EDGECORTIX

SAKURA-II M.2 Modules

User Manual

Revision History

TABLE OF CONTENTS

1. OVERVIEW

This User Manual covers the EdgeCortix SAKURA-II M.2 Modules. The SAKURA-II M.2 Modules feature an EdgeCortix SAKURA-II chip, an edge AI accelerator that is run-time reconfigurable using the EdgeCortix MERA compiler and software framework, and that boasts up to 60 TOPS using EdgeCortix's Dynamic Neural Accelerator (DNA). Using the EdgeCortix MERA compiler and software framework it can run the latest Vision and Generative AI models with market-leading energy efficiency and low latency. EdgeCortix's MERA compiler and software framework provides a robust platform for deploying the latest AI inference models quickly and easily, in an application agnostic manner.

Throughout this manual these will be referred to as "M.2 Module" or just "Module". It comes pre-installed with a fan sink (mandatory for cooling) with a connector for powering the fan. These evaluation platforms are ready to drop into an Ubuntu host PC M.2 Key M slot for software development and AI model evaluation tasks.

This manual is intended for users of stand-alone M.2 Modules that will be inserted into their own systems.

Note: Please ensure that your system can accommodate the module with the fansink integrated (see section 2 for the dimensions of the module with the fansink).

SYSTEM REQUIREMENTS

If the target and development system are the same where the M.2 Module is to be installed, below are the system requirements:

- Intel or AMD x86 based Linux PC with min 32GB of RAM (64GB of RAM is recommended to support compiling of LLMs with large parameters)
- System supporting PCIe Gen 3 and available 2280 M.2 M-Key slot supporting D6 height
- OS: Ubuntu Version 22.04 LTS
- Development software Required: MERA Software Version 2.2 or later

Note: If the target system is different from the development system, then the memory configuration for the target system can be lower; Please contact Egdecortix for recommended configuration.

Follow the instructions in Chapter 3 (Startup Instructions) to properly prepare your PC, install the M.2 Module, and ensure proper boot-up. Once completed, download the MERA software from EdgeCortix Developer Zone. Install the MERA Compiler Framework by referring to the MERA Installation and User Manual available on Developer Zone. See the Appendix for instructions on how to register for the Developer Zone and access the materials.

2. SAKURA-II M.2 MODULE Features

Figure 1 shows an annotated image of the SAKURA-II M.2 Module, with important components identified:

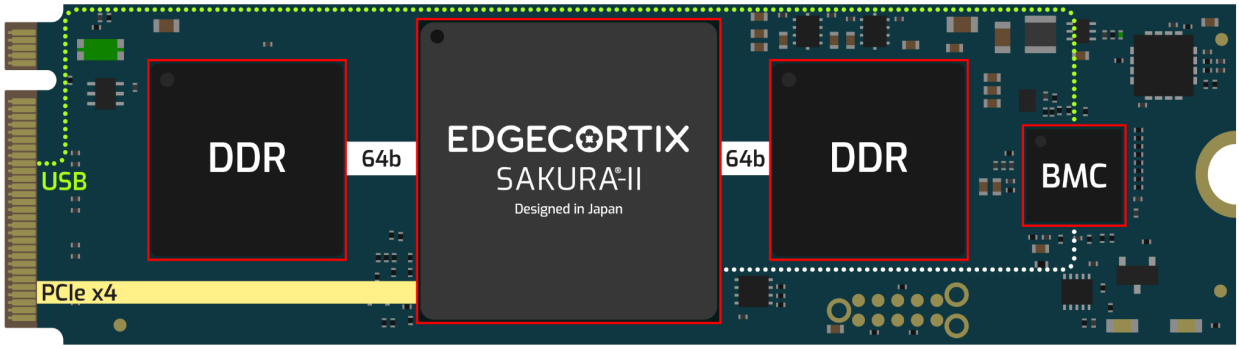


FIGURE 1: SAKURA-II M.2 Module Annotated Image

Specification	M.2 Module
AI Accelerator	Single SAKURA-II
Performance	60 TOPS (INT8) 30 TFLOPS (BF16)
LPDDR4 DRAM	16GB (2 banks of 8GB)
PCIe Interface	Gen 3.0 x4
Board Management Controller (BMC)	Power sequencing, configuration, and reset Voltage, current, and temperature monitoring Protection shut-down SPI Interface to SAKURA-II device
USB Interface	USB-C connector that provides access to BMC for monitoring and control
Cooling Options	Fan Sink (12V Fan)
Electrical	Onboard power derived from M.2 slot (3.3V)
Power Consumption	10W (typical)
Form Factor (without fansink)	2280-D6-M 22 = 22mm width 80 = 80mm length

	D6 = 3.2mm Component Max Height (Top) 1.5mm Component Max Height (Bottom) M = PCIe x4 / SATA / SMBus
Form Factor (with fansink)	24.5mm Width 80mm Length 28.5mm Height
Environmental	-20C to 105C (component operating range) 0 to 95% humidity (non-condensing)

2.1 M.2 MODULES DEVELOPMENT TOOLS

Host Platform	x86-64
Operating System	Ubuntu 22.04 LTS
EdgeCortix Compiler	MERA Compiler Framework
ML Frameworks	PyTorch, ONNX, TensorFlow Lite
Models	Source from Hugging Face or EdgeCortix Model Library

3. STARTUP INSTRUCTIONS

3.1 Preparing the PC

Several steps should be taken prior to installation of the M.2 Module. Please ensure all these steps are completed prior to installation of the M.2 Module.

NOTE: The settings in this section are preliminary and subject to change.

3.1.1 Host PC BIOS Settings

Certain systems must be configured to ensure proper boot up when the M.2 Module is connected. If any of the options below are available, DISABLE them in the BIOS to ensure successful M.2 operation.

1. PCIe ASPM support for the slot in use
2. 4G decoding and BAR resizing settings
3. Fast Boot Up
4. VT-D Support

3.1.2 Operating System Settings

It is necessary to allocate memory for HugePages. The minimum required by the M.2 Module is one HugePage, but if there is more physically contiguous memory available in the system, you can increase the number of HugePages. For example, if the system has 16GB of contiguous RAM, it is reasonable to allocate between 1 and 4 pages.

Instructions for allocating memory for HugePages:

Edit the default command line boot arguments in the file below:

```
/etc/default/grub
```

The original file should look something like this:

```
# If you change this file, run 'update-grub' afterwards to
update
```

```
# /boot/grub/grub.cfg.
# For full documentation of the options in this file, see:
# info -f grub -n 'Simple configuration'

GRUB_DEFAULT T=0
GRUB_TIMOUT_STYLE=hidden
GRUB_TIMOUT=0
GRUB_DISTRIBUTOR='lsb_release -i -s /dev/null || echo Debian'
GRUB_CMDLINE_LINUX_DEFAULT="quiet splash"
GRUB_CMDLINE_LINUX=""
```

Add extra parameters to the variable `GRUB_CMDLINE_LINUX_DEFAULT` as follows:

```
# If you change this file, run 'update-grub' afterwards to
update
# /boot/grub/grub.cfg.
# For full documentation of the options in this file, see:
# info -f grub -n 'Simple configuration'

GRUB_DEFAULT T="0"
GRUB_TIMOUT_STYLE="hidden"
GRUB_TIMOUT="10"
GRUB_DISTRIBUTOR="'lsb_release -i -s /dev/null || echo Debian'"
GRUB_CMDLINE_LINUX_DEFAULT="quiet splash pcie_aspm=off"

# Please modify this variable
GRUB_CMDLINE_LINUX_DEFAULT="quiet splash pcie_aspm=off
default_hugepagesz=1G hugepagesz=1G hugepages=4 iommu=pt"

GRUB_CMDLINE_LINUX=""
```

After modifying the file, update GRUB and reboot.

```
$ sudo update-grub
$ reboot
```

After **restarting** the system, confirm the modifications have been made properly by running:

```
$ grep HugePages_ /proc/meminfo
HugePages_Total:     4
HugePages_Free:      4
HugePages_Rsvd:      0
```



```
HugePages_Surp:       0
```

3.2 SAKURA-II M.2 Module Installation

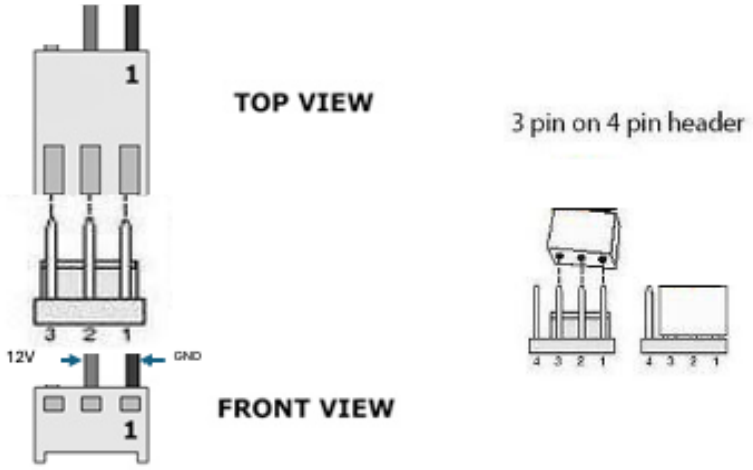
SAKURA-II M.2 Modules are ESD sensitive. Please ensure you take proper precautions to safeguard your board and system. See Appendix for details on ESD.

Install the M.2 Module in an M.2 slot following the guidelines for your specific system. The slot must be Key M.

NOTE: The M.2 module has a fan sink integrated, as shown in the figure and it includes a 3-pin connector for powering the fan. This fan must be connected and running (i.e powered) to ensure proper cooling of the M.2 module to avoid overheating and potential thermal shutdown.

3.2.1 M.2 Module Fan Connection

The fan connector is keyed and fits into a standard 3 or 4-pin fan connector in most PC systems. If your system does not have a way to connect and power the fan, see below for details on the connector and pinout. The fan runs using 12V and below are the connector pinout and recommended connections when using either a 3-pin or 4-pin connector/header.



If you have questions about thermal considerations, please contact EdgeCortix.

3.3 Host PC Boot Up Inspection

Now that the M.2 is connected and the PC boots up properly, open up the Linux terminal and type the following command:

```
sudo lspci
```

The output of the command will be something like this, depending on the PC type and details. Edgecortix has a vendor ID (1FDC) awarded by PCI-SIG and can be found in the output as shown below

```
…
00:1f.5 Serial bus controller [0c80]: …
…
…
# this entry represents the SAKURA evaluation board
01:00.0 Co-processor: Device 1fdc:0001
…
```

Once the M.2 Module has been located, more details can be discovered by typing:

```
sudo lspci -d 01fdc: -vvv
```

This command will display the details of the detected M.2 Module as shown below:

```
01:00.0 Co-processor: Device 1fdc:0001
     Subsystem: Device 1fdc:0001
     Control: I/O- Mem- BusMaster- SpecCycle- MemWINV-
VGASnoop- ParErr- Stepping- SERR- FastB2B- DisINTx-
     Status: Cap+ 66MHz- UDF- FastB2B- ParErr- DEVSEL=fast
>TAbort- <TAbort- <MAbort- >SERR- <PERR- INTx+
     Interrupt: pin A routed to IRQ 255
     IOMMU group: 13
     Region 0: Memory at f6000000 (32-bit, non-prefetchable)
[disabled] [size=8M]
     Region 2: Memory at f5800000 (32-bit, non-prefetchable)
[disabled] [size=8M]
     Region 4: Memory at f5000000 (32-bit, non-prefetchable)
[disabled] [size=8M]
     Capabilities: [80] Power Management version 3
          Flags: PMEClk- DSI- D1- D2- AuxCurrent=0mA
PME(D0+,D1-,D2-,D3hot+,D3cold-)
          Status: D0 NoSoftRst+ PME-Enable- DSel=0 DScale=0
PME-
     Capabilities: [90] MSI: Enable- Count=1/1 Maskable+ 64bit+
```



```
          Address: 0000000000000000  Data: 0000
          Masking: 00000000  Pending: 00000000
     Capabilities: [c0] Express (v2) Endpoint, MSI 00
          DevCap:   MaxPayload 1024 bytes, PhantFunc 0,
Latency L0s <1us, L1 <1us
               ExtTag+ AttnBtn- AttnInd- PwrInd- RBE+ FLReset-
SlotPowerLimit 0.000W
          DevCtl:   CorrErr- NonFatalErr- FatalErr- UnsupReq-
               RlxdOrd- ExtTag+ PhantFunc- AuxPwr- NoSnoop+
               MaxPayload 512 bytes, MaxReadReq 512 bytes
          DevSta:   CorrErr+ NonFatalErr- FatalErr- UnsupReq+
AuxPwr- TransPend-
          LnkCap:   Port #0, Speed 8GT/s, Width x4, ASPM L0s
L1, Exit Latency L0s <256ns, L1 <8us
               ClockPM- Surprise- LLActRep- BwNot-
ASPMOptComp+
          LnkCtl:   ASPM L1 Enabled; RCB 64 bytes, Disabled-
CommClk-
               ExtSynch- ClockPM- AutWidDis- BWInt- AutBWInt-
          LnkSta:   Speed 8GT/s (ok), Width x4 (ok)
               TrErr- Train- SlotClk- DLActive- BWMgmt-
ABWMgmt-
          DevCap2: Completion Timeout: Range B, TimeoutDis+
NROPrPrP- LTR-
                10BitTagComp- 10BitTagReq- OBFF Not Supported,
ExtFmt+ EETLPPrefix-
                EmergencyPowerReduction Not Supported,
EmergencyPowerReductionInit-
                FRS- TPHComp- ExtTPHComp-
                AtomicOpsCap: 32bit- 64bit- 128bitCAS-
          DevCtl2: Completion Timeout: 50us to 50ms,
TimeoutDis- LTR- OBFF Disabled,
                AtomicOpsCtl: ReqEn-
          LnkCap2: Supported Link Speeds: 2.5-8GT/s, Crosslink-
Retimer- 2Retimers- DRS-
          LnkCtl2: Target Link Speed: 8GT/s, EnterCompliance-
SpeedDis-
                Transmit Margin: Normal Operating Range,
EnterModifiedCompliance- ComplianceSOS-
                Compliance De-emphasis: -6dB
          LnkSta2: Current De-emphasis Level: -3.5dB,
EqualizationComplete+ EqualizationPhase1+
                EqualizationPhase2+ EqualizationPhase3+
LinkEqualizationRequest-
                Retimer- 2Retimers- CrosslinkRes: unsupported
     Capabilities: [100 v2] Advanced Error Reporting
          UESta:    DLP- SDES- TLP- FCP- CmpltTO- CmpltAbrt-
UnxCmplt- RxOF- MalfTLP- ECRC- UnsupReq- ACSViol-
          UEMsk:    DLP- SDES- TLP- FCP- CmpltTO- CmpltAbrt-
UnxCmplt- RxOF- MalfTLP- ECRC- UnsupReq- ACSViol-
          UESvrt:   DLP+ SDES+ TLP- FCP+ CmpltTO- CmpltAbrt-
UnxCmplt- RxOF+ MalfTLP+ ECRC- UnsupReq- ACSViol-
```

```
            CESta:    RxErr- BadTLP- BadDLLP- Rollover- Timeout-
AdvNonFatalErr+
            CEMsk:    RxErr- BadTLP- BadDLLP- Rollover- Timeout-
AdvNonFatalErr+
            AERCap:   First Error Pointer: 00, ECRCGenCap+
ECRCGenEn- ECRCChkCap+ ECRCChkEn-
                  MultHdrRecCap- MultHdrRecEn- TLPPfxPres-
HdrLogCap-
            HeaderLog: 00000000 00000000 00000000 00000000
      Capabilities: [150 v1] Device Serial Number
00-00-00-00-00-00-00-00
      Capabilities: [300 v1] Secondary PCI Express
            LnkCtl3: LnkEquIntrruptEn- PerformEqu-
            LaneErrStat: 0
```

Please make sure that these settings are on your lspci command outputs, especially the following lines:

```
01:00.0 Co-processor: Device 1fdc:0001
Subsystem: Device 1fdc:0001
...
      Region 0: Memory at f6000000 (32-bit, non-prefetchable)
      [disabled] [size=8M]
            Region 2: Memory at f5800000 (32-bit, non-prefetchable)
      [disabled] [size=8M]
            Region 4: Memory at f5000000 (32-bit, non-prefetchable)
      [disabled] [size=8M]
...
      Capabilities: [c0] Express (v2) Endpoint, MSI 00
...
      LnkSta: Speed 8GT/s (ok), Width x4 (ok)
...
```

A couple of points to note with the lspci-vvv output:

1. Please note that the lspci command output will be slightly different in each host system. For example:

The 01:00.0 may be a different number, however the device ID and subsystem ID will always be **1fdc**.

The memory locations highlighted in **red** will be different in each system, which is OK. As long as size=8M and the region has


```
a memory allocation (for example, Memory at somevalue) the
system will operate properly.
```

2. LinkSpeed and Width should also be checked:

```
LinkSta: Speed 8GT/s (ok), Width x4 (ok)

These values should match exactly. If the speed is lower than
8GT/s and/or the link width is less than x4, please ensure card
is in the correct M.2 Key M slot that supports PCIe Gen 3 x4.
The M.2 card will still operate if it is lower speed and/or
width (due to slower PCI slot or sharing of the slot), but
performance will be impacted and the extent of the impact will
depend on the model and workload that is being executed.
```

At this point, the M.2 module is verified to have been installed properly and is ready to use. You can start evaluating and developing with SAKURA-II and MERA software.

4. MERA COMPILER FRAMEWORK FEATURES

The MERA Software Stack provides a full end-to-end deployment framework for EdgeCortix SAKURA-II platforms. It provides:

- Ability to import models in PyTorch, TensorFlow Lite, and ONNX formats.
- Support for INT8 and BF16 precision models quantized with the built-in quantization tools of PyTorch and TensorFlow..
- Support for EdgeCortix custom quantization and the ability to quantize FP32 models from PyTorch, TensorFlow Lite, and ONNX using only MERA Quantizer tools.
- Multi-network support that allows the fusing of several models together into a single workload to maximize hardware utilization. Several models can be compiled and optimized together into a single deployment binary artifact.
- Several targets to validate models on increasing level optimizations.
- Interpreters to emulate the DNA platform's internal math with a minimal amount of optimizations.
- Software simulators to perform functional and cycle-accurate simulations of the MERA DNA platforms on x86 hardware.
- Different user configurable levels of optimization for fast development, validation, and testing.

For details on MERA installation and development instructions, refer to the MERA Installation and User Manual.

There may be slight differences in the installation and other processes if using software or hardware other than the ones listed in this document. If you run into any problems, or have any questions about thermal considerations, please contact EdgeCortix.

5. REVISION HISTORY

Revision	Date	Summary
0.71	March 2025	Fixed typo for hugepage in section 3.1.2
0.7	February 2025	Initial Release

All intellectual property rights in and to all of EdgeCortix's trademarks, logos, software, and any and all other intellectual property rights in all material or content that we provide or is contained on our website, including all of our written materials and manuals, shall remain at all times owned by us or, in the cases where we are using such material or content under authority from a third party, in the owner of such material or content. You have no rights in or to any such intellectual property, materials, or content other than as specifically licensed to you by us.

The information contained in these materials is applicable only to the intended uses of the EdgeCortix software and systems. We will periodically update these written materials without notice to you. We encourage you to review the materials on a regular basis to make sure you have the most up-to-date information. Nothing in any of these materials shall be construed as modifying, amending, or supplementing the license agreement between us or any warranties we made.

6. APPENDIX

6.1 Downloading Resources from Developer Zone

As noted, software resources are not included in the shipment and can be downloaded from the EdgeCortix Developer Zone. Please visit the Developer Zone and complete the form to request access to the Developer Zone and associated documentation and MERA software downloads.

If you have any issues accessing the Developer Zone, please contact EdgeCortix.

6.2 ESD Protection and Warnings

SAKURA-II M.2 Modules are populated with electrostatic discharge (ESD) sensitive devices which can be damaged by static charges that can build up on people, tools, and other surfaces. Proper care must be taken in handling these devices and proper grounding must be maintained to ensure that any ESD does not damage any devices on the M.2 Module.

It is beyond the scope of this document to explain and provide specific ESD protection schemes, but users should be familiar with these processes that apply to all ESD-sensitive semiconductor devices. No warranty is provided for improper handling of the SAKURA-II M.2 Module and damage to any devices on the module is the full responsibility of the user.

