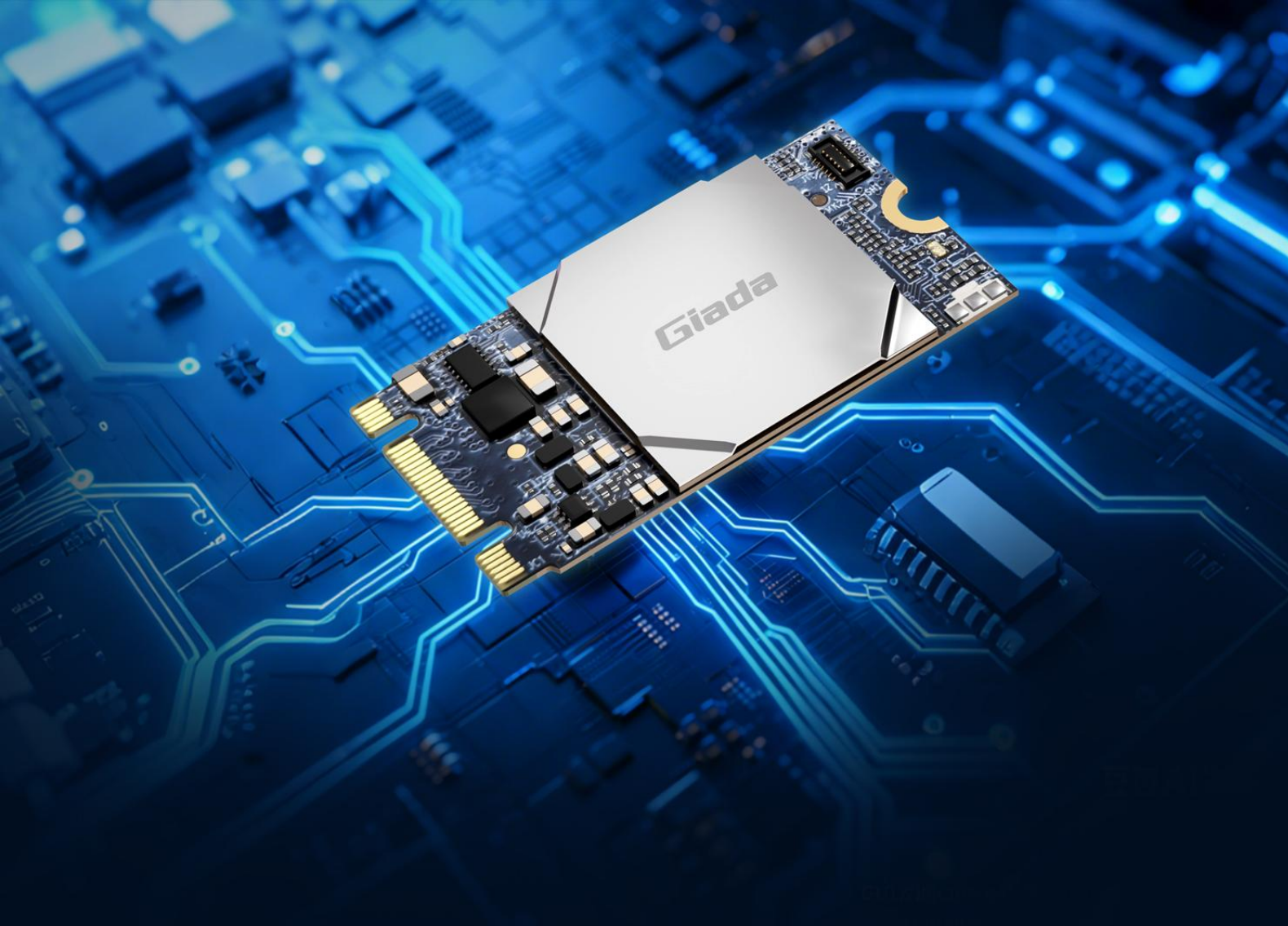


LM2-100

User Manual

Giada

Statement

Safety Notice

• Read the user manual carefully before setting up the Giada product.

• Disconnect the power cord before installing the internal components

• Most electronic components are sensitive to static electrical charge, please wear a wrist-grounding strap when installing the internal components.

• Don't disconnect the power cord when the system is running to avoid damage to the sensitive components by instantaneous surge voltage.

Contact Information

Shenzhen JIEHE Technology Development Co., Ltd.

Website: www.giadatech.com

Phone: +86-755-3330 0326

Email: support@giadatech.com

Address: 37F, Holdfound Sky Plaza Office Building, 11008 Beihuan Blvd.,

Nanshan, SZ, China, 518051

Table of Contents

1. Product Introduction

The LM2-100 computing module is an AI inference acceleration card developed based on the DeepX DX-M1M chip. It features a PCIe 3.0x2 M.2 M+B-Key interface and is equipped with 2GB of LPDDR4x memory. Designed specifically for edge computing scenarios, it offers exceptional computing performance of up to 25 TOPS, along with high accuracy, low power consumption, and a wide operating temperature range.

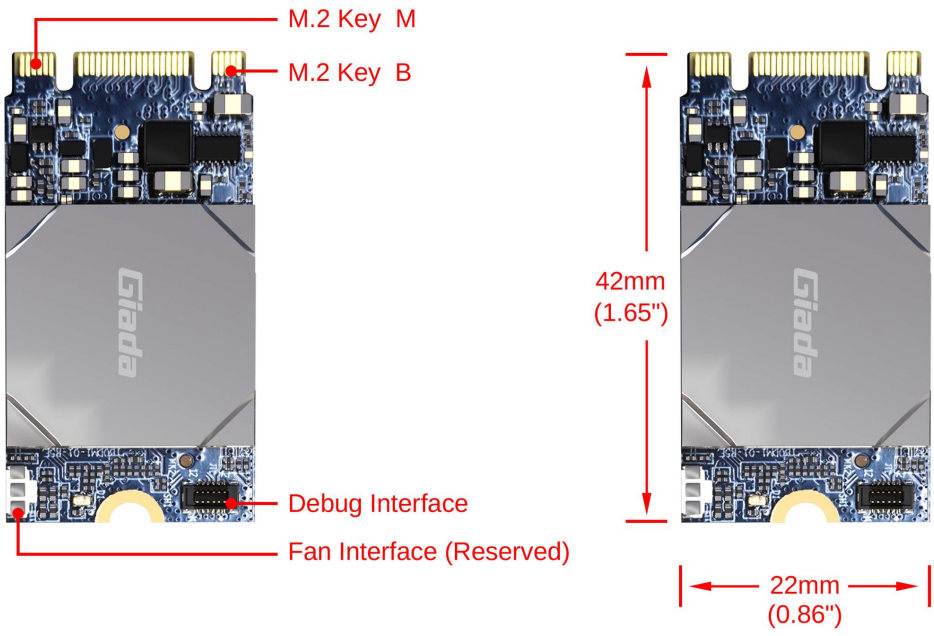
Core Performance:Utilizing IQ8™ (Intelligent Quantization 8-bit Integer) technology, the DX-M1M chip delivers 25 TOPS (INT8) at a typical power consumption of 5W, achieving an impressive energy efficiency ratio of 5 TOPS/W. This enables computational results approaching the precision of FP32.

Interface & Design:The module adopts a standard M.2 2242 B+M-Key form factor and connects to the host via a PCIe Gen 3 x2 interface, facilitating seamless integration into embedded devices such as industrial PCs and single-board computers.

Memory & Compatibility: Featuring 2GB of built-in LPDDR4x memory independent of the main system, the module efficiently runs multiple AI models (such as the YOLO series). It supports mainstream AI frameworks, including PyTorch, ONNX, and TensorFlow, and is compatible with Linux (e.g., Ubuntu) and Windows operating systems.

Application Scenarios: With its high energy efficiency and accuracy, the LM2-100 is well-suited for a variety of edge AI applications, including industrial quality inspection, intelligent surveillance, smart retail, drone vision, and robotics.

2. Product Interface Diagram/Dimensional Drawing



3. Specifications

NPU	DeepX M1M
Memory	2GB LPDDR4x
AI Performance	25 TOPS
I/O Interface	M.2 Key B+M
Host Interface	PCIe 3.0X2
Power Requirement	3.3V upto 3A
PCB	22 x 42 x 4.1 mm (0.86” x 1.65” x 0.16”)
Operating System	Windows, Linux
Operating Temperature	-25°C ~ 65°C
Storage Temperature	-25°C ~ 70°C
AI Frameworks	Supports TensorFlow, TensorFlow Lite, ONNX, Keras, and PyTorch, converted via the Dataflow compiler.

4. Installation Instructions

A.Installation with Extension Bracket (M.2 SSD 2280 Slot)

01.Snap the connecting studs onto the bracket.

02.Snap the 2280 extension plate onto the connecting studs until fully secured.

03.Install the M3 * 3.5 screws.

04.Insert the computing module into the M.2 SSD interface.

01

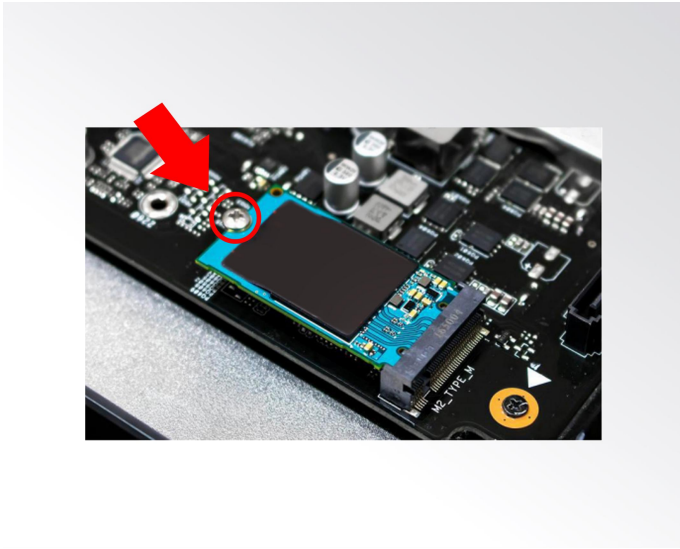
02

03

04

B.Direct Installation (M.2 SSD 2242 Slot)

Secure via the 2242 screw mounting holes.



5. System Installation

Preparation:

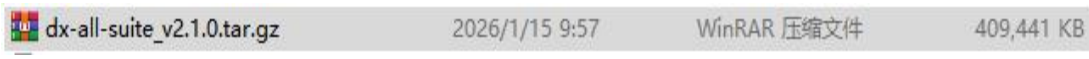
- LM2-100 Module provided by Giada
- PC with screen and installed Ubuntu22.04 and LM2-100 AI computing card .
- SDK dx-all-suite_v.2.2.0.tar and dx-com.Here you can get the latest SDK dx-all-suite_v.2.2.0.tar and dx-com files
  https://mega.nz/file/j4IliboK#yUpA7q5XqdyX0cyDl17TY6NEQDJgjml6GC20wVh5BFo

6. Set up the software environment for an AI computing module.

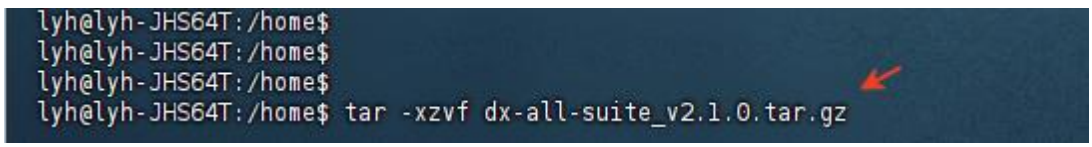
6.1 Open a terminal,enter commands below,and install the necessary software.

```
sudo apt upgrade -y

sudo apt update -y

sudo apt install -y \
gcc-12 g++-12 gcc-aarch64-linux-gnu g++-aarch64-linux-gnu \
git meson ninja-build pkg-config build-essential \
libncurses5-dev libncursesw5-dev libtinfo-dev \
python3-dev python3-pip \
python3-pyqt5 python3-pyqt5.qtquick python3-pyqt5.qtsvg
```

6.2 Copy the dx-all-suite_v2.1.0.tar.gz to home



6.3 Use the command of tar -xzvf to extraction the file



6.4 After the extraction is complete, install the overall environment.

```
cd dx-all-suite/
sudo ./dx-runtime/install.sh --all
```

```
==== DKMS Driver Installed Check ====
[INFO] dxrt-driver-dkms/1.8.0-2, 6.8.0-90-generic, x86_64: installed
[INFO] /lib/modules/6.8.0-90-generic/updates/dkms/dxrt_driver.ko ... OK
[INFO] /lib/modules/6.8.0-90-generic/updates/dkms/dx_dma.ko ... OK
==== Runtime Version Dependency Check ====
[OK] Version Dependency Check
   Runtime Framework Version: 3.1.0
   Device Driver Version: v1.8.0
   PCIe Driver Version: v1.6.0
   Firmware Version: device-id=0 v2.4.0
   ONNX Runtime Version: v1.20.1
==== Runtime Executable File Check ====
[OK] Executable File Check
   Runtime Framework Version: 3.1.0
   Device Driver Version: v1.8.0
   PCIe Driver Version: v1.6.0
   Firmware Version: device-id=0 v2.4.0
   ONNX Runtime Version: v1.20.1
   dxrt-cli ...OK
   run_model ...OK
   parse_model ...OK
   dxtop ...OK
   dxrtd ...OK
   Header: /usr/local/include/dxrt ...OK
==== Runtime dxrtd Service Check ====
[OK] The dxrtd service is running correctly. This is the expected behavior for builds with USE_SERVICE=ON on cmake/dxrt.cfg.cmake

========================================================================
** Sanity check PASSED!
========================================================================
/home/dx-all-suite
---
[INFO] Sanity Check Result:
 ✓PASS: All sanity checks passed successfully.
[HINT] To activate the virtual environment, run:
[HINT]    source /home/dx-all-suite/dx-runtime/venv-dx-runtime/bin/activate
[INFO] [OK] Installing all runtime modules
```

6.5 Installtions the Driver

```
cd ./dx-runtime/dx_rt_npu_linux_driver
Sudo ./install.sh
```

```
dxrt_driver.ko:
Running module version sanity check.
 - Original module
   - No original module exists within this kernel
 - Installation
   - Installing to /lib/modules/6.8.0-94-generic/updates/dkms/

Warning: The post_install script is not executable.

depmod...
Creating modprobe configuration for module load order...
Installing udev rules using /usr/src/dxrt-driver-dkms-2.1.0-2/modules/install-udev_rule.sh...
dxrt permissions will be changed(0666)
Updating initramfs...
update-initramfs: deferring update (trigger activated)
Loading modules...
Restarting dxrt systemd service...

正在处理用于 initramfs-tools (0.140ubuntu13.5) 的触发器 ...##################################################.....................] 
update-initramfs: Generating /boot/initrd.img-6.8.0-94-generic
[INFO] The 'dx_rt_npu_linux_driver' has been installed.
To complete the installation, the system must be restarted.
  Would you like to reboot now? (y/n): y
Start reboot...
```

7. Verification

7.1 Confirm version information.

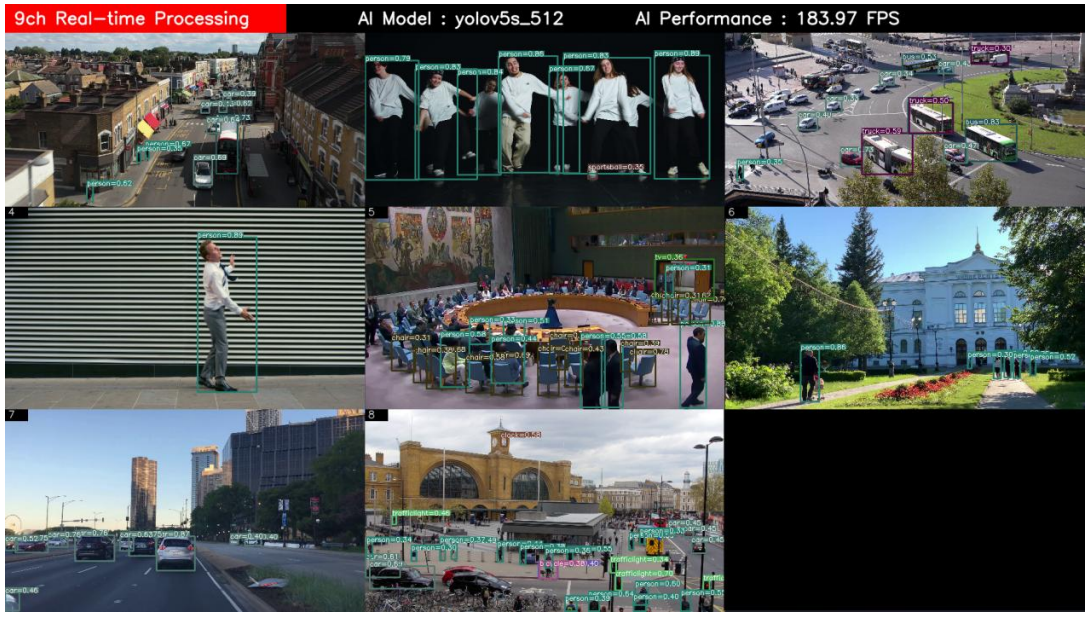
```
cd /home
cd dx-all-suite/dx-runtime/dx_rt
sudo ./bin/dxrt-cli --status
```

```
lyh@lyh-JHS64T:/home/dx-all-suite/dx-runtime/dx_rt$ sudo ./bin/dxrt-cli --status
DXRT v3.1.0
=======================================================
 * Device 0: M1, Accelerator type
---------------------   Version   ----------------------
 * RT Driver version   : v1.8.0
 * PCIe Driver version : v1.6.0
-------------------------------------------------------
 * FW version          : v2.4.0
---------------------- Device Info ----------------------
 * Memory : LPDDR4 4200 Mbps, 1.92GiB
 * Board  : M.2, Rev 0.0
 * Chip Offset : 0
 * PCIe   : Gen3 X2 [02:00:00]

NPU 0: voltage 750 mV, clock 1000 MHz, temperature 38'C
NPU 1: voltage 750 mV, clock 1000 MHz, temperature 38'C
NPU 2: voltage 750 mV, clock 1000 MHz, temperature 38'C
=======================================================
```

If the above information is output as expected,it indicates that the software environment for the LM2-100 is properly configured,and the SDK can be invoked successfully.

7.2 Run the Demon

```
cd /home
cd dx-all-suite/dx-runtime/dx_app
sudo ./bin/yolo_multi -c example/yolo_multi/yolo_multi_demo.json
```



8. Model Conversion (ONNX to DXNN)

DX-COM is the model compilation and conversion tool for the LM2-100, responsible for converting machine learning models from the open standard format ONNX into the DXNN format. Only neural network models in the DXNN format can leverage the powerful computing capabilities of the LM2-100.It is recommended to use Ubuntu OS with at least 16GB of memory for the configuration environment.

3.1 Environment Setup

First, obtain the complete toolchain package for the model compiler. You can get it from the official website. Once obtained, extract it to a local directory by following the steps below:

8.1 Environment Setup:

First, obtain the complete toolchain package for the model compiler. You can get it from the official website. Once obtained, extract it to a local directory by following the steps below:

```
tar -zxvf dx_com_M1_v2.1.0.tar.gz
cd dx_com_M1_v2.1.0/dx_com
```

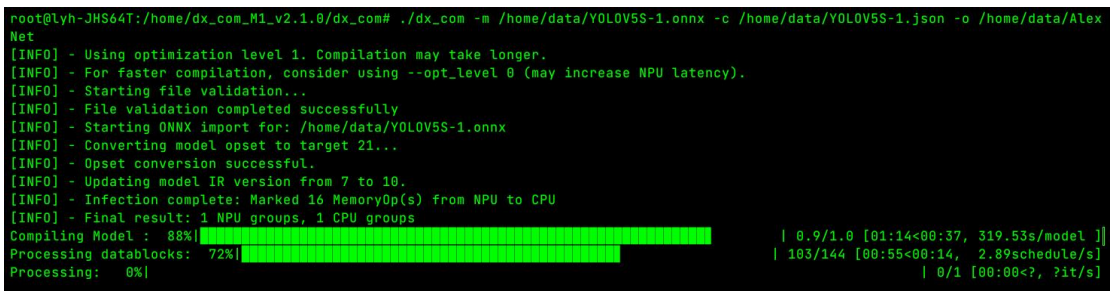
8.2 Compile a neural network model

Command : dx_com -m <MODEL_PATH> -c <CONFIG_PATH> -o <OUTPUT_DIR>

Notes:

-m MODEL_PATH, --model_path
-c CONFIG_PATH, --config_path CONFIG_PATH
-o OUTPUT_DIR, --output_dir OUTPUT_DIR

Case:



8.3 Configuration File Description (config.json)

A typical configuration file only requires the following description:

```
{
  "inputs": {
    "input.1": [      #ONNX input's name
      1,              #Batch Size
      3,              #ONNX input's shape
      224,
      224
    ]
  }
}
```

Through conversion, a neural network model in the DXNN format is obtained. Then, using the provided SDK, the computing power of the LM2-100 can be utilized to run this model. Please refer to the previous

Giada

Shenzhen JIEHE Technology Development Co., Ltd.

Website: www.giadatech.com

Phone: +86-755-3330 0326

Email: support@giadatech.com

Address: 37F, Holdfound Sky Plaza Office Building, 11008 Beihuan Blvd., Nanshan, SZ, China, 518051