1.4 Fueling Semiconductor Innovation and Entrepreneurship in the Next Decade

Lip-Bu Tan

Chairman of Walden International,
Founding Managing partner of Walden Catalyst Ventures,
Senior Advisor & former CEO Cadence Design, San Francisco, CA

2024 IEEE International Solid-State Circuits Conference (ISSCC) | 979-8-3503-0620-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISSCC49657.2024.10454565

1. Introduction

This paper provides a comprehensive overview of the future of semiconductor technology, focusing on the interplay between innovation and entrepreneurship. It discusses the semiconductor industry's current state, entrepreneurship's role in driving innovation, and potential future developments in this field. It highlights the virtuous cycle of innovation and entrepreneurship, where advancements in semiconductor technology fuel new business opportunities, driving further technological progress. This cycle is crucial for creating the next generation of intelligent electronic systems.

Moore's Law has been a driving force behind innovation in the electronics industry and various application market segments (Figure 1.4.1). Intel, as a key participant in the semiconductor industry, has been instrumental in delivering innovation that aligns with Moore's Law. This has paved the way for advancements in areas such as artificial intelligence, autonomous vehicles, cloud computing, and countless other technologies that shape our modern world.

The paper also explores the "stack" of semiconductor technology, from materials and design to manufacturing and generative AI applications. It establishes beneficial relationships between layers of the stack (Figure 1.4.2), with advancements in one area often enabling progress in others. This interconnectedness underscores the importance of a holistic approach to innovation in the semiconductor industry, where advances in different parts of the stack can collectively drive the industry forward.

2. Technologies for A Better World

In today's ever-evolving world, pursuing technologies that can contribute to a better world has become paramount. As we strive to address global challenges and enhance the well-being of individuals, the development and implementation of innovative technologies hold immense potential. In this paper, we will explore a range of technologies that have the power to shape a better world, examining their impact and potential applications.

We will focus on various technological advancements across different sectors, each seeking the path to positive impact (Figure 1.4.3). From sustainable energy solutions to groundbreaking healthcare innovations, from the concept of smart cities to advancements in education, we will delve into how these technologies can pave the way for a brighter future. By harnessing the power of cutting-edge technologies, we can tackle pressing issues such as climate change, healthcare accessibility, urbanization, and education inequality.

By exploring these technologies, we aim to shed light on their potential to transform our world for the better. We can achieve a more sustainable, inclusive, and prosperous future by understanding their capabilities and envisioning their applications. Join us as we delve into the technologies that promise a better world for all.

2.1 Sustainability

As the digital transformation gathers steam, semiconductors are not just components but the DNA of our digital age. We have moved beyond Moore's Law to enter an era of limitless potential enabled by silicon. Take, for example, the 'Net Zero' imperative [1] and the transition to electric vehicles; it is clear that semiconductors are the quintessential enablers of the electric vehicle revolution. Sustainability is not just an industry trend but a societal transformation; Tesla is not just building cars; they are building a sustainable future, empowered by semiconductors. The semiconductor industry is responding to a call for more sustainable investment and results for business and society.

2.2 Generative AI

Generative AI is a transformational technology that can redefine many sectors, from content creation and data analysis to EDA, healthcare, and beyond. By leveraging the capability to generate data, text, images, or even complex circuits, generative AI acts as a force multiplier for human creativity and decision-making. It enables the rapid prototyping of ideas, automates mundane tasks, and opens new avenues for personalized services.

Most intriguing is the potential of generative AI to solve complex problems, such as digital biology, that require synthesizing vast amounts of information, something traditional algorithmic approaches struggle with. As such, generative AI does not just augment existing workflows; it has the potential to create entirely new ones, driving innovation and offering solutions to longstanding challenges. Its role as a catalytic technology makes it not just an incremental improvement over existing systems but a leap forward in our ability to understand and manipulate the world around us.

2.3 Digital Biology

Digital biology revolutionizes our understanding of life sciences, bridging the gap between biological systems and computational models. By leveraging advanced technologies such as gene sequencing, machine learning, and data analytics, digital biology enables a more nuanced understanding of complex biological phenomena, from cellular processes to ecosystems. These technologies offer unprecedented opportunities in medicine, agriculture, and environmental science. For instance, in healthcare, biosimulation can aid in drug discovery, personalized treatments, and the monitoring of diseases in real-time. Beyond its practical applications, digital biology deepens our fundamental understanding of life, thus serving as a cornerstone for future scientific and technological advancements.

3. Generational Semiconductor Drivers

The industry benefits from an unprecedented combination of drivers: AI, Hyperscale Computing, 5G, Autonomous Vehicles, and Industrial IoT (Figure 1.4.4).

3.1 AI – The Computational Workhorse

We are not just improving algorithms but rethinking computer architecture from the ground up. The rising demand for workload-specific AI systems will dramatically transform system design. Traditional serial processing architecture, designed to be general-purpose, often needs more efficiency and speed to process the enormous amounts of data machine-learning models handle. As AI applications diversify from natural language processing to simulation and autonomous driving there will be a growing need for specialized hardware to optimize these workloads. This shift will lead to a new generation of processors that feature a blend of CPUs, GPUs, TPUs, and even custom ASICs, each tailored for particular AI tasks. By aligning hardware more closely with workloads, these architectural changes will likely result in significantly higher performance, reduced energy consumption, and more rapid advancements in AI capabilities. In essence, the specialization of computer architecture based on AI workloads promises to catalyze technological innovation and application-specific optimization.

3.2 Hyperscale Computing – The Backbone of the Digital World

In a hyperscale world, semiconductors are the spine underpinning the vast networks of servers and data centers that power today's digital world. In an environment where speed, reliability, and scalability are paramount, semiconductors ranging from CPUs and GPUs to specialized chips like TPUs and NPUs, as well as matrixed variants provide the computational horsepower necessary for real-time data analytics, machine learning, and other high-demand applications.

As the engine of the hyperscale architecture, processors enable cloud providers to deliver services at unprecedented scales, whether serving billions of search queries, streaming content globally, or providing the computational resources for cutting-edge scientific research. Furthermore, advancements in semiconductor technology, such as smaller transistor sizes and more energy-efficient designs, directly influence the capabilities and economics of hyperscale operations. Ultimately, semiconductors are not just components but the foundational element that makes the hyperscale world function efficiently and effectively.

3.3 5G – The Lifeline of Connectivity

5G is more than just an incremental upgrade over its predecessors; it represents a paradigm shift driven mainly by advancements in semiconductor technology. Unlike traditional mobile networks that primarily focus on voice and essential data services, 5G aims to unify the connectivity fabric for various applications, from IoT devices and autonomous vehicles to industrial automation and beyond.

The high-frequency, low-latency, and energy-efficient semiconductors at the heart of 5G networks make these ambitious goals achievable. These semiconductor innovations enable significantly faster data speeds and allow network slicing, edge computing, and massive machine-type communications. As a result, 5G is poised to be the lifeline of a more interconnected and real-time world, opening new possibilities for services and applications that were previously unimaginable. It is not just about faster smartphones; it is about enabling a unique ecosystem of connected technologies made possible by cutting-edge semiconductor innovations.

3.4 Autonomous Vehicles – Driving Exponential Growth for the Chip Industry
The expected growth of autonomous vehicles will significantly expand the semiconductor market in several ways.

a) Increased Semiconductor Content: Traditional cars typically include $350 - $400 in semiconductors. In contrast, by the end of the current decade, semiconductor content is projected to exceed 20% of the BOM value of premium vehicles [2].

b) Sensor Proliferation: Each autonomous vehicle is expected to have about 200-400 sensors for various functionalities like lidar, radar, and cameras. These sensors primarily rely on semiconductor technology.

c) Data Processing and Storage: Autonomous vehicles generate about four terabytes of data per hour [3], requiring advanced processors and high-capacity storage solutions.

d) Connectivity: 5G and vehicle-to-everything (V2X) communication are vital for real-time data transmission, increasing demand for communication chips.

e) Electric Vehicle Integration: Autonomous vehicles are often electric, requiring complex power management systems and additional semiconductor components like power MOSFETs, IGBTs, and more.

f) Mass Adoption: With an expected revenue CAGR of around 46% between 2023 and 2030 for the autonomous vehicles market [4], the sheer number of such vehicles on the road will necessitate massive quantities of semiconductors.

g) Consumer Electronics: Enhanced in-car entertainment and navigation systems, part of the autonomous vehicle experience, will require additional semiconductor components.

h) Safety and Security: The addition of cybersecurity features to protect autonomous vehicles from hacking will also increase semiconductor usage.

Given these factors, it is clear that the adoption and growth of autonomous vehicles will be a significant driver for the semiconductor market, contributing tens of billions of dollars in additional revenue.

3.5 The IoT Quilt
The Internet of Things (IoT) is fundamentally driven by innovations in silicon technology, serving as the bedrock of an interconnected ecosystem of devices. Miniaturized, power-efficient semiconductor chips enable IoT devices to perform complex tasks while operating in energy-constrained environments, a critical requirement for applications ranging from smart homes and wearables to industrial automation and healthcare monitoring. Advanced sensors, low-power microcontrollers, and specialized communication modules—all made possible through silicon innovation-equip IoT devices with the ability to collect data, process it locally or in the cloud, and interact with other devices in real-time.

The very scalability and affordability of IoT solutions are a direct consequence of advancements in silicon technology, which have reduced both size and cost while exponentially increasing performance. In essence, silicon innovation is the catalyst for transforming the promise of IoT from a conceptual vision into a transformative, ubiquitous reality.

4. System Design with Software 2.0
The importance of system design in a full-stack approach, known as Software 2.0 [5] (Figure 1.4.5), must be considered. Unlike traditional software development methods where explicit programming is the core, Software 2.0 leverages machine learning models as its centerpiece. This paradigm shift necessitates a new approach to system design that integrates not just the application and database layers but also data pipelines, model training, and inferencing components.

A well-designed system in this context optimizes model performance, scalability, and maintainability while ensuring seamless data flow and real-time analytics. The design phase is crucial in crafting the architecture and harmonizing traditional software elements with machine learning components, providing that the system can evolve without significant friction as technology advances. This holistic approach to system design is vital for the success of projects in the Software 2.0 era.

5. Workload Changes and Generative AI Applications
Generative AI applications are significantly altering the landscape of computational workloads, necessitating innovations and upgrades across multiple technological domains. In essence, the emergence of generative AI is a driving force behind technological advancements in semiconductors, networking, and storage and the rapid development of edge infrastructure (Figure 1.4.6). These innovations are crucial for meeting the evolving computational demands and unlocking the full potential of generative AI applications.

5.1 Semiconductor Innovations
The computational complexity of generative AI models requires specialized hardware for efficient processing. Traditional general-purpose CPUs often need to catch up in handling the specific needs of these applications. As a result, there is a surge in the development of specialized semiconductors to meet the high-throughput and low-latency requirements of generative AI. In addition, custom ASICs can handle specific tasks more efficiently in terms of speed and energy consumption.

5.2 Network Scaling and High-Performance Internet Switches
The data-intensive nature of generative AI applications places immense pressure on existing network infrastructures. Increased data transfer rates and low-latency communication are essential for the real-time processing and analysis that these applications often require. High-performance internet switches become critical to handle this upsurge in network traffic and ensure seamless data flow. Furthermore, the architecture might require network scaling to incorporate faster technologies like 5G to maintain real-time data exchange and analytics.

5.3 New Storage Innovations and Software Optimization
Traditional storage solutions like DRAM and SRAM are fast but come with limitations in terms of cost and capacity. As generative AI applications continue to grow in complexity and size, there is a pressing need for new storage solutions. Emerging technologies help address these limitations. On the software side, various optimization techniques are under development to make the most efficient use of available storage and computational resources.

5.4 Real-Time Analytics and Decision-Making
In addition to hardware applications, the need for real-time analytics and decision-making drives the rapid growth of edge infrastructure. Moving computation closer to the data source in edge computing setups reduces latency and conserves bandwidth, enabling generative AI applications to operate more effectively and efficiently.

The Rise of Purpose-Built Silicon
Silicon is experiencing a renaissance, where one size does not fit all, a trend that emerged as a response to the increasingly diverse and complex demands of modern computing. General-purpose processors, while versatile, often lack the efficiency required for specific tasks like machine learning, real-time analytics, and high-speed networking (Figure 1.4.7). As various industries adopt advanced technologies from AI and IoT to 5G and edge computing the need for chips that can perform specialized functions at high speeds and low power consumption has soared.

Purpose-built semiconductors like GPUs for graphics processing, TPUs for machine learning, and ASICs for custom applications offer optimized performance that general-purpose CPUs cannot match. These specialized chips can implement tasks more efficiently, reducing energy costs and latency while increasing throughput. This shift towards specialization is a strategic adaptation to a landscape where computational workloads are becoming more varied and specialized, requiring tailored hardware solutions for optimal performance.

7. Generative AI in Semiconductor Design
The emergence of Generative AI promised to transform many industries, and semiconductor design is already seeing the impact (Figure 1.4.8).

7.1 Design Flow Automation
The state-of-the-art chip design methodology involves a series of transformation, analysis, and optimization steps. Over time, the pressure on performance, the convergence of physics with advanced nodes, and the need to co-design to reduce schedules have introduced loops in the design flow. The number of iterations to stabilize and optimize the design continues to grow. And the demand on the engineering team is becoming untenable.

By adopting AI-driven chip design and implementation tools, it is possible for engineers to concurrently optimize the flow for multiple blocks, which is especially important for the large, complex system-on-chip (SoC) designs needed for today's ever-more powerful electronic systems. Additionally, full-flow reinforcement learning technology significantly improves the engineering team productivity.

An example of an AI tool for chip design is the Cadence® Cerebrus™ Intelligent Chip Explorer. It uniquely automates the flow to optimize performance, power, and area (PPA) of the formerly manually guided tools and optimizations, uses reinforcement learning to make decisions, and incorporates trained models for learning from previous runs.

7.2 Package Design

IC packaging was much simpler ten years ago, so empirical models sufficed, but today's multi-chip models require electrical models. Previously, IC designers did not have to worry much about the package or the PCB. Today, particularly with analog and RF designs, if an IC designer needs to account for what is happening on the PCB or package, chances are that the chip will not work. Innovation in packaging technology has accelerated with the emergence of chiplets and 2.5D/3D stacking (Figure 1.4.9).

AI-driven package layout tools can now reduce the time to produce IC packaging while optimizing signal integrity. By streamlining and automating the flow between the chip, package, and PCB, IC designers gain confidence that their chip will work once placed in the package and then on the PCB before building the design.

An example of an AI-driven package design tool is the Cadence Virtuoso® Studio. It can design the package with the IC design data and consider the PCB requirements to create an optimal solution.

7.3 Verification

Digital simulation has been the workhorse of logic verification for decades. Randomized testbench stimulus has elevated verification effectiveness and created a new category of employee, the verification engineer. The digital verification norms are large-scale regression suites, massive server farms or cloud deployments, and regression management technologies.

AI-driven verification and debugging represent a generational shift from single-run, single-engine simulation regression algorithms to new technology that leverages big data and AI across multiple runs of multiple engines throughout an entire SoC verification campaign. The technology can optimize verification workloads, boost coverage, and accelerate root cause analysis of bugs. Large language model (LLM) technology reduces the time to create clean designs at the start of verification projects.

An example of an AI-driven verification solution is the Cadence Verisium™ AI-Driven Platform. It offers unique capabilities for automating regression triage, mining waveforms for a bug root cause, analyzing reports and log files to determine bug causality in source code check-ins, and analyzing source code changes to rank for likely system behavior disruption.

7.4 PCB

PCB design has long been a manually assisted process, using routing technology for wire placement. Even as PCB complexity grew, the existing technology was used to grind out designs, sometimes running for days.

AI-driven PCB design technology is a next-generation system design technology that offers revolutionary improvements in performance and automation. This generative AI technology reduces placement and routing (P&R) tasks from days to minutes, with equivalent or higher quality than manually designed boards. The technology automates component placement, power plane creation, and routing critical nets.

An example of an AI tool for PCB design is the Cadences Allegro® X AI system design technology. It uses AI to find optimal routing, often an order-of-magnitude faster, enabling exploration of the solution space for a higher quality of results and creating a better starting layout while significantly enhancing the productivity of the PCB designers.

7.5 System Design

Parametric sweeps consume large amounts of analysis time and computational resources. Design engineers need an intelligent, accurate, easy-to-use simulation and analysis solution that reduces repetitive design cycles while increasing user productivity and efficiency.

AI-driven system optimization technology can explore a broader state space, looking beyond local minima and maxima, and approach optimization more efficiently than a brute-force parametric study.

An example of this technology is the Cadence Optimality™ Intelligent System Explorer. It uniquely optimizes system parameters for 3D electromagnetic analysis, high-speed signal and power integrity.

8. Semiconductor Market Growth—The Path to $1T

Forget billions; we aim for a trillion-dollar semiconductor industry revenue by 2030 [6]. The chip market is on an unprecedented growth trajectory, driven by a confluence of factors that extend beyond traditional computing needs. The advent of 5G, IoT, AI, and autonomous vehicles, among other trends, has created a burgeoning demand for advanced, specialized semiconductor chips. Furthermore, as digital transformation sweeps across sectors like healthcare, finance, and manufacturing, the requirement for high-performance, energy-efficient, and secure semiconductors continues to escalate. While classic Moore's Law is slowing down, new "More than Moore" technologies like 2.5D/3D packaging are paving the way for giant performance and manufacturing efficiency leaps. Supply chain dynamics are also evolving, with increasing emphasis on localized production and secure sourcing, potentially driving the value up. Financial investments in research and development (R&D) and semiconductor fabrication facilities are accelerating, signaling strong market confidence. In summary, the path to a $1 trillion valuation by 2030 [6] (Figure 1.4.10) appears plausible and likely, underpinned by technological innovations, diverse application needs, and shifting global economic landscapes.

9. Call to Action

We stand at a pivotal moment in history where our collective innovation has the power to reshape the very fabric of society. From AI and 5G to healthcare and renewable energy, semiconductors are the invisible force driving progress. Yet, to sustain this momentum and push the boundaries of what is possible, we must double down on fostering innovation. Innovation requires more than iterating on existing technologies. It will require investing in groundbreaking research that opens doors to new markets and applications. We must encourage a culture of continuous learning and risk-taking, accelerating the celebration of creativity and breakthroughs.

Equally critical is the urgent need to address our industry's skills shortage. Talent is the lifeblood of innovation, and as our field evolves, so must our workforce. We need to collaborate closely with educational institutions to develop curricula that prepare students for the challenges of tomorrow. Industry internships, apprenticeships, and mentorship programs can bridge academic learning and real-world application. By investing in the next generation of engineers, designers, and thinkers, we are investing in our industry's long-term health and competitiveness.

Lastly, we must ramp up private and public research and development investments. While individual companies play a crucial role, this collective endeavor benefits from government involvement. Public-private partnerships can catalyze research, reduce risks, and quickly bring transformative technologies to market. In a world where geopolitical dynamics and global supply chain complexities increasingly impact our industry, we must stay caught up in research and development. Through sustained, collaborative investment, we will continue to lead in innovation, create high-value jobs, and shape the future. Let us seize this moment together and affirm our commitment to an industry that does not just change the rules but writes them.

References:

[1] "imec.netzero," netzero.imec-int.com. https://netzero.imec-int.com/.
[2] "Intel CEO Predicts Chips Will Be More Than 20% of Premium Vehicle BOM by 2030," www.businesswire.com, Sep. 07, 2021. https://www.businesswire.com/news/home/20210907005345/en/Intel-CEO-Predicts-Chips-Will-Be-More-Than-20-of-Premium-Vehicle-BOM-by-2030
[3] AutoDriveAI, "Autonomous Cars Will Collect Approximately 4 TB of Data Every Hour of Driving," Medium, Aug. 11, 2020. https://medium.com/@autodriveai/autonomous-cars-will-collect-approximately-4-tb-of-data-every-hour-of-driving-3819aba33204
[4] "Autonomous Vehicle Market Size, Share | Industry Report, 2030," www.grandviewresearch.com. https://www.grandviewresearch.com/industry-analysis/autonomous-vehicles-market
[5] A. Karpathy, "Software 2.0," Medium, Mar. 13, 2021. https://karpathy.medium.com/software-2-0-a64152b37c35
[6] O. Burkacky, J. Dragon, and N. Lehmann, "The semiconductor decade: A trillion-dollar industry," www.mckinsey.com, Apr. 01, 2022. https://www.mckinsey.com/industries/semiconductors/our-insights/the-semiconductor-decade-a-trillion-dollar-industry

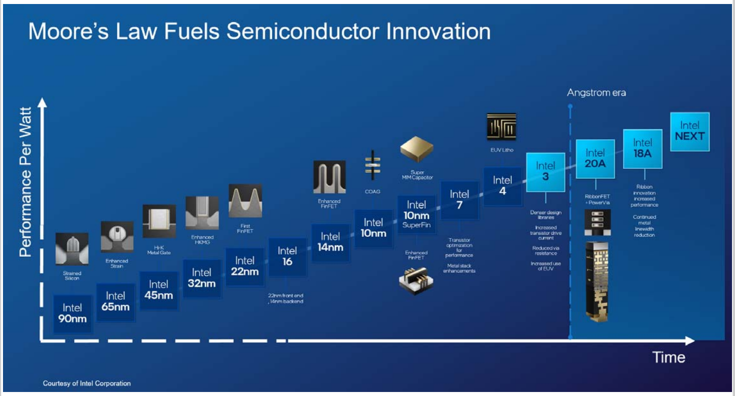

Figure 1.4.1: Moore's Law has been a driving force behind innovation.

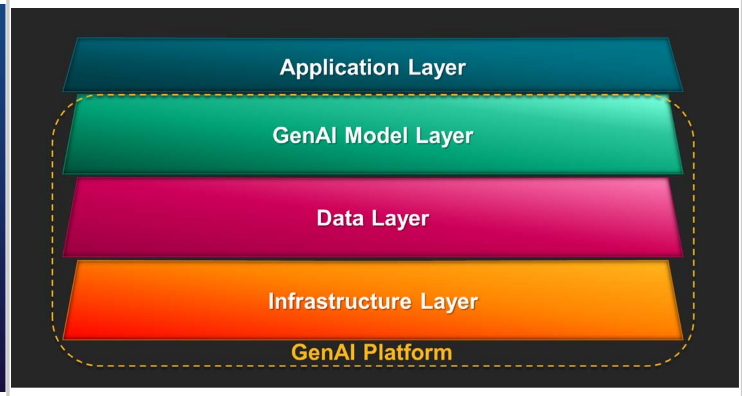

Figure 1.4.2: Semiconductor Technology Stack.

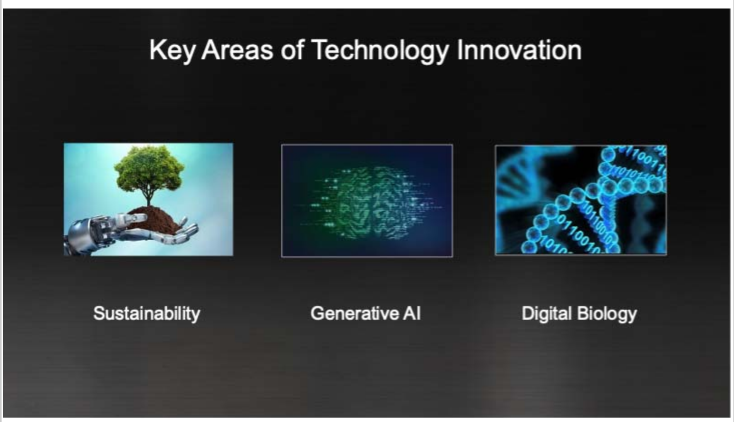

Figure 1.4.3: Key areas of technology innovation for a better world.

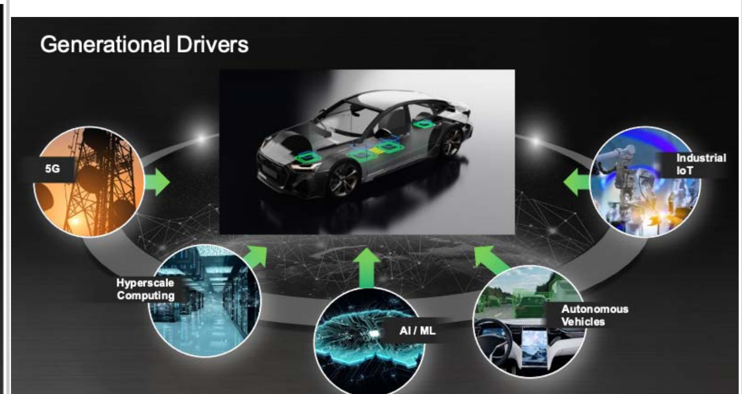

Figure 1.4.4: Generational drivers of the semiconductor industry.

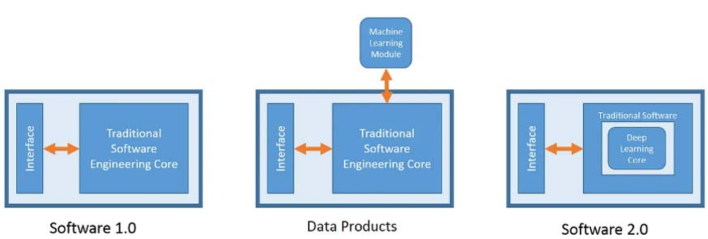

Figure 1.4.5: Software 2.0 (image courtesy of Ahmad Mustapha, Medium, Feb 2021).

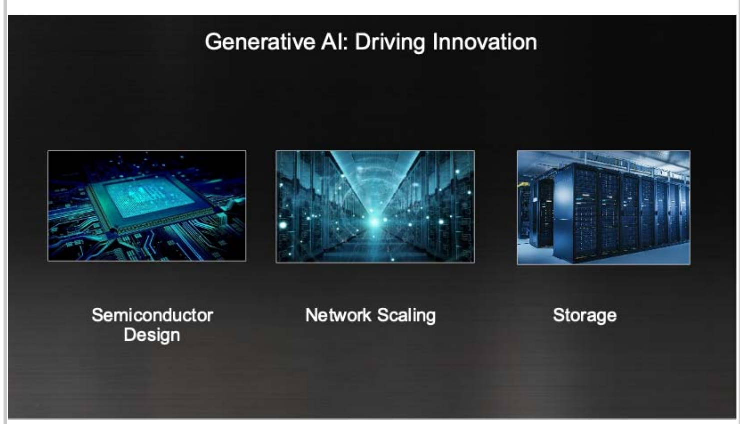

Figure 1.4.6: Generative AI is driving innovation.

979-8-3503-0620-0/24/$31.00 ©2024 IEEE

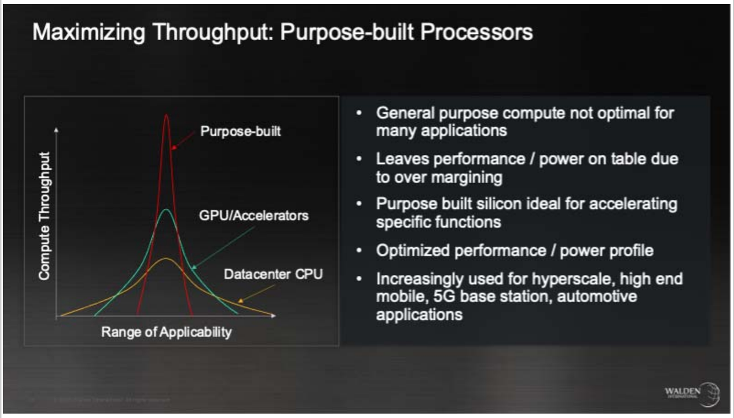

Figure 1.4.7: Purpose-built compute.

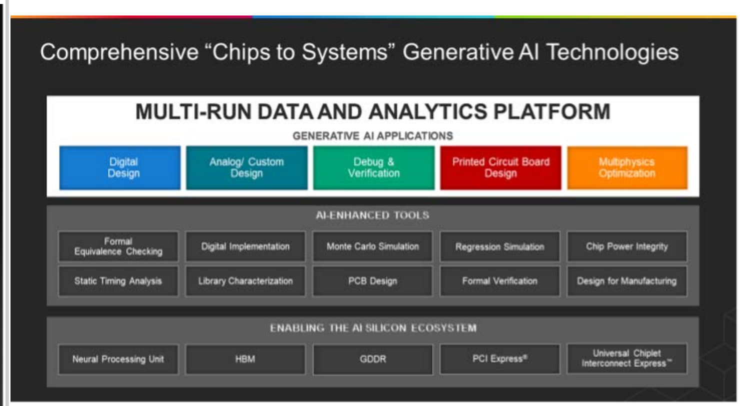

Figure 1.4.8: Generative AI for semiconductor design.

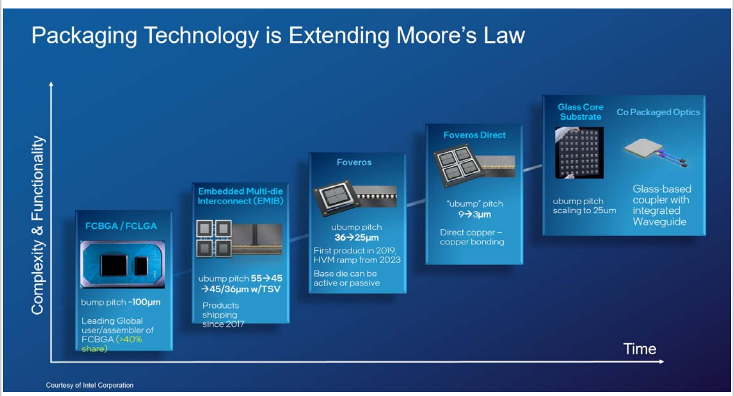

Figure 1.4.9: Packaging Technology Development.

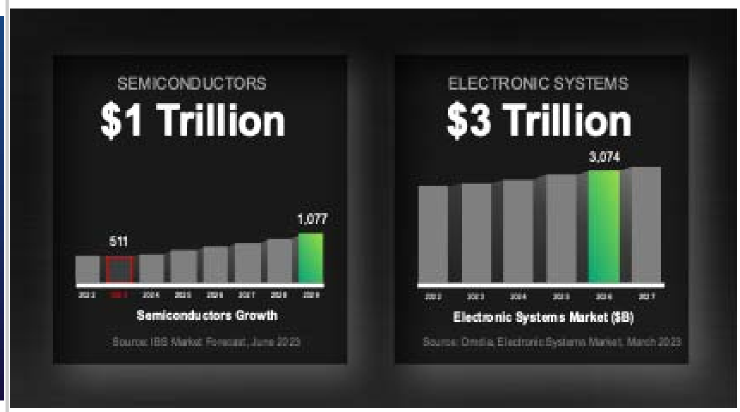

Figure 1.4.10: Semiconductor and electronics systems revenue projections.