Technology Rivalry and Resilience Under Trade Disruptions: The Case of Semiconductor Foundries*

Weiting Miao†

May 2025

Abstract

This paper studies the impact of industrial policies on technology competition and consumer welfare amid rising global trade disruption risks. Distilling key empirical features from novel data on the semiconductor foundry industry, I develop and estimate a dynamic oligopoly model that integrates step-by-step innovation, trade disruption risk, and industrial policies. While distortions from market power and technological externalities justify subsidies, their optimal levels depend on the magnitude of trade disruption risk: when the risk is low, the optimal subsidy rate remains low, as the welfare benefits are distributed globally, but the costs are borne exclusively by the subsidizing government. My quantitative model shows that a 35% trade disruption risk makes the 25% investment subsidy under the U.S. CHIPS Act optimal, resulting in a 6% welfare improvement for the U.S. The paper also analyzes the CHIPS Act's restrictions on investments in rival countries, intended to secure technological leadership against their firms. Its efficacy depends on the strength of technology spillover restrictions and the scale of the rival home market secured for rival firms.

Keywords: Industrial Policy, Technology Competition, Resilience, Trade Disruption

JEL classification: F13, F23, L13, O31, O38

*I am grateful to Laura Castillo-Martinez, Allan Collard-Wexler, Matthias Kehrig, Daniel Yi Xu, and Kei-Mu Yi for their constant guidance and support. I also thank Jeff Thurk and Cheng Chen for their insightful discussions as well as conference and seminar participants. I benefit from conversations with Qiang Miao and Jingyan Zhang for institutional knowledge.

†Duke University, Department of Economics. Email: weiting.miao@duke.edu.

1 Introduction

Amid rising geopolitical tensions and more frequent supply chain disruptions, industrial policies are being used with renewed urgency to pursue technological leadership in critical industries and bolster economic resilience against deglobalization risks. Policymakers are increasingly targeting concentrated and global sectors like semiconductors and electric vehicle (EV) batteries. Yet, the effectiveness of these interventions remains contested: Do they genuinely foster technological advancement for domestic firms, or do they risk creating inefficiencies and allowing benefits to leak abroad through trade? This paper examines these questions, offering new insights into how industrial policies impact both global competition and domestic welfare.

This paper studies industrial policies in the semiconductor industry. How consumer welfare across different countries and technology competition in this highly concentrated industry respond to industrial policy depends on the nature of innovation and production location decisions in the context of possible trade disintegration. To quantitatively assess these policy effects, I develop and estimate a dynamic oligopoly model of step-by-step innovation and capacity investment subsidies across different locations when global trade is potentially disrupted.

The semiconductor foundry industry offers an ideal setting to examine the interaction between trade disruptions and industrial policies, as well as their implications for technology competition. This industry is characterized by continuous innovation, with firms' technology levels being transparently tracked through process names in nanometers. Historically, semiconductor manufacturing thrived in a global free-trade environment, but recent geopolitical tensions have introduced significant trade disruption risks. These challenges have heightened the strategic importance of the industry, prompting a wave of industrial policies, including the U.S. CHIPS Act and the EU CHIPS Act, aimed at bolstering domestic semiconductor production. Studying this industry, therefore, offers both analytical insights and practical relevance for understanding how governments navigate economic resilience and technological competition in a volatile global landscape.

Using comprehensive industry data, I summarize key features of the semiconductor foundry industry, focusing on market structure, innovation patterns, price and investment cost dynamics, and cost structure. The industry is highly concentrated, with continuous consolidation over time, making it essential to consider strategic interactions among a few key players. Several factors drive this concentration. First, most technology upgrades are incremental, with leapfrogging being rare.

As a consequence, firms that fail to innovate for a short period of time are likely to become left behind in the technology race. Second, prices for new technologies drop quickly after introduction, while capacity investment costs remain stable. This dynamic allows only firms with early success to earn returns sufficient to offset their high R&D costs. In contrast, firms far from the technological frontier struggle to catch up, facing slow progress, low profits, which discourages further investment. As a result, no new entrants emerge in advanced chip manufacturing, and firms that exit the competition for advanced technologies rarely reenter. Finally, the industry's cost structure is dominated by R&D and capacity investments, making them central to firms' strategic decisions and the focus of my model.

Building on these stylized facts, I develop a quantitative framework to analyze technology and capacity investment decisions in semiconductor foundries under trade uncertainty. At its core is a dynamic oligopoly model that integrates incremental innovation, cross-firm technology spillovers, capacity investment subsidies, and global trade disruption risks. Firms invest in R&D to advance one step up the technology ladder per period, with follower firms benefiting from lower R&D costs due to spillovers from industry leaders. Given the current technology levels across all firms, those with the relevant technology decide where to build capacity and how much to invest, considering unit costs shaped by local industrial policies. After building capacity, the trade disruption shock is realized, and firms compete in spot markets subject to capacity constraints. During disruptions, firms can only serve markets where they have local production capacities. Without disruptions, an integrated global market allows firms to ship chips freely across locations, subject to total capacity across all their sites.

The model illustrates that both trade disruption risks and industrial policies play a crucial role in shaping firms' innovation incentives and capacity allocation, as they influence market access, investment costs, and expected returns from innovation. These risks also determine how capacity allocation impacts consumer welfare, directly linking trade risks to the design of optimal policies. Even without trade disruptions, market power distortions and technology externalities justify government intervention from a global perspective. However, the optimal subsidy rate of a single country remains low, as the benefits of these subsidies can easily leak abroad through trade, while the cost falls entirely on that country's taxpayers. When trade disruptions become more likely, firms diversify across locations to manage trade shocks, but their profit-maximizing strategies may not align with local consumer welfare. As trade disruption risks increase, the optimal subsidy rate

rises, since subsidies are more likely to incentivize local capacity that directly serves domestic consumers, who become more dependent on local production in the event of disruptions. In terms of technology competition, within an oligopolistic market structure, the impact of policies on firms' innovation behavior depends on the technological standing of all firms. Subsidies for domestic firms provide strategic advantages to those that are already competitive but have limited impact on firms that lag far behind, as catching up is slow, unlikely, and hindered by the absence of leapfrogging innovation.

To quantify the model, I first use the prices of legacy chips as price instruments to estimate the demand for advanced chips. Combing the estimated demand shifter coefficients with shipments to downstream, including personal computers, smartphones, and tablets, I estimated a global market share of 25% for the U.S. and 22% for mainland China. Second, I used external industry reports and investment cost data to calibrate the location-specific and technology-specific capacity costs. My calibration shows that fundamental cost of building capacity in the U.S. is 7.5% higher than in East Asia. Third, using the firm's optimal capacity allocation decisions, I inferred firm productivity from market shares, based on the tight relationship between firm-level capacity costs and market share in the model. Finally, I estimate the dynamic parameters related to innovation, including R&D unit costs across generations and technology spillover from the leading firm to followers, using maximum likelihood estimation based on firm-level technology upgrading history data. The results reveal substantial technology spillovers, with lagging firms only needing to pay 22% of the R&D unit cost incurred by the leading firm.

I use the calibrated model to simulate unilateral capacity investment subsidies in the U.S. and evaluate their impact on consumer welfare in different locations, considering different beliefs about trade disruption risks. I start with a static analysis where technology levels are fixed. For the subsidizing country, investment subsidies boost domestic capacity, reducing prices both with and without trade disruptions. Welfare gains are larger in the disrupted scenario, as home consumers benefit from exclusive access to the additional capacity. The optimal subsidy rate depends on the number of incumbents and the level of trade disruption risk. Fewer incumbents require higher subsidies due to greater market power distortions, while higher disruption risks call for increased subsidies since domestic capacity is less likely to be shared globally. With a 20% trade disruption risk, the optimal capacity investment subsidy rate is 9.4% in the duopoly case and 19.7% in the monopoly case, resulting in welfare improvements of 1.0% and 4.3% in the U.S., respectively. In

other locations, consumers benefit from increased global capacity when trade disruptions do not occur but experience losses during disruptions as capacity shifts to the subsidizing location. The net welfare effect remains positive if the efficiency loss from firm relocation is modest. Therefore, local subsidies are not necessarily beggar-thy-neighbor policies, as their benefits can be shared with other locations through trade when disruptions do not occur. Moreover, firms internalize disruption risks and maintain capacity in other locations when trade risks are high.

I then include endogenous technology upgrading in the welfare analysis to reflect how investment subsidies affect firms' innovation incentives. Because of technology spillovers between firms, there is a tendency to underinvest in R&D, underscoring the need for policy intervention beyond correcting market power distortions. I calculate the optimal subsidy rate for different levels of disruption risk. A disruption risk of 35% implies an optimal 25% capacity investment subsidy rate, which corresponds to the U.S. CHIPS Act, yielding a 6% welfare improvement in the U.S. I further decompose the welfare gains into static gains and dynamic gains driven by accelerated innovation, finding that dynamic gains represent 15% of the overall welfare improvement. This underscores that resilience is the primary factor guiding government subsidy decisions. The innovation channel further enhances consumer welfare gains in other locations from local subsidies. With a 20% trade disruption risk, a 25% subsidy in the U.S. can increase consumer welfare in other locations by 4.6%, with 70% of the gains attributed to innovation.

In addition to direct subsidies, the U.S. CHIPS Act includes guardrails that prohibit recipients from expanding semiconductor investments in "countries of concern", called investment clawbacks. The model enables me to compare the firms' technology trajectories with and without these restrictions. I simulate these investment clawbacks by preventing non-Chinese leading firms from building capacity in mainland China and reducing technology spillovers between Chinese and non-Chinese firms. While these restrictions aim to hinder technological advancement in targeted countries, they can unintentionally help lagging firms in these countries catch up in the technology race by securing domestic market demand during trade disruptions. The overall impact hinges on the effectiveness of technology spillover restrictions, market size of the targeted countries, and the likelihood of trade disruptions.

Related Literature This paper brings together several strands of literature. It relates to the extensive body of work on industrial policies (Brander and Spencer, 1985; Eaton and Grossman,

1986; Maggi, 1996; Choi and Levchenko, 2021; Liu, 2019; Lane, 2022; Juhász, Lane, and Rodrik, 2023; Lashkaripour and Lugovskyy, 2023; Bartelme, Costinot, Donaldson, and Rodriguez-Clare, 2024; Juhász and Lane, 2024), particularly those focused on quantitative analysis of oligopolistic industries(Irwin and Pavcnik, 2004; Barwick, Kalouptsidi, and Zahur, 2021; Chen and Xu, 2023). This paper contributes new insights by structurally analyzing how industrial policies influence innovation dynamics and capacity investment across locations. It also examines how the interaction of these policies with trade disruption risks affects the resilience of local supply and consumer surplus.

Second, this paper contributes to the literature on firm innovation decisions, particularly how industrial and trade policies shape innovation (Miyagiwa and Ohno, 1995; Crowley, 2006; Konings and Vandenbussche, 2008; Ederington and McCalman, 2008; Lileeva and Trefler, 2010; Pierce, 2011). In line with Aghion et al. (2001), the model shows that a firm's incentive to innovate depends on its relative technological position. This paper applies these insights to industrial policies, examining how their effectiveness varies with firms' technological status. The model builds on the empirical industrial organization literature on dynamic oligopoly games of innovation (Pakes and McGuire, 1994; Goettler and Gordon, 2011; Igami, 2017, 2018; Igami and Uetake, 2020; Yang, 2020), integrating capacity investments, subsidies, and trade risks into a unified framework. Its richness, which captures both innovation dynamics and capacity expansion across locations, is essential for evaluating policies like investment restrictions in global technology competition.

Third, this paper engages with the evolving and active literature on supply chain disruption and resilience (Leibovici and Santacreu, 2020; Bonadio, Huo, Levchenko, and Pandalai-Nayar, 2021; Baldwin and Freeman, 2022; Castro-Vincenzi, 2022; Grossman, Helpman, and Lhuillier, 2023; Galdin, 2024). Leibovici and Santacreu (2020) examines ex-post optimal trade and industrial policies to mitigate shortages of critical goods during periods of increased global demand. In contrast, this paper focuses on the role of ex-ante policy interventions in the face of trade disruption risks, emphasizing the role of multinational firms. My paper is closely related to Castro-Vincenzi (2022), which examines how global car manufacturers mitigate climate disruption risks by diversifying plant locations and holding extra capacity. This paper, however, differs by incorporating dynamic innovation decisions and oligopoly competition, with a focus on aggregate trade shocks rather than idiosyncratic productivity shocks. Only a model that includes innovation competition can address how industrial policies influence the technology race across countries. My findings also show that

the innovation channel impacts the optimal local subsidy rate and alters the welfare effects of local subsidies in other regions.

Finally, the paper contributes to empirical studies on the semiconductor industry(Irwin and Klenow, 1994, 1996; Hatch and Mowery, 1998; Flamm, 2019; Thurk, 2020), a critical sector supporting modern computing and receiving significant attention due to national security concerns. Using novel and comprehensive industry data, I summarize key features of the sector. These insights guide the development of a quantitative framework to analyze technology and capacity investment decisions in semiconductor foundries under trade uncertainty. A recent study by Goldberg et al. (2024) explores a similar topic, examining industrial policies in the semiconductor industry through a combination of historical analysis, natural language processing, and a model-based approach to assess the extent and impact of historical government subsidies. While their model specifically emphasizes firm pricing decisions and the role of learning by doing, my model focuses on firm innovation and capacity allocation decisions to explore the policy impacts on technological competition and local supply resilience.

The rest of the paper proceeds as follows. Section 2 provides an institutional background on the semiconductor industry and recent industrial policies in the semiconductor manufacturing sector. Section 3 describes the data and documents key industry features. Section 4 presents a model of innovation and capacity allocation with trade disruption risks. Section 5 outlines the estimation strategy for the parameters of the structural model and presents the estimation results. Section 6 conducts counterfactual exercises to evaluate the ongoing policy. Section 7 concludes.

2 Institutional Background

2.1 Semiconductor Manufacturing Industry

Semiconductors are essential for modern electronic devices such as computers, vehicles, and home appliances. Advancements in semiconductor technology drive technological progress across multiple sectors and fuel to economic growth. Moreover, they are critical for national security, as defense systems rely on these components for operational effectiveness. Semiconductors also underpin emerging technologies like artificial intelligence.

Value Chain The value chain of the semiconductor industry consists of two main components: design and manufacturing. Manufacturing includes wafer fabrication, assembly, packaging, and

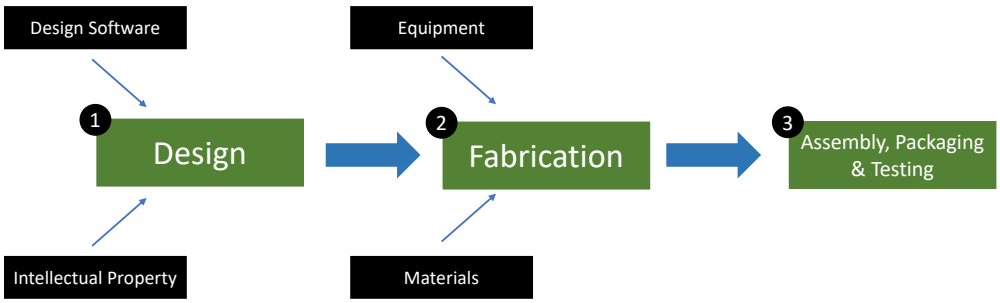


Figure 1: Semiconductor Value Chain

Notes: The plot is adapted from Strengthening the Global Semiconductor Supply Chain in an Uncertain Era (page 5) by Boston Consulting Group (BCG) and the Semiconductor Industry Association (SIA) .

testing, with wafer fabrication being the most technologically intensive and highest value-added process. Fabrication involves creating networks of transistors using layers of different materials on a thin silicon wafer. These wafers can carry many chips that will later be cut and packaged appropriately. This paper focuses on wafer fabrication.

Technology Node Semiconductor fabrication technology can be roughly summarized by the transistor gate length, known as the technology node. These nodes have advanced over time in discrete steps, following a nanometer scale (e.g., 28nm, 14nm, 7nm). Smaller transistor sizes enable greater computing power, faster processing speed, improved energy efficiency, and lower cost per transistor. Since the 28nm generation, the technology node no longer represents the actual physical gate length but remains an informative indicator of technology, especially among foundry providers.[1] There is a formal technology coordination mechanism called the International Technology Roadmap for Semiconductors (ITRS), which was later succeeded by the International Roadmap for Devices and Systems (IRDS). This roadmap provides a unified framework for the industry, ensuring that firms across the supply chain have consistent expectations for future technology developments (Flamm, 2019).

This paper focuses on advanced technology chips for the following reasons: (1) Policies often emphasize leading-edge technology to establish industry leadership; (2) This segment faces higher risks from geographical concentration; (3) Focusing on advanced technology involves interesting dynamics of technological advancement, making the economic analysis more compelling.[2] Advanced chips are primarily used in smartphone and high-performance computing platforms, such as CPUs and

[1] See appendix A.1 for further details.

[2] Another reason is that legacy chips often focus on specialized processes and serve niche markets, making them less comparable even within the same technology node.

GPUs in personal computers, 5G infrastructure, and enterprise data centers. More recently, advanced chips have also been used in automotive platforms for advanced driver-assistance systems.

Fabless-Foundry Model Semiconductor firms historically manage both design and manufacturing, and these firms are known as integrated device manufacturers (IDMs). In the mid-1980s, a new business model emerges with the establishment of Taiwan Semiconductor Manufacturing Company (TSMC). This model introduces contract chip makers, called pure-play foundries, which specialize in fabrication. This allows design-focused firms, known as fabless firms, to avoid the high cost of building and maintaining their own fabrication facilities (called fabs). These dedicated foundries do not design their own products, which alleviates concerns about intellectual property and strategic information for clients, positioning them as service providers rather than competitors. The rise of these third-party foundries, along with the availability of design software and related intellectual property licenses, incentivizes the entry of fabless firms. Meanwhile, as the cost of building and maintaining fabs continues to rise, many IDMs transition to a fabless or fablite model, where the fablite model retains partial manufacturing and outsources the rest. According to Hung et al. (2017), the ultimate market share of fabless foundry model is approximately 45%. Given that this paper focuses on advanced logic chips, the fabless-foundry model is dominant, with Intel and Samsung remaining the only firms that still integrate design and manufacturing.[3] According to SEMI, the capacity share of 12-inch, 32nm and below chips produced by foundries (excluding memory chips) was roughly 65% in 2015, and is projected to grow to 85% by 2026.

Manufacturing Landscape The semiconductor industry was initially pioneered in the U.S., but manufacturing has largely migrated to East Asia since the 1990s. According to Varas et al. (2020), the U.S. and Europe's share of global semiconductor manufacturing capacity plummeted from 80% in 1990 to around 20% in 2020, with East Asia emerging as the dominant hub due to lower costs and more compelling government incentives. Among the top five semiconductor foundries - TSMC, Samsung, GlobalFoundries, United Microelectronics Corporation (UMC), and Semiconductor Manufacturing International Corporation (SMIC) - only GlobalFoundries is based in the U.S., with the rest in East Asia: TSMC and UMC in Taiwan, Samsung in South Korea, and SMIC in mainland China. Except for GlobalFoundries, these companies predominantly concentrate their manufactur-

[3]There are three main types of semiconductors: logic, memory, and discrete, analog, and others (DAO). The figure from Hung et al. (2017) includes memory chips, which are outside the scope of this paper. Memory chips are predominantly produced within the IDM model. DAO chips are highly specialized, difficult to compare across products, and rely mainly on mature technologies.

ing capacity in Asia until recently, when trade disruption risks became a more pressing concern.[4]

Transition to Global Expansion With increasing geopolitical risks and growing concerns over trade disruptions, firms are adjusting their production strategies from concentration to global diversification. TSMC, the industry leader, illustrate this trend. As of 2020, most of its manufacturing facilities were in Taiwan, with only 3 out of 12 fabs located elsewhere, representing less than 10% of its total capacity, none of which featured leading-edge technology. The only fab outside Asia was built in 1996 and received no further investment due to significant cost differences.[5] Reflecting this focus on cost efficiency, TSMC noted during its 2010 Q4 earnings call that fab location was a minor consideration, as free trade was the prevailing belief in the industry at the time.

However, in the face of shifting geopolitical dynamics and evolving industrial and trade policies, TSMC has transitioned to a more global expansion strategy in recent years. In 2020, it announced a new fab in Arizona, followed by plans in 2022 for a second. On the same day it was listed to receive $6.6 billion in CHIPS Act funding, TSMC announced a third Arizona fab. Its international expansion also includes new fabs in Japan, announced in 2021 with additional investments confirmed in 2024, and a 2023 joint venture with European firms to establish fabs in Germany. According to its 2022 Q4 earnings call, TSMC's strategic decisions are driven by customer needs, including the value of geographical flexibility, and the level of available government support.

2.2 Industrial Policies in Semiconductors

Government interventions have been crucial to the development of the semiconductor industry (Liu, 1993; Flamm, 2010; VerWey, 2019a,b). Goldberg et al. (2024) provides a brief history of industrial policy in this sector, noting that its effectiveness varies by region. Japan, South Korea, and Taiwan used infant industry promotion and extensive international technology transfer to successfully catch up to the technology frontier. In contrast, China's decades of industrial policies have yielded modest results, though recent initiatives show greater promise. Goldberg et al. (2024) suggests that access to technology transfer may explain these differences and points out that government intervention

[4]GlobalFoundries expanded its global manufacturing footprint through acquisitions and development, starting with its spin-off from AMD in 2009 and the inheritance of its German fabs. In the same year, it began constructing a new fab in Malta, followed by entering Singapore in 2010 by acquiring Chartered Semiconductor. The expansion continued in 2015 with the acquisition of IBM's semiconductor division, adding facilities in Vermont and New York to its portfolio.

[5]Morris Change, TSMC's founder, noted in a podcast with the Brooking Institution that the manufacturing cost in Oregon is about 50% higher than Taiwan cost.

is likely most effective in the early stages of industry development.

Resurgence The number of industrial policies targeting the semiconductor industry has increased significantly since 2020 (Goldberg et al., 2024). Notably, the U.S. passed the CHIPS Act in 2022, allocating $52 billion in manufacturing grants and research investments over five years, along with a 25% investment tax credit to incentivize domestic semiconductor production. Similarly, the European Chips Act, announced in the same year, aims to mobilize more than €43 billion in public and private investment by 2030. Several factors contribute to this surge, including supply chain disruptions during the COVID-19 pandemic and a heightened focus on resilience. Escalating geopolitical tensions have also prompted governments to seek greater independence in manufacturing to reduce reliance on foreign sources.

Policy Goals What are the main goals of these policies? Goldberg et al. (2024) offers a systematic analysis by manually labeling the stated goals of 58 collected policy measures. The top three identified goals are economic growth and development, enhancing international competitiveness, and improving resilience — aligned with the policy objectives of the U.S. CHIPS Act. According to a Congressional Research Service report (Sargent Jr et al., 2023), the U.S. CHIPS Act includes was motivated by several concerns: the decline in the U.S. position in semiconductor manufacturing and technology, the rise of China's industrial and technological competitiveness, inadequate domestic manufacturing capability to meet national security and economic needs, and reliance on global supply chains concentrated in East Asia. The primary goals are to achieve technological leadership, particularly in competition with China (Liu et al., 2024 discusses the rationale for sabotage policies), and to strengthen supply chain resilience to mitigate losses from disruptions.

Leveraging the quantitative framework developed in the following sections, this paper will evaluate ongoing policies, particularly the U.S. CHIPS Act, in terms of its goals of maintaining technological leadership in the race with China and securing domestic supply resilience. This framework can also be adapted for evaluating other policy measures.

3 Data and Industry Features

3.1 Data

This study utilizes two main datasets. The first is the Pure-Play Foundry Market Tracker complied by Omdia, providing firm-level data on capacity, wafer shipments, and revenue for nearly all pure-play foundries on a quarterly basis from the early 2000s to 2023 Q3[6], along with forecasts for future years. A key feature of this dataset is its breakdown by technology nodes. Given its long time horizon, I use this dataset to characterize long-run industry features, such as the evolution of the number of firms, innovation patterns, and price trends. Also, the node-specific price and quantity data over time allows me to infer demand. However, its coverage of Samsung's foundry operations is limited, despite Samsung's important role in the market.

To complement this, the second dataset is the 300mm Fab Outlook from SEMI, covering a 12-year span starting from 2015 Q1. This fab-level panel dataset includes all fabrication plants using 300mm wafers. It offers insights into the broader semiconductor industry, including developments by IDMs like Samsung and Intel. The dataset provides detailed information on each fab's location, capacity, technology node, and construction and equipment expenditures. This is valuable for estimating capacity investment costs and tracking the technology node of new production lines, which facilitates the estimation of parameters crucial for understanding technology upgrades.

In addition to the primary datasets, complementary data includes firm quarterly reports on revenue and R&D expenditure, industry analyses from sources such as IC Knowledge, and downstream shipment data from sources such as Gartner, IDC, and Canalys.

3.2 Industry Features

This section provides key industry features that reflect market structure, technological progress, and cost structure, guiding the model specification.

Fact 1 (Regional Demand-Supply Imbalance) *U.S. firms generate the largest share of revenue for semiconductor foundries, while East Asia dominates manufacturing capacity.*

Figure 2 shows that around 50% of semiconductor foundry revenue comes from U.S. firms such as Apple, Qualcomm, AMD, and NVIDIA, while the U.S. accounts for less than 10% of pure-play foundry capacity, with a declining trend since 2005. In contrast, most capacity is concentrated

[6]Quarterly data is available starting in 2003.

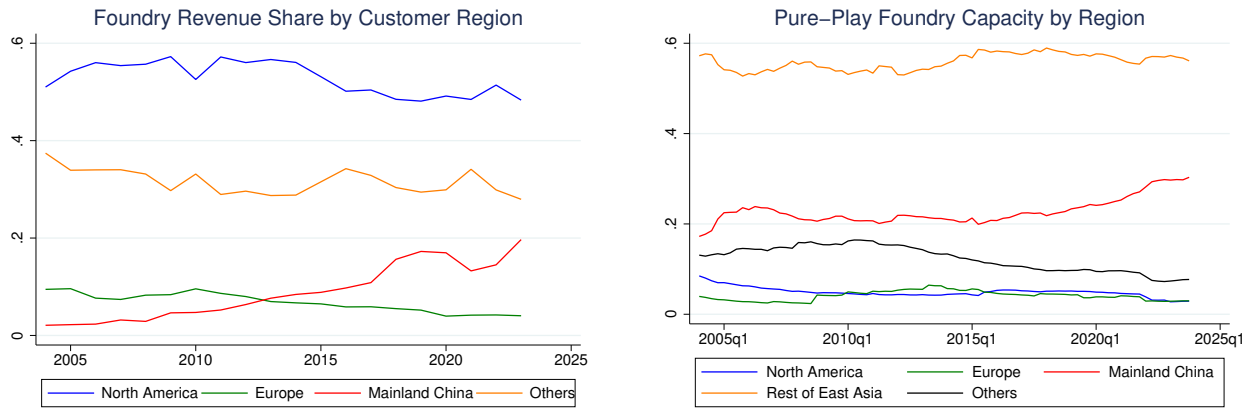


Figure 2: Foundry Demand and Supply Breakdown

Notes: Based on data from Omdia and the author's calculations. The revenue share plot includes revenue from both pure-play and IDM foundries (excluding internal manufacturing). The overall pattern remains similar when specifically focusing on pure-play foundries, which have contributed about 90% of total revenue since 2014. The capacity share plot features only pure-play foundries. Including IDM would raise North America's share, as the U.S. accounted for 12% of the global semiconductor manufacturing capacity in 2020, per Semiconductor Industry Association Factbook 2020.

in East Asia, with an upward trend since 2005, largely driven by China. As of 2023, East Asia's capacity share surpasses 80%, including all technology nodes. The geographical imbalance is even starker for leading-edge capacity. High demand and limited supply in the U.S. leave the country vulnerable to trade disruptions. Likewise, such disruptions could significantly harm the profits of East Asian producers if they lose access to the U.S. market.

Fact 2 (Decreasing Number of Firms in Frontier Technology) *Number of firms with leading-edge manufacturing capabilities decreases as technology advances.*

Figure 3 illustrates the evolution of firms with advanced manufacturing capabilities among pure-play foundries. The trend shows a decline over time in the number of firms capable of producing advanced chips under different definitions of advanced manufacturing capabilities. This pattern remains robust when IDMs are included (see Figure A.2). As of 2023, only four firms (including IDMs) - TSMC, Samsung, SMIC, and Intel - are actively pursuing leading-edge technology development. Given the high concentration of the industry, it is essential to consider the strategic interactions among firms.

Why are fewer firms involved in advanced manufacturing? The answer may lie in the nature of innovation and the interplay of price and investment costs.

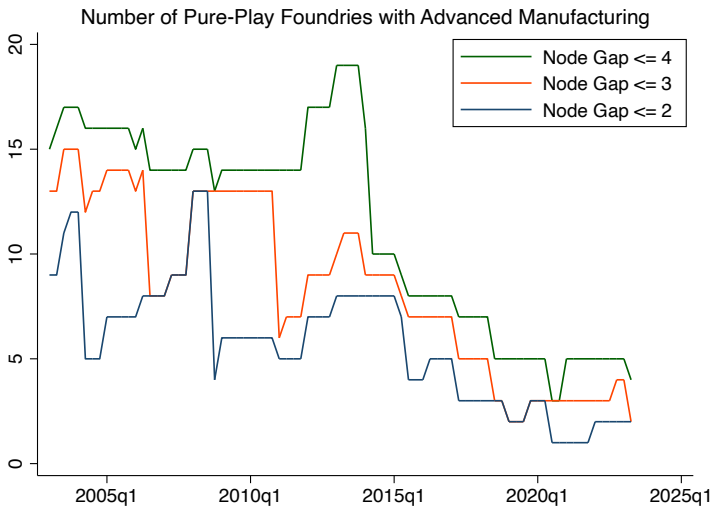


Figure 3: Pure-Play Foundry Industry Evolution

Notes: The plot is based on data from Omdia and the author's calculations. Advanced chips are defined as technology lagging no more than two, three, or four generations behind the frontier, as indicated in the plot.

Fact 3 (Incremental Technology Upgrading) *The technology upgrading at the firm level is predominantly incremental and leapfrogging is rare.*

Figure 4 shows the distribution of step sizes in node improvement across all pure-play foundries since 2003. More than 90% of the upgrades improve the technology node by one step, with very few making jumps of sizes 2 or 3. Among these rare instances of jumps larger than one step, most are upgrades from mature technologies.[7] For firms operating far from the technological frontier, leapfrogging remains an exception, suggesting that bridging the gap is a slow and difficult process—even with strong incentives, let alone in cases where such incentives are absent.

Interviews with industry insiders reveal that technology upgrading relies on two key factors. First, advancements in equipment are essential for achieving smaller technology nodes. Second, as semiconductor fabrication processes grow more complex, precise control and process optimization become increasingly critical. This precision, built through accumulated experience and advanced process control technologies, may explain why leapfrogging is rare.

[7]The only exceptions are UMC, which skipped the 20nm node and jumped directly to 14nm from 28nm, and SMIC, which skipped both 20nm and 10nm nodes and jumped directly from 28nm to 14nm and from 14nm to 7nm. However, the capacities of UMC and SMIC in these nodes are low, accounting for at most approximately 5% of the total capacity for the corresponding technology nodes during the periods sampled. The median capacity share is less than 3%. Also, interestingly, UMC shut down its 14nm production and transitioned to 22nm after several quarters.

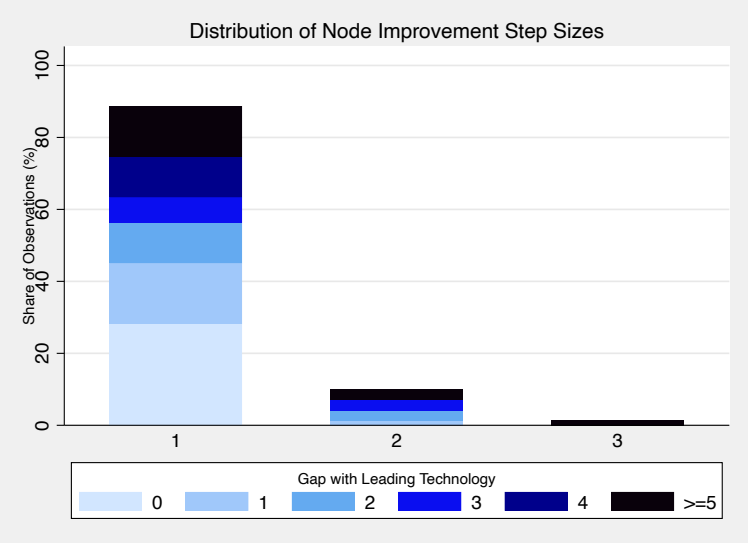


Figure 4: Iterative Innovation

Fact 4 (Declining Margins for Latecomer) *While prices for given nodes fall rapidly post-launch, investment costs stay steady, making high profits achievable only for early entrants.*

Figure 5 shows the price dynamics of different technology vintages. After their introduction, prices of new nodes drop rapidly before stabilizing. This initial decline can be attributed to reduced market power as more manufacturers start producing the new generation of products, expanded capacity among existing firms, and gradual cost reductions as yield rates improve with the maturation of the manufacturing process.

To investigate whether the investment cost of fixed nodes changes over time, I leverage variation in fab construction time to regress investment unit costs on a time trend, controlling for process node fixed effects. The results, shown in Table 1, reveal no significant time-related trend, with the time trend coefficient both small and insignificant. These findings remain robust when controlling for location and company fixed effects.[8]

Since investment costs remain relatively stable while prices decline over time, later entrants face considerably lower profits than technology leaders. Figure A.4 illustrates this, showing that only TSMC, the leading firm, consistently maintains a positive operating margin, while the other top foundries often see margins near zero or negative. As I will discuss in the model section, a possible motivation for firms to endure low profits is the expectation of future high profits if they succeed in

[8]Appendix A.4 provides additional details.

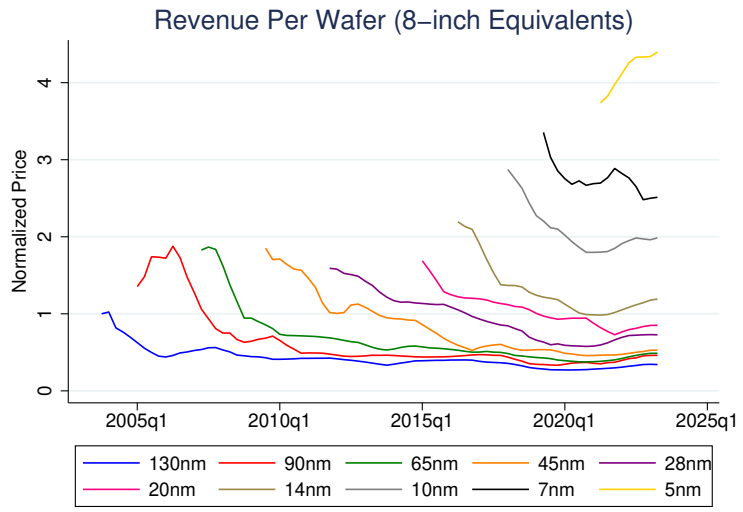


Figure 5: Median Normalized Price Per Wafer

Notes: Revenue per wafer is calculated by dividing the revenue of a company at each node by the estimated wafer shipments, converted to 8-inch equivalents. For each node and period, the median revenue per wafer is computed across firms, and then averaged over the current and preceding three quarters for smoothing. Finally, this price is normalized to 1 for the 130nm node in the initial sample period. This is a naive estimator to illustrate trends in wafer pricing.

Table 1: Fab Investment Cost of a Fixed Node

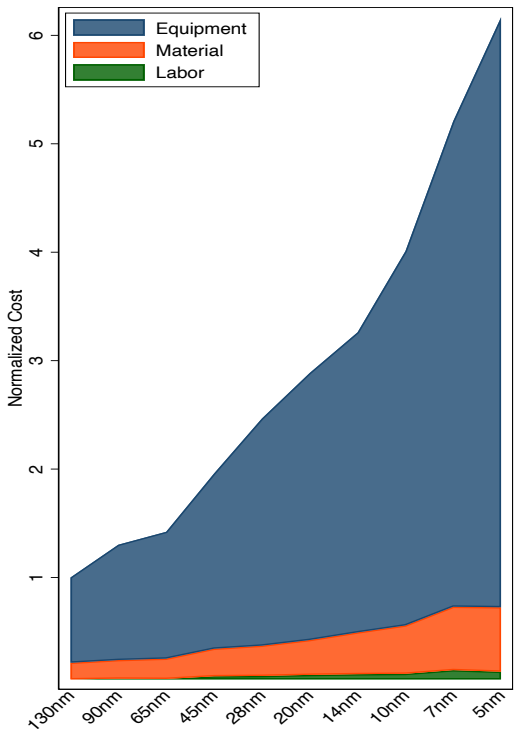
	(1)	(2)	(3)	(4)
Time Trend	0.0002	0.0006	0.0016	0.0004
	(0.0007)	(0.0006)	(0.0012)	(0.0010)
Log(Capacity)		-0.0253*	-0.0186	-0.0079
		(0.0132)	(0.0107)	(0.0128)
Constant	0.1779***	0.4347***	0.3509**	0.2554*
	(0.0131)	(0.1366)	(0.1202)	(0.1274)
Technology Node Fixed Effects	Yes	Yes	Yes	Yes
Firm Fixed Effects	No	No	Yes	No
Country Fixed Effects	No	No	No	Yes
Dependent Variable Mean	0.1808	0.1808	0.1848	0.1789
N	107	107	97	106
R^2	0.9038	0.9082	0.9233	0.9227

Notes: Standard errors are presented in parentheses and are clustered at the firm level. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

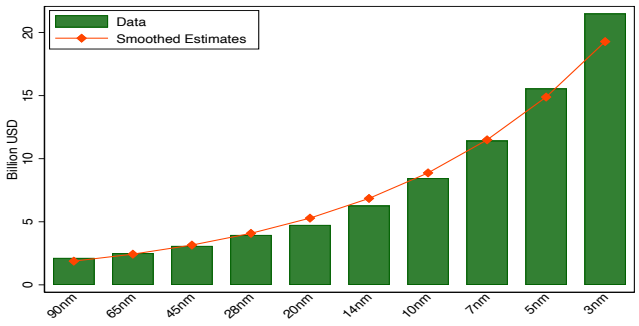
innovating to become industry leaders. However, firms far from the frontier, limited by step-by-step innovation, may endure prolonged low profits with slim chances of catching up, diminishing their

incentive to innovate. This sheds light on why firms exit advanced technology competition with little chance of re-entry, highlighting the importance of maintaining technological continuity and advancement in this industry.

Fact 5 (Cost Structure) *The foundry industry features high R&D intensity, with equipment costs dominating operating expenses. Both investment costs and R&D expenditures increase exponentially with advancing technology nodes.*



(a) Wafer Cost Breakout



(b) Fab Equipment Cost by Technology

(c) TSMC Quarterly R&D Cost

Figure 6: Operating Cost Breakdown and Cost Trajectories by Generation

Notes: Panel 6a shows the decomposition of lifetime operating costs, adapted from the 'Cost Drivers by Node' plot in the IC Knowledge report *Technology and Cost Trends at Advanced Nodes*. I digitized the plot using WebPlotDigitizer. Panel 6b uses data from SMIC's prospectus (page 1-1-142), originally sourced from International Business Strategies (IBS). It shows equipment costs by generation for a 50,000-wafer-per-month fab. Bars represent actual industry data, while the dotted line provides smoothed log regression estimates. Panel 6c uses data from TSMC's quarterly reports to display its quarterly R&D expenses, with the y-axis on a log scale.

The foundry industry has high R&D intensity, with about 10% of sales allocated to R&D, compared to 5% in the U.S. manufacturing sector. Once the technology is established, most operating costs

stem from equipment, determined during the capacity installation stage.[9] Figure 6a shows a wafer cost breakdown by technology node for TSMC, dividing lifetime operating costs into equipment, material, and labor. The plot shows that equipment costs, including depreciation, maintenance, and facilities, account for an average of 84% of total costs.[10]

Figure 6b shows that fab equipment costs by generation increase by approximately 25.9% per generation, with a log regression model achieving an R-squared above 99%. Figure 6c represents the quarterly R&D expenses for TSMC, following an exponential growth trend as technology advances. Given TSMC's longstanding leadership in the industry and regular node upgrades every 2-3 years, assuming exponential growth in R&D costs looks reasonable.[11]

Industry Features Summary Building on the key industry features outlined in this section, the model will incorporate oligopoly competition to reflect the highly concentrated market structure. Given the continuous innovation in technology nodes, the model will account for product innovation, assuming step-by-step advancements due to the rarity of leapfrogging. The industry's cost structure — characterized by high R&D intensity and equipment costs dominating operating expenses — emphasizes the need to focus on R&D and capacity investment decisions. These cost dynamics, as observed in the data, will be integrated into the model. Finally, the model will use capacity investment subsidies as the primary policy tool, reflecting both the importance of capacity costs and the alignment with current policy measures.

4 A Model of Innovation and Capacity Allocation under Trade Disruption Risks

This section introduces a quantitative framework that integrates trade disruption risks into a dynamic oligopoly model featuring innovation and industrial policies. The industry-equilibrium model

[9] The fabrication process relies on a photographic technique, with lithography being a crucial step. The facility must maintain extreme cleanliness to prevent random particles from damaging the integrated circuits (ICs) on wafers. A significant portion of a fab's cost is dedicated to equipment that ensures this cleanliness. This report from BCG provides a good description of the upfront cost and ongoing expenses to the semiconductor manufacturing industry.

[10] The relative importance of equipment costs has increased as technology advances, with equipment costs rising faster than material and labor costs. The equipment cost share increased from 78.6% at 130nm to 88.2% at 5nm, while the material cost share decreased from 15.4% to 9.6%, and labor costs fell from 6.0% to 2.1%.

[11] The log regression model for R&D expenses also demonstrates an excellent fit, with an R-squared near 98%.

builds on the key characteristics of the semiconductor foundry industry outlined earlier to capture firms' innovation and capacity installation decisions across locations. It characterizes firms' responses to trade shocks and location-based subsidies, and quantifies the impact of industrial policies on innovation dynamics and local supply resilience.

4.1 Setup

Given the continuous innovation, turnover, and demand shifts in advanced chips manufacturing, I model the industry's non-stationary evolution over time. Time is modeled as discrete with a finite horizon. After the terminal period T, firms make no further dynamic decisions but continue earning profits from existing technologies. T is set sufficiently large to ensure no incentive for further innovation or that the frontier technology likely reaches its physical limit by then.[12]

The model considers a finite number of firms, N, that differ in core productivity, technology, and home location. Given the oligopoly competition environment, the model needs to keep track of the technology status of all firms, causing the state space to grow rapidly and become intractable as N grows. To manage complexity, the focus is limited to firms engaged in advanced chip manufacturing, defined as those operating within two generations of the frontier technology. To alleviate computational challenges, the model specifically includes five firms: TSMC, Samsung Foundry, GlobalFoundries, UMC, and SMIC, which have collectively held nearly 100% of the market share in advanced chip manufacturing since 2010. Following strategies from Goettler and Gordon (2011) and Igami and Uetake (2020), the state space is bounded by tracking the frontier technology and each firm's gap from it. Specifically, in each period t, the state variables $S_t = \{\bar{n}_t, \mathbf{\Delta n_t}, q_t\}$ include: (1) $\bar{n}_t$, the current frontier technology node, constrained by the physical limit $\bar{N}$; (2) $\mathbf{\Delta n_t}$, the technology status of each firms i relative to the frontier - categorized as frontier ($\Delta n_{it} = 0$), one generation behind ($\Delta n_{it} = 1$), two generations behind ($\Delta n_{it} = 2$), exit ($\Delta n_{it} = 3$);[13] and (3) q_t, the node age of frontier products, reflecting the customer base, which grows as the technology matures, capped by $\bar{q}$.[14]

[12] As transistors become smaller, quantum tunneling effects become increasingly significant, causing unwanted current leakage in classical semiconductor chips (Veerasingam, 2023). Even without considering quantum effects, chip components cannot shrink beyond atomic size, as atoms themselves cannot be reduced further.

[13] The implicit assumption is that firms three generations behind the frontier technology no longer upgrade their technology. This pattern largely aligns with observed data, as firms significantly lagging behind typically exit advanced manufacturing. Section B.5 provides micro-foundations for this assumption through a simplified two-firm model.

[14] To save state space, it is assumed that the customer base for non-frontier technology has reached its

This paper focuses on the dynamic problem of innovation, abstracting from other potential dynamic considerations like capacity and demand (for dynamic demand, see Bertolotti et al., 2024). Innovation is emphasized as it directly relates to the policy goal of technological competition. Ideally, firm capacity in each generation would be treated as a state variable too, but tracking it across multiple generations for all firms is computationally infeasible. To maintain tractability, capacity is assumed to be a static choice with no adjustment cost or irreversibility. [15]

This paper emphasizes product innovation over process innovation.[16] Product innovation in this sector is mostly embedded in capital infrastructure. When firms upgrade their technology, they add the latest product to their portfolio while continuing to use older technologies to serve distinct market segments.

4.2 Timing

This model considers aggregate and transitory trade disruption shocks — such as global trade wars, pandemics, and logistics disruptions — and can be adapted to explore idiosyncratic trade shocks for friend-shoring (see Appendix B.9) and permanent shocks for decoupling risks (see Appendix B.8). This paper assumes fixed trade disruption risks in all future periods, known to all agents, though it can be easily adjusted to account for more flexible beliefs about future risks or unexpected shocks.

To avoid multiple equilibria in strategic industry dynamics, I adopt the stochastically alternating-move game from Igami and Uetake (2020). In this setup, at most one firm per period, randomly selected ex-ante, can make a dynamic R&D decision.

Here is the timeline for firm decisions:

1. At the beginning of each period, nature randomly chooses at most one firm with a probability

upper bound.

[15] Assuming static capacity overlooks an important dynamic: with persistent capacity, the first-mover advantage becomes stronger, as the leading firm has a greater incentive to build more capacity early on to deter entry by competitors. Additionally, more persistent capacity introduces two competing forces that influence the impact of temporary policies. On one hand, it may weaken the effects of such policies because current capacity affects future supply. On the other hand, firms anticipate that once they build more capacity, it will remain for a long time. If maintenance and other per-period costs are substantial, local subsidies aimed at attracting capacity may be less effective initially. However, if investment costs are primarily one-time expenses, the first force might dominate. That said, the impact on more enduring policies should be less significant.

[16] For example, learning-by-doing is a key channel of process innovation, where firms improve yield rates through increased production. However, firm-level yield rate data by technology node is highly confidential and difficult to obtain.

of $p_{pick} = \frac{1}{N}$.

2. The selected firm makes the R&D investment to innovate for the next generation of chips if the physical limit has not been reached. The probability of success depends on the innovation effort and is stochastic.

3. All incumbents compete in the generation-specific spot market, earning flow profits:

   3.1. Given the technological status of all firms, investment costs and market sizes across locations, and the trade disruption risk, firms determine their capacity allocation for each location.

   3.2. The trade disruption shock is realized. Without disruption, firms compete globally under total capacity constraints; with disruption, they compete locally within location-specific constraints. All capacity fully depreciates at the end of the period.

4. This process repeats until the final period T.

4.3 R&D Investment Decisions

Firms face a dynamic problem of innovation and a static problem of allocating capacity and determining shipments across locations each period. After nature selects the potential innovator, if the selected firm has not reached the technology limit, it decides on the amount of R&D effort d to maximize its expected net present value, while non-innovators form rational expectations about their expected values. The Bellman equation for the innovator i's dynamic optimization problem is given by

$$V_{it}(S_t) = \pi_{it}(S_t) + \max_{d \geq 0} \left\{ -c_{i,d}(S_t)d + \beta \left[\begin{array}{c} \rho(d)[(\Lambda_{it+1}(S_{t+1} \mid S_t, a_{it} = 1) \\ +\mathbb{1}(\bar{n}_{t+1} > \bar{n}_t)\pi_{it}^{n=2}(S_t)/(1-\beta)] \\ +[1-\rho(d)]\Lambda_{it+1}(S_{t+1} \mid S_t, a_{it} = 0) \end{array} \right] \right\}, \qquad (1)$$

where $\pi_{it}(S_t)$ represents the per-period profit from advanced technology, $c_{i,d}(S_t)$ is the unit cost of innovation effort, $\rho(d)$ is the probability of successful R&D, a_{it} indicates whether the R&D is successful ($a_{it} = 1$) or not ($a_{it} = 0$). To capture potential technology spillovers across firm, the R&D unit cost varies depending on whether the innovator is the leading firm or a follower. If the innovator is a follower, its R&D unit cost is reduced to a γ portion of the leading firm's unit cost at the same technology. The detailed specification is provided in Section 5. $\pi_{it}^{n=2}(S_t)$ is the per-period profit for a technology deemed advanced this period but not in the next, as $\pi_{it}(.)$ reflects only profits from advanced generations. As the frontier advances, firms with the vintage transitioning

from advanced to mature receive a one-time lump-sum profit proportional to their current earnings from that generation. This is equivalent to assuming that node-specific revenue stabilizes once the technology matures.[17] $\Lambda_{i,t+1}$ represents i's expected value at $t+1$ before nature picks the potential innovator at $t+1$. This is defined as

$$\Lambda_{it}(S_t) = p_{pick}\left[V_{it}(S_t) + \sum_{j\neq i} W_{it}^{j}(S_t)\right].$$

Here, W_{it}^{j} is the value for firm i at t if nature selects the other firm $j \neq i$ at t, formulated as

$$W_{it}^{j}(S_t) = \pi_{it}(S_t) + \beta \begin{bmatrix} \rho(d_j^*(S_t))[\Lambda_{it+1}(S_{t+1} \mid S_t, a_{jt}=1) \\ +\mathbb{1}(\bar{n}_{t+1} > \bar{n}_t)\pi_{it}^{n=2}(S_t)/(1-\beta)] \\ +\left[1-\rho(d_j^*(S_t))\right]\Lambda_{it+1}(S_{t+1} \mid S_t, a_{jt}=0) \end{bmatrix},$$

where $d_j^*(S_t)$ is optimal R&D level chosen by the other firm j when it is the innovator. If the optimal R&D is positive, the optimal R&D satisfies

$$c_{i,d}(S_t) = \beta\rho'(d_i^*)\begin{bmatrix} \Lambda_{it+1}(S_{t+1} \mid S_t, a_{it}=1) + \mathbb{1}(\bar{n}_{t+1} > \bar{n}_t)\pi_{it}^{n=2}(S_t)/(1-\beta) \\ -\Lambda_{it+1}(S_{t+1} \mid S_t, a_{it}=0) \end{bmatrix}. \tag{2}$$

The optimal R&D effort is determined by the unit cost of R&D and the expected value gain from technology upgrading.

States Transition Dynamics The transition dynamics after successful R&D depend on whether the potential innovator is the leading firm or not. If the innovator is the leading firm ($\Delta n_{it} = 0$), the frontier technology advances by one generation, and the customer base resets:

$$\bar{n}_{t+1} = \bar{n}_t + 1, q_{t+1} = 1, \Delta n_{it+1} = 0, \Delta n_{jt+1} = \Delta n_{jt} + 1 \quad \forall j \neq i.$$

If the innovator is not the leading firm ($\Delta n_{it} > 0$), its gap to the frontier narrows:

$$\bar{n}_{t+1} = \bar{n}_t, q_{t+1} = \min\{\bar{q}, q_t + 1\}, \Delta n_{it+1} = \Delta n_{it} - 1, \Delta n_{jt+1} = \Delta n_{jt} \quad \forall j \neq i.$$

If R&D is not successful, only the leading-edge node age evolves as $q_{t+1} = \min\{\bar{q}, q_t + 1\}$, with all other states remaining unchanged.

Two-Firm R&D Dynamics Appendix B.5 provides a simplified two-firm model to illustrate how firms make innovation decisions in the absence of trade disruption risks. The model shows that

[17] Appendix A.6 demonstrates the relative stability of node-specific revenue for mature technologies at TSMC.

in industries with incremental innovation and low profit margins for followers, the lagging firm lacks incentive to catch up if it falls too far behind the industry leader. If the goal is to boost the technology leadership of domestic firms, subsidies provided to them are effective only when they are relatively close to the industry frontier. When the technology gap is too wide, the impact of these subsidies is minimal.

4.4 Capacity Installation and Shipment Decisions

Within each period, following the realization of technology upgrades, firms make decisions in two stages. In the first stage, given all firms' technology status and investment costs across locations, firms make sunk and irreversible capacity investments while considering trade disruption risks. Trade disruptions are realized after capacity installation. If no trade disruption occurs, firms can serve the integrated global market from all fabs without trade costs. If trade disruption occurs, markets become autarkic, and firms can only serve individual markets from local fabs. In both scenarios, firms decide the optimal shipment level for each technology node and destination market, subject to capacity constraints. As noted in Kreps and Scheinkman (1983), when capacity is determined in the first stage, the second stage is equivalent whether firms compete in price or quantity. While I focus on quantity competition here, the results hold under price competition as well.

Chip Demand Demand is modeled in reduced form, with the aggregate demand for technology n at time t is given by

$$\log Q_{nt} = \alpha_0 + \alpha_p \log P_{nt} + D'_{nt}\alpha_D,$$

where Q_{nt} is the aggregate demand for node n in period t, P_{nt} denotes the market price, and D_{nt} is a vector of demand shifters, detailed in section 5.1.

Stage 2: Shipment Decisions For each technology node n and each destination market m, given all firms' capacity allocation, each firm i decides the shipment quantity q_{inm} to maximize profit, subject to the capacity constraint C_{inm}:

$$\begin{aligned} &\max_{q_{inm}} (P_{nm}(q_{inm}, q_{-inm}) - c^s_{inm}) q_{inm} \\ &\text{s.t. } q_{inm} \leq C_{inm}, \end{aligned}$$

where $P_{nm}(q_{inm}, q_{-inm})$ denotes the market level price given the shipment level of all firms, and c^s_{inm} is the unit cost of shipment. Firm i's first-order condition (FOC) is

$$P_{nm} + \frac{dP_{nm}}{dQ_{nm}} q^*_{inm} = P_{nm} + \frac{P_{nm}}{\alpha_P Q_{nm}} q^*_{inm} = c^s_{inm} + \lambda_{inm},$$

where λ_{inm} is the Lagrangian multiplier associated with the capacity constraint in technology n and market m for firm i. It is greater than zero if the capacity constraint binds and equals to zero if it does not. When trade disruption does not occur, m indicates the integrated global market, and C_{inm} is the total capacity of firm i's fabs across all locations. When trade disruption occurs, C_{inm} is the capacity of firm i's local fab in each local market m.

By assuming all costs are determined during the capacity installation stage, i.e., $c^s_{inm} = 0$, it can be easily shown that the firm's optimal shipment equals the installed capacity, $q^*_{inm} = C_{inm}$, when the price elasticity is greater than one, i.e., $\alpha_P < -1$.[18] The intuition is that as firms increase shipment by one unit, the price drops by less than one unit, so all firms aim to sell as much as possible when there are no additional costs. Thus, the expected profit in technology node n given capacity is:

$$\begin{aligned}
\pi_{in}(\mathbf{C}_{in}, \mathbf{C}_{-in}, \psi) &= (1-\psi) P^{int}_n(q^*_{in,tot}, q^*_{-in,tot}) \times q^*_{in,tot} + \psi \sum_m \left[P_{nm}(q^*_{inm}, q^*_{-inm}) \times q^*_{inm} \right] \\
&= (1-\psi) P^{int}_n(C_{in,tot}, C_{-in,tot}) \times C_{in,tot} + \psi \sum_m \left[P_{nm}(C_{inm}, C_{-inm}) \times C_{inm} \right],
\end{aligned}$$

where ψ is the probability of trade disruption, $C_{in,tot} = \sum_m C_{inm}$ is the firm's aggregate capability across all locations, and P^{int}_n is the price level if the global market is integrated.

Stage 1: Fab Capacity Installation The model considers three locations: the U.S., mainland China (CN), and the rest of the world (RoW).[19] Given the technology status of all firms, each firm decides the set of locations and the capacity in each location for each available technology node to maximize the expected profit:

$$\max_{C_{inm} \geq 0} \pi_{in}(\mathbf{C}_{in}, \mathbf{C}_{-in}, \psi) - \sum_m \kappa_{inm} C_{inm}, \tag{3}$$

[18] $\alpha_P < -1$ is a sufficient but not necessary condition. See appendix B.1 for the proof. The majority of costs are indeed determined during the capacity installation stage, making this assumption reasonable. Data shows that the average utilization rate for pure-play foundries has been approximately 86% from 2010 Q1 to 2023 Q4.

[19] The most relevant regions for RoW are Taiwan and South Korea. The model can be easily adapted to include more locations. It also has the potential to address policy-relevant questions like friend-shoring by allowing a more flexible definition of trade disruption, where some locations can still trade while others cannot during disruptions. The main constraint is finding ways to accurately determine the cost parameters for each location.

where κ_{inm} is the unit cost of capacity installation for firm i in technology n at location m. It is given by:

$$\kappa_{inm} = \frac{w_{nm}}{\nu_i}(1 - s_{nm})\delta_{im}, \tag{4}$$

where w_{nm} is the baseline unit cost of capacity in location m for technology n, ν_i is the firm-level productivity,[20] s_{nm} is the subsidy rate of location m, δ_{im} captures potential cost increases when firms build fabs in foreign locations. If the fab is built at home, $\delta_{im} = 1$; otherwise, $\delta_{im} = \delta \geq 1$, indicating the foreign cost shifter.

The optimal capacity allocation equates the marginal benefit of an additional unit of capacity with the unit capacity cost at every location with positive capacity, as characterized by the following FOC, obtained from the derivative of Equation 3:

$$(1-\psi)\left(\frac{P_n^{int}}{\alpha_P Q_n^{int}}\sum_m C_{inm}^* + P_n^{int}\right) + \psi\left(\frac{P_{nm}}{\alpha_P Q_{nm}}C_{inm}^* + P_{nm}\right) = \kappa_{inm} \quad \forall i, n, m. \tag{5}$$

With probability $1-\psi$, the marginal benefit of capacity is uniform across all locations, as free trade ensures equal market access. With probability ψ, trade disruption occurs, causing the marginal benefit to vary with local demand and existing capacity.

Case 1: Without Trade Disruption Risks In the absence of disruption risks, firms concentrate capacity in the most efficient location, typically their home base, as seen over the past two decades. In this case, the firms' problem has a closed-form solution. By combining the firms' FOCs with the demand equation, in the case where the dispersion of κ_{inm} is not too large among incumbents, we have

$$\begin{aligned} P_{nm}^{int} &= \frac{\alpha_P \sum_i \kappa_{inm}}{1 + \alpha_P N_n}, \\ Q_{nm}^{int} &= \exp(\alpha_P \log P_{nm} + \tilde{\alpha}_0 + \log \tilde{N}_{nm}) = \tilde{N}_{nm} \exp(\tilde{\alpha}_{0,n}) \left(\frac{\alpha_P \sum_i \kappa_{inm}}{1 + \alpha_P N_n}\right)^{\alpha_P}, \end{aligned}$$

where N_n is the number of incumbents capable of producing chips in node n, $\tilde{N}_{nm}$ is the market size of region m and node n, and $\tilde{\alpha}_{0,n}$ is the demand shifter related to product quality.[21] The implied

[20] For tractability, ν_i is taken exogenously to match observed firm market shares. While this paper does not address the sources of productivity differences, potential drivers include economies of scale from early market entry or historical government support.

[21] See Appendix B.2 for more details on the optimal capacity solution and Appendix B.3 for more details on the market size definition.

firm level output and profit are

$$q^*_{inm} = -\frac{\alpha_P Q^{int}_{nm}(P^{int}_{nm} - \kappa_{inm})}{P^{int}_{nm}}, \tag{6}$$

$$\pi^*_{im} = (P^{int}_{nm} - \kappa_{inm})q^*_{inm} = -\frac{\alpha_P Q^{int}_{nm}(P^{int}_{nm} - \kappa_{inm})^2}{P^{int}_{nm}}. \tag{7}$$

Case 2: With Trade Disruption Risks When $\psi > 0$, firms are incentivized to build capacities in all locations where their capacity investment costs are not excessively high relative to those of other firms, i.e., $\kappa_{inm} < P_{nm} = \frac{\alpha_P \sum_{j \in I_{nm}} \kappa_{jnm}}{1+\alpha_P N_{nm}}$.[22] Firms' capacity choices across locations are interdependent and must account for competitors' decisions. With non-linear demand, deriving optimal capacity analytically when $\psi > 0$ is challenging. Instead, I employ an iterative algorithm to numerically solve firms' optimal capacity decisions based on their FOCs at the capacity installation stage (Equation 5). Details are provided in Appendix B.4.

4.5 Equilibrium and Solution

Given all the fundamentals - capacity cost by generation and location, firm productivities, R&D baseline costs for each generation, and trade disruption risk ψ - along with the trajectories of demand shifters and the policy path, a sequential industry equilibrium is a set of functions $V_{it}(S_t)$, $W_{it}(S_t)$, $\Lambda_{it}(S_t)$, $d^*_{it}(S_t)$, $C^*_{inmt}(S_t)$, $q^*_{inmt}(\mathbf{C}_{nmt}, \mathbf{1}(\Psi_t < \psi))$, $P_{nmt}(\mathbf{q}^*_{nmt})$ such that for each firm i, technology node n, market m, and period $t \leq T$, the following conditions hold:

1. $V_{it}(S_t), W_{it}(S_t), \Lambda_{it}(S_t)$ satisfy the firms' Bellman equations;
2. $d^*_i(S_t)$ solves the innovator's optimization problem;
3. $C^*_{inmt}(S_t)$ maximizes the firm's expected profit ex-ante;
4. $q^*_{inmt}(\mathbf{C}_{nmt}, \mathbf{1}(\Psi_t < \psi))$ maximizes the firm's flow profit, given the trade disruption realization and capacity constraints $\mathbf{C}_{nmt}$ across all firms and locations;
5. $P_{nmt}(\mathbf{q}^*_{nmt})$ ensures product market clearing.

[22] I_{nm} denotes the set of incumbents that build capacity in technology n in market m. The location-specific investment cost at the firm level depends on the firm's core productivity, foreign cost shifters, and policy interventions. If the dispersion in core productivity among firms is not too large and investment subsidies are location-based rather than ownership-based, all firms will be incentivized to build capabilities in all locations, provided that the foreign cost shifter δ is not too large.

Firms solve two interconnected problems: a dynamic innovation problem and a static capacity allocation and shipment problem across locations in each period. The potential innovator chooses its optimal R&D effort, while non-innovators form rational expectations about their expected values. The static problem involves allocating capacity based on technology status, location costs, and trade risks, followed by optimizing shipments given capacity constraints and trade shocks. I first solve the static problem across all states, feeding the payoffs into the dynamic problem, which is then solved using backward induction.

4.6 Policy Intervention

From a global social planner's perspective, market power distortions cause firms to underinvest in capacity, and technology spillovers across firms lead to underinvestment in R&D, motivating policy intervention. However, with trade, the benefits of expanded capacity and accelerated innovation leak abroad, while the subsidy costs fall entirely on local governments, resulting in a low or even zero optimal subsidy rate. This exemplifies the free-rider problem.

As geopolitical tensions escalate and trade disruptions become more likely, governments have greater incentives to intervene for resilience. Although firms diversify to manage trade shocks without government support, their profit-maximizing behavior may not align with local consumer welfare. When firm profits are irrelevant—such as when only foreign firms are involved— the optimal subsidy balances policy costs with the welfare gains from increased local capacity and innovation. With higher trade risks, the marginal benefit of subsidies rises, since local capacity is more likely to serve domestic consumers and domestic consumers are more reliant on it.

Motivated by the U.S. CHIPS Act, the counterfactual analysis explores two policies: location-specific capacity subsidies and investment restrictions for certain locations. Location-specific capacity investment subsidies influence firms' capacity allocation decisions by affecting the unit cost of capacity installation (see Equation 4). These subsidies also impact the firm's innovation process by altering the expected profit from innovation, as the government covers part of the investment cost for each technology. Investment restrictions, limit where affected firms can install capacity, aiming to prevent technology spillovers to targeted locations. The model reflects this by assuming weaker spillovers to firms in restricted locations.

5 Estimation and Results

This section outlines the calibration and estimation strategy of the model and presents the results. In the first step, I estimate the global demand curve using instrumental variable regressions, leveraging price and shipment data by technology over time. I assume identical price elasticities and demand shifter coefficients across locations, with differences in local demand driven by varying downstream market demand, which are considered in the counterfactual analysis. For the supply side estimation, I assume that firms perceived no trade disruption risks during the sample period, given that the majority of advanced chip manufacturing was domestically concentrated.[23] The supply-side estimation proceeds in two steps. First, I estimate static parameters, including location-specific investment costs and firm-level productivity, using external industry reports or inferred from firm-level capacity and sales shares. Next, conditional on the previously estimated static parameters, the dynamic R&D parameters are estimated using maximum likelihood estimation (MLE) based on firm-level technology upgrade history data.

5.1 Demand Estimation

The demand curve for technology node n is specified as log-linear:

$$\log Q_{nt} = \alpha_0 + \alpha_p \log P_{nt} + D'_{nt}\alpha_D + \epsilon_{nt}, \tag{8}$$

where Q_{nt} represents the total global wafer shipments (measured in 8-inch equivalents) of technology node n, P_{nt} denotes the average wafer price (thousand USD per 8-inch equivalent wafer) among all pure-play foundries, and D_{nt} is a vector of demand shifters. These shifters include the logarithms of worldwide PC shipments and worldwide smartphone and tablet shipments (in million units), technology node n as a proxy for quality (logarithm of transistor density), dummy variables indicating whether n is the frontier technology or one generation behind the leading-edge, and the number of quarters since since the introduction of node n.

Prices are instrumented using the average prices of legacy chips, specifically the prices for 500 nm and 350 nm wafers. The identification assumption is that the price of legacy chips is correlated with the price of advanced chips but uncorrelated with the unobserved demand shock ϵ_{nt}. Legacy

[23] In my sample period, most firms produced advanced chips domestically, except Samsung, which had a fab in Austin for Apple orders. However, when adopting new technology, Samsung built new lines only in South Korea. It is only very recently, due to rising geopolitical tensions, that firms have started establishing a global footprint.

and advanced chips use similar materials, such as raw wafers, chemicals, photoresist, and gases.[24] However, these chips serve different downstream markets. Legacy chips are primarily used in consumer electronics, automotive electronics, industrial electronics, and the Internet of Things, while advanced chips are used mainly in high-performance computing, such as CPUs and GPUs. The identification assumption is valid if the demand shifters for different downstream markets are uncorrelated.

The demand estimation focuses on advanced chip manufacturing, defined as technology no more than two generations behind the leading-edge. The 20 nm and 10 nm nodes are excluded from the estimation as these generations are short-lived. To strengthen the IV assumption by minimizing correlations between downstream markets, the analysis focuses on technology nodes of 90 nm and below. Since the semiconductor foundry market was globally integrated during the sample period, the demand curves are estimated using time-series variation.

Table 2 reports the estimation results. Columns 1 and 2 present OLS estimates with different sets of demand shifters. The adjusted R-squared increases substantially after considering the node age and whether the node is at the frontier or next-to-frontier. Columns 3 and 4 show IV estimates with different instruments, yielding similar results. The estimated price elasticity is approximately -1.5. Demand rises with technological advancement, and, all else equal, it is higher for frontier and next-to-frontier nodes compared to those two generations behind. Figure 7 illustrates the demand shift over node age, showing a strong upward trend during the first two years after introduction, followed by stabilization. This trend may reflect both the accumulation of customers and improvements in node quality over time. Column 3 is used as the preferred specification.

Several caveats should be acknowledged in the demand estimation. First, substitution across different technologies is not captured. Adjacent nodes are likely partial substitutes, which significantly complicates computations and makes it challenging to determine a clear cutoff. In this analysis, I assume that different sectors adopt different technology vintages, with some always adopting the most advanced, others the second most advanced, and so on.[25] Second, the estimation draws on data from pure-play foundries, excluding IDM foundries. Specifically, Samsung has been a sig-

[24]The material quality requirements may vary between legacy chips and advanced chips; however, the identification assumption holds as long as the material prices are correlated. Although the prices for all materials are not available, we can infer the material price correlation from raw wafer prices. The raw wafer prices for 8-inch wafers, used for 350mm and 500mm chips, and 12-inch wafers, used for advanced chips since around 2001, show a correlation as high as 80% based on customer import data in Taiwan.

[25]See Appendix B.6 for the micro-foundation of the demand curve.

Table 2: Demand Estimates

	(1) OLS	(2) OLS	(3) IV	(4) IV
Log price	-1.308***	-0.784***	-1.488***	-1.686***
	(0.394)	(0.186)	(0.413)	(0.457)
Node	0.065	0.120**	0.281***	0.325***
	(0.119)	(0.057)	(0.096)	(0.106)
Log PC shipment	0.858*	0.365	0.538*	0.586*
	(0.442)	(0.250)	(0.292)	(0.317)
Log mobile shipment	0.648***	0.353***	0.223***	0.186**
	(0.155)	(0.071)	(0.081)	(0.085)
Frontier node		0.344**	0.666***	0.757***
		(0.173)	(0.249)	(0.264)
Next-to-frontier node		0.176	0.405**	0.469**
		(0.143)	(0.185)	(0.196)
Node Age Fixed Effects	No	Yes	Yes	Yes
N	163	163	163	163
Adjusted R^2	0.518	0.836	0.819	0.808
First stage results				
Log price for 500nm wafer			1.024***	
			(0.165)	
Log price for 350nm wafer				0.899***
				(0.222)
F-value			166.24	159.67

Notes: Dependent variable is log worldwide wafer shipment. Robust standard errors are in parentheses. Significance levels: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

nificant player in the foundry market since 2010. However, revenue data by technology node for Samsung is unavailable, and imputing it from capacity data is arbitrary, as it requires guessing both utilization and the time lag between manufacturing and shipment.

5.2 Static Parameters

This section calibrates and estimates the parameters related to firm capacity installation cost and the market size for each location.

Capacity Cost The model assumes location specific factor of the capacity investment cost to increase exponentially with the technology node: $w_{ln} = w_{l0}(1 + g^w)^n$, where w_{l0} is the base investment cost in location l, and g^w captures the increasing cost as technology advances. The base cost

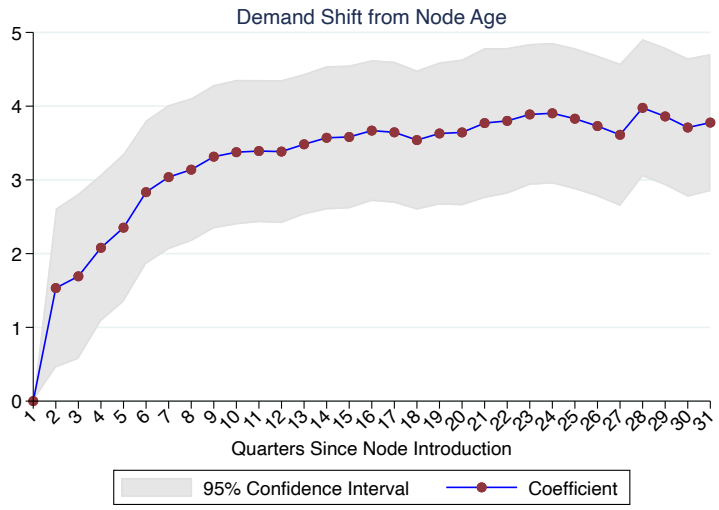


Figure 7: Demand Shift from Node Age

in the U.S. is normalized to $w_{US,0} = 1$. The increasing trend g^w is set to 25.8% to align with the rising equipment costs as technology advances.[26]

Location-specific costs are calibrated from Varas et al. (2020), a report by Boston Consulting Group (BCG) and the Semiconductor Industry Association (SIA). According to the report, building an advanced logic fab costs 78% of the U.S. cost in South Korea and Taiwan, and 72% in mainland China. While the U.S. faces structural disadvantages in some factor costs, particularly labor and construction costs, [27] these constitute a relatively small share of fab operation costs, with the majority coming from equipment costs, which are similar across regions. Cost differences across countries mainly arise from varying levels of government incentives. The report indicates that around 70% of the cost difference between the U.S. and Asian countries is due to government incentives. Without considering these incentives, the cost gap between the U.S. and Asian countries is approximately 7.5% ($= (1 - 75\%) \times (1 - 70\%)$), which aligns with estimates from TechInsights, a leading firm in semiconductor cost and price modeling.[28] Thus, $w_{CN,0} = w_{RoW,0} = 0.925$.

Firm-specific productivity ν_i is estimated using first-order conditions from the model and firm sales

[26] Figure 6b shows the equipment costs by technology, along with the smoothed estimates using a growth rate of 25.8%.

[27] Wendell Huang, the VP and CFO of TSMC, stated during the company's Q1 2023 earnings call that the construction costs in Arizona could be up to five times higher than those in Taiwan.

[28] According to Scotten Jones, president of TechInsights Semiconductor Manufacturing Economics, the cost for TSMC to produce wafers in the U.S. would be 7% higher than in Taiwan if they built a fab of the same size in the U.S. The full article is available at: TSMC Arizona Fab Cost Revisited.

shares. According to equation 6, the model-implied market share depends on a firm's technology and investment costs, which are determined by firm-specific productivity and location-specific costs and subsidies. I calibrate firm-specific productivity to match the average within-technology sales share data when all firms are incumbents. Since firms primarily built within their home locations during the sample period, we are not able to distinguishing higher market shares driven by productivity versus subsidies. I attribute all effects to firm productivity. Alternatively, this could be interpreted as the persistence of historical subsidies at the firm level. Details of this estimation process are available in Appendix C.1.

I consider two cases: (1) each firm has a unique firm-specific productivity, and (2) TSMC is treated as the dominant firm with a distinct productivity level, while all other firms share the same relative unit cost, leading to a higher market share for TSMC.[29] The estimation results are similar across both scenarios. For the baseline specification, I use the TSMC-dominant scenario, where all other firms have a relative unit cost of 1.57 compared to TSMC, resulting in a market share of around 60% for TSMC when all firms are incumbents.

Given that most firms only build fabs for advanced chips within their home location during the sample period, identification of the foreign cost shifter δ is not available. This parameter becomes important only in policy evaluation when trade disruption risk is considered. Therefore, I will test different values for δ in the counterfactual analysis.

Local Market Size I consider two approaches to constructing the market size for each location. The first approach infers market size from the downstream demand for smartphones, tablets, and PCs using the estimated demand curve. Details are provided in Appendix B.3. Based on PC and mobile shipments in 2023, the corresponding market share is approximately 25% for the U.S., around 22% for mainland China, and the remaining 53% for the rest of the world. These results are robust when using data from 2024 Q1.

The second approach utilizes foundry revenue share data by customer region. As shown in Figure 2, North America consistently accounts for about 50% of the revenue share, while mainland China's share has been increasing over the past two decades, reaching around 20% in 2023. However, this measure has two key drawbacks: (1) it captures upstream demand from semiconductor design firms

[29] TSMC's average market share for technologies where all firms are incumbents is approximately 60%. The overall market share of TSMC is higher due to its more advanced technology.

rather than actual consumer demand from downstream products like smartphones and PCs; and (2) it combines demand for both advanced and mature chips without distinction, with the mature chip market involving a larger number of manufacturers that are not accounted for in the model. Given these limitations, I will use the first approach as my baseline measure.

5.3 Dynamic Parameters

This section calibrates and estimates the parameters related to firm dynamic choice.

Assigned Parameters The annual discount factor is calibrated to 0.9, resulting in a quarterly discount factor of $\beta = 0.9^{\frac{1}{4}}$. The terminal technology $\bar{T}$ is set to $\bar{N} = 16$ (sub-0.2nm).[30] It is important to note that what truly matters here are the firms' expectations about the future technology limit. Even if new technologies—such as the invention of new materials that surpass current physical limits—are developed, firms' behavior will only change when they believe these new technologies can be successfully implemented. Given the technology limit, the final period is chosen far enough in the future to ensure that the physical limit is reached or all firms have no further incentive to innovate. The final period is set to 2059 Q4 when $\bar{N} = 16$. I tested different ending periods with a few years of variation, and the results showed minimal differences.

R&D Cost The probability of successful R&D, ρ, is defined as $\rho(d) = \frac{d}{1+d}$ (Pakes and McGuire, 1994). The cost associated with the innovation effort depends on the technology node the firm aims to develop and whether the new node will be the industry frontier. Specifically,

$$c_{i,d}(\Delta n_{it}, \bar{n}_t, q_t) = c_0(1+g^c)^{\bar{n}_t - \Delta n_{it}+1}\gamma^{1(\Delta n_{it}>0)},$$

where c_0 is the base cost, g^c is the growth rate of the per unit R&D cost as technology advances, and $\gamma \leq 1$ captures the lower cost for the followers, reflecting technology spillover effects across firms. I calibrate g^c based on TSMC's R&D growth trend, as it is a longstanding technology leader, consistently upgrading its technology approximately every two years. TSMC's average quarterly R&D growth rate is 3.6%, so I set $g^c = 8 \times 3.6\% = 28.8\%$.

[30] I define the 40nm technology as node 1 in my model. Sub-A2 is the most advanced generation I could find across all roadmaps provided by official sources. This information is from ASML's logic device roadmap released in 2024, according to Tom's Hardware. The most recent IRDS device roadmap ends with 0.5nm(IRDS, 2022). For the next step, I will also compare a pessimistic scenario where the physical limit is near.

I use maximum likelihood estimation (MLE) to estimate the dynamic parameters from the data on firm's technology upgrading history from 2010Q1 to 2023Q4. Denote $a_{it} = 1$ if firm i upgrades its technology in period t. The likelihood function in each period is

$$l_t(a_t \mid S_t; \theta) = \sum_i \hat{p}_{i,pick} \left[\rho\left(d_i^*(S_t)\right)\right]^{a_{it}} \left[1 - \rho\left(d_i^*(S_t)\right)\right]^{1-a_{it}}, \tag{9}$$

where $\hat{p}_{pick}$ is the conditional probability that i is the potential innovator in period t given the observed a_{it}.[31] Following Igami and Uetake (2020), if $a_{it} = 1$ for any i at t, I set $\hat{p}_{i,pick} = \frac{1}{\sum_j a_{jt}}$ for all i where $a_{it} = 1$, and $\hat{p}_{i,pick} = 0$ for all i where $a_{it} = 0$.[32] If $a_{it} = 0$ for all i at t, then $\hat{p}_{i,pick} = p_{pick}$ for all i. The log-likelihood function is

$$\mathcal{L}(\theta) = \sum_{t=1}^{\hat{T}} \ln\left[l_t(a_t \mid S_t; \theta)\right],$$

where $\hat{T}$ is the number of sample periods.

The identification of baseline R&D costs is derived from the frequency of technology upgrades—lower R&D costs lead to stronger R&D efforts and more frequent successful upgrades. The magnitude of technology spillover is inferred from how often lagging firms catch up; stronger spillovers increase the likelihood of these firms upgrading their technology.

Table 3: MLE Results

	Case1 Firm Specific	Case2 TSMC Dominant
$C_0/10^6$	3.0	3.2
	[1.5, 3.5]	[1.7, 3.6]
γ	0.20	0.22
	[0.05, 1.00]	[0.07, 0.92]
log likelihood	-37.23	-35.63

Notes: The 95% confidence intervals are constructed using likelihood-ratio tests. Additional details are provided in Appendix C.2.

Table 3 shows the estimation results for R&D related parameters. The results indicate substantial technology spillover across firms; lagging firms only need to pay about 20% of the unit R&D cost

[31] The data includes a few observations where firms skip a generation: UMC skips 20nm, and SMIC skips 20nm and 10nm. In the model estimation, I assume an unexpected shock that increases their technology by one step before they upgrade.

[32] During the sample period, there were 4 out of 56 quarters with 2 firms upgrading their technology. As a next step, I could use a monthly frequency in the model to avoid this issue. However, this would increase computational costs and require assumptions to convert the quarterly data to a monthly format.

compared to the leading firm. Considering that the leading firm, TSMC, often incurs R&D costs that are 5 to 10 times higher than those of lagging firms like UMC and SMIC, this estimate seems plausible.[33]

5.4 Model Fit

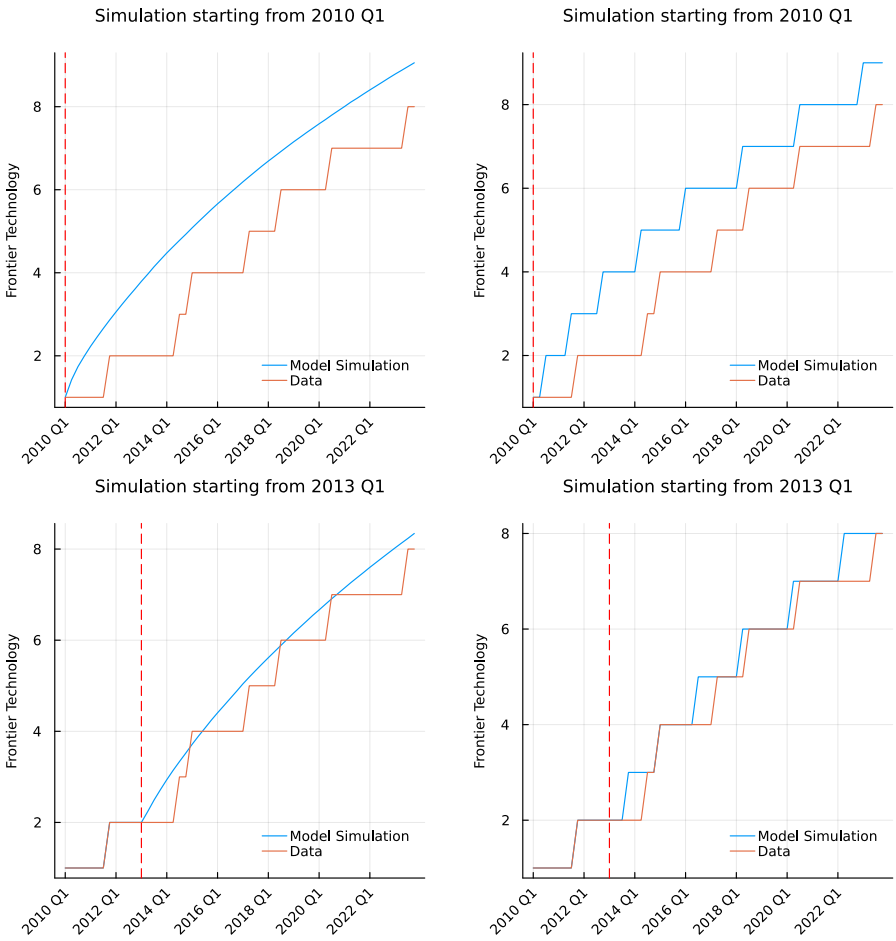


Figure 8: Frontier Technology Trajectories: Model VS Data

Notes: I simulated the model 10,000 times to compare the average frontier technology trajectory with actual data. The left-side plots display this comparison directly, while the right-side plots show a discretized version of the frontier technology in the simulation, based on the average time the frontier spends at each node. The upper two plots present results starting from 2010 Q1, and the lower two plots show results starting from 2013 Q1. See Appendix C.3 for results with different starting periods.

To verify the model fit, Figure 8 compares the simulated frontier technology to the actual observed technology trajectory. The model closely replicates the observed technological advancements, in-

[33]See Appendix A.7 for R&D cost comparison of the top 4 pure-play foundries.

dicating a good fit between the model's predictions and the real-world data. While the model initially predicts quicker technological advancement than what is observed, this discrepancy could be attributed to the fact that the data represents just one realization of the model. To reduce the impact of randomness, I simulated the model with different starting periods and compared the results to the actual data. Appendix C.3 shows the comparison of the simulated frontier technology with observed data when simulating from different initial periods, demonstrating that the model and data align well in most cases.

Another measure of validation is R&D intensity as a share of profit, which is an untargeted data moment. I calculate the total R&D expenditure divided by total profit in the model and compare it to the R&D expenditure-to-profit ratio in the data, where profit is defined as revenue minus the cost of goods sold. From the data, the aggregate R&D to profit share is 18.5%, rising to 27.83% when including other operating expenses.[34] In the model, the share is 23.5% when assuming TSMC has higher firm productivity while other firms have the same productivity, and 22.9% when all firms have specific productivity levels.

6 Policy Evaluation

This section uses the estimated quantitative model to simulate policies under different scenarios to evaluate their effectiveness. Specifically, I examine the implications for consumer welfare and technological progress of unilateral capacity investment subsidies and compute the optimal policy rate under different beliefs about trade disruption risks. Section 6.1 compares static consumer welfare gains and policy costs across different investment subsidy rates and trade disruption risks, taking the technology status as given. Section 6.2 incorporates the endogenous technology upgrading process under different policies and evaluates the optimal investment subsidy rates under varying trade disruption risks. Section 6.3 explores the effects of investment and technology clawbacks on mainland China, comparing the model-predicted technology trajectories of firms with and without these guardrail restrictions. For the policy evaluation, I exclude GlobalFoundries and UMC from the model as they are no longer active players in leading-edge chip manufacturing.[35]

[34] Other operating expenses include Marketing and Sales (M&S) and General and Administrative (G&A) expenses.

[35] GlobalFoundries announced in August 2018 that it would stop all 7nm development to focus on specialized processes (source: GlobalFoundries Press Release. In the same year, UMC announced that it would not rejoin the race to develop 7nm technology (source: Taipei Times).

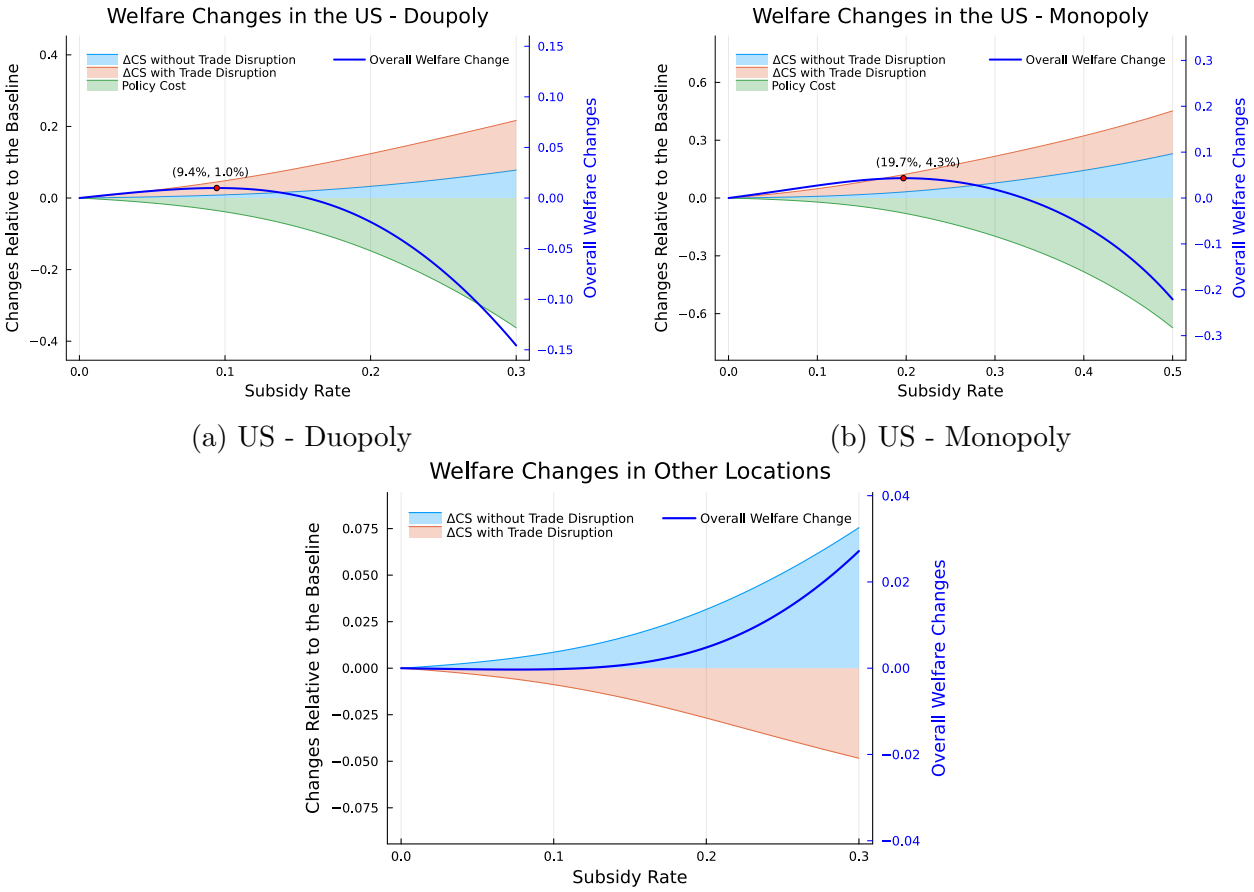
6.1 Static Resilience Gains

This section explores the welfare implications of unilateral capacity investment subsidies under trade disruption risks, taking the technology status of firms as given. I consider the case where Intel has not yet established itself as a significant player in the foundry service market, and there is no investment restrictions to mainland China. Consequently, SMIC has limited motivation to innovate, akin to GlobalFoundries and UMC (see Figure D.18a). Thus, TSMC and Samsung remain the sole competitors in the advanced foundry sector. Since both firms are non-U.S. firms, and this analysis considers a unilateral policy in the U.S., consumer welfare is defined as the consumer surplus minus the policy cost, excluding firm profits. All outcomes discussed below represent percentage changes relative to the no-subsidy baseline welfare, and the results are independent of the product generation.

The welfare change and the optimal subsidy rate in the U.S. depend on the number of incumbents, trade disruption risk (ψ), foreign cost shifter (δ), and subsidy rates of other locations. Establishing the probability of trade disruption is challenging; to set the baseline for policy analysis, I draw from the disasters literature (Barro and Ursúa, 2012; Nakamura et al., 2013) and use the geopolitical risk (GPR) data constructed in Caldara and Iacoviello (2022), along with their estimates of how GPR affects the probability of disaster episodes, selecting 20% as the baseline scenario.[36] For the baseline scenario, the foreign cost shifter is set to $\delta = 1$, and other regions do not offer any location-based subsidies.

At the expense of policy costs, capacity investment subsidies reduce local investment costs, attracting more capacity domestically while decreasing capacity in other regions. As overall costs decline, total global capacity increases. If trade disruptions do not occur, global prices drop, leading to a rise in consumer surplus across all locations. Conversely, if trade disruptions do occur, domestic prices decrease while prices in other regions increase, resulting in higher domestic consumer surplus but lower consumer surplus elsewhere. The ex-ante net welfare change depends on the trade-off between the expected welfare change before the trade shock realization and the policy costs. While market power distortion justifies policy intervention, without trade disruption risks, increased capacity benefits all countries. This reduces the incentive for any single government to intervene, as they can free-ride on the benefits without bearing the costs.

[36] More details are available in Appendix D.1.



(a) US - Duopoly

(b) US - Monopoly

(c) Other Locations

Figure 9: Baseline Static Welfare Implications

Notes: These plots show the baseline case where trade disruption risk $\psi = 20\%$ and foreign cost shifter $\delta = 1$. The decomposition of changes in consumer surplus with and without trade disruption in the plot represents the actual changes in consumer surplus, calculated as the consumer surplus with subsidies minus the consumer surplus without subsidies, divided by the consumer surplus without subsidies, and then multiplied by the probability of trade disruption or no trade disruption. Consumer surplus is derived from the demand curve, which depends on the price level. Details on the consumer surplus calculation are provided in Appendix B.7. The policy cost in the plot is also expressed relative to the consumer surplus without subsidies. The calculation of consumer surplus does not include firm profits.

Figure 9 illustrates the changes in welfare compared to the no-subsidy baseline scenario. Panel 9a and 9b show the welfare changes in the U.S. under either duopoly or monopoly: as the subsidy rate increases, both consumer surplus and policy cost rise. The optimal rate depends on the relative increase of these two factors. The plots also break down the consumer welfare changes based on whether trade disruption occurs. In both scenarios, U.S. consumer prices decrease, but the price drop is more pronounced when a trade disruption happens. Without disruption, the increased

capacity in the U.S. is distributed globally, resulting in less significant price changes. In this baseline scenario, the optimal investment subsidy rate is 9.4%, which improves consumer surplus by 1.0% in the duopoly case and by 4.3% in the monopoly case with a 19.7% subsidy rate.

Panel 9c depicts the changes in consumer surplus in other regions.[37] Local subsidies do not necessarily act as a beggar-thy-neighbor policy due to trade. The increase in consumer surplus when trade disruptions do not occur outweighs the decline in consumer surplus during disruptions, as long as pre-policy investment costs in other locations are not significantly higher than those in the U.S. (i.e., the efficiency loss from firms relocating to foreign locations, δ, is low , and there are no substantial subsidies in other regions). The intuition is that if the initial cost differences across locations are not too large, investment subsidies in the U.S. can significantly lower the overall global investment cost, leading to a sufficient drop in world prices when trade disruptions do not occur. This price reduction can offset the expected losses from reduced domestic capacity during trade disruptions. Conversely, if the initial cost differences are substantial, even with subsidies, subsidies in the U.S. may not sufficiently lower the global price. Instead, the effect is more about shifting capacity from other locations to the U.S., potentially resulting in a negative overall impact. Appendix D.2 further discusses the role of the foreign cost shifter and subsidies in other regions in the static welfare analysis.

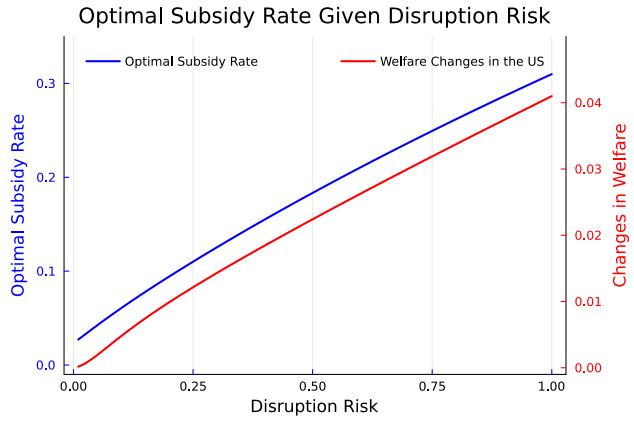


Figure 10: Static Setting: Optimal Subsidy Rate Given Trade Disruption Risks

Notes: The optimal rate, which maximizes U.S. consumer welfare defined as the net change in consumer surplus minus policy cost, is calculated in the baseline case with two incumbents and the foreign cost shifter set to $\delta = 1$.

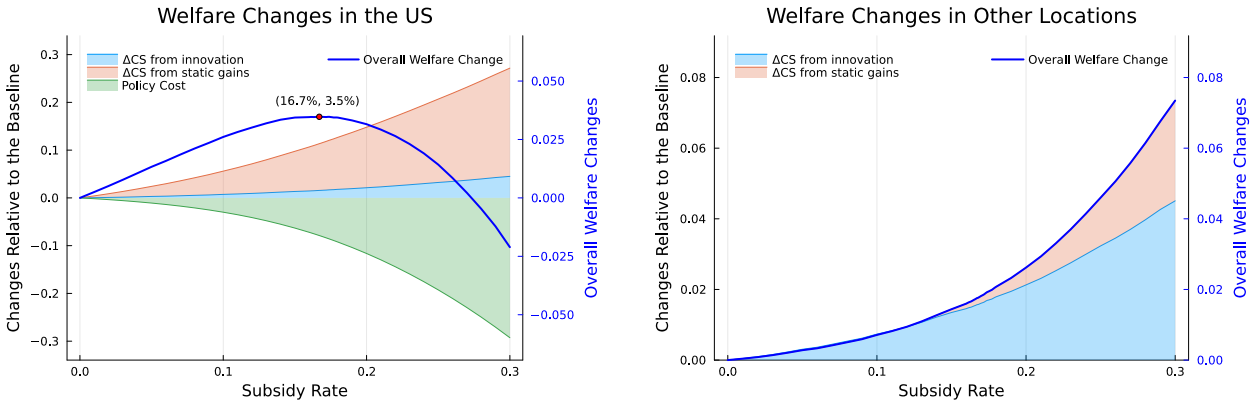
I then examine the optimal investment subsidy rate under varying trade disruption risks (ψ).

[37] When $\delta = 1$, despite different market size, the impact of U.S. subsidies on consumer surplus in CN and the RoW is identical, excluding firm profit.

Figure 10 shows that the optimal subsidy rate increases as ψ rises. With higher disruption risks, the benefits of increased capacity in the U.S. are more likely to be enjoyed exclusively by home consumers rather than being shared globally, leading to a greater increase in U.S. consumer surplus. Regarding the impact on other locations, when ψ is high, firms' capacity allocation decisions become more independent across locations. Even if one location provides substantial subsidies, firms are less likely to reduce capacity in other regions as they would in an integrated global market, where products can be easily shipped across locations. Therefore, the changes in consumer surplus in other locations are smaller when trade disruption risks are higher (see Appendix D.2 for further discussion).

6.2 Dynamic Gains

This section extends the analysis of investment subsidies beyond static welfare implications to explore their impact on consumer welfare by accounting for the dynamic gains or losses resulting from technological changes. In addition to mitigating market power distortions, the presence of industry-level technology spillover externalities, captured by γ in the model, provides further justification for these subsidies.



(a) US (b) Other Locations

Figure 11: Baseline Dynamic Welfare Implications

Notes: These plots show the baseline case where trade disruption risk $\psi = 20\%$ and foreign cost shifter $\delta = 1$. I simulated the model 10,000 times for each subsidy rate and computed the average consumer surplus and policy cost.

In this exercise, I consider a permanent capacity investment subsidy in the U.S., with the policy being known to all firms. Figure 11 shows the changes in welfare compare to the no-subsidy baseline case. Welfare improvements arise from two sources: static gains from addressing market

power distortion and dynamic gains from accelerated innovation. The static gains are calculated as the difference in consumer surplus with and without subsidies, assuming the same firm technology status as in the no-subsidy case; the remainder of the consumer surplus change is attributed to innovation gains. In the baseline case with trade disruption risk of $\psi = 20\%$, the optimal subsidy rate in the U.S. is 16.7%, resulting in a welfare improvement of 3.5%. While dynamic innovation gains further justify the subsidy, the primary benefits come from static gains. Under the optimal subsidy rate, static gains account for approximately 85% of the total gain, while dynamic gains contribute the remaining 15%. Other regions also benefit from U.S. subsidies, particularly through accelerated innovation, as in my model, innovation can be shared within a firm across its fabs globally. A 25% subsidy in the U.S. enhances consumer welfare in other locations by 4.6%, with 70% of the gains stemming from accelerated innovation.

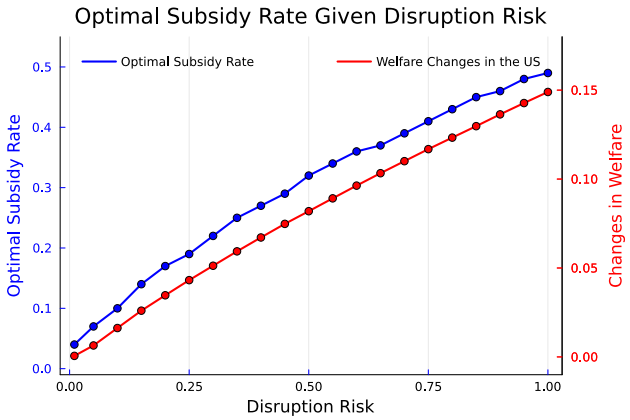


Figure 12: Dynamic Setting: Optimal Subsidy Rate Given Trade Disruption Risks

Notes: The optimal rate refers to the permanent capacity investment subsidy rate that maximizes U.S. consumer welfare, defined as the sum of the present value of the net change in consumer surplus minus policy costs over an infinite horizon. The plot shows the baseline case with the foreign cost shifter set to $\delta = 1$.

Similar to the intuition in the optimal policy for static gains, the optimal subsidy rate, when considering dynamic implications on innovation, also depends on beliefs about trade disruption.[38] As trade disruption risks increase, the optimal subsidy rate rises accordingly (see Figure 12). A 35% trade disruption risk would justify the 25% capacity investment subsidy rate outlined in the U.S. CHIPS Act.

[38] Appendix D.3 provides additional details on the optimal subsidy rate when trade disruption risk is absent or low.

6.3 Investment and Technology Clawbacks

In this section, I examine how investment and technology clawbacks affect the technology race across firms in advanced semiconductor manufacturing. Specifically, I assess whether these policies hinder or accelerate Chinese firms' technological advancement. This exercise simulates three players: TSMC, Samsung, and SMIC.[39] Guardrail restrictions are modeled as: (1) TSMC and Samsung cannot invest in China; (2) technology spillover from non-Chinese to Chinese firms is diminished, with the spillover coefficient becoming $\lambda' = \alpha\lambda$ where $\alpha \geq 1$ reflects the technology blocking magnitude from clawbacks. The simulation starts in 2024Q1 with TSMC and Samsung at 3nm technology, while SMIC is at 7nm (two generations behind). In the baseline scenario, the U.S. subsidy rate is 25%, and China provides no additional subsidy.

Guardrail restrictions are intended to hinder technology spillovers from leading firms to Chinese firms. However, investment clawbacks can also help secure the domestic market for Chinese firms during trade disruptions, enabling them to gain more from technological upgrades.[40] From equation 2 and the R&D specification, the optimal R&D effort is given by:

$$d^*(S_t) = \max\{0, \sqrt{\beta \frac{\Lambda_{gap}(S_t)}{c_d(S_t)}} - 1\},$$

where $\Lambda_{gap}(S_t)$ is the value change from successful R&D, and $c_d(S_t)$ is the unit R&D cost, which depends on the technology spillover λ for non-leading firms. Investment clawbacks increase both Λ_{gap} and c_d for Chinese firms. Which effect dominates hinges on China's market size, disruption risks, and the guardrails' success in reducing technological spillovers.

While estimating α is challenging and outside the scope of this paper, the model allows me to compute Λ^{gap} to infer the minimum α that would lower the Chinese firm's innovation effort. Figure 13 shows the changes in SMIC's value gain from technology upgrades with and without guardrails in the initial period under different disruption risks when $\alpha = 1$. When trade disruption risk is 20%, SMIC's value gain from a technology upgrade with guardrails is 113% higher than without, indicating that α must exceed 2.13 to reduce SMIC's innovation incentive. As disruption risks rise, guardrails ensure a larger effective market for SMIC, making innovation more profitable, thus necessitating a larger α to effectively curb the Chinese firm's innovation. Notably, semiconductor export

[39] SMIC represents the Chinese firm in this model, but it could be another national champion, as non-economic factors might enable other firms to rapidly catch up with national support.

[40] In this model, each vintage corresponds to a separate market, and product innovation allows firms to enter more of these markets.

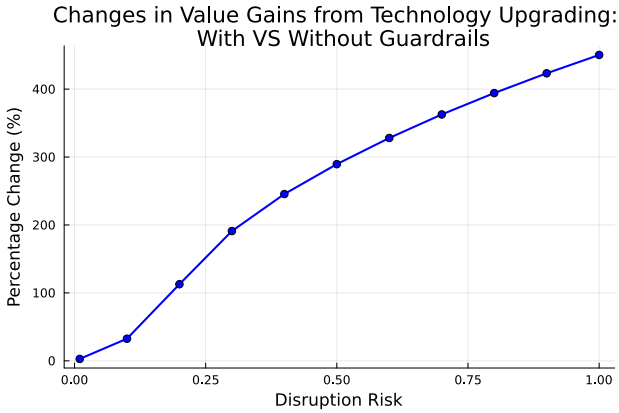


Figure 13: Changes in Value Gains from Technology Upgrading for SMIC from Guardrails

Notes: I calculate the Λ_{gap} for SMIC in 2024Q2 when its technology advances from 7nm to 5nm, comparing scenarios both with and without guardrail restrictions. The technology blocking coefficient is set at $\alpha = 1$, the subsidy rate in the U.S. is set at 25% as outlined in the U.S. CHIPS Act, and the foreign cost shifter is $\delta = 1$. The values displayed in the plot represent the relative change, calculated as $\Lambda_{gap}^{guardrail}/\Lambda_{gap}^{no\_guardrail} - 1$. The plot highlights the minimum value of $\alpha - 1$ at which the presence of guardrails reduces SMIC's incentive to invest in R&D.

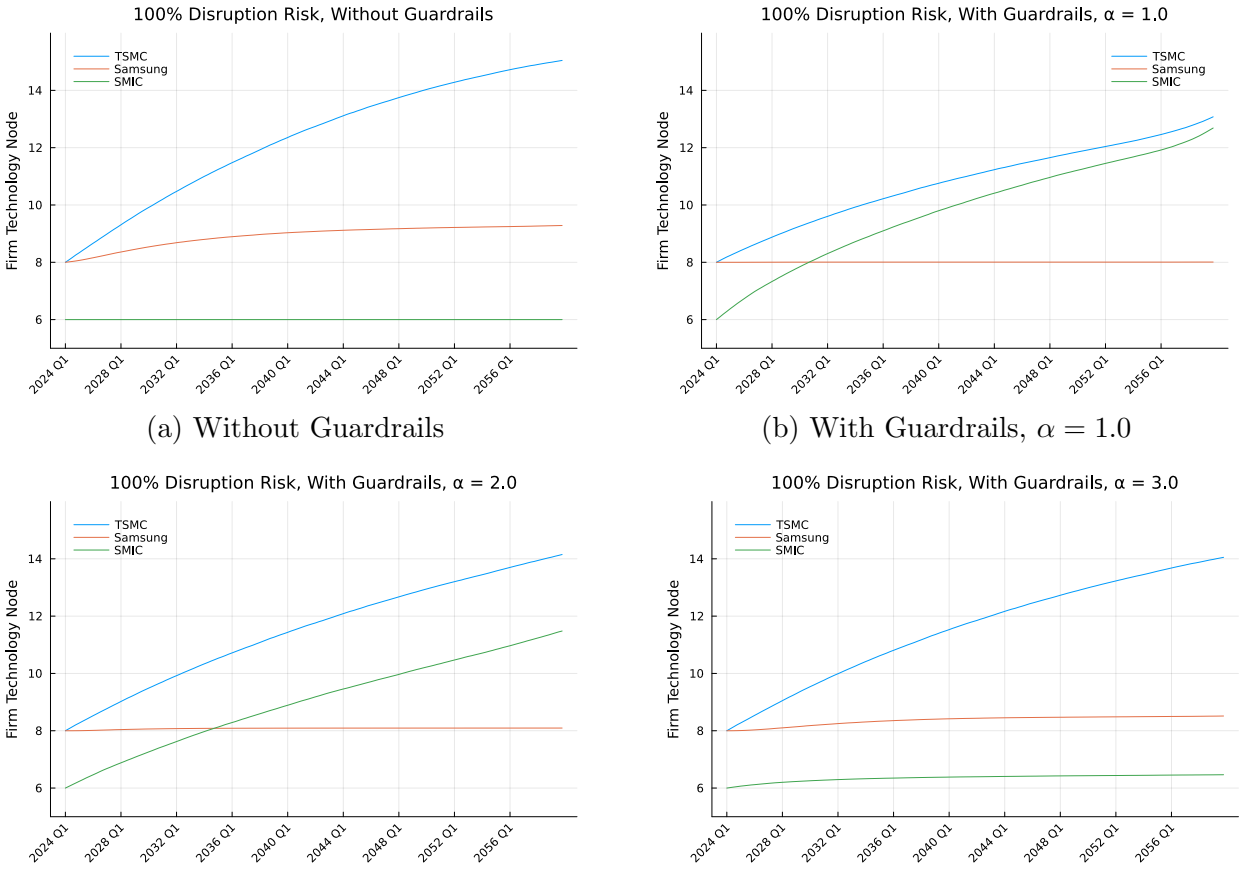
controls to China and the U.S. ban on Huawei-designed chip production are pushing disruption risks in China toward the higher end.

I simulated firm technology trajectories with and without guardrails for different disruption risks (ψ) and technology blocking coefficients (α). Without guardrails, Chinese firms have little incentive to innovate due to low expected returns relative to R&D costs.[41] With guardrails, when disruption risk is low, SMIC's technology growth is limited, even without technology blocking (see Figure D.18). Under high disruption risk, guardrails secure the entire Chinese market for SMIC, boosting its incentive to upgrade as long as spillover dampening isn't too strong. At 100% disruption risk, Figure 14 shows that α must be at least 3 (increasing λ from 22% to 66%) to halt SMIC's advancement.

7 Conclusion

This paper examines how industrial policies affect technology competition and local supply resilience amid trade disruption risks. I explore these questions in the semiconductor foundry industry, which features continuous innovation and has growing strategic importance. I develop and estimate a dynamic oligopoly model with trade uncertainty, where firms innovate and expand capacity across

[41] Samsung's innovation incentives are weak due to its lower inherent productivity compared to TSMC.



(a) Without Guardrails

(b) With Guardrails, $\alpha = 1.0$

(c) With Guardrails, $\alpha = 2.0$

(d) With Guardrails, $\alpha = 3.0$

Figure 14: Simulated Firm Technology Trajectories under 100% Disruption Risk

Notes: These plots show the simulated firm technology trajectories with and without guardrails at a trade disruption risk of $\psi = 100\%$, and foreign cost shifter $\delta = 1$, across different technology blocking coefficients $\alpha \in \{1.0, 2.0, 3.0\}$. I simulated the model 10,000 times for each scenario to compute the average technology trajectory for each firm.

multiple locations, and governments offer subsidies to attract local investments. The model shows that supply resilience creates an additional incentive for government intervention, complementing the need to address traditional market failures, such as market power distortions and technology externalities.

I apply the model to assess the impact of the U.S. CHIPS Act. Investment capacity subsidies enhance local consumer welfare when trade disruptions are a serious concern, without necessarily undermining consumer welfare in other regions. Additionally, the investment restrictions on mainland China may act as effective trade protection for lagging Chinese firms, potentially boosting their R&D incentives.

The objectives of industrial policies can be multifaceted. This paper focuses on their implications on consumer welfare from semiconductor consumption and the technology race between countries. Other important aspects — such as job creation for local households and positive spillovers to other sectors — are beyond the scope of this study. Nonetheless, the paper offers valuable insights into how industrial policies shape technology competition, the production footprint of multinational firms, and consumer welfare in the face of trade disruption risks. While this paper uses the semiconductor manufacturing industry as the context, the model and insights also apply to other industries with similar characteristics, such as the EV battery industry, which features high concentration and continuous innovation.

While this paper focuses on unilateral policy, a natural extension is to explore strategic interactions between governments. There are concerns that the U.S. CHIPS Act could trigger a subsidy race, such as the EU CHIPS Act. Subsidy races tend to drive up optimal subsidy rates across locations, but higher subsidies do not necessarily enhance consumer welfare globally and may instead create a prisoner's dilemma, benefiting only firms. Thus, international coordination has the potential to generate much larger global welfare gains. Additionally, the U.S. CHIPS Act's investment restrictions on China could shift China's optimal policy, incentivizing the Chinese government to further support domestic firms in accessing advanced technologies. This policy shift could influence the competitive dynamics between the U.S. and China in technological development. By extending the framework in this paper to incorporate endogenous government policies, I aim to explore these questions in future research.

References

Aghion, P., S. Bechtold, L. Cassar, and H. Herz (2018). The causal effects of competition on innovation: Experimental evidence. *The Journal of Law, Economics, and Organization 34*(2), 162–195.

Aghion, P., C. Harris, P. Howitt, and J. Vickers (2001). Competition, imitation and growth with step-by-step innovation. *The Review of Economic Studies 68*(3), 467–492.

Baldwin, R. and R. Freeman (2022). Risks and global supply chains: What we know and what we need to know. *Annual Review of Economics 14*(1), 153–180.

Barro, R. J. and J. F. Ursúa (2012). Rare macroeconomic disasters. *Annu. Rev. Econ. 4*(1), 83–109.

Bartelme, D., A. Costinot, D. Donaldson, and A. Rodriguez-Clare (2024). The textbook case for industrial policy: Theory meets data. *Journal of Political Economy*.

Barwick, P. J., M. Kalouptsidi, and N. B. Zahur (2021). *Industrial Policy Implementation: Empirical Evidence from China's Shipbuilding Industry*. Cato Institute Washington, DC, USA.

Bertolotti, F., A. Lanteri, and A. Villa (2024). Investment-goods market power and capital accumulation.

Bonadio, B., Z. Huo, A. A. Levchenko, and N. Pandalai-Nayar (2021). Global supply chains in the pandemic. *Journal of international economics 133*, 103534.

Brander, J. A. and B. J. Spencer (1985). Export subsidies and international market share rivalry. *Journal of international Economics 18*(1-2), 83–100.

Caldara, D. and M. Iacoviello (2022). Measuring geopolitical risk. *American Economic Review 112*(4), 1194–1225.

Castro-Vincenzi, J. (2022). Climate hazards and resilience in the global car industry. *Princeton University manuscript*.

Chen, Y. and D. Y. Xu (2023). A structural empirical model of r&d, firm heterogeneity, and industry evolution. *The Journal of Industrial Economics 71*(2), 323–353.

Choi, J. and A. A. Levchenko (2021). The long-term effects of industrial policy. Technical report, National Bureau of Economic Research.

Crowley, M. A. (2006). Do safeguard tariffs and antidumping duties open or close technology gaps? *Journal of International Economics 68*(2), 469–484.

Eaton, J. and G. M. Grossman (1986). Optimal trade and industrial policy under oligopoly. *The Quarterly Journal of Economics 101*(2), 383–406.

Ederington, J. and P. McCalman (2008). Endogenous firm heterogeneity and the dynamics of trade liberalization. *Journal of International Economics 74*(2), 422–440.

Flamm, K. (2010). *Mismanaged Trade?: Strategic Policy and the Semiconductor Industry*. Brookings Institution Press.

Flamm, K. (2019). Measuring moore's law: evidence from price, cost, and quality indexes. In *Measuring and Accounting for Innovation in the 21st Century*. University of Chicago Press.

Galdin, A. (2024). Resilience of global supply chains and generic drug shortages. *Princeton University manuscript*.

Goettler, R. L. and B. R. Gordon (2011). Does amd spur intel to innovate more? *Journal of Political Economy 119*(6), 1141–1200.

Goldberg, P. K., R. Juhász, N. J. Lane, G. L. Forte, and J. Thurk (2024). Industrial policy in the global semiconductor sector. Technical report, National Bureau of Economic Research.

Grossman, G. M., E. Helpman, and H. Lhuillier (2023). Supply chain resilience: Should policy promote international diversification or reshoring? *Journal of Political Economy 131*(12), 3462–3496.

Hatch, N. W. and D. C. Mowery (1998). Process innovation and learning by doing in semiconductor manufacturing. *Management Science 44*(11-part-1), 1461–1477.

Hung, H.-C., Y.-C. Chiu, and M.-C. Wu (2017). Analysis of competition between idm and fabless–foundry business models in the semiconductor industry. *IEEE Transactions on Semiconductor Manufacturing 30*(3), 254–260.

Igami, M. (2017). Estimating the innovator's dilemma: Structural analysis of creative destruction in the hard disk drive industry, 1981–1998. *Journal of Political Economy 125*(3), 798–847.

Igami, M. (2018). Industry dynamics of offshoring: The case of hard disk drives. *American Economic Journal: Microeconomics 10*(1), 67–101.

Igami, M. and K. Uetake (2020). Mergers, innovation, and entry-exit dynamics: Consolidation of the hard disk drive industry, 1996–2016. *The Review of Economic Studies 87*(6), 2672–2702.

IRDS (2022). The international roadmap for devices and systems: 2022. Technical report, Executive Summary. Technical report, Institute of Electrical and Electronics ....

Irwin, D. A. and P. J. Klenow (1994). Learning-by-doing spillovers in the semiconductor industry. *Journal of political Economy 102*(6), 1200–1227.

Irwin, D. A. and P. J. Klenow (1996). High-tech r&d subsidies estimating the effects of sematech. *Journal of International Economics 40*(3-4), 323–344.

Irwin, D. A. and N. Pavcnik (2004). Airbus versus boeing revisited: international competition in the aircraft market. *Journal of international economics 64*(2), 223–245.

Juhász, R., N. Lane, and D. Rodrik (2023). The new economics of industrial policy. *Annual Review of Economics 16*.

Juhász, R. and N. J. Lane (2024). The political economy of industrial policy. Technical report, National Bureau of Economic Research.

Konings, J. and H. Vandenbussche (2008). Heterogeneous responses of firms to trade protection. *Journal of International Economics 76*(2), 371–383.

Kreps, D. M. and J. A. Scheinkman (1983). Quantity precommitment and bertrand competition yield cournot outcomes. *The Bell Journal of Economics*, 326–337.

Lane, N. (2022). Manufacturing revolutions: Industrial policy and industrialization in south korea. *Available at SSRN 3890311*.

Lashkaripour, A. and V. Lugovskyy (2023). Profits, scale economies, and the gains from trade and industrial policy. *American Economic Review 113*(10), 2759–2808.

Leibovici, F. and A. M. Santacreu (2020). Shortages of critical goods in a global economy: Optimal trade and industrial policy. *FRB St. Louis Working Paper* (2020-10).

Lileeva, A. and D. Trefler (2010). Improved access to foreign markets raises plant-level productivity... for some plants. *The Quarterly journal of economics 125*(3), 1051–1099.

Liu, C.-Y. (1993). Government's role in developing a high-tech industry: the case of taiwan's semiconductor industry. *Technovation 13*(5), 299–309.

Liu, E. (2019). Industrial policies in production networks. *The Quarterly Journal of Economics 134*(4), 1883–1948.

Liu, J., M. Rotemberg, and S. Traiberman (2024). Sabotage as industrial policy. Technical report, National Bureau of Economic Research.

Maggi, G. (1996). Strategic trade policies with endogenous mode of competition. *The American Economic Review*, 237–258.

Miyagiwa, K. and Y. Ohno (1995). Closing the technology gap under protection. *The American Economic Review*, 755–770.

Nakamura, E., J. Steinsson, R. Barro, and J. Ursúa (2013). Crises and recoveries in an empirical model of consumption disasters. *American Economic Journal: Macroeconomics 5*(3), 35–74.

Pakes, A. and P. McGuire (1994). Computing markov-perfect nash equilibria: Numerical implications of a dynamic differentiated product model. *The Rand Journal of Economics 25*(4), 555.

Pierce, J. R. (2011). Plant-level responses to antidumping duties: Evidence from us manufacturers. *Journal of International Economics 85*(2), 222–233.

Sargent Jr, J. F., M. Singh, and K. M. Sutter (2023). Frequently asked questions: Chips act of 2022 provisions and implementation. *Congressional Research Service (CRS) Reports and Issue Briefs*, R47523–R47523.

Thurk, J. (2020). Outsourcing, firm innovation, and industry dynamics in the production of semiconductors. *Department of Economics, University of Notre Dame, Notre Dame, IN*.

Varas, A., R. Varadarajan, J. Goodrich, and F. Yinug (2020). Government incentives and us competitiveness in semiconductor manufacturing. *Boston Consulting Group*, 53–58.

Veerasingam, M. (2023, June). Quantum tunneling and the semiconductors' struggle in the miniaturization race. Medium.

VerWey, J. (2019a). Chinese semiconductor industrial policy: Past and present. *J. Int'l Com. & Econ.*, 1.

VerWey, J. (2019b). Chinese semiconductor industrial policy: Prospects for future success. *J. Int'l Com. & Econ.*, 1.

Yang, C. (2020). Vertical structure and innovation: A study of the soc and smartphone industries. *The RAND Journal of Economics 51*(3), 739–785.

Appendix

A Industry Details

A.1 Technology Node

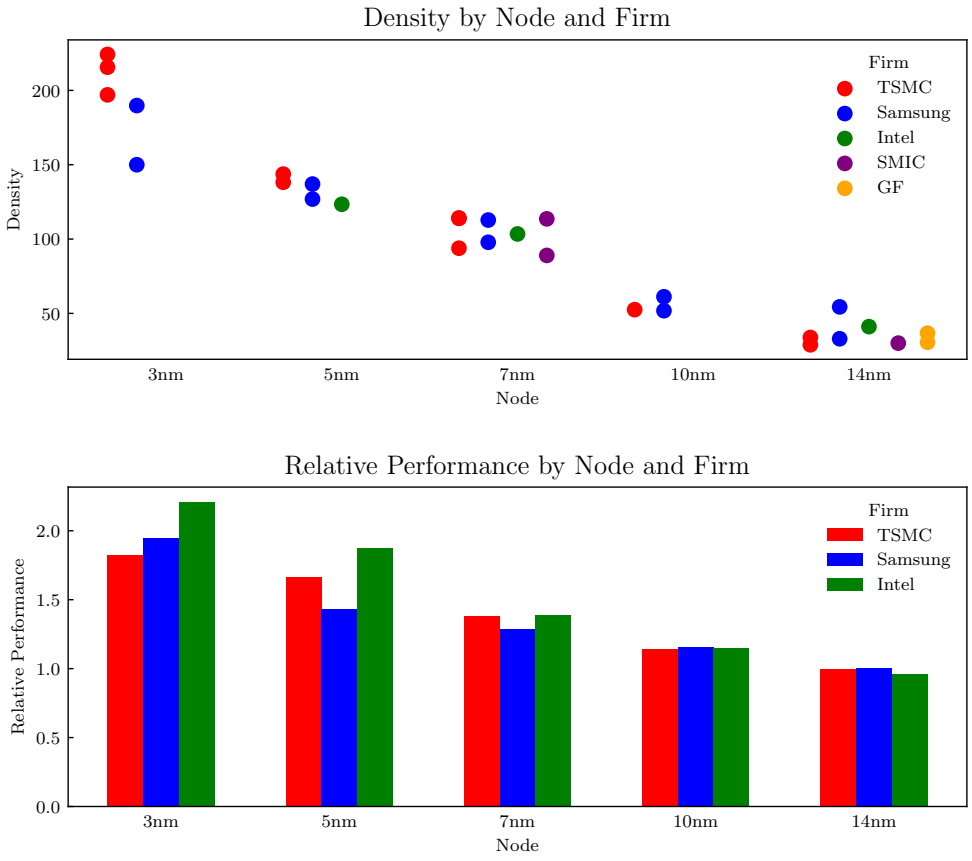


Figure A.1: Semiconductor Industry Evolution

Notes: Transistor density data is sourced from Wikipedia entries such as "3nm process" and "5nm process". The original sources of Wikipedia mainly come from reverse engineering reports of different firms' process nodes. Performance data is adapted from IC Knowledge. The baseline values for different firms are: TSMC/Samsung 16/14nm data from Tom's Hardware and Intel's 10SF, which is considered similar to TSMC's 7nm. These values are then scaled according to company announcements.

Although the technology node no longer reflects actual physical features, it remains an informative indicator for classifying products. Figure A.1 shows the transistor density and the relative performance of different firms by technology node. One firm may have multiple processes within a node, typically upgrading the technology marginally before moving to the next generation. While there is some variation across firms and products within the same node category, the main differences occur across nodes. The performance plot indicates that Intel tends to outperform the two foundries given the technology node. This may be because IDMs optimize technology for their own products,

whereas foundries must develop more general technology for diverse customer products. Overall, the technology node is a valuable indicator for classifying products, particularly among foundries.

A.2 Semiconductor Industry Evolution: Including IDMs

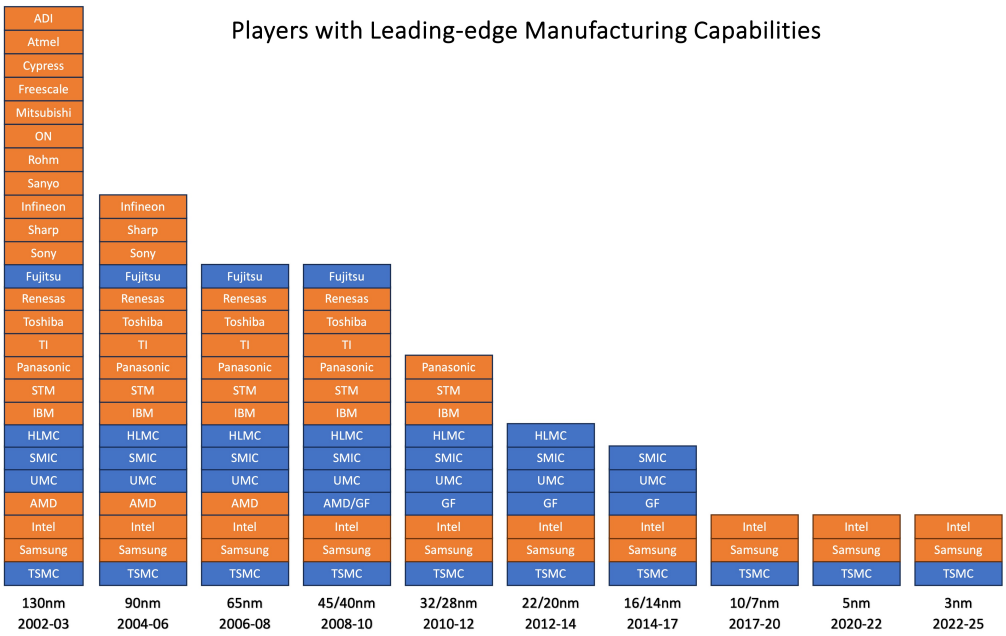


Figure A.2: Semiconductor Industry Evolution

Notes: The plot is adapted from High-End Performance Packaging 3D/2.5D Intergration Report, Yole Development 2020 (page 10). Pure-play foundries are in blue, and integrated device manufacturers (IDM) are in orange. GlobalFoundries (GF) is a spin-off from Advanced Micro Devices (AMD).

The pattern of decreasing number of players holds when including IDMs, as shown in Figure A.2. In the early 2000s, there were more than 20 firms capable of producing the leading-edge chips of that era. However, as technology has advanced, the number of firms able to manufacture leading-edge chips has dwindled. Today, only three firms, i.e., TSMC, Samsung, and Intel, possess the capability to produce these highly advanced chips. Additionally, the firms that currently have advanced technology capabilities are those that also had manufacturing capabilities in the previous generation, , highlighting the step-by-step innovation characteristic of this industry.

A.3 A Simple Model to Rationalize Incremental Innovation

In semiconductor manufacturing, technology upgrading is often embedded in equipment upgrades, which typically require building new production lines or plants. The capital expenditure for constructing new production lines is very high and increases substantially as technology advances. By

skipping some technology nodes and jumping directly to more advanced technologies, firms can potentially save on capital expenditures. However, the success rate of achieving a good yield in advanced technology is often lower compared to more adjacent, incremental technology upgrades.

Denote l as the technology upgrading step of one jump. The success rate of technology upgrading, $\rho(l)$, is a decreasing function of l, while the corresponding cost, $m(l)$, and the expected profit from the new technology, $\pi(l)$, are increasing functions of l. Therefore, the firm's objective function is to maximize the expected net profit:

$$\max_{l \in Z} \rho(l)\pi(l) - m(l).$$

As the success rate of technology upgrading declines rapidly with larger innovation steps, the optimal technology upgrading step is smaller. If $m(l)$ is large, firms have limited opportunities to experiment with multiple trials and may only have one chance to make a technology upgrading decision, making them less likely to take risks. Conversely, if $m(l)$ is small, firms can attempt technology upgrades multiple times until they succeed. This provides a rationale for why leapfrogging is more commonly seen in semiconductor design rather than manufacturing—potentially due to the heavier capital expenditures required in manufacturing.

A.4 Investment Cost Trends for Fixed Nodes

This section explores the dynamics of fab investment costs for fixed nodes, specifically whether these costs increase or decrease over time. The analysis utilizes data from SEMI, which provides equipment and construction costs at the core project level. The data includes 117 IC/Logic core projects with non-missing data on construction and equipment costs. Additionally, the dataset provides information on the timing of equipment installation, the initiation of the first silicon, and the capacity of each core project. However, the corresponding technology nodes are not specified in the data.

To address this gap, I imputed the technology node for each core project using the following criteria: (1) for projects that began construction after 2020, I assigned the planned or current technology node; (2) for projects that started production in 2015 or later, I used the first documented technology node available in the dataset; (3) for projects with production dates before 2015, I referred to core project or fab comments within the dataset or conducted supplementary online searches; (4) in other cases, I assumed the investment was made using the most advanced technology available to the firm in that year.

To assess whether the investment cost of a fixed node declines over time, I conducted the following regression

$$y_{ict} = \alpha + \beta_1 t + \beta_2 X_{ict} + \gamma_c + \epsilon_{ict}.$$

Here, where the outcome variable y_{ict} represents the investment unit cost of project i with technology c at time t, defined as the total investment cost divided by capacity (million USD per wafer per month). The variable of interest, β_1, captures the time trend. X_{ict} includes other control variables such as capacity of the project. Additionally, technology node fixed effects are included to account for varying investment costs across different generations of technology.

The results, as presented in Table 1, indicate that there is no significant time-related trend in investment costs. The coefficients on the time trend are small and insignificant. The results remain robust when controlling for location fixed effects and company fixed effects.

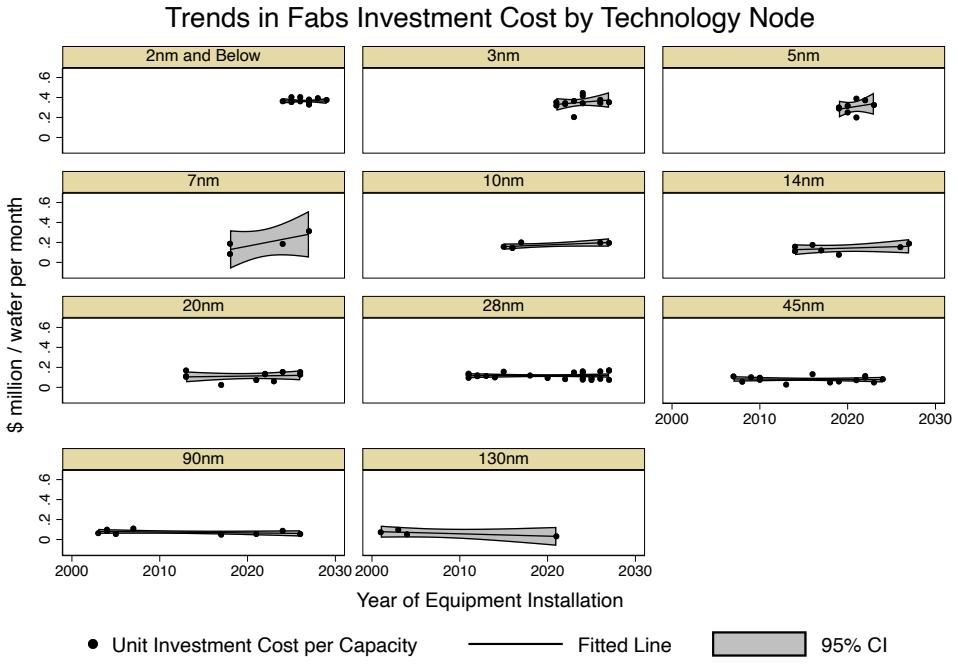


Figure A.3: Trends in Fabs Investment Cost by Technology Node

Notes: Based on data from SEMI and the author's calculations.

To further complement the findings, Figure A.3 presents a raw plot of the unit fab investment cost over time, segmented by technology nodes. This visualization highlights the cost variations across different technology nodes, particularly those with longer lifespans such as 28nm, 40nm, and 90nm.

The plot indicates that there is no significant trend in the investment costs for fixed technology over time, supporting the regression results.

A.5 Operating Margin of Top 4 Pure-Play Foundries

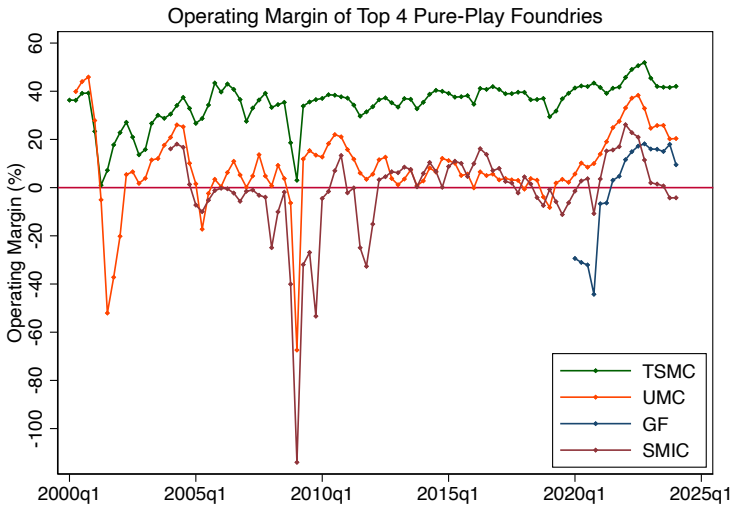


Figure A.4: Operating Margin of Top 4 Pure-Play Foundries

Notes: Data comes from firms' quarterly reports. Operating margin is calculated by dividing the revenue minus the cost of goods sold and operating costs (M&S, G&A, and R&D) by the revenue.

A.6 Mature Technology Revenue

Figure A.5 illustrates TSMC's revenue trend by technology node. The blue line indicates when the technology is considered advanced, while the red dashed line indicates when it is classified as mature. The plot suggests that revenue tends to stabilize once the technology becomes mature.

A.7 R&D Cost Comparison Among Top 4 Pure-Play Foundries

Figure A.6 compares the R&D costs across the top four pure-play foundries. The left panel displays the R&D cost trajectories over time, showing a continuous increase for the leading firm, TSMC, and the catching-up firm, SMIC. UMC's R&D costs have remained relatively stable since 2015, and GF's costs have been stable since 2020 when data became available. Both UMC and GF announced in 2018 that they would not pursue leading-edge technology. The right panel compares their R&D costs relative to TSMC's, showing that the other three top foundries' R&D expenditures are less than half, often around 20% or even lower, compared to TSMC.

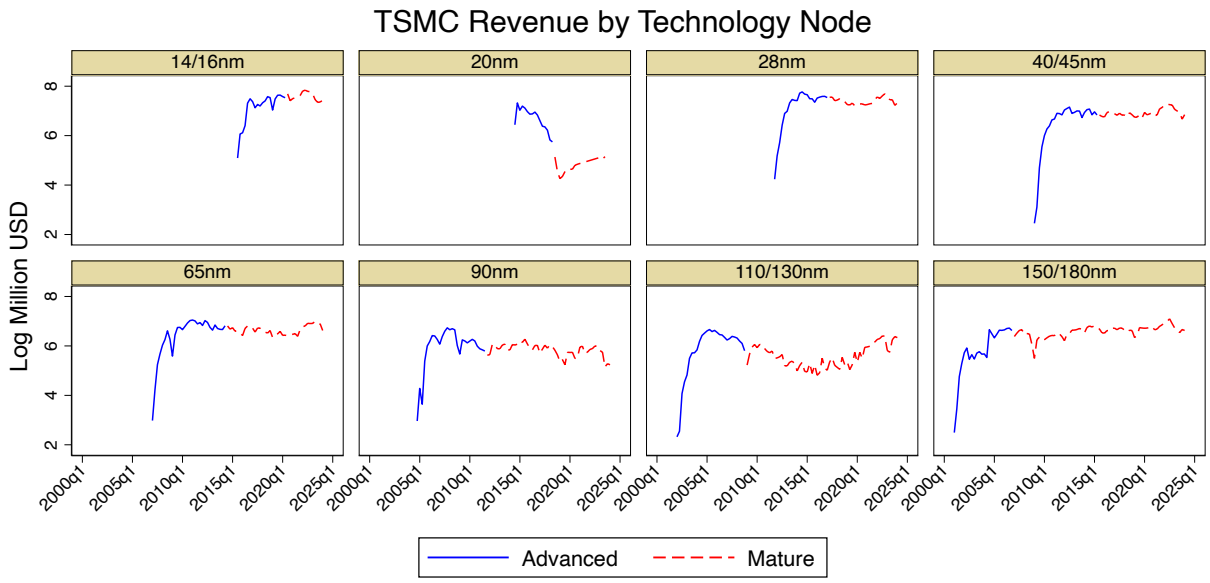


Figure A.5: TSMC Revenue by Technology Node

Notes: Based on TSMC's quarterly report and the author's calculations. Mature technology refers to technology that is at least three generations behind the frontier technology.

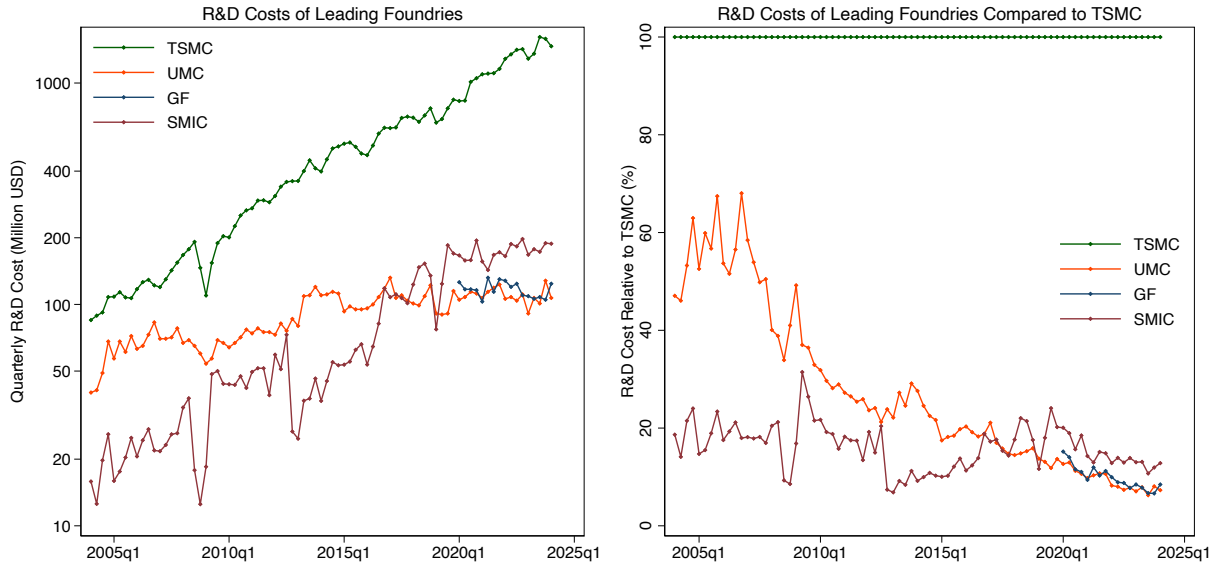


Figure A.6: R&D Cost Comparison Across Top 4 Leading Foundries

Notes: Based on firms' quarterly reports and the author's calculations. GlobalFoundries' quarterly R&D data has been publicly available only since 2020, prior to its IPO in 2021.

B Model Details

B.1 Optimal Shipment Under Capacity Constraints

For each technology node and each destination market, firm i's FOC is

$$P_{nm} + \frac{P_{nm}}{\alpha_P Q_{nm}} q^*_{inm} = c^s_{inm} + \lambda_{inm}$$

with $\lambda_{inm} = 0$ if $q_{inm} < C_{inm}$, and $\lambda_{inm} > 0$ if $q_{inm} = C_{inm}$. The left-hand side can be rewrite as $P_{nm}(1 + s_{inm}/\alpha_P)$, where $s_{inm} \leq 1$ is the market share of firm i in node n and market m. If $\alpha_P < -1$, then $1 + s_{inm}/\alpha_P > 0$, and $P_{nm} > 0$, so the left-hand side is positive. With $c^s_{inm} = 0$, λ_{inm} has to be positive to satisfy the FOC. Thus, $q^*_{inm} = C_{inm}$.

B.2 Spot Market Competition Derivation

The subscripts for technology and destimation markets are omitted for simplicity. For each technology node and each destination market, firm i's FOC is

$$P + \frac{P}{\alpha_P Q} q^*_i = c_i - \lambda_i,$$
$$q^*_i \geq 0,$$

where λ_i is the Lagrangian multiplier, which equals 0 when $q^*_i > 0$ and is greater than 0 when $q^*_i = 0$. Sum across all incumbents, we have

$$nP + \frac{P}{\alpha_P Q} \underbrace{\sum_{i=1}^{n} q^*_i}_{Q} = \sum_{i=1}^{n} (c_i - \lambda_i)$$
$$\implies nP + \frac{P}{\alpha_P} = \sum_{i=1}^{n} (c_i - \lambda_i).$$

When $q^*_i = 0$, FOC implies $P = c_i - \lambda_i < c_i$. Thus,

$$nP + \frac{P}{\alpha_P} = \sum_{i=1}^{n} [\mathbb{1}(c_i < P) c_i + \mathbb{1}(c_i \geq P) P]$$
$$\implies P = \frac{\alpha_P}{1 + \alpha_P \sum_{i=1}^{n} \mathbb{1}(c_i < P)} \sum_{i=1}^{n} \mathbb{1}(c_i < P) c_i.$$

To implement the computation, the following procedure is used:

1. Calculate $P^1 = \frac{\alpha_P}{1+n\alpha_P} \sum_{i=1}^{n} c_i$.

2. Check if $c_i < P^1$ for all i. If this condition is met, set $P = P^1$. If not, update P as follows:

$$P^2 = \frac{\alpha_P}{1 + \alpha_P \sum_{i=1}^{n} \mathbb{1}(c_i < P^1)} \sum_{i=1}^{n} \mathbb{1}(c_i < P^1) c_i.$$

3. Check if $c_i < P^2$ for all i such that $c_i < P^1$. If this condition is met, set $P = P^2$. If not, repeat the update step until $c_i < P^{j+1}$ for all i such that $c_i < P^j$.

B.3 Defining Market Size Across Regions

Suppose the demand for each homogeneous individual is given by:

$$\log q_i = \alpha_0 + \alpha_p \log P \implies q_i = \exp(\alpha_0) P^{\alpha_p}.$$

With N customers, the aggregate demand curve is

$$Q = \sum_{i=1}^{N} q_i = N \exp(\alpha_0) P^{\alpha_p} \implies \log Q = \alpha_0 + \log N + \alpha_p \log P.$$

The estimated aggregate demand curve has the form:

$$\begin{aligned} \log Q &= \tilde{\alpha}_0 + (\alpha_D^{PC} \log Q_{PC} + \alpha_D^{mobile} \log Q_{mobile}) + \alpha_p \log P \\ &= \tilde{\alpha}_0 + \log(Q_{PC}^{\alpha_D^{PC}} Q_{mobile}^{\alpha_D^{mobile}}) + \alpha_p \log P, \end{aligned}$$

where $\tilde{\alpha}_0 = \alpha_0 + D_n' \alpha_D - \alpha_D^{PC} \log Q_{PC} - \alpha_D^{mobile} \log Q_{mobile}$ captures demand shifts related to product quality but not market size. The term $Q_{PC}^{\alpha_D^{PC}} Q_{mobile}^{\alpha_D^{mobile}}$ is used as a proxy for market size. Define $\hat{N}_i = Q_{i,PC}^{\alpha_D^{PC}} Q_{i,mobile}^{\alpha_D^{mobile}}$ and $\tilde{N} = Q_{PC}^{\alpha_D^{PC}} Q_{mobile}^{\alpha_D^{mobile}}$, where $Q_{PC} = \sum_i Q_{i,PC}$ and $Q_{mobile} = \sum_i Q_{i,mobile}$. However, $\tilde{N} \neq \sum_i \hat{N}_i$ in general. Thus, we adjust the market size for each region i as $\tilde{N}_i = \tilde{N} \frac{\hat{N}_i}{\sum \hat{N}_i}$.

B.4 Iterative Algorithm to Solve the Optimal Capacity Allocation

The iterative algorithm is outlined as follows:

1. Take initial guesses for market level price indexes P_n^{int} and P_{nm} for all m and denote them as $\hat{P}$, then compute the corresponding market demand from the demand equations and denote it as $\hat{Q}$.

2. For each firm i, taking the market level prices and outputs as given, compute the optimal capacity level in each location according to the equation 5:

$$\begin{bmatrix} A^{int} + A^{US} & A^{int} & A^{int} \\ A^{int} & A^{int} + A^{CN} & A^{int} \\ A^{int} & A^{int} & A^{int} + A^{RoW} \end{bmatrix} \begin{bmatrix} C_{in,US}^* \\ C_{in,CN}^* \\ C_{in,RoW}^* \end{bmatrix} = \begin{bmatrix} \kappa_{in,US} - \psi \hat{P}_{n,US} \\ \kappa_{in,CN} - \psi \hat{P}_{n,CN} \\ \kappa_{in,RoW} - \psi \hat{P}_{n,RoW} \end{bmatrix} - (1 - \psi) \hat{P}_n^{int},$$

where $A^{int} \equiv (1-\psi)\frac{\hat{P}_n^{int}}{\alpha_P \hat{Q}_n^{int}}$ and $A^m \equiv \psi \frac{\hat{P}_{nm}}{\alpha_P \hat{Q}_{nm}}$. Ensure $C_m^* > 0$ for all m. If any C_m^* is not positive, set the corresponding $C_m^* = 0$ and resolve the system without considering the FOC for that market m.

3. Compute the implied total market shipment from the firm's optimal capacity:

$$\hat{Q}_n^{int} = \sum_{i,m} C_{inm}^*, \hat{Q}_{nm} = \sum_i C_{inm}^*.$$

4. Derive the corresponding price levels using the inverse demand equation.

5. Update the guesses for price indexes and iterate until convergence.

B.5 Two-Firm Model Illustration

This section presents a simplified two-firm model to illustrate the intuition behind firms' innovation decisions in an industry characterized by incremental R&D and oligopoly competition. The model also considers investment subsidies, providing insights into how these subsidies influence the innovation process and firms' relative technology positions. Given the dynamic nature of the model, it can address questions such as whether subsidies can help lagging firms catch up when there are varying technology gaps, and whether the effectiveness of industrial policies depends on their timing.

The simplified model differs from the full model by: (1) focusing on a duopoly scenario where only one firm receives investment subsidies; (2) simplifying the static game by removing trade disruption risks and assuming a specific profit function where only the leading firm earns a positive profit π in all monopoly technology nodes, while firms earn zero profit when both own the technology.

Additional modifications include: (1) removing the assumption that firms significantly lagging behind will no longer upgrade their technology, instead, I use this two-firm model to illustrate the rationale behind this assumption; (2) incorporating a non-zero fixed cost f associated with technology upgrading, which is influenced by investment subsidies.

Set-up Time is discrete with a finite horizon T, at which point the existing technology become obsolete. The model considers two firms located in different countries, each subject to distinct investment subsidies. Initially, both firms possess the capability to produce identical products using first-generation technology. In each period, nature selects one firm at random to undertake R&D.

This random sequential moving structure follows Igami and Uetake (2020) to ensure the model's tractability in a nonstationary environment. The probability of successful R&D, ρ, is a function of the R&D effort d, specifically defined as $\rho(d) = \frac{d}{1+d}$. The cost associated with this innovation effort is $c_d d$. Upon successful R&D, firms must pay a fixed cost f to establish the necessary facilities. The model allows for a maximum of N generations of technology before reaching a physical limit. In the product market, if both firms have the same technology, their profits are zero. Conversely, if technological levels differ, the leading firm realizes a profit of π, while the lagging firm receives no profit.[42] This framework streamlines the analysis by concentrating exclusively on the innovation strategies of firms and the impact of subsidy policies on these strategies and overall market outcomes.

Dynamic Problem Suppose nature chooses firm i at time t, and firm i has not yet reached the technology limit ($x_{it} < N$), the Bellman equation for the firm i's dynamic optimization problem is given by

$$V_{it}(x_{it}, x_{jt}) = \pi_{it}(x_{it}, x_{jt}) + \max_{d \geq 0} \left\{ -c_d d + \beta \left[\begin{array}{c} \frac{d}{1+d} \left(\Lambda_{it+1}(x_{it}+1, x_{jt}) - f \right) \\ + \frac{1}{1+d} \Lambda_{it+1}(x_{it}, x_{jt}) \end{array} \right] \right\}, \tag{10}$$

where the state variables x_{it} and x_{jt} are the technology status of the two firms at time t, and $\Lambda_{i,t+1}$ represents i's expected value at $t+1$ before nature picks the mover at $t+1$, defined as follows

$$\Lambda_{it}(x_{it}, x_{jt}) = \frac{1}{2} \left[V_{it}(x_{it}, x_{jt}) + W_{it}(x_{it}, x_{jt}) \right].$$

Here, $W_{i,t}$ is the value for firm i at t if nature selects the other firm $j \neq i$ at t, formulated as

$$W_{it}(x_{it}, x_{jt}) = \pi_{it}(x_{it}, x_{jt}) + \beta \left[\begin{array}{c} \frac{d_j^*(x_{it}, x_{jt})}{1+d_j^*(x_{it}, x_{jt}+1)} \Lambda_{it+1}(x_{it}, x_{jt}+1) \\ + \frac{1}{1+d_j^*(x_{it}, x_{jt}+1)} \Lambda_{it+1}(x_{it}, x_{jt}) \end{array} \right],$$

where d_j^* is optimal R&D level chosen by the other firm j when it is the mover.

Optimal R&D When the net benefit of upgrading technology exceeds the cost of innovation such that $\Lambda_{it+1}(x_{it}+1, x_{jt}) - f - \Lambda_{it+1}(x_{it}, x_{jt}) > \frac{c_d}{\beta}$, the optimal R&D level, d_i^*, satisfies the first order condition of the Bellman equation 10,

$$c_d = \frac{\beta}{(1 + d_i^*(x_{it}, x_{jt}))^2} \left[\Lambda_{it+1}(x_{it}+1, x_{jt}) - f - \Lambda_{it+1}(x_{it}, x_{jt}) \right]. \tag{11}$$

[42] For example, under the assumption that products are homogeneous and firms compete on price, only the first innovator is able to capture monopoly profits before the entry of a second firm, while subsequent entrants can only earn zero profit.

This equation equates the marginal cost of R&D, c_d, to the discounted marginal benefit of R&D, adjusted for the effect of the R&D success probability, which decreases as d_i^* increases, reflecting the diminishing returns returns on further R&D effort.

The model can be adapted to a continuous time framework, similar to the one described in Aghion et al. (2001). In this setting, there is no discrete R&D success rate; instead, there is a Poisson hazard rate d representing the rate at which technology advances by one step. The cost of R&D is a convex increasing function of the technology upgrading rate d, for example, $c(d) = c_d\frac{d^2}{2}$. The Bellman equation of firm i at time t is

$$V_{it}(x_{it}, x_{jt}) = \max_{d\geq 0}\left\{(\pi_{it}(x_{it}, x_{jt}) - c_d\frac{d^2}{2})\Delta t + e^{-r\Delta t}\begin{bmatrix} d\Delta t\left(V_{it+\Delta t}(x_{it}+1, x_{jt}) - f\right) \\ +d_j^*(x_{it}, x_{jt})\Delta t V_{it+\Delta t}(x_{it}, x_{jt}+1) \\ +(1 - d\Delta t - d_j^*(x_{it}, x_{jt})\Delta t)V_{it+\Delta t}(x_{it}, x_{jt})\end{bmatrix}\right\}.$$

Results from discrete time and continuous time framework are qualitatively same.

Computational Strategy The model can be solved using backward induction. The terminal values associated with the firms' states are simply the static profits the firms gain in this period.

$$\Lambda_{iT}(x_{iT}, x_{jT}) = \begin{cases}(x_{iT} - x_{jT})\pi & \text{if } x_{iT} > x_{jT} \\ 0 & \text{if } x_{iT} \leq x_{jT}\end{cases}.$$

At $T-1$, given the values of $\Lambda_{i,T}$ for all possible states, the mover determines its optimal R&D level according to the equation 11. The values for both the mover and the non-mover at $T-1$ can be computed after knowing the mover's optimal R&D level at $T-1$. Consequently, $\Lambda_{i,T-1}$ can be determined. By iterating this process back to the initial period, the entire game can be solved.

B.5.1 Innovation Patterns

Figure B.7 displays the heatmap of optimal R&D effort given the technology levels of both firms. The heatmap shows that innovation intensities are highest when firms have similar technology levels, echoing the findings of Aghion et al. (2018) that neck-and-neck competition generates the strongest innovation incentives. Conversely, when the technology gap between the two firms exceeds a certain threshold, the lagging firm loses the incentive to catch up and invests nothing in R&D. This provides insights into why significantly lagging firms might opt not to pursue catching up, which aligns with the assumptions outlined in the full model. A key assumption is that innovation is step-by-step, requiring the lagging firm to commit to multi-period efforts to catch up, with positive returns only achievable when it becomes the leading firm. This pattern is also consistent

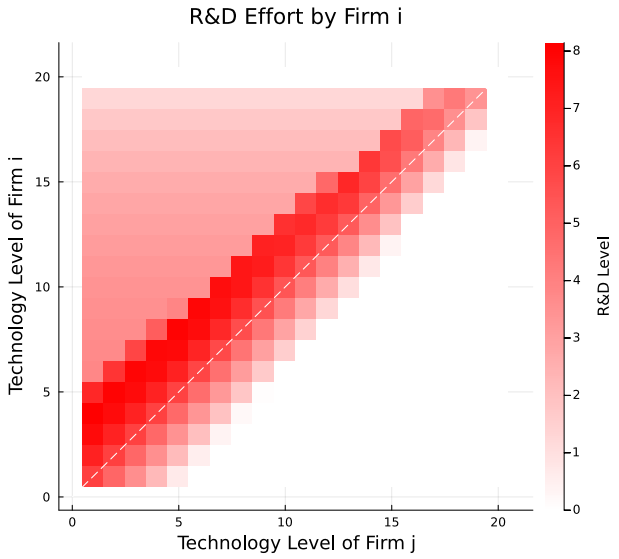


Figure B.7: Optimal R&D Level Given Firms' Technology States

Notes: This graph plots firm i's optimal R&D decisions in the initial period given the technology levels of firm i and j's technology levels. This pattern holds in subsequent periods. The parameter values used for the simulation are: $\beta = 0.9^{1/4}, N = 20, T = 120, \pi = 30, f = 60$ and $c_d = 30$.

with empirical observations that firms rarely attempt technology upgrades when they fall several generations behind frontier technologies.

Additionally, the heatmap shows that when technologies are in their earlier stages, the technology gap that deters the lagging firm from investing in R&D is larger compared to when technologies are more advanced. Advancing to new generations offers option values related to entering subsequent generations; however, this option value diminishes as technologies mature and approach their physical limits.

B.5.2 Effects of Investment Subsidies

This sections explores the impact of industrial policies, specifically focusing on the investment subsidy rate τ that reduces the cost of building facilities from f to $(1-\tau)f$. Consequently, the equation for the optimal R&D level is as follows:

$$c_d = \frac{\beta}{(1 + d_i^*(x_{it}, x_{jt}))^2} \left[\Lambda_{it+1}(x_{it}+1, x_{jt}) - (1-\tau)f - \Lambda_{it+1}(x_{it}, x_{jt})\right]. \tag{12}$$

The effects of investment subsidies on firm profits, after subtracting policy costs, are twofold. On the one hand, they can distort R&D decisions by inducing firms to engage in excessive R&D, where the marginal cost exceeds the marginal benefit when considering policy costs, leading to inefficiency.

On the other hand, higher R&D incentives can influence rival firms' behavior. If the rival reduces its R&D effort, the subsidized firm benefits from both the subsidy and reduced competition, generating strategic gains. Similar insights can be found in the strategic trade literature(Brander and Spencer, 1985). Strategic gains or losses result from the adjustment in the R&D efforts of the rival firm. If the strategic gains surpass the inefficiency losses, it provides a rationale for industrial policies, even abstracting from consumer welfare.

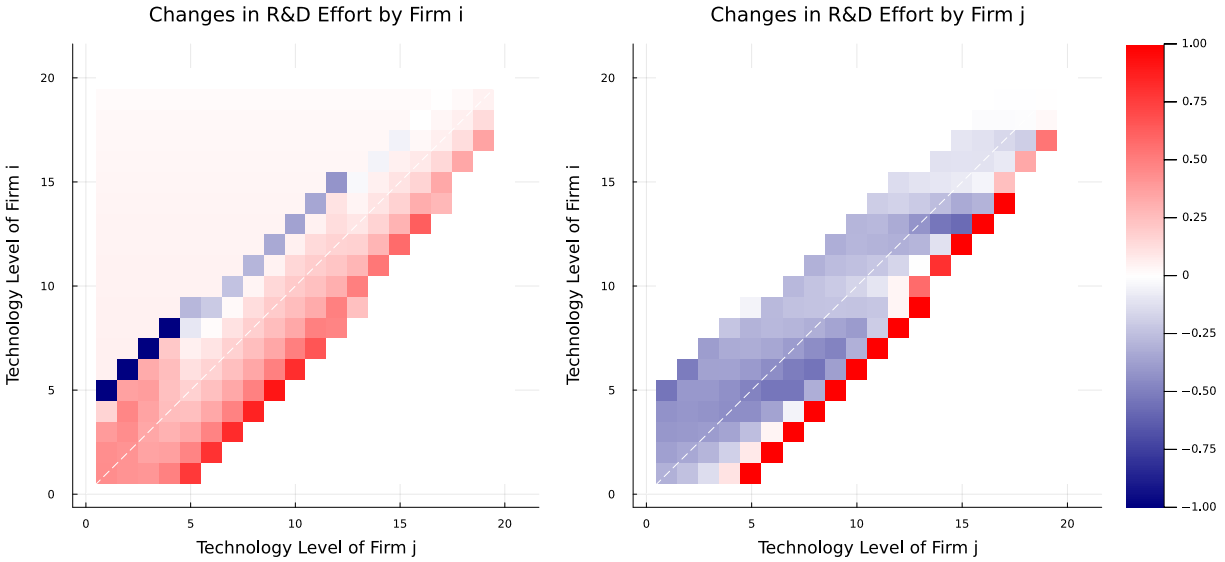


Figure B.8: Changes in Firms' R&D: With and Without Subsidies

Notes: These graphs show the changes in R&D effort by firms i and j when firm i receives a 50% investment subsidy every period compared to no subsidy. The changes are capped at 1 and -1 for clearer visualization.

Figure B.8 illustrates the changes in R&D effort by the domestic firm (i) and the rival firm (j) given the technology positions of both firms when the domestic firm is subsidized versus when it is not. Investment subsidies increase the value of technology advancement by covering part of the fixed cost associated with it, so subsidized firm tends to increase its R&D effort. However, there are scenarios where the subsidized firm might reduce its R&D. This happens because the rival also changes its R&D strategies, affecting the overall strategic environment. When the subsidized firm is leading or closely trailing the rival, the rival firm tends to reduce its R&D investment. In this scenario, the incentive for R&D diminishes for the rival firm because the subsidized firm is likely to invest more heavily in R&D, reducing the rival's chances of gaining or maintaining a technological lead. Consequently, the option value associated with advancing to future generations decreases for the rival. Conversely, if the rival firm holds a substantial technology advantage over the subsidized firm, it may increase its R&D investment to prevent the subsidized firm from catching up, thereby

maximizing potential profits in future generations. Intuitively, with a significant technology lead, the rival firm may perceive no substantial threat from the subsidized firm, even with subsidies, and may choose to invest more in R&D to secure and extend its lead further. Lastly, if the technology gap becomes too large, the R&D response from the both firms diminishes, indicating that the policy becomes ineffective when the gap between the two firms is sufficiently large. The qualitative results of this simulation are robust under many different parameterizations, though the exact magnitudes should not be interpreted literally.

B.5.3 Technology Upgrading Dynamics

To explore how investment subsidies affect technology upgrading dynamics across different initial technology statuses, Figure B.9 shows the technology upgrading dynamics with and without investment subsidies to domestic firms, considering different initial technology gaps. In all cases, the subsidized firm upgrades its technology levels more than in the scenario without subsidies. Meanwhile, the rival firm reduces its R&D efforts, creating strategic gains for the domestic firm. However, the magnitude of the rival's technology level decline decreases as the rival's lead increases, indicating that strategic gains diminish when the domestic firm lags further behind. The simulation also shows that a significantly large technology gap makes it very difficult for the lagging firm to catch up, even with subsidies.

Impact of Policy Timing An interesting question is whether the timing of industrial policies matters. Should we expect similar outcomes when these policies are implemented at different stages of the industry life cycle? To explore this, I examine three cases of investment subsidies with the same rate and duration but applied at different times: during the nascent period of the industry versus more mature stages when there may already be an established leader. Figure B.10 shows that the policy's effectiveness weakens when implemented in a mature industry. As the industry matures, the technology gap between firms widens, making it harder for the policy to reshape the competitive positioning of firms. This aligns with Goldberg et al. (2024), which notes that as industries mature, government involvement decreases, giving way to the private sector. This could also help explain why some industrial policies are successful while others may not be.

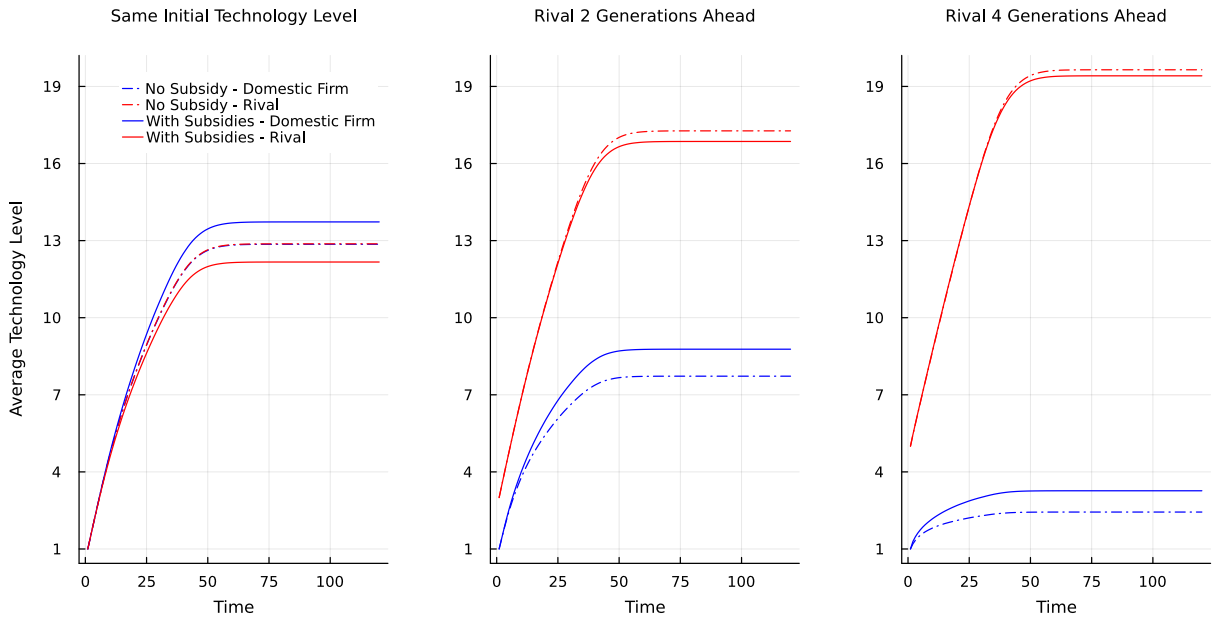


Figure B.9: Technology Upgrading Dynamics - Different Technology Gap

Notes: These graphs show the average technology levels over time from 50,000 simulations across different initial technology levels. In all scenarios with subsidies, a 50% investment subsidy is provided to firm i in all periods. From left to right: the first plot shows both firms starting with the same initial technology level, the second shows the rival firm starting two generations ahead of the domestic firm, and the third shows the rival firm starting four generations ahead.

B.6 Micro-Foundation of the Demand Curve

Each market m consists of a mass of homogeneous consumers L_m demanding goods from multiple sectors, with preferences specified by

$$U_m = \Pi_s (c_m^s)^{\beta^s} \text{ where } \sum_s \beta^s = 1.$$

A share β^s of expenditure is allocated to consumption in sector s. Each sector requires a specific generation of chips and a homogeneous, freely tradable numeraire good. Sectoral consumption is specified as follows:

$$c_m^s = q_0 + \exp(\frac{\alpha_0^s}{-\alpha_p}) \frac{{q_c^s}^{1+\frac{1}{\alpha_p}}}{1+\frac{1}{\alpha_p}},$$

where q_0 denote the numeraire good and q_c^s the semiconductor quantity. α_0^s captures the generation-specific demand shifter, which reflects a combinations of quality, customer base, and preference to frontier technologies. The term $-\alpha_p$ is the price elasticity. The demand for each specific type of

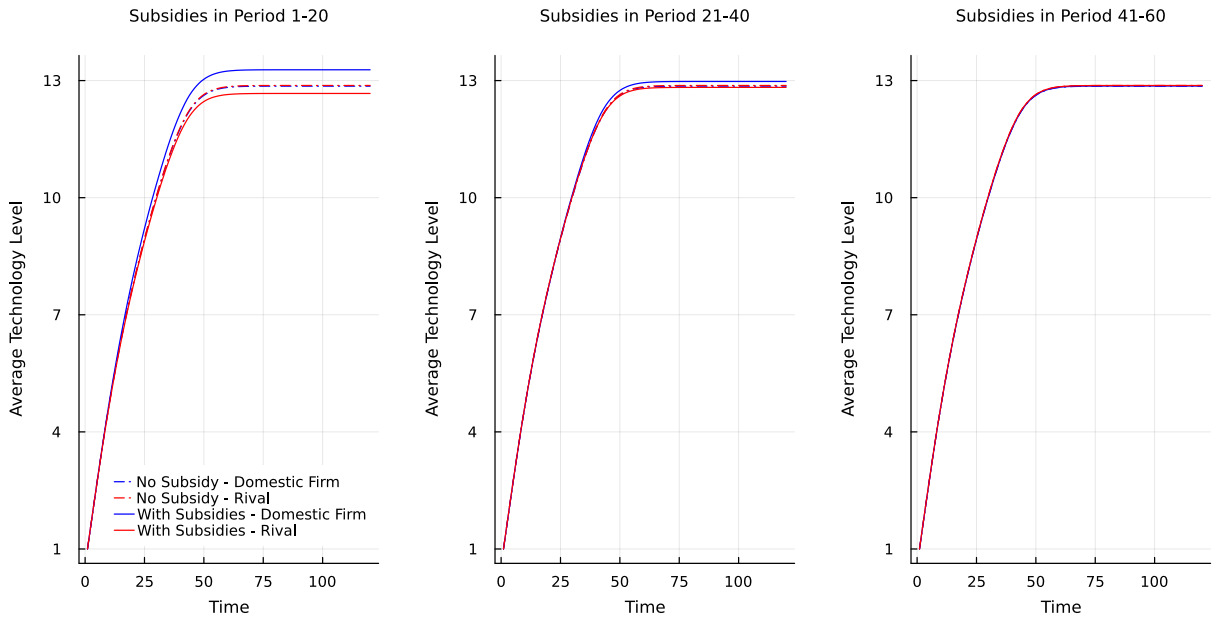


Figure B.10: Technology Upgrading Dynamics - Different Policy Timing

Notes: These graphs show the average technology levels over time from 50,000 simulations across different policy implementation periods. From left to right: the first plot shows investment subsidies to firm i from periods 1 to 20, the second shows unexpected investment subsidies to firm i from periods 21 to 40, and the third shows unexpected investment subsidies to firm i from periods 41 to 60.

chip q_c^s is given by

$$\log(q_c^s) = \alpha_0^s + \alpha_p \log(p_c^s).$$

B.7 Consumer Surplus

The consumer surplus when the price is p_0 is

$$CS(p_0) = \tilde{N}_i \int_{p_0}^{\bar{p}} D(p)dp,$$

where $\bar{p}$ denotes the choke price. Suppose the price changes from p_0 to p_1. The resulting change in consumer surplus is

$$\begin{aligned}
\Delta CS(p_0, p_1) &= \tilde{N}_i(\int_{p_1}^{\bar{p}} D(p)dp - \int_{p_0}^{\bar{p}} D(p)dp) \\
&= \tilde{N}_i \int_{p_1}^{p_0} D(p)dp \\
&= \tilde{N}_i \int_{p_1}^{p_0} \exp(\alpha_0) p^{\alpha_P} dp \\
&= \tilde{N}_i \left. \frac{\exp(\alpha_0) p^{\alpha_P+1}}{\alpha_P + 1} \right|_{p_1}^{p_0} \\
&= \tilde{N}_i \frac{\exp(\alpha_0)}{\alpha_P + 1} \left(p_0^{\alpha_P+1} - p_1^{\alpha_P+1} \right).
\end{aligned}$$

B.8 Decoupling Risk

While the current model assumes *i.i.d.* trade disruption shock, it can be extended to account for persistent shocks. In a decoupling scenario, firms assume 100% disruption risk after decoupling. The model is solved in two steps. First, solve for $\psi = 100\%$ using the current model, denoting the values as Λ^{decoup}. Second, let $100\% > \psi > 0$ represent decoupling risk, and the Bellman equation for innovator becomes:

$$V_{it}(S_t) = \pi_{it}(S_t) + \max_{d \geq 0} \left\{ \begin{array}{c} -c_{i,d}(S_t)d + \beta(1-\psi) \left[\begin{array}{c} \rho(d)[(\Lambda_{it+1}(S_{t+1} \mid S_t, a_{it} = 1) \\ +\mathbb{1}(\bar{n}_{t+1} > \bar{n}_t)\Pi_{it}^{n=2}(S_t)] \\ +[1-\rho(d)]\Lambda_{it+1}(S_{t+1} \mid S_t, a_{it} = 0) \end{array} \right] \\ +\beta\psi \left[\begin{array}{c} \rho(d)[(\Lambda_{it+1}^{decoup}(S_{t+1} \mid S_t, a_{it} = 1) \\ +\mathbb{1}(\bar{n}_{t+1} > \bar{n}_t)\pi_{it}^{n=2,decoup}(S_t)/(1-\beta)] \\ +[1-\rho(d)]\Lambda_{it+1}^{decoup}(S_{t+1} \mid S_t, a_{it} = 0) \end{array} \right] \end{array} \right\},$$

where $\Pi_{it}^{n=2}(S_t)$ represents the present value of the sum of future expected profits for the generation that is three generations behind the frontier technology. Specifically,

$$\Pi_{it}^{n=2}(S_t) = \psi \frac{\pi_{it}^{n=2,decoup}(S_t)}{1-\beta} + (1-\psi) \left[\pi_{it}^{n=2,no\_decoup}(S_t) + \beta \Pi_{it}^{n=2}(S_t) \right].$$

Similarly, the Bellman equation for non-innovator becomes

$$\begin{aligned}
W_{it}^{j}(S_t) =& \pi_{it}(S_t) + \beta(1-\psi) \left[\begin{array}{c} \rho(d_j^*(S_t))[\Lambda_{it+1}(S_{t+1} \mid S_t, a_{jt} = 1) \\ +\mathbb{1}(\bar{n}_{t+1} > \bar{n}_t)\Pi_{it}^{n=2}(S_t)] \\ +\left[1-\rho(d_j^*(S_t))\right]\Lambda_{it+1}(S_{t+1} \mid S_t, a_{jt} = 0) \end{array} \right] \\
&+ \beta\psi \left[\begin{array}{c} \rho(d_j^*(S_t))[\Lambda_{it+1}^{decoup}(S_{t+1} \mid S_t, a_{jt} = 1) \\ +\mathbb{1}(\bar{n}_{t+1} > \bar{n}_t)\pi_{it}^{n=2,decoup}(S_t)/(1-\beta)] \\ +\left[1-\rho(d_j^*(S_t))\right]\Lambda_{it+1}^{decoup}(S_{t+1} \mid S_t, a_{jt} = 0) \end{array} \right].
\end{aligned}$$

The model can be solved using backward induction, and the terminal value now becomes:

$$\Lambda^T(\psi) = \psi \frac{\pi^{decoup}}{1-\beta} + (1-\psi)\left[\pi^{no\_decoup} + \beta\Lambda^T(\psi)\right].$$

The main results of the current model remain qualitatively robust in the decoupling scenario. Quantitatively, a small decoupling risk produces similar magnitudes to a larger *i.i.d.* trade disruption shock.

B.9 Friend-shoring

Friend-shoring can be addressed in two ways. First, assuming zero trade disruption risk within allied regions, capacity allocation involves two steps: first, find the most efficient location within each friendly region, effectively treating allies as one integrated market, and second, allocate capacity optimally as per the current model. The more flexible approach considers asymmetric and idiosyncratic shocks, where friendly regions face lower risks. In this case, we adjust the static step by calculating expected profit by summing across all 2^N possible shock realization outcomes for N locations.

C Estimation Details

C.1 Firm-Level Productivity Estimation

The market share data excludes Samsung, so I first calibrate the relative firm productivity of TSMC and Samsung using capacity data for technology nodes 5nm and below. The average capacity share for TSMC is approximately 73.1%, while Samsung's share is about 26.9%. I aggregated the capacity for 5nm and 3nm nodes because Samsung has converted some of its 5nm capacity to 3nm, resulting in a relatively stable combined capacity share. Capacity data for other firms was not used for estimation because some firms utilize production lines for multiple nodes, and the data typically only reflect the highest-performing node. The calibrated relative unit cost between TSMC and Samsung is 1 versus 1.61. For other firms, I use sales share data by node, focusing on periods when all four firms are incumbents, as market shares are more stable during these times. On average, TSMC holds a 64.2% market share, followed by UMC at 16.9%, GF at 11.7%, and SMIC at 7.2%. From the model, the corresponding unit costs relative to TSMC are 1.50 for UMC, 1.56 for GF, and 1.60 for SMIC.

C.2 Likelihood-Ratio Test

I use likelihood-ratio tests to estimate the confidence intervals for my MLE estimates in Section 5.3. The likelihood-ratio test statistic is calculated as the difference between the log-likelihoods.

$$\lambda_{LR}(\theta_0) = -2[l(\theta_0) - l(\hat{\theta})]$$

where $\hat{\theta}$ is the point estimate that maximizes the likelihood. The set of θ_0 values for which $\lambda_{LR}(\theta_0)$ is smaller than the critical value cannot be rejected. In this context, I consider the 95% confidence interval, so the corresponding critical value is $\chi^2_2(0.05) = 5.991$.

To implement the likelihood-ratio test, I begin by coarsely gridding the θ space and computing the corresponding λ_{LR} values. I then identify the regions near the critical value, refine the grids within these regions, and determine the bounds. For the confidence interval of each parameter, I select conservative bounds: the lower bound is the smallest value in the non-rejected region, and the upper bound is the largest value in the non-rejected region.

C.3 More Model Fit Results

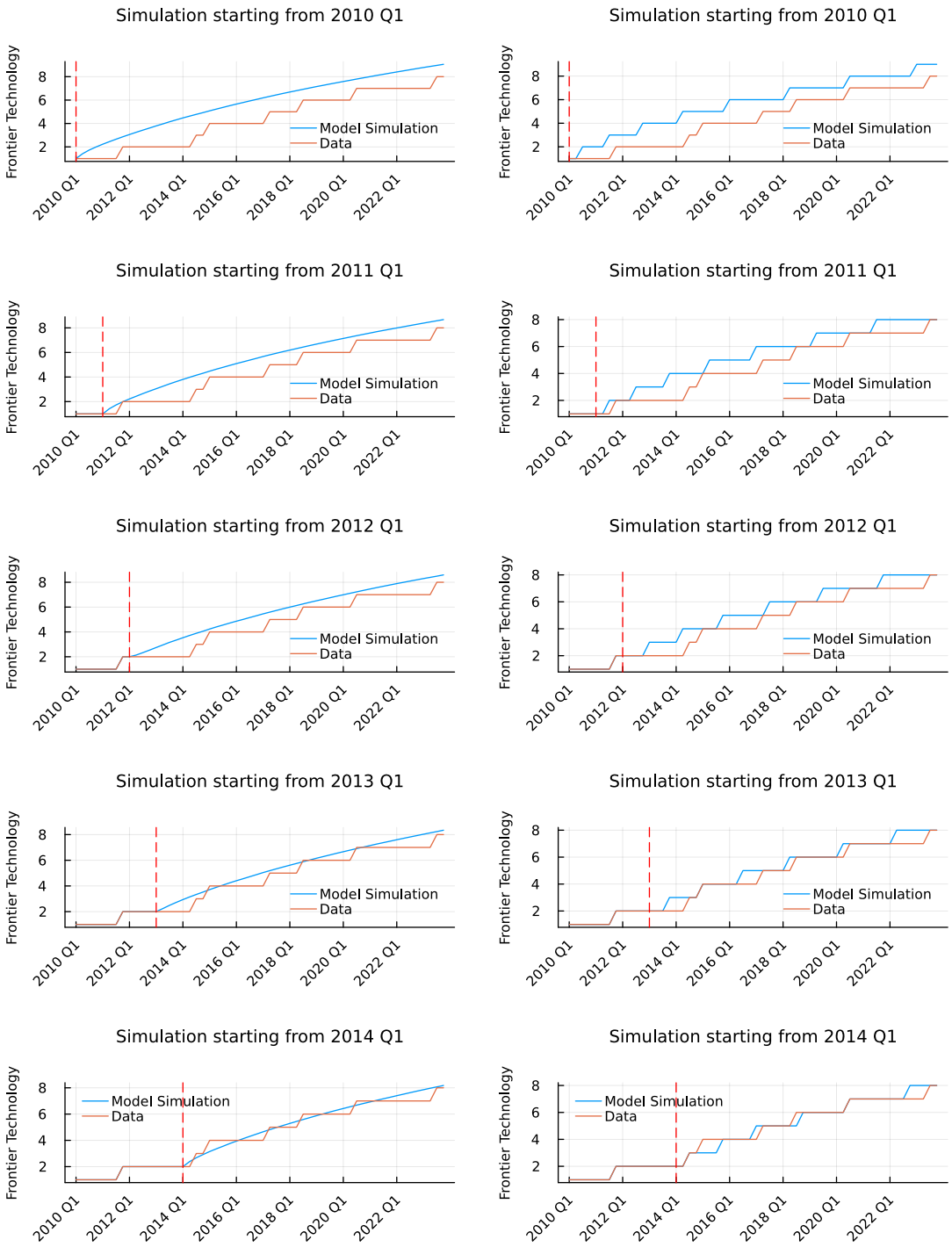


Figure C.11: Frontier Technology Trajectories: Model VS Data; Simulating from 2010-2014

Figures C.11 and C.12 compare the data and model-simulated frontier technology trajectories with different simulation starting periods.

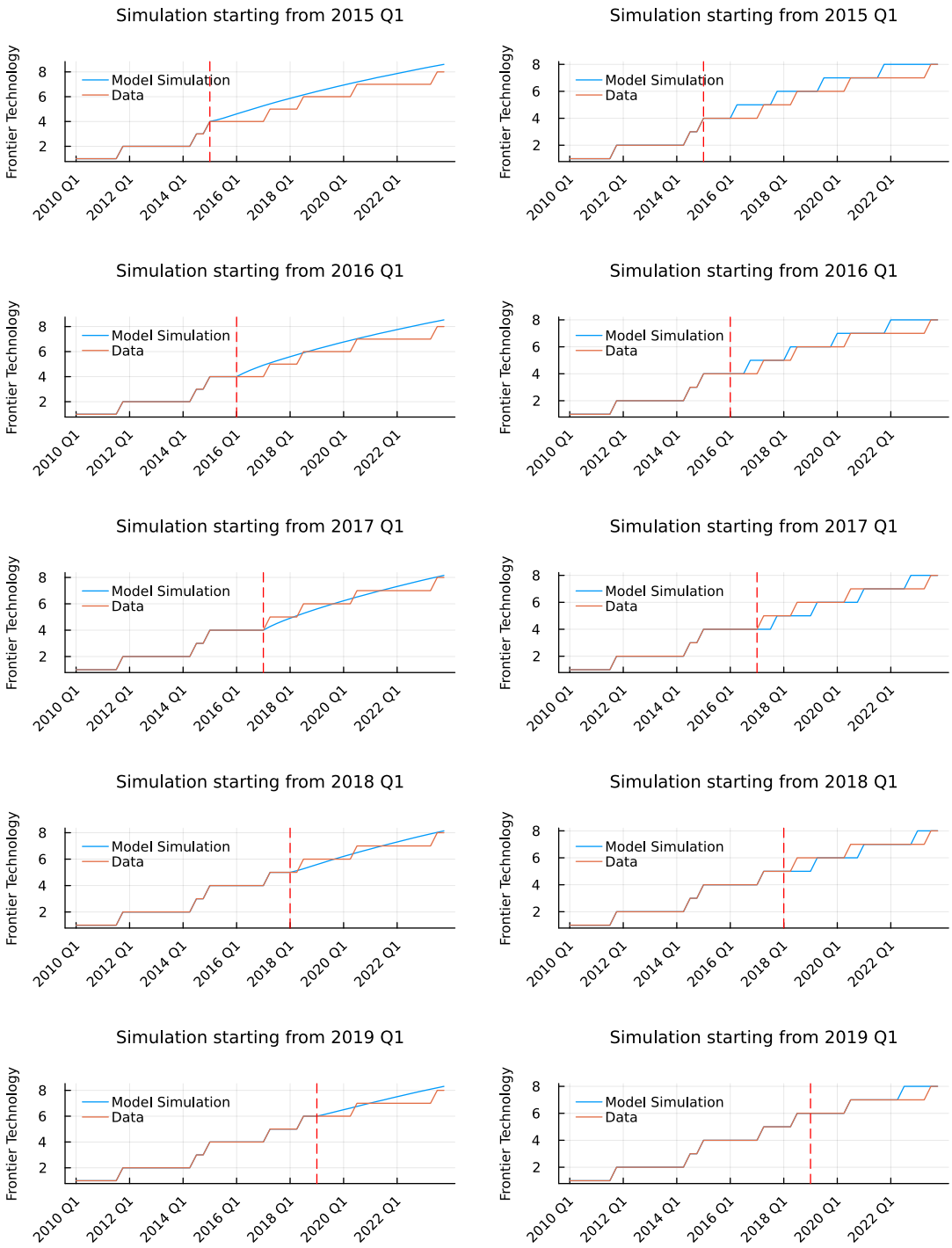


Figure C.12: Frontier Technology Trajectories: Model VS Data; Simulating from 2015-2019

D More Counterfactual Analysis and Details

D.1 Baseline Trade Disruption Risk

Figure D.13 shows the GPR since 2005, highlighting a spike in 2022, with risk levels remaining elevated compared to the pre-2022 period. From Caldara and Iacoviello (2022), GPR spikes increase the disaster probability by 17%. The average standardized global GPR since 2022 is 1.43, implying a 25% increase ($= 1.43 \times 0.1753$) in disaster probability. For the baseline scenario, I select a midpoint value of 20%.

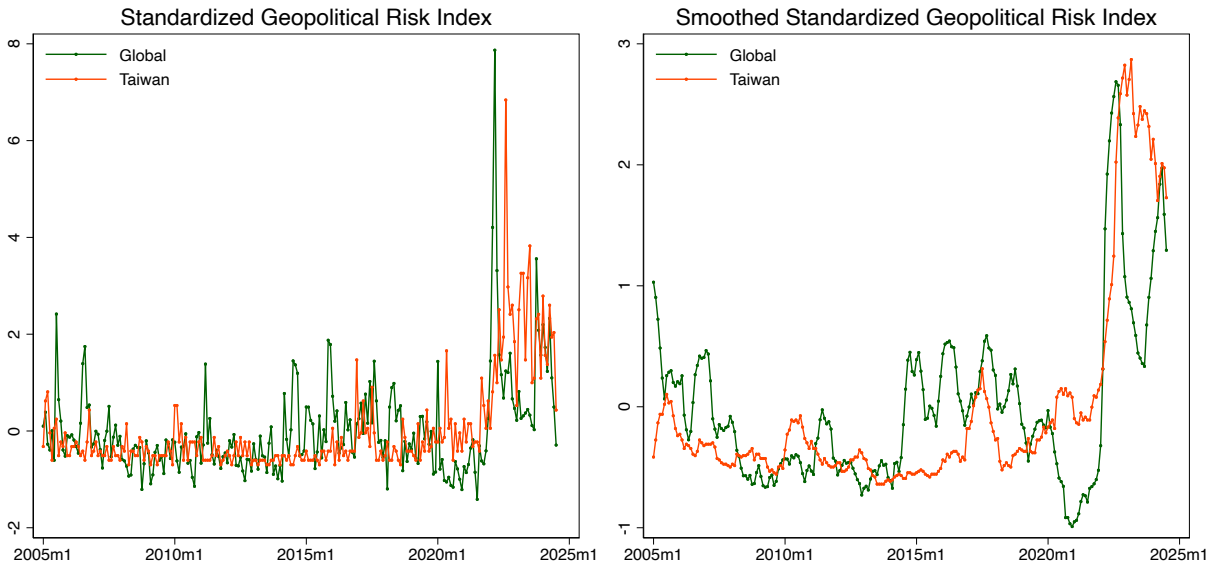


Figure D.13: Geopolitical Risk Trends

Notes: The data is sourced from Caldara and Iacoviello (2022) and has been standardized. The right-hand-side plot shows the smoothed data, with each point representing the average of the past 8 periods.

D.2 Additional Discussion on Static Welfare

D.2.1 Optimal Subsidy Rate in the U.S.

Foreign Cost Shifter With a higher δ, firms invest more in their home location (RoW) without subsidies, resulting in an initially low capacity allocation to the U.S. Consequently, the policy cost given a fixed subsidy rate is relatively low due to the low capacity in the U.S. However, a slight increase in U.S. capacity can substantially improve consumer surplus in the U.S. Therefore, the optimal subsidy rate for the U.S. is higher when δ is higher (see Figure D.14).

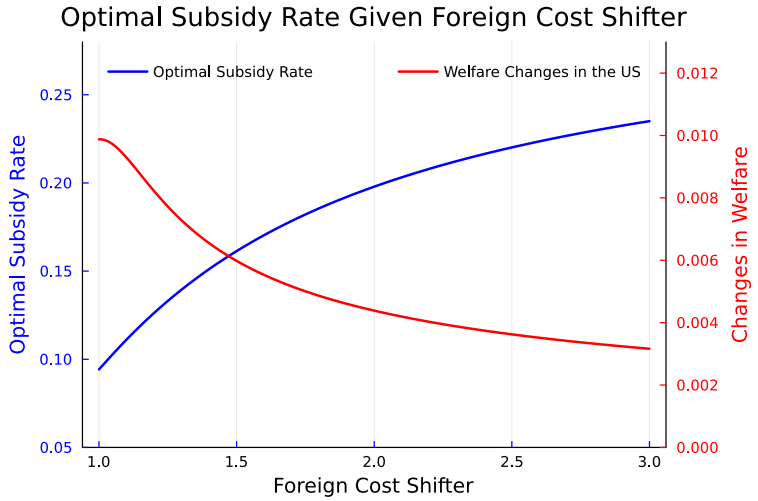


Figure D.14: Static Setting: Optimal Subsidy Rate Given Foreign Cost Shifter

Notes: The optimal rate, which maximizes U.S. consumer welfare defined as the net change in consumer surplus minus policy cost, is calculated in the baseline case with two incumbents and the trade disruption risk set to $\psi = 20\%$.

Subsidies in Other Locations When other locations also provide investment subsidies, the optimal U.S. subsidy rate rises with the subsidy rates of these locations. The intuition mirrors the earlier discussion on δ: high subsidies abroad result in low initial U.S. capacity, keeping policy costs low while offering substantial consumer surplus gains. However, as disruption risk increases, the link between the optimal U.S. subsidy rate and subsidies in other locations weakens.

D.2.2 Impacts of U.S. Subsidies in Other Locations

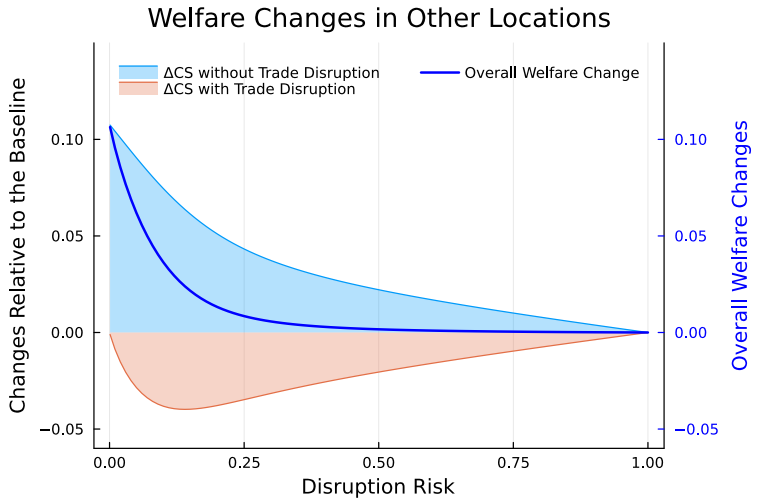


Figure D.15: Welfare Changes in Other Locations Given Trade Disruption Risk

Notes: These plots show the static impact of a 25% capacity investment subsidy in the U.S. on other locations under different trade disruption risks, considering the two incumbent case with the foreign cost shifter set to $\delta = 1.0$.

Trade Disruption Risk How does trade disruption risk affect welfare changes in other locations? Figure D.15 shows that as trade disruption risk rises, the benefits of U.S. subsidies for other locations diminish. In the extreme case of a 100% disruption risk, all markets become autarkic, and U.S. subsidies have no impact on welfare in other locations.

Foreign Cost Shifter When $\delta > 1$, the impact of U.S. subsidies on consumer surplus in CN and the RoW differs. Two main factors attract firms to invest: low costs and large market size. If δ is large, the primary advantage of the RoW is its low cost. U.S. subsidies would mitigate this advantage, leading to a significant transfer of capacity from the RoW to the U.S. While the cost advantage of CN is not as strong as the RoW, the decline in capacity in CN with U.S. subsidies is not as significant. Consequently, the benefit to the RoW is smaller than the benefit to CN in this scenario (see Figure D.19).

Figure D.16 illustrates welfare changes in other locations with a 25% capacity investment subsidy in the U.S. The impact of δ on welfare changes in China and the RoW is non-linear. For China, welfare initially increases as δ rises slightly above 1 because capacity allocation in CN is less affected. However, beyond a certain threshold, welfare change in China decreases with further increases in δ due to efficiency losses from reallocating capacity from the RoW to the U.S., resulting in a

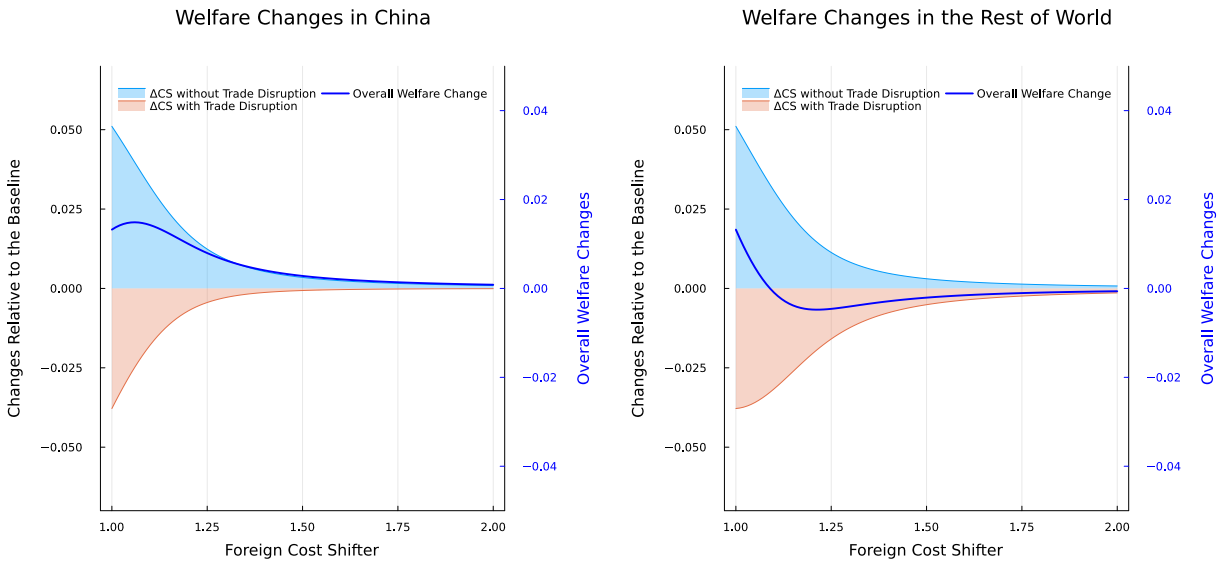


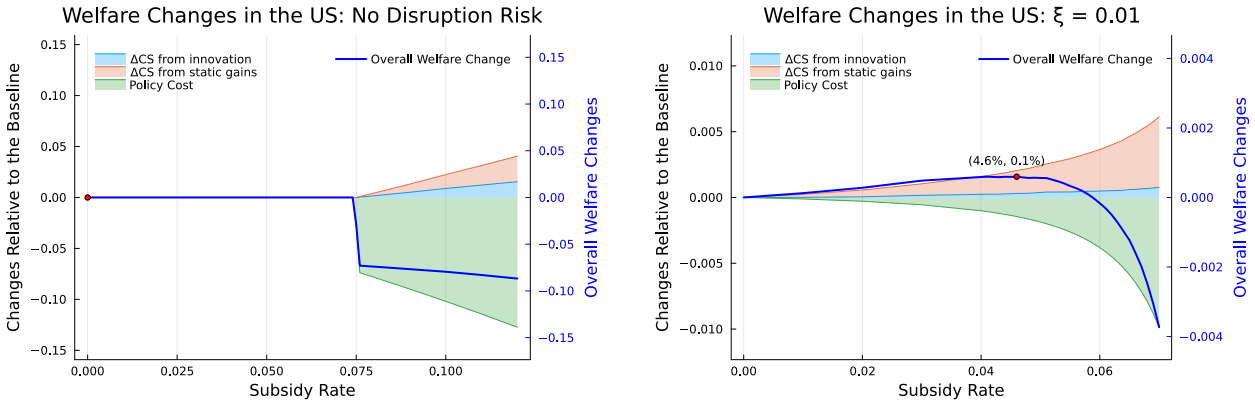
Figure D.16: Welfare Changes in Other Locations Given Foreign Cost Shifter

Notes: These plots show the static impact of a 25% capacity investment subsidy in the U.S. on other locations under different foreign cost shifters, considering the two incumbent case with the trade disruption risk set to $\psi = 20\%$.

lower global capacity gain. For the RoW, welfare initially decreases as δ increases slightly above 1, primarily due to capacity being reallocated from the RoW to the U.S., as previously discussed. As δ surpasses a threshold, welfare in the RoW begins to increase because the policy becomes less effective at diverting production from the RoW to the U.S.

D.3 Additional Discussion on Dynamic Welfare

Figure D.17a shows welfare changes in the U.S. when there is no trade disruption risk. Without this risk, firms lack incentives to diversify production locations and will concentrate in the most efficient location in the world without trade cost. Consequently, the subsidy rate must exceed a certain threshold for a location to become the most efficient and thus effective. If this threshold is high, the optimal policy is to do nothing, as the high subsidy rate would result in excessive policy costs relative to consumer surplus gains. Conversely, as shown in Figure D.17b, even with a very low trade disruption risk, firms have some incentive to diversify production, resulting in an optimal subsidy rate greater than zero. Although the model excludes trade costs for computational feasibility and given the low trade costs in semiconductors, any positive trade cost would intuitively function similarly to a certain level of trade disruption risk.



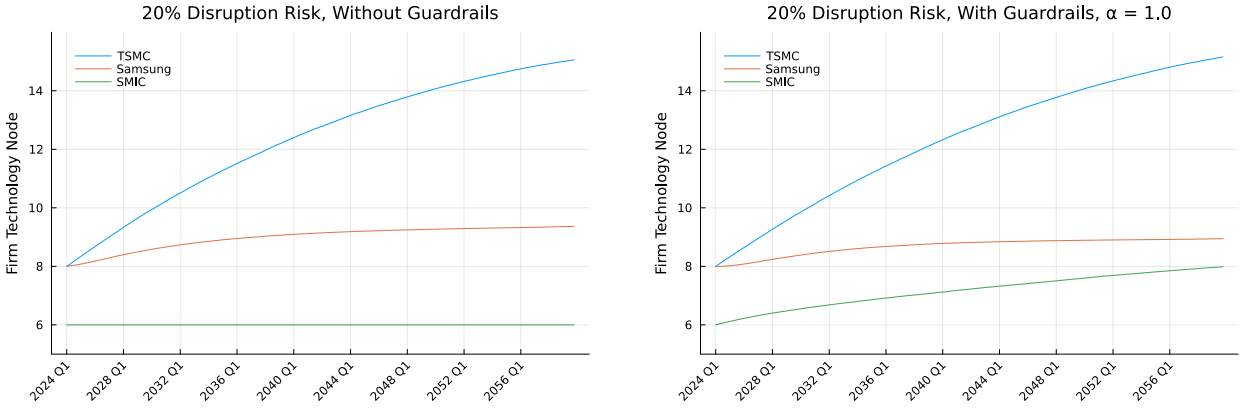
(a) No Disruption Risk

(b) Low Disruption Risk

Figure D.17: Dynamic Welfare Changes in the U.S. with Minimal Disruption Risk

Notes: These plots show the case with the foreign cost shifter set to $\delta = 1$. I simulated the model 10,000 times for each subsidy rate and computed the average consumer surplus and policy cost.

D.4 Additional Discussion on Investment Clawbacks



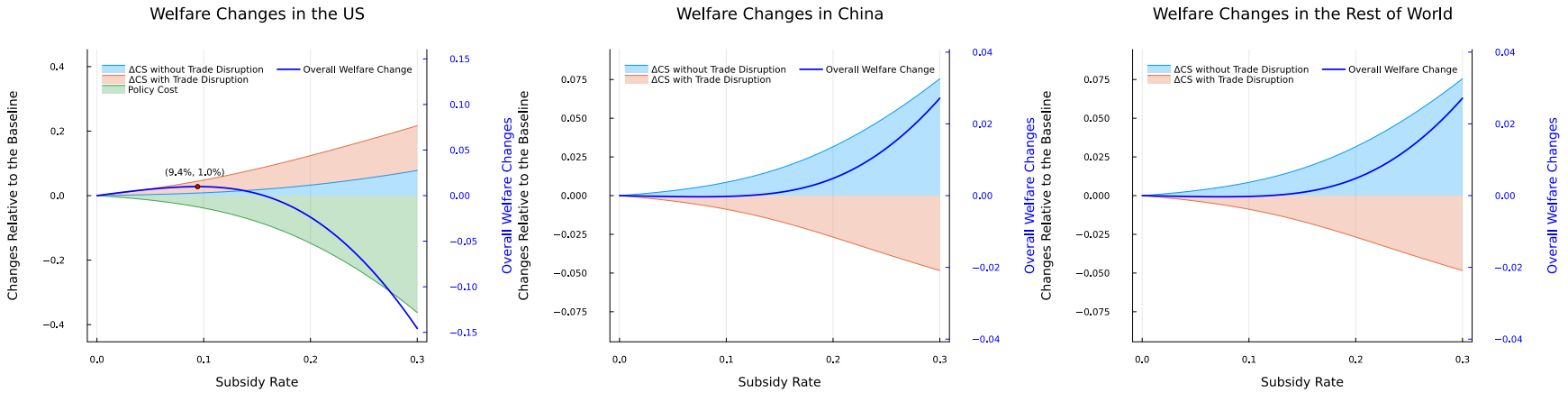
(a) Without Guardrails (b) With Guardrails and $\alpha = 1.0$

Figure D.18: Simulated Firm Technology Trajectories under 20% Disruption Risk

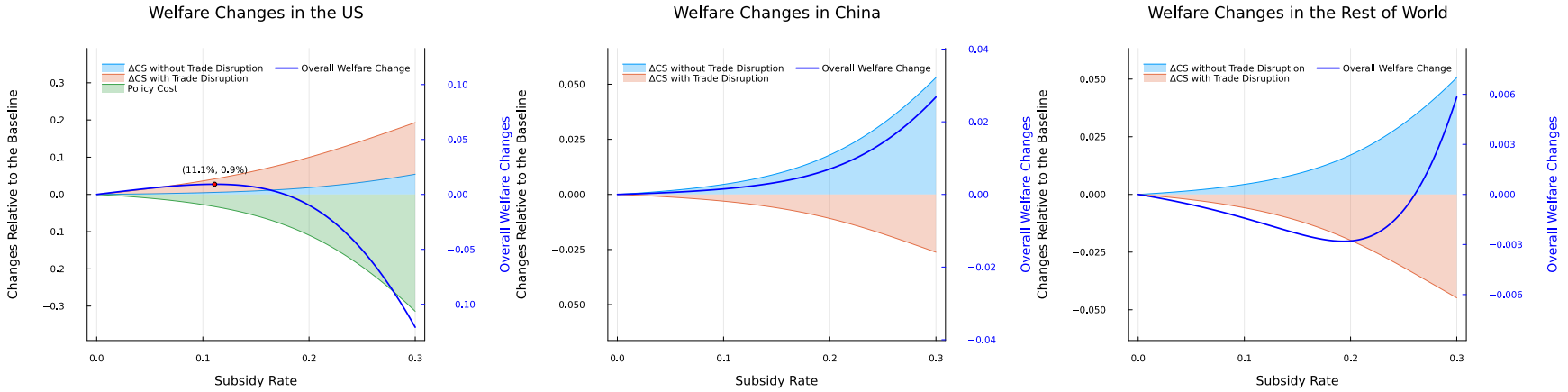
Notes: These plots show the simulated firm technology trajectories with and without guardrails at a trade disruption risk of $\psi = 20\%$, technology blocking coefficient $\alpha = 1.0$, and foreign cost shifter $\delta = 1$. I simulated the model 10,000 times for each scenario to compute the average technology trajectory for each firm.

Figure D.18 shows simulated firm technology trajectories with and without guardrails, assuming a 20% trade disruption risk and no technology blocking. This scenario represents the best-case trajectory for the Chinese firm's technology growth. However, even without technology blocking, the Chinese firm's growth is limited because its secured market is not large enough. With a high probability that trade disruptions do not occur, the leading firms remain competitive in the Chinese market, as they can access it through trade even without direct investment in mainland China.

For future work, I can explore how China might respond to U.S. policies. Intuitively, the guardrail restrictions in mainland China would raise the optimal subsidy rate there for two reasons: (1) from a supply resilience perspective, having SMIC possess the necessary technology becomes more valuable during trade disruptions, as non-Chinese firms can no longer serve the Chinese market; (2) from a policy cost perspective, the guardrail restrictions ensure that the benefits of the subsidies accrue entirely to Chinese firms. These strategic interactions would further shape the technological dynamics among firms.



(a) Baseline case: $\delta = 1$



(b) With foreign cost shifter case: $\delta = 1.1$

Figure D.19: Static Welfare Changes Under Different Foreign Cost Shifter δ

Notes: These plots show the two incumbent case with the trade disruption risk set to $\psi = 20\%$.