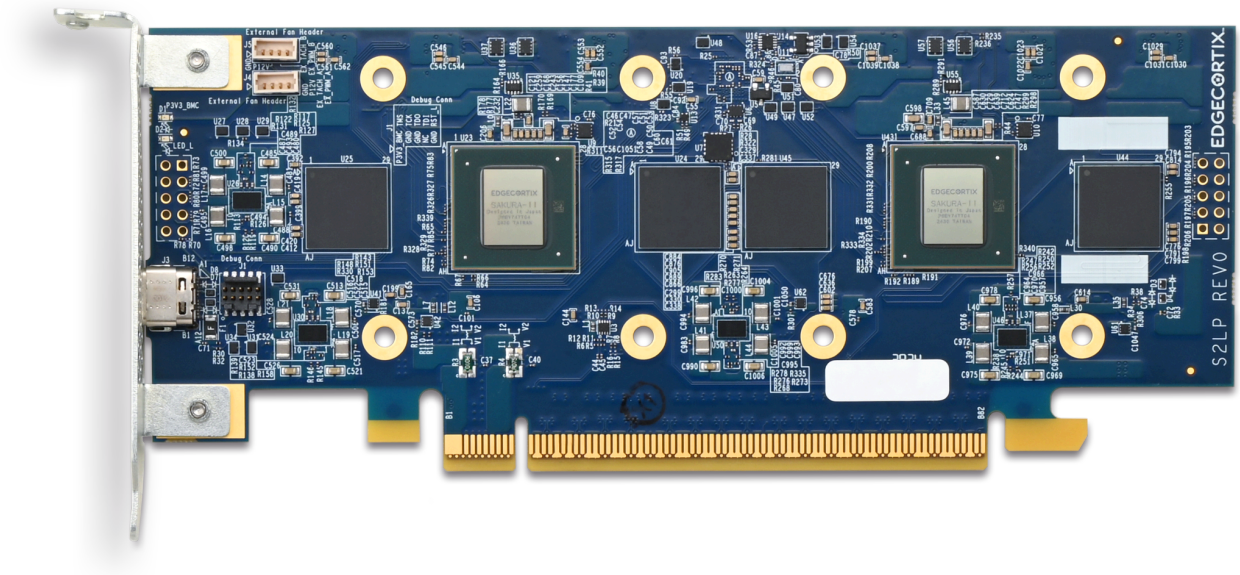
EDGECORTIX

SAKURA-II PCIe Cards

User Manual



Revision History

TABLE OF CONTENTS

1. OVERVIEW

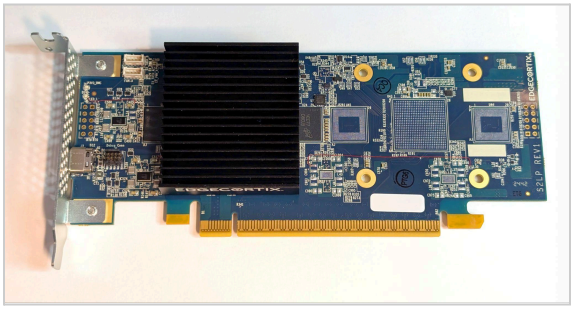
This User Manual covers the EdgeCortix SAKURA-II PCIe Cards. There are two major versions of the Cards,

- Single SAKURA-II Edge AI Accelerator (S2LP-S)
- Dual SAKURA-II Edge AI Accelerator (S2LP-D)

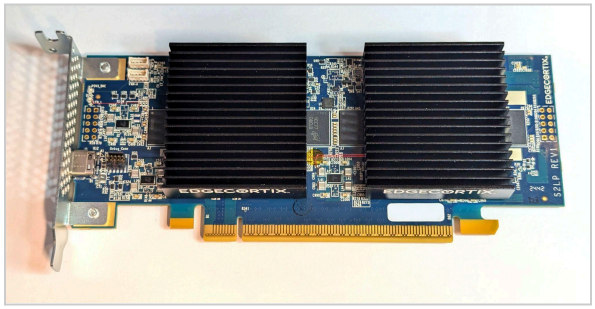
Throughout this manual these will be referred to as “single PCIe Card” and “dual PCIe Card” or just “PCIe Card” for notes that affect both cards.

The PCIe Cards feature SAKURA-II, a high performance edge AI accelerator that boasts 60 TOPS using EdgeCortix’s Dynamic Neural Accelerator (DNA) architecture. It is run-time reconfigurable using the EdgeCortix MERA compiler and software framework to 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.

S2LP-S



S2LP-D



This manual is intended for two target audiences:

- Users of stand-alone PCIe Cards that will be inserted into their own systems
- Users of pre-configured systems

SYSTEM REQUIREMENTS

If the target and development system are the same where the PCIe card is to be installed, below are the system requirements:

- Intel or AMD x86 based Linux PC with minimum 32GB of RAM (64GB of RAM is recommended to support compiling of LLMs with large parameters)
- System supporting PCIe Gen 3 and available PCIe x16 slot
- SAKURA-II single PCIe Card OR SAKURA-II dual PCIe Card. For optimal bandwidth, x8 or x16 electrical is recommended

- For supporting Dual PCIe Card: x16 electrical is required, and must be bifurcated as x8/x8 in order to use both SAKURA-II devices
- Minimum forced airflow is required over the PCIe Card with the passive heatsink
- 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.

1.1 Instructions - Stand-Alone PCIe Cards

If you have received a stand-alone PCIe Card from EdgeCortix, be sure to follow the instructions in Chapter 3 (Startup Instructions) to properly prepare your PC, install the PCIe Card 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.

1.2 Instructions - Pre-Configured Systems

If you have received a pre-configured system from EdgeCortix, these systems come with the PCIe Card and MERA Compiler already installed. You do not need to complete any of the steps in per the Startup Guide and can start using the system immediately. For more information on your specific systems, please refer to the SAKURA-II Edge AI Evaluation System Quick Start Guide.

NOTE: Sections 2 provides detailed information on the PCIe Cards and may also apply to the cards installed in a pre-configured system.

2. SAKURA-II PCIe CARD Features

Figure 1 shows an annotated image of the SAKURA-II single PCIe Card, with important components identified:

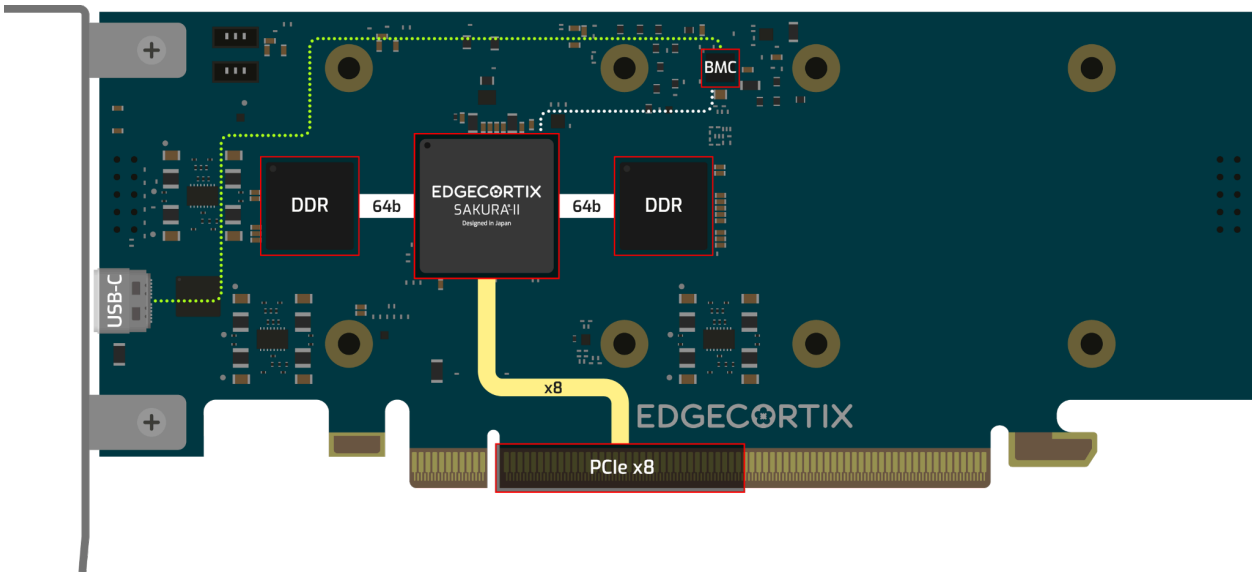


FIGURE 1: SAKURA-II single PCIe Card Annotated Image

Figure 2 shows an annotated image of the SAKURA-II dual PCIe Card:

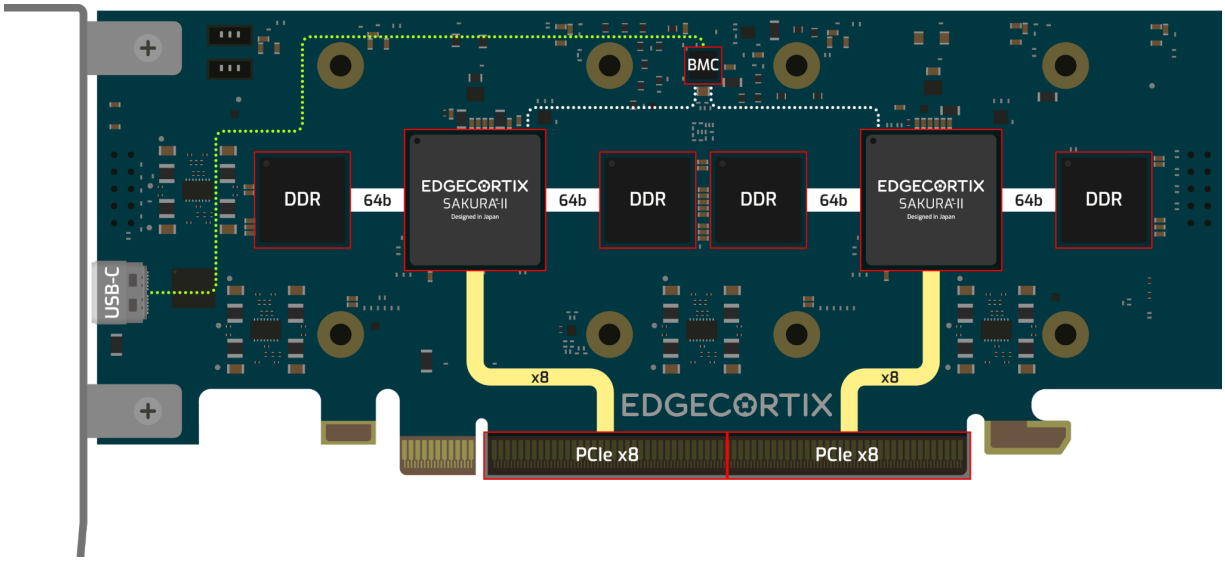


FIGURE 2: SAKURA-II dual PCIe Card Annotated Image

Specification	Single PCIe Card	Dual PCIe Card
AI Accelerator	Single SAKURA-II	Dual SAKURA-II
Performance	60 TOPS (INT8) 30 TFLOPS (BF16)	120 TOPS (INT8) 60 TFLOPS (BF16)
LPDDR4 DRAM	16GB (2 banks of 8GB)	32GB (4 banks of 8GB)
PCIe Interface	Gen 3.0 x8	Gen 3.0 x8/x8 (bifurcated)
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 on PCIe bracket Provides access to BMC for monitoring and control	
Cooling Options	Heatsink	
Electrical	Onboard power derived from PCIe slot (12V and 3.3V)	
Power Consumption (Default mode)	10W (typical)	20W (typical)
Form Factor	PCIe Add-In Card, Low Profile, Half-Length 167.65mm x 68.90mm, Single Slot	
Bracket Options	Low Profile and Full Height	
Environmental	-20C to 105C (component operating range) 0 to 95% humidity (non-condensing)	

2.1 PCIe Card 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 PCIe Card. Please ensure all these steps are completed prior to installation of the PCIe Card.

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 PCIe Card is connected. If any of the options below are available, DISABLE them in the BIOS to ensure successful PCIe 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

When using the dual PCIe Card, in order to use both SAKURA-II devices you must configure the PCIe Bus to support x8/x8 bifurcation.

3.1.2 Operating System Settings

It is necessary to allocate memory for HugePages. The minimum required by the PCIe Card 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 PCIe Card Installation

SAKURA-II PCIe cards are ESD sensitive. Please ensure you take proper precautions to safeguard your board and system. See ESD Protection and Warnings for details on ESD.

Install the PCIe Card in a PCIe slot following the guidelines for your specific system.. A x16 PCIe mechanical slot is required. For the single PCIe Card, x8 or x4 electrical slots can be used. For the dual PCIe Card, the PCIe bus must be bifurcated, and must be configured as x8/x8 to support both SAKURA-II devices. Once the PCIe Card is installed, boot the PC.

NOTE: Minimal forced airflow is recommended over the PCIe Card with the passive heatsink. If you have questions about thermal considerations, please contact EdgeCortix.

3.3 Host PC Boot Up Inspection

Now that the PCIe Card 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 of the above command as shown below

```
00:00.0 Host Bridge…
…

01:00.0 Co-processor: Device 1fdc:0100
# This entry represents the SAKURA-II evaluation board. The 01:00.0
number can vary depending on the PC type, but the device ID will be
the same. If the card is a dual PCIe card, then two instances of the
```



```
device will be displayed, i.e you will see another instance similar
to the line below in the output of the above command
…
03:00.0 Co-procesor: Device 1fdc:0001
…
```

Once the PCIe Card has been located, more details can be discovered by typing:

```
sudo lspci -d 1fdc: -vvv
```

This command will display the details of the detected PCIe card(s) 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: 14
          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 x8, 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 x8 (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 x8 (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:

```
01:00.0 Co-processor: Device 1fdc:0100

Subsystem: Device 1fdc:0001

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 x8 (ok)

These values should match exactly. If the speed is lower than
8GT/s and/or the link width is less than x8, ensure the card is
in the correct slot. If not, remove and replace into the correct
slot for PCIe Gen 3 x16
```



```
If the LinkWidth is lower than x8, it could also mean that the
slot is shared with another PCIe card or a M.2 SSD device which
could bring down the link width. In these systems, an option
would be to move the M.2 SSD to another slot.

The PCIe card will still operate with the lower speed and width,
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 PCIe card 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 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 above. If you run into any problems, or have any questions, please contact EdgeCortix.

9 REVISION HISTORY

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

10 APPENDIX

10.1 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.

10.2 ESD Protection and Warnings

SAKURA-II PCIe Cards 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 PCIe Card.

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 PCIe Card and damage to any devices on the Card is the full responsibility of the user.

10.3 SAKURA-II PCIe Card Board Management Controller

The PCIe card has a Universal Serial Bus Type -C (USB-C) connector that can be used to interact directly with the Board Management Controller (BMC) for board information and status.

10.3.1 Connecting to the USB-C Port

The USB-C connector is used with terminal emulation software such as Minicom under Linux or PuTTY under Windows which can be running on the host machine or an external computer (in case of Windows) such as a laptop or desktop.

The terminal settings are:

1. 115200-N-8-1 (115200 Baud, No parity, 8-bits, 1-stop bit).

2. Local Echo should be turned ON.
3. Enable Implicit CR (carriage return) with every LF (linefeed)

10.3.2 Example Commands

To see board information such as Build Configuration and Serial Number type in the serial terminal window:

cfg

```
EEPROM Stored Parameters
========================
Board Name     = "S2XX-XX"
Serial (text)  = "XXXXX-XXXXXX"
Serial Number  = XXXXXXXX = 0x1234ABCD
Board Revision = 0
ECN Level      = 0
EEPROM Format  = 0
DDR Type       = 0
Sakura Version = 2.00
Mfg Date       = YYYY.MM.DD
ECN Date       = 0000.00.00

PCIe Lane BA   = 0x01
PMode          = 0
vCore          = 550
BMC max temp   = XXX C
Sakura max temp= XX C
board max temp = XX C
```

To see detailed information on software version and total power type:

info

```
Board:       EdgeCortix SAKURA BMC
BMC Software: Revision 0.1.1, built Sep 12 2024 23:58:36
CPUID:       0x410fd214, Cortex-M33, rev 0, patch 4
PROCESSOR:   R7FA4M3AF2CB, 1 MB flash, -40 to 105C temp range, FBGA 64 Pins
             UID 0x42168168 0x35383437 0x9f2a4753 0x4e4b2d03
PLL:         HOCO in 16 MHz, CPU freq 66.666 MHz
Reset:       00 0004 00
upTime:      13:27:08
HEAP:        26960 available, 26960 largest, 1 free blocks
             15 allocates, 0 frees, %17 used, %83 free

Board Temp:   35 C
Voltage_3.3V  3.28 V
Current_3.3V  0.77 A
Input Power   2.52 W
```


10.4 PCIe Bracket Options

The PCIe Cards can come in a variety of configurations, with a pre-installed low profile bracket, a pre-installed full height bracket, or no bracket installed. The next sections show the process to uninstall and install either bracket. Figure 3 shows all of the components, with the low profile bracket installed and the full height bracket off to the side.

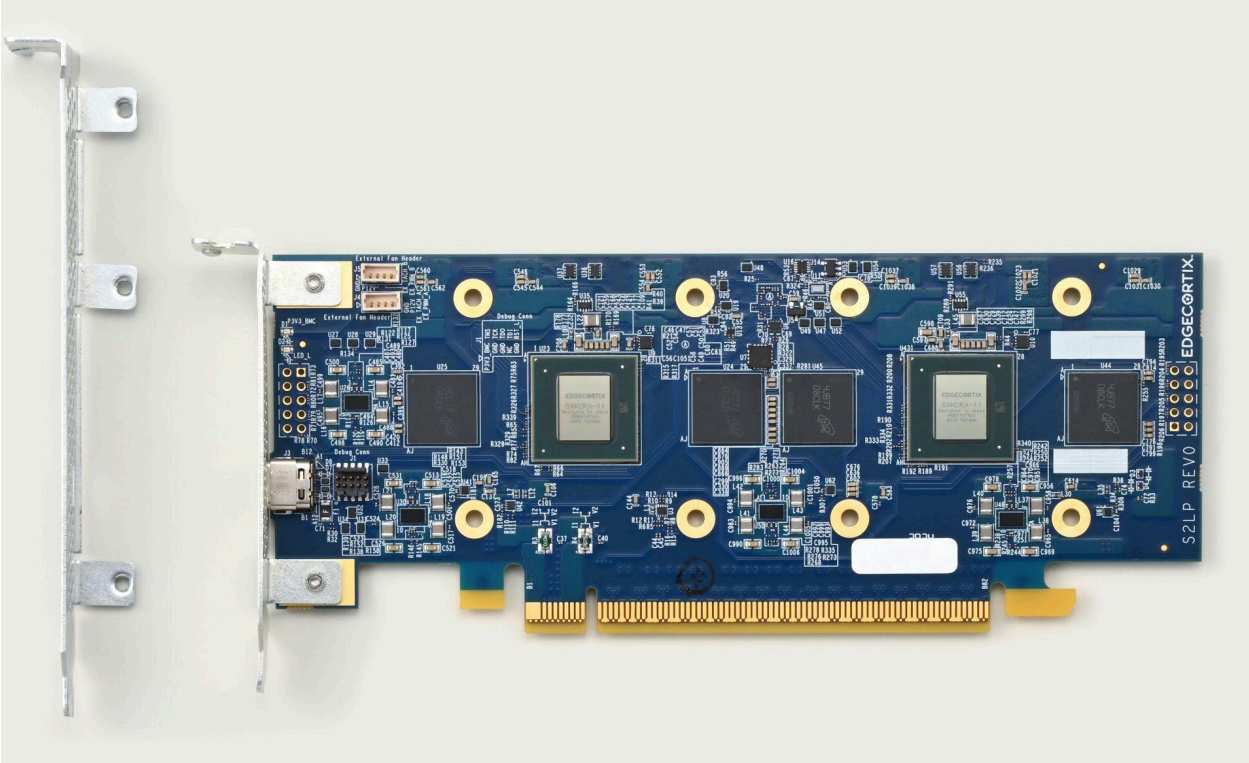


FIGURE 3 – PCIe Card and Brackets

10.4.1 Low Profile Bracket Removal and Full Height Bracket Installation

To remove the low profile bracket, please remove the screws circled in Figure 4.

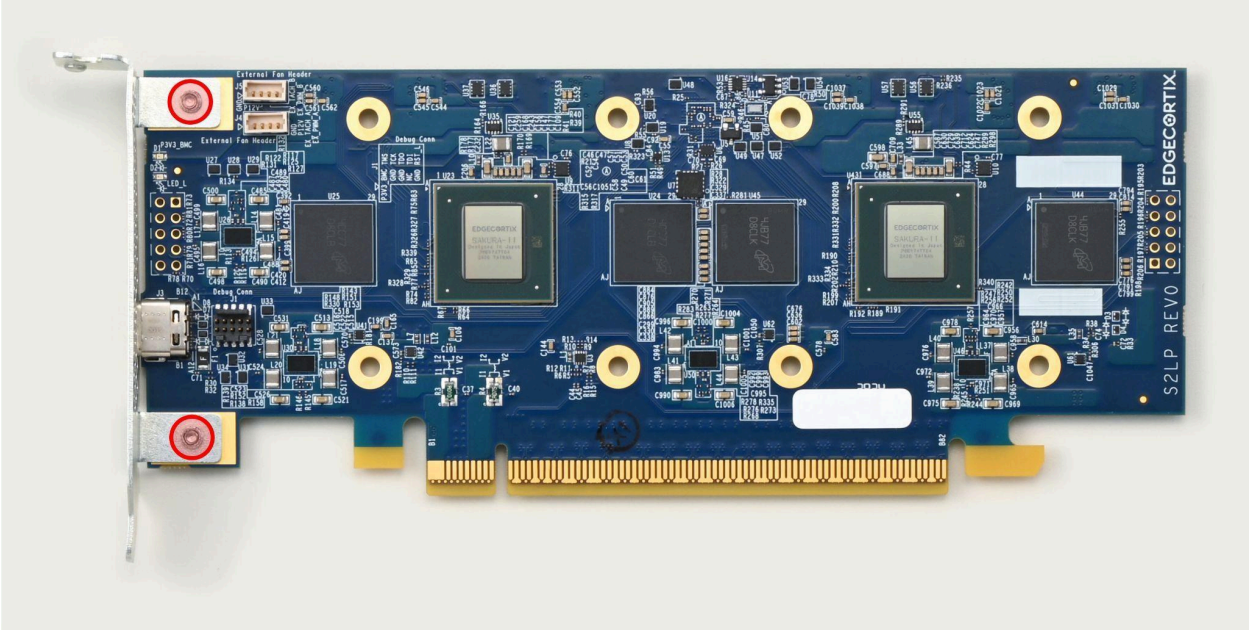


FIGURE 4. PCIe Card with Low Profile Bracket Installed

The full height bracket MUST be installed with the screws installed in the reverse direction, from the top as shown in Figure 5 below.

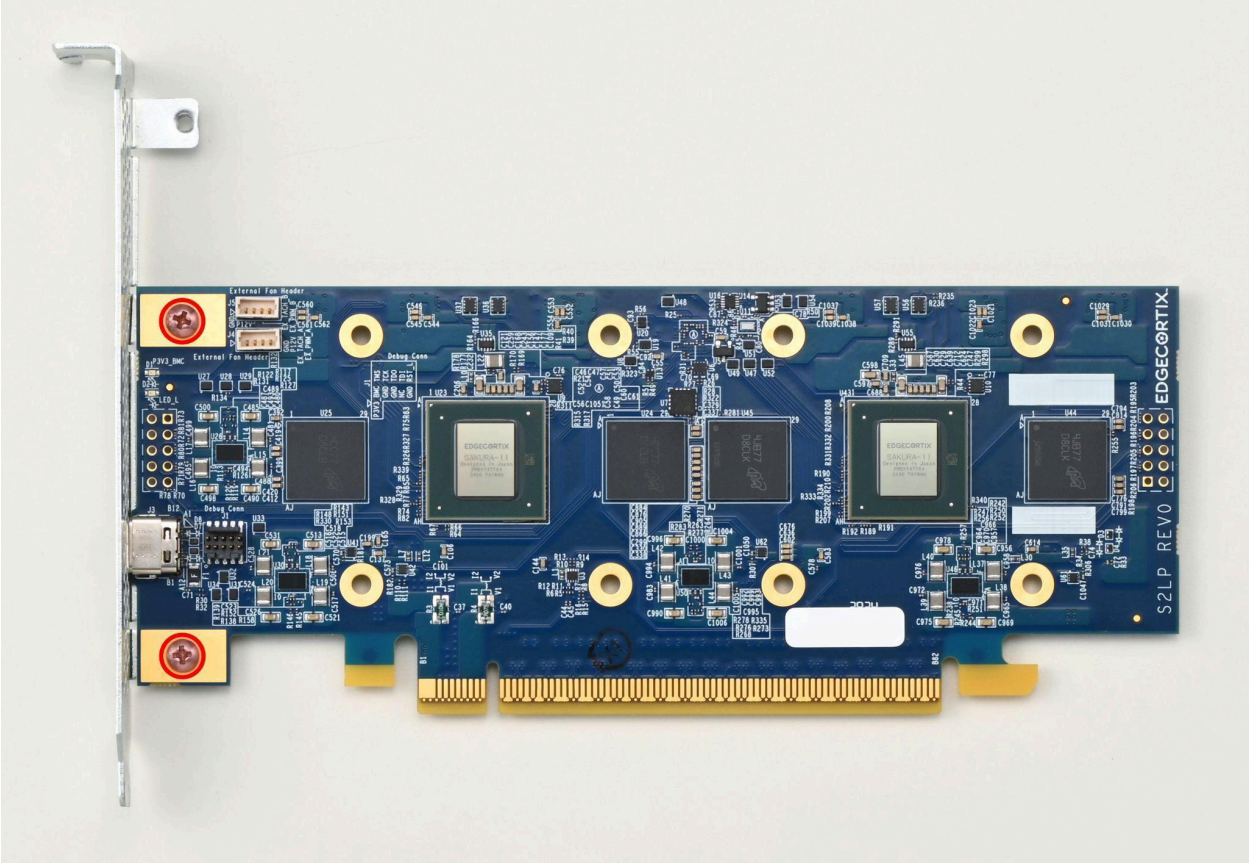


FIGURE 5: PCIe Card with Full Height Bracket Installed

10.4.2 Full Height Bracket Removal and Low Profile Bracket Installation

To remove the full height bracket, please remove the screws circled in Figure 6.

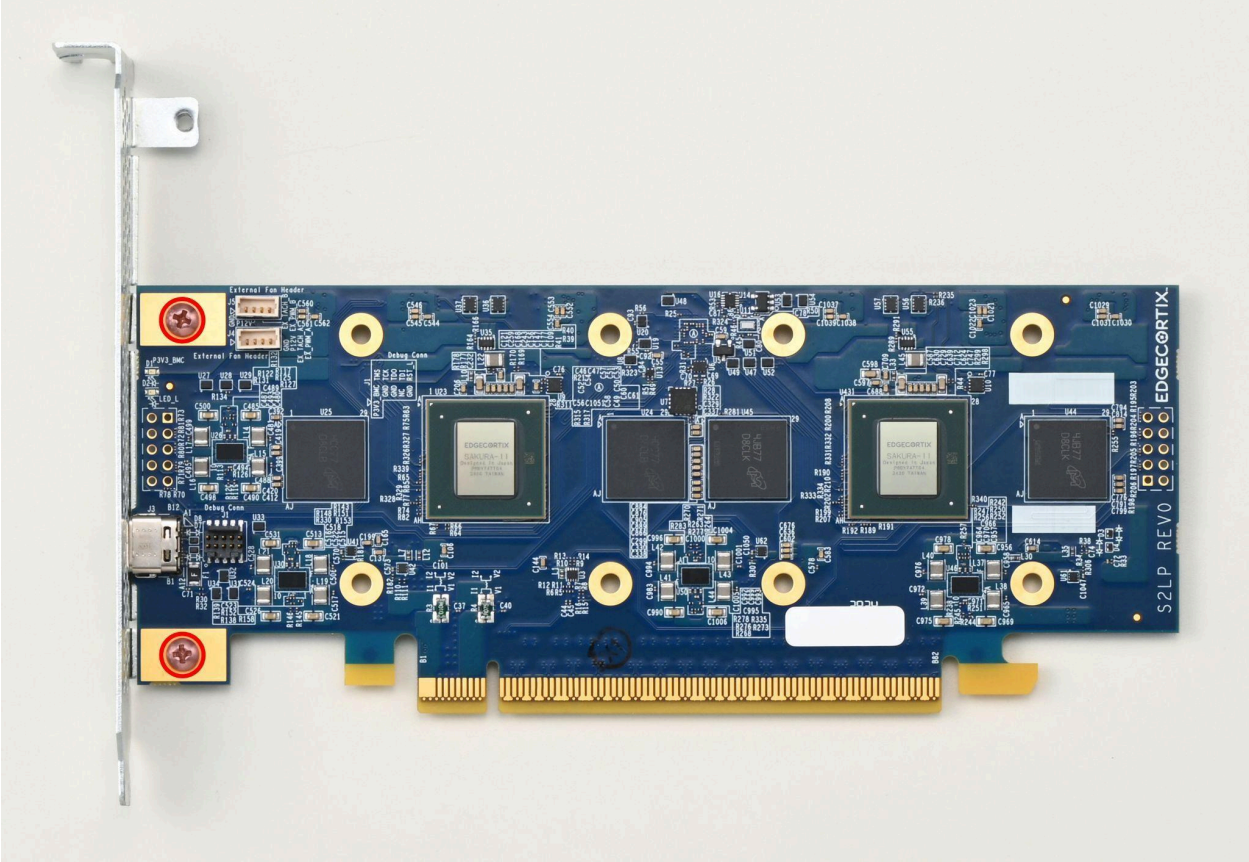


FIGURE 6: PCIe Card with Full Height Bracket Installed (Rev 0)

The low profile bracket MUST be installed with the screws in the reverse direction, from the bottom as shown in Figure 7 below.

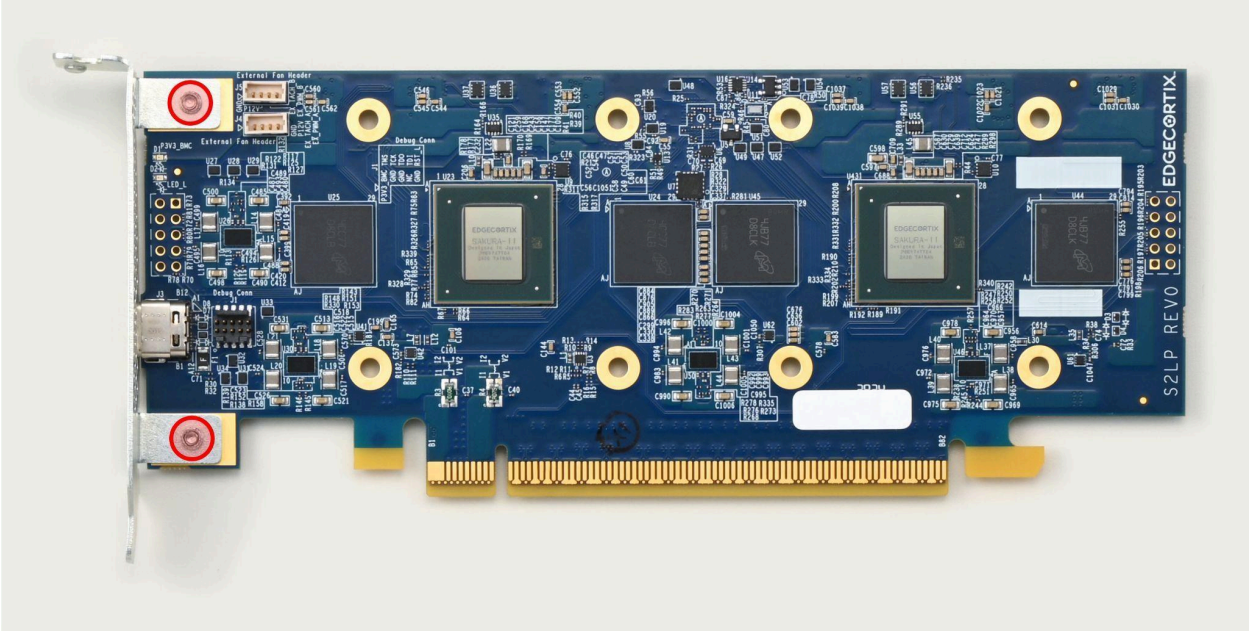


FIGURE 7: PCIe Card with Low Profile Bracket Installed

