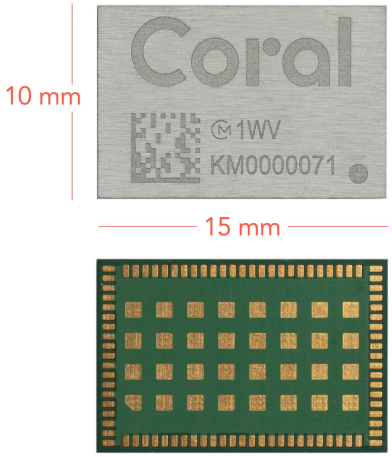
Coral

Accelerator Module datasheet

Version 1.3

Features

- Google Edge TPU ML accelerator
  - 4 TOPS peak performance (int8)
  - 2 TOPS per watt
- Integrated power management
- PCIe Gen2 x1 or USB 2.0 interface
- Surface-mounted (LGA) module
- Size: 15.0 x 10.0 x 1.5 mm
- Weight: 0.67 g
- Operating temp: -40 to +85 °C
- RoHS compliant



Description

The Coral Accelerator Module is a multi-chip module (MCM) designed to perform high-speed inferencing for machine learning (ML) models. It includes the Edge TPU ML accelerator with integrated power control, and it can be connected over a PCIe Gen2 x1 or USB2 interface.

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.

Ordering information

Part number	Description
G313-06329-00	Coral Accelerator Module

See https://coral.ai/products/accelerator-module/.

Table of contents

1 Block diagram

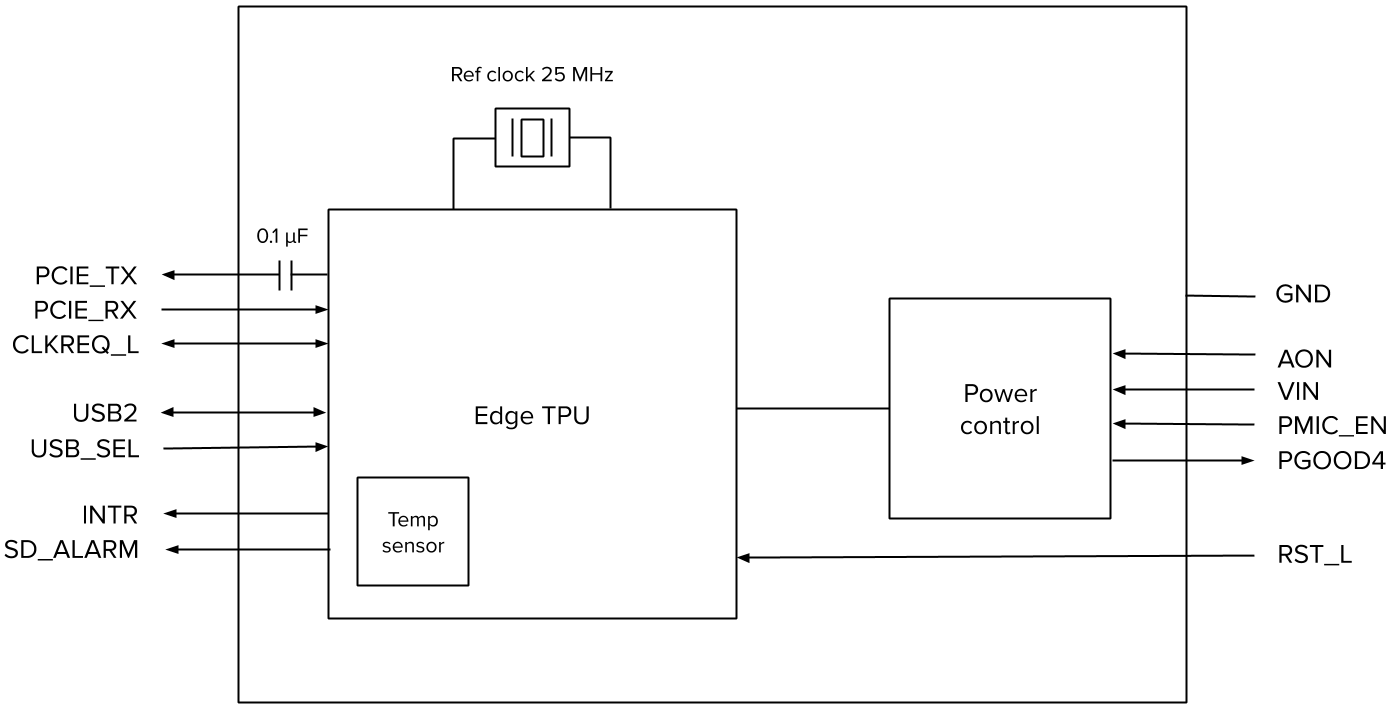


Figure 1. Accelerator Module functional block diagram

2 Electrical characteristics

2.1 Recommended operating conditions

Table 1. Recommended operating conditions

Parameter	Min	Typical	Max
Case operating temperature (T_c) [1]	-20 °C		70 °C
Power supply (VIN)	3.1 V	3.3 V	3.63 V
PMIC digital I/O power supply (AON)	1.7 V	1.8 V	3.63 V

[1] Case temperature is defined as the surface temperature of the module. For details, see 4.5 Thermal management.

2.2 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
Case operating temperature (T_c)	-40 °C	85 °C [1]
Edge TPU junction temperature (T_j)	-40 °C	115 °C
Storage temperature	-40 °C	85 °C
Power supply (VIN)	-0.3 V	6.0 V
PMIC digital I/O power supply (AON)	-0.3 V	6.3 V
PMIC digital I/O [2]	-0.3 V	AON + 0.3 V
Edge TPU digital I/O [3]	-0.3 V	2.1 V

[1] The maximum operating temperature of the case 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.
[2] PMIC digital I/O pins: PGOOD4, PMIC_EN
[3] Edge TPU digital I/O pins: USB_SEL, RST_L, INTR, CLKREQ_L, SD_ALARM

2.3 Logic threshold levels

Table 3. Digital I/O logic thresholds

Parameter	Output		Input	
	Low-level max (V_{OL})	High-level min (V_{OH})	Low-level max (V_{IL})	High-level min (V_{IH})
PMIC digital I/O[1]	0.4 V	AON - 0.4 V	0.5 V	1.35 V

[1] PMIC digital I/O pins: PGOOD4, PMIC_EN

2.4 Power consumption

The power consumed by the 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 4. 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 by the 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.

For more information, see section 4.4 Power delivery network design.

Table 4. 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.

Table 5. Maximum current consumed by the module (for power supply design)

Parameter	**Max**
Power supply current (VIN)	Varies (read above)
PMIC digital I/O power supply current (AON)	10 mA

2.5 Peak performance

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

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

3 Pin layout and description

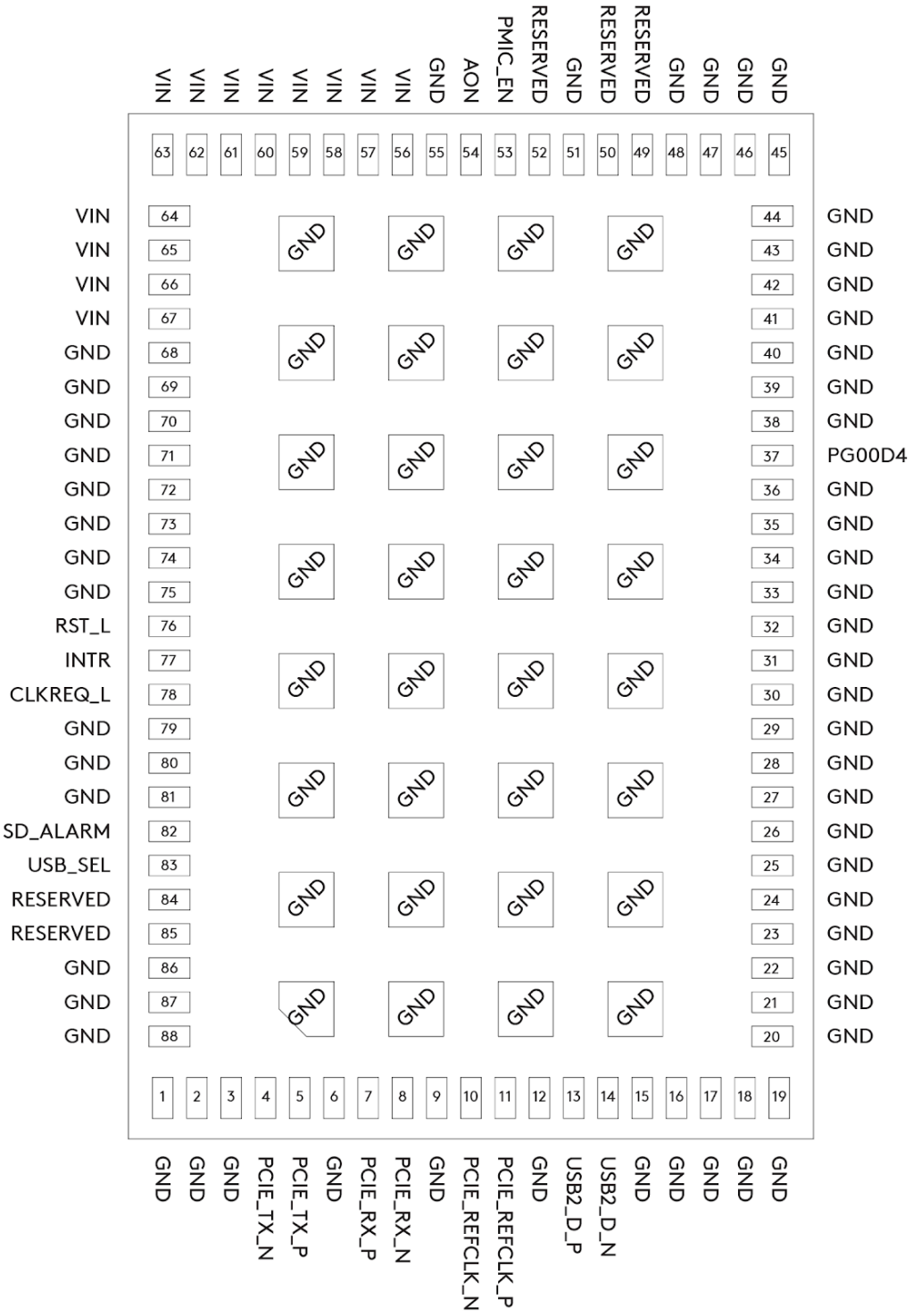


Figure 2. Pin names and numbers (top view)

Table 6. Module pins and descriptions

No.	Pin name	Type	Description	
			PCIe interface	USB2 interface
1 - 3	GND	-	Ground.	
4	PCIE_TX_N	Output	PCIe differential transmit pair. Has internal 0.1 µF DC-blocking cap.	DNC.
5	PCIE_TX_P	Output	PCIe differential transmit pair. Has internal 0.1 µF DC-blocking cap.	DNC.
6	GND	-	Ground.	
7	PCIE_RX_P	Input	PCIe differential receive pair.	Tie low to avoid noise pickup.
8	PCIE_RX_N	Input	PCIe differential receive pair.	Tie low to avoid noise pickup.
9	GND	-	Ground.	
10	PCIE_REFCLK_N	Input	PCIe differential reference clock.	Tie low to avoid noise pickup.
11	PCIE_REFCLK_P	Input	PCIe differential reference clock.	Tie low to avoid noise pickup.
12	GND	-	Ground.	
13	USB2_D_P	I/O	DNC or tie low.	USB2 DP interface.
14	USB2_D_N	I/O	DNC or tie low.	USB2 DM interface.
15 - 36	GND	-	Ground.	
37	PGOOD4	Output	Optional. Power OK signal, active high. Has internal 10K pull-down. See 4.3 Power-on sequence.	
38 - 48	GND	-	Ground.	
49	RESERVED	-	Reserved. DNC.	
50	RESERVED	-	Reserved. DNC.	
51	GND	-	Ground.	
52	RESERVED	-	Reserved. DNC.	
53	PMIC_EN	Input	Power enable input. Must drive high to enable the module.	
54	AON	Power	1.8 V power supply for digital communications (PMIC). Normally connected to 1.8 V but may be tied to VIN.	
55	GND	-	Ground.	
56 - 67	VIN	Power	3.3 V power supply.	
68 - 75	GND	-	Ground.	
76	RST_L	Input	System reset. Active low. Has internal weak pull-down, but you must be sure it's held low during power-up. Size the pull-up strength of the driver accordingly See 4.3 Power-on sequence.	
77	INTR	Output	Optional. This is the first line that interrupts high at a specified Edge TPU junction temperature. Recommended for thermal management. Has internal 100k pull-down. See 4.5 Thermal management.	DNC.
78	CLKREQ_L	I/O	Optional. Low-power mode option for PCIe. This is a bi-directional open drain I/O. It should be implemented as per the PCIe spec, including a proper level translator. On systems that never invoke low power modes, this can be tied low.	DNC or tie low.
79 - 81	GND	-	Ground.	
82	SD_ALARM	Output	Shutdown alarm. This is the second line that interrupts high at a specified Edge TPU junction temperature. This should trigger shutdown. Has internal 100k pull-down. See 4.5 Thermal management.	DNC.
83	USB_SEL	Input	DNC or tie low.	Pull high (1.8 V) to enable USB2 mode. Has internal 4.7k pull-down. Use 0 ohm or tie directly to 1.8V.
84	RESERVED	-	Reserved. DNC.	
85	RESERVED	-	Reserved. DNC.	
86 - 120	GND	-	Ground. Be sure to connect all center pads (89 - 120) to ground for thermal dissipation.	

4 Application details

4.1 Example circuit designs

You can integrate the Accelerator Module into a system design using either the PCIe or USB2 interface with very few supporting components. The following diagrams show typical application circuits with either PCIe or USB2 interfaces.

Note: All ground terminals should be connected, especially the center contacts for thermal dissipation.

4.1.1 PCIe

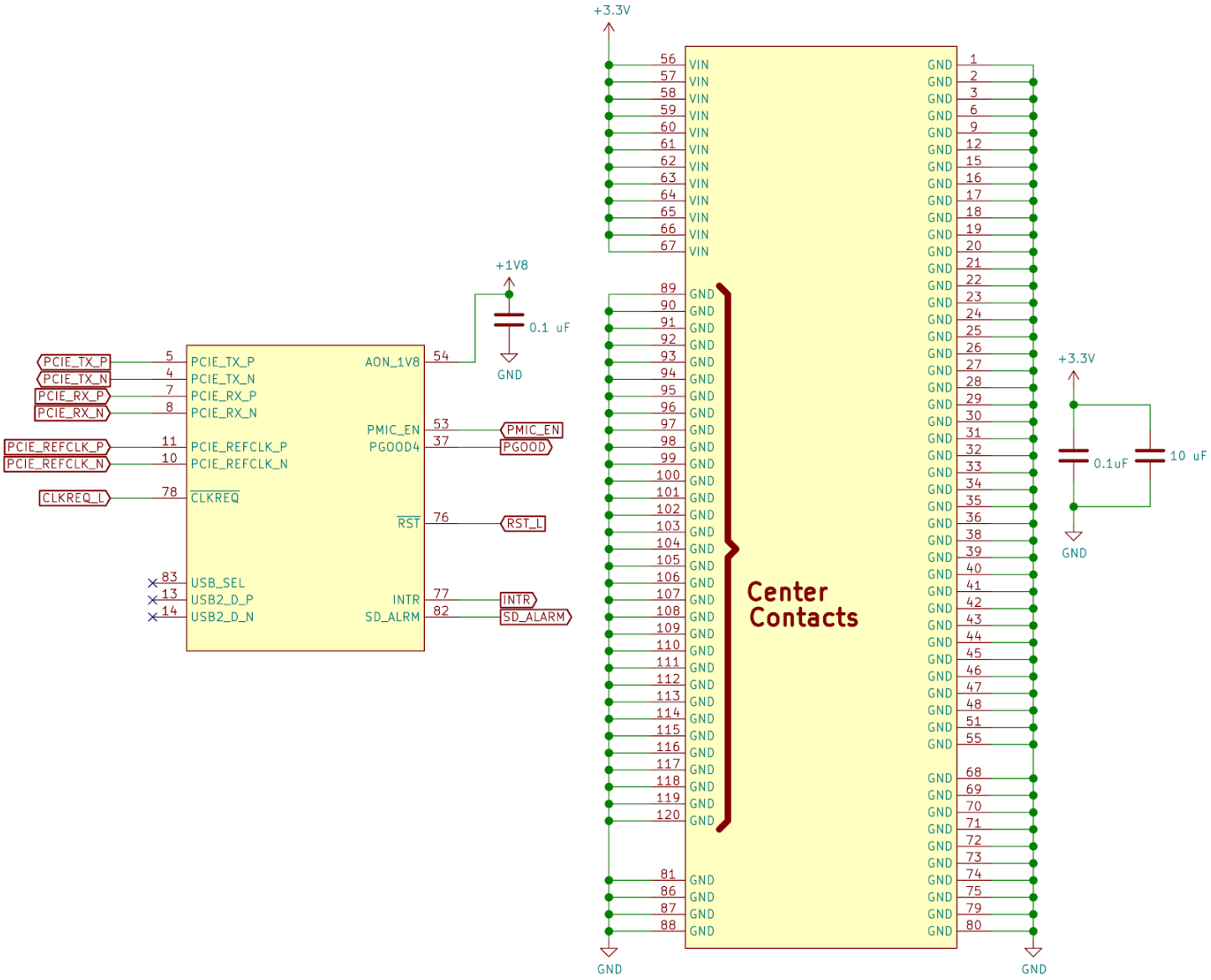


Figure 3. Example PCIe circuit

4.1.2 USB 2.0

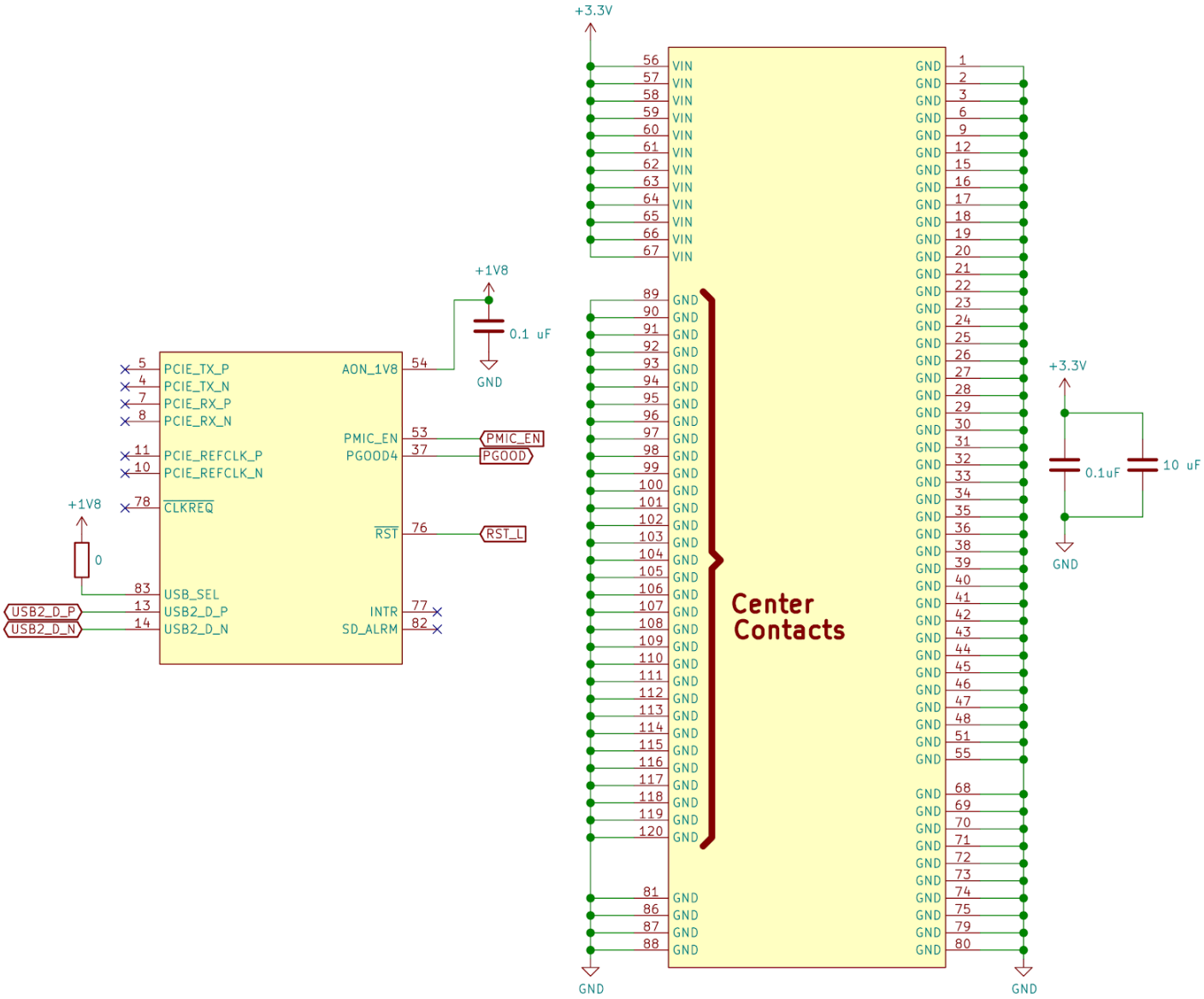


Figure 4. Example USB2 circuit

4.2 Trace length compensation

Table 7 describes the high-speed signals that require each pair to have the same total trace length. Due to space constraints, not all tracings in the module match for each pair, as indicated in the table. You must incorporate any necessary length compensation into your hardware.

Table 7. Pins that must have matching pair lengths, and their internal trace lengths.

Pair	Pin name	Trace length (mils)	Time delay (ps)
PCIe clock	PCIE_REFCLK_P	149.6	26.4
	PCIE_REFCLK_N	158.5	28.0
PCIe TX	PCIE_TX_P	104.0	16.6
	PCIE_TX_N	104.0	16.6
PCIe RX	PCIE_RX_P	125.0	19.9
	PCIE_RX_N	119.0	19.0
USB2 data	USB2_D_P	87.0	13.9
	USB2_D_N	87.0	13.9

4.3 Power-on sequence

4.3.1 PCIe

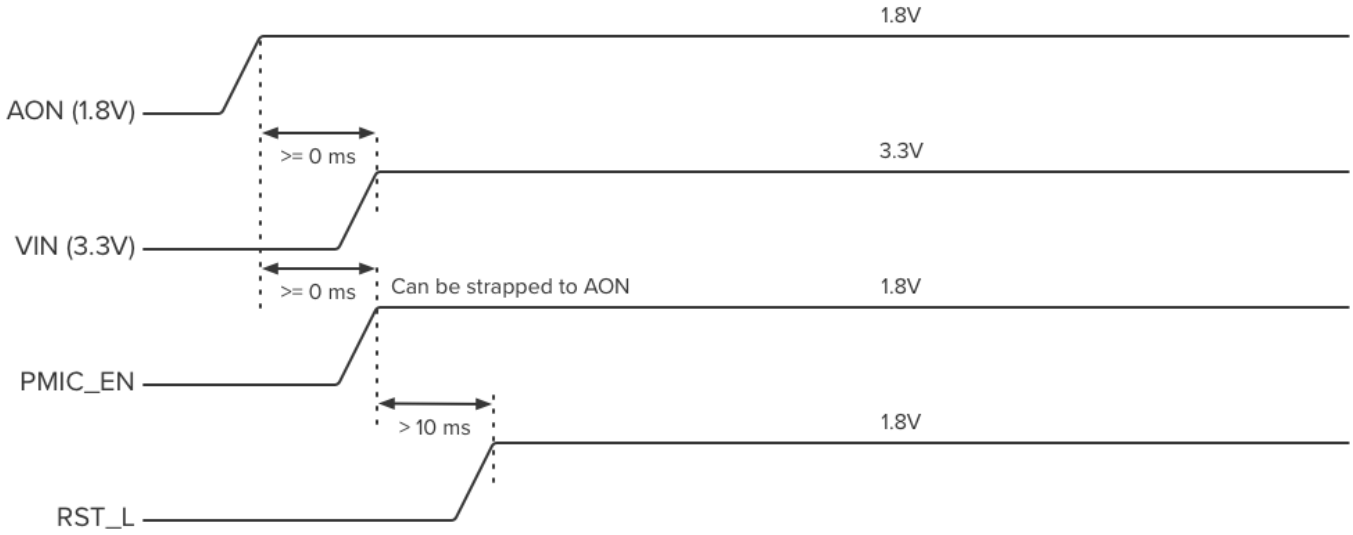


Figure 5. PCIe power-on sequence

4.3.2 USB 2.0

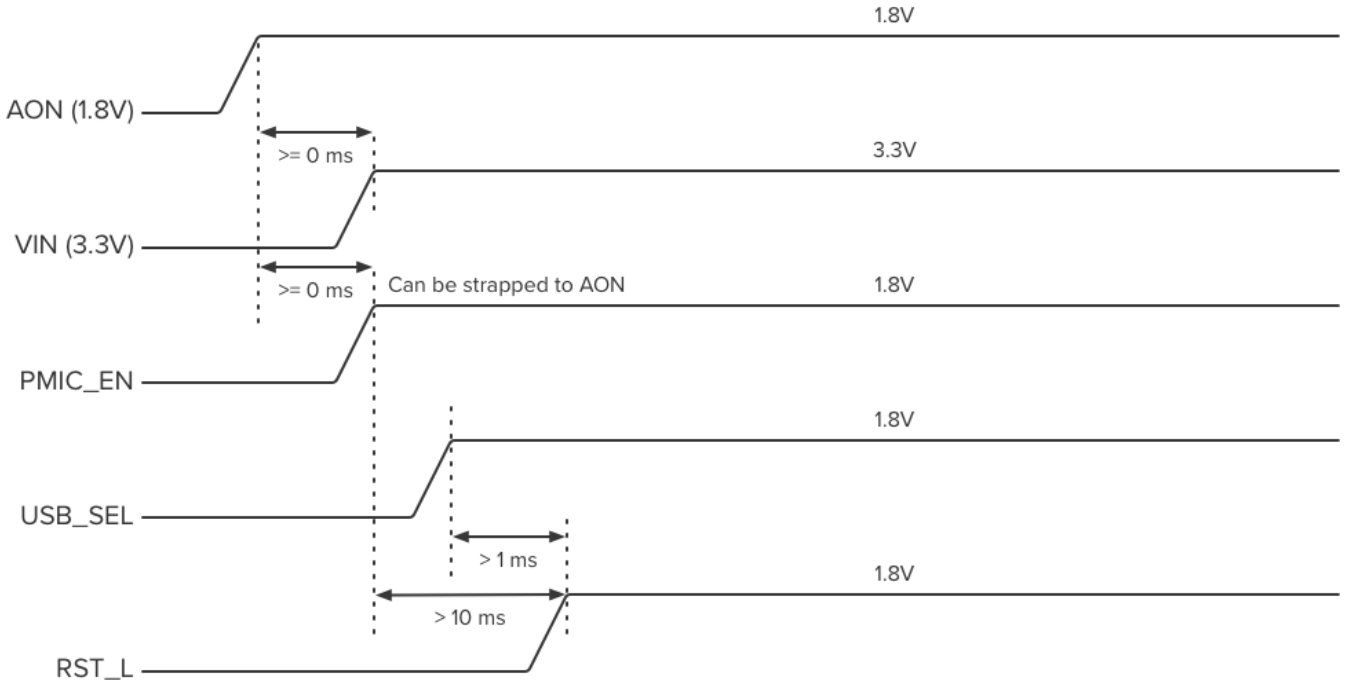


Figure 6. USB2 power-on sequence

Note: The 10 millisecond delay between PMIC_EN and RST_L is a generous estimate. For a more immediate response, monitor PGOOD4 and raise RST_L high 2 milliseconds after PGOOD4 is asserted. Otherwise, the illustrated delay for RST_L should be based on the later of either PMIC_EN or VIN rising.

4.4 Power delivery network design

> **Caution:** If you do not properly design your power delivery network (PDN) to handle peak currents from the Edge TPU, it can easily overwhelm your system and cause brownouts.

As described in section 2.4 Power consumption, the current drawn by the Edge TPU is highly variable and depends on the model being executed, so you must design your power supply based on peak power. Although the average current drawn by the Edge TPU might seem low (less than 500 mA), it can 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. However, these numbers are representative of only the models tested at Google, and your numbers will vary based on your models.

To properly design a PDN for this module, you must consider the current envelopes generated when executing the ML models you'll use in production. The current drawn by the Edge TPU is typically in the form of a few high current peaks, the number of which depends on the model. The burst of high current peaks is usually followed by relatively long periods of inactivity at idle currents. The peaks repeat at regular intervals, depending on the model architecture and number of inferences per second.

The variation of current profile between models makes it very difficult to design a PDN that works for all applications. Ultimately, you must optimize your own PDN based on the models you will use.

In particular, you must fine-tune the loop response in the DC/DC converter so it can absorb the load transients caused by sudden and extreme changes in load current. Likewise, your PDN should maintain a low voltage ripple on the rail and avoid internal overcurrent protection or inductor saturation events. It's important that you validate VIN's PDN performance, such as ripple noise and load step response performance when running your production models.

For more information about how to achieve and test these requirements, refer to the application information from the vendor that provides your DC/DC converter.

4.5 Thermal management

Power dissipation in the Accelerator Module depends on the operating frequency and computational load. As the Edge TPU heats up, performance may be affected, so it's important you design your system to manage thermal variations.

> **Note:** The information in section 2.4 Power consumption includes some sustained power values (table 4) that can help you estimate long-term thermal dissipation. But be sure you perform your own measurements, because total power consumption varies based on the model you're running and other device characteristics.

4.5.1 Thermal limits and resistance

The case temperature T_c and the Edge TPU's junction temperature T_j should stay below the maximum operating specs:

- Maximum case temperature T_c: 85 °C
- 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.

When designing a cooling solution for the module, be sure you consider the thermal behavior of the package when attached to a heatsink. For simulation purposes, you can model the module using these absolute thermal resistance properties:

- Junction-to-case thermal resistance $\theta_{j\text{-}c}$: 10 °C/W
- Junction-to-board thermal resistance $\theta_{j\text{-}b}$: 15 °C/W

These values represent the temperature difference between the Edge TPU's junction and the top/bottom surfaces for a given power flow across the interface, respectively. Figure 7 illustrates these temperature limit locations.

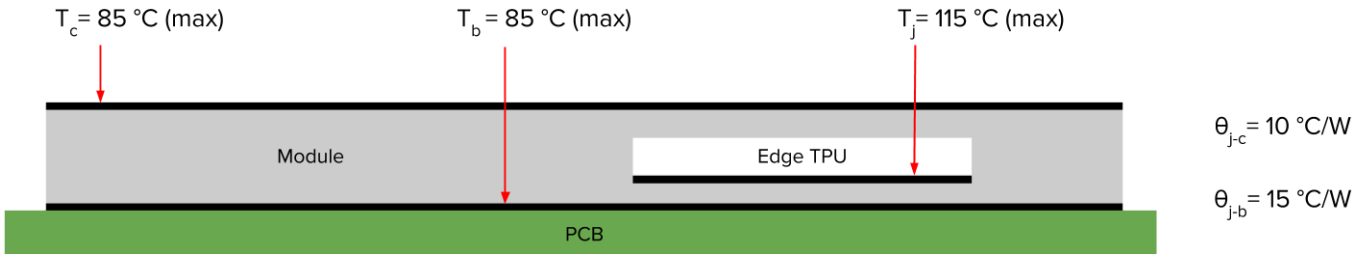


Figure 7. Module cross-section showing Edge TPU junction and module thermal properties

To estimate the effectiveness of your cooling solution—and to calculate a total thermal resistance—you should model $\theta_{j\text{-}c}$ and $\theta_{j\text{-}b}$ thermal impedances in series with the thermal impedances of your interface material and heatsink design.

4.5.2 Temperature warnings and frequency scaling (PCIe only)

The Edge TPU includes an internal temperature sensor to help you make power management decisions. If you're using PCIe, you can manually read the temperature, configure parameters that specify when the INTR and SD_ALARM pins assert based on the current Edge TPU junction temperature, and specify trip-points for dynamic frequency scaling (DFS).

For details, read Manage the PCIe module temperature.

4.5.3 Fixed operating frequency (USB only)

If you connect the Coral Module using the USB interface, then the temperature readings and DFS functionality is not available. Instead, the operating frequency is fixed and you must measure the system temperature yourself.

You can choose to run the Edge TPU at either the "maximum" (500 MHz) or "reduced" (250 MHz) operating frequency when you install the Edge TPU runtime on the host system.

4.6 Software requirements

The Accelerator Module must be operated by the Edge TPU runtime and Coral PCIe driver, which are 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 Package information

5.1 Package and pin dimensions

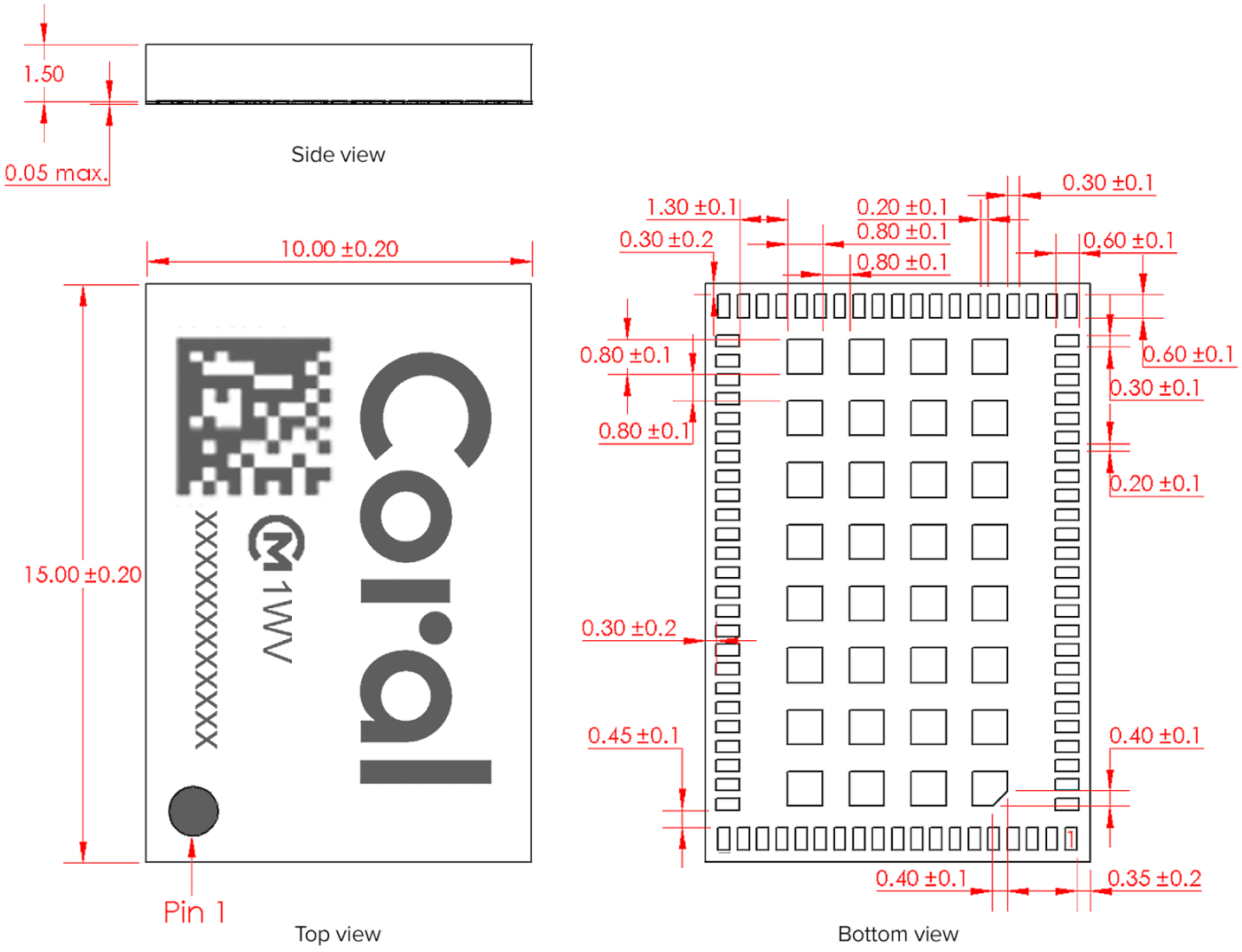


Figure 8. Module dimensions

5.2 Land pattern

To avoid a short circuit due to solder contact with the side shielding, be sure the solder for pads along the module perimeter do not extend to the module outline.

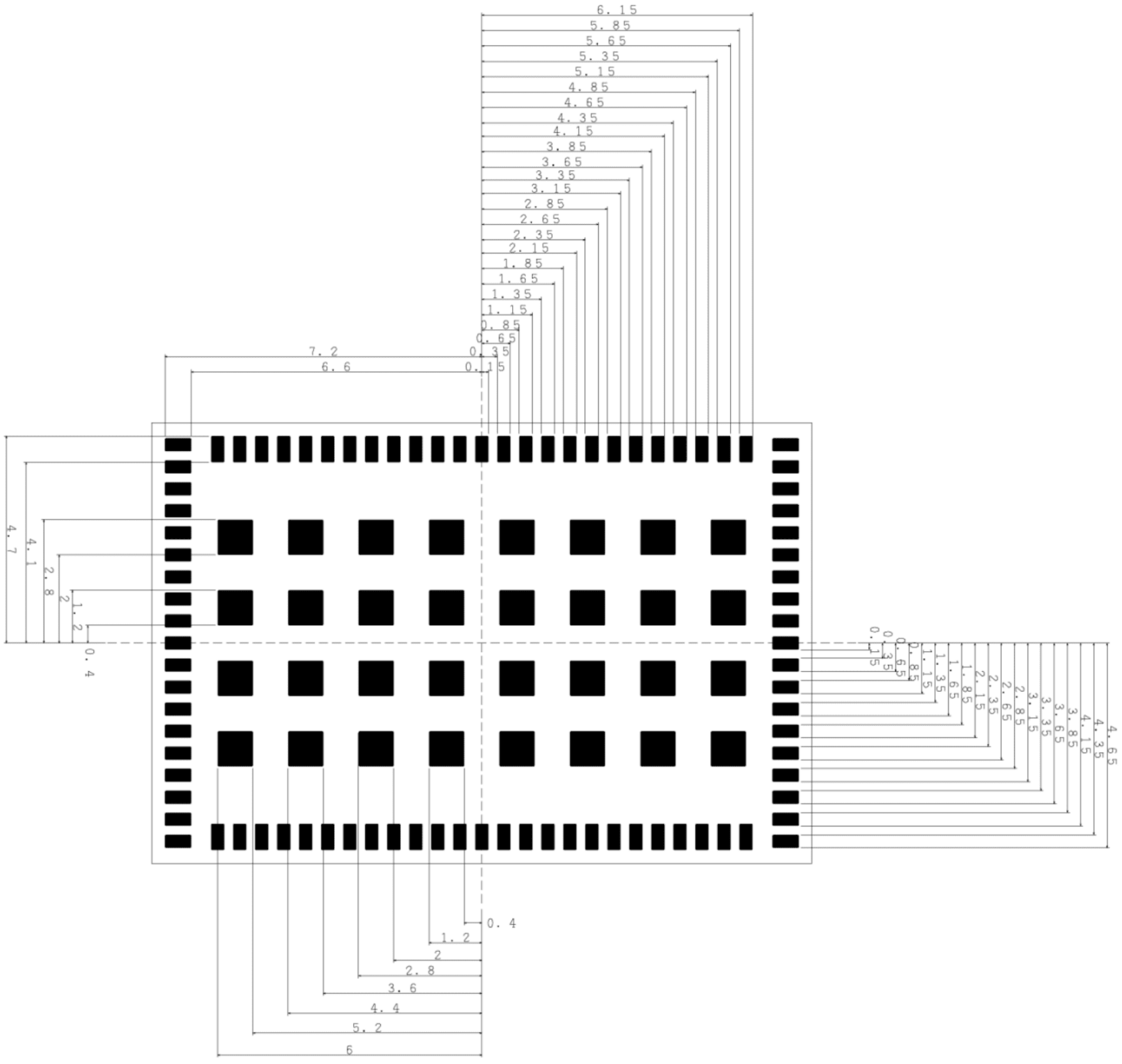


Figure 9. Land pattern dimensions

5.3 Soldering recommendations

- The module requires a no-clean assembly process.
- Set the maximum reflow temperature below 260 °C.
- Do not exceed 2 cycles through reflow.
- Use rosin type flux or weakly active flux with a chlorine content of 0.2 wt % or less.

Caution: Exceeding 260 °C can cause damage to internal components.

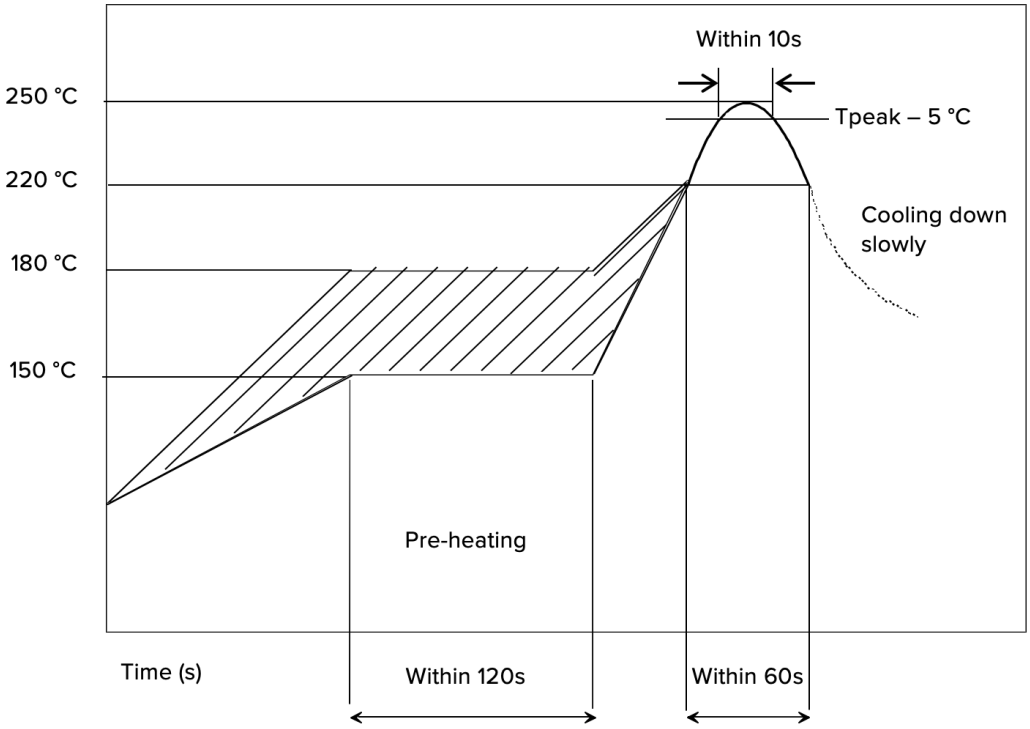


Figure 10. Reflow soldering conditions example

5.4 Tape and reel information

- 1,000 pieces per reel
- Material:
  - Base tape: plastic
  - Reel: plastic
  - Cover tape, cavity tape and reel are made with anti-static processing

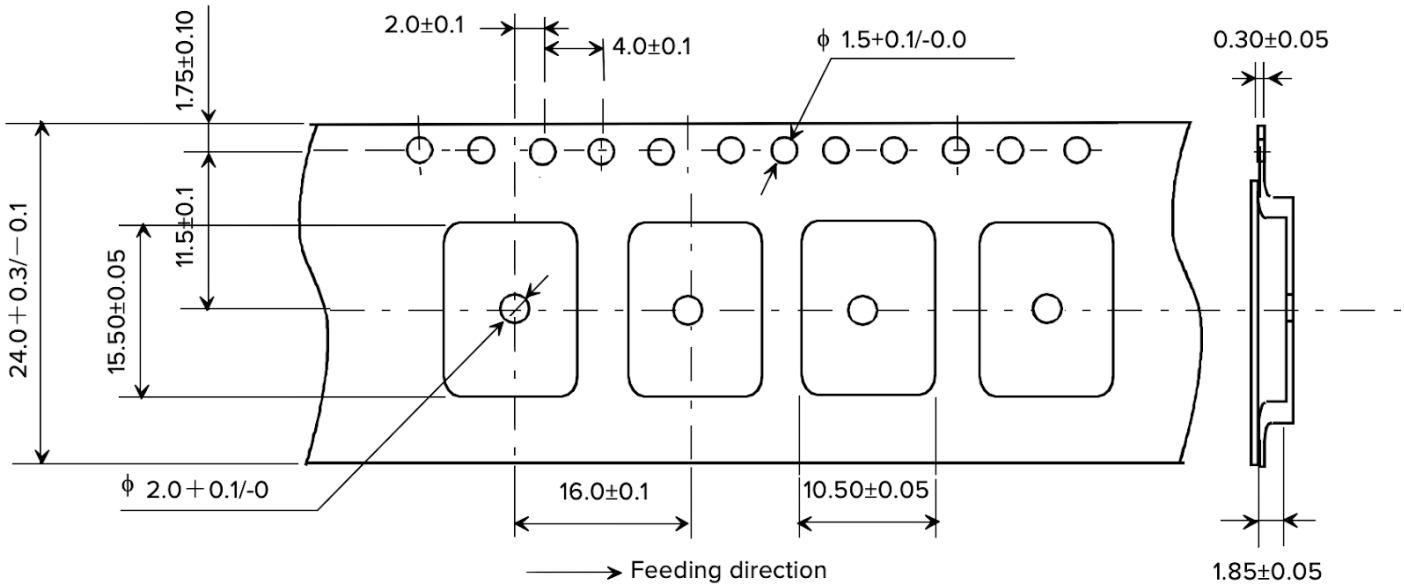


Figure 11. Tape dimensions

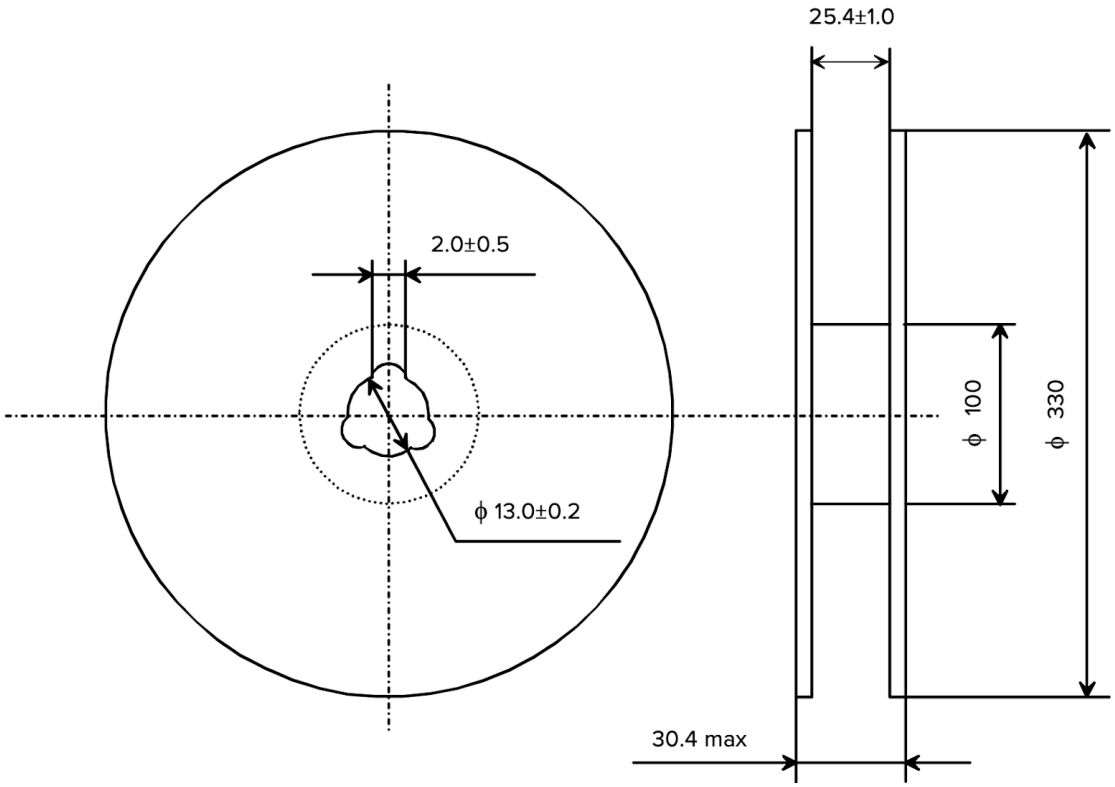


Figure 12. Reel dimensions

The tape is wound clockwise, with feeding holes to the right side as the tape is pulled toward the user.

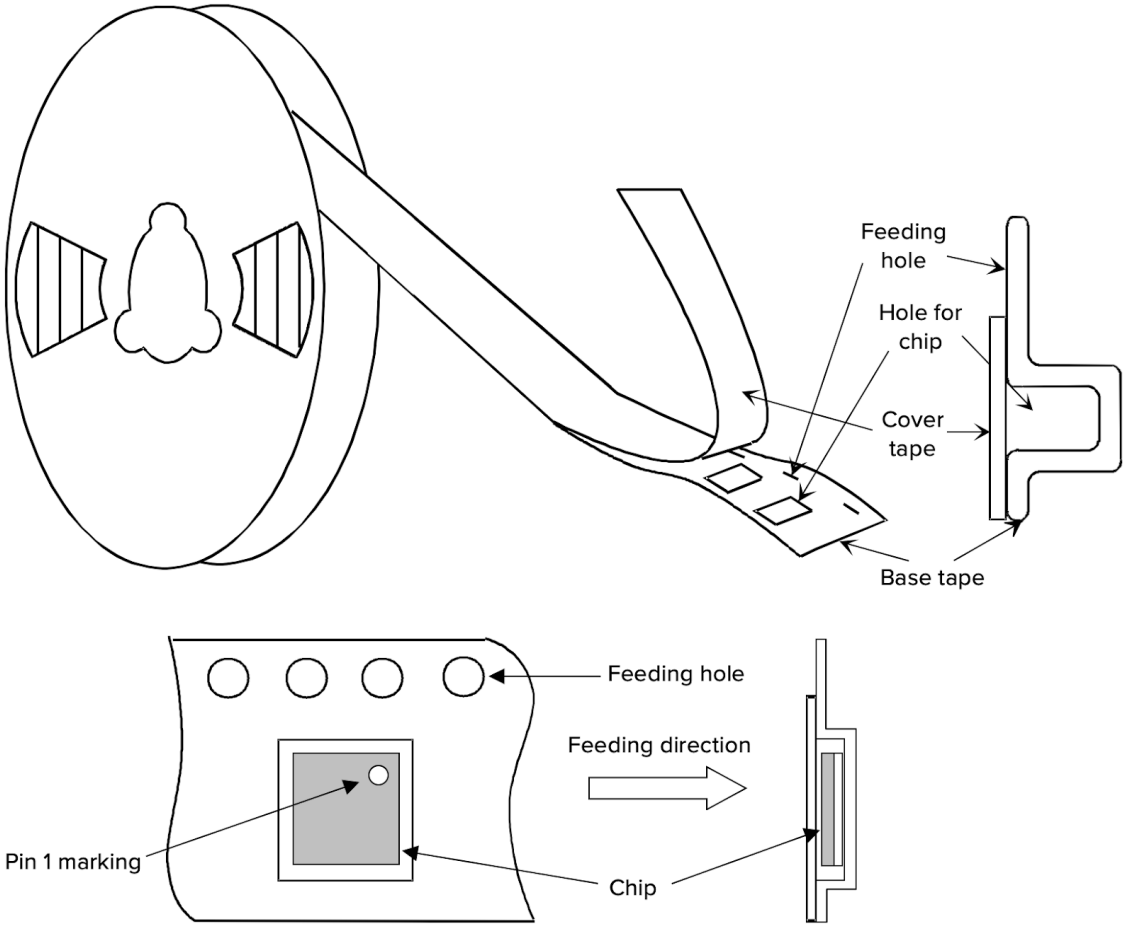


Figure 13. Taping diagram

The cover tape and base tape are not adhered within the "no components" area for 250mm (min).

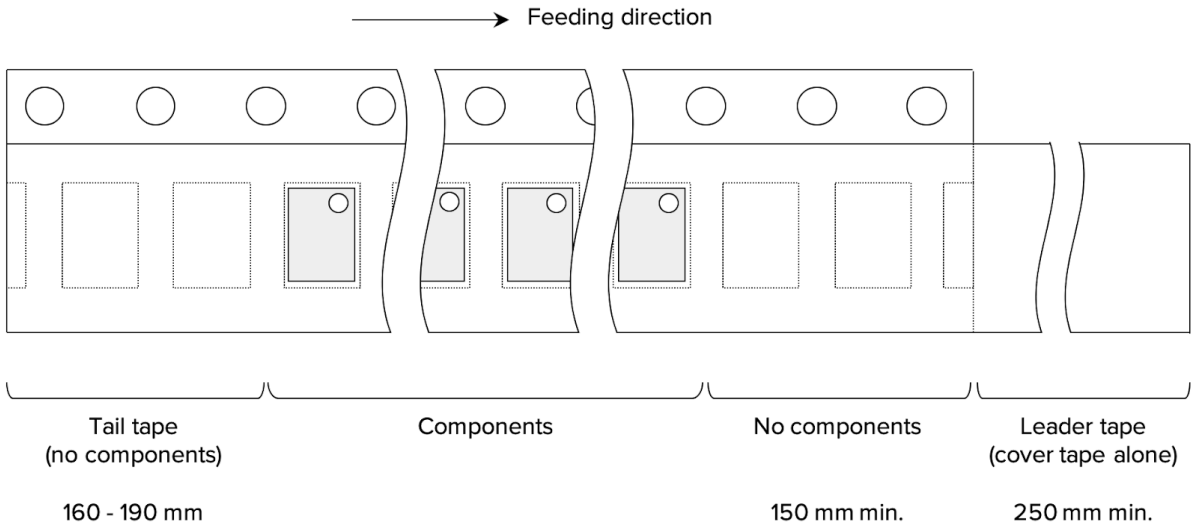


Figure 14. Leader and tail tape diagram

The tear off strength against pulling of cover tape is 5N (min).

The peeling of force is 1.1N (max) in the direction of peeling, as shown in figure 15.

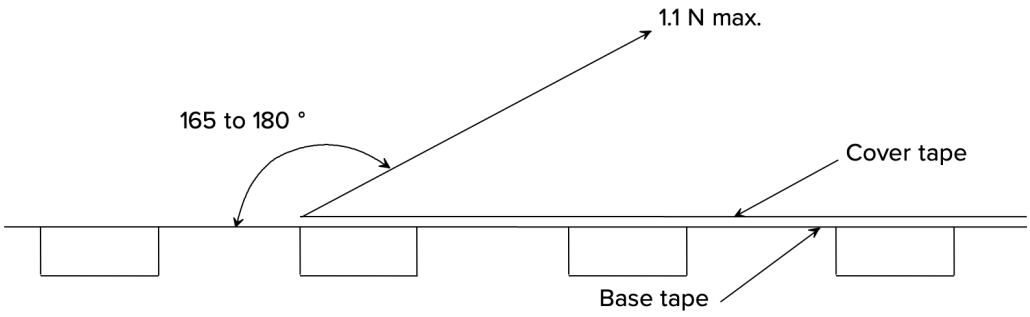


Figure 15. Peeling force diagram

This product is rated to MSL 3. To ensure proper storage conditions, tape and reel must be stored with the provided anti-humidity plastic bag. The bag contains a desiccant and humidity indicator.

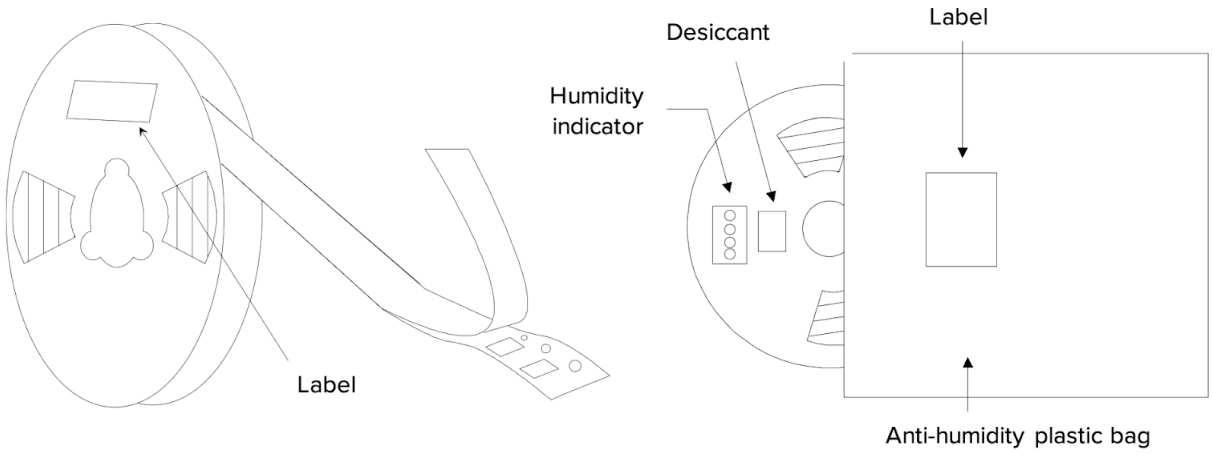


Figure 16. Humidity indicator on plastic package

5.5 Weight

Table 8. Accelerator Module package weight

Component	Weight
Module (one piece)	0.671 g
Taping	144.138 g
Reel	275 g
Total for full reel (approximate)	1.09 kg

5.6 Storage conditions

- Please use this product within 6 months of receipt. If unused for more than 6 months, you must verify the product is ready for soldering.
- While still in the anti-humidity packaging, the product should be stored at an ambient temperature from 5 to 35 °C and humidity from 20 to 70% RH. (Packing materials may deform at temperatures over 40 °C).
- The product must not be stored in a corrosive environment gas (such as Cl2, NH3, SO2, NOx).
- Avoid any mechanical shock, such as dropping the packaging materials.

This product is rated to MSL 3 (JEDEC Standard J-STD-020):

- After the packing is opened, it should be stored at ≤30 °C / ≤60% RH, and used within 168 hours.
- When the color of the humidity indicator on the packing changes, the product should be baked before soldering.

Baking conditions:

- 125+5/-0 °C, 24 hours, 1 time
- Bake on a heat-resistant tray because the materials (base tape, reel tape and cover tape) are not heat-resistant.

6 Document revisions

Table 9. History of changes to this document

Version	Changes
1.3 (September 2020)	Fixed the name and time delay for PCIE_RX_N (table 7).
1.2 (August 2020)	Changed max Edge TPU junction temperature (Tj) to 115 °C (was 125 °C, which is actually used for HTOL and other qualifications).
1.1 (August 2020)	Revised pinout descriptions. Updated example circuit designs. Miscellaneous copy edits.
1.0 (July 2020)	Updated electrical characteristics, power consumption, thermal management, packaging specs, and miscellaneous edits.
DRAFT (May 2020)	Initial release