Exploring Modern Computer Hardware

Julius Chan

Table of Contents

Table of Figures

1.0 Introduction

The coursework will investigate the impact of Moore's Law on the semiconductor industry from its beginnings to current developments. It will examine the growing landscape of applications such as AI, computer vision, and the Internet of Things, as well as scaling difficulties and creative solutions such as chip-embedded cooling.

Furthermore, the research will emphasize the complex relationship between hardware and software by examining cutting-edge processors like as Google's TPU and Apple's SiP. These examples highlight the ongoing need for computer innovation.

Finally, the purpose of this essay is to offer readers with a thorough grasp of how hardware and software work together to satisfy the changing demands of computer systems. It will demonstrate the ongoing creativity necessary to respond to disruptive trends in an era defined by rapid change.

2.0 Moore’s Law

2.1 Introduction to Moore's Law

Moore's Law, stated by Intel co-founder Gordon Moore in 1965 indicates that the number of transistors on an integrated circuit would double every two years while the relative cost goes down. Moore changed his prediction of an annual doubling over a decade to a two-year cycle in 1975. Despite being referred to as a "law," it is a projection based on technology invention rather than a natural occurrence.

Moore's Law has guided the semiconductor industry for almost 60 years, pushing businesses to achieve the double transistor target. Although some have predicted that Moore's Law will eventually break down because of physical limitations, recent developments in packaging technology and materials have kept Moore's Law moving forward, highlighting the significance of continued innovation in the industry.

As seen in Figure 1, the number of transistors jammed onto integrated circuits has increased at an astounding rate since Intel's first 4004 chip in 1971. Moore's law, a 1975 phrase, states that the number should double every two years. However, maintaining the trend is becoming increasingly difficult and costly.

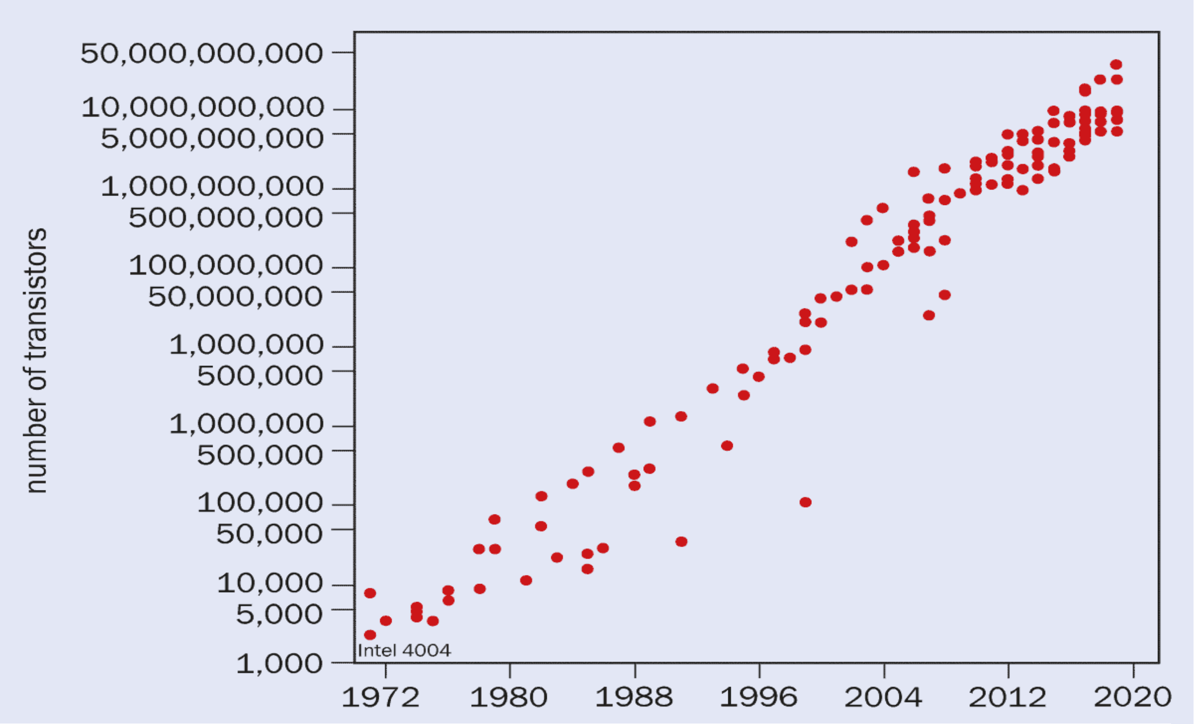


Figure 1 : Moore's Law Graph (https://physicsworld.com/a/moores-law-further-progress-will-push-hard-on-the-boundaries-of-physics-and-economics/, 2023)

2.2 Issues of Scalability

As transistors shrank in size, transistor leakage and heat dissipation became major obstacles to Moore's Law's scalability and heat dissipation appears as a major worry in this setting. This raises concerns regarding the possibility of additional scaling to provide greater chip functionality while also introducing heat management issues.

The power density on the chip rises as chip components decrease and are packed more tightly which results in higher temperatures where excessive heat can degrade performance and shorten the lifespan of the components.

A solution of chip-embedded cooling has been created to tackle the issue of overheating in electronic chips stacking arrangements. With the help of this creative solution, complex isolation techniques are avoided by pumping a nonconductive dielectric fluid between the chips in a stack. An effective cooling loop is created when the fluid takes in heat, boils into vapor, and then condenses again outside the chip. (Electronics Cooling, n.d.)

This chip-embedded cooling method offers a viable way to reduce heat problems in advanced chip layouts and improve computing performance.

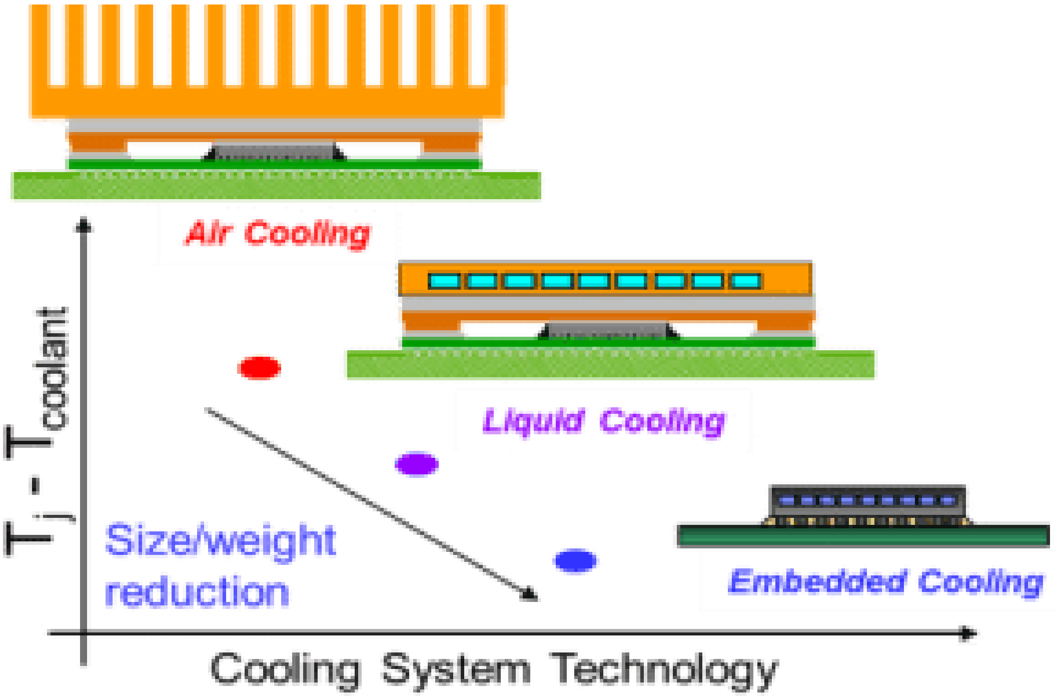


Figure 2 : Cooling Solution of 3D Chip Stacks (https://www.electronics-cooling.com/2018/03/beat-heat-3d-chip-stacks-embedded-cooling/, n.d.)

2.3 Future Development

Moore's Law may see positive changes in the future, according to recent reports that emphasize cutting-edge lithography and packaging technology. According to the Electronics Components and Technology Conference (ECTC), innovative packaging has grown increasingly important.

The industry is expected to grow at a Compound Annual Growth Rate (CAGR) of 10.6% from 2022 to 2028, or US$78.6 billion. By 2028, high-end performance packaging in particular 2.5D/3D platforms is predicted to account for more than 20% of the advanced packaging market. (Yole Group, 2023)

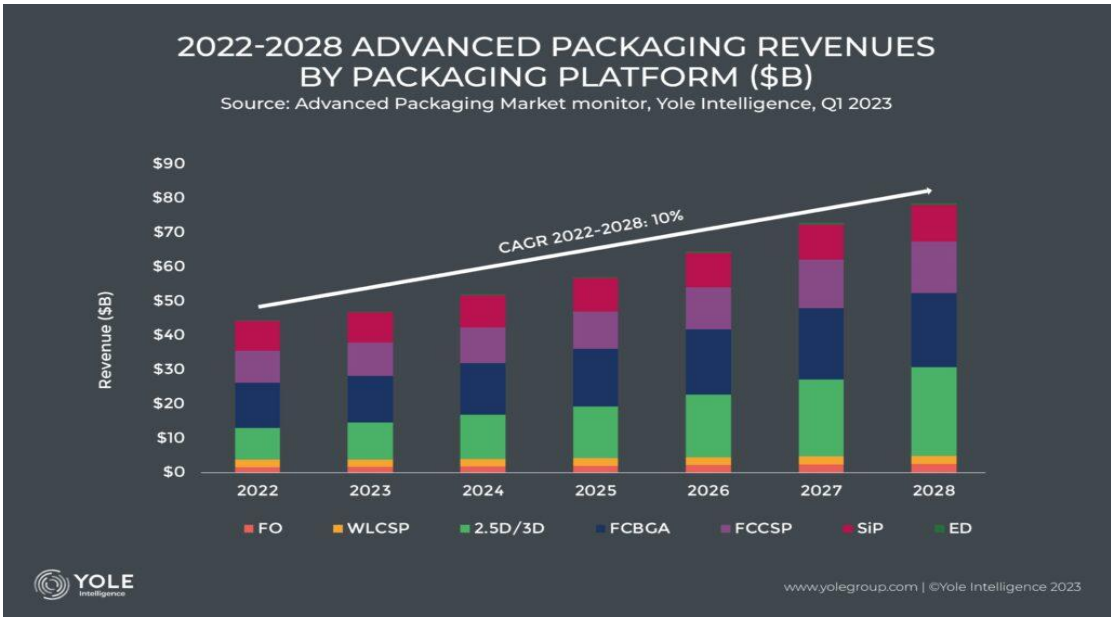


Figure 3 : 2022 -2028 Advanced Packaging Revenue (https://www.yolegroup.com/strategy-insights/innovation-beyond-moores-law-advanced-packaging-explores-new-frontiers/, 2023)

ASML, the sole supplier of extreme ultraviolet (EUV) lithography equipment, intends to offer tools with increased numerical aperture (NA) of 0.55, extending Moore's Law by at least a decade.

This development tackles issues at the 3nm chip process technology node, allowing chipmakers to begin with a low-cost single-exposure EUV process before switching to multi-patterning for more advanced nodes.

These advancements in packaging and lithography technologies are sustaining the endurance of Moore's Law especially in the era of artificial intelligence which indicates an optimistic future for the semiconductor industry.

3.0 Applications Needs

3.1 Artificial Intelligence

Artificial Intelligence (AI) refers to technologies that provide machines the ability to carry out tasks that need intelligence similar to the level of humans. For example, users may communicate with gadgets due to natural language processing (NLP) in virtual personal assistants.

While e-commerce platforms use AI-driven recommendation systems to provide tailored product choices. The healthcare industry uses AI to help with picture identification for medical diagnosis.

Artificial intelligence is used by autonomous cars to make decisions and navigate. These use cases highlight how versatile AI is in improving productivity and user experiences in a variety of industries.

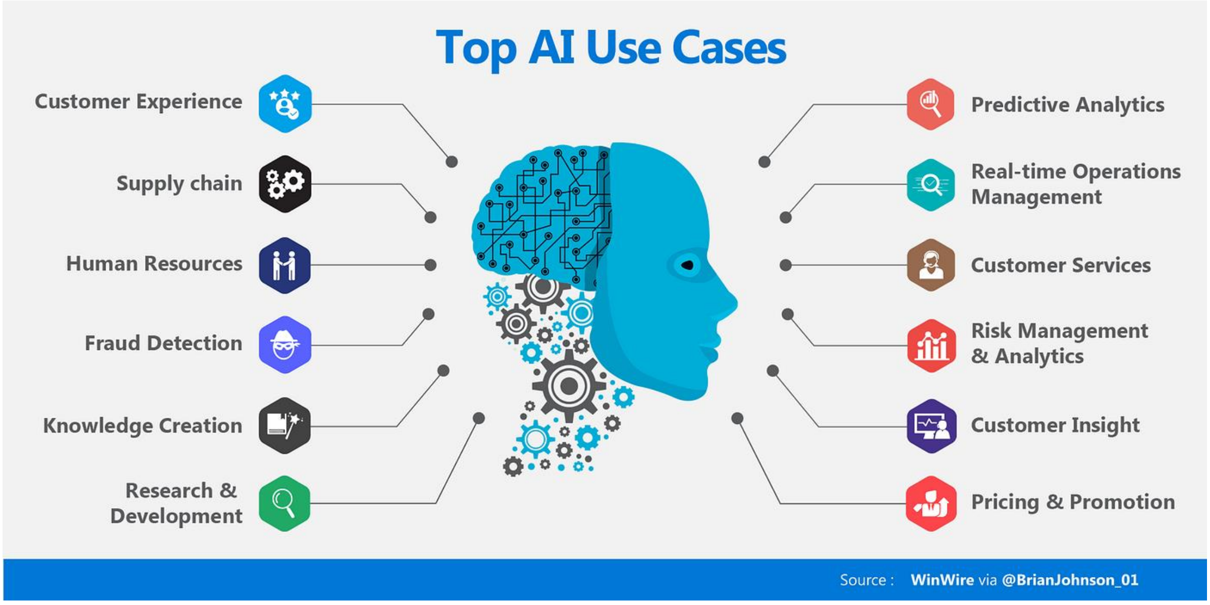


Figure 4 : Top AI Use Cases (https://medium.com/@Brian.johnson_62680/artificial-intelligence-ai-top-use-cases-and-technologies-used-today-3c22e1a63e78, 2018)

3.1.1 Computer Vision

Computer vision (CV) is a major task for current Artificial Intelligence (AI) and Machine Learning (ML) systems. It is accelerating practically every industrial domain, allowing companies to rethink the way equipment and business systems function. (Medium, 2022)

In 2015, Google released TensorFlow as an open-source technology where it became one of the most widely used computer vision software tools and its efficient operation is heavily dependent on well-suited hardware configurations.

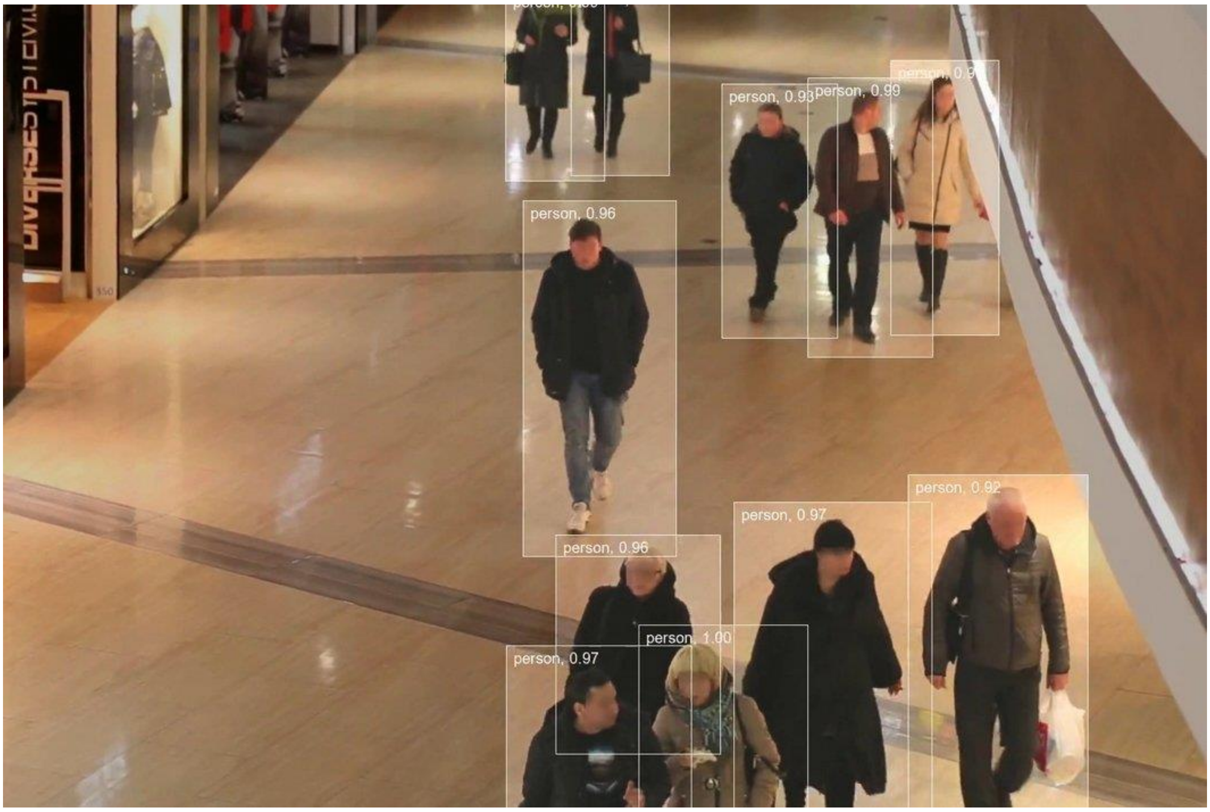


Figure 5 : People Detection with TensorFlow Lite

TensorFlow is built to leverage from Google's Tensor Processing Units (TPUs). TPUs are hardware accelerators created exclusively for machine learning workloads by Google. TensorFlow has introduced enhancements that allow users to use TPUs more effectively for activities like training and inference, especially in large-scale distributed training scenarios.

Even though TensorFlow is compatible with a wide range of hardware configurations, including CPUs and GPUs, the incorporation of TPUs is a remarkable feature that boosts its capabilities to the point where it is quicker than both CPUs and GPUs.

3.2 Internet of Things (IoT)

The concept of adding sensors and intelligence to physical objects was first discussed in the 1980s, when some university students decided to modify a Coca-Cola vending machine to track its contents remotely. (Vision of Humanity, n.d.)

In today's digital age, wearable technology has been a key component of the Internet of Things (IoT), allowing devices with sensors and communication capabilities to be connected for easy data sharing. In particular, wearables such as smartwatches showcase how IoT integration may be used to deliver real-time health data, including heart rate monitoring.

The core hardware components of a typical wearable technology device consist of a System on a Chip (SoC) which serves as the main processing unit and houses the CPU and GPU. Furthermore, these devices include internal storage to hold firmware, applications, and user data, as well as memory (RAM) for quick data access.

As seen in other industries, wearables are not just affected by IoT in terms of health monitoring. Wearable devices with IoT sensors also helps by tracking weather patterns and soil conditions which enable farmers to make well-informed decisions.

3.2.1 Wearable Device (Apple Smart Watch Series)

The hardware components of the Apple Smart Watch have seen a remarkable development, consistently pushing past the limits of innovation, and establishing new benchmarks in the wearable technology industry. When the original Apple Watch was introduced in April 2015, it had a digital crown, Force Touch, Taptic Engine, heart rate sensor, and a custom designed S1 chip that laid the foundation for future development.

The launch of the S3 chip in the Series 3 represented an important event that increase the smartwatch's capabilities by including a faster dual-core CPU and allowing mobile connection in selected models. With the introduction of a 64-bit dual-core CPU that not only increased overall performance but also introduced features such as the ECG app for heart health monitoring.

In 2023, the latest Apple Smart Watch series 9 seen in Figure 6 has continued the heritage of hardware excellence by integrating the latest chip developments while maintaining the 64-bit dual-core CPU with a new four-core Neural Engine performs machine learning tasks up to two times quicker. The new dual-core CPU contains 5.6 billion transistors, which is 60% more than its predecessor, the S8 chip. The series 9 also introduced a larger and more durable crack-resistant display and an increase of storage capacity.

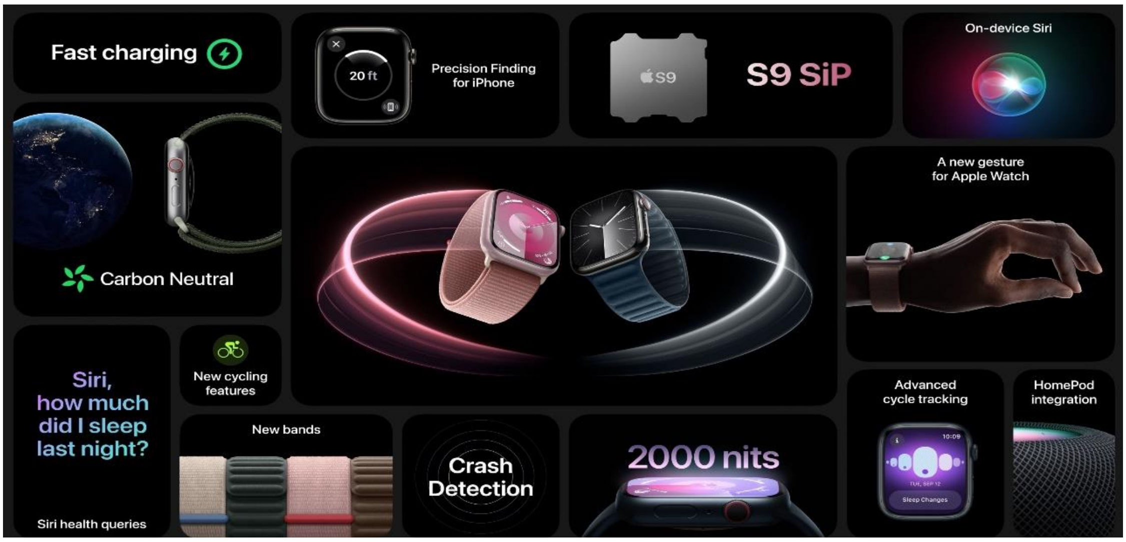


Figure 6 : Apple Smartwatch Series 9 (https://www.cultofmac.com/790236/apple-watch-series-9-launch/, 2023)

The Apple Smart Watch hardware combinations with each succeeding series, demonstrate a thorough approach to drive performance across different levels.

This approach not only meets the demands of the user but also places the gadget at the forefront of technological growth, guaranteeing an exciting future for wearable technology.

4.0 Discussions

4.1 Google TPU

Google TPUv5e (2023)

Key Specifications	Description
Process Node	7 nm
On-Chip Memory	48MiB
Peak compute per chip (bf16)	197 TFLOPs
Peak compute per chip (Int8)	393 TLOPs
High-bandwidth Memory	16GB
Memory Bandwidth	1200/s
TPU Pod Size	256 Chips
Processors	Scalar, Vectors and Matrix Multiplication Units

Google Tensor Processing Units (TPU) are custom-designed AI processors that are geared for big AI model training and inference. They are accessible through rent from Google Cloud and may be used for a variety of applications such as chatbots, code development, media content production, synthetic voice, computer vision, recommendation engines, and customization models.

In 2023, Google announced the general availability of Cloud TPU v5e which is the newest generation AI accelerator from Google Cloud. The v5e Pods are optimal for transformer-based, text-to-image, and CNN-based training.

The device has a peak performance of 393 teraflops of INT8 performance per chip, which is more than the TPU v4's 275 petaflops.

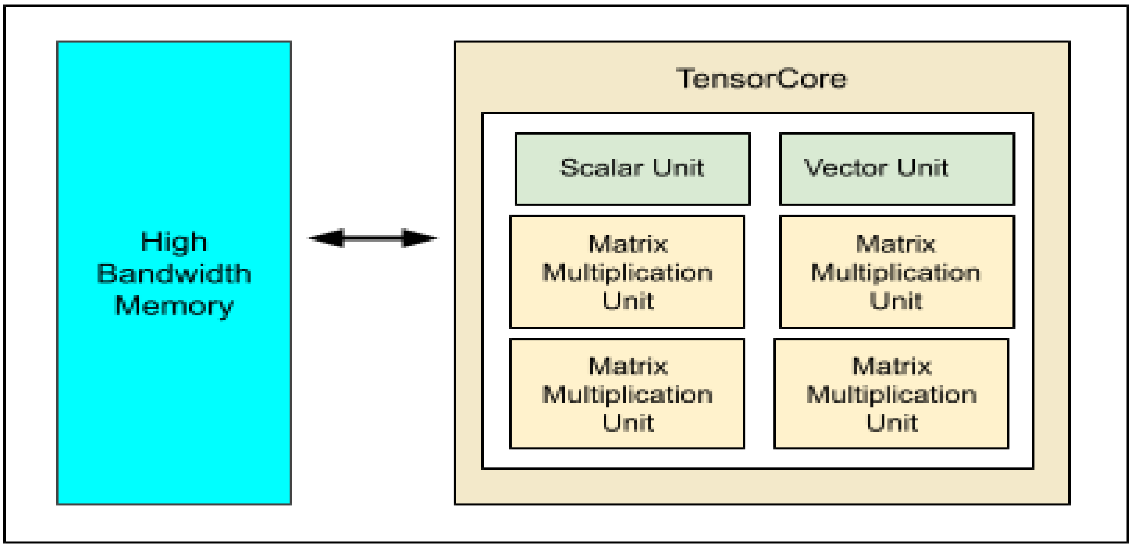

TensorCore
Scalar Unit
Vector Unit
Matrix Multiplication Unit
Matrix Multiplication Unit
Matrix Multiplication Unit
Matrix Multiplication Unit
High Bandwidth Memory

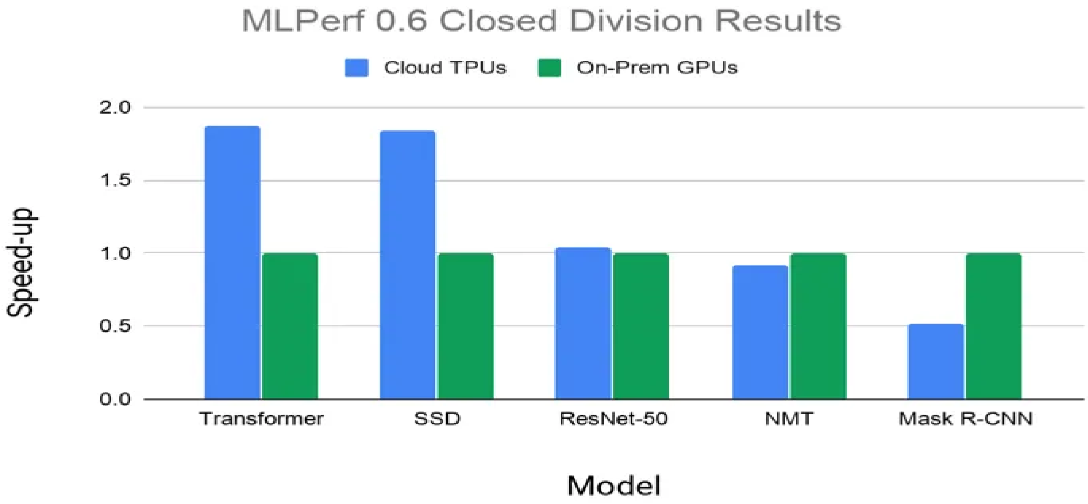
Figure 7 : Google v5e system architecture (https://cloud.google.com/tpu/docs/system-architecture-tpu-vm, n.d.)

Figure 7 illustrates a TPU v5e chip that has one TensorCore. Each TensorCore has four Matrix Multiply Units (MXU), one vector unit, and one scalar unit.

Matrix Multiplication Units are specialized components of a microprocessor that are designed to conduct matrix multiplication tasks quickly. These units are critical in deep learning applications for tasks such as neural network training and inference.

Vector Processing Units (VPUs) are specialized components meant to speed up parallelized operations like multimedia processing. These units are designed to handle vectors, which are data arrays.

A scalar unit is intended to handle single data values at one time. They carry out instructions on discrete chunks of data with no intrinsic parallelism.



Google Cloud TPU v3 Pod speed-ups over the largest-scale on-premise NVIDIA DGX-2h clusters entered in the MLPerf 0.6 Closed Division. The Cloud TPU Pod submissions use 1024, 1024, 1024, 512, and 128 chips respectively; the NVIDIA DGX-2h clusters use 480, 240, 1536, 256, and 192 chips respectively. [1,2]

Figure 8:Training Comparison between accelerators (https://khairy2011.medium.com/tpu-vs-gpu-vs-cerebras-vs-graphcore-a-fair-comparison-between-ml-hardware-3f5a19d89e38, 2020)

Figure 8 shows that when GPUs and TPUs run similar batch sizes in object detection models with the same number of chips, their training performance in SSD and Transformer benchmarks is equivalent. Despite this, Google asserts that in certain workloads, TPUs outperform GPUs by 84%.

The aspect here isn't the TPU's higher power but rather Google's improved tuning of hyperparameters, batch and sub-batch sizes, which results in enhanced training throughput.

While Google's TPU may outperform in some workloads, this does not imply that it is always suitable for every model or training situation. Varied models have varied processing requirements and features and some hardware might do well in certain situations while underperforming in others.

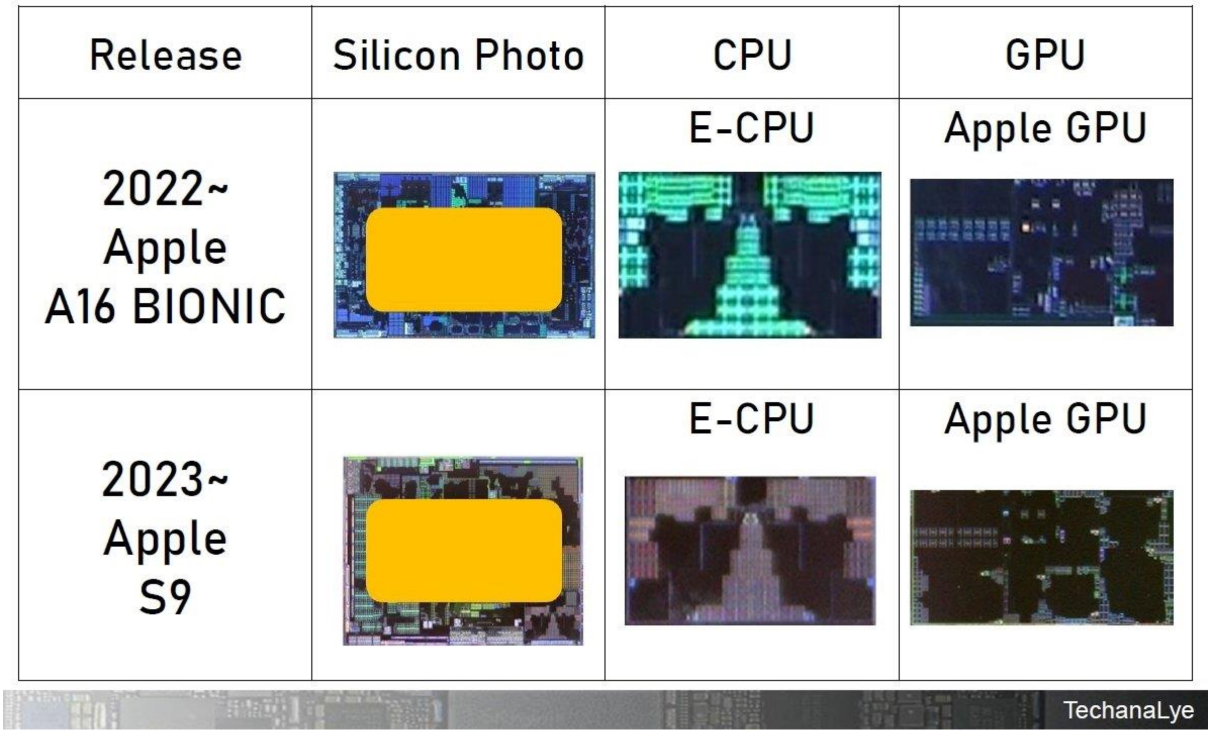
4.2 Apple System in Package

Apple S9 Chip

Key Specifications	Description
CPU Model	Sawtooth Dual-Core with 64-bit Processor
CPU Clock Speed	1.8 GHz
Number of Cores/Threads	2/2
Cache	L1d: 64 KB L2: 4 MB
Number of Transistors	5.6 billion
Processor	ARM
GPU	Custom Apple Silicon

The Apple S9 SiP (System in a Package) is a 64-bit CPU with two cores (Sawtooth Dual-Core) that is utilized in the Apple Watch Series 9 and is a system in a package. The two CPU cores use the same microarchitecture as the Apple A15 SoC. For the first time, the chip contains a Neural Engine (with four cores) for AI applications.

Figure 9: Apple S9 Processor



Release	Silicon Photo	CPU	GPU
2022~ Apple A16 BIONIC		E-CPU	Apple GPU
2023~ Apple S9		E-CPU	Apple GPU

TechanaLye

Figure 10 : Comparison of A16 & S9 (https://wccftech.com/apple-watch-series-9-s9-sip-based-on-the-a16-bionic/ , 2023)

As seen in Figure 10, the S9 System in a Package (SiP) in the Apple Watch Series 9 is based on the A16 Bionic processor from the iPhone 15 series. Despite having fewer CPU and GPU cores than the A16 Bionic, the S9 is optimized for Apple Watch power consumption. This technique highlights

Apple's scalable semiconductor design, which allows them to utilize the same underlying technology for several devices.

This scalability allows Apple to save money on design and production expenses. While the S9 cannot compete with the A16 Bionic in terms of performance, it is customized to the special needs of the Apple Watch Series 9. The S9 was not manufactured using TSMC's cutting-edge 3nm technology because of the expenses involved with such technology.

The Apple Watch's selection of a scaled-down processor shows a focus on customizing technology to the demands of a wearable gadget. Future models, such as the Apple Watch Series 10 may use TSMC's new 3nm 'N3E' process to accommodate more sophisticated chip capabilities.

5.0 Conclusion

In this coursework, I have looked at how software and hardware are tightly related, concentrating on the growing world of parallel architectures in today's computing. Also examining the effects of Moore's Law on semiconductor technology and how it continues to influence lithography and packaging advancements today.

Innovative technologies, such as chip-embedded cooling have arisen to solve issues with fewer transistors and scalability demonstrating the industry's dedication to overcoming obstacles.

Analysing application requirements like AI, computer vision, and IoT demonstrates how adaptable hardware is for a range of sectors. The advancement of wearable technology as seen by the Apple Smart Watch Series demonstrates a dedication to improving performance while accommodating individual user needs.

In conclusion, the interactions between hardware and software will continue to change computing in revolutionary ways as technology develops.

6.0 References

Boesch, G. (2023) *Tensorflow Lite - real-time computer vision on Edge Devices (2024), viso.ai*. Available at: https://viso.ai/edge-ai/tensorflow-lite/ (Accessed: 14 January 2024).

Dlewis (2022) *IOT Technologies explained: History, examples, risks & future, Vision of Humanity*. Available at: https://www.visionofhumanity.org/what-is-the-internet-of-things (Accessed: 11 January 2024).

Eadline, D. (2023) *Google introduces 'hypercomputer' to its AI infrastructure, HPCwire*. Available at: https://www.hpcwire.com/2023/12/11/google-introduces-hypercomputer-to-its-ai-infrastructure/#:~:text=%E2%80%9CCompared%20to%20TPU%20v4%2C%20TPU,TPU%20v4%20bandwidth%20was%201228GBps (Accessed: 13 January 2024).

Electronics Cooling (2018) *Beat the heat in 3D chip stacks with embedded cooling, Electronics Cooling*. Available at: https://www.electronics-cooling.com/2018/03/beat-heat-3d-chip-stacks-embedded-cooling/#:~:text=A%20solution%20to%20this%20problem,to%20carry%20away%20excess%20heat (Accessed: 10 January 2024).

Group, Y. (2023) *Innovation beyond moore's law: Advanced packaging explores new frontiers*. Available at: https://www.yolegroup.com/strategy-insights/innovation-beyond-moores-law-advanced-packaging-explores-new-frontiers/ (Accessed: 10 January 2024).

Intel (no date) *Moore's law, Intel*. Available at: https://www.intel.com/content/www/us/en/newsroom/resources/moores-law.html (Accessed: 09 January 2024).

Khera, V. (2022) *How is the semiconductor industry handling scaling: Is Moore's law still alive?, Corporate and Culture - Cadence Blogs - Cadence Community*. Available at: https://community.cadence.com/cadence_blogs_8/b/corporate/posts/soc-demands-is-moore-s-law-still-alive (Accessed: 09 January 2024).

McKenzie, J. (2023) *Moore's law: Further progress will push hard on the boundaries of physics and economics, Physics World*. Available at: https://physicsworld.com/a/moores-law-further-progress-will-push-hard-on-the-boundaries-of-physics-and-economics/ (Accessed: 09 January 2024).

Patterson, A. (2021) *Moore's law could ride EUV for 10 more years, EE Times*. Available at: https://www.eetimes.com/moores-law-could-ride-euv-for-10-more-years/ (Accessed: 09 January 2024).

PincusUnited States, H.K., Kressel, H. and United States (no date) *Home, Artificial Intelligence in Science : Challenges, Opportunities and the Future of Research | OECD iLibrary*. Available at: https://www.oecd-ilibrary.org/sites/63e48242-en/index.html?itemId=%2Fcontent%2Fcomponent%2F63e48242-en (Accessed: 09 January 2024).

Sohail, O. (2023) *Apple Watch Series 9's S9 SIP is a 4nm part and cut-down version of the A16 Bionic, revealing a scalable architecture for various product lines, Wccftech*. Available at: https://wccftech.com/apple-watch-series-9-s9-sip-based-on-the-a16-bionic/ (Accessed: 12 January 2024).

System architecture | cloud TPU | google cloud (no date) *Google*. Available at: https://cloud.google.com/tpu/docs/system-architecture-tpu-vm (Accessed: 12 January 2024).

Tamplin, T. (2023) *Moore's law: Definition, components, impact, limitations*, *Finance Strategists*. Available at: https://www.financestrategists.com/wealth-management/moores-law/ (Accessed: 09 January 2024).

Tardi, C. (no date) *What is Moore's law and is it still true?*, *Investopedia*. Available at: https://www.investopedia.com/terms/m/mooreslaw.asp#:~:text=Moore's%20Law%20states%20that%20the,will%20pay%20less%20for%20them. (Accessed: 09 January 2024).

Wade, A. (2020) *Viewpoint: Moore's law isn't broken - it's overheated*, *The Engineer*. Available at: https://www.theengineer.co.uk/content/in-depth/viewpoint-moore-s-law-isn-t-broken-it-s-overheated/ (Accessed: 10 January 2024).

7.0 Appendix

Figure 11 : Air Cooling System in 3D Chip Stacks (https://www.electronics-cooling.com/2018/03/beat-heat-3d-chip-stacks-embedded-cooling/ , .n.d.)

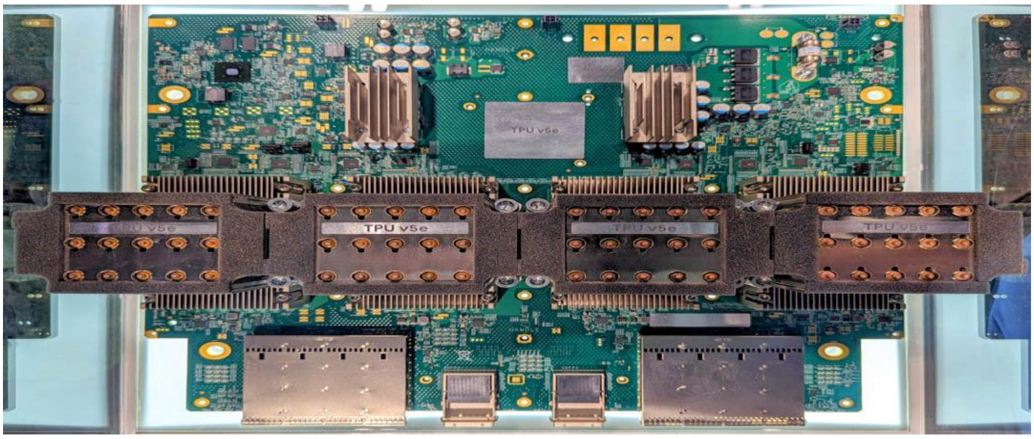


Figure 12 : Google TPU V5e Overhead (https://www.servethehome.com/google-brings-tpu-v5e-ai-board-to-sc23/ , 2023)

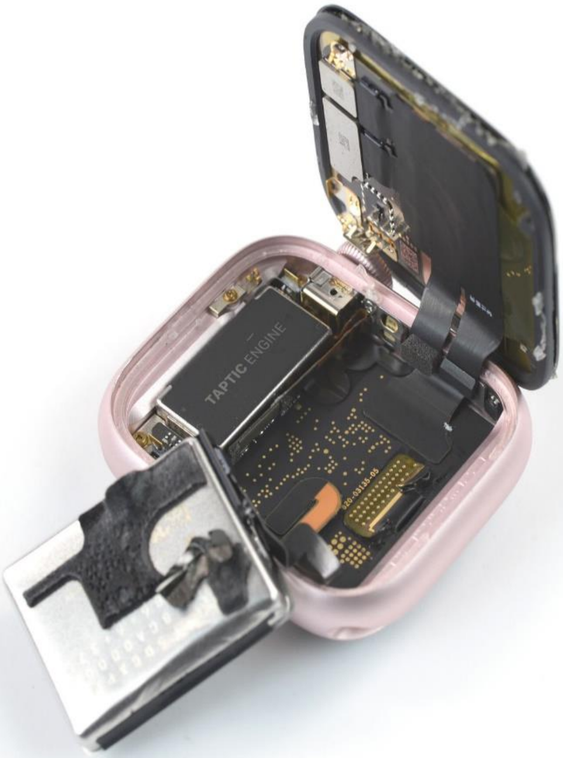


Figure 13 : Teardown of Apple Watch S9 (https://www.servethehome.com/google-brings-tpu-v5e-ai-board-to-sc23/ ,2023)