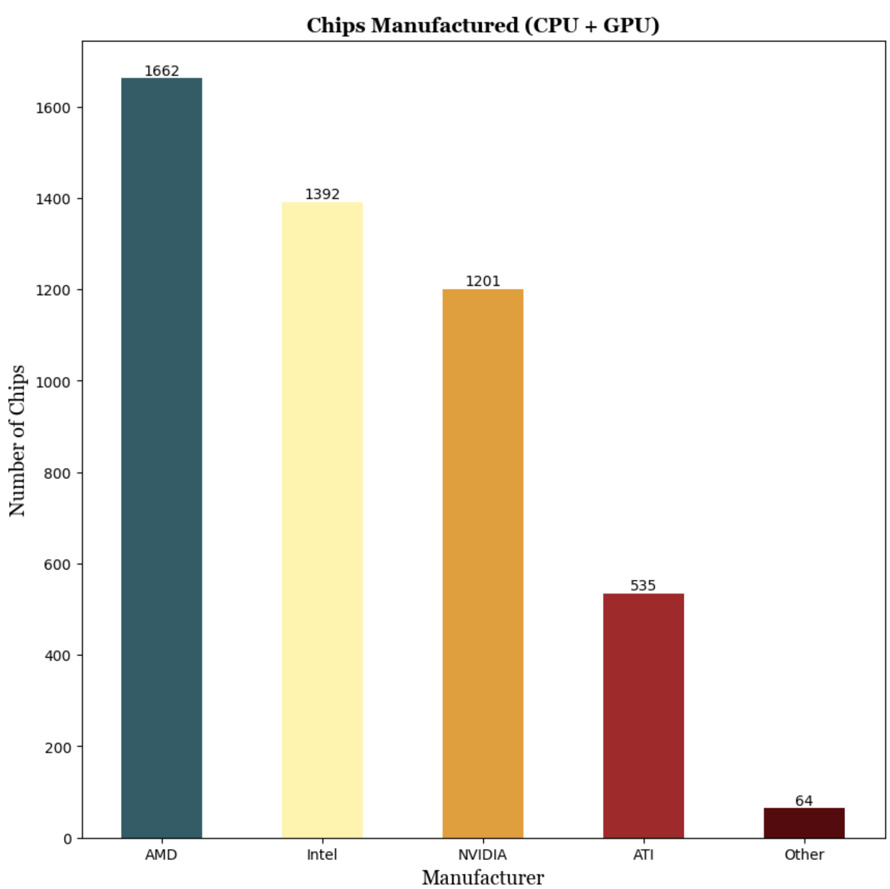
CPU + GPU Exploratory Data Analysis

A Report by: Gabriel Blaise G. Pusta, BSCpE 2A

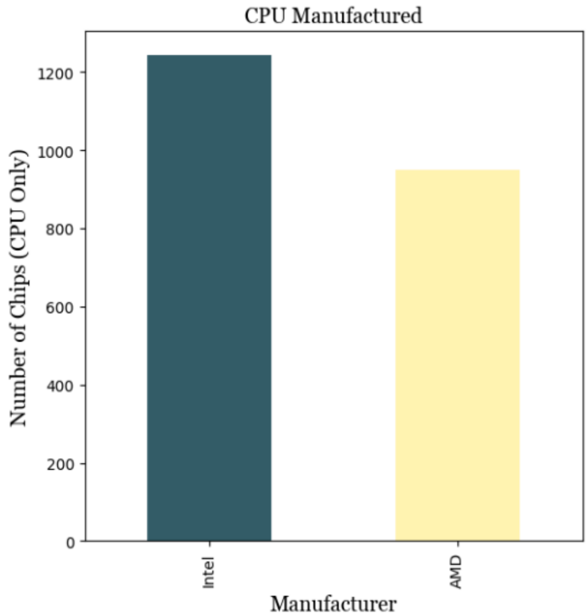
Introduction

This report presents an exploratory data analysis (EDA) of CPUs and GPUs based on various technical specifications. The goal is to uncover meaningful insights into trends over time, particularly in, TDP (Thermal Design Power) process size, die size, transistor count, and frequency.

Overview of CPU and GPU Manufacturers



- AMD has manufactured the highest number of chips (1,662), followed by Intel (1,392), and NVIDIA (1,201



- In terms of CPU production, Intel leads with 1,242 CPUs, compared to AMD's 950.

Observations:

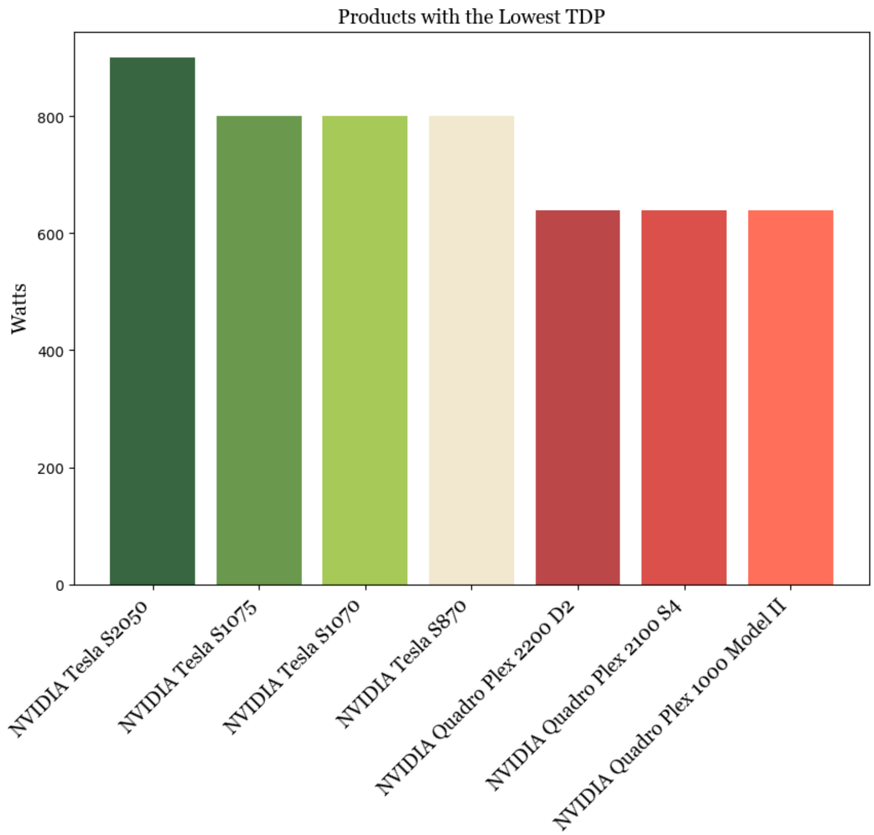
AMD, Intel, and NVIDIA dominate the CPU and GPU market, with AMD producing the highest number of total chips (1,662), followed by Intel (1,392) and NVIDIA (1,201). However, when focusing exclusively on CPUs, **Intel leads the market with 1,242 CPUs**, surpassing AMD's **950 CPUs**, reinforcing Intel's historical dominance in CPU manufacturing.

Overall, this data reflects **Intel's leadership in CPUs, AMD's balanced presence in both markets, and NVIDIA's stronghold in GPUs**, demonstrating the competitive dynamics among leading semiconductor manufacturers.

TDP (Thermal Design Power) Trends

data shows a significant range between high-performance **NVIDIA Tesla and Quadro Plex GPUs** and **low-power Intel Atom CPUs**:

Observations:



Observations:

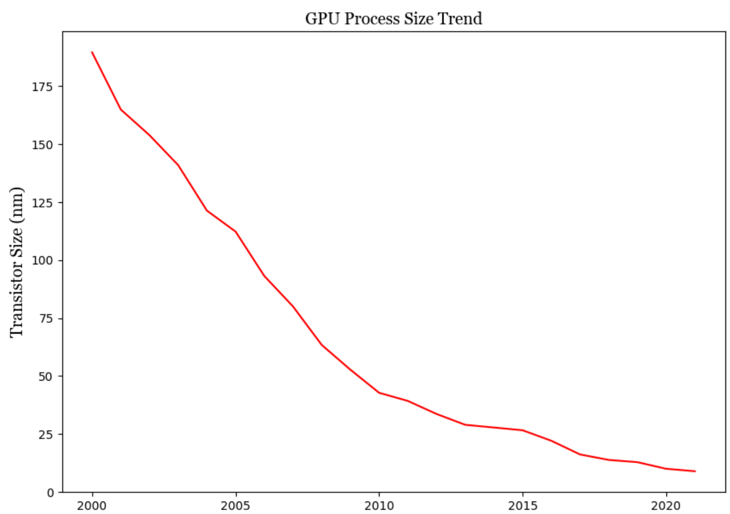
1. **High TDP for NVIDIA Tesla and Quadro Plex GPUs:**
   - The **highest TDP** in the dataset belongs to the **NVIDIA Tesla S2050 at 900W**, followed by other Tesla models (S1075, S1070, S870) at **800W**.
   - The **Quadro Plex series** also has high TDPs, with models like the **Quadro Plex 2200 D2 at 640W**.
   - These GPUs are designed for **high-performance computing (HPC), AI, and professional visualization**, requiring substantial cooling.

2. **Low TDP for Intel Atom CPUs (Not in the graph):**
    - The **Intel Atom Z-series CPUs** have an extremely low TDP of **1W**.
    - These CPUs are meant for **low-power, energy-efficient applications**, such as mobile devices, embedded systems, and lightweight computing.

Process Size Trends

Process size (nm) represents the **smallest feature size** of transistors that can be manufactured on a chip.

GPUs:

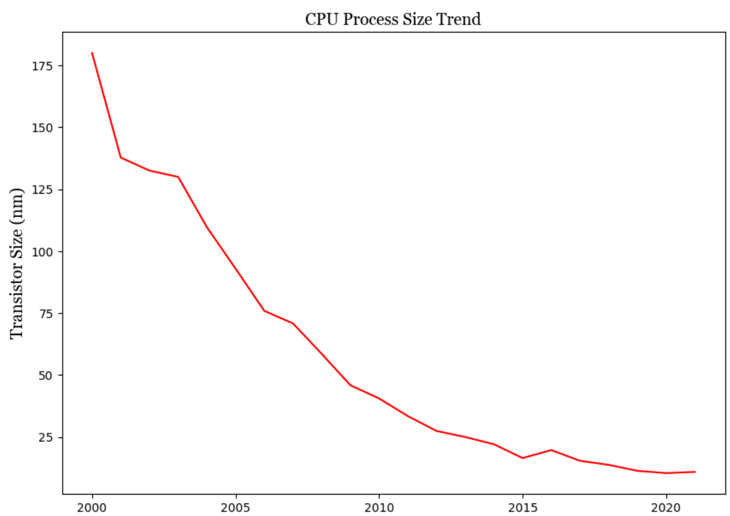


- In **2000**, the average GPU process size was **189.64 nm**, which was quite large.
- A **steady decline** is observed over the years, reaching **63.44 nm in 2008**.
- The most **significant reduction** occurs between **2008 (63.44 nm) and 2012 (33.65 nm)**, showing rapid advancements in GPU fabrication technology.
- By **2016, the process size dropped to 22.09 nm**, and in **2017, it shrank to 16.18 nm**, marking a key shift towards more efficient and powerful GPUs.
- By **2020, GPUs reached 10 nm**, and in **2021, the smallest recorded process size is 8.94 nm**, reflecting the use of advanced **7 nm and 8 nm nodes** in modern GPUs.

Observations for GPUs:

1. **Consistent Shrinkage**: The process size has decreased **every few years**, showing a clear trend toward efficiency.
2. **Rapid Technological Advancements (2008–2012)**: This period saw **major reductions** in process size, indicating significant R&D investment in GPU miniaturization.
3. **Plateau at Sub-10nm**: By **2020–2021, the process reached around 8.94 nm**, indicating that further scaling may be slowing due to physical limitations in transistor design.

CPUs:



- CPUs had a **starting process size of 180 nm in 2000**, slightly smaller than GPUs at the time.
- A **steady decline** occurs, reaching **58.57 nm in 2008**, a similar pattern seen in GPUs.
- A major **reduction occurs between 2008 and 2012**, where the process size dropped from **58.57 nm (2008) to 27.45 nm (2012)**.
- Unlike GPUs, **CPU process size increased slightly in 2016 (19.72 nm) before dropping again in 2017 (15.41 nm)**. This may be due to a temporary shift in design priorities or challenges in manufacturing smaller transistors.
- By **2021, the process size reached approximately 10.92 nm**, showing that CPUs, like GPUs, have approached the limits of silicon fabrication.

Observations for CPUs:

1. **Similar Downward Trend as GPUs**: CPUs follow a process size reduction trend almost parallel to GPUs, though they reached sub-10 nm slightly later.
2. **Slower Shrinkage Compared to GPUs**: After **2015, the rate of decrease slowed down**, likely due to increasing complexity in CPU transistor design.
3. **Fluctuation in 2016**: Unlike GPUs, CPUs had a slight **increase in process size in 2016**, possibly due to manufacturing constraints or shifts in focus towards power efficiency rather than raw size reduction.
4. **Stabilization Around 10 nm**: By **2020–2021, the process size hovers around 10.92 nm**, indicating a slowdown in further shrinkage.

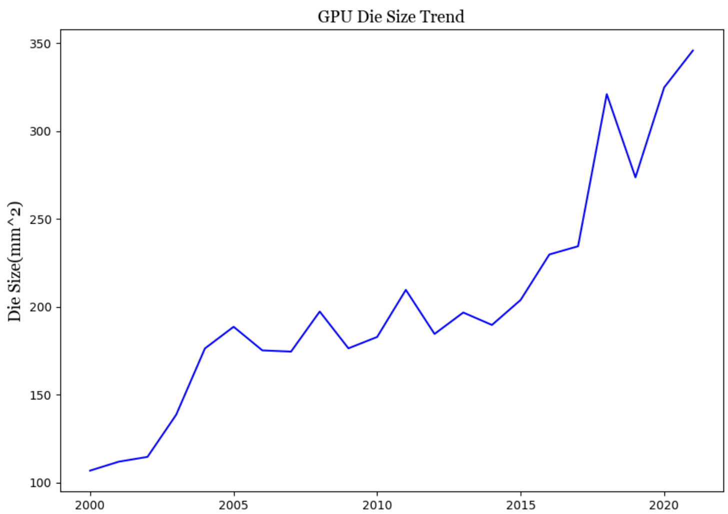
Key Comparisons:

1. **GPUs initially had larger process sizes than CPUs** but eventually caught up by 2015.
2. **CPUs reached smaller process sizes faster (e.g., sub-20 nm by 2015)**, whereas GPUs took longer to drop below **10 nm**.
3. **Both GPUs and CPUs are now reaching process size limits (~10 nm in 2021)**, indicating that further shrinkage may require breakthroughs in new materials (e.g., 3D stacking, nanowires, or quantum computing).

<u>Die Size Analysis</u>

The **die size** refers to the **physical area of the silicon chip** that contains the transistors and circuits. It is measured in **square millimeters (mm²)**.

GPUs:

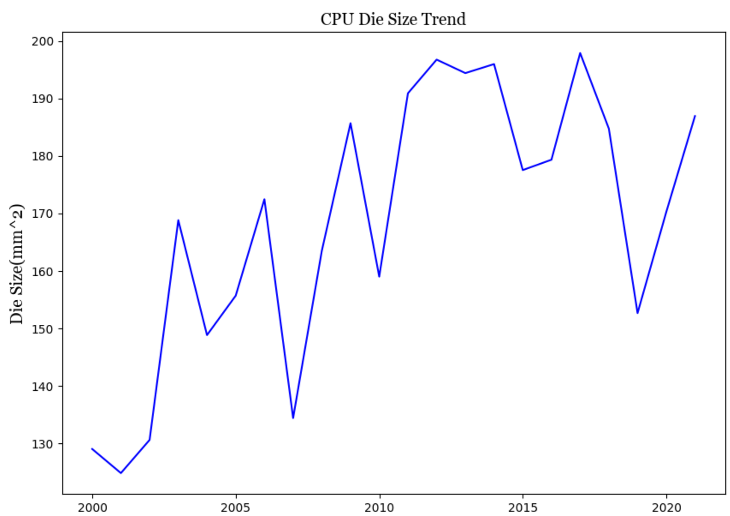


- In **2000**, the average GPU die size was **106.85 mm²**, relatively compact.
- A **steady increase** follows, reaching **176.31 mm² in 2004** and **188.64 mm² in 2005**, indicating a shift towards more powerful GPUs.
- A **fluctuation between 170-200 mm²** occurs from **2006 to 2016**, suggesting that manufacturers were balancing performance and efficiency.
- In **2017, the die size reached 234.43 mm²**, marking a significant growth phase in GPU capabilities.
- A **major leap occurs in 2018**, where the die size jumps to **320.95 mm²**, indicating a rise in **high-performance GPUs with more cores and larger memory buses**.
- By **2020, the die size further increases to 324.74 mm²**, and in **2021, it reaches the highest value of 345.75 mm²**, suggesting the dominance of **powerful GPUs for gaming, AI, and computing**.

Observations for GPUs:

1. **Gradual Growth Until 2016**: GPU die sizes remained within **175-200 mm^2** for about a decade.
2. **Major Expansion in 2018**: The increase to **320+ mm^2** suggests a shift towards larger, more powerful GPUs.
3. **Peak at 2021 (345.75 mm^2)**: Reflects the trend of **high-performance GPUs with massive transistor counts**, driven by gaming, AI, and professional workloads.

CPUs.



- In **2000**, the CPU die size was **129.06 mm^2**, slightly larger than GPUs of that time.
- The size fluctuates **between 130-170 mm^2** until **2007**, indicating steady improvements in efficiency.
- A **spike occurs in 2009 (185.70 mm^2)**, possibly due to increased core counts and cache sizes.
- From **2010 to 2016, the die size remains between 159-196 mm^2**, suggesting manufacturers optimized CPU designs rather than increasing the die size.
- In **2017, CPU die size reaches 197.88 mm^2**, showing an increased emphasis on **multi-core architectures**.
- By **2019, the die size decreases to 152.68 mm^2**, possibly due to process refinements, but increases again in **2020 (170.25 mm^2) and 2021 (186.92 mm^2)**.

Observations for CPUs:

1. **Stable Die Size (2000–2016)**: CPU die sizes stayed between **130-190 mm²**, suggesting **a focus on efficiency and scaling**.
2. **Growth in 2017 (197.88 mm²)**: Indicates a transition to **high-core-count CPUs**.
3. **Reduction in 2019 (152.68 mm²) Followed by an Increase**: The fluctuation suggests a balance between **efficiency and power**.

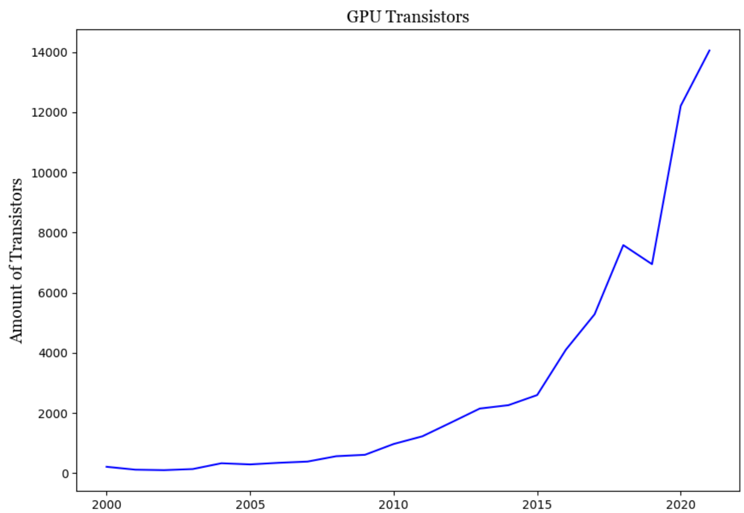
Key Comparisons:

1. **CPUs Had Larger Dies Until ~2007**: Early CPUs had **larger dies than GPUs**, but GPUs surpassed them as graphics demands grew.
2. **GPUs Experienced a Sharp Die Size Increase in 2018**: Unlike CPUs, which remained stable, GPUs **grew significantly** due to gaming, AI, and deep learning demands.
3. **CPUs Show More Stability in Die Size**: While GPU dies **steadily increase**, CPU dies **fluctuate**, reflecting different optimization strategies.

Transistor Count Growth

The **amount of transistors** refers to the **number of microscopic electronic switches** that make up the processing logic of the chip. The more transistors a chip has, the more computational power it can deliver.

GPUs:

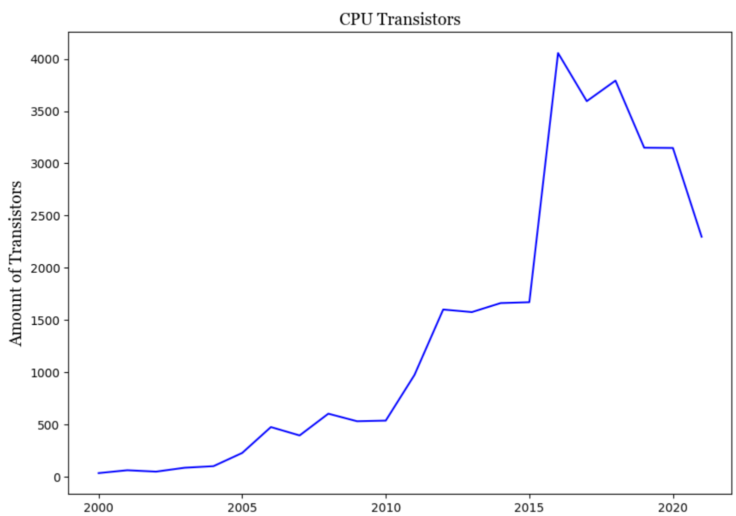


- In **2000**, the average GPU transistor count was **215.53 million**, and it **decreased to 104.27 million in 2002** before beginning a steady rise.
- A **significant jump occurs in 2004 (332.93 million)** and continues increasing annually, reaching **613.83 million in 2009**.
- **The first major spike happens in 2010 (972.98 million transistors)**, coinciding with advancements in **graphics architecture**.
- From **2012 to 2015, transistor counts exceed 1 billion**, rising to **2.6 billion in 2015**.
- A **dramatic surge follows in 2016, reaching 4.1 billion transistors**, reflecting the shift towards **AI and deep learning acceleration** in GPUs.
- **2017 to 2019 show a steady increase**, with transistors peaking at **7.58 billion in 2018** before slightly dropping in 2019.
- **2020 marks a breakthrough with 12.21 billion transistors**, followed by another jump to **14.05 billion in 2021**, reflecting cutting-edge **ray tracing and AI-focused GPUs**.

Observations for GPUs:

1. **Early Growth (2004–2010):** A steady rise as GPUs became more essential for gaming and computational tasks.
2. **Major Spike in 2016 (4.1B transistors):** Likely due to advancements in **deep learning and AI accelerators**.
3. **Peak in 2021 (14.05B transistors):** Driven by **next-gen GPU architectures, high-performance gaming, and AI computing**.

CPUs:



- In **2000, CPUs had only 38 million transistors**, much lower than GPUs.
- Growth was relatively slow until **2005 (230.58 million transistors)**, suggesting improvements in **efficiency rather than raw transistor count**.
- A **major jump occurs in 2006 (478.26 million transistors)**, signaling an increase in **multi-core processors and larger cache sizes**.
- **Between 2008 and 2011, CPU transistor counts steadily increase**, reaching **976.61 million transistors in 2011**.
- **2012 sees a massive rise (1.6 billion transistors)**, maintaining steady growth through **2015 (1.67 billion transistors)**.
- A **huge spike in 2016 (4.06 billion transistors)** coincides with a shift to **multi-core, multi-threaded CPU designs**.

- The CPU transistor count fluctuates after 2016, with a decline in **2019 (3.15 billion transistors)** and **2021 (2.3 billion transistors)**, possibly due to a focus on **process efficiency rather than raw transistor growth**.

Observations for CPUs:

1. **Steady Growth (2000–2010):** Gradual increase as **multi-core processors** became mainstream.
2. **Major Increase in 2016 (4.06B transistors):** Reflects the push towards **higher-core-count architectures**.
3. **Fluctuations after 2017:** Suggests that manufacturers optimized **efficiency and architectural improvements** instead of just increasing transistor counts.

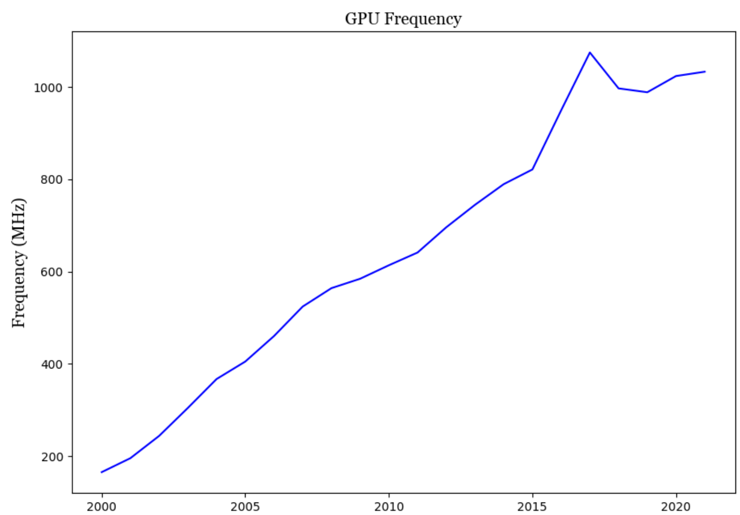
Key Comparisons:

1. **GPUs Surpassed CPUs in Transistor Count After 2006:** Early on, CPUs and GPUs had similar transistor counts, but **GPUs quickly outpaced CPUs** due to the increasing complexity of graphics and AI workloads.
2. **GPU Transistor Growth Accelerates Rapidly After 2016:** While CPU growth slows down, **GPUs surge past 10 billion transistors in 2020**, reflecting the **demand for AI, deep learning, and gaming advancements**.
3. **CPU Transistor Count Peaks and Declines After 2017:** This suggests that **efficiency and architecture optimizations** became a bigger focus than increasing transistor counts.

Frequency Trends

Frequency, also known as **clock speed**, measures how fast a CPU or GPU executes instructions. It is measured in **Hertz (Hz)**, typically in **Gigahertz (GHz)** for modern processors.

GPUs:



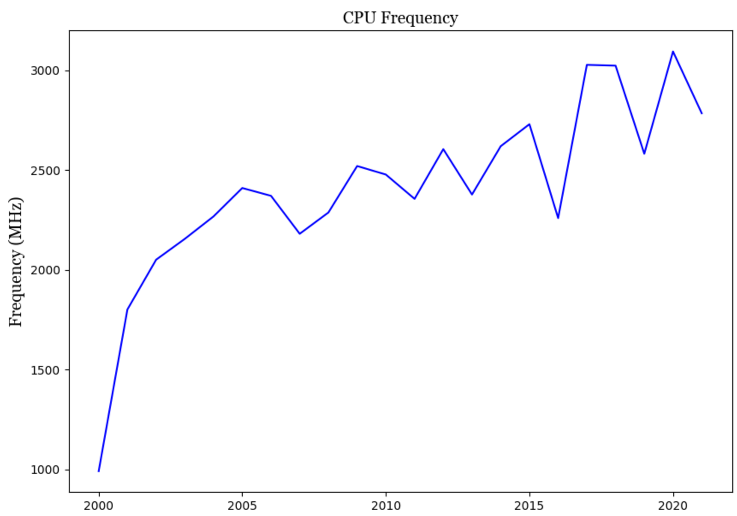
- In **2000**, the average GPU frequency was **165.8 MHz**, increasing steadily each year.
- **The early 2000s saw rapid growth**, reaching **405.27 MHz in 2005**.
- **Between 2006 and 2010, the GPU frequency continued rising**, hitting **613.74 MHz in 2010**.
- **2011–2015 showed steady improvements**, peaking at **821.25 MHz in 2015**.
- A **significant jump occurred in 2016, reaching 949.27 MHz**, likely due to advancements in **parallel computing and deep learning acceleration**.
- **2017 saw the highest recorded increase (1074.54 MHz), but post-2018, frequencies plateaued** around the **1 GHz mark**.

Observations for GPUs:

1. **Rapid Growth (2000–2006):** Early GPUs saw a **continuous increase in clock speeds**, as higher frequencies improved **rendering performance**.

2. **Stable Growth (2010–2015):** Clock speeds continued rising but at a slower pace, likely due to a shift toward **architecture optimizations** rather than raw frequency increases.
3. **Peak in 2017 (1074.54 MHz), Followed by Plateau:** Instead of increasing clock speeds indefinitely, manufacturers focused on **parallelism (CUDA cores, tensor cores, and efficiency improvements)**.

CPUs:



- In **2000, CPU frequency was 992.12 MHz**, much higher than GPUs at the time.
- **A massive jump occurred in 2001 (1.80 GHz) and 2002 (2.05 GHz), showing the rapid progression in CPU technology**.
- **Between 2002 and 2006, CPU frequencies peaked at around 2.4 GHz**, reflecting improvements in **single-threaded performance**.
- **Between 2006 and 2012, frequencies fluctuated but remained around 2.5 GHz**, as the industry shifted towards **multi-core architectures rather than increasing clock speeds**.
- **In 2017, CPU frequencies surged to 3.03 GHz**, likely due to the **rise of high-performance desktop and server processors**.
- **After 2017, CPU frequencies stabilized between 2.5 GHz and 3.1 GHz**, with only small variations.

Observations for CPUs:

1. **Fast Growth (2000–2004):** Frequencies doubled within four years as manufacturers **focused on single-core performance**.
2. **Frequency Plateau (2006–2015):** As power and heat limitations became critical, CPU designs shifted toward **multi-core processing instead of higher frequencies**.
3. **Post-2017 Optimizations:** While peak frequencies hit **3.1 GHz**, focus shifted toward **core count increases and power efficiency rather than raw clock speed gains**.

Key Comparisons:

1. **CPUs Had Much Higher Frequencies Initially:** In **2000, CPUs were already near 1 GHz**, while GPUs were only **165.8 MHz**.
2. **GPU Frequencies Grew Steadily, While CPU Frequencies Plateaued:**
   - GPUs **gradually increased frequency** while focusing on **parallel processing (CUDA, AI accelerators, tensor cores)**.
   - CPUs **peaked around 3.1 GHz in 2017** and have since remained stable due to **thermal and power constraints**.
3. **By 2021, GPU Frequencies Still Lag Behind CPU Frequencies:** Despite gains, **GPUs operate at ~1 GHz**, while **CPUs are ~2.8 GHz** due to their different architectures and workloads.

- **Recommendations for Picking Based on Key Metrics (GPU + CPU):**

TDP (Thermal Design Power)

- When considering TDP, opt for **higher TDP chips** if you need **high performance** for gaming, video editing, or AI tasks, as they often allow higher clock speeds and better multi-threaded performance. However, **lower TDP chips** are ideal for **energy efficiency**, such as in laptops or embedded systems, as they generate less heat and require less cooling.

Process Size

- When examining process size, opt for **smaller process nodes** (e.g., 5nm, 7nm) for **better performance and power efficiency**, particularly for high-performance computing or gaming. However, **larger process sizes** (e.g., 14nm, 28nm) might be **more cost-efficient** for simpler tasks or budget systems, though they consume more power.

Die Size

- A **larger die size** generally means **more transistors**, leading to better parallel processing, making it suitable for **GPUs, AI accelerators, and high-end CPUs**. However, it also increases **power consumption and cost**. Smaller dies are **cheaper to produce and more power-efficient**, making them preferable for **laptops and low-power devices**.

Transistor Count

- Higher transistor count enables **greater computational power and efficiency**, which is beneficial for **AI, gaming, and multi-threaded applications**. However, chips with **fewer transistors** are often **more affordable and power-efficient**, making them better suited for **basic tasks and embedded applications**.

Frequency

- A **higher frequency (GHz)** is beneficial for **gaming and single-threaded applications**, as it results in faster execution of tasks. However, **lower frequencies** are more **energy-efficient**, leading to longer battery life in laptops and mobile devices, though performance may be lower for demanding tasks.