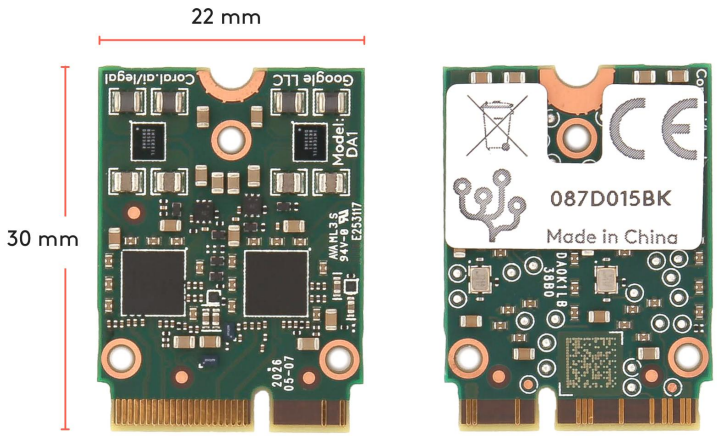
Coral

M.2 Accelerator with Dual Edge TPU datasheet

Version 1.4

Features

- 2x Google Edge TPU ML accelerator
  - 8 TOPS total peak performance (int8)
  - 2 TOPS per watt
- Integrated power management
- 2x PCIe Gen2 x1 interface (one per Edge TPU)
- M.2-2230-D3-E module
- Size: 22.0 x 30.0 x 2.8 mm
- Operating temp: -40 to +85 °C



Description

The Coral M.2 Accelerator with Dual Edge TPU is an M.2 module (E-key) that includes two Edge TPU ML accelerators, each with their own PCIe Gen2 x1 interface.

The Edge TPU is a small ASIC designed by Google that accelerates TensorFlow Lite models in a power efficient manner: each one is 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.

With the two Edge TPUs in this module, you can double the inferences per second (8 TOPS) in several ways, such as by running two models in parallel or pipelining one model across both Edge TPUs.

> **Notice:** Because this module uses two PCIe x1 connections, it is not compatible with all M.2 E-key card slots. The dual Edge TPUs also result in special power requirements that you must carefully review.

Ordering information

Part number	Description
G650-06076-01	Coral M.2 Accelerator with Dual Edge TPU

See https://coral.ai/products/m2-accelerator-dual-edgetpu.

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	22.00 x 30.00 x 2.80 mm
Weight	2.5 g
Host interface	
Hardware interface	M.2 E key (M.2-2230-D3-E)
Serial interface	Two PCIe Gen2 x1
Operating voltage	
DC supply	3.3 V +/- 10 %
Environmental	
Storage temperature	-40 to +85 °C
Operating temperature	-40 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

- 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: 1.00 mm ± 0.10 mm
- Bottom-side component height: 1.00 mm ± 0.10 mm

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

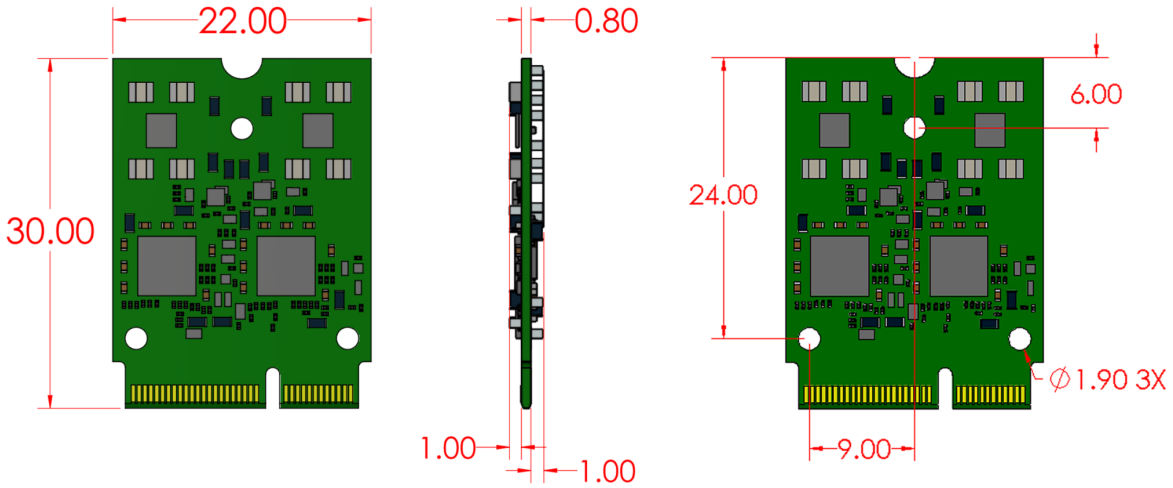


Figure 1. Card module dimensions (in millimeters)

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	-40 °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 each Edge TPU. For some examples of average sustained power consumption from a single Edge TPU, see table 3. However, it's also important that you consider the peak current transients that occur during inferencing.

The maximum current drawn by each 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 by each Edge TPU 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 from a single Edge TPU. 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, and compare the currents when running one Edge TPU or both Edge TPUs in parallel.

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

Table 3. Examples of long-term sustained power during inferencing from **one** Edge TPU

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 both Edge TPUs are running at the maximum operating frequency:

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

4 Connector pinout

Table 4. Card module E-key pinout

Bottom side pins		Top side pins	
Pin	**Signal**	**Signal**	**Pin**
74	3.3V	GND	75
72	3.3V	REFCLKn1	73
70	NC	REFCLKp1	71
68	CLKREQ1# (3.3V)	GND	69
66	PERST1# (3.3V)	PETn1	67
64	NC	PETp1	65
62	ALERT# (3.3V)	GND	63
60	NC	PERp1	61
58	NC	PERn1	59
56	NC	GND	57
54	NC	GND	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	RST_EN	GND	39
36	NC	PERp0	37
34	NC	PERn0	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	NC	NC	15
12	NC	NC	13
10	NC	NC	11
8	NC	NC	9
6	NC	GND	7
4	3.3V	NC	5
2	3.3V	NC	3
		GND	1

Table 5. Product-specific pin descriptions

Pin	Signal	Description
38	RST_EN	Optional module reset. Active high. Has internal 10k pull-down.

5 Application details

5.1 Software requirements

The M.2 Accelerator with Dual Edge TPU 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 on each 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 each Edge TPU is highly variable and depends on the model being executed. Although the average current drawn by a single 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. Of course, if you're running both Edge TPUs simultaneously, you might see even larger combined current spikes. 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 the following:

- Schedule inferencing between the Edge TPUs so they do not draw peak currents simultaneously. In our testing, as little as a millisecond delay between inferences on each Edge TPU is enough to avoid excessive power rail current.
- Underclock one or both of the Edge TPUs to reduce the maximum of all current transients.

5.3 Thermal management

Each 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 two Edge TPUs and two power ICs, indicated in figure 2. During typical operation, approximately 90% of the system power is distributed evenly across the two Edge TPUs, and the remaining 10% dissipates from the two power ICs. Total power dissipation depends on the operating frequency and computational load.

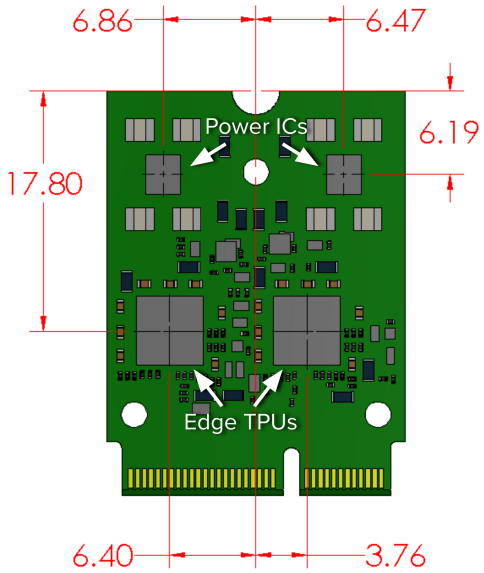


Figure 2. Location of the Edge TPUs and power ICs (PMICs), which are the primary heat sources

5.3.1 Thermal limits

Each 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, we recommended you couple the four heat-producing components to a heat sink or metal enclosure through individual thermal pads. When selecting a thermal pad for the top side of the card module, consider the junction-to-case thermal resistance and component dimensions indicated in table 6.

Table 6. Thermal properties and dimensions for top-side cooling solutions

Component	Top-face dimensions (X-Y)	Top-face height from PCB (Z)	Junction-to-case thermal resistance $\theta_{j\text{-}c}$
Edge TPU (x2)	5.0 x 5.0 mm	0.55 ± 0.03 mm	2.2 °C/W
Power IC (x2)	2.6 x 3.0 mm	0.48 ± 0.03 mm	0.5 °C/W
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: It's important that the heat sink or enclosure has sufficient clearance above the tallest top-side components to prevent the risk of contact and electrical shorting.

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 each Edge TPU and power IC).

5.3.3 Bottom-side cooling options

A secondary thermal path for cooling the Edge TPUs is a thermal epoxy or soft thermal pad on the underside of the card, directly below the Edge TPUs. This can dissipate some of the power through the M.2 card and into the base PCB below.

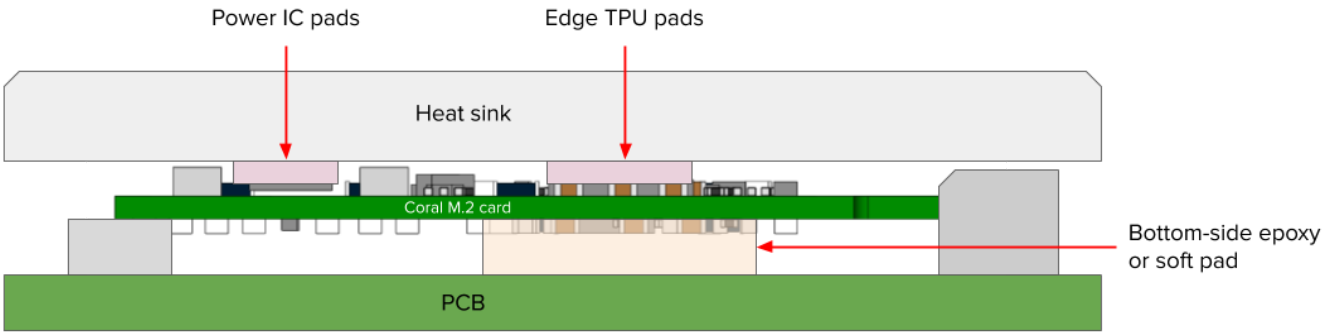


Figure 3. Side view of the card module, connected to host with a top-side heat sink (not included)

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 (x2)	5.0 x 5.0 mm	15 °C/W [1]

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

> **Note:** Card components mounted underneath the Edge TPUs can make cooling through the bottom side of the card more difficult. We recommended using a thermal epoxy or soft thermal pad to fill-in the space around the components and make contact with the bottom side of the M.2 card.

5.3.4 Temperature warnings and frequency scaling

Each 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 each Edge TPU should shut down, and specify trip-points for dynamic frequency scaling (DFS).

For details, read Manage the PCIe module temperature.