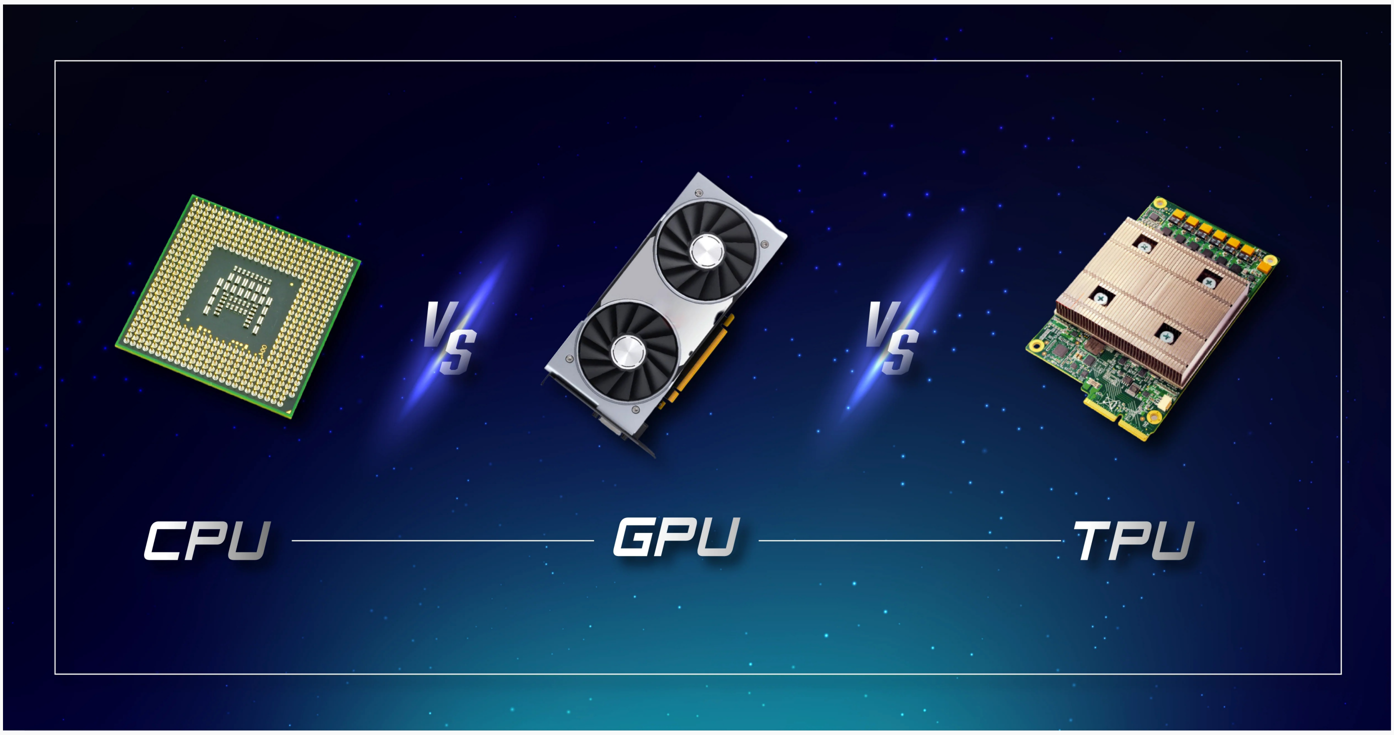
What Is The Difference Between CPU Vs. GPU Vs. TPU? (Complete Overview)

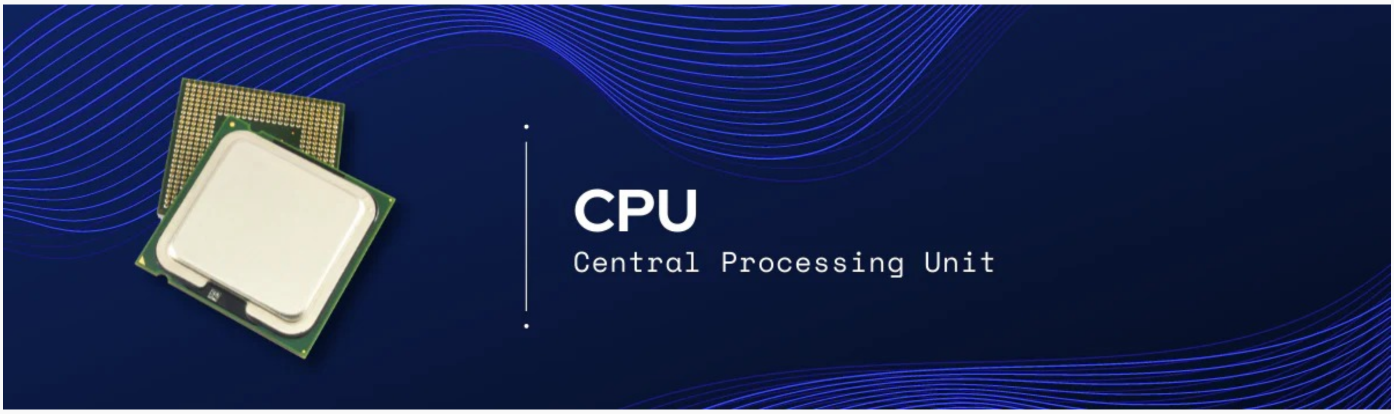
September 09, 2021



Artificial intelligence and machine learning technologies have been accelerating the advancement of intelligent applications. To cope with the increasingly complex applications, semiconductor companies are constantly developing processors and accelerators, including CPU, GPU, and TPU. However, with Moore's law slowing down, CPU performance alone will not be enough to execute demanding workloads efficiently. The problem is, how can companies accelerate the performance of entire systems to support the excessive demands of AI applications? The answer may come via the development of GPUs and TPUs for supplementing CPUs to run deep learning models. That is why it is essential to understand the technologies behind CPU, GPU, and TPU to keep up with the constantly evolving technologies for better performance and efficiency.

Fundamentally, what differentiates between a CPU, GPU, and TPU is that the CPU is the processing unit that works as the brains of a computer designed to be ideal for general-purpose programming. In contrast, GPU is a performance accelerator that enhances computer graphics and AI workloads. While TPUs are Google's custom-developed processors that accelerate machine learning workloads using (specific machine learning framework) TensorFlow. Learn more about performance acceleration in a previous blog post about dedicated computing hardware.

What Is A CPU?



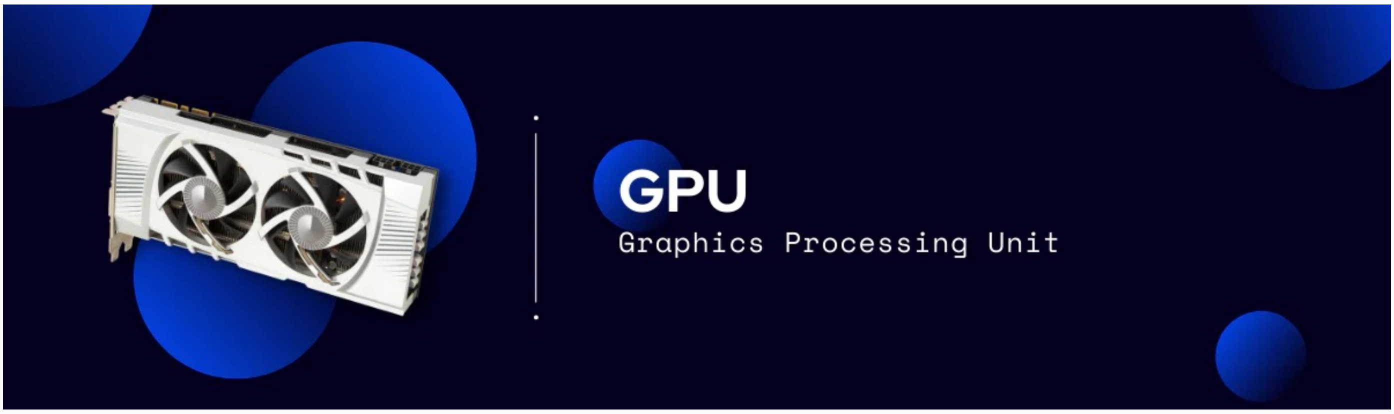
A Central Processing Unit (CPU) is the core processor that exists in all of your smart devices. CPU is a general-purpose processor designed with a few powerful cores and large cache memory that enables it to run a few software threads at once. A CPU is like a conductor in an orchestra; it controls all of the other components from memory to graphics card to perform many processing functions for the system.

A CPU has at least a single processing core but has evolved over time to include more and more cores. Having several cores enables the CPU the ability to perform multithreading, a technology that allows the CPU to perform two lines of execution (threads) at once on a single core. Moreover, modern CPUs now have two to six cores, and some even have eight to 64 cores for enterprise-level CPUs usually reserved for the datacenter.

CPU Features Summary:

- Has Several Cores
- Low Latency
- Specialized in Serial Processing
- Capable of executing a handful of operations at once
- Have the highest FLOPS utilization for RNNs (recurrent neural network)
- Support the largest model thanks to its large memory capacity
- Much more flexible and programmable for irregular computations (e.g., small batches non MatMul computations)

What Is A GPU?



A GPU (graphics processing unit) is a specialized processor that works as a performance accelerator with the CPU. In comparison to a CPU, a GPU has thousands of cores that can break down complex problems into thousands or millions of separate tasks and work them out in parallel. Parallel computing utilizes thousands of GPU cores to optimize various applications, including graphics processing, video rendering, machine learning, and even mining for cryptocurrencies like Bitcoin.

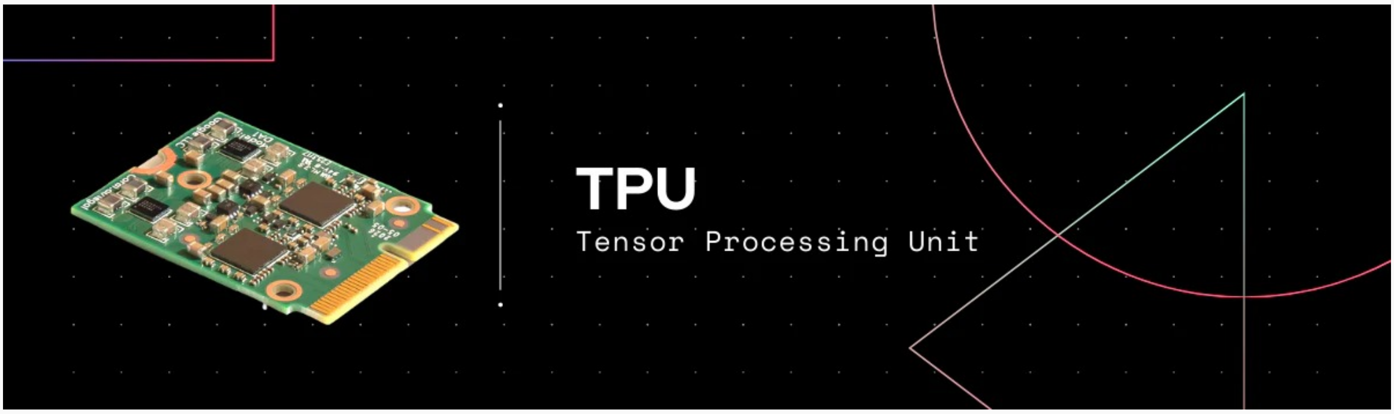
Over the past decade, GPUs have become essential to deep learning development. With the capability to accelerate large matrix operations and perform mixed-precision matrix calculations in a single operation, GPUs can accelerate deep learning at a high-speed rate. This parallel computing technology makes a GPU a crucial part of modern supercomputing that ignited a worldwide AI boom.

GPU Features Summary:

- Has thousands of cores
- High throughput
- Specialized for parallel processing
- Capable of executing thousands of operations at once

Learn More About The Role of GPUs In The Industrial 4.0

What Is A TPU?

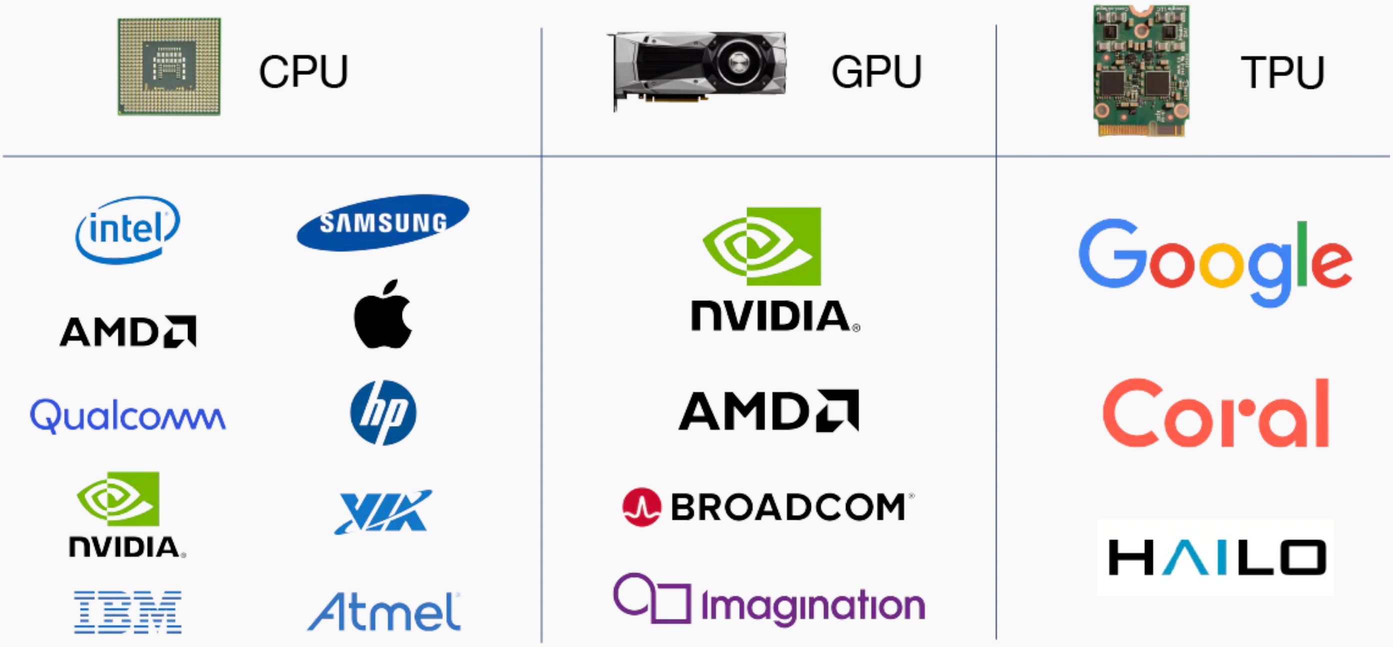


TPUs stand for Tensor Processing Units, which are application-specific integrated circuits (ASICs). TPUs were s designed from the ground up by Google; they started using TPUs in 2015 and made them public in 2018. TPUs are available as a cloud or smaller version of the chip. Cloud TPUs are incredibly fast at performing dense vector and matrix computations to accelerate neural network machine learning on the TensorFlow software. TensorFlow is an open-source machine learning platform built by the Google Brain Team to help developers, researchers, and businesses to run and operate AI models on high-level TensorFlow APIs backed by Cloud TPU hardware. TPUs minimize the time-to-accuracy in training large and complex neural network models. With TPUs, deep learning models that previously took weeks to train on GPUs now only take hours on TPUs.

TPUs Features Summary:

- Special Hardware for Matrix Processing
- High Latency (compared to CPU)
- Very High Throughput
- Compute with Extreme Parallelism
- Highly-optimized for large batches and CNNs (convolutional neural network)

Who Are The Manufactures Of CPUs, GPUs, And TPUs?

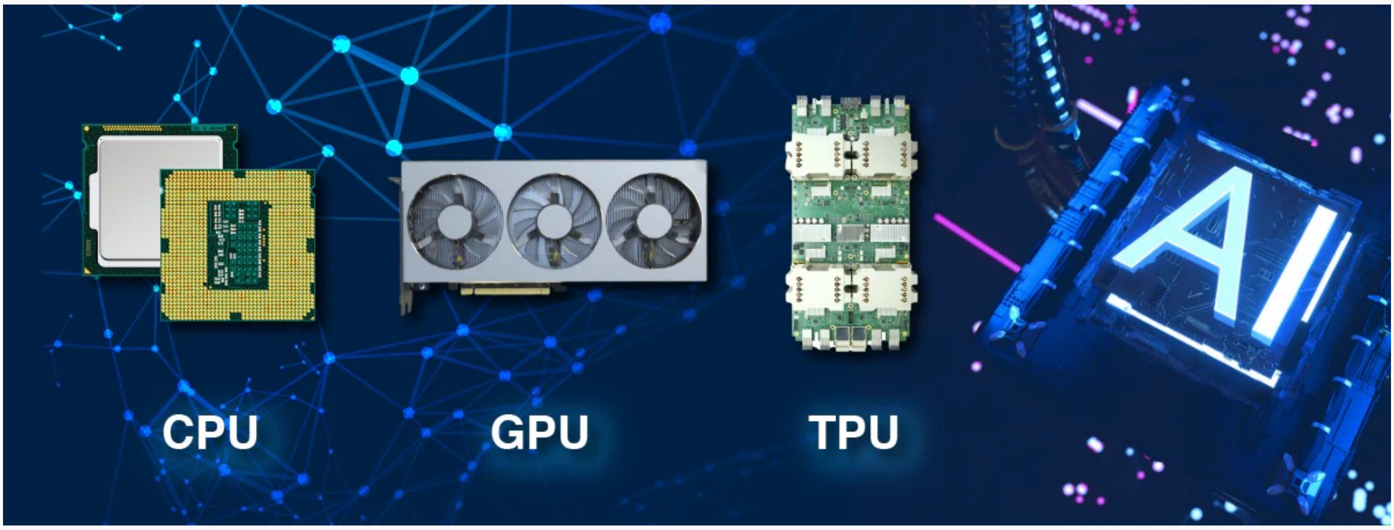


CPU Manufacturers: Intel, AMD, Qualcomm, NVIDIA, IBM, Samsung, Apple, Hewlett-Packard, VIA, Atmel, etc.

GPU Manufacturers: NVIDIA, AMD, Broadcom Limited, Imagination Technologies (PowerVR)

TPU Manufacturer: Google, Coral (owned by Google), HAILO

When To Use CPU, GPU, Or TPU To Run Your Machine Learning Models?



CPUs are general-purpose processors, while GPUs and TPUs are optimized accelerators that accelerate machine learning. It might seem pretty straightforward which one to use to run your machine learning workloads. However, you might want to look closer and consider which type of machine learning models you are running to decide which hardware is best for your workload. Here are some quick guidelines for you to determine which processors are best for your application:

CPUs:

- Prototypes that require the highest flexibility
- Training simple models that do not require a long time
- Training small models with small effective batch sizes
- Mostly written in C++ based on custom TensorFlow operations
- Models with limited I/O or limited system's networking bandwidth

GPUs:

- Models that are too difficult to change or sources that do not exist
- Models with numerous custom TensorFlow operations that a GPU must support
- Models that are not available on Cloud TPU
- Medium or larger size models with bigger effective batch sizes

TPUs:

- Training models using mostly matrix computations
- Training models without custom TensorFlow operations inside the main training loop
- Training Models that require weeks or months to complete

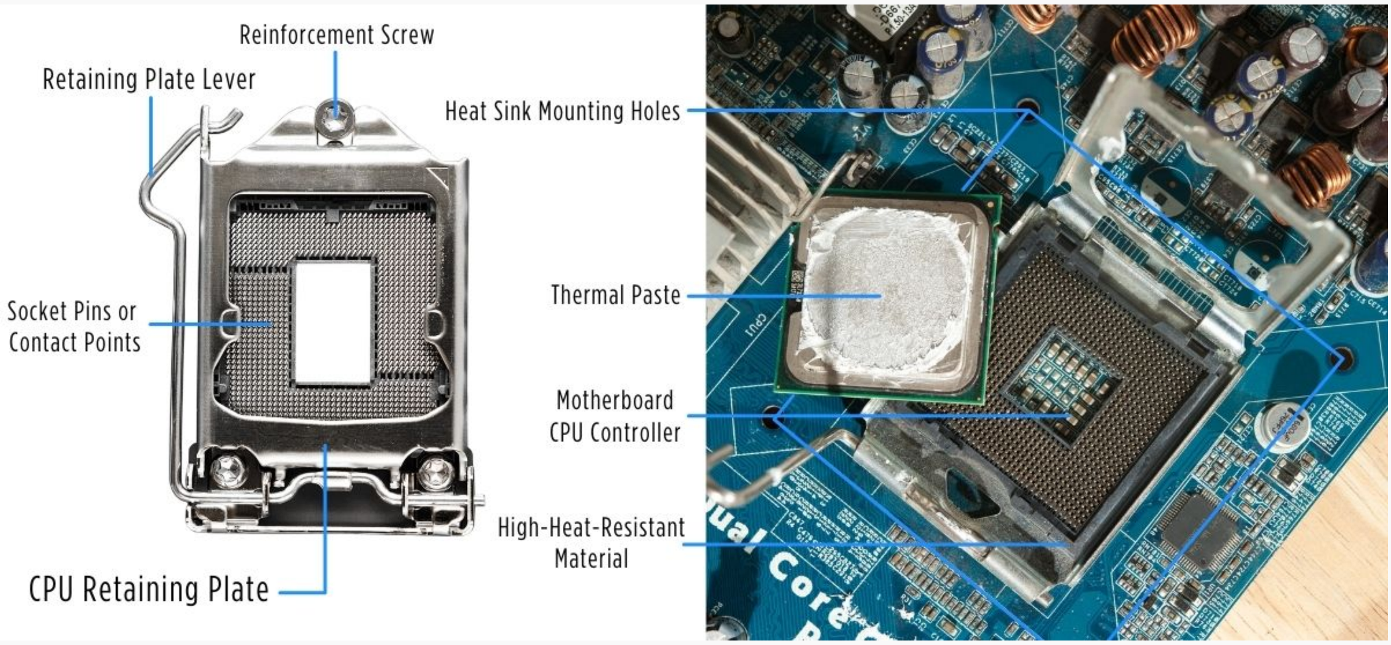
- Training huge models with very large effective batch sizes

Learn More About Deep Learning Embedded Systems

How Do CPUs, GPUs, And TPUs Connect To A Motherboard?

The motherboard is a plastic circuit board (PCB) that contains various computers components such as the CPU, memory, and connectors for other peripherals. The motherboard is where all CPUs, GPUs, and TPUs are connected to communicate with other electronic components of a system.

How Do CPUs Connect To The Motherboard?

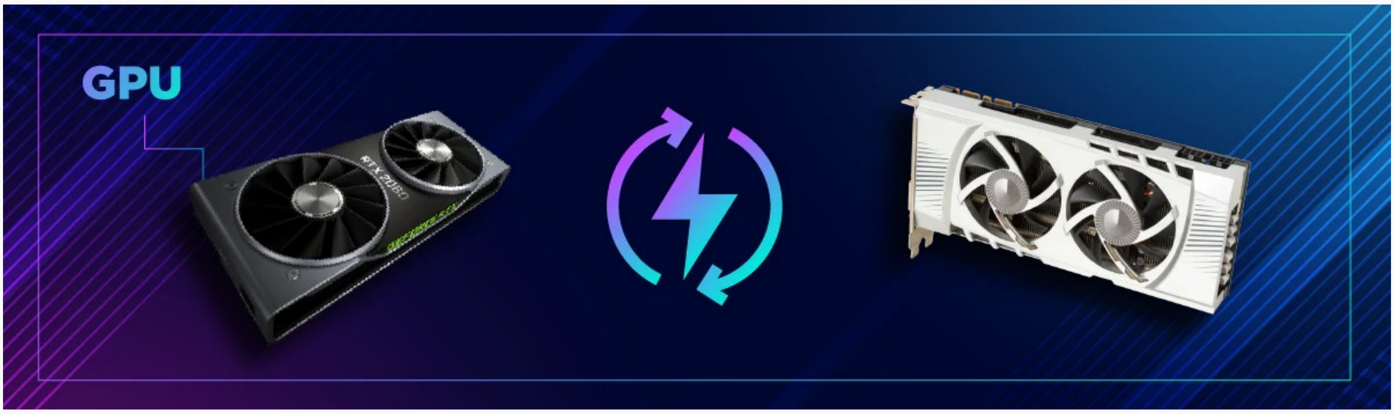


There are two basic types of CPUs, socket CPUs and SoC (system on chip) integrated CPUs. Socket CPUs are installed on the CPU slot at the motherboard. CPU sockets are built with thousands of contact points or metallic pins for power and data transfer between the CPU and the other processors connected on the motherboard. Socket CPUs are typically connected through a ping grid array (PGA) or a land grid array (LGA) CPU slot. On the other hand, SoC is a unique chipset that molded the CPU with other essential peripherals like memory and graphics accelerator into a single silicon chip. SoCs are typically soldered right onto the motherboard with a ball grid array (BGA) connection and provide better power consumption for IoT and mobile applications.

CPU Power Consumption And Its Cooling Solution

CPU power consumption is the value of how much power the CPU needs to run. The power consumption value can be a helpful guide for you to choose the right cooling solution. When choosing a cooling solution, manufacturers often use TDP (thermal design power) to describe how effectively their solution cools down the CPU. TDP value determines how much heat a processor can generate during a heavy workload. CPU's power consumption and TDP are highly related, where higher power consumption leads to higher TDP. The more powerful a CPU is, the more power it consumes, and the more heat it will produce when running a heavy workload. There is a broad range of CPUs with different TDP values available on the market; CPU's TDP typically ranges from 10W to 130W of heat produced. To choose the right cooling solution, your cooling solution and your CPU must have a similar TDP. For example, a CPU with 95W TDP requires a 95W TDP cooling solution. In the industrial sector, fanless rugged computers utilize passive cooling solutions to cool down the CPU. The fanless solution can cool down processors with 10W to 65W of TDP. Anything more than that will require an active cooling solution, usually with a fan.

How Do GPUs Connect To The Motherboard?



GPUs are additional accelerators that are slightly different in the way they connect to a motherboard. GPUs come in two basic types: discrete GPUs and integrated GPUs. A Discrete GPU is an external graphics processor that is apart from the central processing unit. GPU has its own dedicated memory that is separated from the CPU. Discrete GPUs are typically attached to the PCI Express x16 slot on the motherboard. In contrast, integrated GPUs are embedded alongside the CPU on the SoC integrated circuit.

GPU Power Consumption And Its Cooling Solution

Graphics cards also share their power consumption and TDP value measured in watts (W). Typical discrete GPUs consume power around 80W to 250W and produce heat around 100W to 300W based on their TDP rating. Discrete GPUs commonly include built-in cooling solutions that already match their graphics card's TDP. Checking your GPU's power consumption and TDP value helps decide which PSU (power supply unit) to get or if you want to install additional cooling solutions for heavy overclocking applications.

How Do TPUs Connect To The Motherboard?

In early 2019, Google finally released TPU hardware that you can buy from their Coral brand. The current specification for TPU units you can buy can perform 4 trillion operations per second (TOPS), running only on 0.5 watts of power for each TOPS. There are three types of TPU hardware you can get today:

1. The TPU USB accelerators that are featuring the Edge TPU through a USB cable connection.
2. The TPUs that connect through mPCIe or M.2 (A+E and B+M key) connections. M.2 and mPCIe connectors enable the TPUs to be attached directly to the motherboard.
3. The TPU Dev Board option is a single-board computer with a removable system-on-module (SoM) for modular AI applications.

TPU Power Consumption And Its Cooling Solution

Google edge TPU ML accelerator features an 8 TOPS (trillion operations per second) total peak performance with 2 TOPS per watt power consumption. For the cooling solution, you can attach a heat sink or metal enclosure through individual thermal pads on the M.2 TPU to ensure successful long-term operation. Moreover, edge TPU has a high junction temperature Tj, with a maximum junction temperature of Tj: 115℃. Junction temperature is the highest operating temperature of the silicon chip. Edge TPU's junction temperature must stay below the temperature limit for safe operation. Each TPU includes a temperature sensor to monitor the internal temperature and specify trip-point for dynamic frequency scaling (DFS). There is an increasing demand for TPUs for industrial edge applications due to their compact form factor, low power consumption, excellent efficiency, and high temperature-resistant features.

It has been shown that different processing technologies offer various advantages depending on the specific application. Emerging technology is evolving swiftly, making it crucial to stay updated on the latest innovation of computing technologies as the AI and semiconductor industry keeps growing exponentially.

Share f 🐦 in

Related Product

AI Edge Inference Computer

In Vehicle Fanless Computer

Industrial Computer

Machine Vision Computer

Stay Connected

First name*

Last name*

Company name*

Email*

□ I agree that my submitted data is stored and used at Premio Inc to contact me about Premio Inc offerings. [Privacy Policy.](#)*

Subscribe

About Us

- About Us
- Products
- Services
- Blog
- Certifications
- Corporate Social Responsibility
- Terms & Conditions
- Privacy Policy

Services

- Customer Advocate
- Flexible Mfg.
- Lifecycle Mgmt.
- Product Eng.
- On-Demand Logistics
- Quality Assurance
- Customer Service

918 Radecki Court, City of Industry, CA 91748 ☎ 626.839.3100 | 800.977.3646

✉ sales@premioinc.com

CAGE: 0P210 / UEID: Y4N2TTTE7HD5