Commun. Optim. Theory 2026 (2026) 18 https://doi.org/10.23952/cot.2026.18

MATHRES Communications in Optimization Theory
Available online at http://cot.mathres.org

RISE OF CO-EVOLUTIONAL PATHS FOR GROWTH: A STEPWISE STUDY ABOUT THE PURE-PLAY FOUNDRY MODEL

WEI-TORNG JUANG[1], AN-CHI TUNG[2,*], HENRY WAN, JR.[3]

[1]Department of Economics, Soochow University, Taiwan
[2]Institute of Economics, Academia Sinica, Taiwan
[3]Department of Economics, Cornell University, Ithaca, NY, USA

Dedicated to the Memory of Prof. Tapan Mitra

Abstract. In this paper, we review the TSMC saga in three periods against the broader background: pre-TSMC, foundry emergence, and further disintegration. And we give brief accounts of how TSMC upgrades itself technologically and how global geopolitical structure evolves. We then present a series of analytical models to study the nature of the co-evolution of foundries and the fabless. Finally, some implications are drawn.
Keywords. Analytical model; Co-evolution; Pure-play foundry model; Fabless; Game theory.
2020 MSC. 91B38, 91B24, 91A80, 91B06, 91A10.

1. INTRODUCTION

Taiwan Semiconductor Manufacturing Company (TSMC) has caught the attention of the world during the US-China trade war, and even more so with the escalating political tensions across the Taiwan Strait. This is because TSMC provided nearly 60% of global foundry services in 2023[1] [2] and is expected to monopolize the world's AI chip production in 2024, while the semiconductor industry happens to be the driving force behind AI technologies that are bound to decide the prosperity and security of any nation in the years ahead. By mid 2024, TSMC

*Corresponding author.
E-mail address: wjuang@scu.edu.tw (W.-T. Juang), actung@econ.sinica.edu.tw (A.-C. Tung), hyw1@cornell.edu (H. Wan, Jr.)
Received June 9, 2024; Accepted September 14, 2024.

[1]In 5 nm technology, TSMC held a market share of 70-80% in 2023, and over 90% in 3 nm. Note that these market share estimates may differ somewhat by reporting institutions.
[2]Most of our statistics and real-world examples are not listed as references due to space constraints, though all are well-documented. A complete bibliography is available from the authors upon request.

made it onto the world's top 10 largest companies with a market value over 25,000 times as much as its initial batch of capital in 1987[3].

The huge success is widely believed to result from the innovative pure-play foundry model that TSMC pioneered in 1987, which created a market niche for itself in its early years by claiming non-competition with clients. Subsequently, this niche redefined the ecosystem of the semiconductor industry and helped to raise the overall efficiency. Between 2001 and 2022, global GDP tripled, but world semiconductor revenue quadrupled, the foundry services increased by a factor of 14, and TSMC revenue rose by an impressive factor of 21[4].

The fabless sector took off along with the launch of TSMC. Before 1987, chip designers without a fab had to farm out to the IDMs (integrated device manufacturers) for spare capacities, but their orders could be rejected or delayed during high season. Moreover, their technology knowhow might leak out through such deals. In an industry where profitability depends not only on quality but also on time-to-market, these problems are costly. The availability of dedicated foundries helped to stamp out both risks. It follows that more fabless chip designers entered the market with their innovations, including today's giant Nvidia (launched in 1993). The combined share of fabless firms in global IC sales increased dramatically from 0.1% to 34.8% between 1987 and 2021. An unquantifiable benefit of this increase is to make possible the realization of innovations which creates values for the industry and the world.

In the wake of the co-evolution of the fabless and foundry sectors, many IDMs have disintegrated vertically since the late 1990s. Most IDMs nowadays outsource to pure-play foundries more or less – even Intel is TSMC's client, and firms like AMD and IBM have gone further and become fabless. In addition, the horizontal disintegration has also intensified between advanced and mature nodes, with only three chip makers producing the 3 nm node chips in 2024 (Intel, Samsung and TSMC), and just a handful of fabless or system firms using these advanced chips (e.g., Apple and Nvidia).

These developments are all triggered by the emergence of the pure-play foundry model. In economics terms, the launch of TSMC as a novel business model exemplifies the "self-discovery of comparative advantage" of Hausmann and Rodrik [5], based on its sole competitive edge in manufacturing. This innovative business model addresses a common coordination failure problem in the IC industry, by activating a latent productive potential of fabless chip designers who had been denied market entry. In turn, the co-evolution of the fabless and the foundries prompts for the transformation of the other player, the IDMs. Note that the launch of TSMC and ensuing developments did not proceed with a standard Walrasian mechanism in the Debreu style. Rather, it is a dynamic Schumpeterian process of creative destruction. In a nutshell, the TSMC case demonstrates that the innovative dedicated foundry model brings impacts on many more innovations in organization and in technologies, enhancing the overall efficiency of the entire industry.

While we plan to go deeper in future studies, here we conduct an economic analysis of the nature of the rise of co-evolutional paths for growth. Specifically, we look into how the emergence of the pure-play foundry initiates the series of developments. Section 2 begins with

[3]TSMC's market value was estimated at around $ 900 billion by mid 2024, according to companiesmarketcap.com. In 1987, however, it only received NT $ 1.3775 billion (or $ 35 million) as capital (or 24.2% of total authorized capital of $ 145 million).

[4]Global semiconductor revenue is from World Semiconductor Trade Statistics (WSTS), global GDP is from the World Bank, and the rest are from TSMC annual reports.

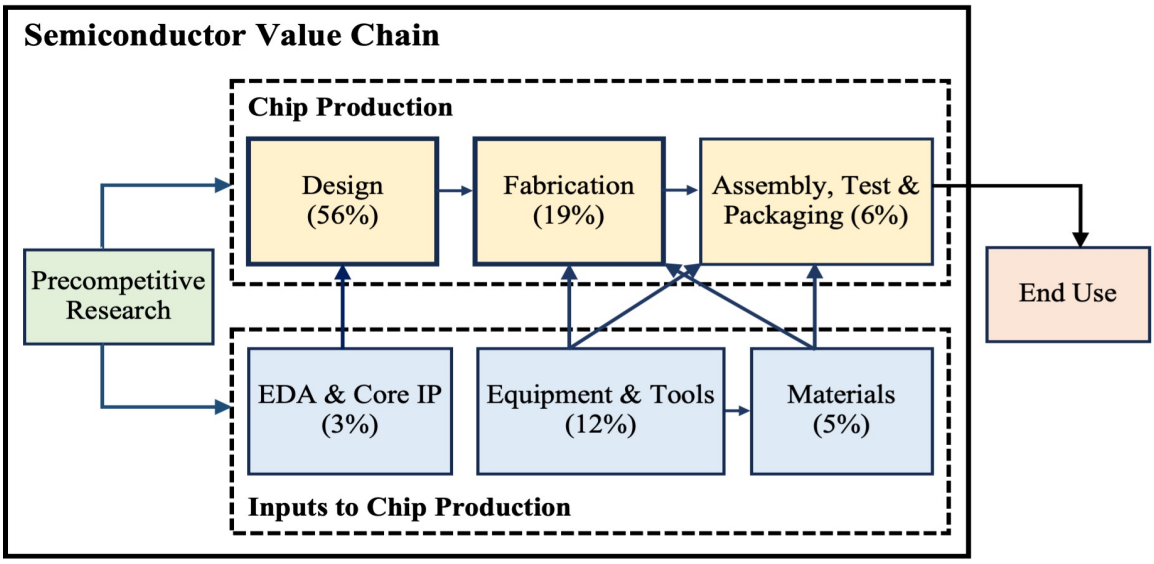

FIGURE 1. Value-added Structure of the Semiconductor Value Chain

an overview of the semiconductor industry and three types of firms, then reviews the TSMC saga in three periods against the broader background: pre-TSMC, advent of the pure-play foundry model, and further disintegration. These are followed by a brief discussion of how TSMC upgrades itself incessantly into a technological leader, and how global geopolitical structure evolves over time. Section 3 presents a series of analytical models to highlight key episodes in the co-evolutional process. Section 4 discusses the implications to conclude the paper.

2. TSMC AND THE SEMICONDUCTOR INDUSTRY

2.1. **Semiconductor ecosystem.** The semiconductor industry comprises an extraordinarily complex supply chain. There are two broad categories of semiconductor components: integrated circuits, or ICs (including logic, memory, microprocessors and analog), and non-ICs (including optoelectronics, sensors, and discrete semiconductors, or OSDs). As global non-IC sales are small relative to IC sales today[5], for simplicity, we use the terms "semiconductor" and "IC" interchangeably in this study.

The semiconductor ecosystem, as illustrated in Figure 1[6], is composed of a number of differentiated activities. In this study, we focus on the two largest segments by value-added, that is, design (56% of total value-added in 2022) and fabrication (19%), the former being R&D-intensive, and the latter capital-intensive. Other segments either are small (e.g., assembly, testing and packaging accounted for only 6%) or have been studied elsewhere (e.g., ASML is the subject of [7]).

2.2. **Three types of semiconductor firms.** In the design and fabrication segments, there are three types of firms, plus a variant. The earliest type of semiconductor firms is an IDM, who

[5]In 2021, non-ICs made only 17% of global semiconductor sales, and ICs 83% (logic chips 27%, microprocessors 14%, memory 28%, and analog 14%).

[6]Figure 1 is adapted from [9]. The numbers in () are percentages of global semiconductor value-added in 2022, which add up to over 100%, due to rounding.

designs and manufactures its own chips, and also does the assembly, testing and packaging. Examples of IDMs include Intel, Samsung, NEC, and Philips, many of which also make consumer or industrial electronics products. Their vertically integrated structure ensures good control over the entire production process, facilitating an effective supply chain. However, by owning a fab, an IDM may encounter the problems of over- or under-capacity during market fluctuation, as well as facility obsolescence when new technologies arrive.

The second type is a fabless firm, who designs and sells its chips, but buys fabrication services from external manufacturers. Examples include those that are born fabless (e.g., Nvidia and Qualcomm) and those who are transformed from IDMs (like AMD and IBM). The main advantage of being fabless is that the firm can concentrate its resources on researches in design, without having to invest in expensive manufacturing facilities, and can therefore adapt more quickly to market changes. Yet relying on existing IDMs for production means the fabless firm runs the risks of uncertainties in capacity and delivery time, and also leakages of business secrets to the external producers as potential rivals. In addition, the fabless may have to pay an extra transaction cost to coordinate with external manufacturers. Note that this transaction cost was prohibitively high in the early days due to technology constraints so that no fabless could survive; once the cost became more affordable, some startups then entered.

The third type is a pure-play foundry, or dedicated foundry, who provides manufacturing services to clients, but neither designs nor sells its own products. The model is pioneered by TSMC in 1987, with a non-competition clause written in its charter to minimize clients' worries about IP leakage and capacity uncertainty [6]. When the new business model was proved successful, more dedicated foundries emerged, such as Taiwan's UMC, all following the non-competition practice. Those firms are good in fabrication and have adequate fund to build and equip fabs, but are weak at product innovation. So far, there are only a small number of pure-play foundries worldwide, who serve multiple clients. In contrast, fabless chip designers are numerous today, who typically deal with just a few external suppliers. For example, Nvidia uses mainly two manufacturers, TSMC and Samsung, while TSMC produced for 532 customers in 2022.

In the ever-evolving landscape of the semiconductor industry, there is a new firm type, the "fab-lite" IDMs. In early days, it is not uncommon that an IDM farmed out some of its mature products so as to concentrate on cutting-edge nodes[7]. But starting from the late 1990s, many IDMs choose to be fab-lite or asset-lite in a different manner that they produce in-house mature-node chips with existing facilities, while outsourcing to third-party foundries for advanced-node chips in which they prefer to be spared of the expensive fab construction cost[8]. As these fab-lite firms can be seen as intermediate cases between IDMs and the fabless, for simplicity, our analysis below focuses on the three main types of semiconductor firms, grouping fab-lite firms under the IDMs.

[7]In 1987, for example, Intel had some of their more mature commodity products produced by external contractors as reported in its 1987 annual report.

[8]For example, Renesas, a major automotive chipmaking IDM based in Japan, produced only 43% of its chips in-house in 2022 (down from 68% in 2017), and relied increasingly on TSMC for both advanced nodes and extra capacity of conventional 40/45 nm node.

	IDM	Fabless	Foundry
Financial indicators (2016-2019)			
gross margin	52%	50%	40%
R&D expenditure / revenue	14%	20%	9%
capital expenditure / revenue	20%	4%	34%
Market concentration (2021)			
market share of #1	18.9%	13.1%	56.6%
combined market share of #2-5	40.1%	32.5%	28.1%
combined market share of the rest	41.0%	54.4%	15.3%

TABLE 1. Comparing Fabless, Foundry and IDM in Selected Years

Table 1[9] compares two sets of statistics of the three types of semiconductor firms in representative years. The upper part of the table shows that, on average, the fabless firms are R&D-intensive, pure-play foundries are capital-intensive, and IDMs are both R&D- and capital-intensive. The lower part of the table depicts the market structure. In the IDM segment where there are a few titans (e.g., Samsung and Intel), the top five firms together commanded over half of the market share in 2021. In the fabless segment, there are several giants (like Nvidia) and many small ones. In the foundry sector, TSMC dominates with over half of the market share (see also Footnote 1).

2.3. **TSMC saga.** TSMC is a game changer in the semiconductor ecosystem. The development of TSMC and the entire IC industry are presented in three stages.

Stage 1. Pre-TSMC (IDM dominance in 1958-1986)

Jack Kilby (Nobel Laureate in Physics in 2000) in Texas Instruments (TI) assembled the first prototype of an IC in 1958; independently, Robert Noyce at Fairchild[10] was granted the first patent for IC in history in 1961. These events mark the start of the IC industry. Because manufacturing methods were mainly proprietary and very expensive to develop at that time, a firm needed to have full control over the entire value chain. Therefore, all semiconductor firms in this period, both startups (e.g., Intel was launched in 1968) and incumbents (e.g., TI), were organized as IDMs.

By the mid 1980s, the standardization of production processes (with CMOS technology) opened a new opportunity for chip designing to decouple from owning manufacturing facilities. This lowered the transaction cost between design and fabrication significantly[11]. It is then possible for small startups with shallow pockets, who had previously been denied access to the

[9]In Table 1, "Foundry" includes foundry services provided by Samsung and Intel. The table is compiled by the authors from various sources.

[10]Noyce co-founded Fairchild Semiconductor in 1957 and Intel in 1968, and both firms serve as examples of Schumpeter's creative destruction [2].

[11]However, the transaction cost is not zero, as the CMOS processes, which are to a great extent standard, still entail a lot of variations and complexities, and require "a very close cooperation between foundries and fabless firms" [1].

market, to live on innovations without having to build, run and fill their own fabs[12]. Chips and Technologies (C&T), for example, had trouble raising the starting capital of $50 million as an IDM, but managed to launch as a fabless with $5 million in 1984. Other examples include Altera and Xilinx, which were set up in 1983 and 1984, respectively [1].

As mentioned above, these new fabless firms had to rely on established IDMs for fabrication, which had many downsides. To begin with, there was always the risk of losing trade secrets. The IDMs might require explicitly the right to use the clients' IPs in their own products, or they might just learn the technological knowhow through the deal[13]. If the fabless could not offer anything valuable to the IDM in exchange, explicitly or implicitly, its orders could get rejected. Furthermore, there is the risk of inadequate production capacity or unpunctual delivery[14], as the IDMs took external orders only to keep their own fabs busy. All in all, the coordination between the fabless and the external manufacturers was not easy to manage[15], hence there were very few fabless firms in the mid 1980s.

Stage 2. Emergence of pure-play foundries (1987 to 1999)

In 1987, TSMC was launched, pioneering the novel idea of an independent foundry that is dedicated to fabricating chips designed by others. By promising never to compete against the clients with a written clause in its charter, TSMC solves the twin problems of the fabless: loss of trade secret, and uncertainty in capacity and delivery time. This innovative model not only carved out a niche for TSMC itself, but also paved the way for both the foundry-fabless co-evolution and the transformation of the IDMs later on.

(a) Self-discovery of comparative advantage by TSMC

The foundry idea had been around since the late 1970s[16], but not realized until Morris Chang founded TSMC[17]. Chang, president of Taiwan's Industrial Technology Research Institute (ITRI, a government research lab) in 1985-1988 after a long career in the US with TI (1958-1983) and General Instrument (1984-1985), undertook a methodic review of the comparative advantage for the development of Taiwan, with GDP per capita at only one fourth of the US level in 1987. He discovered that, in semiconductor, Taiwan was weak in basic R&D, product design, IP and

[12]Most startups have inadequate financial endowment to pay for the fab cost, and are unable to become IDMs. One exception is China's top memory chip maker, Yangtze Memory Technology Corp., a subsidiary of Tsinghua Unigroup, which was launched in 2016 as an IDM with a starting capital of close to $6 billion.

[13]Japanese DRAM makers, who did most foundry services for the world in the 1980s and the early 1990s, always "wanted some production information in return" from the clients, according to Don Brooks, who worked for Fujitsu in 1988-1991 before serving TSMC in 1991-1997. American IDMs, like TI, also traded factory capacity for technology rights.

[14]For example, Actel, a fabless, saw its fourth quarter's profit wiped out in 1993, because it was not able to find a pinch hitter when one of its contracted foundries went awry.

[15]For example, Crystal (a fabless, later to be acquired by Cirrus Logic in 1991) "had to coordinate production done in the fabs of seven different IDMs in order to have access to the best foundry service for a particular product" [4].

[16]Lynn Conway and Carver Mead wrote a book on the potential of separating chip design from fabrication in the late 1970s. They thought this would create "a Gutenberg moment" for semiconductors.

[17]Chang had toyed with the foundry concept when he was with TI around 1976. But TI did not appreciate the idea, as there were no fabless firms around as potential customers at that time.

marketing. The only potential strength was in fabrication[18]. Given that wafer manufacturing was "the least evil choice" in Chang's own words, he set up TSMC as a foundry dedicated to fabrication[19].

To get the firm started, the Taiwan government provided half of the capital funds, but requested Chang to find a multinational semiconductor firm as the anchor investor. Chang's investment invitation was rejected by Intel, TI, Toshiba and many others. The only international player that said yes was Philips[20], though its semiconductor technology was not highly rated by Chang. Philips invested $40 million in cash, transferred its technology and brought in patents and IPs to add to TSMC's technology that came from ITRI. By 1987, TSMC began with a 3000 nm process technology[21], 2 to 3 generations behind the leading technology, along with 100 plus former ITRI employees and a fab rented from ITRI, which saved TSMC two years' fab construction time.

Despite that the problems of capital funding and fab construction had been properly worked out, TSMC was plagued by a more basic issue: where is the market? The fabless firms were rare and small then, the IDMs outsourced only when they were short of capacity or did not want to manufacture certain chips, and TSMC itself had neither credit history nor good technology. Not surprisingly, Gordon Moore called the pure-play foundry model a "bum" idea [8].

TSMC's dilemma in its starting years is illustrated vividly by two episodes. Around 1987 when John East was vice president of AMD (an IDM then), he refused to contract with TSMC, because he had never heard of this firm. Ironically, when East became CEO of Actel (a fabless startup) in 1988, TSMC repeatedly rejected his orders, as the nonstandard custom processes required by Actel were beyond TSMC's capabilities at that time[22].

It follows that, TSMC eked out its existence on tiny and occasional orders from ITRI, and also leftout orders from foreign IDMs, such as Philips, by offering production services at "reasonable" prices[23]. Despite registering a gross income loss in 1987, it gradually climbed the learning curve through customer feedbacks. One deal with Intel was noteworthy. After Andy Grove's visit to TSMC around 1988-1989, TSMC spent more than a year to meet Intel's 200

[18]Chang has observed since his TI days that, the yield rates in ITRI's pilot plant as well as in Japanese IDMs were higher than in TI. He attributed it to the abundance of skilled technicians and engineers, low turnover rate, and high employee dedication, which traits are common in East Asia to date.

[19]UMC's Robert Tsao claimed that he had the idea of a professional foundry in 1984, before the launch of TSMC [6]. However, with hindsight, Tsao's plan to serve only design houses that UMC had invested in does not seem to be encouraging to other fabless firms.

[20]A likely reason that Philips was willing to invest in TSMC is it had some previous investment in Taiwan already, and "wanted to stay on the Taiwanese government's good side" [8].

[21]In 1987, TSMC started by transferring 2000 nm and 3500 nm technologies from ITRI, and customized a 3000 nm technology for Philips. It successfully developed its own technology at smaller nodes the next year on. In contrast, in 1987, Intel was already doing the bulk of its production on 1.5 micron (1500 nm) technologies, and mass-producing its 386 processors with 1000 nm node technology.

[22]TSMC positioned itself as a foundry for standard chips in its initial years, and Chang admitted that the firm was not yet mature enough to serve the fabless clients.

[23]For example, Philips (and its affiliates), who had been TSMC's top client in 1994, was surprised to find TSMC charged lower service fee than its own production cost, due mainly to good yields in TSMC.

plus requests before it finally won the order. The deal itself was money-losing, but helped to improve TSMC's technological capabilities as well as to boost the confidence of potential fabless clients, such as Altera[24].

TSMC's growth speeded up since 1991. The compound annual growth rate of its net revenue was 37% in 1991-1999, and operating income grew at 47% annually. It plowed back the profit into R&D and equipment to shorten the technology gap with industry leaders, particularly Intel, who has been Chang's number one opposing force since his days with TI. By 2001, TSMC has more or less caught up with Intel when both firms mass-produced at 130 nm process node.

(b) Co-evolution of the fabless and foundries

The availability of a dedicated foundry spared the fabless firms of losses incurred from IDM's improper practice, thus they grew like "bamboo shoots" in Chang's own words. As mentioned above, total fabless sales as a portion of global IC sales went up substantially from 0.1% in 1987 to 7.6% in 1999, then to 34.8% in 2021 (the solid line in Figure 2[25]).

Also remarkable are the growths of TSMC as a percentage of global IC sales (dashed line in Figure 2) and of total foundry services (including those provided by IDMs as third-party producers, depicted by the line with circle markers). Notably, between 1987 and the early 1990s, TSMC was the only pure-play foundry in the world. Its market share in foundry service was only 23% in 1994, the rest were provided by the IDMs (Figure 3[26]). However, when this new business model was proved successful, a number of pure-play foundries were launched, such as Tower of Israel (launched in 1993), UMC in Taiwan (transformed in 1995 from an IDM who had served mainly the domestic market), Chartered of Singapore (reorganized in 1995, and acquired in 2009 by GlobalFoundries, or GF, headquartered in the US), and SMIC in China (set up in 2000). When global foundry capacities increase, more fabless designers are induced to enter, bringing in more innovations, and the entire semiconductor market expands, in a virtuous cycle. As TSMC continues to maintain the first mover's advantage, its market share in total foundry services has constantly stayed over 50% since 2014.

Stage 3. Further disintegration (2000 to present)

Parallel to the success of the fabless-foundry co-evolution, one recent development is the transformation of some IDMs into fabless firms. For instance, LSI Logic (started in 1980) went fabless in 2006, AMD (started in 1969) span off GF in 2008, and IBM (started in 1911) sold its manufacturing facilities to GF in 2014. By going fabless, they could focus on product innovation, and have their chips produced by more capable manufacturers, like TSMC or UMC. In AMD's case, the benefit is very pronounced when TSMC became its main partner for the

[24]Also, according to Don Brooks (former president of TSMC), "since wisely Morris Chang put a Westerner at the helm, the confidence of Western fabless designers went up". Before Brooks served as TSMC president in 1991-1997, it was James Dykes in 1987-1988, and Kraus Wiemer in 1988-1991. TSMC presidents after Brooks, beside Chang himself, are all Taiwanese.

[25]In Figure 2, "Foundry" includes the foundry services provided by both pure-play foundries and IDMs. The numbers in the figure are from TSMC annual reports and WSTS.

[26]The numbers in Figure 3 are from TSMC annual reports and IDC.

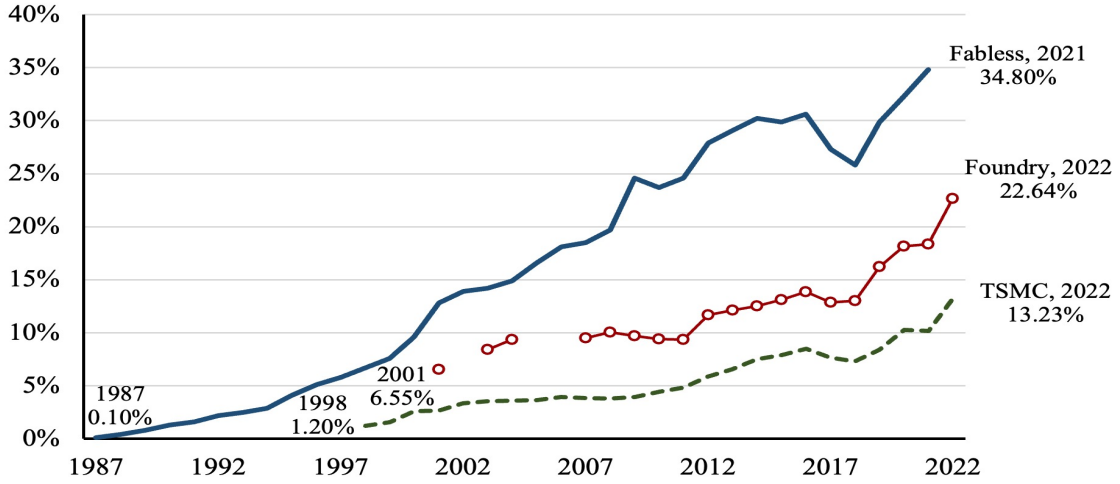

FIGURE 2. Rising Shares of the Fabless, Foundry Services and TSMC in Global IC Sales, 1987-2022

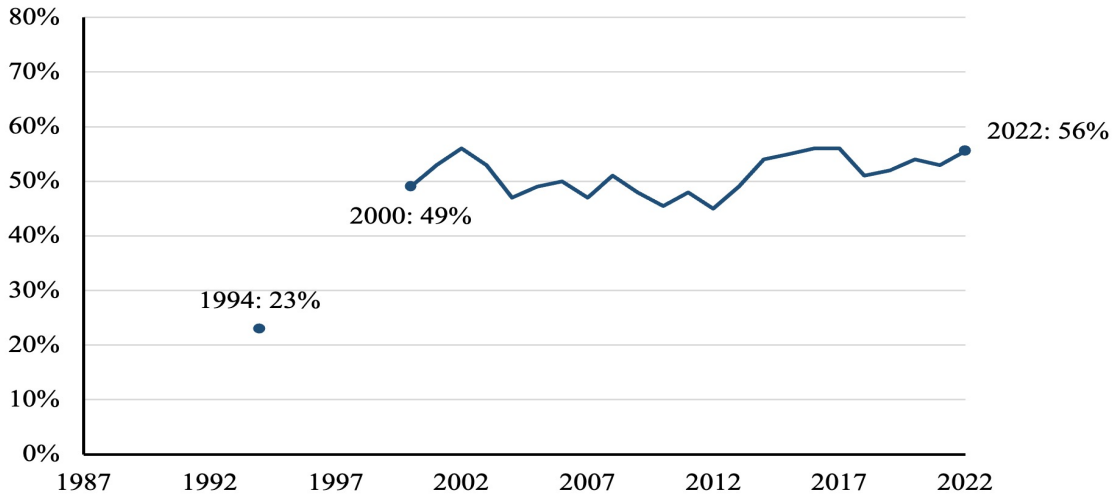

FIGURE 3. TSMC Share in Foundry Service Market, 1994-2022

leading-edge node chips since 2019: its revenue more than tripled in 2019-2023 (from $6,731 million to $22,680 million)[27].

This new development has much to do with the skyrocketing of the foundry cost. The cost to build and equip a foundry has always been on the rise (Figure 4)[28] [29], but the increase is

[27]In 2009, AMD and GF struck a supply agreement mandating that AMD must purchase all of its chips from GF. But after GF announced its exit from 7 nm in 2018, AMD was allowed full flexibility in choosing foundry partners in advanced node chips.

[28]This trend is already predicted by Moore's Second Law (also known as Rock's Law), which says that the cost of a semiconductor chip fabrication plant doubles every four years. It can be seen as the economic flip side to Moore's (First) Law, predicting from observations that the number of transistors in an IC doubles every two years.

[29]In Figure 4, scale adjustment is made for nodes at 5-130 nm by 50k wafers/month, and no adjustment otherwise. The numbers are compiled by the authors from various sources.

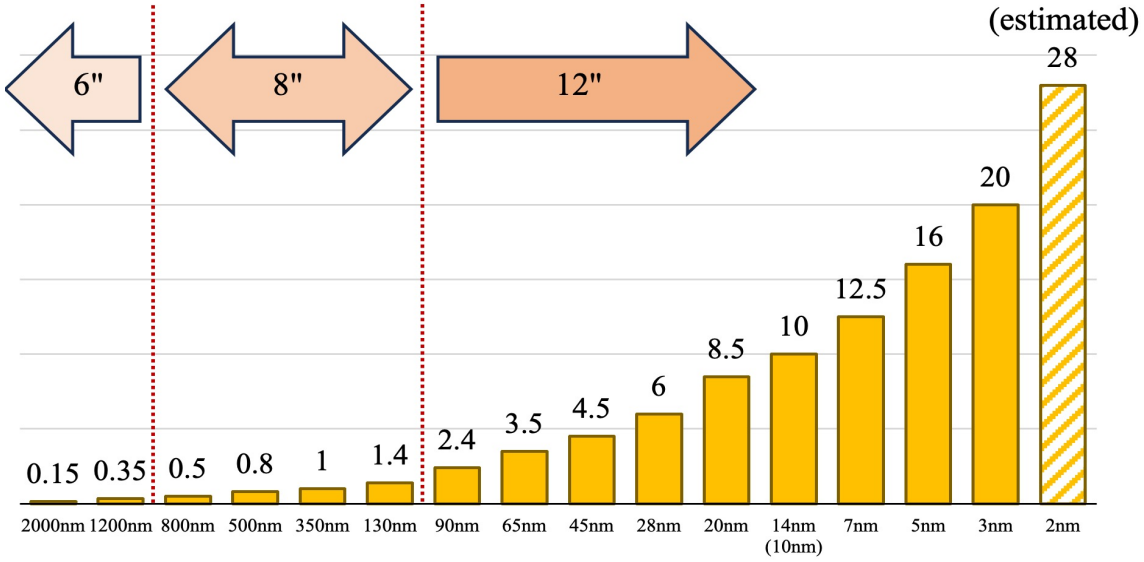

FIGURE 4. Foundry Cost ($ billion)

particularly significant at or below 7 nm, as these advanced nodes rely heavily on the very expensive EUV equipment[30]. By now, a 2 nm fab may cost $28 billion or more[31], whereas it was only about $150 million for a 2000 nm fab in the 1980s.

Design costs also rise with smaller technology node for the fabless and the IDMs. A 65 nm chip would cost about $29 million to design in 2016, but a complex GPU at 3 nm node may take as much as $1.5 billion to develop (including about 40% for software).

The rise in both fab cost and design cost at the advanced nodes brings another new change beside the deepening of vertical disintegration, as illustrated by some IDMs turning fab-lite or fabless. A horizontal disintegration is taking place between process nodes. To be more specific, Apple and Nvidia need advanced chips to compete in the markets of premium phones and GPUs, but firms producing memory chips[32] or automotive chips could use mature nodes for better cost efficiency. On the foundry side, TSMC keeps going into smaller nodes besides the existing mature nodes, but both UMC and GF decided in 2018 to give up investing in 7 nm or below.

Figure 5[33] shows the dramatic reduction of the number of logic chipmakers at the leading-edge node[34]. At 130 nm, there were 26 makers (22 IDMs and four pure-play foundries); at 10

[30]The average selling price of an EUV machine by ASML was about €172 million in 2023, whereas the price of a DUV machine was only about €72 million for an immersion system or €24 million for a dry system, as reported in ASML's 2023 annual report.

[31]Manufacturing equipment cost represents about 65% of total fab cost, construction cost another 20%, testing equipment 7%, packaging equipment 5%, and 3% for the rest.

[32]Memory chips typically do not require the most advanced process node to produce. Samsung, for example, produced its most advanced GDDR7 DRAM at 10 nm in 2023, while manufacturing logic chips at 3 nm or 5 nm node for itself and clients.

[33]Figure 5 is compiled by the authors from IC Insights and various other sources.

[34]Another reason for chipmakers at the technology frontier to be fewer is that some firms initially on the list have been re-organized (e.g., Renesas is from Hitachi, Mitsubishi and NEC), and some have transformed into fab-lite or fabless (e.g., TI and AMD).

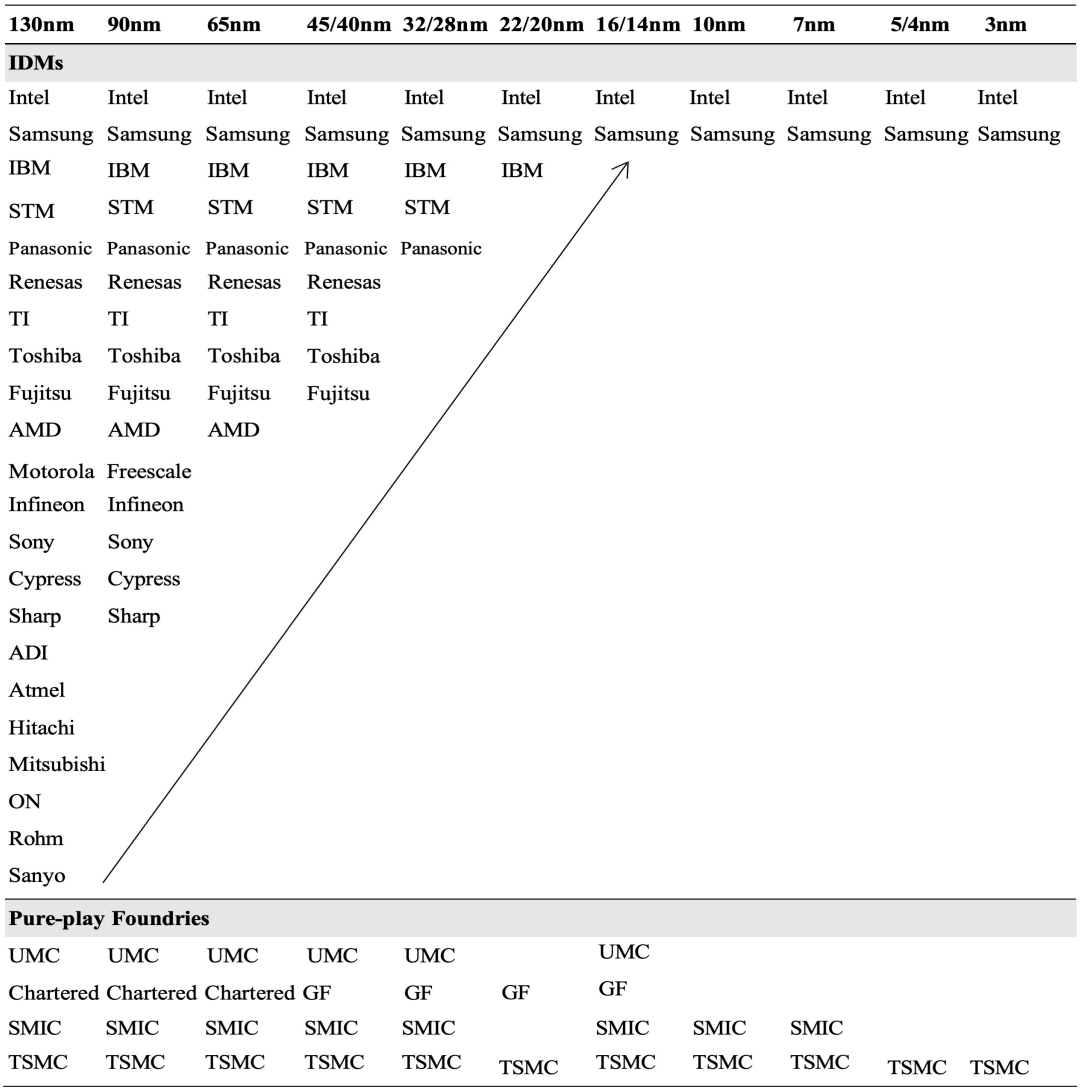

FIGURE 5. IC Manufacturers with Leading-Edge Capabilities

nm and 7 nm, there were four[35]. Currently at 3 nm node or below, which is expected to become the mainstream technology in the AI-induced boom, there are only three capable players, Intel, Samsung and TSMC.

2.4. **From technological laggard to leader.** As mentioned in Subsection 2.3, TSMC started from the process technology a couple of generations behind Intel in 1987, and roughly caught up around 2001. By 2016, Intel still led the leading-edge node non-memory chip (at 14/16 nm) and contributed nearly 70% of the market by revenue, TSMC's share was below 30%. In

[35]SMIC, one of the four remaining players at 7 nm, has supplied the chips for Huawei's Mate 60 phones, and plans to mass produce 5 nm or even 3 nm chips by the end of 2024. But it is often suspected that this achievement is rather a political project than a commercially feasible one, given its low yield and high production cost, with no access to the EUV equipment under US sanction.

2021, at 5 nm, TSMC's market share rose to 85%, the rest went to Samsung, and Intel was zero. By mid 2024, seven major customers (Nvidia, AMD, Intel, Qualcomm, MediaTek, Apple and Google) have chosen TSMC's 3 nm node over that of Samsung, with TSMC's capacity fully booked through 2026. In contrast, Intel's 3 nm node mainly serves internal demand, and Samsung's 3 nm node struggles to get external clients, including Google and Qualcomm on whom it has worked hard but failed to solicit orders, due to poor yield and power inefficiency. The lead of Taiwan (or TSMC, to be more precise) in sub-10 nm logic chips is expected to maintain until at least 2032, with 47% of global capacity[36], as projected by the Semiconductor Industry Association (SIA) [9].

TSMC's achievement is built up by a series of wise, and also lucky, technological decisions over the years. To name a few, in brief, it developed independently the copper process technology at 130 nm node, it chose correctly the gate-last technology over gate-first technology at 28 nm, it adopted EUV equipment earlier than rivals, it successfully speeded up R&D at 10 nm to win Apple's orders from Samsung[37], and it invented the immersion technique as well as invested in advanced chip packaging technology to help lengthening the lifespan of the Moore's Law. Cumulatively, TSMC has upgraded itself into one of the leaders in process technologies.

There are a couple of interrelated elements behind this success[38]. First, TSMC has persistently maintained a "building in-house R&D" strategy since Day One, though it had stayed in the catching-up stage until around 2000. Then in 2001, TSMC launched the semiconductor industry's first 130 nm low-k, copper system-on-a-chip (SoC) process technology. It was an important milestone for TSMC, as the technology was developed in-house, after declining a joint development invitation from IBM, whose project eventually turned out to be a failure regarding reliability. TSMC became recognized worldwide for its technological capabilities afterwards, and also pulled further away from UMC, its closest rival then. Shang-Yi Chiang, a key figure behind many of TSMC's major technological breakthroughs, was told explicitly by Morris Chang when they first met in 1997 that TSMC has determined to be a technology leader, rather than a fast follower.

Second, the high investment is powered by income from hyper-growing clients it carefully selects. In fact, TSMC has been rather "choosy" about partners. For a new startup client, TSMC would look at its business plan and management team to see if it can grow with TSMC. As for high-profile customers who can help TSMC to improve its technological capabilities or to support expensive research investments, TSMC always works hard to win their orders. In

[36]Taiwan's share in global wafer fabrication capacity for sub-10 nm logic chips is expected to drop from 69% in 2022 to 47% in 2030, while the US share will increase from 0% to 28%, as the next biggest supplier.

[37]For instance, TSMC initiated a "Night Hawk Plan" in 2014 to accelerate its 10 nm and 7 nm processes, in which project R&D engineers worked around the clock in three shifts.

[38]In addition, there may be another important component of the chip prowess of TSMC (or more broadly, Taiwan), which is the comprehensive technology ecosystem. These are beyond the scope of the present study, but are factors nearly impossible to duplicate.

addition to Intel's contract around 1988, as reported above, TSMC spent great effort on Apple since the early 2010s[39] [40], resulting in fast rising revenues and stock prices.

Figure 6[41] compares the financial performances of TSMC, Samsung and Intel in 2001-2023. Although TSMC (solid line) was smaller than Intel (circled line) and Samsung (dashed line) by net sales until it outdid Intel in 2022, its profit rate has been higher and more stable than the other two. By end-of-year market value, TSMC has already surpassed both firms, and the gap further widens in 2024 (not shown).

The lackluster financial performances of Intel and Samsung in recent years are not surprising, with issues in the markets of their main products, that is, microprocessors (especially AI chips) for Intel, and memory chips, displays and smartphones for Samsung. These problems aside, they have fallen behind TSMC in process technology. Samsung was the first firm to mass-produce 3 nm chips in 2022, but major clients went to TSMC due to better yield and unrivaled packaging technologies[42]. And Intel has been facing problems due to a series of missteps[43]. For instance, the firm was stuck on the 14 nm process for over six years due to multiple delays in developing the 10 nm node. This delay allowed AMD to launch its 7 nm processors ahead of Intel and gained market share by outsourcing to TSMC. By the end of 2023, Intel's market value shrank below that of AMD, which was only about one fourth of Intel's in 2019. Furthermore, between the end of 2023 and September 11th, 2024, Intel's market value dropped by over 60%.

Note that Intel and Samsung are currently the only two IDMs capable of leading-edge chip manufacturing, while other IDMs have resorted to outsourcing. The main reason why Intel and Samsung decide to stay in the race for smaller nodes is that they have internal demand to serve. To justify financially the heavy capital investment, they have to keep up the production volume to a large enough scale by offering foundry services to external users.

Intel's foundry service began in 2010, discontinued in 2017, re-started in 2021, and renamed into Intel Foundry in 2024. With a late start, the presence of Intel Foundry in the world market was still tiny in 2023 at just 1%[44]. Samsung Foundry was launched in 2005. It now ranks number two in global foundry service market (11% in 2023) ahead of UMC and GF, but is small relative to TSMC as well as to Samsung's other businesses[45]. Although clients might be

[39]To get Apple's A8 chip business, TSMC invested heavily in IP, R&D and new fabs in advance. Moreover, it sent a team of roughly one hundred R&D engineers ("One Team") to station in Apple's headquarter to work out the design together with Apple in 2011. It also joined ASML's Customer Co-Investment Program in 2012 to develop the EUV equipment, upon the request of Apple, so as to push its own technology to the furthest [7].

[40]TSMC won the first Apple contract in 2014 to power the iPhone 6, the best-selling smartphone of all time, with A8 chip. Although Apple dual-sourced TSMC and Samsung for A9 chips, the better performance of TSMC won itself the exclusive contract of A10 and later chips, as well as all the M series chips so far.

[41]In Figure 6, the numbers are from https://companiesmarketcap.com.

[42]For example, CoWoS and InFO, TSMC's advanced packaging technologies, have helped to boost chip performance.

[43]The missteps include missing out on the iPhone, underinvestment in EUV equipment, talent churn, slow response to the AI boom and so on.

[44]Also, because of the heavy investments in building new fabs with expensive equipment, such as the world's first High-NA EUV machine, the Intel Foundry registered a loss of $7 billion in 2023.

[45]In 2022, 9.2% of Samsung's total revenue were estimated to come from System LSI (mainly foundry services), which only contributed 6.9% of total operating profits.

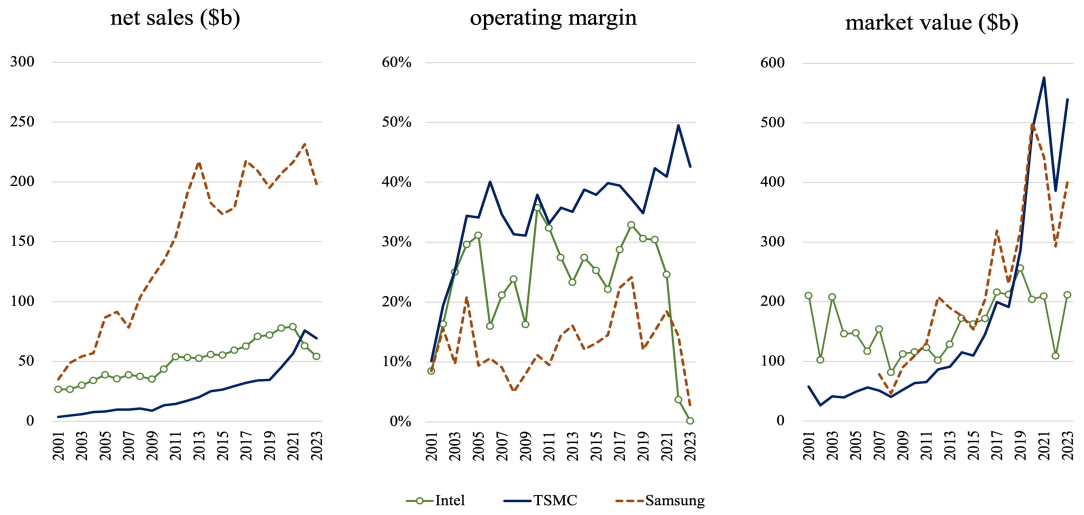

FIGURE 6. Intel, Samsung and TSMC, 2001-2023

happy to secure an additional supplier to have more leverage in price negotiation with TSMC, Samsung often has to offer low prices[46] or package sales[47] to lure away clients from TSMC.

Both Intel and Samsung face a high barrier formed by TSMC. First and foremost, TSMC does not compete with its clients, but Intel and Samsung do. Both Intel and Samsung have separate profit-and-loss statements for their foundry divisions, but still own 100% of these foundries, at least until the mid 2024[48]. That suggests they remain as rivals to many potential customers, such as Apple, Broadcom, and Qualcomm.

Secondly, TSMC has a much larger customer base (500 plus active clients in 2022) than Samsung (around 150) and Intel (just a few). By economies of scope, having more partners means more interactions and learning. Although TSMC has to provide mass-production and small-scale customization at the same time, which is complicated, it has always managed it well. In contrast, Intel has been producing solely for itself for years, it might not be easy to produce anything different from its sophisticated yet niche processes or technologies.

Third, TSMC has built trustful relationships with its fabless customers over the years. As product stability or reliability is valuable, the switching cost for a client to go to a new foundry could be quite dear. For example, if AMD is to leave TSMC for Samsung Foundry, it would need to redesign its chips at a price of hundreds of millions of dollars. Samsung Foundry's CTO, Jeong Ki-tae, estimated that it would take about 3 years to persuade a new client to make an order.

[46]For example, Samsung's price offer to Intel was "reportedly about 60% level of TSMC's", when Intel switched from TSMC to Samsung in 2019 after TSMC started to produce for Intel's rival AMD, according to BusinessKorea.

[47]For example, Samsung uses Qualcomm's chips in some of its smartphones, in exchange for the manufacturing contracts from Qualcomm, who has been its number one strategic partner.

[48]For years, many Wall Street people have urged Intel to spin off its manufacturing arm into a separate company to regain competitiveness in both design and manufacturing.

Although TSMC seems to lead in process technology for the moment, results of the foundry race are not scripted. In addition to forceful support from the US government, Intel has announced an "IDM 2.0" strategy to reclaim its crown in a couple of years, and has received the first High-NA EUV kit from ASML, which represents a key technological advancement. Samsung has also made many new efforts, such as the initiation of an AI semiconductor "turnkey service", managing to encompass design, production, and advanced packaging.

2.5. **Geographic distribution evolves.** Parallel to the long-term trend toward vertical and horizontal disintegration in the semiconductor industry is the ever-changing geographic distribution. It begins with the *US era* (the 1960s and 1970s). Since IC was a US invention, it is no surprise that American IDMs led and dominated the market during these years. American firms also led in related markets, e.g., IBM and Motorola in computer and telecommunication markets, and Perkin-Elmer and GCA in the lithography equipment market.

Then came the *Japan era* ranging throughout the 1980s and into the mid 1990s. This took place through a series of effective Japanese industrial policies, such as the VLSI Project (1976-1980), cheap loans arranged through the government, and so on. Chips made in Japan were of far better quality than US-made chips, according to an HP study in 1980[49]. Facing falling price and loss of profitability, Intel had to withdraw from DRAM in 1985, which product it introduced to the world in 1970 and once accounted for over 90% of its sales revenue [3]. In the peak year 1988, Japan accounted for 50% of worldwide sales of semiconductor, and 6 out of the top 10 semiconductor companies were from Japan according to Gartner.

The *US-East Asia era* began in the late 1990s. The rise of Japan triggered intense reactions from the US firms and government, such as a sequence of dumping law suits and the 1986 US-Japan Semiconductor Agreement. Japan's lead in chips eroded accordingly, so was the case in electronics products and semiconductor equipment. The share of Japanese IC firms in global market dropped from 50% in 1988 to 9% in 2022[50]. Around the same time, the launch of TSMC initiated the co-evolutional path of the fabless (mainly American firms) and the foundries (mostly non-Japanese Asian firms). Henceforth, American IDMs and fabless prospered (along with equipment makers), Taiwanese foundries and design firms thrived. Moreover, Korean IDMs flourished, especially in memory chips to take over Japan's previous shares, and Chinese latecomer IC firms bloomed under strong government support. In this *US-East Asia era*, "most of the world's chips are designed in the US but manufactured in East Asia," as summarized succinctly by Chris Miller in his book, Chip War[51].

Two ongoing events have opened up a new era. Firstly, semiconductor nationalism arises directly from the bitter US-China trade battle, along with the unprecedented pandemic and other events that have disturbed the global supply chain. With the semiconductor industry being the primary area of US-China technological competition, China's rapid catching-up in this industry

[49]American chip makers also found Japan-made lithography machines much better than those made by American firms.

[50]Japanese firms are still strong in chipmaking materials and equipment today. For example, Kanto Denka Kogyo and Resonac Holdings, two SMEs, together commanded over half of the global market for etching gas in 2023.

[51]By value-added, the US contributed 38% of the overall value chain in the world in 2022, followed by Korea (12%), Japan (12%), Taiwan (11%) and China (11%) [9].

seems to slow down, at least for now[52]. But the central government in China did not sit waiting and has continued with generous support for domestic firms. The US, while stepping up its pressure on Chinese IC manufacturers, seeks aggressively to encourage R&D and to revitalize domestic manufacturing, mainly by the CHIPS and Science Act with key incentives amounting to $39 billion in grants. It is projected that its overall share in global wafer fabrication capacity will increase from 10% to 14% during 2022-2032, and its share in sub-10 nm logic chip capacity from 0% to 28% [9]. Other major players (Europe, Japan, Korea, and Taiwan) act in the same direction and some have set similarly ambitious goals [9], while new players, like India, are anxious to join the race.

Secondly, the arrival of AI's "iPhone moment" is reshaping drastically both the demand and supply of the global semiconductor industry. Some firms are obviously doing better than others. For instance, Nvidia, who spearheaded advancements in GPU technology, saw its market value increased more than 7 times between the end of 2022 and the end of June, 2024[53]. Many of its suppliers and partners, including TSMC, also prospered by leaps and bounds. But other firms with weaker links to the AI wave worry to be sidelined, and some are seeking assiduously to strengthen these links. In a nutshell, these new developments suggest the semiconductor race continues to be relentless and it is never certain who will have the last laugh.

3. The Models

3.1. **The basics.** A series of models is presented to highlight some of the critical developments delineated in Section 2. A number of assumptions are made to help focusing on the basic economics behind.

Suppose that there are three players in the semiconductor industry: IDM, Fabless, and (pure-play) Foundry, denoted by I, A, and F, respectively. The production of chips involves design by IDM/Fabless, and fabrication by IDM/Foundry. There are two types of R&D expenditures. Denote the R&D expenditure in design by IDM and Fabless by R_d^I and R_d^A, respectively, and the R&D expenditure in manufacturing by IDM and Foundry by R_m^I and R_m^F, respectively. Assume that the price of chips per unit, p^I (by IDM) and p^A (by Fabless), is determined by how much R&D is spent on design, and the production cost per unit, c^I (by IDM) and c^F (by Foundry), depends on how much R&D is devoted to manufacturing. That is, $p^I = p(R_d^I)$, $p^A = p(R_d^A)$, $c^I = c(R_m^I)$, and $c^F = c(R_m^F)$.

Suppose the market demand is q^I for IDM and q^A for Fabless, and the fee per unit for production service by IDM/Foundry is s, which can be decided by either cost-plus or other pricing rules, but is assumed fixed for now. The construction cost of a chip manufacturing fab, generally massive, is denoted by K (to be paid by IDM or Foundry). When the Fabless designer farms out to a third-party manufacturer, an extra transaction/coordination cost T is incurred. We assume

[52]For example, the EU aims to double its chip production share in the global market to 20% by 2030, from 9.38% in 2023 (calculated by WSTS data). Key incentives amount to $47 billion in grants. And Japan aims for tripling the value of its chip manufacturing from $48.2 billion in 2023 to $112 billion in 2030, with key incentives amounting to $17.5 billion grants [9].

[53]The increase is not monotonic and the fluctuation can be large. Specifically, Nvidia's market value dropped from $3.04 billion by June 28th, 2024 to $2.87 billion on September 11th, 2024, but was still much higher than the end-of-year value at $1.22 billion in 2023, and $364 million in 2022, according to https://companiesmarketcap.com.

	*I*DM*	F*A*bless	*F*oundry
R&D cost in design	R_d^I	R_d^A	
R&D cost in manufacturing	R_m^I		R_m^F
price per unit	$p^I = p(R_d^I)$	$p^A = p(R_d^A)$	
average production cost	$c^I = c(R_m^I)$		$c^F = c(R_m^F)$
units manufactured	q^I	q^A	q^A
foundry service fee per unit		s (paid)	s (received)
capital expenditure	K		K
transaction cost		T	
TOTAL COST	$C^I = q^I c^I + R_d^I + R_m^I + K$	$C^A = q^A s + R_d^A + T$	$C^F = q^A c^F + R_m^F + K$
TOTAL PROFIT	$\Pi^I = q^I(p^I - c^I)$ $-R_d^I - R_m^I - K$	$\Pi^A = q^A(p^A - s)$ $-R_d^A - T$	$\Pi^F = q^A(s - c^F)$ $-R_m^F - K$
endowment constraint	$R_d^I + R_m^I + K \leq \bar{E}^I$	$R_d^A + T \leq \bar{E}^A$	$R_m^F + K \leq \bar{E}^F$

TABLE 2. Relevant Variables and Parameters for IDM, Fabless, and Foundry

that T is paid by the Fabless[54]. Finally, each firm has a financial endowment constraint which the sum of its capital investment and R&D expenses cannot go beyond. The relevant variables and parameters are summarized in Table 2[55].

Next we provide an illustrative description of the development path of the semiconductor industry stated in Section 2.

3.2. **Fabless entrance was once constrained by technology.** In the 1960s and 1970s, all semiconductor firms were IDMs, because design and manufacturing needed to be fully integrated. In other words, it would be too expensive to coordinate between design and fabrication conducted by separate entities.

If there ever were a Fabless A, but without any pure-play foundry F, it would have to outsource to the IDM I. Its profit under the normal practice (without IP leakage and delivery uncertainty) would be as follows, after taking account of foundry service fee, R&D expenditure in design, and the communication cost with the fab (I in this case),

$$\pi^A = q^A(p(R_d^A) - s) - R_d^A - T,$$

assuming for the moment the endowment constraint is met: $R_d^A + T \leq \bar{E}^A$, to which point we will go back later. Assume further that the unit price of chips $p(x)$ is increasing in x, and $p(.)$ is concave. Also assume that, with probability $1-\alpha$, the subcontracted IDM will execute the normal practice, but with probability α, will either steal the trade secret or delay delivery. In the malpractice case, the Fabless will suffer from sales loss at a portion ρ, and obtain profit

$$\pi^A = q^A((1-\rho)p(R_d^A) - s) - R_d^A - T.$$

Note that ρ is the combined share of losses per unit from IP leakage and late or non-delivery. Figure 7 shows the decision tree of the Fabless on whether or not to enter the market.

[54]In reality, the cost may also be shared by both the foundry and the fabless, or by the foundry alone. See footnote 39 for an example of how much effort TSMC has spent before winning Apple's order.

[55]In Table 2, neither the extra cost or the revenue from the foundry service provided by an IDM is included.

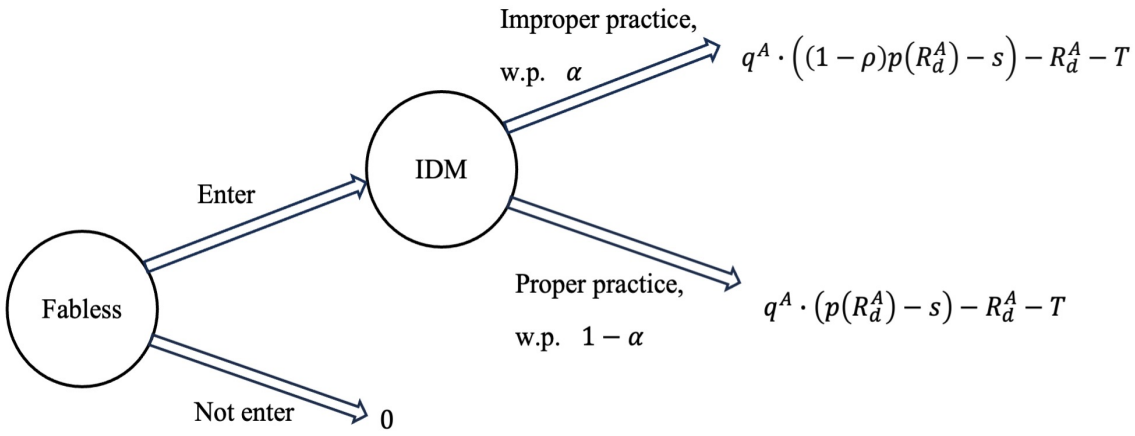

FIGURE 7. Decision Tree of the Fabless

Given the exogenous probability of malpractice by the IDM, the expected profit of the Fabless when entering the market will be

$$\pi^{A,exp} = q^A((1-\alpha\rho)p(R_d^A) - s) - R_d^A - T. \tag{1}$$

Profit maximization requires that

$$(1-\alpha\rho)q^A p'(\bar{R}_d^A) - 1 = 0. \tag{2}$$

Recall that $p(x)$ is increasing in x and $p(.)$ is concave. For any fixed industry and market conditions α, ρ, q^A, and $p(\cdot)$, there exists $\bar{R}_d^A > 0$ such that (2) holds. For any Fabless to enter the industry and survive in the long run, a necessary condition will be

$$\Pi^{A,exp} = q^A((1-\alpha\rho)p(\bar{R}_d^A) - s) - \bar{R}_d^A - T \geq 0. \tag{3}$$

That is, if (3) is satisfied, a Fabless may enter the market and survive in the long run. Now suppose the transaction cost T is prohibitively high, it would be extremely unlikely that (3) holds, and hardly any fabless could survive.

3.3. **Technology progress allowed some fabless to enter.** When the standardization of production processes occurred in the early 1980, the transaction cost was drastically reduced. Although T could still be nonzero, as explained in footnote 11, for illustrative simplicity, we assume that T becomes negligible. The expected profit of the Fabless when entering the industry is now

$$q^A((1-\alpha\rho)p(R_d^A) - s) - R_d^A.$$

Recall (2) that profit maximization requires that

$$(1-\alpha\rho)q^A p'(\bar{R}_d^A) = 1.$$

The condition for the Fabless to make profit, or to survive, becomes

$$q^A((1-\alpha\rho)p(R_d^A) - s) - R_d^A \geq 0. \tag{4}$$

That is, apart from the optimal R&D investment $\bar{R}_d^A$, a Fabless with any R&D investment R_d^A that satisfies (4) will make (non-negative) profit and survive in the industry. For any fixed industry

and market conditions $\alpha, \rho, q^A, p(\cdot)$ and s, let $\underline{R}$ and $\bar{R}$ be such that

$$q^A\big((1-\alpha\rho)p(R_d^A)-s\big)-R_d^A \geq 0 \quad \text{iff} \quad R_d^A \in [\underline{R}, \bar{R}].$$

Now without the huge transaction cost T as an insurmountable hurdle, some fabless firms, with financial endowment between $\underline{R}$ and $\bar{R}$, can enter the industry. This is what happened in the last few years of Stage 1. Within this feasible range, if a firm's financial endowment is high enough to cover the profit-maximizing R&D level, that is, $\bar{E}^A \geq \bar{R}_d^A$, it may invest at the optimal level; or it may spend above the optimal level but not exceeding $\bar{E}^A$, if there are other considerations, such as muscling for market share. If, however, it is unable to raise sufficient fund for the optimal R&D level, it will invest whatever fund it has, as long as it is above the lower bound of survival, $\underline{R}$[56].

3.4. **A fabless would choose a pure-play foundry as a partner over an IDM.** The emergence of a pure-play foundry changes the rule of the game, by distinguishing itself from an IDM with a non-competition guarantee. Hence the fabless does not have to worry about the potential losses from the malpractice of the IDM, if it partners with a pure-play foundry. In such case, the profit earned by the Fabless will be $q^A(p(R_d^A)-s)-R_d^A$. Through profit maximization we obtain the optimal R_d^{A*} such that

$$q^A p'(R_d^{A*}) - 1 = 0. \tag{5}$$

For the Fabless to survive in the market, the necessary condition is

$$q^A(p(R_d^A)-s)-R_d^A \geq 0. \tag{6}$$

That is, apart from the optimal R&D investment R_d^{A*}, a Fabless with any R&D investment R_d^A that satisfies (6) will make (non-negative) profit and survive in the industry. For any fixed industry and market conditions $q^A, p(\cdot)$ and s, let $\underline{R}^*$ and $\bar{R}^*$ be such that

$$q^A\big(p(R_d^A)-s\big)-R_d^A \geq 0 \quad \text{iff} \quad R_d^A \in [\underline{R}^*, \bar{R}^*]. \tag{7}$$

Note that by (2) and (5), we have

$$p'(\bar{R}_d^A) = \frac{1}{(1-\alpha\rho)q^A} \text{ and } p'(R_d^{A*}) = \frac{1}{q^A}.$$

As $\frac{1}{(1-\alpha\rho)q^A} > \frac{1}{q^A}$, by the fact that $p'(\cdot)$ is concave, it follows that $\bar{R}_d^A < R_d^{A*}$. Moreover, it is not difficult to see that any R_d^A that satisfies (4) also satisfies (6). We then have

$$[\underline{R}, \bar{R}] \subset [\underline{R}^*, \bar{R}^*]. \tag{8}$$

That is to say, after the emergence of pure-play foundry, the market may accommodate a broader range of fabless firms regarding feasible R&D investment, from $[\underline{R}, \bar{R}]$ before 1987, to $[\underline{R}^*, \bar{R}^*]$ afterwards. This is illustrated in Figure 8, with Π^A indicated by the difference of the curves and the 45-degree line. Here the curve $C_1 : q((1-\alpha\rho)p-s)-T$ represents the fabless firm's profit before deducting R_d^A in the early part of Stage 1. In this specific graphical example, C_1 lies completely below the 45-degree R_d^A line, suggesting that no fabless firms can survive. Curve $C_2 : q((1-\alpha\rho)p-s)$ refers to Π^A before deducting R_d^A, when $T=0$. As illustrated, some

[56]We allow the R&D investment R_d^A, within the range $[\underline{R}, \bar{R}]$, to differ from the optimal $\bar{R}_d^A$ to accommodate the possibility that a Fabless may choose its R_d^A based on considerations other than profit maximization. Same issues apply to $[\underline{R}^*, \bar{R}^*]$ and $[\underline{R}^{**}, \bar{R}^{**}]$ in Subsections 3.4 and 3.5 below.

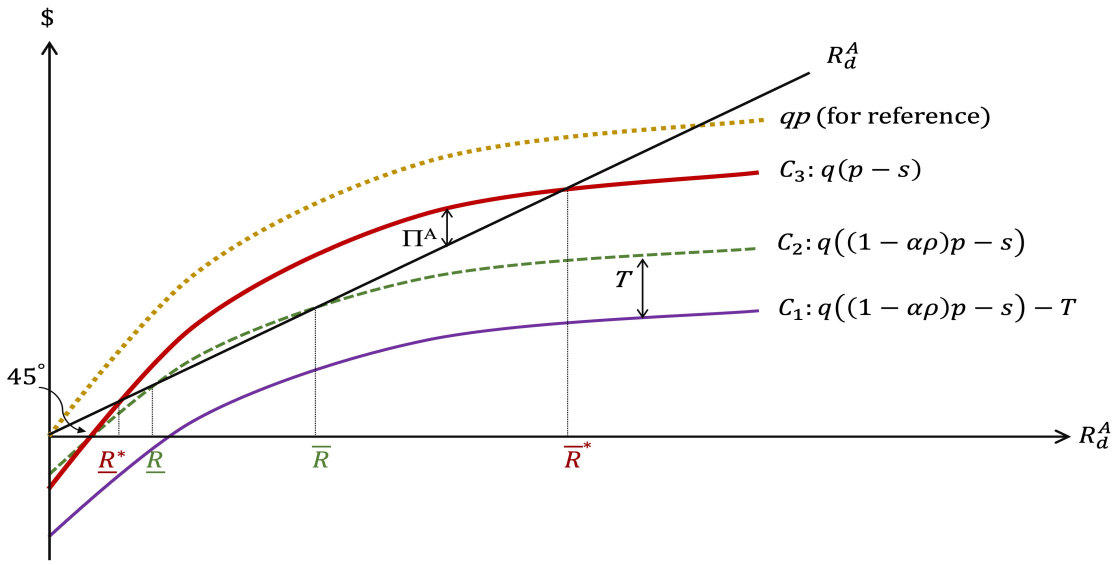

FIGURE 8. Ranges of Fabless with Non-negative Profits

fabless would enter in the later part of Stage 1. Then $C_3 : q(p-s)$ refers to Π^A at Stages 2 and 3 (with no potential losses from IDM's malpractice), before deducting R_d^A; more fabless firms will survive.

3.5. **Co-evolution of the fabless and the foundry.** The expansion of the fabless segment, in firm number and in size, creates positive impacts on foundry firms, which, in turn, provides more incentives for potential fabless firms to enter the market.

Recall that the profit of the Foundry is

$$\Pi^F = q^A\left(s - c(R_m^F)\right) - R_m^F - K.$$

Note that $c(x)$ is a decreasing and convex function of x. Optimization of Π^F requires

$$q^A c'(R_m^{F*}) = -1. \tag{9}$$

When the capacity range for the Fabless's investment in R&D extends from $[\underline{R}, \bar{R}]$ to $[\underline{R}^*, \bar{R}^*]$, the number of fabless firms would increase accordingly. This also implies higher demand for production of chips. Suppose the number of fabless firms grows and induces the demand to rise from q^A to γq^A with $\gamma > 1$. The profit of the Foundry then becomes $\gamma q^A \cdot (s - c(R_m^F)) - R_m^F - K$.

$$\gamma q^A c'(R_m^{F**}) = -1. \tag{10}$$

By the assumption that $c(x)$ is a decreasing and convex function of x, it is not difficult to see from (9) and (10) that $R_m^{F**} > R_m^{F*}$. That is, the Foundry will invest more on manufacturing R&D, and unit production cost will go down[57]. If, however, the Foundry is unable to invest to the optimal level due to endowment constraint, it can make the investment in the next period from profits earned in this period. This is what TSMC has been doing to upgrade itself period after period, as described in Subsection 2.4.

[57]There are other possible channels for unit production cost to go down. One is through scale economies realized by larger capital investment. Another is through learning by doing over time: the more one produces, the more skilled one becomes, and the lower the unit production cost becomes.

Note that higher manufacturing R&D may bring changes other than lowering unit cost, such as better quality (hence higher unit price p) or higher service fee s. Without going further, we assume the increase of quantity from q^A to γq^A raises the marginal profit $(s-c(R_m^F))$, whether s rises or falls.

The lower production cost $c(R_m^{F**})$ may generate positive feedbacks on the industry, as the Foundry may be incentivized to lower the service fee s that we assume fixed so far, so as to grab a larger market share, for example. This may extend further the capacity range for the Fabless's investment in R&D. To see this, recall (7) and replace the service fee s by $s' < s$:

$$q^A(p(R_d^A)-s')-R_d^A \geq 0 \text{ and } R_d^A \in [\underline{R}^{**}, \bar{R}^{**}]. \tag{11}$$

That is, any $R_d^A \in [\underline{R}^*, \bar{R}^*]$ also satisfies (11). We then have

$$[\underline{R}^*, \bar{R}^*] \subset [\underline{R}^{**}, \bar{R}^{**}]. \tag{12}$$

By the above arguments, the feasible range of the Fabless's R&D investment can be described as follows. There exists a feedback loop in the semiconductor industry after the introduction of pure-play foundries: (a) The number of fabless firms increases; (b) This induces more demand for manufacturing service, which results in lower production cost, and perhaps also lower service fee (or higher price due to better quality); (c) Lower unit service fee would expand the feasible range of R&D investment of the fabless, allowing more fabless firms to enter the industry. In short, the introduction of pure-play foundries ignites the co-evolution of fabless firms and pure-play foundries, by forming a self-reinforcing loop between steps (b) and (c).

3.6. **To be an IDM or a Fabless?** "Real men have fabs," AMD founder Jerry Sanders (CEO 1969-2002) made the point that maintaining control over all segments of IC production was necessary for a top-tier semiconductor company. This remark makes perfect sense in the early part of Stage 1, but is proven wrong when AMD turned fabless in 2009 in Stage 3. The question is, if a firm has the option to choose its organization form with no financial constraint, would it rather be an IDM or a Fabless?

Suppose the demand of the chips in the market is q and there are two competing modes of production: the IDM model and the Fabless-Foundry model. The profit for the IDM, Π^I, without counting the income from the foundry service it provides to other firms, will be

$$\Pi^I = q \cdot \big(p(R_d^I) - c^I(R_m^I)\big) - R_d^I - R_m^I - K.$$

The profit for the Fabless differs by stage. In Stage 1, it is

$$\Pi^A = q \cdot \big((1-\alpha\rho)p(R_d^A) - s\big) - R_d^A - T.$$

If T is overly huge, the Fabless will not survive. In the second part of Stage 1 when T reduces to zero as we assume, Π^A becomes

$$\Pi^A = q \cdot \big((1-\alpha\rho)p(R_d^A) - s\big) - R_d^A.$$

Then in Stage 2 and Stage 3, when the Fabless outsources to the pure-play Foundry and is rid of the losses incurred by the malpractice of the IDM, Π^A is

$$\Pi^A = q \cdot \big(p(R_d^A) - s\big) - R_d^A.$$

In any stage, a new entrant would choose the mode that generates more profit, unless it is financially constrained. Specifically, it would choose to be an IDM if $\Pi^I > \Pi^A$, that is, if the combined savings from avoiding third-party subcontractor's improper practice (related to α and

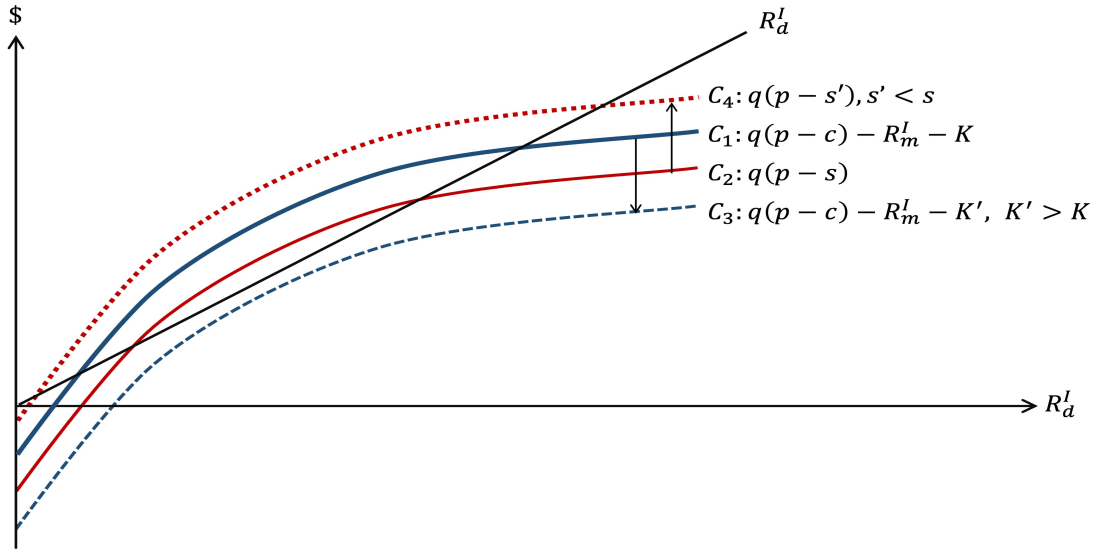

FIGURE 9. When An IDM Decides to Become Fabless

ρ), and from producing in-house rather than farming out (related to s and c), plus the savings of the transaction cost T, are high enough to cover the combined investment expenses on R_m^I and K. With hindsight, the total savings must have been enormous, so that all players, incumbents and new entrants, were IDMs until the mid 1980s, including Intel who was launched in 1968.

As discussed in Subsection 3.3, a handful of fabless firms who previously had no access to the market entered by the end of Stage 1 (e.g., C&T), when the transaction cost significantly reduced. These are the startups who could tolerate the extra costs involved with outsourcing, that is, the sales loss from IDM's malpractice and the markup included in foundry fees, but were unable to pay for the R&D in manufacturing and capital investment. In other words, these firms could survive as fabless firms when $\Pi^A > 0$, but even if $\Pi^I > \Pi^A$, they could not afford to become IDMs due to inadequate financial endowment, that is, $R_d + R_m + K > \bar{E} \geq R_d$.

Then in Stage 2 and Stage 3, the availability of pure-play foundries liberated the fabless from the risks of IP leakage and delivery uncertainty, thus the potential advantage of being IDM over fabless declined. More fabless firms would enter.

What is more, in certain situations, it may be cheaper for an IDM to outsource than to produce in-house. This may take place for various reasons, either the pure-play foundries have better manufacturing technology or enjoy scale economies, or the cost to build a fab gets unaffordable. As mentioned above, some IDMs like AMD, IBM, and LSI Logic have transformed into fabless firms.

Figure 9 is a graphical illustration of what happens in Stage 2 and Stage 3, assuming $\Pi^I > \Pi^A$ so that curve C_1 lies above $C2$. As s becomes smaller (in Stage 2) when the production cost in pure-play foundries declines, but not in IDMs, C_2 may shift upwards to C_4. And if the cost to build a new fab becomes very high (in Stage 3), C_1 shifts downwards to C_3. It would then be rational for a firm to choose to become a fabless (C_4), rather than an IDM (C_3).

4. Concluding Remarks

4.1. **Summary of major findings.** We begin this paper with a brief introduction of the IC industry, by reviewing the semiconductor ecosystem and three types of firms, IDM, fabless, and pure-play foundry, with different functions and market concentration.

We then review the TSMC saga in three periods against the broader background: namely, pre-TSMC period (first IDM domination and then fabless advent), emergence of the pure-play foundry (self-discovery of comparative advantage and coevolution of the fabless and foundries), and further specialization (both vertically and horizontally). These developments are supported by real-world episodes and statistics from firm reports and academic studies. A brief account of how TSMC upgrades itself is then given, followed by a review of how geographical distribution of the global semiconductor value chain advances over time, with some discussions of what may take place next.

A number of critical decisions in this evolutional path for growth are studied by simple models, corresponding to specific phenomena. These include: *when* a fabless decides to enter the market, *which* partner it would choose, *how* the foundries co-evolve with the fabless over time, and *why* an IDM might prefer to go fabless.

4.2. **Implications.** Our findings suggest that there is a rapid growth of the entire semiconductor in the last few decades, and a trend toward vertical as well as horizontal disintegrations in its production value chain. In economics terms, the innovative pure-play foundry model addresses a common coordination failure in the world economy, by activating a latent productive potential of the fabless chip designers. The co-evolution of the foundries and the fabless not only enhances mutual benefits for both the fabless and the foundries, but also prompts the IDMs to transform and specialize, resulting in overall efficiency improvement in the semiconductor industry.

All these began with the founding of TSMC in 1987 as a pure-play foundry, which exemplifies the "self-discovery of comparative advantage", which was also the least evil choice Taiwan could make. Note that the launch of TSMC and ensuing fabless-foundry co-evolution, and the trend toward further disintegration did not proceed with a standard Walrasian mechanism in the Debreu style. Rather, it is a dynamic Schumpeterian process of creative destruction, where entrepreneurship and innovations are indispensable. Creative destruction has, on balance, led to greater prosperity, though the fruits are not distributed equally to all.

The developments in the real world are much more complicated than is analyzed here, these will be the subjects for future studies. To take stock of our analysis, the TSMC case demonstrates that market force keeps the economy growing, but innovation is always essential to upgrade market efficiency to a higher level. Although here we look into the semiconductor industry, the implications should apply to many other modern industries and to many nations as well.

Addendum

This year of 2024 is a key election year in the US. One of the two president candidates, Mr. Donald Trump said in a Businessweek interview in July that "Taiwan took our chip business from us", "it doesn't give us anything," and "we should have never let it happen".

Here we address the issue of whether Taiwan has taken away the IC business from the US, basing on implications from our research. In the 1980s, the semiconductor business of the US was indeed taken away by Japanese firms, who competed head-to-head with American firms in design, manufacturing, final products and equipment. But the US-Taiwan relation in IC, or more precisely the US-Taiwan nexus in IC, is a different story. Rather than taking jobs away from American firms, the launch of TSMC in 1987 unlocked the opportunities for many fabless chip designers to enter the market, and most of them are American firms. Nvidia CEO Jensen Huang once said "Nvidia would not be possible without TSMC". In fact, there are many more Nvidia's, like AMD, who are made more successful by partnering with TSMC. That is to say, the growth of TSMC (and Taiwan, in general) is creating jobs in the US, not replacing jobs.

The next question is, by how much should the US rebuild its IC manufacturing capacities. To strengthen the resilience of wafer fabrication in the US, President Biden has already pushed TSMC and its American alternatives to start the slow remedial process of building fabs in the US. The real question is: what is the reasonable goal to set? According to SIA, with the concerted efforts of the governments and companies, American fab capacities are expected to rise from 10% of world total in 2022 to 14% in 2032, and from 0% to 28% for advanced node logic chips [9]. This expansion, if realized, is surely helpful for American national security and economic resilience. But it comes with a price. Morris Chang estimated that it might be twice as expensive to make chips in the US than in Taiwan. It is not clear if American customers will buy costlier US-made chips over Taiwan-made chips. Furthermore, when chips get more expensive, their pervasiveness will reduce, which would slow down the growth of the IC industry and hold back the growth momentum of the entire world, the US included.

The final question is which area should the US semiconductor firms focus their resources on. The obvious answer is IC design or other R&D-intensive activities, like EDA and core IP, in which the US already has an enormous competitive advantage. The current AI era was led in by American firms, like Nvidia, with their great innovations in design. Earlier on, Intel became the world's top semiconductor firm in 1992-2006, after it turned its focus on the more design-intensive microprocessors, following its forced exit from the less design-intensive DRAM business in 1985. It follows that the next success example is most likely to be based on designs and innovations as well. At the national level, history has proven that industrial policies that give full play to national strengths work best. This is what Chang has figured out when launching TSMC in 1987. And this may be what the US government should do today.

Acknowledgments
The authors are thankful to the editor and the anonymous referee for useful and insightful comments for the improvement of the paper.

REFERENCES

[1] M. Balconi, R. Fontana, Entry and innovation: An analysis of the fabless business, Small Business Economics 37 (2011), 87-106.

[2] L. Berlin, The Man Behind the Microchips: Robert Noyce and the Invention of Silicon Valley, Oxford University Press, New York, 2005.

[3] R. A. Burgelman, Fading memories: A process theory of strategic business exit in dynamic environments, Administrative Science Quarterly, 39 (1994), 24-56.

[4] D. Fuller, A. I. Akinwande, C. G. Sodini, Leading, following or cooked goose? Innovation successes and failures in Taiwan's electronics industry, Industry and Innovation 10 (2003), 179-196.
[5] R. Hausmann, D. Rodrik, Economic development as self-discovery, Journal of Development Economics 72 (2003), 603-633.
[6] A.-C. Tung, H. Wan, Jr., Industrial policy: Chinese debate to Taiwan foundries, Singapore Economic Review 64 (2019), 1057-1080.
[7] A.-C. Tung, H. Wan, Jr., Organisational Investment: The case of ASML – Can the product make the producer? Foreign Trade Review 58 (2023), 176-191.
[8] A. van Agtmael, table From imitators to innovators – Taiwan's TSMC and high tech computer win by reinventing industries and products, The Emerging Markets Century, pp. 119-139, Free Press, New York, 2007.
[9] R. Varadarajan, et al., Emerging Resilience in the Semiconductor Supply Chain, SIA, Washington, DC, 2024.