The Integrated Circuit Industry at a Crossroads: Threats and Opportunities

Salvatore Pennisi

Dipartimento di Ingegneria Elettrica, Elettronica e Informatica (DIEEI)—University of Catania, 95125 Catania, Italy; salvatore.pennisi@unict.it

Abstract: With the outbreak of the COVID-19 pandemic, the persistent chip shortage, war in Ukraine, and U.S.–China tensions, the semiconductor industry is at a critical stage. Only if it is capable of major changes, will it be able to sustain itself and continue to provide solutions for ongoing exponential technology growth. However, the war has undermined, perhaps definitively, a global order that urged the integration of markets above geopolitical divergences. Now that the trend seems to be reversed, the extent to which the costs of this commercial and technological decoupling can be absorbed and legitimized will have to be understood.

Keywords: semiconductors; integrated circuits; fabs; shortage; talent

1. Introduction

The integrated circuit (IC) industry forms the basis of the overwhelming digitalization process, i.e., the most important enabling technology for current and future applications. This has been made possible by the tremendous miniaturization and performance improvement of IC processes—predicted by Moore’s Law—which, starting from about $10^3$ transistors on the first Intel 4004 microprocessor in 1970, reached $10^{11}$ transistors in March 2022 (Apple M1 Ultra) [1]—an unprecedented and unsurpassed rate of improvement, which has enabled, among other inventions, the Internet, mobile telecommunications, and now smart cars. In brief, every industry into which ICs (microchips or simply chips) have been introduced has benefited from greater efficiency, intelligence, and extended functions. Due to this success, chips are today the fourth-most traded product globally (1.15 million semiconductor units shipped last year, 2021)—after crude oil, motor vehicles and their components, and refined oil—in a market that was valued at 0.6 trillion dollars in 2021, with a 26% increase in year-on-year sales, and which is expected to reach 1 trillion dollars in 2035 [2].

Some analysts have gone so far as to call chips the new oil, in that chips ‘power’ applications, by giving the country that is able to produce the highest-performing chips—thanks to cutting-edge technology—greater power than other countries, in terms of computing and communication capabilities, but also from a purely military point of view. One concept that the Russia–Ukraine war has underscored so far is that Ukrainian forces have used small and relatively inexpensive weapons, such as the Javelin and Stinger anti-aircraft missiles, which adopt advanced semiconductors for guidance systems. A single Javelin contains about 250 chips [3]. Western countries have banned the export of semiconductors to Russia, and Russia does not have its own advanced chip production capacity; without imports the Russian military cannot supply itself with precision-guided munitions.

Quite surprisingly, the key role of chips in global economies has only recently been recognized by governments and occupied public debate. In recent decades, global economies have focused more on software and tertiary services, leaving chips as a pure...
commodity. However, the COVID-19 pandemic and the war in Ukraine have highlighted the problem of chip shortage (insufficient production of chips relative to demand), the fragility of the semiconductor supply chain, and the fact that chips are strategic components. As a result, the goal of many governments is currently to strengthen their resilience to external shocks, and to safeguard their technological sovereignty by strongly supporting the integrated circuit industry. Chipmakers, on the other hand, while improving the quality and number of their fabrication plants (fabs) to satisfy the increased demand due to the exponential explosion of applications, have better understood their responsibilities as well as their newly strengthened position of power, which in principle allows them to select their customers, and to determine who can and who cannot get their chips. However, the war in Ukraine and trade tensions between the U.S. and China have made the scenario more complex, deteriorating a global order that preached the integration of markets above geopolitical divergences, and resulting in the end of the age of globalization.

Finally, a difficult problem that can undermine any semiconductor strengthening policy is the shortage of talent. While new fabs can be built in a couple of years with the availability of adequate resources from private/public funding, qualified personnel cannot be found simply by putting up money. It takes many years to train qualified professionals, but first there must be people (young people) willing to invest their future employment in this field. Semiconductor engineers (and, at the top of the list, analog design engineers), are in high demand today, and their scarcity is likely to increase in percentage terms, because the younger generation is less and less interested in hardware, while the number of applications that use electronic components, and thus require hardware skills, is growing.

This paper follows another recent complementary publication by the same author [4], and further analyzes the semiconductor ecosystem, in light of the rapidly changing scenarios. Section 2 describes the semiconductor supply chain and the types of companies associated with making integrated circuits. Section 3 analyzes the key weaknesses and bottlenecks in the chain. Section 4 elaborates on geopolitical and socioeconomic considerations. Conclusions are drawn in Section 5.

2. IC Market, Supply Chain, and Types of Semiconductor Companies

ICs are the major enablers of current and future technologies and applications, such as 5G/6G, smart factories and cars, blockchains, artificial intelligence (AI), and machine learning. The semiconductor supply chain industry makes all this possible. This chain can be broken down into six main stages, which take place in different parts of the world, and involve thousands of companies and millions of people. For the purpose of the following analysis, these stages, and the types of companies, will be summarized in the following subsections. Before doing so, however, let us briefly mention the different segments that IC production can be divided into: Logic; Memory; Analog; MPU (microprocessor unit); MCU (microcontroller unit); Optoelectronics; Sensor/Actuators; Discretes; and DSP (digital signal processor). See Figure 1 for global IC sales in 2021 (in billions of dollars), as reported by the U.S. Semiconductor Industry Association, SIA [2]. In this framework—not specified in the figure—Graphic Processing Units (GPUs) alone have a market of $23.90 billion. Driven by power ICs, discrete semiconductors have had a big boost, as they were previously valued at $23.8 billion in 2020. Memory and logic devices are expected to experience the highest growth rate in the coming years, followed by the analog ICs needed for data conversion, emerging automotive applications, and power management, and by microcontrollers and sensors, due to high-performance IoT applications.
In order of market share, IC applications can be divided into: Communications; Computer; Consumer; (these three C-segments will continue to grow, due to the demand for smartphones and connected devices, as well as games, wearable devices, and the development of the *metaverse*); Automotive (which will grow, due to the demand for electric vehicles with assisted/autonomous driving); Industrial (expected to grow steadily, due to the necessary adaptation of production machines required by the fourth industrial revolution); and Government (which shows limited growth). Figure 2 shows the market share of the above segments for 2021.

2.1. *The Six Phases of the Semiconductor Supply Chain*

(a) **Raw materials and wafer fabrication.** A variety of raw materials are used to make an integrated circuit, ranging in price and availability, from abundant silicon, through more than 100 gases, fluids, photomasks, reagents, etc., to expensive gold and rare earth elements (REEs). At this early stage of manufacturing, ingots are formed from...
pure silicon, and cut into wafers, the size of which has gradually increased over the decades, to improve productivity and reduce costs. The current state of the art uses wafers 300 mm in diameter and 775 mm thick. Over the past two decades, the silicon wafer industry has gone from more than 20 suppliers in the 1990s, to a handful of companies today. As Figure 3 illustrates, Japan’s Shin-Etsu and Sumco are the world’s largest producers of silicon wafers, followed by Taiwan’s GlobalWafers, Germany’s Siltronic, Korea’s SK Siltron, and France’s Soitec [5].

![Figure 3. Principal silicon wafer producers and market share in 2021.](image)

(b) **Design.** The typical design phase of a digital IC includes architectural or system-level design, logic design, circuit design, functional safety, physical design, post-design verification and, finally, preparation of photolithographic masks for the next stage of manufacturing. All these steps are supported by highly sophisticated computer-aided design (CAD) or electronic design automation (EDA) tools, which provide integrated simulation environments and automation, with optimization capabilities to meet IC design specifications in terms of performance, power consumption, area, etc. Digital designs take advantage of more scaled-up technology nodes, while analog and automotive applications adopt more mature, reliable, and robust nodes.

The global EDA market is monopolized by three major companies, as shown in Figure 4: Synopsys and Cadence, from the U.S., and Siemens EDA, from Germany (which acquired U.S. Mentor Graphics in 2017). Each company’s portfolio is very rich, but each has its own peculiarities. Synopsys focuses on digital chip design, static timing verification and confirmation, and System in Package support, neglecting complete process tools. Cadence focuses on analog and mixed-signal platforms and digital back-ends. Siemens EDA focuses on back-end verification, testability design, and optical proximity correction.
Closing verified digital designs, using EDA tools, requires less and less engineering effort. In contrast, the availability of well-trained and experienced circuit designers is a more crucial issue for high-performance analog circuit design, where human knowledge is still mandatory.

(c) **Front-End Fabrication.** Identical integrated circuits (each called a *die*), are fabricated on each wafer in a multistep process, using various techniques and materials (e.g., etching, photolithography, material deposition). Some of the most complicated (and expensive) machines on the planet are used in this step. The global production capacity of integrated circuits, by location of production facilities, is shown in Figure 5, while the global percentage capacity of integrated circuits by technology node is summarized in Table 1. All data refer to 2021.

Figure 5 shows that South Korea, Taiwan, and Japan account for about 60% of global production capacity [6]. China is expanding rapidly, because the cost of building and operating a plant is lower than in any other nation, but currently about half of its wafer capacity is controlled by foreign companies (Korea’s SK Hynix and Samsung, and Taiwan’s TSMC and UMC) and it does not lead the volume fabrication of advanced nodes, as Table 1 shows.
Table 1. Global percentage IC capacity by technology node in 2021 [7].

<table>
<thead>
<tr>
<th>Region</th>
<th>&lt;10 nm</th>
<th>10–22 nm</th>
<th>28–45 nm</th>
<th>&gt;45nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>12%</td>
<td>4%</td>
<td>6%</td>
<td></td>
</tr>
<tr>
<td>Americas</td>
<td>43%</td>
<td>6%</td>
<td>9%</td>
<td></td>
</tr>
<tr>
<td>Japan</td>
<td>3%</td>
<td>5%</td>
<td>13%</td>
<td></td>
</tr>
<tr>
<td>China</td>
<td>92%</td>
<td>28%</td>
<td>47%</td>
<td>31%</td>
</tr>
<tr>
<td>Taiwan</td>
<td>8%</td>
<td>5%</td>
<td>6%</td>
<td>10%</td>
</tr>
<tr>
<td>South Korea</td>
<td>9%</td>
<td>13%</td>
<td>8%</td>
<td></td>
</tr>
</tbody>
</table>

The front-end manufacturing process is the most capital-intensive. Most of the factory’s construction costs are semiconductor manufacturing equipment, with some parts costing more than $100 million each. ASML of The Netherlands is a world leader in the production of advanced photolithography systems (Deep Ultraviolet Lithography, DUVL), and is the only company to have developed the next-generation technique needed for leading-edge nodes, namely Extreme Ultraviolet Lithography (EUVL). To understand the effort required to make EUVL possible, one only has to consider that the major foundries (TSMC, Intel, and Samsung, see Section 2.2) had to invest in ASML for the necessary financial capacity. Each year, ASML is only able to build a few EUVL machines (31 in 2020), because of their complexity. ASML is developing the next High NA (numerical aperture) machine, to be available for early access from 2023.

As a result, building a new semiconductor factory at an advanced node (5 nm) can cost up to $20 billion, and the cost of designing a new chip (tapeout) is more than $500 million. Table 2 compares the price of a processed wafer, the average cost of designing a chip, and the days of work required for different nodes. Exponential growth in cost and labor is observed below 16 nm [8].

High capital costs create barriers to entry or even to staying. We saw that in 2001 the state-of-the-art in processes was 130 nm, and that 26 companies were capable of producing that technology (see Figure 6), but that only 2 companies (TSMC and Samsung) were capable of producing 5-nm processes in 2020. In 2022, Samsung began production of 3-nm nanosheet GAA (Gate All Around) technology [9], and it will be soon followed by TSMC [10], which confirmed that mass production of 2-nm process would start in 2025 [11].

Despite Intel’s aggressive roadmap to recover lost positions, leading technologies are dominated by the two Asian companies. In 2021, TSMC produced 92% of global logic ICs [12].

Table 2. Wafer cost, chip cost, and labor requirements for different technology nodes [8].

<table>
<thead>
<tr>
<th>Technology Node (nm)</th>
<th>Foundry Sale Price per Wafer ($)</th>
<th>Average chip Design Cost (Million $)</th>
<th>Person Days (×1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>1937</td>
<td>28</td>
<td>45</td>
</tr>
<tr>
<td>40</td>
<td>2274</td>
<td>36</td>
<td>58</td>
</tr>
<tr>
<td>28</td>
<td>2891</td>
<td>40</td>
<td>80</td>
</tr>
<tr>
<td>22</td>
<td>3677</td>
<td>69</td>
<td>110</td>
</tr>
<tr>
<td>16</td>
<td>3984</td>
<td>104</td>
<td>166</td>
</tr>
<tr>
<td>10</td>
<td>5992</td>
<td>174</td>
<td>278</td>
</tr>
<tr>
<td>7</td>
<td>9346</td>
<td>297</td>
<td>475</td>
</tr>
<tr>
<td>5</td>
<td>16,988</td>
<td>540</td>
<td>864</td>
</tr>
</tbody>
</table>
(d) **Back-End Fabrication** (Assembly, Test, and Packaging—ATP). Several steps are required to obtain a finished chip. Firstly, an Optical/E-beam inspection of the wafer is conducted, to identify defects—such as metal shorts (up to 10 nm and 3 nm, respectively)—and eventually repair them. Secondly, a Wafer Probe/Test, which is the first time that the chips are tested to see if they work as designed: highly accurate non-destructive measurement of a test element group (TEG)—including transistors, interconnects and other devices—is conducted through a probe board that interfaces the wafer and the test equipment.

Thirdly, Wafer Dicing (individual cuts are made on the wafers). Fourthly, Die Bond (the process of attaching the bare die to a substrate, which provides electrical connectivity to the outside, and to the base of the package). Fifthly, Wire Bond/Solder Bump, in which each die pad is connected to a corresponding pad on the substrate via a thin gold wire, or through flip chip technology. Sixthly, encapsulation, when the die is packed. Finally, testing, to detect defects that may have occurred during the assembly process; an integrated circuit socket is used in the final testing, which plays the crucial role of connecting the device to the tester, similar to a probe board in wafer testing. Each individual integrated circuit must be tested with a custom test socket.

Automated Test Equipment (ATE) is a computerized machine that uses test tools to perform and evaluate the results of functionality, performance, quality, and stress tests performed on integrated circuits. The ATE requires minimal human interaction, and is directly responsible for ensuring not only that the IC functions as intended, but also that the IC does not cause hazards as a result of its use. The growth of the ATE market is driven by the significant use of test equipment in the automotive industry, which utilizes, among other things, microcontrollers, sensors, and radar chips, and which has ‘Zero Defects’ goals, to ensure very high levels of reliability and safety in automobiles. This means that integrated circuit manufacturers are shifting their specifications from defects per million to defects per billion (DPB) [13]. Teradyne (U.S.), LTX–Credence–Xcerra (U.S.), and Advantest (Japan) hold the majority of the ATE global market share. The probe cards market is dominated by FormFactor (U.S.), Technoprobe (Italy), and Micronics (Japan).

Semiconductor packaging technology has evolved to minimize costs and improve the overall performance of integrated circuits (counteracting heating, mechanical damage, radio frequency noise emission, electrostatic discharge, etc.), while providing higher speeds, smaller footprints, higher pin counts, and lower profiles. The semiconductor packaging market is under constant pressure to provide innovative solutions, in terms of size, performance, and ‘time-to-market’. The market is moderately fragmented, with no dominant players, but is mainly concentrated in East Asia, due to lower labor costs. Taiwanese companies hold about 50% of the global market for Outsourced Semicon-
ductor Assembly and Testing services (OSAT). Taiwan’s ASE is the world’s largest supplier, with more than 24% of the market, followed by Amkor (U.S.), JCET (China), and SPIL (Taiwan). These companies hold about 70% of the market share. In addition, Malaysia accounts for more than 10% of the global packaging trade.

Initially, OSAT companies required significantly less investment in plant, equipment, EDA tools, and R&D than foundries and IDMs, but their profit margins were also lower. This picture changed substantially with the advent of the system-in-a-package approach, particularly 3D Flip Chip. InFO (Integrated Fan-Out) packaging is a wafer-level system-integration technology platform, featuring high-density RDL (Redistribution Layers) and TIV (Through InFO Via) for high-density interconnect, which requires complex and expensive processes. In this framework, chip-on-wafer-on-substrate (CoWoS) technology and chiplets will strongly influence the advanced development of high-performance computing [14].

(e) **Product manufacturing.** Finished chips are sent or sold to electronics manufacturers, or other types of manufacturers, and are incorporated into products. In the past, most large-scale electronics manufacturing was handled by in-house assembly. The division of labor in the electronics industry has led to the emergence of electronics manufacturing services (EMS) or electronics contract manufacturing (ECM) companies. These new companies offer flexibility and large economies of scale in manufacturing, raw material sourcing and resource sharing, industrial design expertise, and the creation of value-added services, such as warranty and repair. This market is dominated by Taiwanese companies Foxconn (Hon Hai), Pegatron, and Wistron (all three being contract manufacturers for Apple). Foxconn—which also produces products for Amazon, Cisco, Dell, Nintendo, Nokia, Acer, Xiaomi, etc.—has 12 plants in China, where it is the largest private employer, with about 1.3 million employees [15].

At this final stage, printed circuit boards (PCBs) are also needed in almost all electronic products, to secure integrated circuits in specific locations, and to provide reliable electrical connections between component terminals. PCBs can be produced in-house by many large companies, or they can be outsourced.

(f) **Sales.** IC components, as well as final products with IC content, are sold to consumers.

### 2.2. Types of Semiconductor Companies

In the past, semiconductor companies’ production facilities were mostly in-house: that is, almost the entire process, from research and design to assembly and testing. In the early 2000s, profit margins were low at semiconductor companies, with most generating returns below the cost of capital; therefore, due to financial and time-to-market constraints, IC manufacturing companies began to outsource segments of their manufacturing operations to subcontractors. Today, we can find companies that design integrated circuits, and may or may not produce their own chips, and companies that produce chips but may or may not design them. All these companies can be identified primarily in fabless, IDM, and foundry, as specified below.

An **Integrated Device Manufacturer (IDM)** carries out chip design, fabrication, and ATP in-house. IDMs include Intel (whose CEO, Pat Gelsinger, recently shared his IDM2.0 vision for the company [16]), IBM, Samsung, NEC, SK Hynix, Micron, Texas Instruments, Toshiba, Sony, STMicroelectronics, NXP, and Onsemi. IDMs can also provide contract fabrication services for other firms, or can a outsource consistent part of their production cycles to ‘pure-play foundries’, or simply foundries, like TSMC, Samsung Foundry, UMC, GlobalFoundries, and SMIC.

A **fabless** semiconductor company, on the other hand, focuses exclusively on chip design, and outsources the various manufacturing steps to foundries and IDMs (to produce the designed chips), to OSAT (to assemble, package, and test the chips), and to EMS companies (to integrate the packaged chips into devices). Examples of fabless companies include Broadcom, Qualcomm, AMD, Media Tek, Nvidia, and Xilinx. Fabless semicon-
ductor companies need less capital, and have generally higher and less volatile profit margins than IDMs, but quality control and ensuring on-time production can be an issue for them. Between IDMs and fabless, a fab-lite semiconductor manufacturing model allows in-house production targeted only at specific low-cost technology nodes that are still in high demand.

In this list, we can also include large technology companies that have the economic ability and convenience to design their own chips in-house for their specific applications, for competitive differentiation, preventing replication and ensuring consistency across different devices [17]. For example, Apple develops custom chips for the iPhone and iPad, Facebook (now Meta) designs chips optimized for the types of content it stores and processes on its servers, Amazon’s Graviton and Inferentia and Google’s (now Alphabet) Tensor Processing Unit (TPU) are AI-based IC accelerators for cloud computing, and Tesla has developed the D1 Dojo Chip to train AI models. It may sound surprising, but Apple can be considered the third largest fabless player in the world, behind Broadcom and Qualcomm. Moreover, besides Tesla, many automakers are collocating semiconductor engineers to develop new chips. They are understood to be part of the semiconductor industry, as the average IC content per vehicle will exceed $1000 by 2026 [18].

IC designers often rely on other companies (sometimes referred to as design houses) for IP cores, which are reusable units of logic design, cells, or IC layout (software) that are the ‘intellectual property’ of one party, and can be licensed to another party. This is especially true for start-ups that, due to limited resources, focus their efforts on a specific design with unique features, while referring to IP cores for standard functions. IP cores include CPUs, GPUs, embedded memory compilers, interface, and interconnect technologies. The semiconductor IP market is dominated by three companies that cover more than two-thirds of the market: ARM (UK-based), with a share of more than 40%, Synopsys (U.S.), and Cadence (U.S.).

Figure 7 summarizes the types of companies that form the complete IC ecosystem. Research is not explicitly included in this diagram, because in almost all cases it comes from specific departments in the companies, and from government-funded institutions and academic laboratories. No other industry has the same intensity of R&D: about 22% of annual sales ($90 billion), compared to the 21% of the pharmaceutical and biotechnology sector.

![Figure 7. The ecosystem of semiconductor companies.](image-url)
2.3. Remarks

In recent years, application domains have evolved from divergent, in which less sophisticated products with specific and minimal functions have been developed (e.g., cell phones or cars used only for communication and transportation, respectively), to convergent, in which the technological convergence of data processing, telecommunications, and energy management is fully exploited (e.g., in smartphones or smart cars). In parallel, the semiconductor industry’s value chain has shifted from a vertical integration model, in which four of the six stages of the chain discussed earlier (with the exception of wafer and product fabrication) were carried out in-house and integrated, to a vertical disintegration model, in which specialized companies have emerged for each stage. This business model was motivated by the rapid advancement of technology, with the continuous reduction in chip size and the diversification of device features through special processes, which led to an exponential increase in the design complexity and manufacturing costs of integrated circuits. This business model has also benefited from the global economy, which has required the integration of markets above geopolitical divergences.

Both the converging application domains and the vertical disintegration model promote supply chain disruption. In fact, a system that incorporates an increasing number of chips is increasingly exposed to a shortage, because it cannot be completed for the lack of even one chip. In contrast, a long, often monopolistic, globally dispersed supply chain can be easily disrupted by a single social or geopolitical event. All these aspects are explored in the following.

3. Supply Chain Bottlenecks, Global Shortages, and Counteracting Measures

It can be understood from Section 2.3 that the semiconductor supply chain is extremely fragile, and has several bottlenecks that can facilitate disruptions. In addition, the 2020 COVID-19 pandemic and the U.S.–China trade war highlighted and accelerated many of these problems, quickly causing a chip shortage that persists to this day. The shortage, along with economic and military conflicts, has convinced companies and governments to take action. The bottlenecks, the causes of the silicon shortage, and some industry and government responses are discussed below.

3.1. Bottlenecks

We have seen that there are only a few companies, or even one, that dominate the entire global market. Samsung and Intel lead the overall semiconductor market, TSMC the foundry sector for cutting-edge nodes (<10 nm), Qualcomm and Nvidia the fabless sector, ASML the manufacturing of EUVL machinery, ARM the design of IP cores, FormFactor and Technoprobe the probe card market, Foxconn the manufacturing of products, etc. In addition, geographic regions, or even a single country, have specialized in the production of certain raw materials (for example, in 2021, China supplied more than 85% of the world’s refined Rare Earth Elements, followed by the rest of Asia at 13%, and Europe at 2%), or in specific manufacturing processes and technologies. Due to the lockdown of factories, many OSAT companies accumulated orders. The average lead time for packaging was 8 weeks, pre-COVID; now it is 20 weeks (and can even be 50 weeks for prototypes).

In general, Figure 1 shows that East Asia supplies more than 75% of global IC production capacity (led by Samsung, TSMC, and SK Hynix [2]), and is the hub of semiconductor manufacturing, including ATEs, wafers, and IC substrates. This dominance is expected to grow, especially with increased investment in China by foreign and domestic companies, and government engagement. The U.S. has lost its primacy in IC manufacturing (from 37% in 1990 to 12% today), and also, partially, in leading research (European R&D front-runners in semiconductor technologies are IMEC, CEA-Leti, and the Fraunhofer Institute-FMD), but the U.S. still leads global chip sales, as the last available data of 2021 from SIA show [19], and as summarized in Figure 8. The success of the U.S. is due to
its engineering workforce, even if the U.S. position has declined slightly in recent years, with a parallel increase in China (whose global chip sales figures are close to those of Taiwan, and also to Europe and Japan).

Figure 8. Global chip sales in 2021.

3.2. The Pandemic and Other Calamitous Events

The vulnerability of the supply chain became clear during the COVID-19 pandemic of 2020 and the two years that followed, when cities or entire countries shut down or severely slowed production. China’s ‘zero-COVID policy’ led to a total shutdown of activities for weeks (it seems, however, that SMIC was not stopped in Wuhan city during the first block in 2020, nor in Shanghai city in 2022 [20,21]). However, the scenario was also exacerbated by events such as:

(a) the worst drought in Taiwan in 56 years, in 2021 (chip makers use large amounts of ultra-pure water to clean factories and wafers) [22].
(b) fires in plants (an Asahi Kasei semiconductor plant, in October 2020; a Renesas Electronics—which supplies 30% of the global market for microcontrollers used in automobiles—plant in March 2021 [23], and ASML’s Berlin plant, producing EUVL equipment, in January 2022 [24]).
(c) winter storms (forcing the closure of two Samsung and NXP semiconductor plants for several months in 2021, in Austin, Texas [25]).

All of these causes combined, along with the increased demand for integrated circuits, have repeatedly and at multiple points disrupted the chain, preventing global chip production from meeting the demands of the different types of industries, of which there are 169, according to a Goldman Sachs study which included industries that spend more than 1% of their Gross Domestic Product on chips.

The pandemic played a primary role in initiating the 2020 chip shortage for the automotive industry, as the global lockdown initially reduced personal mobility and demand for cars and, consequently, the automotive industry’s demand for chips. At the same time, the pandemic has accelerated digital transformation and the adoption of remote work, remote study, movie streaming, and e-commerce technologies worldwide, greatly increasing semiconductor demand in the consumer, telecommunications, and personal computer sectors. For example, in the fourth quarter of 2020, sales of mainstream computers grew 26.1% year-on-year. As a result, chip makers shifted their production to where demand was strongest, and failed to meet the needs of the automotive industry, which recovered rapidly, and sooner than expected.
Automakers cannot return to pre-2020 supply levels, and the problem will continue in the coming years, with components such as microcontrollers, image sensors, power-management units, power MOSFETs, and display drivers. Many automakers are therefore delaying vehicle deliveries, and suspending new orders for some models, to reduce production due to the global chip shortage. General Motors said in November 2021 that it would temporarily suspend the inclusion of heated and ventilated seats in several models, although it was working on a plan to retrofit those vehicles when parts became available. In the first seven months of 2022, automakers in North America skipped the assembly of more than 1 million vehicles [26].

Eight major Japanese automakers said they had assembled about 3.4 million vehicles in the first half of 2022—down more than 14% from the previous year. The world’s largest automaker, Toyota (based in Japan), reported a 31% drop in profits in the January–March quarter, compared with the same period a year earlier, and cut its production plans in June 2022 by “tens of thousands of units globally”; it also announced the suspension of production at various times in May and June, due to a shortage of spare parts caused by the pandemic lockdown in Shanghai [27].

According to the Society of Motor Manufacturers and Traders (SMMT), U.K. car production shrank by one-third (32.4%) in the first three months of 2022, with nearly 100,000 fewer cars than in the same period last year. Jaguar Land Rover’s latest financial results for 2022 revealed a loss of more than £500 million for the British company, despite a record order book [28,29].

The automotive sector exploits stable, mature semiconductor processes. But the chip shortage also affected more advanced nodes for consumer, industrial, and medical device manufacturing industries, with products such as video game consoles, graphic cards (the rise of cryptocurrency mining in 2021 has further boosted demand), memories, and processors.

3.3. U.S.–China Trade War

Other important reasons that have exacerbated the chip shortage are related to U.S. sanctions against China, which have caused a reduction in global production capacity, and an increase in inventories for Chinese companies. The case of Huawei is a clear example.

Huawei has been accused by the United States of putting backdoors in its equipment, that could be exploited for espionage purposes. In 2019, Huawei was blacklisted by the U.S., and placed on the so-called ‘Entity List’, which prohibits American companies (such as Google with its Android operating system) from exporting specific technologies to companies on the list. In 2020, Huawei was not able to source the cutting-edge chips it needed for its smartphones, because the U.S. was preventing Chinese companies from using advanced EDA tools and foundry services from companies (e.g., TSMC) that exploit U.S. intellectual property. Huawei, and many other Chinese companies on the Entity List, therefore had three to six months of chip inventories with which to try to secure their business [30]; however, Huawei founder, Ren Zhengfei, recently warned of tough times ahead for the firm, and committed to ensuring its survival [31].

Filling inventories reduced the number of chips available on the market in parallel with the onset of the pandemic. In addition, Chinese foundries and IDMs were unable to purchase new EUVL machines from ASML. In fact, in 2020, the U.S. forced the Dutch government to ban the export of ASML’s EUVL machines to China, because ASML also uses American intellectual property, according to Washington’s arguments (and the U.S. is now pushing to also ban the export of mainstream machines [32]). In August 2022, the Bureau of Industry and Security of the U.S. Department of Commerce issued a new provisional rule on a wider range of technologies export restrictions, although not mentioning China directly, and involving technologies for substrates of gallium oxide (Ga2O3) and diamonds, as well as EDA software “specially designed for the development of integrated circuits with Gate-All-Around Field-Effect Transistor (GAAFET) structure”, used to design 3-nm and more advanced chips. While the U.S. wished to maintain an
economic edge, its primary motivation was military, as advanced chips are used in advanced weapons.

3.4. Response of Companies and Governments to the Chip Shortage

The semiconductor industry has responded to the shortage of and the rapid increase in demand for chips, with a substantial increase in capital expenditures to support plans to build new fabs and, thus, to increase global chip production capacity in the near future. The most significant example is TSMC, which will increase capital spending from $30 billion in 2021 to $44 billion in 2022, and to a total of $100 billion in three years [33].

Intel has announced the construction of a 1000-acre mega chip production site in Ohio, with an initial build-out of two fabs of more than $20 billion [34]. Intel has also announced that it plans to spend more than $36.2 billion to build new semiconductor manufacturing facilities and research centers in the EU. According to the company, total investment could reach 80 billion euros over the next decade. Micron Technology has also announced that it plans to invest $40 billion in its U.S. manufacturing operations through to the end of the decade [35], while China’s largest semiconductor company, SMIC, has invested an average of $9.5 billion over the 2021–2022 period [36]. These are just a few examples that demonstrate the industry’s large investments in cutting-edge and in more mature technologies around the world.

In this context, manufacturers have also received, or will receive, external funding through government programs. Indeed, governments in several major economies are trying to incentivize local chip design, research, and especially chip production. For example, in 2014, China was the first to start providing government subsidies to its semiconductor ecosystem; these subsidies, it is estimated, will be around $100 billion by 2030. This has had, and will continue to have, an impact on chip supply, but most likely only for less advanced nodes, even if SMIC recently announced that it is able to process 7-nm chips [37].

The European Chips Act is a legislative proposal that would allocate more than 43 billion euros to the integrated circuit industry [38].

The U.S. CHIPS and Science Act aims to restore U.S. leadership in chip production, including for leading-edge technologies [39]. It has allocated about $52.7 billion in subsidies for companies building additional semiconductor fabs in the U.S., with $2 billion of the total to be used to build additional capacity for legacy chips, and about $13.2 billion also available for R&D and workforce development programs.

The Indian government, in December 2021, launched an incentive program, worth about $10 billion, to attract international semiconductor and display manufacturers, and to make the country a global manufacturing hub [40].

TSMC has also announced its plans for further investment in the United States and Japan, and potentially in the EU in the future [41]. Noteworthy is the construction of a 5-nm factory (Fab 21), completed in Arizona in July 2022. Similarly, Foxconn recently announced its plan to invest in India [42]. All such countermeasures, however, must take into account the current socioeconomic and geopolitical situation, and the changing scenarios that are unfolding. In addition, subsidizing the current industry could make it harder for potentially more innovative start-ups to be successful. The biggest gains for the U.S. chip industry come from disruptive new technologies replacing old ones, as the history of Intel has demonstrated [43].

4. Analysis and Perspectives

The semiconductor industry has grown over the past two decades, at an average of 8.5% per year (see Figure 9, which shows an exponential market growth, as the vertical axis is logarithmic), fundamentally because the industry’s extraordinary efforts have been based on strong and open scientific, technological, and commercial cooperation with both the global market and the supply chain, albeit with all the critical issues highlighted for the latter. In the semiconductor ecosystem, the capital-intensive segment of
the fabs (especially for the extreme technology nodes) was hence separated from that of
the IC design, so that fabless and fab-light houses could potentially generate a greater
revenue. The packaging step—once responsible for only a few percent of the final chip
cost—became more and more sophisticated, especially for systems in a package (SIP),
and for advanced RF chips, and was separated and specialized; furthermore, being the
most labor-intensive, and often based in low-cost countries such as China or Malaysia, it
was also favored by manufacturing incentives offered by the governments.

Figure 9. Global semiconductor market from 1987 to 2022 (solid line), compared to a constant 8.5% year-on-year growth (dashed line).

This ecosystem, highly dependent on business cooperation among nations, produced a relatively stable market, where most of the demand came from consecutive killer applications such as computers, smartphones, and now automotive and industrial IoT. Because of this stability, the oversight of supply chain inventory, production, sales, and even R&D was quite predictable, and the scalability of technology nodes was ‘only’ a matter of reducing optical size while keeping the transistor structure almost unchanged. Within this framework, semiconductor shortages have occurred cyclically, sometimes due to the emergence of a new killer application or exacerbated by external shocks, such as the technology bubble or the 2009 recession. In general, however, cycles from underproduction to overproduction have been repeatedly observed in leading-edge ICs and memory ICs. In the alternation of these semiconductor cycles, large fabless companies did not take on real financial risks and stresses. Even the risks were outsourced to foundries and, ultimately, to TSMC, which became the single supplier.

However, the semiconductor ecosystem is undergoing a dramatic change in its
structure, for a number of important concomitant reasons that will be discussed in the
remainder of this section; these reasons open up new scenarios in terms of market growth
and opportunities, but also present potential dangers, and require the development of
new policies and business models.

4.1. Unprecedented Market Growth and Profitability

We begin this analysis by observing an unprecedented level of demand for ICs, driven
by AI, 5G, IoT, health, and automotive applications. 2017 and 2021 were record years for
the semiconductor industry, with 22% and 25% year-on-year improvement [44,45].

2018 was identified as the year when the amount of data generated by humans was
equaled and surpassed by that generated by machines, which has since grown exponentially [46].

Of course, data must be transmitted, processed, and stored electronically, with specific integrated circuits. The high demand for semiconductors has only been exacerbated
by the pandemic, but stems from the sudden increase of all these new applications that require, and for many years will require, much greater capacity from the manufacturing industry. The scenario is radically different from the usual alternating inventory cycles seen above. In addition, the advent of AI, IoT, and autonomous vehicles requires the integrated circuit industry to be more flexible, more focused on R&D, and with shorter production times. Moreover, increased demand is leading to a resurgence of integrated circuit design start-ups, and perhaps more importantly, many big technology and automotive companies have begun to design chips in-house. The semiconductor industry has been faced with a new type of competitor.

4.2. Technological Breakpoints

Another major deal-breaker is the fact that, as we approach the atomic scale, optical shrinking techniques no longer work. In other words, the scaling road map is no longer marked. For each new nanoscale generation, new paths have to be worked out in terms of materials, processes, and transistor architecture. Every technological advance requires exponentially increasing expenditures in R&D, plant, machinery, and tools, which explains why there are only a few state-of-the-art foundries in the world today. Another observation needs to be made, about the breaking of the link between mature and leading nodes. The state-of-the-art in lithography has moved to 7, 5, and 3 nm, while microcontrollers, analog, IoT, and automotive, because of the functionality and reliability needed, are still implemented at 40–180 nm. As a result, when the next node is released, the previous one does not find suitable applications, contributing to increased foundry risks.

4.3. Foundry-Customer Agreements

The huge, high-risk expenditures discussed above can no longer be borne by foundries alone. Because large fabless companies have a vital need for new technology nodes, fabless must share these risks with foundries, by entering into long-term agreements (LTAs) or non-cancelable, non-refundable orders (NCNRs). Just as ASML required strong co-investment from its customers to realize EUVL machines, foundries require co-investment and capacity risk-sharing with their customers, to manage the semiconductor cycles. This is an indicator of the changing structure of the industry, which also explains TSMC’s present investment of $100 billion over three years, based on consultation with customers in anticipation of their needs [10].

At first glance, this radical change that occurs in agreeing on orders should not involve chipmakers with older technologies. After all, TSMC has a monopoly, while there is much competition among legacy-node companies. However, the recent chip shortage has shown that even mature technologies can be in short supply, and has revealed the new power and privileged position of the foundries, which can now take advantage of NCNR (no longer accepting the usual just-in-time order policy) with customers who want to ensure constant supply. Automakers are an example of such customers, who have had to partially abandon the well-known lean manufacturing system (Toyota’s system) [47], aimed at minimizing inventory, but exposing production to fluctuations in the IC market.

An agreement between chipmakers and automakers (and also with other industries) is also necessary for another important reason. The majority of automotive and industrial chips are realized in 40 nm or above, utilizing 200-mm wafers. Therefore, 200-mm wafer manufacturing is still important for these sectors and, incidentally, also for RF, MEM’s Analog, and Power Management ICs. Most of the 200-mm fabs are today fully depreciated, which happens after at least four years of operation, and this means that the manufacturing cost of a chip made in one of these fabs is very low, and that the final price is (or can be) very low too. However, the addition of fabrication capacity means that companies have to build new 200-mm fabs (from around 200 fabs worldwide in 2010, to 220 in 2025, as estimated by SEMI [48]) and have to buy new machines and tools to produce chips whose market price is a fraction of what they would really cost, because they are
produced in a new, non-depreciated fab. It clearly seems uneconomical, and consequently the investment in foundry is justified if customers sign LTA and NCNR, or if the foundry is supported by strong government subsidies (many of these new 200-mm fabs are in fact in China), or if several companies join forces [49]).

4.4. Trade War

Another extremely important factor that is blocking previously observed cooperation between nations is, of course, the already mentioned U.S.–China trade war. China is currently the main importer of semiconductors (35% of the global demand for semiconductors), which cost the country more than foreign crude oil [50]. China plans to develop its chip design and manufacturing industry (reaching 70% onshore chip manufacturing by 2025) by supporting its industry with large sums of money. In reply, the U.S. government not only is (perhaps belatedly) replicating this funding policy with the Chip and Science Act, but is also attempting to set barriers in China’s way to developing advanced semiconductors. Companies receiving subsidies from the U.S. cannot build advanced (<28 nm) chip fabs in China.

The U.S. Chip Act tries also to push foreign companies to take sides in the war, in order to obtain U.S. subsidies, surely affecting small and medium-sized businesses (TSMC and Samsung are not dependent on U.S. subsides). U.S. House Speaker Nancy Pelosi’s recent visit to East Asia—particularly South Korea, Taiwan and Japan—was an attempt to influence adherence to the U.S. geopolitical policy; but the effect of the U.S. bans, as SMIC and its 7-nm technology has shown, has been to push Chinese companies to do even more massive research. China could also choose to block deals with large U.S. tech companies such as Qualcomm, Broadcom, and Micron, cutting more than 50% of their revenues, and further reducing the U.S. market share. As another countermeasure, China could limit or ban Rare Earth Elements exports to the U.S. and its allies. In fact, U.S.–China tensions are accelerating the decoupling of the two supply chains and markets, raising concerns that Western semiconductor capacity could be oversized if the Chinese market is cut off.

4.5. The Ukraine Conflict

The effects of the war on commodity prices, supply chain constraints, and overall uncertainty will influence chip manufacturers and consumers. Ukraine is not only the world’s largest supplier of neon gas (70% of global supply)—critical for lasers used in lithography—but also the world’s largest supplier of xenon and krypton gas, also critical for chip production [51]. Roughly 54% of Ukrainian neon comes from two companies, Ingas (based in Mariupol) and Cryoin (based in Odessa), both cities highly affected by Russian attacks. Moreover, Russia holds 40% of the market for palladium, 15% for titanium, 12% for platinum, and 10% for copper, all of which are important for printed circuit boards, and for sensors and plating processes in chip production, as well as for high-tech products such as catalytic converters and ion batteries [52]. Sanctions imposed on Russia make the latter an uncertain source for such supplies in the near future [53].

In the short term, the impact of the conflict on semiconductor manufacturing should be manageable, as the largest foundries have great purchasing power, and access to stocks that can span two months or more. However, many other smaller companies do not have this type of reserve. In the medium–long term, the industry may consider making larger investments—for example, in neon recycling technologies.

Another effect of the conflict has been to create instability in world energy markets, raising energy costs, and driving oil and gas prices to their highest levels in nearly a decade. This could have a major impact on chipmakers, depending on local electricity prices. In fact, large semiconductor factories consume up to 100-MWh of energy per hour [54].

In addition, rising energy and fuel costs, combined with rising inflation, taxes, and interest rates, are putting pressure on consumers’ disposable income [55]. We are already seeing weakness in semiconductor end markets, particularly those exposed to consumer
spending. After a period of record revenues for chipmakers, peaking in the first quarter of 2022 after five consecutive quarters of record revenues, a slowdown is expected (but not for data centers and the automotive sectors). In this context, an increase in component prices due to rising energy costs could amplify market uncertainty.

4.6. Technology Issues vs. Political Goals

One problem that is not being adequately addressed in this context is that many political actions are taken by technologically illiterate people. Examples of policy mistakes are numerous, and range from U.K. prime minister Theresa May’s satisfaction with the acquisition of ARM by the Japanese SoftBank group, which was judged to be a good reaction of the markets in the aftermath of Brexit [56,57]. According to a study by the U.S.–China Business Council, the trade policies of former President Donald Trump have cost the U.S. 245,000 jobs [58]. The U.S. Chip Act extends trade limits to 10 years—an enormous time frame that doesn’t take into account the speed at which semiconductor technology is advancing. Furthermore, manufacturing costs in the U.S. and the EU are much higher than in China and Taiwan. Subsidizing the fabs is an attempt to offset these higher costs, but it does not seem to be adequate. In any case, if technological issues become intertwined and subordinated to unrelated political goals, then further pressure could be added to an increasingly fragile relationship, which could lead to military escalation; this could give mainland China an excuse to renew its threat to attack and annex Taiwan, which separated from mainland China in 1949. To this end, TSMC Chairman Mark Liu declared in August 2022 that if China were to invade Taiwan, the fabs would be rendered “non-operational”, implying that “war brings no winners, everyone is a loser” [59].

Another key factor that is not taken into account by the policies of different governments is the legal protection of inventions, something that should be focused on instead of resorting to blacklists and bans. While large companies have the strength to protect their intellectual property, the many small businesses and startups that contribute to the evolution of the art are unable, in the event of litigation, to resort to lengthy and costly international legal procedures. In this sense, state intervention would be crucial to help companies to develop and enforce patents.

4.7. Reshoring

In response to trade tensions, many Japanese, South Korean, and Taiwanese manufacturers plan to move some or all of their operations out of China [60]. In addition, the disruption of the supply chain is reversing the trend of recent years to relocate, even far away, in view of significant savings on labor costs. Bringing the production of assembly components closer (reshoring) is a new trend. In this regard, less and less manpower is needed in the new fabs and OSAT services for the high degree of automation. For instance, the new 300-mm fab of Infineon, which opened in Villach (Austria) in 2021, requires only 400 workers.

4.8. Talent War

Whereas all types of industrial and information engineers are in demand, those that work on enabling technology should be prioritized. Both research for new lithography nodes and IC design require an increasing number of well-educated engineers, possibly with a Ph.D., as these activities are talent-intensive. Moreover, we have seen that chip designers are recently being competed for also by big tech companies and carmakers. In this context, the continuing shortage of young talent interested in semiconductor engineering is of great concern, and a talent war is already underway.

China has struggled to find 230,000 more semiconductor engineers by 2022, to meet its already mentioned targets. TSMC and other chipmakers in Taiwan are working with the government to strengthen the local semiconductor workforce, by investing $300 million in new graduate schools for semiconductor chip technology programs [61,62]. Intel
plans to address the workforce shortage with a $100 million investment over the next decade, to build a talent pipeline and to support research, education, and workforce development in the United States. Approximately $50 million of Intel’s investment will go toward funding grants to Ohio higher education institutions for new undergraduate and graduate programs, faculty training, laboratory equipment upgrades, research aimed at improving semiconductor manufacturing, and opportunities for students, including internships. Lately, Intel and many other companies have been hiring college and undergraduate students as a desperate tentative measure, to fill vacancies. Intel airs TV commercials during soccer games to promote job positions.

In terms of educational global effort and contributions in the field of IC design, the Free and Open Source Silicon Foundation (FOSSI), [63], the free Skywater’s 130nm PDK [64], and the free tapeouts sponsored by Google for low-income countries [65] should be mentioned.

This author has already emphasized in [4] the urgency of clear and effective information campaigns on the role of semiconductor engineers and on the job opportunities offered by the semiconductor chip industry; he also noted the total absence of the keywords ‘semiconductors’ and ‘electronics’ (either micro- or nano-) in the various analyses and reports on the professions of the future, and summarized a series of actions to be taken to reverse the talent shortage [4,66]. However, this problem also requires industry players to change their hiring and staffing policies. Currently, salaries for chip-related hardware research and design are low, compared to Internet companies. Better wages are in fact needed to make this profession more attractive. Actually, the war for talent has led to slight wage increases, which is inevitable if semiconductor engineers and IC designers are in short supply and the market needs new fabs with additional labor. The limited, but significant, data available to the author show that the entry-level salary of an electronic engineer in Italy has increased by an average of 2.5% every year since 2017 [67].

4.9. Branding

The IC industry now exhibits a structure with high barriers to entry, and significant pricing power. Selling an IC for the price of a few cents, even if it comes from a depreciated fab, seems rather unbelievable, if we think of the expensive technology and the knowledge necessary to produce it. Chip manufacturers should use value-based pricing, and avoid downward quotations if they want to increase revenues, draw investors, and offer better salaries. This could happen more easily if they do not operate in a stagnant market, but instead move into emerging and more profitable ones. Working in new markets improves the company’s reputation, and attracts more customers and talent. The best chipmakers need to brand themselves.

5. Conclusions

Demand for integrated circuits is steadily rising, due to the emergence of an increasing number of new applications, supporting exponential market growth that involves both advanced and mature technology nodes. A general boost in chip production capacity is being prepared, which in turn requires a new kind of strong commercial agreement between manufacturers and customers, made possible by the new power position of foundries and the general consensus that chips are not commodities, but key and strategic elements. However, if new fabs are built by prior agreement with customers, then labor should also be secured in advance. New fabs do indeed require new workers and engineers, which are already in short supply; not to mention that other types of industries are developing in-house chip design capabilities, and will compete in the talent war. Therefore, it will be necessary for governments, companies, universities, and associations, to work jointly on actions such as orientation and communication for young people, promotion of STEM (Science, Technology, Engineering, and Mathematics) studies for girls, grants to universities to create new degree programs, new positions and new laboratories, and study exchanges with foreign countries, in order to reinforce passion.
among the young generations [4]. However, strengthening passion cannot be the only answer. Raising salaries, improving career opportunities, compensation, and benefits are necessary actions to reverse the trend of young people interested in engineering who now massively prefer software to the lab. The other major unknown that will affect the future of the IC industry concerns the trade war between the U.S. and China and, to a lesser extent, the Ukrainian crisis. Developments in these tensions are leading to partial or even full supply chain decoupling from markets that are no longer global, causing major concerns for the profitability of foundries (especially those at the forefront) facing such severe constraints and market limitation. The effects of the Ukrainian war are on commodity and energy prices, supply chain constraints, and overall uncertainty. The future of the semiconductor ecosystem, however potentially disruptive it may continue to be, is closely tied to these critical issues, whose effects will need to be carefully considered.

**Funding:** This research was partially funded by the University of Catania, under grant PIACERI 2022.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The author declares no conflict of interest.

**References**


