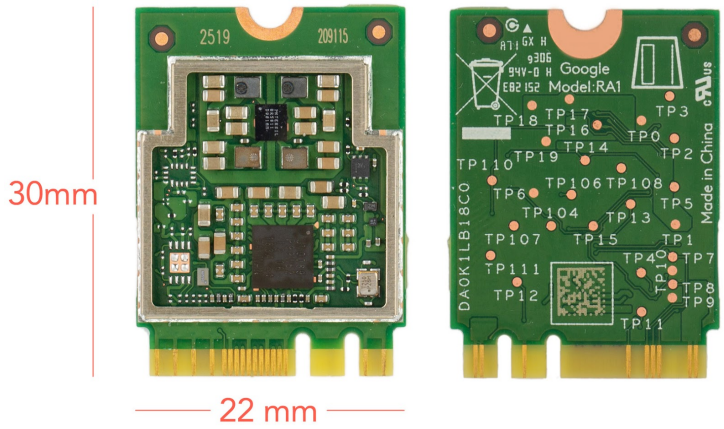
Coral

M.2 Accelerator datasheet

Version 1.5

Features

- Google Edge TPU ML accelerator
  - 4 TOPS total peak performance (int8)
  - 2 TOPS per watt
- Integrated power management
- PCIe Gen2 x1 interface
- Available in two M.2 form factors:
  - M.2-2230-A-E-S3 (A+E key)
  - M.2-2280-B-M-S3 (B+M key)
- Operating temp: -20 to +85 °C



Description

The Coral M.2 Accelerator is an M.2 module (either A+E or B+M key) that brings the Edge TPU ML accelerator to existing systems and products.

The Edge TPU is a small ASIC designed by Google that accelerates TensorFlow Lite models in a power efficient manner: it's capable of performing 4 trillion operations per second (4 TOPS), using 2 watts of power—that's 2 TOPS per watt. For example, one Edge TPU can execute state-of-the-art mobile vision models such as MobileNet v2 at almost 400 frames per second. This on-device ML processing reduces latency, increases data privacy, and removes the need for a constant internet connection.

The M.2 form-factor allows you to add local ML acceleration to products such as embedded platforms, mini-PCs, and industrial gateways that have a compatible M.2 card slot.

Ordering information

Part number	Description
G650-04527-01	Coral M.2 Accelerator with A+E key
G650-04686-01	Coral M.2 Accelerator with B+M key

See https://coral.ai/products/m2-accelerator-ae.

Table of contents

1 Specifications

For in-depth mechanical details, refer to the PCI-SIG's PCI Express M.2 specification.

Table 1. Technical specifications

Physical specifications	
Dimensions	A+E key: 22.00 x 30.00 x 2.35 mm B+M key: 22.00 x 80.00 x 2.35 mm
Weight	A+E key: 3.1 g B+M key: 5.8 g
Host interface	
Hardware interface	M.2 A+E key (M.2-2230-A-E-S3) or M.2 B+M key (M.2-2280-B-M-S3)
Serial interface	PCIe Gen2 x1
Operating voltage	
DC supply	3.3 V +/- 10 %
Environmental	
Storage temperature	-40 to +85 °C
Operating temperature	-20 to +85 °C [1]
Relative humidity	0 to 90% (non-condensing)
Mechanical (non-op)	
Shock	100 G, 11 ms (persistent) 1000 G, 0.5 ms (stress) 1000 G, 1.0 ms (stress)
Vibration (random/sinusoidal)	0.5 Grms, 5 - 500 Hz (persistent) 3 Grms, 5 - 800 Hz (stress)
Compliance	
Countries [2]	Unit shipped as a component. Final system certification/compliance to be done by the customer.
ESD [3]	1 kV HBM, 250 V CDM

[1] The max operating temperature depends on the power consumption and thermal management in your system.
[2] We can provide a certification example to show that a reasonably designed system can meet certification requirements.
[3] Always handle in a static safe environment.

2 Dimensions

2.1 A+E key dimensions

- PCB width: 22.00 mm ± 0.15 mm
- PCB height: 30.00 mm ± 0.15 mm
- PCB thickness: 0.80 mm ± 0.08 mm
- Top-side component height: 1.55 mm ± 0.10 mm
- Bottom-side component height: 0 mm

For in-depth mechanical specs, refer to the PCI Express M.2 Specification.

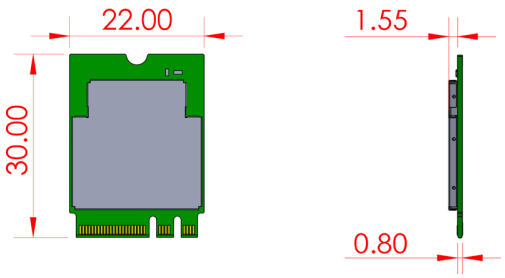


Figure 1. A+E key card dimensions (in millimeters)

2.2 B+M key dimensions

- PCB width: 22.00 mm ± 0.15 mm
- PCB height: 80.00 mm ± 0.15 mm
- PCB thickness: 0.80 mm ± 0.05 mm
- Top-side component height: 1.55 mm ± 0.10 mm
- Bottom-side component height: 0 mm

For in-depth mechanical specs, refer to the PCI Express M.2 Specification.

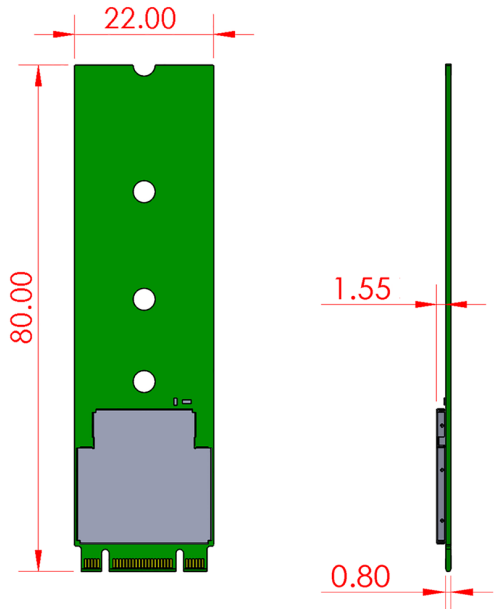


Figure 2. B+M key card dimensions (in millimeters)

3 Electrical characteristics

3.1 Absolute maximum ratings

Exceeding the absolute ratings can cease operation and possibly cause permanent damage. Exposure to absolute ratings for extended periods of time can also adversely affect reliability.

Table 2. Absolute maximum ratings

Parameter	Min	Max
Storage temperature	-40 °C	85 °C
Operating temperature	-20 °C	85 °C [1]
Edge TPU junction temperature (T_j)	-40 °C	115 °C
Power supply (3.3 V)	-0.3 V	6.0 V

[1] The maximum operating temperature is for the entire assembly and assumes that the Edge TPU junction temperature (T_j) does not exceed its absolute maximum rating, which depends on the power consumption and thermal management in your system.

3.2 Power consumption

The power consumed by the card module depends on the ML model, the number of inferences per second, and the operating frequency of the Edge TPU. For some examples of average sustained power consumption, see table 3. However, it's also important that you consider the peak current transients that occur during inferencing.

The maximum current drawn by the Edge TPU is typically much higher than the average current. That's because when the Edge TPU executes an ML model, it repeatedly activates a large number of arithmetic logic units (ALUs) simultaneously, resulting in a pattern of brief but large current transients. Each model architecture also activates a different set and different number of ALUs, meaning the magnitude and the shape of the transient current very much depends on the model.

Although the average current drawn from the 3.3V supply is typically less than 500 mA, brief current transients that occur during inferencing can reach roughly 3 A. These spikes also occur suddenly: even a simple model can generate current transients in excess of 1 A/µs. However, these numbers are representative of only the models tested at Google, and your numbers will vary. To determine the actual peak supply current, you should observe the current when running the models you will deploy in production.

For more information, see section 5.2 Power delivery and management.

Table 3. Examples of long-term sustained power during inferencing

Model [1]	Low operating frequency 125 MHz	Reduced operating frequency 250 MHz	Max operating frequency 500 MHz
MobileNet v2	0.6 W (7.1 ms @ 141 fps)	0.9 W (3.9 ms @ 256 fps)	1.4 W (2.4 ms @ 416 fps)
Inception v3	0.5 W (58.7 ms @ 17 fps)	0.6 W (51.7 ms @ 19.3 fps)	0.7 W (48.2 ms @ 20.7 fps)

[1] Pre-compiled models were tested using models_benchmark.cc

Typical idle power consumption is 375 - 400 mW.

3.3 Peak performance

Peak performance when the Edge TPU is running at the maximum operating frequency:

- 4 trillion operations per second (TOPS), 8-bit fixed-point math
- 2 TOPS per watt

4 Connector pinout

4.1 A+E key pinout

Table 4. A+E key card pinout

Bottom side pins		Top side pins	
Pin	**Signal**	**Signal**	**Pin**
74	3.3V	GND	75
72	3.3V	NC	73
70	NC	NC	71
68	NC	GND	69
66	NC	NC	67
64	NC	NC	65
62	NC	GND	63
60	NC	NC	61
58	NC	NC	59
56	NC	GND	57
54	NC	NC	55
52	PERST0# (3.3V)	CLKREQ0# (3.3V)	53
50	NC	GND	51
48	NC	REFCLKn0	49
46	NC	REFCLKp0	47
44	NC	GND	45
42	NC	PETn0	43
40	NC	PETp0	41
38	NC	GND	39
36	NC	PERn0	37
34	NC	PERp0	35
32	NC	GND	33
30	Key E Slot	Key E Slot	31
28	Key E Slot	Key E Slot	29
26	Key E Slot	Key E Slot	27
24	Key E Slot	Key E Slot	25
22	NC	NC	23
20	NC	NC	21
18	GND	NC	19
16	NC	NC	17
14	Key A Slot	Key A Slot	15
12	Key A Slot	Key A Slot	13
10	Key A Slot	Key A Slot	11
8	Key A Slot	Key A Slot	9
6	NC	GND	7
4	3.3V	NC	5
2	3.3V	NC	3
		GND	1

4.2 B+M key pinout

Table 5. B+M key card pinout

Bottom side pins		Top side pins	
Pin	**Signal**	**Signal**	**Pin**
74	3.3V	GND	75
72	3.3V	GND	73
70	3.3V	GND	71
68	NC	NC	69
66	Key M Slot	NC	67
64	Key M Slot	Key M Slot	65
62	Key M Slot	Key M Slot	63
60	Key M Slot	Key M Slot	61
58	NC	Key M Slot	59
56	NC	GND	57
54	NC	REFCLKp0	55
52	CLKREQ0# (3.3V)	REFCLKn0	53
50	PERST0# (3.3V)	GND	51
48	NC	PERp0	49
46	NC	PERn0	47
44	NC	GND	45
42	NC	PETp0	43
40	NC	PETn0	41
38	NC	GND	39
36	NC	NC	37
34	NC	NC	35
32	NC	GND	33
30	NC	NC	31
28	NC	NC	29
26	NC	GND	27
24	NC	NC	25
22	NC	NC	23
20	NC	GND	21
18	Key B Slot	Key B Slot	19
16	Key B Slot	Key B Slot	17
14	Key B Slot	Key B Slot	15
12	Key B Slot	Key B Slot	13
10	NC	NC	11
8	NC	NC	9
6	NC	NC	7
4	3.3V	NC	5
2	3.3V	GND	3
		GND	1

5 Application details

5.1 Software requirements

The M.2 Accelerator must be operated by the Edge TPU runtime and Coral PCIe driver, which is compatible with the following systems:

- Linux:
  - 64-bit version of Debian 10 or Ubuntu 16.04 (or newer)
  - x86-64 or ARMv8 system architecture
- Windows:
  - 64-bit version of Windows 10
  - x86-64 system architecture
- All systems require support for MSI-X as defined in the PCI 3.0 specification

5.2 Power delivery and management

> **Caution:** If you do not carefully consider the power demands of the ML models running the Edge TPU, along with the ability of your host to handle the corresponding current transients, the peak currents might cause brownouts or other abnormal behavior in the upstream power regulator.

As described in section 3.2 Power consumption, the current drawn by the Edge TPU is highly variable and depends on the model being executed. Although the average current drawn by the Edge TPU might seem low (less than 500 mA), it can repeatedly and rapidly spike up to 3 A, depending on the model you're running. These spikes also occur suddenly: even a simple model can generate current transients in excess of 1 A/µs, which can last several tens of microseconds.

Ideally, your host system and M.2 socket can be designed to tolerate these higher currents, and your power supply can provide fast transient response performance. Alternatively, you may use some software strategies to mitigate the effects of the peak currents, such as underclocking the Edge TPU.

5.3 Thermal management

The Edge TPU dissipates power roughly proportional to its computational load. The resulting heat in the Edge TPU die must be safely and reliably conducted away to avoid excessive die temperatures that can affect performance and reliability.

The primary heat-generating components on the card are the Edge TPU and power IC, located under the shield can as indicated in figure 3. The shield can provides thermal coupling to these components with thermal pads—there is no air gap between the components and the shield can. (For thermal resistance detail, see section 5.3.2 Top-side cooling options.)

During typical operation, approximately 90% of the system power dissipates from the Edge TPU, and the remaining 10% dissipates from the power IC. Total power dissipation depends on the operating frequency and computational load.

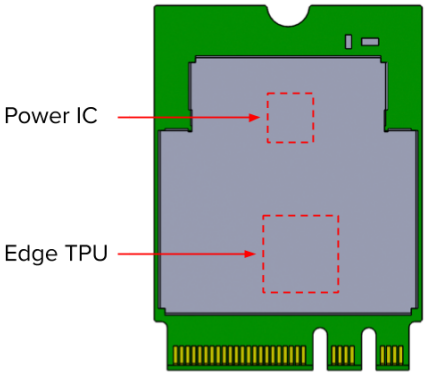


Figure 3. Approximate location of the power IC (PMIC) and Edge TPU (coupled with the shield can using thermal pads)

5.3.1 Thermal limits

The Edge TPU's junction temperature T_j must stay below the maximum operating specification:

- Maximum Edge TPU junction temperature T_j: 115 °C

Warning: Exceeding the maximum temperature can result in permanent damage to the Edge TPU and surrounding components, and can possibly cause fire and serious damage, injury, or death.

For information about how to read the Edge TPU temperature, see Manage the PCIe module temperature.

5.3.2 Top-side cooling options

To ensure successful long-term operation, you might want to add a cooling solution on the top-side of the card module, on top of the shield can. When selecting a thermal solution for the top, consider the following thermal resistance properties with the shield can in place:

- Edge TPU junction-to-shield-can thermal resistance $\theta_{j\text{-}s}$: 11 °C/W

Although many applications can sustain proper thermal levels with the shield can in place, you can achieve higher thermal dissipation (if necessary) by removing the shield can and placing a thermal solution in direct contact with the Edge TPU. If you choose to do so, then consider the junction-to-case thermal resistance and component dimensions indicated in table 6.

Table 6. Thermal properties and dimensions for cooling solutions with the shield can removed

Component	Top-face dimensions (X-Y)	Top-face height from PCB (Z)	Junction-to-case thermal resistance $\theta_{j\text{-}c}$
Edge TPU	5.0 x 5.0 mm	0.55 ± 0.03 mm	2.2 °C/W
Power IC	2.6 x 3.0 mm	0.48 ± 0.03 mm	0.5 °C/W
Shield can frame	N/A	~1.35 mm	N/A
Other	N/A	1.00 ± 0.10 mm	N/A

Notice that other top-side components are taller than the primary heat-producing components, so your heat sink or other enclosure must clear those components. For improved thermal conductivity, consider adding metal stubs that extend from the heat sink to the surface of the Edge TPU, and fill the remaining gap to the Edge TPU with a thermal coupling material.

> **Caution:** If you remove the shield can, it's important that your added heat sink or enclosure has sufficient clearance above the tallest top-side components to prevent the risk of contact and electrical shorting.

If you remove the shield can, be sure to consider the distance between the PCB and heat sink or enclosure. This distance determines the minimum allowable thermal pad thickness, as well as the maximum compressive force that can be exerted on the card. To ensure safe operation, the sustained compressive pressure onto each component from the thermal pads should not exceed 30 PSI (assuming there is an air gap below the card, and thermal pads on the entire top face of the Edge TPU and power IC).

5.3.3 Bottom-side cooling options

A secondary thermal path for cooling the Edge TPU is a thermal epoxy or soft thermal pad on the underside of the card, directly below the Edge TPU. This may dissipate some of the power through the card module and into the base PCB below.

The bottom-side cooling solution is less effective than the top-side solution and should be considered a supplemental thermal path. In order to approximate the effectiveness of a bottom-side thermal path, you should use the junction-to-board thermal resistance $\theta_{j\text{-}b}$ indicated in table 7.

Table 7. Thermal properties for bottom-side cooling solutions

Component	Top-face dimensions (X-Y)	Junction-to-board thermal resistance $\theta_{j\text{-}b}$
Edge TPU	5.0 x 5.0 mm	15 °C/W [1]

[1] In this case, $\theta_{j\text{-}b}$ is the temperature difference between the Edge TPU junction and the surface of the card module when measured from the bottom of the card, directly underneath the Edge TPU.

5.3.4 Temperature warnings and frequency scaling

The Edge TPU includes an internal temperature sensor to help you make power management decisions. You can manually read the temperature, configure parameters that specify when the Edge TPU should shut down, and specify trip-points for dynamic frequency scaling (DFS).

For details, read Manage the PCIe module temperature.

6 Document revisions

Table 8. History of changes to this document

Version	Changes
1.5 (August 2020)	Changed max Edge TPU junction temperature (Tj) to 115 °C (was 125 °C, which is actually used for HTOL and other qualifications). Changed minimum operating temperature to -20 °C (was -40 °C).
1.4 (August 2020)	Added information about power consumption and thermal management. Updated operating temperature and system requirements. Removed description of DFS; added a link to a more detailed app note. Miscellaneous edits. Restructured document to match similar products.
1.3 (April 2020)	Updated system architecture requirements
1.2 (December 2019)	Revised dimensions and added tolerances
1.1 (October 2019)	Added max power consumption
1.0 (August 2019)	Initial release