Compute and Antitrust: Regulatory implications of the AI hardware supply chain, from chip design to foundation model APIs

Haydn Belfield [1] **Shin-Shin Hua** [1]

Abstract

We argue that the antitrust and regulatory literature to date has failed to pay sufficient attention to compute, despite compute being a key input to AI progress and services (especially with the advent of powerful new generative AI systems), the potentially substantial market power of companies in the supply chain, and the advantages of compute as a 'unit' of regulation in terms of detection and remedies. We explore potential topics of interest to competition law under merger control, abuse of dominance, state aid, and anti-competitive agreements (cartels and collusion). Major companies and states increasingly view the development of AI over the coming decades as core to their interests, due to its profound impact on economies, societies, and balance of power. If the rapid pace of AI progress is sustained over the long-term, these impacts could be transformative in scale. This potential market power and policy importance, particularly in the generative AI field, should make compute an area of significant interest to antitrust and other regulators.

1. Introduction

Tensions in the Taiwan Strait, queues of container ships outside ports, consumer electronics shortages and inflation, immersive video games and ChatGPT. These seemingly disparate features of our contemporary world are all part of, or influenced by, the AI hardware supply chain.

Scales and precisions across this supply chain can take on a science fiction quality. The mirrors used in advanced photolithography must be so perfectly flat that if the mirror were scaled to the size of Germany, the biggest flaw on the mirror would be less than one-tenth of a millimetre high [1]. The lasers are as precise as shining a laser pointer from the Moon and hitting a thumb [3]. AlphaGo Zero played 4.9 million games of Go against itself [19]. The biggest public AI model, Wu Dao 2.0, had 1.75 trillion parameters, similar to the number of synapses in a mouse brain [10, 17]. Some models will soon have 100 trillion parameters [21].

Compute or computing power refers to a "specialised stack of software and hardware (inclusive of processors, memory and networking) engineered to support AI-specific workloads or applications" [7], [8]. Rather than personal levels of compute at 'the edge' in smartphones or laptops, we focus on the industrial scale of cloud data centres and supercomputers. AI relies on large amounts of specialised compute - large 'computing clusters' with particular types of computer chips: graphics processing units (GPUs) for training, field-programmable gate arrays (FPGAs) for inference, and application-specific integrated circuits (ASICs) for both. These chips increasingly have specific design features useful for AI applications [12, 11].

2. Compute is important for antitrust

The intersection of compute and competition law is an important, tractable yet relatively under-analysed topic. We argue the recent antitrust and regulatory literature to date has failed to pay sufficient attention to compute in a holistic manner, across the entire supply chain. This is despite compute being a key input to AI progress and services, especially the 'eye-watering' compute costs required to train and deploy generative AI and LLMs. GPT-4, for example, may have been trained with 10^25 FLOP and cost $100m just in compute costs. In addition, the literature does not yet fully take account of the likely market power of firms across the compute supply chain, and the regulatory advantages of compute in terms of detection and remedies. We argue that this gap is significant and may mean regulators and academia fail to identify and address risks such as how the market structures and behaviours in the compute supply chain could lead to AI development. This failure could further concentrate market power held by Big Tech.

[1]University of Cambridge, Leverhulme Centre for the Future of Intelligence. Correspondence to: Haydn Belfield <hb492@cam.ac.uk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

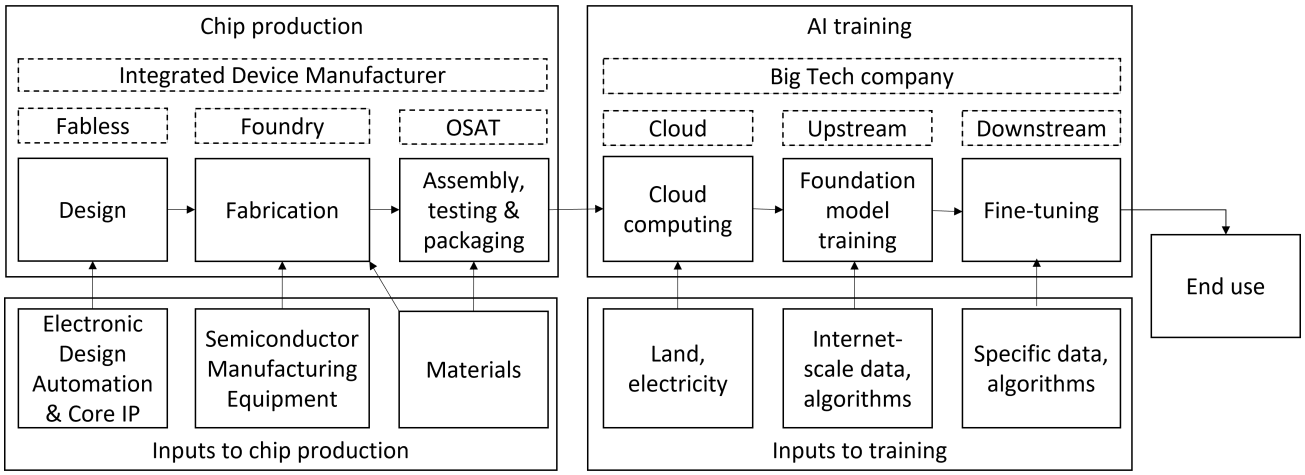

Figure 2: The semiconductor supply chain.

Solid: chain segment; dashed: business model.

This new era in computing is distinct from previous hardware cases such as the IBM mainframe cases of the 1970s, and the Microsoft personal computing cases of the 2000s [4, 6]. Much of the antitrust focus on Big Tech in recent years has addressed adjacent areas such as *data issues* or abusive pricing practices by *online platforms* or advertising markets, rather than compute [2, 20]. Recent regulatory attention on cloud computing is welcome, but also fails to take full account of the importance of compute as an input to AI progress and services. We argue that a systematic and holistic view of the compute supply chain taking into account the links between e.g. cloud, chips and AI systems such as generative AI and LLMs, is relatively underexplored from current academic literature and policy thinking.

Failing to fully consider 'compute as an AI input' is potentially a significant gap in regulatory enforcement and the related academic literature.

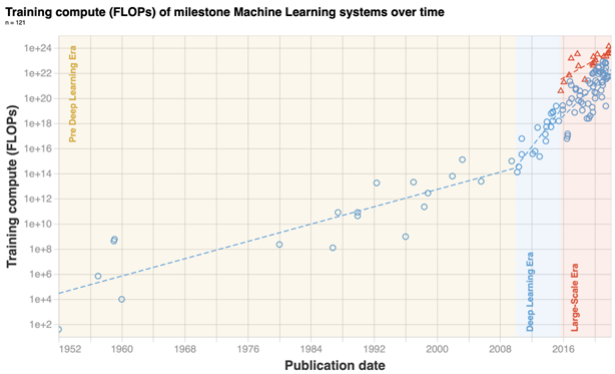

Figure 1: Trends in n = 121 milestone ML models between 1952 and 2022. [18]

First, compute is a major input to and driver of AI progress, and its importance as an input continues to grow. Compute is one of the key bottlenecks for AI development alongside data and talent. The recent 'AI summer' period of dramatic progress in ML is commonly said to have started a decade ago with 2012's 'AlexNet' [14]. From 2010 to 2021, the amount of training compute used in the largest AI training runs has been doubling every 6 months [18]. Recent frontier models used 55 million times more compute than AlexNet (see Figure 1). Recent breakthroughs in AI/ML have been exceedingly costly, above $10-100 million per training run. 'Scaling laws' indicate that this trend is likely to continue.

Second, the compute supply chain is typified by high barriers to entry and remarkable concentration and likely market power (See Figure 2). Only one company produces highly advanced photolithography machines for fabs: ASML. There are only three providers of GPUs: Intel, AMD and most importantly NVIDIA. Only three companies are able to manufacture advanced chips: Intel, Samsung and TSMC – and TSMC (with with 54% of the world's overall logic foundry market) is uniquely capable of producing the most advanced chips. Setting up an advanced chip fab costs around $10 billion and takes several years. TSMC spent $44bn on capital expenditure in 2022 (more than e.g. Exxon), and more than $100bn over three years. The Biden export controls have raised barriers to entry yet hgigher. There are only three main cloud providers in the West (Amazon Web Services 32%, Microsoft Azure 21% and Google Cloud 8%) and three in China (Alibaba, Huawei and Tencent) [9]. Setting up a new cloud 'region' (collection of data centres that works together) also costs $10 billion. Training a foundation model is largely out of reach for most academic groups and smaller companies. This trend is pronounced in generative AI, where smaller AI companies such as OpenAI and Anthropic have entered into 'compute partnerships' with Microsoft and Google respectively. This dependency could extend Big Tech's dominance to new markets [15, 5]. These foundation models underpin thousands of down-

stream applications, either fine-tuned or built on a structured access API. Foundation model developers may play a gatekeeper role, controlling access on an ongoing basis.

Third, regulatory enforcement in compute offers several advantages to a regulator. 'Effective enforceability' that is a regulator's ability to detect breaches and to remedy and sanction may be higher for compute relative to other key inputs. Compared to other sources of market power such as 'talent' or data, compute is generally more legible and quantifiable, and is more amenable to structural remedies. The amount of talent or data, for example, is difficult to measure or compare between companies. Assessing the market power of data depends on further factors, such as a dataset's uniqueness, quality, permitted/consented uses and how recent it is. 'How much' data and of what kind could constitute the threshold for dominance in a market? Rather than being a relatively simple quantitative exercise, it is rather complex, context-dependent, and possibly imprecise. Remedies are also harder to design and implement for talent or data – for example divesting or transferring talent is vulnerable to employee 'flight risk'. By comparison, compute may help with the ability to detect breaches, because it is more easily measured and quantified. Compute is physical, discrete - instantiated in particular equipment and chips. Indeed, it is usually large and bulky - located in large fabs or data centres. The relationship between compute and performance is better understood, so a link to market power may be easier to demonstrate. State aid for compute would look more like tax breaks or subsidies for infrastructure, a situation much more familiar to competition regulators. Compute also helps with the ability to remedy and sanction breaches, because it may be easier to transfer (e.g. a structural remedy that splits a cloud computing provider into two companies with similar, quality-adjusted numbers of data centres) or give access to (e.g. mandating competitor access to manufacturing equipment or chips).

Overall, we recommend regulators take into account the importance of compute as an AI input (1) in prioritising whether to take enforcement action in this space and (2) in their existing regulatory enforcement.

3. What is competition law and why is it an important regulatory lever?

In the preceding sections, we demonstrated the high levels of concentration and high barriers to entry along the compute supply chain. This means both that there are potential regulatory concerns in the compute supply chain that may be of particular interest to competition regulators as a key input to AI, and that these concerns may be more effectively addressed through taking enforcement action in the compute market compared to other key AI inputs.

But what is the role of competition law and why is it the right regulatory lever? Competition law seeks to protect the process of competition on a market. When firms compete for customers, they are incentivised to produce the best quality products at the minimum price, which benefits consumers and society at large. Therefore, a core objective of competition law is to prohibit firms from engaging in conduct which may distort the competitive process and harm competition. The four main categories of antitrust are preventing firms with a powerful position on a market from abusing their market power or dominant position, preventing firms from engaging in anti-competitive agreements (cartels and collusion), preventing governments from unfairly privileging certain firms (state aid), and preventing anti-competitive consequences of concentrations (i.e. mergers and acquisitions).

Compute could be a useful regulatory lens across each of these, as can be illustrated by four case-studies: the Federal Trade Commission (FTC)'s investigation of Amazon Web Services (AWS); potential agreements on hardware security; state aid aspects of the European Chips Act; and the NVIDIA/Arm acquisition - which we now highlight.

3.1. Case study: NVIDIA/Arm & Merger Control

Competition regulators might scrutinise compute through mergers and acquisitions that need competition approval (merger control). The central concern in merger control is whether the combination of companies reduces competition in the market by creating or strengthening a dominant player.

The decisions of competition regulators as to whether to approve, approve conditional on remedies, or block transactions has several implications for the future competitive landscape for AI. Where a competition regulator identifies competition concerns with the merger it can make the clearance conditional on 'remedies' from the parties, including access remedies, granting competitors use of a key input, or structural remedies: separating parts of the business that hold a key input in order for that part(s) to act as an independent market participant competing with the incumbent [68]. It may also include behavioural remedies, though these are not generally preferred due to ongoing monitoring costs. Ultimately, a competition regulator could also block a merger from going ahead.

One useful case study is the now-abandoned Nvidia/Arm transaction. Whilst every merger assessment will be fact-specific, there are themes in the regulators' assessments that could be applied more broadly to other future merger control reviews. We focus on the merger assessments by the EU and UK regulators.

A merger of companies on the compute supply chain is likely to have a number of significant implications. First,

the compute supply chain mostly consists of global markets, which is likely to trigger merger review in multiple jurisdictions. This means that if just one regulator opposes the deal it would likely scupper it [16]. Say, for example, that Japan blocks a deal. The merging parties could try to carve out part of the transaction so that the effects on trade in Japan are unaffected. However, it is extremely difficult (or commercially unattractive) in practice to do this because their business functions are likely integrated to serve the world or regions of the world, rather than being easily carved up into a Japan-focused division.

Another pragmatic consideration is timing. Mergers are often under intense time pressure and have drop-dead dates such that if delayed past a certain date, the parties will have the right to abandon the deal. If a merger is blocked and the parties need to carve out part of the transaction it is likely that this will delay the closing beyond the drop-dead date and therefore the parties will choose to abandon the deal. Indeed, the intensiveness even of a merger going from a Phase I investigation to an in-depth, Phase II investigation can have a similar effect because of its delaying of the transaction completion.

Even if a competition regulator does not block a deal, it may raise competition concerns and seek remedies from the parties to address the concerns. In relation to compute, these remedies may include behavioural remedies such as those offered by Nvidia/Arm. Nvidia offered commitments to maintain Arm as a neutral technology supplier to address the key concern that Arm has substantial market power for the licensing of CPU IP for use in processor products and, therefore, that the combined Nvidia/Arm could restrict or degrade the access to Arm's technology of companies who compete with Nvidia in providing processor products [7, 8]. The EC was also concerned that the deal could stifle innovation by making licensees of Arm IP reluctant to continue sharing commercially sensitive information with the combined entity because they are also Nvidia's competitors or refocusing Arm's RD spending on products that are most profitable for Nvidia downstream, to the detriment of players reliant on Arm's IP in other product areas. The EC rejected Nvidia's commitments as inadequate to address its concerns and opened an in-depth investigation. The deal also faced opposition from the US FTC and the UK CMA. Nvidia eventually decided to abandon the merger after seeing "little to no progress in winning [regulatory] approval" [13].

4. Conclusion

The pandemic has made clear to all of us the complex, fragile, worldwide supply chains on which our economies and societies rely. One of the most important and interesting of these is the compute supply chain. It stretches from chip designers in Cambridge (UK); photolithography manufacturers in Veldhoven (Netherlands); fabs in Hsinchu (Taiwan); data centres in Ashburn (Northern Virginia); and AI developers in San Francisco (California) to end users and affected communities everywhere in the world. The importance of compute as a key input to powerful generative AI models underscores that the question of who controls the supply of compute, and how they do so, is a critical question that has wide-ranging social and economic implications.

In light of radical developments in generative AI, regulators and those interested in competition law are highly likely to pay increasing attention to the compute supply chain. Those engaged in AI governance in academia, civil society, and industry should be aware of this coming interest and advocate for particular regulatory focus on the significance of compute as an AI input. We hope this paper has provided some ideas for further development and research. Antitrust is a powerful tool to shape markets, corporate behaviour and outcomes for consumers. The intersection between compute, AI and competition law is a major topic for our times.

References

[1] ASML. *Lenses & mirrors - Lithography principles*. en. URL: https://www.asml.com/en/technology/lithography-principles/lenses-and-mirrors (visited on 03/01/2022).

[2] Miriam Caroline Buiten. "Exploitative abuses in digital markets: between competition law and data protection law". In: *Journal of Antitrust Enforcement* 9.2 (July 2021), pp. 270–288. ISSN: 2050-0696. DOI: 10.1093/jaenfo/jnaa041. URL: https://doi.org/10.1093/jaenfo/jnaa041 (visited on 02/14/2022).

[3] Max A. Cherney. *We toured Intel's advanced chip making site in Oregon. Here's how it works.* en. Section: Enterprise. Dec. 2021. URL: https://www.protocol.com/enterprise/intel-euv-fab-chips (visited on 03/01/2022).

[4] William F Colby Jr. "The Multinational's Dilemma: The IBM Proceeding in Europe". en. In: *Penn State International Law Review* 3.2 (1985), p. 27. URL: https://elibrary.law.psu.edu/cgi/viewcontent.cgi?article=1026&context=psilr.

[5] Diane Coyle. "Preempting a Generative AI Monopoly — by Diane Coyle". en. In: *Project Syndicate* (Feb. 2023). URL: https://www.project-syndicate.org/commentary/preventing-tech-giants-from-monopolizing-artificial-intelligence-chatbots-by-diane-coyle-2023-02 (visited on 02/24/2023).

[6] ECJ Court of First Instance (Grand Chamber). *Judgment of the Court of First Instance (Grand Chamber) of 17 September 2007. Microsoft Corp v Commission of the European Communities*. en. Sept. 2007. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62004TJ0201 (visited on 02/28/2022).

[7] European Commission. *Mergers: Commission opens in-depth investigation into proposed acquisition of Arm by NVIDIA*. en. Text. Brussels: European Commission, Oct. 2021. URL: https://ec.europa.eu/commission/presscorner/detail/en/ip_21_5624 (visited on 02/28/2022).

[8] Federal Trade Commission. *Complaint In the Matter of Nvidia, Softbank and Arm*. Dec. 2021. URL: https://www.ftc.gov/system/files/documents/cases/d09404_part_3_complaint_public_version.pdf.

[9] Gartner. *Gartner Says Worldwide IaaS Public Cloud Services Market Grew 40.7% in 2020: Amazon, Microsoft and Alibaba Led the 2020 Race to the Cloud*. en. Tech. rep. Stamford Conneticut: Gartner, June 2021. URL: https://www.gartner.com/en/newsroom/press-releases/2021-06-28-gartner-says-worldwide-iaas-public-cloud-services-market-grew-40-7-percent-in-2020 (visited on 02/28/2022).

[10] Suzana Herculano-Houzel, Bruno Mota, and Roberto Lent. "Cellular scaling rules for rodent brains". In: *Proceedings of the National Academy of Sciences of the United States of America* 103.32 (Aug. 2006), pp. 12138–12143. ISSN: 0027-8424. DOI: 10.1073/pnas.0604911103. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1567708/ (visited on 02/26/2022).

[11] Saif Khan and Alexander Mann. *AI Chips: What They Are and Why They Matter*. en. Tech. rep. Center for Security and Emerging Technology, Apr. 2020. DOI: 10.51593/20190014. URL: https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter/ (visited on 02/28/2022).

[12] Saif M. Khan, Alexander Mann, and Dahlia Peterson. *The Semiconductor Supply Chain: Assessing National Competitiveness*. en. Tech. rep. Center for Security and Emerging Technology, Jan. 2021. DOI: 10.51593/20190016. URL: https://cset.georgetown.edu/publication/the-semiconductor-supply-chain/ (visited on 02/28/2022).

[13] Ian King, Giles Turner, and Peter Elstrom. "Nvidia Quietly Prepares to Abandon $40 Billion Arm Bid". In: *Bloomberg* (Jan. 2022). URL: https://www.bloomberg.com/news/articles/2022-01-25/nvidia-is-said-to-quietly-prepare-to-abandon-takeover-of-arm (visited on 02/28/2022).

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc., 2012. URL: https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (visited on 02/28/2022).

[15] Madhumita Murga. "Big Tech companies use cloud computing arms to pursue alliances with AI groups". In: *Financial Times* (Feb. 2023). URL: https://www-ft-com.ezp.lib.cam.ac.uk/content/5b17d011-8e0b-4ba1-bdca-4fbfdba10363.

[16] Lauren O'Brien. *The Long Arm of Competition law? NVIDIA/Arm : can the deal overcome innovation concerns?* en. Oct. 2021. URL: https://techinsights.linklaters.com/post/102h9il/the-long-arm-of-competition-law-nvidia-arm-can-the-deal-overcome-innovation-c (visited on 02/28/2022).

[17] Alberto Romero. *GPT-3 Scared You? Meet Wu Dao 2.0: A Monster of 1.75 Trillion Parameters*. en. June 2021. URL: https://towardsdatascience.com/gpt-3-scared-you-meet-wu-dao-2-0-a-monster-of-1-75-trillion-parameters-832cd83db484 (visited on 03/01/2022).

[18] Jaime Sevilla et al. "Compute Trends Across Three Eras of Machine Learning". In: *arXiv:2202.05924 [cs]* (Feb. 2022). arXiv: 2202.05924. URL: http://arxiv.org/abs/2202.05924 (visited on 02/15/2022).

[19] David Silver et al. "Mastering the game of Go without human knowledge". en. In: *Nature* 550.7676 (Oct. 2017). Number: 7676 Publisher: Nature Publishing Group, pp. 354–359. ISSN: 1476-4687. DOI: 10.1038/nature24270. URL: http://www.nature.com/articles/nature24270 (visited on 02/28/2022).

[20] Toshiaki Takigawa. "Super platforms, big data, and competition law: the Japanese approach in contrast with the USA and EU". In: *Journal of Antitrust Enforcement* 9.2 (July 2021), pp. 289–312. ISSN: 2050-0696. DOI: 10.1093/jaenfo/jnaa053. URL: https://doi.org/10.1093/jaenfo/jnaa053 (visited on 02/14/2022).

[21] Ce Zhang and Cristian Mezzanotte. *Training a recommender model of 100 trillions parameters on Google Cloud*. en. May 2022. URL: https://cloud.google.com/blog/products/ai-machine-learning/training-a-recommender-model-of-100-trillions-parameters-on-google-cloud (visited on 02/24/2023).