysis) SDK; the function debugging tools can fulfill all the requirements from different levels of programming framework and function library; the performance optimization tools include tools for performances analysis and system monitoring.

The Cambricon inference acceleration engine (MagicMind) provides end-to-end model representation, model optimization and deployment capabilities, supports multiple frameworks, algorithm models in multiple business scenarios, and supports multiple AI computing hardware platforms (MLU&CPU).

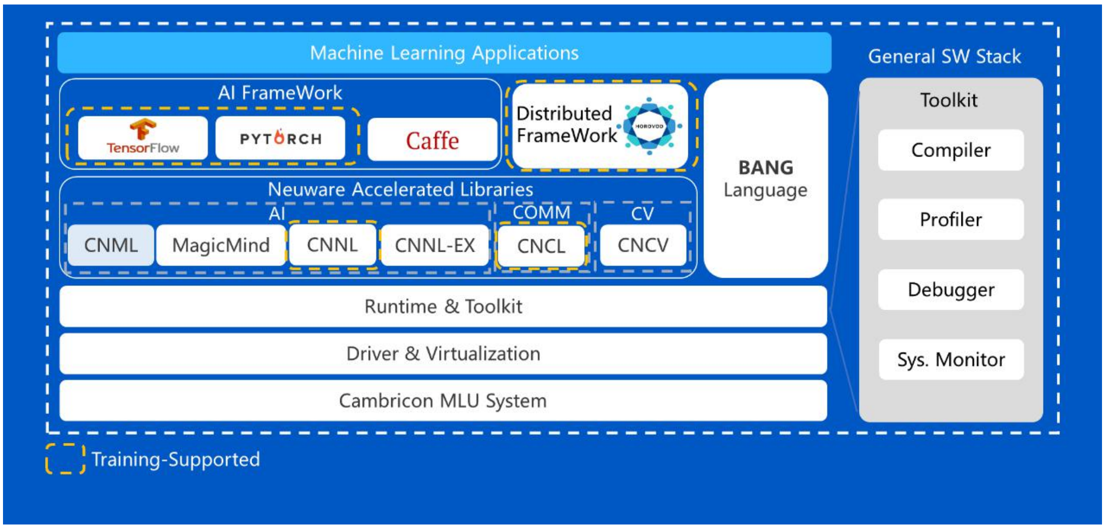

Fig. 4.1 Cambricon NeuWare

For more information, please visit www.cambricon.com

Tel: 86-10-83030003

Email: business@cambricon.com

Address: 11th Floor, Block D, Truth Plaza, No. 7 Zhichun Road, Haidian District, Beijing, China

Cambricon
寒 武 纪

5. Compliance

The MLU370-X Series is compliant with the regulations listed in this section. Compliance marks, including the FCC ID numbers, can be found on the labels of each devices.

United States

Federal Communications Commission (FCC)

This device complies with Part 15 of the FCC Rules.

Operation is subject to the following two conditions: (1) This device may not cause harmful interference, and (2) this device must accept any interference received, including interference that may cause undesired operation.

This equipment has been tested and found to comply with the limits for a Class B digital device, pursuant to Part 15 of the FCC Rules. These limits are designed to provide reasonable protection against harmful interference in a residential installation. This equipment generates, uses and can radiate radio frequency energy and, if not installed in accordance with the instructions, may cause harmful interference to radio communications. However, there is no guarantee that interference will not occur in a particular installation.

If this equipment does cause interference to radio or television reception, which can be determined by turning the equipment off and on, the user is encouraged to try to correct the interference by one or more of the following measures:

- Reorient or relocate the receiving antenna
- Increase the separation between the equipment and receiver
- Connect the equipment into an outlet on a circuit different from that to which the receiver is connected

- Consult the dealer or an experienced radio/TV technician for help

Caution: Any changes or modifications not expressly approved by the party responsible for compliance could void the user's authority to operate this equipment.

Underwriters Laboratories (UL)

UL Listed Product Logo for MLU370-X Series Intelligent Processing Cards，model name MLU370-X.

C
UL
US
LISTED