Chapter 1

# Hardware

T

his chapter describes the hardware used to realize out the project. Where the use of the Raspberry Pi single board computer designed to have high performance both in education and science coupled with 8 mega-pixel camera module CMOS sensor was an almost conditioned choice. Shows that scientific and engineering–grade imagery can be produced with the Raspberry Pi 3 and its V2.1 camera module. It also presents a brief introduction to thermography and its physical principles before introducing Lepton 2.5 thermal camera module, made by the company FLIR which is currently one of the top manufactures of thermal camera solutions. The Coral Dev Board is a single-board computer that contains an Edge TPU coprocessor. It’s ideal for prototyping new projects that demand fast on-device inferencing for machine learning models. The Coral Dev Board is ideal when you need to perform fast machine learning (ML) inferencing in a small form factor.

## 1.1 Raspberry Pi 3

The Raspberry Pi is a high-performance single-board computer designed to experiment and solve real-world problems. This small computer supports a camera module that uses a Sony IMX219 8 mega-pixel CMOS sensor.

### 1.1.1 Specification

As of early 2016, over 8 million Raspberry Pi’s had been sold, making it one of the most popular single-board computers on the market.[1]

The Raspberry Pi credit-card-sized computer supports several accessories, including a camera module containing the Sony IMX219 sensor. This computer and camera configuration is of particular interest since it can provide raw-data format imagery that can

be used for a multitude of applications, including computer vision, biophotonics, medical testing, remote sensing, astronomy, improved image quality, high dynamic range (HDR) imaging, and security monitoring. The Raspberry Pi 3 is the third generation single board Raspberry Pi computer and became available to consumers in February 2016. Some of the more significant Raspberry Pi attributes, including interfaces, are described in Table 1.1. The Raspberry Pi Foundation provides several operating systems for the Raspberry Pi 3, including Raspbian and a Debian-based Linux distribution, as well as third-party Ubuntu, Windows 10 IOT Core, RISC OS, and specialized distributions for download.[2]

Table 1.1: Raspberry Pi 3 Specification.

CPU 1.2 GHz 64-bit ARM Cortex-A53

1 GB of RAM LPDDR2 (900 MHz)

Wireless N and Blue-tooth 4.1 communication

Four USB ports

HDMI interface

Ethernet port

MicroSD card slot

40 GPIO pins

Camera interface

Composite video audio jack

## 1.2 Raspberry Pi, Camera Board V2

The camera is based on the Sony IMX219 silicon CMOS back-lit sensor and produces 8 mega-pixel images that are 3280 *◊* 2464 pixels in size.

The IMX219 sensor operates in the visible spectral range from 400 to 700 nm).[3] Sensor specifications are detailed in Table 1.2.

### 1.2.1 Specification

The V2 camera module operates at a fixed focal length (3.04 mm) and single *f*-number (F2*.*0) typically focused from the near-field to infinity. Images can be captured at ISO settings between 100 and 800 in manually set increments of 100 and camera exposure times between 9µs and 6 s using a rolling shutter. Some of the more significant camera specifications are shown in Table 1.3. In addition to still photos, the Raspberry Pi Sony IMX219 sensor supports a cropped 1080p format at 30 frames per second (fps) and full-frame imaging video at up to 15 fps, but not in raw-data format. The entire camera board is small 25 mm *◊* 25 mm *◊* 9 mm and weighing about 3 g.

Table 1.2: Sony IMX219 sensor chip specifications.

|  |  |
| --- | --- |
| Sensor parameter | Specification |
| Image sensor type | Back-lit CMOS |
| Image size | Diagonal 4.60 mm (type 1/4.0) |
| Number of active pixels | 3280 (H) *◊* 2464 (V) *≥* 8.08 mega-pixels |
| Chip size | 5.095 mm (H) *◊* 4.930 mm(V) (w/ Scribe) |
| Unit cell size (pixel) | 1.12 µm (H) *◊* 1.12 µm(V) |
| Substrate material | Silicon |
| Bit depth | 10-bit A/D converter on chip |
| Data output | CSI2 serial data output (selection of 4 lane/ 2 lane) |
| Communication | 2-wire serial communication circuit on chip |
| Max full-frame frame rate | 30 frames/s |
| Pixel rate | 280 mega-pixel/s (all-pixels mode) |
| Data rate | Max. 755 Mbps/lane (at 4 lane), 912 Mbps / lane(at 2 lane) |

It connects directly to the Raspberry Pi 3 through a 15 pin mobile industry processor interface (MIPI) camera serial interface and is shown alongside a Raspberry Pi 3 in

Figure 1.1.[1, 4]

Table 1.3: Raspberry Pi camera Specification.

|  |  |
| --- | --- |
| Camera parameter | Specification |
| Lens focal length | 3.04 mm |
| *f*-number | 2.0 |
| Instantaneous field of view | 0.368 mrad |
| Full-frame field of view | 59.17 °(H) 58.3 ° (V) |

*◊*



Figure 1.1: Raspberry Pi 3 and camera module V2.1.

## 1.3 Thermal imaging theory

The story of infra-red imaging started in 1800, when Herschel discovered infra-red radiation experimentally at long wavelengths just outside the visible spectrum of sun light. The quantitative explanation of incandescent infra-red radiation in 1900 by Max Planck started a development, which today has resulted in modern infra-red technologies with infra-red camera systems- These are also the result of scientific developments in semiconductor physics and micro-system technologies.[5]

Since its birth, it is possible to recognize three generations of infra-red cameras[6]: the first generation cameras were characterized by a single element detector, combined with two scanning mirrors to create infra-red images. Their main disadvantage was that they su↵ered from saturation problems. Saturation indicates the limit of the highest irradiation that can be measured by a detector. For digital sensors, since incident photoelectrons are converted in charges, each detector can store a maximum amount of charges known as the full well capacity.[5]

The second generation cameras were characterized by an increase in the number of detectors, positioned in a large linear array or in two small 2-D array. The third generation cameras, i.e., the ones currently used, are characterized by large focal plane array (FPA) detectors, thus increasing the reliability and sensitivity of such infra-red systems.[7]

### 1.3.1 Electromagnetic radiation

Electromagnetic radiation is all around (and within, and throughout) us and is comprised of everything from gamma radiation on the high frequency end to radio waves on the low frequency end. So the Figure (1.2) give an overview of EM waves, ordered according to their wave-length or frequency. This spectrum consists of a great variety of di↵erent waves. All of them can be observed in nature, and many have technical applications. Starting from the left of the figure, for example, *“*-rays have the highest frequencies, that is, the shortest wavelengths.[8]

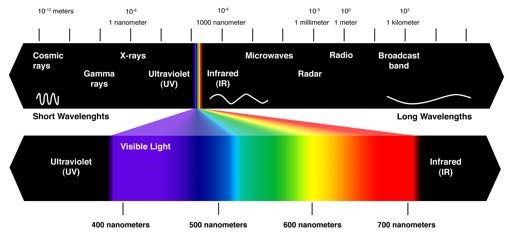


Figure 1.2: Composition of the spectrum of electromagnetic waves. Source: [https://socratic.org](https://socratic.org/questions/what-is-the-wavelength-of-a-photon-of-blue-light-whose-frequency-is-6-3-10-14-s-)

The visible light, defined by the sensitive range of the light receptors in our eyes, only covers a very small range within this spectrum, with wavelengths from 380 to 780 nm. The adjacent spectral region with wavelengths from 780 nm up to 1 mm is usually called infra-red. This range is followed by microwaves, RADAR, and all EM waves that are used for radio, TV, and so on. While most imaging sensors detect radiation in the visible spectrum (wavelengths from 380 to 700nm), long wave infra-red sensors detect radiation the infra-red spectrum, and it accounts for most of the thermal radiation emitted by objects near room temperature. Then for IR imaging, only a small range of the IR spectrum is used. It is shown in an expanded view in Figure (1.3).

Typically, three spectral ranges are defined for thermography: long-wave (LW) region from around 8 to 14 µm, mid-wave (MW) region from around 3 to 5 µm, short-wave (SW) region from 0*.*9 to 1*.*7 µm.

The origin of naturally occurring EM radiation is manifold. The most important phenom for thermography is the thermal radiation. In brief, the thermal radiation implies that every body or object at a temperature T *>* 0 K (*≠*273*.*15*¶*C) emits EM radiation.

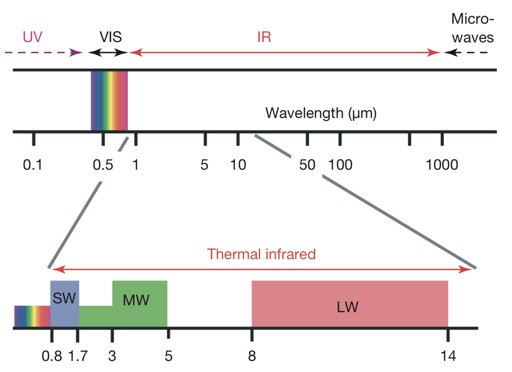


Figure 1.3: Infrared (IR) and adjacent spectral regions and expanded view of so-called thermal IR. This is the region where IR imaging systems for short-wave (SW), mid-wave (MW), and long- wave (LW) cameras exist. Special systems have extended ranges. Source: [8]

The amount of radiation and its distribution as a function of wavelength depend on temperature and material properties.[8] For temperatures in the range of natural and technological processes, this radiation is in the thermal IR spectral region. This is known as the infra-red spectrum, and it accounts for most of the thermal radiation emitted by objects near room temperature.

### 1.3.2 Modern detectors

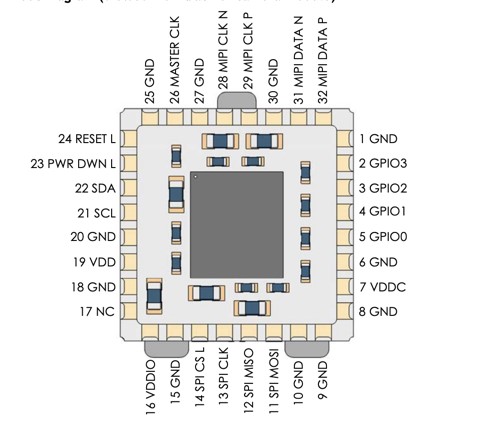
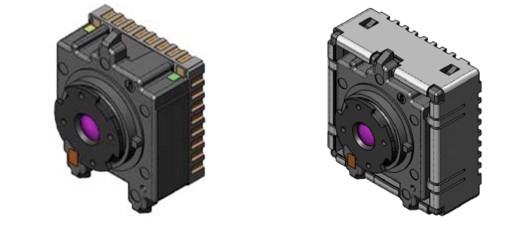
The main di↵erence between a thermal imaging and normal image capturing is the sensor used. One type of the sensor used in thermal cameras is called a microbolometer. Essentially its functionality is similar to the CMOS or CCD sensor, just in di↵erent wavelengths. Unlike many other thermal sensor the microbolometer doesn’t need an active cooling system to function a long period of times. Modern uncooled detectors use sensors whose working mechanism is based on a change of resistance, voltage or current when heated by IR radiation. Uncooled detectors are mostly composed by pyroelectric and ferroelectric materials or based on microbolometer technology. The thermal signal depends upon the radiant power but not upon its spectral content, i.e., it is wavelength independent.[5, 7]

## 1.4 Thermal Camera module Lepton 2.5

The thermal camera module we use in this project its made by FLIR, we will go through its technical specification, modes of capture and communication protocol for both video frame transfer and camera control. Information about the camera are extracted from o cial documentation.

### 1.4.1 Specification

Lepton is an infra-red camera system that integrates a fixed-focus lens assembly, an 80 *◊* 60 long-wave infra-red (LWIR) micro-bolometer sensor array, and signal-processing electronics. Easy to integrate and operate, Lepton is intended for mobile devices as well as any application requiring very small footprint, very low power, and instant-on operation. Lepton can be operated in its default mode or configured into other modes through a command and control interface (CCI). The e↵ective frame rate of the camera is only 8.6 Hz, however for our needs this is not a problem. The camera only requires low voltage supply and has small power consumption of around 140 mW.



(a) *with and without socket*. (b) *Pinout Diagram*.

Figure 1.4: Lepton Camera sketch.

Source: [9]

See table 1.4 for more specifications. For better manipulation with the camera module we use a breakout board (figure 1.5) with a housing for the Lepton camera module. The breakout board provides better physical accessibility, improves heat dissipation and increases input voltage supply range to 3–5 V, as it has its own regulated power supply. This power supply provides the camera module with three necessary voltages: 1.2, 2.8 and 2.5–3.1 V. The breakout board also supplies the camera with master clock signal.[10] The camera uses two interfaces for communication:

* SPI for transferring video frames from the camera to a SPI master device.
* I2C for receiving control commands from the I2C master device.

Table 1.4: Thermal camera specification

|  |  |  |
| --- | --- | --- |
|  | FLIR Lepton 2.5 |  |
| Resolution (h x w) | 80 *◊* 60 | pixels |
| Spectral Range | 8 to 14 | µm |
| Horizontal Field of View | 51 | ° |
| Thermal Sensitivity | *<* 50 | mK |
| Frame Rate | *<* 9 | Hz |
| Control Interface | I2C |  |
| Video Interface | SPI |  |
| Promised Time to Image | *<* 0.5 | s |
| Integral Shutter | yes |  |
| Radiometry | 14-bit pixel value |  |
| Operating Power | 150 | mW |



Figure 1.5: Breakout board.

Source: [https://groupgets.com/manufacturers/getlab/products/flir-lepton-breakoutboard-v1-4](https://groupgets.com/manufacturers/getlab/products/flir-lepton-breakout-board-v1-4)

### 1.4.2 System Architecture

The lens assembly focuses infrared radiation from the scene onto an 80 *x* 60 array of thermal detectors with 17-micron pitch. Each detector element is a vanadium-oxide (VOx) microbolometer whose temperature fluctuates in response to incident flux. The change in temperature causes a proportional change in each microbolometer’s resistance. VOx provides a high temperature coe cient of resistance (TCR) and low 1*/f* noise, resulting in excellent thermal sensitivity and stable uniformity. The microbolometer array is grown monolithically on top of a readout integrated circuit (ROIC) to comprise

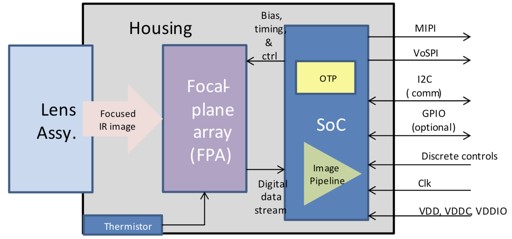


Figure 1.6: Lepton Architecture.

Source: [9]

the complete focal plane array (FPA). Once per frame, the ROIC senses the resistance of each detector by applying a bias voltage and integrating the resulting current for a finite period of time called the integration period. The serial stream from the FPA is received by a system on a chip (SoC) device, which provides signal processing and output formatting.

### 1.4.3 Video Pipeline

The video pipeline includes non-uniformity correction (NUC), defect replacement, spatial and temporal filtering, automatic gain correction (AGC), and colourization.

Non-uniformity

correction(NUC)

Defect

Replacement

Spatial/Temporal

Filtering

AGC

Colorize

rawdatain

AGCdisabled

AGCenabled

AGC/Colorizationenabled

Figure 1.7: Lepton video pipeline block diagram.

Source: [9]

The non-uniformity correction (NUC) block applies correction terms to ensure that the camera produces a uniform output for each pixel when imaging a uniform thermal scene. Factory-calibrated terms are applied to compensate for temperature effects, pixel response variations, and lens-illumination roll-off. To compensate for temporal drift, the NUC block also applies an offset term that can be periodically updated at runtime via a process called flat-field correction (FFC). The FFC process is further described in FFC States, (1.4.5).

The defect-replacement block substitutes for any pixels identified as defective during factory calibration or during runtime. The replacement algorithm assesses the values of neighboring pixels and calculates an optimum replacement value. The typical number of defective pixels is *<* 1.

Temporal Filtering the image pipeline includes a number of sophisticated image filters designed to enhance signal-to-noise ratio (SNR) by eliminating temporal noise and residual non-uniformity. The filtering suite includes a scene-based non-uniformity correction (SBNUC) algorithm which relies on motion within the scene to isolate fixed pattern noise (FPN) from image content.

The AGC algorithm for converting the full-resolution (14-bit) thermal image into a contrast-enhanced image suitable for display is a histogram-based non-linear mapping function. See (1.4.6).

The colorize block takes the contrast-enhanced thermal image as input and generates a 24-bit RGB color output.

### 1.4.4 Power States

Lepton currently provides five power states. As depicted in the state diagram shown in Figure 6, most of the transitions among the power states are the result of explicit action from the host. The automatic transition to and from the overtemp state is an exception. In the figure, transitions that require specific host-side action are shown in bold. Automatic transitions are not bolded.

* Off: When no voltage is applied, Lepton is in the o↵ state. In the off state, no camera functions are available.
* Uninitialized: In the uninitialized state, all voltage forms are applied, but Lepton has not yet been booted and is in an indeterminate state. It is not recommended to leave Lepton in this state as power is not optimized; it should instead be booted to the on-state (and then transitioned back to standby if imaging is not required).
* On: In the on state, all functions and interfaces are fully available.
* Standby: In the standby state, all voltage forms are applied, but power consumption is approximately 4 mW. In the standby state, no functions are available, but it is possible to transition to the on state via the start-up sequence. The shutdown sequence is the recommended transition back to the standby state. It is also possible

to transition between standby and on states via software commands, as further defined in the software IDD.

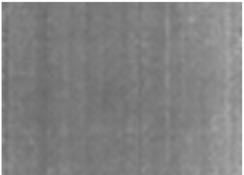
* Overtemp: The overtemp state is automatically entered when the Lepton senses that its temperature has exceeded approximately 80 °C. Upon entering the overtemp state, Lepton enables a “*shutdown imminent*” status bit in the telemetry line and starts a 10 s counter. If the temperature of the Lepton falls below 80 °C before the counter times out, the“*shutdown imminent*”bit is cleared and the system transitions back to the on state. If the counter does time out, Lepton automatically transitions to the standby state.

### 1.4.5 FFC States

Lepton is factory calibrated to produce an output image that is highly uniform, such as shown in Figure 1.8a, when viewing a uniform-temperature scene. However, drift ejects over long periods of time degrade uniformity, resulting in imagery which appears more grainy (Figure 1.8b) and/or blotchy (Figure 1.8c). Operation over a wide temperature range (for example, powering on at -10°C and heating to 65°C) will also have a detrimental effect on image quality. For scenarios in which there is ample scene movement, such as most handheld applications, Lepton is capable of automatically compensating for drift effects using an internal algorithm called scene-based non-uniformity correction (scene-based NUC or SBNUC). However, for use cases in which the scene is essentially stationary, such as fixed-mount applications, scene-based NUC is less effective. In those applications, it is recommended to periodically perform a flat-field correction (FFC). FFC is a process whereby the NUC terms applied by the camera’s signal processing engine are automatically recalibrated to produce the most optimal image quality. The sensor is briefly exposed to a uniform thermal scene, and the camera updates the NUC terms to ensure uniform output. The entire FFC process takes less than a second.

The current FFC state is provided through the telemetry line. There are three FFC states, as illustrated in Figure 1.9:

1. FFC not commanded (default): In this state, Lepton applies by default a set of factory-generated FFC terms.
2. FFC in progress: Lepton enters this state when FFC is commanded. The default FFC duration is nominally 23 frames.
3. FFC complete: Lepton automatically enters this state whenever FFC is completed. Lepton also provides an“FFC desired”flag in the telemetry line. The“FFC desired” flag is asserted at start-up, when a specified period (default = 3 minutes) has elapsed



(a) *Highly uniform image*. (b) *Grainy image (high-spatial* (c) *Blotchy image (low-spatial frequency noise)*. *frequency noise)*.

Figure 1.8: Examples of Good Uniformity, Graininess, and Blotchiness

Source: [9]

since the last FFC, or when the sensor temperature has changed by a specified value (default = 3 °C) since the last FFC. The“FFC desired”flag is intended to indicate to the host to command an FFC at the next possible opportunity.

Leptonpoweredon

FFCNot

Commanded

FFCInProgress

FFCComplete

FFCCommanded

FFCComplete

FFCCommanded

Figure 1.9: FFC States.

### 1.4.6 AGC Modes

There are two AGC modes:

* AGC disabled (default)
* AGC enabled

AGC is a process whereby the large dynamic range of the infrared sensor is collapsed to a range more appropriate for a display system. For Lepton, this is a 14-bit to 8bit conversion. In its most simplistic form, AGC can be a linear mapping from 14-bit to 8-bit; however, a simple linear AGC is generally incapable of providing pleasing imagery in all imaging conditions. For example, when a scene includes both cold and hot regions (for example, a hot object in front of a cold background as illustrated in 1.11, linear AGC can produce an output image in which most pixels are mapped to either full black or full white with very little use of the gray shades (8-bit values) in between. Because of this limitation of linear AGC, a more sophisticated algorithm is preferred. Similar to most AGC algorithms that optimize the use of gray shades, Lepton’s is histogram-based. Essentially a histogram counts the number of pixels in each frame that have a given 14-bit value. Figure 1.10 the concept for a 3x3 pixel area. Classic histogram equalization uses the cumulative histogram as a mapping function

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | 8192 | 8189 | 8191 | | 8193 | 8195 | 8192 | | 8194 | 8192 | 8193 | | 8189  8190  8191  8192  8193  8194  8195  0  1  2  3  4  NumberofOccurence |

Figure 1.10:Source: [9] Illustration of a Histogram for a 3 *◊* 3 Pixel Area.

between 14-bit and 8-bit. The intent is to devote the most gray shades to those portions of the input range occupied by the most pixels. For example, an image consisting of 60% sky devotes 60% of the available gray shades to the sky, leaving only 40% for the remainder of the image. By comparison, linear AGC“wastes”gray shades when there are gaps in the histogram, whereas classic histogram equalization allocates no gray shades to the gaps. This behavior is in principle an e cient use of the available gray shades, but there are a few drawbacks:

* The resulting contrast between an object and a much colder (or hotter) background can be rendered poor by the fact the algorithm “collapses” the separation between such that the object is only one step gray shade above the background. This phenomenon is illustrated in 1.11.
* Too much emphasis can be placed on background clutter, particularly when a mostly isothermal background comprises a large fraction of the total image area. This is also illustrated in 1.11. The Lepton AGC algorithm is a modified version of classic histogram equalization that mitigates these shortcomings. One such modification

is a parameter called “clip limit high”. It clips the maximum population of any single bin, limiting the influence of heavily populated bins on the mapping function. Another parameter utilized by the Lepton algorithm is called “clip limit low”. It adds a constant value to every non-zero bin in the histogram, resulting in additional contrast between portions of the histogram separated by gaps. Figure 1.11 is an example showing the benefit of the Lepton clip parameters.



(a) *Linear AGC*. (b) *Classic Histogram Equaliza-* (c) *Lepton’s Variant of His-*

*tion*. *togramEqualization*.

Figure 1.11: Comparison of Linear AGC and Classic–Lepton

variant of Histogram Equalization Source: [9]

A high value of clip limit high results in a mapping more like classic histogram equalization, whereas a low value results in mapping more like linear AGC. For clip limit low, the opposite is true: a high value results in a mapping more like linear AGC, whereas a low value results in a mapping more like classic histogram equalization. The default values of both parameters produce a good compromise between the two; however, because optimum AGC is highly subjective and often application dependent, customers are encouraged to experiment to find settings most appropriate for the target application. By default, the histogram used to generate Lepton’s 14-bit to 8-bit mapping function is collected from the full array. In some applications, it is desirable to have the AGC algorithm ignore a portion of the scene when collecting the histogram. For example, in some applications it may be beneficial to optimize the display to a region of interest (ROI) in the central portion of the image. When the AGC ROI is set to a subset of the full image, any scene content located outside of the ROI is not included in the histogram and therefore does not affect the mapping function (note: this does not mean the portion outside of the ROI is not displayed or that AGC is not applied there, only that those portions outside the

AGC ROI do not influence the mapping function).[9]

## 1.5 Interface

The Raspberry Pi seen in section (1.1) has bi-directional I/O pins, which you can use to drive LEDs, spin motors, or read button presses. The board offers its GPIO over a standard male header on the board. Over the years the header has expanded from 26 pins to 40 pins while maintaining the original pinout. There are at least two, different numbering schemes you may encounter when referencing pin numbers: 1. Broadcom chip-specific pin numbers

2. P1 physical pin numbers.

Here’s a figure 1.12 showing all 26 pins on the P1 header, including any special function they may have, and their dual numbers.

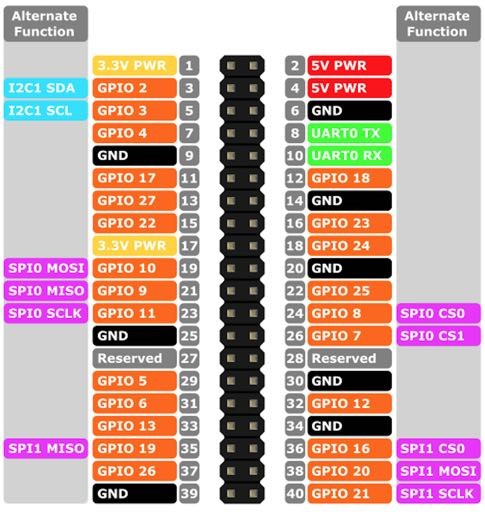


Figure 1.12: Raspberry Pi 3 Pin Mappings. Source: [Microsoft Windows Dev Center](https://docs.microsoft.com/es-es/windows/iot-core/learn-about-hardware/pinmappings/pinmappingsrpi)

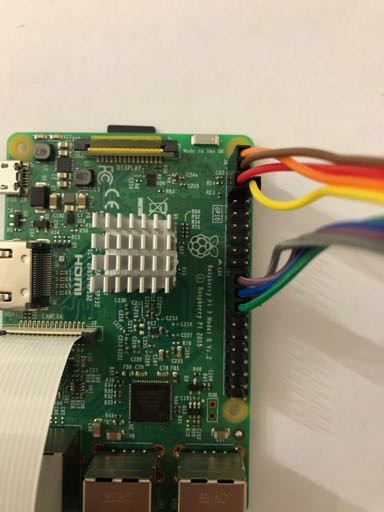
The camera uses two interfaces for communication:

* SPI for transferring video frames from the camera to a SPI[[1]](#footnote-1) master device.
* I[[2]](#footnote-2)C for receiving control commands from the I2C2 master device.

The Raspberry Pi’s CPU has enough processing power to maintain smooth operation without delays, which turned out to be crucial for maintaining synchronization with the camera module when transferring video frames. Below is reported the connection scheme used between the FLIR breakout and the Raspberry Pi ’s GPIO according to the figures (1.13) and the table (1.5).

|  |  |  |  |
| --- | --- | --- | --- |
| Raspberry GPIO | Breakout board | Alternative function | PIN |
| +3.3V | VIN | SPI0 | 1 |
| SDA | SDA | SPI0 | 3 |
| SCL | SCL | SPI0 | 5 |
| GND | GND | SPI0 | 6 |
| MOSI | MOSI | GND | 19 |
| MISO | MISO | 3.3V | 21 |
| SCLK | CLK | I2C1 | 23 |
| CE0 | CS | I2C1 | 24 |

Table 1.5: Schematic connection GPIO



(a) *Breakout board detail*. (b) *Top view*.



(c) *Frontal view*.

Figure 1.13: Realization of the connection between the Raspberry Pi 3 GPIO and Break out board

## 1.6 Coral Dev-Board

Nowadays some manufacturers replace the GPU for a TPU, a Tensor Processing Unit. Driven by the exponential interest for deep learning in different field such as research, industries, robotics. The Coral Dev Board is a single-board computer that contains an Edge TPU coprocessor. It’s ideal for prototyping new projects that demand fast on-device inferencing for machine learning models. Coral Dev Board cannot train a neural network because the TPU is specific designed to work with special pre-compiled model and to be small and energy e efficient.

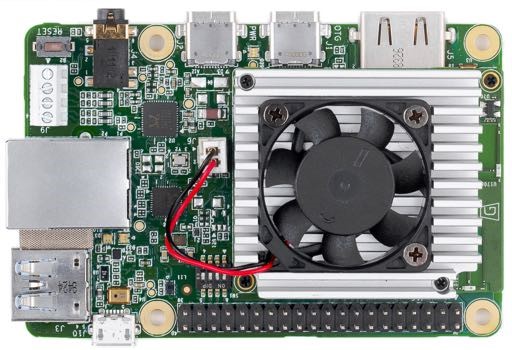


Figure 1.14: Coral Dev Board.

Source: <https://coral.ai/docs/dev-board/get-started/>

### 1.6.1 Overview

The Coral Dev Board is a single-board computer that’s ideal when you need to perform fast machine learning (ML) inferencing in a small form factor. You can use the Dev Board to prototype your embedded system and then scale to production using the onboard Coral System-on-Module (SoM) combined with your custom PCB hardware. The SoM provides a fully-integrated system, including NXP’s iMX 8M system-on-chip (SoC), eMMC memory, LPDDR4 RAM, Wi-Fi, and Bluetooth, but its unique power comes from Google’s Edge TPU coprocessor. The Edge TPU is a small ASIC designed by Google that provides high performance ML inferencing with a low power cost. For example, it can execute state-of-the-art mobile vision models such as MobileNet v2 at almost 400 FPS, in a power efficient manner. The baseboard provides all the peripheral connections you need to prototype a project, including USB 2.0/3.0 ports, DSI display interface, CSI-2 camera interface, Ethernet port, speaker terminals, and a 40-pin I/O header.

Key benefits of the Dev Board:

* High-speed and low-power ML inferencing (4 TOPS @ 2 W)
* A complete Linux system (running Mendel, a Debian derivative)
* Prototyping and evaluation board for the small Coral SoM (40 x 48 mm)

Edge TPU System-on-Module (SoM)

NXP i.MX 8M SoC (Quad-core Arm Cortex-A53, plus Cortex-M4F)

Google Edge TPU ML accelerator coprocessor

Cryptographic coprocessor

Wi-Fi 2x2 MIMO (802.11b/g/n/ac 2.4/5 GHz)

Bluetooth 4.2

8 GB eMMC

1 GB LPDDR4

USB connections

USB Type-C power port (5 V DC)

USB 3.0 Type-C OTG port

USB 3.0 Type-A host port

USB 2.0 Micro-B serial console port

Audio connections

3.5 mm audio jack (CTIA compliant)

Digital PDM microphone (x2)

2.54 mm 4-pin terminal for stereo speakers

Video connections

HDMI 2.0a (full size)

39-pin FFC connector for MIPI DSI display (4-lane)

24-pin FFC connector for MIPI CSI-2 camera (4-lane)

MicroSD card slot

Gigabit Ethernet port

40-pin GPIO expansion header

Supports Mendel Linux (derivative of Debian)

Table 1.6: Coral Dev Board Features.

The Google Coral with a special TPU chip performing all tensor calculations works with special pre-compiled TensorFlow Lite networks. If the topology of the neural network and its required operations can be described in TensorFlow it may work well on the Google Coral.

### 1.6.2 8 bits integers

An alternative to reduce compute time is the use of integers 8-bit instead float 32-bit. The difference in the use of amount of memory allocated is reduced. The advantage derive from the Neural Network ’s accuracy insensibility of number represented, while maintaining the same precision. On the basis of this we can deduce that will save memory, furthermore an cut-off of number of transistor in the chip because is not more necessary take into account floating point. Consequently the this permit to achieve processor powerful, small and energy efficient.

## 1.7 TPU explained in depth

The Google Coral has a TPU on board which speeds up the tensor calculations enormously. These tensor calculations are used in deep learning and neural networks. The Google Coral Dev board use an ASIC made by the Google team called the Edge TPU. It is a much lighter version of the well-known TPUs used in Google’s datacenter. It also consumes very little power, so it is ideal for small embedded systems. Nevertheless, the similarities in applied technology are significant.[11]

### 1.7.1 Neural Node

Neural networks used in deep learning consists of many neural nodes. They are all connected together in a defined way. The way these nodes are wired is called the topology of the network. This topology determines the function the network performs. See this list for a selection of several types of deep learning networks. Each node has always three basic components. A multiplier multiplies all the inputs with their respective so-called weight, the synapses. An adder who accumulates all the individual multiplications. And an activation function that shapes the output given the addition. A schematic view below

(1.15).

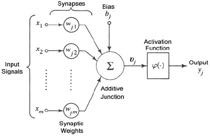


Figure 1.15: Artificial Neuron models and its parts. Source: Adapted from Haykin (1994)

The formula can be written as follows (1.1).

*yi* = *„*ÿ*wij · xi*B (1.1) *n*

A

*i*=1

Deep neural network is consist of millions of neural nodes distribute over many layers; despite the simplicity of the operation, only on multiplication and one addition, require long compute time to obtain a result. keep in mind that for training a network requires many epochs. In other words, the main problem is time. The algorithms is well suited for parallel execution. Although distributing the algorithm over the processes and threads can be an option actually provides disappoint results for the presence of bottlenecks due to architecture general purpose. On the other hand use of Graphical Processing Unit, the GPU on video card designed for efficient manipulation memory. Their highly parallel structure make them more efficient respect CPU especially for algorithms that process large blocks of data in parallel. The best choice is the use of the Tensor Processing Unit, the TPU. This device has been specially designed for the above neural node algorithm.

### 1.7.2 The adder

Generally a strategy adopted when software is tool slow is modify the hardware to reach the better achievements. The structure inside TPU is realized by three main components derived from neural node, then keeping in mind the scheme in figure (1.15): the multiplier, the adder and the activation function must be included in the hardware. Start with analysing the diagram of the 4-bit adder realized in figure (1.16). Where: A and B are the inputs. If the output overflows C4 the carry out is set. C0 is the carry-in of a previous phase.[11] Signals A and B propagate through the circuit and generate the result A + B.

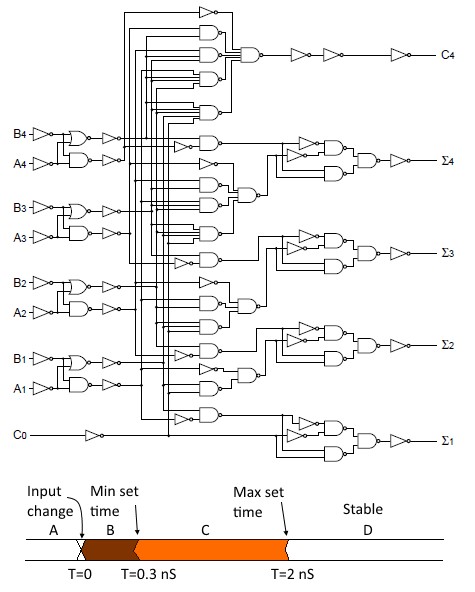


Figure 1.16: 4-bit adder

Source: [Q-engineering](https://qengineering.eu/google-corals-tpu-explained.html)

Changing one of them alters almost immediately the output. This happens extremely very fast, within a few nanoseconds. This propagation time depends on the number of digital ports whose output changes. The propagation time is therefore not fixed, but lies between two limits, a minimum and a maximum time, see diagram at the bottom of the drawing (1.16). By the way, al mentioned times are illustrative and have no relationship to any device.[11]

### 1.7.3 Pipeline

Consider the example where: the propagation time of one adder is 2ns, therefore the maximum clock rate can be 500MHz. Taking into account that a neural node can have many inputs, also hundreds, that must be accumulated this affect the propagation time that increase dramatically. This implies that the last adder in the chain must be attend all intermediate result before its output becomes stable. If we consider a chain of 250 input each of one with 2ns delay, we obtain total time of 500ns waiting.

Consequently clock results extremely slow 2MHz, then solution adopted is structured pipeline where between each adder is inserted a memory element that keep keeps the result stable for the next adder. As show in figure (1.17).

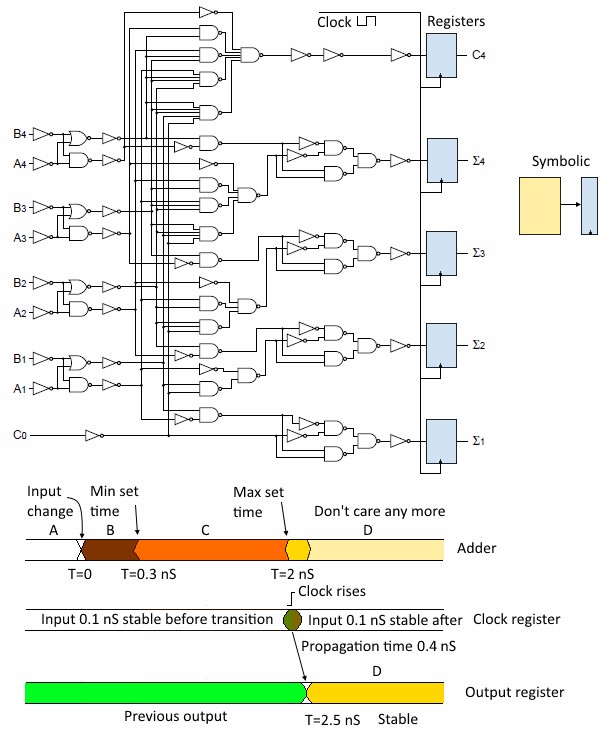


Figure 1.17: 4 bit adder with register

Source: [Q-engineering](https://qengineering.eu/google-corals-tpu-explained.html)

The output of the registers is updated at the rising edge of the clock signal. That is the only time the output can change. When the clock signal is high or low, the inputs cannot manipulate the output, then it remains stable. Just before the clock rises, the input must be stable for a minimum time, also just after the rise time. In contrast to the previous example, now consider a delay of 0*.*1ns and adding the register’s propagation time of (0*.*4ns), the total propagation reach for a new stable signal is now 2*.*5ns, which results increasing up to clock speed of 400MHz. Observable in figure (1.18).

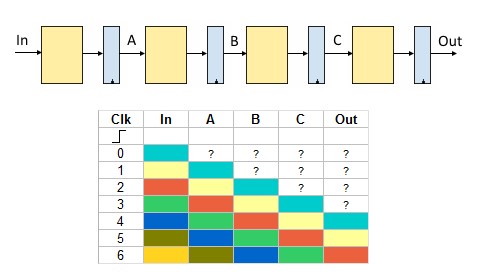


Figure 1.18: Digital pipeline. Source: [Q-engineering](https://qengineering.eu/google-corals-tpu-explained.html)

Taking into account the table in figure (1.18,) each color represents a value. After four clock cycles, the value at the input has propagated through the network and appears at the output. Because the registers are updated simultaneously a new input is accepted every 2*.*5ns. The propagation time has been restored. The time it takes to travel through the whole pipeline is called latency time and is 10ns in the above diagram (4 *x* 2*.*5ns). Every digital component is built on large pipelines which guarantee the required speed.[11]

### 1.7.4 The mul-add cell

The input signal of every neural node has effect on final result. This gives a simple multiplication for an input mediate by a weight value. The operation of multiplication can be implemented with same logic of an adder, then are necessary more gates. Remembering that the neural node has multiplied his value for weight value. Thus the representation assume the scheme in figure (1.19).

Are necessary many important clarifications about the input registers X1 and X2 displayed in figure (1.20). They generate a delay line in the *mul-add* chain, thus can permit to correct synchronize. Considering the second cell *mul-add* who sum the output from

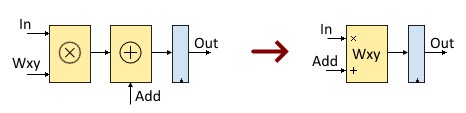


Figure 1.19: Mul-Add register.

Source: [Q-engineering](https://qengineering.eu/google-corals-tpu-explained.html)

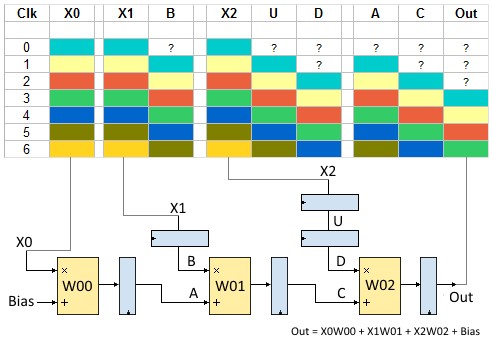


Figure 1.20: Schematic for a three input neuron. Source: [Q-engineering](https://qengineering.eu/google-corals-tpu-explained.html)

previous cell (signal A) retarded by one clock cycle. Consequently also the multiplication between the value in register X1 and weight value must be delayed the same amount of time. The schema reported in figure (1.20) shows for each colour that represent the vector of input [*X*0*,X*1*,X*2] while the input A and B, represented by lines straight, maintains always same colour. This means that they appear simultaneous, i.e. are synchronised.

The same must be valid for input C and D.

### 1.7.5 Systolic array

Starting from structure, just examined, is quite easy extend it to other neurons, taking into account that every single input is connected with all neurons in the next layer, as shows in figure (1.21). The adoption of this solution, called systolic array, permits to have an increment of speed in parallel calculation. Since the propagation time remains unchanged for an spread of systolic array in depth or width, while only the latency time

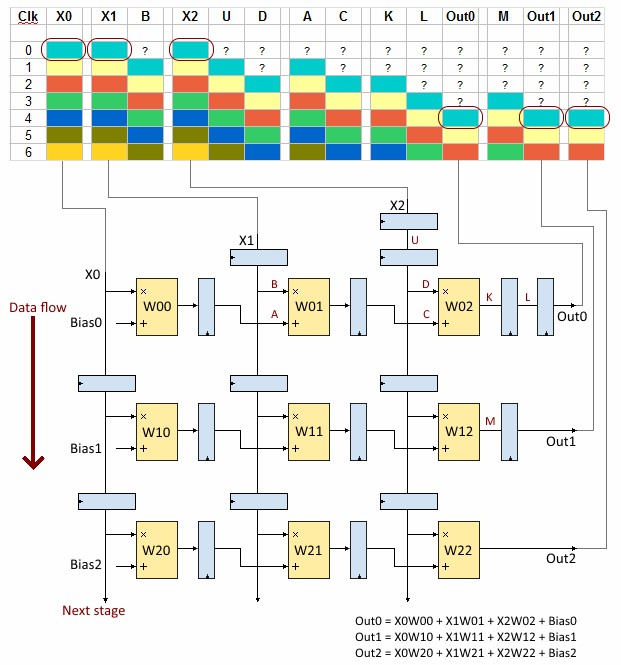


Figure 1.21: Systolic Array Edge TPU. Source: [Q-engineering](https://qengineering.eu/google-corals-tpu-explained.html) grows up this represent the solution widely used for neural network hardware.

### 1.7.6 Activation unit

Once the output is available, it is sent to the activation unit. This module within the Edge TPU applies the activation function to the output. It is hardwired. In other words, you cannot alter the function, it works like a ROM. Probably it is a ReLU function as it is nowadays the most used activation function and it is very easy to implement in hardware.[11]

1. SPI – Serial peripheral interface bus [↑](#footnote-ref-1)
2. I2C (Inter-integrated circuit) [↑](#footnote-ref-2)